

TALN 2015

Table des matières

Préface

Multilinguality at Your Fingertips : BabelNet, Babelfy and Beyond !.....	i-i
Pourquoi construire des ressources terminologiques et pourquoi le faire différemment ?.....	ii-ii

Session Extraction d'information

Apprentissage par imitation pour l'étiquetage de séquences : vers une formalisation des méthodes d'étiquetage « easy-first ».....	1-12
Stratégies de sélection des exemples pour l'apprentissage actif avec des champs aléatoires conditionnels.....	13-24
Identification de facteurs de risque pour des patients diabétiques à partir de comptes-rendus cliniques par des approches hybrides.....	25-36
Oublier ce qu'on sait, pour mieux apprendre ce qu'on ne sait pas : une étude sur les contraintes de type dans les modèles CRF.....	37-48

Session Compréhension et paraphrase

Analyse d'expressions temporelles dans les dossiers électroniques patients.....	49-58
Compréhension automatique de la parole sans données de référence.....	59-70

Session Désambiguïsation

Désambiguïsation d'entités pour l'induction non supervisée de schémas événementiels.....	71-82
Création rapide et efficace d'un système de désambiguïsation lexicale pour une langue peu dotée.....	83-94

Session Opinions et sentiments

Méthode faiblement supervisée pour l'extraction d'opinion ciblée dans un domaine spécifique.....	95-106
Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité.....	107-118

Session Sémantique

Estimation de l'homogénéité sémantique pour les Questionnaires à Choix Multiples.....	119-130
Extraction automatique de relations sémantiques dans les définitions : approche hybride, construction d'un corpus de relations sémantiques pour le français.....	131-145
Déclasser les voisins non sémantiques pour améliorer les thésaurus distributionnels.....	146-157

Session Syntaxe et paraphrase

Grammaires phrastiques et discursives fondées sur les TAG : une approche de D-STAG avec les ACG.....	158-169
Noyaux de réécriture de phrases munis de types lexico-sémantiques.....	170-181
Extraction automatique de paraphrases grand public pour les termes médicaux.....	182-193
Analyse syntaxique de l'ancien français : quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage ?.....	194-207

Session Classification et Alignement

Attribution d'Auteur : approche multilingue fondée sur les répétitions maximales.....	208-219
Mesurer la similarité entre phrases grâce à Wikipédia en utilisant une indexation aléatoire.....	220-231
Typologie automatique des langues à partir de treebanks.....	232-243

Session Traduction

Utilisation de mesures de confiance pour améliorer le décodage en traduction de parole.....	244-254
Multialignement vs bialignement : à plusieurs, c'est mieux !.....	255-266
Apprentissage discriminant des modèles continus de traduction.....	267-278

Session Plénière

Utiliser les interjections pour détecter les émotions.....	279-292
Comparaison d'architectures neuronales pour l'analyse syntaxique en constituants.....	293-304
...des conférences enfin disons des causeries... Détection automatique de segments en relation de paraphrase dans les reformulations de corpus oraux.....	305-316

Posters 1

Une méthode discriminant formation simple pour la traduction automatique avec Grands Caractéristiques.....	317-322
Natural Language Reasoning using Coq: Interaction and Automation.....	323-329
Vous aimez ?...ou pas ? LikeIt, un jeu pour construire une ressource lexicale de polarité.....	330-336
Étude des verbes introducteurs de noms de médicaments dans les forums de santé.....	337-343
Initialisation de Réseaux de Neurones à l'aide d'un Espace Thématique.....	344-349
FDTB1: Repérage des connecteurs de discours en corpus.....	350-356
ROBO, an edit distance for sentence comparison Application to automatic summarization.....	357-363
Classification d'entités nommées de type « film ».....	364-370
A critical survey on measuring success in rank-based keyword assignment to documents.....	371-376
Effects of Graph Generation for Unsupervised Non-Contextual Single Document Keyword Extraction.....	377-383
Adaptation par enrichissement terminologique en traduction automatique statistique fondée sur la génération et le filtrage de bi-segments virtuels.....	384-390
Une mesure d'intérêt à base de surreprésentation pour l'extraction des motifs syntaxiques stylistiques.....	391-396
Une Approche évolutionnaire pour le résumé automatique.....	397-403
Identification des unités de mesure dans les textes scientifiques.....	404-410
Évaluation intrinsèque et extrinsèque du nettoyage de pages Web.....	411-417
CANÉPHORE : un corpus français pour la fouille d'opinion ciblée.....	418-424
Extraction de Contextes Riches en Connaissances en corpus spécialisés.....	425-431
Traitement automatique des formes métriques des textes versifiés.....	432-438
Apprentissage automatique d'un modèle de résolution de la coréférence à partir de données orales transcrites du français : le système CROC.....	439-445
Vers un diagnostic d'ambiguïté des termes candidats d'un texte.....	446-452
Augmentation d'index par propagation sur un réseau lexical Application aux comptes rendus de radiologie.....	453-459
Détection automatique de l'ironie dans les tweets en français.....	460-465
Dictionnaires morphologiques du français contemporain : présentation de Morfetik, éléments d'un modèle pour le TAL.....	466-472

Posters 2

Une métagrammaire de l'interface morpho-sémantique dans les verbes en arabe.....	473-479
Création d'un nouveau treebank à partir de quatrièmes de couverture.....	480-486
Entre écrit et oral ? Analyse comparée de conversations de type tchat et de conversations téléphoniques dans un centre de contact client.....	487-493
Construction et maintenance d'une ressource lexicale basées sur l'usage.....	494-500
Utilisation d'annotations sémantiques pour la validation automatique d'hypothèses dans des conversations téléphoniques.....	501-507
Etiquetage morpho-syntaxique en domaine de spécialité: le domaine médical.....	508-514
Vers une typologie de liens entre contenus journalistiques.....	515-521
CDGFr, un corpus en dépendances non-projectives pour le français.....	522-528
Utilisation des réseaux de neurones récurrents pour la projection interlingue d'étiquettes morpho-syntaxiques à partir d'un corpus parallèle.....	529-536
Segmentation et Titrage Automatique de Journaux Télévisés.....	537-543
Un système hybride pour l'analyse de sentiments associés aux aspects.....	544-550
La ressource EXPLICADIS, un corpus annoté spécifiquement pour l'étude des relations de discours causales.....	551-557
La séparation des composantes lexicale et flexionnelle des vecteurs de mots.....	558-564
Traitements pour l'analyse du français préclassique.....	565-571
Classification de texte enrichie à l'aide de motifs séquentiels.....	572-578
Le traitement des collocations en génération de texte multilingue.....	579-585
Médicaments qui soignent, médicaments qui rendent malades : étude des relations causales pour identifier les effets secondaires.....	586-592
Exploration de modèles distributionnels au moyen de graphes 1-PPV.....	593-599
Apport de l'information temporelle des contextes pour la représentation vectorielle continue des mots.....	600-606
Etiquetage morpho-syntaxique de tweets avec des CRF.....	607-613
Caractériser les discours académiques et de vulgarisation : quelles propriétés ?.....	614-620
Extraction et analyse automatique des comparaisons et des pseudo-comparaisons pour la détection des comparaisons figuratives.....	621-627
Proposition méthodologique pour la détection automatique de Community Manager. Étude multilingue sur un corpus relatif à la Junk Food.....	628-634

Démonstrations

MEDITE : logiciel d'alignement de textes pour l'étude de la génétique textuelle.....	635-636
Phœbus : un Logiciel d'Extraction de Réutilisations dans des Textes Littéraires.....	637-639
YADTK : Une plateforme open-source à base de règles pour développer des systèmes de dialogue oral.....	640-641
TermLis : un contexte d'information logique pour des ressources terminologiques.....	642-643
Etude de l'image de marque d'entités dans le cadre d'une plateforme de veille sur le Web social.....	644-645
Building a Bilingual Vietnamese-French Named Entity Annotated Corpus through Cross-Linguistic Projection.....	646-647
Recherche de motifs de graphe en ligne.....	648-649
Un patient virtuel dialogant.....	650-651
Intégration du corpus des actes de TALN à la plateforme ScienQuest.....	652-653
Une aide à la communication par pictogrammes avec prédiction sémantique.....	654-656
Un système expert fondé sur une analyse sémantique pour l'identification de menaces d'ordre biologique.....	657-658
DisMo : un annotateur multi-niveaux pour les corpus oraux.....	659-661

Multilinguality at Your Fingertips: BabelNet, Babelfy and Beyond!

Roberto Navigli
Dipartimento di Informatica
Sapienza Università di Roma
Viale Regina Elena 295, 00161 Roma, Italy
navigli@di.uniroma1.it

Abstract.

Multilinguality at Your Fingertips : BabelNet, Babelfy and Beyond !

Multilinguality is a key feature of today's Web, and it is this feature that we leverage and exploit in our research work at the Sapienza University of Rome's Linguistic Computing Laboratory, which I am going to overview and showcase in this talk.

I will start by presenting BabelNet 3.0, available at <http://babelnet.org>, a very large multilingual encyclopedic dictionary and semantic network, which covers 271 languages and provides both lexicographic and encyclopedic knowledge for all the open-class parts of speech, thanks to the seamless integration of WordNet, Wikipedia, Wiktionary, OmegaWiki, Wikidata and the Open Multilingual WordNet.

Next, I will present Babelfy, available at <http://babelfy.org>, a unified approach that leverages BabelNet to jointly perform word sense disambiguation and entity linking in arbitrary languages, with performance on both tasks on a par with, or surpassing, those of task-specific state-of-the-art supervised systems.

Finally I will describe the Wikipedia Bitaxonomy, available at <http://wibitaxonomy.org>, a new approach to the construction of a Wikipedia bitaxonomy, that is, the largest and most accurate currently available taxonomy of Wikipedia pages and taxonomy of categories, aligned to each other. I will also give an outline of future work on multilingual resources and processing, including state-of-the-art semantic similarity with sense embeddings.

Mots-clés : multilinguisme, dictionnaire, sémantique, désambiguïsation, taxonomie.

Keywords: multilinguality, dictionary, semantic, disambiguation, taxonomy.

Pourquoi construire des ressources terminologiques et pourquoi le faire différemment ?

Marie-Claude L'Homme
Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal, C.P. 6128, succ. Centre-ville
Montréal (Québec), H1T 3L9, Canada
mc.lhomme@umontreal.ca

Résumé. Dans cette présentation, je défendrai l'idée selon laquelle des ressources terminologiques décrivant les propriétés lexico-sémantiques des termes constituent un complément nécessaire, voire indispensable, à d'autres types de ressources. À partir d'exemples anglais et français empruntés au domaine de l'environnement, je montrerai, d'une part, que les ressources lexicales générales (y compris celles qui ont une large couverture) n'offrent pas un portrait complet du sens des termes ou de la structure lexicale observée du point de vue d'un domaine de spécialité. Je montrerai, d'autre part, que les ressources terminologiques (thésaurus, ontologies, banques de terminologie) souvent d'obédience conceptuelle, se concentrent sur le lien entre les termes et les connaissances dénotées par eux et s'attardent peu sur leur fonctionnement linguistique. Je présenterai un type de ressource décrivant les propriétés lexico-sémantiques des termes d'un domaine (structure actantielle, liens lexicaux, annotations contextuelles, etc.) et des éléments méthodologiques présidant à son élaboration.

Abstract.

Why compile terminological resources and why do it differently?

In this talk, I will argue that terminological resources that account for the lexico-semantic properties of terms are a necessary – perhaps even indispensable – complement to other kinds of resources. Using examples taken from the field of the environment, I will first show that general lexical resources (including those that have a large coverage) do not give a complete picture of the meaning of terms or the lexical structure observed from the point of view of a special subject field. I will then proceed to show that terminological resources (thesauri, ontologies, term banks), that often take a conceptual approach, focus on the relationship between terms and the knowledge they convey and give little information about the linguistic behavior of terms. I will present a type of resource that describes the lexico-semantic properties of terms (argument structure, lexical relationships, contextual annotations, etc.) and some methodological considerations about its compilation.

Mots-clés : ressources terminologiques, ressources lexicales, liens lexicaux, corpus spécialisé, structure actantielle, annotations contextuelles

Keywords: terminological resources, lexical resources, lexical relationships, specialized corpora, argument structure, contextual annotations

Apprentissage par imitation pour l'étiquetage de séquences : vers une formalisation des méthodes d'étiquetage « *easy-first* »

Elena Knyazeva^{1,2} Guillaume Wisniewski^{1,2} François Yvon²
(1) Université Paris Sud, 91 403 Orsay CEDEX
(2) LIMSI-CNRS, 91 403 Orsay CEDEX
{nom.prénom}@limsi.fr

Résumé. De nombreuses méthodes ont été proposées pour accélérer la prédiction d'objets structurés (tels que les arbres ou les séquences), ou pour permettre la prise en compte de dépendances plus riches afin d'améliorer les performances de la prédiction. Ces méthodes reposent généralement sur des techniques d'inférence approchée et ne bénéficient d'aucune garantie théorique aussi bien du point de vue de la qualité de la solution trouvée que du point de vue de leur critère d'apprentissage.

Dans ce travail, nous étudions une nouvelle formulation de l'apprentissage structuré qui consiste à voir celui-ci comme un processus incrémental au cours duquel la sortie est construite de façon progressive. Ce cadre permet de formaliser plusieurs approches de prédiction structurée existantes. Grâce au lien que nous faisons entre apprentissage structuré et apprentissage par renforcement, nous sommes en mesure de proposer une méthode théoriquement bien justifiée pour apprendre des méthodes d'inférence approchée. Les expériences que nous réalisons sur quatre tâches de TAL valident l'approche proposée.

Abstract.

Imitation learning for sequence labeling: towards a formalization of easy-first labeling methods.

Structured learning techniques, aimed at modeling structured objects such as labeled trees or strings, are computationally expensive. Many attempts have been made to reduce their complexity, either to speed up learning and inference, or to take richer dependencies into account. These attempts typically rely on approximate inference techniques and usually provide very little theoretical guarantee regarding the optimality of the solutions they find.

In this work we study a new formulation of structured learning where inference is primarily viewed as an incremental process along which a solution is progressively computed. This framework generalizes several structured learning approaches. Building on the connections between this framework and reinforcement learning, we propose a theoretically sound method to learn to perform approximate inference. Experiments on four sequence labeling tasks show that our approach is very competitive when compared to several strong baselines.

Mots-clés : Apprentissage par Imitation ; Apprentissage Structuré ; Étiquetage de Séquences.

Keywords: Imitation Learning ; Structured Learning ; Sequence Models.

1 Introduction

L'apprentissage structuré a pour objectif de prédire des objets composés de parties¹ inter-dépendantes, comme c'est le cas pour des arbres ou des séquences, en exploitant les dépendances entre parties pour améliorer les performances des prédictions. La prédiction d'objets structurés est au cœur de nombreuses tâches de Traitement Automatique des Langues (TAL) : analyse syntaxique en dépendances, reconnaissance d'entités nommées, traduction automatique, etc. La quasi totalité des modèles d'apprentissage structuré, tels que les grammaires hors-contexte probabilistes (PCFG) ou les champs aléatoires conditionnels (CRF), reposent sur une généralisation de la classification multi-classes : ils visent en effet à apprendre une fonction de score (par exemple une probabilité jointe ou conditionnelle) mesurant l'adéquation entre une observation et chacune des structures possibles. Une fois apprise, cette fonction permet, lors de la prédiction, de distinguer la meilleure solution parmi toutes les hypothèses considérées. Notre travail propose une formulation alternative

1. Dans la suite, nous considérons que chaque partie de la structure à prédire peut être représenté par une étiquette choisie dans un inventaire fini.

de l'apprentissage structuré, issue du modèle de Collins & Roark (2004). Plutôt que de modéliser les propriétés d'une « bonne » solution, puis de chercher cette solution parmi un ensemble de candidats, cette formulation consiste à modéliser directement le processus permettant de *construire* une solution correcte. Elle permet de contourner plusieurs limites des modèles d'apprentissage structuré classiques (Daumé III & Marcu, 2005).

En effet, les modèles d'apprentissage structuré classiques souffrent de deux problèmes majeurs. Premièrement, leur expressivité est limitée : pour pouvoir résoudre le problème (combinatoire) de la recherche de la meilleure sortie, il est en effet nécessaire de choisir une paramétrisation idoine de la fonction de score, qui ne considère que des dépendances *locales* entre étiquettes (hypothèse de Markov). La programmation dynamique permet alors de trouver efficacement, par exemple à l'aide de l'algorithme de Viterbi, la solution de meilleur score. En conséquence, la plupart des méthodes d'apprentissage structuré de l'état de l'art ne peuvent pas tirer pleinement profit de toutes les informations contenues de structure. Deuxièmement, même lorsque seules des dépendances locales sont considérées, la complexité des méthodes d'apprentissage structuré reste souvent élevée, surtout pour des tâches présentant un grand nombre d'étiquettes.

Les méthodes de décodage approché reposant, par exemple, sur des stratégies de recherche gloutonne ou sur des algorithmes comme A^* , constituent une manière naturelle d'éviter ces deux problèmes : elles permettent soit de considérer des paramétrisations du score pour lesquelles il n'existe pas de méthode d'inférence efficace, soit, lorsque seules des dépendances locales sont considérées, d'accélérer le décodage. Les performances en prédiction de ces approches sont toutefois limitées par le phénomène de propagation des erreurs : comme le choix des étiquettes est, en partie, fondé sur le choix des étiquettes précédentes (qui n'est généralement pas remis en cause), une erreur de prédiction complique les choix futurs et la probabilité de commettre une erreur augmente donc rapidement au fur et à mesure que les décisions sont prises.

Deux familles de méthodes ont été proposées dans la littérature pour limiter la propagation d'erreurs lors d'un décodage avec une méthode heuristique. Les méthodes de la première famille cherchent à modifier la distribution des exemples d'apprentissage en fonction des erreurs qui sont faites lors du décodage, afin d'apprendre à faire des prédictions correctes même en présence d'erreurs dans les décisions passées. Ainsi, Finkel *et al.* (2006) proposent une stratégie fondée sur la modélisation de trajectoires engendrées par le décodeur en utilisant une méthode de Monte Carlo. Cette approche bénéficie de garanties théoriques fortes, mais sa complexité rend sa mise en œuvre prohibitive. Daumé III *et al.* (2009) décrivent un modèle similaire dans un cadre d'apprentissage par renforcement. Notre travail s'inscrit dans la continuité de cette étude.

Les méthodes de la seconde famille, généralement qualifiées de « plus simple d'abord » (*easy first*), modifient le décodage afin de retarder les décisions les moins sûres, limitant ainsi le risque de propagation d'erreurs. Par exemple, Shen *et al.* (2007) décrivent un analyseur morpho-syntaxique capable de prédire les étiquettes dans un ordre libre ; Yamada & Matsumoto (2003) proposent un analyseur en dépendances qui s'appuie sur une méthode d'inférence gloutonne et Goldberg & Elhadad (2010) généralisent cette approche pour construire les dépendances en ordre libre. Finalement, Gesmundo & Henderson (2014) appliquent ce principe à la traduction automatique hiérarchique (Chiang, 2007) en relâchant la contrainte de réaliser l'inférence de manière ascendante. Bien que ces méthodes aient obtenu des résultats expérimentaux concluants, elles ne bénéficient d'aucune garantie théorique et leur mise en œuvre repose sur plusieurs heuristiques.

En prenant pour exemple la tâche d'étiquetage de séquences, nous proposons, dans ce travail, un nouveau cadre qui permet de réunir ces deux familles dans un formalisme commun (§ 2). En faisant le lien entre ce cadre et l'apprentissage par renforcement, nous proposons une méthode théoriquement bien justifiée pour apprendre des méthodes de décodage *easy first* (§ 3). Les expériences réalisées sur quatre tâches de TAL (§ 4) montrent que la méthode proposée permet d'améliorer légèrement les résultats de l'état-de-l'art, tout en réduisant la complexité de l'inférence et de l'apprentissage.

2 Apprentissage structuré incrémental

Dans cette section, nous introduisons un cadre général pour l'apprentissage structuré incrémental (§ 2.1) qui permet de reformuler le problème de l'apprentissage structuré ; nous présentons ensuite deux modèles d'étiquetage de séquences (§ 2.2) quiinstancient ce cadre.

2.1 Principes

Le principe fondamental des algorithmes d'apprentissage structuré *incrémental* est de considérer un espace de recherche regroupant l'ensemble des structures possibles, que l'ensemble des sous-parties de celles-ci. Par exemple, dans le cas de l'étiquetage de séquences, l'espace de recherche considéré regroupera tous les étiquetages possibles de la séquence d'ob-

servations complète, ainsi que tous les étiquetages *partiels* de celle-ci (c.-à-d. dans lesquels seules certaines observations sont étiquetées).

L'espace de recherche, muni de la relation d'ordre suivante : xRy si x est une sous-partie de y , définit alors un semi-treillis. Cette définition permet de reformuler la recherche de la structure de plus grand score comme un processus de construction : dans le treillis décrivant l'espace de recherche, chaque nœud correspond à une solution partielle et une arête dénote un moyen d'*étendre* cette solution vers une solution plus complète. Il est ainsi possible de considérer l'ensemble des arrêtes issues d'un nœud comme un ensemble d'actions réalisables à partir de ce nœud : le parcours de l'espace de recherche peut alors être décrit comme un processus de construction qui va incrémentalement étendre une solution partielle vers une solution complète. L'inférence se résume alors à une suite d'*actions* ou de *décisions* (quelle solution partielle étendre ? comment l'étendre ? etc.)

Plus formellement, un espace de recherche peut être défini, en intension, par un Processus de Décision Markovien (PDM) (Sutton & Barto, 1998). Dans ce travail ², un PDM correspond à un quintuplet $\langle \mathcal{S}, \mathcal{A}, R, \mathcal{S}_f, s_i \rangle$:

- \mathcal{S} est l'ensemble des états possibles (l'ensemble des séquences d'étiquettes complètes ou non) ;
- \mathcal{A}_s est l'ensemble des actions réalisables dans l'état s ; chacune de ces actions permet de passer dans un état s' en « étendant » la structure partielle décrite par l'état s ; on notera $s' = s \oplus a$ pour noter l'opération d'extension ;
- r est une fonction de $\mathcal{S} \times \mathcal{A}$ dans \mathbb{R} qui définit la *récompense* $r(a, s)$ reçue lorsque l'action a est exécutée dans l'état s ;
- $\mathcal{S}_f \subset \mathcal{S}$ est l'ensemble des états finaux ;
- $s_i \in \mathcal{S}$ est un état initial (supposé unique).

La fonction de récompense peut être choisie de manière arbitraire. Il est cependant naturel de la lier à la qualité (telle qu'évaluée par la fonction de coût du problème) de la solution partielle que permet d'obtenir l'action a . Dans ce cas toutefois, les récompenses ne peuvent être connues que pour des données étiquetées, puisqu'elles s'appuient sur l'évaluation (partielle) de la fonction de coût. Déterminer un apprenant capable de prédire ces récompenses pour les données de test est au cœur de l'approche que nous proposons ici (§3).

Une *politique* $\pi : \mathcal{S} \mapsto \mathcal{A}$ est une fonction qui détermine l'action $a_t = \pi(s_t)$ qui doit être choisie dans un état s_t donné. Elle spécifie une suite d'actions a_1, \dots, a_T , qui permet d'atteindre un état final à partir de l'état initial en construisant une séquence d'étiquettes y . Une politique, et par conséquent, la séquence d'étiquettes qu'elle engendre, est associée à une *récompense cumulée* définie par :

$$V(\pi) = \sum_{t \in \llbracket 0, T \rrbracket} r(s_t, \pi(s_t)) \quad (1)$$

où s_0 est l'état initial et s_T est un état final.

Étant donné un PDM, l'apprentissage par renforcement cherche à déterminer la *politique optimale*, c'est-à-dire celle dont la récompense cumulée est la plus grande. Il existe, pour les « petits » PDM, des méthodes, fondées sur la programmation dynamique, comme les algorithmes *Value Iteration* et *Policy Iteration* (Sutton & Barto, 1998) qui permettent de construire la politique optimale efficacement. Mais, dans le cas général, quand l'espace de recherche ne peut plus être exploré entièrement, il faut s'en remettre à des méthodes de recherche approchée. Nous adoptons, dans ce travail, une méthode qui repose sur la politique gloutonne :

$$a_t = \pi(s_t) = \arg \max_{a \in \mathcal{A}_{s_t}} f_\pi(s_t, a) \quad (2)$$

où $f_\pi(s_t, a)$ est une fonction de score associée à la politique π , dépendant uniquement de l'action a et de l'état courant s_t . La fonction de score doit être choisie de sorte que la réalisation de la politique associée obtienne une récompense cumulée $V(\pi)$ élevée. Nous verrons, à la section 3 comment l'apprentissage par imitation permet d'estimer conjointement la fonction de score et la fonction de récompense.

Avec une stratégie gloutonne, la réalisation d'une politique (qui permet d'engendrer la séquence d'étiquettes) possède une complexité linéaire par rapport au nombre d'états et d'actions, tant que la fonction de score n'utilise que des informations contenues dans l'état courant. Elle n'offre toutefois aucune garantie concernant l'optimalité de la solution trouvée.

2. Dans la définition usuelle des PDM, lors de la réalisation d'une action a dans un état s le système passe dans un état s' avec une certaine probabilité. Cette généralisation permet de modéliser une interaction plus réaliste avec l'environnement. Par exemple, en robotique, lorsque le système décide de réaliser l'action « tourner à gauche de 30° » il est possible qu'à causes des frottements ou d'obstacles il ne se retrouve pas exactement dans l'état envisagé. Par souci de clarté, nous simplifions nos définitions au cas déterministe qui est suffisant pour décrire notre travail.

2.2 Modèles d'étiquetage de séquences

Nous allons introduire, dans cette section, deux PDM permettant de réaliser un étiquetage de séquences. Dans la suite, nous noterons $\mathbf{x} = (x_t)_{t=1}^T$ une séquence de T observations, $\mathbf{y} = (y_t)_{t=1}^T$ la séquence d'étiquettes correspondante avec $x_t \in \mathcal{X}$ et $y_t \in \mathcal{Y}$. Nous utiliserons également les notations \mathbf{x}_θ , où θ est un ensemble d'entiers compris entre 1 et T , pour désigner les éléments de \mathbf{x} dont les indices sont dans θ ; et $i:j$ pour l'ensemble des entiers compris entre i et j (inclus).

2.2.1 Étiquetage (monotone) gauche-droite

Le premier modèle d'étiquetage considéré correspond au PDM suivant :

- $\mathcal{S} = \{(\mathbf{x}_{1:T}, \mathbf{y}_{1:t}), \mathbf{x}_{1:T} \in \mathcal{X}^T, \mathbf{y}_{1:t} \in \mathcal{Y}^t, 0 \leq t \leq T\}$: chaque état est décrit par l'observation et une sortie partielle dont seules les t premières étiquettes sont prédites ;
- $\mathcal{A}_s = \{y, y \in \mathcal{Y}\}$: chaque action ajoute une nouvelle étiquette à une séquence dont les t premières étiquettes sont connues et permet de passer de l'état $s = (\mathbf{x}_{1:T}, \mathbf{y}_{1:t})$ à l'état $s' = (\mathbf{x}_{1:T}, \mathbf{y}_{1:t+1})$ avec $\mathbf{y}_{1:t+1} = \mathbf{y}_{1:t} \oplus y$;
- $\mathcal{S}_f = \{(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}), \mathbf{x}_{1:T} \in \mathcal{X}^T, \mathbf{y}_{1:T} \in \mathcal{Y}^T\}$: les états finaux correspondent à l'ensemble des séquences complètement étiquetées ;
- $s_i = (\mathbf{x}_{1:T}, \emptyset)$: l'état initial correspond à une séquence d'étiquettes vide.

La fonction de récompense sera définie au § 3.1. Ce PDM comporte $\sum_{t=0}^T |\mathcal{Y}|^t$ états et $\sum_{t=0}^T |\mathcal{Y}|^t \times |\mathcal{Y}|$ actions ($|\mathcal{Y}|$ actions par état). Il permet de construire une solution en choisissant, successivement de gauche à droite, l'étiquette de chaque observation et décrit un espace de recherche qui correspond à celui considéré par l'algorithme de Viterbi (et plus généralement par les autres algorithmes utilisant la programmation dynamique comme l'algorithme *forward-backward*). La Figure 1 en donne un exemple.

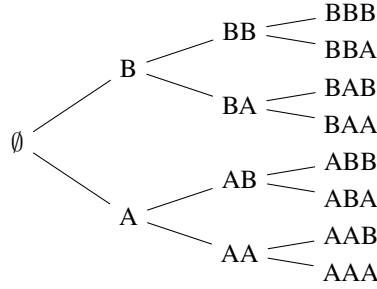


FIGURE 1 – Espace de recherche correspondant à l'étiquetage gauche-droite d'une séquence de trois observations lorsqu'il y a deux étiquettes possibles (A et B). Pour simplifier, la représentation des états n'inclut pas la séquence d'observations.

Ce modèle et le cadre d'apprentissage associé offrent deux avantages par rapport aux modèles de séquences classiques comme les HMM ou les CRF. Premièrement la complexité de l'inférence (en $\mathcal{O}(T \cdot |\mathcal{Y}|)$) est plus faible que celle des modèles qui se basent sur un décodage de Viterbi (en $\mathcal{O}(T \cdot |\mathcal{Y}|^2)$). Deuxièmement, il est possible d'introduire, sans changer la complexité de l'inférence ou de l'apprentissage, des caractéristiques décrivant toutes les décisions passées (ces informations sont directement déductibles des informations stockées dans chaque état) alors que, dans un CRF linéaire d'ordre 1, seules des caractéristiques décrivant l'étiquette précédente peuvent être incluses³. L'inférence sera toutefois approximative : rien ne garantit que la solution trouvée correspond effectivement à celle qui a le plus grand score.

2.2.2 Étiquetage en ordre libre

Le second modèle d'étiquetage que nous considérons est une généralisation du modèle précédent, *qui n'impose plus l'ordre dans lequel les positions étiquetées sont choisies*. Il correspond au PDM suivant :

- $\mathcal{S} = \{(\mathbf{x}_{1:T}, \mathbf{y}_\theta), \theta \in \wp([1 : T])\}$ où $\wp(E)$ est l'ensemble des parties de E : chaque état correspond à une séquence dont seuls les éléments dont les indices sont dans θ sont étiquetés ; les positions étiquetées ne sont pas nécessairement contiguës ;

3. Inclure des caractéristiques d'ordre supérieur dans des CRF augmente la complexité de l'apprentissage et de l'inférence exponentiellement en l'ordre du modèle.

- $\mathcal{A}_s = \{(y, t), y \in Y, t \notin \theta\}$ pour $s = (\mathbf{x}_{1:T}, \mathbf{y}_\theta)$: une action choisit *une des positions non étiquetées et son étiquette* et permet de passer de l'état $(\mathbf{x}_{1:T}, \mathbf{y}_\theta)$ à l'état $(\mathbf{x}_{1:T}, \mathbf{y}_{\theta'})$ avec $\theta' = \theta \cup \{t\}$ et $y_t = y$;
- $\mathcal{S}_f = \{(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}), \mathbf{x}_{1:T} \in \mathcal{X}^T, \mathbf{y}_{1:T} \in \mathcal{Y}^T\}$: les états finaux correspondent à l'ensemble des séquences totalement étiquetées ;
- $s_0 = (\mathbf{x}_{1:T}, \emptyset)$: l'état initial correspond à une séquence d'étiquettes vide.

La fonction de récompense est définie en § 3.1. Ce modèle permet d'étiqueter une séquence en choisissant l'ordre dans lequel les étiquettes sont attribuées. Il repose sur l'intuition que certaines étiquettes sont plus faciles à prédire que d'autres et que la connaissance de celles-ci facilitera le choix des autres étiquettes. Comme pour le modèle gauche-droite, les informations stockées dans chaque état permettent de définir des caractéristiques plus riches (dépendances longues, étiquettes « futures », ...) que celles prises en compte dans les modèles standard de prédiction de séquences.

La figure 2 donne un exemple de l'espace de recherche considéré lors d'un étiquetage en ordre libre. De manière générale, celui-ci comporte $\sum_{t=0}^T \binom{t}{T} |\mathcal{Y}|^t$ états et $\sum_{t=0}^T \binom{t}{T} |\mathcal{Y}|^t \times |\mathcal{Y}|$ actions. On notera que lorsque la contrainte sur l'ordre des étiquettes est relâchée, plusieurs séquences d'actions peuvent conduire dans le même état.

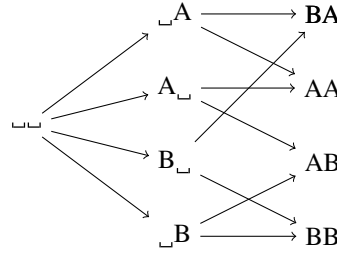


FIGURE 2 – Espace de recherche d'un étiquetage en ordre libre d'une séquence de deux observations lorsqu'il y a deux étiquettes possibles. Les positions dont l'étiquette n'est pas encore prédite sont indiquées par le symbole \square .

3 Apprentissage de la fonction de score

3.1 Apprentissage par imitation

Dans le cadre de l'apprentissage structuré incrémental, l'objectif de l'apprentissage est d'estimer la fonction de récompense à partir d'un ensemble de séquences étiquetées, puis la fonction de score f sur laquelle la politique gloutonne est fondée. La fonction de score devra être choisie de sorte que *i)* la récompense cumulée de la politique soit d'autant plus élevée que la solution est de bonne qualité (telle qu'évaluée par une fonction de coût dépendant de la tâche, comme le score F_1 ou la distance de Hamming) *ii)* en intégrant le fait que les décisions sont prises de manière gloutonne et, donc que le score doit inclure une estimation des coûts futurs pour que la solution trouvée soit proche de la solution optimale.

Plusieurs méthodes répondant à ces deux objectifs ont été proposées en apprentissage par renforcement. Dans ce travail nous nous intéressons aux méthodes dites d'*apprentissage par imitation (imitation learning)* (Abbeel & Ng, 2004; Ross & Bagnell, 2010). Ces méthodes estiment la fonction de score et de récompense conjointement en *réduisant* ces deux problèmes à un problème de classification multi-classes (Langford & Zadrozny, 2005) : le parcours de l'espace de recherche est vu comme une suite de problèmes de classification dont l'objectif est de retrouver, dans chaque état, la meilleure action possible, correspondant à celle que prendrait un oracle. Il est raisonnable de supposer que cette action est celle dont le choix permettra d'obtenir, *dans l'état final*, une solution dont la récompense *cumulée* sera optimale. L'exemple canonique motivant ce cadre est l'apprentissage d'un système conduisant une voiture à partir de l'observation du comportement d'un conducteur (Ross & Bagnell, 2010).

Plus précisément, nous considérons, pour choisir l'action à effectuer dans un état s , un classifieur linéaire multi-classes. La fonction de score de la règle de décision décrite dans l'équation (2) s'écrit alors :

$$f(s, a) = \langle \phi(a, s) | \mathbf{w} \rangle \quad (3)$$

où $\langle \cdot | \cdot \rangle$ est le produit scalaire usuel, \mathbf{w} est le vecteur de paramètres du classifieur, $\phi(a, s)$ est un vecteur de caractéristiques décrivant l'action a et l'état courant s . Comme l'état encode toutes les informations concernant la séquence d'observations

et la solution partiellement étiquetée, il est possible de définir des caractéristiques riches prenant, par exemple, en compte des historiques de grande taille ou, dans le cas du modèle en ordre libre, les étiquettes des voisins déjà prédites.

L'apprentissage de ce classifieur repose sur la connaissance de la *politique oracle* qui associe, à chaque état, la meilleure action pouvant être choisie. Il suffit alors d'utiliser le corpus d'apprentissage pour engendrer l'ensemble des états et des actions correspondants, d'étiqueter ceux-ci grâce à la politique oracle (c.-à-d. de déterminer, pour chaque exemple, quelle est l'action qui aurait dû être choisie) puis d'estimer les paramètres du classifieur à l'aide d'un algorithme d'apprentissage supervisé standard. La mise en œuvre de ce principe soulève toutefois plusieurs problèmes :

1. Étant donnée la taille de l'espace de recherche, il est impossible de considérer tous les états possibles comme exemples lors de l'apprentissage ;
2. Dans le cadre proposé, le modèle de séquences est appris en optimisant une fonction de coût 0/1 qui permet de caractériser la capacité de l'apprenant à choisir la bonne action dans chaque état. Cette fonction objectif n'a aucun lien direct avec la qualité des séquences prédites et les garanties offertes par la théorie de l'apprentissage statistique⁴ ne s'appliquent donc pas. L'optimisation d'un coût 0/1 en apprentissage est-elle théoriquement fondée ?
3. Il n'est pas toujours possible, lors de l'apprentissage, d'avoir accès directement à un oracle indiquant quelle action choisir dans un état donné ; celui-ci devra être reconstruit à partir de la référence.

La plupart des méthodes *easy first* de la littérature ignorent ces questions ou n'y apportent qu'une réponse expérimentale. Le lien que nous faisons ici avec l'apprentissage par imitation, et plus généralement l'apprentissage par renforcement, permet d'obtenir des garanties théoriques. En effet, les deux premiers problèmes sont des problèmes intrinsèques à l'apprentissage par imitation et plusieurs solutions (Langford & Zadrozny, 2005; Daumé III *et al.*, 2009; Ross & Bagnell, 2010) ont été proposées dans la littérature. Nous détaillons une de ces solutions, SEARN, dans la section suivante. Le dernier problème est un problème spécifique à l'application de l'apprentissage par imitation à la prédiction de séquences auquel nous proposerons une nouvelle solution présentée dans la section 3.3.

3.2 SEARN

Nous résumons, dans cette section, le principe de SEARN et les garanties que présente cette méthode d'apprentissage⁵. Nous supposons, dans la suite, avoir accès à une politique oracle π^{oracle} capable de déterminer, parmi toutes les actions réalisables dans un état donné, l'action (supposée unique) permettant d'obtenir la solution avec la meilleure récompense cumulée.

SEARN (Daumé III *et al.*, 2009) est un algorithme générique d'apprentissage par imitation utilisé pour apprendre les paramètres d'une politique gloutonne reposant sur la règle de décision décrite par l'équation (2). L'idée centrale de SEARN est de construire l'ensemble d'apprentissage (les états à parcourir, les actions à effectuer) de manière itérative en commençant par suivre la distribution des états engendrée par la politique oracle et en introduisant progressivement les états engendrés par la politique apprise. Ce processus permet de limiter le nombre d'états considérés lors de l'apprentissage tout en garantissant que les exemples sur lesquels la politique est apprise seront similaires aux cas qui seront vus en test : si les exemples n'étaient engendrés qu'à partir de la politique oracle, le système ne saurait pas quelle décision prendre dès qu'il s'en éloignerait puisqu'il n'aurait jamais été confronté à une telle situation en apprentissage.

Plus précisément, comme le montre l'algorithme 1, SEARN apprend une suite de politiques $\pi^0, \pi^1, \dots, \pi^N$ où N , le nombre d'itérations, est un hyper-paramètre de l'algorithme. La politique initiale π^0 est la politique oracle π^{oracle} ; puis, à la i^{e} itération, la politique courante est utilisée pour engendrer un nouvel ensemble d'apprentissage : pour chaque séquence d'observations du corpus, π^i est utilisée pour prédire une séquence d'étiquettes et π^{oracle} pour déterminer la bonne action⁶ qui aurait dû être prise dans chaque état visité. Concrètement, dans chaque état visité, on ajoute à l'ensemble d'apprentissage en cours de construction un exemple qui associe à l'état l'action qu'aurait pris l'oracle. À la fin de l'itération, les paramètres d'une politique ρ^{i+1} sont estimés sur cet ensemble d'apprentissage à l'aide d'un algorithme de

4. Lorsque les exemples sont i.i.d., la fonction qui minimise, à l'intérieur d'une classe de fonctions donnée, l'erreur sur l'ensemble d'apprentissage est aussi celle qui minimise l'erreur sur l'ensemble de test avec une forte probabilité.

5. Par souci de simplification, nous avons présenté SEARN dans le cas où la contribution d'une action à la fonction de coût est soit 0 (lorsque la prédiction résultant de l'action est correcte), soit 1 (lorsque cette prédiction est erronée). Cette situation correspond, par exemple, à une évaluation par une distance de Hamming. La généralisation à d'autres fonctions de coût nécessite est possible en considérant une généralisation de la classification multi-classes, la classification multi-classes *pondérée* (*cost-sensitive classification*) qui ordonne les erreurs selon leur « gravité ». Cette généralisation ne change pas substantiellement l'esprit de l'algorithme, ni notre discussion.

6. La bonne action à prendre est choisie de cette manière dans SMILE, qui est la version de SEARN présentée dans (Ross & Bagnell, 2010).

classification multi-classes standard :

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \sum_{s, a \in \mathcal{S}} \mathbb{1} \left\{ \arg \max_{a'} \langle \phi(s, a') | \mathbf{w} \rangle = a \right\} \quad (4)$$

La $(i + 1)^{\text{e}}$ politique π^{i+1} est alors définie comme un *mélange stochastique* de la politique π^i et de la politique ρ^{i+1} :

$$a = \pi^{i+1}(s) = \begin{cases} \rho^{i+1}(s) & \text{avec une probabilité } \beta \\ \pi^i(s) & \text{avec une probabilité } 1 - \beta \end{cases} \quad (5)$$

où β , la probabilité de choisir la politique que l'on vient d'apprendre, est le second hyper-paramètre de l'algorithme. Ce mélange stochastique permet de contrôler la vitesse à laquelle on s'éloigne de la politique oracle : la probabilité d'appliquer π^{oracle} à la i^{e} itération est égale à $(1 - \beta)^i$.

Algorithme 1 : Principe de SEARN / SMILE

Entrées : un ensemble de séquences étiquetées $\mathcal{T} = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$, un nombre d'itérations N , $\beta \in [0, 1]$ vitesse à laquelle on s'éloigne de la politique oracle

```

1 Initialiser  $\pi^0$ ;
2 for  $i = 0 : N$  do
3    $S \leftarrow \emptyset$ ;                                ▷ Ensemble des exemples collectés
4   for  $\mathbf{x}, \mathbf{y} \in \mathcal{T}$ ;                                ▷ Décodage avec la politique  $\pi^{i-1}$ 
5   do
6      $\langle \mathcal{S}, \mathcal{A}, \mathcal{S}_f, s_0 \rangle \leftarrow \text{search\_space}(\mathbf{x})$ ;    ▷ Crée le PDM associé à l'exemple
7      $s \leftarrow s_0$ ;                                    ▷ État courant
8     while  $s \notin \mathcal{S}_f$  do
9        $S \leftarrow S \cup \{(s, \pi^{\text{oracle}}(s))\}$ ;
10       $s = \pi^i(s)$ ;                                ▷ Passe à l'état suivant
11    end
12  end
13  apprendre une nouvelle politique  $\rho^{i+1}$  sur  $S$ ;
14   $\pi^{i+1} = (1 - \beta) \cdot \pi^i + \beta \cdot \rho^{i+1}$ ;          ▷ Mélange stochastique
15 end
Sortie :  $\tilde{\pi}^N$ 

```

À la fin de l'apprentissage, le résultat de l'algorithme est la politique $\tilde{\pi}^N$ qui s'obtient en remplaçant dans la politique π^N la composante correspondant à la politique initiale π_0 (la politique oracle) par la politique qui choisit au hasard une des actions accessibles dans un état. En effet, la politique oracle n'est connue que pour les données d'entraînement ; la politique $\tilde{\pi}^N$ pourra être appliquée à des séquences d'observations dont les étiquettes de référence sont inconnues.

La version présentée ici de SEARN possède des garanties théoriques (Ross & Bagnell, 2010) bornant l'erreur en généralisation du classifieur structuré qui reposent sur deux caractéristiques de l'algorithme. Premièrement, la première politique appliquée est la politique oracle ; le premier classifieur appris permet alors de retrouver la meilleure séquence d'actions. Deuxièmement, grâce à l'utilisation du mélange stochastique, les états parcourus lors du décodage changent lentement, ce qui permet de limiter l'effet de propagation d'erreur lors du passage d'une politique π^i à la politique π^{i+1} .

3.3 Détermination de la politique oracle

La mise en œuvre de SEARN pour l'apprentissage de modèles de séquences suppose que l'on ait accès à une politique oracle capable de déterminer quelle action doit être effectuée dans un état donné. Déterminer la politique oracle lorsque l'on considère le modèle d'étiquetage gauche-droite et la distance de Hamming comme fonction de coût est aisé : à chaque état, l'oracle renvoie l'action qui associe la bonne étiquette (selon la référence) à la position suivante.

Par contre, dans le cas d'un décodage en ordre libre (toujours évalué par une distance de Hamming), plusieurs séquences d'actions peuvent engendrer la séquence d'étiquettes de référence et, à chaque état, plusieurs actions pourront être considérées comme optimales : quelle que soit la fonction de coût considérée, il y a toujours autant d'actions optimales que de positions non-étiquetées. Cela est directement dû au fait que l'ordre dans lequel la séquence d'étiquettes est générée est

une variable cachée : seul l’étiquetage final est observé. L’oracle sera donc une fonction de \mathcal{S} dans $\wp(\mathcal{A}_s)$ et renverra non plus une unique action optimale mais un ensemble d’actions optimales. Suivant Goldberg & Nivre (2013), nous qualifions ces oracles de *non-déterministes*.

La prise en compte d’oracles non-déterministes dans SEARN nécessite d’adapter l’algorithme d’apprentissage : il n’est, en effet, plus possible de réduire directement le problème de prédiction structuré à un problème de classification multi-classes, puisqu’il n’y a plus dans chaque état une unique bonne réponse. Deux stratégies peuvent être envisagées pour résoudre ce problème.

Suppression de l’ambiguïté La première stratégie consiste à supprimer l’ambiguïté en choisissant, parmi toutes les actions optimales identifiées par l’oracle, $\pi^{\text{oracle}}(s)$, une étiquette \hat{a} qui sera considérée comme l’étiquette devant être prédite et vers laquelle devra être faite une éventuelle mise à jour du vecteur de paramètres. C’est également cette action qui déterminera dans quel état ira le système (c.-à-d. celle qui sera utilisée pour effectuer la ligne 10 de l’algorithme 1).

Cette stratégie a été proposée par (Shen *et al.*, 2007) dans un contexte similaire⁷ et a été reprise, entre autres, dans (Goldberg & Nivre, 2012; Ma *et al.*, 2012; Gesmundo & Henderson, 2014). Shen *et al.* (2007) préconisent de prendre comme action de référence, l’action optimale qui possède le plus grand score :

$$\hat{a} = \arg \max_{a \in \pi^{\text{oracle}}(s)} f(s, a) \quad (6)$$

En effet, intuitivement, cette action correspond à l’action optimale dont la prédiction nécessitera la plus petite mise à jour et donc la plus petite remise en cause du vecteur de paramètres courant.

Le principal avantage de cette stratégie est de ne pas nécessiter de modification de l’algorithme d’apprentissage. Mais les hypothèses sur lesquelles reposent les garanties théoriques de SEARN sont violées. En particulier, il n’est plus possible, dans cette stratégie de contrôler la vitesse à laquelle on s’éloigne de la politique optimale.

Réduction à un problème d’ordonnancement Nous introduisons, dans ce travail, une stratégie alternative qui consiste à considérer toutes les actions optimales identifiées par l’oracle comme étant correctes et, en cas d’erreur, à renforcer chacune de ces actions.

Plus précisément, nous proposons de réduire le problème de prédiction structuré, non plus à un problème de classification multi-classes, mais à un problème d’ordonnancement bipartite (Liu, 2009) dont l’objectif est de distinguer un ensemble d’exemples positifs d’un ensemble d’exemples négatifs en assurant que tous les exemples positifs ont un score plus grand que tous les éléments négatifs. La collecte par SEARN des exemples d’apprentissage (ligne 9 de l’algorithme 1) correspond alors à l’opération :

$$S \leftarrow S \cup \{(s, a), \forall a \in \pi^{\text{oracle}}(s)\} \quad (7)$$

Lors du décodage, l’état suivant est déterminé en exécutant aléatoirement une des actions de $\pi^{\text{oracle}}(s)$. On peut montrer qu’avec cette stratégie, les garanties théoriques de SEARN restent valables.

Une comparaison expérimentale de ces deux stratégies est présentée à la section 4.

4 Expériences

Dans cette section, nous étudions les performances des deux modèles introduits dans ce travail, le modèle gauche-droite et le modèle ordre libre, sur différentes tâches de TAL. Nous commencerons par présenter notre protocole expérimental avant de rapporter et discuter nos résultats.

4.1 Protocole expérimental

Dans toutes nos expériences nous utilisons, comme classifieur multi-classes, une machine à vecteurs supports (SVM) avec un noyau linéaire et une régularisation L2 ; la valeur du paramètre contrôlant la régularisation est systématiquement

⁷. (Shen *et al.*, 2007) introduit cette idée pour l’apprentissage d’un analyseur morpho-syntaxique *en ligne* ; dans ce travail, nous ne considérons que des méthodes d’apprentissage *batch*.



FIGURE 3 – Exemple de données pré-segmentées pour la reconnaissance de l’écriture manuscrite.

aerodrome	E-rxdrom-	1- < 0 > 2 < -
aeronaut	E-rxnc-t	1- < 0 > 2- <
aeronautics	E-rxnc-tIks	2- < 0 > 1- < 0 <<
aeroplane	E-rxpren-	1- < 0 > 2 < -

FIGURE 4 – Extrait de NETTALK : à chaque mot est associé une séquence de phonèmes (avec ‘-’ un symbole « NULL », et structure prosodique (0,1,2 marquent différents degrés d’accentuation pour les voyelles, > et < marquent respectivement les attaques et coda de syllabes).

déterminée par validation croisée. Les résultats de nos deux modèles sont comparés à un SVM⁸ multi-classes « simple » n’utilisant aucune information sur les étiquettes du voisinage et à un CRF linéaire⁹ considérant uniquement l’étiquette précédente comme information de structure et réalisant une recherche exacte de la solution optimale.

Ces méthodes d’étiquetage de séquences sont comparées sur quatre tâches différentes d’étiquetage de séquences, qui ont été choisies car elles mettent en jeu un grand nombre d’étiquettes, ce qui rend l’apprentissage et l’inférence computationnellement coûteux :

Reconnaissance de l’écriture manuscrite Le corpus¹⁰ utilisé contient 44 images de 150 mots (soit 6 600 exemples au total). Il s’agit d’un corpus artificiel, très structuré (la plupart des combinaisons d’étiquettes sont interdites et la connaissance d’une étiquette désambiguise fortement les lettres voisines) qui est généralement utilisé pour tester les méthodes d’apprentissage structuré.

Chaque image est pré-segmentée en séquence d’images de lettres de taille 8×16 pixels. Quelques exemples de données pré-segmentées sont représentées sur la Figure 3. Les données sont réparties en 10 paquets ; nous utilisons 9 paquets pour l’entraînement et 1 paquet pour le test et nous faisons la validation croisée sur les 10 configurations. Nous utilisons, comme caractéristiques, la valeur des 144 pixels ainsi que les 9 étiquettes précédentes.

Prononciation automatique L’objectif de cette tâche est de déterminer automatiquement la prononciation d’un mot à partir de la suite de lettres le composant. L’information de prononciation est constituée de deux parties qui seront apprises et évaluées séparément : une suite de phonèmes et une description de la structure prosodique.

Dans nos expériences, nous utilisons le corpus NETTALK¹¹ (Sejnowski & Rosenberg, 1987), qui contient 20 008 mots anglais accompagnés d’une information de prononciation. Pour chaque mot anglais contenant T lettres, la représentation phonologique correspondante est encodée par deux séquences de T symboles : une séquence de phonèmes, utilisant un alphabet de 51 étiquettes (dont un symbole NULL) et une séquence décrivant la structure prosodique (pour les consonnes, la position dans la syllabe, pour les voyelles le degré d’accentuation) avec un alphabet de 6 symboles (dont un symbole NULL). Un extrait de NETTALK est représenté sur la Figure 4. Les données sont partitionnées en 10 paquets. Nous utilisons 8 paquets pour l’entraînement, 1 paquet pour l’optimisation de l’hyper-paramètre et 1 paquet pour le test.

Nous utilisons comme caractéristiques les 4-grammes de lettres dans une fenêtre de taille ± 9 par rapport à la position courante. Les caractéristiques de structure correspondent aux étiquettes décodées (phonèmes ou marques prosodiques) ainsi que leurs combinaisons avec les caractéristiques de base.

Analyse morpho syntaxique de l’allemand Pour cette tâche nous avons utilisé le corpus TIGER¹², contenant 50 000

8. Nous avons utilisé l’implémentation des SVM fournie dans la bibliothèque SCIKIT-LEARN (Pedregosa *et al.*, 2011)

9. Implémenté dans Wapiti (Lavergne *et al.*, 2010) : <http://wapiti.limsi.fr>

10. <http://www.seas.upenn.edu/~taskar/ocr/>

11. [https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+\(Nettalk+Corpus\)](https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+(Nettalk+Corpus))

12. <http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.en.html>

	Écriture	Phonèmes	Accents	POS tags
SVM non-struct	74,97%	93,17%	91,49%	96,34%
Gauche-droite	91,93%	92,51%	90,78%	96,48%
Ordre libre	97,28%	93,32%	92,28%	96,91%
CRF	90,28%	93,02%	91,71%	97,00%
Gauche-droite (sans prop.)	96,56%	94,41%	94,31%	96,73%
Ordre libre (sans prop.)	98,65%	94,10%	94,12%	97,11%

TABLE 1 – Performances (% d’étiquettes correctes) des différents apprentis considérés sur 4 tâches standard.

phases allemandes, soit 900 000 mots étiquetés avec leur catégorie syntaxique. Ce corpus distingue 54 catégories différentes. Nous utilisons le partitionnement standard de ce corpus (80% des données pour l’entraînement, 10% pour la validation et 10% pour le test).

Dans nos expériences nous utilisons les caractéristiques suivantes : des informations de surface du mot courant (présence de majuscules, de chiffres, ...), les préfixes et les suffixes de taille de 1 à 4, les mots dans une fenêtre de taille 2 par rapport au mot considéré. Le mot courant est également combiné avec les 9 étiquettes précédentes.

4.2 Mise-en-œuvre de SEARN

Afin de réduire la complexité de l’apprentissage, nous avons, dans notre implémentation de SEARN, simplifié la définition du mélange stochastique défini par l’équation (5) en ne considérant que les deux dernières politiques apprises. Formellement, nous supposons que $\forall 0 < j < i, \rho^j = \rho^{i-1}$. La politique stochastique à l’itération i mélange donc simplement le dernier classifieur appris ρ^{i-1} (avec probabilité $1 - (1 - \beta)^i$) et la politique optimale π_0 . Par conséquent, la politique finale $\tilde{\pi}^N$ ne comporte plus qu’une composante ρ^N .

Cette approximation s’appuie sur le fait que les classifieurs les plus probables sont les classifieurs les plus récents, qui diffèrent peu du dernier classifieur appris. Des expériences préliminaires ont montré que cette simplification n’avait pas d’impact sur les performances de l’approche.

Suivant Daumé III *et al.* (2009), nous avons fixé β à $\frac{1}{T}$ où T est la longueur de la phrase considérée. Considérer des valeurs de β plus petite n’améliore pas les performances mais augmente significativement la vitesse de convergence ; des valeurs plus grandes de β dégradent les performances.

Sauf mention contraire, dans toutes nos expériences, nous avons utilisé pour déterminer la politique oracle la stratégie consistant à conserver l’ambiguïté (la seconde des stratégies présentées à la section 3.3). En effet, celle-ci a obtenu de meilleurs résultats dans nos expériences préliminaires.

4.3 Résultats et discussion

Les principaux résultats expérimentaux sont résumés dans la Table 1.

Ces résultats montrent, conformément à l’intuition, que la recherche exacte (mise en œuvre uniquement dans le CRF) améliore, légèrement, les résultats par rapport à une approche gloutonne sauf quand les sorties sont très structurées (comme pour la tâche de reconnaissance de l’écriture). En effet, dans ce cas la connaissance d’un historique riche (les 9 dernières étiquettes prédites alors que le CRF ne considère des dépendances qu’entre deux étiquettes consécutives) apporte une information suffisante pour compenser les erreurs induites par l’algorithme de recherche approchée. Ce phénomène apparaît également lorsque l’on considère un espace de recherche plus grand : le modèle « ordre libre » obtient des performances au moins aussi bonnes que le CRF et arrive même souvent à obtenir des résultats meilleurs.

Pour évaluer la capacité du modèle ordre libre à s’éloigner du décodage monotone, nous avons calculé la différence moyenne entre l’indice de la décision et la position concernée dans le mot. Pour la tâche de prosodie, cette valeur est de 2,7 ; pour la tâche de reconnaissance de phonèmes, elle est de 2,9. Ces deux observations montrent que l’ordre du décodage est, en pratique, assez éloigné de l’ordre monotone.

Une étude qualitative de l’ordre dans lequel le système effectue les actions, montre que la « facilité » de l’action correspond souvent à notre intuition. Par exemple, dans le cadre de la tâche de prosodie, on peut imaginer que l’accent secondaire

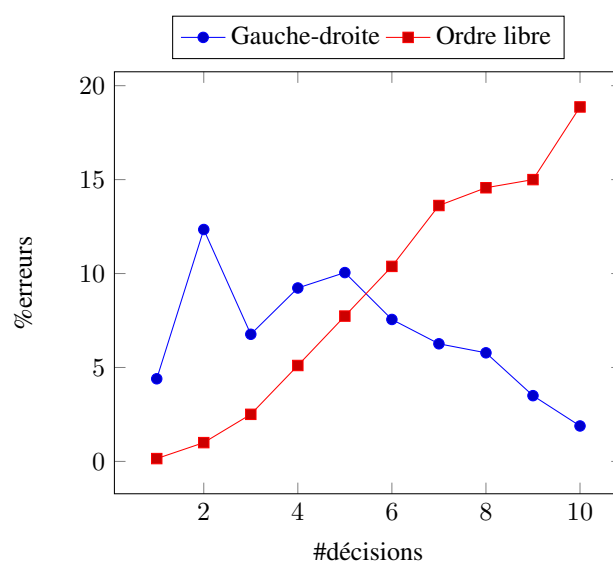


FIGURE 5 – Conversion graphème-phonème. Taux d’erreur en fonction de l’indice de la décision : modèle gauche-droite (à gauche) et modèle ordre libre (à droite)

peut être plus facile à détecter quand on connaît déjà l’information relative à l’accent primaire. En analysant l’ordre d’attribution des étiquettes, on peut voir que l’accent secondaire (quand il existe) dans un mot est mis après l’accent primaire dans 70% des cas. De manière générale, le système a tendance à décoder les consonnes (sur lesquels on commet 31% du nombre total d’erreurs) avant les voyelles (62% de toutes les erreurs) ; le reste des erreurs est commis sur les sons muets (7%).

Pour illustrer la capacité du modèle « ordre libre » à commencer par choisir les étiquettes les plus faciles en premier, nous avons représenté, à la figure 5, le taux d’erreur en fonction de l’indice de la décision dans la séquence pour la tâche graphème-phonème. Cette figure montre clairement que, dans le modèle « ordre libre », les erreurs apparaissent beaucoup plus tardivement lors du décodage, ce qui limite le problème de la propagation d’erreurs et permet d’obtenir les bonnes performances observées.

Pour quantifier l’effet de la propagation d’erreur pour les méthodes d’apprentissage par imitation, nous avons effectué l’expérience de contrôle suivante : le décodage incrémental est toujours effectué par la politique apprise (et les erreurs de prédictions ont toujours lieu), mais l’historique est systématiquement remis à jour avec l’étiquette de référence (et ne contient donc aucune erreur). Les résultats pour les différentes tâches, présentés dans la seconde partie de la Table 1, montre que l’impact de la propagation des erreurs (entre 0,3 et 5 points suivant les tâches) est loin d’être négligeable et la propagation des erreurs doit effectivement être contrôlée.

Les résultats présentés Table 1 ont été obtenus avec l’oracle conservant l’ambiguïté. Les performances obtenues par l’oracle levant l’ambiguïté sont nettement moins bonnes : ce dernier obtient un score de 90,01% (-2,27%) sur la tâche de prédiction de la structure prosodique et un score de 91,80% (-1,52%) sur la tâche de reconnaissance des phonèmes.

5 Conclusion

Nous avons présenté dans ce travail un nouveau formalisme pour la prédiction d’objets structurés, comme les arbres et les séquences. Ce formalisme, qui consiste à voir l’apprentissage structuré comme un processus incrémental au cours duquel la sortie est progressivement construite, permet d’inscrire l’apprentissage structuré dans le cadre de l’apprentissage par renforcement. Grâce à ce lien, nous avons pu introduire une méthode théoriquement bien justifiée pour apprendre des méthodes d’inférence approchée et ainsi proposer des méthodes d’étiquetage de séquences rapides et capables de prendre en compte des dépendances riches. Cette approche est validée par des résultats équivalents ou supérieurs aux méthodes état de l’art sur quatre tâches variées du TAL.

L’apprentissage structuré incrémental est un cadre d’apprentissage général qui n’est pas limité à la prédiction de séquences

et nous envisageons d'appliquer les méthodes d'apprentissage décrites dans ce travail à des problèmes plus complexes comme l'analyse en dépendances ou l'apprentissage des modèles de séquences factorisés capables de prédire, de manière jointes, plusieurs étiquettes pour chaque observations. Nous envisageons également d'approfondir les résultats théoriques offertes par les méthodes d'apprentissage incrémental afin de mieux comprendre les principes sur lesquels les méthodes de prédiction *easy first*.

Références

- ABBEEL P. & NG A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-first International Conference on Machine Learning, ICML '04*, p. 1–, New York, NY, USA : ACM.
- CHIANG D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, **33**(2), 201–228.
- COLLINS M. & ROARK B. (2004). Incremental parsing with the perceptron algorithm. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, p. 111–118, Barcelona, Spain.
- DAUMÉ III H., LANGFORD J. & MARCU D. (2009). Search-based structured prediction. *Machine Learning Journal*.
- DAUMÉ III H. & MARCU D. (2005). Learning as search optimization : approximate large margin methods for structured prediction. In *ICML '05 : Proceedings of the 22nd international conference on Machine learning*, p. 169–176, New York, NY, USA : ACM Press.
- FINKEL J. R., MANNING C. D. & NG A. Y. (2006). Solving the problem of cascading errors : Approximate bayesian inference for linguistic annotation pipelines. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, p. 618–626, Sydney, Australia.
- GESMUNDO A. & HENDERSON J. (2014). Undirected machine translation with discriminative reinforcement learning. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, p. 10–19, Gothenburg, Sweden : Association for Computational Linguistics.
- GOLDBERG Y. & ELHADAD M. (2010). An efficient algorithm for easy-first non-directional dependency parsing. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 742–750.
- GOLDBERG Y. & NIVRE J. (2012). A dynamic oracle for arc-eager dependency parsing. In *Proceedings of COLING 2012*, p. 959–976 : The COLING 2012 Organizing Committee.
- GOLDBERG Y. & NIVRE J. (2013). Training deterministic parsers with non-deterministic oracles. *Transactions of the Association of Computational Linguistics*, **1**, 403–414.
- LANGFORD J. & ZADROZNY B. (2005). Relating reinforcement learning performance to classification performance. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513.
- LIU T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, **3**(3), 225–331.
- MA J., XIAO T., ZHU J. & REN F. (2012). Easy-first chinese pos tagging and dependency parsing. In *Proceedings of COLING 2012*, p. 1731–1746 : The COLING 2012 Organizing Committee.
- PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.
- ROSS S. & BAGNELL J. A. D. (2010). Efficient reductions for imitation learning. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- SEJNOWSKI T. J. & ROSENBERG C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, **1**, 145–168.
- SHEN L., SATTI G. & JOSHI A. (2007). Guided learning for bidirectional sequence classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, p. 760–767, Prague, Czech Republic.
- SUTTON R. & BARTO A. (1998). *Reinforcement learning : an introduction*. MIT Press.
- YAMADA H. & MATSUMOTO Y. (2003). Statistical dependency analysis with support vector machines. In *Proceedings of IWPT*.

Stratégies de sélection des exemples pour l'apprentissage actif avec des champs aléatoires conditionnels

Vincent Claveau Ewa Kijak

IRISA – CNRS – Univ. Rennes 1, Campus de Beaulieu, 35042 Rennes cedex

Vincent.Claveau@irisa.fr, Ewa.Kijak@irisa.fr

Résumé. Beaucoup de problèmes de TAL sont désormais modélisés comme des tâches d'apprentissage supervisé. De ce fait, le coût des annotations des exemples par l'expert représente un problème important. L'apprentissage actif (*active learning*) apporte un cadre à ce problème, permettant de contrôler le coût d'annotation tout en maximisant, on l'espère, la performance de la tâche visée, mais repose sur le choix difficile des exemples à soumettre à l'expert. Dans cet article, nous examinons et proposons des stratégies de sélection des exemples pour le cas spécifique des champs aléatoires conditionnels (*Conditional Random Fields*, CRF), outil largement utilisé en TAL. Nous proposons d'une part une méthode simple corrigeant un biais de certaines méthodes de l'état de l'art. D'autre part, nous détaillons une méthode originale de sélection s'appuyant sur un critère de respect des proportions dans les jeux de données manipulés. Le bien-fondé de ces propositions est vérifié au travers de plusieurs tâches et jeux de données, incluant reconnaissance d'entités nommées, *chunking*, phonétisation, désambiguïsation de sens.

Abstract.

Strategies to select examples for Active Learning with Conditional Random Fields

Nowadays, many NLP problems are modeled as supervised machine learning tasks. Consequently, the cost of the expertise needed to annotate the examples is a widespread issue. Active learning offers a framework to that issue, allowing to control the annotation cost while maximizing the classifier performance, but it relies on the key step of choosing which example will be proposed to the expert.

In this paper, we examine and propose such selection strategies in the specific case of Conditional Random Fields (CRF) which are largely used in NLP. On the one hand, we propose a simple method to correct a bias of certain state-of-the-art selection techniques. On the other hand, we detail an original approach to select the examples, based on the respect of proportions in the datasets. These contributions are validated over a large range of experiments implying several tasks and datasets, including named entity recognition, chunking, phonetization, word sens disambiguation.

Mots-clés : CRF, champs aléatoires conditionnels, apprentissage actif, apprentissage semi-supervisé, test statistique de proportion.

Keywords: CRF, conditional random fields, active learning, semi-supervised learning, statistical test of proportion.

1 Introduction

De nombreuses tâches de TAL reposent désormais sur des approches d'apprentissage artificiel supervisé. Parmi les techniques couramment employées, les champs aléatoires conditionnels (*Conditional Random Fields*, CRF) ont montré d'excellentes performances pour tout ce qui relève de l'annotation de séquences (*tagging*, reconnaissance d'entités nommées et extraction d'information, translittération...). Cependant, comme pour tous les problèmes supervisés, le coût d'annotation des séquences pour entraîner les modèles est un critère important à considérer. Pour des problèmes simples, comme l'étiquetage en parties-du-discours, des études ont montré que ce coût est relativement faible (Garrette & Baldridge, 2013), mais la plupart des problèmes cités précédemment nécessitent au contraire un très grand nombre d'annotations (cf. section 5.2).

Pour limiter ce coût, les approches semi-supervisées exploitent, en plus des exemples annotés, des exemples non-annotés qui sont eux plus facilement disponibles. Parmi ces approches, l'apprentissage actif (*Active learning*) permet à l'expert d'annoter des exemples supplémentaires de manière itérative, contrôlant ainsi le compromis coût d'annotation/performance du classifieur. Un classifieur peut ainsi être appris ou amélioré à chaque itération, et peut servir à guider le choix des prochains exemples à annoter. Dans cet article, nous nous intéressons à ce problème d'apprentissage actif, et plus précisément au problème de la sélection des exemples qui sont proposés à l'expert, dans le cas particulier des CRF.

Il existe bien sûr déjà de nombreuses méthodes de sélection, soit génériques, soit propres aux CRF. Dans cet article, nous montrons que certaines méthodes très classiques de l'état de l'art comportent un biais tendant à favoriser le choix d'exemples longs, et donc coûteux à annoter. Nous proposons une technique simple pour lever ce biais. Mais notre contribution principale porte sur la proposition d'une technique de sélection originale, utilisant la représentation qui est faite des données par les CRF, et s'appuyant sur un critère de respect des proportions d'attributs dans les jeux de données. Ces différentes propositions sont évaluées expérimentalement sur plusieurs jeux de données et tâches classiques des CRF.

L'article est structuré de la façon suivante. La section 2 rappelle quelques notions de bases autour des CRF et de l'apprentissage actif, et présente des travaux connexes ainsi que les données servant à nos expérimentations. Nous revisitons ensuite en section 3 certaines techniques habituellement utilisées pour en montrer les limites. Dans la section 4, nous présentons une nouvelle technique pour la sélection des exemples à annoter. La section 5 présente et commente les expérimentations menées, et la dernière section présente quelques perspectives ouvertes par ce travail.

2 Contexte et état de l'art

2.1 Notions de base

Les champs aléatoires conditionnels ou *Conditional Random Fields* (Lafferty *et al.*, 2001) sont des modèles graphiques non dirigés qui représentent la distribution de probabilités d'annotations y sur des observations x . Ils sont très employés en TAL ; x est alors une séquence de lettres ou de mots et y la séquence correspondante de labels. Dans ce cadre, la probabilité conditionnelle $P(y|x)$ se définit à travers la somme pondérée de fonctions dites caractéristiques (*feature functions*) f_j :

$$P(y|x, \theta) = \frac{1}{Z_\lambda(x)} \exp \left(\sum_j \sum_t \lambda_j f_j(x, y_t, y_{t-1}, t) \right)$$

où $Z_\lambda(x)$ est un facteur de normalisation et θ est le vecteur des poids λ_j . Les fonctions caractéristiques sont souvent binaires, renvoyant 1 lorsqu'une certaine combinaison de labels et d'attributs des observations est satisfaite, 0 sinon. Elles sont appliquées à chaque position t de la séquence et leur poids λ_j reflète leur importance pour déterminer la classe. Il est important de noter qu'en pratique, ce n'est pas tout le vecteur x qui est considéré mais juste une certaine combinaison d'attributs sur les objets autour de la position t . Ces combinaisons sont définies par l'utilisateur, le plus souvent indirectement par un ensemble de patrons $\{\text{Pat}_i\}$ dont les réalisations à chaque position t de chaque séquence x ($\text{Pat}_i(x, t)$), ajoutées aux informations de labels correspondantes (y_{t-1} et y_t), définissent l'ensemble des fonctions possibles.

L'apprentissage d'un CRF consiste à estimer les poids λ_j à partir de données dont les labels sont connus. On cherche alors

le vecteur θ qui maximise la log-vraisemblance \mathcal{L} du modèle sur les m séquences annotées :

$$\mathcal{L}(\theta) = \sum_m \log P_{\theta}(y^{(m)} | x^{(m)}, \theta)$$

En pratique, on ajoute souvent des contraintes sur la taille du vecteur θ pour éviter le sur-apprentissage. Ce problème d'optimisation peut être résolu en utilisant des algorithmes de type quasi-Newton, comme L-BFGS (Schraudolph *et al.*, 2007). Une fois le CRF appris, l'application du CRF à des nouvelles données consiste à trouver, pour une séquence d'observations x , la séquence de labels la plus probable, notée y^* dans la suite de cet article, par exemple avec un algorithme de Viterbi.

Grâce à leur capacités à prendre en compte l'aspect séquentiel et les descriptions riches des textes, les CRF ont été utilisés avec succès dans de nombreuses tâches s'exprimant comme des problèmes d'annotation. Ils sont ainsi devenus des outils standard pour l'extraction d'information, la reconnaissance d'entités nommées, le tagging, etc. (Wang *et al.*, 2006; Pranjali *et al.*, 2006; Constant *et al.*, 2011; Raymond & Fayolle, 2010, *inter alia*).

2.2 Apprentissage semi-supervisé

L'apprentissage semi-supervisé consiste à utiliser conjointement des données annotées (notées \mathcal{T} dans la suite de l'article) et des données non-annotées (\mathcal{N}). Son but est de réduire le nombre d'annotations et donc le coût de l'annotation, et/ou d'améliorer les performances du classifieur à coût d'annotation identique. Différentes approches d'apprentissage semi-supervisé ont déjà été explorées pour les CRF. Plusieurs travaux utilisent les données non étiquetées directement dans l'apprentissage du modèle en modifiant l'expression de l'entropie. Cette modification rend la fonction objectif non-concave et nécessite donc d'adapter la procédure d'apprentissage.

Une autre famille de travaux a consisté à adapter les procédures d'apprentissage et de décodage pour que les CRF soient capables d'exploiter des connaissances sur les séquences autres que l'annotation complète de la séquence. Il peut s'agir par exemple d'annotations partielles des séquences, c'est-à-dire dont les étiquettes ne portent que sur quelques mots (Salakhutdinov *et al.*, 2003). Il peut également s'agir de connaissances a priori sur la distribution des étiquettes sachant certains attributs (Mann & McCallum, 2008).

Bien que cela ne relève pas strictement de l'apprentissage semi-supervisé, il convient également d'évoquer les travaux utilisant des techniques annexes sur les données non annotées pour améliorer l'apprentissage sur les données annotées. Par exemple, (Miller *et al.*, 2004) et (Freitag, 2004) font du *clustering* sur les données non-annotées pour proposer de nouveaux attributs – en l'occurrence, des classes de mots – ensuite utilisés pour mieux décrire les données annotées. Dans cette veine, il convient également de citer les travaux de (Ando & Zhang, 2005) et ceux de (Smith & Eisner, 2005). Ces derniers exploitent une proximité entre une séquence annotée et d'autres séquences pour influencer sur l'estimation des paramètres du CRF. Bien que là encore ces travaux ne se placent pas dans le même cadre que nos travaux, ceux-ci partagent néanmoins l'idée d'exploiter la ressemblance des séquences vues comme des ensembles d'attributs.

2.3 Apprentissage actif

Dans notre cas, nous nous plaçons dans un cadre spécifique d'apprentissage semi-supervisé qualifié d'apprentissage actif (*active learning*). Son principe est que la supervision est effectuée par l'expert de manière itérative et interactive (Settles, 2010). Cela est souvent mis en œuvre dans un algorithme dont les grandes lignes sont les suivantes :

1. apprendre un classifieur à partir de \mathcal{T}
2. appliquer le classifieur à \mathcal{N}
3. sélectionner des exemples de \mathcal{N}
4. annoter ces exemples et les ajouter à \mathcal{T}
5. retourner en 1

Ce processus est ainsi répété jusqu'à ce qu'un critère d'arrêt soit atteint. Ce critère peut être que le coût de l'annotation maximal est atteint, que la performance du classifieur minimale est atteinte, ou que \mathcal{N} est vide.

Le point crucial de ces algorithmes d'apprentissage actif est l'étape 3 de sélection des exemples à faire annoter à l'expert. On cherche à choisir les exemples les plus bénéfiques pour l'apprentissage, ceux permettant d'obtenir les meilleures performances de classification, et pour ce faire, on s'appuie souvent sur l'étape 2. Beaucoup de travaux ont été proposés sur ce

point, notamment dans le domaine du TAL (Olsson, 2009) où ces problèmes d’annotation sont courants. Indépendamment des classifieurs utilisés, plusieurs familles de stratégies ont été proposées. La plus courante est la sélection par incertitude dans laquelle on utilise le résultat de l’étape de 2 pour choisir les exemples pour lesquels le classifieur courant est le moins sûr (cf. section 3). Un défaut connu de cette approche est qu’au début du processus, quand il y a peu d’exemples annotés, les mesures d’incertitude du classifieur ne sont pas fiables.

Une autre famille très usuelle est la sélection par comité. Son principe est d’apprendre non pas un mais plusieurs classifieurs à l’étape 1, de les appliquer à \mathcal{N} , et de sélectionner les exemples sur lesquels ils sont le plus en désaccord. Cette approche est souvent mise en œuvre par des techniques de *bagging* et/ou de *boosting* (Abe & Mamitsuka, 1998), ou par des représentations complémentaires des données sur lesquelles sont appris des classifieurs différents (Pierce & Cardie, 2001). En plus du coût calculatoire plus important généré par ces apprentissages multiples, ces techniques souffrent du même problème que la sélection par incertitude : les classifieurs sont peu fiables dans les premiers tours de l’itération avec $|\mathcal{T}|$ petit.

Une dernière famille usuelle est la sélection basée sur la modification attendue du modèle. Le principe est ici de sélectionner l’exemple qui impacterait le plus le modèle, en supposant que cet impact résulterait en une amélioration des performances. L’intuition sous-jacente est que l’exemple choisi couvre des cas non traités par les exemples de \mathcal{T} . La mise en œuvre de cette approche dépend beaucoup du classifieur utilisé. Settles & Craven (2008) a proposé plusieurs variantes de cette approche pour les CRF, dont seulement l’une, appelée *Information Density*, a donné quelques résultats positifs. Celle-ci repose simplement sur le choix de la séquence dans \mathcal{N} la plus différente des séquences de \mathcal{T} . Pour évaluer cette différence, les auteurs représentent les séquences comme un vecteur des combinaisons d’attributs capturés par les fonctions caractéristiques. Les labels des séquences de \mathcal{N} étant inconnus, il faut bien noter que ces sont les attributs sur x qui sont considérés. La séquence la plus dissimilaire est simplement définie comme celle ayant le cosinus moyen avec les séquences de \mathcal{T} le plus faible.

Ces derniers travaux sont les plus proches de ceux que nous présentons dans cet article. Nous en reprenons d’ailleurs en partie la représentation des séquences, vues comme des ensembles d’attributs, bien que le critère que nous proposons se veut plus performant que celui proposé dans ces travaux (cf. section 4). Par ailleurs, la méthode d’évaluation utilisée par Settles & Craven (2008) ne rend pas compte correctement de l’effort d’annotation fourni à chaque itération : les auteurs évaluent les performances en fonction des séquences, sans considérer que certaines peuvent être beaucoup plus longues que d’autres. Pour notre part, l’effort d’annotation est mesuré en terme de mots annotés, ce qui a des conséquences sur les stratégies de sélection classiques testées par ces auteurs (cf. section suivante).

2.4 Contexte expérimental

Dans la suite de cet article, nous allons valider nos propositions de sélection des séquences sur différentes tâches pour lesquelles les CRF sont classiquement utilisés. Nous décrivons brièvement ces tâches et ces données ci-dessous ; pour plus de détails, le lecteur intéressé peut se reporter aux références indiquées.

Nous utilisons le jeu de données de la tâche de reconnaissance d’entités nommées de la campagne ESTER (Gravier *et al.*, 2005). Il contient des transcriptions d’émissions de radio en français, soit 55 000 groupes de souffle, dont les entités nommées sont annotées selon 8 classes (personne, lieu, temps...). Le jeu CoNLL2002 contient les données utilisées pour la tâche de reconnaissance d’entités nommées en néerlandais proposée dans le cadre de CoNLL 2002 (Tjong Kim Sang, 2002). Il contient 4 labels d’entités différents et nous utilisons 14 000 séquences (phrases) dans les expériences rapportées dans la suite de l’article. Le jeu CoNLL2000 est composé de textes de journaux en anglais annotés en chunks (Tjong Kim Sang & Buchholz, 2000). Il contient environ 11 000 phrases et 4 classes (3 types de chunks et un label ‘autre’). Nous utilisons également les données de désambiguïsation de sens de SensEval-2 (Edmonds & Cotton, 2001). Ce jeu de données porte sur la désambiguïsation de *hard*, *line*, *serve*, *interest*, chacun des sens étant représenté par un label différent. Il contient environ 16 000 phrases. Une tâche un peu différente sur laquelle nous nous testons est celle de la phonétisation de mots isolés en anglais fournis par les données Nettealk. Le but est de transcrire ces mots dans un alphabet phonétique spécifique. Cette tâche est donc vue comme une tâche d’annotation lettre par lettre. On a ainsi 18 000 mots et 52 labels différents correspondant à l’alphabet phonétique. Une étape préliminaire des données a consisté à aligner les mots avec leur phonétisation et donc à introduire le cas échéant des symboles ‘vide’.

Les données sont décrites de manière habituelle pour ces tâches, avec les parties du discours, lemmes, information sur la présence de majuscule, etc., et le schéma d’annotation BIO est adopté lorsque nécessaire (ESTER, CoNLL2002, CoNLL2000). Tous ces corpus de données ont été divisés en neuf dixièmes pour l’entraînement (ensemble \mathcal{T} et \mathcal{N}) et

un dixième pour l'évaluation des performances. Dans la plupart des cas, la mesure de performance utilisée est le taux de précision par mot (label correct ou non), sauf pour la tâche de phonétisation, où c'est le taux de précision par séquence (le mot doit être entièrement et correctement phonétisé). Cette mesure est réalisée à chaque itération et rapportée à l'effort d'annotation, c'est-à-dire au nombre de mots auxquels l'expert a ajouté le label.

L'implémentation des CRF que nous utilisons est WAPITI (Lavergne *et al.*, 2010), avec ses paramètres par défaut sauf si indiqué autrement. Il convient de noter que des tests avec d'autres réglages (algorithmes d'optimisation, normalisation...), non rapportés dans l'article, ne modifient pas les conclusions présentées.

3 Sélection par incertitude

Comme nous l'avons vu, une solution classique pour la sélection des exemples à annoter à chaque itération est de proposer à l'oracle ceux pour lesquels le classifieur appris à l'issue de l'itération précédente est le moins sûr. Avec des CRF, cela se traduit par choisir la séquence x sur la base des probabilités $P(y|x; \theta)$.

3.1 Confiance minimale et entropie de séquence

Parmi les différentes façons de procéder, Settles & Craven (2008) montre que deux stratégies de cette famille obtiennent de bons résultats dans la plupart des cas. Il s'agit de la sélection par confiance minimale et de la sélection par entropie de séquence. La première consiste simplement à choisir dans \mathcal{N} la séquence obtenant la probabilité minimale avec le modèle courant :

$$x = \operatorname{argmin}_{x \in \mathcal{N}} P(y^*|x, \theta)$$

La méthode par entropie consiste à choisir la séquence x de plus grande entropie sur l'ensemble des labels possibles y de cette séquence :

$$x = \operatorname{argmax}_{x \in \mathcal{N}} \left(- \sum_y P(y|x, \theta) \log P(y|x, \theta) \right)$$

3.2 Biais de longueur

L'un des problèmes de ces approches de l'état-de-l'art est qu'elles ont tendance à choisir les séquences les plus longues, celles-ci ayant des probabilités souvent plus faibles que des séquences courtes. Or, le coût d'annotation est proportionnel à la longueur des séquences, c'est donc un comportement potentiellement indésirable si l'on cherche à maximiser la performance pour un coût d'annotation minimal. Pour illustrer cela, nous reportons dans les figures 1 et 2 la longueur des séquences du jeu de données ESTER selon leur probabilité donnée par un modèle entraîné sur respectivement 20 et 10 000 séquences choisies aléatoirement. Dans les deux cas, on observe bien le biais attendu : en moyenne, la probabilité de la séquence donnée par le CRF est linéairement corrélée à sa longueur. Cela est notamment plus marqué quand le modèle est appris sur peu de séquences (figure de gauche). Or c'est justement le cas pour les premières itérations de l'apprentissage actif. Ce critère est donc particulièrement peu adapté en début d'apprentissage actif.

À l'inverse, une normalisation brutale par la longueur des séquences a tendance à privilégier les séquences très courtes n'apportant pas d'informations utiles à l'apprentissage.

3.3 Normalisation

Sur la base des constatations précédentes, il semble important de normaliser selon la longueur des séquences. Il serait possible d'utiliser les coefficients des droites de régression, mais comme on peut le constater sur les figures précédentes, celles-ci décrivent un comportement globale du nuage de points mais ne sont pas forcément adaptées pour décrire les points autour d'une longueur de séquence fixée.

Nous proposons à la place une méthode de normalisation adaptative plus locale, s'appuyant sur l'observation de la probabilité moyenne des séquences pour une longueur donnée. Cela revient à étudier le comportement du nuage de point, c'est-à-dire la répartition des probabilités, sur une tranche verticale des figures précédentes. Nous proposons pour cela

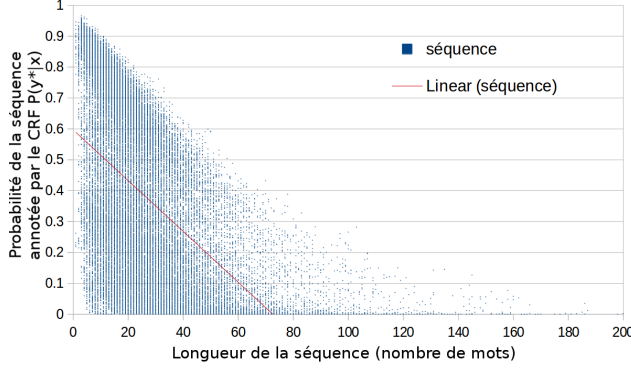


FIGURE 1 – Probabilités des séquences ($P(y^*|x)$) du jeu de données ESTER selon leur longueur obtenues avec un modèle appris sur 20 séquences, et droite de régression linéaire.

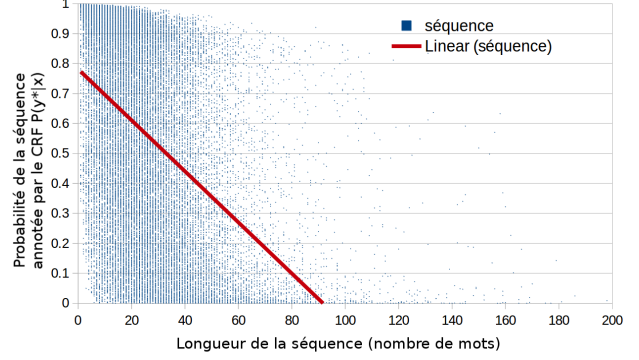


FIGURE 2 – Probabilités des séquences ($P(y^*|x)$) du jeu de données ESTER selon leur longueur, obtenues avec un modèle appris sur 10 000 séquences, et droite de régression linéaire.

une méthode de normalisation s’inspirant des méthodes d’estimation par fenêtres de Parzen (Parzen, 1962; Wasserman, 2005). L’idée sous-jacente est que pour une longueur de séquence fixée (ou à plus ou moins ϵ), les scores de probabilités devraient être distribués uniformément entre 0 et 1. Pour une séquence x de \mathcal{N} de longueur l , nous estimons la moyenne $\hat{\mu}_l$ et l’écart-type $\hat{\sigma}_l$ des probabilités obtenues à cette itération sur toutes les séquences de \mathcal{N} de longueur $l \pm \epsilon$, c’est-à-dire de l’ensemble $\{P(y^*|x') \mid x' \in \mathcal{N}, |x'| = |x| \pm \epsilon\}$. Ces valeurs sont alors utilisées pour centrer et réduire les probabilités utilisées dans les stratégies de sélection précédentes. Par exemple, pour la sélection par confiance minimale, on a :

$$x = \operatorname{argmin}_{x \in \mathcal{N}} \left(\frac{P(y^*|x, \theta) - \hat{\mu}_l}{\hat{\sigma}_l} \right)$$

Cela doit ainsi permettre, pour chaque longueur de clause considérée, d’améliorer la dispersion des probabilités des séquences de cette longueur, et donc d’annuler le biais de longueur des séquences observé précédemment.

En pratique, dans les expériences rapportées en section 5, les séquences de longueur comparables ne sont pas trouvées à ϵ près mais par voisinage : la moyenne est calculée sur un nombre fixé de séquences dont la longueur s’approche le plus de celle visée. Cette approche inspirée de l’estimation par k-plus-proches voisins permet de traiter les cas de séquences *outlier* aux longueurs très différentes pour lesquelles un voisinage défini à ϵ près ne couvrirait aucune autre séquence.

4 Représentativité des fonctions caractéristiques

La proposition principale de cet article est de considérer que la distribution des attributs, tels que capturés par les fonctions caractéristiques, peut guider la sélection des exemples à faire annoter dans un cycle d’apprentissage actif. Pour étayer cette intuition, nous étudions tout d’abord dans la sous-section 4.1 comment ces attributs sont distribués en terme de fréquence et en terme d’utilisation dans les modèles. La sous-section 4.2 propose une méthode originale pour sélectionner les séquences à faire annoter en se basant sur ces considérations.

4.1 Étude préliminaire

Les fonctions caractéristiques encodent les relations entre la description des séquences et les classes. Il est intéressant d’en observer les fréquences d’apparition dans les données, mais aussi de voir quelles sont parmi elles les fonctions effectivement utilisées pour la prédiction. Pour cela, nous calculons la distribution des occurrences de toutes les fonctions caractéristiques constructibles sur les données ESTER, soit plus formellement :

$$\operatorname{occ}(f_j) = |\{f_j(x^{(m)}, y_{t-1}^{(m)}, y_t^{(m)}, t) = 1 \mid \text{tout exemple } m, \text{ toute position } t\}|$$

Nous entraînons d’autre part également deux modèles sur l’ensemble des données ESTER (supervision complète) afin d’étudier les fonctions effectivement utilisées pour la prédiction dans les modèles. Pour chaque modèle, nous extrayons

donc de l'ensemble des fonctions caractéristiques celles ayant un poids $|\lambda_j| > 0$. Les paramètres d'apprentissage, notamment la stratégie de normalisation en L1 ou L2, influent beaucoup sur le nombre de fonctions caractéristiques de poids non nuls. Nous entraînons donc un modèle avec une normalisation L1 et un autre avec une normalisation *elastic-net* (mariant également L1 et L2). Nous rapportons dans la figure 3 ces trois distributions : celle de l'ensemble des fonctions caractéristiques possibles à partir des données et celles effectivement utilisées dans les modèles appris L1/L2 et L1 sur ces données.

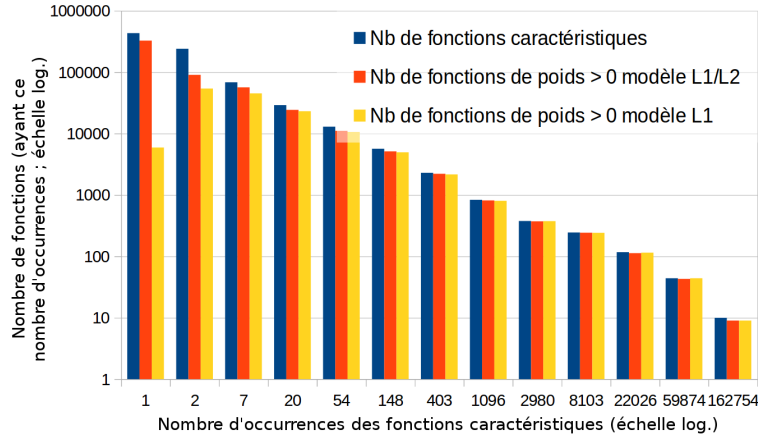


FIGURE 3 – Distribution des fonctions caractéristiques (nombre de fonctions selon leur nombre d'occurrences ; échelle logarithmique sur les deux axes) sur les données ESTER et distribution des fonctions utilisées dans le modèle.

On observe que ces trois distributions sont très similaires sauf pour les fonctions caractéristiques les plus rares, notamment avec le modèle L1. La plupart des combinaisons d'attributs/labels des données apparaissent donc comme utiles (car de poids $|\lambda_j| > 0$) pour les prédictions dans nos deux modèles. Cela signifie que les modèles CRF que nous utilisons exploitent une très grande majorité des combinaisons attributs/labels présentes dans les données, que ces combinaisons soient très fréquentes ou plus rares (à l'exception des très rares configurations pour les modèles L1), et de manière proportionnelle à leur fréquence dans les données. Pour construire un jeu d'entraînement plus petit mais menant à des modèles ayant des caractéristiques similaires, il semble important d'offrir le maximum de variété de configurations en respectant ces proportions, c'est-à-dire en respectant au mieux la distributions des combinaisons d'attributs/labels du jeu d'entraînement complet.

Dans le cas semi-supervisé, la majorité des données n'est pas annotée. Il est donc important de vérifier si les conclusions précédentes sont également vraies en ne regardant pas les labels. On examine donc la distribution des fonctions caractéristiques sans considération des labels, c'est-à-dire uniquement en regardant les attributs relevant de x . La figure 4 illustre ainsi la distribution non plus exactement des fonctions caractéristiques, mais de leurs occurrences quel que soit leur label, ce que l'on note f_j^* . Formellement, on a donc :

$$\text{occ}(f_j^*) = |\{f_j(x^{(m)}, y_1, y_2, t) = 1 \mid \text{tout exemple } m, \text{ toute position } t, \text{ toutes classes } y_1, y_2\}|$$

On observe les mêmes tendances que précédemment. Ces différentes expériences suggèrent l'importance d'avoir un jeu d'entraînement varié et représentatif de l'ensemble des combinaisons d'attributs définis par les fonctions caractéristiques. C'est sur la base de ce critère que nous proposons la stratégie de sélection présentée ci-après.

4.2 Test de proportion

À chaque itération de l'apprentissage actif, on souhaite avoir l'ensemble d'entraînement le plus représentatif des données que l'on va traiter. Par représentatif, on veut dire dont la distribution des séquences, telles que vues par le CRF via les fonctions caractéristiques, soit la plus proche de celles de $\mathcal{T} \cup \mathcal{N}$. On se place donc comme précédemment dans un cadre où chaque séquence est vue comme l'ensemble des fonctions caractéristiques qu'elle permet de générer, labels non compris.

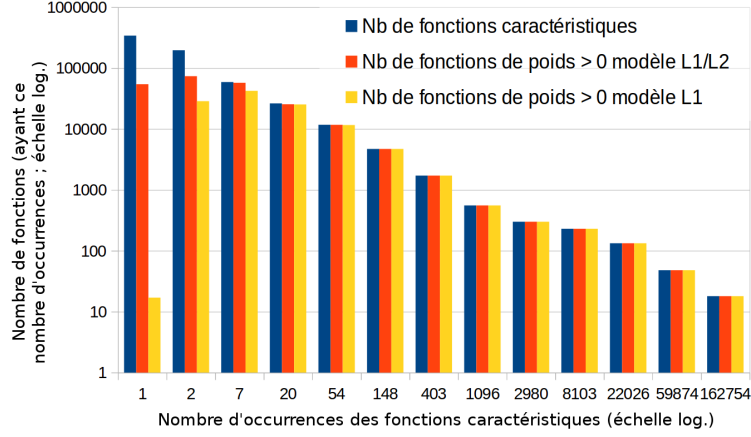


FIGURE 4 – Distribution des fonctions caractéristiques sans indication de label (nombre de fonctions selon leur nombre d’occurrences ; échelle logarithmique sur les deux axes) sur les données ESTER et distribution des fonctions utilisées dans le modèle sans indication de classe.

Pour choisir la séquence x à ajouter à l’ensemble d’entraînement à chaque itération (et donc à faire annoter par l’oracle), nous voulons donc mesurer combien le jeu d’entraînement résultant $\mathcal{T} \cup \{x\}$ se rapproche de l’ensemble des données à notre disposition (annotées ou non, i.e. $\mathcal{T} \cup \mathcal{N}$). Pour cela, pour chaque fonction caractéristique possible, nous proposons d’examiner simplement si la proportion de cette fonction observée dans l’échantillon $\mathcal{T} \cup \{x\}$ est comparable à celle de l’échantillon $\mathcal{T} \cup \mathcal{N}$. Ces échantillons ne sont pas indépendants, mais peuvent être considérés comme tel dès lors que $|\mathcal{N}| \gg |\mathcal{T}|$, ce qui est assuré dans tous les cas aux premières itérations de l’apprentissage actif.

Plus formellement, nous effectuons un test statistique de proportion entre les deux échantillons $\mathcal{T} \cup \{x\}$ et $\mathcal{T} \cup \mathcal{N}$, respectivement notés 1 et 2 et de taille n_1 et n_2 . Soit $\hat{p}_1^j = r_1^j/n_1$ l’estimateur de proportion d’occurrences d’une certaine fonction caractéristique f_j apparaissant r_1^j fois dans l’échantillon 1, et $\hat{p}_2^j = r_2^j/n_2$ l’estimateur de proportion de la même fonction pour l’échantillon 2. On peut alors calculer le z -score suivant :

$$z_{j,x} = \frac{\hat{p}_1^j(f_j) - \hat{p}_2^j(f_j)}{\sqrt{\hat{p}^j * (1 - \hat{p}^j) * (1/n_1 + 1/n_2)}} \quad \text{avec} \quad \hat{p}^j = \frac{r_1^j + r_2^j}{n_1 + n_2}$$

Ce z -score suit une loi normale centrée réduite, ce qui nous permet de calculer la probabilité $P(z_{j,x})$ d’observer une telle différence de proportion entre nos deux échantillons. Dans notre cadre, une probabilité élevée traduit intuitivement que l’échantillon 1 contient une proportion comparable à celle de l’échantillon 2 de la fonction caractéristique visée f_j .

Il faut bien sûr combiner ces probabilités pour toutes les fonctions caractéristiques. Pour cela, nous faisons une hypothèse simplificatrice qui est de considérer que les observations des fonctions caractéristiques sont indépendantes. Même s’il est évident que cette hypothèse est invalidée dans la plupart des cas, elle nous permet d’estimer simplement la probabilité globale de l’échantillon sur l’ensemble des fonctions caractéristiques comme le produit des $P(z_{j,x})$ pour toutes les fonctions caractéristiques f_j . Finalement, le choix de la séquence à ajouter à l’ensemble des séquences annotées est donc celle maximisant cette probabilité :

$$x^* = \operatorname{argmax}_{x \in \mathcal{N}} \prod_j P(z_{j,x})$$

5 Expérimentations

Dans cette section, nous comparons expérimentalement les différentes stratégies évoquées de sélection des exemples pour l’apprentissage actif. Celles-ci sont rappelées ci-dessous, et les courbes d’apprentissage obtenues sont présentées dans la sous-section 5.2.

5.1 Contexte

Les stratégies de sélection de que nous testons sont d’une part celles de la littérature, qui nous servent ainsi de point de comparaison. Il s’agit de la sélection par confiance minimale, entropie, et *information density*. Nous ajoutons également une stratégie *baseline* consistant à choisir les séquences au hasard (*random*). Nous testons d’autre part également notre stratégie de normalisation sur la confiance minimale (confiance minimale normalisée) et la méthode basée sur la proportion. Nous ne rapportons pas de résultats avec des techniques de sélection par comité, celles-ci obtenant des résultats plus faibles que les précédentes dans la quasi-totalité des cas (Settles & Craven, 2008).

Toutes ces méthodes sont testées dans les mêmes conditions (paramètres du CRF, patrons...). À l’initialisation, une séquence est tirée au hasard pour servir de premier exemple (le même pour toutes les méthodes de sélection). À chaque itération, un unique exemple est choisi pour être annoté et le classifieur est ré-entraîné sur l’ensemble des données annotées (il ne s’agit donc pas d’une mise à jour du CRF).

5.2 Résultats

Les figures 5, 6, 7 et 8 présentent les courbes d’apprentissage sur nos différents jeux de données. La performance des classifieurs appris à chaque itération est donc exprimée en fonction du coût de l’annotation cumulé de l’ensemble \mathcal{T} . Dans les figures, ce coût est rapporté sur une échelle logarithmique qui permet de bien apprécier les différents cas (peu d’annotations vs. beaucoup d’annotations).

Plusieurs observations en ressortent. D’une part, ces courbes ont des allures très différentes d’un jeu de données à l’autre. Cela s’explique par les caractéristiques des tâches et des données, impliquant que certaines soient plus facilement faisables avec de bonnes performances en peu d’annotations (CoNLL2000) ou non (CoNLL2002). On note au passage que pour certaines tâches, l’allure des courbes laisse penser que plus d’exemples améliorerait encore les performances ; les conclusions de (Garrette & Baldridge, 2013) ne sont onc pas à généraliser. Pour tous les jeux de données sauf Nettealk, les différences observées, notamment lorsque le coût d’annotation est petit, sont sensibles. Concernant Nettealk, il est plus difficile de faire ressortir une méthode de sélection meilleure que les autres. Cela s’explique certainement par la difficulté de la tâche due notamment au très grand nombre de labels possibles, et donc au très grand nombre de configurations attributs/labels possibles qui nécessite dans tous les cas un nombre extrêmement important d’exemples pour couvrir toutes ces configurations.

Deuxièmement, on observe que les trois stratégies de la littérature offrent des performances moyennes, parfois peu éloignées de la stratégie *random*. Les stratégies de confiance minimale et entropie sont même parfois nettement en deçà du hasard (SenseEval-2), visiblement pénalisées par leurs biais discutés en section 3. Ce point est important à noter puisqu’il est souvent occulté par les évaluations ne prenant en compte que le nombre de séquences, comme nous l’avons déjà souligné pour le travail de (Settles & Craven, 2008).

Troisièmement, on constate le bien fondé de notre proposition de normalisation puisque cette stratégie nous permet d’obtenir des résultats meilleurs ou identiques à la version non normalisée. Elle obtient notamment les meilleurs résultats lorsque le nombre d’annotation est important (ESTER, CoNLL2002, SenseEval-2), même si l’échelle logarithmique cache ici un peu cette longue domination.

Enfin, notre proposition de sélection basée sur le respect des proportions obtient de très bons résultats dans nos différents cas d’étude. Elle se comporte globalement mieux que les autres techniques de sélection, y compris l’*information density* dont elle se rapproche conceptuellement. On peut noter que notre stratégie apporte un gain notable lorsque le coût d’annotation considéré est petit. Ce résultat attendu s’explique par le fait que la méthode ne repose pas sur les prédictions, peu fiables à ce stade, du classifieur courant. En revanche, ce gain est moindre voire nul par rapport aux autres méthodes lorsque la quantité de données annotées devient très importante. Cela montre la limite de notre approche qui n’exploite aucune information issue du classifieur, mais permet aussi d’imaginer des stratégies mixtes dans lesquelles ces informations de classification seraient également exploitées à d’un certain nombre d’annotations.

6 Conclusions

À l’heure où la plupart des problèmes du TAL sont exprimés en tâches d’apprentissage supervisé, le coût des annotations des exemples par l’expert représente un problème important. L’apprentissage actif apporte un cadre à ce problème, per-

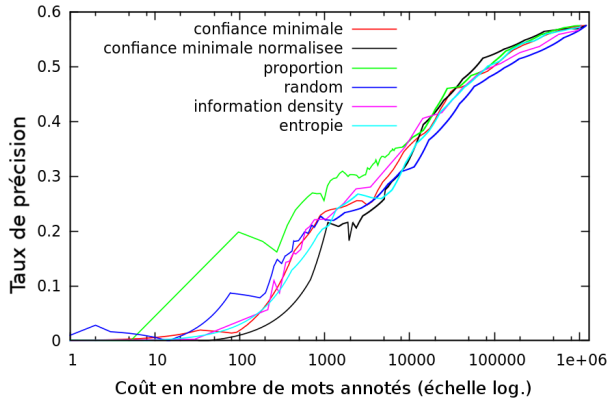


FIGURE 5 – Courbe d'apprentissage sur les données ES-TER : taux de précision selon le coût d'annotation en mots (échelle log)

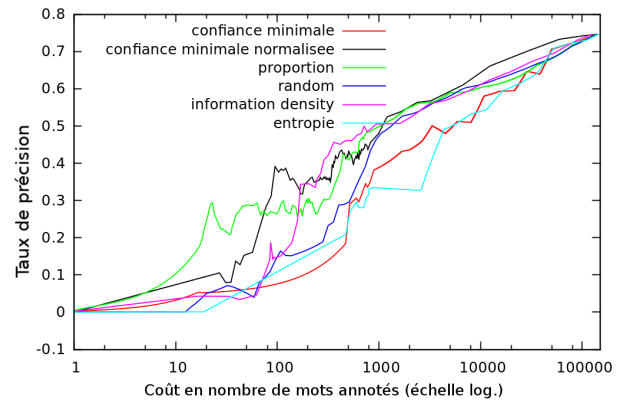


FIGURE 6 – Courbe d'apprentissage sur les données CoNLL2002 : taux de précision selon le coût d'annotation en mots (échelle log)

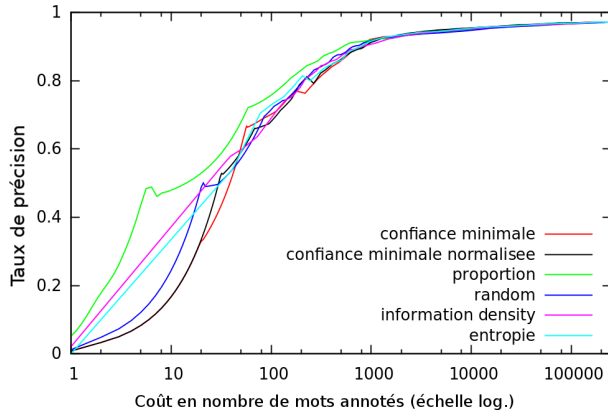


FIGURE 7 – Courbe d'apprentissage sur les données CoNLL2000 : taux de précision selon le coût d'annotation en mots (échelle log)

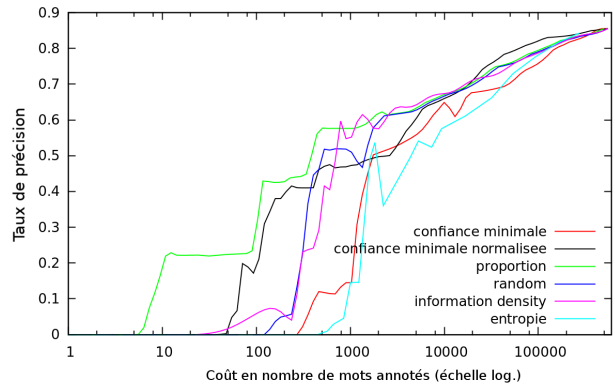


FIGURE 8 – Courbe d'apprentissage sur les données SensEval-2 : taux de précision selon le coût d'annotation en mots (échelle log)

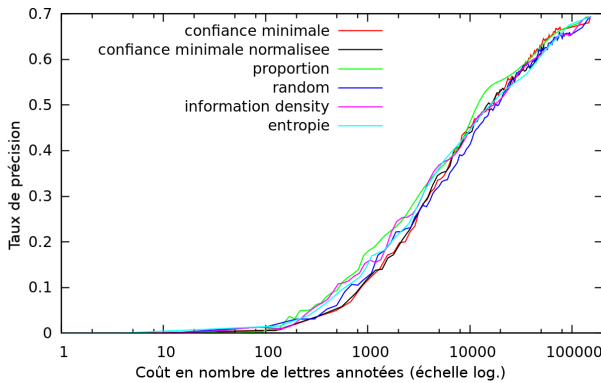


FIGURE 9 – Courbe d'apprentissage sur les données Net-talk : taux de précision (mots correctement phonétisés) selon le coût d'annotation en nombre de lettres (échelle log)

mettant de contrôler le coût d'annotation tout en maximisant, on l'espère, la performance de la tâche visée. Comme nous l'avons vu, cela est en fait largement dépendant de la stratégie de sélection des exemples. Dans cet article, nous avons examiné quelques unes de ces stratégies pour lesquelles nous avons mis en évidence un biais dégradant le ratio coût d'annotation/performance. La normalisation que nous avons proposée permet de lever ce biais de manière très simple tout en offrant un gain de performance notable. Lorsque les coûts d'annotation sont limités, la nouvelle stratégie que nous avons proposée, s'appuyant sur un critère original de proportionnalité, s'avère la plus avantageuse.

Bien sûr, beaucoup de variantes, d'améliorations et de pistes de recherche sont envisageables. Parmi celles-ci, nous souhaitons essayer de prendre en compte la dépendance entre les fonctions caractéristiques. Dans notre proposition actuelle, elles sont abusivement considérées comme indépendantes, ce qui n'est jamais le cas en pratique. Ces dépendances peuvent même être très importantes puisque les patrons permettant de construire les fonctions caractéristiques font souvent appel plusieurs fois aux mêmes éléments (lemme du mot courant, PoS du mot courant...) et que ces éléments sont eux-mêmes en relation de dépendance. Tout ceci peut donc fortement impacter l'estimation de nos probabilités et fausser finalement le choix de l'exemple.

Une autre piste prometteuse est de mélanger ces différentes techniques de sélection pour en combiner les avantages. Il est bien sûr possible de les intégrer simplement (vote, produit des scores ou des rangs...), mais il nous semble plus intéressant de viser des combinaisons plus complexes, que l'on pourrait par exemple obtenir avec des techniques d'apprentissage d'ordre (*learning to rank*) (Liu, 2009).

Enfin, dans notre cadre actuel, les séquences sélectionnées sont annotées complètement. Il serait intéressant d'étudier le cas des annotations partielles, avec les mêmes contraintes d'optimisation du ratio coût/performance, en nous inspirant par exemple des travaux de Salakhutdinov *et al.* (2003).

Références

- ABE N. & MAMITSUKA H. (1998). Query learning strategies using boosting and bagging. In *Proceedings of the Fifteenth International Conference on Machine Learning*, Madison, Wisconsin, USA : Morgan Kaufmann Publishers Inc.
- ANDO R. K. & ZHANG T. (2005). A high-performance semi-supervised learning method for text chunking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, p. 1–9, Stroudsburg, PA, USA : Association for Computational Linguistics.
- CONSTANT M., TELLIER I., DUCHIER D., DUPONT Y., SIGOGNE A. & BILLOT S. (2011). Intégrer des connaissances linguistiques dans un CRF : Application à l'apprentissage d'un segmenteur-étiqueteur du français. In *Traitement Automatique du Langage Naturel (TALN'11)*, Montpellier, France.
- EDMONDS P. & COTTON S. (2001). Senseval-2 : Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, p. 1–5 : Association for Computational Linguistics.
- FREITAG D. (2004). Trained named entity recognition using distributional clusters. In *Proceedings of the conference EMNLP*.
- GARRETTE D. & BALDRIDGE J. (2013). Learning a part-of-speech tagger from two hours of annotation. p. 138–147.
- GRAVIER G., BONASTRE J.-F., GEOFFROIS E., GALLIANO S., TAIT K. M. & CHOUKRI K. (2005). ESTER, une campagne d'évaluation des systèmes d'indexation automatique. In *Actes des Journées d'Étude sur la Parole, JEP, Atelier ESTER2*.
- LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*.
- LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.
- LIU T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, **3**(3), 225–331.
- MANN G. S. & MCCALLUM A. (2008). Generalized expectation criteria for semi-supervised learning of conditional random fields. In *Proceedings of ACL-08 : HLT*, p. 870–878, Columbus, Ohio, USA.
- MILLER S., GUINNESS J. & ZAMANIAN A. (2004). Name tagging with word clusters and discriminative training. In *Proceedings of the conference ACL*.

- OLSSON F. (2009). *A literature survey of active machine learning in the context of natural language processing*. Rapport interne Swedish Institute of Computer Science, Swedish Institute of Computer Science.
- PARZEN E. (1962). On estimation of a probability density function and mode. *Ann. Math. Stat.*, **33**, 1065–1076.
- PIERCE D. & CARDIE C. (2001). Limitations of co-training for natural language learning from large datasets. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, Pittsburgh, Pennsylvania, USA.
- PRANJAL A., DELIP R. & BALARAMAN R. (2006). Part Of speech Tagging and Chunking with HMM and CRF. In *Proceedings of NLP Association of India (NLPAI) Machine Learning Contest*.
- RAYMOND C. & FAYOLLE J. (2010). Reconnaissance robuste d’entités nommées sur de la parole transcrite automatiquement. In *Actes de la conférence Traitement Automatique des Langues Naturelles*, Montréal, Canada.
- SALAKHUTDINOV R., ROWEIS S. & GHAHRAMANI Z. (2003). Optimization with EM and Expectation-Conjugate-Gradient. In *Proceedings of the conference ICML*.
- SCHRAUDOLPH N. N., YU J. & GÜNTHER S. (2007). A stochastic quasi-Newton method for online convex optimization. In *Proceedings of 11th International Conference on Artificial Intelligence and Statistics*, volume 2 of *Workshop and Conference Proceedings*, p. 436–443, San Juan, Puerto Rico.
- SETTLES B. (2010). *Active Learning Literature Survey*. Computer sciences technical report 1648, University of Wisconsin–Madison.
- SETTLES B. & CRAVEN M. (2008). An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1069–1078 : ACL Press.
- SMITH N. & EISNER J. (2005). Contrastive estimation : Training log-linear models on unlabeled data. In *Proceedings of ACL*.
- TJONG KIM SANG E. F. (2002). Introduction to the conll-2002 shared task : Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, p. 155–158 : Taipei, Taiwan.
- TJONG KIM SANG E. F. & BUCHHOLZ S. (2000). Introduction to the conll-2000 shared task : Chunking. In C. CARDIE, W. DAELEMANS, C. NEDELLEC & E. TJONG KIM SANG, Eds., *Proceedings of CoNLL-2000 and LLL-2000*, p. 127–132 : Lisbon, Portugal.
- WANG T., LI J., DIAO Q., WEI HU Y. Z. & DULONG C. (2006). Semantic event detection using conditional random fields. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW '06)*.
- WASSERMAN L. (2005). *All of Statistics : A Concise Course in Statistical Inference*. Springer Texts in Statistics.

Identification de facteurs de risque pour des patients diabétiques à partir de comptes-rendus cliniques par des approches hybrides

Cyril Grouin¹ Véronique Moriceau^{1, 2} Sophie Rosset¹ Pierre Zweigenbaum¹

(1) LIMSI-CNRS, UPR 3251, rue John von Neumann, 91400 Orsay

(2) Université Paris-Sud, Campus universitaire d'Orsay, 91400 Orsay

{prenom.nom}@limsi.fr

Résumé. Dans cet article, nous présentons les méthodes que nous avons développées pour analyser des comptes-rendus hospitaliers rédigés en anglais. L'objectif de cette étude consiste à identifier les facteurs de risque de décès pour des patients diabétiques et à positionner les événements médicaux décrits par rapport à la date de création de chaque document. Notre approche repose sur (i) HeidelTime pour identifier les expressions temporelles, (ii) des CRF complétés par des règles de post-traitement pour identifier les traitements, les maladies et facteurs de risque, et (iii) des règles pour positionner temporellement chaque événement médical. Sur un corpus de 514 documents, nous obtenons une F-mesure globale de 0,8451. Nous observons que l'identification des informations directement mentionnées dans les documents se révèle plus performante que l'inférence d'informations à partir de résultats de laboratoire.

Abstract.

Risk factor identification for diabetic patients from clinical records using hybrid approaches

In this paper, we present the methods we designed to process clinical records written in English. The aim of this study consists in identifying risk factors for diabetic patients and to define the temporal relation of those medical events wrt. the document creation time. Our approach relies (i) on HeidelTime to identify temporal expressions, (ii) on CRF and post-processing rules to identify treatments, diseases and risk factors, and (iii) on rules to determine the temporal relation of each medical event. On a corpus of 514 documents, we achieved a 0.8451 global F-measure. We observe we performed best on the identification of information mentioned in the text than information inference from lab results.

Mots-clés : Comptes-rendus hospitaliers, extraction d'information, apprentissage statistique.

Keywords: Electronic Health Records, Information Extraction, Machine Learning.

1 Introduction

Les documents cliniques contiennent des informations personnelles (description de l'environnement familial et social) et cliniques (examens, maladies, traitements) qui sont structurées et redondantes. Les comptes-rendus hospitaliers, notamment américains, sont structurés selon le modèle SOAP¹ (informations subjectives, informations objectives, résultats, conclusion/conduite à tenir) afin d'assurer l'interopérabilité entre comptes-rendus hospitaliers. Les informations les plus utiles pour établir le diagnostic d'un patient sont généralement répétées dans les différents documents qui constituent le dossier médical personnel d'un patient. La redondance et la dispersion de ces informations entre plusieurs documents nécessitent la mise au point d'outils informatiques permettant d'analyser le contenu de ces documents pour en extraire les informations pertinentes pour l'aide au diagnostic. Le résultat de ces traitements automatiques vise à découvrir de nouvelles informations (interactions médicamenteuses, effets secondaires, facteurs de risque, etc.) qui échappent à l'humain en raison de leur nombre.

Le diabète (terme générique couramment employé pour désigner le "diabète sucré") est une maladie qui se traduit par la perturbation de la régulation des sucres de l'organisme par l'insuline². Cette maladie se traduit par une augmentation du taux de sucre dans le sang. L'organisation mondiale de la santé (OMS) rapporte que le diabète de type 2 constitue la

1. L'acronyme renvoie aux quatre principales sections présentes dans un compte-rendu clinique : *Subjective, Objective, Assessment and Plan*.

2. <http://www.chu-rouen.fr/page/diabete>

forme de diabète la plus répandue dans le monde (90% des diabètes rencontrés³). Qualifié de “diabète de la maturité”, il survient chez les adultes âgés et augmente de 50% le risque de décès par une maladie cardio-vasculaire⁴. Les facteurs de risque qui augmentent les risques de décès sont connus et documentés (hypercholestérolémie, hypertension, obésité, tabagisme).

L’objectif du travail décrit dans cet article consiste à repérer automatiquement les facteurs de risques de développement de maladies cardio-vasculaires par des patients diabétiques, depuis les documents cliniques au format textuel. Les méthodes que nous avons développées ainsi que les données utilisées s’inscrivent dans le cadre de la campagne d’évaluation internationale i2b2 dont l’édition 2014 portait notamment sur cette problématique (Stubbs *et al.*, 2014a), à partir de documents cliniques rédigés en anglais.

2 État de l’art

Les informations pertinentes pour établir des diagnostics cliniques sont exprimées de deux manières différentes dans les comptes-rendus cliniques. Soit de manière explicite dans le cas où l’information est directement mentionnée (*le patient est connu pour une histoire de diabète*), soit de manière implicite par le biais d’événements médicaux (mode de vie : *tabagisme actif, boit occasionnellement de l’alcool*, résultats de laboratoire : *pression artérielle de 146/88*) qu’il importe d’analyser pour en inférer des informations (*146/88 mm/hg* → hypertension). Si la majorité des informations cliniques concernent le patient, certains événements peuvent se rapporter à la famille du patient (*son père avait du diabète, était fumeur, et est décédé d’un infarctus du myocarde à 65 ans*). Il importe alors de déterminer à qui se rapportent les informations mentionnées. D’autre part, l’existence continue et répétée dans le temps de certains événements médicaux influe sur la nature du diagnostic. Il est donc essentiel de tenir compte du positionnement temporel de ces événements par rapport à la date de consultation mentionnée en début de compte-rendu clinique. Plusieurs éditions récentes des campagnes d’évaluation internationales organisées par l’institut i2b2⁵ ont porté sur ces aspects. La comparaison des méthodes employées nous permet de mettre en évidence les méthodes les plus pertinentes au vu des objectifs poursuivis et des résultats obtenus.

2.1 Détection d’événements médicaux

La détection des événements médicaux (examens, maladies, modes de vie, traitements) depuis des comptes-rendus clinique est possible, soit par des approches à bases de règles et de projection de lexiques, soit par des approches par apprentissage statistique. Le choix de l’approche dépend, d’une part de la disponibilité de corpus annotés (pré-requis indispensable pour l’apprentissage statistique), et d’autre part du type d’informations à traiter (repérage d’entités nommées réalisable au moyen des deux approches par opposition à l’inférence d’informations uniquement possible avec des règles).

Afin de détecter dans des comptes-rendus hospitaliers les occurrences de traitements médicaux (noms de médicaments) et les informations associées (dosage, mode d’administration, durée, fréquence, etc.), (Doan *et al.*, 2010) ont produit un système qui découpe les documents en sections et en phrases, puis qui applique des règles d’étiquetage sémantique. Le découpage permet de repérer les passages porteurs d’informations, mais également de permettre le calcul de rattachements d’informations dans le cas de reprises pronominales. Sur un corpus de 251 documents issu de la campagne i2b2 2009 (Uzuner *et al.*, 2010), pour une évaluation au niveau des tokens⁶, les auteurs rapportent une F-mesure globale de 0,821 avec une précision (0,840) supérieure au rappel (0,803). En l’absence de corpus annoté, tous les participants de cette campagne ont produit des systèmes à base de règles.

En matière de détection du tabagisme chez des patients (*non fumeur, fumeur, ancien fumeur, statut inconnu*) depuis des comptes-rendus cliniques, (Clark *et al.*, 2008) ont mis en place un système en deux étapes de manières à (i) classer les documents selon qu’ils contiennent des indices sur le tabagisme, et (ii) pour les documents contenant de tels indices, effectuer une analyse linguistique du contenu pour associer ces indices à des expressions temporelles. Pour réaliser cette deuxième étape, les auteurs ont utilisé une approche à base de SVM. Sur le corpus i2b2 2006 composé de 104 documents (Uzuner *et al.*, 2008), et pour une évaluation au niveau du document⁷, les auteurs rapportent une exactitude de 93,6% pour la détermination du statut du tabagisme et une exactitude de 100% concernant la première étape de filtrage des documents.

3. <http://www.who.int/mediacentre/factsheets/fs312/fr/>

4. <http://www.who.int/mediacentre/factsheets/fs138/fr/>

5. Integrating Informatics and Biology to the Bedside, <https://www.i2b2.org/NLP/>

6. Une évaluation au niveau des tokens prend en compte toutes les occurrences d’une même forme : pour un nom d’un médicament répété plusieurs fois dans un document, l’évaluation prendra en compte chaque apparition.

7. Dans le cas d’une évaluation au niveau du document, une seule occurrence de l’information traitée est attendue.

Enfin, de manière à identifier les facteurs de comorbidité (*asthme, attaque cardiaque, dépression, diabète, hypercholestérolémie, hypertension, hypertriglycémie, maladie cardio-vasculaires, obésité, etc.*), (Childs *et al.*, 2008) ont produit un système composé de 281 règles et 9 étapes dans le but de reproduire les “signaux” médicaux qu’un expert humain considérerait comme pertinents pour décider de l’existence (présent, possible, absent, inconnu) de chaque facteur dans un compte-rendu clinique. Parmi les étapes appliquées figurent notamment l’application de NegEx (Chapman *et al.*, 2001) pour marquer la négation et l’incertitude. Sur le corpus i2b2 2008 contenant 8044 facteurs à détecter (Uzuner, 2009), les auteurs ont obtenu une micro F-mesure de 0,9773.

2.2 Positionnement des événements dans la chronologie des patients

Le positionnement d’événements médicaux dans la chronologie du patient constitue un champs de recherche récent, dans lequel les approches à base d’apprentissage sont largement employées, parfois complétées par des règles (Sun *et al.*, 2013). L’identification des expressions temporelles et la normalisation de ces expressions selon un format pivot est généralement réalisée au moyen d’outils de repérage de ce type d’expressions, tels que GUTime (Verhagen *et al.*, 2005), HeidelTime (Strötgen & Gertz, 2010), ou SUTime (Chang & Manning, 2012).

A partir de documents déjà annotés en expressions temporelles et événements médicaux, (Cherry *et al.*, 2013) ont combiné des approches à base d’apprentissage (entropie maximale et SVM) et de règles et lexiques pour repérer les relations temporelles (avant, pendant, après) qui existent, soit entre deux événements médicaux (maladie, examen, traitement, etc.), soit entre un événement médical et une expression temporelle (date, heure). Afin de traiter les particularités des relations temporelles, les auteurs ont défini quatre systèmes permettant de repérer : (i) les relations temporelles locales, (ii) les relations temporelles qui existent entre sections d’un document, (iii) les relations distantes dans le document entre événements se produisant au même moment, et (iv) les relations distantes qui entrent dans le cadre d’un lien de causalité. Sur un corpus de 120 documents, les auteurs rapportent une F-mesure globale de 0,6837 avec une précision (0,7537) supérieure au rappel (0,6449).

2.3 Détection des événements médicaux et positionnement temporel

Sur la campagne d’évaluation i2b2 2014 dans laquelle nous inscrivons ce travail, plusieurs participants ont utilisés des annotations complémentaires, soit en annotant manuellement des données issues de leur organisme (Roberts *et al.*, 2014), soit en réutilisant les données de l’édition 2006 sur le tabagisme (Cormack *et al.*, 2014). Ces annotations permettent de s’assurer de la cohérence globale des annotations et de garantir leur pertinence pour la méthode utilisée. Les participants ayant obtenu les meilleurs résultats sont ceux qui, en plus de l’utilisation d’annotations complémentaires, ont conçu plusieurs classifieurs selon les types d’information à traiter, un classifieur global et un deuxième dédié au tabagisme (Torii *et al.*, 2014, F=0,9209), ou en enchaînant plusieurs étapes (Roberts *et al.*, 2014, F=0,9277) incluant identification des concepts au moyen de lexiques, filtrage de ces concepts et positionnement temporel au moyen d’approches par apprentissage telles que les modèles de langue.

3 Objectifs

3.1 Présentation

L’objectif global que nous poursuivons consiste à repérer les facteurs de risque de développement de maladies cardiaques par des patients diabétiques parmi huit catégories (Stubbs *et al.*, 2014b,a). Le moment où apparaissent ces différents facteurs de risque dans la chronologie de la consultation constitue également une information capitale et permet d’étudier la progression des maladies cardiaques dans le temps. Il existe trois types d’information pertinentes pour répondre à cette problématique :

- les maladies connues : *diabète, maladie coronaro-artérielle (CAD)*,
- les facteurs de risque associés : *cholestérol et hyperlipidémie, hypertension, obésité, tabagisme, histoire familiale de maladies coronaro-artérielles*,
- et des indices annexes : *médicaments*.

Pour chacune de ces huit catégories existent des informations associées, dont le tableau 1 renseigne des différentes valeurs possibles en anglais (langue utilisée dans les documents de notre corpus, voir section 4) : pour les médicaments, la classe

pharmacologique ; pour les maladies, un indicateur de la manière dont l’information est présentée dans le document : soit l’information est directement mentionnée (mention), soit elle doit être inférée à partir de résultats de laboratoire (A1C, high LDL, etc.) ; et pour le tabagisme, le statut.

(a)	Élément	Classes pharmacologiques
	Médicament	insulin, metformin, calcium channel blocker, statin, aspirin, ACE inhibitor, beta blocker, nitrate, diuretic, ezetimibe, ARB, sulfonyleureas, fibrates, thienopyridine, niacin, thiazolidinedione, DPP4 inhibitors
(b)	Élément	Indicateurs
	Diabète	mention, A1C, glucose
	CAD	mention, event, symptom, test
	Hyperlipidémie	mention, high LDL, high chol.
	Hypertension	mention, high bp
	Obésité	mention, BMI
(c)	Élément	Statut
	Tabagisme	current, past, ever, never, unknown

TABLE 1 – Classes pharmacologiques des médicaments (a), indicateurs associés aux maladies (b), statut du tabagisme (c)

Si les maladies ne sont pas directement mentionnées dans le document (par exemple, les occurrences “hypertension” et “HTN” constituent des mentions et doivent être notées comme telles en tant qu’indicateur), elles doivent être inférées depuis des résultats de laboratoire, uniquement si les valeurs de ces résultats dépassent des seuils prédéfinis⁸ :

- Diabète :
 - dosage de l’hémoglobine A1c supérieur à 6,5 mmol/L,
 - ou deux valeurs successives de glycémie à jeun supérieures à 126 mg/dL ;
- Hyperlipidémie :
 - taux de cholestérol total supérieur à 240 mg/dL,
 - ou taux de cholestérol LDL (également appelé “mauvais cholestérol”) supérieur à 100 mg/dL ;
- Hypertension : pression sanguine supérieure à 140/90 mm/hg.

3.2 Réalisation

Nous poursuivons donc l’objectif global de : (i) repérage des éléments médicaux parmi les huit catégories précédemment listées, (ii) spécification de la manière dont l’information est représentée dans chaque catégorie (classe pharmacologique, indicateur, statut), et (iii) de détermination du positionnement temporel de ces facteurs de risque par rapport à une date de référence (dans le cas présent, la date de création du document (DCT) a été retenue) parmi trois valeurs possibles (avant, pendant, après la DCT).

L’identification de ces différents éléments se fait au niveau du document. Ainsi, il importe uniquement de connaître les médicaments, les maladies et les facteurs de risque d’un patient, quel que soit le nombre d’occurrences de chacun de ces éléments. Cependant, parce que chaque occurrence d’un élément peut renvoyer à différents moments de la chronologie de la consultation, plusieurs positionnements temporels peuvent être affectés à un même facteur de risque. Un patient pourra, par exemple, avoir pris un traitement médical avant, pendant et après la consultation à laquelle renvoie le document clinique. Cette particularité renvoie donc à une tâche de classification multi-labels.

4 Corpus

Le corpus que nous avons utilisé provient de l’édition 2014 du challenge i2b2. Il se compose de documents cliniques rédigés en anglais, issus de la base de données MIMIC-II (Saeed *et al.*, 2011). Ils conservent la forme d’origine du document papier, en particulier une largeur de colonne fixe, la présence de lignes blanches entre deux lignes de texte pour reproduire un double espacement, le positionnement exact des éléments sur la page (tabulation, espaces multiples), la présence de symboles particuliers pour représenter les séparateurs de colonnes de tableau (accent circonflexe, barre verticale), etc. Nous observons que seuls certains documents ont fait l’objet d’un pré-traitement pour rétablir chaque

8. Les seuils des valeurs numériques correspondent aux seuils communément admis dans la communauté médicale. Ces seuils figurent dans le guide d’annotation qui a été utilisé par les annotateurs et fourni par les organisateurs aux participants.

phrase sur une même ligne, en supprimant les sauts de ligne à l'intérieur d'une phrase. Aucune tokénisation n'a été réalisée dans le corpus.

Le corpus comprend 1 304 documents cliniques relatifs à 296 patients distincts. Les dossiers de chaque patient intègrent entre 3 et 5 documents, renvoyant à différentes consultations dans le temps. Trois cohortes de patients constituent ce corpus : (i) les patients qui ont déjà une maladie coronaro-artérielle, (ii) les patients qui n'ont pas de maladie coronaro-artérielle, et (iii) les patients qui développent une maladie coronaro-artérielle pendant la période couverte par leur dossier. Les documents cliniques sont de différents types : document d'admission, document de transfert, lettre de sortie et lettres de correspondance entre chirurgien et médecin, sans que le type de document soit clairement spécifié dans chaque fichier. Nous donnons en figure 1 un extrait de document issu du corpus, en mettant en évidence les informations pertinentes pour notre double problématique de détection de facteurs de risque et de positionnement temporel de ces événements.

Record date : 2086-11-07

(...)

I just had the pleasure of seeing Ms. Benitez for a follow up cardiovascular examination. She has been reasonably stable since I saw her last **in July**. She has had several episodes of chest discomfort characterized as a tightness which has been quite transient. The frequency has been less than one time per month. The last occurred **yesterday** while she was walking in a shopping mall.

(...)

She has had no symptom of cardiovascular ischemia, denying transient hemiparesis, hemiparesthesia, or amaurosis fugax.

(...)

Her current medical regimen includes **losartan** 75 mg p.o. b.i.d., **nifedipine** XL 60 mg p.o. q.d., **aspirin** 325 mg p.o. q.d., **atorvastatin** 40 mg p.o. q.d., **metoprolol** 12.5 mg p.o. b.i.d., (...)

On examination she appears well. Weight is 142 pounds. **Blood pressure is 140/60**. Heart rate is 50 and regular.

(...)

Her electrocardiogram shows sinus bradycardia with first degree AV block and findings suggestive of a **prior anterolateral myocardial infarction**.

FIGURE 1 – Extrait de dossier patient. Les éléments en gras constituent des informations essentielles pour déterminer les facteurs de risque de développement de maladies cardio-vasculaires pour des patients diabétiques

Le corpus d'entraînement contient 790 documents (178 patients) tandis que le corpus de test se compose de 514 documents (118 patients). Les patients sont différents entre les deux corpus. Nous avons segmenté aléatoirement le corpus d'entraînement en sous-corpus d'apprentissage (390 documents pour 89 patients) pour la mise au point de nos méthodes, sous-corpus de développement (131 documents pour 30 patients) pour configurer le système, et sous-corpus de test interne (269 documents pour 59 patients) pour évaluer le résultat de nos méthodes. Nous représentons sur la figure 2 la répartition initiale des documents entre corpus d'entraînement et corpus de test officiel, fournis par les organisateurs, et la segmentation que nous avons faite du corpus d'entraînement en sous-corpus d'apprentissage, de développement et de test interne.

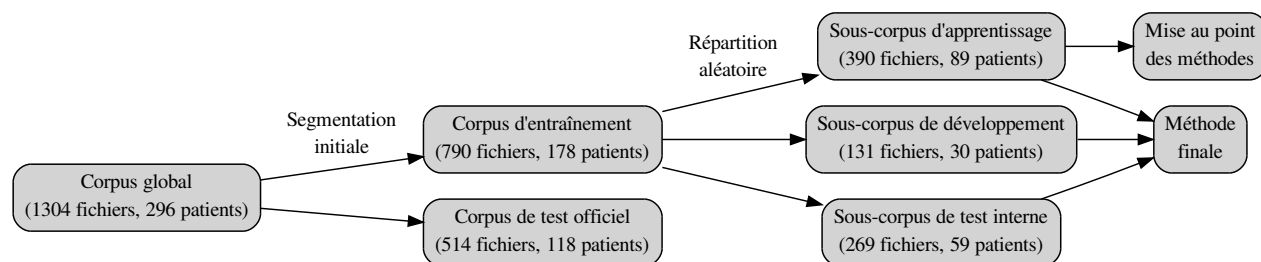


FIGURE 2 – Répartition des documents dans les différents corpus (entraînement et test officiel), et segmentation du corpus d'entraînement en sous-corpus d'apprentissage, de développement et de test interne

Nous donnons dans le tableau 2 des exemples d'événements médicaux à extraire pour chacune des huit catégories d'information, ainsi que les valeurs des informations associées (voir tableau 1 pour les différents types et valeurs possibles d'informations associées propre à chaque catégorie).

Phrase	Catégorie	Information associée	Temporalité
He was admitted to the hospital for <i>BG's of 400's</i>	Diabetes	<i>high glucose</i> (indicateur)	before
The patient is noted to have a history of mixed systemic conditions including <i>diabetes, coronary artery disease, depressive disorder...</i>	Diabetes	<i>mention</i> (indicateur)	before, during, after
	CAD	<i>mention</i> (indicateur)	before, during, after
pt had dissection and thus <i>2cd stent was placed</i>	CAD	<i>event</i> (indicateur)	during
She has occasional episodes of <i>angina</i>	CAD	<i>symptom</i> (indicateur)	before
<i>Father: extensive CAD</i> , with first MI in 50 s	Family History	<i>present</i> (indicateur)	—
Her HCL is still 36 and <i>LDL 118</i>	Hyperlipidemia	<i>high LDL</i> (indicateur)	before
The patient demonstrates a blood pressure of <i>146/88</i>	Hypertension	<i>high bp</i> (indicateur)	during
Medications on admission: ASA, <i>Lipitor 20, Lopresor 50 bid</i>	Medication	<i>aspirin</i> (classe)	before, during, after
	Medication	<i>statin</i> (classe)	before, during, after
	Medication	<i>beta blocker</i> (classe)	before, during, after
Vital signs: weight 241 lb, <i>BMI 37.8</i>	Obese	<i>BMI</i> (indicateur)	before, during, after
The patient <i>denies active tobacco</i> or alcoholic usage	Smoker	<i>never</i> (statut)	—

TABLE 2 – Exemples d'événements médicaux et informations associées (indicateur, classe pharmacologique, statut, temporalité) à extraire. Les passages en italiques désignent les indices permettant d'identifier ou d'inférer les événements

5 Méthodes

L'approche que nous avons retenue repose sur les étapes suivantes :

- Puisque les expressions temporelles constituent une information importante dans la tâche, nous commençons par identifier les expressions temporelles, nous les normalisons, et les réutilisons dans les étapes suivantes ;
- L'identification des maladies, des facteurs de risque et des médicaments se fait en trois étapes : (i) un pré-traitement du texte afin de réaliser une représentation vectorielle des caractéristiques de chaque token du document, (ii) une classification supervisée pour détecter les facteurs de risques directement mentionnés, ainsi que les noms de traitements médicaux, et (iii) un post-traitement à base de règles pour identifier certains facteurs de risque supplémentaires, tels que les résultats de laboratoire dont les valeurs supérieures à certains seuils déclenchent l'identification d'un facteur (par exemple, une pression sanguine supérieure à 140/90 mm/hg est signe d'hypertension et doit être identifiée).
- Enfin, l'identification du positionnement temporel est réalisé au moyen de règles produites manuellement.

Le schéma 3 décrit la succession de ces différentes étapes.

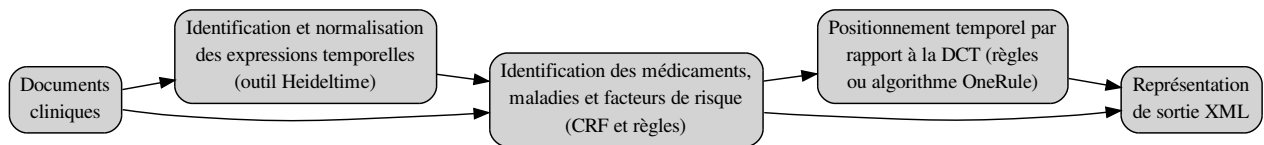


FIGURE 3 – Enchaînement des étapes suivies pour identifier les médicaments, maladies, facteurs de risque et déterminer le positionnement temporel de ces éléments par rapport à la date de création du document (DCT)

5.1 Normalisation des expressions temporelles

Nous avons utilisé l'outil à base de règles HeidelTime (Strötgen & Gertz, 2013) pour identifier les expressions temporelles absolues et relatives contenues dans les documents cliniques. La normalisation des expressions temporelles repose à la fois sur des lexiques et sur le temps des verbes présents dans la phrase de l'expression temporelle. Cette normalisation se fait en référence à la date de création du document (DCT). Dans les documents cliniques du corpus, nous avons pris pour référence la date introduite par la mention *Record date* présente dans chaque document.

Nous avons adapté l'outil aux caractéristiques du corpus (Moriceau & Tannier, 2014), d'une part en ajoutant une dizaine

de règles, notamment pour gérer certains formats particuliers de dates (M/JJ ou M/AA), et d'autre part en définissant des déclencheurs (*past*, *last*, *next*, *ago*) qui vont permettre de situer temporellement des durées qui ne peuvent être normalisées au format JJ-MM-AAAA. Par exemple, l'expression "2 weeks ago" ("il y a 2 semaines") est normalisée par HeidelTime en P2W et ce format de normalisation, qui n'indique pas si la durée est passée ou future, ne permet pas de situer directement cette expression par rapport à la DCT : la présence de tels déclencheurs permet alors de le faire.

Une fois réalisée la normalisation de toutes les expressions temporelles, nous avons calculé les relations temporelles de chaque expression (avant, pendant, après) en comparant la valeur normalisée de ces expressions avec la DCT.

5.2 Identification des maladies, médicaments et facteurs de risque

Afin d'identifier les maladies, médicaments et facteurs de risques, nous avons construit une chaîne de traitements reposant principalement sur une approche par apprentissage, complétée par des règles de post-traitements. Nous avons ainsi utilisé l'outil Wapiti (Lavergne *et al.*, 2010) fondé sur le formalisme des champs aléatoires conditionnels (CRF) (Sutton & McCallum, 2006). De manière à comparer les résultats obtenus par cette approche, nous avons constitué une approche basique (*baseline*) reposant sur la projection, sur les corpus de test, des observations effectuées sur les corpus d'entraînement.

5.2.1 Approche par apprentissage statistique

Nous avons construit nos modèles CRF en nous fondant sur les caractéristiques suivantes :

- **Caractéristiques lexicales** : le token ;
 - **Caractéristiques typographiques** :
 - longueur du token en nombre de caractères,
 - casse typographique du token,
 - présence de signes de ponctuation dans le token,
 - présence de chiffres dans le token ;
 - **Caractéristiques morpho-syntaxiques** : l'étiquette en partie du discours du token telle que fournie par l'outil Tree Tagger (Schmid, 1994) ;
 - **Caractéristiques sémantiques** :
 - si le token est un nom de médicament, la classe pharmacologique du token d'après une liste constituée à partir des annotations présentes dans le corpus d'entraînement (voir tableau 1a pour les 17 classes pharmacologiques utilisées) ;
 - la normalisation des expressions temporelles identifiées dans la même phrase que le token, telle que fournie par l'outil HeidelTime ;
 - **Caractéristique de structure** : la section dans laquelle apparaît le token, parmi 21 sections manuellement définies d'après les structures les plus fréquemment observées en corpus (*allergies*, *assessment and plan*, *chief complaint*, *family history*, *history of present illness*, *medications*, *physical exam*, *review of system*, *vital signs*, *social history*, etc.).
- Pour certaines caractéristiques (token, casse typographique, partie du discours), nous avons également défini des bigrammes de caractéristiques. Nous n'avons réalisé aucune validation croisée pour construire notre modèle. Nous avons cependant configuré la pénalité laplacienne $l1$ implémentée dans l'outil Wapiti de manière à réduire le sur-apprentissage des catégories les plus représentées dans le corpus.

Nous avons défini deux modèles CRF. Le premier prend en compte la normalisation des expressions temporelles identifiées par HeidelTime dans la même phrase que le token (modèle CRF complet), tandis que le deuxième ne tient pas compte de cette normalisation (modèle CRF simplifié). Les modèles que nous avons appliqués sur le sous-corpus de test interne ont été constitués à partir du sous-corpus d'apprentissage (390 documents) lors de l'étape de mise au point des méthodes, alors que les modèles appliqués sur le corpus de test officiel ont été constitués sur l'ensemble du corpus d'entraînement (790 documents, regroupant sous-corpus d'apprentissage, de développement et de test interne) une fois la méthode finalisée (voir figure 2).

5.2.2 Règles de post-traitement

Puisque le système CRF permet principalement d'identifier les facteurs de risque directement mentionnés dans le texte, nous avons complété notre approche par une douzaine de règles de post-traitement de manière à identifier les facteurs de risque représentés sous la forme de résultats de laboratoire. Dans ce dernier cas, seules les valeurs supérieures à des seuils

prédéfinis par les médecins constituent effectivement un facteur de risque et doivent être identifiées comme tel (les valeurs inférieures à ces seuils sont considérées comme normales et ne constituent donc pas des facteurs de risque, voir section 3).

En ce qui concerne la présence de maladies coronaro-artérielles dans la famille du patient, une étude du corpus nous a permis de constater un trop faible nombre de cas pour que nous puissions les traiter de manière efficace. Nous avons donc choisi de systématiquement considérer qu’il n’existe pas d’histoire familiale de ce type de maladie.

5.3 Calcul du positionnement temporel par rapport à la DCT

5.3.1 Approche à base de règles

Afin de calculer le positionnement temporel de chaque élément médical précédemment identifié par rapport à la date de création du document, nous avons réalisé une étude statistique du corpus d’entraînement. Nous avons pu observer que le positionnement temporel dépend à la fois de la catégorie et de l’indicateur associé au facteur de risque. Sur cette base, nous avons défini cinq règles générales qui nous permettent de traiter rapidement le positionnement temporel des événements médicaux : (i) pour *Medication*, les trois valeurs (“before”, “during”, “after”) sont systématiquement associées ; (ii) pour *CAD*, *Diabetes*, *Hyperlipidemia*, *Hypertension*, *Obese*, la valeur “before” est sélectionnée dans tous les cas ; (iii) pour les cinq mêmes types d’événements, la valeur “during” est sélectionnée dans tous les cas sauf pour la catégorie *Hyperlipidemia* si “indicator” a pour valeur *high chol.* ; (iv) pour ces cinq événements encore, la valeur “after” est associée pour certaines valeurs seulement de l’attribut “indicator”⁹ ; et (v) pour la catégorie *Smoker*, si le statut n’a pas été prédit par le CRF, nous associons la valeur “unknown” par défaut.

5.3.2 Approche par apprentissage statistique

Nous avons également défini une deuxième approche, inspirée de celle élaborée lors de notre participation à la campagne d’évaluation ShARe/CLEF eHealth 2014 (Hamon *et al.*, 2014) pour affecter des valeurs d’attribut temporel aux concepts médicaux présents dans les documents cliniques. Cette méthode repose sur un apprentissage supervisé, fondé sur la distribution des relations dans les différentes sections du documents, complétée par l’utilisation de caractéristiques déterminantes décrites plus bas. Pour traiter la classification multi-labels, nous prenons comme classes cibles la concaténation des valeurs temporelles présentes pour un événement (par exemple, “before+during+after” si un événement est associé en même temps à ces trois valeurs). Nous avons testé plusieurs algorithmes (arbres de décision, Naïve Bayes, OneRule) implémentés dans l’outil Weka (Hall *et al.*, 2009) et avons retenu l’algorithme OneRule en raison des bons résultats produits sur le corpus d’entraînement.

Puisque la tâche définit six facteurs de risque pour lesquels une relation temporelle doit être identifiée (*CAD*, *Diabetes*, *Hyperlipidemia*, *Hypertension*, *Medications*, *Obese*), nous avons créé un modèle distinct pour chacun de ces six facteurs. Les catégories *Smoker* et *Family history* n’impliquant pas un positionnement temporel des événements par rapport à la date de création du document¹⁰ (voir tableau 2), nous ne les traitons pas ici. Les six modèles créés reposent sur les caractéristiques suivantes :

Information associée aux événements médicaux : classe pharmacologique ou valeur de l’indicateur ;

Informations de structure :

- position relative d’un événement dans le texte découpé en cinq blocs de taille égale (*relative_position* = 0 . . . 4) ;
- section dans laquelle l’événement est identifié (*section_type*, selon les 21 sections manuellement définies).

La structure d’un document est modélisée au travers de sections et de positions relatives. Une étude du corpus d’entraînement nous a permis de mettre en évidence l’existence d’une corrélation entre les sections et la distribution des positionnements temporels. D’autre part, le positionnement “before” intervient souvent au début du document tandis que le positionnement “after” apparaît davantage vers la fin des documents. Nous avons donc utilisé ces informations de structure lors de la construction de nos modèles.

Les règles apprises par le classifieur OneRule sont indiquées dans le tableau 3 : pour chaque événement, le tableau présente l’attribut sélectionné pour déterminer le positionnement temporel de ce type d’événement, puis pour chaque valeur possible de cet attribut, la décision de positionnement retenue.

9. Pour *CAD* : event, mention, symptom ; *Diabetes* : mention ; *Hyperlipidemia* : mention ; *Hypertension* : mention ; et pour *Obese* : BMI et mention.

10. Dans le cadre de la campagne i2b2 2014, l’information de tabagisme (*Smoker*) n’est pas positionnée par rapport à la date de création du document. En revanche, elle est inscrite dans le temps par le biais du *statut* du tabagisme parmi cinq valeurs (*en cours*, *passé*, *toujours*, *jamais*, *inconnu*).

Événement	Attribut : Valeurs possibles	Positionnement temporel
CAD	indicator : <i>mention</i> <i>event, symptom, test</i>	before, during, after before
Diabetes	indicator : <i>mention</i> <i>A1C, glucose</i>	before, during, after before
Hyperlipidemia	indicator : <i>high chol., high LDL</i> <i>mention</i>	before before, during, after
Hypertension	indicator : <i>high bp</i> <i>mention</i>	during before, during, after
Medication	section_type : <i>Allergies, Discharge, HPI, Medications, Medications_On_Admission,</i> <i>Past_Medical_History, Plan, Problems, Subjective</i> <i>Assessment, Brief_Hospital_Course, Chief_Complaint, Conclu-</i> <i>sions, Consultations, Diagnosis, Family_History, Follow_Up,</i> <i>General, Hospital_Course, Impression_and_Plan, Interpretation,</i> <i>Major_Surgical_Or_Invasive_Procedure, Microbiology, Objective,</i> <i>Patient_Test_Information, Pertinent_Results, Physical_Examination,</i> <i>Prologue, Reason_For_This_Examination, Review_of_Systems,</i> <i>Social_History, Underlying_Medical_Condition, Vital_Signs</i>	before, during, after after
Obese	relative_position : <i>0, 1, 2</i> <i>3, 4</i>	before, during, after during

TABLE 3 – Règles apprises par le classifieur OneRule pour le calcul du positionnement temporel

5.4 Configurations des expériences

Nous avons défini trois configurations expérimentales distinctes de manière à tester différentes hypothèses de travail. Nous résumons ces configurations dans le tableau 4.

Config	Identification des maladies, facteurs de risque et traitements	Calcul de la temporalité	Hypothèse testée
S-1R	Modèle CRF simplifié et règles de post-traitement	OneRule	Approche statistique de base
C-1R	Modèle CRF complet et règles de post-traitement	OneRule	Les informations temporelles apportent une information utile
C-E	Modèle CRF complet et règles de post-traitement	Règles empiriques	Le calcul de la temporalité par des règles améliore la précision

TABLE 4 – Résumé des trois configurations expérimentales et hypothèse testée

La figure 4 représente les configurations testées en mettant en évidence les approches utilisées pour identifier les maladies, facteurs de risque et traitements (boîtes sur fond rouge) et pour calculer la temporalité (boîtes sur fond bleu). Sur ce schéma, la boîte de l’outil HeidelTime permet de rappeler en quoi diffère la construction des deux modèles CRF.

6 Résultats

Nous indiquons dans le tableau 5 les résultats que nous avons obtenus, sur les sous-corpus de développement et de test interne (modèles CRF construits sur le corpus d’apprentissage de 390 documents) et sur le corpus de test officiel (modèles

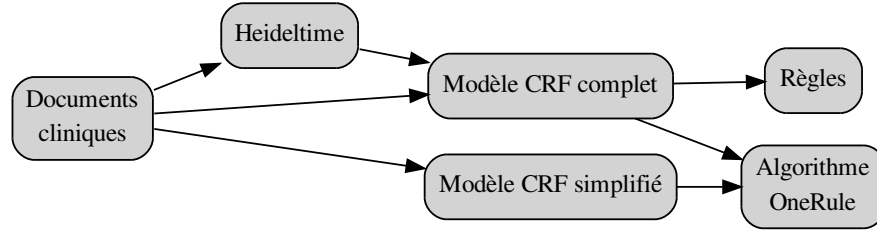


FIGURE 4 – Représentation graphique des trois configurations expérimentales

CRF construits sur l'ensemble du corpus d'entraînement, soit 790 documents). Les résultats sont donnés en terme de micro F-mesure, calculés par le script d'évaluation fourni par les organisateurs de la campagne d'évaluation.

Corpus	Test interne (269 docs)				Test officiel (514 docs)							
Modèle CRF	Corpus d'apprentissage (390 docs)				Corpus d'entraînement (790 docs)							
Expérience	Baseline	S-1R	C-1R	C-E	Baseline	S-1R				C-1R	C-E	
Précision	0,6206	0,7609	0,7609	0,8037	0,6503	0,9057	DCT before during after			0,9069	0,8753	
Rappel	0,8849	0,9445	0,9445	0,9484	0,9005	0,7922				0,7621	0,7689	
F-mesure	0,7295	0,8428	0,8428	0,8700	0,7552	0,8451				0,8282	0,8187	
CAD	0,4742	0,5558	0,5558	0,6015	0,4500	0,7387	0,6625	0,7637	0,8607	0,6971	0,6387	
Diabetes	0,7580	0,9158	0,9158	0,9257	0,7679	0,8996	0,8826	0,8921	0,9281	0,8689	0,8528	
Family_Hist	0,0131	1,000	1,000	1,000	0,9630	0,9630	—	—	—	0,9630	0,9630	
Hyperlipidemia	0,7749	0,8386	0,8386	0,8639	0,7955	0,8315	0,8043	0,8416	0,8517	0,8199	0,8167	
Hypertension	0,8668	0,8399	0,8399	0,9212	0,8806	0,9172	0,9139	0,9012	0,9444	0,9190	0,8753	
Medication	0,8128	0,8927	0,8927	0,9050	0,8401	0,8389	0,8435	0,8353	0,8378	0,8208	0,8208	
Obese	0,8013	0,5536	0,5536	0,7440	0,8259	0,6991	0,6331	0,8046	0,6331	0,6800	0,7817	
Smoker	0,5838	0,7514	0,7514	0,7454	0,5857	0,7237	—	—	—	0,7083	0,7096	

TABLE 5 – Résultats globaux et détaillés (micro mesures) sur le sous-corpus de test interne et sur le corpus de test officiel, pour chacune des trois configurations expérimentales envisagées (colonnes nommées S-1R, C-1R, C-E, voir tableau 4). L'évaluation du positionnement temporel est fournie pour l'expérience produisant les meilleurs résultats. Les meilleurs résultats sont représentés en gras

7 Discussion

Nous obtenons nos meilleurs résultats au moyen de la première configuration, fondée sur la combinaison du CRF avec les règles de post-traitements pour le calcul du positionnement temporel. Alors que nous avons obtenu de meilleurs résultats sur le sous-corpus de développement au moyen de la deuxième configuration (c.-à-d. en prenant les normalisations d'expressions temporelles fournies par HeidelTime comme caractéristiques pour construire notre modèle CRF), cette configuration s'est révélée la moins efficace sur le corpus de test officiel. Cette différence peut s'expliquer : (i) par le fait que le modèle CRF utilisé pour identifier les facteurs de risque n'a pas été construit sur le même ensemble de fichiers (390 documents pour le test interne vs. 790 documents pour le test officiel), et/ou (ii) parce que les propriétés présentes dans le sous-corpus de test interne et le corpus de test officiel ne se retrouvent pas selon les mêmes distributions, malgré la répartition aléatoire des documents lors de la constitution de nos sous-corpus de travail (voir section 4).

Contrairement aux résultats obtenus sur le sous-corpus de développement, l'utilisation des normalisations fournies par l'outil HeidelTime n'est pas pertinente pour le corpus de test officiel ($F_{C-1R} = 0,8282 < F_{S-1R} = 0,8451$). De manière similaire aux résultats obtenus sur le sous-corpus de développement, le calcul des relations temporelles permet d'améliorer les résultats ($F_{C-E} = 0,8187 < F_{C-1R} = 0,8282$).

Dans le détail, notre approche permet de traiter efficacement les facteurs de risque relatifs à deux catégories : *hypertension* ($F=0,9190$) et *diabète* ($F=0,8996$). Ces résultats s'expliquent par le nombre élevé de mentions pour ces deux catégories

dans le corpus d'entraînement, ce qui permet de produire des modèles CRF, pour la prédiction des facteurs de risque, et OneRule, pour la prédiction du positionnement temporel, particulièrement robustes. Le résultat élevé ($F=0,9630$) pour le facteur *family history of CAD* n'est pas pertinent dans la mesure où nous n'avons pas traité ce facteur de risque et avons simplement utilisé la valeur par défaut "not present" sur chacun des documents cliniques. Nous obtenons nos moins bons résultats sur les catégories *obese* ($F=0,6991$, jusqu'à 0,7817 sous la troisième configuration) et *smoker* ($F=0,7237$).

En ce qui concerne le calcul de la valeur du positionnement temporel associé à chaque facteur de risque, nous observons que les résultats par valeurs temporelles ("before", "during", "after") divergent selon les facteurs de risque : (i) pour les médicaments, les résultats sont homogènes entre les trois valeurs temporelles ; (ii) pour l'obésité, nous obtenons de meilleurs résultats avec la relation "during", et (iii) pour les autres facteurs de risque, nous réalisons de meilleures performances sur la relation "after". Une deuxième observation concernant le facteur d'obésité est que nous obtenons une F-mesure plus élevée avec la configuration C-E ($F=0,7817$), c'est-à-dire en calculant la temporalité au moyen de règles. Cela signifie que le modèle OneRule n'est pas efficace sur ce facteur puisqu'il conduit à dégrader les résultats de 10 points de F-mesure : l'information de position relative dans le texte est moins prédictive que les règles définies empiriquement. La comparaison des configurations C-1R et C-E pour les autres facteurs montre cependant que les modèles OneRule (C-1R) sont plus efficaces pour les catégories *CAD* (+5.8pt), *Hypertension* (+4.4pt), et *Diabetes* (+1.6pt).

Sur l'identification des facteurs de risque, nous obtenons des résultats différents en fonction de la manière dont sont représentés ces facteurs dans les documents. Nous avons ainsi mieux identifié les mentions que les informations devant être inférées de valeurs numériques supérieures à des seuils prédéterminés. Pour le facteur *hyperlipidemia*, nous obtenons par exemple des F-mesures de 0,8517 sur les "mention" (711 entités dans le corpus de test), de 0,4444 sur les valeurs "high cholesterol" (un taux de cholestérol total supérieur à 240 mg/dL, soit seulement 11 entités), et de 0,2941 pour les valeurs "high LDL" (un taux de cholestérol LDL supérieur à 100 mg/dL, soit 29 entités).

8 Conclusion

Dans cet article, nous avons présenté les expériences que nous avons menées pour répondre à la double problématique de (i) l'identification des facteurs de risque de développement de maladies cardio-vasculaires pour des patients diabétiques et (ii) de positionnement de ces facteurs de risque par rapport à la date de consultation, depuis les informations exprimées dans des documents cliniques. Sur l'identification des facteurs de risque, notre approche repose sur un système par apprentissage statistique (CRF) complété par des règles de post-traitements. Nous avons déterminé le positionnement temporel des éléments précédemment identifiés vis à vis de la date de création du document (DCT) au moyen d'un ensemble de six règles. L'enchaînement de ces deux étapes nous permet d'obtenir une micro F-mesure globale de 0,8451 sur le corpus de test. Il est difficile de dire à quelle distance ce score est de l'optimum atteignable par rapport au corpus d'entraînement dans la mesure où les meilleurs systèmes ($F=0,9277$) ont été entraînés par des équipes qui ont d'abord complété les annotations de ce corpus. Un travail de réannotation de corpus constitue une étape utile au vu de ces résultats.

Le calcul de la temporalité par une méthode d'apprentissage, même simple comme OneRule, conduit à de meilleurs résultats que les cinq règles définies empiriquement. Nous comptons tester une autre approche du problème de classification multi-labels en entraînant séparément un classifieur pour chaque positionnement temporel. Nous envisageons également d'adopter des méthodes différentes pour le traitement de certains facteurs comme le statut de tabagisme, en considérant qu'il s'agit également d'une tâche de classification que la présence d'indices dans les documents pourrait aider à résoudre.

Remerciements

Ce travail a été financé par l'Agence Nationale de Sécurité du Médicament (ANSM) dans le cadre du projet Vigi4MED (Vigilance dans les forums sur les Médicaments) ANSM-2013-S-060 et par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet Accordys (Agrégation de Contenus et de Connaissances pour Raisonner à partir de cas dans la DYSmorphologie fœtale) ANR-12-CORD-0007-03.

Références

CHANG A. X. & MANNING C. D. (2012). SUTime : a library for recognizing and normalizing time expressions. *Language Resources and Evaluation*.

- CHAPMAN W. W., BRIDEWELL W., HANBURY P., COOPER G. F. & BUCHANAN B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*, **34**(5), 301–10.
- CHERRY C., ZHU X., MARTIN J. & DE BRUIJN B. (2013). A la recherche du temps perdu : extraction of temporal relations from medical text in the 2012 i2b2 NLP challenge. *J Am Med Inform Assoc*, **20**(5), 843–48.
- CHILDS L. C., TAYLOR R. J., SIMONSEN L., HEINTZELMAN N. H., KOWALSKI K. M. & TAYLOR R. J. (2008). Description of a rule-based system for the i2b2 challenge in natural language processing for clinical data. *J Am Med Inform Assoc*, **16**(4), 571–5.
- CLARK C., GOOD K., JEZIERNY L., MACPHERSON M., WILSON B. & CHAJEWSKA U. (2008). Identifying smokers with a medical extraction system. *J Am Med Inform Assoc*, **15**(1), 36–9.
- CORMACK J., NATH C., MILWARD D., RAJA K. & JONNALAGADDA S. (2014). Agile text mining for the i2b2 2014 cardiac risk factors challenge. In *i2b2 Work Proc*, Washington, DC.
- DOAN S., BASTARACHE L., KLIMKOWSKI S., DENNY J. C. & XU H. (2010). Integrating existing natural language processing tools for medication extraction from discharge summaries. *J Am Med Inform Assoc*, **17**(5), 528–31.
- HALL M. A., FRANK E., HOLMES G., PFAHRINGER B., REUTEMANN P. & WITTEN I. H. (2009). The WEKA data mining software : An update. *SIGKDD Explor Newsl*, **11**(1).
- HAMON T., GROUIN C. & ZWEIGENBAUM P. (2014). Disease and disorder template filling using rule-based and statistical approaches. In *Working notes of the ShARe/CLEF eHealth Evaluation Lab*.
- LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proc of ACL*, p. 504–13, Uppsala, Sweden.
- MORICEAU V. & TANNIER X. (2014). French resources for extraction and normalization of temporal expressions with HeidelTime. In *Proc of LREC*, p. 3239–43, Reykjavik, Iceland.
- ROBERTS K., SHOOSHAN S. E., RODRIGUEZ L., ABHYANKAR S., KILICOGU H. & DEMNER-FUSHMAN D. (2014). NLM : Machine learning methods for detecting risk factors for heart disease in EHRs. In *i2b2 Work Proc*, Washington, DC.
- SAEED M., VILLAROE M., REISNER A. T., CLIFFORD G., LEHMAN L.-W., MOODY G., HELDT T., KYAW T. H., MOODY B. & MARK R. G. (2011). Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) : A public-access intensive care unit database. *Crit Care Med*, **39**(5), 952–60.
- SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proc of International Conference on New Methods in Language*.
- STRÖTGEN J. & GERTZ M. (2010). Heideltime : high quality rule-based extraction and normalization of temporal expressions. In *Proc of SemEval*.
- STRÖTGEN J. & GERTZ M. (2013). Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, **47**(2), 269–298.
- STUBBS A., KOTFILO C., XU H. & UZUNER O. (2014a). Practical applications for NLP in clinical research : the 2014 i2b2/UTHealth shared tasks. In *Proc of i2b2/UTHealth NLP Challenge*.
- STUBBS A., UZUNER O., KUMAR V. & SHAW S. (2014b). *Annotation guidelines : risk factors for heart disease in diabetic patients*. i2b2/UTHealth NLP Challenge.
- SUN W., RUMSHISKY A. & UZUNER O. (2013). Temporal reasoning over clinical text : the state of the art. *J Am Med Inform Assoc*, **20**(5), 814–9.
- SUTTON C. & MCCALLUM A. (2006). An introduction to conditional random fields for relational learning. In L. GETOOR & B. TASKAR, Eds., *Introduction to Statistical Relational Learning*. MIT Press.
- TORII M., WEI FAN J., LI YANG W., LEE T., WILEY M. T., ZISOOK D. & HUANG Y. (2014). De-identification and risk factor detection in medical records. In *i2b2 Work Proc*, Washington, DC.
- UZUNER O. (2009). Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc*, **16**(4), 561–70.
- UZUNER O., GOLDSTEIN I., LUO Y. & KOHANE I. (2008). Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc*, **15**(1), 14–24.
- UZUNER O., SOLT I. & CADAG E. (2010). Extracting medication information from clinical text. *J Am Med Inform Assoc*, **17**(5), 514–518.
- VERHAGEN M., MANI I., SAURI R., KNIPPEN R., JANG S. B., LITTMAN J., RUMSHISKY A., PHILLIPS J. & PUSTEJOVSKY J. (2005). Automating temporal annotation with TARSQI. In *Proc of ACL, Interactive Poster and Demonstration Sessions*, Stroudsburg, PA.

Oublier ce qu'on sait, pour mieux apprendre ce qu'on ne sait pas : une étude sur les contraintes de type dans les modèles CRF

Nicolas Pécheux^{1,2} Alexandre Allauzen^{1,2} Thomas Lavergne^{1,2}
Guillaume Wisniewski^{1,2} François Yvon²
(1) Université Paris-Sud, 91 403 Orsay CEDEX
(2) LIMSI-CNRS, 91 403 Orsay CEDEX
{prenom.nom}@limsi.fr

Résumé. Quand on dispose de connaissances *a priori* sur les sorties possibles d'un problème d'étiquetage, il semble souhaitable d'inclure cette information lors de l'apprentissage pour simplifier la tâche de modélisation et accélérer les traitements. Pourtant, même lorsque ces contraintes sont correctes et utiles au décodage, leur utilisation lors de l'apprentissage peut dégrader sévèrement les performances. Dans cet article, nous étudions ce paradoxe et montrons que le manque de contraste induit par les connaissances entraîne une forme de sous-apprentissage qu'il est cependant possible de limiter.

Abstract.

Ignore what you know to better learn what you don't : a case study on type constraints for CRFs

When information about the possible outputs of a sequence labeling task is available, it may seem appropriate to include this knowledge into the system, so as to facilitate and speed-up learning and inference. However, we show in this paper that using such constraints at training time is likely to drastically reduce performance, even when they are both correct and useful at decoding. In this paper, we study this paradox and show that the lack of contrast induced by constraints leads to a form of under-fitting, that it is however possible to partially overcome.

Mots-clés : Étiquetage Morpho-Syntaxique ; Apprentissage Statistique ; Champs Markoviens Aléatoires.

Keywords: Part-of-Speech Tagging ; Statistical Machine Learning ; Conditional Random Fields.

1 Introduction

De nombreux problèmes de Traitement Automatique des Langues (TAL) peuvent être formalisés comme des problèmes d'étiquetage de séquences, bénéficiant de ce fait de méthodes et de résultats établis en apprentissage automatique. Il serait souhaitable pour certaines applications de pouvoir introduire des contraintes sur les étiquetages possibles, de manière implicite ou explicite afin d'introduire des connaissances linguistiques ou de réduire le temps de calcul pour des problèmes de grande dimension. Par exemple, dans une tâche de segmentation utilisant un encodage BIO¹ des étiquettes, on peut vouloir imposer qu'une étiquette 'O' ne précède jamais une étiquette 'I'. Les contraintes linguistiques peuvent introduire des connaissances provenant de règles syntaxiques ou de dictionnaires, comme c'est le cas dans certaines tâches d'analyse morpho-syntaxique (Li *et al.*, 2012). De manière plus pragmatique, l'analyse morpho-syntaxique pour les langues à morphologie riche implique de prédire une étiquette parmi des ensembles comprenant des centaines, voire des milliers, d'étiquettes : les problèmes de désambiguïsation associés sont donc à la fois plus difficiles et computationnellement plus coûteux, au point de rendre inopérantes les méthodes standard (Müller *et al.*, 2013).

Cette étude s'intéresse donc à l'introduction de contraintes lors de l'apprentissage d'un étiqueteur morpho-syntaxique. Pour cela, nous supposons disposer d'un dictionnaire associant à chaque mot un sous-ensemble des étiquettes possibles. Ce dictionnaire peut refléter une connaissance linguistique préalable, par exemple être extrait automatiquement de WIKI-TIONNAIRE ou encore être déduit des données d'apprentissage. Sous l'hypothèse que ce dictionnaire est correct, il semble naturel de vouloir prendre cette information en compte, afin d'une part, d'accélérer l'apprentissage et l'inférence, d'autre part, d'améliorer la qualité des prédictions. En réduisant l'ensemble des étiquettes pouvant être prédites et donc la taille

1. Pour début (B) ; intérieur (I) ; et en dehors (O).

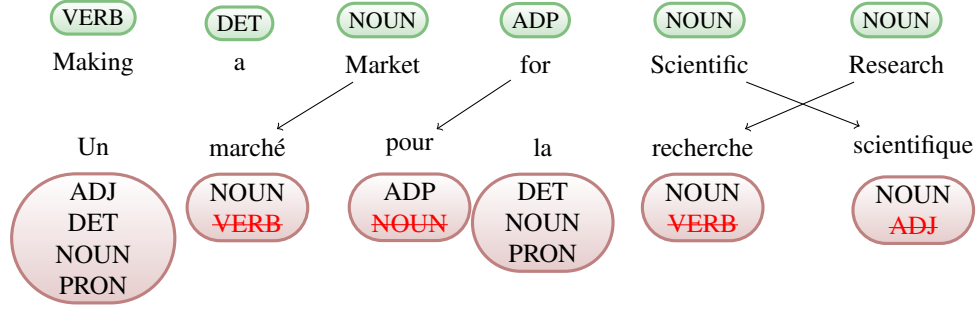


FIGURE 1: Instance d'apprentissage obtenue par transfert cross-lingue à partir d'une phrase source (haut) vers une phrase cible (bas). Les étiquettes autorisées par un dictionnaire sont représentées en rouge. Les étiquettes de la phrase source (en vert) sont *projetées* à travers les liens d'alignement de manière à désambiguïser les étiquettes cibles. Ces dernières constituent la référence (en noir), éventuellement ambiguë, pour le problème d'analyse morpho-syntaxique. À l'apprentissage, comme au décodage, on peut considérer, pour construire l'espace de recherche, que les 12 étiquettes sont possibles (cas non représenté) ou bien que seules celles qui sont proposées par le dictionnaire de type (encadrées en rouge) sont licites.

des espaces de recherche associés, les contraintes devraient permettre en un certain sens de simplifier la tâche du modèle. En effet, ce dernier peut alors concentrer son apprentissage sur la discrimination entre des hypothèses réalistes, en nombre réduit, plutôt que de considérer des configurations qui ne peuvent pas se produire.

Dans cet article, nous montrons que cette intuition n'est pas toujours correcte et qu'ajouter une telle information, même lorsqu'elle est pertinente et exacte, peut conduire à une dégradation de la capacité de généralisation du système, comme l'illustre le résultat paradoxal décrit à la section 2. La contribution de ce travail est d'apporter des explications à ce comportement inattendu, afin de pouvoir y remédier. Cette étude se concentre sur les modèles log-linéaires qui sont présentés à la section 3. En analysant théoriquement l'effet de l'inclusion des contraintes dans le modèle (section 4), il est possible de mettre en lumière les relations complexes qui existent entre les contraintes, la régularisation et le sous-apprentissage. Les résultats expérimentaux présentés à la section 5 montrent, en effet, que l'introduction de contraintes peut entraîner une forme de sous-apprentissage de certaines caractéristiques, qu'il est possible d'éviter. En particulier, il semble important de limiter, lors de l'apprentissage, l'impact des contraintes afin de garder une forme de contraste.

2 Un résultat paradoxal

Le point de départ de cette étude est une tentative de reproduire les résultats de Täckström *et al.* (2013). Ces derniers s'intéressent à une tâche d'analyse morpho-syntaxique pour des langues cibles peu dotées, pour lesquelles deux types de ressources sont disponibles : d'une part un dictionnaire (WIKTIONNAIRE) permettant de connaître, pour un mot, l'ensemble de ses catégories morpho-syntaxiques possibles ; d'autre part, un corpus parallèle aligné mot-à-mot et dont la partie source a été étiquetée automatiquement. En combinant ces deux ressources, comme illustré à la figure 1, il est possible d'apprendre un analyseur morpho-syntaxique, même lorsque l'on ne dispose pas de données cibles annotées.

Dans cette approche, le dictionnaire joue un rôle central : d'une part, en validant les étiquettes projetées au travers des liens d'alignement pour créer la référence ; d'autre part, en restreignant l'espace de recherche de l'analyseur morpho-syntaxique : lors de l'apprentissage et du décodage, la liste des étiquettes possibles pour chaque mot peut alors être réduite à un ensemble d'alternatives (les étiquettes autorisées par le dictionnaire) bien plus restreint que l'ensemble des étiquettes définies dans le schéma d'annotation.

Le tableau 1 rassemble les taux d'erreur obtenus par notre ré-implémentation du meilleur modèle décrit dans Täckström *et al.* (2013). En comparant la première et la deuxième ligne, on s'aperçoit qu'il est intéressant d'ajouter de manière explicite les contraintes de dictionnaire lors du décodage : ceci oblige le modèle à choisir, lors du décodage, l'une des étiquettes possibles et permet ainsi d'éviter certaines erreurs. Il est donc utile, du moins dans ce cas, d'utiliser l'information sur les étiquettes possibles d'un mot. En revanche, de manière surprenante, la troisième ligne de ce tableau montre qu'introduire ces contraintes lors de l'apprentissage dégrade sévèrement les performances.

Nous sommes donc, en apparence, face à un double paradoxe : (a) inclure des contraintes pourtant informatives pénalise le modèle ; (b) reproduire des conditions similaires à l'entraînement et au test n'est pas la meilleure configuration. Dans cet article, nous proposons d'expliquer ce paradoxe aussi bien d'un point de vue théorique (§ 4) qu'expérimental (§ 5).

appr.	test	cs	de	el	es	fi	fr	id	it	sv
✗	✗	17.3	13.3	16.8	14.7	19.2	14.1	14.8	13.3	12.5
✗	✓	16.7	11.8	16.3	12.4	17.4	13.7	14.6	12.7	12.0
✓	✓	21.2	15.8	17.6	15.5	27.4	23.1	27.9	15.1	14.7

TABLE 1: Une série de résultats surprenants : taux d'erreur (en %) pour neuf langues, obtenus par un modèle CRF partiellement observé sur une tâche d'analyse morpho-syntaxique par transfert cross-lingue à partir de l'anglais, selon que l'on utilise les contraintes de type pour définir l'espace de recherche à l'apprentissage (appr.) et/ou au test (test). L'intégration de contraintes à l'apprentissage dégrade systématiquement les performances. Les contraintes de type sont obtenues en prenant l'union d'un dictionnaire déduit des alignements et d'un dictionnaire extrait du WIKTIONNAIRE (voir la section 5.2 pour plus de détails sur les conditions expérimentales et les langues considérées).

3 Cadre classique : modèles et espaces de recherche

Cette section introduit un cadre général qui permettra de mieux comprendre le rôle des différents espaces de recherches manipulés dans notre problème d'apprentissage structuré ; la question des contraintes sera ensuite abordée au § 4.

3.1 Espaces de recherche et de référence

Dans un problème d'apprentissage générique, on dispose de l'ensemble \mathcal{X} des *entrées* possibles, ainsi que celui \mathcal{Y} des *sorties* possibles. Une manière classique de formuler un problème d'apprentissage pour le TAL (Smith, 2011, p. 23) est de considérer que pour chaque entrée $\mathbf{x} \in \mathcal{X}$, l'ensemble des sorties possibles est restreint à un sous-ensemble $\mathcal{Y}(\mathbf{x}) \subseteq \mathcal{Y}$. Cet espace $\mathcal{Y}(\mathbf{x})$ est appelé *l'espace de recherche*.

Exemple 3.1. Dans le cas de l'analyse morpho-syntaxique, notons \mathcal{V} le vocabulaire et \mathcal{T} l'ensemble des étiquettes morpho-syntaxiques possibles pour la tâche considérée. On a alors $\mathcal{X} = \bigcup_{n \in \mathbb{N}^*} \mathcal{V}^n$ et $\mathcal{Y} = \bigcup_{n \in \mathbb{N}^*} \mathcal{T}^n$. On considère usuellement, pour un $\mathbf{x} \in \mathcal{X}$ donné, uniquement les séquences d'étiquettes de même longueur que \mathbf{x} , c'est-à-dire que $\forall \mathbf{x} \in \mathcal{X}, \mathcal{Y}(\mathbf{x}) = \mathcal{T}^{|\mathbf{x}|}$.

Dans le cadre de l'apprentissage supervisé, on dispose d'un ensemble de données d'apprentissage $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1..N}$ supposées i.i.d sous une certaine distribution inconnue \mathbb{D} , où $\mathbf{y}_i \in \mathcal{Y}(\mathbf{x}_i)$ est la sortie de référence pour l'entrée \mathbf{x}_i . De manière plus générale, on peut considérer que pour chaque exemple \mathbf{x}_i on dispose d'un sous-ensemble $\mathcal{Y}^r(\mathbf{x}_i) \subset \mathcal{Y}(\mathbf{x}_i)$, que l'on appelle *espace de référence*. Tous les éléments de $\mathcal{Y}^r(\mathbf{x}_i)$ peuvent être complètement corrects, — ainsi, en traduction automatique, plusieurs traductions d'une même phrase peuvent être également bonnes — ; ou bien seulement partiellement corrects comme dans le cadre de l'apprentissage partiellement supervisé, dans lequel on dispose d'une connaissance incomplète de la véritable référence.

3.2 Modèle log-linéaire

Dans ce travail, on s'intéresse aux modèles conditionnels log-linéaires qui, connaissant les entrées $\mathbf{x} \in \mathcal{X}$, définissent une distribution de probabilité sur les sorties $\mathbf{y} \in \mathcal{Y}(\mathbf{x})$ possibles comme :

$$p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_{\boldsymbol{\theta}}(\mathbf{x})} \exp(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}, \mathbf{y})) \quad (1)$$

où $\boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d$ est un vecteur de d caractéristiques, $\boldsymbol{\theta} \in \mathbb{R}^d$ un vecteur de paramètres et $Z_{\boldsymbol{\theta}}(\mathbf{x})$ est le terme de normalisation. La log-vraisemblance des paramètres $\boldsymbol{\theta}$ s'écrit alors :

$$\ell_r(\boldsymbol{\theta}, \mathcal{D}) = \sum_{i=1}^N \log p_{\boldsymbol{\theta}}(\mathcal{Y}^r(\mathbf{x}_i) | \mathbf{x}_i), \text{ avec} \quad (2)$$

$$p_{\boldsymbol{\theta}}(\mathcal{Y}^r(\mathbf{x}) | \mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}^r(\mathbf{x})} p_{\boldsymbol{\theta}}(\mathbf{y} | \mathbf{x}). \quad (3)$$

Remarquons que lorsque pour chaque $\mathbf{x} \in \mathcal{X}$, $|\mathcal{Y}^r(\mathbf{x})| = 1$, on retrouve le cadre classique de l'apprentissage supervisé. Le principe de maximum de vraisemblance consiste à choisir :

$$\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta}} \ell_r(\boldsymbol{\theta}, \mathcal{D}) - \lambda_1 \|\boldsymbol{\theta}\|_1 - \frac{1}{2} \lambda_2 \|\boldsymbol{\theta}\|_2^2, \quad (4)$$

où l'on introduit souvent une régularisation \mathcal{L}_1 (pondérée ici par λ_1) et/ou une régularisation \mathcal{L}_2 (pondérée par λ_2). L'approche du maximum de vraisemblance conduit à augmenter la masse de probabilité de l'espace de référence $\mathcal{Y}^r(\mathbf{x})$ au sein de l'espace de recherche $\mathcal{Y}(\mathbf{x})$ (équation (3)). Dans le cas d'un CRF linéaire du premier ordre (Lafferty *et al.*, 2001), les caractéristiques se décomposent sur les paires d'étiquettes voisines selon :

$$\phi(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{|\mathbf{x}|} \phi(y_i, y_{i-1}, \mathbf{x}). \quad (5)$$

La complexité de l'apprentissage et du décodage d'un CRF sont quadratiques en la taille du jeu d'étiquettes à l'ordre 1 et croissent exponentiellement avec l'ordre. Cette complexité justifie qu'on se limite en général à des jeux d'étiquettes restreints (typiquement quelques dizaines) et à des modèles d'ordre faible (1 ou 2).

4 Contraintes

Étant donné le cadre de l'apprentissage structuré et les modèles log-linéaires décrits à la section 3, nous allons maintenant introduire formellement la notion de contrainte.

4.1 Fonction de contrainte et espaces restreints

Nous modélisons les contraintes par la notion de *fonction de contrainte* : $c : \mathbf{x} \in \mathcal{X} \rightarrow \mathcal{Y}^c(\mathbf{x}) \subseteq \mathcal{Y}(\mathbf{x})$. Ces fonctions sont déterministes et ne font pas partie du modèle.

Exemple 4.1. Dans le cas de l'analyse morpho-syntaxique, soit $t : \mathcal{V} \rightarrow 2^T$ un dictionnaire associant à chaque mot un ensemble d'étiquettes, on considère la fonction « contrainte dictionnaire » suivante, que l'on note abusivement également t :

$$t : \mathbf{x} = (x_1, \dots, x_{|\mathbf{x}|}) \in \mathcal{X} \rightarrow \mathcal{Y}^t(\mathbf{x}) = t(x_1) \times t(x_2) \times \dots \times t(x_{|\mathbf{x}|})$$

qui n'autorise que les séquences d'étiquettes respectant, pour chaque mot, les contraintes données par le dictionnaire.

4.2 Modèle log-linéaire avec contraintes

On peut maintenant étendre les notations de la section 3 en prenant en compte des contraintes sur l'espace de recherche, données par une fonction de contrainte c :

$$\forall \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}^c(\mathbf{x}), p_{\boldsymbol{\theta}}^c(\mathbf{y}|\mathbf{x}) = \frac{1}{\mathcal{Z}_{\boldsymbol{\theta}}^c(\mathbf{x})} \exp(\boldsymbol{\theta}^T \phi(\mathbf{x}, \mathbf{y})), \quad (6)$$

où le terme de normalisation devient :

$$\mathcal{Z}_{\boldsymbol{\theta}}^c(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}^c(\mathbf{x})} \exp(\boldsymbol{\theta}^T \phi(\mathbf{x}, \mathbf{y})). \quad (7)$$

Les contraintes influencent uniquement le calcul de la fonction de partition dans le modèle exponentiel : tout se passe comme si les sorties impossibles selon les contraintes avaient une probabilité nulle.

À l'apprentissage, en notant a la fonction de contrainte utilisée, l'équation (2) s'écrit :

$$\ell_r^a(\boldsymbol{\theta}, \mathcal{D}) = \sum_{i=1}^N \log p_{\boldsymbol{\theta}}^a(\mathcal{Y}^r(\mathbf{x}_i) | \mathbf{x}_i). \quad (8)$$

Tout comme à l'apprentissage, si l'on a accès à une fonction de contrainte de bonne qualité, il peut être avantageux d'exploiter celle-ci pour réduire les candidats possibles lors du décodage, et ainsi de diminuer les risques d'erreur tout en augmentant la vitesse d'inférence. En notant d cette fonction de contrainte, cela revient à considérer la règle de décision :

$$\mathbf{y}^* = f(\mathbf{x}) = \arg \max_{\mathbf{y} \in \mathcal{Y}^d(\mathbf{x})} p_{\theta}^d(\mathbf{y}|\mathbf{x}). \quad (9)$$

Intuitivement, il semble préférable d'utiliser le même espace de recherche lors de l'apprentissage et du décodage, mais il est important de bien voir que rien ne l'impose. Nous avons d'ailleurs vu à la section 2 un exemple où il était préférable de ne pas considérer la même fonction de contrainte lors de l'apprentissage et lors du décodage (deuxième ligne du tableau 1). Remarquons que l'on pourrait même envisager de mettre des contraintes plus strictes à l'apprentissage que lors du décodage². La question principale soulevée par cette étude est de se demander comment choisir optimalement $\mathcal{Y}^a(\mathbf{x})$ pour l'apprentissage et $\mathcal{Y}^d(\mathbf{x})$ lors du décodage.

4.3 Contraintes comme caractéristiques

Il est intéressant de noter qu'il est possible de représenter explicitement les contraintes, jusqu'ici externes au modèle, comme des caractéristiques particulières associées à des poids qui en dissuadent la violation. Soit c une fonction de contrainte et supposons qu'il existe un ensemble de caractéristiques $I \subset 2^d$ permettant d'encoder exactement le complémentaire de l'espace de recherche donné par l'application des contraintes :

$$\forall \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}(\mathbf{x}) \quad \mathbf{y} \notin \mathcal{Y}^c(\mathbf{x}) \Leftrightarrow \exists k \in I, \phi_k(\mathbf{x}, \mathbf{y}) \neq 0.$$

Il suffit alors de fixer le poids de toutes ces caractéristiques à $-\infty$ pour obtenir un modèle sans contraintes équivalent. Par exemple, dans le cas des contraintes de type, on peut considérer l'ensemble des caractéristiques (mot, étiquette) qui ne figurent pas dans le dictionnaire³. On note que pour parvenir à cette représentation équivalente, il est nécessaire d'associer à ces caractéristiques un poids de $-\infty$, une valeur qu'il n'est pas possible d'atteindre dans la configuration sans contraintes du fait de la régularisation. L'utilisation explicite de contraintes revient donc, dans ce cas, à ignorer la régularisation pour une certaine classe de caractéristiques, ce qui peut donc conduire à du surapprentissage.

5 Expériences

Nous considérons dans cette section expérimentale deux tâches d'analyse morpho-syntaxique. En premier lieu, nous étudions en détail l'analyse morpho-syntaxique de l'allemand en considérant différents jeux de caractéristiques et d'étiquettes possibles, et en utilisant des contraintes extraites de différentes manières du corpus d'entraînement (section 5.1). Cette première tâche nous permet d'étudier les phénomènes de manière précise et contrôlée. En second lieu, nous nous intéressons à l'analyse morpho-syntaxique par transfert cross-lingue, dans laquelle les contraintes de type apparaissent naturellement. Dans cette seconde tâche, nous mesurons l'importance que peut prendre l'intégration des contraintes, si l'on souhaite obtenir des performances satisfaisantes.

5.1 Analyse morpho-syntaxique supervisée

5.1.1 Conditions expérimentales

Corpus et tâches On s'intéresse dans cette section à la tâche d'analyse morpho-syntaxique de l'allemand à partir de données annotées. Nous utilisons le corpus arboré TIGER (Brants *et al.*, 2004) avec le même partitionnement que Fraser *et al.* (2013), contenant 50 472 phrases, soit 888 238 mots étiquetés avec leur catégorie morpho-syntaxique. Les étiquettes pour cette tâche sont structurées en différents champs : la catégorie syntaxique (CS) pouvant prendre 54 valeurs possible, ainsi que des traits morphologiques (CM) : cas, nombre, genre, personne, temps, mode, pouvant prendre respectivement 4, 2, 3, 3, 2, 3 valeurs⁴. Ainsi, le mot *legendären* peut être étiqueté $cs=ADJ, cas=gen, num=sg, gen=masc, pers=X, tmp=X, mode=X$. Sur les 1 373 étiquettes possibles, 619 sont observées sur le corpus d'apprentissage. Nous étudions les tâches consistant à prédire l'étiquette syntaxique (CS) et l'étiquette complète (CS+ CM).

2. Mais comme on peut s'y attendre, nous avons observé alors de très mauvaises performances, allant jusqu'à 90% d'erreurs.

3. Ces caractéristiques font typiquement partie des modèles, ce qui montre que ces derniers sont capable d'encoder implicitement les contraintes.

4. Ainsi que les valeurs « non applicable » et « ambigu » que l'on traite ici comme des catégories à part entière.

contraintes			MaxEnt				CRF d'ordre 1			
type	appr.	test	global	MDV	MHV	amb	global	MDV	MHV	amb
X	X	X	10.7	6.6	49.8	10.9	2.9	2.2	9.4	3.0
corpus	X	✓	10.7	6.6	49.8	10.9	2.9	2.3	8.8	3.0
	✓	✓	15.5	6.6	100.0	11.0	8.2	2.6	61.4	3.5
corr.	X	✓	10.7	6.6	49.8	10.9	2.4	1.7	9.0	2.8
	✓	✓	15.4	6.5	100.0	11.0	7.7	2.1	61.6	3.4
oracle	X	✓	6.1	6.6	1.0	10.9	1.6	1.7	0.4	2.8
	✓	✓	6.0	6.5	1.1	11.0	1.6	1.7	0.4	2.9

TABLE 2: Taux d'erreur (%) pour les modèles MaxEnt et CRF entraînés de façon supervisée pour la tâche d'analyse morpho-syntaxique sur le corpus TIGER, en fonction de différentes contraintes considérées à l'apprentissage (appr.) et/ou au test (test) : aucunes (**X**) ; contraintes de type extraites du corpus d'apprentissage (corpus) ; corrigées en utilisant le corpus de test (corr.) ; et complétées (oracle). Le taux d'erreur est donné en prenant en compte tous les mots (global) ; les mots dans le vocabulaire (MDV) ; les mots hors vocabulaire (MHV) ; et les mots ambigus (amb).

Modèle Nous utilisons un CRF linéaire avec différents jeux de caractéristiques qui sont décrits par la suite. L'équation (4) est optimisée en utilisant 30 itérations de l'algorithme de propagation résiliente (Riedmiller & Braun, 1993). Nous utilisons une régularisation \mathcal{L}_1 et \mathcal{L}_2 dont les hyperparamètres sont choisis par *grid search*, pour chaque expérience, dans $\{0, 0.1, 1\}^2$ de manière à maximiser les performances sur le corpus de développement. Différents choix des contraintes de type impliquent un nombre très variable de caractéristiques et choisir la régularisation adaptée à chaque configuration est important pour ne pas interpréter à tort des différences de résultats qui seraient dues à une régularisation inappropriée.

Contraintes de type Nous envisageons trois manières différentes d'obtenir des contraintes de type à partir du corpus annoté : « *corpus* », « *corrigées* » et « *oracle* ». Les contraintes de *corpus* sont obtenues en considérant, pour chaque mot-type, l'ensemble des étiquettes auxquelles il est associé dans le corpus d'apprentissage. Par exemple « amüsiert » a pour seule étiquette ADJ dans le corpus. Cette méthode délivre cependant un dictionnaire incomplet (les mots hors du vocabulaire du corpus d'apprentissage ne sont pas couverts) et incorrect (certains mots ambigus ont pu n'être observés qu'avec une seule étiquette sur les données d'apprentissage). En effet, « amüsiert » apparaît également avec l'étiquette VERB en test. Afin d'étudier l'impact de ces deux problèmes sur les phénomènes étudiés, nous considérons deux conditions oracles, au sens où nous utilisons pour les définir les données de développement et de test. La première consiste à corriger le dictionnaire ainsi extrait : pour chaque mot dans le vocabulaire d'apprentissage, on s'assure que toutes les étiquettes observées en développement et en test sont bien incluses ; si ce n'est pas le cas, on les ajoute (contraintes « *corrigées* »). On associe donc à « amüsiert » les étiquettes ADJ et VERB. Dans la deuxième, on extrait les contraintes sur l'ensemble des données (de développement et de test) et non sur les seules données d'apprentissage (contraintes « *oracle* »). Cela revient également à considérer les contraintes corrigées auxquelles on a également ajouté les contraintes de type pour les mots hors du vocabulaire d'apprentissage.

Évaluation Les performances sont évaluées en utilisant le taux d'erreur standard (rapport du nombre d'occurrences incorrectes sur le nombre total d'occurrences) (global). Afin d'affiner davantage nos analyses, nous donnons aussi les taux d'erreur pour les mots connus (MDV) et pour les mots inconnus (MHV), et au sein de ces derniers, les taux pour les mots ambigus (c'est-à-dire observés dans le corpus d'apprentissage avec au moins deux étiquettes différentes) (amb).

5.1.2 Modèle MaxEnt simple

Nous commençons par un modèle log-linéaire très simple (MaxEnt) comprenant deux patrons de caractéristiques, le premier pouvant tester les associations (mot, étiquette) pour le mot et l'étiquette courante et le second testant l'étiquette courante seule (étiquette). Notons que, dans ce modèle, il n'y a pas de dépendance entre étiquettes.

Les résultats obtenus par le modèle MaxEnt sont détaillés dans le tableau 2. Si, conformément à l'intuition (on prédit toujours l'étiquette la plus fréquente associée à un mot), ajouter les contraintes de type au test ne change rien aux résultats, on

peut s'étonner de l'impact négatif obtenu lorsque celles-ci sont incluses lors de l'apprentissage. Une analyse plus précise des résultats montre que cette baisse de performances est entièrement due aux mots inconnus : lorsque les contraintes oracles (qui incluent les étiquettes de tous les mots du corpus de test) sont considérées, les mots hors-vocabulaire sont systématiquement bien reconnus et les performances avec et sans contraintes au décodage sont équivalentes.

Cette expérience suggère que le principal problème lié à l'introduction des contraintes de type à l'apprentissage est de désambiguïser abusivement de trop nombreuses occurrences. En effet, une grande majorité des mots-formes du corpus d'apprentissage ne présentent pas d'ambiguïté et sont donc complètement désambiguïsés par les contraintes de type. Pour ces exemples, on a $p_{\theta}^a(\mathcal{Y}^n(\mathbf{x})|\mathbf{x}) = 1$ et l'occurrence ne contribue plus au gradient ni à l'optimisation de l'équation (4), ce qui implique qu'aucun paramètre n'est mis à jour. Dans le cas présent, cela entraîne en particulier que les paramètres relatifs aux *a priori* des catégories ne sont plus calculés que sur les mots ambigus. Or, les mots inconnus sont souvent plus proches des mots rares, eux-mêmes le plus souvent non-ambigus. Ainsi, l'étiquette associée à la caractéristique ayant le plus fort poids en l'absence de contraintes à l'apprentissage est NOUN, correspondant à 50% des mots inconnus, alors que l'étiquette ayant le plus grand *a priori* en appliquant les contraintes lors de l'apprentissage devient APPRART⁵, qui ne correspond à aucun mot inconnu. On retrouve le fait que le filtrer des étiquettes équivaut à relâcher la régularisation sur certaines caractéristiques, ce qui peut conduire à sous-apprendre d'autres caractéristiques utiles, ici les caractéristiques relatives aux *a priori* des étiquettes.

5.1.3 Prédiction de la catégorie syntaxique avec un CRF

On considère maintenant un modèle CRF d'ordre 1 comportant un jeu de caractéristiques standard. Pour les mots courant, précédent et suivant, on considère : le mot en minuscule, ses préfixes jusqu'à une taille de 5, ses suffixes jusqu'à une taille de 2, s'il est en majuscule, s'il contient un trait d'union, s'il ne contient que des nombres, s'il contient un chiffre, sa forme obtenue en identifiant majuscules, minuscules et symboles, avec et sans répétitions (par exemple pour 'États-Unis' on a 'Xxxx.Xxxx' et 'Xx.Xx'). On considère également les associations des mots courant et précédent, des mots courant et suivant. Toutes ces caractéristiques sont considérées conjointement avec chaque étiquette possible, ce à quoi on ajoute l'étiquette courante seule et les bigrammes associant l'étiquette courante et les étiquettes suivante et précédente.

Les résultats obtenus par ce modèle sont dans le tableau 2 et sont au niveau de l'état de l'art (Müller *et al.*, 2013). Comme pour le modèle MaxEnt, les mots hors-vocabulaire constituent une part importante des erreurs. À nouveau, l'ajout des contraintes de type apprises sur le corpus d'apprentissage n'améliore pas les performances. En effet, comme illustré à la section 4.3, les caractéristiques (mot, étiquette) permettent d'apprendre les mêmes contraintes de manière endogène au modèle : on voit ici que cela est fait sans erreur. On observe, cependant, que corriger les contraintes issues de l'apprentissage permet d'obtenir des gains substantiels (réduction des erreurs de 2.9% à 2.4%). Cependant, que l'on corrige ou non les contraintes, les utiliser lors de l'apprentissage multiplie le taux d'erreur par un facteur d'environ trois. Le résultat paradoxal observé à la section 2 n'est donc pas spécifique au cadre du transfert cross-lingue ou de l'apprentissage partiellement supervisé. Ici encore, la dégradation observée pour les MHV explique une grande partie de la baisse des performances. On observe toutefois également une dégradation pour les mots ambigus présents dans le vocabulaire d'apprentissage. Contrairement à l'intuition initiale, réduire les candidats possibles pour permettre au modèle de n'avoir à discriminer que les étiquettes plausibles n'apporte, dans cette expérience du moins, aucun avantage. De manière intéressante, le phénomène disparaît dans la condition « oracle ». Comme le modèle est le même pour ces contraintes et pour les contraintes corrigées (puisque la seule différence est la prise en charge des MHV au test), on en conclut que savoir désambiguïser correctement les mots inconnus permet également de mieux prédire des mots voisins connus mais ambigus.

Deux hypothèses, mutuellement non-exclusives, peuvent expliquer ces résultats. La première, déjà évoquée à la section 5.1.2, met l'accent sur les mots complètement désambiguïsés par les contraintes de type à l'apprentissage. En effet, ces mots sont alors ignorés, alors que leurs statistiques et surtout les caractéristiques qu'ils partagent avec d'autres occurrences pourraient être utiles à d'autres endroits. Une seconde hypothèse est que l'introduction de contraintes rend les conditions d'apprentissage et de tests différentes, puisqu'à l'apprentissage tous les mots sont connus, ce qui n'est pas le cas au test. Cette incohérence entre l'apprentissage et test pourrait également contribuer à la dégradation des performances.

Pour tester ces deux hypothèses, nous avons effectué deux expériences de contrôle. La première essaie de résoudre le second problème en introduisant des mots inconnus lors de l'apprentissage. Les mots rares (c'est-à-dire de fréquence faible) ont souvent un comportement syntaxique proche des mots inconnus (Jurafsky & Martin, 2000, chap. 6). Nous proposons donc de ne pas utiliser les contraintes de type pour ces mots rares et de leur assigner, uniquement pour l'apprentissage, l'ensemble des étiquettes possibles. Dans nos expériences, nous considérons qu'un mot est rare si sa fréquence d'appa-

5. Préposition avec article.

contraintes	CS			CS + CM		
	global	MHV	amb	global	MHV	amb
χ	2.9	8.8	3.0	?	?	?
hapax10	3.0	9.3	3.1	?	?	?
hapax5	3.0	9.4	3.1	?	?	?
hapax1	3.2	11.2	3.1	14.4	37.2	14.0
min10	3.2	10.9	3.1	16.6	45.7	14.9
min4	3.3	12.8	3.0	17.5	53.4	15.2
min2	3.6	15.6	3.1	18.1	58.6	15.3
corpus	8.2	61.4	3.5	19.9	74.7	15.8

TABLE 3: Taux d’erreur (en %) d’un CRF supervisé pour la tâche d’analyse morpho-syntaxique sur le corpus TIGER avec le jeu d’étiquettes syntaxiques seules (CS) ou syntaxiques et morpho-syntaxiques (CS + CM), en considérant à l’apprentissage les contraintes de type : uniquement pour les mots de fréquence supérieur à 10 (hapax10), 5 (hapax5), 1 (hapax1); en s’assurant que toute position comprend un minimum d’étiquettes (min10, min4, min2); ou pour tous (corpus). Lors du test on utilise systématiquement les contraintes *corpus*. Pour la tâche d’analyse morpho-syntaxique complète, il n’est, dans certains cas, pas possible de faire l’expérience en un temps raisonnable.

rition est inférieure à un (hapax1), à cinq (hapax5) ou à dix (hapax10). Le tableau 3 montre que cette heuristique permet partiellement de résoudre le problème observé. Pour le modèle simple (MaxEnt), cette heuristique suffit à ramener le modèle avec contraintes de type à l’apprentissage au même niveau que le modèle sans contrainte. Pour les modèles CRF, plus riches en caractéristiques, la dégradation est faible dans le cas des conditions hapax5 et hapax10. On observe encore une fois que l’amélioration des performances résulte principalement d’un meilleur traitement des mots inconnus.

Le problème de l’approche précédente est que pour les mots rares, toutes les étiquettes sont considérées, ce qui reste problématique dans des tâches où cet ensemble est très grand. Comme la difficulté semble surtout provenir des mots complètement désambiguïsés à l’apprentissage, dans une deuxième expérience, nous ajoutons pour chaque mot complètement désambiguïsé un certain nombre d’étiquettes aléatoires de manière à s’assurer qu’il y a au moins i compétiteurs (min- i) au total (en comptant l’étiquette de référence) à chaque position. Les résultats présentés dans le tableau 3 montrent que les performances obtenues sont bien meilleures que lorsque l’on applique les contraintes de base, et légèrement moins bonnes que lorsque l’on autorise toutes les étiquettes pour les mots rares. Il est enfin possible de n’ajouter des étiquettes aléatoires que pour les mots *rare*s (et désambiguïsés par les contraintes), mais il s’avère que cela détériore légèrement les résultats. Il semble donc que le nombre de concurrents joue également un rôle important pour les performances.

5.1.4 Analyse morpho-syntaxique pour le jeu d’étiquette complet

Nous considérons ensuite la tâche de prédiction de l’étiquette morpho-syntaxique complète. Les étiquettes morpho-syntaxiques sont structurées, au sens où les catégories morphologiques possibles dépendent de la catégorie syntaxique. Une approche possible est donc de découpler le problème en apprenant d’abord les étiquettes syntaxiques, puis en utilisant celles-ci pour filtrer les traits morphologiques (Müller *et al.*, 2013). Dans ce travail, nous nous intéressons toutefois uniquement à la prédiction de l’étiquette morpho-syntaxique complète.

Pour cette tâche, le nombre total d’étiquettes rend prohibitive l’utilisation du modèle CRF précédent, en l’état, et diverses heuristiques doivent être envisagées pour réduire l’espace de recherche. Il est, de plus, nécessaire de limiter le nombre d’étiquettes candidates pour les mots inconnus. Une première approche consiste à se limiter à l’ensemble des étiquettes observées (619 au lieu de 1373), ou bien encore aux étiquettes dites « ouvertes »⁶, ce qui limite les étiquettes possibles à 435, ou enfin, selon l’approche retenue dans cet article, de ne prendre en compte que celles qui sont observées avec des mots rares (de fréquence 1), ce qui ramène ce nombre à 204. Nous considérons les mêmes caractéristiques que pour le modèle de la section 5.1.3, à ceci près que chaque fois que l’on considérerait une caractéristique portant sur une étiquette (catégorie syntaxique), nous considérons maintenant à la fois l’étiquette complète, la catégorie syntaxique et les combinaisons impliquant la catégorie syntaxique et chacune des catégories morphologiques. On aura donc, par exemple, une

6. Estimées en partitionnant les données d’apprentissage et en imposant que la fréquence à laquelle une étiquette est vue avec un nouveau mot soit supérieure à un seuil (e.g. 10^{-4}).

caractéristique testant à la fois la catégorie syntaxique et le cas des étiquettes courante et précédente. À notre connaissance, seuls Müller *et al.* (2013) et Silfverberg *et al.* (2014) ont également utilisé des caractéristiques internes aux étiquettes.

En utilisant les contraintes de type, il est possible d'entraîner un modèle CRF standard sans avoir besoin d'utiliser d'autre heuristique simplificatrice (contrairement, par exemple, à Müller *et al.* (2013)). Les résultats obtenus, dans le tableau 3, montrent que l'on retrouve le même comportement que pour la tâche d'analyse syntaxique simple. Garantir un minimum de compétiteurs à chaque position permet un bon compromis entre vitesse d'apprentissage et performances en généralisation. Ceci est insuffisant cependant pour obtenir les mêmes performances qu'en omettant les contraintes pour les mots rares (hapax1), résultat alors état de l'art pour un modèle d'ordre 1 (Müller *et al.*, 2013), mais un peu plus de dix fois plus lent à entraîner. On peut imaginer augmenter encore les performances au prix d'un entraînement plus long. De meilleures techniques qui permettraient de limiter les dégradations dues à l'introduction de contraintes de type, tout en conservant leur bénéfice computationnel, restent à trouver.

5.2 Analyse morpho-syntaxique ambiguë

La tâche d'analyse morpho-syntaxique de la section 5.1 nous a permis d'étudier le problème dans un cadre bien contrôlé. Dans ce cadre, les contraintes de type, même lorsqu'elles ne sont utilisées qu'au décodage, ne permettent jamais d'améliorer les performances. Il s'avère en fait que le modèle est capable de les apprendre presque parfaitement, et leur seul intérêt provient du gain important en vitesse qu'elles permettent. Ces contraintes étant exclusivement extraites du corpus d'apprentissage lui-même, elles n'apportent toutefois aucune information nouvelle, et peuvent donc être responsables du surapprentissage observé. Nous considérons ici un autre exemple, dans lequel les contraintes de type apparaissent de manière naturelle, sont extraites indépendamment du corpus d'apprentissage et se révèlent utiles pour améliorer les performances.

On considère la tâche d'analyse morpho-syntaxique faiblement supervisée, introduite à la section 2 et décrite en détail dans (Täckström *et al.*, 2013). Nous reproduisons la configuration de Wisniewski *et al.* (2014) en utilisant leurs ressources ainsi que le code fourni⁷. Pour chaque langue, Wisniewski *et al.* (2014) utilisent deux sources de contraintes de type : d'une part un dictionnaire automatiquement extrait de WIKTIONNAIRE et d'autre part des contraintes extraites du corpus d'apprentissage, annoté indirectement à partir de l'anglais à travers des liens d'alignement. Les contraintes extraites des bitextes jouent un rôle analogue aux contraintes extraites des corpus de la section 5.1, alors que les contraintes issues du WIKTIONNAIRE reflètent une connaissance linguistique externe que l'on souhaiterait exploiter. En plus d'être utilisées pour apprendre la référence ambiguë, c'est-à-dire pour construire $\mathcal{Y}^r(\mathbf{x})$, les contraintes de type c peuvent restreindre l'espace de recherche $\mathcal{Y}^c(\mathbf{x})$ à l'apprentissage et au décodage, configuration retenue par Täckström *et al.* (2013) et Wisniewski *et al.* (2014). Les tableaux 1 et 4 montrent pourtant que comme pour l'apprentissage supervisé, inclure les contraintes à l'apprentissage nuit aux performances, avec dans certains cas une différence drastique, par exemple pour le finnois (fi) ou l'indonésien (id) pour lesquels le taux d'erreur est quasiment doublé. En omettant les contraintes de type lors de l'apprentissage, ce simple changement permet d'obtenir un gain moyen de 6.0% sur les langues considérées par rapport au modèle⁸ état-de-l'art de Täckström *et al.* (2013), ce qui montre l'importance de prendre en compte le problème.

La section 5.1 permet de comprendre en quoi les contraintes de type posent problème lors de l'apprentissage. En effet, pour toutes les positions sans lien d'alignement, les étiquettes de référence sont les mêmes que les étiquettes possibles : $\mathcal{Y}^r(\mathbf{x}) = \mathcal{Y}^a(\mathbf{x})$ et donc $p_\theta^a(\mathcal{Y}^r(\mathbf{x})|\mathbf{x}) = 1$ (voir l'équation (8)). Le modèle n'apprend donc rien pour cette position. Dans la figure 1, par exemple, le premier mot 'Un' est associé à quatre étiquettes références possibles, qui sont aussi les étiquettes possibles lorsque les contraintes de type sont appliquées à l'apprentissage. Cette observation est toujours vraie si les contraintes de type permettent de désambigüiser complètement un mot. Utiliser les contraintes de type revient donc à ignorer une grande partie des exemples. Une solution possible est alors d'utiliser à l'apprentissage les contraintes de type uniquement si les étiquettes ainsi restreintes sont strictement plus nombreuses que les étiquettes de référence. Par exemple, dans la figure 1, pour la première position, 'Un' possède quatre étiquettes références possibles ; on utilise donc l'ensemble des douze étiquettes possibles pour définir l'espace de recherche. En revanche, pour la deuxième position, 'marché', seule

7. <http://perso.limsi.fr/wisniewski/ambiguous>

8. En réalité, à cause d'une erreur d'implémentation, les résultats publiés dans Täckström *et al.* (2013) correspondent au cas où l'on applique aucune contrainte de type pour réduire l'espace de recherche (premières lignes de chaque block dans le tableau 4). Les résultats corrigés ont été publiés par la suite dans un errata disponible ici <http://www.dipanjanadas.com/files/erratum.pdf>. Bien que les auteurs considèrent que les résultats des deux configurations sont semblables, leurs résultats montrent pourtant clairement une différence de l'ordre de 2% en moyenne pour le cas des contraintes bitexte seules et de l'union, mais pas dans le cas des contraintes issues de WIKTIONNAIRE seules. Nous observons dans nos expériences une différence pour ce dernier cas également, que nous attribuons à la manière dont sont extraits les dictionnaires par Wisniewski *et al.* (2014) qui utilisent toutes les informations de forme de WIKTIONNAIRE, et donc obtiennent des contraintes de type plus puissantes (et donc plus dangereuses à l'apprentissage).

contraintes	appr.	test	cs	de	el	es	fi	fr	id	it	sv
bitexte	✗	✗	17.3	13.6	17.0	14.8	19.2	14.3	14.8	13.5	12.4
	✗	✓	17.3	12.3	17.5	14.4	18.1	14.9	15.0	13.3	12.8
	⊕	✓	17.2	12.4	18.3	14.7	18.8	18.6	16.0	13.4	13.3
	✓	✓	23.3	17.2	23.8	19.9	34.3	24.9	30.2	15.2	19.4
wiki	✗	✗	7.8	9.5	8.3	11.4	12.6	9.8	11.2	9.5	9.7
	✗	✓	7.3	8.2	9.8	9.4	10.9	9.7	11.2	9.8	9.3
	⊕	✓	7.3	9.0	14.5	9.8	11.4	9.6	12.2	12.4	9.6
	✓	✓	8.8	10.7	16.9	10.3	12.1	10.9	13.9	13.4	10.1
wiki ∩ bitexte	✗	✗	8.3	9.7	8.4	11.2	12.7	10.0	11.1	9.4	9.5
	✗	✓	8.0	8.4	9.9	9.2	10.5	10.3	11.3	9.8	9.6
	⊕	✓	8.0	8.8	12.6	9.3	11.4	11.9	11.9	10.8	9.7
	✓	✓	12.8	13.2	14.0	12.0	22.4	14.7	20.5	14.7	14.6

TABLE 4: Taux d’erreur (%) obtenus par un modèle CRF partiellement observé sur la tâche d’analyse morpho-syntaxique par transfert cross-lingue, selon que l’on utilise : aucune contrainte (✗) ; les contraintes de type uniquement lorsqu’elles sont différentes des contraintes de référence ⊕ ; ou les contraintes de type pour tous les mots (✓) pour définir l’espace de recherche à l’apprentissage (appr.) et/ou au test (test). Les contraintes de type sont obtenues en combinant un dictionnaire tiré des bitextes (bitexte) et un dictionnaire issue d WIKTIONNAIRE (wiki) (voir le tableau 1 pour le cas de l’union).

l’étiquette NOUN est référence, on peut donc utiliser les contraintes de type et laisser le modèle apprendre à préférer NOUN à VERB uniquement.

Cette stratégie, indiquée par le symbole ⊕ dans le tableau 4, ne permet en fait pas d’améliorer les performances. Au contraire, elle les dégrade pour plusieurs langues. Il semble donc qu’au-delà du fait que les contraintes de type permettent d’accélérer considérablement la vitesse d’apprentissage (d’un facteur 15 environ), elle ne permettent pas de simplifier la tâche du modèle, et même, au contraire, dégradent à nouveau les résultats.

Remarquons enfin que l’impact négatif des contraintes lors de l’apprentissage ne concerne pas seulement les CRF : nous observons le même comportement pour le modèle à base d’historique, HBAL de (Wisniewski *et al.*, 2014), dont la mise à jour est semblable à celle d’un perceptron, entraîné dans les mêmes conditions.

6 État de l’art

L’utilisation de contraintes de type pour l’analyse morpho-syntaxique a surtout été proposé dans le contexte de l’apprentissage non-supervisé (Merialdo, 1994), que ces contraintes soient extraites de corpus (Goldberg *et al.*, 2008; Ravi & Knight, 2009; Naseem *et al.*, 2009) ou issues de ressources externes comme WIKTIONNAIRE (Li *et al.*, 2012).

Dans le cadre de l’apprentissage supervisé, filtrer les candidats possibles lors du décodage pour accélérer la vitesse de l’analyse morpho-syntaxique est une pratique standard (Ratnaparkhi, 1996; Moore, 2014). Hajic (2000) considère différentes manières d’obtenir des dictionnaires de type pour l’analyse morpho-syntaxique, similaires à nos conditions « corpus », « complétées » et « oracle », et constate que l’utilisation de cette dernière permet d’obtenir davantage de gains que n’en procure un accroissement des données d’apprentissage. Plus récemment, Moore (2014) propose une forme de lissage inspirée du lissage Kneyser-Ney utilisé pour les modèles de langues (Chen & Goodman, 1998), qui permet d’augmenter le rappel des contraintes extraites du corpus, et donc se rapprocher de ce que nous avons appelé les contraintes « corrigées ». À notre connaissance, seuls Smith *et al.* (2005), Waszczuk (2012) et Östling (2013) font état d’utilisation explicite des contraintes lors de l’apprentissage supervisé. Östling (2013) utilise une condition qui serait semblable à ce que nous aurions appelé hapax3.

Smith *et al.* (2005); Waszczuk (2012) séparent l’analyse morpho-syntaxique, pour un jeu d’étiquettes complexe, en deux étapes : une étape de *proposition*, pour laquelle on utilise un module externe proposant un certain nombre d’étiquettes — ce qui revient à construire des contraintes de type ; et une étape de *désambiguïsation*, consistant à prédire en contexte la bonne étiquette parmi les propositions — ce qui revient à effectuer l’apprentissage en incluant les contraintes de type à

l'apprentissage. Pour les tâches considérées, il n'est pas envisageable de se passer des contraintes de type⁹.

Müller *et al.* (2013) considèrent une autre manière de filtrer l'espace de recherche, dans le but d'utiliser un modèle CRF d'ordre plus important. Ces auteurs proposent ainsi d'utiliser une cascade de modèles de complexité croissante (Charniak & Johnson, 2005), en réduisant à chaque étape les étiquettes autorisées à chaque position.

Enfin, d'autres manières d'intégrer des contraintes dans un modèle ont été proposées. Dans la régularisation *a posteriori* (Ganchev *et al.*, 2010) la distribution est choisie de manière à maximiser la log-vraisemblance mais également à respecter, en moyenne, certaines contraintes. Un autre cadre permettant d'inclure des contraintes de manière déclarative est celui des modèles conditionnels sous contraintes (Chang *et al.*, 2010, 2012). D'autres manières d'intégrer l'information linguistique dans l'apprentissage supervisé et, plus généralement, de s'interroger sur la meilleure manière de choisir les compétiteurs lors de l'apprentissage, sont l'estimation contrastive (Smith & Eisner, 2005) et ses extensions récentes (Gimpel & Bansal, 2014).

7 Conclusion

Dans cet article, nous avons exploré ce qui nous apparaissait comme un paradoxe, en essayant de répondre à la question suivante : pourquoi l'utilisation de contraintes utiles lors du décodage dégrade les performances lorsque ces mêmes contraintes sont utilisées pour « aider » l'apprentissage ? Nous avons vu que l'intégration des contraintes lors de l'apprentissage conduit à ignorer la contribution de nombreux exemples, à savoir ceux qui seraient pleinement désambiguïsés par les contraintes, et que ceci nuit à la capacité de généraliser à des mots hors-vocabulaire.

Les contraintes de type, permettant d'accélérer considérablement l'apprentissage et le décodage des CRFs, ne peuvent être utilisées telles quelles pendant l'apprentissage sous peine de sévères dégradations des performances, même lorsque ces contraintes sont utilisées lors du décodage. Nous avons proposé quelques pistes permettant de limiter les effets négatifs tout en préservant les bénéfices en temps de calcul, qui est indispensable pour de nombreuses applications.

De manière plus générale, lorsque l'on dispose d'informations linguistiques, il semble important de faire attention à la manière dont on les intègre au modèle, car même une approche en apparence « inoffensive » peut se révéler néfaste pour les performances. En particulier, il peut ne pas être bon d'utiliser cette information, même pertinente lors de l'apprentissage, pour obliger le modèle à tirer au mieux parti du corpus d'entraînement. La meilleure manière d'intégrer une information linguistique externe reste donc une problématique intéressante à étudier.

Références

- BRANTS S., DIPPER S., EISENBERG P., HANSEN-SCHIRRA S., KÖNIG E., LEZIUS W., ROHRER C., SMITH G. & USZKOREIT H. (2004). Tiger : Linguistic interpretation of a german corpus. *Research on Language and Computation*, 2(4), 597–620.
- CHANG M.-W., GOLDWASSER D., ROTH D. & SRIKUMAR V. (2010). Discriminative learning over constrained latent representations. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, p. 429–437, Stroudsburg, PA, USA : Association for Computational Linguistics.
- CHANG M.-W., RATINOV L. & ROTH D. (2012). Structured Learning with Constrained Conditional Models. *Machine Learning*, 88(3), 399–431.
- CHARNIAK E. & JOHNSON M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, p. 173–180, Stroudsburg, PA, USA : Association for Computational Linguistics.
- CHEN S. F. & GOODMAN J. T. (1998). *An Empirical Study of Smoothing Techniques for Language Modeling*. Rapport interne TR-10-98, Computer Science Group, Harvard University.
- FRASER A., SCHMID H., FARKAS R., WANG R. & SCHÜTZE H. (2013). Knowledge sources for constituent parsing of german, a morphologically rich and less-configurational language. *Volume 39, Issue 1 - March 2013*.
- GANCHEV K., GRAÇA J. A., GILLENWATER J. & TASKAR B. (2010). Posterior regularization for structured latent variable models. *J. Mach. Learn. Res.*, 11, 2001–2049.

9. Même avec celles-ci, Smith *et al.* (2005) indiquent des temps d'apprentissage de plusieurs jours.

- GIMPEL K. & BANSAL M. (2014). Weakly-supervised learning with cost-augmented contrastive estimation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1329–1341 : Association for Computational Linguistics.
- GOLDBERG Y., ADLER M. & ELHADAD M. (2008). Em can find pretty good hmm pos-taggers (when given a good start). In *Proceedings of ACL-08 : HLT*, p. 746–754 : Association for Computational Linguistics.
- HAJIC J. (2000). Morphological tagging : Data vs. dictionaries. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.
- JURAFSKY D. & MARTIN J. H. (2000). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ, USA : Prentice Hall PTR.
- LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, p. 282–289 : Morgan Kaufmann, San Francisco, CA.
- LI S., GRAÇA J. A. V. & TASKAR B. (2012). Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12*, p. 1389–1398, Stroudsburg, PA, USA.
- MÉRIALDO B. (1994). Tagging English text with a probabilistic grammar. *Computational Linguistics*, **20**(2), 155–172.
- MOORE R. (2014). Fast high-accuracy part-of-speech tagging by independent classifiers. In *Proceedings of the 25th International Conference on Computational Linguistics : Technical Papers, COLING'14*, p. 1165–1176, Dublin, Ireland : Dublin City University and Association for Computational Linguistics.
- MÜLLER T., SCHMID H. & SCHÜTZE H. (2013). Efficient Higher-Order CRFs for Morphological Tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, number October, p. 322–332, Seattle, Washington, USA : Association for Computational Linguistics.
- NASEEM T., SNYDER B., EISENSTEIN J. & BARZILAY R. (2009). Multilingual part-of-speech tagging : Two unsupervised approaches. *Journal of Artificial Intelligence Research*, **36**.
- ÖSTLING R. (2013). Stagger : an open-source part of speech tagger for swedish. *Northern European Journal of Language Technology (NEJLT)*, **3**, 1–18.
- RATNAPARKHI A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing, EMNLP'96* : Association for Computational Linguistics.
- RAVI S. & KNIGHT K. (2009). Minimized models for unsupervised part-of-speech tagging. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, ACL '09*, p. 504–512, Stroudsburg, PA, USA : Association for Computational Linguistics.
- RIEDMILLER M. & BRAUN H. (1993). A direct adaptive method for faster backpropagation learning : The RPROP algorithm. In *Proc. ICNN*, p. 586–591.
- SILFVERBERG M., RUOKOLAINEN T., LINDÉN K. & KURIMO M. (2014). Part-of-speech tagging using conditional random fields : Exploiting sub-label dependencies for improved accuracy. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, p. 259–264 : Association for Computational Linguistics.
- SMITH A. N. & EISNER J. (2005). Contrastive estimation : Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, p. 354–362 : Association for Computational Linguistics.
- SMITH N. A. (2011). *Linguistic Structure Prediction*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.
- SMITH N. A., SMITH D. A. & TROMBLE R. W. (2005). Context-Based Morphological Disambiguation with Random Fields. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, p. 475–482, Vancouver, British Columbia, Canada : Association for Computational Linguistics.
- TÄCKSTRÖM O., DAS D., PETROV S., MCDONALD R. & NIVRE J. (2013). Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging. *Transactions of the Association for Computational Linguistics*, **1**, 1–12.
- WASZCZUK J. (2012). Harnessing the CRF Complexity with Domain-Specific Constraints. The Case of Morphosyntactic Tagging of a Highly Inflected Language. In *Proceedings of COLING 2012*, number December 2012, p. 2789–2804, Mumbai, India : The COLING 2012 Organizing Committee.
- WISNIEWSKI G., PÉCHEUX N., GAHBICHE-BRAHAM S. & YVON F. (2014). Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1779–1785, Doha, Qatar : Association for Computational Linguistics.

Analyse d'expressions temporelles dans les dossiers électroniques patients

Mike Donald Tapi Nzali¹ * Aurélie Névéol¹ Xavier Tannier^{2,1}

(1) LIMSI-CNRS, Campus Universitaire d'Orsay, bât 508, 91405 ORSAY, FRANCE

(2) Université Paris-Sud, 91403 ORSAY, FRANCE

prenom.nom@limsi.fr

Résumé. Les références à des phénomènes du monde réel et à leur caractérisation temporelle se retrouvent dans beaucoup de types de discours en langue naturelle. Ainsi, l'analyse temporelle apparaît comme un élément important en traitement automatique de la langue. Cet article présente une analyse de textes en domaine de spécialité du point de vue temporel. En nous appuyant sur un corpus de documents issus de plusieurs dossiers électroniques patient désidentifiés, nous décrivons la construction d'une ressource annotée en expressions temporelles selon la norme TimeML. Par la suite, nous utilisons cette ressource pour évaluer plusieurs méthodes d'extraction automatique d'expressions temporelles adaptées au domaine médical. Notre meilleur système statistique offre une performance de 0,91 de F-mesure, surpassant pour l'identification le système état de l'art HeidelTime. La comparaison de notre corpus de travail avec le corpus journalistique FR-Timebank permet également de caractériser les différences d'utilisation des expressions temporelles dans deux domaines de spécialité.

Abstract.

An analysis of temporal expressions in Electronic Health Records in French

References to phenomena occurring in the world and their temporal characterization can be found in a variety of natural language utterances. For this reason, temporal analysis is a key issue in natural language processing. This article presents a temporal analysis of specialized documents. We use a corpus of documents contained in several de-identified Electronic Health Records to develop an annotated resource of temporal expressions relying on the TimeML standard. We then use this corpus to evaluate several methods for the automatic extraction of temporal expressions. Our best statistical model yields 0.91 F-measure, which provides significant improvement on extraction, over the state-of-the-art system HeidelTime. We also compare our medical corpus to FR-Timebank in order to characterize the uses of temporal expressions in two different subdomains

Mots-clés : Extraction d'Information ; Analyse Temporelle ; Développement d'un Corpus Annoté.

Keywords: Information Extraction, Temporal Analysis, Development of Annotated Corpus.

1 Introduction

1.1 Contexte et motivation

Des informations importantes sur des questions de santé publique se trouvent dans les dossiers électroniques patient rédigés en langue naturelle. Ces dossiers patient constituent un corpus considérable qui croît au rythme de plusieurs centaines de milliers de documents chaque année pour un seul établissement hospitalier. L'analyse rétrospective des parcours de santé permet aux professionnels de santé d'avoir une vue synthétique sur la prise en charge d'une pathologie au sein de leur établissement. Elle permet de comparer les prises en charge effectives avec les recommandations de bonne pratique afin d'évaluer la qualité des parcours de soin, d'identifier des éléments du parcours demandant une prise en charge spécifique, d'améliorer la démarche diagnostique ou thérapeutique. Le but à long terme de ce travail est de développer des méthodes de traitement automatique de la langue biomédicale permettant de faciliter la synthèse d'informations sur les parcours de soin dans le cadre d'analyses rétrospectives. Notre travail porte sur le contenu textuel des dossiers patients, notamment les compte-rendus de séjours et compte-rendus d'actes.

*. Nouvelle affiliation: Université de Montpellier, LIRMM, I3M - Montpellier, France - mike-donald.tapi-nzali@lirmm.fr

Le potentiel des méthodes de traitement automatique de la langue biomédicale (bioNLP) pour exploiter ces documents a été démontré (Demner-Fushman *et al.*, 2009) et fait l’objet de nombreux travaux aussi bien sur l’anglais (par exemple, Friedman *et al.* (1994); Savova *et al.* (2010)) que sur le français (par exemple, Deléger *et al.* (2010); Grouin *et al.* (2011)). Le dossier électronique patient contient plusieurs types de documents (comptes-rendus de séjours, comptes-rendus d’actes, correspondance entre professionnels de santé, ordonnances...) permettant de retracer le parcours du patient dans l’hôpital depuis la première admission, les examens diagnostiques, la prise en charge et le suivi thérapeutique. Un même événement de l’historique médical du patient peut ainsi être mentionné à plusieurs reprises dans divers documents. Il est donc nécessaire de détecter les mentions d’événements ainsi que les mentions d’expressions temporelles afin de pouvoir les relier entre eux et d’agréger les informations associées.

1.2 Travaux antérieurs en analyse temporelle

L’analyse des expressions temporelles dans les textes est une problématique du traitement automatique des langues qui a connu un intérêt grandissant ces dernières années. Les efforts ont d’abord et principalement porté sur les textes journalistiques en langue anglaise, grâce à la création du corpus TimeBank (Pustejovsky *et al.*, 2003), utilisé dans les campagnes d’évaluation TempEval, dont la première édition a eu lieu en 2007 (Verhagen *et al.*, 2007). Plus récemment, ces travaux ont été étendus à d’autres langues (Li *et al.*, 2014; Strötgen *et al.*, 2014a), dont le français (Moriceau & Tannier, 2014), ainsi qu’à d’autres domaines (SMS, textes historiques, résumés d’essais cliniques), montrant des différences intéressantes dans l’extraction et la normalisation des expressions temporelles (Strötgen & Gertz, 2012), ces différences variant d’ailleurs selon la langue étudiée (Strötgen *et al.*, 2014b). Ces constats soulignent le besoin de considérer l’analyse temporelle à la fois sous l’angle du domaine et de la langue considérée.

Le domaine clinique, quant à lui, a été pris en compte à partir de la campagne d’évaluation *i2b2* 2012 (Sun *et al.*, 2013b,a), dans une tâche dédiée à l’extraction de relations temporelles dans des notes cliniques en anglais. Par la suite, d’autres travaux dans ce domaine ont été menés pour les langues anglaise (Jindal & Roth, 2013) et suédoise (Velupillai, 2014), donnant lieu notamment à la création d’un guide détaillé pour la création d’annotations temporelles, ainsi qu’à une étude des spécificités du domaine clinique (Styler IV *et al.*, 2014).

La tâche principale de l’analyse temporelle consiste en l’extraction et la normalisation des expressions temporelles. La normalisation est l’opération de transformation d’une expression (par exemple, “hier” ou “le 1^{er} janvier 2015”) en une représentation formatée et entièrement spécifiée (en particulier, indiquant la valeur absolue des dates relatives ; par exemple, selon le contexte, l’expression “hier” pourrait être normalisée en une date telle que “2015-01-01”). Plusieurs outils réalisant une telle annotation temporelle ont été créés ces dernières années, parmi lesquels on peut citer SUTime (Chang & Manning, 2012), TIMEN (Llorens *et al.*, 2012) pour la langue anglaise, et XIP (Hagège & Tannier, 2008; Bittar & Hagège, 2012) et HeidelTime (Strötgen & Gertz, 2013) pour des adaptations multilingues.

Le travail que nous décrivons ici s’appuie en partie sur l’outil HeidelTime, que nous décrivons donc avec plus de précisions ci-dessous.

1.3 HeidelTime

HeidelTime est un système libre, à base de règles, d’étiquetage d’expressions temporelles, qui a déjà été décliné dans plusieurs langues et, pour l’anglais, dans plusieurs domaines de spécialités (journalistique, scientifique) (Strötgen & Gertz, 2013). Il a notamment obtenu les meilleurs résultats pour l’extraction et la normalisation des expressions temporelles pour l’anglais, dans le contexte des campagnes TempEval-2 et TempEval-3 (Verhagen *et al.*, 2010; UzZaman *et al.*, 2013). HeidelTime est disponible pour 11 langues ¹.

HeidelTime produit des annotations dans le format ISO-TimeML (Pustejovsky *et al.*, 2010), norme devenue un standard en la matière, que nous avons également respectée dans les travaux décrits ici. En particulier, cette norme fait la distinction entre quatre catégories d’expressions temporelles : les *dates*, les *heures*, les *durées* et les *fréquences*.

1. <http://code.google.com/p/heideltime/>

1.4 Objectif et contribution

Dans cet article, nous nous intéressons à la phase d'extraction des expressions temporelles dans des dossiers électroniques patient en français. Nous présentons un nouveau corpus clinique annoté en expressions temporelles normalisées (section 2). Ensuite, nous utilisons cette ressource pour évaluer plusieurs méthodes d'extraction automatique d'expressions temporelles adaptées au domaine médical (sections 3 et 4).

La principale contribution scientifique de ce travail est le développement d'une importante ressource annotée pour le domaine biomédical, ainsi que d'outils d'analyse temporelle nous permettant de traiter des textes du domaine et de caractériser le domaine de spécialité de manière globale. Par ailleurs, notre travail offre une contribution méthodologique en ce qui concerne l'adaptation d'outils d'analyse temporelle à un nouveau domaine de spécialité, notamment en quantifiant la charge de travail nécessaire à l'adaptation pour des méthodes symboliques et statistiques, et les résultats que l'on peut en attendre.

2 Développement du corpus clinique annoté pour les expressions temporelles

Pour constituer le corpus de travail, nous avons sélectionné trois dossiers patient complets, issus de patients du service de Néphrologie ayant subi une greffe de rein. Le choix de dossiers complets est guidé par la perspective à long terme d'étudier les parcours de soin, qui nécessite de pouvoir reconstituer l'historique médical complet d'un patient. Le contexte de transplantation rénale est suffisamment complexe pour fournir des exemples de parcours de soin riches du point de vue chronologique.

2.1 Protocole de développement du corpus

Deux dossiers ont été utilisés comme corpus d'entraînement pour les expériences décrites en section 3, et le troisième comme corpus de test. L'ensemble des 361 documents ont été désidentifiés selon le protocole établi par Grouin & Névélol (2014) : douze types d'informations identifiantes sont marqués par deux annotateurs indépendants. Après adjudication, les informations identifiantes sont remplacées automatiquement par des substituts plausibles en veillant à conserver la cohérence et l'aspect naturel des textes.

Le corpus a ensuite été pré-annoté automatiquement à l'aide de l'outil HeidelbergTime, qui a fourni des annotations automatiques pour les expressions temporelles de type *Date*, *Durée*, *Fréquence* et *Heure*. Dans un deuxième temps, ces annotations ont été révisées manuellement par trois annotateurs, qui ont validé les annotations correctement proposées par HeidelbergTime, corrigé les annotations erronées et ajouté les annotations manquantes. Nous avons également annoté les "signaux" (*Signal*), qui selon la norme TimeML sont des prépositions et conjonctions temporelles, ou des caractères spéciaux qui explicitent les relations temporelles entre entités. Nous anticipons une utilité de ces annotations dans une phase ultérieure du travail, à savoir la détection de relations entre événements et expressions temporelles.

Un premier jeu de 20 documents (issus du corpus d'entraînement), utilisé comme jeu d'entraînement pour l'annotation, a été annoté indépendamment par les trois annotateurs (les auteurs de cet article). Les désaccords ont ensuite été résolus lors d'une réunion de consensus. Le reste du corpus a ensuite été divisé de façon à ce que chaque document soit annoté indépendamment par deux annotateurs, les désaccords étant ensuite résolus dans une réunion de consensus. Dans cette deuxième phase, la pré-annotation utilisée a bénéficié de la mise en place de quelques règles spécifiques au domaine clinique dans l'outil HeidelbergTime comme indiqué en section 3.1.

La Figure 1 présente un extrait du corpus annoté selon la norme TimeML.

Pour réaliser l'annotation, nous avons utilisé l'outil libre BRAT rapid annotation tool (BRAT (Stenetorp *et al.*, 2012)). À la différence d'HeidelbergTime, outil dédié à l'annotation automatique d'expressions temporelles, BRAT est un outil générique qui permet de visualiser des annotations existantes et de les modifier manuellement. Selon une revue récente d'outils d'annotations utilisés dans le domaine biomédical, BRAT présente l'avantage d'être facile à installer et à prendre en main et de permettre l'utilisation de pré-annotations (Neves & Leser, 2012) – dans notre travail, les pré-annotations utilisées sont issues du traitement du corpus par HeidelbergTime.

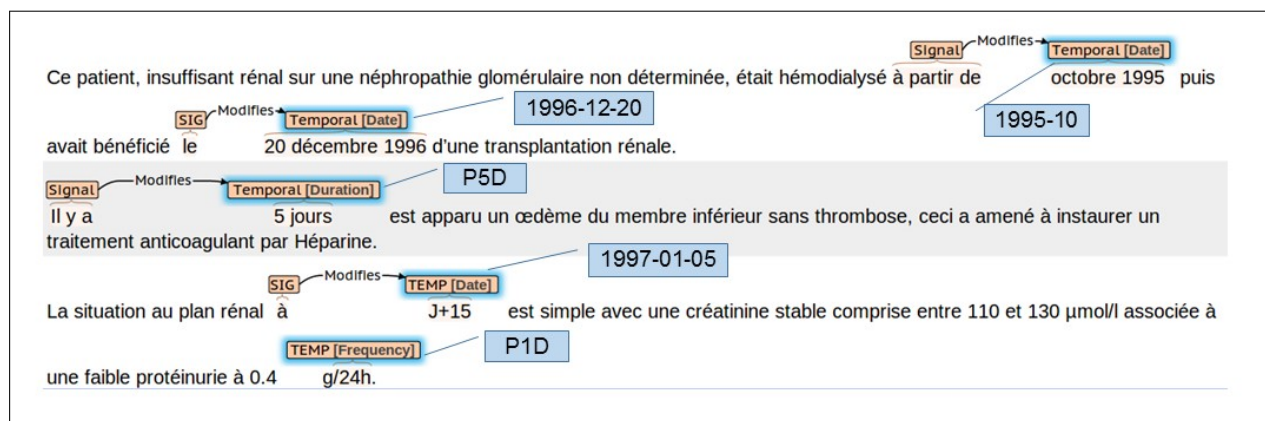


FIGURE 1 – Extrait du corpus clinique annoté en expressions temporelles. Pour cet exemple, nous montrons également la normalisation des expressions qui fait partie du travail en cours (cf section 5). Toutes les informations identifiantes ainsi que les dates ont été remplacées par des substituts plausibles.

2.2 Détails quantitatifs concernant le corpus

La Table 1 présente la distribution des expressions temporelles annotées dans notre corpus de travail, comparée à la distribution des expressions temporelles dans le corpus journalistique FR-TimeBank (Bittar *et al.*, 2011).

Par ailleurs, l'accord inter-annotateur (mesuré en termes de F-mesure) sur le corpus clinique est supérieur à 0,90 pour l'ensemble des paires d'annotateurs. Le temps d'annotation moyen observé est de 25 minutes pour 100 annotations (pour un annotateur), plus 9 minutes consacrées à l'adjudication (lors des réunions de consensus auxquelles participent deux annotateurs). Cela nous permet d'estimer l'annotation en double du corpus complet à environ 45 heures de travail.

	FR-TimeBank		Corpus Clinique	
	#	%	#	%
Date	227	53,41	2594	65,14
Durée	52	12,24	343	8,61
Fréquence	16	3,76	994	24,96
Heure	130	30,59	51	1,28

TABLE 1 – Distribution des expressions temporelles dans deux domaines de spécialité

La comparaison de notre corpus clinique avec le corpus journalistique FR-TimeBank montre qu'il existe une différence de taille considérable entre les deux ressources : 109 documents comprenant un total de 16 208 tokens (soit en moyenne, 148 tokens par document) annotés avec 425 expressions temporelles pour FR-TimeBank et 361 documents comprenant un total de 141 811 tokens (soit en moyenne, 393 tokens par document) annotés avec 3 982 expressions temporelles pour le corpus clinique. Il faut remarquer que la méthode de sélection des documents et la finalité envisagée des deux corpus diffèrent également. Le corpus FR-TimeBank a été construit pour être représentatif du genre journalistique et regroupe sept types d'articles. Le corpus clinique n'a pas cherché à représenter de diversité thématique, mais s'est plutôt attaché à sélectionner des dossiers patients complets afin de permettre à terme une étude des parcours de soin. Par ailleurs, le corpus FR-TimeBank est actuellement beaucoup plus riche que le corpus clinique car il contient des annotations des événements et des relations temporelles en plus des expressions temporelles normalisées.

En terme de distribution des expressions temporelles, on constate d'après la table 1 que les dates et les durées sont présentes dans les deux corpus de manière équivalente. Cependant, la distribution des heures et des fréquences révèle des différences considérables entre les deux domaines de spécialité en présence. En effet, les fréquences sont très présentes dans le corpus clinique (24,96 % vs. 3,76 %) alors que les heures sont prépondérantes dans le corpus journalistique (30,59 % vs. 1,28 %). En pratique, cela s'explique par la nécessité de décrire un événement de l'actualité avec une précision incluant l'horaire de survenue, alors que pour les événements médicaux, une granularité de l'ordre de la journée semble suffisante. En revanche, la prescription de médicaments et de traitements implique d'indiquer la fréquence de prise par le patient, alors que les événements de l'actualité ont lieu ponctuellement.

3 Extraction automatique des expressions temporelles

Nous avons mis en œuvre trois approches pour l'adaptation de l'extraction d'expressions temporelles dans notre corpus de dossiers cliniques patients. La première est un enrichissement manuel des règles de l'outil HeidelTime ; la deuxième est une méthode supervisée de prédiction de séquences à base de CRF (champs aléatoires conditionnels) ; enfin, la troisième est une approche hybride utilisant les deux premières.

3.1 Approche symbolique

L'adaptation d'un système symbolique d'extraction d'information temporelle à un nouveau domaine nécessite un réglage de l'outil pour tenir compte de phénomènes nouveaux ou différents (Strötgen & Gertz, 2012). Nous avons ainsi enrichi les règles francophones d'HeidelTime pour le traitement d'un certain nombre d'expressions temporelles spécifiques au domaine clinique, telles que *J+1* (qui désigne par exemple le lendemain d'une opération chirurgicale), *J11* (qui désigne par exemple le onzième jour après le début d'un protocole de traitement), des expressions particulières de fréquences ou de durée (les abbréviations étant bien plus usitées que dans le domaine journalistique – par exemple, *5 fois/j*). Au total, 14 règles d'extraction et de normalisation ont été ajoutées, et aucune modifiée ou supprimée². Nous estimons le temps de travail pour développer les règles permettant d'adapter HeidelTime au domaine biomédical à environs 8 heures.

3.2 Approche supervisée (CRF)

Les champs aléatoires conditionnels (CRF (Lafferty *et al.*, 2001)) linéaires se sont imposés récemment comme l'une des approches les plus efficaces et robustes pour l'étiquetage supervisé de séquences, appliquée avec succès pour des tâches telles que l'étiquetage morphosyntaxique (Lafferty *et al.*, 2001), l'extraction d'entités nommées (McCallum & Li, 2003), l'extraction d'informations structurées (Pinto *et al.*, 2003).

Les CRF sont des modèles probabilistes graphiques non dirigés conçus pour définir une distributions de probabilités conditionnelles sur des séquences d'étiquettes, étant données des séquences observées. Cette nature conditionnelle démarque les CRF des modèles qui nécessitent une hypothèse d'indépendance des variables, tels que les modèles de Markov cachés (HMM). En pratique, une qualité des modèles CRF est leur robustesse sur des ensembles de données de petite taille. Une introduction détaillée aux CRF peut être trouvée dans Sutton & McCallum (2006)

Nous avons créé un modèle CRF pour la détection des expressions temporelles dans les textes, ainsi que pour leur typage (date, durée, fréquence, heure).

3.2.1 Modèle CRF

Les traits utilisés pour l'apprentissage sont d'ordres morphologique, syntaxique et sémantique.

Traits morphologiques

- Capitalisation du token ;
- Longueur du token ;
- Présence d'un chiffre dans le token ;
- Présence de ponctuation dans le token ;

Traits syntaxiques : Étiquettes morphosyntaxiques. Nous avons utilisé TreeTagger (Schmid, 1994) mais nous avons modifié l'étape de segmentation des mots. En effet, pour tenir compte en particulier du grand nombre d'abréviations présentes dans les documents, et de l'absence fréquente d'espaces entre ces abréviations (*5 fois/j*) ou dans les dates (*21/05/05*), les marques de ponctuation ou les chiffres insérés dans les tokens sont considérés comme des séparateurs de mots.

Traits sémantiques :

- Le token lui-même ;

2. Pour un détail sur le fonctionnement des règles, voir Strötgen & Gertz (2013) ou Moriceau & Tannier (2014) pour le français.

- La présence du token dans des listes de déclencheurs construits au préalable (indices temporelles comme les noms de mois ou de jours de la semaine, ainsi qu’unités de mesures) ;
- Identifiant du cluster de Brown associé au token (Liang, 2005).

3.2.2 Représentation des données

Les données sont représentées selon le format tabulaire BIO standard pour les CRF. Chaque token est étiqueté par les traits décrits ci-dessus ainsi que par une classe de la forme *B-Type* (début d’un segment de classe *Type*), *I-Type* (intérieur d’un segment) ou *O* (extérieur à toutes les classes étudiées), où *Type* prendra les valeurs des catégories d’expressions temporelles considérées (*Date*, *Duration*, *Set*, *Time*).

La figure 2 illustre ce format, qui est à la fois l’entrée et la sortie du modèle CRF utilisé.

Token	POSTreeTagger	Length	IsCapitalized	IsPunctuation	IOB tags
à	PRP	1	O	NO_PUNCT	O
partir	VER	6	mm	NO_PUNCT	O
de	PRP	2	mm	NO_PUNCT	B-Signal
octobre	NOM	4	Mm	NO_PUNCT	B-Date
1995	NUM	4	O	NO_PUNCT	I-Date
,	PUN	1	O	PUNCT	O
puis	ADV	4	mm	NO_PUNCT	O
avait	VER	5	mm	NO_PUNCT	O
[...]					
À	PRP	1	MM	NO_PUNCT	B-Signal
J+1	NAM	3	Mm	NO_PUNCT	B-Date

FIGURE 2 – Exemple de fichier tabulaire au format BIO.

3.2.3 Mise en œuvre expérimentale

Nous avons utilisé l’outil Wapiti (Lavergne *et al.*, 2010) pour la mise en œuvre du modèle CRF, avec l’algorithme RPROP (*resilient backpropagation*), qui donne généralement de meilleurs résultats sur ce type de tâches (Grouin, 2013, p. 158). La valeur de régularisation $L1$ a été déterminée à 1, 8 par validation croisée 10 fois sur le corpus d’apprentissage. Les résultats présentés à la section 4 sont ceux calculés sur le corpus de test à partir du modèle entraîné finalement sur l’ensemble du corpus d’apprentissage.

3.3 Approche hybride

L’approche hybride consiste à utiliser à la fois les résultats de la méthode symbolique (HeidelTime) et de la méthode statistique. Notre choix a été d’utiliser simplement la sortie d’HeidelTime comme trait supplémentaire dans notre modèle CRF, en utilisant le format BIO décrit ci-dessus.

3.4 Évaluation

Pour chacune des trois méthodes, nous avons utilisé deux dossiers (soit 246 documents) comme corpus d’entraînement, et un dossier (soit 115 documents) comme corpus de test.

Les performances de l’extraction d’expressions temporelles ont été mesurées en terme de précision, rappel et F-mesure pour chaque type d’expression temporelles. Nous avons également calculé la micro-moyenne sur les catégories afin d’obtenir une mesure de performance globale.

4 Résultats

Dans cette section, nous présentons le résultat de nos expérimentations sur l'extraction d'expressions temporelles dans le corpus clinique, puis nous discutons de la question de l'adaptation de méthodes d'extraction d'expressions temporelles en français à un nouveau domaine de spécialité. En raison de la nature sensible du corpus, nous précisons qu'il n'est actuellement pas possible de le diffuser librement. En revanche, le jeu de règles spécifiques ajouté dans Heidelberg sera disponible dans la prochaine mise à jour de la distribution du logiciel.

4.1 Extraction automatique d'expressions temporelles

La Table 2 montre les performances globales (micro-moyenne) des différents systèmes évalués sur le corpus d'entraînement (en validation croisée pour les modèles statistiques) et de test.

	Corpus d'entraînement			Corpus de test		
	Précision	Rappel	F-Mesure	Précision	Rappel	F-Mesure
HeidelTime	0,5167	0,6433	0,5731	0,5291	0,6283	0,5744
HeidelTime avec ajout de règles	0,8280	0,8504	0,8391	0,7666	0,7941	0,7801
CRF	0,9458	0,9745	0,9599	0,8068	0,9231	0,8610
CRF avec traits Heidelberg	0,9470	0,9789	0,9627	0,8837	0,9403	0,9111

TABLE 2 – Performances globales de l'extraction d'expressions temporelles

Nous constatons que les performances des différentes méthodes augmentent avec la quantité d'information temporelle et d'information spécifique au domaine fournie à l'outil : la méthode symbolique Heidelberg standard n'inclut aucune information spécifique au domaine clinique et offre les moins bonnes performances avec une F-mesure de 0,57. Ensuite, Heidelberg enrichi de règles spécifiques au domaine biomédical obtenues en analysant le corpus d'entraînement, ainsi que le modèle CRF simple offre de bonnes performances avec des F-mesures de 0,78 et 0,86, respectivement. Les meilleures performances sont obtenues avec le modèle hybride qui utilise des informations spécifiques au domaine *via* le corpus d'entraînement, et des informations temporelles *via* la sortie d'Heidelberg. Il est intéressant de remarquer que les performances de cet outil sont comparables à l'accord inter-annotateur pour l'annotation d'expressions temporelles, ce qui indique que la limite supérieure de performance que l'on peut attendre d'un tel outil est atteinte.

Par ailleurs, les performances obtenues sur le corpus d'entraînement se transfèrent avec peu de perte sur le corpus de test, ce qui indique d'une part qu'il n'y a pas de surentraînement des méthodes statistiques, et d'autre part que les corpus d'entraînement et de test sont relativement équilibrés, bien que cet aspect n'ait pas été formellement contrôlé au moment de la construction du corpus (car nous avons privilégié la cohésion des parcours de soin en incluant des dossiers complets).

Les Tables 3 et 4 présentent le détail des résultats de Heidelberg (avec les règles spécifiques au domaine clinique) et du modèle CRF hybride sur chaque type d'expression temporelle dans le corpus de test.

	Précision	Rappel	F-mesure
Date	0,9144	0,9334	0,9238
Durée	0,8182	0,8090	0,8136
Fréquence	0,4242	0,8209	0,5593
Heure	0,4688	0,0798	0,1364
Global	0.7666	0.7941	0.7801

TABLE 3 – Performance de l'extraction d'expressions temporelles sur le corpus de test avec l'outil Heidelberg adapté pour le domaine clinique

La table 4 indique que les performances du modèle statistique sont inférieures pour la catégorie *Heure* par rapport aux autres catégories. Il est probable que cette différence s'explique par le faible nombre d'expressions temporelles de ce type dans le corpus (seulement 1,28 %). On constate par ailleurs que les performances obtenues par l'outil Heidelberg sur les expressions temporelles de type *Date* et *Durée* sont comparables à celles du modèle statistique hybride (0,92 vs. 0,95 pour les dates, 0,81 vs. 0,83 pour les durées). En revanche, pour les *Heure* et les *Fréquence*, qui semblent concentrer l'essentiel des différences d'utilisation des expressions temporelles entre le domaine journalistique et le domaine clinique,

	Précision	Rappel	F-mesure
Date	0,9262	0,9774	0,9511
Durée	0,7500	0,9296	0,8302
Fréquence	0,8535	0,8601	0,8568
Heure	0,3750	0,8571	0,5217
Global	0,8837	0,9403	0,9111

TABLE 4 – Performance de l’extraction d’expressions temporelles sur le corpus de test avec le modèle statistique hybride

les performances d’HeidelTime sont d’au moins trente points inférieures à celle du modèle statistique. Il faut également remarquer que l’accord inter-annotateur est moins élevé sur la catégorie *Heure* que sur les autres. Outre le fait que cette catégorie est relativement peu présente dans le corpus clinique, certaines expressions la représentant sont ambiguës par rapport à la catégorie *Fréquence*. Par exemple, l’expression “le soir” sera considérée comme une *Fréquence* dans (1) mais comme une *Heure* dans (2).

(1) Outre le SKENAN et l’ACTISKENAN, le malade prend du LAROXYL 10 gouttes **le soir**.

(2) Le patient a été admis aux Urgences **le soir du 7 juillet**.

4.2 Adaptation à un nouveau domaine de spécialité

Nos expériences montrent qu’il n’est pas possible d’utiliser directement des outils d’extraction d’expressions temporelles conçus pour le domaine journalistique à un nouveau domaine de spécialité. En cela, notre travail sur le français rejoint les conclusions de Strötgen *et al.* (2014b) sur l’anglais.

Cependant, nous montrons qu’il est possible d’adapter l’extraction d’expression temporelles à un nouveau domaine de spécialité grâce à un effort modéré (charge de travail estimée à 8 heures). En effet, l’ajout de 14 règles à l’outil HeidelTime (qui en comporte déjà 154) suffit à augmenter ses performances sur le domaine clinique de 20 points de F-mesure, et d’obtenir des résultats tout à fait honorables (F-mesure de 0,78). Notons cependant que le corpus n’est représentatif que d’une spécialité médicale et d’un hôpital ; même si les règles ajoutées ne sont pas spécifiques à ces paramètres, il est possible que d’autres pratiques médicales fassent émerger des expressions temporelles de nature différente.

La mise en œuvre de méthodes statistiques demande un effort supplémentaire plus conséquent avec le développement d’un corpus annoté. Il faut néanmoins souligner que les traits utilisés par nos modèles CRF sont entièrement génériques et ne comportent aucune information spécifique au domaine biomédical qui ne soit pas transposable directement à un autre domaine (typiquement, les clusters de Brown peuvent être obtenus à partir d’un corpus de tout domaine de spécialité). Ainsi, nous pouvons faire l’hypothèse que, moyennant la disponibilité d’un corpus d’entraînement, notre modèle statistique est transposable à un autre domaine.

5 Conclusion et perspectives

Ce travail a permis de développer une nouvelle ressource annotée en expressions temporelles pour un domaine de spécialité, le domaine clinique. À partir de cette ressource, nous avons construit un modèle hybride de reconnaissance des expressions temporelles qui offre de très bonnes performances, comparables à l’accord-inter annotateur.

Une perspective à court terme de ce travail est d’étendre l’annotation et la reconnaissance des expressions temporelles à la normalisation des expressions, selon la norme TimeML. L’enrichissement du corpus avec la normalisation des expressions est en cours. Nous souhaitons ensuite tester la généralisabilité des méthodes à d’autres sous-domaines, et nous intéresser à la détection des événements et des relations entre événements et expressions temporelles.

Remerciements

Nous remercions le Service d’Informatique Biomédicale (SIBM) ainsi que l’équipe CISMeF du CHU de Rouen qui nous ont permis d’utiliser le corpus LERUDI pour cette étude. Ce travail a bénéficié d’une aide de l’Agence Nationale de la

Recherche portant la référence CABeRneT³ ANR-13-JS02-0009-01.

Références

- BITTAR A., AMSILI P., DENIS P. & DANLOS L. (2011). French timebank : An iso-timeml annotated reference corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : Short Papers - Volume 2*, HLT '11, p. 130–134, Stroudsburg, PA, USA : Association for Computational Linguistics.
- BITTAR A. & HAGÈGE C. (2012). Un annotateur automatique d'expressions temporelles du français et son évaluation sur le TimeBank du français. In *Actes de la conférence TALN 2012*.
- CHANG A. X. & MANNING C. (2012). SUTime : A library for recognizing and normalizing time expressions. In (LREC2012, 2012).
- DELÉGER L., GROUIN C. & ZWEIGENBAUM P. (2010). Extracting medication information from french clinical texts. In *Stud Health Technol Inform*, volume 160, p. 949–953.
- DEMNER-FUSHMAN D., CHAPMAN W. W. & McDONALD C. J. (2009). What can natural language processing do for clinical decision support ? *J Biomed Inform*, **42**, 760–772.
- FRIEDMAN C., ALDERSON P., AUSTIN J., CIMINO J. & JOHNSON S. (1994). A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc*, **1**, 161–174.
- GROUIN C. (2013). *Anonymisation de documents cliniques : performances et limites des m'ethodes symbolique et par apprentissage statistique*. PhD thesis, Université Pierre et Marie Curie (Paris VI).
- GROUIN C., DELÉGER L., ROSIER A., TEMAL L., DAMERON O., VAN HILLE P., BURGUN A. & ZWEIGENBAUM P. (2011). Automatic computation of cha2ds2-vasc score : information extraction from clinical texts for thromboembolism risk assessment. In *AMIA Annu Symp Proc*, p. 501–510.
- GROUIN C. & NÉVÉOL A. (2014). De-identification of clinical notes in french : towards a protocol for reference corpus developpement. In *J Biomed Inform*.
- HAGÈGE C. & TANNIER X. (2008). XTM : A Robust Temporal Text Processor. In *Computational Linguistics and Intelligent Text Processing, proceedings of 9th International Conference CICLing 2008*, p. 231–240, Haifa, Israel : Springer Berlin / Heidelberg.
- JINDAL P. & ROTH D. (2013). Extraction of events and temporal expressions from clinical narratives. *Journal of Biomedical Informatics (JBI)*.
- LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.
- LAVERGNE T., CAPPÉ O. & YVON F. C. (2010). Practical very large scale crfs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 504–513, Uppsala, Sweden.
- LI H., STRÖTGEN J., ZELL J. & GERTZ M. (2014). Chinese temporal tagging with heideltime. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2 : Short Papers*, p. 133–137 : Association for Computational Linguistics.
- LIANG P. (2005). *Semi-supervised learning for natural language*. PhD thesis.
- LLORENS H., DERCZYNSKI L., GAIZAUSKAS R. & SAQUETE E. (2012). TIMEN : An Open Temporal Expression Normalisation Resource. In (LREC2012, 2012).
- LREC2012 (2012). *Proceedings of the Eighth International Language Resources and Evaluation (LREC'2012)*, Istanbul, Turkey.
- MCCALLUM A. & LI W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, CONLL '03*, p. 188–191, Stroudsburg, PA, USA : Association for Computational Linguistics.
- MORICEAU V. & TANNIER X. (2014). French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavík, Iceland.

3. CABeRneT : Compréhension Automatique de Textes Biomédicaux pour la Recherche Translationnelle

- NEVES M. & LESER U. (2012). A survey on annotation tools for the biomedical literature. *Briefings in Bioinformatics*.
- PINTO D., MCCALLUM A., WEI X. & CROFT W. B. (2003). Table extraction using conditional random fields. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, p. 235–242, New York, NY, USA : ACM.
- PUSTEJOVSKY J., HANKS P., SAUR R., SEE A., GAIZAUSKAS R., SETZER A., RADEV D., SUNDHEIM B., DAY D., FERRO L. & LAZO M. (2003). The timebank corpus. *Corpus Linguistics*, p. 647–656.
- PUSTEJOVSKY J., LEE K., BUNT H. & ROMARY L. (2010). ISO-TimeML : An International Standard for Semantic Annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, (LREC'10)*, p. 394–7, La Valette, Malta.
- SAVOVA G., MASANZ J., OGREN P., ZHENG J., SOHN S., KIPPER-SCHULER K. & CHUTE C. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes) : architecture, component evaluation and applications. *J Am Med Inform Assoc*, **17**, 507–513.
- SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*.
- STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). Brat : A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, p. 102–107, Stroudsburg, PA, USA : Association for Computational Linguistics.
- STRÖTGEN J., ARMITI A., VAN CANH T., ZELL J. & GERTZ M. (2014a). Time for more languages : Temporal tagging of arabic, italian, spanish, and vietnamese. *ACM Transactions on Asian Language Information Processing (TALIP)*, **13**(1), 1–21.
- STRÖTGEN J., BÖGEL T., ZELL J., ARMITI A., CANH T. V. & GERTZ M. (2014b). Extending heideltime for temporal expressions referring to historic dates. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 2390–2397 : European Language Resources Association (ELRA).
- STRÖTGEN J. & GERTZ M. (2012). Temporal tagging on different domains : Challenges, strategies, and gold standards. In N. C. C. CHAIR, K. CHOUKRI, T. DECLERCK, M. U. DOGAN, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey : European Language Resources Association (ELRA).
- STRÖTGEN J. & GERTZ M. (2013). Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, **47**(2), 269–298.
- STYLER IV W. F., BETHARD S., FINAN S., PALMER M., PRADHAN S., DE GROEN P. C., ERICKSON B., MILLER T., LIN C., SAVOVA G. & PUSTEJOVSKY J. (2014). Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, p. 143–154.
- SUN W., RUMSHISKY A. & ÖZLEM UZUNER (2013a). Annotating temporal information in clinical narratives. *J Biomed Inform*, **46**, Suppl :S5–12.
- SUN W., RUMSHISKY A. & ÖZLEM UZUNER (2013b). Evaluating temporal relations in clinical text : 2012 i2b2 challenge. *J Am Med Inform Assoc.*, **20**, 806–813.
- SUTTON C. & MCCALLUM A. (2006). An Introduction to Conditional Random Fields for Relational Learning. In L. GETOOT & B. TASKAR, Eds., *Introduction to Statistical Relational Learning*. MIT Press. [http ://people.cs.umass.edu/mccallum/papers/crf-tutorial.pdf](http://people.cs.umass.edu/mccallum/papers/crf-tutorial.pdf).
- UZZAMAN N., LLORENS H., DERCZYNSKI L., VERHAGEN M., ALLEN J. & PUSTEJOVSKY J. (2013). Semeval-2013 task 1 : Tempeval-3 : Evaluating time expressions, events, and temporal relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, p. 1–9 : ACL.
- VELUPILLAI S. (2014). Temporal expressions in swedish medical text – a pilot study. In *Proceedings of BioNLP 2014*, p. 88–92, Baltimore, Maryland : Association for Computational Linguistics.
- VERHAGEN M., GAIZAUSKAS R., SCHILDER F., HEPPLER M. & PUSTEJOVSKY J. (2007). Semeval-2007 task 15 : Tempeval temporal relation identification. In *SemEval-2007 : 4th International Workshop on Semantic Evaluations*.
- VERHAGEN M., SAURI R., CASELLI T. & PUSTEJOVSKY J. (2010). SemEval-2010 - 13 : TempEval-2. In *Proceedings of SemEval workshop at ACL*, Uppsala, Sweden.

Compréhension automatique de la parole sans données de référence

Emmanuel Ferreira Bassam Jabaian Fabrice Lefèvre
Université d'Avignon, CERI-LIA, France
{*prénom.nom*}@univ-avignon.fr

Résumé. La majorité des méthodes état de l'art en compréhension automatique de la parole ont en commun de devoir être apprises sur une grande quantité de données annotées. Cette dépendance aux données constitue un réel obstacle lors du développement d'un système pour une nouvelle tâche/langue. Aussi, dans cette étude, nous présentons une méthode visant à limiter ce besoin par un mécanisme d'apprentissage sans données de référence (zero-shot learning). Cette méthode combine une description ontologique minimale de la tâche visée avec l'utilisation d'un espace sémantique continu appris par des approches à base de réseaux de neurones à partir de données génériques non-annotées. Nous montrons que le modèle simple et peu coûteux obtenu peut atteindre, dès le démarrage, des performances comparables à celles des systèmes état de l'art reposant sur des règles expertes ou sur des approches probabilistes sur des tâches de compréhension de la parole de référence (tests des Dialog State Tracking Challenges, DSTC2 et DSTC3). Nous proposons ensuite une stratégie d'adaptation en ligne permettant d'améliorer encore les performances de notre approche à l'aide d'une supervision faible et ajustable par l'utilisateur.

Abstract.

Spoken language understanding without reference data

Most recent state-of-the-art spoken language understanding models have in common to be trained on a potentially large amount of data. However, the required annotated corpora are not available for a variety of tasks and languages of interest. In this work, we present a novel zero-shot learning method for spoken language understanding which alleviate the need of any annotated or in-context data. Instead, it combines an ontological description of the target domain and the use of a continuous semantic space trained on large amounts of unannotated and unstructured found data with neural network algorithms. We show that this very low cost model can reach instantly performance comparable to those obtained by either state-of-the-art carefully hand crafted rule-based or trained statistical models on reference spoken language understanding tasks (test sets of the second and the third Dialog State Tracking Challenge, DSTC2,DSTC3). Eventually we extend the approach with an online adaptative strategy allowing to refine progressively the initial model with only a light and adjustable supervision.

Mots-clés : Compréhension automatique de la parole, espace sémantique continu, apprentissage sans données de référence, données d'apprentissage hors domaine.

Keywords: Spoken language understanding, continuous semantic space, zero-shot learning, out-of-domain training data.

1 Introduction

Dans un système de dialogue homme-machine, le module de compréhension automatique de la parole (Spoken Language Understanding, SLU) joue un rôle intermédiaire entre le module de reconnaissance de la parole et le gestionnaire de dialogue. Son rôle est d'extraire une liste d'hypothèses de séquence d'étiquettes sémantiques à partir d'une transcription automatique de la requête de l'utilisateur. Actuellement, les systèmes état de l'art pour la compréhension sont basés sur des approches probabilistes et sont appris grâce à différentes méthodes d'apprentissage automatique afin de pouvoir attribuer des étiquettes sémantiques aux entrées des utilisateurs.

Les techniques d'apprentissage supervisé nécessitent un grand nombre de phrases annotées sémantiquement par des utilisateurs experts. Ces corpus annotés sont coûteux (en expertises humaines et en temps de construction) et sont dépendants du domaine d'application (souvent restreint) et de la langue utilisée.

Plusieurs études ont comparé les différentes approches probabilistes pour la compréhension de la parole, e.g. (Hahn *et al.*, 2010; Lefèvre, 2007; Deoras & Sarikaya, 2013). Les approches état de l’art utilisent le plus souvent des modèles statistiques discriminants, tels que les champs aléatoires conditionnels de Markov (conditional random fields, CRF) (Wang & Acero, 2006) ou les réseaux de neurones profonds (Deep Neural Networks, DNN) (Deoras & Sarikaya, 2013). Malgré leurs bonnes performances, ces approches sont très dépendantes des données et sont donc difficilement généralisables.

Pour faire face à cette limitation, plusieurs recherches ont proposé un processus d’annotation non-supervisé, e.g. en se basant sur des allocations latentes de Dirichlet (Camelin *et al.*, 2011). D’autres travaux ont porté sur l’utilisation d’algorithmes d’apprentissage non-supervisé (Tur *et al.*, 2011; Lorenzo *et al.*, 2013) ou semi-supervisé (Celikyilmaz *et al.*, 2011; Hakkani-Tur *et al.*, 2011) pour palier à l’absence de ressources annotées en exploitant notamment le web sémantique pour permettre une recherche de données d’apprentissage supplémentaires afin d’améliorer les performances des classificateurs employés.

Un autre groupe d’études s’est intéressé à proposer des techniques visant à réduire le temps de collecte, de transcription et d’annotation de nouveaux corpus. Par exemple, dans (Gao *et al.*, 2005) ou encore dans (Sarikaya, 2008), il a été proposé de construire en premier lieu un petit corpus pour apprendre un système pilote et d’utiliser ce système pour poursuivre la collecte de nouvelles données. D’autres travaux ont employé des techniques issues de l’apprentissage actif (active learning) pour réduire le temps nécessaire à l’annotation et à la vérification d’un corpus, e.g. (Tur *et al.*, 2003) et (Tur *et al.*, 2005). Plus récemment, plusieurs recherches ont été conduites pour diminuer le coût et l’effort de collecte de données par l’étude de portabilité de systèmes à travers les langues (Lefèvre *et al.*, 2010; Jabaian *et al.*, 2013), et les domaines (Lefèvre *et al.*, 2012).

En outre, (Dauphin *et al.*, 2014) proposent l’adoption d’un algorithme d’apprentissage dit sans données de référence (zero-shot learning) pour une classification sémantique d’énoncés. Cette méthode tente de trouver un lien entre les catégories et les énoncés dans un espace sémantique. Ce dernier est appris par un réseau de neurones profond sur une grande quantité de données non-annotées et non-structurées.

Dans le même esprit, dans cet article, nous présentons une méthode visant à limiter la dépendance aux données annotées par l’utilisation d’un mécanisme similaire. En effet, notre méthode repose sur une description ontologique minimale de la tâche visée et sur l’utilisation d’un espace sémantique continu appris par des approches à base de réseaux de neurones sur des données génériques non-annotées (facilement disponible sur web). Notre étude expérimentale a été menée sur une tâche de compréhension de la parole en utilisant les données de la seconde et de la troisième campagne d’évaluation Dialog State Tracking Challenge¹ (DSTC2 and DSTC3) (Henderson *et al.*, 2014a,b). Nous montrons que la technique proposée offre des performances comparables à celles obtenues par des systèmes à base de règles expertes d’une part et appris sur des données annotées d’autre part.

Cependant, une telle approche est dépendante de la qualité de la description ontologique fournie mais aussi de l’espace sémantique continu considéré (sa capacité à modéliser la richesse sémantique du domaine cible). Pour faire face à ces limites, nous proposons l’ajout d’une stratégie d’adaptation « en ligne ». Cette approche a pour objectif d’introduire une faible supervision dans l’optique de raffiner de façon incrémentale la définition de notre connaissance ontologique et de mieux exploiter l’espace sémantique considéré.

Cet article est organisé comme suit : dans la section 2 nous décrivons la tâche de compréhension de la parole. La section 3 présente les approches proposées pour l’apprentissage sans données de référence puis pour l’adaptation en ligne du système, suivie d’une présentation de quelques travaux connexes. Nous présentons notre étude expérimentale dans la section 5 et nous concluons enfin par quelques remarques et perspectives.

2 Compréhension automatique de la parole

Le rôle du module de compréhension de la parole est d’extraire une séquence de m étiquettes sémantiques, également appelées concepts, $C = c_1, c_2, \dots, c_m$ d’une phrase d’utilisateur de n mots, $W = w_1, w_2, \dots, w_n$. Si classiquement chaque étiquette sémantique c_i est définie par un couple champ/valeur, comme par exemple *food=Italian* ou encore *destination=Boston*, dans cet article, nous adopterons le standard d’annotation sémantique employé dans les corpus en langue anglaise des campagnes d’évaluation DSTC2 et DSTC3 (Henderson *et al.*, 2014a).

Dans ces corpus, les étiquettes sémantiques correspondent à des actes de dialogue de la forme `acttype (champ=valeur)`

1. <http://camdial.org/mh521/dstc/>

où `acttype` représente le nature de l'acte de dialogue considéré, à savoir son intention dialogique (e.g. la confirmation ou la réfutation). Par exemple, la phrase utilisateur « hello i am looking for a french restaurant in the south part of town » sera associée à la séquence d'actes de dialogue suivante « `hello()`, `inform(food=french)`, `inform(area=south)` ».

Les combinaisons possibles de `acttype` (`champ=valeur`) sont déterminées sur la base d'un inventaire ontologique des différents types d'actes de dialogue, des champs ainsi que de leurs valeurs respectives.

Les différents types d'actes de dialogue sont en grande partie indépendants de la tâche visée. Ils peuvent se diviser en quatre grands groupes : ceux ayant pour but de transmettre de l'information (`inform`), ceux représentant différents types de requêtes (`request`, `reqalts`, `reqmore`), ceux relatifs aux confirmations (`confirm`, `affirm`, `negate`, `deny`) et les formules de politesse (`hello`, `thankyou`, `bye`).

L'ensemble des couples champs/valeurs est quant à lui très lié à la tâche de dialogue, chaque couple correspond généralement à une entrée spécifique dans la base de données utilisée pour répondre aux requêtes des utilisateurs (e.g. contraintes de recherche).

3 Apprentissage sans données de référence pour la compréhension automatique de la parole

L'apprentissage sans données de référence (zero-shot learning), proposé pour la première fois dans (Palatucci *et al.*, 2009), correspond à un cas particulier d'apprentissage où certaines valeurs de l'ensemble des sorties possibles, Y , ne sont pas présentes dans l'ensemble d'exemples du corpus d'apprentissage.

Dans cette étude, nous examinons le problème de prédire la séquence d'actes de dialogue d'une phrase utilisateur sans avoir vu au préalable un exemple de phrase utilisateur dans le contexte de l'interaction et donc sans avoir vu un exemple d'actes de dialogue dans ce dit contexte.

Pour ce faire, une source de connaissance sémantique doit être exploitée pour extrapoler ces sorties à partir de leur définition. Notre méthode se base donc sur trois composants principaux :

- un espace sémantique continu noté F qui peut être défini comme un espace de dimension d à même de coder les différentes propriétés des étiquettes sémantiques ;
- une base de connaissances K qui peut être vue comme un dictionnaire d'exemples dans F utilisé pour relier l'espace sémantique à l'espace de sortie du système ;
- l'analyseur sémantique qui extrait une liste ordonnée des meilleures hypothèses de séquence d'étiquettes sémantiques à partir d'un transducteur à états finis représentant l'ensemble des hypothèses pour une phrase utilisateur (scorées par des informations issues de F et de K).

Dans la suite de cette section nous décrivons plus en détails ces différents composants. Cependant, les choix faits quant à leurs implémentations concrètes pour la tâche visée seront donnés dans la partie expérimentale.

3.1 Espace sémantique continu

De récentes avancées sur les réseaux de neurones, ont permis d'envisager l'apprentissage de diverses représentations vectorielles compactes de mots (word embedding) présentant des régularités notables avec les propriétés syntaxiques et sémantiques des mots qu'elles modélisent (Mikolov *et al.*, 2013a; Bian *et al.*, 2014). Des travaux ont déjà pu montrer l'intérêt de considérer ce type de représentation sur différentes tâches de traitement automatique des langues naturelles (Bengio & Heigold, 2014; Clinchant & Perronnin, 2013).

L'objectif du module de compréhension étant d'extraire des informations sémantiques à partir d'entrées utilisateur en langage naturel, l'utilisation d'une telle représentation pour définir l'espace sémantique continu offre des possibilités de généralisation d'un grand intérêt.

De plus, ce type de représentation présente l'avantage de ne pas reposer sur l'exploitation de données liées à la tâche, mais au contraire sur un apprentissage réalisé sur une très grande quantité de données (de large couverture) souvent plus facilement accessible (e.g. dump wikipedia). Cependant différentes techniques, comme celle présentée dans (Zou *et al.*, 2013) par exemple, permettent d'adapter/de transférer le modèle ainsi appris pour une tâche spécifique ou encore pour

une autre langue.

3.2 Base de connaissance sémantique

La base de connaissance sémantique K est définie comme la matrice d'affectation représentant les informations ontologiques du domaine visé, qui dans cette étude se limitent à la liste des étiquettes sémantiques et aux exemples de formes de surface qui leurs sont associées. Dans cette matrice (illustrée Figure 1), chaque ligne correspond à un vecteur d'exemple de dimension d dans F et chaque colonne à une étiquette sémantique. Ainsi la valeur de chaque cellule de la matrice (notée $c_{i,j}$ et appelée valeur d'affectation par la suite) indique s'il existe une éventuelle affectation entre le $i^{\text{ème}}$ vecteur dans l'espace sémantique F et la $j^{\text{ème}}$ étiquette sémantique.

Les exemples (entrées de la matrice) sont obtenus en projetant dans F un certain nombre de formes de surface associées à la description ontologique du domaine. Ces formes de surface peuvent être composées d'un ou plusieurs mots. Par exemple, « what food is served? » pour `request(food)`, « yes » pour `affirm()` ou encore « french food » pour `inform(food=french)`.

Elles peuvent être facilement obtenues automatiquement en se basant sur l'ontologie du domaine (e.g. guide d'annotation), la base de données associée à la tâche (e.g. extraction des valeurs possibles pour chaque champ) ainsi que sur un certain nombre d'exemples illustrant les différents types d'actes de dialogue (e.g. données par un expert). Il est à noter que la méthode employée ne nécessite en aucun cas d'être exhaustive lors de la définition de ces exemples (contrairement à une approche à base de règles expertes - grammaire), en effet le recours à l'espace sémantique F permettra leur généralisation après coup.

Sur la Figure 1, les lignes et colonnes de K sont respectivement étiquetées par les formes de surface et les étiquettes sémantiques pour en faciliter la lecture. Les valeurs d'affectation sont d'abord initialisées par des valeurs binaires, 1 si affectation et 0 sinon. Il est important de noter que nous ne contraignons pas la représentation actuelle de K par une correspondance unique entre une forme de surface et une étiquette sémantique. Ainsi, plusieurs valeurs d'affectation peuvent être mise à 1 sur une même ligne. Par exemple la forme de surface « Paris » pourrait très bien être à la fois affectée à l'étiquette sémantique `inform(location=Paris)` et à `inform(name=Paris)` si un établissement présent dans la base de données avait pour nom Paris.

3.3 Analyseur sémantique

En phase de décodage, pour chaque nouvelle phrase utilisateur que l'on cherche à étiqueter, toutes les séquences de mots contiguës (formes de surface) sont considérées. Par exemple pour la phrase « yeah downtown », trois formes de surface différentes sont extraites : « yeah », « downtown » et « yeah downtown ». Ces formes de surface sont ensuite projetées dans l'espace sémantique F (cercles bleus dans la Fig. 1) pour être comparées aux vecteurs associés aux exemples de la base de connaissance K (croix noires dans la Fig. 1). Pour ce faire un critère de similarité (e.g. similarité cosinus) entre ces vecteurs est employé.

L'algorithme des k plus proches voisins est ensuite utilisé pour associer à chaque forme de surface extraite de la phrase utilisateur une liste ordonnée d'hypothèses sémantiques. Ces dernières sont ensuite utilisées pour construire un transducteur à états finis dans lequel les formes de surface et leurs hypothèses sémantiques sont respectivement les entrées et les sorties des arcs, eux-même pondérées par les distances (dédites des similarités).

Un processus de repondération (e.g. pénalité appliquée pour chaque mot présent sur un arc) permet de régler l'influence de la longueur des formes de surface considérées. L'algorithme du plus court chemin est appliqué sur l'automate à états finis obtenu pour générer des hypothèses ordonnées de séquences d'étiquettes sémantiques (le plus court chemin étant mis en gras sur la Figure 1).

3.4 Adaptation en ligne

La méthode proposée pour une adaptation en ligne permet de mettre à jour les valeurs d'affectation de K en fonction des retours utilisateurs suite à l'interrogation du module de compréhension (en ligne).

En effet, une association directe entre mot (ou séquence de mots) et étiquette sémantique peut être retrouvée dans le

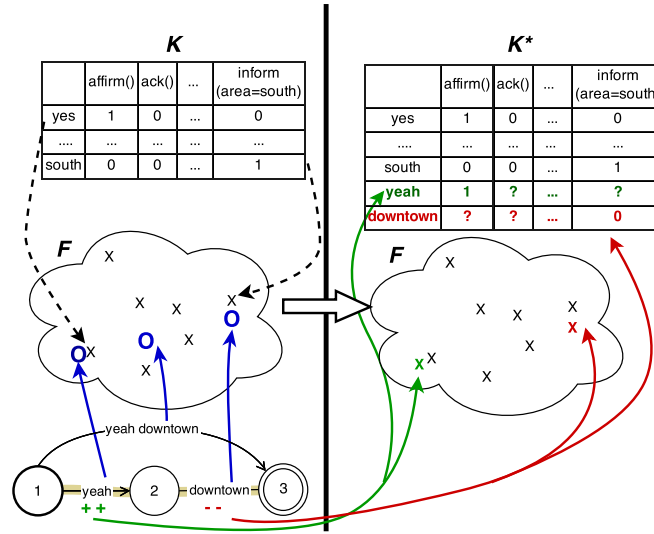


FIGURE 1 – Illustration d'un décodage sémantique basé sur une technique d'apprentissage sans données de référence

transducteur produit par l'analyseur sémantique. Ceci offre la possibilité d'exploiter cette information pour adapter le modèle dynamiquement.

Dans le but de minimiser l'effort de supervision, un scénario dans lequel la supervision est limitée à un ensemble de retours binaires (validation/rejet) sur les étiquettes sémantiques produites est proposé. Compte tenu du fait que ce scénario ne nécessite pas une correction manuelle des étiquettes de la part des utilisateurs, il peut être facilement intégré au sein d'une plate-forme de dialogue existante en utilisant des retours simples de l'utilisateur (oui/non).

Un ratio coût/amélioration peut ainsi être contrôlé en déterminant une politique de demande de retours aux utilisateurs. La définition d'une stratégie optimale pour gérer ce ratio fera l'objet de travaux ultérieurs. Dans l'étude actuelle, nous considérons un utilisateur (simulé) donnant à chaque tour un retour sur chaque étiquette sémantique produite par le système. Ces retours sont ensuite utilisés pour mettre à jour K en K^* .

Un exemple de ce processus est donné dans la Figure 1. Ce dernier illustre un cas où les véritables étiquettes sémantiques de la phrase utilisateur sont mal reconnues par l'analyseur sémantique : ici la phrase « yeah downtown » est étiquetée comme `affirm()`, `inform(area=south)` au lieu de `affirm()`, `inform(area=centre)`.

Les m retours utilisateurs constituent un jeu de m tuples $U = ((c_k, T_k, f_k))_{1 \leq k \leq m}$, où (c_k, T_k) est le couple forme-de-surface/étiquette-sémantique proposé à l'utilisateur et f_k est son retour (1 si positif, 0 si négatif). L'algorithme 1 (partiellement illustré sur la Figure 1) est utilisé pour mettre à jour K en K^* en fonction de K et U .

Chaque cellule (i, j) dans K est constituée de 4 valeurs distinctes : $p_{i,j}$ et $n_{i,j}$ représentant respectivement le nombre de retours positifs et négatifs observés jusqu'alors, $knn_{i,j}$ la valeur obtenue par une addition par élément des k plus proches lignes voisines (repondérées par un produit scalaire des similarités normalisées, voir Algorithme 1.16) et $c_{i,j}$ la valeur d'affectation présentée ci-dessus qui est aussi la valeur utilisée par notre analyseur sémantique. Lors d'une mise à jour l'ensemble des ces valeurs peut être impacté. Ainsi, l'algorithme 1 montre les conditions et la nature de leurs mises à jour mais aussi comment K peut être étendue (ajout d'une nouvelle ligne en présence d'une séquence de mots inconnus, par exemple).

Dans un premier temps, K^* est initialisé avec une copie de K . Puis toutes les nouvelles formes de surface c de U qui ne figurent pas parmi nos exemples connus sont ajoutées dans K^* (cf. Algorithme 1.4-6). Ensuite tous les comptes sur les retours sont mis à jour en se basant sur les informations contenues dans U (cf. Algorithme 1.8-9). Pour ce faire deux facteurs d'échelles α_p et α_n ont été définis afin de permettre d'ajuster l'importance d'une nouvelle observation par rapport aux connaissances courantes au regard de sa valence (on pourrait par exemple choisir de faire plus confiance aux retours positifs). Pour les couples forme-de-surface/étiquette-sémantique initiaux (issus de notre définition ontologique initiale

Algorithm 1 Mise à jour de la base de connaissance K

```

1: Sachant :  $K$  et  $U$  Sortie :  $K^*$ 
2:  $K^* \leftarrow K$ 
3: for all  $(c, T, f) \in U$  do
4:   if  $c \notin K^*$  then
5:     ajouter une nouvelle ligne pour  $c$  dans  $K^*$  avec valeurs de cellule initialisées par défaut
6:      $m_{last} = 1$ 
7:      $i \leftarrow$  identifiant ligne  $c$ ,  $j \leftarrow$  identifiant colonne  $T$ 
8:      $p_{i,j} \leftarrow p_{i,j} + f \times \alpha_p$ 
9:      $n_{i,j} \leftarrow n_{i,j} + (1 - f) \times \alpha_n$ 
10:    if  $p_{i,j} + n_{i,j} > 0$  then
11:       $old_c \leftarrow c_{i,j}$ 
12:       $c_{i,j} \leftarrow \frac{p_{i,j}}{p_{i,j} + n_{i,j}}$ 
13:      if  $c_{i,j} - old_c < 0$  then  $m_i \leftarrow 1$ 
14:    else  $c_{i,j} \leftarrow 0$ 
15:  for all  $c_{i,j} \in K^*$  do
16:    calculer  $knn_{i,j}$ 
17:  for all  $c_{i,j} \in K^*$  do
18:    if  $p_{i,j} + n_{i,j} = 0$  et  $m_i = 1$  then  $c_{i,j} \leftarrow knn_{i,j}$ 

```

du domaine), les valeurs $p_{i,j}$ sont initialisées avec une valeur a priori p_0 .

Dans le cas général, la valeur d'affectation à une étiquette sémantique est obtenue par un simple ratio entre les retours positifs et négatifs associés à la cellule concernée (voir Algorithme 1.12).

Pour chaque modification de ligne, un marqueur m_i est employé afin de détecter si une connaissance à priori (affectation positive) est remise en question par de nouvelles observations (détecté par une baisse de la valeur d'affectation $c_{i,j}$, cf. Algorithme 1.13). Dans ce cas, les affections pour lesquelles il n'y a eu aucune observation pour cette forme de surface (autres cellules sur la même ligne) ont leurs valeurs d'affectation correspondant à la valeur knn à la place de 0. Ainsi, de nouvelles propositions pourront être testées et évaluées par l'utilisateur si la forme de surface venait à se représenter (processus d'exploration de l'espace d'affectation).

4 Travaux connexes

Le problème de l'apprentissage sans données de référence a été déjà abordé par la communauté de l'apprentissage automatique. On peut notamment citer les premiers travaux de Larochelle et al. (Larochelle *et al.*, 2008) qui ont introduit ce type spécifique d'apprentissage pour résoudre une tâche de reconnaissance optique de caractères.

En parallèle les auteurs de (Palatucci *et al.*, 2009) ont proposé une approche similaire pour apprendre un classifieur qui prédit des classes omises dans l'ensemble d'apprentissage. Cet algorithme utilise également une base de connaissances de propriétés sémantiques des classes connues afin d'explorer de nouvelles classes (généralisation).

En tant qu'application de cette technique dans le domaine du traitement naturel de la langue, notre proposition s'inscrit dans une même ligne que la proposition faite dans les travaux de Dauphin et al. (Dauphin *et al.*, 2014). Cependant, la nature et la manière dont nous définissons notre représentation sémantique se distinguent de ces travaux. En effet dans notre cas nous n'utilisons pas de données reliées à notre domaine mais au contraire nous utilisons une représentation généraliste. De plus, la tâche considérée n'est pas identique puisque notre objectif est d'avoir une annotation sémantique complète d'une phrase utilisateur (séquence d'étiquettes) et non pas une simple classification globale de la phrase en catégorie.

Ayant toujours le même objectif de minimiser le besoin de données d'apprentissage coûteuses en temps et en expertises humaines, différentes approches ont déjà été appliquées pour exploiter le web sémantique pour des tâches de classification d'énoncés.

Par exemple, les auteurs de (Heck & Hakkani-Tur, 2012) ont proposé une approche d'apprentissage non-supervisé pour la

compréhension de la parole basées sur l'utilisation des connaissances sémantiques du Web sémantique. Ces propositions reposent sur une combinaison d'un système de recherche d'information du web et d'un analyseur de dépendance basé sur des informations syntaxiques.

Anastasakos et Deoras (Anastasakos & Deoras, 2014) ont également proposé d'exploiter un espace continu pour modéliser les mots. Ils ont proposé une approche pour obtenir des représentations vectorielles spécifiques à des tâches et des domaines précis afin d'apprendre un système de compréhension en utilisant un algorithme d'apprentissage non-supervisé. Ils ont également proposé de transférer ces représentations d'une langue à une autre permettant l'apprentissage d'un système de compréhension multilingue.

Notre technique proposée pour l'adaptation en ligne rejoint également celles de récents travaux ayant pour but d'adapter des modèles pour la tâche de compréhension. Par exemple, dans (Bayer & Riccardi, 2013) une approche basée sur les exemples est proposée pour l'adaptation en ligne du modèle sémantique. Une autre solution présentée dans (Gotab *et al.*, 2010) utilise une méthode supervisée qui permet de mettre à jours les modèles avec une supervision limitée effectuée par les utilisateurs du système.

5 Expérimentations et résultats

5.1 Description de données

Toutes les expériences présentées dans cet article sont basées sur les corpus DSTC2 et DSTC3 (Henderson *et al.*, 2014a,b). Ces corpus ont été construits pour un défi de recherche dédié à la détection du but de l'utilisateur tout au long d'un dialogue oral (et non pas uniquement l'étiquetage sémantique des énoncés de l'utilisateur au fur et à mesure). Cependant, dans notre étude expérimentale, nous exploitons ces données (transcriptions, annotation sémantique, etc.) comme un ensemble de test pour évaluer notre approche d'apprentissage sans données de référence pour l'étiquetage sémantique sur deux configurations de dialogues réalistes.

Le défi DSTC2 couvre le domaine de la recherche d'informations sur des restaurants alors que DSTC3 étend le domaine et couvre également la recherche d'informations touristiques plus générale en incluant notamment des nouveaux types d'établissement (pubs, coffee shops) mais aussi de nouveaux champs et valeurs. Dans notre expérience, seules les données de test de ces deux corpus sont utilisées (9890 énoncés d'utilisateurs pour DSTC2 et 18715 pour DSTC3). Chaque ensemble est évalué en deux modes différents : transcriptions manuelles et n-meilleures transcriptions automatiques des entrées de l'utilisateur.

5.2 Évaluation de l'approche proposée

Afin de constituer notre espace sémantique, un modèle word2vec (Mikolov *et al.*, 2013a) a été utilisé pour apprendre une représentation vectorielle des mots sur 300 dimensions. Ce modèle a été appris avec l'algorithme *Skip-gram* (avec une fenêtre de 10 mots) sur une grande quantité de données² en langue anglaise disponibles librement et présentant une grande couverture thématique.

Ce type de représentation présente certaines régularités avec les propriétés syntaxiques et sémantiques des mots comme celles montrées dans Mikolov (Mikolov *et al.*, 2013b) ainsi qu'une structure linéaire permettant la combinaison des représentations des mots par une simple addition vectorielle élément par élément. Cette technique est donc utilisée pour projeter nos formes de surface vers leur représentation sémantique vectorielle de type word2vec vue comme une somme des représentations individuelles de chaque mot les constituant.

Plusieurs travaux état-de-l'art ont montré que la similarité cosinus est une métrique pertinente pour comparer différents vecteurs de mots word2vec (Mikolov *et al.*, 2013a,b). De ce fait, nous avons également utilisé cette métrique dans l'algorithme de type k plus proche voisins pour la prédiction sur les formes de surface et l'adaptation de la base de connaissance. Ainsi, dans les expériences considérées, $k = 1$ pour l'analyse sémantique et 20 pour les valeurs knn dans la matrice d'affectation. De plus, nous avons utilisé l'algorithme du plus court chemin pour parcourir le graphe sémantique (transducteur) avec une métrique de distance cosinus (voir la section 3.3).

2. enwik9, One Billion Word Language Modelling Benchmark, Brown corpus, English GigaWord de 1 à 5 - soit plus de 4 milliards de mots en contexte

Tâche	Modèle	Entrée	F-score	P	R
DSTC2	S-règles	n-meilleures	0,782	0,900	0,691
	S-appris	n-meilleures	0,802	0,846	0,762
	ZSSP	manuelle	0,919	0,898	0,942
		n-meilleures	0,794	0,796	0,792
DSTC3	S-règles	n-meilleures	0,824	0,852	0,797
	ZSSP	manuelle	0,899	0,873	0,928
		n-meilleures	0,826	0,806	0,849

TABLE 1 – Evaluation des performances de l’analyseur sémantique basé sur l’apprentissage sans données de référence en termes de F-score, **P**récision et **R**appel.

Les bases de connaissances liées à la tâche utilisées dans les expériences sont extraites de la description ontologique du domaine fournie dans le challenge (e.g. listes des champs/valeurs) ainsi que d’un ensemble d’informations de dialogue générique en suivant la procédure automatique décrite dans la section 3.2.

La sémantique du domaine est représentée par 8 champs et 215 valeurs dans DSTC2 et par 13 champs et 279 valeurs dans DSTC3. Pour les deux tâches 16 actype différents sont considérés, en résultent 663 différentes étiquettes sémantiques pour DSTC2 et 855 pour DSTC3.

Nous avons définis manuellement 53 formes de surface pour modéliser les types d’actes de dialogue, par exemple « say again » pour l’acte de demande de répétition. Cet effort est commun aux deux tâches cibles. Dans les deux descriptions ontologiques considérées, les champs et les valeurs ont des noms significatifs (lexicalisés) et ils peuvent être directement utilisés dans les formes de surface (par exemple « address », « french », « has tv »). Au total, 4160 formes de surface ont été ainsi générées complètement automatiquement et sont utilisées pour DSTC2, 6555 pour DSTC3.

Pour évaluer nos propositions, nos résultats sont comparés avec deux systèmes état de l’art : le premier est un système à base de règles expertes utilisé dans le défi DSTC et le second est un système présenté dans (Williams, 2014), appris sur les données d’apprentissage du DSTC2 (nommé SLU1 dans l’article de Williams). Ces deux systèmes sont respectivement référencés par « S-règles » et « S-appris » dans la suite de cet article.

Les résultats de nos expériences (présentés dans le tableau 1) montrent que l’approche proposée, nommé ZSSP (pour Zero-Shot Semantic Parser) par la suite, atteint un niveau de performance (en termes de F-score) légèrement meilleur que celui de l’approche à base de règles (0,794 contre 0,782 sur DSTC2 et 0,826 contre 0,824 sur DSTC3) et comparable à celui d’un modèle appris (0,794 contre 0,802 sur DSTC2). Ainsi le modèle proposé atteint au démarrage des performances état-de-l’art sans utilisation de nombreuses règles spécifiques manuellement établies (coût d’experts humains) ni de données d’apprentissage (coûts de collecte et d’annotation).

Cependant, afin de mesurer l’impact de la représentation sémantique choisie sur la performance globale de l’approche, un système qui n’utilise pas ce type de représentation a été construit. Un F-score de 0,839 (contre 0,919 en configuration normale) est obtenu sur les transcriptions manuelles du DSTC2 par une simple stratégie de détection de patrons de mots à partir des exemples de la même base de connaissances K . Cette dernière observation confirme l’avantage d’avoir recours à une représentation sémantique riche apprise sur une grande quantité de données non annotées. En effet, cette dernière permet une meilleure généralisation des connaissances lexicales initiales (qui elles peuvent être assez limitées).

5.3 Adaptation en ligne

Comme mentionné précédemment, un mécanisme l’adaptation en ligne a également été proposé dans la section 3.4 pour améliorer les performances de l’approche ZSSP. Ainsi, les énoncés transcrits du corpus d’apprentissage du DSTC2 sont utilisés pour simuler des retours de validation utilisateurs et donc pour adapter notre base de connaissance K dynamiquement (en évitant le bruit dû à des erreurs de transcription automatique). Nous utiliserons l’ensemble de test DSTC2 (comme précédemment) pour juger de l’évolution des performances de l’approche ZSSP avec cette mise à jour. Pour positionner notre approche par rapport à l’état de l’art, les mêmes systèmes de référence que précédemment sont utilisés.

Ainsi, pour rendre possible la phase d’adaptation, les évaluations/retours des utilisateurs sont simulées en comparant la meilleure hypothèse du modèle avec l’étiquette sémantique de référence des phrases utilisateurs dans le corpus d’apprentissage DSTC2. Toutes les formes de surface de notre meilleure hypothèse ayant une étiquette sémantique présente dans

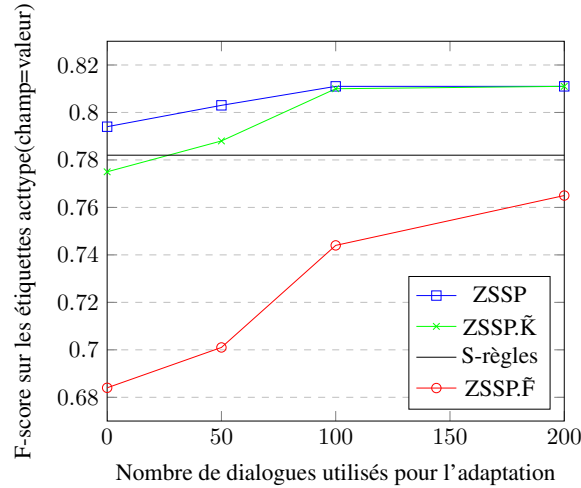


FIGURE 2 – Performances de diverses configurations de la méthode ZSSP en termes de F-score en fonction du nombre de dialogues utilisées pour l'adaptation.

l'annotation de référence sont considérées comme positives et toutes les autres comme négatives. K est mise à jour à la fin de chaque tour en suivant l'algorithme présenté dans la section 3.4 (avec $\alpha_p = \alpha_n = 1$).

Dans le but de quantifier l'influence de l'espace sémantique considéré F et de la base de connaissance initiale K sur l'approche proposée dans ce papier, nous avons fait le choix d'étudier trois configurations différentes de cette dernière. Nous distinguerons donc de l'approche ZSSP classique (base de connaissance K de qualité et un espace sémantique reposant sur une représentation word2vec apprise sur une grande quantité de données) deux variantes : la première, notée ZSSP. \tilde{F} , utilise une représentation sémantique « dégradée » et réduite à 50 dimensions, à savoir une représentation word2vec apprise avec l'algorithme *Skip-gram* (avec une fenêtre de 5 mots) sur des données non annotées issues du corpus d'apprentissage du DSTC2 (190366 mots en contexte) ; la seconde, notée ZSSP. \tilde{K} utilise une version « dégradée » de K où 10% des formes de surface (exemples de types d'actes de dialogues) ont été retirés.

Les résultats présentés dans la figure 2 montrent l'évolution du F-score en fonction du nombre de dialogues utilisés pour l'adaptation. Même avant l'adaptation ZSSP (0, 794) et ZSSP. \tilde{K} (0, 775) atteignent des performances proches d'un système à base de règle (0, 782). Mais un espace sémantique appris sur une petite quantité de données peut avoir un impact significatif sur cette performance (comme montré avec ZSSP. \tilde{F} , 0, 684) dû à la fois à des mots hors vocabulaire et des mauvaises propriétés de généralisation de cet espace sémantique.

Néanmoins, dans toutes les configurations de ZSSP, la performance augmente conjointement avec le nombre de dialogues d'adaptation. En effet, à la fois ZSSP et ZSSP. \tilde{K} obtiennent, après seulement 100 dialogues, des performances nettement meilleures que les modèles de références (0.811 contre 0.782 pour S-règles et 0.803 pour S-appris³).

En outre, l'écart entre ZSSP. \tilde{F} et le modèle à base de règles est nettement réduit tout au long du processus d'adaptation en ligne (de 0, 098 à 0, 017 après 200 dialogues). Cette observation montre que la méthode proposée peut aussi fonctionner avec un espace sémantique bruité. Ces résultats confirment l'avantage de la méthode d'adaptation en ligne proposée pour faire face aux limites de la couverture initiale de K et à la robustesse de l'espace sémantique F .

5.4 Généralisation

L'avantage majeur de l'utilisation d'un modèle word2vec par rapport à un simple modèle de détection par mots clés est l'intégration d'une représentation continue des mots dans le processus de décodage. Cette caractéristique confère au système une capacité de généralisation inhérente permettant de couvrir des mots inconnus correspondant à des valeurs non présentes dans l'ontologie définie du domaine ou de la tâche. Par exemple, dans le contexte d'un domaine de recherche de restaurant, il est intéressant pour un système de dialogue de détecter certaines situations où un utilisateur parle d'un type

3. les performances de S-appris n'ont pas été reportées sur la figure. 2 dans le but d'éviter une possible confusion (au regard de l'axe des abscisses) sachant qu'il utilise beaucoup plus de données d'apprentissage

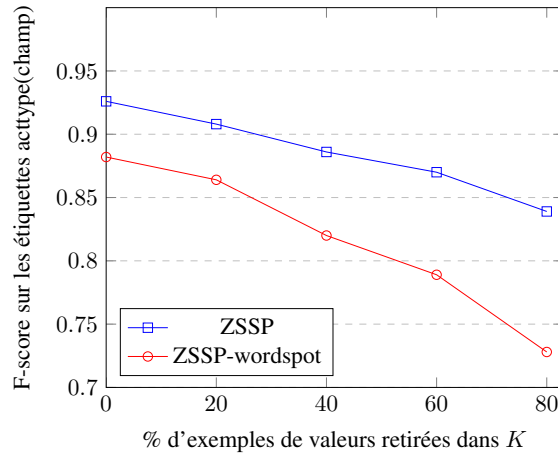


FIGURE 3 – Capacité de généralisation de l’approche ZSSP sur la corpus de test DSTC2 exprimée en termes de F-score sur la détection d’actes de dialogue génériques (i.e. $actype(champ)$) en fonction du pourcentage d’exemples de valeurs retirées dans K

d’aliment inconnu jusqu’alors par le système (si ce dernier n’est pas dans la base de données d’origine) ou au moins être en mesure de proposer une alternative en conséquence (en exploitant par exemple la proximité dans l’espace sémantique).

Afin d’évaluer la capacité de généralisation de notre système, nous avons volontairement supprimé de la base de connaissances de DSTC2 des formes de surface correspondant aux différents pourcentages des valeurs possibles de certains champs spécifiques. Dans cette étude préliminaire, nous avons choisi d’étudier l’impact sur les champs *food*, *area* et *pricerange*. Les performances du modèle sur les transcriptions manuelles ont été évaluées en termes de F-score pour $actype(champ)$ uniquement au lieu de $actype(champ=valeur)$ afin d’évaluer la détection des concepts de haut niveau.

Ainsi, nous comparons la performance de ZSSP avec une autre configuration de l’analyseur, notée ZSSP-wordspot. Ce dernier étiquette uniquement les segments qui atteignent un degré de similarité très élevé (une correspondance quasi parfaite - 0,94). Vu que ce modèle est capable d’exploiter l’espace sémantique, cette configuration peut être assimilée à une stratégie robuste de détection de mots clés.

Les résultats (présentés dans la figure 3) montrent clairement une légère baisse de performances lorsque le pourcentage de valeurs retirées est grand. La différence entre les deux configurations est de 0,044 à 0% et de 0,111 à 80%. Cela confirme que l’approche proposée est tolérante à une faible densité de données dans K . Cette caractéristique peut être utile pour développer un système de dialogue générique permettant une évolution transparente de la base de connaissances contenant une base de données croissante.

6 Conclusions et perspectives

Dans cet article nous avons présenté une approche d’apprentissage sans données de référence pour la compréhension de la parole. Cette dernière repose à la fois sur l’utilisation d’une représentation sémantique riche apprise sur des données généralistes et sur une description ontologique minimale décrivant la tâche de compréhension visée. Nous avons montré que cette approche, bien que peu coûteuse, est tout de même comparable en termes de performances à des méthodes statistiques apprises sur de grande quantité de données annotées et aussi à un système à bases de règles expertes. De plus, la méthode proposée montre une meilleure tolérance à des valeurs de concept manquantes et donc offre des propriétés de généralisation pouvant être employées notamment dans l’extension de domaine en ligne.

De plus nous avons montré qu’un processus d’adaptation simple et ajustable en ligne permet de répondre aux deux limites de l’approche, à savoir la qualité de la base de connaissance K et de l’espace sémantique employé F . L’effort de supervision reste acceptable puisque l’utilisateur se contente de confirmer les hypothèses faites par le système et donc n’est pas contraint de corriger explicitement les erreurs du système. La comparaison avec d’autres techniques d’apprentissage active et la généralisation de cette technique par l’adaptation d’une vision plus probabiliste et dynamique sont planifiées pour de futurs travaux, de même que son évaluation dans le contexte d’interactions complètes.

Remerciements

Le travail présenté dans cet article a été partiellement financé par le projet ANR MaRDI (Man Robot Dialogue), ANR-12-CORD-0021. Vous trouvez plus d'informations concernant le projet sur <http://mardi.metz.supelec.fr>.

Références

- ANASTASAKOS T. & DEORAS A. (2014). Task specific continuous word representations for mono and multi-lingual spoken language understanding. In *ICASSP*.
- BAYER A. & RICCARDI G. (2013). On-line adaptation of semantic models for spoken language understanding. In *ASRU*.
- BENGIO S. & HEIGOLD G. (2014). Word embeddings for speech recognition. In *INTERSPEECH*.
- BIAN J., GAO B. & LIU T. (2014). Knowledge-powered deep learning for word embedding. In *ECML*.
- CAMELIN N., DETIENNE B., HUET S., QUADRI D. & LEFÈVRE F. (2011). Unsupervised concept annotation using latent dirichlet allocation and segmental methods. In *EMNLP Workshop on Unsupervised Learning in NLP*.
- CELIKYILMAZ A., TUR G. & HAKKANI-TUR D. (2011). Leveraging web query logs to learn user intent via bayesian latent variable model. In *ICML*.
- CLINCHANT S. & PERRONNIN F. (2013). Aggregating continuous word embeddings for information retrieval. In *Workshop on Continuous Vector Space Models and their Compositionality*.
- DAUPHIN Y., TUR G., HAKKANI-TUR D. & HECK L. (2014). Zero-shot learning and clustering for semantic utterance classification. *arXiv preprint arXiv:1401.0509*.
- DEORAS A. & SARIKAYA R. (2013). Deep belief network based semantic taggers for spoken language understanding. In *INTERSPEECH*.
- GAO Y., GU L. & KUO H. (2005). Portability challenges in developing interactive dialogue systems. In *ICASSP*.
- GOTAB P., DAMNATI G., BÉCHET F. & DELPHIN-POULAT L. (2010). Online slu model adaptation with a partial oracle. In *INTERSPEECH*.
- HAHN S., DINARELLI M., RAYMOND C., LEFÈVRE F., LEHNEN P., DE MORI R., MOSCHITTI A., NEY H. & RICCARDI G. (2010). Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE TASLP*, **19**(6), 1569–1583.
- HAKKANI-TUR D., HECK L. & TUR G. (2011). Exploiting query click logs for utterance domain detection in spoken language understanding. In *ICASSP*.
- HECK L. & HAKKANI-TUR D. (2012). Exploiting the semantic web for unsupervised spoken language understanding. In *SLT*.
- HENDERSON M., THOMSON B. & WILLIAMS J. (2014a). The second dialog state tracking challenge. In *SIGDIAL*.
- HENDERSON M., THOMSON B. & WILLIAMS J. (2014b). The third dialog state tracking challenge. In *SLT*.
- JABAIA B., BESACIER L. & LEFÈVRE F. (2013). Comparison and Combination of Lightly Supervised Approaches for Language Portability of a Spoken Language Understanding System. *IEEE TASLP*, **21**(3), 636–648.
- LAROCHELLE H., ERHAN D. & BENGIO Y. (2008). Zero-data learning of new tasks. In *Conference on Artificial Intelligence*.
- LEFÈVRE F. (2007). Dynamic Bayesian networks and discriminative classifiers for multi-stage semantic interpretation. In *ICASSP*.
- LEFÈVRE F., MAIRESSE F. & YOUNG S. (2010). Cross-lingual spoken language understanding from unaligned data using discriminative classification models and machine translation. In *INTERSPEECH*.
- LEFÈVRE F., MOSTEFA D., BESACIER L., ESTEVE Y., QUIGNARD M., CAMELIN N., FAVRE B., JABAIA B. & ROJAS-BARAHONA L. (2012). Robustness and portability of spoken language understanding systems among languages and domains : the PORT-MEDIA project. In *LREC*.
- LORENZO A., ROJAS-BARAHONA L. & CERISARA C. (2013). Unsupervised structured semantic inference for spoken dialog reservation tasks. In *SIGDIAL*.

- MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.
- MIKOLOV T., YIH W. & ZWEIG G. (2013b). Linguistic regularities in continuous space word representations. In *NAACL-HLT*.
- PALATUCCI M., POMERLEAU D., HINTON G. E. & MITCHELL T. M. (2009). Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems 22*, p. 1410–1418.
- SARIKAYA R. (2008). Rapid bootstrapping of statistical spoken dialogue systems. *Speech Communication*, **50**(7), 580–593.
- TUR G., HAKKANI-TUR D., HILLARD D. & CELIKYILMAZ A. (2011). Towards unsupervised spoken language understanding : Exploiting query click logs for slot filling. In *INTERSPEECH*.
- TUR G., HAKKANI-TUR D. & SCHAPIRE R. (2005). Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, **45**(2), 171–186.
- TUR G., RAHIM G. & HAKKANI-TUR D. (2003). Active labeling for spoken language understanding. In *EUROSPEECH*.
- WANG Y. & ACERO A. (2006). Discriminative models for spoken language understanding. In *ICSLP*.
- WILLIAMS J. D. (2014). Web-style ranking and slu combination for dialog state tracking. In *Meeting of the Special Interest Group on Discourse and Dialogue*.
- ZOU W., SOCHER R., CER D. & MANNING C. (2013). Bilingual word embeddings for phrase-based machine translation. In *EMNLP 2013*.

Désambiguïsation d'entités pour l'induction non supervisée de schémas événementiels

Kiem-Hieu Nguyen^{1, 2} Xavier Tannier^{3, 1} Olivier Ferret² Romaric Besançon²

(1) LIMSI-CNRS

(2) CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, F-91191, Gif-sur-Yvette

(3) Univ. Paris-Sud

{nguyen,xtannier}@limsi.fr, {olivier.ferret,romaric.besancon}@cea.fr

Résumé. Cet article présente un modèle génératif pour l'induction non supervisée d'événements. Les précédentes méthodes de la littérature utilisent uniquement les têtes des syntagmes pour représenter les entités. Pourtant, le groupe complet (par exemple, "un homme armé") apporte une information plus discriminante (que "homme"). Notre modèle tient compte de cette information et la représente dans la distribution des schémas d'événements. Nous montrons que ces relations jouent un rôle important dans l'estimation des paramètres, et qu'elles conduisent à des distributions plus cohérentes et plus discriminantes. Les résultats expérimentaux sur le corpus de MUC-4 confirment ces progrès.

Abstract.

Entity disambiguation for event template induction

In this paper, we present an approach for event induction with a generative model. This model makes possible to consider more relational information than previous models, and has been applied to noun attributes. By their influence on parameter estimation, this new information make probabilistic topic distribution more discriminative and more robust. We evaluated different versions of our model on MUC-4 datasets.

Mots-clés : Événements, modèle génératif, désambiguïsation d'entités, échantillonnage de Gibbs.

Keywords: Event Induction, Generative Model, Entity Disambiguation, Gibbs Sampling.

1 Introduction

L'extraction d'information s'est initialement définie et se définit toujours en grande partie au travers des tâches prescrites par les évaluations MUC (Message Understanding Conferences (Grishman & Sundheim, 1996)) et plus particulièrement sa tâche de remplissage de formulaire. Dans un tel contexte, l'objectif est d'extraire à partir de textes les événements d'un certain type ainsi que les éléments permettant de les caractériser. Le formulaire constitue la représentation du type d'événement considéré et par là même, la spécification des informations à extraire. Un formulaire rassemble donc l'ensemble de ces informations, prenant la forme dans un certain nombre de cas d'entités nommées, en précisant le rôle occupé par chaque information vis-à-vis de l'événement auquel elle se rattache. Pour un événement comme un tremblement de terre par exemple, le formulaire regroupera typiquement des informations (rôles) comme sa localisation, sa date, sa magnitude et les dégâts qu'il a pu occasionner (Jean-Louis *et al.*, 2011).

Malgré les efforts entrepris pour définir des tâches génériques, comme la reconnaissance d'entités nommées, la tâche de remplissage de formulaire reste très dépendante du type d'événement considéré. Ainsi, le travail réalisé pour développer un système concernant un type d'événement donné est pour une bonne part à recommencer pour un autre type d'événement. Quelques travaux se sont néanmoins attachés à limiter l'effort nécessaire à la définition d'un nouveau système d'extraction d'événements. Freedman *et al.* (2011) abordent la question par le biais d'une association entre des méthodes génériques fondées sur l'apprentissage et des règles définies manuellement. Grishman & He (2014) optent quant à eux pour la conjugaison d'un nombre restreint d'exemples fournis manuellement et de méthodes, en particulier distributionnelles, permettant d'étendre automatiquement la couverture de ces exemples.

D'autres travaux ont poussé plus loin cette logique en voulant offrir aux utilisateurs des modes d'extraction de l'infor-

mation plus souples et plus ouverts quant à la spécification de leur besoin informationnel. Ainsi, l'approche *On-demand information extraction* (Sekine, 2006), préfigurée dans Hasegawa *et al.* (2004) et concrétisée par la *Preemptive Information Extraction* (Shinyama & Sekine, 2006), vise à induire l'équivalent d'un formulaire à partir d'un ensemble de documents représentatifs des informations à extraire, documents typiquement obtenus par le biais de requêtes soumises à un moteur de recherche. Ce courant de recherche s'est ensuite davantage orienté vers l'extraction de relations, avec notamment Kathrin Eichler & Neumann (2008), Rosenfeld & Feldman (2007) et plus récemment Min *et al.* (2012), que vers l'extraction d'événements.

Dans cet article, nous nous plaçons dans la voie tracée initialement par Hasegawa *et al.* (2004) et Sekine (2006) en considérant la possibilité d'induire des représentations d'événements à partir de textes. Plus globalement, nous cherchons à construire une base de connaissances événementielles à partir de larges corpus journalistiques afin d'offrir des moyens d'accès structurés à ces corpus. Nous nous concentrons dans le cadre du travail présenté dans cet article sur le processus d'induction de schémas d'événements.

2 Objectif

Notre objectif global est de modéliser les événements décrits dans un corpus journalistique et d'identifier sans supervision les schémas (ou formulaires) récurrents ainsi que les rôles associés permettant de représenter ces événements. L'idée initiale de l'approche est de regrouper les entités¹ correspondant à certains rôles dans des événements en fonction de leurs relations avec les mêmes prédicats. Par exemple, dans un corpus sur des attentats terroristes, les mots qui sont objets des verbes *kill*, *attack* peuvent être regroupés et caractérisés par un rôle *VICTIM*. Le résultat de cette identification est donc constitué par un ensemble de groupes (*clusters*) contenant des mots et des relations associés à une probabilité d'appartenance à ce groupe (voir un exemple plus loin, Figure 4). Ces groupes ne sont pas nommés mais représentent chacun un rôle d'événement.

L'approche que nous proposons ici a pour objectif d'améliorer cette approche initiale en aidant à la désambiguïsation des entités. En effet, certaines entités ambiguës, comme "*man*" ou "*soldier*", peuvent correspondre à deux rôles différents (victime ou auteur de l'attaque). Une entité comme "*terrorist*" peut se retrouver mêlée aux victimes lorsque les articles relatent qu'un terroriste est tué par la police (et est donc également objet de *kill*). Notre hypothèse est que le contexte proche des entités est porteur d'information pour aider à leur désambiguïsation. Par exemple, le fait que l'entité "*man*" soit associée à "*armed*", "*dangerous*", "*heroic*" ou "*innocent*" peut conduire à une meilleure attribution et donc définition des rôles. Nous introduisons donc dans le modèle, en plus des relations avec les prédicats, les relations des entités avec leurs attributs (modificateurs syntaxiques).

Dans la pratique, nous utilisons un modèle génératif proche des "modèles de sujet" (*topic models*), mais sans modéliser la notion de document, habituellement centrale. Ces modèles permettent une catégorisation "douce" (*soft clustering*), c'est-à-dire que chaque mot et chaque relation peuvent apparaître dans plusieurs groupes. Ceci permet de traiter efficacement les nombreux mots et relations ambigus, qui peuvent selon le contexte tenir des rôles différents. En pratique, un modèle génératif considère que les observations (ici, les entités et les relations du corpus) peuvent être générées à partir des rôles. Pour cela, les rôles sont définis par des distributions probabilistes sur les prédicats, les entités et leurs arguments syntaxiques. L'enjeu est d'apprendre les paramètres de ces distributions (ici, par échantillonnage de Gibbs) pour que le résultat soit le plus proche possible des observations initiales (*maximum a posteriori* – MAP).

Pour évaluer la qualité de notre approche, nous comparons les groupes de mots produits avec des formulaires et des rôles de référence, qui sont pour leur part nommés et contiennent des mots du corpus associés aux phrases dans lesquelles ils apparaissent. Pour effectuer cette comparaison, nous utilisons une stratégie automatique et empirique d'association entre les rôles du système et ceux de la référence, de façon similaire aux travaux précédents dans le domaine.

N.B. : la terminologie anglophone pour le domaine considéré étant plus répandue et plus stabilisée, nous précisons les termes anglais correspondant aux termes français que nous avons choisis : formulaire : *template* ; rôle : *slot* ; modèle de sujet : *topic model* ; relation-prédicat : parfois nommée *event trigger*, ou *verb path*, ou tout simplement *relation*.

Après une brève présentation des travaux reliés (Section 3), nous précisons la représentation que nous avons choisie pour les entités et les relations que nous extrayons (Section 4). Nous décrivons ensuite le modèle génératif mis en œuvre pour le regroupement en rôles (Section 5), avant de proposer plusieurs expérimentations et évaluations (Section 6).

1. Dans la pratique, tout groupe nominal est une entité candidate pour un rôle.

3 Travaux reliés

La problématique de l'induction de schémas d'événements à partir de textes n'est pas nouvelle. Elle plonge en effet ses racines dans les travaux sur l'acquisition des schémas (Lebowitz, 1983) utilisés par les systèmes de compréhension de textes de la fin des années 70 et du début des années 80 (DeJong, 1982). Cette même perspective se retrouve dans Ferret & Grau (1997). Une des premières introductions de cette problématique pour la création automatique de templates dans le domaine de l'extraction d'information est le fait de Collier (1998). Elle apparaît ensuite dans Harabagiu (2004) sous la forme de la structuration de thèmes événementiels tels que ceux considérés dans les évaluations *Topic Detection and Tracking* (Wayne, 1998). Les schémas ainsi formés ont ensuite été utilisés à la fois en extraction d'information et en résumé automatique. L'intérêt pour cette problématique s'est aussi étendu au domaine du question-réponse par la volonté de créer automatiquement une représentation des événements à partir de textes (Filatova *et al.*, 2006) pour améliorer la recherche de réponse à des questions événementielles (Filatova, 2008). L'essor plus récent des approches faiblement supervisées a enfin vu le développement de plusieurs travaux importants abordant le sujet selon différents angles. Qiu *et al.* (2008) l'ont ainsi envisagé comme un problème de clustering dans un graphe de propositions construit à partir des textes. Regneri *et al.* (2010) l'ont abordé comme un problème d'alignement de séquences d'événements. Bejan (2008) a adopté pour sa part une approche générative fondée sur le paradigme de l'allocation de Dirichlet latente (LDA) en considérant qu'un document est représenté comme une distribution de probabilité sur un ensemble de schémas d'événements et que chacun de ces schémas est lui-même défini comme une distribution de probabilité sur un ensemble de frames sémantiques de type FrameNet (Baker *et al.*, 1998). Des modèles génératifs plus spécifiques au problème considéré que la simple transposition d'une approche LDA ont été ensuite proposés dans des travaux comme Chambers (2013), Cheung *et al.* (2013) et dernièrement Frermann *et al.* (2014). Enfin, Chambers & Jurafsky (2008), Chambers & Jurafsky (2009), Chambers & Jurafsky (2011), amélioré par Balasubramanian *et al.* (2013), et Chambers (2013) se sont plus particulièrement focalisés sur l'émergence de rôles et la découverte de chaînes d'événements pour induire des schémas narratifs à partir de textes en se fondant sur la résolution de corréférence et en prenant en compte la dimension temporelle pour l'ordonnement des événements au sein de ces schémas. Le travail que nous présentons dans cet article se situe dans le prolongement de Chambers (2013). Nous aurons d'ailleurs l'occasion de décrire son approche plus en détails et de la comparer à la nôtre en termes qualitatifs et quantitatifs dans la Section 6.

4 Représentation des entités et des relations

Une entité est représentée par un triplet comprenant un mot (la tête de l'entité), une liste de "relations-attributs" et une liste de "relations-prédicats". Considérons l'exemple suivant :

(1) Two armed men attacked the police station and killed a policeman. An innocent young man was also wounded.

Comme illustré par la Figure 1, quatre entités sont distinguées, représentées par quatre triplets. Les têtes d'entités sont extraites des syntagmes.

Une relation-prédicat est composée d'un prédicat (*attack*, *kill*, *explosion*) et d'un type de dépendance (sujet – *nsubj*, objet – *dobj*, etc.). Une relation-attribut est composée d'un argument (*armed*, *police*, *young*) et d'un type de dépendance (modifieur adjectival – *amod*, nominal – *nn* ou verbal – *vmod*). Notons que le mot de l'entité gouverne la relation-attribut mais est gouverné dans la relation-prédicat. Nous utilisons l'analyseur de Stanford (Manning *et al.*, 2014) pour l'analyse syntaxique et la résolution des corréférences.

Une *entité* est extraite pour chaque groupe nominal dont la tête (nom commun ou propre) est liée à au moins un prédicat. Les pronoms ne sont pas considérés. Une *relation-prédicat* est quant à elle extraite pour chaque verbe ou nom d'événement.

Triplets	Relations-attributs	Tête	Relations-prédicats
#1	armed:amod	man	[attack:nsubj, kill:nsubj]
#2	police:nn	station	attack:dobj
#3	-	policeman	kill:dobj
#4	[innocent:amod, young:amod]	man	wound:dobj

FIGURE 1 – Représentation des entités comme des triplets de ([relations-attributs], tête, [relations-prédicats])

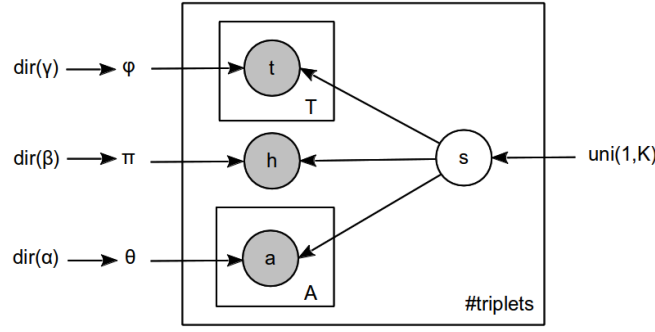


FIGURE 2 – Modèle génératif pour l’induction des formulaires d’événements

ment lié à une entité. Les noms d’événements sont les noms entrant dans les catégories *noun.EVENT* et *noun.ACT* dans WordNet (Miller, 1995). Une entité peut être liée à plus d’une relation-prédicat. Cette multiplicité peut avoir une origine intra-phrastique, comme c’est le cas avec la coordination présente dans la première phrase de notre exemple, ou inter-phrastique dans le cas des coréférences. Nous fusionnons en effet les différentes mentions d’une chaîne de coréférence en une seule entité, à laquelle sont rattachées les relations-prédicats de chaque mention. Il serait même possible de fusionner des mentions inter-document si un système traitait ce type de coréférences avec une précision satisfaisante. Enfin, une *relation-attribut* est extraite pour chaque adjectif, nom ou verbe jouant le rôle de modifieur adjectival, nominal ou verbal d’un nom. Si plusieurs modifieurs existent, seul l’élément le plus proche de la première mention de l’entité est conservé. Cette heuristique permet de ne pas conduire à une influence supérieure des attributs par rapport aux relations-prédicats (les attributs multiples étant plus fréquents). Elle est appropriée pour la langue anglaise, où elle permet d’omettre un grand nombre d’adjectifs non discriminants pour l’entité, mais pourrait être rediscutée pour une autre langue. Les expériences finales montrent de meilleures performances lorsque l’on se limite à un attribut.

5 Modèle génératif

Nous présentons dans cette section le modèle de sujet que nous avons mis en œuvre, avec une description algorithmique et graphique de son processus de génération puis un détail de l’estimation de ses paramètres.

5.1 Description du modèle

La représentation graphique de notre modèle est présentée à la Figure 2. Pour chaque triplet représentant une entité e , le modèle choisit un rôle s pour cette entité à partir d’une distribution uniforme $uni(1, K)$, où K est le nombre de rôles. La tête h du syntagme de l’entité est ensuite générée à partir d’une distribution multinomiale π_s . Chaque relation-prédicat t de l’ensemble T_e des relations-prédicats est générée à partir d’une distribution multinomiale ϕ_s . Enfin, chaque relation-attribut a de l’ensemble A_e des relations-attributs est également générée à partir d’une distribution multinomiale θ_s . Les distributions θ , π et ϕ sont générées par les lois de Dirichlet $dir(\alpha)$, $dir(\beta)$ et $dir(\gamma)$, respectivement.

Étant donné un ensemble d’entités E , notre modèle (π, ϕ, θ) est défini par :

$$P_{\pi, \phi, \theta}(E) = \prod_{e \in E} P_{\pi, \phi, \theta}(e) \quad (2)$$

sans faire de distinction entre les documents contenant les entités. La probabilité de chaque entité e est définie par :

$$\begin{aligned}
P_{\pi, \phi, \theta}(e) &= P(s) \\
&\times P(h|s) \\
&\times \prod_{t \in T_e} P(t|s) \\
&\times \prod_{a \in A_e} P(a|s)
\end{aligned} \tag{3}$$

Le processus génératif est le suivant :

```

for rôle  $s \leftarrow 1$  to  $K$  do
  Générer une distribution de têtes d'entités  $\pi_s$  à partir d'une loi de Dirichlet  $\text{dir}(\beta)$  ;
  Générer une distribution de relations-attributs  $\theta_s$  à partir d'une loi de Dirichlet  $\text{dir}(\alpha)$  ;
  Générer une distribution de relations-prédicats  $\phi_s$  à partir d'une loi de Dirichlet  $\text{dir}(\gamma)$  ;
end
for entité  $e \in E$  do
  Générer un rôle  $s$  à partir d'une distribution uniforme  $\text{uni}(1, K)$  ;
  Générer une tête  $h$  à partir d'une distribution multinomiale  $\pi_s$  ;
  for  $i \leftarrow 1$  to  $|T_e|$  do
    Générer une relation-prédicat  $t_i$  à partir d'une distribution multinomiale  $\phi_s$  ;
  end
  for  $j \leftarrow 1$  to  $|A_e|$  do
    Générer une relation-attribut  $a_j$  à partir d'une distribution multinomiale  $\theta_s$  ;
  end
end

```

5.2 Apprentissage des paramètres

L'estimation des paramètres du modèle est effectuée par la méthode d'échantillonnage de Gibbs (Griffiths, 2002). La variable s du rôle est échantillonnée par intégration de toutes les autres variables du modèle. Les modèles précédents (Cheung *et al.*, 2013; Chambers, 2013) sont fondés sur une modélisation des sujets au niveau du document, dans la tradition des modèles comme l'allocation de Dirichlet latente (*Latent Dirichlet Allocation*, LDA (Blei *et al.*, 2003)). Notre modèle s'affranchit de cette notion de document puisque ce niveau de structure n'a finalement pas d'impact sur les rôles qu'il contient. L'entrée du modèle est donc une chaîne continue de triplets d'entités. Les frontières du document sont uniquement utilisées à l'issue de l'apprentissage pour l'étape de filtrage (décrite à la Section 6.5). La distribution des rôles est donc globale et non spécifique à chaque document, ce qui est plus en phase avec l'induction de structures à l'échelle du corpus. De plus, la distribution *a priori* des rôles est ignorée en initialisant avec une distribution uniforme, cas particulier d'une distribution de Bernoulli généralisée.

L'attribution des rôles par échantillonnage dépend des états initiaux et du hasard. Dans notre implémentation de l'échantillonnage de Gibbs, nous utilisons 2 000 itérations de *burn in* sur un total de 10 000 itérations. Cette étape de *burn in* permet de s'assurer que les paramètres convergent vers un état stable avant d'utiliser les échantillons pour l'estimation des distributions de probabilités. De plus, après le *burn in*, un intervalle de 100 itérations est appliqué entre deux estimations des paramètres pour éviter une trop grande proximité entre deux échantillons successifs.

Enfin, les attributs étant moins porteurs de sens mais venant plutôt "en support", la phase de *burn in* ne les considère pas et n'échantillonne que les entités et les prédicats. L'état stable obtenu est ensuite utilisé comme initialisation pour un échantillonnage sur les trois éléments. Ainsi, les éléments non ambigus sont peu affectés par les attributs tandis que les éléments plus "sensibles", comme par exemple les entités ambiguës "*man*", "*soldier*" ou les relations "*kill:doj*" et "*attack:nsubj*" voient leurs probabilités modifiées. Sans les attributs, les "terroristes" ("*guerrilla*", "*terrorists*", etc.) sont souvent mélangés aux victimes, car souvent "tués" (c'est-à-dire, en relation avec *kill:doj*) par la police. Grâce à l'ajout des adjectifs et des attributs, la séparation entre les "*perpetrators*" (qui sont par exemple "*armed*" et "*dangerous*") et les "*victims*" (plutôt "*heroic*" ou "*innocent*") est facilitée. On voit par exemple à la Figure 3, qui montre l'évolution des probabilités de certains éléments à mesure des itérations de *burn in*, que la probabilité de la relation "*kill:doj*" diminue dans le rôle correspondant aux victimes, de même que celle de l'entité "*terrorist*" dans ce même rôle.

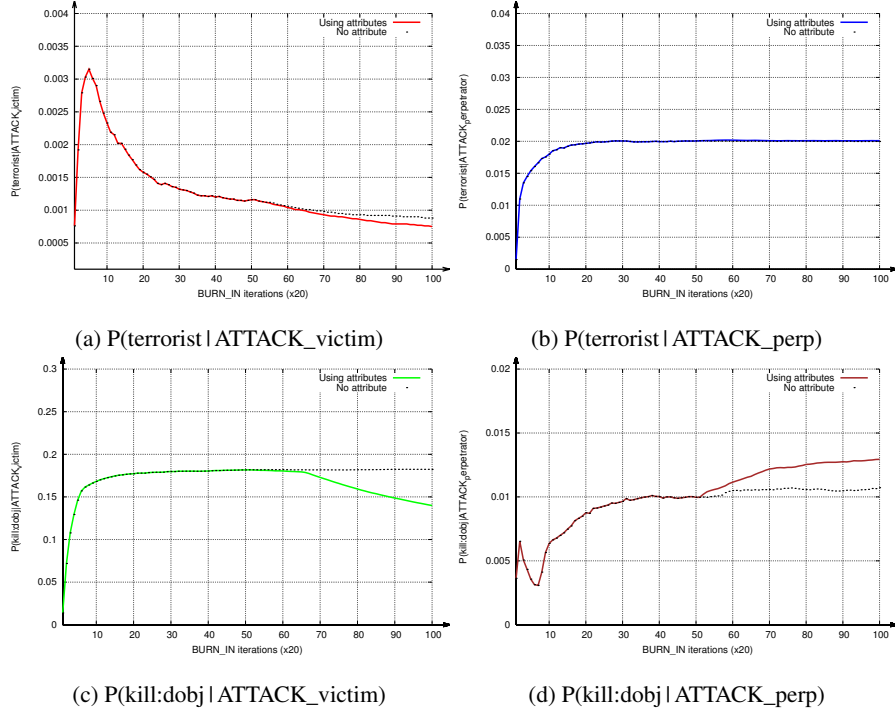


FIGURE 3 – Les 2 000 premières itérations (*burn in*) de l'échantillonnage pour 4 rôles, avec les attributs ajoutés à mi-parcours (50 %, ligne continue), ou sans les attributs (ligne pointillée)

6 Évaluation

Pour l'évaluation, nous utilisons le corpus MUC-4 (Message Understanding Conference (Sundheim, 1991)) afin de nous comparer aux autres systèmes dont les résultats sont généralement obtenus avec ce corpus. Nous utilisons les métriques traditionnelles de précision, rappel et F-mesure. Dans ce qui suit, nous présentons d'abord le corpus utilisé (Section 6.1), puis les paramètres du système (6.2), avant de détailler la méthode d'association entre les rôles appris par le système et les rôles du corpus de référence (6.3). Par la suite, nous présentons une première expérience (Section 6.4) montrant l'intérêt d'utiliser les attributs des entités pour la modélisation des rôles. La seconde expérience (Section 6.5) étudie l'impact d'une étape de classification des documents. Enfin, nous comparons nos résultats avec ceux faisant référence, et plus particulièrement à Chambers (2013), tant sur des aspects quantitatifs que qualitatifs (Section 6.6).

6.1 Corpus

Le corpus MUC-4 contient 1 700 articles journalistiques en langue anglaise concernant des incidents et des attaques terroristes en Amérique Latine. Le corpus est divisé en 1 300 documents pour le développement et quatre ensembles de test contenant chacun 100 documents. Nous suivons les règles établies par les travaux antérieurs pour garantir la comparabilité de nos résultats avec ces travaux (Patwardhan & Riloff, 2007; Chambers & Jurafsky, 2011). Quatre types de formulaires sont présents : *Arson*, *Attack*, *Bombing*, et *Kidnapping* ; pour chacun de ces formulaires, quatre rôles peuvent être pertinents : *Instrument*, *Target*, *Victim*, *Perpetrator* (fusion de *Perpetrator_Individual* et *Perpetrator_Organization*). L'association entre les réponses du système et la référence est fondée sur l'association entre les têtes des groupes nominaux uniquement, la tête étant définie comme le nom le plus à droite du syntagme, ou comme le dernier nom avant une préposition "of". Certains formulaires et rôles dit "optionnels" sont ignorés lors du calcul du rappel. Les types des formulaires sont ignorés pour l'évaluation, ce qui signifie que le *perpetrator* d'un *bombing* et celui d'un *attack* au niveau de la référence peuvent se retrouver associés dans un même rôle induit.

BOMBING_instrument		
Attributs	Têtes	Relations avec prédicats
car:nn	bomb	explode:nsubj
powerful:amod	fire	hear:dobj
explosive:amod	explosion	place:dobj
dynamite:nn	blow	cause:nsubj
heavy:amod	charge	set:dobj
KIDNAPPING_victim		
Attributs	Têtes	Relations avec prédicats
several:amod	people	arrest:dobj
other:amod	person	kidnap:dobj
responsible:amod	man	release:dobj
military:amod	member	kill:dobj
young:amod	leader	identify:prep_as

FIGURE 4 – Distributions apprises par le modèle que nous nommerons plus tard *HT+A* pour les rôles *BOMBING_instrument* et *KIDNAPPING_victim* (les noms des rôles étant attribués après l’association décrite à la Section 6.3)

6.2 Paramètres du système

Les hyperparamètres² sont fixés selon les meilleurs résultats obtenus sur l’ensemble de développement. Le nombre de rôles est ainsi fixé à $s = 35$. Les paramètres *a priori* des lois de Dirichlet sont fixés à $\alpha = 0, 1$, $\beta = 1$ et $\gamma = 0, 1$. Le modèle est appris sur l’ensemble des données. L’association des rôles (Section 6.3) est effectuée sur les deux premiers ensembles de test. Les sorties des deux derniers ensembles de test sont ensuite évaluées suivant cette association. Par ailleurs, l’échantillonnage étant fondé sur des tirages aléatoires, les valeurs de précision, rappel et F-mesure indiquées sont des moyennes sur 10 exécutions du système.

6.3 Association des rôles du système avec le corpus de référence

Notre modèle apprend K rôles et assigne chaque entité d’un document à l’un de ces rôles. L’association des rôles consiste à faire correspondre chaque rôle de la référence à un rôle équivalent appris par le modèle. Notons que parmi les K rôles appris, certains sont peu pertinents, et d’autres, parfois de très bonne qualité, contiennent des entités ne faisant pas partie de la référence (informations spatio-temporelles, contexte des protagonistes, etc.). Il n’est pas donc illogique que le nombre de rôles appris soit très supérieur au nombre de rôles attendus pour un événement.

De la même façon que les travaux précédents sur le même type de tâche, nous effectuons l’association des rôles de façon automatique et empirique. Nous réservons une partie du corpus de développement à l’opération. Chaque rôle de référence est associé au rôle appris par le système qui permet d’obtenir la meilleure F-mesure. Ici, deux rôles ayant le même nom mais appartenant à deux formulaires différents sont considérés de façon distincte.

À titre d’exemple, la Figure 4 montre les mots obtenant le poids le plus élevé pour les rôles *BOMBING_instrument* et *KIDNAPPING_victim*. L’association ainsi créée est conservée pour l’application sur le corpus de test.

6.4 Utilisation des attributs des entités

Deux versions différentes de notre modèle sont évaluées : *HT* utilise uniquement les têtes des entités et leurs relations avec des événements tandis que *HT+A* introduit également la distribution des attributs des entités.

Pour estimer plus extensivement le gain du modèle avec les attributs, nous avons également fait varier la richesse de l’entrée du modèle, notamment pour tester les éventuelles interactions entre les attributs et l’utilisation des informations

2. Tandis que les *paramètres* du système sont des variables à apprendre par le système, les *hyperparamètres* sont des constantes fixées manuellement par l’expérimentateur.

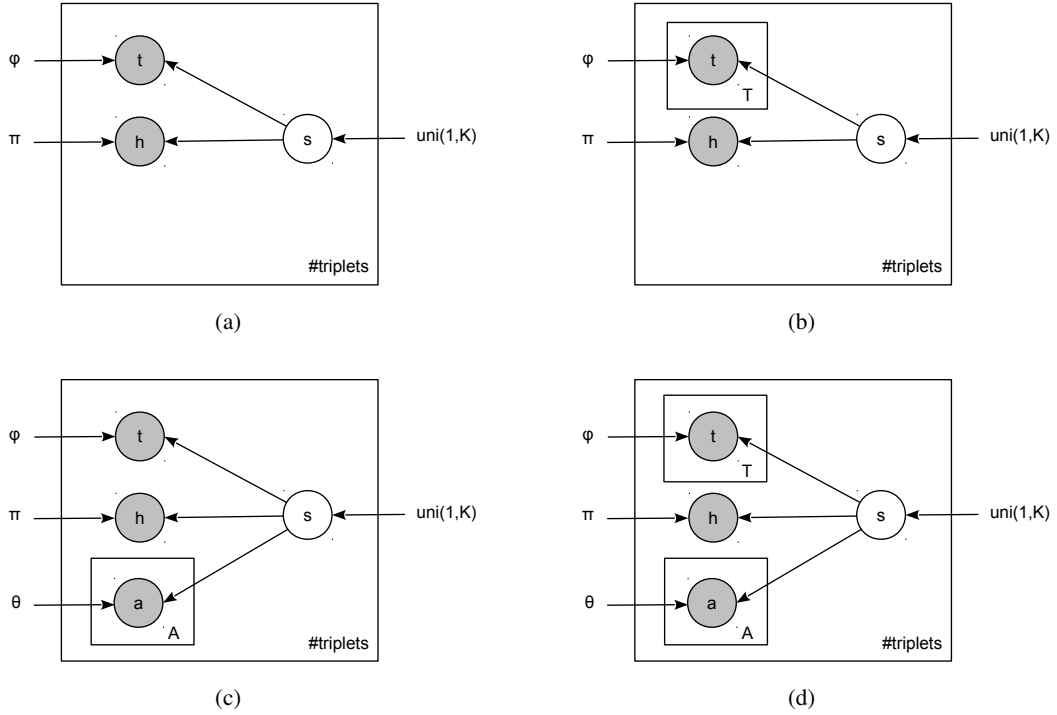


FIGURE 5 – Variations du modèle (les distributions de Dirichlet sont omises) : 5a) modèle HT avec entrée simple. Ce modèle est équivalent à 5b) si on n'a qu'un seul prédicat lié à chaque entité ($T=1$) ; 5b) modèle HT avec entrée multiple ; 5c) modèle HT+A sur entrée simple ; 5d) modèle HT+A sur entrée multiple

Data	HT			HT+A		
	P	R	F	P	R	F
Single	29,59	51,17	37,48	30,22	52,41	38,33
Multiple	29,32	52,21	37,52	30,82	51,68	38,55
Coref	39,99	53,53	40,01	32,42	54,59	40,62

TABLE 1 – L'amélioration apportée par l'utilisation des attributs d'entités

de coréférence. L'entrée dite "simple" ne permet de relier qu'un seul événement à chaque entité. Dans cette configuration, l'énoncé "an armed man attacked the police station and killed a policeman" produit deux triplets pour l'entité "man" : $(armed:amod, man, attack:nsubj)$ et $(armed:amod, man, kill:nsubj)$. Dans l'entrée "multiple", cette duplication de l'entité n'est pas nécessaire puisque plusieurs relations peuvent être reliées à une même entité, conduisant par exemple à $(armed:amod, man, [attack:nsubj, kill:nsubj])$. Enfin, la sortie appelée "coref" ajoute au modèle précédent la prise en compte des relations multiples provenant d'une relation de coréférence détectée par l'analyseur. Par exemple, si "an armed man" est l'antécédent de "he" dans "He was arrested three hours later", alors l'entrée du modèle sera $(armed:amod, man, [attack:nsubj, kill:nsubj, arrest:doobj])$.

Notons que ces différences dans les données fournies au modèle ont une influence sur l'organisation du modèle lui-même. Le Figure 5 montre les variations de ce modèle selon que l'entrée est "simple" ou "multiple".

La Table 1 montre une amélioration systématique apportée par l'utilisation des attributs, que ce soit avec ou sans la résolution de coréférence. La meilleure performance (F-mesure de 40,62) est obtenue par le modèle le plus complexe. Ainsi, l'ajout commun des attributs et de la coréférence conduit à un gain de 3 points de F-mesure.

6.5 Classification des documents

Dans cette seconde expérience, nous avons ajouté à notre modèle une étape de classification des documents après la phase d'apprentissage.

Le corpus MUC-4 contient un grand nombre de documents "non pertinents", c'est-à-dire ne comprenant aucune annotation selon les formulaires décrits plus haut. 567 documents sur les 1 300 de l'ensemble de développement sont non pertinents. Ils sont pourtant difficiles à éliminer automatiquement car ils contiennent de nombreuses allusions au terrorisme, avec les mots "bomb", "force", "guerrilla" et bien d'autres. Dans notre modèle comme dans les précédents, l'annotation erronée d'entités dans ces documents est la cause d'un grand nombre de faux positifs. Le but de notre classification de documents *a posteriori* est donc d'augmenter la précision en limitant la réduction du rappel.

Étant donné un document d dont les entités ont été assignées à des rôles du modèle, nous devons décider si ce document est pertinent ou pas. Nous définissons le score de pertinence du document de la façon suivante :

$$pertinence(d) = \frac{\sum_{e \in d: s_e \in S_m} \sum_{t \in T_e} P(t|s_e)}{\sum_{e \in d} \sum_{t \in T_e} P(t|s_e)} \quad (4)$$

où e est une entité dans le document d ; s_e (de la liste S_m de tous les rôles) est le rôle attribué à e ; t est un prédicat de la liste T_e de tous les prédicats (rappelons que chaque entité et chaque prédicat possède une probabilité pour chaque rôle).

L'équation (4) définit donc le score d'une entité comme la somme des probabilités conditionnelles des prédicats qui lui sont associés, dans le rôle qui lui a été assigné. Le score de pertinence du document est alors proportionnel au score des entités associés aux rôles qui ont été sélectionnés. Si le score d'un document est supérieur à un certain seuil λ , alors le document est considéré comme pertinent. La valeur de λ est fixée à $\lambda = 0,02$, valeur qui maximise la F-mesure sur l'ensemble de développement.

La Table 2 nous montre l'amélioration apportée par l'application de la classification des documents. La précision augmente nettement et le rappel diminue peu, conduisant à une hausse de la F-mesure. Nous nous comparons également avec un classifieur "oracle", virtuel, qui filtrerait parfaitement les documents non pertinents en conservant les pertinents. Les performances de notre modèle avec ce classifieur oracle montrent qu'une marge d'amélioration non négligeable existe encore pour exploiter la classification des documents.

Le filtrage des éléments non pertinents se retrouve aussi bien au niveau des méthodes d'extraction d'information non supervisées que supervisées. Dans ce dernier cas, il est souvent appliqué à un niveau plus fin que le document, typiquement la phrase (Patwardhan & Riloff, 2009, 2007). Pour ce qui est des méthodes non supervisées, Chambers (2013) réalise ce filtrage de façon indirecte par la prise en compte des documents au sein de son modèle de sujet tandis que Chambers & Jurafsky (2011) et Cheung *et al.* (2013) utilisent comme nous les *clusters* appris pour classer les documents. La pertinence d'un document est estimée dans ce cas à partir de statistiques sur les prédicats.

6.6 Comparaison avec l'état de l'art

Nous avons enfin comparé nos résultats avec ceux des systèmes d'extraction non supervisée de formulaires de référence. Nous avons en particulier ré-implémenté la méthode proposée par Chambers (2013), en ajoutant également à son modèle une distribution des attributs comparable à la nôtre (cf. Figure 6).

Les différences principales entre ce modèle et le nôtre sont les suivantes :

	P	R	F
HT + A	32,42	54,59	40,62
HT + A + classification	35,57	53,89	42,79
HT + A + classification oracle	44,58	54,59	49,08

TABLE 2 – Amélioration apportée par la classification des documents

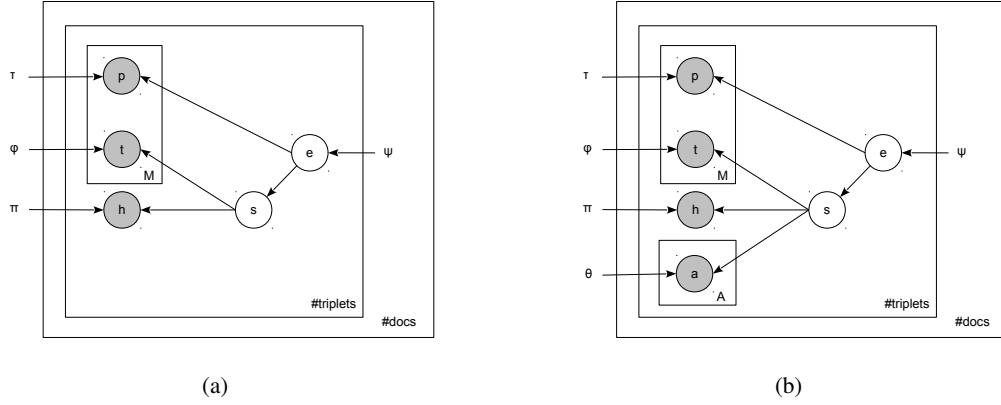


FIGURE 6 – Modèle de Chambers (2013) : 6a) version originale ; 6b) avec une distribution supplémentaire pour les attributs

Système	P	R	F
(Cheung <i>et al.</i> , 2013)	32	37	34
(Chambers & Jurafsky, 2011)	48	25	33
Nest (Chambers, 2013)	41	41	41
Nest (réimplémenté)	39	43	41
Nest (réimplémenté) + Attributs	39	44	41
HT + A + doc. classification	36	54	43

TABLE 3 – Comparaison avec les systèmes non-supervisés état-de-l’art

1. Chambers (2013) ajoute une distribution ψ reliant les événements aux documents. Celle-ci rend le modèle plus complexe et peut-être moins intuitif puisque la notion de document ne devrait pas être *a priori* en lien avec les formulaires et les rôles : un document contient possiblement des références à plusieurs formulaires et l’association entre entités et rôles ne dépend pas du niveau du document. Cependant, cette distribution permet d’éviter le recours à une classification des documents séparée.
2. Chaque entité est liée à une variable d’événement p . Cet événement génère un prédicat pour chaque mention d’entité (rappelons que les mentions d’une entité sont toutes les occurrences d’une entité dans un document, y compris dans une chaîne de coréférence). Notre travail se concentre davantage sur le fait de prendre en compte une certaine pluralité des informations liées aux entités. Les relations-prédicats et les relations-attributs multiples sont ainsi traitées de la même façon. Cette multiplicité n’a pas seulement pour origine les chaînes de coréférence mais résulte aussi de la multiplicité "naturelle" des relations attachées à une entité (qui pourraient être des relations autres que syntaxiques). En conséquence, il est possible d’avancer que notre modèle offre une plus souplesse d’extension que celui de Chambers (2013), à la fois en termes de modélisation et de données d’entrée.
3. Enfin, Chambers (2013) ajoute une contrainte heuristique lors de l’échantillonnage imposant que le sujet et l’objet d’un même prédicat (par exemple, *kill:nsbj* et *kill:doobj*) ne soient pas distribués dans le même rôle. Notre modèle ne nécessite pas une telle heuristique.

Certains aspects concernant le prétraitement des données et les paramètres du modèle n’étant pas totalement précisés (Chambers, 2013), notre implémentation (menée sur les mêmes données) conduit à des résultats légèrement différents de ceux publiés. C’est pourquoi nous présentons ici les deux informations.

La Table 3 montre que notre modèle dépasse tous les autres modèles considérés, en particulier par un rappel bien supérieur, conduisant à une meilleure F-mesure. Il est également intéressant de constater que la prise en compte des attributs améliore le modèle de Chambers (2013), illustrant en cela la généralité de l’idée sous-jacente.

7 Conclusion et perspectives

Nous avons présenté dans cet article un modèle génératif permettant de modéliser les rôles tenus par des entités dans des schémas d'événements. Nous avons mis l'accent sur l'utilisation du contexte des entités protagonistes et nous avons proposé un modèle à la fois plus simple et plus efficace que ceux de l'état de l'art. Nous avons évalué ce modèle en comparant les rôles obtenus à ceux définis pour le corpus MUC-4. Une évaluation sur d'autres langues disposant d'analyseurs syntaxiques performants, comme le français, serait intéressante. Cependant, les corpus associés restent à construire.

Même si les résultats sont meilleurs que ceux obtenus précédemment, l'approche non supervisée est encore loin des résultats présentés par les approches supervisées. Les pistes d'amélioration sont donc nombreuses. En premier lieu, les caractéristiques du corpus MUC-4 sont un facteur limitant. Il est de petite taille et les rôles sont presque les mêmes pour chaque formulaire d'événement, ce qui n'est pas représentatif de la réalité. Un corpus de plus grande taille, même partiellement annoté, et présentant des schémas d'événements plus variés, permettrait d'envisager d'autres approches.

Comme nous l'avons montré, notre modèle a l'avantage de permettre une représentation unifiée de tous les types de relations ; une extension possible de notre travail serait donc de multiplier les relations (syntaxiques, sémantiques, thématiques, etc.) pour affiner les groupes générés.

Enfin, le protocole d'évaluation, devenu une sorte de standard de fait, est très imparfait. Notamment, la façon de réaliser l'alignement final avec les rôles de la référence peut influencer fortement sur les résultats.

Remerciements

Ce projet a été partiellement financé par la Fondation de Coopération Scientifique "Campus Paris-Saclay" à travers le projet Digiteo ASTRE N° 2013-0774D.

Références

- BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of ACL-COLING 1998*, p. 86–90, Montréal, Québec, Canada.
- BALASUBRAMANIAN N., SODERLAND S., MAUSAM & ETZIONI O. (2013). Generating Coherent Event Schemas at Scale. In *2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, Seattle, USA.
- BEJAN C. A. (2008). Unsupervised Discovery of Event Scenarios from Texts. In *Twenty-First International Florida Artificial Intelligence Research Society Conference (FLAIRS 2008)*, p. 124–129, Coconut Grove, Florida.
- BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993–1022.
- CHAMBERS N. (2013). Event Schema Induction with a Probabilistic Entity-Driven Model. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, p. 1797–1807, Seattle, USA.
- CHAMBERS N. & JURAFSKY D. (2008). Unsupervised Learning of Narrative Event Chains. In *ACL-08 : HLT*, p. 789–797, Columbus, Ohio.
- CHAMBERS N. & JURAFSKY D. (2009). Unsupervised Learning of Narrative Schemas and their Participants. In *Proceedings of ACL-IJCNLP 2009*, p. 602–610, Suntec, Singapore.
- CHAMBERS N. & JURAFSKY D. (2011). Template-Based Information Extraction without the Templates. In *49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL 2011)*, p. 976–986, Portland, Oregon, USA.
- CHEUNG K. J. C., POON H. & VANDERWENDE L. (2013). Probabilistic Frame Induction. In *Proceedings of NAACL-HLT 2013*, p. 837–846.
- COLLIER R. (1998). *Automatic Template Creation for Information Extraction*. PhD thesis, University of Sheffield.
- DEJONG G. (1982). An overview of the FRUMP system. In W. LEHNERT & M. RINGLE, Eds., *Strategies for natural language processing*, p. 149–176. Lawrence Erlbaum Associates.
- FERRET O. & GRAU B. (1997). An Aggregation Procedure for Building Episodic Memory. In *15th International Joint Conference on Artificial Intelligence (IJCAI-97)*, p. 280–285, Nagoya, Japan.

- FILATOVA E. (2008). *Unsupervised Relation Learning for Event-Focused Question-Answering and Domain Modelling*. PhD thesis, Columbia University.
- FILATOVA E., HATZIVASSILOGLOU V. & MCKEOWN K. (2006). Automatic Creation of Domain Templates. In *COLING-ACL 2006*, p. 207–214, Sydney, Australia.
- FREEDMAN M., RAMSHAW L., BOSCHÉ E., GABBARD R., KRATKIEWICZ G., WARD N. & WEISCHEDEL R. (2011). Extreme Extraction – Machine Reading in a Week. In *2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, p. 1437–1446, Edinburgh, Scotland, UK.
- FRERMANN L., TITOV I. & PINKAL M. (2014). A Hierarchical Bayesian Model for Unsupervised Induction of Script Knowledge. In *14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, p. 49–57, Gothenburg, Sweden.
- GRIFFITHS T. (2002). *Gibbs sampling in the generative model of Latent Dirichlet Allocation*. Rapport interne, Stanford University.
- GRISHMAN R. & HE Y. (2014). An Information Extraction Customizer. In P. SOJKA, A. HORÁK, I. KOPEČEK & K. PALA, Eds., *17th International Conference on Text, Speech and Dialogue (TSD 2014)*, volume 8655 of *Lecture Notes in Computer Science*, p. 3–10. Springer International Publishing.
- GRISHMAN R. & SUNDHEIM B. (1996). Message Understanding Conference-6 : A Brief History. In *16th International Conference on Computational linguistics (COLING’96)*, p. 466–471, Copenhagen, Denmark.
- HARABAGIU S. (2004). Incremental Topic Representation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING’04)*, Geneva, Switzerland.
- HASEGAWA T., SEKINE S. & GRISHMAN R. (2004). Discovering Relations among Named Entities from Large Corpora. In *42nd Meeting of the Association for Computational Linguistics (ACL’04)*, p. 415–422, Barcelona, Spain.
- JEAN-LOUIS L., BESANÇON R. & FERRET O. (2011). Text Segmentation and Graph-based Method for Template Filling in Information Extraction. In *5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, p. 723–731, Chiang Mai, Thailand.
- KATHRIN EICHLER H. H. & NEUMANN G. (2008). Unsupervised Relation Extraction From Web Documents. In *6th Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco.
- LEBOWITZ M. (1983). Generalization from natural language text. *Cognitive Science*, **7**(1), 1–40.
- MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKEY D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 55–60, Baltimore, USA.
- MILLER G. A. (1995). WordNet : a lexical database for English. *Communication of the ACM*, **38**(11), 39–41.
- MIN B., SHI S., GRISHMAN R. & LIN C.-Y. (2012). Ensemble Semantics for Large-scale Unsupervised Relation Extraction. In *2012 Joint Conference EMNLP-CoNLL*, p. 1027–1037, Jeju Island, Korea.
- PATWARDHAN S. & RILOFF E. (2007). Effective Information Extraction with Semantic Affinity Patterns and Relevant Regions. In *Proceedings of the 2007 Joint Conference EMNLP-CoNLL*, p. 717–727, Prague, Czech Republic.
- PATWARDHAN S. & RILOFF E. (2009). A Unified Model of Phrasal and Sentential Evidence for Information Extraction. In *Proceedings of EMNLP 2009*, p. 151–160, Singapore.
- QIU L., KAN M.-Y. & CHUA T.-S. (2008). Modeling Context in Scenario Template Creation. In *Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, p. 157–164, Hyderabad, India.
- REGNERI M., KOLLER A. & PINKAL M. (2010). Learning Script Knowledge with Web Experiments. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, p. 979–988, Uppsala, Sweden.
- ROSENFELD B. & FELDMAN R. (2007). Clustering for unsupervised relation identification. In *Sixteenth ACM conference on Conference on information and knowledge management (CIKM’07)*, p. 411–418, Lisbon, Portugal.
- SEKINE S. (2006). On-demand information extraction. In *Proceedings of COLING-ACL 2006*, p. 731–738, Sydney, Australia.
- SHINYAMA Y. & SEKINE S. (2006). Preemptive Information Extraction using Unrestricted Relation Discovery. In *HLT-NAACL 2006*, p. 304–311, New York City, USA.
- SUNDHEIM B. M. (1991). Third Message Understanding Evaluation and Conference (MUC-3) : Phase 1 Status Report. In *Proceedings of the 4th DARPA Workshop on Speech and Natural Language*, p. 301–305, San Diego, California, USA.
- WAYNE C. (1998). Topic Detection & Tracking : A Case Study in Corpus Creation & Evaluation Methodologies. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain.

Création rapide et efficace d'un système de désambiguïsation lexicale pour une langue peu dotée

Mohammad Nasiruddin, Andon Tchechmedjiev, Hervé Blanchon, Didier Schwab

LIG-GETALP

Univ. Grenoble Alpes

prenom.nom@imag.fr

<http://getalp.imag.fr/WSD>

Résumé. Nous présentons une méthode pour créer rapidement un système de désambiguïsation lexicale (DL) pour une langue L peu dotée pourvu que l'on dispose d'un système de traduction automatique statistique (TAS) d'une langue riche en corpus annotés en sens (ici l'anglais) vers L. Il est, en effet, plus facile de disposer des ressources nécessaires à la création d'un système de TAS que des ressources dédiées nécessaires à la création d'un système de DL pour la langue L. Notre méthode consiste à traduire automatiquement un corpus annoté en sens vers la langue L, puis de créer le système de désambiguïsation pour L par des méthodes supervisées classiques. Nous montrons la faisabilité de la méthode et sa genericité en traduisant le *SemCor*, un corpus en anglais annoté grâce au *Princeton WordNet*, de l'anglais vers le bangla et de l'anglais vers le français. Nous montrons la validité de l'approche en évaluant les résultats sur la tâche de désambiguïsation lexicale multilingue de Semeval 2013.

Abstract.

Rapid Construction of Supervised Word Sense Disambiguation System for Lesser-resourced Languages

We introduce a method to quickly build a Word Sense Disambiguation (WSD) system for a lesser-resourced language L, under the condition that a Statistical Machine Translation system (SMT) is available from a well resourced language where semantically annotated corpora are available (here, English) towards L. We argue that it is less difficult to obtain the resources mandatory for the development of an SMT system (parallel-corpora) than it is to create the resources necessary for a WSD system (semantically annotated corpora, lexical resources). In the present work, we propose to translate a semantically annotated corpus from English to L and then to create a WSD system for L following the classical supervised WSD paradigm. We demonstrate the feasibility and genericity of our proposed method by translating *SemCor* from English to Bangla and from English to French. *SemCor* is an English corpus annotated with *Princeton WordNet* sense tags. We show the feasibility of the approach using the Multilingual WSD task from Semeval 2013.

Mots-clés : clarification de texte, désambiguïsation lexicale, langues peu dotées, traduction automatique, portage d'annotations.

Keywords: clarification of texts, word sense disambiguation, under resourced languages, machine translation, annotation transfert.

1 Introduction

La clarification de texte est une tâche centrale pour le traitement automatique des langues. Elle peut, en effet, permettre d'améliorer de nombreuses applications comme l'extraction d'informations multilingues, le résumé automatique ou encore la traduction automatique. Il s'agit de lever, manuellement ou automatiquement, un certain nombre d'ambiguïtés : déterminer les différents acteurs qui interviennent dans l'énoncé, leurs rôles ou déterminer le sens des mots utilisés parmi un inventaire pré-défini (désambiguïsation lexicale). Par exemple, dans «*La souris mange le fromage*», l'animal devrait être préféré au dispositif électronique de pointage et serait traduit en malais, par exemple, par «*tikus*» plutôt que par «*tetikus*».

Les méthodes de désambiguïsation lexicale supervisée ont besoin de corpus annotés en sens de mot pour être entraînées. Malheureusement de tels corpus n'existent que dans très peu de langues, ce qui rend souvent impossible la désambiguï-

sation lexicale supervisée.

Dans cet article, nous présentons une méthode pour créer rapidement un système de désambiguïsation lexicale supervisée pour une langue L peu dotée. Cette méthode nécessite un système de traduction automatique statistique (TAS) d'une langue riche en corpus annotés en sens vers L . Il est, en effet, plus facile de disposer des ressources nécessaires à la création d'un système de TAS (des corpus alignés) que des ressources dédiées nécessaires à la création d'un système de désambiguïsation lexicale pour la langue L . Le système de DL pour la langue L sera alors construit en utilisant les traductions annotées dans la langue L produite.

Dans cet article, nous présentons la désambiguïsation lexicale en fonction des ressources qu'elle utilise et montrons que beaucoup de langues sont trop peu dotées pour permettre la construction d'un système de DL supervisée. Nous présentons notre méthode de construction de corpus annotés par traduction automatique et l'illustrons avec le français et le bengali. Enfin, nous évaluons la méthode sur le corpus de la tâche de désambiguïsation lexicale multilingue de Semeval 2013.

2 Désambiguïsation lexicale et langue

2.1 Processus général de la désambiguïsation lexicale

La mise en place d'un système de désambiguïsation lexicale se déroule en trois étapes :

1) constitution de ressources génériques : plusieurs ressources non dédiées à la désambiguïsation lexicale sont envisageables : dictionnaires, encyclopédies, corpus non annotés, corpus annotés, bases lexicales, ... Certaines sont construites automatiquement parfois en utilisant d'autres ressources. Cette étape est optionnelle et est souvent réalisée par des équipes spécialisées. Elle est celle qui demande le plus de supervision humaine.

2) constitution d'une ressource dédiée à la DL : utilisation d'une ou plusieurs ressources génériques pour associer une représentation informatique à chacun des sens d'un mot. Ces sens sont soit directement définis à partir de l'expertise humaine pour certaines ressources génériques comme les bases lexicales, soit induits à partir des contextes d'utilisation dans les textes (induction de sens). Structuellement, la ressource peut être, par exemple, un graphe, des chaînes de caractères ou des représentations vectorielles ;

3) utilisation de la ressource dédiée pour désambiguïser des textes ; il s'agit de l'algorithme de désambiguïsation proprement dit. Plusieurs paramètres peuvent être définis pour le traitement. Certains sont communs à tous les algorithmes comme la taille du contexte considéré pour un mot cible (par exemple quelques mots avant ou après la cible, la phrase qui contient la cible, voire le texte) tandis que d'autres dépendent du type d'algorithme mis en œuvre (par exemple la limite à considérer pour la profondeur de la recherche dans un graphe ou encore les paramètres à considérer pour des algorithmes stochastiques).

Ainsi, selon ce point de vue, (Schwab *et al.*, 2013) utilisent WordNet comme ressource générique, une représentation sous forme de sacs de mots issus des définitions des sens et de leurs liens comme ressource dédiée, un algorithme à colonies de fourmis et une mesure de proximité entre les sacs de mots comme algorithme de désambiguïsation lexicale. Roberto Navigli et son équipe (Navigli & Ponzetto, 2012) utilisent *BabelNet* comme ressource générique, une représentation sous forme de graphe issu des sens et de leurs liens comme ressource dédiée, des algorithmes de graphes (*Pagerank*, *Degree*, ...) comme algorithmes de désambiguïsation.

2.2 Ressources pour la Désambiguïsation lexicale

En désambiguïsation lexicale, deux types de ressources sont importantes : des corpus manuellement annotés par des sens et des sources de connaissances. Les campagnes d'évaluation sur l'anglais ont globalement montré que plus un système utilise de corpus annotés, meilleurs sont les résultats. De même, meilleures sont les sources de connaissances en termes de quantité et de qualité, meilleurs sont les résultats. Dans le processus d'informatisation d'une langue, avant de pouvoir construire un corpus annoté manuellement par des sens, il faut disposer d'un inventaire de sens. Aucune autre langue que l'anglais ne bénéficie d'autant de corpus de textes manuellement annotés par des sens et de connaissances lexicales. La figure 1 permet d'illustrer de manière parlante l'état des ressources nécessaires pour la DL librement accessibles pour un certain nombre de langues. Il est donné pour que le lecteur puisse se faire une idée de la situation actuelle. Un recensement plus précis serait difficile à obtenir et il faut ainsi interpréter les positions des langues les unes par rapport aux autres plutôt

que de manière absolue sauf pour l'anglais que nous avons placé le plus en haut à droite. De fait si on peut considérer que la quantité de données annotées est un paramètre quantifiable (par exemple en nombre moyen d'occurrences par terme du lexique), la richesse des sources de connaissances disponibles est, elle, plus floue. C'est en particulier le cas entre deux langues différentes puisque la taille de leur vocabulaire est différente. Il faut noter également que certaines langues peuvent bénéficier de données provenant d'autres langues par des alignements (comme c'est le cas dans BabelNet, par exemple).

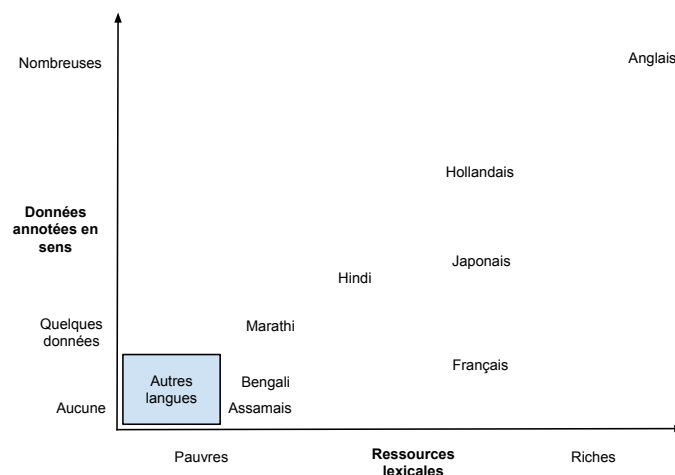


FIGURE 1: Données disponibles pour la désambiguïsation lexicale en fonction de la langue

2.2.1 Bases lexicales

WordNet Avant les années 1990, la désambiguïsation lexicale n'était pratiquement réalisée qu'à partir de dictionnaires électroniques. Le *Princeton WordNet* (Fellbaum, 1998), initié au milieu des années 1980, a permis la mise à disposition d'une ressource utilisable librement. Elle est devenue rapidement très populaire et a rapidement conduit à la disparition de l'usage des dictionnaires électroniques en désambiguïsation lexicale.

Le *Princeton WordNet* est organisé autour de la notion d'ensemble de synonymes (*synsets*) décrits par une partie du discours (nom, verbe, adjectif, adverbe), une définition et leurs liens (hyperonyme, hyponyme, antonyme, ...). Chaque sens d'un item lexical (entrée) correspond à un *synset*. La version courante du *Princeton WordNet*, la 3.0, comprend 155 287 items lexicaux pour un total de 117 659 *synsets*. Des versions pour d'autres langues existent mais, faute de moyens humains équivalents, leur qualité est encore inférieure à celle de l'anglais. Bien souvent, les mots de ces langues sont décrits grâce à des *synsets* du *Princeton WordNet*. La *Global WordNet Association* établit la liste des wordnets existants¹.

BabelNet BabelNet (Navigli & Ponzetto, 2012) est une ressource lexicale à grande échelle construite par alignement automatique des *synsets*, issus de *Princeton WordNet* et de pages Wikipedia correspondantes. BabelNet introduit la notion de *Babel Synset*, qui contient tout le contenu du *synset* correspondant dans le *Princeton WordNet*, ainsi qu'un ensemble de pages Wikipedia similaires. Cette correspondance entre *synsets* WordNet et pages Wikipédia se fait par un algorithme de désambiguïsation automatique. Les pages Wikipédia reliées par des hyperliens internes à Wikipédia ainsi que les articles associés dans les autres langues disponibles dans Wikipedia sont liés aux pages correspondantes. Pour toutes les pages dans les autres langues, si il n'y a pas de définition disponible ou extraite de la page, la définition anglaise *Princeton WordNet* ou un extrait venant de *SemCor* est traduit par *Google Translate* pour servir de définition.

BabelNet, dans sa dernière version, la 2.5.1, comprend 271 langues, 13 789 332 *Babel synsets*, 117 204 438 sens, 354 538 633 relations lexico-sémantiques et 40 328 194 définitions textuelles. Pour l'anglais il y a 11 812 887 entrées, 6 670 627 *Babel synsets*, 16 741 223 sens de mots et un degré de polysémie de 7,51. Pour le français, il y a 5 295 221

1. <http://globalwordnet.org/wordnets-in-the-world/>

entrées, 4 120 733 Babel synsets, 7 076 728 sens de mots et un degré de polysémie de 1,72. Pour le bengali il y a 188 511 entrées, 34 832 Babel synsets, 233 163 sens de mots et un degré de polysémie de 6,69.

2.2.2 Corpus annotés

Selon Benoît Habert, «*un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue*» (Habert *et al.*, 1998). Généralement, un corpus contient jusqu'à une douzaine de millions de mots et peut être lemmatisé et annoté avec des informations concernant les parties du discours. Parmi ces corpus, on trouve le *British National Corpus* (Burnard, 1998) (100 millions de mots) et le *American National Corpus* (Ide & Macleod, 2001) (20 millions de mots). Les textes proviennent de diverses sources comme des journaux, des livres, des encyclopédies ou du Web.

Exemples de corpus annotés En désambiguïsation lexicale, plusieurs corpus annotés en sens sont utilisés. On peut citer, par exemple :

1. La *Defense Science Organisation* (Ng & Lee, 1996) a produit un corpus non disponible librement. 192 800 mots ont été annotés avec des *synsets* du *Princeton WordNet*. L'annotation se concentre sur 121 noms (113 000 occurrences) et 70 verbes (79 800 occurrences) qui ont été choisies parmi les plus fréquents et les plus ambigus de l'anglais. Selon les auteurs, la couverture correspond à environ 20% des occurrences de noms et de verbes en anglais.
2. Le *SemCor* (Miller, 1995) est un sous-ensemble du Corpus de Brown (Francis & Kučera, 1964). Sur les 700 000 mots de ce dernier, environ 230 000 sont annotés avec des *synsets* du *Princeton WordNet*. L'annotation porte au total sur 352 textes. Pour 186 d'entre eux, 192 639 mots (soit l'ensemble des noms, verbes, adjectifs et adverbes) sont annotés. Sur les 166 autres, seulement 41 497 verbes sont annotés.
3. Les corpus issus des campagnes d'évaluation. Depuis 1998, il y a eu plusieurs campagnes (semeval-senseval) destinées à évaluer la désambiguïsation lexicale. La plupart ont concerné l'anglais mais également le japonais, l'espagnol, le chinois ou le français. La taille de ces corpus est de l'ordre d'une centaine de fois plus petite que celle des deux précédents corpus, soit quelques milliers de mots.

Difficultés liées à la construction d'un corpus annoté Il n'existe que peu de données manuellement annotés. La *Global WordNet Association* dresse la liste des 26 corpus annotés avec un wordnet². Ces corpus concernent 17 langues. Seules trois d'entre elles (l'anglais, le hollandais et le bulgare) atteignent les 100 000 annotations. À notre connaissance, il n'existe pas de donnée annotée pour le bengali et très peu pour le français (environ 3600 mots annotés avec le dictionnaire Larousse pour la campagne Romanceval 1998 et 1656 mots annotés avec des sens de BabelNet pour la tâche 12 de la campagne SemEval 2013). Ces deux langues qui nous intéressent plus particulièrement sont donc peu dotées en ce domaine.

La construction d'un corpus manuellement annoté en sens est réputée comme une tâche très difficile par comparaison à d'autres tâches d'annotation. En effet, s'il n'y avait que 45 annotations possibles pour le *Penn Treebank* (Marcus *et al.*, 1993), un corpus annoté en parties du discours, il y en a autant que de *synsets* (117 000) pour une annotation en sens issus du *Princeton WordNet*. Ainsi, pour l'annotation du corpus de la *Defense Science Organisation*, alors que les conditions étaient plus favorables que celles des annotateurs du *SemCor* (uniquement 191 mots différents pour seulement 1 800 annotations possibles), le taux d'annotation était seulement de 150 à 250 mots par heure (1 homme-année pour les 192 800 occurrences de mots) tandis que les annotateurs du *Penn Treebank* réalisaient 6 000 annotations par heure.

Dans de telles conditions, on comprends mieux pourquoi assez peu de corpus annotés existent. Des recherches ont visé à faciliter cette annotation. Par exemple, (Vossen *et al.*, 2011) utilisent, pour le hollandais, un algorithme de désambiguïsation automatique dont les annotations les moins sûres sont vérifiées/modifiées par les annotateurs et (Mihalcea & Chklovski, 2003) utilisent des méthodes de *crowdsourcing* pour augmenter le nombre d'annotateurs.

Les corpus de grande taille annotés en sens sont pourtant le seul moyen de mettre en œuvre un processus de désambiguïsation lexicale supervisée.

2. <http://globalwordnet.org/wordnet-annotated-corpora/> consultée le 4 février 2015. Il existe d'autres corpus annotés avec des *synsets* de wordnets comme ces corpus de domaines annotés avec des *synsets* de l'*Hindi WordNet* http://www.cfilt.iitb.ac.in/wsd/annotated_corpus/

2.3 Désambiguïsation lexicale supervisée

Le principe, issu de l'apprentissage automatique, consiste à entraîner un classifieur pour chaque mot cible, afin de prédire le sens le plus vraisemblable en fonction de son contexte. Dans les termes utilisés à la section 2.1, la ressource générique est le corpus annoté en sens, la ressource dédiée est construite au moyen de classifieurs utilisés pour discriminer le contexte de chaque mot afin de déterminer son meilleur sens.

Ces approches (dites supervisées), sont très populaires pour l'anglais. Dans les campagnes d'évaluation Senseval-SemEval, elles ont obtenu, de loin, les meilleures performances sur l'anglais. La limitation principale est la nécessité de disposer de grands corpus annotés ce dont peu de langues disposent comme nous l'avons vu dans la partie précédente. Dans cet article, nous proposons une méthode fondée sur la traduction automatique statistique et le portage d'annotations d'un corpus comme nous le montrons dans la suite.

3 Construction d'un corpus annoté par traduction automatique

3.1 Principe général

Dans ce travail, nous proposons une méthode de traduction automatique et de transfert direct des annotations qui nous permet d'obtenir des corpus dans toutes les langues disposant d'un système de traduction avec comme source une langue possédant un corpus annoté en sens. Notre méthode a été mise en œuvre sur le *SemCor*, corpus en anglais annoté avec des sens issus du *Princeton WordNet* (voir section 2.2.2) à l'aide d'un système de traduction construit avec la boîte à outils *Moses*, de l'anglais vers le français et de l'anglais vers le bengali. Nous avons ainsi obtenu un corpus annoté en sens issus du *Princeton WordNet* pour le français et un autre pour le bengali.

3.2 Transfert d'annotations

À notre connaissance, le transfert d'annotations linguistiques a été utilisé à partir des années 1990 (Brown *et al.*, 1991). Dans cette approche, le principe consistait à exploiter des corpus parallèles (source, cible) annotés à la source et de construire un modèle d'alignements qui permet de transférer ces annotations vers la cible.

De tels transferts d'annotations ont été appliqués à une large gamme d'annotations : les parties du discours (Yarowsky & Ngai, 2001), les dépendances syntaxiques (Hwa *et al.*, 2005), les *chunks* (Yarowsky *et al.*, 2001), les rôles sémantiques (Padó & Lapata, 2009), *etc.* De plus récents travaux comme (van der Plas & Apidianaki, 2014) n'exploitent pas directement les données parallèles mais utilisent des techniques de désambiguïsation par traductions pour déterminer les alignements de mots et projeter les annotations. De leur côté, (Wang & Manning, 2014) utilisent un corpus parallèle et un système de désambiguïsation lexicale multilingue pour obtenir les traductions en contexte de chaque mot du corpus, ce qui leur permet d'ensuite transférer une annotation en rôle sémantique depuis l'anglais vers le français.

Dans le contexte plus spécifique de la désambiguïsation lexicale, (Diab & Resnik, 2002) utilisent un système de traduction automatique commercial pour traduire un corpus en anglais vers une langue cible, les mots anglais provenant de la source sont alors utilisés comme annotations sémantiques dans la cible. Le projet *MultiSemCor*, (Padó & Lapata, 2009) est très proche de l'approche que nous présentons ici puisqu'il s'agit de faire traduire en italien par des traducteurs professionnels une sous-partie du *SemCor*. Notre approche consiste à simultanément traduire le corpus et à porter les annotations de la source grâce à un système de TAS construit avec la boîte à outil *Moses*.

3.3 Traduction automatique statistique

La traduction automatique (TA) est le processus qui consiste à traduire un énoncé d'une langue naturelle (langue source) à une autre (langue cible). L'énoncé peut-être soit écrit ou oral mais nous ne nous intéresserons qu'à l'écrit dans cet article.

La traduction automatique statistique (TAS) est actuellement une approche très largement utilisée en TA. Schématiquement, un système est, dans un premier temps, entraîné sur des corpus parallèles langue source - langue cible. Il s'agit d'obtenir des informations statistiques permettant de calculer quelles sont les meilleures traductions pour un mot ou une

suite de mots (approche dite *phrase-based*). Dans un second temps, ces informations sont exploitées pour produire des traductions de la langue source à la langue cible.

Moses³ (Hoang & Koehn, 2008) est une boîte à outils pour la traduction automatique statistique. Sa licence est libre (LGPL licence) ce qui facilite l’obtention d’un système statistique complet à l’état-de-l’art.

3.4 Mise en œuvre de notre approche

Notre approche consistant à traduire le *SemCor* et à porter ses annotations a pu être mise en œuvre grâce à un premier système de traduction de l’anglais vers le français et un second système depuis l’anglais vers le bengali que nous présentons maintenant.

3.4.1 Système anglais–français

Le système de traduction statistique anglais–français est celui mis au point par le Laboratoire d’Informatique de Grenoble pour participer à la campagne 2012 d’IWSLT (*International Workshop on Spoken Language Translation*) (Besacier *et al.*, 2012). Ce système a été construit avec des données alignées usuelles (*Europarl Parallel Corpus*, *United Nations Parallel Corpus*, ...). Il a été évalué avec la métrique BLEU — score de 24,85 — ainsi que par des juges humains — score de 11.

3.4.2 Système anglais–bengali

Notre système anglais–bengali n’a pas encore fait l’objet d’une publication, nous le décrivons donc de manière plus détaillée. Il est basé sur la boîte à outil *Moses* (version 2.1). Les données proviennent de différentes sources :

- Corpus parallèles : EMILLE corpora, OPUS corpus (KDE4, GNOME), INDIC, OpenSubtitles2013, Tanzil, Ta-toeba, Bibel, Jehova Witness
- Corpus monolingues pour le bengali : l’ensemble des corpus parallèles cités ci-dessus, extraction du Wikipedia bengali du 28 décembre 2014 qui comporte 28393 articles.

Les ressources ont été pré-traitées comme suit : (1) filtrage des marqueurs HTML, (2) conversion de tous les caractères en UTF-8, (3) application de la normalisation forme D, (4) application de la normalisation des ponctuations (retrait des marques de ponctuation redondantes, conversion vers la version canonique des caractères, *etc.*), (5) tokenisation, (6) suppression des phrases de plus de 50 mots (7) suppression des phrases trop longues (ratio supérieur à 9).

Après ce pré-traitement, nous obtenons ces statistiques :

données parallèles

phrases	679 019		
	tokens	types	mots
anglais	12 096 756	165 765	132 404
bengali	11 288 581	229 634	14 497

données monolingues

phrases	1 107 776		
	tokens	types	mots
bengali	21 273 100	317 265	18 632

Nous avons utilisé 85% de ces données pour l’entraînement, 10% pour l’optimisation du système et 5% pour le tester. Le score bleu de ce système est de 27,21.

3.4.3 Traduction du *SemCor* et portage des annotations

Pour traduire le *SemCor* (Miller *et al.*, 1993) (v. 3.0) et porter ses annotations vers la langue cible, nous utilisons nos systèmes de traduction automatique statistique. Un extrait du *SemCor* est présenté dans la figure 2. En même temps que le système traduit le texte, phrase par phrase, nous extrayons la correspondance mot à mot que nous fournit le décodeur *Moses* et nous l’utilisons pour transférer les annotations d’un mot source au mot correspondant dans le texte cible suivant

3. <http://www.statmt.org/moses/>

```

<contextfile concordance=brown>
<context filename=br-a01 paras=yes>
  <p pnun=1>
    <s snun=1>
      <wf cmd=ignore pos=DT>The</wf>
      <wf cmd=done rdf=group pos=NNP lemma=group wnsn=1 lexs=1:03:00:: pn=group>
      Fulton_County_Grand_Jury</wf>
      <wf cmd=done pos=VB lemma=say wnsn=1 lexs=2:32:00::>said</wf>
      <wf cmd=done pos=NN lemma=friday wnsn=1 lexs=1:28:00::>Friday</wf>
      <wf cmd=ignore pos=DT>an</wf>
      <wf cmd=done pos=NN lemma=investigation wnsn=1 lexs=1:09:00::>
      investigation</wf>
      <wf cmd=ignore pos=IN>of</wf>
      <wf cmd=done pos=NN lemma=atlanta wnsn=1 lexs=1:15:00::>Atlanta</wf>
      [...]
    <punc>.</punc>
  </s>
</p>
[... ]
</context>
</contextfile>

```

FIGURE 2: *SemCor*(v. 3.0) dans le format SGML (Standard Generalized Markup Language).

l’algorithme 1. La traduction et le transfert des annotations a pris environ une semaine en utilisant un seul cœur sur un serveur 32-cœur Intel Xeon E5-2650, 2.0 GHz et 256 GB de mémoire physique.

4 Systèmes de désambiguïsation lexicale

À partir de ces corpus annotés en sens, nous pouvons utiliser des méthodes de désambiguïsation supervisée pour construire un ou plusieurs systèmes de désambiguïsation automatique. Dans cette section, nous présentons une méthode supervisée, un classifieur bayésien naïf et son évaluation sur le corpus de *SemEval 2013*.

4.1 Principe

Un corpus annoté contient un grand nombre de textes segmentés en phrases et en mots puis lemmatisés et annotés sémantiquement. Ainsi, pour chaque mot du corpus, nous pouvons obtenir, une liste de phrases dans lesquelles ce mot apparaît dans ses différents sens et en extraire des attributs du contexte prédicteurs du sens. Les prédicteurs habituels incluent les catégories grammaticales des mots du cotexte, les lemmes de ces mots ainsi que leur position dans la phrase.

Un classifieur est un algorithme nécessitant d’apprendre un modèle de prédiction afin d’affecter une série d’instance à un ensemble fini de classes. Pour cela, il faut fournir au classifieur une série d’instances annotées avec les classes auxquelles elle appartiennent afin de construire le modèle. Dans le cas de la désambiguïsation supervisée, les classes à prédire sont les identifiants des sens et les instances sont les attributs prédicteurs extraits.

4.2 Classifieur bayésien naïf

Pour un ensemble d’instances $I_x = (x_1, x_2, \dots, x_n)$ annotées avec N classes $C_k, k \in 1..N$, un classifieur bayésien naïf estime la probabilité d’obtenir une classe C_k en fonction d’un certain ensemble d’attributs x_1, \dots, x_n : $P(C_k | x_1, \dots, x_n) = p(C_k)p(x_1 | C_k)p(x_2 | C_k, x_1) \dots p(x_n | C_k, x_1, \dots, x_n)$. L’algorithme fait la supposition de l’indépendance des attributs les uns par rapport aux autres, ce qui implique l’approximation suivante, avec $Z = p(x)$ un facteur de normalisation :

$$P(C_k | x_1, \dots, x_n) \simeq \frac{1}{Z} p(C_k) \prod_{i=1}^n p(x_i | C_k)$$

Algorithm 1: Processus de traduction du *SemCor* et portage de ses annotations

Entrées: SC : SemCor
Sorties: TSC : SemCor traduit annoté
decoder : chemin absolu vers le décodeur (dans notre cas, *Moses*);
model : chemin absolu vers le fichier du modèle entraîné (*moses.ini*);
-config : chemin absolu vers le fichier de configuration (option de *Moses*);
-print-alignment-info : sortie alignement mot à mot vers la sortie standard, séparée des traductions par ||| (option de *Moses*);
initialization;
SemCorTranslation()

```

    pour tous les répertoires  $D \in SC$  faire
        créer un nouveau répertoire  $TD$ ;
        pour tous les fichiers  $F \in D$  faire
            phrase  $S \leftarrow \{\}$ ;
            créer un nouveau fichier  $TF$ ;
            pour tous les ligne  $L \in F$  faire
                pour tous les WF entrée  $E \in L$  faire
                     $WF[cmd, rdf, lemma, pos, lexs, wns, pn, ot, word, punc] \leftarrow$ 
                        ExtractWFInfoValue( $E$ );
                fin
                 $S \leftarrow S \cup \{WFInfo.word\}$ ;
            fin
            target-words  $TWs \leftarrow$  GetAlignedWords( $S$ );
            pour  $i \in 1 \dots |TWs|$  faire
                 $S[i].word \leftarrow TWs[i]$ ;  $S[i].lemma \leftarrow TWs[i]$ ;
            fin
            target-sense  $TS \leftarrow$  SenseMapper( $pos, lexs, word$ );
             $S.lexs \leftarrow TS$ ;
            écrire  $S$  dans le fichier  $TF$ ;
        fin
    fin

```

ExtractWFInfoValue(E)

```

    WF Information  $WFI \leftarrow \{\}$ ; WF Valeur  $WV \leftarrow \{\}$ ;
    pour tous les item  $I \in E$  faire
        split  $I$  by "=";
         $WFI[] \leftarrow I[left\_item]$ ;  $WV[] \leftarrow I[right\_item]$ ;
    fin
    retourner  $WFI[], WV[]$ ;

```

GetAlignedWords(S)

```

    alignements de mots  $WAs \leftarrow \{\}$ ; target-words  $TWs \leftarrow \{\}$ ;
    sortie de traduction  $TO \leftarrow$  MosesDecoder( $S$ );
    split  $TO$  by "|||";
     $WAs[] \leftarrow TO[right\_item]$ ;
    pour tous les word alignment  $WA \in WAs$  faire
        split  $WA$  by "-";
         $TWs[] \leftarrow WA[right\_item]$ ;
    fin
    retourner  $TWs[]$ ;

```

MosesDecoder(S)

```

    sortie de la traduction avec alignement de mots  $TO \leftarrow$  -config model  $S$  -print-alignment-info;
    retourner  $TO$ ;

```

L'estimation du modèle est souvent faite par estimation de vraisemblance. Lorsque le modèle est construit, la décision (classification) est faite avec la règle du maximum *a posteriori* :

$$\hat{C}_k = \arg \max_{k \in 1..N} p(C_k) \prod_{i=1}^n p(x_i | C_k)$$

Ce classifieur, simple à mettre en œuvre et extrêmement performant en termes de complexité d'apprentissage et de classification, est utilisé dans de nombreuses tâches, comme la détection de pourriels. Il s'utilise aussi couramment comme base de comparaison dans les travaux portant sur la classification supervisée. Nous l'avons utilisé ici car il offre de bonnes performances en désambiguïsation de l'anglais même avec peu de données d'entraînement (voir section 4.3) et permet de comprendre facilement d'où viennent les différences de performance dans la classification entre deux langues.

4.3 Systèmes implantés

Pour construire le système de désambiguïsation supervisée nous nous sommes basés sur les caractéristiques du système arrivé premier sur la tâche 7 de la campagne d'évaluation Semeval 2007, NUS-PT (Chan *et al.*, 2007). Suivant les travaux de (Lee & Ng, 2002) qui proposent une étude sur les meilleurs attributs à utiliser, NUS-PT en extrait trois types :

- *Les collocations locales* qui sont des séquences de mots autour du mot ciblé (inclus). Onze combinaisons du type $C_{i,j}$ sont utilisées, où i et j représentent les bornes de la collocation par rapport au mot ciblé : $C_{-1,-1}$; $C_{1,1}$; $C_{-2,-2}$; $C_{2,2}$; $C_{-2,-1}$; $C_{-1,1}$; $C_{1,2}$; $C_{-3,-1}$; $C_{-2,1}$; $C_{-1,2}$; $C_{1,3}$. Par exemple, dans la phrase «*Le chaton blanc est très heureux*», si le mot ciblé est «blanc», alors $C_{-1,1}$ correspond à la collocation «*chaton blanc est*», et $C_{1,3}$ correspond à «*est très heureux*» ;
- *Les catégories grammaticales* avoisinantes qui correspondent aux catégories grammaticales de trois mots sur la gauche et de trois mots sur la droite ;
- *les mots du cotexte* correspondant à la même fenêtre de $-3, 3$.

Dans le système que nous avons implanté, nous avons réutilisé les mêmes attributs. Toutefois, pour cette expérience, qui consiste uniquement à examiner la faisabilité de l'approche, nous n'utilisons que le *SemCor* contrairement à NUS-PT qui utilise également le corpus de la DSO et des exemples extraits de corpus de tests parallèles.

Dans les cas où le classifieur ne retourne pas de réponse (pas assez d'exemples d'entraînement, mots non existants dans *SemCor*), nous utilisons la solution de repli classique, utilisée par exemple par NUS-PT, qui consiste à assigner le premier sens (le plus fréquent).

Trois systèmes ont été créés :

- un système permettant de désambiguïser de l'anglais avec des sens issus du *Princeton WordNet* ou des sens issus de *BabelNet* grâce aux alignements *synsets-Babel Synsets*. Il s'agit de notre système de référence construit directement à partir du *SemCor* ;
- un système permettant de désambiguïser du bengali avec des sens issus du *Princeton WordNet*. Ce système est construit à partir de la traduction vers le bengali du *SemCor*. Le bengali ne disposant pas de corpus annotés en sens à notre connaissance, une évaluation classique directe de ses performances (*in vitro*) n'est pas possible. Il fait l'objet, en revanche, d'étude de ses performances *in vivo*, c'est-à-dire utilisé dans une application, mais ce n'est pas l'objet de cet article. L'évaluation *in vitro* du troisième système présenté ici nous permettra d'avoir une idée des performances de cet annotateur du bengali ;
- un système permettant de désambiguïser du français avec des sens issus du *Princeton WordNet* ou des sens issus de *BabelNet* grâce aux alignements *synsets-BabelSynsets*. Ce système, construit à partir de la traduction vers le français du *SemCor*, est comparé au système de référence pour l'anglais sur la tâche de désambiguïsation multilingue (tâche 12) de Semeval 2013.

Ces trois systèmes sont mis à la disposition de la communauté et accessibles à l'adresse <http://getalp.imag.fr/static/wsd/TALN2015/>.

5 Évaluation de la méthode

Pour tester notre approche et vérifier qu'elle est générique, nous nous plaçons dans un contexte absolument similaire pour le système du bengali et les systèmes du français. La seule source de données utilisée dans les deux cas est donc une

traduction du *SemCor* (voir partie 3.4.3).

Nous allons chercher à estimer la perte de performance liée à l'utilisation de la traduction du corpus. Cela est rendu possible par l'utilisation du corpus de la tâche 12 (désambiguïsation lexicale multilingue) de Semeval 2013 qui est un corpus d'évaluation traduit en 5 langues (anglais, français, allemand, italien, espagnol) pour lequel nous utilisons les textes anglais et français.

5.1 Corpus de Semeval 2013 tâche 12 : désambiguïsation lexicale multilingue

Comme mentionné ci-dessus, le corpus de Semeval 2013 (tâche 12) comporte 5 langues dans lesquelles il a été traduit ainsi que les annotations sémantiques transférées puis validées manuellement. Il comprend 13 textes de différents domaines (politique, commentaire sportif, domaine général).

Puisque *SemCor* est annoté avec les sens de *Princeton WordNet* et que l'évaluation se fait avec des sens BabelNet, nous avons réalisé une conversion en utilisant les alignements de BabelNet avec WordNet⁴. Nous avons d'abord récupéré les offsets des synsets correspondant aux annotations de sens dans WordNet puis avons interrogé l'API java de BabelNet pour obtenir les identifiants BabelNet correspondants. Nous avons obtenu une correspondance de **96,10%** sur l'ensemble de *SemCor*. Nous avons ensuite réalisé l'apprentissage pour obtenir les systèmes présentés en 4.3 puis les avons appliqués sur notre corpus d'évaluation.

5.2 Mesures d'évaluation

La tâche de désambiguïsation lexicale multilingue de SemEval 2013 utilise les mesure classiques de précision P , de rappel R et de score F_1 qui correspond à la moyenne harmonique de P et R . La précision se définit comme $P = \frac{\text{annotés correctement}}{\text{total annotés}}$, le rappel comme $R = \frac{\text{annotés correctement}}{\text{total à annoter}}$ et le score F_1 comme $F_1 = \frac{2 \cdot P \cdot R}{P + R}$.

5.3 Résultats et analyse

Rappelons, tout d'abord, qu'il ne s'agissait pas ici d'obtenir des meilleurs scores possibles mais de montrer qu'il est possible de créer des systèmes de désambiguïsation lexicale pour des langues qui n'ont pas (ou trop peu) de données annotées en sens.

Système	Précision	Rappel	Score F1
SUP-EN	64,80%	64,70%	64,75%
MFS-EN	66,90%	66,60%	66,64%
SUP-FR	51,60%	51,50%	51,55%
MFS-FR	45,60%	45,10%	45,34%

TABLE 1: Les résultats (Précision, Rappel, score F1) de l'expérience. MFS-EN/FR sont respectivement les résultats de la référence avec sens le plus fréquent en anglais et en français. SUP-EN/SUP-FR sont les résultats pour le système supervisé en anglais (apprentissage directement sur le *SemCor*) et en français (apprentissage sur une traduction du *SemCor*).

Le tableau 1 présente les résultats de l'expérience. Les systèmes MFS-EN et MFS-FR sont les résultats obtenus en assignant systématiquement le sens le plus fréquent (*Most Frequent Sense* - heuristique du sens le plus fréquent) comme réponse. Nous pouvons déjà constater que cette référence classique en évaluation de la désambiguïsation lexicale, fournit des résultats pour le français bien inférieurs à ceux de l'anglais.

SUP-EN correspond à notre système de référence, pour l'anglais, construit directement à partir du *SemCor*. Il obtient un score inférieur d'environ 2% sur l'heuristique du sens le plus fréquent. Le système créé à partir de la traduction vers le français du *SemCor*, SUP-FR obtient 51,6% soit une perte de 13,2% si on compare à l'anglais mais il s'agit d'un score qui est supérieur de 6.21% de la référence du sens le plus fréquent.

À notre connaissance, il n'existe pas d'autre expérience utilisant un système supervisé sur ce corpus. Il convient donc de considérer avec précaution les comparaisons avec les systèmes ayant participé à la tâche car ils bénéficiaient de la

4. BabelNet est construit sur la base de Wordnet 3.1 alors que *SemCor* 3.0 est basé sur *Princeton WordNet* 3.0

richesse des informations de BabelNet. 3 équipes ont participé à la campagne. Notre système référence sur l'anglais aurait été entre la première (−3, 7%) et la seconde (+4, 4%) tandis que le système sur le français aurait terminé entre la seconde (−2, 25%) et la troisième (+3, 25%).

Nous voyons ici qu'avec un système supervisé non optimisé avec des données d'entraînement qui peuvent être facilement étoffées en ajoutant d'autres corpus, nous obtenons déjà des résultats qui valident notre approche.

6 Conclusion et perspectives

Dans cet article, nous avons montré qu'il existe des corpus annotés dans certaines langues (en anglais par exemple) alors qu'il en existe peu ou pas dans la plupart des langues. Ces corpus sont pourtant essentiels à la création de systèmes de désambiguïsation lexicale supervisée. Nous avons proposé une méthode qui consiste à traduire des corpus annotés et à porter ces annotations dans la langue pour laquelle on veut un système de désambiguïsation. Nous avons utilisé le même script sur un système *Moses* traduisant de l'anglais vers le bengali et de l'anglais vers le français pour créer un système de désambiguïsation du bengali et un système de désambiguïsation du français. Nous avons ainsi montré la faisabilité et la généralité de l'approche. Nous avons aussi montré qu'en utilisant un apprentissage supervisé naïf, sur assez peu de données, qui plus est traduites automatiquement, on obtient des performances suffisantes pour valider cette approche.

Dans l'avenir, nous envisageons d'utiliser plus de corpus annotés libres ou gratuits pour la recherche mais également d'acquérir et d'utiliser le DSO qui, lui, ne l'est pas. Nous souhaitons également comparer différents algorithmes supervisés voire les combiner par une fusion tardive, par exemple. Enfin une dernière piste d'amélioration consistera à essayer d'améliorer les attributs du contexte permettant l'apprentissage automatique en intégrant, par exemple, les attributs utilisés par d'autres systèmes.

Références

- BESACIER L., LECOUEUX B., AZOUZI M. & LUONG NGOC Q. (2012). The LIG English to French Machine Translation System for IWSLT 2012. In *In proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT)*, p. 102–108, Unknown.
- BROWN P., PIETRA S. D., PIETRA V. D. & MERCER R. (1991). Word-sense disambiguation using statistical methods. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, p. 264–270 : Association for Computational Linguistics.
- BURNARD L. (1998). *The British National Corpus*.
- CHAN Y. S., NG H. T. & ZHONG Z. (2007). Nus-pt : exploiting parallel texts for word sense disambiguation in the english all-words tasks. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, p. 253–256 : Association for Computational Linguistics.
- DIAB M. & RESNIK P. (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, p. 255–262, Stroudsburg, PA, USA : Association for Computational Linguistics.
- FELLBAUM C. (1998). *WordNet*. Wiley Online Library.
- FRANCIS W. N. & KUČERA H. (1964). *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown)*. Rapport interne, Brown University, Providence, Rhode Island.
- HABERT B., FABRE C. & ISSAC F. (1998). *DE L'ECRIT AU NUMERIQUE. Constituer, normaliser et exploiter les corpus électroniques*. Number ISBN : 2-225-82953-5. ELSEVIER MASSON.
- HOANG H. & KOEHN P. (2008). Design of the moses decoder for statistical machine translation. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, p. 58–65 : Association for Computational Linguistics.
- HWA R., RESNIK P., WEINBERG A., CABEZAS C. & KOLAK O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Nat. Lang. Eng.*, **11**(3), 311–325.
- IDE N. & MACLEOD C. (2001). The american national corpus : A standardized resource of american english. In *Proceedings of Corpus Linguistics 2001*, volume 3.

- LEE Y. K. & NG H. T. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, p. 41–48, Stroudsburg, PA, USA : Association for Computational Linguistics.
- MARCUS M. P., MARCINKIEWICZ M. A. & SANTORINI B. (1993). Building a large annotated corpus of english : The penn treebank. *Computational linguistics*, **19**(2), 313–330.
- MIHALCEA R. & CHKLOVSKI T. (2003). *Building sense tagged corpora with volunteer contributions over the Web*, p. 357–402. John Benjamin Publishing Compagny.
- MILLER G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, **38**(11), 39–41.
- MILLER G. A., LEACOCK C., TENGI R. & BUNKER R. T. (1993). A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, p. 303–308 : Association for Computational Linguistics.
- NAVIGLI R. & PONZETTO S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**, 217–250.
- NG H. T. & LEE H. B. (1996). Integrating multiple knowledge sources to disambiguate word sense : An exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, p. 40–47 : Association for Computational Linguistics.
- PADÓ S. & LAPATA M. (2009). Cross-lingual annotation projection of semantic roles. *J. Artif. Int. Res.*, **36**(1), 307–340.
- SCHWAB D., GOULIAN J. & TCHECHMEDJIEV A. (2013). Worst-case complexity and empirical evaluation of artificial intelligence methods for unsupervised word sense disambiguation. *International Journal of Web Engineering and Technology*, **8**(2), 124–153.
- VAN DER PLAS L. & APIDIANAKI M. (2014). Cross-lingual word sense disambiguation for predicate labelling of french. In *Proceedings of TALN 2014 (Volume 1 : Long Papers)*, p. 46–55 : Association pour le Traitement Automatique des Langues.
- VOSSEN P., GÖRÖG A., LAAN F., VAN GOMPEL M., IZQUIERDO-BEVIA R. & VAN DEN BOSCH A. (2011). Dutch-semco : building a semantically annotated corpus for dutch. In *Electronic lexicography in the 21st century : New Applications for New Users : Proceedings of eLex 2011, Bled, 10-12 November 2011*, p. 286–296.
- WANG M. & MANNING D. C. (2014). Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association of Computational Linguistics – Volume 2, Issue 1*, p. 55–66.
- YAROWSKY D. & NGAI G. (2001). Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, p. 1–8, Stroudsburg, PA, USA : Association for Computational Linguistics.
- YAROWSKY D., NGAI G. & WICENTOWSKI R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, p. 1–8 : Association for Computational Linguistics.

Méthode faiblement supervisée pour l'extraction d'opinion ciblée dans un domaine spécifique

Romaric Besançon

CEA, LIST, Laboratoire Vision et Ingénierie des Contenus
CEA Saclay Nano-INNOV, PC no 173, 91191 Gif-sur-Yvette CEDEX
romaric.besancon@cea.fr

Résumé. La détection d'opinion ciblée a pour but d'attribuer une opinion à une caractéristique particulière d'un produit donné. La plupart des méthodes existantes envisagent pour cela une approche non supervisée, sans hypothèse sur les caractéristiques visées. Or, les utilisateurs ont souvent une idée *a priori* des caractéristiques sur lesquelles ils veulent découvrir l'opinion des gens. Nous proposons dans cet article une méthode pour une extraction d'opinion ciblée qui exploite cette information minimale sur les caractéristiques d'intérêt. Ce modèle s'appuie sur une segmentation automatique des textes, un enrichissement des données disponibles par similarité sémantique et une annotation de l'opinion par classification supervisée. Nous montrons l'intérêt de l'approche sur un cas d'étude dans le domaine des jeux vidéo.

Abstract.

A Weakly Supervised Approach for Aspect-Based Opinion Mining in a Specific Domain.

The goal of aspect-based opinion mining is to associate an opinion with fine-grain aspects of a given product. Most approaches designed in this purpose use unsupervised techniques, whereas the information of the desired targeted aspects can often be given by the end-users. We propose in this paper a new approach for targeted opinion detection that uses this minimal information, enriched using several semantic similarity measures, along with topical segmentation and supervised classification. We prove the interest of the approach on an evaluation corpus in the specific domain of video games.

Mots-clés : Analyse d'opinion, classification supervisée, similarité sémantique.

Keywords: Opinion analysis, classification, semantic similarity.

1 Introduction

L'analyse automatique de l'opinion est une tâche importante pour beaucoup d'entreprises qui souhaitent connaître l'accueil public de leur marque ou de leur produit et a, de ce fait, donné lieu à de nombreuses recherches ces dernières années (Pang & Lee, 2007; Liu, 2012). Dans toute sa généralité, la problématique de la détection d'opinion se compose de plusieurs tâches, qu'il peut être utile ou non de mettre en œuvre selon les applications visées : la détection de la présence ou non d'une opinion ; la classification de la polarité de l'opinion (positif, négatif, neutre) ; la classification de l'intensité de l'opinion ; l'identification de l'objet de l'opinion (ce sur quoi porte l'opinion) ; l'identification de la source de l'opinion (qui exprime l'opinion). Toutes ces tâches peuvent se pratiquer à différents niveaux : au niveau global du texte, au niveau très précis d'une caractéristique particulière ou à des niveaux intermédiaires comme la phrase ou le paragraphe. À ces différents aspects s'ajoute le problème du domaine thématique des textes étudiés. En effet, la détection de l'opinion s'appuie sur le sens et la polarité des mots, qui peuvent changer selon les domaines considérés (Marchand *et al.*, 2014).

La problématique à laquelle nous nous intéressons dans cet article est celle de la classification de l'opinion associée à certaines caractéristiques précises de l'objet d'étude, dans un domaine spécifique. Par exemple, dans le cas des hôtels, les différentes caractéristiques seraient le prix, la taille des chambres, la propreté etc. Dans le domaine de notre cas d'étude sur les jeux vidéo, ces caractéristiques sont le graphisme, le *gameplay*, la profondeur de jeu etc.

Dans les recherches précédentes sur le sujet, ce problème est le plus souvent abordé avec des approches non supervisées visant à extraire des couples d'objets et de caractéristiques de ces objets, puis à chercher à associer une polarité d'opinion à ces caractéristiques en utilisant des mesures d'association (Popescu & Etzioni, 2005). Ces associations entre objets et caractéristiques et entre caractéristiques et opinion s'appuient souvent sur des analyses complexes prenant en particu-

lier en compte les dépendances syntaxiques au niveau des phrases (Qiu *et al.*, 2011; Cataldi *et al.*, 2013). Par ailleurs, ces méthodes s'appuyant sur des analyses locales, des solutions doivent également être mises en œuvre pour pallier les problèmes de résolution de coréférences pour les cibles de l'opinion (Ding & Liu, 2010). Une autre famille d'approches pour répondre à ce problème s'appuie sur des modèles probabilistes génératifs de type LDA (*Latent Dirichlet Allocation*), pour essayer de modéliser ces différentes dépendances, entre l'objet et ses caractéristiques d'une part et avec entre ces caractéristiques et les expressions d'opinion d'autre part (Moghaddam & Ester, 2013; Xueke *et al.*, 2013; Titov & McDonald, 2008). Les caractéristiques extraites de cette façon sont des parties de l'objet, des propriétés ou des concepts liés et peuvent être relativement nombreuses : par exemple, (Hu & Liu, 2004) compte en moyenne près de 70 caractéristiques pour différentes classes d'objets (appareils photos, téléphones mobiles, lecteurs MP3 etc.)

Le problème que nous abordons ici est un peu différent : nous cherchons à nous appuyer sur une connaissance *a priori* des caractéristiques qui nous intéressent et sur lesquelles on veut connaître les opinions des utilisateurs. En effet, dans la pratique, les entreprises, qui sont spécialistes de leur domaine et de leurs produits, ont une assez bonne idée de ce sur quoi elles veulent avoir une opinion. D'une part, cette connaissance est une contrainte parce qu'elle cible les caractéristiques à chercher, à l'opposé des approches non supervisées ; d'autre part, même si cette connaissance est très limitée (elle peut se limiter, en fait, aux seuls noms de caractéristiques visées), il est possible de l'utiliser pour guider directement la détection ciblée de l'opinion.

Pour répondre à ce problème, une approche possible est d'appliquer les techniques complètement non supervisées citées plus haut pour extraire des caractéristiques sur lesquelles des opinions sont exprimées, puis d'essayer de relier ces caractéristiques extraites à celles visées. Ce type de démarche peut être qualifiée d'*ascendante* : on part d'informations locales dans le texte pour les relier aux caractéristiques visées. C'est, par exemple, l'approche envisagée dans la tâche *Aspect-based Sentiment Analysis* de SemEval 2014 (Pontiki *et al.*, 2014). Nous proposons ici une approche différente, de type *descendante* : on utilise directement la connaissance des caractéristiques visées pour identifier dans les documents les segments de textes relatifs à ces différentes caractéristiques et leur attribuer une opinion, en utilisant des techniques de classification d'opinion au niveau du segment de texte. Ce type d'approche descendante nous semble en effet plus robuste, en ce qu'elle s'appuie sur des segments de texte plus importants, à la fois pour la qualification de l'opinion et pour la reconnaissance du lien avec les caractéristiques visées, ce qui permet d'exploiter plus d'indices contextuels et éventuellement d'être moins dépendant du domaine. Cette méthode est néanmoins destinée à l'analyse de textes plutôt longs et structurés, *e.g.* des articles critiques, et n'est pas adaptée pour des textes très courts (*e.g.* des *tweets*).

Nous présentons notre approche plus en détails dans la section suivante, et en montrons une évaluation sur un corpus de critiques en anglais, dans le domaine des jeux vidéo, dans la section 3.

2 Approche proposée pour la détection d'opinion ciblée

2.1 Vue générale de l'approche

La problématique à laquelle nous voulons répondre est de n'utiliser qu'une information minimale (les noms des différentes caractéristiques) pour développer un système de reconnaissance automatique d'une opinion ciblée sur ces caractéristiques. L'approche que nous proposons est la suivante :

- un découpage automatique des documents en différents segments de texte ;
- l'association de ces segments à une des caractéristiques visées ;
- une annotation automatique de l'opinion effectuée indépendamment sur ces différents segments ;
- une annotation finale de l'opinion sur chaque caractéristique combinant les annotations sur chaque segment.

L'idée de cette approche est de mettre en place un système de reconnaissance de l'opinion sur les différentes caractéristiques visées de la façon la moins supervisée possible, c'est-à-dire en ne considérant que les données du problème telles qu'on l'a posé ci-dessus, sans autre connaissance structurée ou données d'apprentissage supplémentaires. En effet, pour la première étape, on pourrait envisager de faire une segmentation manuelle d'un ensemble de documents selon ces différentes caractéristiques pour entraîner, de façon automatique, un segmenteur spécialisé. De même, l'association d'une caractéristique à chaque segment de texte peut être facilitée par une ontologie du domaine associant un vocabulaire spécifique à chacune des caractéristiques. Or, même si on trouve de nombreuses ressources lexico-ontologiques pour les langues bien dotées, elles sont souvent généralistes et leur spécialisation sur un domaine (par extraction et enrichissement d'une partie des informations) demande souvent des interventions manuelles. Or, ces opérations manuelles de construction ou d'adaptation de ressources sont coûteuses et doivent être refaites pour chaque nouveau domaine ou chaque nouvelle caractéristique considérée. Dans l'optique d'avoir un système adaptable facilement à de nouveaux domaines (*i.e.* de façon

complètement automatique), nous choisissons de mettre en place, pour cette étude, une approche moins supervisée.

2.2 Reconnaissance automatique d'une opinion au niveau du texte

Dans l'approche que nous proposons, la détection de l'opinion se fait au niveau d'un segment de texte. Nous nous plaçons ici dans un paradigme de reconnaissance automatique de l'opinion par classification automatique, sans utilisation de connaissances extérieures, comme des lexiques d'opinion, toujours dans la logique que des ressources lexicales spécialisées ne sont pas forcément disponibles.

2.2.1 Prétraitement

Concernant le prétraitement des documents, nous avons volontairement choisi de ne garder qu'un prétraitement minimal. En effet, pour la reconnaissance automatique de l'opinion, il est important de garder les informations de flexion : l'opinion suggérée par « I love » (présent) n'est pas forcément la même que celle suggérée par « I loved » (au passé : si l'objet a été aimé, ce n'est plus forcément le cas). On se limite donc ici à faire une segmentation du texte en mots, en gardant tous les mots dans leur forme de surface originale (en opérant néanmoins une normalisation par la mise en minuscules des mots). Nous avons également utilisé un filtre par anti-dictionnaire (*stoplist*) pour supprimer les mots grammaticaux non porteurs d'opinion. Nous avons pour cela repris les *stopwords* fournis par (Blitzer *et al.*, 2007) avec le corpus *Multi-Domain Sentiment Dataset*.

Nous n'avons pas effectué de traitement linguistique spécifique pour prendre en compte la négation mais nous utilisons également des n-grams de mots pour la représentation des documents (de sorte que des expressions composées comme « *not bad* » pourront être prises en compte). Dans la pratique, nous considérerons des n-grams de taille 3.

2.2.2 Classification automatique de l'opinion

Nous optons dans cette étude pour une approche de détection de l'opinion par classification supervisée. Plus précisément, nous avons fait des tests en utilisant des classifieurs standard de type SVM (*Support Vector Machine*) (Vapnik, 1995) et AdaBoost (*Adaptive Boosting*).

Les SVM ont souvent été utilisés avec succès pour la tâche de la classification d'opinion (Pang *et al.*, 2002; Rushdi-Saleh *et al.*, 2011), parfois en s'appuyant sur des fonctions noyaux complexes ((Wu *et al.*, 2009) utilisent par exemple des *tree kernels*). Dans cette expérimentation, nous utilisons pour notre part des noyaux linéaires standards.

Le *boosting* (Schapire, 1999) est une technique d'apprentissage s'appuyant sur l'idée qu'on peut construire un classifieur efficace par une combinaison pertinente de classifieurs faibles. Dans le cas de l'analyse de texte, ces classifieurs faibles sont les mots ou des n-grams de mots. Cette méthode de classification statistique a déjà montré de bons résultats sur la tâche de détection d'opinion, par exemple lors de la campagne d'évaluation DEFT'07 (Torres-Moreno *et al.*, 2007), et présente par ailleurs l'avantage d'avoir un modèle relativement explicite : on identifie les mots ou n-grams les plus discriminants pour la classification, ce qui peut permettre d'acquérir automatiquement des ressources lexicales d'opinion dans le domaine considéré¹.

Les implémentations utilisées de ces classifieurs statistiques sont SVMlight (Joachims, 2002) et BoosTexter (Schapire & Singer, 2000).

2.3 Reconnaissance automatique d'une opinion ciblée

2.3.1 Segmentation automatique des documents

Pour la segmentation automatique des documents, nous utilisons la méthode *LCseg* de segmentation thématique automatique à base de chaînes lexicales (*Lexical Chain Segmenter* (Galley *et al.*, 2003)). Cette méthode de segmentation s'appuie sur le repérage de chaînes lexicales (suites de termes répétés, éloignés d'une distance inférieure à une valeur

1. Dans le cas d'un SVM linéaire, on peut également retrouver les traits les plus influents en utilisant une régularisation L1, mais l'approche est moins immédiate.

donnée) et l'utilisation d'une mesure de cohésion lexicale, définie sur des frontières données (changements de phrase) et dépendant des chaînes lexicales qui traversent ces frontières. La segmentation consiste à trouver les changements de phrases qui minimisent cette mesure de cohésion lexicale. Une des forces de cette méthode est son indépendance par rapport au domaine, ce qui en fait une bonne base générique de segmentation. De plus, cette technique de segmentation a montré de meilleures performances que des techniques de découvertes automatique de thèmes par LDA (*Latent Dirichlet Allocation*), par exemple pour la segmentation thématique d'emails (Joty *et al.*, 2010).

Pour évaluer l'intérêt d'une segmentation thématique des documents, nous comparons les résultats avec une segmentation effectuée sur la base des phrases (chaque phrase est considérée comme un segment).

2.3.2 Association d'une caractéristique à chacun des segments

Après avoir segmenté le texte selon des segments thématiquement cohérents, on cherche à associer une caractéristique à chaque segment. La seule information dont nous disposons est le nom des différentes caractéristiques. Comme cette seule information n'est certainement pas suffisante, la première étape consiste donc à l'enrichir en considérant d'autres mots associés.

La détermination de mots associés s'appuie sur une forme de similarité sémantique entre les mots. Dans cette étude, nous utilisons, pour établir cette similarité sémantique, des informations distributionnelles construites automatiquement. Plus précisément, nous avons testé plusieurs méthodes d'expansion :

- par les mots les plus fréquemment présents avec les mots désignant les caractéristiques (mots *co-occurents*) ;
- par des *voisins distributionnels* des caractéristiques, calculés selon un modèle de thésaurus distributionnel ;
- par des *voisins « représentationnels »* des caractéristiques, selon un modèle de représentations lexicales distribuées.

Pour la construction d'un thésaurus distributionnel, nous avons suivi la méthode de (Ferret, 2010) : dans un corpus donné, chaque mot est représenté par ses co-occurrences avec les autres mots ; une mesure de similarité entre les mots est alors définie par la similarité entre leurs profils de co-occurrence (mesurée par le cosinus entre les vecteurs représentant ces profils). Cette méthode est paramétrée par la taille de la fenêtre des mots co-occurents, *i.e.* nombre de mots pleins (noms, verbes ou adjectifs) considérés comme co-occurents de chaque côté du mot considéré (une fenêtre de taille 1 correspond donc à considérer dans le profil de co-occurrence un mot à gauche et un mot à droite). La taille de cette fenêtre change en général la nature de la similarité sémantique : une fenêtre de petite taille a tendance à favoriser les similarités de type *synonyme* (les mots partagent des contextes locaux proches, *e.g.* ils apparaissent comme objets des mêmes verbes), alors qu'une fenêtre de plus grande taille génère des regroupements plus thématiques.

Les représentations lexicales distribuées sont construites selon la méthode de (Mikolov *et al.*, 2013), en utilisant son outil disponible *word2vec*². L'idée de cette méthode est d'utiliser un réseau de neurones pour apprendre une représentation vectorielle des mots qui arrive à capter une similarité sémantique entre les mots, fondée sur leur contexte. Deux approches sont en fait proposées : l'une, dite en sac-de-mots continu (*continuous bag-of-words* ou CBOW), cherche à maximiser la probabilité d'un mot en fonction de son contexte, alors que l'autre (*skip-gram*) cherche, au contraire, à prédire le contexte sachant le mot. En plus du modèle utilisé, cette approche utilise également d'autres paramètres, dont la taille de la fenêtre de contexte considérée et la taille du vecteur de représentation des mots.

La qualité des données ainsi construites (co-occurrences simples ou thésaurus distributionnels) dépend de la taille de la collection : dans notre cas, nous avons utilisé les collections spécialisées sur les jeux vidéo dont nous disposons, pour un corpus total d'environ 500 000 mots³. Pour les représentations lexicales distribuées, on comparera les résultats à ceux obtenus avec des voisins utilisant des représentations pré-calculées, fournies avec l'outil, sur un corpus généraliste très important (corpus *Google News*, de 3 milliards de mots).

Pour ces différentes méthodes, un score $w_{\text{assoc}}(c, t)$ est attribué à l'association du nom de la caractéristique c considérée avec chaque terme t utilisé pour son expansion : dans le cas des co-occurents, ce score est lié à la fréquence de co-occurrence, dans le cas des voisins sémantiques, ce score est lié à la distance entre les profils de co-occurrence ou les représentations lexicales distribuées.

Un poids d'association $w(s, c)$ d'une caractéristique c à un segment $s = \{t_1, \dots, t_n\}$ est alors défini par la somme, pour

2. <https://code.google.com/p/word2vec/>

3. Nous utilisons donc, pour partie, notre corpus de test dans les données sur lesquelles la construction des réseaux de co-occurrences est faite, ce qui présente en fait un biais dans l'évaluation. Néanmoins, même en intégrant ces données, le corpus reste relativement petit pour ce type de tâche, et nous avons donc décidé de les conserver. Idéalement, d'autres collections du domaine, non annotées, devraient être collectées pour construire ces ressources sémantiques.

chacun des termes du segment, du poids d'association de ce terme avec la caractéristique, selon le modèle d'expansion choisi. On attribue alors à chaque segment s la caractéristique $\text{car}(s)$ de plus haut poids :

$$\text{car}(s) = \underset{c}{\operatorname{argmax}} \sum_{t_i \in s} w_{\text{assoc}}(c, t_i)$$

2.3.3 Annotation automatique de l'opinion associée aux caractéristiques

Pour déterminer l'opinion de chacun des segments, nous utilisons simplement un modèle de reconnaissance de l'opinion générale, présentée en section 2.2, entraîné sur des textes entiers annotés en opinion et appliqué sur le segment de texte. Nous obtenons ainsi, pour un segment s , un score pour chaque décision d'opinion (positive/négative), notés $w_{op}(\text{pos}, s)$ et $w_{op}(\text{neg}, s)$.

Si on note $S(c)$ l'ensemble des segments liés à la caractéristique c (i.e. $S(c) = \{s | \text{car}(s) = c\}$), l'annotation globale de l'opinion sur c est alors obtenue en faisant simplement la moyenne des scores d'opinion sur chacun des segments de cet ensemble et en prenant la décision de l'opinion (positive/négative) sur la base de ces moyennes.

$$\text{opinion}(c) = \begin{cases} \text{positive} & \text{si } w_{op}(\text{pos}, c) > w_{op}(\text{neg}, c) \\ \text{negative} & \text{si } w_{op}(\text{neg}, c) > w_{op}(\text{pos}, c) \end{cases}$$

$$\text{avec } w_{op}(x, c) = \frac{1}{|S(c)|} \sum_{s \in S(c)} w_{op}(x, s)$$

3 Évaluation

Les expériences pour l'évaluation de la méthode proposée ont été réalisées sur plusieurs corpus de documents en anglais⁴, dans le domaine des jeux vidéo. Nous avons retenu les caractéristiques suivantes : le graphisme (*Graphics*), le son (*Sound*), le *Gameplay*, la profondeur du jeu (*Depth*), la présentation (*Presentation*). Ces caractéristiques nous ont été suggérées par les systèmes de notations adoptés par les critiques de différents sites de jeux vidéo, en particulier le site *videogamesdaily.com*, à partir duquel nous avons construit nos données de référence, mais d'autres sites utilisent des critères similaires : par exemple, sur les archives du site de critiques en français *jeuxvideo.com*, on trouve une séparation de la note selon des caractéristiques proches : *Graphismes*, *Jouabilité*, *Durée de vie*, *Bande son*.

3.1 Corpus d'évaluation

3.1.1 Corpus MDSD

Le corpus *Multi-Domain Sentiment Dataset* a été utilisé pour la première fois par (Blitzer *et al.*, 2007). C'est une collection de critiques en anglais, pour différents produits dans des domaines différents, récupérées sur le site Amazon. Ce corpus a été utilisé en particulier pour étudier la détection automatique de l'opinion sur des textes en fonction du domaine. Ces critiques sont des petits textes (en moyenne une centaine de mots) associés à une indication de la satisfaction de l'utilisateur sous la forme d'une note sur 5 (nombre d'étoiles). Pour bien distinguer les critiques positives et négatives, les documents ayant une note de 4 ou 5 sont considérés positifs, ceux ayant une note de 1 ou 2 négatifs (les documents ayant une note de 3 sont ignorés). Le nombre de documents visé pour chaque domaine est de 1000 documents positifs et autant de négatifs, mais certains domaines sont moins représentés et la collection contient au final 38548 documents.

Dans le cadre de cette étude, nous nous sommes focalisés sur le domaine des jeux vidéo. Nous avons donc extrait du corpus *Multi-Domain Sentiment Dataset* un sous-ensemble de critiques portant sur les jeux vidéo. Plus précisément, ces critiques ont été prises dans le domaine « *Computer & Video Games* » et ont été filtrées sur la base de mots clés pour supprimer les produits liés à l'informatique qui ne sont pas des jeux vidéo. On a ainsi retenu 1229 critiques, dont 848 positives et 381 négatives. Ce corpus est noté MDSD-JV dans les résultats des expériences.

4. Les techniques mises en œuvre ne s'appuyant pas sur des ressources explicites, elles sont indépendantes de la langue : il est néanmoins nécessaire, pour traiter une langue différente, de disposer au moins d'un corpus annoté en opinion dans cette langue, soit du domaine considéré, soit d'un domaine général.

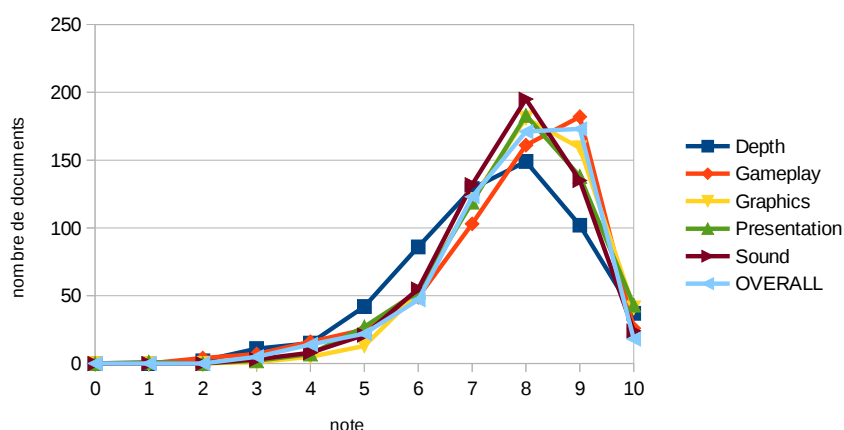


FIGURE 1 – Distribution des notes dans le corpus *videogamesdaily* selon les différentes caractéristiques (les notes sont arrondies à la valeur entière la plus proche).

3.1.2 Corpus *videogamesdaily*

Pour l'évaluation de la détection d'opinion ciblée, nous avons construit un corpus à partir des critiques des archives du site *videogamesdaily.com*, dans lequel les articles de critiques de jeux vidéo se concluent par une grille d'appréciation selon différents points : le graphisme, le son, le *gameplay*, la profondeur, la présentation, avec, pour chacun de ces points, une note décimale comprise entre 1 et 10.

Nous avons collecté et reformaté ces archives pour former une collection annotée pour chacune des caractéristiques. Nous avons ainsi obtenu 573 articles. Les documents de ce corpus étant des articles critiques d'un site spécialisé, ils sont plus détaillés que des critiques Amazon et ont en particulier une taille bien plus importante que les documents du corpus MDSD : 460 mots en moyenne contre 200 mots pour le corpus MDSD-JV.

En ce qui concerne la distribution des notes, présentée à la figure 1, selon les différentes caractéristiques considérées, on constate une sur-représentation des notes positives (qui est également présente dans le corpus initial des critiques Amazon) et également un décalage de l'échelle de notation : les notes de 1 à 3 ne sont quasiment pas utilisées, les courbes sont plutôt centrées sur des notes moyennes autour de 8. De façon similaire au corpus extrait d'Amazon, on choisit comme exemples de documents négatifs des documents ayant des notes de 6 ou moins et des documents ayant des notes de 9 à 10 comme exemples de documents positifs. Les notes entre 6.1 et 8.9 sont ignorées. Le nombre de documents positifs et négatifs ainsi obtenus, pour chacune des caractéristiques, est présenté dans le tableau 1. Notons que, si l'on a entre 199 et 257 documents pour chacune des caractéristiques, ces documents couvrent au total 462 documents différents du corpus (ce qui signifie qu'il n'y a pas énormément de recoupement entre les différents ensembles : en pratique, seuls 49 documents sont communs à tous les ensembles). Par ailleurs, dans la plupart des cas, les critiques ont souvent des notes sur les différentes caractéristiques qui vont dans le même sens, mais on compte néanmoins environ 12% des documents qui ont au moins une caractéristique négative et une caractéristique positive.

	nb docs positifs	nb docs négatifs
<i>Gameplay</i>	156	101
<i>Depth</i>	100	155
<i>Presentation</i>	135	90
<i>Graphics</i>	142	70
<i>Sound</i>	113	86

TABLE 1 – Nombre de documents positifs et négatifs selon chacune des caractéristiques considérées.

work minutes what#was rent repetitive
 bought#game cannot didn't boring nothing
 believe waste was don't 360 not#much#fun
 money gave poor completely not horrible
 unless look needs why returned worst i#got
 terrible

note music also never liked light d rocks variety
 different knows live 7 amazing must own
 if#you#don't love inside shooter ps2 enough
 close side who best fun better#than well
 original

FIGURE 2 – Les mots discriminants pour l’opinion, appris par BoosTexter sur le corpus MDSD-JV (les mots en rouge sont négatifs, les mots en vert positifs, la taille des mots est proportionnelle à l’importance accordée à ces mots par le modèle)

3.2 Résultats de classification d’opinion générale

Les premières expérimentations évaluent la qualité de la détection de l’opinion au niveau du texte sur ces collections. Pour chaque collection, les tests ont été réalisés par validation croisée (*10-fold*). Les résultats présentés dans le tableau 2 sont des moyennes obtenues sur les 10 essais. Les mesures d’évaluation utilisées pour la détection de l’opinion sont l’exactitude des résultats (*accuracy*), définie par le rapport des annotations correctes sur le nombre total d’annotations, ainsi que les précision/rappel/f-score, calculées comme des moyennes des scores pour les opinions positives et négatives.

	BoosTexter		SVM	
	videogamesdaily	MDSD-JV	videogamesdaily	MDSD-JV
Exactitude	77,98%	90,16%	77,55%	88,72%
F-score	77,49%	90,03%	77,12%	88,59%
Précision	78,96%	90,56%	78,49%	89,42%
Rappel	78,62%	90,17%	79,43%	88,71%

TABLE 2 – Évaluation de la reconnaissance automatique de l’opinion par apprentissage automatique

On voit dans ces résultats que les deux types de classifieurs donnent des résultats relativement bons et comparables, même si les résultats obtenus par BoosTexter sont un peu meilleurs que ceux obtenus avec les SVM. Les résultats sont meilleurs pour le corpus MDSD-Jeux vidéo, ce qui n’est pas étonnant étant donné que c’est le corpus de taille la plus importante (le corpus *videogamesdaily* est relativement limité pour des expérimentations par apprentissage automatique).

Un avantage de l’approche par Boosting est de construire un modèle qui s’appuie explicitement sur les mots comme classifieurs faibles et permet donc de visualiser le vocabulaire utilisé par le classifieur pour différencier les textes positifs et négatifs. La figure 2 présente par exemple les 30 mots les plus discriminants pour le corpus MDSD-Jeux vidéo. On voit dans ces listes des mots qui sont porteurs d’opinion (*horrible*, *terrible*, *repetitive*, *not#much#fun* pour le lexique négatif, *amazing*, *best*, *fun* pour le lexique positif), ce qui montre qu’on arrive à capter de façon automatique des expressions d’opinion. On trouve également d’autres mots spécifiques au domaine (e.g. *360* a une connotation négative alors que *PS2* a une connotation positive dans les critiques considérées).

Afin de tester la capacité des classifieurs à construire un modèle suffisamment général, nous entraînons des modèles sur un corpus et les testons sur l’autre. Ces expérimentations ont seulement été faites avec le modèle BoosTexter, qui donnait plutôt de meilleurs résultats sur les corpus pris isolément. Par ailleurs, (Mansour *et al.*, 2013) a montré qu’un classifieur entraîné sur un grand nombre de données de différents domaines permettait parfois d’avoir de meilleurs résultats qu’un classifieur entraîné sur le domaine visé. Dans cette optique, nous reportons également les résultats en entraînant le classifieur sur la totalité des données du corpus MDSD et en faisant le test sur nos deux corpus spécialisés. Les résultats de ces tests croisés, en exactitude, sont présentés dans le tableau 3 (les résultats des corpus sur eux-mêmes sont ceux obtenus en validation croisée, alors que pour les résultats croisés, la totalité des documents du corpus source a été utilisée pour l’apprentissage et la totalité du corpus cible pour le test).

On remarque à nouveau dans ces résultats, l’impact de la taille du corpus, le modèle appris sur le corpus *videogamesdaily* menant à un modèle plus spécialisé qui ne s’adapte pas bien sur d’autres collections. Par contre, pour le corpus MDSD-JV, le modèle appris permet d’obtenir des résultats intéressants sur un autre corpus de nature différente.

Le dernier résultat montre également qu’on arrive à avoir une meilleure généralisation en apprenant sur une grande quan-

		testé sur	
		videogamesdaily	MDSD-JV
appris sur	videogamesdaily	77,98%	52,62%
	MDSD-JV	66,67%	90,16%
	MDSD	77,98%	83,73%

TABLE 3 – Résultats d’exactitude en entraînant sur un corpus et en testant sur un autre corpus.

waste not#buy return bad **disappointing**
 money broke useless **boring** not#worth **great** solid wonderful easy favorite my#only love
worst junk way#too horrible ok returned highly job pleased not#too delicious perfectly
 beware nothing poorly **poor** disappointed perfect excellent better#than great#product
 not#good not#recommend unfortunately amazing awesome i#like price best loves fun
 disappointment awful don't#buy not your#money grill not#disappointed nice comfortable you#not enjoy
terrible

FIGURE 3 – Les mots discriminants pour l’opinion, appris par BoosTexter sur le corpus MDSD complet (les mots en rouge sont négatifs, les mots en vert positifs, la taille des mots est proportionnelle à l’importance accordée à ces mots par le modèle).

tité de données, même en incluant des documents de domaines différents. Il faut noter que d’autres études, comme (Garcia-Fernandez *et al.*, 2014), montrent, à l’opposé, qu’à taille de corpus égal, un corpus homogène reste préférable. (Mansour *et al.*, 2013) rapporte des résultats moyens de l’ordre de 90 % sur chacun des domaines du corpus MDSD pris séparément. On a donc ici des résultats un peu moins bons, mais qui restent relativement comparables. Si l’on regarde les mots discriminants appris sur ce corpus (présentés à la figure 3), on constate qu’on apprend en effet des mots porteurs d’opinion plus génériques, sans mots spécifiques du domaine.

3.3 Résultats de classification d’opinion ciblée

Nous évaluons la méthode proposée pour la classification d’opinion ciblée sur le corpus *videogamesdaily*, qui contient des annotations de référence selon les différentes caractéristiques. Comme les annotations de référence ne sont conservées que pour ces scores inférieurs à 6 ou supérieurs à 9, les documents n’ont pas d’annotations pour toutes les caractéristiques. Réciproquement, il est possible que la procédure d’annotation ne produise pas d’annotation sur un document pour une caractéristique donnée. Les scores d’exactitude présentés sont, de ce fait, calculés en considérant seulement l’intersection des annotations de référence avec les annotations automatiques. De ce fait, on rapporte également, pour les différents résultats, le nombre d’annotations qui sont prises en compte pour l’évaluation.

La segmentation thématique des documents par LCseg produit entre 1 et 22 segments par document avec une moyenne d’environ 9 segments. Les segments font en moyenne 86 mots (après suppressions des mots outils). Une segmentation par phrase mène à un découpage moyen des documents en 54 segments, d’une taille moyenne de 15 mots.

Pour les résultats, nous avons plusieurs paramètres à étudier :

- l’influence de la segmentation, en phrases ou en segments thématiques ;
- l’influence du modèle d’association des mots pour l’expansion : mots co-occurents, voisins sémantiques par thésaurus distributionnels ou par représentation lexicale distribuée ;
- la taille de la fenêtre pour la prise en compte du contexte ;

Le tableau 4 présente les résultats obtenus, en termes d’exactitude, sur les différentes caractéristiques considérées, pour un découpage en phrases ou en segments thématiques, en utilisant les voisins sémantiques calculés sur la base d’un thésaurus distributionnel construit en utilisant plusieurs taille de fenêtre de voisinage (en prenant 1,2,5 ou 10 mots de contextes).

Ces résultats mettent en évidence l’intérêt de l’utilisation d’une segmentation thématique par rapport à une simple segmentation en phrases pour obtenir une plus grande précision de détection d’opinion. L’amélioration observée peut également être partiellement due au fait que les segments thématiques sont de taille plus importante et supportent donc plus d’élé-

	segments thématiques				phrases			
	w1	w2	w5	w10	w1	w2	w5	w10
Depth	55,7%	65,2%	58,4%	62,2%	45,7%	52,0%	49,6%	46,8%
Gameplay	69,9%	70,6%	68,8%	71,0%	61,7%	62,1%	60,9%	62,1%
Graphics	68,4%	67,9%	74,7%	69,7%	70,1%	67,3%	69,2%	67,8%
Presentation	66,3%	68,2%	63,3%	61,3%	68,3%	61,5%	63,9%	63,8%
Sound	69,2%	71,7%	71,6%	70,5%	60,8%	60,9%	66,7%	63,8%
average	65,9%	68,7%	67,4%	66,9%	61,3%	60,8%	62,1%	60,9%
annotations	706	680	697	694	1059	1052	1043	1055

TABLE 4 – Résultats (exactitude) pour la détection d’une opinion ciblée, en comparant un découpage en phrases ou en segments thématiques, pour des voisins distributionnels calculés avec différentes tailles de contexte.

	Co-occurents				Voisins
	w1	w2	w5	w10	w2
Depth	57,7%	47,9%	50,6%	59,2%	65,2%
Gameplay	56,4%	57,4%	62,7%	56,9%	70,6%
Graphics	53,3%	60,0%	64,9%	58,6%	67,9%
Presentation	65,5%	69,0%	67,7%	71,5%	68,2%
Sound	62,4%	66,3%	73,5%	70,1%	71,7%
average	59,0%	60,1%	63,9%	63,3%	68,7%
annotations	732	742	668	612	680

TABLE 5 – Résultats (exactitude) pour la détection d’une opinion ciblée : comparaison d’expansion sémantique entre co-occurents et voisins distributionnels.

ments pour l’application du modèle de détection de l’opinion. En effet, l’attribution d’une opinion à un texte fonctionne en général d’autant mieux que le texte contient plus de mots susceptibles de porter une information de polarité d’opinion : si l’on attribue à chaque document une opinion calculée comme la moyenne des opinions détectées pour chacun des segments qui le composent, on obtient un score d’exactitude sur l’annotation globale des documents qui est de 66,7 % pour le découpage en segments thématiques et de 59,8 % pour le découpage en phrases. En comparant avec le score obtenu à partir de la totalité des documents (77,98 %), on remarque effectivement une baisse de performance due, vraisemblablement, à l’application de la détection d’opinion sur des documents de plus petite taille.

D’autre part, cette différence entre segments et phrases s’explique aussi par le fait que plus d’annotations sont prises en compte pour l’évaluation avec la segmentation en phrases : en effet, dans ce cas, la probabilité qu’au moins une des phrases soit associée à une caractéristique est plus importante, et on a donc globalement moins de caractéristiques pour lesquelles aucune annotation n’est fournie.

Par ailleurs, on remarque un comportement globalement meilleur pour une taille de fenêtre de 2, ce qui semble un bon compromis entre une mesure plus proche de la notion de synonymie et une fenêtre suffisamment large pour prendre en compte la petite taille du corpus sur lequel les co-occurrences sont apprises.

Le tableau 5 présente les résultats en comparant les meilleurs résultats obtenus avec les voisins sémantiques (sur un contexte de deux mots) à ceux obtenus avec les mots co-occurents pour l’attribution des caractéristiques à chaque segment, pour différentes tailles de fenêtres de co-occurrence (les résultats sont obtenus avec la segmentation thématique). Ce tableau montre que l’utilisation de voisins sémantiques donne globalement de meilleurs résultats que l’utilisation des simples co-occurents. On constate néanmoins que pour certaines des caractéristiques, les résultats peuvent être meilleurs avec les co-occurents. En particulier, pour la caractéristique « *Presentation* », les résultats sont de façon générale moins bons avec les voisins sémantiques : cette tendance était également vérifiée dans le tableau précédent, avec le découpage en phrases, ce qui montre que, pour cette caractéristique, les voisins sémantiques ne fournissent pas une bonne expansion (ce qui peut s’expliquer par la nature particulièrement polysémique du mot « *presentation* »). De façon générale, néanmoins, les valeurs moyennes sont plus élevées avec les voisins sémantiques.

De façon complémentaire, on peut voir des différences entre les attributions des caractéristiques aux différents segments selon qu’on utilise des co-occurents ou des voisins sémantiques. Le tableau 6 montre le nombre de segments attribués à chacune des catégories pour les deux modèles. On remarque plusieurs choses en observant ces distributions : d’une

	segments thématiques				phrases			
	voisins		co-occurents		voisins		co-occurents	
	w2	w10	w2	w10	w2	w10	w2	w10
Depth	183	243	1805	1244	1460	2130	7165	2183
Gameplay	3107	3110	171	310	12471	12878	1027	711
Graphics	1213	1242	190	69	6892	7249	983	232
Presentation	183	171	1356	985	1508	1446	5638	1758
Sound	405	378	1407	303	2273	2186	5826	610
No category	216	163	378	2396	6232	4947	10197	25342

TABLE 6 – Nombre de segments/phrases attribués à chaque caractéristique, selon le modèle d’expansion sémantique choisi.

part, l’utilisation de voisins sémantiques ou de co-occurents change l’attribution des caractéristiques aux segments, en favorisant les caractéristiques *Gameplay* et *Graphics* avec les voisins sémantiques alors qu’elles sont minoritaires en considérant les co-occurents ; d’autre part, l’augmentation de la taille de la fenêtre de co-occurrence tend à accentuer cette tendance. Enfin, ces deux observations sont conservées que ce soient avec les segments thématiques ou les phrases. Ces tendances peuvent être dues aux différences de nature des mots, *Gameplay* et *Graphics* étant fortement liés au domaine thématique alors que les autres mots ont aussi des sens plus généraux.

Le tableau 7 contient les résultats obtenus en utilisant les représentations lexicales distribuées pour le calcul des voisins sémantiques, d’une part en calculant ces représentations sur notre corpus spécialisé dans le domaine des jeux vidéo, pour différents voisinages considérés (des contextes de tailles 1,2,3,5 et 10 ont été testés)⁵ et, d’autre part, en utilisant les représentations lexicales pré-calculées sur le corpus général *Google News*, beaucoup plus important.

	corpus Jeux Vidéo								corpus Google	
	segments thématiques				phrases				segments	phrases
	w1	w2	w5	w10	w1	w2	w5	w10		
Depth	60,6%	61,6%	61,1%	54,4%	57,2%	53,0%	50,3%	49,8%	60,8%	53,8%
Gameplay	66,7%	67,7%	71,2%	70,0%	62,0%	61,7%	63,8%	65,8%	68,1%	68,8%
Graphics	77,8%	72,5%	70,0%	70,1%	71,9%	69,7%	67,3%	67,3%	82,3%	74,6%
Presentation	67,3%	66,1%	54,8%	58,0%	63,9%	63,9%	63,3%	67,7%	66,2%	63,6%
Sound	69,4%	73,8%	74,5%	72,9%	65,0%	66,3%	65,7%	63,8%	65,1%	65,4%
average	68,3%	68,4%	66,3%	65,1%	64,0%	62,9%	62,1%	62,9%	68,5%	65,2%
annotations	551	661	651	698	801	929	978	1037	632	776

TABLE 7 – Résultats (exactitude) pour la détection d’une opinion ciblée, avec utilisation des représentations lexicales distribuées.

Ce tableau montre qu’on peut obtenir, avec ce modèle, des résultats moyens comparables avec l’approche par thésaurus distributionnel et confirme l’intérêt de la segmentation thématique. On remarque néanmoins que les écarts entre les scores pour les différentes caractéristiques sont plus importants qu’avec les voisins distributionnels (écart-type de 4,95 comparé à un écart-type de 2,37 pour les voisins distributionnels). On remarque aussi qu’avec ce modèle, l’augmentation de la taille de la fenêtre tend à diminuer la qualité des résultats. Pour les résultats avec les représentations générales, on remarque à nouveau avec des écarts de scores plus importants et en particulier un score bien supérieur pour la caractéristique *Graphics*, qui a peut-être une distribution dans ce corpus particulièrement liée au domaine des jeux vidéos. Des tests supplémentaires devraient également être menés pour prendre ces représentations générales comme point de départ de l’entraînement sur le corpus spécialisé. Ces résultats montrent néanmoins, de façon intéressante, que pour un domaine spécifique, on n’a pas forcément besoin d’une collection de documents très importante pour construire des modèles intéressants. A contrario, cela montre aussi qu’avec un corpus très important pour entraîner un modèle, on peut construire un modèle qui aura des performances acceptables, même en domaine spécifique.

En ce qui concerne l’établissement des voisins sémantiques, le tableau 8 présente, pour illustration, les 10 voisins sémant-

5. Ces résultats sont obtenus avec le modèle *CBOW* et une taille de représentation de 500, en utilisant le *negative sampling*. Nous avons fait des tests avec les différents paramètres (modèle, taille de la représentation), dont nous ne présentons pas tous les résultats faute de place. Les résultats présentés offrent un bon compromis entre le nombre d’annotations et la qualité des annotations.

tiques les plus proches pour chacune des caractéristiques considérées, obtenus d’une part avec un thésaurus distributionnel (fenêtre de taille 2) et d’autre part avec les représentations lexicales distribuées (modèle CBOW, fenêtre de taille 2 et taille de représentation de 500).

	voisins distributionnels	voisins word2vec
Depth	<i>flavour, layer, ton, challenge, amount, something, scope, strategy, variety, backdrop</i>	<i>breadth, wealth, layer, originality, longevity, astounding, periphery, fidelity, limitless, several</i>
Gameplay	<i>experience, play, control, mechanic, combat, storey, storyline, action, predecessor, meat</i>	<i>subplot, game, shoehorned, sorely, mechanics, multiplayer, monotony, superficial, unoriginal, fundamental</i>
Graphics	<i>background, animation, texture, detail, game, environment, model, version, great, engine</i>	<i>horrendous, replayability, outdated, psx, pros, portrait, summary, integrated, youre, looker</i>
Presentation	<i>visual, package, highlight, prowess, mixture, flair, standpoint, tt, quality, acting</i>	<i>accompagnement, aural, markedly, fidelity, prowess, inspiring, blemish, assured, ditty, nxc</i>
Sound	<i>effect, voice, music, soundtrack, guncon, noise, lighting, chaos, robot, sublime</i>	<i>ambient, particle, narration, music, stereo, crisp, 5.1, echo, sharp, echoes</i>

TABLE 8 – Voisins sémantiques des noms des caractéristiques considérées.

3.3.1 Limitations

Nous avons indiqué, dans la présentation du corpus, que les critiques contenant à la fois des jugements positifs et négatifs sur les différentes caractéristiques sont relativement peu nombreuses (environ 12% des documents). L’approche naïve consistant à donner à chacune des caractéristiques l’orientation d’opinion globale du document donne donc, statistiquement, de bons résultats. En pratique, on obtient avec cette approche *baseline* un score moyen de 73,54%, qui est en effet supérieur aux scores obtenus ici. Néanmoins, si l’on restreint l’évaluation aux seuls documents contenant effectivement les deux orientations, on obtient un score moyen de 57.47% avec cette approche naïve alors qu’on a un score de 67,47% pour la meilleure approche proposée dans cet article. Ce type d’approche est donc à privilégier si les opinions exprimées sont effectivement variées selon les caractéristiques dans les données considérées.

4 Conclusion

Nous présentons dans cet article une approche pour la détection d’opinion ciblée qui, à l’inverse des méthodes existantes non supervisées, utilise une faible supervision sur la nature des caractéristiques visées. Cette information minimale est étendue de façon automatique à l’aide de similarités sémantiques et est exploitée dans un cadre de classification automatique de l’opinion. L’approche est évaluée dans le domaine des jeux vidéo et donne des résultats intéressants, avec presque 70 % de bonne détection de l’opinion sur chacune des caractéristiques visées. Ces résultats sont d’autant plus prometteurs qu’ils s’appuient sur une méthode très peu supervisée, demandant très peu de connaissances externes. Concernant la similarité sémantique utilisée, on montre que des approches à base de thésaurus distributionnels ou de représentations lexicales distribuées donnent des résultats moyens comparables, même si les approches par représentations distribuées semblent donner des résultats moins réguliers sur les différentes caractéristiques.

Parmi les développements futurs de cette étude, on peut envisager d’enrichir les traits pris en compte pour la détection de l’opinion, en intégrant par exemple des lexiques d’opinion, même généralistes, dans les modèles. Il faudrait également comparer l’approche proposée aux approches non supervisées, par exemple utilisées dans la campagne SemEval, sur les mêmes données. Enfin, une autre extension possible de ce travail serait de déterminer si le fait de disposer d’une évaluation de l’opinion pour chacune des caractéristiques permet d’améliorer la détermination de l’opinion globale.

Références

- BLITZER J., DREDZE M. & PEREIRA F. (2007). Biographies, bollywood, boom-boxes and blenders : Domain adaptation for sentiment classification. In *Association for Computational Linguistics*, Prague, Czech Republic.
- CATALDI M., BALLATORE A., TIDDI I. & AUFAURE M.-A. (2013). Good location, terrible food : detecting feature sentiment in user-generated reviews. *Social Network Analysis and Mining*, 3(4), 1149–1163.

- DING X. & LIU B. (2010). Resolving object and attribute coreference in opinion mining. In *Proceedings of COLING '10*, p. 268–276, Stroudsburg, PA, USA : Association for Computational Linguistics.
- FERRET O. (2010). Testing semantic similarity measures for extracting synonyms from a corpus. In *Proceedings of LREC'10*.
- GALLEY M., MCKEOWN K., FOSLER-LUSSIER E. & JING H. (2003). Discourse segmentation of multi-party conversation. In *Proceedings of ACL '03*, p. 562–569, Stroudsburg, PA, USA : Association for Computational Linguistics.
- GARCIA-FERNANDEZ A., FERRET O. & DINARELLI M. (2014). Evaluation of different strategies for domain adaptation in opinion mining.
- HU M. & LIU B. (2004). Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artificial Intelligence, AAAI'04*, p. 755–760 : AAAI Press.
- JOACHIMS T. (2002). *Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms*. Kluwer/Springer.
- JOTY S., CARENINI G., MURRAY G. & NG R. T. (2010). Exploiting conversation structure in unsupervised topic segmentation for emails. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10*, p. 388–398, Stroudsburg, PA, USA : Association for Computational Linguistics.
- LIU B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- MANSOUR R., REFAEI N., GAMON M., ABDUL-HAMID A. & SAMI K. (2013). Revisiting the old kitchen sink : Do we need sentiment domain adaptation ? In *Proceedings of RANLP 2013*, p. 420–427.
- MARCHAND M., BESANÇON R., MESNARD O. & VILNAT A. (2014). Influence des marqueurs multi-polaires dépendant du domaine pour la fouille d'opinion au niveau du texte. In *Proceedings of TALN 2014*, p. 1–12.
- MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.
- MOGHADDAM S. & ESTER M. (2013). The flida model for aspect-based opinion mining : Addressing the cold start problem. In *Proceedings of the 22Nd International Conference on World Wide Web, WWW '13*, p. 909–918.
- PANG B. & LEE L. (2007). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, **2**.
- PANG B., LEE L. & VAITHYANATHAN S. (2002). Thumbs up ? : Sentiment classification using machine learning techniques. In *Proceedings of EMNLP '02*, p. 79–86, Stroudsburg, PA, USA : Association for Computational Linguistics.
- PONTIKI M., GALANIS D., PAVLOPOULOS I., PAPAGEORGIOU H., ANDROUTSOPOULOS I. & MANANDHAR S. (2014). SemEval 2014 Task 4 : Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) at (COLING 2014)*, p. 27–35, Dublin, Ireland.
- POPESCU A.-M. & ETZIONI O. (2005). Extracting product features and opinions from reviews. In *Proceedings of HLT '05*, p. 339–346, Stroudsburg, PA, USA : Association for Computational Linguistics.
- QIU G., LIU B., BU J. & CHEN C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, **37**, 9–27.
- RUSHDI-SALEH M., MARTÍN-VALDIVIA M. T., RÁEZ A. M. & LÓPEZ L. A. U. (2011). Experiments with SVM to classify opinions in different domains. *Expert Syst. Appl.*, **38**(12), 14799–14804.
- SCHAPIRE R. E. (1999). A brief introduction to boosting. In *Proceedings of IJCAI'99*, p. 1401–1406, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.
- SCHAPIRE R. E. & SINGER Y. (2000). Boostexter : A boosting-based system for text categorization. *Mach. Learn.*, **39**(2-3), 135–168.
- TITOV I. & McDONALD R. (2008). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web, WWW '08*, p. 111–120, New York, NY, USA : ACM.
- TORRES-MORENO J. M., EL-BÈZE M., BÉCHET F. & CAMELIN N. (2007). Comment faire pour que l'opinion forgée à la sortie des urnes soit la bonne ? application au défi defit 2007. *Actes du troisième DÉfi Fouille de Textes*, p. 129.
- VAPNIK V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc.
- WU Y., ZHANG Q., HUANG X. & WU L. (2009). Phrase dependency parsing for opinion mining. In *Proceedings of EMNLP '09*, p. 1533–1541, Stroudsburg, PA, USA : Association for Computational Linguistics.
- XUEKE X., XUEQI C., SONGBO T., YUE L. & HUAWEI S. (2013). Aspect-level opinion mining of online customer reviews. *Communications, China*, **10**(3), 25–41.

Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité

Egle Eensoo¹ Mathieu Valette¹

(1) ERTIM, INALCO, 2 rue de Lille, 75343 PARIS cedex 07
egle.eensoo@inalco.fr, mathieu.valette@inalco.fr

Résumé. Cet article entend dresser, dans un premier temps, un panorama critique des relations entre TAL et linguistique. Puis, il esquisse une discussion sur l'apport possible d'une sémantique de corpus dans un contexte applicatif en s'appuyant sur plusieurs expériences en fouille de textes subjectifs (analyse de sentiments et fouille d'opinions). Ces expériences se démarquent des approches traditionnelles fondées sur la recherche de marqueurs axiologiques explicites par l'utilisation de critères relevant des représentations des acteurs (composante dialogique) et des structures argumentatives et narratives des textes (composante dialectique). Nous souhaitons de cette façon mettre en lumière le bénéfice d'un dialogue méthodologique entre une théorie (la sémantique textuelle), des méthodes de linguistique de corpus orientées vers l'analyse du sens (la textométrie) et les usages actuels du TAL en termes d'algorithmiques (apprentissage automatique) mais aussi de méthodologie d'évaluation des résultats.

Abstract.

A method of corpus semantics applied to opinion mining and sentiment analysis: the impact of dialogical and dialectical features on the expression of subjectivity.

This paper first aims to provide a critical overview of the relationship between NLP and linguistics, and then to sketch out a discussion on the possible contribution of corpus semantics in an application-based context based on several subjective text mining studies (sentiment analysis and opinion mining). These studies break away from traditional approaches founded on the detection of axiological markers. Instead, they use explicit criteria related to the representation of actors (dialogical component) and argumentative or narrative structures (dialectical component). We hope to highlight the benefit of a methodological dialogue between theory (text semantics), meaning-oriented methods of corpus linguistics (i.e. textometrics) and NLP current practices in terms of algorithmic (machine learning) and assessment methodology.

Mots-clés : Textométrie, Sémantique de corpus, Fouille d'opinion, Analyse des sentiments

Keywords: Textometry, corpus semantics, opinion mining, sentiment analysis

1 Introduction

Avec l'essor dans le TAL des méthodes par apprentissage automatique et la relative désaffection pour les méthodes symboliques à base de règles linguistiques formelles dans le monde académique¹, les linguistes sont aujourd'hui contraints de repenser leur rôle dans un contexte où dominent les méthodes mathématiques. Si l'annotation requise pour la constitution des données d'apprentissage nécessite un savoir-faire et une connaissance experte parfois adossée à des présupposés théoriques, les spécialistes de la fouille de textes, par exemple, montrent peu d'intérêt pour les théories linguistiques, vraisemblablement à raison, tant se creuse le fossé entre les préoccupations minutieuses mais *ad hoc* de certains linguistes et celles des talistes, guidées par un principe de réalité : la masse de données textuelles accessibles.

Cet article propose un panorama critique des relations entre TAL et linguistique et esquisse, au moyen d'exemples commentés issus d'applications en fouille de textes (analyse de sentiments et fouille d'opinions), une discussion sur l'apport possible d'une réflexion linguistique dans ce contexte applicatif. Nous souhaitons en particulier mettre en

¹ (Tanguy, 2012) relate plusieurs études (Church, 2011, Hall *et al.*, 2008) où a été observé que la proportion d'articles de l'*Association for Computational Linguistics* intégrant une section statistique a progressé de 30 à 90 % du début des années 90 à la fin des années 2000.

lumière le bénéfice potentiel d'un dialogue méthodologique entre des méthodes de linguistique de corpus orientées vers l'analyse du sens (la textométrie), l'exploitation de concepts de la sémantique textuelle (Rastier, 2001, 2011) et les usages actuels du TAL en termes d'algorithmiques mais aussi de pratiques évaluatives.

L'article est construit en quatre parties. Le paragraphe 2 offre une lecture optimiste des relations qu'entretiennent le TAL et la linguistique et de leur réunion possible autour de l'objet *texte*. Le paragraphe 3 procède à l'examen en miroir des outils et méthodes nécessaires à l'établissement d'une sémantique instrumentée, en mettant notamment en vis-à-vis la textométrie et le TAL. Le paragraphe 4 présente les concepts linguistiques et la méthodologie adoptés par les auteurs pour une tâche de fouille de textes subjectifs. Enfin, le paragraphe 5 présente, à des fins illustratives, trois expérimentations adossées à la méthodologie décrite dans le paragraphe précédent.

2 Le statut contemporain du texte dans le TAL

Longtemps unis par des objets formels similaires sinon communs (la proposition, la phrase) et un même positionnement référentialiste, la linguistique et le TAL ont vu leurs rapports se distendre depuis une quinzaine d'années. Les modèles théoriques de la linguistique formelle se sont en effet avérés peu adaptés à la prise en compte de l'évolution rapide de la demande applicative à laquelle le TAL a été confronté. Jusqu'au début des années 2000, la plupart des applications concernaient la thématique, le lexique ou la terminologie. Les tâches correspondantes nécessitant une automatisation (résolution d'anaphore, désambiguïsation lexicale, identification des parties du discours) relevaient d'une sémantique de la phrase. Rapidement, les technologies de l'information et la *redocumentarisation* du monde (Pédauque, 2007) ont actualisé le statut d'objet scientifique du *texte* – statut que la linguistique ne lui accorde encore que marginalement et au sein de certains courants seulement (analyse du discours, linguistique textuelle). Des tâches telles que la classification de textes et la fouille de textes ont émergé, rendant nécessaire une approche macroscopique et à grande échelle des productions langagières plus en phase avec l'unité *texte* qu'avec l'unité *phrase*. Les modèles formels de la sémantique de la phrase, avec leurs analyses « profondes » mais très locales apparaissent moins efficaces pour l'analyse de grands corpus, notamment en termes de rappel, bien qu'elles proposent encore des solutions pertinentes pour l'extraction d'information *précise*, liée aux applications telles que l'interface homme/machine (système de question-réponse, ou réponse à des questions formulées en langue dite naturelle) (Zweigenbaum *et al.*, 2008). Par ailleurs, les méthodes symboliques sont plébiscitées dans l'industrie où beaucoup d'applications nécessitent un haut taux de précision sans que le rappel soit déterminant. Enfin, la tendance actuelle est à l'hybridation dans le monde académique comme dans l'industrie. Le couplage de données produites à partir de méthodes à base de règles et de technique apprentistes permet d'améliorer les performances de systèmes de manière significative (Villena-Román *et al.*, 2011).

L'essor, dans le courant des années 2000, des applications en fouille de textes subjectifs (fouille d'opinion, analyse des sentiments, détection des émotions, etc.) implique également une évolution des tâches : alors que le TAL privilégiait les unités *référentielles* et souvent lexicales (entités nommées, concepts, termes, thèmes), il est aujourd'hui confronté à des *valeurs*. Certes, les méthodes d'extraction et de classification n'ont guère évolué : dans beaucoup d'applications, les adjectifs sont aux textes subjectifs ce que les substantifs sont aux concepts (Strapparava & Valitutti, 2004) et on a tendance à appliquer aux premières les méthodes qui ont fait leur preuve sur les secondes. Dépasser le « lexicalisme » du TAL est un des enjeux de la linguistique car l'inventaire des objets de la linguistique susceptibles d'être appréhendés par le TAL est, en effet, loin d'être clos. Il est par exemple probable que les contraintes de genres, de discours, que la structure actancielle des textes, que le schéma de la communication, soient utiles à l'interprétation des émotions, sentiments ou des opinions².

En somme, tout se passe comme si les questions qui se posent au TAL évoluaient d'une problématique logico-formelle dominée par le primat référentiel et le choix historique de la phrase (et son avatar : l'énoncé) comme unité d'analyse, vers une problématique herméneutique et interprétative dont l'objet est la réception et l'interprétation des textes considérés comme des unités de sens complexes déterminées par un projet de communication. La proposition a notamment été formulée par (Rastier, 2001) et oppose, *in fine*, deux paradigmes, la linguistique des langues et la linguistique des textes. Ce moment de flottement paradigmatique est l'occasion d'esquisser des méthodes fondées non pas sur les présupposés théoriques du paradigme logico-grammatical mais sur un paradigme herméneutique et interprétatif peu exploré encore en TAL.

Dans le paragraphe suivant, nous procédons donc à l'examen contrastif des présupposés épistémologiques et méthodologiques du TAL (et plus particulièrement de la fouille de textes) d'une part, et d'une sémantique de corpus dédiées à l'interprétation des textes d'autre part.

² On pourra utilement consulter (Micheli *et al.*, 2013) à ce propos.

3 La textométrie, ou l'interprétation assistée par ordinateur³

L'analyse statistique des données textuelles (ADT), ou *textométrie*, est un ensemble particulier de pratiques relevant du champ général de la linguistique de corpus. Elle comprend des traitements statistiques (analyse factorielle des correspondances, spécificités fondées sur le modèle hypergéométrique, etc.) et des outils de visualisation des corpus (nuages de mots, histogrammes, etc.) et documentaires (concordanciers) destinés à l'aide à l'interprétation des textes⁴.

3.1 La textométrie et le TAL

La linguistique de corpus et la textométrie ne relèvent pas du TAL. En dépit de quelques traits communs (les corpus numériques, les algorithmes mathématiques informatisés) et d'affinités intercommunautaires ponctuelles, elles se distinguent à tous les égards. On dresse ci-après l'inventaire de ces différences.

Du point de vue épistémologique – Le TAL, fondamentalement, vise l'automatisation des processus, l'élimination de la part de l'humain dans les traitements, tandis que la textométrie repose sur une itération entre l'analyse des sorties logicielles et la consultation des textes ou de fragments ; en cela, il s'agit davantage d'une linguistique assistée par ordinateur. Par ailleurs, le TAL est utilitariste et a pour finalité des applications informatiques, ce qui implique une recherche de performance et d'optimisation ; la textométrie a des objectifs épistémiques : accroître les connaissances et participer à l'interprétation d'un corpus. Enfin, à la différence du TAL où la mise en place d'un protocole d'évaluation est indispensable, l'évaluation et la reproductibilité ne sont pas problématisées par la textométrie. Les études textométriques sont validées par homologation, c'est-à-dire par l'assentiment d'une communauté qui, dans le meilleur des cas, est distante (par exemple, communauté de la critique littéraire pour l'analyse textométrique de textes littéraires), mais parfois n'est peut-être qu'un avatar du « jugement d'acceptabilité » contre lequel s'est pourtant dressée la linguistique de corpus.

Du point de vue des applications – Comme on l'a vu, les applications ne sont que marginalement un enjeu en textométrie, même si certains travaux sont susceptibles d'applications (constitution de ressources par exemple) alors que la demande socio-économique détermine dans une large mesure les tâches auxquelles le TAL s'attelle. Cette demande implique le renouvellement des problématiques de recherche : il y a 15 ans, l'extraction d'informations lexicales ou syntagmatiques destinées à alimenter des bases de connaissances (mémoires de traduction, terminologies de métier, système de question-réponse, etc.) structurait le champ. Puis, avec l'essor des réseaux sociaux sur le web, des applications en fouille d'opinion, analyse des sentiments, analyse du buzz, etc. se sont développées. La traduction automatique, historiquement liée au TAL symbolique, connaît également un regain d'intérêt motivé par l'efficacité des méthodes statistiques.

Du point de vue des documents – Les pratiques de la textométrie et celles du TAL opposent les notions de *sources* et de *ressources* : les documents analysés en textométrie sont variés et souvent caractérisés avec précision. À la fin des années 1990, les œuvres littéraires dominaient (romans, poésie, théâtre) mais on étudiait aussi des enquêtes ouvertes, des textes politiques, syndicaux, etc. Au milieu des années 2000, les nouveaux genres de l'Internet font leur apparition (mails, puis forums de discussion, tweets). On retrouve en partie ces types documentaires en TAL (très rarement les textes littéraires), mais les textes à vocation technique ou encyclopédique (telles que Wikipédia) apparaissent privilégiés. Surtout, davantage que des sources précises (i.e. des œuvres ou des éditeurs électroniques, des sites web, etc.), ce sont des ressources générales qui sont désignées : Internet, Web, Google, Google Books, Facebook. Les corpus, en TAL sont avant tout des réservoirs d'objets linguistiques infratextuels (termes, structures prédicatives, etc.). L'établissement philologique du corpus en TAL est souvent réduit à quelques valeurs quantitatives (nombre d'occurrences de mots, nombre de textes) quand les textomètres présentent leur corpus de manière plus qualitative (description des auteurs, des genres textuels, etc.).

Du point de vue des méthodes – c'est probablement au niveau des méthodes d'analyse que la différence entre la textométrie et le TAL est la plus visible. À la différence des talistes, les textomètres ne sont pas des informaticiens mais,

³ Les observations faites dans ce paragraphe s'appuient en partie sur une l'analyse contrastive des actes de deux conférences communautaires francophones emblématiques : les Conférences en Traitement Automatique de la Langue Naturelle (TALN) et les Journées internationales d'Analyses statistiques des Données Textuelles (JADT). L'étude, menée sur un échantillon de 8 volumes d'actes de TALN et 8 volumes d'actes de JADT (de 1999 à 2014) donnera lieu à une publication ultérieure.

⁴ Les actes des Journées Internationales d'Analyses statistiques des données textuelles (JADT) donneront au lecteur un aperçu des pratiques textométriques : <http://lexicometrica.univ-paris3.fr/jadt/>.

en règle générale, des utilisateurs finaux de logiciels dotés d'interface graphique (Hyperbase, Lexico 3, TXM, Iramuteq, TextObserver, etc.) lesquels s'adossent de plus en plus souvent aux outils que les talistes développent ou utilisent pour leurs propres tâches : bibliothèques de traitements linguistiques (par exemple NLTK, Stanford NLP), langages de programmation (par exemple, Perl ou Python pour la manipulation de textes ; R ou Matlab pour le calcul), etc. En bref, les textomètres sont dépendants d'outils qu'ils conçoivent et parfois qu'ils implémentent. On a là une différence de culture remarquable : l'essentiel des efforts en matière de création d'outils en textométrie se porte actuellement sur l'ergonomie logicielle et la visualisation des données. Les méthodes mathématiques employées, qui satisfont le plus grand nombre, évoluent peu depuis 30 ans mais les heuristiques et les savoir-faire analytiques sont déterminants. Souvent les talistes s'étonnent du peu de variété des méthodes statistiques des textomètres, et leur opposent d'impressionnantes bibliothèques de traitements. C'est qu'ils ne prennent pas la mesure des tâches herméneutiques qui font la spécificité de l'ADT. L'interprétation des résultats d'analyse, en TAL, est non cruciale et souvent occultée au profit de deux types de commentaires : commentaires sur les performances de la méthode utilisée d'une part, commentaires sur les résultats d'évaluation qui suivent des méthodes normalisées. L'évaluation des performances du système repose en effet sur les mesures de congruence entre le résultat de la classification et le corpus de test annoté (taux d'exactitude, précision, rappel, f-score, etc.). Or, comme l'observe (Yvon 2006, 41), d'autres évaluations sont possibles (analyse sémantique des valeurs discriminantes sélectionnées par l'algorithme, adéquation avec une théorie linguistique, plausibilité cognitive, etc.) mais les alternatives sont rares et peu valorisées en termes académiques. Mieux encore, les données langagières proprement dites sont jugées encombrantes et, pour des raisons éditoriales sans doute, mais peut-être par manque d'outils intellectuels pour les appréhender, on ne les montre guère (Hall *et al.*, 2008).

Du point de vue de ce qu'est un corpus – L'inclination apprentiste qu'a suivi le TAL ces dernières années a profondément accentué les différences liées à l'utilisation et la fonction du corpus. Les méthodes d'apprentissage automatique dit « supervisé », lesquelles sont encore privilégiées en TAL, consistent à créer un modèle reproduisant la configuration optimale des données du corpus, quelles qu'elles soient. Si, dans une tâche de classification de textes par exemple, un corpus est composé de deux classes, l'entraînement du modèle consistera à sélectionner les critères (par exemple les mots-formes) qui caractérisent de façon appropriée les textes d'une classe par rapport à l'autre, quand bien même ces critères ne seraient nullement interprétables d'un point de vue linguistique.

Le corpus en textométrie est conçu comme un mode de contextualisation à échelle multiple des phénomènes observables, de la cooccurrence, « forme minimale du contexte » (Mayaffre, 2008) au corpus intégral qui objective l'intertexte (Rastier, 1998) et qui, à mesure qu'il s'élargit, tend vers le contexte extralinguistique qu'il simule. Ainsi, les sous-corpus construits ont toujours une fonction différentielle. On distinguera principalement le *corpus de référence* « constituant le contexte global de l'analyse, ayant le statut de référentiel représentatif, et par rapport auquel se calcule la valeur de paramètres (pondérations...) et se construit l'interprétation des résultats » et le *corpus de travail*, « ensemble des textes pour lesquels on veut obtenir une caractérisation » (Rastier & Pincemin, 1999, 84). Cette approche du corpus, est indubitablement plus sophistiquée en termes d'analyse et d'interprétation des données, mais elle écarte toute instance de validation. En bref, les concepts de corpus en TAL et en textométrie sont fondamentalement distincts.

La textométrie, comme la linguistique de corpus, demeure un ensemble de techniques et d'heuristiques qui nécessite un guidage théorique pour assurer sa pleine mesure. L'analyse de discours (Charaudeau, 1992) en est un exemple. Nous porterons notre attention, pour notre part, sur la sémantique textuelle de (Rastier, 2001, 2011) dont les rapports avec le TAL sont déjà anciens et ont donné lieu, en particulier à la fin des années 90 et au début des années 2000, à plusieurs instantiations⁵.

3.2 La textométrie et la sémantique des textes

Les affinités de la textométrie et de la sémantique des textes ont été identifiées précocement (Rastier, éd., 1995). La plupart ont été explicitées par (Mayaffre, 2008) et de façon systématique par (Pincemin, 2010) à laquelle nous renvoyons le lecteur. En voici les principaux éléments susceptibles d'alimenter notre discussion.

Le texte ne fait l'objet d'aucune préconception réductrice – Les signes qui composent le texte ne sont pas hiérarchisés (les substantifs ne sont pas préférés *a priori* aux mots grammaticaux ou aux signes de ponctuation) et ne sont pas substituables par des constructions artefactuelles, en particulier si elles sont de haut niveau, tels les concepts, les hyperonymes, les synonymes. Or, l'annotation de corpus au moyen de ressources variées est non seulement très courante en TAL mais ne fait guère l'objet de réflexion critique. Pourtant, même le traitement basique qui consiste à lemmatiser un corpus, parce qu'elle en factorise les formes, fait l'objet de débats circonspects en textométrie (Brunet, 2000) comme en sémantique des textes (Bourion, 2001).

⁵ Par exemple (Beust, 1998), (Thlivitis, 1998), (Perlerin, 2004), (Rossignol, 2005).

Le retour au texte est la condition de l'interprétation – L'analyse en textométrie comme en sémantique textuelle repose sur une itération entre l'analyse des sorties logicielles et la consultation des textes ; en d'autres termes, la connaissance des textes est une condition nécessaire à leur analyse, elle est notamment génératrice d'hypothèses interprétatives.

Le contexte global construit par le corpus de référence joue un rôle déterminant dans l'interprétation des faits sémantiques – C'est le principe souvent répété de détermination du global sur le local, qui relativise, sans les exclure, les unités linguistiques inférieures comme la phrase. Du côté de la textométrie, la constitution d'un corpus de référence et d'un corpus de travail en est une mise en œuvre.

« Dans la langue, il n'y a que des différences » – Héritée de la tradition saussurienne (Saussure, 2002), le différentialisme fonde la sémantique textuelle et est sans doute un aspect remarquable de la textométrie dans le contexte général de la linguistique de corpus. Le succès jamais démenti des mesures de spécificités (tests χ^2 ou d'écart réduit, modèle hypergéométrique) destinées à contraster une partie d'un corpus avec une autre de manière à en faire émerger les singularités, en atteste.

3.3 Synthèse

Nous prenons acte (i) de l'hypothétique évolution du TAL vers une problématique herméneutique intéressée par l'interprétation des textes et non plus seulement par l'extraction des données discrètes qu'ils recèlent ; (ii) de l'inadéquation des modèles linguistiques dominants, préoccupés par des phénomènes relevant de la langue et non du texte ; (iii) des hiatus épistémologiques et de la complémentarité méthodologiques observés entre le TAL et la textométrie ; (iv) des affinités entre celle-ci et la sémantique textuelle. Nous formulons le projet général de jeter un pont entre la sémantique textuelle et le TAL par le truchement de la textométrie, afin de mutualiser les avantages d'une association entre celles-ci et les standards du TAL, c'est-à-dire l'évaluation à partir de méthode par apprentissage supervisé. Nous illustrerons notre propos à partir d'une tâche de fouille de textes subjectifs.

4 Sémantique de corpus pour la fouille de textes subjectifs

4.1 Principales méthodes du champ applicatif

Nous distinguerons quatre types d'approche en fouille de textes subjectifs : (i) les approches apprentistes, qui ne sont pas spécifiques à la fouille d'opinion ou l'analyse des sentiments mais sont utilisées dans différentes tâches (recherche d'information, traduction automatique, étiquetage morphosyntaxique, classification thématique, etc.). Appliquées à la fouille d'opinion, elles ont tendance à privilégier l'accumulation massive de descripteurs et ne nécessitent pas une connaissance linguistique approfondie des textes (par exemple, Pang *et al.*, 2002 ; Lin & Hauptmann, 2006) ; (ii) les approches cognitivistes, qui font appel à des ressources lexicales supposant l'existence de catégories cognitives préétablies et indépendantes des langues, par exemple des ressources dérivées de Wordnet (Ghorbel & Jacot, 2011 ; Lavalley *et al.*, 2010 ; Kim *et al.*, 2010 ; Liu *et al.*, 2003) ou des ressources basées sur la théorie *Appraisal* (Whitelaw *et al.*, 2005, Bloom & Argamon, 2010) ; (iii) les approches opportunistes, qui exploitent des phénomènes linguistiques de surface détectables automatiquement comme des patrons morphosyntaxiques (Turney, 2002 ; Yi *et al.*, 2003), des parties du discours (Hatzivassiloglou & Wiebe, 2000), etc. ; (iv) les approches linguistiques théoriques qui revendiquent un cadre linguistique à des fins heuristiques. (Vernier *et al.*, 2009a, 2009b), par exemple, s'inspirent des catégories évaluatives de (Charaudeau, 1992). Pour un état de l'art plus détaillé, on pourra lire (Eensoo & Valette, 2014b).

C'est dans un cadre méthodologique relevant de cette quatrième approche que nous situons la méta-étude présentée ici.

4.2 Concepts et méthodologie de sémantiques de corpus

Concepts – Nous formulons en effet l'hypothèse que les discours axiologiques se construisent par des interactions entre différentes composantes sémantiques qui ne relèvent pas du strict vocabulaire des valeurs. Nous proposons ci-dessous une synthèse basée sur trois expériences montrant, dans différentes tâches d'analyse des sentiments et de fouille d'opinion, par méthodes d'apprentissage, que les descripteurs classifiants les plus efficaces peuvent ne relever que de deux classes de valeurs sémantiques appelées *composantes* sémantiques par (Rastier 2001)⁶ : la composante dialogique et la composante dialectique.

⁶ (Rastier, 2001) inventorie quatre composantes sémantiques : dialogique, dialectique, thématique, tactique. La composante thématique est abordée par (Eensoo & Valette, 2012, 2014a, 2014b) mais nous n'en ferons pas état ici.

- La *composante dialogique* concerne la représentation des acteurs, le positionnement énonciatif et la distribution des rôles actanciels. Elle actualise essentiellement les pronoms personnels, les pronoms possessifs et certaines entités nommées.
- La *composante dialectique* est une catégorie sémantique dédiée à la représentation du temps et du déroulement aspectuel, des structures argumentatives et de certaines modalités. Le vocabulaire la caractérisant est plus varié. Il peut s'agir de marqueurs de structuration (adverbes tels que *enfin, donc, cependant*), des verbes modaux (*falloir, devoir*, etc.), et des indicateurs rhétoriques (emphases, points d'interrogation, mots interrogatifs, etc.).

Cette grille interprétative a permis à (Eensoo & Valette, 2012, 2014a, 2014b) de mettre en évidence que l'expression subjective pouvait être caractérisée avec un nombre restreint de marqueurs relevant des différentes composantes sémantiques sans nécessairement recourir à un vocabulaire subjectif. Ils élaborent le concept d'*agoniste* comme « une classe d'acteurs stéréotypés correspondant à une position ou à la défense d'une valeur (ou d'un ensemble de valeurs) » (Eensoo & Valette, 2014b, 116). L'*agoniste* est une construction textuelle (et non psychologique ou cognitive) reposant sur une combinaison d'éléments linguistiques relevant des composantes sémantiques.

Méthodologie générale – La méthodologie de sémantique de corpus adoptée repose sur deux étapes : l'analyse textométrique de corpus préalablement annotés d'une part, la validation par apprentissage supervisé des critères textométriques obtenus et qualifiés sémantiquement au moyen de la grille interprétative adoptée, d'autre part. L'analyse textométrique effectuée en amont de toute classification automatique permet d'identifier des critères de classification linguistiquement explicables et suffisamment robustes pour servir comme descripteurs aux méthodes d'apprentissage supervisé. L'hypothèse est que les critères de classification *interprétables* sont plus robustes que les descripteurs trouvés par des méthodes d'apprentissage, souvent non signifiants d'un point de vue textuel et incidents au corpus d'apprentissage. Ainsi, lors de l'étape de sélection de critères, le textomètre écarte les critères liés à l'échantillon du corpus et choisit les critères textuels cohérents avec les composantes sémantiques (ici, la composante dialogique et la composante dialectique) actualisées dans le corpus.

À des fins expérimentales, nous avons écarté de cette étude les critères textuels relevant de la composante *thématique* (*i.e.* les thèmes et les domaines actualisés) de façon à évaluer l'autonomie des positions agonistiques par rapports aux thématiques des trois expériences exposées ci-après : médicaux et sanitaire en 5.1, idéologiques en 5.2 et politiques et législatifs en 5.3.

4.3 Élaboration textométrique des critères sémantiques de catégorisation

Pour les expérimentations présentées dans le paragraphe 5, ont été utilisés plusieurs types de critères : (i) unités isolées : un choix de formes, lemmes ou catégories morphosyntaxiques, (ii) collocations (n-grammes) de taille variée (de 2 à 4 unités) et (iii) cooccurrences phrastiques multiniveaux (combinant les éléments de différents niveaux de description linguistique : formes, lemmes ou catégories morphosyntaxiques). Tous les critères ont été sélectionnés selon quatre principes : leur caractère spécifique à un sous-corpus, leur répartition uniforme dans le sous-corpus, leur fréquence et leur pertinence linguistique.

L'analyse du corpus et l'identification des critères ont été effectuées avec deux logiciels textométriques – Lexico 3 (Salem *et al.*, 2003) et TXM (Heiden *et al.*, 2010) – qui implémentent les algorithmes de spécificités (Lafon, 1980) et de cooccurrences (Lafon, 1981). Les deux premiers types de critères sont choisis selon la procédure suivante :

1. calcul des spécificités des items isolés (formes, lemmes et catégories morphosyntaxiques) et de leurs n-grammes (fonction « Segments Répétés » de Lexico 3) pour chaque sous-corpus ;
2. analyse des contextes d'apparition des items spécifiques (au moyen de concordances textuelles) afin de s'assurer de leur pertinence textuelle et de l'unicité de leur fonction (les critères ayant une seule fonction et signification ont été privilégiés) ;
3. vérification de la répartition uniforme des items dans le sous-corpus (fonctionnalité « Carte de Sections » du Lexico 3) ;

La sélection des cooccurrences est faite comme suit :

1. calcul des cooccurrences (fonction « Cooccurrences » de TXM) des items spécifiques fréquents et uniformément repartis sur la totalité du corpus ;

2. analyse des contextes d'apparition de ces cooccurrences ;
3. sélection des cooccurrences spécifiques à un sous-corpus ;

Dans les deux cas, les critères de classification pour chaque texte sont des fréquences absolues car d'une part, il a été démontré que les fréquences relatives sont moins performantes que les valeurs booléennes (Pang & Vaithyanathan, 2002), d'autre part, nous avons constaté que les fréquences absolues sont plus performantes que les fréquences relatives.

4.4 Classification par apprentissage supervisé à partir des critères sémantiques élaborés

La deuxième étape consiste à utiliser des algorithmes d'apprentissage supervisé pour classer les textes. En utilisant la plateforme WEKA (Hall *et al.* 2009)⁷, plusieurs algorithmes, de familles différentes, ont été testés : les arbres de décision (J48), *Naive Bayes*, *Naive bayes multinomial* et les Machines à Vecteurs de Support (SMO). L'objectif est d'observer les différences et similitudes au niveau des performances en changeant la nature et la quantité des critères. Dans le présent article, ne sont mentionnerons que les résultats des algorithmes les plus efficaces pour les tâches choisies. A l'exception de la troisième expérience (5.3) pour laquelle nous disposons d'un corpus de test (Grouin *et al.*, 2007), les évaluations sont opérées suivant la méthode de la validation croisée sur 10 sections.

5 Trois expériences de sémantique de corpus

5.1 Agonistes dysphoriques et euphoriques dans les forums de discussions médicales et sanitaires⁸

(Eensoo & Valette, 2012, 2014a) disposent d'un corpus de 300 ego-documents (témoignages, récits d'histoires vécues) postés par les internautes sur différents forums de discussion à dominante médico-sanitaire (aufeminin.com, doctissimo.fr, etc.) et catégorisé en deux classes : les textes dysphoriques et les textes euphoriques. La référence de la catégorisation est établie par l'agrégateur des documents Samestory⁹. Nous ne disposons pas de guide d'annotation mais en analysant un échantillon du corpus nous avons pu déduire la stratégie d'annotation. Un témoignage « dysphorique » est (i) une histoire qui fini mal, (ii) un témoignage exprimant des doutes, des interrogations, ou sollicitant de l'aide. Un témoignage « euphorique » est (i) une histoire triste qui finit bien, (ii) un témoignage modulant la gravité d'une situation en en soulignant les points positifs, (iii) un conseil. Pour les besoins de l'application d'analyse de sentiments, ils i identifient et inventorient 70 critères sémantiques à partir de l'analyse textométrique puis les caractérisent en fonction des composantes sémantiques.

Il en résulte la construction de deux agonistes. D'un point de vue dialogique, l'agoniste dysphorique apparaît égocentré (surreprésentation de la 1^{er} personne du singulier) et enclos sur son univers intime, il exprime un univers impressif et non factuel (« *Je ne sais pas*¹⁰ comment cela va évoluer »). Du point de vu dialectique, on constate une excentration de l'action : (« *On me dit* que les causes de cette maladie ne sont pas encore précises »). L'*agoniste euphorique* est élaboré sur un noyau sémique inverse. Du point de vue de la composante dialogique, c'est un acteur-énonciateur altruiste qui s'adresse à un tiers (surreprésentation de la 2^{er} personne du singulier) (« *Alors tu* vois il faut avoir espoir »). L'agoniste euphorique construit des univers alternatifs en faisant part de son expérience à des fins d'édification (« *Je tenais à faire part de mon expérience* ») et en intertextualisant son témoignage (« *Je te file une adresse : <http://www. ...>* »). Le caractère le plus remarquable des textes euphoriques réside au niveau de la composante dialectique. À la différence de l'agoniste dysphorique, l'agoniste euphorique élabore un texte séquencé, descriptif ou argumentatif (« *J'avais déjà quelques éruptions qui ont débuté après avoir pris la décision de déménager* », « *Par contre j'étais soignée à l'homéopathie* »).

Parmi les critères construits, 30 relèvent de la composante dialectique et 16 de la composante dialogique. L'évaluation de la capacité classificatrice des critères qualifiés a été réalisée au moyen d'une classification de textes effectuée en utilisant un algorithme d'apprentissage automatique de la famille des *Machines à vecteurs de support* – SMO (Platt, 1998).

⁷ <http://www.cs.waikato.ac.nz/ml/weka/>

⁸ Pour un exposé complet des résultats, nous invitons le lecteur à se reporter à l'étude originale correspondante.

⁹ <http://www.same-story.com/>

¹⁰ Désormais, tous les éléments en italique sont des exemples de critères de catégorisation.

Types de critères	Exactitude <i>validation croisée</i>
Mots simples (10 700 critères) <i>ligne de comparaison</i>	68 %
Critères dialogiques (16 critères)	64 %
Critères dialectiques (30 critères)	73 %
Critères dialectiques et dialogiques (46 critères)	77 %

TABLE 1 : Résultat de la classification : agonistes dysphoriques et euphoriques

Le tableau 1 donne à voir quelques résultats de la classification. La ligne de comparaison (*baseline*) est la classification sur formes simples (sans changement de casse ni de lemmatisation), qui permet d'obtenir un taux d'exactitude, c'est-à-dire le pourcentage de textes bien classés, de 68 %. Le cumul des critères dialectiques et dialogiques permet de s'élever de 9 points au dessus de la ligne de comparaison (77 %). Ce résultat est intéressant car ce sont ces composantes qui se démarquent le plus nettement des pratiques en fouille de textes, lesquelles, en général, privilégient des descripteurs thématiques ou thymiques.

5.2 Agonistes hostiles et non hostiles aux Roms dans un corpus de commentaires d'articles

(Eensoo & Valette, 2014b) étudient un corpus constitué de 644 commentaires d'articles de presse de 2013 ayant pour objet la communauté Rom en France. Les commentaires sont écrits par les lecteurs-internautes. Ils proviennent de quatre quotidiens : *Le Monde*, *Libération*, *Le Figaro* et *Le Parisien*. Ces commentaires ont été classés en deux supercatégories composées de 445 commentaires hostiles pour la première supercatégorie et 199 commentaires non hostiles pour la seconde. Les supercatégories ont elles-mêmes été divisées en cinq catégories plus fines que nous n'aborderons pas dans cet article¹¹. La catégorisation manuelle des documents a été effectuée par les auteurs de l'étude selon une lecture et interprétation macroscopique des textes excluant l'identification des éléments lexicaux discrets qui pourraient se confondre avec les critères de classification automatique. Par la suite, les auteurs inventorient 42 critères dialectiques et 11 critères dialogiques à partir de l'analyse textométrique effectuée sur ce corpus.

Types de critères	Exactitude <i>validation croisée</i>
Mots simples (6075 critères) <i>ligne de comparaison</i>	70 %
Critères dialogiques (11 critères)	69 %
Critères dialectiques (42 critères)	71 %
Critères dialectiques et dialogiques (53 critères)	72 %

TABLE 2 : Résultat de la classification : agonistes hostiles et non hostiles aux Roms

Les critères ont été évalués au moyen de l'algorithme Naïve Bayes Multinomial (Mccallum & Nigam 1998). Comme dans l'expérience précédente, la ligne de comparaison demeure la classification sur formes simples, qui permet d'obtenir un taux d'exactitude de 70 %. L'élément marquant ici est que le résultat de la classification, avec seulement 11 critères dialogiques, égale pratiquement la ligne de comparaison, quand les critères dialectiques la dépassent, tout comme la combinaison des deux catégories. On a la démonstration que des marqueurs énonciatifs en très petit nombre – essentiellement quelques pronoms *je*, *nous*, *vous*, des adjectifs possessifs, et le tag *NAM* (pour noms propres, obtenue au moyen d'un étiquetage Treetagger) – peuvent suffire à obtenir une classification, certes perfectible, mais comparable à celle effectuée sur les formes simples. C'est l'indice selon nous que la seule posture énonciative observable dans la sélection des marques de personnes, suffit, dans certaines tâches de classification, à identifier la position idéologique des énonciateurs.

¹¹ Voir note 8.

5.3 Agonistes favorable (pour) et défavorable (contre) dans des débats parlementaires

La troisième expérience, réalisée pour les besoins de cet article, s'appuie sur le corpus de débats parlementaires mis à la disposition de la campagne DEFT 2007 (Grouin *et al.* 2007). Ce corpus regroupe 28 832 interventions de députés à l'Assemblée Nationale extraites des débats. Le corpus d'apprentissage totalise 17 299 interventions. Il est divisé en deux classes : 6 899 interventions favorables à la loi en cours d'examen ; 10 400 interventions défavorables à la loi en cours d'examen. Le corpus de test, quant à lui, est composé de 11 533 interventions au total, 4 961 intervention favorables et 6 572 défavorables. La référence est établie en considérant le vote effectif (favorable ou défavorable à la loi en examen) des intervenants. L'application de la méthodologie exposée ici a permis d'identifier 26 critères dialogiques et 64 critères dialectiques.

Les critères dialogiques favorables à la loi en examen sont le pronom personnel et possessifs de 1^e personne (*je, mon, ma, mes*), la mention de partis politiques et des verbes porteurs de la fonction expressive au sens jakobsonien (« C'est pourquoi *nous saluons* le travail accompli par la commission », « ce dont *je me réjouis* », etc.). Les critères dialectiques sont notamment des verbes modaux (*il faut, il doit*) et quelques éléments de structuration argumentative, par exemple, énumératifs (« *Enfin*, ce projet répond aux attentes de nos concitoyens », « Il serait *également* souhaitable que, etc. »). Les critères dialogiques défavorables sont les pronoms personnels et possessifs de 2^e personne du pluriel (*vous, votre, vos*) ou encore l'impersonnel *on* qui dénotent des prises de paroles interlocutoires plus marquées que pour les parlementaires favorables. Parmi les marqueurs dialectiques défavorables, on relève une forte saillance des marques de négation (*non, ne, pas, jamais, rien*). Des stratégies rhétoriques plus agressives sont également observables via des adverbes d'interrogation (*comment, quand*), le point d'interrogation et divers marqueurs argumentatifs d'opposition (*Or, Mais, pourtant*).

Types de critères	Exactitude sur corpus de test DEFT 2007	Exactitude validation croisée
Mots simples (5 832 critères) <i>ligne de comparaison</i>	70 %	76 %
Critères dialogiques (26 critères)	61 %	65 %
Critères dialectiques (64 critères)	65 %	68 %
Critères dialectiques et dialogiques (90 critères)	66 %	70 %

TABLE 3 : Résultat de la classification : débats parlementaires « pour » et « contre » la loi en examen

La classification effectuée comme précédemment avec SMO donne un taux d'exactitude de l'ensemble de nos critères textométriques de 70 % en validation croisée sur 10 sections et 66% sur le corpus de test fourni par DEFT 2007 ; les résultats sont en deçà des lignes de comparaison mais présentent la particularité de n'avoir été obtenus qu'à partir des formes et des n-grammes de formes, sans lemmatisation, sans étiquetage morphosyntaxique, sans normalisation de la casse, ni recherche de patrons de cooccurrences. On notera que l'écart entre le taux d'exactitude obtenu avec nos critères et celui de la ligne de comparaison est moins important sur le corpus de test que par validation croisée, ce qui témoigne d'une certaine robustesse.

6 Conclusion

En couplant la sémantique textuelle, la textométrie et des méthodes d'apprentissage automatique, nous avons tenté de valider la pertinence applicative du concept de *composante sémantique* dans le cadre de différentes tâches de classification de textes subjectifs. La méthodologie présentée permet d'identifier un très petit nombre de critères textuels de classification qui sont pertinents et surtout non triviaux pour de telles tâches, et de les interpréter suivant une grille de lecture linguistiquement contrôlée. Sans viser le dépassement des différentes méthodes évoquées dans l'état de l'art (paragraphe 4.1.), la méta-étude effectuée apporte la démonstration que les critères relevant des seules composantes dialectique (construction narrative et argumentative) et dialogique (positionnements énonciatifs, acteurs), permettent d'obtenir des résultats de classifications approchant (expérience 5.3), voisinant (expérience 5.2) ou surpassant (expérience 5.1) une ligne de comparaison simulant les techniques apprentistes standard. Mieux encore, l'étude souligne que ces critères textuels classifiants, identifiés selon une méthode d'extraction ascendante (la textométrie), ne ressortissent nullement aux catégories traditionnellement proposées, souvent au moyen de méthodes descendantes par application de modèles cognitifs ou issus de la psychologie, notamment.

Références

- BEUST, P. (1998). *Contribution à un modèle interactionniste du sens. Amorce d'une compétence interprétative pour les machines*, Thèse de doctorat, Caen.
- BLOOM, K. & ARGAMON, S. (2010). "Unsupervised extraction of appraisal expressions", *Proceedings of the 23rd Canadian conference on Advances in Artificial Intelligence (AI'10)*, Atefeh Farzindar and Vlado Kešelj (Eds.). Springer-Verlag, Berlin, Heidelberg, p. 290-294.
- BOMMIER-PINCEMIN, B. (1999). *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat, Université Paris IV Sorbonne.
- BOURION, E. (2001). *L'aide à l'interprétation des textes électroniques*, Thèse de doctorat, Université Nancy 2.
- BRUNET E. (2000). « Qui lemmatise dilemme attise », *Lexicometrica*, 2, 1-19.
- CHARAUDEAU, P. (1992). *Grammaire du sens et de l'expression*. Hachette Education.
- CHURCH, K. (2011). "A pendulum swung too far", *Linguistic Issues in Language Technology*, 6.
- EENSOO E. & VALETTE M. (2012). « Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments », dans G. Antoniadis, H. Blanchon, G. Sérasset, Eds, *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, vol. 2, *TALN, 4-8 juin 2012*, Grenoble, p. 367-374.
- EENSOO E. & VALETTE M. (2014a). « Sémantique textuelle et TAL : un exemple d'application à l'analyse des Sentiments », dans D. Ablali, S. Badir, D. Ducard, Eds., *Documents, textes, œuvres*, Presses Universitaires de Rouen, Collection Rivages linguistiques.
- EENSOO E. & VALETTE M. (2014b). « Approche textuelle pour le traitement automatique du discours évaluatif », dans A. Jackiewicz, (éd.), *Études sur l'évaluation axiologique, Langue française*, décembre 2014, 184, p. 107-122.
- GHORBEL H. & JACOT D. (2011). "Further Experiments in Sentiment Analysis of French Movie Reviews", E. Mugellini, P. Szczepaniak, M. Pettenati, M. Sokhn, Eds., *Advances in Intelligent Web Mastering – 3*, Berlin / Heidelberg : Springer, 86, p. 19-28
- GROUIN C., BERTHELIN JB, EL AYARI S, HEITZ T, HURAUULT-PLANTET M, JARDINO M, KHALIS Z & LASTES M. (2007). Présentation de DEFT'07. *Actes de l'atelier de clôture du 3ème Défi Fouille de Textes*, 2007. Grenoble, France. p. 1–8.
- HATZIVASSILOGLU V. & WIEBE J. (2000). "Effects of adjective orientation and gradability on sentence subjectivity", *Proceedings of the International Conference on Computational Linguistics (COLING)*.
- HALL M., EIBE F. HOLMES G., PFAHRINGER B., REUTEMANN P. & WITTEN I. H. (2009). « The WEKA Data Mining Software: An Update », *SIGKDD Explorations*, Volume 11, Issue 1.
- HALL D., JURAFSKY D. & MANNING C. D. (2008). "Studying the history of ideas using topic models", *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 363–371.
- HEIDEN S., MAGUE J.-P. & PINCEMIN B. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », I. C. Sergio Bolasco (éd.), *JADT 2010*, vol. 2, p. 1021-1032.
- KIM, S.M., VALITUTTI, A. & CALVO, R.A. (2010). "Evaluation of unsupervised emotion models to textual affect recognition", *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (CAAGET '10)*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 62-70.
- LAFON, P. (1980). « Sur la variabilité de la fréquence des formes dans un corpus ». *Mots*, 1, p. 127-165.
- LAFON P. (1981), « Analyse lexicométrique et recherche des cooccurrences », *Mots*, 3, p. 95-148.

- LAVALLEY, R.; CLAVEL, C. & BELLOT, P. (2010). « Extraction probabiliste de chaînes de mots relatives à une opinion », *Traitement Automatique des Langues*, 51, p. 101-130.
- LIN W.-H. & HAUPTMANN, A. (2006). "Are these documents written from different perspectives? a test of different perspectives based on statistical distribution divergence", *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL-44)*. Association for Computational Linguistics, Stroudsburg, PA, USA, p. 1057-1064.
- LIU, H., LIEBERMAN, H. & SELKER, T. (2003). "A model of textual affect sensing using real-world knowledge", *Proceedings of the 8th international conference on Intelligent user interfaces (IUI '03)*, ACM, New York, NY, USA, p. 125-132.
- MCCALLUM A. & NIGAM K. (1998). "A Comparison of Event Models for Naive Bayes, Text Classification", *AAAI-98 Workshop on 'Learning for Text Categorization'*, p. 41-48
- MAYAFFRE D. (2008). « De l'occurrence à l'isotopie. Les cooccurrences en lexicométrie », dans M. Valette (éd), *Textes, documents numériques, corpus. Pour une science des textes instrumentée, Syntaxe et sémantique*, 9, p. 53-72.
- MICHELI R., HEKMAT I. & RABATEL A., Eds. (2013). *Les émotions argumentées dans les médias, Le discours et la langue*, 4/1, EME Éditions, 222 p.
- PANG P., LEE L. & VAITHYANATHAN, S. (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques", *Proceedings of EMNLP*, p. 79-86.
- PEDAUQUE R. T., Coll. (2007). *La redocumentarisation du Monde*, Paris, Éditions Cepadues, 213 p.
- PERLERIN, V. (2004). *Sémantique légère pour le document. Assistance personnalisée pour l'accès au document et l'exploration de son contenu*, Thèse de doctorat, Caen.
- PINCEMIN B. (2010). "Semántica interpretativa y textometría", dans C. Duteil-Mougél et V. Cárdenas, Eds., *Semántica e interpretación, Tópicos del Seminario*, 23, Enero-junio 2010, p. 15-55.
- PLATT J. (1998). "Machines using Sequential Minimal Optimization", dans B. Schoelkopf, C. Burges et A. Smola (éds), *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MIT Press.
- RASTIER F. (éd.) (1995). *L'analyse thématique des données textuelles : l'exemple des sentiments*, Paris, Didier, collection Études de sémantique lexicale, 270 p.
- RASTIER F. (1998). « Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage », *Langages*, 129, p. 97-111.
- RASTIER F. (2001). *Arts et sciences des textes*, Paris, PUF, 303 p.
- RASTIER F. (2011). *La mesure et le grain. Sémantique de corpus*, Paris, Honoré Champion, 272 p.
- RASTIER F. & PINCEMIN B. (1999). « Des genres à l'intertexte », I. Kanellos (éd.), *Cahiers de Praxématique*, 33, *Sémantique de l'intertexte*, p. 83-111.
- ROSSIGNOL M. (2005). *Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle*, Thèse de doctorat, Université de Rennes 1.
- SALEM A., LAMALLE C., MARTINEZ W., FLEURY S., FRACCHIOLLA B., KUNCOVA A. & MAISONDIEU A. (2003). *Lexico3 – Outils de statistique textuelle, Manuel d'utilisation*, Université de la Sorbonne nouvelle – Paris 3.
- SAUSSURE, F. DE (2002). *Écrits de linguistique générale*, Paris, Gallimard.
- TANGUY L. (2012). *Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes*, Mémoire d'habilitation à diriger des recherches, Université Toulouse-Le Mirail, Toulouse.

TANGUY L. & FABRE C. (2014). « Évolutions de la linguistique outillée : méfaits et bienfaits du TAL », *L'information grammaticale*, 142, p. 15-23.

THLIVITIS, T. (1998). *Sémantique Interprétative Intertextuelle : assistance informatique anthropocentrée à la compréhension des textes*, Thèse de doctorat, Rennes 1.

VERNIER M., MONCEAUX L., DAILLE B. & DUBREIL E. (2009a). « Catégorisation des évaluations dans un corpus de blogs multi-domaine », *Revue des nouvelles technologies de l'information (RNTI)*, p. 45-70.

VERNIER M., MONCEAUX L. & DAILLE B. (2009b). « DEFT'09 : détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique », *Actes de l'atelier de clôture de la 5ème édition du Défi Fouille de Textes*.

VILLENA-ROMAN, J., COLLADA-PEREZ, S., LANA-SERRANO, S., GONZALEZ-CRISTOBAL, J. C. (2011). "Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization", *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*, p. 323-328.

WHITELAW C., GARG N. & ARGAMON S. (2005). "Using appraisal groups for sentiment analysis", ACM (éd.), *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, p. 625-631.

YVON F. (2006). *Des apprentis pour le traitement automatique des langues*, Mémoire d'habilitation à diriger des recherches, Université Pierre et Marie Curie, Paris.

ZWEIGENBAUM P., BELLOT P., GRAU B., LIGOZAT A.-L., ROBBA I., ROSSET S., TANNIER X ? et VILNAT A. (2008). « Apports de la linguistique dans les systèmes de recherche d'informations précises », *Revue française de linguistique appliquée* 1/ 2008 (Vol. XIII), p. 41-62.

Estimation de l'homogénéité sémantique pour les Questionnaires à Choix Multiples

Van-Minh Pho^{1,2} Brigitte Grau^{1,3} Anne-Laure Ligozat^{1,3}
(1) LIMSI-CNRS, Orsay
(2) Université Paris-Sud, Orsay
(3) ENSIIE, Evry
prenom.nom@limsi.fr

Résumé. L'homogénéité sémantique stipule que des termes sont sémantiquement proches mais non similaires. Cette notion est au cœur de travaux relatifs à la génération automatique de questionnaires à choix multiples, et particulièrement à la sélection automatique de distracteurs. Dans cet article, nous présentons une méthode d'estimation de l'homogénéité sémantique dans un cadre de validation automatique de distracteurs. Cette méthode est fondée sur une combinaison de plusieurs critères de voisinage et de similarité sémantique entre termes, par apprentissage automatique. Nous montrerons que notre méthode permet d'obtenir une meilleure estimation de l'homogénéité sémantique que les méthodes proposées dans l'état de l'art.

Abstract.

Semantic homogeneity estimation for MCQs

Semantic homogeneity states that terms are semantically close but not similar. This notion is the focus of work related to multiple-choice test generation, and especially to automatic distractor selection. In this paper, we present a method to estimate semantic homogeneity within a framework of automatic distractor validation. This method is based on a combination of several criteria of semantic relatedness and similarity between terms, by machine learning. We will show that our method allows to obtain a better estimation of semantic homogeneity than methods proposed in related work.

Mots-clés : similarité, voisinage sémantique, classification de termes.

Keywords: similarity, semantic relatedness, term ranking.

1 Introduction

La notion de similarité sémantique permet d'estimer l'intensité du lien sémantique reliant deux concepts. Cette notion est au cœur de travaux relatifs à la génération automatique de Questionnaires à Choix Multiples (QCM), et particulièrement à la sélection automatique de distracteurs (mauvais choix de réponse), qui se doivent d'être sémantiquement homogènes pour respecter les règles de construction d'un QCM. Ces travaux (Karamanis *et al.*, 2006; Mitkov *et al.*, 2009) abordent la problématique de la sélection automatique de distracteurs comme un problème de similarité sémantique entre les différentes options (choix de réponse incluant les distracteurs et la réponse correcte). Cependant, la problématique de la sélection automatique de distracteurs est différente d'un problème de similarité sémantique : en effet, bien que les options doivent être sémantiquement proches, elles ne doivent pas être de sens similaires.

Nous proposons donc une définition différente de la notion d'homogénéité sémantique qui stipule que les options sont sémantiquement proches mais non similaires. L'objectif de cet article est de proposer une méthode d'estimation de l'homogénéité sémantique des options en combinant différentes mesures de voisinage et de similarité sémantique, dans une perspective applicative de validation ou de génération automatique de QCM. Afin de présenter en détail la notion d'homogénéité sémantique et la méthode que nous proposons, il est nécessaire d'exposer le contexte des QCM.

Les QCM sont largement utilisés dans de nombreux contextes d'apprentissage et d'évaluation. Les principales raisons en sont que leur évaluation peut être automatisée et que leur pertinence, ainsi que leur objectivité dans l'évaluation des compétences de l'apprenant, ont été prouvées (Haladyna *et al.*, 2002). Cependant, la rédaction de QCM est coûteuse

en temps, et la qualité des QCM est cruciale si l'on veut s'assurer que les résultats des apprenants correspondent à leurs compétences. Ainsi, la réalisation d'applications permettant l'évaluation automatique de la qualité de QCM ou la génération automatique de ceux-ci est nécessaire pour aider les enseignants.

Un QCM est un ensemble de *questions*, chacune d'entre elles étant composée de deux parties (cf. exemple ci-dessous) : l'*amorce* et les *options*, incluant la *réponse* (option correcte) et un ou plusieurs *distracteurs* (options incorrectes).

Amorce : De quel pays est originaire le kimchi ?

Réponse : Corée

Distracteur : Japon

Distracteur : Chine

Distracteur : Mongolie

La sélection des distracteurs est une tâche difficile lors de la création de QCM : la qualité d'une question dépend principalement de la qualité de ses options (Rodriguez, 2005). Le critère principal de la qualité des options est l'homogénéité sémantique de celles-ci, c'est-à-dire que les options doivent partager différents traits sémantiques. Cependant, elles doivent avoir des contenus sémantiques suffisamment différents pour constituer des réponses plausibles mais non possibles. Cette notion d'homogénéité sémantique découle des règles de rédaction de QCM «*Rendre la formulation des options homogène en contenu et en structure grammaticale*» et «*Rendre les options indépendantes les unes des autres : le sens de l'une ne doit pas être inclus dans le sens de l'autre*» (Haladyna *et al.*, 2002). Aussi, étant donné des candidats, extraits de différentes ressources, l'estimation de leur homogénéité sémantique par rapport à la réponse correcte permet de les filtrer et de ne garder que ceux qui sont assez homogènes pour constituer des distracteurs pertinents.

Nous proposons d'évaluer l'homogénéité sémantique en comparant chaque distracteur à la réponse et en calculant plusieurs critères fondés sur l'utilisation de différentes ressources sémantiques pour couvrir de nombreuses relations sémantiques, mesures qui sont combinées dans un modèle d'apprentissage automatique. Nous étendons les travaux existants (Karamanis *et al.*, 2006; Lee & Seneff, 2007; Mitkov *et al.*, 2009) en proposant une plus large palette de mesures, étendant ainsi leur couverture et en traitant plus de types de distracteurs, tels que les chunks (syntagmes de plus bas niveau) et les entités nommées, sans limitation de domaine. Dans cet article, nous montrerons par une évaluation de corpus que notre combinaison de mesures de voisinage et de similarité sémantique permettent d'obtenir une meilleure estimation de l'homogénéité sémantique que les méthodes présentées par les travaux précédents.

2 État de l'art

Il existe différentes stratégies d'estimation de l'homogénéité sémantique : mesure fondée sur la fréquence des collocations (Lee & Seneff, 2007), mesures de voisinage ou de similarité distributionnel (Karamanis *et al.*, 2006; Mitkov *et al.*, 2009), sélection des distracteurs appartenant à un document de référence à partir duquel la question (amorce et réponse) est créée (Mitkov *et al.*, 2009), mesures fondées sur des bases de connaissances hiérarchiques (Mitkov *et al.*, 2009) ou sur la similarité phonétique (Mitkov *et al.*, 2009). La mesure calculée par (Lee & Seneff, 2007) privilégie les termes apparaissant le plus fréquemment avec les contextes gauche et/ou droit de la réponse extraite d'une phrase. (Karamanis *et al.*, 2006) et (Mitkov *et al.*, 2009) calculent un score de similarité distributionnelle entre les candidats et la réponse. (Mitkov *et al.*, 2009) calculent notamment une mesure dont les co-occurents comparés sont les mots liés aux mots comparés par une relation de dépendance dans un grand corpus de documents. (Mitkov *et al.*, 2009) utilisent d'autres stratégies : la première consiste à privilégier les candidats apparaissant dans le document de référence. Les autres stratégies sont les mesures de voisinage et de similarité sémantique fondées sur WordNet exposées à la section 3.3 et une mesure de similarité phonétique fondée sur Soundex, un algorithme d'indexation phonétique de mots.

L'estimation de l'homogénéité sémantique est évaluée à travers la qualité des distracteurs sélectionnés. Cette évaluation est effectuée par des apprenants (à travers des tests psychométriques) (Lee & Seneff, 2007; Mitkov *et al.*, 2009) ou par le jugement d'experts du domaine (Karamanis *et al.*, 2006).

Les principales limitations de ces travaux sont qu'ils sont limités à une application de domaine spécifique (médecine, apprentissage des prépositions) et/ou par les types syntaxiques des réponses (mots, syntagmes nominaux limités aux chunks nominaux suivis ou non d'un chunk prépositionnel à tête nominale (Mitkov & Ha, 2003)) tandis que notre travail n'est pas spécifique à un domaine et qu'il couvre tout type de chunk et d'entité nommée. De plus, aucun des travaux de l'état de l'art n'évalue l'homogénéité sémantique sur un corpus de QCM de référence.

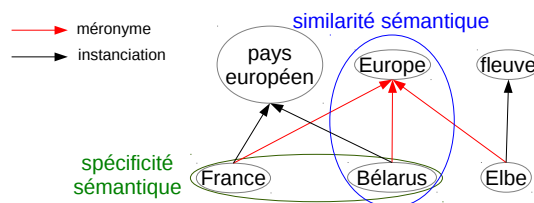


FIGURE 1 – Caractérisation sémantique de paires de nœuds

3 Homogénéité sémantique

L'homogénéité sémantique stipule que les options partagent des caractéristiques sémantiques communes. Dans l'exemple de la section 1, toutes les options sont des pays asiatiques. L'homogénéité sémantique des différentes options signifie que celles-ci doivent être sémantiquement voisines, sans partager le même contenu sémantique.

Nous donnons la définition de plusieurs notions utiles à la définition de l'homogénéité sémantique en faisant référence à une organisation des connaissances sous la forme d'un graphe hiérarchique (figure 1) contenant des concepts typés et des relations sémantiques.

La définition du *voisinage sémantique* est la suivante : «Le voisinage sémantique indique dans quelle mesure deux concepts sont sémantiquement distants dans un réseau ou une taxonomie en utilisant toutes les relations entre eux (c'est-à-dire des relations d'hyponymie/hyperonymie¹, d'antonymie², de méronymie³ et toutes sortes de relations fonctionnelles incluant *is-made-of*, *is-an-attribute-of*, etc.)» (Ponzetto & Strube, 2007). Le voisinage sémantique est établi entre deux termes lorsqu'il existe un chemin entre les concepts auxquels ils se réfèrent, et le degré de voisinage est dépendant de la longueur du chemin. Dans la figure 1, tous les concepts peuvent être considérés comme sémantiquement voisins.

Nous définissons la *similarité sémantique* comme un cas particulier de voisinage sémantique : deux termes sont similaires s'ils partagent le même sens (c'est-à-dire qu'ils sont des synonymes) ou un sens partiel, c'est-à-dire que les concepts auxquels ils se réfèrent sont liés par une chaîne ascendante ou descendante de relations *is-a* ou de méronymie, tels que «Bélarus» et «Europe» dans la figure 1.

Nous définissons la *spécificité sémantique* comme un cas particulier de voisinage sémantique entre deux termes dont les concepts auxquels ils se réfèrent partagent un ancêtre commun direct, à l'instar de «France» et «Bélarus» dans la figure 1.

Nous proposons enfin la définition de l'*homogénéité sémantique* comme étant un cas particulier de *voisinage sémantique* qui considère toutes les relations entre les concepts comparés mais exclut la notion de *similarité sémantique* : deux options ne peuvent être similaires. Enfin, une meilleure homogénéité est atteinte si la *spécificité sémantique* est respectée.

Pour évaluer l'homogénéité sémantique des options, nous comparons les distracteurs à la réponse sur différents traits sémantiques. Notre objectif n'est pas d'apprendre une décision indiquant une homogénéité sémantique ou non car celle-ci est plus une question de degré qu'une décision binaire. De ce fait, nous considérons l'estimation automatique de l'homogénéité sémantique comme un problème d'ordonnancement dans lequel les candidats sémantiquement homogènes, soit les candidats classés dans les premiers rangs, sont les distracteurs. Les candidats à classer, excepté les distracteurs, sont sélectionnés selon des critères syntaxiques.

Pour estimer l'homogénéité sémantique entre un distracteur et la réponse, nous calculons plusieurs scores de voisinage et de similarité sémantique. Ces mesures sont calculées à partir de différents types de ressources : des représentations sémantiques hiérarchiques, et des représentations distributionnelles. Les représentations hiérarchiques permettent de prendre en compte les relations sémantiques explicites pour comparer deux concepts. Cependant, les ressources correspondant à ces représentations sont souvent limitées par leur couverture. Les mesures de voisinage distributionnel, construites selon le principe stipulant que des termes ayant des contextes similaires dans un corpus sont sémantiquement voisins, ont une couverture plus large mais la nature des relations sémantiques et les informations sur les représentations hiérarchiques sont inconnues. Dans les sections suivantes, nous présentons chacune des mesures choisies.

1. Deux concepts dont le premier a un sens plus spécifique/général que le second.
 2. Deux concepts dont les sens sont opposés.
 3. Deux concepts dont le premier est une partie ou un membre du second.

Les mesures fondées sur des représentations hiérarchiques ou des concepts explicitement typés sont :

- la similarité des types d’entité nommée ;
- la similarité des types spécifiques provenant de la base de connaissances DBpédia ;
- plusieurs mesures de voisinage sémantique fondées sur la base de données lexicale WordNet.

Les mesures de voisinage distributionnel sont :

- la comparaison des liens de Wikipédia ;
- l’Analyse Sémantique Explicite.

3.1 Similarité des types d’entité nommée

La première mesure de voisinage sémantique que nous présentons est fondée sur une annotation de texte en entités nommées, c’est-à-dire classiquement les noms de lieux, personnes et organisations. Nous considérons que l’identité de type d’entité nommée de deux termes est théoriquement un critère nécessaire pour qu’ils soient sémantiquement homogènes. Ce critère n’est pas suffisant car les types d’entité nommée choisis sont des catégories très générales, et qu’il convient également d’exclure les termes similaires.

Afin de mesurer ce critère d’homogénéité, nous considérons les trois grandes catégories : *lieu*, *organisation* et *personne*. Pour comparer le type d’entité nommée de deux termes, nous utilisons la mesure binaire *meme_type_EN* (équation (1)) qui indique simplement si les termes sont de même type d’entité nommée ou non.

$$meme_type_EN(t_1, t_2) = \begin{cases} 1 & \text{si } EN(t_1) = EN(t_2) \wedge t_1 \text{ est une entité nommée} \wedge t_2 \text{ est une entité nommée} \\ 0 & \text{sinon} \end{cases} \quad (1)$$

où t_1 et t_2 sont deux termes et $EN(t)$ est le type d’entité nommée du terme t .

3.2 Similarité des types sémantiques provenant de DBpédia

En plus de mesurer la similarité de types d’entité nommée généraux, nous souhaitons comparer les types sémantiques des termes à un niveau de granularité plus fin : des types plus spécifiques permettent de vérifier plus précisément l’homogénéité sémantique. Ainsi dans l’exemple de la section 1, les options sont toutes des pays asiatiques, ce qui donne une indication plus précise de leur homogénéité sémantique que de vérifier simplement qu’elles sont des lieux.

Cependant, tandis que les types d’entité nommée peuvent être reconnus indépendamment d’une ressource, les types spécifiques doivent être reconnus à partir d’une taxonomie hiérarchique. Pour cela, nous utilisons DBpédia⁴, une ressource hiérarchique construite à partir des pages de Wikipédia. Les entités de DBpédia sont associées à des types sémantiques qui représentent les classes de l’ontologie de DBpédia, organisées en une taxonomie⁵. Cette ressource a l’avantage d’avoir une large couverture pour le domaine ouvert.

Pour calculer l’homogénéité sémantique entre deux termes fondée sur leur type DBpédia et leur position dans leur taxonomie, nous utilisons la mesure de Wu et Palmer (Wu & Palmer, 1994), $wup(t_1, t_2)$.

$$wup(t_1, t_2) = \frac{2 \times profondeur(lcs(t_1, t_2))}{profondeur(type(t_1)) + profondeur(type(t_2))} \quad (2)$$

où $type(t)$ est le type DBpédia du terme t , $profondeur(u)$ est la profondeur du type u dans la taxonomie ($profondeur(u) = 1$ si u est la racine de la taxonomie) et $lcs(t_1, t_2)$ est l’hyperonyme commun le plus spécifique de $type(t_1)$ et $type(t_2)$. La mesure de Wu et Palmer est fondée sur le chemin le plus court entre les deux concepts pondéré par leur profondeur dans la taxonomie. Ainsi, deux concepts profonds de parent commun obtiennent un meilleur score que deux concepts moins profonds de parent commun.

4. <http://dbpedia.org/About>

5. <http://mappings.dbpedia.org/server/ontology/classes/>

3.3 Mesures de voisinage sémantique fondées sur WordNet

Les mesures précédentes sont fondées sur la similarité des types sémantiques des termes. Dans cette sous-section, nous présentons des mesures de voisinage et de similarité sémantique fondées sur le sens des termes et les relations qui les lient.

Pour mesurer l'homogénéité sémantique sur tout type de termes et particulièrement les termes qui ne sont pas des entités nommées, nous utilisons des mesures définies pour WordNet⁶, un réseau lexical qui groupe les mots synonymes en *synsets* liés par des relations sémantiques. Une glose (définition) est associée à chaque synset. WordNet contient également des entités nommées, mais restreintes à quelques types d'entité nommée (grandes villes, pays, continents...). Nous utilisons les quatre mesures sélectionnées par (Mitkov *et al.*, 2009) dans leur travail de sélection automatique des distracteurs : la mesure de recoupement étendu de gloses (*extended gloss overlap measure*) (Banerjee & Pedersen, 2003) ; la mesure de Leacock et Chodorow (Leacock & Chodorow, 1998) fondée sur le plus court chemin entre les synsets ; les mesures de Jiang et Conrath (Jiang & Conrath, 1997), et de Lin (Lin, 1997), fondées sur le *contenu d'information* (*information content*).

La mesure de recoupement étendu de gloses (Banerjee & Pedersen, 2003) (*simREG*) prend en compte les gloses des synsets comparés, ainsi que les gloses de leurs hyperonymes et de leurs hyponymes.

$$\begin{aligned} \text{simREG}(s_1, s_2) = & \text{score}(\text{def}(s_1), \text{def}(s_2)) + \text{score}(\text{hype}(s_1), \text{hype}(s_2)) + \text{score}(\text{hypo}(s_1), \text{hypo}(s_2)) \\ & + \text{score}(\text{hype}(s_1), \text{def}(s_2)) + \text{score}(\text{def}(s_1), \text{hype}(s_2)) \end{aligned} \quad (3)$$

où $\text{def}(s)$ est la glose du synset s , $\text{hype}(s)$ est la glose de l'hyperonyme de s (si s a plusieurs hyperonymes, alors $\text{hype}(s)$ est la concaténation de ces gloses) et $\text{hypo}(s)$ est la glose de l'hyponyme de s (si s a plusieurs hyponymes, alors $\text{hypo}(s)$ est la concaténation de ces gloses). $\text{score}(g(s_1), g(s_2))$ est la somme des longueurs des chaînes communes des gloses de s_1 et s_2 au carré.

La mesure de Leacock et Chodorow (Leacock & Chodorow, 1998) (*simLCH*) est fondée sur le plus court chemin entre deux synsets dans la taxonomie, c'est-à-dire le nombre minimal d'arêtes entre ces synsets. Cette mesure ne prend en compte que les relations d'hyponymie et d'hyperonymie.

$$\text{simLCH}(s_1, s_2) = -\log\left(\frac{\text{len}(s_1, s_2)}{2 \times \text{MAX}}\right) \quad (4)$$

où $\text{len}(s_1, s_2)$ est le nombre minimal d'arêtes entre les synsets s_1 et s_2 , et MAX est la profondeur de la taxonomie.

Les mesures de Jiang et Conrath (Jiang & Conrath, 1997) et de Lin (Lin, 1997) sont fondées sur le contenu d'information des synsets et de leur hyperonyme commun ($\text{lcs}(s_1, s_2)$). Le contenu d'information $\text{IC}(s)$ représente l'importance de l'information relative à un synset s dans un contexte donné.

$$\text{IC}(s) = -\log(p(s)) \quad (5)$$

où $p(s)$ est la probabilité d'apparition de s et de ses hyponymes dans un corpus de référence. Ainsi, pour deux synsets s et s' tels que s' est un hyponyme de s , $p(s) \geq p(s')$ et donc $\text{IC}(s) \leq \text{IC}(s')$.

Les formules de la mesure de Jiang et Conrath (*simJCN*) et de Lin (*simLin*) sont présentées ci-dessous.

$$\text{simJCN}(s_1, s_2) = \frac{1}{\text{IC}(s_1) + \text{IC}(s_2) - 2 \times \text{IC}(\text{lcs}(s_1, s_2))} \quad (6)$$

$$\text{simLin}(s_1, s_2) = \frac{2 \times \text{IC}(\text{lcs}(s_1, s_2))}{\text{IC}(s_1) + \text{IC}(s_2)} \quad (7)$$

où $\text{lcs}(s_1, s_2)$ est l'hyperonyme commun le plus spécifique des synsets s_1 et s_2 . Ces mesures comparent le contenu d'information des synsets s_1 et s_2 avec celui de leur hyperonyme commun le plus spécifique. Ces mesures combinent les connaissances sémantiques de WordNet avec les distributions des concepts dans un grand corpus. Elles privilégient les

6. <http://wordnet.princeton.edu/>

concepts dont l'hyperonyme commun le plus spécifique est proche, et dont les contenus d'informations (calculés à partir des distributions des concepts dans le corpus) sont proches de leur hyperonyme commun le plus spécifique.

Les termes pouvant avoir plusieurs sens, nous les associons à plusieurs synsets. Ainsi, pour calculer le voisinage sémantique entre deux termes, nous calculons les mesures sur toutes les paires de synsets associées aux termes et nous gardons le score maximal.

Ces mesures se complètent car elles calculent le score de similarité ou de voisinage sémantique selon différents critères : tandis que la mesure de Leacock et Chodorow se fonde uniquement sur les relations sémantiques explicites entre les concepts, la mesure de recouplement étendu de gloses se fonde sur la proximité textuelle des gloses des concepts et les mesures de Jiang et Conrath et de Lin se fondent sur un corpus de textes pour comparer l'importance (représenté par la fréquence d'apparition) des concepts.

Afin d'avoir également des mesures s'appliquant à tout type de termes sans la limite de leur présence ou non dans une ressource sémantique hiérarchique, nous avons sélectionné des mesures pouvant être calculées sur de grands corpus. Cette famille de mesures ne prennent pas en compte les relations sémantiques explicites, et sont dédiées à l'estimation du voisinage sémantique à l'instar des mesures précédentes concernant la mesure du voisinage sémantique. L'hypothèse qui a été considérée pour la conception de ces mesures est que les termes ont des sens sémantiquement proches s'ils partagent un contexte similaire.

3.4 Comparaison des liens de pages de Wikipédia

Une représentation contextuelle possible d'un terme est l'ensemble des liens entrants et sortants associés à une page de Wikipédia. Nous considérons les pages dont le titre correspond au terme d'une option. Les liens entrants et sortants représentent les pages de Wikipédia associées au corps d'une page de Wikipédia. L'outil Wikipedia Miner (Milne & Witten, 2013) calcule un score appris sur ces liens à partir de dumps de Wikipédia. Ce score est une combinaison de huit attributs représentant quatre mesures calculées sur les liens entrants et sortants :

- l'union des liens des pages comparées ;
- l'intersection des liens des pages comparées ;
- la *normalized link distance*, adaptée de la *normalized Google distance* (Cilibrasi & Vitanyi, 2007), calculant la distance sémantique de deux pages p_1 et p_2 en comparant les pages de Wikipédia où apparaissent les liens associés à p_1 et p_2 . Si p_1 et p_2 sont liées aux mêmes pages, cela signifie un fort voisinage sémantique entre p_1 et p_2 , tandis que si p_1 et p_2 sont liées à des pages différentes, cela signifie un faible voisinage sémantique entre p_1 et p_2 . La formule de la *normalized link distance* est la suivante : $nld(p_1, p_2) = \frac{\log(\max(|P_1|, |P_2|)) - \log(|P_1 \cap P_2|)}{\log(|W|) - \log(\min(|P_1|, |P_2|))}$ où P_1 et P_2 sont les ensembles de pages reliant respectivement p_1 et p_2 , et W est l'ensemble de toutes les pages de Wikipédia ;
- la similarité vectorielle des liens (*link vector similarity*), inspirée de la mesure du TF-IDF mais appliquée aux liens des pages comparées vers les pages de Wikipédia, au lieu des mots traités par le TF-IDF. Les dimensions des vecteurs comparés sont calculées à partir du $lf \times iaf$ (*link frequency* \times *inverse article frequency*). La fréquence des liens (*link frequency*) donne une estimation de l'importance du lien l_i dans une page p_j : $lf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ où $n_{i,j}$ est le nombre d'occurrences du lien l_i dans la page p_j . La fréquence inverse d'article (*inverse article frequency*) mesure l'importance générale d'un lien : $iaf_i = \log\left(\frac{|W|}{|p:l_i \in p|}\right)$ où W est l'ensemble des pages de Wikipédia. Le score de voisinage sémantique est l'angle des vecteurs des pages comparées.

3.5 Analyse Sémantique Explicite

Une autre représentation contextuelle des termes est leur distribution à travers les documents d'un corpus. Deux termes dont les distributions (c'est-à-dire les fréquences d'apparition) sont proches dans les mêmes documents ont une forte probabilité d'être sémantiquement voisins. Afin de comparer les distributions de deux termes, nous calculons une mesure fondée sur l'Analyse Sémantique Explicite (*Explicit Semantic Analysis*, ESA) (Gabrilovich & Markovitch, 2007). L'ESA est fondée sur une représentation vectorielle de textes (d'un mot à un document entier) dont les dimensions sont les poids du texte dans chaque document du corpus. Un mot est représenté par un vecteur de poids et un texte contenant plusieurs mots est représenté par le barycentre des vecteurs de poids représentant chaque mot du texte. Le poids d'un mot m dans un document d correspond au TF-IDF de m dans d . Le score de voisinage de deux textes est le cosinus des vecteurs

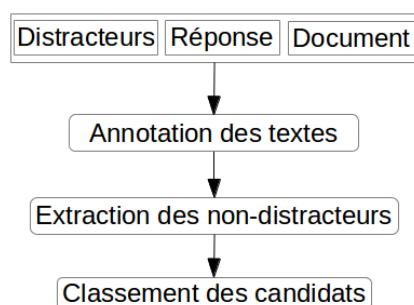


FIGURE 2 – Architecture montrant les différentes étapes de la validation automatique de distracteurs

représentant ces textes. Dans notre cas, le corpus de documents est Wikipédia. Pour calculer la mesure fondée sur l’ESA, nous utilisons l’outil ESALib⁷.

4 Évaluation de la qualité des distracteurs

Nous appliquons l’estimation de l’homogénéité sémantique des options à la validation automatique de distracteurs. Pour ce faire, nous joignons les distracteurs corrects de la question à des non-distracteurs (termes n’ayant pas été manuellement sélectionnés pour être les distracteurs de la question). Nous appelons *candidats* l’ensemble des distracteurs et des non-distracteurs. Notre objectif est d’apprendre un modèle d’ordonnancement capable de classer les distracteurs dans les premiers rangs, étant donné qu’ils devraient être plus homogènes à la réponse que les non-distracteurs. De plus ce modèle doit être capable de reconnaître les non-distracteurs similaires à la réponse mais non reconnus comme tels par la ressource WordNet afin de ne pas les sélectionner.

Les questions que nous traitons sont associées à un document de référence à partir duquel les amorces sont conçues. Les non-distracteurs sont sélectionnés dans ce document selon un critère d’homogénéité syntaxique. La figure 2 montre les différentes étapes de la validation automatique des distracteurs. Une première étape consiste à annoter le document afin d’obtenir des informations syntaxiques et sémantiques sur les options. Les non-distracteurs sont ensuite sélectionnés selon deux méthodes différentes (chacune de ces méthodes est évaluée séparément) : la première méthode consiste à extraire les non-distracteurs à partir des annotations du document (evalNDdocument), et la seconde consiste à sélectionner les options des autres questions du corpus (evalNDoptions). Les candidats obtenus sont classés afin de valider les distracteurs.

4.1 Annotation du document et des options

Pour extraire les non-distracteurs et calculer les différentes mesures, les candidats et les réponses doivent être annotés par des informations syntaxiques et sémantiques. L’annotation est de meilleure qualité si ces extraits sont analysés dans leur contexte. Ainsi, nous effectuons quatre annotations du document dans l’ordre suivant :

1. une analyse syntaxique avec l’outil Stanford Parser (Klein & Manning, 2003) ;
2. une annotation en entité nommées avec l’outil Stanford Named Entity Recognition (Finkel *et al.*, 2005) ;
3. une annotation en types spécifiques pour trouver les entités associées à des entités de DBpédia (et, par extension, des pages de Wikipédia). Cette annotation est effectuée avec l’annotateur DBpedia Spotlight (Daiber *et al.*, 2013), qui associe des entités de DBpédia aux termes correspondants du document, et désambiguïse ces termes si nécessaire. Cependant, certains termes ne sont pas annotés (à tort) par DBpedia Spotlight. Nous associons ces termes à toutes les entités de DBpédia dont les titres correspondent à ces termes, donc sans désambiguïsation ;
4. une annotation en synsets de WordNet, visant à associer les termes (candidats et réponse) avec un ou plusieurs synsets de WordNet, comme suit :

7. <http://ticcky.github.io/esalib/>

- si le terme apparaît dans WordNet, le terme est associé à ses synsets correspondants ;
- si le terme n’apparaît pas dans WordNet et n’est pas une entité nommée, elle est associée aux synsets correspondants à sa tête syntaxique (par exemple, le chunk «the little cat» est associé aux synsets de WordNet de nom «cat»).

Les annotations des options sont extraites de leurs correspondances dans le document. Si une option n’apparaît pas dans le document, elle est annotée hors contexte de la même manière que l’est le document.

4.2 Extraction et annotation des non-distracteurs

L’extraction des non-distracteurs est différente selon le type d’évaluation des distracteurs (evalNDdocument et evalNDoptions).

Dans le cas de l’évaluation evalNDdocument, les non-distracteurs sont extraits de la même source que la question (car les questions sont créées à partir d’un document de référence). Ces non-distracteurs sont tous syntaxiquement homogènes à la réponse. Si celle-ci est une entité nommée, les non-distracteurs sont les entités nommées du document annotées avec le Stanford Named Entity Recognition, étant donné que (Pho *et al.*, 2014) ont montré que les distracteurs ont généralement le même type d’entité nommée que la réponse. Néanmoins, afin de prendre en compte la métonymie, nous sélectionnons tous les non-distracteurs qui sont des entités nommées, quel que soit leur type d’entité nommée. Si la réponse est un chunk et n’est pas une entité nommée, les non-distracteurs sont les chunks du document de même type syntaxique que la réponse. Les chunks sont sélectionnés à partir des arbres de constituants des phrases du document, avec Tregex (Levy & Andrew, 2006), un outil permettant de sélectionner des nœuds d’arbres syntaxiques à partir de patrons. À ces non-distracteurs, nous associons leur type d’entité nommée, leurs entités de DBpédia et leurs synsets de WordNet annotés dans le document.

Dans le cas de l’évaluation evalNDoptions, les non-distracteurs sont les options des autres questions du corpus. Si la réponse à la question n’est pas une entité nommée, seuls les non-distracteurs de même type syntaxique que la réponse sont gardés.

Pour les deux évaluations, un dernier filtrage consiste à retirer les non-distracteurs similaires à une option, afin d’éviter les chevauchements sémantiques : deux éléments sont considérés comme similaires s’ils sont associés aux mêmes entités de DBpédia ou s’ils réfèrent aux mêmes synsets dans WordNet. Parmi les non-distracteurs sélectionnés, certains d’entre eux pourraient être assez pertinents pour être des distracteurs mais ne le sont pas car la question contient assez de distracteurs, ou sont des distracteurs d’autres questions. Dans cet article, nous traitons ces non-distracteurs comme des non-distracteurs normaux mais nous envisageons de faire une annotation manuelle des non-distracteurs afin d’écarter ces cas.

4.3 Ordonnancement sémantique

Le classement des candidats selon les différents critères d’homogénéité sémantique est effectué avec SVMRank⁸, un outil d’ordonnancement automatique par apprentissage supervisé fondé sur un modèle SVM (*Séparateur à Vaste Marge* ou *Support Vector Machine*). Un SVM est un classifieur discriminant défini par un hyperplan séparant les données des différentes classes. L’outil SVMRank compare les couples de distracteurs-non-distracteurs d’une même question et apprend les poids des critères tels que pour chaque couple de distracteur-non-distracteur (d, nd) , $svm(d) > svm(nd)$, où $svm(c)$ est le score attribué au candidat c à partir de la combinaison des critères et des poids de chacun de ces critères, appris par SVM. Pour l’évaluation evalNDoptions, nous ajoutons un critère supplémentaire indiquant si le candidat apparaît dans le document de référence de la question ou non.

5 Expériences

5.1 Corpus

Afin d’évaluer notre méthode, nous utilisons un corpus de QCM en langue anglaise extrait de différentes sources : des tests d’évaluation de systèmes de compréhension automatique de textes fournis par QA4MRE⁹ (ensemble qa4mre) et

8. <http://www.cs.cornell.edu/people/tj/svmlight/svmrank.html>

9. <http://www.celct.it/newsReader.php?idnews=74>

corpus	ensemble	# q.	# opt.	(# q.)/opt.	objectif
tousQCM	qa4mre	341	1531	4,5	compréhension automatique de textes
	evalAnglais	394	1292	3,3	évaluation de la langue
	total	735	2823	3,8	
qcmEN	qa4mre	56	252	4,5	compréhension automatique de textes
	evalAnglais	47	150	3,2	évaluation de la langue
	total	103	402	3,9	
qcmNonEN	qa4mre	51	239	4,7	compréhension automatique de textes
	evalAnglais	100	342	3,8	évaluation de la langue
	total	151	581	3,8	

TABLE 1 – Caractéristiques des corpus : nom des corpus, nom des ensembles, nombre de questions, nombre d'options, nombre moyen d'options par question et objectif

plusieurs sites web d'apprentissage de la langue anglaise (ensemble evalAnglais). L'ensemble qa4mre a été conçu pour évaluer des machines, mais (Pho *et al.*, 2014) montrent qu'ils respectent des critères de formation de QCM. À partir de ce corpus, nous avons établi deux sous-corpus : le premier est constitué de questions dont les réponses sont des entités nommées (corpus qcmEN présenté au tableau 1), à l'instar de la question suivante :

Énoncé : Which Japanese city was the first to try limit convenience store hours ?

Réponse : Kyoto

Distracteur : Saitama

Distracteur : Tokyo

et le second est constitué de questions dont les réponses sont des chunks qui ne sont pas des entités nommées (corpus qcmNonEN présenté au tableau 1), à l'instar de la question suivante :

Énoncé : Trade union officials fear that this new campaign might end up unjustly penalizing _____, by driving the employers further underground.

Réponse : workers

Distracteur : employers

Distracteur : the "Mr. Bigs"

Le tableau 1 montre plusieurs caractéristiques du corpus (tousQCM), ainsi que des deux sous-corpus sur lesquels nous travaillons : qcmEN et qcmNonEN.

Le corpus qcmEN comporte 14 % des questions du corpus tousQCM et le corpus qcmNonEN comporte environ 20 % du corpus tousQCM. Les questions que nous traitons (chunks et EN) composent plus d'un tiers du corpus d'origine, ce qui montre que ces types de questions sont couramment posées lors de tests. Sur chacun des sous-corpus qcmEN et qcmNonEN, l'apprentissage du modèle a été effectué séparément.

5.2 Méthode d'évaluation

Nous considérons que les distracteurs sont sémantiquement plus proches de la réponse que les non-distracteurs et, par conséquent, devraient avoir un meilleur rang. Afin d'évaluer cela, nous calculons la précision (équation (8)) et le rappel (équation (9)) moyens au rang n en fonction du nombre de distracteurs, ainsi que la f-mesure (équation (10)).

$$PM = \frac{\sum_i^{nbQ} P_{i,nbD}}{nbQ} \quad (8)$$

$$RM = \frac{\sum_i^{nbQ} R_{i,nbD}}{nbQ} \quad (9)$$

$$F = 2 \times \frac{PM \times PR}{PM + PR} \quad (10)$$

où nbQ est le nombre de questions du corpus, nbD le nombre de distracteurs de la question évaluée, et $P_{i,nbD}$ et $R_{i,nbD}$ sont la précision (équation (11)) et le rappel (équation (12)) de la question i .

$$P_{i,nbD} = \frac{\#D \text{ de rang } \leq nbD}{\#C \text{ de rang } \leq nbD} \quad (11)$$

$$R_{i,nbD} = \frac{\#D \text{ de rang } \leq nbD}{nbD} \quad (12)$$

où D signifie distracteurs et C signifie candidats.

	qcmEN			qcmNonEN		
	<i>R</i>	<i>P</i>	<i>F</i>	<i>R</i>	<i>P</i>	<i>F</i>
<i>meme_type_EN</i>	0,83	0,26	0,40			
<i>wup</i> sur les types de DBpédia	0,70	0,34	0,46	0,94	0,14	0,24
<i>simREG</i>	0,67	0,25	0,36	0,37	0,23	0,28
<i>simLCH</i>	0,73	0,27	0,39	0,42	0,22	0,29
<i>simJCN</i>	0,83	0,23	0,36	0,40	0,18	0,25
<i>simLin</i>	0,84	0,23	0,36	0,40	0,18	0,25
Comparaison des liens de Wikipédia	0,41	0,32	0,36	0,76	0,22	0,34
ESA	0,40	0,34	0,37	0,35	0,24	0,28
Modèle d'ordonnancement	0,48	0,46	0,47	0,39	0,36	0,37

TABLE 2 – Résultats des méthodes de voisinage sémantique avec l'évaluation evalNDdocument

Cette évaluation constitue une proposition originale, les travaux étant usuellement évalués par des utilisateurs, ne permettant pas leur reproductibilité.

La précision et le rappel sont calculés pour chacune des mesures de voisinage sémantique, ainsi que pour le modèle d'ordonnancement. Nous évaluons le classement par une validation croisée en 7 sous-ensembles, c'est-à-dire que chacun des sous-ensembles du corpus est évalué selon le modèle appris à partir des autres sous-ensembles du corpus.

5.3 Résultats

Dans le cas de l'évaluation evalNDdocument, le tableau 2 montre que le modèle d'ordonnancement obtient un meilleur équilibre entre le rappel et la précision que les mesures individuelles, quel que soit le corpus. Le modèle donne une meilleure précision que les autres mesures et de meilleurs résultats que les mesures fondées sur WordNet, utilisées par (Mitkov *et al.*, 2009).

Certaines mesures évaluées donnent un meilleur rappel que le modèle d'ordonnancement. Nous distinguons deux cas : le premier concerne les mesures fondées sur les types (d'entité nommée et spécifiques) qui sont plus efficaces pour filtrer les candidats que pour sélectionner les distracteurs. Le second cas concerne les mesures dont la couverture des ressources est faible (WordNet dans le corpus qcmEN et Wikipédia dans le corpus qcmNonEN).

Les mesures donnent globalement des résultats inférieurs dans le corpus qcmNonEN. La raison principale est que les candidats et les réponses qui ne sont pas des entités nommées sont associés à moins d'informations sémantiques que les entités nommées, particulièrement sur les types sémantiques.

Dans le corpus qcmEN, la plupart des cas où les non-distracteurs ont un meilleur rang que les distracteurs sont dus au fait que les distracteurs et la réponse ne sont pas typés par un type (de DBpédia) très spécifique. Parmi les non-distracteurs restants, ceux-ci sont assez pertinents pour être des distracteurs ou sont similaires à la réponse, donc ne peuvent être des distracteurs.

La majorité des non-distracteurs du corpus qcmNonEN de meilleur rang que les distracteurs sont clairement des non-distracteurs mais certaines mesures (particulièrement celles fondées sur WordNet) considèrent que ces non-distracteurs sont sémantiquement plus voisins que les distracteurs. Parmi les non-distracteurs restants, certains d'entre eux ne sont pas sémantiquement proche de la réponse dans le contexte courant (document de référence) ou sont assez pertinents pour remplacer les distracteurs.

Dans le cas de l'évaluation evalNDoptions, le tableau 3 montre que les mesures individuelles donnent une précision plus faible que pour l'évaluation evalNDdocument. Cela est dû à deux causes principales. Premièrement, cette évaluation extrait plus de non-distracteurs que l'évaluation evalNDdocument, donc les mesures de faible couverture et/ou fondées sur les types donnent un très fort rappel et une très faible précision. Deuxièmement, un grand nombre de non-distracteurs sont sémantiquement plus proches de la réponse que les distracteurs, mais n'ont pas été sélectionnés manuellement car ils n'apparaissent pas dans le contexte de la question, soit le document de référence. En revanche, quel que soit l'évaluation, le modèle d'ordonnancement donne les mêmes résultats pour le corpus qcmEN, contrairement au corpus qcmNonEN où l'évaluation evalNDdocument donne de meilleurs résultats.

Les résultats montrent que l'appartenance au document de référence est un critère important pour ordonner les entités

	qcmEN			qcmNonEN		
	<i>R</i>	<i>P</i>	<i>F</i>	<i>R</i>	<i>P</i>	<i>F</i>
<i>meme_type_EN</i>	0,83	0,03	0,05			
<i>wup</i> sur les types de DBpédia	0,51	0,07	0,13	0,89	0,04	0,07
<i>simREG</i>	0,58	0,10	0,17	0,32	0,15	0,20
<i>simLCH</i>	0,65	0,08	0,15	0,36	0,13	0,19
<i>simJCN</i>	0,73	0,05	0,09	0,37	0,12	0,18
<i>simLin</i>	0,72	0,05	0,10	0,36	0,11	0,17
Comparaison des liens de Wikipédia	0,34	0,24	0,28	0,67	0,15	0,28
ESA	0,27	0,21	0,23	0,30	0,18	0,22
Modèle d'ordonnancement	0,43	0,42	0,42	0,24	0,22	0,22

TABLE 3 – Résultats des méthodes de voisinage sémantique avec l'évaluation evalNDOptions

nommées, contrairement aux chunks non entité nommée. En effet, un grand nombre de candidats sont des mots ou des n-grammes «communs» qui se retrouvent dans plusieurs documents comme le mot «track», ce qui fait que le critère d'appartenance à un document n'améliore pas le modèle d'apprentissage au niveau des chunks non entité nommée.

6 Conclusion

Dans cet article, nous avons proposé une méthode d'estimation de l'homogénéité sémantique fondée sur la combinaison par apprentissage de plusieurs mesures de voisinage et de similarité sémantique. Dans le cadre d'application à la validation automatique de QCM, nous obtenons des résultats supérieurs aux méthodes de l'état de l'art. Les mesures fondées sur la similarité du type des options permettent de donner une indication sur la similarité de leurs catégories sémantiques et les mesures fondées sur les relations sémantiques entre les termes ainsi que les mesures de voisinage distributionnel permettent d'affiner la reconnaissance de l'homogénéité sémantique. Dans des travaux futurs, nous souhaitons adapter notre approche à tout type de réponse.

7 Remerciements

Ce travail a été financé par Digiteo dans le cadre du projet Aneth.

Références

- BANERJEE S. & PEDERSEN T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI*, volume 3, p. 805–810.
- CILIBRASI R. L. & VITANYI P. M. (2007). The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, **19**(3), 370–383.
- DAIBER J., JAKOB M., HOKAMP C. & MENDES P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, p. 121–124 : ACM.
- FINKEL J. R., GRENNAGER T. & MANNING C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, p. 363–370 : Association for Computational Linguistics.
- GABRILOVICH E. & MARKOVITCH S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, p. 1606–1611.
- HALADYNA T. M., DOWNING S. M. & RODRIGUEZ M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, **15**(3), 309–333.
- JIANG J. J. & CONRATH D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

- KARAMANIS N., HA L. A. & MITKOV R. (2006). Generating multiple-choice test items from medical text : A pilot study. In *Proceedings of the fourth international natural language generation conference*, p. 111–113 : Association for Computational Linguistics.
- KLEIN D. & MANNING C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, p. 423–430 : Association for Computational Linguistics.
- LEACOCK C. & CHODOROW M. (1998). Combining local context and wordnet similarity for word sense identification. *WordNet : An electronic lexical database*, **49**(2), 265–283.
- LEE J. & SENEFF S. (2007). Automatic generation of cloze items for prepositions. In *INTERSPEECH*, p. 2173–2176.
- LEVY R. & ANDREW G. (2006). Tregex and tsurgeon : tools for querying and manipulating tree data structures. In *Proceedings of the fifth international conference on Language Resources and Evaluation*, p. 2231–2234 : Citeseer.
- LIN D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, p. 64–71 : Association for Computational Linguistics.
- MILNE D. & WITTEN I. H. (2013). An open-source toolkit for mining wikipedia. *Artificial Intelligence*, **194**, 222–239.
- MITKOV R. & HA L. A. (2003). Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2*, p. 17–22 : Association for Computational Linguistics.
- MITKOV R., HA L. A., VARGA A. & RELLO L. (2009). Semantic similarity of distractors in multiple-choice tests : extrinsic evaluation. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, p. 49–56 : Association for Computational Linguistics.
- PHO V.-M., ANDRÉ T., LIGOZAT A.-L., GRAU B., ILLOUZ G. & FRANÇOIS T. (2014). Multiple choice question corpus analysis for distractor characterization.
- PONZETTO S. P. & STRUBE M. (2007). Knowledge derived from wikipedia for computing semantic relatedness. *J. Artif. Intell. Res.(JAIR)*, **30**, 181–212.
- RODRIGUEZ M. C. (2005). Three options are optimal for multiple-choice items : A meta-analysis of 80 years of research. *Educational Measurement : Issues and Practice*, **24**(2), 3–13.
- WU Z. & PALMER M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, p. 133–138 : Association for Computational Linguistics.

Extraction automatique de relations sémantiques dans les définitions : approche hybride, construction d'un corpus de relations sémantiques pour le français

Emmanuel Cartier
Université Paris 13 Sorbonne Paris Cité, LIPN UMR 7030, équipe RCLN
emmanuel.cartier@lipn.univ-paris13.fr

Résumé. Cet article présente une expérimentation visant à construire une ressource sémantique pour le français contemporain à partir d'un corpus d'environ un million de définitions tirées de deux ressources lexicographiques (Trésor de la Langue Française, Wiktionary) et d'une ressource encyclopédique (Wikipedia). L'objectif est d'extraire automatiquement dans les définitions différentes relations sémantiques : hyperonymie, synonymie, méronymie, autres relations sémantiques. La méthode suivie combine la précision des patrons lexico-syntaxiques et le rappel des méthodes statistiques, ainsi qu'un traitement inédit de canonisation et de décomposition des énoncés. Après avoir présenté les différentes approches et réalisations existantes, nous détaillons l'architecture du système et présentons les résultats : environ 900 000 relations d'hyperonymie et près de 100 000 relations de synonymie, avec un taux de précision supérieur à 90% sur un échantillon aléatoire de 500 relations. Plus de 2 millions de prédications définitoires ont également été extraites.

Abstract.

Automatic Extraction of Semantic Relations from Definitions : en experiment in French with an Hybrid Approach

This article presents an experiment to extract semantic relations from definitions. It is based on approximately one million definitions from two general dictionaries (Trésor de la Langue Française, French Wiktionary) and from the collaborative Wikipedia. We aim at extracting from these data several semantic relations : hyperonymy, synonymy, meronymy and other semantic relations. The methodological approach combines the precision of lexico-syntactic patterns and the recall of statistical analysis. After a survey of the state-of-the-art methods in this area, we detail our system and give the overall outcomes : about 900 000 hypernymy and 100 000 synonymy relations are extracted with a precision above 90% on a sample of 500 pairs for each relation. About 2 millions of definitory predicates are also extracted.

Mots-clés : relations sémantiques, patrons lexico-syntaxiques, distributionnalisme, prédication, hyperonymie, synonymie, méronymie, définition.

Keywords: semantic relations, lexico-syntactic patterns, distributionnalism, predication, hypernymy, synonymy, meronymy, definition.

1 Introduction

Aujourd'hui, nous pourrions dresser le tableau suivant des recherches de TAL en sémantique : depuis une dizaine d'années, la disponibilité de corpus de plus en plus imposants a permis aux approches distributionnelles de gagner du terrain dans différents domaines, mais sans vision globale de leur potentiel et de leurs limites ; les approches par apprentissage automatique, qui ont pris le pas sur les approches symboliques, ne permettent pas d'en induire un modèle sémantique unifié et sont extrêmement hermétiques à toute interprétation théorique. Les approches symboliques, enfin, ont réduit leurs prétentions initiales pour focaliser sur des moyens d'expressions spécifiques - spécialement les patrons lexico-syntaxiques liés à telle ou telle information linguistique, avec un certain succès ; par ailleurs, différents modèles et réalisations de la structuration sémantique du lexique ont émergé, autour de WordNet, de FrameNet et de ses dérivés, du modèle sens-texte, de la théorie des qualia ou encore autour des projets lexico-encyclopédiques de type BabelNet ou NELL.

Cet article détaille une expérience visant à repérer automatiquement dans un corpus de définitions tirées de deux dictionnaires et d'une encyclopédie collaborative, différentes relations sémantiques. Pour ce faire, nous utiliserons une méthode hybride guidée par une analyse fréquentielle de patrons dénotant l'hyperonymie, l'hyponymie, la méronymie, la synonymie et d'autres relations sémantiques.

En dehors de proposer un corpus de relations sémantiques du français contemporain, cet article cherche aussi à s'interroger sur la structuration sémantique du lexique et sa modélisation, d'une part, et à évaluer les méthodes de

repérage automatique, en choisissant finalement une méthode liant l'expertise linguistique et un calcul statistique sur corpus.

2 Etat de l'art sur le repérage de relations sémantiques

Dans cette section, nous parcourons les différentes approches et travaux utilisés jusqu'ici pour mettre au jour les relations sémantiques entre lexies, en commençant par l'exploitation des ressources sémantiques existantes et des modèles sous-jacents. Nous présentons ensuite les approches basées sur les contextes dénotant des relations sémantiques, puis l'approche distributionnelle.

2.1 Ressources sémantiques et encyclopédiques existantes pour le français

Les dictionnaires du français ont une longue et riche histoire, et il est imaginable d'en faire l'exploitation informatique. Dans le cadre de cette étude, nous n'évoquerons que les ressources existantes liées au français contemporain, accessibles numériquement et librement pour la recherche.

2.1.1 Méthode lexicographique manuelle

Il s'agit de la méthode traditionnelle. Trois ouvrages, initialement sous format papier, sont exploitables pour le français : le Trésor de la Langue Française¹, le Littré² et les différentes versions du dictionnaire de l'Académie Française³. Ces ressources présentent les défauts classiques de ce type d'ouvrage : les informations sémantiques n'y sont pas décrites systématiquement, ni dans un langage formalisé. Par ailleurs, ces ouvrages n'ont pas de mise à jour disponible pour la période la plus contemporaine⁴. On peut tout de même imaginer exploiter pour le TAL les "définitions" de ces ouvrages, dans un objectif de modélisation de ce type d'informations, et pour extraire des informations sémantiques pour le français dans une perspective diachronique.

Deux ressources sémantiques plus récentes ont été spécifiquement développées dans le cadre du TAL. D'une part *Wordnet* (Miller, 1990), qui propose un réseau lexical hiérarchisé sur la base d'un noyau de relations sémantiques (hyponymie, hyponymie, synonymie, antonymie), ensuite étendu à d'autres relations (méronymie principalement). Pour le français, une seule réalisation manuelle, EuroWordnet (Vossen, 1998), réunissant une quinzaine de langues européennes, a été produite sur la base du modèle Wordnet, avec une couverture très limitée et un droit d'accès restreint.

Sur la base du modèle Sens-Texte, plusieurs réalisations lexicographiques ont été produites manuellement, toutes très limitées quantitativement étant donné la richesse descriptive du modèle. Pour le français, le projet RELIEF mené à l'ATILF vise à produire une ressource sémantique centrée sur un noyau de fonctions lexicales, pour un noyau de lexies. (Lux-Pogodalla et al., 2011; Polguère, 2014).

Un autre modèle a été utilisé, rendant compte non plus des relations sémantiques mais des structures argumentales des lexies cibles. Cette ressource, issue des principes de la linguistique cognitive (Fillmore, 1982 par exemple) a donné naissance à FrameNet (Baker et al., 1998), une ressource construite manuellement à partir d'une analyse de corpus, et d'un modèle des rôles argumentaux. Une ressource similaire n'est actuellement pas disponible pour le français, mais un projet est en cours (Candito et al., 2014).

2.1.2 Méthode collaborative

Un second moyen de mettre en place une ressource lexicographique consiste à la construire collaborativement. Nous renvoyons à (Simko et Bielikova, 2014) pour une présentation détaillée de cette approche, qui ne sera pas explorée dans cet article.

2.1.3 Méthode automatique à base de ressources lexicales

Une troisième méthode consiste à "transformer" une ressource existante pour la rendre informatiquement exploitable. Concernant le français, deux ressources ont ainsi été exploitées : d'une part, *WordNet*, d'autre part le *Wiktionnaire*.

¹ <http://atilf.atilf.fr>

² <http://www.littre.org>

³ <http://atilf.atilf.fr/academie9.htm>

⁴ Le TLF a une couverture qui s'arrête dans les années 1970; le Littré est un ouvrage de la fin du XX^{ème} siècle; la huitième édition du dictionnaire de l'académie date de 1932-1935 et la neuvième est en cours de numérisation

WordNet est la ressource sémantique de référence en TAL, même si le modèle adopté, basé sur des relations sémantiques hors contexte, prête à de nombreuses critiques. Les chercheurs ont développé différentes techniques pour construire un *WordNet* français automatiquement. La technique de base consiste à traduire les lexies de WordNet vers la langue cible au moyen de lexiques bilingues (Sagot et Fišer, 2008 ; Mouton et de Chalendar, 2010). La difficulté réside dans les mots polysémiques, puisque les lexiques bilingues ne reprennent pas la nomenclature de WordNet. JAWS (Mouton et de Chalendar, 2010) utilise des modèles syntaxiques distributionnels pour désambiguïser les différents sens des lexies dans WordNet. Un autre système, WOLF (Sagot et Fišer, 2011) effectue la désambiguïssation des polysèmes de WordNet au moyen de corpus parallèles. Depuis, les auteurs ont produit de nombreuses améliorations (Apidianaki et Sagot, 2012 ; Sagot et Fišer, 2012), mais la ressource est encore loin d'être exploitable.

L'exploitation du Wiktionnaire a été faite principalement par (Sajous et al., 2014). Le Wiktionnaire comprend environ 2 millions d'entrées, dont environ 1,4 millions d'entrées lexicales pour 186 000 lemmes. Parmi les nombreuses informations que chaque article peut contenir, figurent dans la très grande majorité des cas des définitions et, de manière sporadique, des relations sémantiques. L'écueil principal concerne le format non systématique de la source, qui rend la ressource difficilement exploitable. Cependant, les auteurs sont parvenus à extraire un certain nombre d'informations phonétiques et morphosyntaxiques utiles. Dans le cadre du présent travail, nous avons développé un programme permettant d'extraire un certain nombre d'informations à valeur sémantique (définitions et relations sémantiques) à partir de cette ressource.

Il faut enfin noter de nombreux travaux concernant la compilation et l'unification de différentes ressources (dictionnaires et encyclopédies) : BabelNet (Navigli et Ponzetto, 2012), Freebase (Bollacker et al., 2008), YAGO (Suchanek et al., 2007) et DBpedia (Auer et al., 2007). Ces outils génèrent des ressources multilingues à grande échelle, qui fournissent malheureusement des résultats encore trop bruités.

2.1.4 Méthodes automatiques à base de corpus textuels

Ces approches, privilégiées étant donné le coût prohibitif de la confection manuelle d'une ressource sémantique, visent à identifier automatiquement des relations sémantiques entre lexies sur gros corpus. Plusieurs techniques ont été utilisées.

Les approches symboliques reposent sur le postulat qu'il existe des propriétés linguistiques internes (par exemple des radicaux communs explicitant des dérivations lexicales) et surtout externes (contextuelles : patrons lexico-syntaxiques) signalant des relations sémantiques. On identifie alors les patrons dénotant les relations cibles, on les formalise, puis on reconnaît sur corpus les occurrences de ces patrons. La première réalisation de cette technique revient à (Hearst, 92), qui a particulièrement travaillé sur la relation d'hyperonymie. Elle indique quelles doivent être les propriétés de ces patrons :

« We identify a set of lexico-syntactic patterns that are easily recognizable, that occur frequently and across text genre boundaries, and that indisputably indicate the lexical relation of interest. » ((Hearst, 92, p.539)

Pour « découvrir » les patrons, elle propose un algorithme d'amorçage (*bootstrapping*) dont nous parlerons plus loin. En français, de nombreux travaux ont été menés par l'équipe de Jean-Pierre Desclès depuis 1990, avec une technique similaire⁵, pour reconnaître des relations aussi diverses que les relations causales (Garcia, 1998), la méronymie (Jackiewicz, 1999), les relations temporelles et aspectuelles (Battisteli, 2009) ou les expressions définitives (Cartier, 2004). (Morin et Jacquemin, 2004) ont également repris les travaux de Hearst, en automatisant la découverte des patrons, pour mettre en place un système de repérage de relations sémantiques. Cette technique a les inconvénients suivants : difficulté à expliciter des relations « génériques », ambiguïté de certains patrons, coût de développement.

Les méthodes semi-supervisées utilisent des corpus annotés pour apprendre, dans le même esprit mais de manière automatique des propriétés linguistiques signalant les relations visées. Les algorithmes d'apprentissage sont variés : Machine à Vecteur de Support (SVM) (Zhao and Grishman, 2005; Bunescu and Mooney, 2006), régression logistique (Kambhatla, 2004), parsing augmenté (Miller et al, 2000), Champs Conditionnels Aléatoires (CRF) (Culotta et al, 2006). Ces méthodes présentent deux défauts principaux : temps de développement des corpus d'apprentissage, problème d'adaptabilité à de nouveaux domaines. Voir (Bach et Badaskar, 2007) pour une présentation détaillée de ces méthodes.

Amorçage : (Hearst, 1992) en présentait déjà le fonctionnement générique : on identifie plusieurs termes représentatifs de la relation sémantique visée (appelés *seeds*), on repère en corpus les occurrences des termes (proximité en nombre de mots, ou en terme de distance syntaxique), puis on récupère les patrons correspondants - candidats patrons pour la relation étudiée-, le processus étant itératif, jusqu'au point d'arrêt (nombre de patrons atteint, découverte de patrons épuisée, etc.). Sur cette base de nombreux travaux ont été menés (par exemple Pantel et Pennachioti, 2006 ; Kozareva, Riloff, and Hovy 2008 ; 2010).

⁵ Similaire seulement, car la méthode d'exploration contextuelle comprend deux temps : le premier consiste à marquer dans le texte des indicateurs de relations sémantiques, puis le second applique des patrons lexico-syntaxiques pour repérer les arguments de la relations sémantique. Voir (Desclès, 2006)

Ce type de système a l'avantage de ne pas nécessiter une intervention humaine prohibitive, mais il présente un inconvénient majeur, un faible rappel. En effet, certains patrons peuvent être ambigus, un patron ambigu récupérant de mauvais exemplaires, qui vont à leur tour récupérer de mauvais patrons. A notre connaissance, aucune méthode n'a encore été trouvée pour gérer ces glissements de sens.

Utilisation des contextes définitoires : une autre voie pour accéder aux relations sémantiques lexicales consiste à utiliser les contextes définitoires dans les textes. A l'évidence, en corpus, ces fragments textuels sont quantitativement moins importants que les patrons lexico-syntaxiques exprimant telle ou telle relation sémantique, mais on peut penser que les informations y sont plus fiables.

Plusieurs travaux ont été menés dans cette direction, dans différentes langues et sur différents corpus. (Pearson, 1998) décrit les différents propriétés linguistiques des contextes définitoires en anglais. (Meyer, 2001) fait de même dans un contexte terminologique. (Storror et Wellinchoff, 2006 ; Walter et Pinkal, 2006) ont travaillé sur l'allemand, (Pinto et Oliveira, 2004) sur le portugais, (Leu et Ko, 2007) sur le chinois. Plusieurs systèmes ont été mis en place : Definder (Klavans et Muresan, 2001) sur les textes médicaux anglais, (Alarcón et al., 2008, 2009) sur l'espagnol. (Navigli et al., 2007, 2010) ont développé *GlossExtractor*. Les auteurs proposent de travailler à partir des *glosses*⁶ de Wikipedia. Ce projet a donné lieu à des thésaurus en cinq langues. (Cartier, 2004) a étudié la structure lexico-syntaxique des définitions en français en partant d'un corpus issu de l'Encyclopédie Universalis et d'articles techniques et scientifiques. Des travaux ont été menés pour extraire automatiquement du TLF des relations hyperonymiques à partir des définitions (Barque et al., 2010). (Rebeyrolle, 2000) a également travaillé sur les contextes définitoires.

La plupart des travaux élaborent manuellement les patrons lexico-syntaxiques signalant des contextes définitoires. Mais certains auteurs automatisent le processus. Par exemple (Navigli, 2010) utilisent les Word Classes Lattices. (Serra, 2009) et (Cartier, 2004) insistent sur la dispersion des contextes définitoires et le continuum qui existe entre une expression définitoire proprement dite, un contexte propre à telle ou telle relation sémantique, et même toute mention du terme défini. Au final, les deux auteurs considèrent que « toute occurrence d'une lexie est de facto définitoire », reprenant la notion de *définition implicite* établie par (Gergonne, 1818).

Méthodes distributionnelles : depuis les années 90, le paradigme statistique, issu des intuitions du distributionnalisme (Harris, 1954) puis de la linguistique de corpus (Firth, 1957; Miller et Charles, 1991), domine les travaux en Traitement Automatique des Langues (TAL), avec des réalisations convaincantes dans tous les domaines : reconnaissance automatique des unités (poly-)lexicales, des parties du discours, des relations sémantiques, modèles probabilistes du langage, etc. Ces études ont mis à jour des phénomènes linguistiques au travers des notions de collocations, de « multiword expressions », de « word sketches » (Kilgariff et al., 2004) ou encore de « similarité sémantique ».

Ces différentes approches se basent sur l'hypothèse que le meilleur moyen d'accéder au fonctionnement des langues est de se baser sur les traces matérielles, et la répétition des séquences constitutives. Plusieurs hypothèses « sémantiques » en sont dérivées (Turney et Pantel, 2010 ; Baroni et Lenci, 2010 ; Clark, 2015) : tout d'abord, le calcul des répétitions de séquences permet de mettre au jour les préférences sélectives, et donc l'usage, des lexies. Par exemple (Kilgariff, 2004) déduit, sur cette base, les structures argumentales les plus fréquentes des verbes, qu'il appelle *Word Sketches*. (Hanks, 2013) théorise cette approche avec la *Théorie des Normes et des Exploitations*, les normes (usages) pouvant être décrites par les constructions les plus fréquentes utilisées avec une lexie donnée, et les exploitations (ruptures ou évolutions d'usage) par des constructions attestées moins fréquentes. Une autre hypothèse dérivée se résume ainsi : deux lexies partageant un grand nombre de contextes sont « similaires » sémantiquement (Harris, 1954). C'est ce flou dans l'identification exacte de la relation sémantique, et l'absence actuelle de mesure pour les distinguer qui rend la notion de « similarité » difficilement opérationnalisable.

Cependant, il est évident que le calcul des répétitions nous informe *en quelque manière* sur le sens des lexies, et la répétition de séquences dénotant des relations sémantiques nous informe également sur les patrons les plus employés. Dans ce cadre nous utiliserons l'outil SDMC (Béchet et al., 2013), outil venu de la fouille de données, qui permet de repérer les séquences répétitives en utilisant une combinaison de traits (forme linguistique, lemme et partie du discours) pour découvrir les patrons lexico-syntaxiques les plus fréquents. Nous utiliserons cet outil pour guider la recherche manuelle des patrons dénotant des relations sémantiques dans les définitions.

L'approche suivie dans l'expérimentation qui suit se positionne donc dans la perspective de la découverte de patrons lexico-syntaxiques, étant donné la précision qu'ils permettent d'obtenir. Pour éviter l'écueil de la dispersion des patrons sur corpus, nous focaliserons sur des énoncés privilégiés, les définitions. Pour contourner la difficulté liée au temps de développement prohibitif des ressources, nous utiliserons SDMC, un outil permettant d'extraire les patrons lexico-syntaxiques les plus fréquents sur corpus.

⁶ les premières phrases de chaque article dont l'objectif éditorial est d'expliciter les caractéristiques essentielles du concept décrit

3 Expérimentation : extraction automatique des relations sémantiques dans les définitions

Nous présentons ci-après l'expérimentation menée pour repérer des relations sémantiques dans les définitions en français. Après avoir brièvement présenté le modèle définitoire retenu, nous détaillons le corpus, l'architecture du système et les différents processus constitutifs. Nous terminons par la présentation des résultats et leur évaluation.

3.1 Modélisation de la définition

La notion de *définition* a été étudiée par la philosophie, la logique, la linguistique et la psychologie. Nous partons ici du modèle « classique » de la définition, représenté par (Arnauld et Nicole, 1662). Dans cette conception, qui a impacté et impacte fortement la lexicographie, une définition (de mot) est une proposition dénotant les propriétés essentielles d'une unité lexicale. On appelle *definiendum* (DFN) l'unité définie, et *definiens* (DFS) le segment textuel définissant. Appelons également *relateur définitoire* (DFR) la séquence linguistique mettant en relation d'identification les deux éléments précédents. Par exemple, en partant de :

(1) *Le chat domestique (Felis silvestris catus) est un mammifère carnivore de la famille des félidés. Il est l'un des principaux animaux de compagnie et compte aujourd'hui une cinquantaine de races différentes reconnues par les instances de certification. (Wikipedia)*

nous aboutissons à :

DFN : Chat domestique DFR : .. est un... DFS : mammifère carnivore de la famille des félidés

DFN : Chat domestique DFR : ...est un des principaux... DFS : animaux de compagnie

Ici, deux définitions « classiques » peuvent être identifiées, avec deux relateurs différents. Le dernier segment (*et compte... certification.*) peut être caractérisé comme une prédication *essentielle*, étant donné le contexte définitoire. Nous appellerons ce type de segment *prédication définitoire*. Le segment nominal entre parenthèses (*Felis silvestris catus*) est un synonyme du terme défini, dont le relateur est complexe : DFN (SYN).

Le *definiens* se compose, dans la vision classique de la définition, de deux parties : un hyperonyme et des différences spécifiques. Ce modèle utilise les notions problématiques et floues de « différences » (ou encore propriétés) et « spécifiques », que nous tenterons de préciser après l'expérimentation. Ce modèle classique, aristotélien, du *definiens*, a été étendu à d'autres informations définitoires : définition synonymique, méronymique, causale, téléique, etc. Voir par exemple (Martin, 1992) pour une typologie des définitions lexicographiques.

L'énoncé définitoire, comme toute énoncé, comprend également une prise en charge énonciative et une modalisation implicites ou explicites : *Selon Sisley, un Composant G est un... / XXX considère que le composant G est un...* Dans le corpus qui est le nôtre, ces prises en charge énonciatives resteront implicites.

Enfin, une restriction de domaine peut limiter la portée de la définition : *En astrophysique, on appelle XXX ...*

3.2 Corpus

Les définitions se rencontrent principalement dans les dictionnaires et les encyclopédies. Le corpus utilisé devra également être dans le domaine public, être disponible sous format électronique, et décrire le vocabulaire général du français ; enfin, étant donné que cette expérimentation vise à mettre en place un prototype d'extracteur, nous avons restreint notre corpus aux définitions de noms.

Les trois ressources suivantes seront donc utilisées :

- **Trésor de la Langue Française informatisé (TLFI)** : le laboratoire ATILF a cordialement accepté de nous fournir une version électronique du TLFI, qui comprend 61 234 unités lexicales nominales pour un total de 90 348 définitions ;
- **Version française du Wiktionnaire (FRWIK)** : nous avons mis au point un utilitaire permettant d'en extraire les termes définis, la partie du discours, ainsi que les différentes définitions. Au total, ce corpus est constitué de 140 784 noms, pour un total de 187 041 définitions ;
- **Version française de Wikipedia (WIKP)** : nous avons également utilisé Wikipedia, même s'il s'agit d'une ressource de type « encyclopédie collaborative » ; mais d'autres chercheurs (Navigli and al, 2008) ont montré l'intérêt des premières phrases de chaque article, qui « définissent » les aspects essentiels du concept décrit. Nous avons donc extrait de cette ressource 610 013 entrées nominales⁷ et la première phrase de chaque article.

3.3 Architecture du système

Le système mis en place comprend cinq étapes de traitement :

1. Nettoyage du corpus : il s'agit ici de transformer le fichier de départ en un fichier texte propre comportant une définition par ligne ; cette étape a aussi consisté à convertir le wiktionnaire de son format web à un format

⁷ A noter que (Navigli et al. 2008) n'utilisent que 410 000 définitions car ils opèrent un filtrage sur la présence du marqueur définitoire *être un*.

- exploitable, en extrayant les seules définitions ; il s'agit également de préparer le corpus pour l'étape suivante, notamment par segmentation en mots en en phrases ;
2. Analyse morphosyntaxique du corpus : elle a été conduite avec TreeTagger (Schmid, 1994);
 3. Identification des marqueurs définitoires : il s'agit de marquer dans le texte les lexies marquant une relation sémantique donnée ;
 4. Simplification/ canonisation des phrases : cette étape vise à simplifier les énoncés définitoires pour faciliter l'application ultérieure des patrons;
 5. Extraction des relations sémantiques au moyen de patrons lexico-syntaxiques.

Dans cet article, nous ne détaillerons pas les deux premières étapes, nous concentrant sur les trois dernières.

3.3.1 Identification des marqueurs définitoires

Cette étape consiste à identifier et marquer dans la phrase source les unités lexicales (simples ou composées) permettant d'identifier un énoncé définitoire ou l'un de ses composants. Reprenant des travaux précédents (Rebeyrolle, 2000 ; Cartier, 2004), nous avons utilisé la liste de marqueurs suivants :

Type de marqueur	Échantillon de marqueurs	Nombre total
Relateur définitoire 1 (REL1)	<i>être, ":", "c'est-à-dire, ou...</i>	11
Relateur définitoire 2 (REL2)	<i>se définir comme, vouloir dire, s'appeler, signifier...</i>	34
Marqueur de catégorisation (CAT)	<i>catégorie, famille, espèce...</i>	47
Marqueur de spécification (SPEC)	<i>exemple, exemplaire, prototype...</i>	23
Marqueur de propriété (PROP)	<i>se caractériser, avoir pour caractéristique...</i>	42

TABLEAU 1 : liste des marqueurs linguistiques de la définition en français

Le nombre de marqueurs est relativement bas. Comme nous le verrons dans la suite, leur couverture est maximale. Il faut néanmoins noter que la mise en place de cette liste nécessite un travail de dépouillement et une expertise non négligeables.

3.3.2 Simplification des phrases

La simplification/canonisation des phrases est une étape innovante dans le cadre des travaux sur l'extraction d'information à base de patrons lexico-syntaxiques. Son objectif principal est de réduire les énoncés à une forme « canonique », en repérant et supprimant les éléments accessoires et en ne conservant que les éléments essentiels au modèle définitoire. Cet objectif est réalisé en trois étapes :

1. Identification et suppression des éléments circonstanciels et accessoires des phrases, pour autant qu'ils ne font pas partie d'une partie constitutive du modèle définitoire adopté ;
2. Identification, *extraction* et suppression des informations définitoires périphériques : restrictions de domaine, marqueurs énonciatifs, relations sémantiques « annexes » au prédicat définitoire principal : ces informations sont conservées puis retirées de l'énoncé définitoire à analyser ;
3. Unification des groupes nominaux, puisqu'ils sont dans le cadre de cette expérimentation, les éléments cibles de la partie définissante, et que leur variété est un facteur de perte de précision pour les étapes suivantes.

Prenons un exemple de ces différents traitements. Soit la définition de *abaisse* dans le Wiktionnaire, après analyse morphosyntaxique et identification des marqueurs définitoires :

(2) *en/P cuisine/NC et/CC en/P pâtisserie/NC ./PONCT un/DET DEFINIENDUM être/V/DEF_REL un/DET pièce/NC de/P pâte/NC aplatis/VPP ./PONCT généralement/ADV au/P+D rouleau/NC à/P pâtisserie/NC ou/CC un/DET laminoir/NC ./PONCT*

Elle sera réduite à la séquence suivante :

un/DET DEFINIENDUM être/V un/DET pièce/NC de/P pâte/NC aplatis/VPP ./PONCT au/P+D rouleau/NC à/P pâtisserie/NC ou/CC un/DET laminoir/NC ./PONCT

Et nous récupérons la restriction de domaine : *en/P cuisine/NC et/CC en/P pâtisserie/NC*.

La suppression concerne les adverbiaux, les compléments circonstanciels et les propositions subordonnées circonstancielles, qui sont des informations accessoires. La difficulté consiste à distinguer entre les informations circonstancielles portant sur l'énoncé dans son entier, qui sont la cible de ce processus, et les informations circonstancielles qui sont liées à l'un des composants de la définition, et qu'il faut conserver.

Par exemple, nous souhaitons supprimer la parenthèse dans :

(3) *DEFINIENDUM (/PONCT proche/ADJ de/PREP la/DET capitale/NC)/PONCT être/V un/DET atoll/NC de/P le/DET république/NC du/P+D Kiribati/NPP ./PONCT*

Mais pas la proposition relative dépendante du *definiens* :

DEFINIENDUM être/V du/P+ enzymes/NC qui/PROREL contrôler/V le/DET structure/NC topologique/ADJ de/P l'ADN/NC...

Pour identifier les structures lexico-syntaxiques correspondant aux différents éléments à supprimer, nous avons calculé avec SDMC les patrons lexico-syntaxiques les plus fréquents en deux positions : au début des énoncés (c'est-à-dire avant le *definiendum*) et entre le terme défini et le relateur définitoire. La zone du *definiens* comprend généralement des structures définissantes qui ne sont pas touchées par le processus.

Trois cas méritent mention :

- Dans Wikipedia, la grande majorité des éléments qui s'insèrent entre le terme défini et le relateur définitoire explicitent une relation sémantique synonymique : *DEFINIENDUM (/PONCT parfois/ADV Apaiang/NPP ,/PONCT même/ADJ prononciation/NC)/PONCT être/V/DEF_REL un/DET...*
- Les morphèmes adverbiaux de négation doivent être conservés ;
- Un certain nombre de compléments circonstanciels dénotent une restriction de domaine, notamment lorsqu'ils sont placés en tête d'énoncé. Pour les repérer et enregistrer leur contenu avant de les retirer de l'énoncé initial, nous avons modélisé sous forme d'expressions régulières, à partir des résultats de SDMC les patrons les plus fréquents :

$$\wedge((?:en/dans/\grave{a}/sur/selon/pour/chez/par).\{5,150\}?)t, \vee/PONCT/t/$$

$$DEFINIENDUM/t, \vee/PONCT/t((?:en/dans/\grave{a}/sur/selon/pour/chez/par).\{5,150\}?)t, \vee/PONCT\}$$

3.3.3 Unification/réduction de la variété des syntagmes nominaux

Intuitivement, la grande majorité des définitions obéit au modèle hyperonymique, où le *definiens* est composé d'un hyperonyme suivi des différences spécifiques exprimées sous forme adjectivale (doté ou non de ses expansions) et/ou de proposition relative. Le premier élément du *definiens* est donc l'hyperonyme de la lexie définie. Mais plusieurs phénomènes complexifient l'identification de l'hyperonyme nominal et principalement : les déterminants composés (*est une des ... ; est un ensemble de ...*), des adjectifs axiologiques (*est le principal moyen de....*), des termes marquant une relation spécifique (*est un genre de ...*).

Pour traiter ces exceptions, nous avons listé la plupart des déterminants composés qui sont pris en compte lors de la segmentation en mots, et les marqueurs de relation sémantique sont identifiés en tant que tels afin d'éviter d'être les cibles de la relation sémantique à extraire.

Ensuite, les différentes formes de noms simples ou composés sont unifiées. Pour cela, en nous basant sur un certain nombre d'études (Ramish, 2015 ; Mathieu-Colas, 1996), nous reconnaissons les formes les plus fréquentes en français (ici sous forme d'expressions régulières):

NC (ADJ) ? de/P NC (ADJ) ?
NC (ADJ){0,3}
NPP+

Ce traitement est important car, d'une part, il facilite l'extraction des cibles argumentales des relations sémantiques définitoires principales (hyperonymie ou synonymie), et, d'autre part, il rend plus efficace l'extraction ultérieure des autres relations sémantiques.

3.3.4 Identification des relations sémantiques par patrons lexico-syntaxiques

La mise au point des patrons lexico-syntaxiques est la dernière étape, permettant de repérer automatiquement les relations sémantiques suivantes : hyperonymie, synonymie, méronymie, et prédications définitoires. Les traitements précédents ont fortement simplifié cette reconnaissance.

La mise au point des patrons a combiné l'expertise linguistique et l'analyse fréquentielle du corpus par SDMC. Cet outil est lancé sur les séquences définitoires issues des traitements précédents, éventuellement tronquées (voir plus loin). La recherche de séquences répétées a été faite en utilisant le paramétrage suivant⁸ : recherche d'un maximum de cinq cooccurents, fenêtre de quinze mots, combinant les traits: forme graphique, lemme, partie du discours, avec présence d'un marqueur de la relation en question. Le calcul statistique nous a permis d'aboutir pour chaque corpus à un noyau de patrons, correspondant à chaque type de relations sémantiques. L'expertise linguistique, à partir des résultats fréquentiels bruts, a permis d'une part de ne retenir que les patrons les plus fréquents et les moins ambigus (par sondage), puis, parmi les patrons les moins fréquents, de ne retenir que les patrons non ambigus pour la relation sémantique considérée.

La recherche fréquentielle de patrons est précédée de réductions de la phrase source selon la relation sémantique visée. Pour ce qui concerne les relations d'hyperonymie, de synonymie et de méronymie, pour s'assurer que la relation liait le *definiendum* à un syntagme nominal, le calcul a été effectué en tronquant l'énoncé définitoire sur la droite au niveau de

⁸ Le choix des paramètres n'a pas de justification objective : il s'agit de paramètres maximisant la couverture des patrons repérés. Nous renvoyons à (Kiela et Clark, 2014) pour une évaluation des paramètres optimaux, notamment fenêtre de recherche et nombre de cooccurents., qui recoupe les intuitions choisies ici.

la séquence définissante, à la première occurrence d'un pronom relatif, d'un adjectif suivi de ses arguments ou d'un complément nominal doté d'un déterminant (DEFINIENDUM est un GN | PROPREL ; DEFINIENDUM est un GN | ADJ PREP ; DEFINIENDUM est un GN | PREP DET ...). Ces éléments signalent en effet le début d'une autre information définitoire liée à la prédication, et fournissent une frontière droite aux groupes nominaux à rechercher.

Pour ce qui concerne la relation de prédication, la recherche statistique s'est faite après la reconnaissance des autres relations sémantiques, permettant ainsi de cibler la recherche sur le reste de la définition. Par exemple, pour le définiendum *Baglage*, à partir de la phrase source :

(4) DEFINIENDUM :/PONCT redevance/NC au/P+D officier/NC préposer/VPP à/P
ce/PRO arrangement/NC ./PONCT

Nous obtenons, après marquage de la relation d'hyperonymie :

DEFINIENDUM :/PONCT/DEF_REL GN(redevance)/HYPER au/P+D GN(officier/NC) préposer/VPP
à/P ce/PRO GN(arrangement/NC) ./PONCT

Ce qui permet de faire une recherche seulement dans ce qui complète l'hyperonyme, où sont explicitées les différences spécifiques de la lexie. La recherche des patrons de prédications définitoires ne s'appuie sur aucun marqueur spécifique, et aboutit à identifier trois patrons très génériques correspondant aux expansions syntaxiques d'un nom en français : complément prépositionnel de nom ((1) *de la famille des félidés* ; (4) : *à l'officier*) groupe adjectival suivi ou non de ses expansions nominales ou verbales ((1) *carnivore* ; (5) ... *affecté au transport urbain*), proposition relative. phrase coordonnée sans marqueur ((1) ... *et compte aujourd'hui...*). Le coordination d'éléments a été partiellement modélisée.

3.4 Résultats et analyse

Nous présentons ci-après les résultats globaux puis les résultats par relation sémantique.

3.4.1 Résultats globaux

Le tableau 2 présente les totaux de repérage pour chaque relation sémantique et pour chaque corpus.

	Wiktionnaire (WIKT) : 186 502 lexies-définitions				TLF : 90 190 lexies-définitions				Wikipedia (WIKIP) : 610 013 lexies-définitions			
	Nbre	Moyenne par déf.	Nbre de déf. sans extraction	% sans relation	Nbre	Moyenne par déf.	Nbre de déf. sans extraction	% sans relation	Nbre	Moyenne par déf.	Nbre de déf. sans extraction	% sans relation
hyperonymie	187934	1,02	2435	1,31%	90605	1,03	2364	2,62%	592311	1,03	35459	5,81%
synonymie	785	1,53	185988	99,72%	376	1,2	89876	99,65%	83276	1,24	542911	89,00%
méronymie	1313	1	185189	99,30%	1733	1,11	88627	98,27%	0	0	610013	100,00%
prédications	469325	2,55	2435	1,31%	232300	2,65	2364	2,62%	1398178	2,43	35459	5,81%
domaine	9413	1,02	177245	95,04%	5779	1,43	86143	95,51%	31834	1,05	579598	95,01%

TABLEAU 2 : résultats globaux d'extraction des relations sémantiques

Ces chiffres appellent plusieurs commentaires :

- la relation d'hyperonymie est la plus représentée dans le corpus, de très loin devant la synonymie et la méronymie ; on note tout de même une plus grande représentation de la synonymie dans Wikipedia, sachant que dans un certain nombre de cas, les gloses comprennent à la fois un synonyme et un ou des hyperonymes ; ce constat est à rapprocher d'un modèle définitoire comprenant dans la quasi totalité des cas un hyperonyme et des propriétés spécifiques ;
- la moyenne d'extraction d'une relation donnée par définition est intéressante, puisqu'on note que, tandis que l'hyperonymie, la méronymie, et les domaines sont proches d'une instance par définition, les synonymes ont une tendance à se multiplier dans une même définition (1,53 pour WIKT) ; les prédications sont évidemment multiples dans chaque définition, avec une moyenne de 2,5 prédications extraites, le TLF ayant la plus grande diversité ;
- les restrictions de domaine sont présentes de manière stable entre les corpus : environ 5% des définitions comprennent cette information⁹ ;
- Parmi les définitions pour lesquelles aucune extraction d'hyperonyme n'a eu lieu (colonnes 3 et 4 pour chaque ressource), les causes en sont diverses :

- absence du patron qui aurait permis le repérage : *abondement/NC :/PONCT/DEF_REL lorsque/CC du/P+D GN(salarié/NC) acheter/V du/P+D GN(action/NC) de/P leur/DET GN(proprie/ADJ société/NC) ./PONCT le/DET GN(abondement/NC) correspondre/V/DEF_REL au/P+D GN(versement/NC complémentaire/ADJ) verser/VPP par/P le/DET GN(société/NC) ./PONCT*

⁹ A noter que le TLF explicite cette information dans un champ spécifique différent de la définition

- énoncé définitoire incomplet : *absence/NC* :/PONCT/DEF_REL *se/PRO* *employer/V/DEF_REL*
absolument/ADV ./PONCT
- énoncé définitoire trop complexe pour être analysé : *alloux/NC* :/PONCT/DEF_REL *en/P* *GN(droit/NC foncier/ADJ)*
allodial/ADJ *un/DET* *GN(alloux/NC)* *être/V/DEF_REL* *le/DET* *GN(inverse/NC)* *de/P* *un/DET* *GN(fief/NC)* *en/P* *GN(droit/NC*
féodal/ADJ) ./PONCT
- erreur d'analyse morphosyntaxique : *acide/NC* :/PONCT/DEF_REL *liquider/VS* *capable/ADJ* *de/P* *attaquer/VINF* *et/CC*
de/P *dissoudre/VINF* *le/DET* *GN(métal/NC)* ./PONCT *certain/PRO* *GN(roche/NC)* ./PONCT

3.4.2 Hyperonymie

Les résultats d'extraction par patrons sont détaillés dans le tableau 3.

On constate :

- La dispersion des patrons est plus importante dans le corpus Wikipedia, ce qui s'explique par le caractère collaboratif et moins normé de cette ressource.
- Dans les corpus dictionnaires, en additionnant les patrons dérivés (par exemple, la coordination de groupes nominaux pour le patron générique *DEF être un GN*, ou encore la formulation à pronom relatif présentatif : *celui qui/ce qui...*), un seul patron couvre plus de 90% des occurrences : le patron *DEF être un/le GN/HYPER*. Seulement trois patrons se partagent la totalité des occurrences. En plus du précédent : *DEF : genre de GN/HYPER*, *DEF : nom de GN/HYPER*.
- Le corpus encyclopédique a une bien plus grande dispersion d'expressions : le patron avec *être* représente un peu plus de 50% des cas, et on rencontre également le patron *nom de GN*. La particularité de ce corpus ressortit à l'utilisation massive d'une hyperonymie par apposition et parenthésage après le terme défini, qui représente environ 35% des cas. Mais, de manière globale, on constate que le nombre de patrons est limité.

Patron générique	Patron	Nb	%	Nb	%	Nb	%
		TLF		WIKT		WIKP	
DEF_ :_DEF_REL_(DET) ?_GN/HYPER	DEF_ :_(le) ?_GN	62969	78,45%	93676	80,67%	570	
	DEF_ :_(le) ?_GN_ _GN_et_GN	573	0,71%	526	0,45%		
	DEF_ :_(le)_GN_et_GN	5601	6,98%	4858	4,18%		
	DEF_ :_(tout)_CE_PROREL	1290	1,61%	800	0,69%		
	DEF_ :_*_DEF_REL_*_GN	3966	4,94%	5087	4,38%		
	DEF_ :_CELUI_PROREL	1221	1,52%	2945	2,54%		
	DEF_ :_PROREL	1938	2,41%	170	0,15%		
	DEF_REL/DEF_ _GN_ ,	225	0,28%	224	0,19%	61099	
	DEF_(ou) ?_GN_)					55638	
	DEF_(_GN_)					75219	
	DEF_ _ (ou) ?_GN_ , ?					24804	
	DEF_ _GN					3744	
	GN_(de)_DEF					29668	
	GN_OU_DEF					1385	
	Être_..._GN					26668	
	être_un_GN					277803	
	être_un_GN_et_un_GN					28861	
	être_un_GN_ _un_GN_et_un_GN					8162	
	Être_un_ADJ_GN					6322	
	Être_un_ADJ_et_ADJ_GN					5	
(DEF_ :_) ?_*_(genre, espèce...) *_GN/HYPER	DEF_ :_genre_de_GN	700	0,87%	2432	2,09%		
	être_un_genre_de_GN					14991	
	DEF_ :_genre_de_GN_ _de_GN_et_de_GN	3	0,00%	9	0,01%		
	DEF_ :_genre_de_GN_et_de_GN	40	0,05%	93	0,01%		
	genre_de_(ADJ)_GN	129	0,16%	544	0,47%		
	genre_de_GN_ _de_GN_et_de_GN	2	0,00%	14	0,01%		
	genre_de_GN_et_de_GN	33	0,04%	47	0,04%		
(DEF_ :_) ?_*_(nom,	DEF_REL_*_nom_*_de_GN	326	0,41%			3272	

dénomination...)_*_GN/H YPER							
	DEF_REL_*_nom_donne_*_a_GN	138	0,17%	622	0,54%	274	
	nom_(de)"_GN_"	10	0,01%	27	0,02%		
	nom_*_de_GN	1005	1,25%	1813	1,56%		
	nom_*_donne_*_a_GN	62	0,08%	322	0,28%		
	DEF_REL_*_nom_*_de_GN			1768	1,52%		
signifier	signifier"_GN_"	35	0,04%	136	0,12%		
	signifier"_GN_"_ou_"_GN_"			4	0,00%		
	Total	80266		187934		592311	

TABLEAU 3 : distribution des patrons de l'hyperonymie sur les trois corpus

3.4.3 Méronymie/holonymie

La méronymie est généralement exprimée comme information principale de l'énoncé définitoire, dans un patron générique du type : *DEF est un(e) (partie, ...) de GN*.

L'holonymie est exprimée de manière converse par le patron : *DEF est (composée, constituée de...) GN*.

Il est également possible de rencontrer un troisième patron combinant expression hyperonymique et méronymique : *DEF est un(e) HYPER (composé de) GN...*

Comme montré dans le tableau 2, très peu de relations méronymiques ont été repérées, pourtant, comme l'évaluation le montrera, la précision est très bonne. Une analyse plus fine devra être menée pour recueillir plus de patrons et les affiner.

3.4.4 Synonymie

Cette relation sémantique est l'une des techniques de définition dans le cadre lexicographique, mais exclut le plus souvent une définition par hyperonyme, dans les sources lexicographiques.

Dans Wikipedia, les gloses contiennent très souvent à la fois une définition par hyperonyme et une ou des expressions synonymes, principalement par le biais d'une apposition ou d'une indication entre parenthèses qui suit le terme défini (voir exemple 1).

Dans le TLF et le Wiktionnaire, l'expression synonymique est exclusive de l'expression hyperonymique. Dans ces cas, la définition est une énumération (parfois réduite à un seul membre) de groupes nominaux coordonnés.

3.4.5 Prédicats définitoires

Ce que nous avons appelé *prédicats définitoires* dénotent les différences spécifiques et ont été repérés sur la base des constructions syntaxiques expansions de groupes nominaux. Elles ne sont donc pas typées *a priori* sémantiquement mais sont articulées autour d'un prédicat verbal ou adjectival, qui exprime la relation sémantique. Un nombre important de patrons dénotent la relation sémantique de *fonction* (*servant à, utilisé pour, etc.*). Ces prédictions devront dans un prochain travail être consolidées afin de mettre au jour la structuration sémantique des termes définis.

3.5 Evaluation

En l'absence de ressource de référence, nous avons mis au point un protocole d'évaluation manuelle par trois linguistes. Cinq cent définitions ont été extraites du corpus, et les différentes informations définitoires (hyperonymie, synonymie, méronymie, domaine, prédictions) ont été annotées par trois annotateurs experts. Les instructions étaient les suivantes : à partir des extractions automatiques (type de relation sémantique, lexies mises en relation) et de la définition source (où se trouve la *référence*), indiquer si la relation sémantique est correcte ou erronée ; dans le second cas, il était également demandé aux évaluateurs d'indiquer en commentaire les raisons de l'erreur et, le cas échéant, les bonnes lexies dans la relation. donnée La totalité des annotations divergentes entre annotateurs (trente-deux cas) ont été résolues d'un commun accord. Au total, le corpus de référence, la précision et le rappel des extractions sont les suivants :

Information définitoire	Corpus de référence	Extractions automatiques	Extractions correctes	Précision	Rappel
hyperonymes	512	489	477	0.975	0.931
synonymes	137	123	109	0.886	0.795
méronymes	67	43	35	0.813	0.522
prédications définitoires	976	1012	953	0.941	0.976
domaine	18	16	16	100	0.888

TABLEAU 4 : résultats de l'évaluation manuelle sur un échantillon de 500 définitions

On notera le rappel assez faible pour la relation de méronymie, qui s'explique par des patrons trop imprécis et lacunaires. Les extractions fautives pour les synonymes et les hyperonymes proviennent le plus souvent d'une reconnaissance partielle des groupes nominaux correspondants (*animal* au lieu de *animal de compagnie*, *filles* au lieu de *jeune fille*), ainsi que d'erreurs d'analyse morphosyntaxique. Pour les prédications, le découpage sur la base de patrons génériques génèrent un taux d'erreur minimal qui pourrait encore être amélioré par une analyse en dépendance.

Extractions communes entre dictionnaires : nous avons comparé les extractions entre les différentes ressources pour les mêmes lexies, en partant de l'hypothèse que si des relations sémantiques étaient repérées plusieurs fois dans plusieurs sources, cela avait toute chance de valider la relation extraite. Cette comparaison a été faite pour les relations de synonymie et d'hyperonymie, les autres informations extraites pouvant difficilement être comparées étant donné la variabilité des expressions et l'absence de généralisation effectuée. Les résultats sont présentés dans le tableau 5.

	Total éléments	1 dictionnaire	%	2 dictionnaires	%	3 dictionnaires	%
Lexies	886705	806497	90,95%	33358	3,76%	46850	5,28%
Hyperonymes	870850	781864	89,78%	52802	6,06%	36184	4,16%
Synonymes	84437	79978	94,72%	2600	3,08%	1859	2,20%

TABLEAU 5 : extractions communes entre dictionnaires

On note un recoupement assez faible des informations sémantiques entre dictionnaires. Pourtant, comme l'a montré l'évaluation manuelle, la précision des repérages est importante. Cela signifie que les dictionnaires expriment des relations sémantiques en utilisant soit des variantes (pour la synonymie et l'hyperonymie), soit spécifient des hyperonymes en se plaçant à différents niveaux d'abstraction. De cette évaluation, nous pouvons tirer un corpus de relations sémantiques consolidées de référence, puisque nous avons, en cumulant les éléments communs à deux ou trois dictionnaires, 88 986 (10,22%) hyperonymes communs et 4 459 (5,28%) synonymes communs.

Extractions communes avec d'autres ressources : une autre expérimentation a consisté à comparer les extractions de SemDef avec celles contenues dans Wolf et dans le Wiktionnaire. Ces deux ressources n'ont elles-mêmes pas été validées, mais nous pouvons partir de la même hypothèse que pour l'évaluation précédente. Pour le Wiktionnaire, nous sommes partis de l'extraction présentée dans cet article ; parmi les lexies étiquetées « nom commun »¹⁰, 114004 ont au moins une relation sémantique explicitée ; nous avons conservé celles qui s'apparentent aux deux relations étudiées ici¹¹, soit 38 018 relations de synonymie et 124152 relations d'hyperonymie. Pour Wolf, nous sommes partis de la version 1.0b4 au format XML disponible en ligne. Nous en avons extrait les synsets ayant des réalisations en français de type nominaux (42427) : la relation de synonymie provient des réalisations en français d'un synset (soit 50130), et les relations d'hyperonymie ont été extraites directement (58359). Le tableau 6 détaille les résultats de la comparaison des ressources, en partant de la ressource cumulée SemDef obtenue par l'expérience précédente. On constate que le cumul donne des résultats comparables à ceux obtenus entre dictionnaires.

	Total éléments	1 dictionnaire	%	2 dictionnaires	%	3 dictionnaires	%
Lexies	300895	271799	90,33%	20531	6,82%	8565	2,85%
Hyperonymes	693356	682259	98,40%	10281	1,48%	816	0,12%
Synonymes	145248	136617	94,06%	7957	5,48%	674	0,46%

TABLEAU 6 : extractions communes entre SemDef, Wolf et Wiktionnaire-relations-sémantiques

¹⁰ Nous insistons sur ce point car certains noms communs sont en réalité des noms propres, ou en sont dérivés, comme les gentils.

¹¹ Les relations explicitées comme hyperonymes ou hyponymes ont été assimilées à la relation d'hyperonymie, et les relations suivantes à la synonymie : Synonymes, Quasi-synonymes, Noms_vernaculaires, Variantes, Variantes_dialectales, Variantes_orthographiques, Abréviations, Anciennes_orthographes, Diminutifs, Synonymes_pour_la_définition

4 Conclusions, perspectives

Dans cet article, nous avons présenté une expérimentation visant à extraire automatiquement des relations sémantiques dans des définitions issues de différents corpus. A partir de près d'un million de définitions de noms extraites du TLF, du Wiktionnaire et de Wikipedia, nous avons pu extraire près de 900 000 relations d'hyponymie et près de 100 000 relations de synonymie, avec un taux de précision supérieur à 90% sur un échantillon aléatoire de 500 relations. Nous avons également extrait près de 2 millions de prédictions dont le statut définitoire est avéré étant donné le contexte définitoire, mais qui demandent une étude approfondie. Les différentes ressources seront mises à disposition de la communauté scientifique à l'occasion du colloque TALN, ainsi que les corpus source Wiktionnaire et Wikipedia.

Cette expérimentation a combiné deux approches complémentaires : une approche symbolique centrée sur la notion de patron lexico-syntaxique, ici dénotant une relation sémantique lexicale, avec pour objectif d'en décrire les différentes formes dans le corpus choisi ; une approche statistique basée sur le calcul des répétitions de séquence, combinant différents niveaux (forme graphique, lemme, partie du discours, marque sémantique) qui a permis d'accélérer la mise au point manuelle des patrons. Notre approche a été bonifiée par deux prétraitements : un prétraitement linguistique des énoncés définitoires, afin de les canoniser, en éliminant les informations accessoires à la prédication définitoire principale ; une décomposition des contextes à étudier statistiquement selon l'information recherchée, afin de réduire le bruit de calculs statistiques bruts sur corpus. Cette combinaison de techniques nous semble prometteuse pour tout type de relations sémantiques, même si elle doit être éprouvée sur corpus libre, et étendue à d'autres types d'informations sémantiques.

Du point de vue de la structuration sémantique du lexique, les résultats obtenus montrent que le réseau lexical ne se limite pas aux relations classiques, « classificatoires » (hyponymie-hyponymie, synonymie, antonymie) qui sont le cœur d'un réseau lexical de type Wordnet : les définitions explicitent certes clairement ces relations de catégorisation, mais d'autres relations relient les lexies entre elles : méronymie-holonymie, fonction, et prédictions définitoires qui décrivent différents aspects essentiels. Ces derniers éléments sont peut-être la matière la plus intéressante du corpus étudié ici, car elles expriment des propriétés essentielles des lexies, mais, à ce stade de l'étude, elles ne peuvent pas être catégorisées.

Cette étude nous engage à mener des travaux complémentaires dans différentes directions : d'une part, préparer une évaluation à plus grande échelle, afin de valider les relations sémantiques extraites et fournir à la communauté scientifique un corpus de référence plus conséquent encore. Cette évaluation pourra être conduite à la fois par une validation collaborative en ligne, mais également en étudiant les relations sémantiques sur gros corpus à partir de l'hypothèse distributionnelle.

Remerciements

Merci aux évaluateurs de leurs commentaires et suggestions, qui ont permis de grandement bonifier cet article.

Références

- ALARCÓN R., BACH C. AND SIERRA G. (2008) "Extracción de contextos definatorios en corpus especializados: Hacia una elaboración de una herramienta de ayuda terminográfica". *Revista Española de Lingüística*. Madrid, 2008. 247-278.
- APIDIANAKI M. ET SAGOT B. (2014) Data-driven Synset Induction and Disambiguation for Wordnet Development. *Language Resources and Evaluation Journal*, Springer Netherlands, Vol. 48(4), pp. 655-677.
- ARNAULD A. ET NICOLE P. (1662) *La logique ou l'art de penser*, édition critique par D. Descotes, Paris: Champion, 2011.
- AUER S, BIZER C, KOBILAROV G, LEHMANN J, IVES Z (2007) DBpedia: A nucleus for a web of open data. In: *Proceedings of 6th International Semantic Web Conference*, Springer, Busan, Korea, pp 11-15
- BACH, N., AND BADASKAR S. (2007) "A Review of Relation Extraction," Literature review for Language and Statistics II, 2007
- BAKER, C.F., FILLMORE C.J., AND LOWE J.B. (1998) The Berkeley FrameNet project." COLING-ACL '98: Proceedings of the Conference. Montreal, Canada 1998. 86-90.
- BARONI, M., AND LENCI A. (2010) "Distributional Memory : A General Framework for Corpus-Based Semantics," Computational Linguistics, 36-4 (2010), 50
- BARQUE L., NASR A., POLGUÈRE A. (2010) From the Definitions of the Trésor de la Langue Française to a Semantic Database of the French Language, Proceedings of the 14th EURALEX International Congress, Leewarden.

BATTISTELLI D. (2009) *La temporalité linguistique : circonscrire un objet d'analyse ainsi que des finalités à cette analyse*. HDR, Université de Nanterre - Paris X, 2009.

BÉCHET N., CELLIER P., CHARNOIS T., CRÉMILLEUX B., QUINIOU S. (2013). SDMC : un outil en ligne d'extraction de motifs séquentiels pour la fouille de textes. Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'13), Jan 2013, Toulouse, France.

BOLLACKER K, EVANS C, PARITOSH P, STURGE T, TAYLOR J (2008) Freebase: A collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, ACM, New York, NY, USA, SIGMOD '08, pp 1247–1250

BUNESCU R, MOONEY R (2006) Subsequence kernels for relation extraction. In: Weiss Y, Scholkopf B, Platt J (eds) Advances in Neural Information Processing Systems 18, MIT Press, Cambridge, MA, pp 171–178

CANDITO, M. AMSILI, P., BARQUE, L., BENAMARA, F., CHALENDAR, G., DJEMAA, M., HAAS, P., HUYGHE, R., MATHIEU, Y., MULLER, P., SAGOT, B. & VIEU, L., (2014), Developing a French FrameNet: methodology and first results, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), Reykjavik, Iceland, 2014

CARTIER E. (2004) : Repérage automatique des expressions de finitoyres : modélisation de l'information de finitoyre, méthode d'exploration contextuelle, méthodologie de développement des ressources linguistiques, description des expressions du français contemporain, mise en œuvre informatique, thèse de doctorat, Université Paris IV-Sorbonne.

CLARK, S. (2015) “Vector Space Models of Lexical Meaning,” in *Handbook of Contemporary Semantics*, second edition, ed. by Shalom Lappin and Chris Fox (Wiley-Blackwell, 2015), pp. 1–43

CULOTTA A, MCCALLUM A, BETZ J (2006) Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Association for Computational Linguistics, New York, New York, pp 296–303

DESCLÈS J.-P. (2006) «Contextual Exploration Processing for Discourse Automatic Annotations of Texts», FLAIRS 2006, Melbourne, Floride, 11-13 mai, Invited Speakers, p. 281-284.

FILLMORE, C. J. (1982) Frame semantics. *Linguistics in the Morning Calm*. Seoul, South Korea: Hanshin Publishing Co., 1982. 111-137

FIRTH, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pp. 1–32. Blackwell, Oxford.

GARCIA, D. (1998), Analyse automatique des textes pour l'organisation causale des actions, Réalisation du système Coatis. The sé d'informatique, Université Paris IV.

GERGONNE J. (1818) « Variétés, essai sur la théorie des définitions », *Annales de Mathématiques pures et appliquées*, tome 9 (1818-1819), p.1-35.

HANKS, P. (2013) *Lexical Analysis: Norms and Exploitations*. Cambridge. MIT Press.

HARRIS, Z. (1954). Distributional structure. *Word*, 10(23), 146–162.

HEARST M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the 15th International Conference on Computational Linguistics (COLING 1992), p. 539–545, Nantes.

JACKIEWICZ A. (1999), « La causalité dans les textes », in *Semantyka i konfrontacja jezykowa (Sémantique et Confrontation des Langues)*, Varsovie, Pologne, SOW, Académie des Sciences de la Pologne, 1999, vol.2, pp.147-164

KAMBHATLA N. (2004) Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In: Proceedings of the ACL 2004, Association for Computational Linguistics, Morristown, NJ, USA.

KIELA D. AND CLARK S. (2014) A Systematic Study of Semantic Vector Space Model Parameters, in *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, 2014, pp. 21–30

KILGARRIFF, A., RYCHLY P., SMRZ P., AND TUGWELL D.. (2004). The Sketch Engine. In Proceedings of Euralex, pages 105–116, Lorient.

- KLAVANS J. AND MURESAN S. (2001) "Evaluation of the DEFINDER System for Fully Automatic Glossary Construction". In Proceedings of the American Medical Informatics Association Symposium. ACM Press, New York, 2001. 252- 262
- KOZAREVA Z, HOVY E (2010) Learning arguments and supertypes of semantic relations using recursive patterns. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Uppsala, Sweden, pp 1482–1491
- KOZAREVA Z, RILOFF E, HOVY E (2008) Semantic class learning from the web with hyponym pattern linkage graphs. In: Proceedings of ACL-08: HLT, Association for Computational Linguistics, Columbus, Ohio, pp 1048–1056
- LEU F. AND KO C. (2007) "An Automated Term Definition Extraction using the Web Corpus in Chinese Language". In Proceedings of the Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'07), 2007. 435-440.
- LUX-POGODALLA V. & POLGUÈRE A. (2011). Construction of a French Lexical Network : Methodological Issues. In Proceedings of the First International Workshop on Lexical Resources, WoLeR 2011. An ESSLLI 2011 Workshop, p. 54–61, Ljubljana.
- MARTIN R. (1992) *Pour une logique du sens*, Paris, P.U.F., coll. Linguistique nouvelle, 2e édition revue et augmentée, 1992
- MATHIEU-COLAS M. (1996) « Essai de typologie des noms composés français », Cahiers de lexicologie, 69, 1996-II, pp. 71-125.
- MEYER I. (2001) "Extracting Knowledge-rich Contexts for Terminography". In Recent Advances in Computational Terminology. D. Bourigault, C. Jacquemin and M.C. L'Homme (eds.).. John Benjamin's, Amsterdam, 2001. 278- 302.
- MILLER G. A., BECKWITH R., FELLBAUM C., GROSS D. & MILLER K. J. (1990). Introduction to WordNet : An On-line Lexical Database. International Journal of Lexicography, 3(4), 235–244.
- MILLER S, FOX H, RAMSHAW L, WEISCHEDEL R (2000) A novel use of statistical parsing to extract information from text. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, Morgan Kaufmann Publishers Inc., Seattle, Washington, pp 226–233
- MILLER, G. AND CHARLES W. (1991) « Contextual correlates of semantic similarity. » *Language and Cognitive Processes*, 6:1–28
- MORIN E. AND JACQUEMIN C. (2004). Automatic Acquisition and Expansion of Hypernym Links. Computers and the Humanities (CHUM), Kluwer, 38(4), 363–396.
- MOUTON, C. AND DE CHALENDAR, G. (2010). JAWS: Just Another WordNet Subset. In Proc. of TALN'10, Montreal, Canada.
- NASTASE V., NAKOV P., SÉAGHDHA D.O., AND SZPAKOWICZ S. (2013), *Semantic Relations Between Nominals*, Synthesis Lectures on Human Language Technologies, April 2013, Vol. 6, No. 1 , Pages 1-119
- NAVIGLI R. AND S. PONZETTO (2012) BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, 2012, pp. 217-250
- NAVIGLI R. AND P. VELARDI (2010) Learning Word-Class Lattices for Definition and Hypernym Extraction. Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden, July 11-16, 2010, pp. 1318-1327
- NAVIGLI R. AND VELARDI P. (2007) "GlossExtractor: A Web Application to Automatically Create a Domain Glossary". In Lecture Notes in Computer Science 4733, 2007. 339-349
- PANTEL, P. AND PENNACCHIOTTI M.. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In Proceedings of 44th Annual Meeting of the Association for Computational Linguistics joint with 21st Conference on Computational Linguistics (COLING-ACL), pages 113–120, Sydney
- PEARSON J. (2001) Terms in Context. John Benjamin's, Amsterdam.
- POLGUÈRE, A. (2014) "Principes de Modélisation Systémique Des Réseaux Lexicaux," in 21ème Conférence TALN, pp. 79–90

- RAMISCH C. (2015) "Multiword Expressions Acquisition: A Generic and Open Framework", Theory and Applications of Natural Language Processing series XIV, Springer, ISBN 978-3-319-09206-5, 230 p., 2015.
- REBEYROLLE, J. (2000) *Forme et fonction de la définition en discours*, Thèse de doctorat, Université Toulouse-Le Mirail.
- SAGOT B. & FIŠER D. (2008). Construction d'un WordNet libre du français à partir de ressources multilingues. In Actes de TALN 2008 (Traitement automatique des langues naturelles), Avignon : LIA.
- SAGOT, B., AND FIŠER D. (2011), "Extending Wordnets by Learning from Multiple Resources," LTC'11: 5th Language and Technology Conference, 2011
- SAGOT B. ET FIŠER D. (2012). Cleaning noisy wordnets. In Proceedings of LREC 2012, Istanbul, Turquie
- SAJOUS, F., HATHOUT N., CALDERONE B. (2014), "Ne Jetons Pas Le Wiktionnaire Avec L'Oripeau Du Web ! Études et Réalisations Fondées Sur Le Dictionnaire Collaboratif," 8 (2014), 663–680
- SCHMID H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
- SERRA, G. (2009) "Extracción de Contextos Definitorios En Textos de Especialidad a Partir Del Reconocimiento de Patrones Lingüísticos," *Linguamatica*, 2009, 13–38
- SIMKO J. ET BIELIKOVA M. (2014) *Semantic Acquisition Games :Harnessing Manpower for Creating Semantics*, Springer International Publishing Switzerland,
- STORRER A. AND WELLINGHOFF S. (2006) "Automated Detection and Annotation of Term Definitions in German Text Corpora". In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06). Genève, 2006. 2373- 2376.
- SUCHANEK F.M., KASNECI G, WEIKUM G (2007) YAGO: A core of semantic knowl- edge. In: Proceedings of WWW-07, pp 697–706
- TURNERY, P.D., AND P. PANTEL. (2010) "From Frequency to Meaning: Vector Space Models of Semantics." *Journal of Artificial Intelligence Research* 37 (1): 141–188.
- VOSSEN, P. (1998) EuroWordNet: Building a Multilingual Database with Wordnets for European Languages. In: K. Choukri, D. Fry, M. Nilsson (eds), *The ELRA Newsletter*, Vol3, n1, 1998. ISSN: 1026-8200.
- WALTER S. AND PINKAL M. (2006) "Automatic Extraction of Definitions from German Court Decisions". In Proceedings of the Workshop on Information Extraction Beyond the Document. 21st International Conference on Computational Linguistics (COLING'2006). Sydney, 2006. 20–28.
- ZHAO S, GRISHMAN R (2005) Extracting relations with integrated information using kernel methods. In: ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp 419–426

Déclasser les voisins non sémantiques pour améliorer les thésaurus distributionnels

Olivier Ferret

CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, Gif-sur-Yvette, F-91191 France.

olivier.ferret@cea.fr

Résumé. La plupart des méthodes d'amélioration des thésaurus distributionnels se focalisent sur les moyens – représentations ou mesures de similarité – de mieux détecter la similarité sémantique entre les mots. Dans cet article, nous proposons un point de vue inverse : nous cherchons à détecter les voisins sémantiques associés à une entrée les moins susceptibles d'être liés sémantiquement à elle et nous utilisons cette information pour réordonner ces voisins. Pour détecter les faux voisins sémantiques d'une entrée, nous adoptons une approche s'inspirant de la désambiguïsation sémantique en construisant un classifieur permettant de différencier en contexte cette entrée des autres mots. Ce classifieur est ensuite appliqué à un échantillon des occurrences des voisins de l'entrée pour repérer ceux les plus éloignés de l'entrée. Nous évaluons cette méthode pour des thésaurus construits à partir de cooccurents syntaxiques et nous montrons l'intérêt de la combiner avec les méthodes décrites dans (Ferret, 2013b) selon une stratégie de type vote.

Abstract.

Downgrading non-semantic neighbors for improving distributional thesauri

Most of the methods for improving distributional thesauri focus on the means – representations or similarity measures – to detect better semantic similarity between words. In this article, we propose a more indirect approach focusing on the identification of the neighbors of a thesaurus entry that are not semantically linked to this entry. This identification relies on a discriminative classifier trained from unsupervised selected examples for building a distributional model of the entry in texts. Its bad neighbors are found by applying this classifier to a representative set of occurrences of each of these neighbors. We evaluate more particularly the interest of this method for thesauri built from syntactic co-occurents and we show the interest of associating this method with those of (Ferret, 2013b) following an ensemble strategy.

Mots-clés : Sémantique lexicale, similarité sémantique, thésaurus distributionnels.

Keywords: Lexical semantics, semantic similarity, distributional thesauri.

1 Introduction

Les ressources distributionnelles sont utilisées dans un ensemble de tâches de plus en plus important, allant de l'extraction de relations (Min *et al.*, 2012) à l'analyse syntaxique (Henestroza Anguiano & Candito, 2012). Le travail sur lequel se focalise cet article concerne plus spécifiquement les thésaurus distributionnels, qui associent à un mot un ensemble de voisins dits sémantiques, généralement ordonnés selon l'ordre décroissant de leur similarité avec ce mot, à l'image des exemples du tableau 1. À la suite de (Grefenstette, 1994), la façon la plus répandue de construire de tels thésaurus à partir d'un corpus est de caractériser chaque mot du corpus par l'ensemble de ses contextes d'occurrence et d'évaluer le niveau de similarité de deux mots en fonction d'une mesure de similarité reposant sur les contextes qu'ils partagent. Cette mesure permet alors de sélectionner les plus proches voisins d'un mot. Ce schéma général se retrouve sous diverses variantes dans des travaux comme (Lin, 1998), (Curran & Moens, 2002), (Weeds, 2003) ou (Heylen *et al.*, 2008).

Au-delà du problème spécifique de la construction de thésaurus, cette façon d'aborder le problème de la similarité sémantique des mots est caractéristique de la mise en œuvre traditionnelle de l'approche distributionnelle. Cette mise en œuvre a fait depuis quelques temps l'objet de nombreux développements. Une partie d'entre eux se sont attachés à améliorer l'approche de (Grefenstette, 1994), mais sans la changer en profondeur, en s'attachant à la pondération des éléments constituant les contextes distributionnels, à l'instar de (Broda *et al.*, 2009), (Zhitomirsky-Geffet & Dagan, 2009) ou (Yamamoto & Asakura, 2010). (Kazama *et al.*, 2010) a pour sa part adopté un point de vue bayésien pour aborder la question.

advance	gain [0,13], surge [0,10], progress [0,09], improvement [0,09], increase [0,09], decline [0,09] ...
distress	pain [0,14], anguish [0,13], anxiety [0,13], discomfort [0,12], grief [0,12], hardship [0,11] ...
inquisitor	good-cop [0,06], uranium-enrichment [0,06], misanthrope [0,05], interviewer [0,05] ...
insomniac	angler [0,07], tonsillitis [0,07], procrastinator [0,06], shuffler [0,06], grenadian [0,06] ...

TABLE 1 – Premiers voisins de quelques entrées du thésaurus distributionnel de la section 2

D’autres travaux ont envisagé des changements plus radicaux. Les modèles à base d’exemples (Erk & Pado, 2010) ou de prototypes multiples (Reisinger & Mooney, 2010), dans lesquels la représentation d’un mot est fondée sur un ensemble d’exemples caractéristiques au lieu d’une agrégation de contextes d’occurrence, en sont une manifestation. Les méthodes s’appuyant sur la construction de représentations lexicales distribuées en sont une autre, que ce soit par le biais de méthodes de factorisation de matrice comme l’analyse sémantique latente (Padó & Lapata, 2007) ou la factorisation de matrice non négative (Van de Cruys, 2010), de méthodes fondées sur la notion de hachage comme le Random Indexing (Kanerva *et al.*, 2000) ou plus récemment des méthodes issues du Deep Learning pour la construction de représentations de type *word embedding* (Huang *et al.*, 2012; Mikolov *et al.*, 2013) ou du modèle GloVe de (Pennington *et al.*, 2014),

En dehors des avancées réalisées globalement dans le champ de la sémantique distributionnelle, certains travaux se concentrent sur des voies d’amélioration plus spécifiques aux thésaurus distributionnels. Même s’ils ne traitent pas explicitement de cette notion de thésaurus, (Zhitomirsky-Geffet & Dagan, 2009) et (Yamamoto & Asakura, 2010) relèvent de cette problématique. Ils s’appuient en effet sur un mécanisme d’amorçage dans lequel la première étape consiste à trouver des voisins sémantiques selon une approche du type (Grefenstette, 1994), le résultat ne constituant rien d’autre qu’un thésaurus distributionnel. Ces voisins sont utilisés dans un second temps pour repondérer les éléments constitutifs des contextes distributionnels et aboutir ainsi à une version améliorée du thésaurus initial. Une telle forme d’amorçage se retrouve également au niveau de (Ferret, 2012) et de (Ferret, 2013b). Dans ce cas, le thésaurus initial est à la base de la sélection non supervisée d’exemples positifs et négatifs de mots sémantiquement liés, exemples servant ensuite à entraîner un classifieur permettant de réordonner le thésaurus initial. Dans le cas de (Ferret, 2012), cette sélection s’appuie purement sur le thésaurus initial en exploitant ses relations de symétrie tandis que (Ferret, 2013b) utilise en complément un thésaurus distributionnel de mots composés. Claveau *et al.* (2014) proposent quant à eux plusieurs façons de généraliser l’idée avancée dans (Ferret, 2012) de l’exploitation des relations à l’échelle du thésaurus.

Le travail que nous présentons dans cet article reprend l’optique, développée dans les travaux du paragraphe précédent, d’une amélioration d’un thésaurus distributionnel fondé sur un processus de réordonnement de ses voisins et s’inscrit de ce point de vue dans une nouvelle approche visant à identifier les voisins les moins susceptibles d’être en relation sémantique avec leur entrée afin de les déclasser. Plus précisément, nous verrons comment cette approche, appliquée dans le contexte de thésaurus construits sur la base de cooccurrences graphiques (Ferret, 2013a), c’est-à-dire extraits sur la base d’une fenêtre glissante de taille fixe, peut également être appliquée avec succès à des thésaurus fondés sur des cooccurrences syntaxiques. Nous montrerons également l’intérêt de combiner cette approche avec celle présentée dans (Ferret, 2013b).

2 Thésaurus initial

Avant de présenter plus avant la méthode d’amélioration des thésaurus distributionnels que nous proposons, nous présenterons en premier lieu la façon dont nous construisons de tels thésaurus et la façon dont nous les évaluons, en les comparant de ce point de vue aux travaux de référence et aux dernières avancées dans ce domaine.

2.1 Construction du thésaurus initial

Le processus de construction du thésaurus initial que nous avons suivi est très comparable à celui décrit dans (Grefenstette, 1994), (Lin, 1998) ou (Curran & Moens, 2002) pour la construction de thésaurus fondés sur des cooccurrences syntaxiques. Il reprend très concrètement le processus décrit dans (Ferret, 2010) pour les cooccurrences graphiques, avec quelques adaptations. Le point de départ est constitué comme dans (Ferret, 2010) par les 380 millions de mots du corpus AQUAINT-2 provenant d’articles de journaux écrits en anglais. Dans le cas présent, le prétraitement linguistique du corpus a été réalisé par l’analyse syntaxique MINIPAR (Lin, 1994) permettant de fournir trois types d’informations exploités pour la construction du thésaurus : la forme lemmatisée des mots, leur catégorie morphosyntaxique et plus spécifiquement

type cooc.	réf.	#mots éval.	#syn. / mot	rappel	R-préc.	MAP	P@1	P@5	P@10	P@100
syntaxiques 14 117	W	10 178	2,9	31,5	11,5	13,3	15,6	6,9	4,5	0,9
	M	9 060	50,4	12,8	9,4	4,8	30,6	21,7	17,3	6,5
	WM	11 887	39,4	13,2	10,8	7,9	29,4	18,9	14,6	5,2
graphiques 14 670	W	10 473	2,9	24,6	8,2	9,8	11,7	5,1	3,4	0,7
	M	9 216	50,0	9,5	6,7	3,2	24,1	16,4	13,0	4,8
	WM	12 243	38,7	9,8	7,7	5,6	22,5	14,1	10,8	3,8

TABLE 2 – Évaluation du thésaurus initial fondé sur des cooccurents syntaxiques et comparaison par rapport à l’usage de cooccurents graphiques

les relations de dépendance syntaxique qui les unissent. Pour la constitution des données distributionnelles, la principale différence avec (Ferret, 2010) réside au niveau des éléments constitutifs des contextes distributionnels, prenant la forme de cooccurents syntaxiques. Plus précisément, chaque cooccurrent est représenté sous la forme de la paire (*relation syntaxique*, *lemme cooccurrent*), en se limitant aux noms, verbes et adjectifs. Le filtrage fréquentiel des contextes, éliminant les cooccurents de fréquence 1, la pondération des cooccurents par l’information mutuelle ramenée aux valeurs positives (PPMI : *Positive Pointwise Mutual Information*) et l’adoption de la mesure *Cosinus* pour évaluer la similarité des contextes sont pour leur part similaires aux paramètres sélectionnés par (Ferret, 2010) ainsi qu’aux conclusions d’études récentes et plus exhaustives couvrant aussi les cooccurents syntaxiques, comme (Kiela & Clark, 2014)¹. Un filtre fréquentiel a en outre été appliqué aux mots cibles et à leurs cooccurents et ne conserve que les mots de fréquence supérieure à 10.

La construction proprement dite du thésaurus a été réalisée de façon classique en sélectionnant les voisins sémantiques les plus proches de chaque entrée considérée selon la mesure de similarité entre contextes. Plus précisément, cette mesure a été calculée entre chaque entrée et l’ensemble de ses voisins possibles et ceux-ci ont été ordonnés selon l’ordre décroissant des valeurs calculées. Seuls les 100 premiers voisins ont alors été conservés pour chaque entrée du thésaurus. Entrées et voisins se limitent dans le cas présent à des noms².

2.2 Évaluation et mise en perspective du thésaurus initial

Le tableau 2 donne les résultats de l’évaluation du thésaurus distributionnel obtenu (lignes *syntaxiques*) et les met en perspective avec ceux du thésaurus construit dans (Ferret, 2010) sur la base du même corpus mais avec des cooccurents graphiques (lignes *graphiques*). Cette évaluation est réalisée comme dans (Ferret, 2010) en comparant les voisins sémantiques extraits à deux ressources de référence complémentaires : les synonymes de WordNet [W] (Miller, 1990), dans sa version 3.0, qui permettent de caractériser une similarité fondée sur des relations paradigmatiques et le thésaurus Moby [M] (Ward, 1996), qui regroupe des mots liés par des relations plus diverses de proximité sémantique. Comme l’illustre la 4^{ème} colonne du tableau, ces deux ressources sont aussi très différentes en termes de richesse. Le but étant d’évaluer la capacité à extraire des voisins sémantiques, elles sont filtrées pour en exclure les entrées et les voisins non présents dans le vocabulaire du corpus AQUAINT-2 (cf. la différence entre le nombre de mots de la 1^{ère} colonne et le nombre de mots effectivement évalués de la 3^{ème} colonne). Une fusion de ces deux ressources a également été faite [WM]. Compte tenu de l’inévitable incomplétude de ces ressources de référence vis-à-vis des notions de similarité et de proximité sémantique, les chiffres du tableau 2 doivent être considérés comme des minima.

Ces résultats se déclinent sous la forme de différentes mesures, à commencer à la 5^{ème} colonne par le taux de rappel par rapport aux ressources considérées pour les 100 premiers voisins de chaque nom. Ces voisins étant ordonnés, il est en outre possible de réutiliser les métriques d’évaluation classiquement adoptées en recherche d’information en faisant jouer aux mots cibles le rôle de requêtes et aux voisins celui des documents. Les dernières colonnes du tableau 2 rendent compte de ces mesures : la R-précision (R-préc.) est la précision obtenue en se limitant aux R premiers voisins, R étant le nombre de synonymes dans la ressource de référence pour l’entrée considérée ; la MAP (Mean Average Precision) est la moyenne des précisions pour chacun des rangs auxquels un synonyme de référence a été identifié ; enfin, sont données les précisions pour différents seuils de nombre de voisins sémantiques examinés (précision après examen des 1, 5, 10 et 100 premiers voisins). Pour une meilleure lisibilité, toutes ces valeurs sont données sur une échelle de 0 à 100.

1. Claveau *et al.* (2014) ont montré qu’une version ajustée de la pondération Okapi-BM25 permet de dépasser les performances de PPMI pour des cooccurents graphiques mais les différences par rapport au point de comparaison adopté ne permettent pas d’attribuer de façon sûre cette amélioration des performances au type de pondération.

2. Le thésaurus construit est disponible sur le site <https://github.com/osf9018/a2st>.

De façon non surprenante, le premier constat qu'impose le tableau 2 est la supériorité des cooccurents syntaxiques sur les cooccurents graphiques, même si cette supériorité s'accompagne d'un plus grand nombre d'entrées dépourvues de voisins. La différence de 553 entrées est toutefois faible au regard du nombre total d'entrées et s'explique *a priori* par la densité moindre des cooccurents syntaxiques par rapport aux cooccurents graphiques, aboutissant parfois à une intersection vide des contextes distributionnels. En revanche, les observations faites dans (Ferret, 2010) à propos des cooccurents graphiques se confirment ici avec des cooccurents syntaxiques. En particulier, les résultats obtenus sont fortement sensibles à la ressource de référence utilisée pour l'évaluation, à la fois du point de vue du nombre de voisins par entrée et du type de ces voisins. Ainsi, la précision à différents rangs est significativement plus forte avec Moby comme référence, qui fournit beaucoup de voisins de différents types pour chaque entrée, qu'avec WordNet, qui ne donne pour chaque entrée qu'un ensemble restreint de synonymes. La MAP et la R-précision montrent une tendance inverse pour les mêmes raisons. Il est à noter que la richesse des références utilisées explique au moins en partie pourquoi des travaux comme (Curran & Moens, 2002) ou plus récemment (Riedl & Biemann, 2013) affichent des valeurs élevées pour la précision au rang 1 par exemple.

méthode	#mots éval.	#syn./ mot	rappel	R-préc.	MAP	P@1	P@5	P@10	P@100
A2ST-SYNT	11 887	39,4	13,2	10,8	7,9	29,4	18,9	14,6	5,2
(Lin, 1998)	9 823	44,5	12,7	11,6	8,1	36,1	23,7	18,2	5,6
(Huang <i>et al.</i> , 2012)	10 537	42,6	3,8	1,9	0,8	7,1	5,0	4,0	1,6
(Mikolov <i>et al.</i> , 2013)	12 326	38,6	6,2	5,5	4,2	16,3	9,5	7,0	2,4
(Baroni <i>et al.</i> , 2014)-count	12 052	39,3	13,6	12,5	9,8	31,9	19,6	15,2	5,3
(Baroni <i>et al.</i> , 2014)-predict	12 052	39,3	11,3	10,9	8,5	30,3	18,4	13,8	4,4

TABLE 3 – Comparaison de plusieurs approches pour la construction de thésaurus distributionnels

Le tableau 3 permet de mettre en perspective les résultats de notre thésaurus (A2ST-SYNT) avec ceux de plusieurs autres thésaurus, en utilisant [WM] comme référence. (Lin, 1998) est le thésaurus mis à disposition par Lin³, construit comme A2ST-SYNT grâce à des cooccurents syntaxiques obtenus par l'analyseur MINIPAR. L'évaluation de ce thésaurus donne de meilleurs résultats que pour A2ST-SYNT, ce qui peut s'expliquer par deux facteurs : d'une part, le corpus utilisé par Lin, d'une taille de 1,5 milliards de mots, est beaucoup plus important que le corpus AQUAINT-2 ; d'autre part, du fait des entrées disponibles, l'évaluation du corpus de Lin a été réalisée sur un plus petit ensemble d'entrées, en moyenne de plus forte fréquence comme le montre le nombre de synonymes par entrée. (Huang *et al.*, 2012) et (Mikolov *et al.*, 2013) correspondent quant à eux à deux approches récentes fondées sur la construction de représentations distribuées de mots, appelées *word embeddings*, par des réseaux de neurones. Ces représentations sont utilisées en lieu et place des contextes distributionnels classiques lors de la construction des thésaurus. Dans le cas de (Huang *et al.*, 2012), nous avons utilisé les représentations construites à partir de Wikipédia⁴ fournies par les auteurs tandis que dans le cas de (Mikolov *et al.*, 2013), nous avons calculé ces représentations à partir du corpus AQUAINT-2 en utilisant les meilleurs paramètres du modèle *Skip-gram* sélectionnés par (Mikolov *et al.*, 2013). Dans les deux cas, les résultats sont significativement inférieurs à ceux de A2ST-SYNT, avec un niveau particulièrement bas pour (Huang *et al.*, 2012) qui peut s'expliquer au moins en partie par la différence de corpus. Ces résultats suggèrent néanmoins que l'utilisation de ce type de représentations distribuées n'est pas encore l'option la plus intéressante pour la construction de thésaurus distributionnels. Ce constat est renforcé par les deux dernières lignes du tableau 3, qui donnent les résultats des thésaurus construits avec les vecteurs de contexte mis à disposition par Baroni *et al.* (2014)⁵ : (Baroni *et al.*, 2014)-*predict* correspond à des vecteurs construits grâce au modèle CBOW de (Mikolov *et al.*, 2013) tandis que (Baroni *et al.*, 2014)-*count* correspond à des vecteurs de cooccurents obtenus de façon classique par une fenêtre graphique. Le niveau des résultats obtenus, rivalisant et même dépassant pour bon nombre de mesures les résultats de A2ST-SYNT et ceux du thésaurus de Lin, confirme l'observation faite à propos du thésaurus de Lin de la grande importance de la taille du corpus initial sur les résultats, égale à 2,8 milliards de mots dans le cas de (Baroni *et al.*, 2014). Mais l'observation la plus importante est ici la supériorité de (Baroni *et al.*, 2014)-*count* par rapport à (Baroni *et al.*, 2014)-*predict*, ce qui vient limiter le constat général fait par Baroni *et al.* (2014) de la supériorité de l'approche *predict* par rapport à l'approche *count*. Ce constat n'est visiblement pas vérifié dans le cas des thésaurus distributionnels.

3. <http://webdocs.cs.ualberta.ca/lindek/Downloads/sim.tgz>

4. http://nlp.stanford.edu/~socherr/ACL2012_wordVectorsTextFile.zip

5. <http://clic.cimec.unitn.it/composes/semantic-vectors.html>

3 Principes et vue d'ensemble

Dans ce travail, nous adoptons une approche indirecte de l'amélioration des thésaurus distributionnels en nous focalisant sur la détection des voisins les moins sémantiquement liés à leur entrée. En dehors du tableau 2, les exemples du tableau 1 nous permettent d'appréhender plus qualitativement ces voisins « pas très sémantiques ». Ainsi, le mot *misanthrope* pour l'entrée *inquisitor* ou le mot *procrastinator* pour l'entrée *insomniac* sont des exemples assez évidents que certains voisins sont à rejeter sans équivoque. En détectant ces voisins et en les repoussant en queue de liste, les mots véritablement liés à l'entrée considérée, lorsqu'ils sont présents dans la liste de ses voisins, voient leur rang augmenter de façon mécanique. Il en résulte un thésaurus plus intéressant à exploiter dans la mesure où en pratique, seuls les tout premiers voisins d'une entrée sont utilisables compte tenu de l'augmentation importante du taux d'erreur en fonction du rang.

L'approche que nous proposons pour identifier ces voisins pas très sémantiques repose, comme la construction du thésaurus initial, sur l'hypothèse distributionnelle. Sa mise en œuvre est en revanche différente. Cette hypothèse stipule dans les grandes lignes que la signification d'un mot peut être caractérisée par l'ensemble des contextes dans lesquels ce mot est rencontré. De ce fait, deux mots sont considérés comme sémantiquement similaires s'ils sont observés dans un ensemble suffisant large de contextes communs. Dans les travaux du type de (Curran & Moens, 2002), cette approche est implémentée en associant à chaque mot un contexte distributionnel rassemblant tous les mots avec lesquels il cooccure dans un large corpus. Suivant que les cooccurents sont graphiques ou syntaxiques, un tel contexte est une représentation peu structurée de type « sac de mots » ou « sac de paires (mot, relation syntaxique) ». (Kazama *et al.*, 2010) propose une variante de ce schéma en modélisant la représentation distributionnelle d'un mot sous la forme d'une distribution multinomiale, sans que cela n'induisse de nouvelles possibilités de représentation.

Or, cette approche présente l'inconvénient de n'autoriser qu'une faible diversité et une faible structuration des éléments utilisés pour construire les modèles. Des traits comme des ngrammes de mots ou de catégories morphosyntaxiques ne sont ainsi pas utilisés pour la construction de thésaurus alors qu'ils le sont largement pour des tâches comme la désambiguïsation sémantique par exemple. Les inclure dans une représentation de type « sac de mots » conduirait en effet à une inflation très significative des modèles, sans compter que des mesures de similarité telles que la mesure *Cosinus* ne sont pas adaptées à des représentations hétérogènes (Alexandrescu & Kirchhoff, 2007). Pour dépasser cette limite, nous proposons donc de construire un modèle discriminant pour représenter les contextes d'un mot, ce type de modèle étant naturellement capable d'intégrer une large diversité et un grand nombre de traits. Un tel modèle a plus précisément pour objectif de permettre la discrimination sémantique d'un mot en contexte, c'est-à-dire une occurrence de ce mot dans un corpus, vis-à-vis de tous les autres mots, et plus particulièrement de ceux de ses voisins issus d'un thésaurus distributionnel les plus éloignés de lui sur le plan sémantique. L'hypothèse sous-jacente s'appuie sur l'idée distributionnelle : un mot et un synonyme de ce mot doivent apparaître dans les mêmes contextes et sont donc représentables par les mêmes traits puisque ceux-ci sont extraits de ces contextes. Un modèle reposant sur ces traits capable d'identifier l'occurrence d'un mot en contexte doit donc être à même d'identifier une occurrence d'un de ses synonymes et plus généralement une occurrence d'un mot qui peut se substituer à lui sur le plan paradigmatique. Comme nous le démontrerons dans ce qui suit, un tel modèle est en pratique particulièrement bien adapté à la détection des mots qui ne sont pas sémantiquement liés au mot cible de ce modèle.

Le fait de vouloir capturer les contextes d'un mot au travers d'un modèle discriminant pose néanmoins un problème spécifique. Dans le cas d'un modèle de type « sac de mots », les contextes de deux mots peuvent être directement comparés par le biais d'une mesure de similarité. Dans le cas d'un modèle discriminant, une telle comparaison directe n'est pas possible. Ils ne peuvent en fait être comparés que le biais d'une application à un ensemble d'exemples, c'est-à-dire d'occurrences de mots dans le cas présent. Ainsi, pour déterminer si un voisin d'une entrée d'un thésaurus est un « mauvais » voisin du point de vue sémantique, nous avons choisi d'appliquer un modèle discriminant appris pour cette entrée à un échantillon d'occurrences de ce voisin issu du corpus utilisé pour construire le thésaurus. D'un point de plus global, la méthode d'amélioration d'un thésaurus distributionnel proposée se caractérise donc par l'application des cinq étapes suivantes pour chaque entrée du thésaurus :

- construction d'un classifieur déterminant si une occurrence d'un mot s'identifie ou non à une occurrence de l'entrée considérée ;
- sélection d'un échantillon d'occurrences pour chacun des voisins de l'entrée dans le thésaurus. En pratique, chaque occurrence correspond à une phrase ;
- application du classifieur de la première étape à ces échantillons ;
- détection des mauvais voisins en fonction de ces résultats de classification ;
- réordonnement des voisins de l'entrée considérée par déclassement des mauvais voisins détectés.

4 Améliorer un thésaurus distributionnel

4.1 Construction des modèles discriminants des mots en contexte

La première, et la plus importante étape de notre processus d'amélioration d'un thésaurus distributionnel est la définition d'un modèle déterminant dans quelle mesure l'occurrence O_i d'un mot V peut être une occurrence d'un mot de référence E , en faisant bien entendu abstraction du rattachement effectif de O_i à V . Cette tâche peut également être envisagée comme une tâche d'étiquetage dans laquelle les occurrences d'un mot cible V ont deux étiquettes possibles : E et $nonE$. Dans le contexte plus général qui nous est propre, l'intérêt de cette tâche n'est pas à prendre au premier degré mais vient plutôt du fait qu'un tel classifieur est susceptible de modéliser les contextes dans lesquels E est observé et donc, si l'on s'attache à l'hypothèse distributionnelle, de modéliser également son sens.

En allant un pas plus loin dans cette logique, un tel classifieur peut être vu comme un moyen de tester si un mot a le même sens que E . Il s'agit là d'un problème très proche de la notion de pseudo désambiguïsation sémantique, au sens de (Gale *et al.*, 1992) : un pseudo-mot est créé avec deux sens, E et $nonE$, $nonE$ prenant la forme d'un ou plusieurs mots dont le sens est supposé représentatif de sens autres que celui de E . L'objectif est alors de construire un classifieur permet de distinguer en contexte les pseudo-sens E et $nonE$. De cette vision, découle la décision d'adopter, pour construire notre classifieur, les mêmes traits que ceux utilisés les plus couramment en désambiguïsation sémantique. Nous avons sur ce plan suivi (Lee & Ng, 2002), un travail de référence dans ce domaine, en choisissant un classifieur à base de Machine à Vecteurs de Support (SVM) avec un noyau linéaire et les trois catégories de traits suivantes pour caractériser chaque occurrence de E et des mots représentatifs de $nonE$ ⁶ :

- les mots environnants ;
- la catégorie morphosyntaxique des mots environnants ;
- les collocations locales.

Pour les *mots environnants*, nous avons retenu tous les mots pleins (noms communs et noms propres, verbes et adjectifs) ainsi que les adverbes présents dans la même phrase qu'une occurrence de E . Chaque mot environnant est représenté par son lemme sous la forme d'un trait binaire dont la valeur est égale à 1 en cas de présence dans la même phrase qu'une occurrence de E . Pour le deuxième type de traits, nous avons considéré la catégorie morphosyntaxique des trois mots précédant et des trois mots suivant une occurrence de E . Chaque couple {catégorie, position} donne lieu à un trait binaire. Le symbole spécial *empty* est utilisé pour remplacer la catégorie morphosyntaxique lorsque la position se situe au-delà de la fin de la phrase ou précède son début. Enfin, les collocations locales correspondent à des couples de mots présents dans le voisinage d'une occurrence de E . Une collocation est notée $C_{i,j}$, avec i et j faisant référence aux positions respectives du premier et du second mot de la collocation. Dans notre cas, i et j prennent leurs valeurs dans l'intervalle $[-3, +3]$, à l'image des catégories morphosyntaxiques. Plus précisément, les 11 collocations suivantes sont extraites pour chaque occurrence de E :

- $C_{-3,-1}$, $C_{-2,-2}$, $C_{-2,-1}$, $C_{-1,-1}$
- $C_{-2,1}$, $C_{-1,1}$, $C_{-1,2}$
- $C_{1,1}$, $C_{1,2}$, $C_{1,3}$, $C_{2,2}$

Comme dans le cas des catégories morphosyntaxiques, le symbole spécial *empty* est utilisé pour les mots précédant le début de la phrase ou suivant la fin de celle-ci et à l'instar des mots environnants, les mots des collocations sont donnés sous une forme lemmatisée. Chaque instance de l'un de ces 11 schémas de collocations est représentée par un tuple $\langle \text{lemme1, position1, lemme2, position2} \rangle$, donnant lieu également à un trait binaire pour le classifieur SVM.

Conformément au processus global décrit à la section précédente, un classifieur SVM spécifique a été entraîné pour chaque entrée de notre thésaurus initial, ce qui requiert la sélection d'un ensemble d'exemples positifs et négatifs. Pour les exemples positifs, nous avons simplement choisi de manière aléatoire un nombre fixe de phrases issues du corpus AQUAINT-2 contenant au moins une occurrence de l'entrée, la première occurrence dans une phrase étant retenue comme exemple positif. La sélection des exemples négatifs a quant à elle été guidée par notre thésaurus, dans la perspective de s'appuyer sur des critères de proximité sémantique par rapport à l'entrée. De ce point de vue, le choix d'un voisin de l'entrée éloigné en termes de rang aurait garanti un faible nombre de faux exemples négatifs, c'est-à-dire de mots similaires à l'entrée⁷, dans la mesure où les performances décroissent rapidement à mesure que le rang des voisins

6. Dans ce qui suit nous ferons référence aux traits associés aux occurrences de E mais les mêmes traits sont associés aux occurrences des mots représentatifs de $nonE$.

7. Formellement, les exemples sont des occurrences de mots mais nous parlerons parfois de mots pour simplifier l'expression.

augmente, comme le montre le tableau 2. En pratique, prendre comme exemples négatifs des voisins ayant un rang plus faible est une meilleure option dans la mesure où ils sont plus utiles en termes de discrimination, étant plus proches de la zone de transition entre exemples positifs et négatifs. Pour limiter les risques de faux exemples négatifs, nous avons réparti notre sélection sur trois rangs, en l'occurrence les rangs 10, 15 et 20, sans optimisation particulière. Pour chacun de ces exemples négatifs, un nombre fixe de phrases ont été sélectionnées de la même façon que pour les exemples positifs. En moyenne, le nombre d'exemples négatifs est donc trois fois plus important que le nombre d'exemples positifs, ce qui reflète la présence très majoritaire de voisins non sémantiquement liés à leur entrée dans le thésaurus initial.

4.2 Identification des mauvais voisins et réordonnement du thésaurus

Pour l'identification des « mauvais » voisins d'une entrée, le classifieur développé pour reconnaître les occurrences de cette entrée est appliqué à un nombre fixe d'occurrences représentatives de chacun de ses voisins. La sélection de ces occurrences s'effectue de la même façon que la sélection des exemples positifs et négatifs décrite à la section précédente. L'application de ce classifieur vise à déterminer si le contexte de l'occurrence considérée est compatible avec le contexte d'une occurrence de l'entrée. En pratique, la décision de ce classifieur est rarement positive, ce qui n'est pas complètement surprenant : même si deux mots sont sémantiquement équivalents, chacun d'entre eux est caractérisé par des usages spécifiques, en particulier dans un corpus donné, et certains traits utilisés par notre classifieur, comme les collocations, sont plus susceptibles que les contextes distributionnels classiques de capturer de telles spécificités. De ce fait, nous faisons l'hypothèse qu'une décision positive du classifieur est un indice fort de la présence d'un voisin sémantiquement similaire à l'entrée et nous considérons un voisin comme potentiellement « bon » si au moins un nombre minimal G de ses occurrences sélectionnées sont classées positivement. À l'inverse, un voisin est considéré comme « mauvais » si le nombre de décisions positives du classifieur est inférieur à G . Les voisins identifiés comme mauvais ne sont pas complètement écartés mais font l'objet d'une rétrogradation en queue de la liste des voisins de l'entrée considérée, leur ordre initial relatif restant inchangé. Il est enfin à noter que le classifieur de mots en contexte n'est pas appliqué aux voisins dont certaines occurrences ont été utilisées en tant qu'exemples négatifs lors de son entraînement. Ces voisins se verraient en effet très souvent rétrogradés alors qu'ils occupent de faibles rangs et qu'ils sont de ce fait liés de façon effective à l'entrée dans une proportion de cas non négligeable.

5 Expérimentations et évaluation

5.1 Mise en œuvre

La mise en œuvre de la méthode que nous avons présentée ci-dessus nécessite de préciser plusieurs points. L'un d'entre eux concerne la taille des échantillons de phrases à sélectionner à la fois pour les entrées du thésaurus et pour leurs voisins. Ces phrases, et plus précisément les occurrences de mots qu'elles contiennent, sont utilisées à la fois pour l'entraînement du classifieur de mots en contexte et pour l'identification des mauvais voisins. En pratique, nous avons sélectionné aléatoirement un échantillon de 250 phrases pour chaque mot de notre vocabulaire et nous avons exploité l'ensemble ainsi constitué pour les deux tâches. Cet échantillonnage a été réalisé sur la base de la forme lemmatisée des mots du vocabulaire. Par ailleurs, le chiffre de 250 est une limite haute dans la mesure où beaucoup de mots du vocabulaire ont une fréquence inférieure à cette limite, la fréquence minimale étant égale à 11 et la fréquence médiane à 249. Cette limite représente une forme de moyen terme entre les 385 exemples d'entraînement en moyenne de la tâche *Lexical Sample* de l'évaluation *Senseval 1* et les 118 exemples de la même tâche de *Senseval 2*.

Les modalités d'entraînement de nos classifieurs de mots en contexte sont également à préciser. Ceux-ci étant des SVM linéaires, seul le paramètre de régularisation C peut être optimisé. Néanmoins, le nombre de ces classifieurs étant égal au nombre d'entrées du thésaurus, le coût d'une telle optimisation n'est pas complètement négligeable. Nous avons donc évalué dans un premier temps ces classifieurs par le biais d'une procédure de validation croisée à 5 volets, avec une valeur de C par défaut égale à 1. Le tableau 4 donne l'exactitude moyenne de ces classifieurs, ainsi que leur écart-type, pour l'ensemble des entrées du thésaurus ainsi que pour une segmentation selon trois tranches fréquentielles (basses : fréquence < 100 ; moyennes : $100 < \text{fréquence} \leq 1000$; hautes : fréquence > 1000).

Ce tableau montre que le niveau général de résultat de ces classifieurs est élevé, avec des valeurs d'exactitude très similaires pour les différentes tranches fréquentielles⁸. Par ailleurs, les tentatives d'optimisation du paramètre C que nous

8. L'écart-type pour les fréquences basses est un peu plus élevé mais peut s'expliquer par le nombre d'exemples assez faible pour ces fréquences, ce

	toutes	hautes	moyennes	basses
exactitude	86,2	86,1	86,0	86,5
écart-type	6,1	4,2	5,7	7,6

TABLE 4 – Résultats des classifieurs de mots en contexte

avons réalisées pour quelques entrées du thésaurus n’ont pas montré d’amélioration possible. Nous avons donc décidé, à l’échelle de toutes les entrées du thésaurus, de conserver le paramètre C à la valeur de 1.

	3	5	7	10	15
R-préc.	11,2	11,1	11,1	11,0	10,8
P@5	19,4	19,5	19,5	19,4	19,3

TABLE 5 – R-précision et précision au rang 5 pour différentes valeurs de G , avec la référence [WM]

Le dernier paramètre à fixer dans la méthode considérée ici est le paramètre G déterminant le nombre d’occurrences d’un voisin classées négativement par le classifieur en contexte de l’entrée en dessous duquel ce voisin est considéré comme mauvais. Le tableau 5 illustre l’influence de ce paramètre en donnant les résultats obtenus pour différentes valeurs de G en termes de R-précision et précision au rang 5 par rapport à la référence WM. Globalement, il laisse apparaître que cette influence est faible. Les valeurs les plus élevées de G dégradent les résultats globaux mais cette tendance n’est pas très marquée. Cette relative insensibilité aux valeurs de G laisse à penser que l’assimilation d’une occurrence d’un voisin à une entrée par le classifieur en contexte de celle-ci est un indice très fort du fait que ce voisin n’est probablement pas un mauvais voisin. Compte tenu de ce constat, la valeur $G = 3$ a été retenue pour les résultats de la section suivante.

5.2 Évaluation après réordonnancement

Le tableau 6 donne les résultats de l’évaluation, selon les mêmes modalités qu’à la section 2.2, du thésaurus obtenu après réordonnancement de ses voisins selon la méthode présentée ci-dessus. Chaque mesure est accompagnée de sa différence en valeur par rapport à la mesure correspondante pour le thésaurus initial. Les résultats du nouveau thésaurus (lignes *syntaxiques*) sont comme précédemment mis en balance avec les résultats déjà obtenus pour les cooccurents graphiques.

La première observation suscitée par ce tableau est le fait qu’au niveau global, toutes les mesures sont améliorées de façon significative⁹, ce qui distingue la méthode considérée favorablement par rapport à (Ferret, 2012) et (Ferret, 2013b) qui enregistraient certaines baisses avec WordNet comme référence, même si ces baisses étaient souvent non significatives. Cet état de fait trouve son explication probable dans le très faible taux d’erreur de la méthode d’identification des mauvais voisins. En prenant WordNet comme référence, seuls 670 voisins répartis sur 570 entrées ont été faussement identifiés comme de mauvais voisins, ce qui ne représente que 4,7% des voisins dégradés. Ce résultat valide en outre, au moins partiellement, la possibilité de capturer par le biais d’un classifieur discriminant intervenant au niveau des occurrences de mots une partie au moins des propriétés distributionnelles de ceux-ci. Il est d’ailleurs à noter que les traits utilisés dans le cas présent sont adaptés à la désambiguïsation sémantique, ce qui ne les prédispose pas pour autant à rendre compte de tous les aspects sémantiques des mots. Des traits plus spécifiques à la tâche considérée pourraient éventuellement conduire à de meilleurs résultats. Sur un autre plan, il faut préciser que les améliorations sont obtenues pour toutes les fréquences, y compris les plus faibles. Le petit nombre d’exemples en contexte pour ces entrées aurait pu faire craindre des améliorations moindres, voire des dégradations, ce qui n’est pas le cas. Même si l’influence du nombre d’exemples fournis reste à analyser, cette observation suggère que l’approche n’est pas très sensible à la valeur de ce paramètre.

Le deuxième grand enseignement apporté par le tableau 6 concerne le type des relations sémantiques. Les résultats avec Moby comme référence se trouvent en effet améliorés de façon plus importante que ceux obtenus avec WordNet comme référence. Cette tendance peut paraître surprenante au premier abord dans la mesure où la méthode présentée, en mettant l’accent sur la caractérisation en contexte d’un mot par opposition à tous les autres, pourrait *a priori* sembler favoriser les voisins les plus directement substituables à lui, donc des synonymes. Mais il faut aussi considérer que la synonymie n’est pas la seule relation sémantique de nature paradigmatique et que Moby abrite aussi beaucoup de ces relations. Ainsi, si

qui est source d’instabilité.

9. La significativité statistique des différences avec le thésaurus initial est évaluée grâce à un test de Wilcoxon pour échantillons appariés avec un seuil de significativité de 0,01.

type cooc.	réf.	R-préc.	MAP	P@1	P@5	P@10
syntaxiques	W	11,8 (0,3)	13,7 (0,4)	16,1 (0,6)	7,1 (0,2)	4,6 (0,1)
	M	9,6 (0,2)	4,9 (0,1)	31,6 (1,0)	22,3 (0,6)	17,8 (0,5)
	WM	11,2 (0,5)	8,2 (0,3)	30,3 (0,9)	19,4 (0,5)	15,1 (0,5)
graphiques	W	9,1 (0,9)	10,7 (0,9)	12,8 (1,1)	5,6 (0,5)	3,7 (0,3)
	M	7,2 (0,5)	3,5 (0,3)	26,5 (2,4)	17,9 (1,5)	14,0 (1,0)
	WM	8,4 (0,7)	6,1 (0,5)	24,8 (2,3)	15,4 (1,3)	11,7 (0,9)

TABLE 6 – Résultat du réordonnement des voisins sémantiques

l'on utilise WordNet comme moyen d'analyse des relations présentes dans Moby, on constate que la cohyponymie y est la relation la plus fréquente et que l'hyponymie et l'hyperonymie directes y occupent respectivement les 6^{ème} et 7^{ème} rangs.

Le tableau 6 permet enfin de constater que la méthode décrite dans cet article est plus efficace pour les cooccurrents graphiques que pour les cooccurrents syntaxiques, même si dans ce dernier cas, elle apporte tout de même un plus significatif. Le niveau initial plus élevé des résultats dans le cas des cooccurrents graphiques est l'explication la plus évidente de ce constat mais il faut noter que l'absence de prise en compte des relations syntaxiques au niveau du classifieur de mots en contexte est une source d'améliorations non exploitée ici. Il faut aussi souligner que le thésaurus « syntaxique » est moins riche que le thésaurus fondé sur des cooccurrents graphiques en entrées de faible fréquence, entrées qui bénéficient précisément le plus des améliorations mises en œuvre.

WordNet	catastrophe, calamity, tragedy, disaster
<u>Moby</u>	accident, apoplexy, blow, breakdown, breakup, calamity, casualty, catastrophe, climax, collapse, collision, convulsion, crash, debacle + 35 mots supplémentaires
initial	divisionism, <u>calamity</u> , upheaval, <u>catastrophe</u> , schism, landslip, <u>disaster</u> , devastation, conflagration, <u>tragedy</u> , deterioration, shakeout ...
réordonnés	<u>calamity</u> , upheaval, <u>catastrophe</u> , schism, <u>disaster</u> , devastation, conflagration, <u>tragedy</u> , deterioration, shakeout, maneuvering, displacement ...

TABLE 7 – Impact de notre réordonnement pour l'entrée *cataclysm*

Le tableau 7 apporte quant à lui une vue plus qualitative des résultats de la procédure de réordonnement présentée en l'illustrant pour l'entrée *cataclysm*. Les lignes **WordNet** et Moby donnent les voisins de référence dans ces deux ressources, la ligne *initial*, les plus proches voisins présents dans le thésaurus initial « syntaxique » et la ligne *réordonnés*, ceux dans le thésaurus après le réordonnement. Au niveau des premiers voisins qui sont montrés ici, les « bons » voisins sont communs aux deux ressources et l'on constate globalement que les voisins présents dans le thésaurus ne sont pas aberrants, même lorsqu'ils ne font pas partie des ressources de référence. La procédure de réordonnement permet néanmoins d'améliorer la situation en éliminant le premier voisin, non pertinent, tout en préservant les voisins pertinents. De ce fait, le rang de ces « bons » voisins augmente mécaniquement, ce qui était l'effet recherché. On notera néanmoins que le voisin *schism*, lié sémantiquement au voisin dégradé *divisionism*, conserve sa place relative.

5.3 Combinaison des approches

La méthode appliquée ici et celles proposées par Ferret (2013b) s'appuient sur des critères très différents les uns des autres. Dans le prolongement de (Curran, 2002), il apparaît *a priori* intéressant de combiner les résultats de ces différentes méthodes de réordonnement de thésaurus. Chaque thésaurus résultat donnant pour chacune de ses entrées une liste de voisins ordonnés selon l'ordre décroissant de leur proximité avec leur entrée, la solution la plus évidente est de procéder pour chaque entrée à une fusion de la liste des voisins issue de chacun des thésaurus résultat en adoptant une méthode classique de vote. Le tableau 8 donne les résultats que nous avons obtenus avec quatre de ces méthodes. Trois d'entre elles, *Borda*, *Condorcet* (Nuray & Can, 2006) et *Reciprocal Rank Fusion* (RRF, avec le paramètre $k = 60$ de (Cormack *et al.*, 2009)), s'appuient uniquement sur les rangs tandis que *CombSum*, utilisée ici avec une normalisation des valeurs de type *Zero-one* (Wu *et al.*, 2006), exploite les valeurs de similarité. Quatre thésaurus sont ainsi fusionnés : le thésaurus initial, en l'occurrence celui construit à partir des cooccurrents graphiques puisqu'il est commun à tous les travaux considérés ; le thésaurus réordonné grâce au critère de symétrie de (Ferret, 2013b) ; celui réordonné grâce aux mots composés, toujours

Thésaurus	R-préc.	MAP	P@1	P@5	P@10
initial	7,7	5,6	22,5	14,1	10,8
symétrie	+0,3	+0,1	+2,1	+0,8	+0,6
composé	+0,1	-0,1	+2,0	+0,9	+0,6
déclassement	+0,7	+0,5	+2,3	+1,3	+0,9
RRF	+0,9	+0,7	+4,2	+2,3	+1,6
Borda	+0,8	+0,7	+4,1	+2,1	+1,5
Condorcet	+0,9	+0,8	+3,4	+2,4	+1,7
CombSum	+1,2	+1,0	+5,1	+2,6	+1,8

TABLE 8 – Évaluation de la fusion des différentes méthodes d’amélioration, avec la référence [WM]

issu de (Ferret, 2013b) ; enfin, le thésaurus produit grâce à la détection des mauvais voisins présentée dans cet article.

Outre les résultats pour ces quatre méthodes de fusion, le tableau 8 rappelle les résultats pour les thésaurus fusionnés. Ces résultats, de même que ceux issus des fusions, sont donnés en différence de valeur par rapport au thésaurus initial. Un premier constat d’évidence s’impose : les méthodes de fusion permettent toutes de dépasser les résultats de chacun des quatre thésaurus fusionnés. Les gains en termes de R-précision et de MAP apparaissent modestes mais la référence étant [WM], le nombre de voisins de référence est important, ce qui a un impact direct sur ces deux mesures. En revanche, les gains sont nettement plus substantiels concernant la précision aux rangs 1, 5 et 10. Dans une optique applicative, cette tendance est la plus importante : seuls les voisins des tout premiers rangs sont en effet utilisés dans un tel contexte comme nous l’avons déjà indiqué précédemment. Parmi l’ensemble des méthodes de fusion, *CombSum* se détache clairement pour toutes les mesures, l’effet étant particulièrement notable pour la précision au rang 1. L’utilisation des valeurs de similarité, dont la normalisation est indispensable dans le cas présent, s’avère donc supérieure à celle des rangs. Parmi les méthodes exploitant les rangs, *Borda* est l’option la moins bonne pour toutes les mesures. *Condorcet* se comporte quant à elle de façon intéressante mais souffre d’une faiblesse au rang 1, ce qui conduit à lui préférer RRF pour un usage général.

6 Discussion et travaux liés

Comme nous l’avons vu en introduction, une première façon d’améliorer un thésaurus distributionnel est d’intervenir sur le contenu des contextes distributionnels des mots ou sur la pondération de celui-ci. Les méthodes concernées sont donc dépendantes de la représentation de ces contextes. En définissant une méthode d’amélioration indépendante de la méthode initiale de constitution du thésaurus, l’approche que nous adoptons s’affranchit d’hypothèses *a priori* sur la représentation de l’information distributionnelle adoptée par cette méthode et peut donc en principe être appliquée à tout thésaurus distributionnel, quel que soit son mode de construction.

Le problème de l’amélioration d’un thésaurus distributionnel a aussi été abordé en exploitant ses spécificités, au-delà des données distributionnelles ayant permis sa constitution initiale. Ainsi, (Ferret, 2012) sélectionne de façon non supervisée un ensemble d’exemples positifs et négatifs de mots sémantiquement similaires en s’appuyant sur l’hypothèse de la symétrie de la relation de similarité sémantique entre mots. Cette sélection permet d’entraîner un classifieur utilisé ensuite pour réordonner les voisins du thésaurus. (Ferret, 2013b) s’inscrit dans le même cadre mais fait appel à un principe de sélection des exemples différent utilisant le contexte de mots composés similaires pour mettre au jour la similarité de leurs constituants. Dans les deux méthodes, les exemples sont de plus sélectionnés parmi les entrées de plus forte fréquence afin de minimiser les cas d’erreur. Plus globalement, l’idée sous-jacente est de produire un ensemble d’exemples comportant le moins d’erreurs possible afin de transposer, au travers de l’entraînement d’un classifieur statistique, les performances observées pour les entrées de forte fréquence vers les entrées de fréquence plus faible.

Cette approche se heurte néanmoins à une difficulté intrinsèque : les exemples étant issus du thésaurus, l’application d’un tel classifieur aux entrées de forte fréquence a tendance à faire décroître les résultats pour ces entrées. L’effet se comprend aisément : un classifieur statistique peut difficilement dépasser les performances de son ensemble d’apprentissage. Le fait de s’appuyer fortement sur le thésaurus initial a également pour effet négatif de biaiser les améliorations obtenues en faveur des voisins relevant de la proximité sémantique par opposition à ceux relevant davantage de la similarité sémantique du fait de la prévalence nette des premiers par rapport aux seconds dans le thésaurus initial. La méthode d’amélioration que nous considérons dans cet article ne présente pas ces deux difficultés : le critère de réordonnement des voisins

qu'elle exploite est extérieur au thésaurus distributionnel puisqu'il repose sur les occurrences des mots et leur contexte environnant et non sur une agrégation des informations liées à ces occurrences. (Claveau *et al.*, 2014) représente une autre façon de résoudre les problèmes posés par (Ferret, 2012) et (Ferret, 2013b) : l'idée est ici de représenter un thésaurus comme un graphe de voisinage distributionnel et d'exploiter les relations de réciprocité dans ce graphe, soit par l'entremise de fonctions d'agrégation, soit pour calculer un score de confiance d'une liste de voisins permettant ensuite de réordonner ces voisins. Les améliorations obtenues, données en pourcentage, sont à peu près comparables à celles observées dans notre cas pour les cooccurents graphiques et tendent à dépasser celles associées aux cooccurents syntaxiques. Plus globalement, il serait intéressant d'intégrer ces résultats dans les approches de fusion que nous avons déjà étudiées.

7 Conclusion et perspectives

Dans cet article, nous avons présenté une approche de réordonnement des voisins d'un thésaurus distributionnel fondée sur une modélisation discriminante du contexte distributionnel des mots. Plus précisément, cette modélisation repose sur la construction de classifieurs permettant de différencier en contexte un mot des autres mots dans une optique de tâche de pseudo-désambiguïsation sémantique. Dans le cas des voisins sémantiques d'une entrée de thésaurus, le classifieur construit pour l'entrée permet de détecter les voisins dont les contextes d'occurrence ne sont pas jugés compatibles avec les contextes de l'entrée et qui sont donc, en vertu de l'hypothèse distributionnelle, considérés comme sémantiquement distants. Le réordonnement des voisins s'effectue en final par le déclassement des voisins détectés comme non similaires à l'entrée. La méthode présentée a été testée pour l'anglais sur un large thésaurus de noms constitué à partir de cooccurents syntaxiques et a montré son efficacité par une amélioration significative des résultats dans le cadre d'une évaluation intrinsèque. Nous avons montré en outre que la combinaison de cette méthode avec les méthodes présentées dans (Ferret, 2012) et (Ferret, 2013b) selon une approche de type vote permet d'exploiter de façon intéressante les complémentarités de ces différentes méthodes.

Nous envisageons d'étendre ce travail en y intégrant la notion de sens de mot, à l'instar de (Reisinger & Mooney, 2010) ou de (Huang *et al.*, 2012). En l'absence de cette différenciation, le ou les sens majoritaires d'une entrée du thésaurus dans le corpus considéré ont tendance à être représentés de façon également très majoritaire parmi les voisins de cette entrée et à en limiter la diversité sur le plan sémantique. Dans le contexte de notre travail, cette extension devrait être assez directe puisqu'elle consisterait pour l'essentiel à transformer nos classifieurs de mots en contexte en véritables classifieurs de désambiguïsation sémantique.

Références

- ALEXANDRESCU A. & KIRCHHOFF K. (2007). Data-driven graph construction for semi-supervised graph-based learning in NLP. In *NAACL HLT 2007*, p. 204–211, Rochester, New York.
- BARONI M., DINU G. & KRUSZEWSKI G. (2014). Don't count, predict ! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL 2014*, Baltimore, Maryland.
- BRODA B., PIASECKI M. & SZPAKOWICZ S. (2009). Rank-Based Transformation in Measuring Semantic Relatedness. In *22nd Canadian Conference on Artificial Intelligence*, p. 187–190.
- CLAVEAU V., KIJAK E. & FERRET O. (2014). Improving distributional thesauri by exploring the graph of neighbors. In *COLING 2014*, p. 709–720, Dublin, Ireland.
- CORMACK G. V., CLARKE C. L. A. & BUETTCHER S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *SIGIR'09*, p. 758–759.
- CURRAN J. (2002). Ensemble methods for automatic thesaurus extraction. In *EMNLP 2002*, p. 222–229.
- CURRAN J. R. & MOENS M. (2002). Improvements in automatic thesaurus extraction. In *Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, p. 59–66, Philadelphia, USA.
- ERK K. & PADO S. (2010). Exemplar-based models for word meaning in context. In *ACL 2010*, p. 92–97, Uppsala, Sweden.
- FERRET O. (2010). Testing semantic similarity measures for extracting synonyms from a corpus. In *LREC'10*, Valletta, Malta.
- FERRET O. (2012). Combining bootstrapping and feature selection for improving a distributional thesaurus. In *ECAI 2012*, p. 336–341, Montpellier, France.

- FERRET O. (2013a). Identifying bad semantic neighbors for improving distributional thesauri. In *ACL 2013*, p. 561–571, Sofia, Bulgaria.
- FERRET O. (2013b). Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel. In *TALN 2013*, p. 48–61, Les Sables d’Olonne, France.
- GALE W. A., CHURCH K. W. & YAROWSKY D. (1992). Work on statistical methods for word sense disambiguation. In *AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, p. 54–60.
- GREFENSTETTE G. (1994). *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.
- HENESTROZA ANGUIANO E. & CANDITO M. (2012). Probabilistic lexical generalization for french dependency parsing. In *ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, p. 1–11, Jeju, Republic of Korea.
- HEYLEN K., PEIRSMANY Y., GEERAERTS D. & SPEELMAN D. (2008). Modelling Word Similarity : An Evaluation of Automatic Synonymy Extraction Algorithms. In *LREC 2008*, Marrakech, Morocco.
- HUANG E. H., SOCHER R., MANNING C. D. & NG A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *ACL’12*, p. 873–882.
- KANERVA P., KRISTOFERSON J. & HOLST A. (2000). Random indexing of text samples for latent semantic analysis. In *CogSci 2000*, p. 103–6 : Lawrence Erlbaum.
- KAZAMA J., DE SAEGER S., KURODA K., MURATA M. & TORISAWA K. (2010). A bayesian method for robust estimation of distributional similarities. In *ACL 2010*, p. 247–256, Uppsala, Sweden.
- KIELA D. & CLARK S. (2014). A systematic study of semantic vector space model parameters. In *2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, p. 21–30, Gothenburg, Sweden.
- LEE Y. K. & NG H. T. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *EMNLP 2002*, p. 41–48.
- LIN D. (1994). PRINCIPAR : An efficient, broad-coverage, principle-based parser. In *COLING’94*, p. 42–48, Kyoto, Japan.
- LIN D. (1998). Automatic retrieval and clustering of similar words. In *ACL-COLING’98*, p. 768–774, Montréal, Canada.
- MIKOLOV T., YIH W.-T. & ZWEIG G. (2013). Linguistic regularities in continuous space word representations. In *NAACL HLT 2013*, p. 746–751, Atlanta, Georgia.
- MILLER G. A. (1990). WordNet : An On-Line Lexical Database. *International Journal of Lexicography*, **3**(4).
- MIN B., SHI S., GRISHMAN R. & LIN C.-Y. (2012). Ensemble semantics for large-scale unsupervised relation extraction. In *EMNLP-CoNLL 2012*, p. 1027–1037, Jeju Island, Korea.
- NURAY R. & CAN F. (2006). Automatic ranking of information retrieval systems using data fusion. *Information Processing and Management*, **42**(3), 595–614.
- PADÓ S. & LAPATA M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, **33**(2), 161–199.
- PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global vectors for word representation. In *EMNLP 2014*, p. 1532–1543, Doha, Qatar.
- REISINGER J. & MOONEY R. J. (2010). Multi-prototype vector-space models of word meaning. In *HLT-NAACL 2010*, p. 109–117, Los Angeles, California.
- RIEDL M. & BIEMANN C. (2013). Scaling to large³ data : An efficient and effective method to compute distributional thesauri. In *EMNLP 2013*, p. 884–890, Seattle, Washington, USA.
- VAN DE CRUYS T. (2010). *Mining for Meaning. The Extraction of Lexico-semantic Knowledge from Text*. PhD thesis, University of Groningen, The Netherlands.
- WARD G. (1996). Moby thesaurus. Moby Project.
- WEEDS J. (2003). *Measures and Applications of Lexical Distributional Similarity*. PhD thesis, Department of Informatics, University of Sussex.
- WU S., CRESTANI F. & BI Y. (2006). Evaluating score normalization methods in data fusion. In *AIRS’06*, p. 642–648 : Springer-Verlag.
- YAMAMOTO K. & ASAKURA T. (2010). Even unassociated features can improve lexical distributional similarity. In *Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010)*, p. 32–39, Beijing, China.
- ZHITOMIRSKY-GEFFET M. & DAGAN I. (2009). Bootstrapping Distributional Feature Vector Quality. *Computational Linguistics*, **35**(3), 435–461.

Grammaires phrastiques et discursives fondées sur les TAG : une approche de D-STAG avec les ACG *

Laurence Danlos^{1,2,3} Aleksandre Maskharashvili^{4, 5, 6} Sylvain Pogodalla^{4, 5, 6, 7}

(1) Université Paris Diderot (Paris 7), Paris, F-75013, France

(2) ALPAGE, INRIA Paris–Rocquencourt, Paris, F-75013, France

(3) Institut Universitaire de France, Paris, F-75005, France

(4) INRIA, Villers-lès-Nancy, F-54600, France

(5) Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54500, France

(6) CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54500, France

(7) Heinrich Heine Universität, Düsseldorf, Allemagne

{laurence.danlos}{aleksandre.maskharashvili}{sylvain.pogodalla}@inria.fr

Résumé. Nous présentons une méthode pour articuler grammaire de phrase et grammaire de discours qui évite de recourir à une étape de traitement intermédiaire. Cette méthode est suffisamment générale pour construire des structures discursives qui ne soient pas des arbres mais des graphes orientés acycliques (DAG). Notre analyse s’appuie sur une approche de l’analyse discursive, Discourse Synchronous TAG (D-STAG), qui utilise les Grammaires d’Arbres Adjoint (TAG). Nous utilisons pour ce faire un encodage des TAG dans les Grammaires Catégorielles Abstraites (ACG). Cet encodage permet d’une part d’utiliser l’ordre supérieur pour l’interprétation sémantique afin de construire des structures qui soient des DAG et non des arbres, et d’autre part d’utiliser les propriétés de composition d’ACG pour réaliser naturellement l’interface entre grammaire phrastique et grammaire discursive. Tous les exemples proposés pour illustrer la méthode ont été implantés et peuvent être testés avec le logiciel approprié.

Abstract.

Sentential and Discourse TAG-Based Grammars: An ACG Approach to D-STAG

This article presents a method to interface a sentential grammar and a discourse grammar without resorting to an intermediate processing step. The method is general enough to build discourse structures that are direct acyclic graphs (DAG) and not only trees. Our analysis is based on Discourse Synchronous TAG (D-STAG), a Tree-Adjoining Grammar (TAG)-based approach to discourse. We also use an encoding of TAG into Abstract Categorical Grammar (ACG). This encoding allows us to express a higher-order semantic interpretation that enables building DAG discourse structures on the one hand, and to smoothly integrate the sentential and the discourse grammar thanks to the modular capability of ACG. All the examples of the article have been implemented and may be run and tested with the appropriate software.

Mots-clés : Syntaxe, sémantique, discours, grammaire, grammaire d’arbres adjoints, TAG, D-LTAG, D-STAG, grammaire catégorielle abstraite, ACG.

Keywords: Syntax, semantics, discourse, grammar, Tree-Adjoining Grammar, TAG, D-LTAG, D-STAG, Abstract Categorical Grammar, ACG.

1 Introduction

On considère généralement que la structure interne d’un texte, construite en particulier par l’utilisation de relations de discours, joue un rôle important dans son interprétation. Différentes techniques peuvent être mises en œuvre afin de calculer automatiquement cette structure. Certaines techniques se fondent sur la segmentation du discours en unités discursives élémentaires, puis en l’identification des relations susceptibles de les relier (Marcu, 2000; Soricut & Marcu, 2003). D’autres techniques utilisent des *grammaires du discours*, et en particulier des grammaires d’arbres (Polanyi & van den Berg, 1996; Gardent, 1997; Schilder, 1997). Les propriétés de ces dernières, notamment le besoin d’une opération d’adjonction, a conduit à utiliser le même formalisme grammatical des Grammaires d’Arbres Adjoints (TAG) (Joshi *et al.*, 1975; Joshi

*. Ce travail a bénéficié d’une aide de l’Agence Nationale de la Recherche portant la référence (ANR-12-CORD-0004).

& Schabes, 1997) pour la description des grammaires phrastiques et des grammaires discursives, notamment avec TAG pour le discours (D-LTAG) (Webber & Joshi, 1998; Forbes *et al.*, 2003; Webber, 2004; Forbes-Riley *et al.*, 2006).

Ces approches présentent deux caractéristiques importantes. Premièrement, bien qu’elles s’appuient sur un seul et unique formalisme grammatical, deux grammaires différentes sont utilisées d’une part pour l’analyse syntaxique et d’autre part pour l’analyse discursive. Cette dichotomie, qui implique la présence d’une étape intermédiaire de traitement, complique la modélisation des marqueurs discursifs dont l’usage peut être ambigu entre rôle syntaxique et rôle discursif. Il empêche également le traitement de ces ambiguïtés par les méthodes classiques adoptées en analyse. Deuxièmement, les structures discursives sont données directement par les arbres de dérivation. Or cette représentation sous forme d’arbres ne permet pas de rendre compte d’analyses pour lesquelles une structure de graphe orienté acyclique (DAG) pourrait sembler plus adaptée. Le formalisme grammatical réduit donc le choix des théories linguistiques modélisables.

Pour obvier au premier problème, Nakatsu & White (2010) proposent une grammaire unique rendant compte à la fois des phénomènes phrastiques et des phénomènes discursifs. L’approche se fonde sur une Grammaire Catégorielle Combinatoire du Discours (DCCG), une extension à la Grammaire Catégorielle Combinatoire (CCG) (Steedman, 2001; Steedman & Baldridge, 2011). Elle propose également une interface vers la représentation sémantique. Cependant, elle ne propose pas de modélisation pour les structures discursives représentées par un DAG (les structures « multi-parents » représentées aux figures 1(b) et 1(c)).

Pour répondre au deuxième problème et permettre la construction de structures multi-parents, Danlos (2009, 2011) définit Discourse Synchronous TAG (D-STAG), un formalisme fondé sur les TAG offrant une interprétation sémantique avec de l’ordre supérieur. Ainsi, D-STAG peut engendrer par exemple les structures discursives de la figure 1 qui correspondent aux exemples (1–4) tirés de (Danlos, 2009). Cependant, cette approche nécessite elle aussi un traitement en trois étapes. À la première étape, chaque phrase est analysée. À la deuxième étape, de même qu’en D-LTAG, une suite de clauses et de marqueurs de discours est extraite de la suite des analyses des phrases du texte pour construire ce qui est appelé une « forme normale discursive ». À la troisième étape, une grammaire TAG synchrone discursive analyse cette forme normale, produisant les représentations sémantiques souhaitées.

- (1) [Fred est grognon]₁ parce qu’[il a perdu ses clefs]₂. De plus, [il a raté son permis de conduire]₃.
- (2) [Fred est grognon]₁ parce qu’[il a mal dormi]₄. [Il a fait un cauchemar]₅.
- (3) [Fred est allé au supermarché]₆ parce que [le frigo était vide]₇. Ensuite, [il est allé au cinéma]₈.
- (4) [Fred est grognon]₁ parce que [sa femme est absente une semaine]₉. [Ceci prouve qu’il l’aime beaucoup]₁₀.

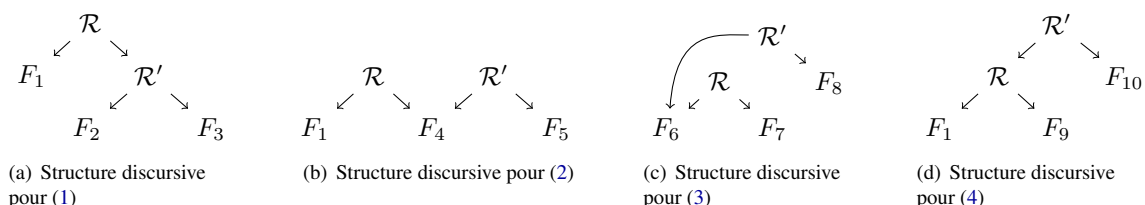


FIGURE 1 – Structures discursives possibles reliant trois unités de discours

Dans cet article, nous présentons une nouvelle approche permettant l’articulation d’une grammaire phrastique et d’une grammaire discursive à la D-STAG. Le but est d’éviter le recours à une étape intermédiaire de traitement résultant de l’indépendance complète entre les deux grammaires. Pour ce faire, nous utilisons le codage des TAG dans les Grammaires Catégorielles Abstraites (ACG) (de Groote, 2001), un cadre grammatical défini à partir de la théorie des types et du λ -calcul permettant l’encodage de nombreux formalismes grammaticaux.

Deux caractéristiques des ACG sont mises en œuvre : d’une part l’encodage des opérations d’adjonction et de substitution des grammaires TAG en ACG et l’interface syntaxe-sémantique qu’il offre pour les TAG (Pogodalla, 2004, 2009), et d’autre part la modularité de la composition d’ACG. Cette dernière est cruciale pour interfacer les deux grammaires et rendre compte des propriétés syntaxiques aussi bien que discursives des connecteurs sans recourir à une analyse en trois étapes. Nous reprenons la forme des arbres élémentaires discursifs des connecteurs de D-STAG, qui tiennent compte d’hypothèses psycho-linguistiques sur l’incrémentalité de l’extension du discours (par adjonction sur le discours déjà construit). Lors de l’interfaçage avec la grammaire phrastique, cette contrainte conduit à l’utilisation d’opérations qui sont au-delà de celles proposées par TAG mais qui s’expriment naturellement dans les ACG.

Cette intégration entre grammaire phrastique et grammaire discursive en une seule étape présente trois avantages. Premièrement, elle permet de réutiliser aussi bien des grammaires phrastiques TAG que des grammaires TAG discursives (nous ne connaissons pas de grammaire CCG à large couverture pour le français). Deuxièmement, elle permet d’éviter le découpage en plusieurs étapes et permet d’envisager à terme l’utilisation des techniques classiques de désambiguïsation prenant en compte à la fois le niveau phrastique et le niveau discursif. Les grammaires discursives sont en effet très ambiguës. Notre approche ne change pas cela. Elle présente la même ambiguïté que les approches en plusieurs étapes. Mais avoir un seul traitement d’analyse permet d’utiliser les techniques de désambiguïsation déjà existantes plutôt que de considérer l’interaction entre un modèle au niveau phrastique et un modèle au niveau discursif. Un troisième avantage est d’offrir un même cadre et une même grammaire non seulement pour l’analyse, mais aussi pour la génération de textes de plusieurs phrases reliées par des connecteurs de discours. Les ACG dites de second ordre sont en effet réversibles (Kanazawa, 2007). Un texte structuré peut donc s’obtenir par réalisation de surface.

Le choix de D-STAG nous permet aussi de construire des structures discursives sous forme de DAG. Pour ce faire, nous nous appuyons au niveau sémantique sur les analyses proposées dans (Danlos, 2009, 2011), et en particulier sur l’utilisation de l’ordre supérieur. C’est pourquoi nos exemples illustrent la chaîne d’interprétation jusqu’au calcul d’une représentation sémantique s’appuyant sur la théorie de représentation du discours segmentée (SDRT) (Asher & Lascarides, 2003) pour laquelle les structures ne sont pas nécessairement des arbres. Mais cela ne restreint pas la généralité de l’approche qui pourrait sans doute construire des représentations fondées sur la théorie de la structure rhétorique (RST) (Mann & Thompson, 1988) et des structures sous forme d’arbre, comme pour D-LTAG. Les λ -termes exemples, notamment sémantiques, pouvant être complexes (même si l’objet de l’article n’est pas d’expliquer comment ils fonctionnent, mais de décrire l’interface entre grammaire phrastique et grammaire discursive), nous proposons des grammaires qui les implantent. Ils peuvent donc être vérifiés en utilisant le logiciel de développement d’ACG proposé par l’équipe Sémagramme du LORIA¹.

Nous présentons brièvement les approches D-LTAG et D-STAG à la section 2. Nous montrons en particulier les difficultés de la prise en compte des propriétés discursives des adverbiaux médiaux qui ont conduit ces deux approches à recourir à un traitement intermédiaire. En section 3 nous présentons les ACG. Nous pouvons alors montrer en section 4 comment les TAG sont encodées dans les ACG, et, en suivant les mêmes principes, comment encoder D-STAG dans les ACG. La section 5 décrit la réalisation de l’interface entre ces grammaires dans les ACG. La section 6 présente finalement la manière d’obtenir la représentation sémantique associée.

2 Grammaires TAG pour le discours

Nous renvoyons le lecteur à (Webber & Joshi, 1998; Forbes *et al.*, 2003; Webber, 2004; Forbes-Riley *et al.*, 2006) pour une présentation en profondeur de D-LTAG et à (Danlos, 2009, 2011) pour une présentation de D-STAG et une comparaison entre ces deux approches. Ici, nous ne soulignons que leurs différences principales et leurs limitations.

D-LTAG D-LTAG propose trois familles principales d’arbres élémentaires pour rendre compte des différents effets des marqueurs discursifs sur la structure du discours. La première famille concerne les conjonctions de subordination. En s’appuyant sur leur structure prédicat-argument au niveau discursif, D-LTAG les modélise à l’aide d’arbres initiaux avec deux nœuds à substitution pour chacun des arguments comme le montre la figure 2(a)². Cela contraste avec la modélisation au niveau phrastique (syntaxique) de ces connecteurs, qui sont habituellement encodés par des arbres *auxiliaires* car ils ne font pas partie du domaine de localité des verbes des clauses dans lesquelles ils apparaissent. Il sont en effet adjoints aux nœuds S ou V de ces derniers. La seconde famille de connecteurs est utilisée pour étendre une clause avec des arbres auxiliaires ancrés par des conjonctions de coordination (ou par le connecteur vide ϵ). Le premier argument sémantique de la relation de discours correspond alors à l’unité de discours à laquelle l’arbre du connecteur est adjoint. Le second argument sémantique correspond à la clause qui est substituée au nœud pied de l’arbre du connecteur, comme le montre la figure 2(b). La troisième famille consiste également en arbres auxiliaires. Mais ces derniers sont associés à *une unique* clause comme le montre la figure 2(c). L’autre argument des connecteurs qui ancrent de tels arbres s’obtient par résolution anaphorique vers une unité de discours précédente (Webber *et al.*, 2003).

1. Les fichiers ACG d’exemples peuvent être téléchargés depuis <https://hal.inria.fr/hal-01145994/file/taln-2015-dstag-acg-examples.zip>. Le logiciel est disponible à <http://www.loria.fr/equipes/calligramme/acg/#Software>.

2. Le type Du correspondant à une unité de discours. Cette unité peut être élémentaire lorsque c’est une simple clause au niveau phrastique (exemples (2)) et (3), ou complexe lorsqu’il s’agit de plusieurs clauses elles-mêmes reliées par des relations de discours (exemples (1) et (4)).

Le traitement en trois étapes fonctionne de la manière suivante : tout d’abord, chaque phrase reçoit une analyse TAG (arbre dérivé et arbre de dérivation) par une TAG phrastique standard. Puis, dans un deuxième temps, chaque arbre de dérivation est traité pour identifier les connecteurs de discours (DC) possibles et leurs arguments d’un point de vue syntaxique. Ces derniers (un ou deux suivant le connecteur) sont ajoutés comme arbres initiaux de racine Du à la grammaire du discours, de même que l’arbre élémentaire ancré par le connecteur détecté (voir figure 3). Un procédé similaire réalise l’extraction des connecteurs adverbiaux médiaux. Enfin, dans un troisième temps, la séquence ainsi obtenue est analysée à l’aide de la grammaire discursive.

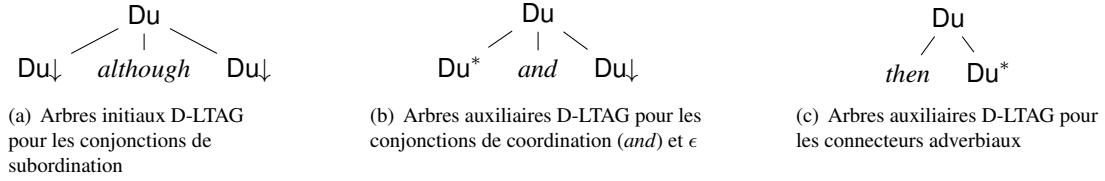


FIGURE 2 – Schémas des arbres élémentaires D-LTAG

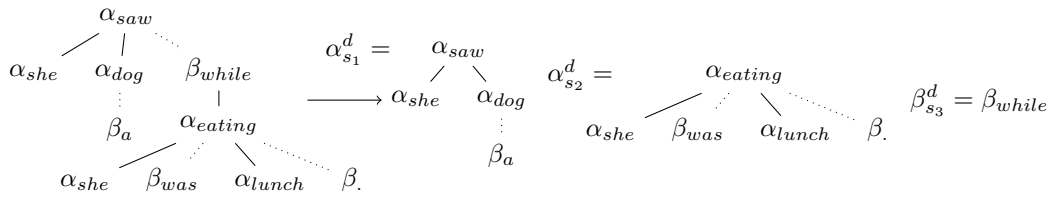


FIGURE 3 – Extraction des arbres élémentaires de la grammaire du discours

D-STAG En D-STAG, tous les connecteurs de discours sont modélisés avec des arbres auxiliaires qui sont adjoints au texte qu’ils étendent. Le contenu de la clause ajoutée est le second argument du connecteur et s’obtient par substitution. La figure 4 montre quelques schémas de ces arbres auxiliaires³. Suivant le principe des TAG synchronisées, chaque arbre élémentaire est associé à un arbre sémantique pour former une paire (arbre syntaxique, arbre sémantique). Ici, les arbres sémantiques ont une racine qui est soit t , le type des propositions, soit le type d’ordre supérieur $(t \rightarrow t) \rightarrow t$ nécessaire pour l’obtention des structures DAG. (Danlos, 2009) introduit deux termes $\Phi'_{\mathcal{R}}$ et $\Phi''_{\mathcal{R}}$ comme en (5) et la sémantique d’un connecteur associé à un arbre tel que ceux de la figure 4 s’obtient alors par : $\lambda p_4 p_3 p_2 p_1 p_0. p_4 (\phi (p_3 p_0) (p_2 p_1))$ où $\phi \in \{\Phi'_{\mathcal{R}}, \Phi''_{\mathcal{R}}\}$ et p_0 correspond à la sémantique de la clause auquel l’arbre est adjoint, p_1 celle de la clause qui est substituée, p_2, p_3 , et p_4 celles des arbres auxiliaires éventuellement adjoints au nœuds ②, ③, et ④ de l’arbre. En D-STAG, les connecteurs ancrent deux telles paires : une dont la partie sémantique met en jeu Φ' , et une dont la partie sémantique met en jeu Φ'' . (Danlos, 2009) détaille les raisons qui permettent à ces formules de rendre compte des phénomènes que nous souhaitons modéliser.

$$(5) \quad \begin{aligned} \Phi'_{\mathcal{R}} &= \lambda X Y. X(\lambda x. Y(\lambda y. \mathcal{R}(x, y))) \\ \Phi''_{\mathcal{R}} &= \lambda X Y P. X(\lambda x. Y(\lambda y. \mathcal{R}(x, y) \wedge P(x))) \end{aligned}$$

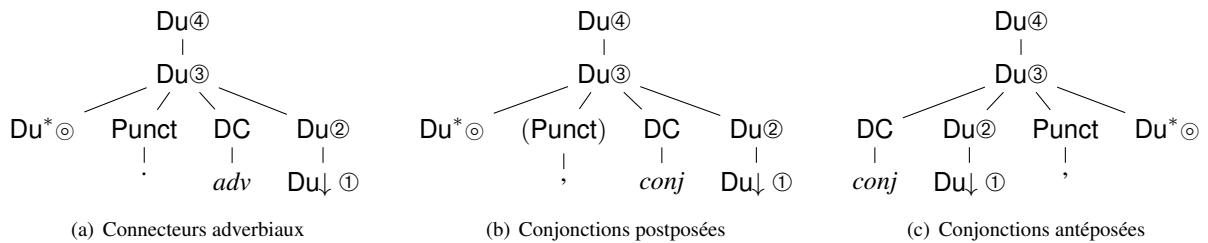


FIGURE 4 – Arbres élémentaires D-STAG

Comme en D-LTAG, avant que la grammaire discursive ci-dessus ne puisse s’appliquer, il est nécessaire de réaliser un pré-traitement pour obtenir une *forme normale discursive* qui peut seulement ensuite être analysée. Elle s’obtient de

3. Le connecteur vide ϵ est traité comme un adverbial.

manière similaire à celle exposée pour D-LTAG et est essentiellement motivée par les mêmes raisons : les connecteurs de discours et les unités de discours élémentaires ont besoin d’être identifiés pour être ancrés dans la grammaire discursive, et l’extraction des adverbiaux médiaux doit s’opérer lors de cette phase.

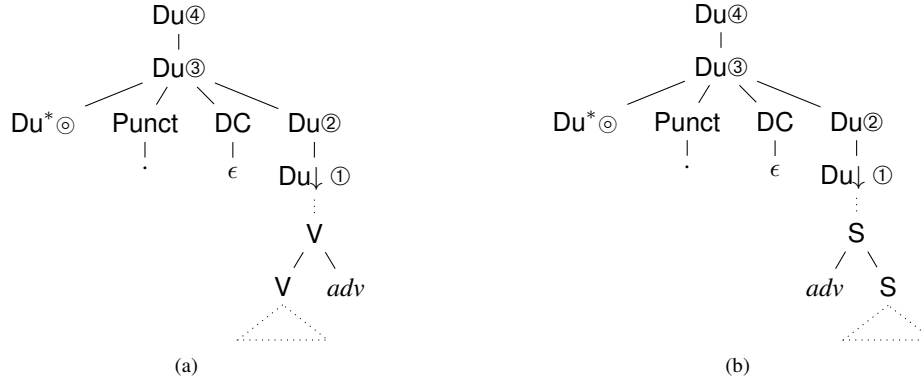


FIGURE 5 – Arbres auxiliaires pour les connecteurs de discours

Extraction des adverbiaux médiaux En observant les arbres de la figure 4 (le problème est similaire en D-LTAG), on constate que la clause hôte du connecteur de discours est *substituée* dans l’arbre élémentaire (au nœud $\text{Du}\downarrow$). Or, au niveau phrastique, les adverbiaux sont *adjoints* à cette clause. Pour les connecteurs en position initiale, les arbres D-STAG permettent d’engendrer les mêmes formes de surface, le connecteur étant placé juste *avant* la clause hôte, comme pour l’analyse phrastique. Ceci n’est plus vrai si le connecteur n’est pas en position initiale, par exemple en position médiale, et l’arbre discursif ne permet pas d’obtenir la même forme de surface que l’arbre phrastique. Dès lors, une forme intermédiaire comme la forme normale discursive est nécessaire : elle permet de déplacer les connecteurs adverbiaux médiaux en position initiale. Pour s’en passer, il faudrait pouvoir décrire une opération qui substitue une clause dans l’arbre discursif du connecteur vide et simultanément lui adjoint l’arbre auxiliaire de l’adverbial. La figure 5(a) décrit une telle opération. Les lignes en pointillé représenteraient une contrainte de dominance requise dans l’arbre qui est substitué en $\text{Du}\downarrow$. La description d’une telle contrainte ne nous semble pas possible en TAG⁴. Il est alors naturel d’utiliser la même approche pour les adverbiaux en position initiale afin qu’ils apparaissent également dans leur clause hôte (sous le nœud S, voir la figure 5(b)).

3 Grammaires Catégorielles Abstraites

Les ACG appartiennent à la famille des grammaires de types logiques. Plutôt qu’un formalisme grammatical propre, elles offrent un cadre grammatical dans lequel différents formalismes grammaticaux peuvent être encodés (de Groote & Pogodalla, 2004), notamment les TAG (de Groote, 2002). La définition des ACG s’appuie sur un petit nombre d’opérations de la théorie des types et du λ -calcul. Ces opérations se combinent à l’aide de règles simples de composition, offrant aux ACG une grande souplesse. En particulier, les ACG engendrent des langages de λ -termes linéaires, généralisant à la fois les langages de chaînes de caractères et les langages d’arbres.

Une caractéristique des ACG est de considérer les structures d’analyse des grammaires, le *langage abstrait*, de manière explicite plutôt que comme produit dérivé (ainsi les arbres de dérivation des grammaires non contextuelles ne sont-ils pas définis de manière intrinsèque. D’une manière similaire pour les TAG, les arbres de dérivation résultent de l’analyse mais ne définissent pas les langages engendrés). Ces structures sont ensuite interprétées à l’aide d’un *lexique* pour obtenir le *langage objet* des formes de surfaces. On appelle *vocabulaire* les *signatures d’ordre supérieur* définissant les éléments atomiques des langages (types atomiques à partir desquels sont construits inductivement les types implicatifs $\alpha \rightarrow \beta$ et constantes à partir desquelles sont construits les λ -termes). Étant donné un tel vocabulaire Σ , l’ensemble des λ -termes typés construits sur la signature est $\Lambda(\Sigma)$. Pour une ACG \mathcal{G} de lexique \mathcal{L} , on notera indifféremment $\mathcal{G}(a) = o$, $\mathcal{L}(a) = o$, ou $a := o$ l’interprétation du terme (resp. du type) abstrait a par le terme (resp. le type) objet o . On appelle *analyse ACG* l’opération qui permet de retrouver les structures (termes) abstraites a à partir d’un terme objet o . Cette opération consiste

4. C’est possible a priori avec les D-Tree Substitution Grammars (Rambow et al., 2001), mais les dépendances indiquées dans l’arbre de dérivation seraient différentes, l’interface syntaxe-sémantique des grammaires synchrones serait à définir, et les propriétés de réversibilité à établir.

à trouver le ou les antécédents a de o par le lexique (inversion du lexique). En adoptant ce point de vue, les arbres de dérivation des TAG sont représentés par des termes d'un langage abstrait, alors que les arbres dérivés ou le *yield* (la chaîne de caractères qui s'obtient en lisant les feuilles d'un arbre dérivé) sont représentés par des termes de différents langages objets. Il s'agit d'un langage objet d'arbres pour la représentation des arbres dérivés, et d'un langage objet de chaînes pour le *yield*. La classe des ACG de *second ordre*⁵ permet des analyses polynomiales dont les bornes de complexité correspondent aux meilleures connues (Kanazawa, 2008).

En définissant la manière dont les structures abstraites sont interprétées, le lexique joue un rôle crucial dans notre proposition. Tout d'abord, on peut noter qu'il est possible que deux interprétations partagent un même vocabulaire abstrait. Ainsi, un même terme peut être projeté sur deux structures objets différentes reposant sur deux vocabulaires objets différents, typiquement une forme de surface (chaîne de caractères) et une forme sémantique (formule logique). Cette *composition* permet par exemple l'interprétation sémantique des arbres de dérivation. Elle est illustrée par $\mathcal{G}_{derived\ trees}$ et $\mathcal{G}_{TAG\ sem.}$ qui partagent le vocabulaire Σ_{TAG} dans la figure 6. Ensuite, un autre mode de composition, correspondant à la composition fonctionnelle, est possible. Il s'agit de faire interpréter une deuxième fois, par un deuxième lexique, une structure intermédiaire résultant d'une première interprétation. Cette fois-ci, c'est le vocabulaire objet d'une première ACG qui est aussi le vocabulaire abstrait de la deuxième ACG. Cette composition est illustrée par $\mathcal{G}_{yield} \circ \mathcal{G}_{derived\ trees}$ à la figure 6. Nous utilisons cette composition pour relier les arbres de dérivation du niveau discursif aux arbres de dérivation phrastiques par le biais de $\mathcal{G}_{disc-clause\ int.}$, et, par la suite, aux arbres dérivés et aux chaînes de caractères.

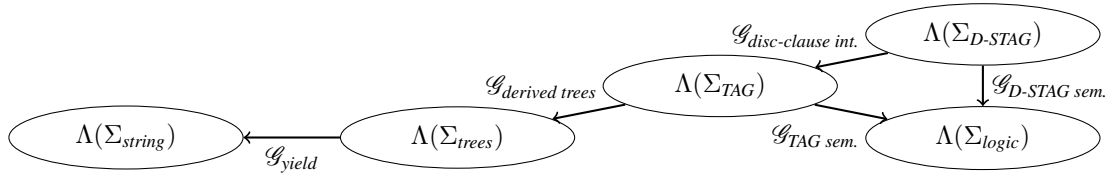


FIGURE 6 – Architecture ACG pour l'interface entre grammaire phrastique et grammaire discursive

4 Encodages des grammaires

TAG dans les ACG Pour encoder une TAG dans une ACG⁶, nous utilisons une signature d'ordre supérieur Σ_{TAG} dont les types atomiques comprennent les types suivant : $S, V, NP, S_A, V_A, \dots$ pour tout symbole non terminal X de la grammaire TAG qui peut faire l'objet d'une adjonction ou d'une substitution. Les types X sont utilisés pour les substitutions (soit pour le type de l'arbre qui est substitué, c'est-à-dire le type de sa racine, soit pour le type paramètre représentant un site de substitution), et les types X_A sont utilisés pour les adjonctions (soit pour le type de l'arbre auxiliaire de racine et nœud pied X , soit pour le type paramètre indiquant qu'une adjonction est possible). Pour chaque arbre élémentaire $\gamma_{entrée\ lex.}$, il y a une constante $C_{entrée\ lex.}$ dont le type dépend des sites de substitution et d'adjonction au sein de $\gamma_{entrée\ lex.}$, comme le montre la table 1 donnant l'encodage dans Σ_{TAG} des arbres de la figure 7⁷. Cette signature comprend également pour chaque type X_A une constante $I_X : X_A$ utilisée pour représenter une adjonction factice, c'est-à-dire lorsque que dans une dérivation TAG, aucune adjonction ne se produit sur un nœud X alors que ce dernier pourrait recevoir une adjonction.

Les termes construits sur cette signature sont interprétés par $\mathcal{G}_{derived\ trees}$ dans une autre signature dont le seul type atomique est le type τ des arbres. Dans cette signature, pour chaque symbole X d'arité n du vocabulaire permettant la construction

des arbres, il y a une constante $X_n : \overbrace{\tau \rightarrow \dots \rightarrow \tau}^{n\text{ fois}} \rightarrow \tau$. Nous ne décrivons pas ici l'interprétation définie par \mathcal{G}_{yield} qui interprète directement les arbres avec leur *yield*. Elle consiste simplement à interpréter les arbres comme des chaînes de caractères et chaque constante n -aire par la concaténation de ses n arguments.

Le lexique de la table 1 permet par exemple d'interpréter comme des chaînes de caractères deux dérivations avec une

5. L'ordre d'un type atomique a est $\text{ord}(a) = 1$, l'ordre d'un type complexe est $\text{ord}(\alpha \rightarrow \beta) = \max(\text{ord}(\alpha) + 1, \text{ord}(\beta))$. L'ordre d'un terme est l'ordre de son type, et l'ordre d'une ACG est l'ordre maximum de ses constantes abstraites. Dans une ACG d'ordre 2, tous les arguments des constantes sont des termes de type atomique, et pas des fonctions.

6. Pour une explication plus détaillée de cet encodage, ainsi que de l'interface syntaxe-sémantique, nous renvoyons le lecteur à (de Groote, 2001, 2002; Pogodalla, 2009).

7. Une constante $C_{entrée\ lex.}$ a comme type résultat le type X de la racine de $\gamma_{entrée\ lex.}$ si ce dernier est un arbre initial, et X_A si c'est un arbre auxiliaire. Comme paramètres, chaque site d'adjonction de label Y (y compris le nœud racine le cas échéant) introduit un paramètre de type Y_A et chaque site de substitution de label Y introduit un paramètre de type Y . Par convention, l'ordre des paramètres est donné en parcourant en premier les sites d'adjonction, avec un parcours en largeur, puis les sites de substitution, avec un parcours de gauche à droite.

adjonction au nœud S et au nœud V, ainsi que le montrent les équations (6) et (7). À noter que les types X de Σ_{TAG} sont interprétés comme des arbres (voir par exemple C_{Fred} et γ_{Fred}) alors que les types X_A sont interprétés comme des fonctions des arbres dans les arbres, permettant la modification du sous-arbre auquel elles sont appliquées.

$$(6) \quad t_1 = C_{allé \ à} (C_{ensuite}^S I_S) (C_{être \ aux.} I_V) C_{Fred} C_{Paris}$$

$$\mathcal{G}_{yield} \circ \mathcal{G}_{derived \ trees}(t_1) = ensuite + , + Fred + est + allé + à + Paris$$

$$(7) \quad t_2 = C_{allé \ à} I_S (C_{être \ aux.} (C_{ensuite}^V I_V)) C_{Fred} C_{Paris}$$

$$\mathcal{G}_{yield} \circ \mathcal{G}_{derived \ trees}(t_2) = Fred + est + ensuite + allé + à + Paris$$

Si l'on représente les termes t_1 et t_2 comme des arbres (un nœud étant étiqueté par le foncteur et ses fils par ses arguments qui sont éventuellement d'autres arbres), on obtient les arbres de la figure 8. Ceux-ci sont tout à fait similaires aux arbres de dérivation TAG à ceci près que toutes les adjonctions sont représentées (éventuellement par l'adjonction de l'identité I_X) et que l'ordre des paramètres est fixe, correspondant à chacun des sites, plutôt que variable avec l'adresse de Gorn du site concerné indiquée sur l'arête. L'arbre de gauche correspond aux adjonctions de l'adverbe et de l'auxiliaire sur le verbe, et l'arbre de droite correspond à une adjonction de l'adverbe sur l'auxiliaire, le tout étant adjoint au verbe.

Remarque. La notion de forme de surface f qui ancre un arbre γ_f en TAG correspond au fait que la constante objet f est un sous-terme de l'interprétation d'une constante abstraite C_f (par exemple « ensuite » ancre $C_{ensuite}^S$ car c'est un sous-terme de $\mathcal{G}_{yield} \circ \mathcal{G}_{derived \ trees}(C_{ensuite}) = \lambda S x. S (ensuite + x)$).

Constantes et types de Σ_{TAG}	Leur interprétation par $\mathcal{G}_{derived \ trees}$
$S, V, NP \dots$	τ
$S_A, V_A, NP_A \dots$	$\tau \rightarrow \tau$
$C_{Fred} : NP$	$\gamma_{Fred} : \tau$ $\gamma_{Fred} = NP_1 Fred$
$C_{être \ aux.} : V_A \rightarrow V_A$	$\gamma_{est} : (\tau \rightarrow \tau) \rightarrow \tau \rightarrow \tau$ $\gamma_{est} = \lambda A x. V_2 (A (V_1 est)) x$
$C_{allé \ à} : S_A \rightarrow V_A \rightarrow NP \rightarrow NP \rightarrow S$	$\gamma_{allé \ à} : (\tau \rightarrow \tau) \rightarrow (\tau \rightarrow \tau) \rightarrow \tau \rightarrow \tau \rightarrow \tau$ $\gamma_{allé \ à} = \lambda S A s c. S(S_2 s (VP_2(A (V_1 allé)) (PP_2 (Prep_1 à) c)))$
$C_{ensuite}^S : S_A \rightarrow S_A$	$\gamma_{ensuite}^S : (\tau \rightarrow \tau) \rightarrow \tau \rightarrow \tau$ $\gamma_{ensuite}^S = \lambda A x. A (S_2 (Adv_1 ensuite) x)$
$C_{ensuite}^V : V_A \rightarrow V_A$	$\gamma_{ensuite}^V : (\tau \rightarrow \tau) \rightarrow \tau \rightarrow \tau$ $\gamma_{ensuite}^V = \lambda A x. A (V_2 x (Adv_1 ensuite))$

TABLE 1 – Exemple de lexique pour l'encodage de la grammaire TAG de la figure 7

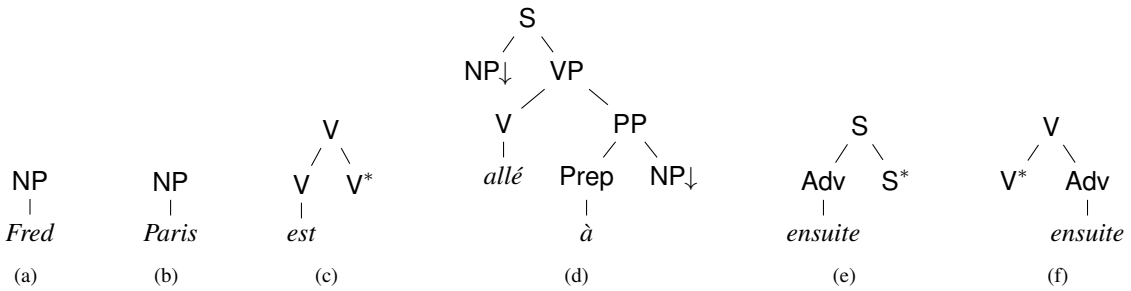


FIGURE 7 – Exemples d'arbres de la grammaire phrastique TAG

D-STAG dans les ACG L'encodage de D-STAG dans les ACG suit les mêmes principes que ceux énoncés dans le paragraphe précédent. Les arbres dérivés et de dérivation correspondent maintenant aux arbres de la grammaire de discours de la figure 4. On obtient donc l'équivalent des arbres de dérivation de la grammaire D-STAG. Les différences principales par rapport à l'encodage TAG apparaissent dans les interprétations :

- $\mathcal{G}_{disc-clause \ int.}$ implante l'interface entre la grammaire discursive et la grammaire phrastique en évitant, par simple composition, l'étape intermédiaire d'extraction et de construction de la forme normale de discours.



FIGURE 8 – Termes abstraits et arbres de dérivation TAG

- $\mathcal{G}_{TAG\ sem.}$ implante l'interprétation des structures de discours. Par rapport à (Danlos, 2011), afin d'avoir un système plus unifié, tous les discours sont interprétés avec un ordre supérieur. De plus, nous introduisons un nouveau type atomique ℓ pour les labels. De cette façon, nous pouvons construire des formules comme $l_1: \text{grumpy}(F) \wedge \exists x. l_2: \text{keys}(x) \wedge l_2: \text{lose}(F, x) \wedge l_3: \phi_{\text{Expl.}}(l_1, l_2)$ ⁸. De cette manière, ce sont les quantificateurs qui ont portée sur les relations de discours, plutôt que l'inverse comme dans une formule du type $\phi_{\text{Expl.}}(\text{grumpy}(F), \exists x. \text{keys}(x) \wedge \text{lose}(F, x))$.

5 Interface entre grammaires phrastiques et discursives

Le vocabulaire abstrait Σ_{D-STAG} utilise les mêmes types atomiques que Σ_{TAG} pour la grammaire phrastique (NP, V, V_A etc.), ainsi que de nouveaux types atomiques propres à la description du niveau discursif : Du , le type pour les unités de discours, et le type Du_A correspondant pour les sites d'adjonction. Une constante abstraite de Σ_{D-STAG} introduisant un marqueur discursif comme $d_{\text{parce que}}$ a typiquement le type $Du_A \rightarrow Du_A \rightarrow Du_A \rightarrow Du \rightarrow Du_A$ (que l'on notera par définition $DC \triangleq Du_A \rightarrow Du_A \rightarrow Du_A \rightarrow Du \rightarrow Du_A$). Ce type reflète la forme des arbres auxiliaires de D-STAG de la figure 4. Nous utilisons également un type T pour un texte (par opposition à phrase).

Le cœur de notre proposition réside dans l'ancrage de l'arbre, ou du terme, d'un marqueur de discours $d_{\text{conn.}}$ de type DC non pas par la chaîne de caractère conn. mais par l'arbre auxiliaire correspondant $C_{\text{conn.}}$ (voir la remarque de la section 4) par l'intermédiaire de l'ACG $\mathcal{G}_{\text{disc-clause int.}}$. Cela nous permet de spécifier que, dans l'arbre qui est substitué (l'argument de type Du qui est paramètre de $d_{\text{conn.}}$, ou, pour le dire dans les termes de D-STAG, celui qui est substitué au nœud Du_{\downarrow}), il faut qu'il y ait également adjonction de l'arbre $C_{\text{conn.}}$ au niveau phrastique. Cela se traduit par exemple dans $\mathcal{G}_{\text{disc-clause int.}}$ (voir table 2) par le fait que l'interprétation du terme abstrait d_{ensuite}^V fait appel à une adjonction sur son paramètre s de l'arbre auxiliaire C_{ensuite}^V . C'est ainsi que l'arbre discursif d_{ensuite}^V est relié à l'arbre phrastique C_{ensuite}^V de l'adverbe en position médiale. On rend ainsi compte de la contrainte de dominance requise (voir figure 5) dans l'arbre qui est substitué en Du_{\downarrow} . Cette contrainte, qui n'est pas à notre connaissance exprimable en TAG, l'est naturellement avec les ACG.

Pour permettre cette adjonction, le type Du des unités de discours est interprété comme une clause qui peut encore être l'objet d'une adjonction d'un arbre auxiliaire de conjonction de subordination ou d'un adverbial en position initiale ou médiale. Cela revient à l'interpréter comme un type de second ordre $S_A \rightarrow V_A \rightarrow S$ où des adjonctions sur les nœuds S et V sont encore possibles⁹. Plus précisément, nous interprétons Du comme $S_A \rightarrow (V_A \rightarrow V_A) \rightarrow S$ pour rendre compte de la possibilité de ce que les adverbiaux médiaux peuvent être entre d'autres adverbes¹⁰. Dès lors, au niveau de la grammaire discursive, les verbes intransitifs par exemple auront le type $S_A \rightarrow V_A \rightarrow V_A \rightarrow NP \rightarrow S$ plutôt que $S_A \rightarrow V_A \rightarrow NP \rightarrow S$ pour permettre l'adjonction au-dessus ou au-dessous du marqueur discursif adverbial. On obtient alors les types et les interprétations décrites dans la table 2¹¹. Bien qu'un même symbole (S par exemple) puisse apparaître aussi bien à gauche qu'à droite du signe $:=$, il est important de noter que le symbole de gauche appartient au

8. La notation $l:p$ se lit l est un label de p . Un même label peut étiqueter plusieurs propositions. Cela signifie que ces dernières appartiennent à une même clause. Par ailleurs, les prédicats des formules sémantiques sont donnés en anglais pour les distinguer de la partie syntaxiques en français.

9. Une autre solution serait de définir $DC \triangleq Du_A \rightarrow Du_A \rightarrow Du_A \rightarrow (S_A \rightarrow V_A \rightarrow Du) \rightarrow Du_A$ avec un type fonctionnel comme quatrième paramètre, au lieu de Du . Mais l'ACG ne serait plus de second ordre. L'analyse ACG resterait décidable car la grammaire serait toujours lexicalisée pour les constantes d'ordre supérieur, et donc inversible. Cependant, la complexité d'analyse de ces grammaires n'est en général pas polynomiale.

10. Une analyse plus précise pour vérifier que cette configuration existe en français devra être menée. Cela généralise en tout cas la possibilité pour l'adverbe marqueur de discours de se trouver aussi bien avant qu'après un autre adverbe (comme les occurrences de *généralement ensuite* ou de *ensuite doucement* sur le web semblent l'indiquer) et permet également de traiter les auxiliaires par une adjonction (*a ensuite doucement ...*). Bien entendu, il faudrait faire de même pour les adverbiaux en position initiale, mais pour des raisons de concision nous ne le faisons pas ici.

11. mod et cons sont deux opérateurs qui ne servent qu'à juxtaposer les arbres TAG de dérivation phrastique des unités de discours élémentaires. Ils sont interprétés de la manière suivante : $\text{mod} := \lambda s\ m.m\ s$ (c'est-à-dire qu'il réalise l'adjonction effective sur l'arbre dérivé) et $\text{cons} := \lambda s_1\ s_2\ s_3\ s\ x.s_1(s_2(S_3\ x \cdot (s_3\ s)))$ qui construit un arbre dérivé en insérant un point (quand c'est nécessaire) entre les arbres dérivés correspondant aux unités élémentaires de discours.

vocabulaire abstrait Σ_{D-STAG} tandis que les types ou termes de droite sont construits sur le vocabulaire objet Σ_{TAG} .

Si d'un point de vue théorique la complexité de l'analyse augmente à cause de l'interprétation de types atomiques par des types d'ordre plus élevés, l'étude précise de la complexité reste à faire. La complexité est polynomiale puisque les ACG de second ordre sont équivalentes aux grammaires non contextuelles multiples (MCFG) et aux systèmes de réécriture linéaires (m -LCFRS). Mais d'autres propriétés comme le bon parenthésage (*well-nestedness*), le rang, ou l'étendue (*fan-out*) de la nouvelle grammaire pourrait permettre de raffiner ce résultat.

	$NP ::= NP$	$N ::= N$	$Du ::= S_A \rightarrow (V_A \rightarrow V_A) \rightarrow S$
	$NP_A ::= NP_A$	$N_A ::= N_A$	$Du_A ::= S_A$
	$V ::= V$	$T ::= S$	$S ::= S_A \rightarrow (V_A \rightarrow V_A) \rightarrow S$
	$V_A ::= V_A \rightarrow V_A$		$S_A ::= S_A \rightarrow S_A$
I_X	$: X_A$		$:= \lambda P.P$
d_{Fred}	$: NP$		$:= C_{Fred}$
$d_{être\ aux.}$	$: V_A \rightarrow V_A$		$:= \lambda a\ v. C_{être\ aux.}(a\ v)$
$d_{allé\ à}$	$: S_A \rightarrow V_A \rightarrow V_A \rightarrow NP \rightarrow NP \rightarrow S$		$:= \lambda s\ a_1\ a_2\ s\ o\ c\ m. C_{allé\ à}(S\ c)(a_2(m(a_1\ I_V)))\ s\ o$
$d_{initial\ anchor}$	$: S \rightarrow Du_A \rightarrow T$		$:= \lambda s\ m. mod(s\ I_S(\lambda x.x))\ m$
d_{anchor}	$: S \rightarrow Du_A \rightarrow Du$		$:= \lambda s\ m\ d_s\ d_v. mod(s\ d_s\ d_v)\ m$
$d_{ensuite}^S$	$: DC$		$:= \lambda d_1\ d_2\ d_3\ s. cons\ d_1\ d_2\ d_3\ (s\ (C_{ensuite}^S\ I_S)(\lambda x.x))$
$d_{ensuite}^V$	$: DC$		$:= \lambda d_1\ d_2\ d_3\ s. cons\ d_1\ d_2\ d_3\ (s\ I_S\ C_{ensuite}^V)$

TABLE 2 – Interprétation par $\mathcal{G}_{disc-clause\ int.}$ pour l'interface entre phrase et discours

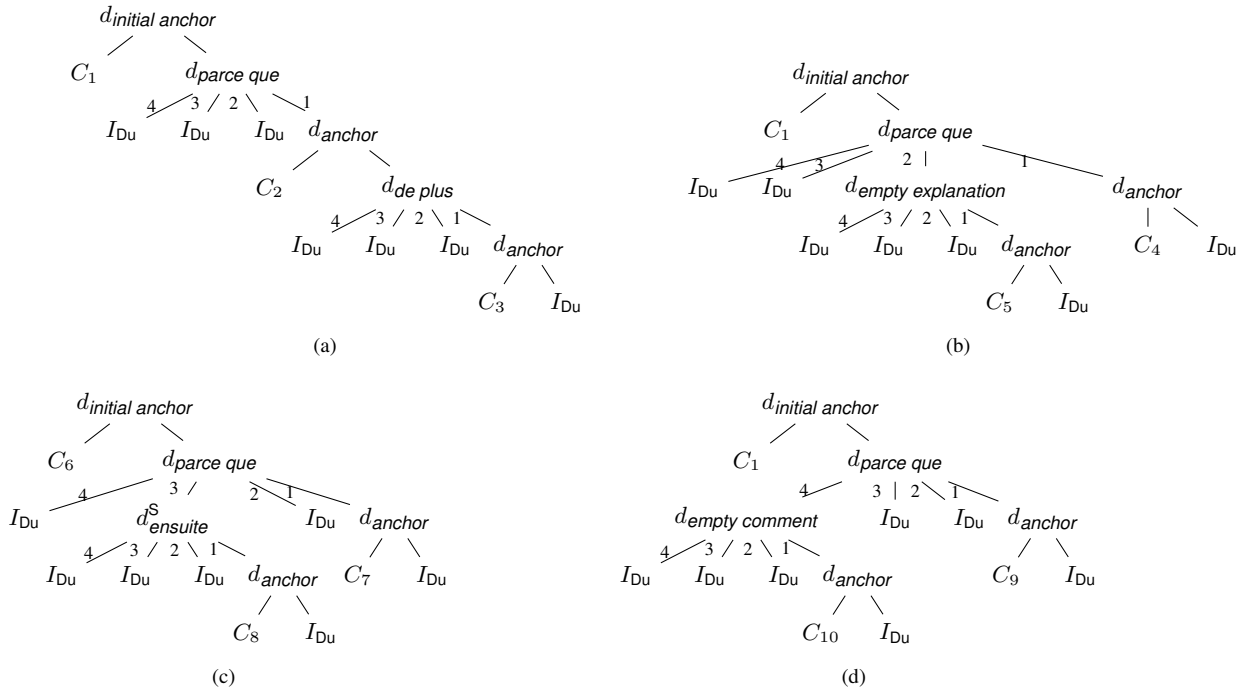


FIGURE 9 – Arbres de dérivation pour le discours. $\mathcal{G}_{D-STAG\ sem.}$ les interprète, le cas échéant, en DAG (voir (21))

Nous ne pouvons donner ici la définition de tous les termes C_i qui définissent les arbres de dérivation discursifs pour chacune des clauses correspondant à l'unité discursive i dans les exemples (1–4) car il faudrait définir le lexique entier¹². Nous ne donnons que C_8'' en (8) (qui diffère de C_8 seulement par le complément de lieu, ici un nom propre). On voit qu'en appliquant le résultat à $(C_{ensuite}^S\ I_S)$ et $(\lambda x.x)$ (resp. à I_S et $C_{ensuite}^V$), c'est-à-dire en remplaçant adv_s et adv_v , comme le fait $d_{ensuite}^S$ (resp. $d_{ensuite}^V$), on obtient le terme t_1 donné en (6) (resp. t_2 donné en (7)).

$$\begin{aligned}
 (8) \quad C_8'' &= d_{allé\ à}\ I_S\ I_V\ (d_{être\ aux.}\ I_V)\ d_{Fred}\ d_{paris} \\
 &:= \lambda^0 adv_s\ adv_v. C_{allé\ à}\ adv_s\ (C_{être\ aux.}\ (adv_v\ I_V))\ C_{Fred}\ C_{paris}
 \end{aligned}$$

12. Cela est fait dans les fichiers d'exemples.

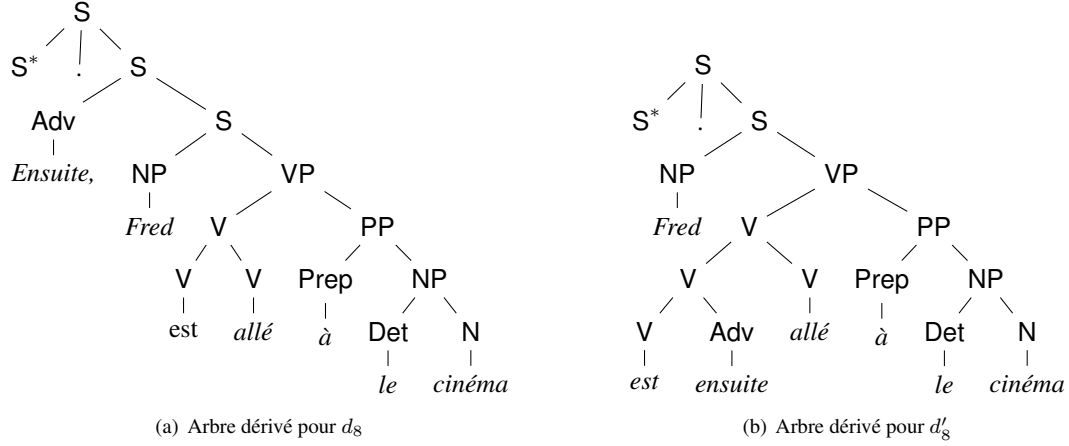


FIGURE 10 – Interprétation comme arbres dérivés de d_8 et d'_8

À l'aide des lexiques définis dans les tables 1 et 2, et avec les définitions (9) and (10), on peut vérifier que l'on obtient les interprétations (11)–(16). Le terme d_8 (resp d'_8) correspond à l'arbre auxiliaire discursif (comme le montre le type Du_A) qui étend le discours précédent par *Ensuite, il est allé au cinéma* (resp. *Il est ensuite allé au cinéma*).

- (9) $d_8 = d_{\text{ensuite}}^S I_{\text{Du}} I_{\text{Du}} I_{\text{Du}} (d_{\text{anchor}} C_8 I_{\text{Du}}) : \text{Du}_A$
- (10) $d'_8 = d_{\text{ensuite}}^V I_{\text{Du}} I_{\text{Du}} I_{\text{Du}} (d_{\text{anchor}} C_8 I_{\text{Du}}) : \text{Du}_A$
- (11) $\mathcal{G}_{\text{disc-clause int.}}(d_8) = \text{cons } I_S I_S I_S (\text{mod } (C_{\text{allé à}} (C_{\text{ensuite}}^S I_S) (C_{\text{être aux.}} I_V) C_{\text{Fred}} (C_{\text{le}} (C_{\text{cinéma}} I_N))) I_S) : \text{S}_A$
- (12) $\mathcal{G}_{\text{derived trees}} \circ \mathcal{G}_{\text{disc-clause int.}}(d_8) = [\text{voir la représentation figure 10(a)}]$
- (13) $\mathcal{G}_{\text{yield}} \circ \mathcal{G}_{\text{derived trees}} \circ \mathcal{G}_{\text{disc-clause int.}}(d_8) = \lambda x.x + . + \text{Ensuite} + , + \text{Fred} + \text{est} + \text{allé} + \text{à} + \text{le} + \text{cinéma} : \sigma \rightarrow \sigma$
- (14) $\mathcal{G}_{\text{disc-clause int.}}(d'_8) = \text{cons } I_S I_S I_S (\text{mod } (C_{\text{allé à}} I_S (C_{\text{être aux.}} (C_{\text{ensuite}}^V I_V)) C_{\text{Fred}} (C_{\text{le}} (C_{\text{cinéma}} I_N))) I_S) : \text{S}_A$
- (15) $\mathcal{G}_{\text{derived trees}} \circ \mathcal{G}_{\text{disc-clause int.}}(d'_8) = [\text{voir la représentation figure 10(b)}]$
- (16) $\mathcal{G}_{\text{yield}} \circ \mathcal{G}_{\text{derived trees}} \circ \mathcal{G}_{\text{disc-clause int.}}(d'_8) = \lambda x.x + . + \text{Fred} + \text{est} + \text{ensuite} + \text{allé} + \text{à} + \text{le} + \text{cinéma} : \sigma \rightarrow \sigma$

De la même manière, les arbres de dérivation discursifs de la figure 9¹³, correspondant aux analyses des exemples (1–4), sont donnés par les termes d_1 , d_2 , d_3 et d_4 des équations (17)–(20) ($d_{\text{empty explanation}}$ et $d_{\text{empty comment}}$ ancrent tous les deux le connecteur vide, mais avec la sémantique *Explanation* pour le premier, et *Comment* pour le second).

- (17) $d_1 = d_{\text{initial anchor}} C_1 (d_{\text{parce que}} I_{\text{Du}} I_{\text{Du}} I_{\text{Du}} (d_{\text{anchor}} C_2 (d_{\text{de plus}} I_{\text{Du}} I_{\text{Du}} I_{\text{Du}} (d_{\text{anchor}} C_3 I_{\text{Du}}))))$
- (18) $d_2 = d_{\text{initial anchor}} C_1 (d_{\text{parce que}} I_{\text{Du}} I_{\text{Du}} (d_{\text{empty explanation}} I_{\text{Du}} I_{\text{Du}} I_{\text{Du}} (d_{\text{anchor}} C_5 I_{\text{Du}})) (d_{\text{anchor}} C_4 I_{\text{Du}}))$
- (19) $d_3 = d_{\text{initial anchor}} C_6 (d_{\text{parce que}} I_{\text{Du}} (d_{\text{ensuite}}^S I_{\text{Du}} I_{\text{Du}} I_{\text{Du}} (d_{\text{anchor}} C_8 I_{\text{Du}})) I_{\text{Du}} (d_{\text{anchor}} C_7 I_{\text{Du}}))$
- (20) $d_4 = d_{\text{initial anchor}} C_1 (d_{\text{parce que}} (d_{\text{empty comment}} I_{\text{Du}} I_{\text{Du}} I_{\text{Du}} (d_{\text{anchor}} C_{10} I_{\text{Du}})) I_{\text{Du}} I_{\text{Du}} (d_{\text{anchor}} C_9 I_{\text{Du}}))$

6 Sémantique

Pour l'interprétation sémantique, nous avons modifié les types (et par conséquent les formules) proposés dans (Danlos, 2009). Ainsi, en plus des types e pour les entités, t pour les valeurs de vérité, et ℓ pour les labels, nous définissons $qnp \triangleq (e \rightarrow \ell \rightarrow t) \rightarrow \ell \rightarrow t$ et $\ell tt \triangleq (\ell \rightarrow t) \rightarrow t$. Nous utilisons les connecteurs logiques habituels, et les constantes sémantiques sont données en anglais. La table 3 montre quelques interprétations¹⁴. L'objet de cet article n'étant pas de montrer en quoi les formules de (Danlos, 2009, 2011) fonctionnent, nous laissons le lecteur vérifier, éventuellement à

13. On a ajouté sur les arêtes les numéros de sites d'adjonction ou de substitution des arbres élémentaires des connecteurs de la figure 4.

14. Notons que d_{ensuite}^S et d_{ensuite}^V ont maintenant la même interprétation.

l'aide des fichiers exemples, que l'égalité (21) est correcte. Elle correspond à la structure discursive de la figure 1(c) : les propositions étiquetées par l_6 (resp. l_7 et l_8) correspondent à la clause C_6 (resp. C_7 et C_8), et $l_{\mathcal{R}}$ et $l_{\mathcal{R}'}$ aux relations \mathcal{R} et \mathcal{R}' . Cela décrit bien un DAG. Nous ne montrons pas ici les trois autres dérivations qui peuvent également être interprétées par $\mathcal{G}_{D-STAG\ sem.}$ avec des formules représentant les trois autres structures de la figure 1.

$$(21) \quad \mathcal{G}_{D-STAG\ sem.}(d_3) = \exists_{\ell} l_6 \ l_{\mathcal{R}} \ l_{\mathcal{R}'} . \exists ! x. l_6 : \text{supermarket}(x) \wedge l_6 : \text{go_to}(\text{fred}, x) \wedge \\ (\exists_{\ell} l_8 . (\exists ! x. l_8 : \text{movies}(x) \wedge l_8 : \text{go_to}(\text{fred}, x)) \wedge ((\exists_{\ell} l_7 . (\exists ! x. l_7 : \text{fridge}(x) \wedge \\ l_7 : \text{empty}(x)) \wedge (l_{\mathcal{R}} : \phi_{\text{Expl.}}(l_6, l_7) \wedge \top)) \wedge l_{\mathcal{R}'} : \phi_{\text{Nar}}(l_6, l_8)))$$

NP	:= qnp	N	:= e \rightarrow ℓ \rightarrow t	Du	:= ℓ tt
NP _A	:= qnp \rightarrow qnp	N _A	:= (e \rightarrow ℓ \rightarrow t) \rightarrow e \rightarrow ℓ \rightarrow t	Du _A	:= ℓ tt \rightarrow ℓ tt
T	:= t	V _A	:= t \rightarrow t	S	:= (t \rightarrow t) \rightarrow ℓ \rightarrow t
				S _A	:= t \rightarrow t
I_X	:= $\lambda P. P$				
d_{Fred}	:= $\lambda P \ l. P \ \text{fred} \ l$				
$d_{\text{allé à}}$:= $\lambda S \ a_1 \ a_2 \ s \ o \ m \ l. S \ (s \ (\lambda x \ l_1. o \ (\lambda y \ l_2. a_2 \ (m \ (a_1 \ (l_2 : \text{go_to} \ x \ y)))) \ l_1) \ l)$				
$d_{\text{initial anchor}}$:= $\lambda s \ m. \exists_{\ell} l. m \ (\lambda Q. (s \ (\lambda x. x) \ l) \wedge (Q \ l)) \ (\lambda l'. \top)$				
d_{anchor}	:= $\lambda s \ m \ P. \exists_{\ell} l. m \ (\lambda Q. (s \ (\lambda x. x) \ l) \wedge (Q \ l)) \ P$				
d_{ensuite}^S	:= $\lambda d_4 \ d_3 \ d_2 \ s \ f. d_4 \ (\lambda P. \exists_{\ell} l. d_3 \ f \ (\lambda x. d_2 \ s \ (\lambda y. (P \ x) \wedge (l : \phi_{\text{Nar}}(x, y)))))$				
d_{ensuite}^V	:= $\lambda d_4 \ d_3 \ d_2 \ s \ f. d_4 \ (\lambda P. \exists_{\ell} l. d_3 \ f \ (\lambda x. d_2 \ s \ (\lambda y. (P \ x) \wedge (l : \phi_{\text{Nar}}(x, y)))))$				

TABLE 3 – Interprétation par $\mathcal{G}_{D-STAG\ sem.}$ pour la sémantique du discours

7 Conclusion

Cet article montre comment interfacier deux grammaires phrastiques et discursives fondées sur les TAG sans recourir à un découpage du processus d'analyse. Cette approche est suffisamment générale pour permettre d'engendrer des structures de graphe qui ne soient pas de simples arbres. Nous avons aussi montré comment implanter cette interface avec des grammaires ACG pour associer formes de surface et représentations sémantiques pour des textes de plusieurs phrases. Ainsi, nous bénéficions d'un mécanisme autorisant l'analyse et la génération prenant en compte les relations de discours. Nous n'avons pas exposé ici comment il est possible de modéliser les modifications des connecteurs (... *probablement parce qu'il pleut*). Notre travail futur portera sur les connecteurs multiples (... *parce qu'ensuite il s'est rendu compte qu'il était fauché*) et sur la manière d'intégrer des modèles de désambiguïsation phrastiques et discursifs pour tirer partie de l'analyse en une seule phase. Enfin, nous pensons que l'utilisation d'une même grammaire permettra de mieux modéliser les problèmes à l'interface entre syntaxe et sémantique lorsqu'on prend en compte les verbes d'attitude propositionnelle et leur position dans la phrase (Danlos, 2013).

Références

- ASHER N. & LASCARIDES A. (2003). *Logics of conversation*. Cambridge University Press.
- DANLOS L. (2009). D-STAG : un formalisme d'analyse automatique de discours basé sur les TAG synchrones. *T.A.L.*, **50**(1), 111–143. <http://hal.inria.fr/inria-00524743/en/>.
- DANLOS L. (2011). D-STAG : a formalism for discourse analysis based on SDRT and using Synchronous TAG. In P. DE GROOTE, M. EGG & L. KALLMEYER, Eds., *14th conference on Formal Grammar - FG 2009*, volume 5591 of *LNCS/LNAI*, p. 64–84 : Springer. doi :10.1007/978-3-642-20169-1_5.
- DANLOS L. (2013). Connecteurs de discours adverbiaux : Problèmes à l'interface syntaxe-sémantique. *Linguisticae Investigationes*, **36**(2), 261–275. doi :10.1075/li.36.2.05dan.
- DE GROOTE P. (2001). Towards Abstract Categorical Grammars. In *Association for Computational Linguistics, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Conference*, p. 148–155. <http://aclweb.org/anthology/P01-1033>.
- DE GROOTE P. (2002). Tree-Adjoining Grammars as Abstract Categorical Grammars. In *Proceedings of the Sixth International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6)*, p. 145–150 : Università di Venezia. <http://www.loria.fr/equipes/calligramme/acg/publications/2002-tag+6.pdf>.

- DE GROOTE P. & POGODALLA S. (2004). On the expressive power of Abstract Categorical Grammars : Representing context-free formalisms. *Journal of Logic, Language and Information*, **13**(4), 421–438. doi :[10.1007/s10849-004-2114-x](https://doi.org/10.1007/s10849-004-2114-x).
- FORBES K., MILTSAKAKI E., PRASAD R., SARKAR A., JOSHI A. K. & WEBBER B. L. (2003). D-LTAG system : Discourse parsing with a Lexicalized Tree-Adjoining Grammar. *Journal of Logic, Language and Information*, **12**(3), 261–279. doi :[10.1023/A:1024137719751](https://doi.org/10.1023/A:1024137719751). Special Issue : Discourse and Information Structure.
- FORBES-RILEY K., WEBBER B. L. & JOSHI A. K. (2006). Computing discourse semantics : The predicate-argument semantics of discourse connectives in D-LTAG. *Journal of Semantics*, **23**(1), 55–106. doi :[10.1093/jos/ffh032](https://doi.org/10.1093/jos/ffh032).
- GARDENT C. (1997). *Discourse Tree Adjoining Grammar*. CLAUS Report 89, Universit, Saarbr. <ftp://ftp.coli.uni-sb.de/pub/coli/claus/claus89.ps>.
- JOSHI A. K., LEVY L. S. & TAKAHASHI M. (1975). Tree adjunct grammars. *Journal of Computer and System Sciences*, **10**(1), 136–163. doi :[10.1016/S0022-0000\(75\)80019-5](https://doi.org/10.1016/S0022-0000(75)80019-5).
- JOSHI A. K. & SCHABES Y. (1997). Tree-adjoining grammars. In G. ROZENBERG & A. K. SALOMAA, Eds., *Handbook of formal languages*, volume 3, chapter 2. Springer.
- KANAZAWA M. (2007). Parsing and generation as datalog queries. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, p. 176–183, Prague, Czech Republic : Association for Computational Linguistics. <http://www.aclweb.org/anthology/P07-1023>.
- KANAZAWA M. (2008). A prefix-correct earley recognizer for multiple context-free grammars. In *Proceedings of the Ninth International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+9)*, p. 49–56, Tuebingen, Germany. <http://tagplus9.cs.sfu.ca/papers/Kanazawa.pdf>.
- MANN W. C. & THOMPSON S. A. (1988). Rhetorical Structure Theory : Toward a functional theory of text organization. *Text*, **8**(3), 243–281. doi :[10.1515/text.1.1988.8.3.243](https://doi.org/10.1515/text.1.1988.8.3.243).
- MARCU D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press.
- NAKATSU C. & WHITE M. (2010). Generating with discourse combinatory categorical grammar. *Linguistic Issues in Language Technology*, **4**. <http://elanguage.net/journals/lilt/article/view/1277>.
- POGODALLA S. (2004). Computing Semantic Representation : Towards ACG Abstract Terms as Derivation Trees. In *Seventh International Workshop on Tree Adjoining Grammar and Related Formalisms - TAG+7*, p. 64–71, Vancouver, BC, Canada. <https://hal.inria.fr/inria-00107768>.
- POGODALLA S. (2009). Advances in Abstract Categorical Grammars : Language Theory and Linguistic Modeling. ESSLLI 2009 Lecture Notes, Part II. <https://hal.inria.fr/hal-00749297>.
- POLANYI L. & VAN DEN BERG M. H. (1996). Discourse structure and discourse interpretation. In P. J. E. DEKKER & M. STOKHOF, Eds., *Proceedings of the Tenth Amsterdam Colloquium : ILLC/Department of Philosophy, University of Amsterdam*. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.56.221>.
- RAMBOW O., VIJAY-SHANKER K. & WEIR D. (2001). D-Tree Substitution Grammars. *Computational Linguistics*, **27**(1), 87–121. doi :[10.1162/089120101300346813](https://doi.org/10.1162/089120101300346813).
- SCHILDER F. (1997). Tree discourse grammar or how to get attached to a discourse ? In *Proceedings of the Tilburg Conference on Formal Semantics (IWCS-1997)*, p. 261–273. <ftp://ftp.informatik.uni-hamburg.de/pub/unihh/informatik/WSV/schild97a.ps.gz>.
- SORICUT R. & MARCU D. (2003). Sentence level discourse parsing using syntactic and lexical information. In M. HEARST & M. OSTENDORF, Eds., *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, p. 149–156. <http://aclweb.org/anthology/N03-1030>.
- STEEDMAN M. (2001). *The Syntactic Process*. MIT Press.
- STEEDMAN M. & BALDRIDGE J. (2011). Combinatory categorical grammar. In R. BORSLEY & K. BÖRJARS, Eds., *Non-Transformational Syntax : Formal and Explicit Models of Grammar*, chapter 5. Wiley-Blackwell.
- WEBBER B. L. (2004). D-LTAG : extending lexicalized TAG to discourse. *Cognitive Science*, **28**(5), 751–779. doi :[10.1207/s15516709cog2805_6](https://doi.org/10.1207/s15516709cog2805_6).
- WEBBER B. L. & JOSHI A. K. (1998). Anchoring a lexicalized tree-adjoining grammar for discourse. In M. STEDE, L. WANNER & E. HOVY, Eds., *Proceedings of the ACL/COLING workshop on Discourse Relations and Discourse Markers*. <http://aclweb.org/anthology/W98-0315>.
- WEBBER B. L., STONE M., JOSHI A. K. & KNOTT A. (2003). Anaphora and discourse structure. *Computational Linguistics*, **29**(4), 545–587. <http://aclweb.org/anthology/J03-4002>.

Analyse syntaxique de l'ancien français : quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage ?

Gael Guibon ¹ Isabelle Tellier ^{1,2}
Sophie Prevost ¹ Matthieu Constant ³ Kim Gerdes ^{2,4}

(1) Lattice CNRS
(2) université Paris 3 - Sorbonne Nouvelle
(3) université Paris-Est, LIGM
(4) LPP CNRS

E-mails: gael.guibon@gmail.com, isabelle.tellier@univ-paris3.fr,
sophie.prevost@ens.fr, Matthieu.Constant@u-pem.fr, kim@gerdes.fr

Résumé. L'article présente des résultats d'expériences d'apprentissage automatique pour l'étiquetage morpho-syntaxique et l'analyse syntaxique en dépendance de l'ancien français. Ces expériences ont pour objectif de servir une exploration de corpus pour laquelle le corpus arboré SRCMF sert de données de référence. La nature peu standardisée de la langue qui y est utilisée implique des données d'entraînement hétérogènes et quantitativement limitées. Nous explorons donc diverses stratégies, fondées sur différents critères (variabilité du lexique, forme Vers/Prose des textes, dates des textes), pour constituer des corpus d'entraînement menant aux meilleurs résultats possibles.

Abstract.

Old French parsing : Which language properties have the greatest influence on learning quality ?

This paper presents machine learning experiments for part-of-speech labelling and dependency parsing of Old French. Machine learning methods are used for the purpose of corpus exploration. The SRCMF Treebank is our reference data. The poorly standardised nature of the language used in this corpus implies that training data is heterogenous and quantitatively limited. We explore various strategies, based on different criteria (variability of the lexicon, Verse/Prose form, date of writing) to build training corpora leading to the best possible results.

Mots-clés : étiquetage morpho-syntaxique, analyse en dépendance, ancien français, apprentissage automatique, exploration de corpus.

Keywords: POS labelling, Dependency Parsing, Old French, machine learning, corpus exploration.

1 Introduction

L'ancien français a donné lieu à de nombreux travaux linguistiques, mais il a été jusqu'à présent très peu exploré dans une perspective "TAL". Il existe pourtant depuis peu un corpus arboré permettant cette exploration : le SRCMF (Syntactic Reference Corpus of Medieval French), (Stein & Prevost, 2013). Ce corpus, décrit en détail en section 2, contient des textes de divers domaines (littéraire, historique, religieux...), formes (vers/prose), époques (du 10^{ème} au 13^{ème} siècle) et dialectes (normand, champenois, picard...). La langue de ces textes étant beaucoup moins normalisée que maintenant, le SRCMF présente une variabilité et une hétérogénéité qui n'a pas d'équivalent pour le français contemporain.

Les données étiquetées du corpus SRCMF peuvent être utilisées par des techniques d'apprentissage automatique, pour acquérir un étiqueteur morpho-syntaxique et/ou un analyseur en dépendance (le format adopté pour les analyses syntaxiques dans SRCMF) de l'ancien français. Dans (Guibon *et al.*, 2014) nous avons décrit une première série d'expériences exploitant ce corpus. L'emploi de CRF (Lafferty *et al.*, 2001) pour la couche d'annotation en parties du discours, associé à l'utilisation de Mate (Bohnet, 2010) pour les analyses en dépendance, ont ainsi permis d'améliorer les expériences préliminaires réalisées (notamment avec TreeTagger ¹) par Achim Stein (Stein, 2014). Ces séries d'expériences procédaient

1. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

toutes à des "apprentissages croisés" consistant à apprendre à partir d'un des textes à l'exclusion de tous les autres, et à tester chacun des programmes appris sur chacun des autres textes. L'objectif visé était d'étudier les éventuelles corrélations entre les performances de ces programmes et des notions de proximités linguistiques entre textes. Mais ces "apprentissages croisés" étaient aussi en quelque sorte suggérés par les données elles-mêmes. En effet, du fait de la grande hétérogénéité de la langue d'un texte à un autre, procéder à des expériences par "simples" validations croisées (par texte ou en les mélangeant sans discernement) aurait peu de sens. Dans le contexte de ce corpus, il faut porter une plus grande attention que d'habitude aux relations entre les données d'apprentissage et les données de test.

Dans cet article, nous approfondissons ces premières approches en cherchant toujours à obtenir les meilleurs étiqueteurs et analyseurs syntaxiques possibles pour un texte (ou un ensemble de textes) donné(s). Les méthodes d'apprentissage automatique employées ici sont les mêmes que dans (Guibon *et al.*, 2014), mais il ne s'agit plus d'améliorer les résultats obtenus initialement par Stein puisque d'une expérience à une autre nous faisons varier le mode de constitution des données d'entraînement et de test, rendant toute comparaison impossible. Nous cherchons à caractériser finement les propriétés des corpus d'entraînement qui influencent le plus les résultats en test. L'objectif final serait, pour un texte nouveau dont on connaîtrait les propriétés principales (en termes de domaine, forme, date d'écriture...) et qu'on voudrait étiqueter au mieux, de constituer pour lui un corpus d'entraînement "sur mesure" à partir des données de SRCMF.

Nous souhaitons ainsi développer une méthodologie applicable lors de l'exploration de n'importe quel corpus hétérogène, en jouant sur les propriétés ou métadonnées qui caractérisent les textes qu'il contient. L'apprentissage automatique est la base de cette méthodologie d'exploration de corpus. Elle gravite ici autour de quatre propriétés ou métadonnées pertinentes pour les données de SRCMF : l'homogénéité lexicale des textes (section 3), la quantité de données en entraînement (section 4), la forme (vers/prose, section 5) et enfin la date d'écriture (section 6) des documents.

2 Le corpus SRCMF

2.1 Présentation du corpus

SRCMF est un corpus d'ancien français annoté syntaxiquement dans le cadre d'un projet ANR-DFG, dirigé par A. Stein² (ILR, U. Stuttgart) et S. Prevost³ (Lattice, CNRS/ENS/Paris3) et associant, outre les laboratoires des porteurs, l'ICAR (CNRS/ENS de Lyon). L'objectif de ce projet était la constitution d'une ressource syntaxiquement annotée pour le français médiéval pouvant être utilisable pour l'entraînement ultérieur d'analyseurs syntaxiques. Les ressources initiales utilisées pour la constitution de ce corpus sont la Base de Français Médiéval (BFM⁴) (Guillot *et al.*, 2007) et le Nouveau Corpus d'Amsterdam (NCA⁵) (Stein *et al.*, 2006). La sélection des textes destinés à intégrer le corpus SRCMF s'est faite à partir de différents critères : le caractère « incontournable » de certains textes (la *Chanson de Roland* par exemple), la fiabilité des éditions, la diversité des textes en termes de date, domaine, forme et dialecte et, enfin, le fait que les textes étaient déjà étiquetés morpho-syntaxiquement. Dans un souci d'équilibre entre les textes, ceux qui comprenaient plus de 40 000 mots ont été échantillonnés (début, milieu, fin). La Table 1 présente les principales caractéristiques des textes de SRCMF.

Parmi les 15 textes (245 000 mots) de SRCMF, 10 (soit 201 465 mots) ont été retenus pour nos expériences. Evoquons tout d'abord quelques spécificités de ces données. L'ancien français est un état de langue qui connaît une forte variation morphologique, y compris dans un même texte. Par exemple, l'adverbe 'ainsi' se rencontre sous les graphies suivantes : *ainsi*, *ainsin*, *ainsinc*, *ainssy*... tandis que l'on recense au moins 17 formes différentes pour le pronom personnel 'je' (*je*, *gié*, *jou*, *gel*...). Il intègre un grand nombre de formes contractées (bien plus que le français moderne), qui résultent de phénomènes d'enclise (prise d'appui accentuel d'un mot sur un mot le précédant : *ne + les > nes*, *je + le > jel*). Sur le plan syntaxique, l'expression du sujet n'est pas obligatoire et l'ordre des mots est assez souple (le sujet peut être postverbal et l'objet nominal préverbal), souplesse favorisée par l'existence d'une déclinaison bi-casuelle héritée du latin, mais en voie d'étiollement dès cette époque. Enfin, on y rencontre beaucoup plus de syntagmes discontinus qu'en français moderne :

(1) *et mes sires Gauvains lor demande coment il l'ont puis fet **que** il se partirent de cort (Graal, 1230)*
*Et Messire Gauvain leur demande comment ils ont fait **depuis qu'**ils ont quitté la court*

(2) *Si jurroient li compaignon tel serement come **cil font qui en queste doivent entrer***
Les compaignons faisaient un serment comme font ceux qui doivent entrer en quête

2. <http://www.uni-stuttgart.de/lingrom/stein/>

3. <http://www.lattice.cnrs.fr/Sophie-Prevost,229>

4. <http://bfm.ens-lyon.fr/>

5. <http://www.uni-stuttgart.de/lingrom/stein/corpus/>

Texte	Date	Nb mots	Nb clauses	Forme	Dialecte	Domaine
<i>Vie Saint Légier</i>	Fin 10e s.	1388	192	vers	nd (non défini)	religieux
<i>Vie de Saint Alexis</i>	1050	4804	562	vers	normand	didactique
<i>Chanson de Roland</i>	1100	28 766	3857	vers	normand	littéraire
<i>Lapidaire en prose</i>	Milieu 12e s.	4708	468	prose	anglo-normand	didactique
<i>Yvain</i> , Chretien de Troyes	1177-1181	41 305	3788	vers	champenois	littéraire
<i>La Conquête de Constantinople</i> de Robert de Clari	>1205	33 534	2308	prose	picard	historique
<i>Queste del Saint Graal</i>	1220	40 417	3078	prose	nd	littéraire
<i>Aucassin et Nicolette</i>	Fin 12e s.- début 13e s.	9844	1101	vers & prose	picard	historique
<i>Miracles de Gautier de Coinci</i>	1218-1227	17 360	1422	vers	picard	religieux
<i>Roman de la Rose</i> de Jean de Meun	1269-1278	19 339	1449	vers	nd	didactique

TABLE 1 – Textes du SRCMF utilisés dans nos expériences

2.2 Enrichissement linguistique du corpus

2.2.1 Etiquetage morpho-syntaxique

Le jeu d'étiquettes morpho-syntaxiques utilisé en annotation comprend 60 valeurs⁶, structurées en 2 champs : Catégorie et Type, les “catégories” correspondant aux traditionnelles parties du discours (nom, verbe, adjectif, pronom, ...) et les “types” correspondant à la spécification de ces ‘catégories’ (verbe conjugué : VERcjk, nom propre : NOMpro,...). Pour les formes contractées (phénomènes d'enclise), il existe des étiquettes complexes, qui associent les valeurs des 2 unités linguistiques contractées en une seule unité graphique. Par exemple : *nel* = contraction de *ne* (adverbe de négation) + *le* (pronom personnel) = ADVneg.PROper. Toute unité graphique a une étiquette morpho-syntaxique.

2.2.2 Annotation syntaxique

L'annotation a été réalisée manuellement, en double aveugle, avec le logiciel NotaBene⁷ (Mazziotta, 2010), selon un modèle dépendantiel, inspiré de (Tesnière, 1959) et (Polguère *et al.*, 2009). Le modèle utilisé dans SRCMF hiérarchise d'une part des unités syntaxiques : structures (qui ont une tête verbale, nominale, adjectivale ...), nœuds (regroupés dans des structures dépendanciels) et groupes (pour le traitement des faits de coordination). Le modèle hiérarchise d'autre part des fonctions (sujet, objet, modifieur, ...), qui précisent la relation entre le nœud tête et les structures qui en dépendent. Toute relation dépendancielle est ainsi exprimée par un triplet : (nœud mère, nœud fille, relation dépendancielle).

La structure maximale de l'analyse syntaxique est la “phrase” (“Snt” pour Sentence), définie par la présence d'un verbe fléchi, qui la gouverne, et par le fait qu'elle n'a pas de fonction. La “phrase” correspond, dans la grammaire traditionnelle, à une proposition indépendante ou principale. A l'inverse, les propositions subordonnées ont une fonction vis-à-vis d'un élément de la proposition dont elles dépendent (Objet, Complément, Circonstant, Modifieur attaché d'un nom (c'est le cas des propositions relatives)). Dans la terminologie SRCMF, ce sont des “non-phrases”. Dans la suite de cet article, nous utiliserons le terme “clause” pour désigner l'ensemble des structures à tête verbale (verbe fléchi), qu'il s'agisse de “phrases” (principales/indépendantes) ou de “non-phrases” (subordonnées). Par ailleurs il n'y a pas de coordination de “phrases”, on découpe au maximum. Dans l'exemple suivant, il y a ainsi 4 phrases :

(3) *Lors entre li preudons en sa chapele / et prent .i livre et une estole / et vient fors / et comence a conjurer l' anemi*
Alors, l'homme entre dans la chapelle et prend un livre et une étole, et ressort et commence à exorciser L'ennemi.
 (Graal, folio 188b)

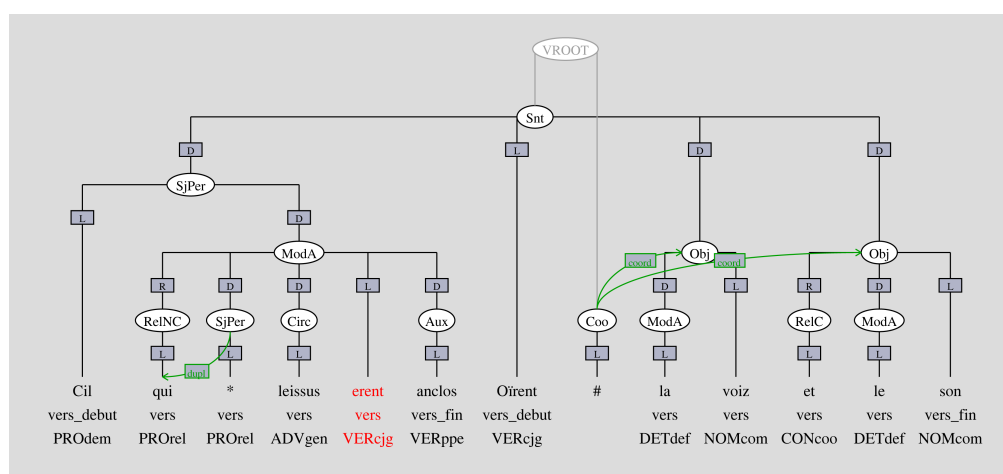
Toutes les structures sont annotées, chaque unité a une étiquette syntaxique. Dans les structures non verbales (par exemple SN), on a nécessairement une tête, et l'on peut avoir des ModA (= modifieurs attachés : déterminants, adjectifs, subordonnées relatives) et/ou un RelC/RelNC (relateur (non) coordonnant : préposition, conjonction de coordination ou de subordination). Parmi les fonctions principales des dépendants du verbe, on distingue des Actants (sujet personnel/impersonnel (SjPer/SjImp), Objet (Obj), Complément (Cmpl), Attribut du sujet (AtSj)...), des Auxiliés (Auxilié ac-

6. http://bfm.ens-lyon.fr/article.php3?id_article=176

7. <http://sourceforge.net/projects/notabene/>

et vous leur demanderez conseil à propos de la venue du roi (Yvain, 1848-1849)

Le logiciel NotaBene exporte les annotations selon deux formats : CoNLL (utilisé pour nos expériences) et TigerXML (utilisé par le logiciel de requête TigerSearch (Lezius, 2002), qui permet une visualisation en graphes). La Figure 1 présente l'arbre au format TigerXML, selon un graphe plus riche en informations que ceux utilisés pour nos expériences.



Les graphes correspondent chacun à une phrase pouvant comprendre plusieurs clauses. Il s’agit de clauses verbales gravitant autour d’un verbe conjugué, ce qui engendre des séquences plus courtes que dans des corpus arborés de langue contemporaine tels que le *French TreeBank* (Abeillé *et al.*, 2003), qui sont segmentés en phrases délimitées par une ponctuation. La ponctuation, peu présente dans les fichiers originaux, est absente du SRCMF.

3.1 Propriété lexicales

173

Corpus	nb unités distinctes	nb mots	nb mots/nb unités	nb répétitions par lemme	taille moy. des clauses
Alexis	1415	4804	3,39	1.98	8
Aucassin	1962	9844	5,02	2.38	9
Coinci	3085	17 360	5,63	2.19	12
Conq.	3424	33 534	9,79	2.9	14
Graal	3874	40 417	10,43	2.56	12
Lapidaire	1197	4708	3,93	2.39	10
Legier	578	1388	2,4	2.08	7
Roland	4304	28 766	6,68	2.00	7
Rose	4097	19339	4,72	2.13	13
Yvain	5040	41305	8,19	2.17	10

TABLE 2 – Quelques propriétés des textes

	Rol.	Graal	Yvain	Conq.
Aucassin	449 / 5818 7.72%	681 / 5147 13.23%	700 / 6289 11.13%	665 / 4714 14.11%
Roland	/	678 / 7473 9.07%	706 / 8601 8.21%	452 / 7259 6.23%
Graal	/	/	1543 / 7340 21.0% ²	878 / 6411 13.70%
Yvain	/	/	/	836 / 7612 10.98%

TABLE 3 – Vocabulaire partagé entre deux textes (calcul | %)

Dans cet article, toutes les expériences ont été réalisées selon ce protocole suivant :

- La lemmatisation a été réalisée en utilisant le fichier de paramètres du NCA⁸ pour *TreeTagger*, appris par Stein et utilisé lors de ses expériences précédentes (Stein, 2014).
- Les Étiquetages morpho-syntaxiques ont été appris avec *Wapiti* (version 1.4.0) (Lavergne *et al.*, 2010) en utilisant les lemmes prédits précédemment. Les patrons utilisés prennent en compte la vérification du contexte proche, des lemmes avec leur contexte, et de la terminaison des mots, comme dans (Guibon *et al.*, 2014).
- Les analyses en dépendance ont été réalisées avec *Mate* (GraphBased anna-3.61 avec les configurations par défaut) en utilisant les données prédites précédemment.

Nous commençons par reproduire, pour les analyser à l’aune de la variabilité lexicale, quelques résultats d’“apprentissage croisés”, dans lesquels un des textes sert d’apprentissage et un autre sert de test. Ces expériences ont été menées d’une part pour l’apprentissage d’un analyseur en parties du discours, d’autre part pour l’acquisition d’un analyseur syntaxique en dépendances utilisant les résultats de l’analyseur en parties du discours.

3.2 Lexique et étiquetage morpho-syntaxique

La prédiction des étiquettes morpho-syntaxiques par “apprentissage croisés” s’est faite sur les cinq plus grands textes disponibles. *Miracles* (Coinci) a été toutefois écarté au profit d’*Aucassin*, car ce dernier présente un mélange intéressant de prose et de vers. Chacun des textes a été découpé en sous-parties d’environ seize mille mots prélevés aléatoirement afin d’équilibrer leurs tailles et donc de limiter autant que possible les biais liés à ces tailles. Les entraînements ont donc toujours utilisé des textes de même taille, chaque modèle appris a été testé sur toutes les autres sous-parties du corpus de test. Les résultats sont présentés dans la Table 4. Chaque valeur est la moyenne des exactitudes des différents modèles.

Dans la Table 4, on voit que pour chaque texte considéré indépendamment en test (en observant le tableau colonne par colonne), le texte permettant d’atteindre en général la meilleure exactitude quand il est utilisé en apprentissage est *Graal*. Ce texte n’est pas le plus riche en nombre d’unités distinctes, mais c’est celui qui présente le plus de répétitions des mêmes mots. Mais *Conquest*, qui présente un taux de répétition (nombre de mots/nombre d’unités) comparable, mène, lui, à des étiqueteurs nettement moins efficaces, ce qui discrédite ce seul indice comme significatif. Plus que la variabilité intrinsèque de chaque texte, c’est sans doute leur proximité deux à deux, lisible dans la Table 3, qui favorise l’apprentissage. Les corrélations sont effectivement assez bonnes, mais parfois irrégulières. Ces premiers résultats contredisent l’hypothèse d’une influence majeure qu’aurait pu avoir un lexique diversifié sur l’étiquetage morpho-syntaxique. Il convient désormais d’aborder son influence sur l’analyse syntaxique afin d’en identifier les spécificités.

8. http://bfm.ens-lyon.fr/article.php?id_article=324

Train \ Test		Auc.	Rol.	Graal	Yvain	Conq.
Aucassin	Exactitude		80.00	85.76	80.03	87.86
	% Inconnus / Connus		48.19 51.81	29.24 70.76	30.77 69.23	26.99 73.01
	Exactitude I/C		71.43 87.97	72.78 91.07	67.45 85.63	79.31 91.02
Roland	Exactitude	80.48		82.66	78.20	84.13
	% Inconnus / Connus	41.62 58.38		39.90 60.10	40.05 59.95	45.86 54.14
	Exactitude I/C	73.26 85.42		74.45 87.90	73.13 81.42	77.24 89.85
Gaal	Exactitude	85.38	80.58		82.70	86.84
	% Inconnus / Connus	30.01 69.99	52.75 47.25		23.11 76.89	32.69 67.31
	Exactitude I/C	73.55 90.45	73.62 88.35		72.78 85.69	78.49 90.88
Yvain	Exactitude	83.13	80.22	89.05		82.11
	% Inconnus / Connus	29.98 70.02	52.00 48.00	19.81 80.19		35.04 64.96
	Exactitude I/C	64.11 91.27	71.20 89.97	74.63 92.59		66.52 90.52
Conq.	Exactitude	80.48	74.51	79.98	71.04	
	% Inconnus / Connus	31.91 68.09	56.63 43.37	33.43 66.57	37.12 62.88	
	Exactitude I/C	67.42 86.64	67.03 84.28	65.83 87.11	60.78 77.17	

TABLE 4 – Moyenne des différentes expériences d’étiquetages morpho-syntaxiques en “apprentissage croisé”

3.3 Lexique et analyse en dépendance

La diversité du lexique ne semble pas être très influente non plus sur l’analyse syntaxique. Deux méthodes d’évaluation sont utilisées : l’exactitude du gouverneur du mot (*Unlabelled Attachment Score* : UAS) et l’exactitude du gouverneur associé à la bonne fonction syntaxique (*Labelled Attachment Score* : LAS). La Table 5 montre ainsi que plus la méthode d’évaluation est stricte (LAS) plus il apparaît une corrélation entre la diversité lexicale d’un texte et les résultats.

Corpus de test	UAS moyen	LAS moyen	UAS mots inconnus / connus	LAS mots inconnus / connus
Aucassin	75.71 %	61.92 %	67.73% 78.61%	44.41% 63.62%
Roland	76.54 %	58.77 %	63.76% 77.28%	44.10% 60.36%
Gaal	77.62 %	63.94 %	71.53% 82.74%	49.37% 69.07%
Yvain	71.14 %	56.03 %	70.53% 80.75%	46.41% 65.69%
Conq.	77.67 %	65.67 %	62.98% 79.08%	41.24% 66.49%

TABLE 5 – Moyenne des résultats de l’analyse en dépendance par corpus de test

Toutefois l’influence de la diversité lexicale n’est pas isolée et est toujours conjointe avec le nombre de mots des corpus utilisés. La Table 2 montre la nécessité d’étudier ensemble ces deux données. Malgré le fait qu’un corpus d’entraînement ayant un plus grand nombre d’unités différentes devrait être un meilleur candidat à l’entraînement qu’un corpus pauvre en variété, dans les faits le rapport entre le nombre de mots et le nombre d’unités ne permet pas de préjuger d’une bonne analyse en dépendances. Le lexique et le rapport entre le nombre de mots et le nombre d’unités lexicales sur un corpus d’entraînement semble avoir une influence légère sur le score d’évaluation le plus restrictif, le LAS. Au travers des expériences suivantes nous cherchons quelles autres particularités linguistiques du corpus peuvent être les plus influentes.

4 Leave one out

Nous avons d’abord reproduit une expérience de (Stein, 2014) qui consiste à sélectionner un texte faisant office de corpus de test et à lui appliquer le modèle appris conjointement sur les neufs autres textes. Notre expérience, dont les résultats figurent dans la Table 6, diffère cependant de celle de Stein : alors qu’il n’avait pas utilisé le *Roman de la Rose* et avait regroupé ensemble les textes de date inférieure au dixième siècle, nous avons tenu à utiliser également les textes *Legier*, *Alexis* et *Lapidaire*. Autant, dans les expériences précédentes, nous avons privilégié les corpus de taille raisonnable (30 000 mots environ minimum), autant pour celle-ci il nous a paru intéressant d’utiliser aussi les corpus de petite taille. Ces textes, une fois “cumulés”, permettent d’obtenir le plus grand corpus d’entraînement possible. C’est cette taille qui compte essentiellement, celle des corpus de test important beaucoup moins. Nous espérons identifier par cette expérience certaines “fractures” (différences significatives) entre un texte particulier et l’ensemble des autres.

XP	Mots inconnus / Mots connus	Exactitude	UAS	LAS
9 sur Alexis	20.05% 79.95% /	85.91 % 79.01% 87.12%	81.10 % 72.17% 83.34%	69.86% 55.56% 73.44%
9 sur Aucassin	13.92% 86.08% /	91.21 % 83.58% 92.45%	86.87 % 74.74% 88.84%	77.20% 59.56% 80.06%
9 sur Clari	15.92% 84.08% /	92.55 % 88.40% 93.33%	87.12 % 79.52% 88.63%	78.35% 66.12% 80.66%
9 sur Coinci	12.96% 87.04% /	89.72 % 75.51% 91.84%	79.91 % 66.89% 81.85%	69.27% 49.64% 72.20%
9 sur Lapidaire	17.99% 82.01% /	88.89 % 77.69% 91.35%	84.88 % 74.26% 87.21%	75.57% 55.61% 79.95%
9 sur Legier	59.44% 40.56% /	66.64 % 52.58% 76.24%	61.74 % 54.53% 66.67%	46.04% 33.93% 54.30%
9 sur Graal	7.17% 92.83% /	93.58 % 85.51% 94.19%	89.51 % 80.79% 90.19%	80.82% 66.06% 81.97%
9 sur Roland	22.70% 77.30% /	90.74 % 85.36% 92.32%	87.91 82.39% 89.54%	76.23 % 63.82% 79.88%
9 sur Rose	14.04% 85.96% /	90.74 % 80.63% 92.96%	81.56 % 68.89% 83.64%	70.94% 52.36% 73.98%
9 sur Yvain	10.55% 89.45% /	89.61 % 86.11% 90.02%	84.19 73.76% 85.42%	74.08% 58.08% 75.97%

TABLE 6 – Resultats des tests sur un corpus par le modèle appris sur les neuf autres

Les résultats de la Table 6 montrent de fait une fracture visible pour *Legier*, pour lequel aucune des expériences effectuées ne donne de résultats satisfaisants en comparaison avec ceux obtenus avec les autres textes. Avec seulement 66.64 % de taux d'exactitude sur les étiquettes morpho-syntaxiques, il ne faut pas s'attendre à une analyse en dépendances correcte. Là où, pour toutes les autres expériences, la différence entre UAS et LAS avoisine 10 % (réduction également conforme aux résultats cités dans (Stein, 2014)), pour *Legier*, et seulement pour lui, la diminution atteint 15.7 %. On peut imputer cette singularité au fait que ce texte est le plus ancien du corpus. Il se distingue des autres par de grandes différences lexicales et morphologiques, qui influencent directement les résultats de nos expériences.

Cette expérience confirme l'hétérogénéité de l'ancien français. Elle montre qu'un grand corpus d'entraînement n'engendre pas obligatoirement une bonne reconnaissance en test, si les langues diffèrent trop entre les deux. La Table 6 illustre aussi que plus il y a de mots inconnus dans un texte, plus l'écart entre les performances sur les mots connus et inconnus est grand, et plus les résultats sont globalement faibles. Elle invite à chercher d'autres critères, d'autres caractéristiques que la seule taille pour constituer des corpus d'apprentissage "sur mesure". C'est ce que nous proposons dans la suite.

5 Quelles différences d'influence entre vers et prose ?

Nous disposons dans SRCMF de textes pour la plupart en vers, mais également de certains en prose, voire constitués d'un mélange des deux pour *Aucassin*. Cette caractéristique nous a amenés à essayer de quantifier les différences que peut engendrer la forme des textes (vers ou prose) sur l'étiquetage morpho-syntaxique et l'analyse en dépendances.

Corpus	Textes	Nb de mots	Nb unités
Prose	Clari + Graal + Lapidaire	78 904	7271
Vers	Alexis + Coinci + Legier + Roland + Rose + Yvain	113 086	13891
Vers (réduit)	Alexis + Legier + Roland + Yvain	76385	9710
Prose [entraînement]	/	41 910	4320
Vers (réduit) [entraînement]	/	41 907	6840
Prose [test]	/	36 749	4370
Vers (réduit) [test]	/	34 478	4417

TABLE 7 – Corpus de différents types de textes

Le regroupement des textes en vers et en prose conduit à deux corpus de tailles différentes comme on peut le montrer la Table 7. Nous avons donc sélectionné un sous-ensemble des textes en Vers permettant un équilibrage des corpus. Cette table nous permet aussi de constater que, même s'ils ont des nombres de mots comparables, les deux corpus Prose/Vers (une fois ce dernier réduit) diffèrent au niveau du lexique. Les textes en vers présentent une plus grande variété lexicale.

5.1 Analyse de l'étiquetage morpho-syntaxique

La Table 8 présente les résultats obtenus en procédant à la division de chaque corpus en deux parties à peu près égales : l'une servant de corpus d'entraînement et l'autre servant de corpus de test pour tous les autres corpus. Il est ainsi possible de comparer l'efficacité de tous les corpus d'entraînement, y compris sur un test de la même forme (Vers ou Prose).

Train \ Test		Prose [test]	Vers réduit [test]
prose	UAS	85.47%	76.33%
	LAS	74.96%	62.96%
	ACC	91.36%	83.61%
	Mots inconnus / Mots connus	16.49% 83.51%	21.26% 78.74%
	Lexique différent / commun	57.02% 42.98%	77.05% 22.95%
	UAS Mots inconnus / Mots connus	73.76% 87.78%	65.87% 79.15%
	LAS Mots inconnus / Mots connus	55.48% 78.81%	46.37% 67.44%
	ACC Mots inconnus / Mots connus	77.33% 94.14%	76.78% 85.46%
vers réduit	UAS	83.12%	82.79%
	LAS	71.52%	71.40%
	ACC	90.06%	90.78%
	Mots inconnus / Mots connus	18.81% 81.19%	14.03% 85.97%
	Lexique différent / commun	66.47% 33.53%	42.52% 57.48%
	UAS Mots inconnus / Mots connus	73.43% 85.37%	72.39% 84.49%
	LAS Mots inconnus / Mots connus	55.45% 75.24%	55.62% 73.98%
	ACC Mots inconnus / Mots connus	81.02% 92.15%	84.13% 91.86%

TABLE 8 – Résultats des tests entre vers et prose en deux sous-corpus

Nous constatons que le taux d'exactitude reste en moyenne conforme aux expériences faites par apprentissage croisé entre textes. Si l'on prend par exemple les résultats de l'étiquetage morpho-syntaxique appris sur les textes en vers (corpus réduit) et testés sur ceux en prose, les noms communs et les noms propres sont souvent reconnus et bien étiquetés (6.7% d'erreur pour les premiers et 14.6% d'erreur pour les seconds) tandis que ce n'est pas du tout le cas, par exemple, des pronoms possessifs (58% d'erreurs). On observe le même phénomène pour l'expérience inverse avec 10 % d'erreur pour les noms communs, 9% pour les noms propres et 75 % pour les pronoms possessifs.

Surtout, le tableau 8 contredit l'hypothèse selon laquelle la forme d'un texte induirait des analyses syntaxiques différentes. La corrélation entre les résultats ne se fait pas tant en fonction de la forme du texte qu'en fonction du nombre de mots connus et de la diversité lexicale du corpus d'entraînement, comme nous l'évoquions déjà en partie 3. En effet, le corpus d'entraînement en vers comporte davantage de mots communs avec le corpus de test en prose qu'avec le corpus de test en vers. Toutefois cette caractéristique ne se retrouve pas dans le cas du corpus d'entraînement en prose qui offre de meilleurs résultats sur son propre corpus de test (mais ce résultat est toujours lié au nombre de mots connus).

Si l'on regarde le type d'erreurs récurrentes, on voit que ce sont les étiquettes les plus fréquentes qui apparaissent le plus, telle que VERconj (verbe conjugué). Du coup, la fréquence d'erreurs pour les verbes conjugués est basse (2% dans l'expérience du corpus en vers sur le corpus en prose), puisqu'il s'agit d'une des étiquettes attribuées "par défaut". La grande variabilité des formes de l'ancien français joue sûrement un rôle dans ces résultats. Par exemple, les adverbes généraux (ADVgen) se retrouvent régulièrement étiquetés en verbes conjugués, qu'il s'agisse de *Ich* (ici) ou encore *adont* (donc, alors) comme on le voit dans l'exemple de la Table 9. Ceci s'explique par le fait que 'i' et 'ont' sont des désinences verbales. Plus généralement, de nombreux verbes conjugués à la troisième personne du pluriel partagent la terminaison *-ent* avec de nombreux adverbes, engendrant ainsi un mauvais étiquetage. Dans le tableau 9, *COMMENCHE* est étiqueté comme nom propre car la casse de la première lettre d'un mot constitue un des patrons utilisés pour différencier les mots. Retrouver un mot en majuscule étiqueté en tant que nom commun est une erreur régulière, mais, vue la rareté des mots tout en majuscule, enlever ce patron diminuerait les performances globales de l'étiqueteur morpho-syntaxique.

Adont VERconj	si ADVgen	fu VERconj	croisés VERppe	li DETdef	cuens NOMcom	Thiebaus NOMpro	de PRE
Champagne NOMpro	ICHI VERconj	COMMENCHE NOMpro	LI DETdef	PROLOGUES NOMpro	DE PRE	COUSTANTINOBLE NOMpro	

TABLE 9 – Exemples de mauvais étiquetages

5.2 Résultats des analyses en dépendances

Le tableau 8 présente des résultats très bons en comparaison de ceux obtenus lors de la validation inter-textes. Ainsi les scores de l’UAS sont tout particulièrement élevés puisque lors des expériences précédentes, l’UAS maximal obtenu avoisinait les 81%. Plusieurs explications sont possibles : les corpus d’entraînement sont plus grands et couvrent donc davantage de cas différents, ce qui rejoint la question de l’importance du lexique du corpus d’entraînement, sa diversité étant bien plus grande que ce que peut apporter un texte isolé (maximum de 5040 mots différents).

En faisant le rapprochement entre les tableaux 5 et 8, il est possible de conclure sur l’importance d’un corpus d’entraînement au lexique varié, d’autant plus lorsqu’il s’agit de traiter de l’ancien français aux nombreuses variations de formes. Cette variabilité des formes, nous n’avons pu la mesurer que sur le texte de *Yvain*, seul texte pour lequel nous possédions des lemmes vérifiés. Mais, au vu des résultats de l’analyse en dépendances, en particulier le LAS, nous pouvons estimer avoir une plus grande variété de formes dans les textes en vers. Certes, la majorité des textes de l’époque sont en vers, mais en utilisant des corpus de tailles proches nous pouvons observer une diversité lexicale moindre dans les textes en prose. Pour en mesurer l’influence il convient de prendre en considération le lexique commun entre le corpus d’entraînement et le corpus de test, comme cela est montré dans le tableau 8. Ce tableau 8 montre également une corrélation entre taux d’exactitude (accuracy) et UAS obtenue lors de l’analyse en dépendance. Ce n’est toutefois pas toujours le cas pour le LAS, qui semble davantage dépendre de la qualité du lexique plutôt que de la qualité de l’étiquetage morpho-syntaxique.

Train \ Test	Aucassin
PROSE	UAS
	85.72 %
	LAS
	75.08%
	ACC
	90.89%
	Mots inconnus / Mots connus
	18.54% 81.46%
vers (réduit)	Lexique différent / commun
	53.00% 47.00%
	UAS Mots inconnus / Mots connus
	74.74% 88.22%
	LAS Mots inconnus / Mots connus
	56.82% 79.24%
	ACC Mots inconnus / Mots connus
	81.04% 93.13%
	UAS
	83.14%
	LAS
	72.48%
	ACC
	88.74%
	Mots inconnus / Mots connus
	21.26% 78.74%
	Lexique différent / commun
	56.67% 43.33%
	UAS Mots inconnus / Mots connus
	71.62% 86.25%
	LAS Mots inconnus / Mots connus
	56.28% 76.85%
	ACC Mots inconnus / Mots connus
	76.73% 91.99%

TABLE 10 – Tests des corpus en prose et en vers sur *Aucassin*

Cette expérience sur des corpus de types différents confirme l’influence de la richesse lexicale d’un corpus d’entraînement. En effet, le corpus d’entraînement en vers offre le taux moyen minimal de mots inconnus, sans doute grâce à sa plus grande diversité lexicale (tableau 7). Parmi les trois expériences dépassant le taux de 80% de mots connus, l’entraînement sur le corpus en vers entraîne la meilleure reconnaissance syntaxique (LAS) sur les mots inconnus : 2.5% de mieux que lors de l’utilisation du corpus en prose pour l’apprentissage avec test sur d’autres textes en prose, alors que ces derniers comptent une moins grande part de mots inconnus. Nos résultats montrent aussi que lors de l’utilisation d’*Aucassin* (seul texte de SRCMF mélangeant les deux formes prose et vers) en corpus de test, l’entraînement sur la prose donne généralement des résultats meilleurs, corrélés avec le taux plus élevé des mots reconnus. N’ayant pas la quantité exacte de prose et de vers dans *Aucassin*, cette expérience ne peut totalement confirmer les précédentes. Elle montre un score en LAS très proche malgré l’utilisation de corpus d’entraînement distincts avec un taux de mots connus différents.

Il ressort de nos expériences que les textes en vers semblent plus aptes à servir de corpus d'entraînement, tout du moins dans le cadre du corpus SRCMF. Mais les trois textes en prose dont nous disposons couvrent moins d'un siècle de différence, tandis que ceux en vers couvrent presque trois siècles. Cette couverture temporelle différente pour les corpus de chaque forme permet sans doute d'expliquer en partie leur écart conséquent en diversité lexicale. La diversité lexicale si prompte à améliorer les résultats, serait alors liée à l'écart temporel des textes. Ceci nous amène maintenant à explorer l'impact de la période d'écriture sur la capacité à constituer un bon corpus d'entraînement.

6 Une évolution langagière perceptible ?

A l'époque où les textes du SRCMF ont été écrits, peu de personnes savaient lire et écrire. Les échanges se faisaient majoritairement par voie orale, la langue n'était donc que très peu standardisée, d'où les variations d'écriture entre les textes, et sans doute aussi une évolution plus rapide et importante qu'aujourd'hui. C'est particulièrement frappant pour les graphies, certains auteurs utilisant plusieurs orthographes pour un seul mot au sein d'un même texte.

Nous avons tenté de normaliser les formes rencontrées mais avons abandonné cette solution pour deux raisons. La première est linguistique : il est discutable de modifier le corpus traité lorsque nous désirons justement l'explorer tel qu'il est. Une normalisation basée sur nos ressources actuelles devrait être modifiée à chaque ajout de nouvelles données, changeant en même temps le sens des résultats précédents. La seconde raison est pratique : les méthodes simples testées (à base de dictionnaires et de distance d'édition (Myers, 1986)) n'ont pas donné de meilleurs résultats pour les apprentissages.

A la variation lexicale s'ajoute l'existence d'une déclinaison, qui se traduit en particulier par la présence d'un -s désintentionnel à la fin de bon nombre de noms masculins singuliers au cas sujet (mais pas au cas oblique (= objet)). Selon la fonction, on peut donc trouver le nom propre *Yvain* (ou *Yvein*) décliné en *Yvains* ((ou *Yveins*), sachant que dès cette époque la déclinaison commence à s'effriter (on a des cas sujets singuliers sans '-s'). De plus la graphie des formes a en partie changé au cours des siècles et, du fait que le SRCMF s'étend sur plusieurs siècles, il nous a paru intéressant d'effectuer une expérience d'apprentissage utilisant des corpus de textes de siècles différents. Cependant, *Vie de Saint Alexis* et *Vie Saint Légier* étant les uniques représentants de la période la plus ancienne, nous ne pouvions regrouper que des textes de deux siècles différents de taille satisfaisante : le 12ème siècle et le 13ème siècle, décrits dans la Table 11.

Corpus \ Infos	Textes	Nb de mots	Nb unités	Taille moy. des clauses
Corpus du 12ème siècle	Lapidaire + Roland + Yvain	74 779	9258	9
Corpus du 13ème siècle	Aucassin + Clari + Coinci + Graal	101 155	9232	12
Corpus 12ème siècle [entraînement]	/	31 488	4901	7
Corpus 13ème siècle [entraînement]	/	50 740	5772	12
Corpus 12ème siècle [test]	/	43 291	5580	9
Corpus 13ème siècle [test]	/	50 415	5226	12

TABLE 11 – Corpus de siècles différents

Pour nos expériences nous avons donc utilisé deux corpus de siècles différents dont la taille et la longueur moyenne des clauses différent, tandis que la diversité lexicale est semblable, ce qui n'était pas le cas en partie 5. Afin de pouvoir comparer leur influence en tant que corpus d'entraînement, nous avons procédé à la même séparation que dans la partie précédente, à savoir la division de chaque corpus en un corpus d'entraînement et un de test mais, contrairement à l'expérience précédente, elle n'induit pas une diversité lexicale très différente entre les différents sous-corpus. Même si le corpus d'entraînement issu des textes du 13ème siècle est apparemment plus "apte" à l'entraînement que celui issu du 12ème siècle de par sa taille et sa richesse lexicale, l'objectif des expériences présentées dans le tableau 12 est de chercher s'il y a une différence notable entre l'utilisation de textes de siècles différents. Plus exactement, il s'agit moins de comparer les corpus d'entraînement que de comparer leurs résultats sur les différents corpus de tests.

Il apparaît ainsi que l'utilisation d'un corpus d'entraînement d'un siècle donné pour un corpus de test du même siècle ne produit pas toujours de meilleurs résultats que son application sur un corpus d'un siècle différent, ce qui peut sembler inattendu. C'est le cas pour l'apprentissage sur le corpus d'entraînement composé de textes du 12ème siècle, qui donne de moins bons résultats sur son propre corpus de test que sur le corpus de test de textes du 13ème siècle. Compte tenu des expériences précédentes, on pourrait penser que la qualité moindre de ces résultats est la conséquence d'une relation

Train \ Test		12ème Siècle [test]	13ème Siècle [test]
12ème Siècle	UAS	66.07%	66.72%
	LAS	50.71%	52.89%
	ACC	71.88%	75.19%
	Mots inconnus / Mots connus	30.88% 69.12%	32.33% 67.67%
	Lexique différent / commun	58.13% 41.87%	62.52% 37.48%
	UAS Mots inconnus / Mots connus	51.41% 72.62%	52.00% 73.75%
	LAS Mots inconnus / Mots connus	33.24% 58.52%	33.82% 62.00%
	ACC Mots inconnus / Mots connus	55.83% 79.05%	57.88% 83.46%
13ème Siècle	UAS	72.31%	81.15%
	LAS	57.57 %	69.52%
	ACC	78.28%	87.32%
	Mots inconnus / Mots connus	22.52% 77.48%	17.65% 82.35%
	Lexique différent / commun	70.17% 29.83%	52% 48%
	UAS Mots inconnus / Mots connus	59.44% 76.06%	68.51% 83.86%
	LAS Mots inconnus / Mots connus	39.32% 62.87%	50.11% 73.69%
	ACC Mots inconnus / Mots connus	62.20% 82.95%	67.78% 91.51%

TABLE 12 – Resultats des tests entre deux périodes divisées en deux sous-corpus

moins forte entre les deux textes. Ce n'est cependant pas le cas, le tableau 12 montrant bien un taux de mots connus et de lexique commun toujours plus grand lors de l'application sur un corpus de test du même siècle. Ceci confirme une plus grande relation entre les résultats obtenus et la fréquence des mots déjà connus en entraînement. Cela laisse penser qu'il n'y a donc pas de réelle corrélation entre la proximité temporelle d'un texte et l'efficacité de l'étiquetage morpho-syntaxique et de l'analyse en dépendance. Une autre caractéristique de ces deux corpus pourrait expliquer cet écart au niveau des résultats obtenus : la longueur moyenne des clauses. Nous pouvons nous demander si une longueur moyenne des clauses moindre dans le corpus d'entraînement que dans le corpus de test peut engendrer une difficulté accrue. C'est effectivement le cas en ce qui concerne l'analyse en dépendances, pour laquelle le taux d'erreurs est plus important avec des clauses plus longues dans le corpus de test, mais ceci n'explique pas le taux d'exactitude rencontré.

7 Conclusion

Dans cet article nous avons procédé à l'exploration d'un corpus d'ancien français en utilisant des méthodes d'apprentissage automatique. Diverses expériences ont été menées en vue de quantifier l'impact des caractéristiques de textes de l'ancien français sur les résultats de méthodes d'apprentissage automatique. Nous espérons également tirer de ces expériences une compréhension meilleure, ou en tout cas différente, de la langue utilisée dans ces textes.

Dans un premier temps, nous avons constaté que la diversité lexicale de la langue n'influe pas clairement sur les résultats et qu'il est nécessaire de l'associer avec le nombre de mots et d'unités présents dans les textes. Dans un second temps, nous avons testé des cas extrêmes d'hétérogénéité (en prenant la totalité des textes hors celui de test en apprentissage), ce qui nous a permis d'isoler le texte le plus ancien et le plus atypique linguistiquement comparé à l'ensemble du SRCMF. Avec les expériences suivantes, nous montrons que la forme (Vers/Prose) d'un texte ne semble pas influencer l'étiquetage morpho-syntaxique et l'analyse en dépendances de l'ancien français. Elles nous ont toutefois permis de constater certaines récurrences au niveau du type d'erreurs et, surtout, de voir l'apparition de ces erreurs dans les deux formes, bien que les textes aient une couverture diachronique différente. Enfin, en essayant de quantifier l'évolution de la langue au cours d'un siècle (évolution qui au final ne s'est pas avérée perceptible), nous observons une diminution de la variété lexicale.

Les résultats du tableau 12 contredisent quelque peu l'influence de la proximité des corpus d'entraînement et de test, observée dans les expériences précédentes. En l'état actuel, nous ne pouvons donc proposer d'explications pleinement satisfaisantes quant aux causes exactes des résultats de l'analyse en dépendances de l'ancien français. Il convient donc d'envisager des expériences complémentaires, afin d'évaluer l'influence possible du domaine et du dialecte des textes.

De manière générale, cet article illustre que l'apprentissage automatique de différents niveaux d'analyse de la langue est une méthodologie puissante pour explorer les propriétés d'un corpus hétérogène. Les stratégies employées dans nos expériences diffèrent uniquement par le choix des corpus servant en entraînement et en test. Les résultats obtenus, de par

leur grande variabilité, donnent des indices précieux pour caractériser les propriétés les plus discriminantes qui distinguent un texte (ou un ensemble de textes) d'un autre (ou d'un ensemble d'autres).

La méthodologie proposée ici est de fait applicable à d'autres corpus puisqu'elle permet d'explorer en isolant les caractéristiques d'un texte de façon ordonnée et hiérarchique à partir du lexique, des étiquettes morpho-syntaxiques, et des structures syntaxiques. Nos expériences suggèrent ainsi que, face à un nouveau texte de l'ancien français, il conviendrait de suivre la méthode suivante :

- Dans un premier temps procéder à une approche générale en effectuant une analyse lexicale complète (lexique et rapport entre nombre d'unité et de mots) et en comparant ce texte aux autres par l'expérience "leave one out" (partie 4).
- Dans un second temps y appliquer une analyse guidée par des caractéristiques linguistiques ou des métadonnées connues du corpus : la date de parution, forme du texte, le dialecte utilisé, etc.

Savoir, par exemple, s'il est préférable de disposer d'un très grand nombre de données annotées ou, au contraire, de données annotées moins nombreuses mais plus proches des données cibles est une question commune à de nombreux domaines (Rafrafi *et al.*, 2013; Tellier *et al.*, 2013). Avec cet article, nous esquissons une méthode générale qui pourra trouver des applications dans d'autres contextes faisant intervenir des corpus hétérogènes

Références

- ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In *Treebanks*, p. 165–187. Springer.
- BOHNET B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *The 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China.
- GUIBON G., TELLIER I., CONSTANT M., PRÉVOST S. & GERDES K. (2014). Parsing poorly standardized language dependency on old french. In *13th Treebank and Language Theory (TLT)*.
- GUILLOT C., LAVRENTIEV A. & MARCHELLO-NIZIA C. (2007). La base de français médiéval (bfm) : états et perspectives. *Le nouveau corpus d'Amsterdam : actes de l'atelier de Lauterbad, 23-26 février 2006*, p. 143–152.
- LAFFERTY J., MCCALLUM A. & PEREIRA F. C. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, p. 282–289, Seattle, Washington.
- LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.
- LEZIUS W. (2002). Tigersearch ein suchwerkzeug fr baumbanken. In *Proceedings of the 6. Konferenz zur Verarbeitung natürlicher Sprache (6th Conference on Natural Language Processing, KONVENS 2002)*, Saarbrücken, Germany.
- MAZZIOTTA N. (2010). Logiciel notabene pour l'annotation linguistique. annotations et conceptualisations multiples. In *Recherches qualitatives. Hors-série Les actes*, volume 9.
- MYERS E. (1986). An o(nd) difference algorithm and its variations. *Algorithmica*, p. 251–266.
- POLGUÈRE A. *et al.* (2009). *Dependency in linguistic description*, volume 111. John Benjamins Publishing.
- RAFRAFI A., GUIGUE V. & GALLINARI P. (2013). Classification de sentiments multi-domaines en contexte hétérogène et passage à l'échelle. In *Conférence en Recherche d'Information et Applications (CORIA)*.
- STEIN A. (2014). Parsing heterogeneous corpora with a rich dependency grammar. In N. C. C. CHAIR, K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland : European Language Resources Association (ELRA).
- STEIN A. *et al.* (2006). Nouveau corpus d'amsterdam. corpus informatique de textes littéraires d'ancien français (ca 1150-1350), établi par anthonij dees (amsterdam 1987), remanié par achim stein, pierre kunstmann et martin-d. gleßgen.
- STEIN A. & PRÉVOST S. (2013). Syntactic annotation of medieval texts : the syntactic reference corpus of medieval french (srcmf). In T. NARR, Ed., *New Methods in Historical Corpus Linguistics*. Corpus Linguistics and International Perspectives on Language, CLIP Vol. 3, P. Bennett, M. Durrell, S. Scheible and R. Whitt (eds).
- TELLIER I., DUPONT Y., ESHKOL I. & WANG I. (2013). Adapt a text-oriented chunker for oral data : How much manual effort is necessary ? In *The 14th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'2013)*.
- TESNIÈRE L. (1959). *Éléments de syntaxe structurale*. Librairie C. Klincksieck.

Noyaux de réécriture de phrases munis de types lexico-sémantiques

Martin Gleize^{1,2} et Brigitte Grau^{1,3}

(1) LIMSI-CNRS, Rue John von Neumann, 91405 Orsay CEDEX, France

(2) Université Paris-Sud, Orsay

(3) ENSIIE, Evry

gleize@limsi.fr, bg@limsi.fr

Résumé. De nombreux problèmes en traitement automatique des langues requièrent de déterminer si deux phrases sont des réécritures l'une de l'autre. Une solution efficace consiste à apprendre les réécritures en se fondant sur des méthodes à noyau qui mesurent la similarité entre deux réécritures de paires de phrases. Toutefois, ces méthodes ne permettent généralement pas de prendre en compte des variations sémantiques entre mots, qui permettraient de capturer un plus grand nombre de règles de réécriture. Dans cet article, nous proposons la définition et l'implémentation d'une nouvelle classe de fonction noyau, fondée sur la réécriture de phrases enrichie par un typage pour combler ce manque. Nous l'évaluons sur deux tâches, la reconnaissance de paraphrases et d'implications textuelles.

Abstract.

Enriching String Rewriting Kernels With Lexico-semantic Types

Many high level natural language processing problems can be framed as determining if two given sentences are a rewriting of each other. One way to solve this problem is to learn the way a sentence rewrites into another with kernel-based methods, relying on a kernel function to measure the similarity between two rewritings. While a wide range of rewriting kernels has been developed in the past, they often do not allow the user to provide lexico-semantic variations of words, which could help capturing a wider class of rewriting rules. In this paper, we propose and implement a new class of kernel functions, referred to as type-enriched string rewriting kernel, to address this lack. We experiment with various typing schemes on two natural sentence rewriting tasks, paraphrase identification and recognizing textual entailment.

Mots-clés : fonction noyau, variations sémantiques, réécriture de phrase, reconnaissance de paraphrases, implication textuelle.

Keywords: kernel methods, semantic variations, sentence rewriting, paraphrase identification, textual entailment.

1 Introduction

De nombreuses applications en traitement automatique des langues (TAL) reposent sur le fait de savoir reconnaître que des phrases possèdent des sens proches, que ce soit la reconnaissance d'implication textuelle (RTE) (Dagan *et al.*, 2006), de paraphrases (Dolan *et al.*, 2004) ou de similarité sémantiques (Agirre *et al.*, 2012). Ces problèmes sont généralement représentés comme des problèmes de classification résolus par des méthodes d'apprentissage supervisé reposant sur des représentations différentes des phrases et des phénomènes linguistiques à gérer. Dans (Wan *et al.*, 2006; Lintean & Rus, 2011; Jimenez *et al.*, 2013), les phrases sont représentées par des sacs de mots ou de n-grammes, et reposent essentiellement sur un appariement lexical exact. La reconnaissance de variations lexicales entre deux énoncés de sens proche est traitée par l'ajout de ressources externes, telles WordNet (Mihalcea *et al.*, 2006; Islam & Inkpen, 2009). La prise en compte d'une représentation structurée des phrases pour traiter les variations de formes de surface est fondée sur la comparaison d'arbres syntaxiques (Calvo *et al.*, 2014). (Heilman & Smith, 2010) introduisent un modèle de distance d'édition sur ces arbres. (Socher *et al.*, 2011) utilisent des auto-encodeurs récursifs opérant sur les arbres de constituants pour apprendre à identifier les paraphrases. Les meilleures systèmes combinent différentes méthodes, comme le méta-classifieur de (Madnani *et al.*, 2012), reposant sur des métriques de traduction automatique, et ayant à ce jour les meilleurs résultats sur la reconnaissance de paraphrases. Certaines méthodes détaillées dans la section suivante utilisent des fonctions noyau pour apprendre ce qui rend deux couples de phrases similaires. (Zanzotto *et al.*, 2007) proposent une fonction noyau de comparaison de couples d'arbres syntaxiques, étendue ensuite à des graphes (Zanzotto *et al.*, 2010). (Bu *et al.*, 2012) introduisent un noyau de réécriture de chaînes de caractères (*string rewriting kernel*).

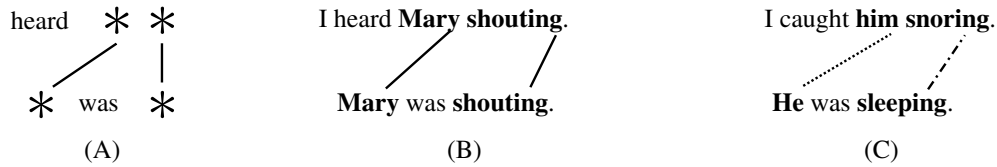


FIGURE 1 – La règle de réécriture (A) réécrit (B) mais pas (C).

Dans cet article, nous proposons d'étendre ces méthodes, en donnant la possibilité de spécifier de manière fine comment deux mots peuvent être appariés et nous définissons ainsi un nouveau noyau de réécriture de phrases enrichi par du typage. Nous détaillons comment calculer ce type de noyau efficacement et nous l'évaluons sur deux tâches de TAL. Notre méthode obtient des résultats analogues à l'état de l'art sur la reconnaissance de paraphrases et dépasse les méthodes de reconnaissance d'implication textuelle fondées sur des approches de même nature.

2 État de l'art des méthodes à noyau pour la réécriture de phrases

Les *fonctions noyau* (*kernel* en anglais) mesurent la similarité de deux éléments. Utilisées dans des méthodes d'apprentissage supervisé telles que les SVM (Vapnik, 2000), elles permettent d'apprendre des fonctions de décision complexes. L'objectif d'une fonction noyau adaptée à ces méthodes est d'avoir une valeur élevée pour deux instances de même étiquette, et une valeur faible pour deux instances d'étiquette différente (Schölkopf & Smola, 2002).

Des méthodes à noyau ont rapidement été employées en traitement automatique des langues. (Lodhi *et al.*, 2002) utilisent le noyau de chaînes (*string kernel*) pour compter le nombre de sous-séquences communes entre deux textes et l'appliquent à la classification de textes.

Classifier des réécritures de phrases revient toutefois à classifier des couples de phrases et requièrent de capturer deux formes de liens : le lien d'une phrase avec l'autre dans un même couple de réécriture, et le lien d'un couple avec un autre. (Zanzotto *et al.*, 2007) proposent une fonction noyau de comparaison de paires d'arbres syntaxiques, étendue ensuite à des graphes (Zanzotto *et al.*, 2010). Leur méthode calcule dans un premier temps le meilleur appariement des noeuds des arbres d'une même paire, pour capturer les entités sur lesquelles portent les deux phrases et former un unique arbre. Un noyau d'arbres –*tree kernel*, introduit par (Moschitti, 2006)– compte dans un second temps les sous-arbres communs des deux paires.

(Bu *et al.*, 2012) introduisent un noyau de réécriture de chaînes de caractères (*string rewriting kernel*) afin de capturer les dépendances syntaxiques sur des paires de phrases vues comme des chaînes de mots, ce qui permet d'apprendre des types de réécriture complexes. Là où l'alignement des mots des deux phrases d'un même couple est réalisé *a priori* dans (Zanzotto *et al.*, 2010) pour réduire le coût de calcul, la contribution de (Bu *et al.*, 2012) propose un algorithme efficace pour intégrer le calcul optimal de ces liens au calcul final du noyau.

Toutes ces méthodes sont toutefois incapables d'introduire des variations lexicales entre mots ou un typage sémantique de ceux-ci, limitant ainsi les types de réécritures apprises. C'est le problème que nous proposons de résoudre dans cet article, en introduisant la notion de *type* pour enrichir les noyaux de réécriture de phrases de (Bu *et al.*, 2012).

3 Noyaux de réécriture de phrases

Définis récemment, les noyaux de réécriture de phrases dénombrent les réécritures communes entre deux couples de phrases vues comme séquences de leurs mots (Bu *et al.*, 2012). La figure 1 présente un exemple de règle de réécriture (A), qui peut être vue comme une paraphrase sous-phrastique avec variables liées (Madnani & Dorr, 2010). La règle (A) réécrit la première phrase de (B) en sa seconde, mais elle ne réécrit pas les phrases de (C). Or, il pourrait être intéressant que la règle (A) se déclenche aussi sur (C), afin d'augmenter les similarités entre réécriture. C'est la motivation de notre contribution : nous présentons et implémentons des noyaux de réécriture de phrases avec types, qui prennent en compte les variations lexico-sémantiques dans les couples de mots.

Dans la suite, nous entendons par *phrase* une séquence de mots, sans y ajouter plus de contraintes linguistiques. On note $(s, t) \in (\Sigma^* \times \Sigma^*)$ une instance de réécriture de phrases, avec sa phrase source s et sa phrase cible t , toutes deux des séquences finies d'éléments de Σ un ensemble fini de mots. Supposons que l'on dispose d'instances étiquetées par

$\{+1, -1\}$ –pour paraphrase/non-paraphrase ou implication/non-implication dans les applications. Il est possible d’appliquer une méthode à noyau pour entraîner un système à classer automatiquement les instances non étiquetées. Un noyau sur des instances de réécriture de phrases est une fonction :

$$K : (\Sigma^* \times \Sigma^*) \times (\Sigma^* \times \Sigma^*) \rightarrow \mathbb{R}$$

telle que pour tous les $(s_1, t_1), (s_2, t_2) \in \Sigma^* \times \Sigma^*$,

$$K((s_1, t_1), (s_2, t_2)) = \langle \Phi(s_1, t_1), \Phi(s_2, t_2) \rangle \quad (1)$$

où Φ projette chaque instance dans l’espace de Hilbert de grande dimension de ces caractéristiques. Les fonctions noyaux permettent d’éviter la représentation explicite potentiellement coûteuse de Φ par l’intermédiaire d’un produit scalaire. Le rôle des noyaux de réécriture de phrases est de mesurer la similarité de deux couples de phrases en comptant le nombre de règles de réécriture d’un ensemble de règles R qu’elles partagent. Φ est donc naturellement définie par $\Phi(s, t) = (\phi_r(s, t))_{r \in R}$ avec chaque caractéristique $\phi_r(s, t) = n$ le nombre de couples de sous-séquences de (s, t) que r réécrit. Suivant la définition de R , Φ peut être de dimension non bornée et n’est pas calculable directement, d’où l’intérêt de cette approche par noyau.

4 Noyaux de réécriture de phrases avec types

4.1 Règles de réécriture typées

Soit le domaine joker $D \subseteq \Sigma^*$ l’ensemble des phrases qui peuvent être remplacées par un joker $*$. Nous présentons maintenant le formalisme des noyaux de réécriture de phrases avec types.

Soit Γ_p l’ensemble des *types motif* et Γ_v l’ensemble des *types variable*.

On associe à un type $\gamma_p \in \Gamma_p$ la *relation de type* $\gamma_p \subseteq \Sigma \times \Sigma$.

On associe à un type $\gamma_v \in \Gamma_v$ la relation de type $\gamma_v \subseteq D \times D$.

Munis des relations de type associées, on désignera l’association de Γ_p et Γ_v par *schéma de types*.

Soit Σ_p défini par

$$\Sigma_p = \bigcup_{\gamma \in \Gamma} \{[a|b] \mid \exists a, b \in \Sigma, a \approx_\gamma b\} \quad (2)$$

Définissons enfin les règles de réécriture typées. Une *règle de réécriture typée* est un triplet $r = (\beta_s, \beta_t, \tau)$, où $\beta_s, \beta_t \in (\Sigma_p \cup \{*\})^*$ désignent les motifs typés source et cible et $\tau \subseteq \text{ind}_*(\beta_s) \times \text{ind}_*(\beta_t)$ définit les alignements entre jokers dans les deux motifs. $\text{ind}_*(\beta)$ désigne l’ensemble des indices des jokers de β .

On dit que la règle de réécriture (β_s, β_t, τ) *réécrit* un couple de phrases (s, t) si et seulement si les conditions suivantes sont vérifiées :

- Le motif β_s , resp. β_t , peut être transformé en s , resp. t , en :
 - substituant chaque élément $[a|b]$ de Σ_p dans le motif par a ou b ($\in \Sigma$)
 - substituant chaque joker dans le motif par un élément du domaine joker D
- $\forall (i, j) \in \tau$, s , resp. t , substitue les jokers à l’indice i , resp. j , par $s_* \in D$, resp. t_* , tel qu’il existe un type variable $\gamma \in \Gamma_v$ avec $s_* \approx_\gamma t_*$.

Un noyau de réécriture de phrases avec types (TESRK) est simplement un noyau de réécriture de phrases comme défini à l’équation 1 mais avec R un ensemble de règles de réécriture typées. Cette classe de fonctions noyau dépend du domaine joker D et de l’ensemble R , qui peut être choisi de façon à permettre plus de flexibilité dans l’appariement de couples de mots dans les réécritures.

Suivant ce formalisme, le noyau de réécriture de phrases bijectif sur k-grammes (kb-SRK) est défini par le domaine joker $D = \Sigma$ et les règles

$$R = \{(\beta_s, \beta_t, \tau) \mid \beta_s, \beta_t \in (\Sigma_p \cup \{*\})^k, \tau \text{ bijective}\}$$

sous le schéma de types $\Gamma_p = \Gamma_v = \{id\}$ avec $a \approx^{id} b \Leftrightarrow a \approx b \Leftrightarrow a = b$.

4.2 Exemple

Dans cette section, nous présentons un exemple d’application de kb-SRK à un couple réel de phrases, en mettant en avant les limites du noyau et comment il est possible de les dépasser en changeant de schéma de types. Reprenons la figure 1 :

(A) est une règle de réécriture avec $\beta_s = (\text{heard}, *, *)$, $\beta_t = (*, \text{was}, *)$, $\tau = \{(2, 1); (3, 3)\}$. Chaque motif a la même longueur, et les couples de jokers dans les deux motifs sont alignés de manière bijective. Elle est donc une règle valide de kb-SRK. Elle réécrit le couple de phrases (B) : chaque couple de jokers peut en effet être substitué dans les phrases source et cible par le même mot et les motifs de (A) peuvent ainsi être transformés en couple de sous-phrases de (B). Cependant, (A) ne peut pas réécrire (C) avec la définition originale de kb-SRK. Redéfinissons alors Γ_p en $\{\text{hypernym}, \text{id}\}$ où $a \overset{\text{hypernym}}{\approx} b$ si et seulement si a et b ont un hypernyme commun dans WordNet. Changeons aussi Γ_v en $\Gamma_v = \{\text{same_pronoun}, \text{entailment}, \text{id}\}$ où $a \overset{\text{same_pronoun}}{\approx} b$ si et seulement si a et b sont un pronom de même personne et même nombre, et où $a \overset{\text{entailment}}{\approx} b$ si et seulement si le verbe a a une relation sémantique *entailment* avec b dans WordNet. En redéfinissant ainsi le schéma de types, la règle (A) peut maintenant réécrire (C).

5 Calcul de TESRK

5.1 Formulation du problème

La fonction noyau kb-SRK peut être calculée efficacement (Bu *et al.*, 2012). Cette section montre qu’il est possible d’en calculer efficacement une version enrichie de types. S’il est impossible de conserver les mêmes bornes théoriques de complexité en temps, les expériences menées montrent que le temps de calcul est du même ordre de grandeur. Le kb-SRK avec types est paramétré par k la longueur des k-grammes, et par son schéma de types constitué des ensembles Γ_p and Γ_v et des relations associées. Nous omettrons dans la suite d’annoter K_k et \bar{K}_k de Γ_p and Γ_v par souci de clarté et parce que ces paramètres resteront d’ordinaire constants pour des valeurs de k variables dans nos expériences. Réécrivons le produit scalaire de l’équation 1 pour mieux refléter les contraintes imposées par les k-grammes :

$$K_k((s_1, t_1), (s_2, t_2)) = \sum_{\substack{\alpha_{s_1} \in k\text{-grams}(s_1) \\ \alpha_{t_1} \in k\text{-grams}(t_1)}} \sum_{\substack{\alpha_{s_2} \in k\text{-grams}(s_2) \\ \alpha_{t_2} \in k\text{-grams}(t_2)}} \bar{K}_k((\alpha_{s_1}, \alpha_{t_1}), (\alpha_{s_2}, \alpha_{t_2})) \quad (3)$$

où $\bar{K}_k((\alpha_{s_1}, \alpha_{t_1}), (\alpha_{s_2}, \alpha_{t_2}))$ est le nombre de règles de réécriture différentes qui réécrivent à la fois le couple de k-grammes $(\alpha_{s_1}, \alpha_{t_1})$, et le couple de k-gramme $(\alpha_{s_2}, \alpha_{t_2})$ (la même règle ne peut en effet pas se déclencher deux fois dans des paires de k-grammes) :

$$\bar{K}_k((\alpha_{s_1}, \alpha_{t_1}), (\alpha_{s_2}, \alpha_{t_2})) = \sum_{r \in R} \mathbb{1}_r(\alpha_{s_1}, \alpha_{t_1}) \mathbb{1}_r(\alpha_{s_2}, \alpha_{t_2}) \quad (4)$$

avec $\mathbb{1}_r$ la fonction indicatrice de la règle r : 1 si r réécrit le couple de k-grammes considéré, 0 sinon.

K_k n’est évidemment pas calculable efficacement en suivant sa définition à l’équation 3. La somme contient $\mathcal{O}((n - k + 1)^4)$ termes, avec n la longueur de la phrase la plus longue, et chaque terme implique d’énumérer toutes les règles de réécriture de R .

5.2 Calcul de \bar{K}_k pour kb-SRK avec types

L’énumération elle-même des règles de réécriture, à l’équation 4, n’est pas calculable en temps raisonnable : il y a $|\Sigma|^{2k}$ règles sans jokers et sans autre relation de type que l’identité, et $|\Sigma|$ sera sans doute la taille d’un lexique d’une langue dans toute application pertinente. En réalité, il suffit de générer de manière constructiviste les règles dont les motifs sont simultanément correctement substitués par $(\alpha_{s_1}, \alpha_{t_1})$ et $(\alpha_{s_2}, \alpha_{t_2})$.

Soit l’opérateur \otimes tel que $\alpha_1 \otimes \alpha_2 = ((\alpha_1[1], \alpha_2[1]), \dots, (\alpha_1[k], \alpha_2[k]))$. Cette opération est en général appelée *zip* en programmation fonctionnelle. Il est possible grâce à la fonction *CompterCouplagesParfaits* calculée par l’algorithme 1 de dénombrer récursivement les règles de réécriture réécrivant simultanément $(\alpha_{s_1}, \alpha_{t_1})$ et $(\alpha_{s_2}, \alpha_{t_2})$. Nous présentons la formule que nous utilisons pour calculer \bar{K}_k , et nous expliquons sa correction par le détail de l’algorithme :

$$\bar{K}_k((\alpha_{s_1}, \alpha_{t_1}), (\alpha_{s_2}, \alpha_{t_2})) = \text{CompterCouplagesParfaits}(\alpha_{s_1} \otimes \alpha_{s_2}, \alpha_{t_1} \otimes \alpha_{t_2}) \quad (5)$$

L’algorithme 1 prend en entrée le reste des couples de mots de $\alpha_{s_1} \otimes \alpha_{s_2}$ et $\alpha_{t_1} \otimes \alpha_{t_2}$ et produit en sortie le nombre de façons communes qu’ils ont de se réécrire.

Premièrement (lignes 2 et 3), le cas de base où les deux entrées sont vides est géré. Il y a exactement 1 façon de réécrire l'ensemble vide en lui-même : c'est de ne rien faire.

Puis, aux lignes 4 à 9, il ne reste plus de couples de mots source, donc l'algorithme continue d'épuiser les couples cible tant qu'ils ont un type motif commun. Si un couple cible n'a pas de type motif commun, c'est que ces deux mots dépariés devraient substituer un joker dans une règle valide, mais comme la source est vide, ce joker ne peut pas être aligné et l'algorithme retourne 0.

Dans le cas général (lignes 11 à 19), on considère le premier couple de mots (a_1, a_2) dans le reste de $\alpha_{s_1} \otimes \alpha_{s_2}$ à la ligne 12. Le reste du calcul dépend de ses types. Tout couple de mots dans $\alpha_{t_1} \otimes \alpha_{t_2}$ qui peut s'associer par ses types variable avec (a_1, a_2) (lignes 15 à 19) est un nouvel alignement commun de jokers potentiel donc l'algorithme teste tous les alignements possibles et continue récursivement le calcul après avoir retiré les deux couples alignés. Et si (a_1, a_2) sont des mots avec un type motif commun, ils ne sont pas forcés de substituer un joker (lignes 13 et 14) et nous pouvons donc choisir de ne pas créer de nouvel alignement à cette étape, mais juste de continuer la récursion en "oubliant" le couple motif.

Cet algorithme énumère essentiellement toutes les configurations telles que chaque couple de mots est assuré d'avoir un type motif en commun ou d'avoir un alignement exclusif avec un couple de mots dont il partage les types variable (condition de la ligne 15), ce qui est exactement la définition d'une réécriture de règle réécrivant avec succès dans TESRK.

Algorithm 1: Dénombrement naïf de couplages parfaits

```

1 Function CompterCouplagesParfaits (remS, remT)
   Data:
   remS : couples de mots restants dans la source
   remT : couples de mots restants dans la cible
   graph :  $\alpha_{s_1} \otimes \alpha_{s_2}$  et  $\alpha_{t_1} \otimes \alpha_{t_2}$  comme graphe biparti, omis dans les arguments pour éviter d'alourdir les appels
   récursifs
   ruleSet :  $\Gamma_p$  et  $\Gamma_v$ 
   Result: Nombre de règles de réécriture réécrivant  $(\alpha_{s_1}, \alpha_{t_1})$  et  $(\alpha_{s_2}, \alpha_{t_2})$ 
2 if remS ==  $\emptyset$  and remT ==  $\emptyset$  then
3   | return 1 ;
4 else if remS ==  $\emptyset$  then
5   | (b1, b2) = remT.first() ;
6   | if  $\exists \gamma \in \Gamma_p \mid b_1 \overset{\gamma}{\approx} b_2$  then
7   |   | return CompterCouplagesParfaits( $\emptyset$ , remT - {(b1, b2)}) ;
8   | else
9   |   | return 0 ;
10 else
11   | result = 0 ;
12   | (a1, a2) = remS.first() ;
13   | if  $\exists \gamma \in \Gamma_p \mid a_1 \overset{\gamma}{\approx} a_2$  then
14   |   | res += CompterCouplagesParfaits(remS - {(a1, a2)}, remT) ;
15   | for (b1, b2) ∈ remT |  $\exists \gamma \in \Gamma_v \mid a_1 \overset{\gamma}{\approx} b_1$  and  $a_2 \overset{\gamma}{\approx} b_2$  do
16   |   | res += CompterCouplagesParfaits(
17   |     | remS - {(a1, a2)},
18   |     | remT - {(b1, b2)})
19   |   | );

```

Le nom de la fonction CompterCouplagesParfaits est justifié car ce problème est en fait équivalent au dénombrement des couplages parfaits dans le graphe biparti des jokers potentiels, i.e. le graphe ayant pour sommets les jokers potentiels et avec une arête entre deux couples de mots si ils vérifient la condition de la ligne 15 sur leurs types variable. (Valiant, 1979) démontre que ce problème n'est pas calculable efficacement. Notre implémentation représente le graphe par sa matrice de biadjacence, et si on suppose nos relations de type calculables en temps constant par rapport à k , la fonction a une complexité temporelle de $\mathcal{O}(k)$ en ignorant les appels récursifs. Le nombre d'appels récursifs peut dépasser $k!^2$ qui est le nombre de couplages parfaits d'un graphe biparti complet à $2k$ sommets. A la section 6 nous montrons toutefois que sur des données linguistiques, l'algorithme effectue un nombre linéaire d'appels récursifs pour des faibles valeurs de k ,

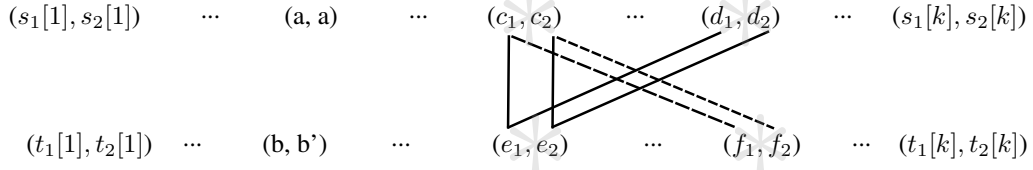


FIGURE 2 – Graphe biparti des couples de mots, avec des arêtes entre les jokers potentiels

et jusqu'à quadratique pour des valeurs de k supérieures à 10 – valeurs pour lesquelles le noyau devient de toute façon inefficace.

La figure 2 montre un exemple de k -grammes source et cible zippés vus comme graphe biparti et avec des arêtes reliant les potentiels jokers. En supposant que les sommets (a, a) et (b, b') ont un type motif commun, ils peuvent être ignorés dans le calcul comme aux lignes 7 et 14 de l'algorithme 1. En revanche, les couples de mots (c_1, c_2) à (f_1, f_2) doivent substituer des jokers dans une règle réécrivant les deux instances. Dans le cas où la ligne 16 aligne (c_1, c_2) avec (e_1, e_2) , l'appel récursif retourne 0 car les deux autres couples ne peuvent pas être alignés. Une règle valide est générée seulement si les c sont liés aux f et les d aux e . Le kb-SRK n'avait pas à tester toutes ces possibilités grâce à la transitivité de son seul type (l'identité) (Bu *et al.*, 2012). Pour kb-SRK enrichis par des types, il y a moins de contraintes sur les appariements de jokers, ce qui donne une meilleure expressivité mais entraîne aussi l'explosion combinatoire du calcul.

5.3 Calcul de K_k

Même avec une méthode efficace pour calculer \bar{K}_k , l'implémentation de K_k en appliquant directement l'équation 3 reste coûteuse. L'idée principale de notre algorithme est de déterminer efficacement un ensemble de taille raisonnable \mathbb{C} d'éléments $((\alpha_{s_1}, \alpha_{t_1}), (\alpha_{s_2}, \alpha_{t_2}))$ ayant la propriété fondamentale d'inclure tous les éléments tels que

$$\bar{K}_k((\alpha_{s_1}, \alpha_{t_1}), (\alpha_{s_2}, \alpha_{t_2})) \neq 0$$

Par définition de \mathbb{C} , il suit qu'il est possible de calculer efficacement :

$$K_k((s_1, t_1), (s_2, t_2)) = \sum_{((\alpha_{s_1}, \alpha_{s_2}), (\alpha_{t_1}, \alpha_{t_2})) \in \mathbb{C}} \bar{K}_k((\alpha_{s_1}, \alpha_{t_1}), (\alpha_{s_2}, \alpha_{t_2})) \quad (6)$$

Il existe de multiples façons de réaliser cette réduction de domaine, avec un équilibre à trouver entre temps de calcul et taille du domaine \mathbb{C} conservé. Notre méthode suit la propriété suivante : $\bar{K}_k((\alpha_{s_1}, \alpha_{t_1}), (\alpha_{s_2}, \alpha_{t_2})) = 0$ si il existe un couple de mots $(a_1, a_2) \in \alpha_{s_1} \otimes \alpha_{s_2}$ sans type motif en commun tel qu'on ne trouve pas un couple $(b_1, b_2) \in \alpha_{t_1} \otimes \alpha_{t_2}$ lui étant compatible pour un type variable, c'est-à-dire avec $a_1 \rightsquigarrow b_1$ et $a_2 \rightsquigarrow b_2$ pour un $\gamma \in \Gamma_v$. C'est évidemment aussi applicable pour un couple de $\alpha_{t_1} \otimes \alpha_{t_2}$. Plus simplement, les mots qui ne relèvent pas d'un même type motif sont départiés et doivent nécessairement substituer un joker dans une règle réécrivant à la fois $(\alpha_{s_1}, \alpha_{t_1})$ et $(\alpha_{s_2}, \alpha_{t_2})$. Si il advient que l'on ne peut pas trouver d'alignement de jokers pour un tel couple de mots, aucune règle ne réécrit les deux paires de k -grammes et il est possible d'ignorer ces entrées dans le calcul de K_k . Ce filtrage peut être effectué en temps raisonnable et l'ensemble \mathbb{C} produit ne contient uniquement qu'un nombre linéaire d'entrées en fonction de n d'après nos expériences.

L'algorithme 2 calcule un ensemble \mathbb{C} à utiliser dans l'équation 6 pour obtenir la valeur finale de la fonction noyau K_k . Toutes les *maps* de l'algorithme désignent une implémentation à base de tables de hachage et sont des maps vers des multiensembles (*multiset* en anglais). Les multiensembles sont utilisés tout au long du calcul : ce sont des extensions des ensembles où les éléments peuvent apparaître en plusieurs exemplaires, ce nombre d'exemplaires étant appelé la *multiplicité*. Classiquement implémenté par des tables de hachage vers des entiers, ils permettent de récupérer en temps constant le nombre d'un élément donné. Les opérations d'union et d'intersection ont des définitions spéciales pour les multiensembles, qu'il est bon de rappeler puisque ces opérations sont utilisées dans l'algorithme. Si $\mathbb{1}_A(x)$ désigne la multiplicité de x dans A , on a $\mathbb{1}_{A \cup B}(x) = \max(\mathbb{1}_A(x), \mathbb{1}_B(x))$ et $\mathbb{1}_{A \cap B}(x) = \min(\mathbb{1}_A(x), \mathbb{1}_B(x))$.

Commentons le déroulement de l'algorithme. Aux lignes 1 à 4, il indexe les mots des phrases source par les mots des phrases cible qui ont des types variable communs, et vice versa. Cela permet aux lignes 15 à 19 d'associer efficacement un

Algorithm 2: Calcul d'un ensemble incluant toutes les entrées telles que $\bar{K}_k \neq 0$

Data: s_1, t_1, s_2, t_2 phrases, et k un entier
Result: Ensemble \mathbb{C} qui inclut toutes les entrées sur lesquelles $\bar{K}_k \neq 0$

```

1 Initialize maps  $e_{s \rightarrow t}^i$  and maps  $e_{t \rightarrow s}^i$ , for  $i \in \{1, 2\}$ ;
2 for  $i \in \{1, 2\}$  do
3   for  $a \in s_i, b \in t_i \mid a \overset{\gamma}{\rightsquigarrow} b, \gamma \in \Gamma_v$  do
4      $e_{s \rightarrow t}^i[a] += (b, \gamma); e_{t \rightarrow s}^i[b] += (a, \gamma);$ 
5  $w_{s \rightarrow t}, aP_t = \text{InclusionJoker}(s_1, s_2, t_1, t_2, e_{s \rightarrow t}^1, e_{s \rightarrow t}^2);$ 
6  $w_{t \rightarrow s}, aP_s = \text{InclusionJoker}(t_1, t_2, s_1, s_2, e_{t \rightarrow s}^1, e_{t \rightarrow s}^2);$ 
7 Initialize multiset res;
8 for  $(\alpha_{s_1}, \alpha_{s_2}) \in aP_s$  do
9   for  $(\alpha_{t_1}, \alpha_{t_2}) \in aP_t$  do
10     $\text{res} += ((\alpha_{s_1}, \alpha_{s_2}), (\alpha_{t_1}, \alpha_{t_2}));$ 
11  $\text{res} = \text{res} \cup w_{s \rightarrow t} \cup w_{t \rightarrow s}.map(\text{swap});$ 
12 return res;

```

Function $\text{InclusionJoker}(s_1, s_2, t_1, t_2, e^1, e^2)$

```

15 Initialize map  $d$  multisets resWildcards, resAllPatterns;
16 for  $(\alpha_{s_1}, \alpha_{s_2}) \in kgrams(s_1) \times kgrams(s_2)$  do
17   for  $(b_1, b_2) \mid \exists \gamma \in \Gamma_v, (a_1, a_2) \in \alpha_{s_1} \otimes \alpha_{s_2}, (b_i, \gamma) \in e^i[a_i] \forall i \in \{1, 2\}$  do
18      $d[(b_1, b_2)] += (\alpha_{s_1}, \alpha_{s_2});$ 
19 for  $(\alpha_{t_1}, \alpha_{t_2}) \in kgrams(t_1) \times kgrams(t_2)$  do
20   for  $(b_1, b_2) \in \alpha_{t_1} \otimes \alpha_{t_2} \mid b_1 \overset{\gamma}{\neq} b_2 \forall \gamma \in \Gamma_p$  do
21     if compatWKgrams not initialized then
22       Initialize multiset compatWKgrams =  $d[(b_1, b_2)]$ ;
23     compatWKgrams =  $\text{compatWKgrams} \cap d[(b_1, b_2)]$ ;
24     if compatWKgrams not initialized then
25        $\text{resAllPatterns} += (\alpha_{t_1}, \alpha_{t_2});$ 
26     for  $(\alpha_{s_1}, \alpha_{s_2}) \in \text{compatWKgrams}$  do
27        $\text{resWildcards} += ((\alpha_{s_1}, \alpha_{s_2}), (\alpha_{t_1}, \alpha_{t_2}));$ 
28 return (resWildcards, resAllPatterns);

```

Type	Relation de type entre les mots (a, b)	Outils/ressources
id	mots ayant la même forme de surface et même POS tag	OpenNLP tagger
idMinusTag	mots ayant la même forme de surface	OpenNLP tokenizer
lemma	mots ayant le même lemme	WordNetStemmer
stem	mots ayant la même racine	Porter stemmer
synonym, antonym	mots ayant la relation [type]	WordNet
hypernym, hyponym entailment, holonym	b est [type] de a	WordNet
lvhsn	mots ayant une distance d'édition de 1	Levenshtein distance

TABLE 1 – Types définis pour les expérimentations

couple de mots avec l'ensemble des paires de k -grammes opposées contenant un couple avec des types variable communs, c'est-à-dire les couples de k -grammes avec lesquels il pourrait substituer un joker aligné. Aux lignes 20 à 28, seuls les quadruplets de k -grammes dont les couples de mots sans type motif commun d'un côté ont chacun un couple associé pour les types variable de l'autre côté. A la ligne 26, il n'y a pas de couple de mots sans type motif commun ; nous sauvegardons donc ce pur motif dans "all-Patterns". Les k -grammes purs motifs sont appareillés deux à deux aux lignes 8 à 10. Enfin, à la ligne 11, l'algorithme calcule l'union multiensemble des entrées satisfaisant le test d'inclusion de jokers ; l'appel de *swap* dans un cas est nécessaire pour toujours avoir les sources du côté gauche et les cibles du côté droit.

6 Expériences

6.1 Présentation des systèmes

Nos expérimentations portent sur deux tâches : la reconnaissance de paraphrases et la reconnaissance d'implications textuelles. Nous avons utilisé la même configuration dans les différents tests, en faisant varier quelques paramètres que nous avons voulu étudier : le nombre d'exemples d'apprentissage, k , le schéma de types. Nous avons implémentés deux fonctions noyau, le noyau initial kb-SRK de (Bu *et al.*, 2012), dénommé par la suite *SRK*, et notre noyau enrichi, dénommé *TESRK*. L'étiquetage morpho-syntaxique est réalisée avec OpenNLP (Baldrige & G., 2010) et les mots sont racinisés par l'algorithme de Porter (Porter, 2001) dans le cas de *SRK*. Différents prétraitements sont réalisés pour *TESRK*, pour définir les types. Ils sont détaillés Table 1. Nous avons utilisé LIBSVM (Chang & Lin, 2011) pour entraîner un classifieur SVM binaire avec chacun des deux noyaux. L'algorithme de LIBSVM par défaut utilise un paramètre C , qui peut être grossièrement vu comme un paramètre de régularisation. Nous optimisons ce paramètre pour le f -score par validation croisée sur les données d'entraînement. Tous les noyaux ont été normalisés par $\hat{K}(x, y) = \frac{K(x, y)}{\sqrt{K(x, x)}\sqrt{K(y, y)}}$. Nous notons "+" une somme de noyaux, avec normalisation avant et après la somme. Nous avons suivi le protocole expérimental de Bu *et al.* (Bu *et al.*, 2012), et avons introduit un noyau de vecteurs supplémentaire, dénommé *PR*, composé de deux traits représentant la *précision sur des uni-grammes* et le *rappel*, définis dans (Wan *et al.*, 2006). Dans nos tests, le noyau linéaire fournit de meilleurs résultats.

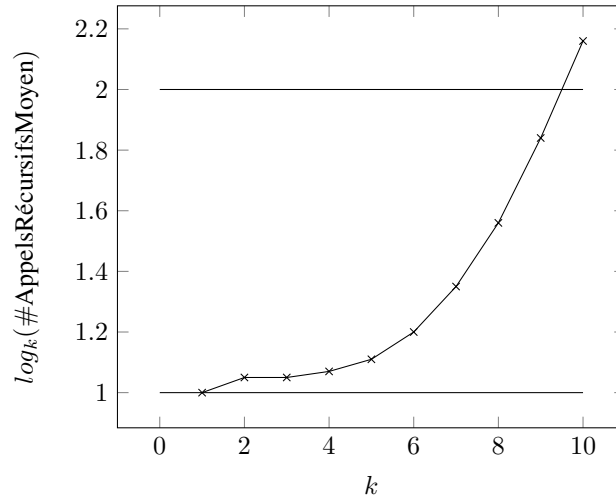
6.2 Reconnaissance de paraphrase

La reconnaissance de paraphrases consiste à déterminer si deux phrases ont le même sens. Le jeu de données que nous avons utilisé pour évaluer nos systèmes est le MSR Paraphrase Corpus (Dolan & Brockett, 2005), qui contient 4076 couples de phrases d'entraînement et 1725 couples de test en langue anglaise. Par exemple, les phrases "An injured woman co-worker also was hospitalized and was listed in good condition." et "A woman was listed in good condition at Memorial's HealthPark campus, he said." sont des paraphrases dans ce corpus. En revanche, "There are a number of locations in our community, which are essentially vulnerable," Mr Ruddock said. et "There are a range of risks which are being seriously examined by competent authorities," Mr Ruddock said. ne sont pas des paraphrases.

La table 2 présente nos meilleurs résultats, avec le système *TESRK + PR*, défini par la somme de *PR* et de kb-SRK de k allant de 1 à 4, munis des types $\Gamma_p = \Gamma_v = \{\text{stem}, \text{synonym}\}$. Nous constatons que nos résultats sont comparables à l'état de l'art sur cette tâche et en particulier, ils sont meilleurs que ceux de kb-SRK original. Nous avons aussi essayé

Système Paraphrase	Accuracy	F-score
All paraphrase	66.5	79.9
Wan et al. (2006)	75.6	83.0
Bu et al. (2012)	76.3	N/A
Socher et al. (2011)	76.8	83.6
Madnani et al. (2012)	77.4	84.1
PR	73.5	82.1
SRK + PR	76.2	83.6
TESRK	76.6	83.7
TESRK + PR	77.2	84.0

TABLE 2 – Résultats d'évaluation sur MSR Paraphrase Corpus

FIGURE 3 – Evolution du nombre d'appels récursifs à CompterCouplagesParfaits en fonction de k

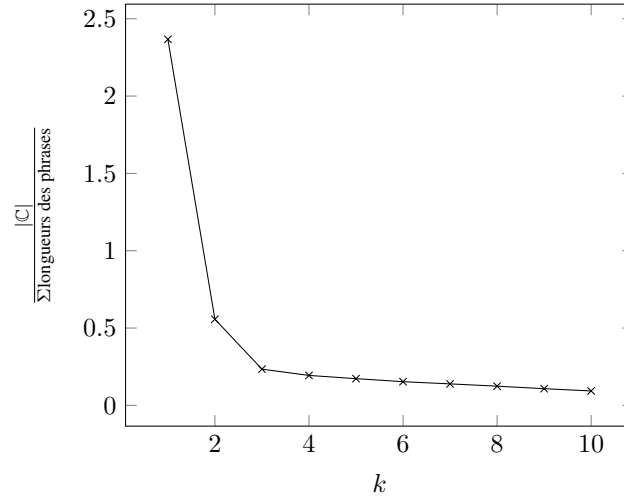
d'autres combinaisons de types mais elles ne donnaient pas mieux : on peut probablement attribuer cela à la nature du corpus de MSR, qui ne semble pas contenir des variations lexico-sémantiques des mots très avancées. Nous avons fait notre possible pour reproduire les performances du kb-SRK original (Bu *et al.*, 2012) : notre implémentation et la leur devraient théoriquement être équivalente.

La figure 3 représente le nombre d'appels récursifs moyens à CompterCouplagesParfaits pendant l'exécution de TESRK en fonction de k . Comme nous avons composé ce nombre avec \log_k , il est très facile de voir si la complexité observée est plus proche de $\mathcal{O}(k)$ ou de $\mathcal{O}(k^2)$. Nous observons ainsi que la complexité est linéaire pour les faibles valeurs de k mais semblent exploser rapidement quand k dépasse 7. Heureusement, compter le nombre de règles de réécriture communes sur des paires de (7 à 10)-grammes donnent rarement des résultats non nuls, donc il n'est pas judicieux d'utiliser ces valeurs de k .

La figure 4 représente la taille moyenne de l'ensemble \mathbb{C} produit par l'algorithme 2 en fonction de k , divisée par la somme des tailles des 4 phrases impliquées dans le calcul du noyau. On peut voir que cette quantité est linéaire par rapport à la taille des entrées, avec un pic pour les petites valeurs, ce qui n'est pas un problème, puisque le calcul de \bar{K}_k est très rapide pour ces valeurs de k .

6.3 Reconnaître l'implication textuelle

Reconnaître l'implication textuelle consiste à déterminer si une phrase *hypothèse* peut être raisonnablement déduite en lisant une phrase *texte*. Le jeu de données que nous avons utilisé pour évaluer nos systèmes est RTE-3 (Dagan *et al.*, 2006), en langue anglaise. Comme les travaux similaires (Heilman & Smith, 2010; Bu *et al.*, 2012), nous avons gardé en entraînement l'intégralité des couples texte-hypothèse de RTE-1 et 2, combinés à l'ensemble d'entraînement de RTE-3,


 FIGURE 4 – Evolution du nombre d'éléments de \mathbb{C} produits par l'algorithme 2 en fonction de k

Système RTE	Accuracy
All entailments	51.2
Heilman and Smith (2010)	62.8
Bu et al. (2012)	65.1
Zanzotto et al. (2007)	65.8
Hickl et al. (2006)	80.0
PR	61.8
SRK + PR	63.8
TESRK (All)	62.1
TESRK (Syn) + PR	64.1
TESRK (All) + PR	66.1

TABLE 3 – Résultats d'évaluation sur RTE-3

ce qui donne 3767 couples de phrases. Pour tester, nous avons simplement pris l'ensemble de test de RTE-3 contenant 800 couples de phrases.

Un exemple d'implication textuelle valide trouvé dans ce jeu de données est le couple de phrases *"In a move widely viewed as surprising, the Bank of England raised UK interest rates from 5% to 5.25%, the highest in five years."* et *"UK interest rates went up from 5% to 5.25%."* : la première implique la seconde. En revanche, les phrases *"Former French president General Charles de Gaulle died in November. More than 6,000 people attended a requiem mass for him at Notre Dame cathedral in Paris."* et *"Charles de Gaulle died in 1970."* ne constituent pas d'implication textuelle.

La table 3 présente nos meilleurs résultats, avec le système *TESRK (All) + PR*, défini comme la somme de PR, 1b-SRK (the original kb-SRK for $k = 1$) et des kb-SRK avec types pour k de 2 à 4. Les types utilisés sont $\Gamma_p = \{\text{stem, synonym}\}$ and $\Gamma_v = \{\text{stem, synonym, hypernym, hyponym, entailment, holonym}\}$. Il semble intéressant de comparer nos résultats uniquement avec les systèmes utilisant des techniques et ressources de même nature, mais nous incluons tout de même le meilleur système à RTE-3 pour référence (Hickl *et al.*, 2006). Cette fois nous n'avons pas réussi à reproduire fidèlement les performances de (Bu *et al.*, 2012), mais nous observons tout de même que kb-SRK avec types améliore significativement les performances du kb-SRK de base, allant même jusqu'à faire mieux que l'implémentation originale. Nous avons aussi expérimenté avec des types moins riches pour le système *TESRK (Syn) + PR* en enlevant tous les types WordNet sauf les synonymes, ce qui aboutit cependant à des performances moins élevées. Cela semble vouloir indiquer qu'un système de types riche aide effectivement à capturer des réécritures plus complexes.

Notons le besoin pour $k = 1$ de remplacer TESRK par SRK, sans quoi nos performances chutaient considérablement. Notre hypothèse est qu'inclure des types riches n'aide vraiment que si ils sont capturés au sein d'un contexte d'au moins quelques mots.

7 Conclusion

Nous avons développé une extension de types pour une classe déjà expressive de noyaux de réécriture de phrases. Les types fournissent une plus grande flexibilité dans le décompte des règles de réécritures communes et peuvent aussi ajouter une couche sémantique aux couples de phrases. Nous avons détaillé une méthode efficace pour calculer le noyau de réécriture de phrases bijectif sur les k-grammes muni de types. Un classifieur SVM utilisant ces noyaux enrichis de relations de type provenant de ressources lexico-sémantiques obtient des performances similaires à ou meilleures que l'état de l'art en reconnaissance de paraphrases et implications textuelles.

Nous aimerions nous pencher dans l'avenir sur les applications de ce noyau à d'autres tâches, comme la réponse automatique à des questions.

Références

- AGIRRE E., DIAB M., CER D. & GONZALEZ-AGIRRE A. (2012). Semeval-2012 task 6 : A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, p. 385–393 : Association for Computational Linguistics.
- BALDRIGE, J. M. T. & G. B. (2010). Opennlp.
- BU F., LI H. & ZHU X. (2012). String re-writing kernel. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers-Volume 1*, p. 449–458 : Association for Computational Linguistics.
- CALVO H., SEGURA-OLIVARES A. & GARCÍA A. (2014). Dependency vs. constituent based syntactic n-grams in text similarity measures for paraphrase recognition. *Computación y Sistemas*, **18**(3), 517–554.
- CHANG C.-C. & LIN C.-J. (2011). Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **2**(3), 27.
- DAGAN I., GLICKMAN O. & MAGNINI B. (2006). The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising textual entailment*, p. 177–190. Springer.
- DOLAN B., QUIRK C. & BROCKETT C. (2004). Unsupervised construction of large paraphrase corpora : Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, p. 350 : Association for Computational Linguistics.
- DOLAN W. B. & BROCKETT C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proc. of IWP*.
- HEILMAN M. & SMITH N. A. (2010). Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 1011–1019 : Association for Computational Linguistics.
- HICKL A., WILLIAMS J., BENSLEY J., ROBERTS K., RINK B. & SHI Y. (2006). Recognizing textual entailment with lcc's groundhog system. In *Proceedings of the Second PASCAL Challenges Workshop*.
- ISLAM A. & INKPEN D. (2009). Semantic similarity of short texts. *Recent Advances in Natural Language Processing V*, **309**, 227–236.
- JIMENEZ S., BECERRA C., GELBUKH A., BÁTIZ A. J. D. & MENDIZÁBAL A. (2013). Softcardinality : hierarchical text overlap for student response analysis. In *Proceedings of the 2nd joint conference on lexical and computational semantics*, volume 2, p. 280–284.
- LINTEAN M. C. & RUS V. (2011). Dissimilarity kernels for paraphrase identification. In *FLAIRS Conference*.
- LODHI H., SAUNDERS C., SHAW-TAYLOR J., CRISTIANINI N. & WATKINS C. (2002). Text classification using string kernels. *The Journal of Machine Learning Research*, **2**, 419–444.
- MADNANI N. & DORR B. J. (2010). Generating phrasal and sentential paraphrases : A survey of data-driven methods. *Computational Linguistics*, **36**(3), 341–387.
- MADNANI N., TETREAU J. & CHODOROW M. (2012). Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 182–190 : Association for Computational Linguistics.

- MIHALCEA R., CORLEY C. & STRAPPARAVA C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, p. 775–780.
- MOSCHITTI A. (2006). Efficient convolution kernels for dependency and constituent syntactic trees. In *Machine Learning : ECML 2006*, p. 318–329. Springer.
- PORTER M. F. (2001). Snowball : A language for stemming algorithms.
- SCHÖLKOPF B. & SMOLA A. J. (2002). *Learning with kernels : Support vector machines, regularization, optimization, and beyond*. MIT press.
- SOCHER R., HUANG E. H., PENNIN J., MANNING C. D. & NG A. Y. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems*, p. 801–809.
- VALIANT L. G. (1979). The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, **8**(3), 410–421.
- VAPNIK V. (2000). *The nature of statistical learning theory*. Springer Science & Business Media.
- WAN S., DRAS M., DALE R. & PARIS C. (2006). Using dependency-based features to take the “para-farce” out of paraphrase. In *Proceedings of the Australasian Language Technology Workshop*, volume 2006.
- ZANZOTTO F. M., DELL’ARCIPRETE L. & MOSCHITTI A. (2010). Efficient graph kernels for textual entailment recognition. *Fundamenta Informaticae*.
- ZANZOTTO F. M., PENNACCHIOTTI M. & MOSCHITTI A. (2007). Shallow semantics in fast textual entailment rule learners. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, p. 72–77 : Association for Computational Linguistics.

Extraction automatique de paraphrases grand public pour les termes médicaux

Natalia Grabar¹ Thierry Hamon²

(1) CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France
natalia.grabar@univ-lille3.fr

(2) LIMSI-CNRS, BP133, Orsay; Université Paris 13, Sorbonne Paris Cité, France
hamon@limsi.fr

Résumé. Nous sommes tous concernés par notre état de santé et restons sensibles aux informations de santé disponibles dans la société moderne à travers par exemple les résultats des recherches scientifiques, les médias sociaux de santé, les documents cliniques, les émissions de télé et de radio ou les nouvelles. Cependant, il est commun de rencontrer dans le domaine médical des termes très spécifiques (*e.g.*, *blépharospasme*, *alexitymie*, *appendicectomie*), qui restent difficiles à comprendre par les non spécialistes. Nous proposons une méthode automatique qui vise l'acquisition de paraphrases pour les termes médicaux, qui soient plus faciles à comprendre que les termes originaux. La méthode est basée sur l'analyse morphologique des termes, l'analyse syntaxique et la fouille de textes non spécialisés. L'analyse et l'évaluation des résultats indiquent que de telles paraphrases peuvent être trouvées dans les documents non spécialisés et présentent une compréhension plus facile. En fonction des paramètres de la méthode, la précision varie entre 86 et 55 %. Ce type de ressources est utile pour plusieurs applications de TAL (*e.g.*, recherche d'information grand public, lisibilité et simplification de textes, systèmes de question-réponses).

Abstract.

Automatic extraction of layman paraphrases for medical terms.

We all have health concerns and sensibility to health information available in the modern society through modern media, such as scientific research, health social media, clinical documents, TV and radio broadcast, or novels. However, medical area conveys very specific notions (*e.g.*, *blepharospasm*, *alexitymia*, *appendectomy*), which are difficult to understand by people without medical training. We propose an automatic method for the acquisition of paraphrases for technical medical terms. We expect that such paraphrases are easier to understand than the original terms. The method is based on the morphological analysis of terms, syntactic analysis of texts, and text mining of non specialized texts. An analysis of the results and their evaluation indicate that such paraphrases can indeed be found in non specialized documents and show easier understanding level. According to the setting of the method, precision of the extractions ranges between 86 and 55%. This kind of resources is useful for several Natural Language Processing applications (*e.g.*, information retrieval for lay people, text readability and simplification, question and answering systems).

Mots-clés : Domaines de spécialité, terminologie médicale, composition, analyse morphologique, paraphrase, compréhension.

Keywords: Specialized Area, Medical Terminology, Compounds, Morphological Analysis, Paraphrasis, Understanding.

1 Introduction

Nous sommes tous concernés par notre état de santé et restons sensibles aux informations de santé disponibles dans la société moderne à travers par exemple les résultats des recherches scientifiques, les médias sociaux de santé, les documents cliniques, les émissions de télé et de radio ou les nouvelles. Cependant, il est commun de rencontrer dans le domaine médical des termes très spécifiques, comme ceux présentés en exemple (1). Si la compréhension de ces termes est aisée pour certaines catégories du personnel médical (*e.g.*, médecins, étudiants en médecine, infirmiers, pharmaciens), les citoyens ordinaires non spécialistes du domaine médical peuvent avoir des difficultés de compréhension et d'utilisation de tels termes.

(1) *blépharospasme, alexitymie, appendicectomie, desmorrhexie, lombalgie*

La compréhension de tels termes est importante pour les patients et il a été montré qu'elle joue un rôle crucial pour un processus de santé réussi (AMA, 1999; McCray, 2005; Eysenbach, 2007). Toutefois, il a été également montré que ces notions ne peuvent pas être correctement maîtrisées par les patients dans plusieurs situations réelles :

- compréhension des étapes nécessaires à la préparation et la prise de médicaments (Patel *et al.*, 2002) ;
- compréhension des notices de médicaments et des informations fournies aux patients dans les brochures et les consensus informés. Par exemple, parmi 2 600 patients recrutés dans deux hôpitaux, 26 à 60 % ne peuvent pas comprendre les informations de santé fournies dans ces sources (Williams *et al.*, 1995) ;
- compréhension d'informations de santé disponibles sur les sites web à destination des patients (Berland *et al.*, 2001; Hargrave *et al.*, 2003; Kusec, 2004), et ceci en différentes langues (anglais, espagnol, français).

Ces constats peuvent avoir un impact négatif sur la communication entre les patients et les médecins, et le soins offerts aux patients (Tran *et al.*, 2009). Le contexte présenté correspond à la motivation principale de notre travail : proposer une méthode pour l'acquisition automatique de paraphrases pour expliquer les termes médicaux techniques. Plus particulièrement, nous proposons de nous concentrer sur les termes formés par la composition néoclassique (Booij, 2010; Iacobini, 1997; Amiot & Dal, 2005), comme exemplifié en (1). Une des particularités de ces termes est qu'ils impliquent souvent les bases venant du latin ou du grec (voir les exemples en (2) et (3)), ce qui les rends sémantiquement opaques et plus difficiles à comprendre que les mots formés avec les bases existant dans la langue française (*{anatomie; anatomique}*, *{livre; livresque}*). En effet, avant que le terme puisse être compris, il est d'abord nécessaire de le décomposer et de faire le lien avec la langue générale.

(2) *myocardiaque* est formé avec une base latine *myo* (*muscle*) et une base grecque *cardia* (*cœur*)

(3) *cholecystectomie* est formé avec deux bases grecques *chole* (*bile*) et *ectomy* (*ablation chirurgicale*), et une base latine *cystis* (*vessie*)

Nous présentons d'abord des travaux liés de l'état de l'art (section 2) et précisons les objectifs de notre travail (section 3). Nous présentons ensuite le matériel utilisé (section 4), et les étapes de la méthode (section 5). Nous décrivons et discutons les résultats obtenus (sections 6 et 7), et concluons avec des orientations pour les travaux futurs (section 8).

2 État de l'art

Notre travail est lié à plusieurs champs de recherche en TAL : lisibilité (section 2.1), simplification lexicale (section 2.2), construction de ressources dédiées (section 2.3) et décomposition de composés néoclassiques (section 2.4). Ces travaux ont un lien entre eux et, vus tous ensemble, présentent un problème de recherche assez complexe.

2.1 Lisibilité

Les travaux en lisibilité étudient la facilité avec laquelle un texte peut être compris. Deux types de mesures de lisibilité sont distingués : classiques et computationnelles (François, 2011). Les mesures classiques sont essentiellement basées sur le calcul du nombre de caractères et/ou syllabes dans les mots, phrases ou documents, et sur les modèles de régression linéaire (Flesch, 1948; Gunning, 1973; Dubay, 2004). Les mesures computationnelles peuvent impliquer les modèles vectoriels et une grande variété de descripteurs et de leurs combinaisons (Wang, 2006; Zeng-Treiler *et al.*, 2007; Leroy *et al.*, 2008; François & Fairon, 2013).

2.2 Simplification lexicale

La simplification lexicale aide à rendre un texte plus facile à comprendre. Par exemple, en 2012, la compétition *SemEval*¹ proposait une tâche de simplification de textes de la langue générale anglaise. Pour un texte court et un mot cible, et plusieurs substitutions possibles pour ce mot et satisfaisant le contexte, l'objectif était de trier ces substitutions selon leur degré de simplicité (Specia *et al.*, 2012). Plusieurs critères ont été exploités par les participants : lexicale extrait d'un

1. <http://www.cs.york.ac.uk/semeval-2012/>

corpus oral et de la Wikipédia, n-grammes de Google, WordNet (Sinha, 2012) ; longueur de mots, nombre des syllabes, information mutuelle et fréquence de mots (Jauhar & Specia, 2012) ; fréquence dans la Wikipédia, longueur de mots, n-grammes, complexité syntaxique des documents (Johannsen *et al.*, 2012) ; n-grammes, fréquence dans la Wikipédia, n-grammes de Google (Ligozat *et al.*, 2012) ; WordNet et fréquences de mots (Amoia & Romanelli, 2012). Les critères liés à la fréquence de mots sont parmi les plus efficaces pour la tâche. Notons cependant qu'une étape préalable à la simplification concerne la détection de mots ou passages difficiles (Grabar *et al.*, 2014) qui devraient être simplifiés avec les méthodes proposées plus haut par exemple.

2.3 Ressources dédiées

Des ressources spécifiques sont nécessaires pour effectuer la simplification des textes. Dans les domaines de spécialité, comme dans le domaine médical, ces ressources se présentent souvent sous forme de lexiques où les termes sont mis en correspondance avec les expressions non spécialisées correspondantes, comme dans les exemples (4) à (7). La première initiative de ce type est apparue avec le travail collaboratif Consumer Health Vocabulary (CHV) (Zeng & Tse, 2006) (exemples (4)). Une des méthodes proposées consiste à utiliser les requêtes médicales les plus fréquentes et à les aligner avec les termes d'UMLS (Unified Medical Language System) (Lindberg *et al.*, 1993). Ensuite, les alignements sont validés manuellement. Un autre travail a exploité un petit corpus et plusieurs mesures d'association statistique pour construire un lexique de termes techniques alignés avec leurs équivalents non techniques (Elhadad & Sutaria, 2007), les deux ensembles de termes étant fournis par l'UMLS et donc possiblement dérivés du Consumer Health Vocabulary (exemples (7)). Des travaux similaires dans d'autres langues ont suivi. En français, l'acquisition de variations morpho-syntaxiques à partir d'un corpus comparable spécialisé et non spécialisé (Deléger & Zweigenbaum, 2008; Cartoni & Deléger, 2011) a fourni des équivalences verbe/nom (exemples (5)) et un ensemble de variations syntaxiques plus large (exemples (6)). Dans ces deux travaux, la correspondance avec les terminologies médicales n'est pas établie. Notons aussi que les travaux en acquisition de variantes terminologiques (Hahn *et al.*, 2001), de synonymes (Fernández-Silva *et al.*, 2011) et de paraphrases (Max *et al.*, 2012) sont aussi pertinents pour cette thématique de recherche.

- (4) {*myocardial infarction; heart attack*}, {*abortion; termination of pregnancy*}, {*acrodynia; pink disease*}
- (5) {*consommation régulière; consommer de façon régulière*}, {*gêne à la lecture; empêche de lire*}, {*évolution de l'affection; la maladie évolue*}
- (6) {*retard de cicatrisation; retarder la cicatrisation*}, {*apports caloriques; apport en calories*}, {*calculer les doses; doses sont calculées*}, {*efficacité est renforcée; renforcer son efficacité*}
- (7) {*myocardial infarction; heart attack*}, {*SBP; systolic blood pressure*}, {*atrial fibrillation; arrhythmia*}, {*hypercholesterolemia; high cholesterol*}, {*mental stress; stress*}

2.4 Décomposition de composés néoclassiques

La décomposition de composés néoclassiques consiste à détecter leurs composants morphologiques. Dans les travaux de TAL, la décomposition est exploitée pour améliorer les résultats en indexation et recherche d'information (Lovis *et al.*, 1995; Schulz *et al.*, 1999; Hahn *et al.*, 2001) ou en traduction automatique (Loginova-Clouet & Daille, 2013). En effet, il peut être intéressant de décomposer un terme comme *iridochoroidite* en ses composants (*inflammation*, *iris* et *choroïde*) pour trouver plus de documents ou de traductions pertinents. D'autres travaux s'intéressent de plus à l'établissement de relations sémantiques entre les composants de termes de manière manuelle (Pacak *et al.*, 1980; Dujols *et al.*, 1991; Wolff, 1987) ou automatique (Daille, 2003; Grabar & Hamon, 2006). Par exemple, dans le composé *iridochoroidite* nous pouvons établir la relation de *localisation*, car une *inflammation* est localisée dans l'*iris* et le *choroïde*. La décomposition automatique des termes exploite souvent des méthodes à base de règles ou des approches probabilistes en corpus (McCray *et al.*, 1988; Namer, 2003; Loginova-Clouet & Daille, 2013; Claveau & Kijak, 2014).

3 Objectifs

Le travail que nous proposons est lié à plusieurs travaux de l'état de l'art : la décomposition de composés néoclassiques (section 2.4) et à la construction de ressources spécifiques (section 2.3). Notre objectif est de développer une méthode qui permet d'acquérir des paraphrases non spécialisées pour des termes techniques composés du domaine médical. De tels objectifs sont rarement poursuivis dans les travaux existants : seuls les exemples en (4) et (7) provenant de CHV contiennent ce type de paraphrases en anglais. Nous travaillons avec le matériel en français. Contrairement aux travaux existants, nous ne travaillons pas avec des corpus comparables spécialisés et non spécialisés, mais exploitons les termes fournis par des terminologies médicales existantes et les articles de la Wikipédia. Nous supposons que la Wikipédia peut contenir les paraphrases recherchées, comme dans {*myocardiaque*; *muscle du cœur*}, {*cholecystectomie*; *ablation de la vésicule biliaire*}. Par rapport à nos travaux précédents (Grabar & Hamon, 2014a,b), nous nous concentrons sur l'exploitation de la Wikipédia qui fournit des paraphrases plus riches (par rapport aux forums de discussion, où les paraphrases extraites sont très redondantes et offrent donc moins de couverture) et exploitons l'analyse syntaxique des textes et non pas des fenêtres de mots, ce qui permet d'extraire des paraphrases mieux fondées linguistiquement et de faire des comparaisons et évaluations plus précises des données acquises.

4 Données linguistiques

Trois types de données sont utilisés : les termes médicaux que nous voulons paraphraser (section 4.1), le corpus duquel les paraphrases sont extraites (section 4.2), et les ressources linguistiques (*i.e.* morphologie, synonymie, supplétion) qui aident à établir le lien entre les termes et le corpus (section 4.3).

4.1 Termes médicaux

Les termes médicaux proviennent de la Snomed International (Côté *et al.*, 1997)² et de la partie française d'UMLS (Lindberg *et al.*, 1993). Ces terminologies contiennent des termes syntaxiquement simples (*e.g.* *acrodynie*) et complexes (*e.g.* *infarctus du myocarde*). Nous utilisons l'ensemble des termes disponibles. Les termes syntaxiquement complexes sont segmentés en mots. Le seul filtre appliqué consiste à éliminer les mots contenant des nombres car ceux-ci correspondent le plus souvent à des composés chimiques et sont gérés par un autre type de compositionnalité (Klinger *et al.*, 2008; Jessop *et al.*, 2011). Dans ce qui suit, *mot* et *terme* sont échangeables et peuvent signifier soit l'unité graphique obtenue suite à la segmentation des termes syntaxiquement complexes, soit la notion médicale.

4.2 Corpus

Nous exploitons les articles de la Wikipédia liés au Portail de la Médecine (version de janvier 2015). Ce corpus contient 18 434 articles (15 235 219 occurrences). Le corpus contient des informations encyclopédiques sur plusieurs notions médicales. Les contributeurs ont en général une bonne connaissance des sujets abordés. L'objectif est entre autre de présenter les notions techniques et de les rendre accessibles au grand public. Nous nous attendons à ce que ces articles contiennent des paraphrases de termes techniques présentant un niveau de compréhension accessibles pour les non spécialistes.

4.3 Ressources linguistiques

Ressources morphologiques. Les ressources morphologiques comportent 155 468 paires de mots couvrant les dérivations {*aorte*; *aortique*} et les flexions {*aortique*; *aortiques*}. Elles sont issues des travaux précédents (Grabar & Zweigenbaum, 2000). Ces ressources permettent de traiter l'aspect morphologique de la variation terminologique.

Ressources de synonymes. Les ressources de synonymes proviennent également des travaux précédents (Grabar *et al.*, 2009) et ont été complétées par les synonymes simples d'UMLS. Ces ressources sont adaptées à la langue médicale. Elles contiennent 14 914 paires de synonymes, comme {*embolie*; *thrombose*}, {*tumeur*; *fibrome*}. Ces ressources sont également utilisées pour traiter la variation des termes.

2. Agence des Systèmes d'Information Partagés de Santé : esante.gouv.fr/asip-sante

Ressources supplétives. Ces ressources contiennent des paires de mots au format *{base supplétive; mot du français}*. Ce sont les ressources qui permettent de faire le lien entre les bases latines et grecques et les mots du français moderne. Ces ressources ont été construites lors des travaux précédents (Namer, 2003; Zweigenbaum & Grabar, 2003). Elles ne sont pas dédiées aux expériences présentées ici, mais elles restent néanmoins spécifiques au matériel traité que sont les termes médicaux. Ces ressources fournissent 1 022 paires, comme dans ces exemples : *{andr; mâle}*, *{ectomie; ablation}*, *{myo; muscle}*, *{para; contre}*, *{peri; autour}*.

5 Méthode

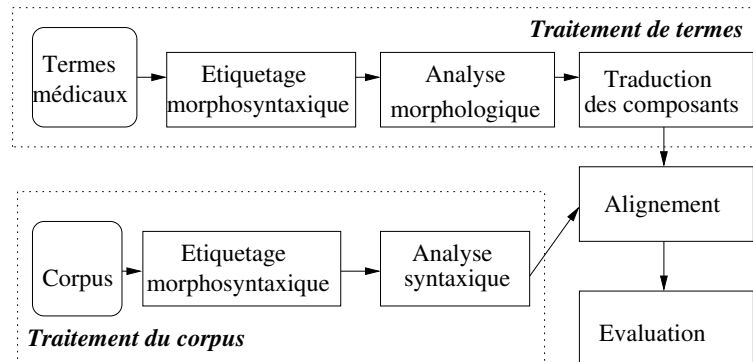


FIGURE 1 – Méthodologie générale de l'extraction de paraphrases grand public pour les termes composés.

La méthodologie est définie afin de pouvoir effectuer l'analyse des composés médicaux néoclassiques et de trouver ensuite les paraphrases correspondantes et non techniques dans les corpus. Dans certains cas, les paraphrases apparaissent dans les corpus dans des contextes définitoires (exemple (8)), auquel cas les paraphrases cooccurrent avec les termes techniques correspondant, ou bien de manière libre et sans être accompagnés de leur terme technique (exemple (9)). C'est ce deuxième type de contextes qui nous intéresse plus spécifiquement car il n'est pas contraint par l'occurrence du terme technique, et dans lequel nous pouvons en effet trouver la paraphrase *inflammation des cellules* qui correspond au terme *cellulite* dans l'exemple (9).

- (8) *La cellulite est une infection grave qui se propage sous la peau et s'attaque aux tissus mous comme la peau elle-même et les graisses sous-jacentes.*
- (9) *L'infection virale cause une inflammation des cellules nerveuses, conduisant à la destruction partielle ou totale du ganglion des motoneurones.*

La méthode est composée de quatre grandes étapes présentées à la figure 1 : le traitement de termes (section 5.1), le traitement du corpus (section 5.2), l'alignement des termes et des segments du corpus pour l'extraction de paraphrases grand public (section 5.3), et l'évaluation des extractions (section 5.4).

5.1 Traitement de termes médicaux

Pour accéder aux informations morphologiques des termes, nous effectuons trois traitements :

1. *Étiquetage morpho-syntaxique et lemmatisation des termes.* Les termes sont étiquetés morpho-syntaxiquement et lemmatisés avec *Cordial* (Laurent *et al.*, 2009). L'étiquetage morpho-syntaxique est effectué en contexte des termes. Si un mot donné reçoit plus d'une étiquette, c'est la plus fréquente qui est retenue. À cette étape, nous obtenons les lemmes des termes avec leurs parties du discours (exemple (10)).

- (10) *myocardique/A*
cholécystectomie/N
polyneuropathie/N

acromégalie/N
galactosémie/N

2. *Analyse morphologique.* Les lemmes sont ensuite analysés morphologiquement par *Dérif* (Namer, 2009). Cet outil effectue une analyse des lemmes afin de calculer leur structure morphologique, de les décomposer en leurs composants (bases et affixes), et de les analyser sémantiquement. Nous présentons des exemples de l'analyse morphologique de quelques termes en (11).

- (11) *myocardique/A* : [[[*myo N**] [*carde N**] *NOM*] *ique ADJ*]
cholécystectomie/N : [[*cholécysto N**] [*ectomie N**] *NOM*]
polyneuropathie/N : [*poly*] [[*neur N**] [*pathie N**] *NOM*] *NOM*
acromégalie/N : [[*acr N**] [*mégal N**] *ie NOM*]
galactosémie/N : [[*galactose NOM*] [*ém N**] *ie NOM*]

Les bases et affixes calculés sont associés avec les catégories syntaxiques (*NOM*, *ADJ*, *V*). Lorsqu'une base est supplétive (elle est empruntée au latin ou grec et n'existe pas en français moderne), *Dérif* lui assigne la catégorie la plus probable (e.g. *N** pour les noms, *A** pour les adjectifs). Par exemple, l'analyse de *myocardique/A* indique que ce mot contient deux bases supplétives nominales *myo N** (*muscle*) et *carde N** (*cœur*) et un affixe adjectival *-ique/ADJ*. À cette étape, les mots sont décomposés en leur composants morphologiques. Nous pouvons observer que certaines bases (e.g. *galactose* et *cholécysto*) peuvent être décomposées encore plus finement, en *galact* (*lait*) et *ose* (*sucres*), et *chole* (*bile/biliaire*) et *cystis* (*vésicule*), respectivement. Nous considérons que les mots qui contiennent plus d'une base sont des composés. Ils sont traités lors des étapes suivantes de la méthode. Comme présenté dans les exemples en (12), *Dérif* fournit également des gloses pour expliquer le sens des composés analysés.

- (12) *myocardique/A* : "(Partie de – Type particulier de) cœur en rapport avec le(s) muscle"
cholécystectomie/N : "ablation (de – vers) le(s) vésicule biliaire"
polyneuropathie/N : "neuropathies multiples, nombreux"
acromégalie/N : "Affection liée au(x) grandeur en rapport avec le(s) extrémité"
galactosémie/N : "Affection liée au(x) sang en rapport avec le(s) galactose"

3. *Association des composants morphologiques avec les mots du français.* Les bases obtenues suite à la décomposition sont associées avec (ou traduites en) mots du français moderne. Nous utilisons pour ceci la ressource de données supplétives présentées à la section 4.3. Les exemples en (13) présentent les données obtenues à cette étape.

- (13) *myocardique/A* : *myo=muscle*, *carde=cœur*
cholécystectomie/N : *cholécysto=vésicule biliaire*, *ectomie=ablation*
polyneuropathie/N : *poly=nombreux*, *neuro=nerf*, *pathie=maladie*
acromégalie/N : *acr=extrémité*, *mégal=grandeur*
galactosémie/N : *galactose=galactose*, *ém=sang*

Nous pouvons voir que, suite à cette traduction, certains mots restent techniques (e.g., *galactose*, *vésicule biliaire*), tandis que d'autres perdent tout leur sens technique (e.g. *mégal=grandeur*, *poly=nombreux*).

5.2 Traitement du corpus

Le corpus est traité par *Cordial* pour effectuer l'étiquetage morpho-syntaxique, la lemmatisation et l'analyse syntaxique. L'analyse syntaxique est utilisée pour définir les frontières des syntagmes.

5.3 Extraction de paraphrases grand public correspondant aux termes techniques

À cette étape, les mots du français qui correspondent à la décomposition morphologique des termes sont projetés sur le corpus pour en extraire les syntagmes qui contiennent les paraphrases. Nous considérons tout syntagme syntaxique, de même que les bigrammes, les trigrammes et les quadrigrammes de syntagmes. Dans l'exemple (14), un des groupes nominaux contient les mots *muscle* et *cœur*, qui correspondent aux composants morphologiques de *myocardique* (exemple (13)). Ce groupe nominal est donc un bon candidat pour fournir une paraphrase de termes *myocarde* ou *myocardique*.

- (14) *Les causes de tachycardie ventriculaire sont superposables à celles des extrasystoles ventriculaires : infarctus du myocarde, insuffisance cardiaque, hypertrophie du muscle du cœur et prolapsus de la valve mitrale.*

Nous effectuons plusieurs expériences d'extraction de paraphrases en faisant varier quatre paramètres :

- la taille de la fenêtre, qui varie d'un à quatre syntagmes syntaxiques, ce qui permet de récupérer les segments avec des paraphrases plus ou moins grandes, et donc de paraphraser des termes avec plus de composants,
- les ressources linguistiques pour gérer la variation terminologique. Nous avons alors trois possibilités : utilisation de formes brutes du corpus, utilisation de ressources morphologiques pour normaliser les flexions et les dérivations vers des lemmes, utilisation de la ressource de synonymes pour gérer les relations de synonymie au sein des paraphrases. Actuellement, nous n'effectuons pas la combinaison de ressources morphologiques et de synonymie,
- le taux d'alignement des termes techniques, ce qui permet de contrôler si tous les composants de ces termes sont alignés,
- le taux d'alignement des syntagmes syntaxiques, ce qui permet de contrôler si tous les mots des syntagmes sont alignés avec les composants.

Comme *baseline*, nous utilisons les contextes définitoires où les termes apparaissent. Les définitions (comme en (8)) sont extraites grâce aux patrons proposés dans la littérature (Péry-Woodley & Rebeyrolle, 1998), comme *est un*, *défini comme*. Avec cette approche, nous devons d'abord détecter le terme technique et ensuite le contexte définitoire correspondant. Si le test est positif, la phrase entière est extraite.

5.4 Évaluation

L'évaluation vise à vérifier si la méthode proposée permet d'acquérir les paraphrases de termes médicaux spécialisés. Les extractions sont évaluées manuellement, ce qui nous permet de calculer la précision. Pendant l'évaluation, nous distinguons quatre situations :

1. la paraphrase est correcte : *e.g. myocardique* paraphrasé en *muscle du cœur* ;
2. l'extraction de la paraphrase est basée sur une analyse morphologique incorrecte (*{sanglot; lot sang}*), la traduction vers le français n'est pas satisfaisante (*antisolaire* associé avec *sol* et *contre*), ou bien le terme traité n'est pas compositionnel et ses composants ne traduisent pas sa sémantique (*ostéodermie*, associé avec *peau* et *os* signifie *une structure d'écailles, de plaques osseuses ou d'autres compositions dans les couches dermiques de la peau, comme chez les lézards ou dinosaures*) ;
3. la paraphrase contient les informations correctes au milieu d'autres informations ou bien des informations partielles. Par exemple *endophtalmie* est paraphrasé en *interne de l'œil*, alors que son explication complète est plus large *inflammation des tissus internes de l'œil* ;
4. l'extraction est fautive et ne contient pas les informations utiles.

Ce type d'évaluation permet de calculer trois mesures :

- précision stricte $P_{stricte}$: seulement les paraphrases correctes sont considérées (cas 1) ;
- précision lâche P_{lache} : les paraphrases correctes et possiblement correctes sont considérées (cas 1 et 3) ;
- le taux d'erreurs évalue le taux d'extractions fautes (cas 4).

Les résultats sont présentés dans la section 6. Ils sont ensuite analysés du point de vue de la qualité des extractions (sections 7.1 à 7.3). Nous comparons aussi les résultats de la *baseline* avec les paraphrases extraites par la méthode proposée (section 7.4). Nous effectuons également une comparaison avec les travaux de l'état de l'art (section 7.5). Nous examinons finalement les termes qui ne reçoivent pas de paraphrases (section 7.6).

6 Résultats

Les 274 131 termes d'UMLS et de la Snomed International fournissent 76 536 mots qui ne contiennent pas de nombres. De ces mots, 15 121 sont analysés par *Dérif* et décomposés en deux bases au moins. Ces 15 121 composés correspondent donc au matériel traité par notre méthode pour l'acquisition de paraphrases. Il est possible de distinguer quatre ensembles quant à l'appariement entre les termes décomposés et les syntagmes, que nous exemplifions avec *myopathie* décomposé en *muscle* et *maladie* (les segments alignés sont soulignés dans les exemples) :

E1 : les deux unités, le terme et le syntagme, sont complètes dans l'alignement : {myo pathie; maladie du muscle}

E2 : le terme est complet mais le syntagme est partiel dans l'alignement : {myo pathie; maladie du muscle cardiaque}

$E3$: le terme est partiel mais le syntagme est complet dans l'alignement : {*myopathie*; la *maladie*}

$E4$: le terme et le syntagme sont partiels dans l'alignement : {*myopathie*; l' *origine de la maladie*}

Nous pouvons gérer cet aspect grâce aux taux d'alignement calculés. Pour la tâche visée, nous considérons qu'il est plus intéressant d'avoir un alignement complet du terme avec un alignement complet ou partiel du syntagme, ce qui correspond aux ensembles $E1$ et $E2$. L'ensemble $E1$ est le plus optimisé car il propose l'information recherchée plus exactement. Cependant, $E2$ est aussi à prendre en compte car il est possible de déduire, à partir du syntagme, la paraphrase requise.

Nombre de	unigrammes			bigrammes			trigrammes			quadrigrammes		
	<i>b</i>	<i>l</i>	<i>s</i>	<i>b</i>	<i>l</i>	<i>s</i>	<i>b</i>	<i>l</i>	<i>s</i>	<i>b</i>	<i>l</i>	<i>s</i>
<i>syntagmes</i>	9854	16093	22110	11875	18504	27670	7936	12284	19984	4701	7542	12804
<i>termes uniques</i>	1513	1947	2090	1780	2260	2463	1523	1966	2231	1079	1515	1922
<i>syntagmes_{E1}</i>	2681	4163	5370	1109	1611	2521	403	634	988	326	510	793
<i>termes uniques_{E1}</i>	668	1023	1051	492	670	962	239	358	472	204	297	419
<i>syntagmes_{E2}</i>	3893	6486	8876	3937	6290	9590	2154	3380	5138	1171	1947	3241
<i>termes uniques_{E2}</i>	1015	1358	1508	1025	1482	1693	752	1038	1401	517	768	1047

TABLE 1 – Résultats d'extraction de paraphrases pour les termes techniques.

Le tableau 1 présente les résultats d'extraction de paraphrases grand public. Nous indiquons d'abord le nombre des syntagmes extraits (*Nombre de syntagmes*) et le nombre de types de termes paraphrasés (*Nombre de termes uniques*) pour l'ensemble des résultats. Nous distinguons plusieurs expériences en fonction de la taille de la fenêtre syntaxique (*unigrammes*, *bigrammes*, *trigrammes* et *quadrigrammes*) et des ressources utilisées (*b* sans les ressources, *l* ressources pour la normalisation morphologique et *s* ressources pour la normalisation de synonymes). Nous indiquons ensuite les informations correspondantes pour les ensembles $E1$ et $E2$.

Nombre de	unigrammes			bigrammes			trigrammes			quadrigrammes		
	<i>b</i>	<i>l</i>	<i>s</i>	<i>b</i>	<i>l</i>	<i>s</i>	<i>b</i>	<i>l</i>	<i>s</i>	<i>b</i>	<i>l</i>	<i>s</i>
<i>syntagmes_{E1}</i>	2681	4163	5370	1109	1611	2521	403	634	988	326	510	793
<i>termes uniques_{E1}</i>	668	1023	1051	492	670	962	239	358	472	204	297	419
<i>correct</i>	549	785	644	378	517	461	195	290	257	175	254	235
<i>pos. correct</i>	39	32	67	22	45	75	10	19	41	7	13	35
<i>ttt termes</i>	47	60	44	28	28	46	9	10	26	9	9	26
<i>incorrect</i>	33	146	296	64	80	380	25	39	148	13	21	123
$P_{stricte}$	82	77	61	77	77	48	82	81	55	86	86	56
P_{lache}	88	80	68	81	84	40	86	86	63	89	90	64
$\%_{incorrect}$	5	14	28	13	12	39	11	11	31	6	7	29

TABLE 2 – Évaluation de l'ensemble $E1$.

Dans le tableau 2, nous indiquons les résultats d'évaluation de l'ensemble $E1$: le nombre de paraphrases correctes (*correct*), le nombre de paraphrases possiblement correctes (*pos. correct*), le nombre de paraphrases dont l'analyse morphologique ou la "traduction" doivent être améliorées (*ttt termes*), et le nombre de paraphrases incorrectes (*incorrect*). La précision varie en fonction des ressources exploitées : elle est la plus élevée lorsqu'aucune ressource n'est utilisée et la moins élevée avec les synonymes, où le risque de générer des alignements erronés est plus important. La précision stricte varie entre 86 et 55 %, et la précision lâche entre 90 et 40 %. Le taux d'erreurs varie entre 5 et 39 %. Au total, ces expériences fournissent 1 031 paraphrases correctes et 1 128 paraphrases correctes et possiblement correctes.

7 Discussion

7.1 Analyse morphologique des termes

La décomposition et analyse morphologique de *Dérif* peuvent fournir quelques erreurs ou ambiguïtés. Nous avons par exemple des décompositions ambiguës, où il existe plus d'une décomposition possible mais dont une seule est correcte.

Par exemple, *posturographie* est décomposé en : *[post [[uro N*] [graphie N*] NOM] NOM]*, ce qui peut être glosé *contrôle pendant la période qui suit la thérapie faite sur le système urinaire*. Cependant, la décomposition correcte est *[[posturo N*] [graphie N*] NOM]*, glosée *définition de la position optimale du corps en posture assise ou debout*. Certaines décompositions sont erronées, comme par exemple *sanglot* décomposé en *lot* et *sang* ou *exotique* décomposé en *externe* et *oreille*. Les décompositions erronées génèrent des paraphrases erronées à l'étape suivante. Ces erreurs ne sont pas comptabilisées dans les taux d'erreurs présentés dans le tableau 2.

7.2 Extraction de paraphrases et leur évaluation

Nous pouvons extraire plusieurs paraphrases correctes et intéressantes, comme celles présentées en (15).

- (15)
- *podalgie* : *douleur du pied* (termes bruts)
 - *mastite* : *inflammation du sein* (termes bruts)
 - *desmorrhexie* : *rupture des ligaments* (variation morphologique)
 - *bronchite* : *inflammation des bronches, inflammation bronchique* (variation morphologique : bronche->bronches, bronche->bronchique)
 - *dentalgie* : *douleurs dentaires* (variation morphologique : dents->dentaires)
 - *cystoprostatectomie* : *ablation de la vessie et de la prostate* (termes bruts)
 - *aclasie* : *absence de fracture* (variation de synonymie : cassure->fracture)
 - *enterectomie* : *résection des intestins* (variation de synonymie : ablation->résection)

Parmi les paraphrases erronées, nous trouvons parfois des erreurs de relations sémantiques entre les composants. Il s'agit typiquement de proposer la coordination entre les composants qui sont en relation de subordination (exemples en (16)). Mais le plus souvent, les corpus fournissent les relations sémantiques correctes entre les composants. Ceci correspond à un grand avantage de la méthode basée sur l'analyse syntaxique. En effet, dans le travail précédent (Grabar & Hamon, 2014b), qui exploitait des corpus plus grands et dont la méthode d'extraction était basée sur la fenêtre graphique d'une largeur donnée, le taux d'erreurs était beaucoup plus élevé, pouvant atteindre 59 % pour un nombre de termes paraphrasés moindre (273 termes avec des paraphrases correctes et 343 termes avec des paraphrases incorrectes et possiblement correctes).

- (16) *hematospermie* : *le sang ou le sperme* au lieu de *le sperme dans le sang*

D'autres paraphrases incorrectes concernent les termes qui ne sont pas compositionnels, comme *ostéodermie* ou *causalgie*, et dont le sens précis ne peut plus être dérivé de leurs composants.

Une importante partie de termes paraphrasés sont des termes à deux composants. Les termes à trois composants ou plus restent rares. L'augmentation de la fenêtre syntaxique permet justement d'augmenter la taille des termes paraphrasés. Il nous reste cependant à analyser l'ensemble *E2* pour apprécier mieux l'effet de la fenêtre syntaxique. Actuellement, la longueur moyenne des termes paraphrasés varie entre 2,002 et 2,125 composants : la plupart des termes paraphrasés contiennent deux composants. De manière générale, il est difficile de calculer le rappel des résultats obtenus. Nous pensons que la manière la plus appropriée de l'évaluer est de prendre en compte le nombre de termes analysés morphologiquement. Dans ce cas, les termes paraphrasés (1 128) couvrent 7,5 % des 15 121 termes analysés morphologiquement.

L'augmentation de la couverture est une perspective importante de notre travail.

7.3 Utilisation de ressources linguistiques

Comme nous pouvons le voir dans le tableau 2, l'utilisation de ressources linguistiques complémentaires permet d'augmenter la couverture car plus de propositions sont alors extraites, par contre cela diminue la précision car les propositions risquent alors d'apporter du bruit. Nous pouvons aussi voir que l'utilisation de ressources de synonymie mène à l'extraction d'un plus grand pourcentage de propositions erronées. Comme dans d'autres tâches en recherche et extraction d'information, l'explication principale est que les synonymes correspondent souvent à des valeurs contextuelles : selon les contextes ils sont plus ou moins acceptables. En revanche, les ressources morphologiques contiennent des paires de mots dont la substituabilité contextuelle est plus évidente : la variation flexionnelle ou dérivationnelle n'apporte que peu de changements sémantiques. En (17), nous présentons quelques exemples d'erreurs dues à l'utilisation de synonymes. Pour

un composé néoclassique donné (*i.e. cardialgie*), nous indiquons sa sémantique attendue (*douleur de cœur*) et parfois sa décomposition. Nous présentons ensuite la ou les paraphrases erronées extraites pour ce composé (*plaie du cœur*) et la raison de cette extraction. Dans l'exemple cité, il s'agit de l'utilisation de la paire de synonymes {*douleur; plaie*}. Ces synonymes sont corrects et mutuellement substituables dans plusieurs contextes, mais pas dans le contexte de la paraphrase de *cardialgie*. Notons que dans la compétition de simplification proposée par *SemEval* (Specia *et al.*, 2012), les candidats à remplacement étaient pré-sélectionnés et satisfaisaient le contexte. Tandis que dans notre travail, la pré-sélection des ressources n'est pas effectuée.

- (17) - *cardialgie (douleur de cœur) : plaie du cœur – {douleur; plaie}*
 - *cheiropathie (maladie des mains) : Le syndrome main – {maladie; syndrome}*
 - *choroïde (est décomposé en forme et membrane, et signifie une des couches de la paroi du globe oculaire) : aspect de l'épithélium – {forme; aspect}, {membrane; épithélium}*
 - *cinépathie (est décomposé en mouvement et maladie, est aussi connue sous le terme de mal des transports) : évolution du syndrome – {mouvement; évolution} et {maladie; syndrome}*

Comme nous l'avons noté, nous n'effectuons pas actuellement la combinaison de ressources morphologiques et de synonymie pour deux raisons : le coût de calcul devient alors très élevé et de plus cela multiplie les erreurs dues à la synonymie. Lorsque nous pourrions gérer mieux les valeurs contextuelles des synonymes, la combinaison de ces deux types de ressources pourra apporter des solutions pour augmenter la couverture de termes médicaux paraphrasés.

7.4 Comparaison avec les contextes définitoires

Le nombre total de définitions extraites avec les patrons définitoires est 2 037, portant sur 1 286 termes uniques. Le patron le plus fréquent *est un* est reconnu le plus fréquemment. D'autres patrons, comme également *appelé et peut être défini comme*, sont aussi trouvés mais avec une fréquence moindre. Nous distinguons les définitions correctes (exemple (18)) et les définitions incorrectes ou apportant des informations non suffisantes pour la compréhension (exemple (19)). Comme pour la méthode principale, le calcul de la précision stricte est basé sur les définitions correctes, tandis que la précision lâche accepte aussi les définitions possiblement correctes. La précision stricte est de 52,5 %, et la précision lâche de 68 %.

- (18) *L'angiographie est une technique d'imagerie médicale portant sur les vaisseaux sanguins qui ne sont pas visibles sur des radiographie s standards.*
La néphrite est une inflammation du rein (du grec : nephro- , le rein, et -itis , inflammation).
- (19) *L'angiographie est un examen invasif.*
Les deux principales causes de néphrite sont les infections ou les maladies auto-immunes.

Les contextes considérés comme corrects fournissent les définitions pour 849 termes, alors que nous obtenons des définitions correctes ou possiblement correctes pour 1 028 termes. Parmi ces termes, nous trouvons :

1. termes composés : *achillodynie, clinodactylie, dyslexie, bronchodilatateur,*
2. termes affixés : *choroïde, amaigrissement, surmoi,*
3. termes morphologiquement simples : *acide, hypnose, deni.*

En relation avec la méthode d'acquisition de paraphrases de composés néoclassiques, seules les définitions pour les termes composés sont comparables directement. Comme les définitions portant sur les composés néoclassiques correspondent à la majorité des termes définis, nous prenons en compte toutes les définitions extraites. La qualité des définitions est variable. Certains termes sont sous-définis (*L'adénomyose est un type d'endométriose interne*) ou bien gardent des définitions très techniques. Par exemple, les trois définitions de *péricarde* qui suivent ont des niveaux de lisibilité variables. À notre avis, la première définition est la plus appropriée pour les non experts :

- *La couche extérieure du cœur est appelée péricarde.*
- *Le péricarde est un sac à double paroi contenant le cœur et les racines des gros vaisseaux sanguins.*
- *Le péricarde est un organe de glissement, formé de deux feuillets limitant une cavité virtuelle, la cavité péricardique, qui permet les mouvements cardiaques.*

En comparaison avec la méthode principale, nous pouvons observer que les paraphrases couvrent un nombre légèrement plus important de termes. Nous nous sommes attendus à ce résultat car la méthode d'extraction de paraphrases ne requiert

pas la présence du terme analysé dans le texte, mais seulement de ses composants. Concernant la précision, elle est ainsi également plus élevée avec la méthode de paraphrase. Quant à l'utilité de ces définitions, nous pensons qu'elles peuvent être utilisées telles quelles ou bien transformées en paraphrases. Dans les deux cas, elles sont supplémentaires aux paraphrases extraites. Les deux ensembles (paraphrases issues de l'ensemble *E1* et contextes définitoires) fournissent 1 827 termes définis ou paraphrasés correctement et 2 089 termes définis ou paraphrasés correctement ou possiblement correctement.

7.5 Comparaison avec les travaux existants

Nous pouvons comparer les résultats obtenus avec ceux présentés dans trois travaux existants (Deléger & Zweigenbaum, 2008; Cartoni & Deléger, 2011; Elhadad & Sutaria, 2007) :

- *Types de termes*. Dans notre étude, nous travaillons surtout avec les termes composés, qui sont assez difficiles à comprendre par les locuteurs, et pour lesquels les paraphrases grand public apportent des informations nécessaires à leur compréhension. Dans les travaux existants (exemples (4) à (6)), seul le travail sur l'anglais (Elhadad & Sutaria, 2007) fournit des paraphrases des termes composés, tandis que les deux autres travaux (Deléger & Zweigenbaum, 2008; Cartoni & Deléger, 2011) se concentrent sur la variation morpho-syntaxique des termes ;
- *Nombre de paraphrases extraites*. Dans notre étude, nous extrayons 1 031 paraphrases correctes et 1 128 paraphrases correctes et possiblement correctes. Comme nous l'indiquons dans la section 7.4, les définitions améliorent la couverture. Dans les travaux existants, nous pouvons noter l'extraction de 65 et 82 paraphrases (Deléger & Zweigenbaum, 2008), de 109 paraphrases (Cartoni & Deléger, 2011), et de 152 paraphrases (Elhadad & Sutaria, 2007) ;
- *Précision des résultats*. Dans notre étude, les valeurs de la précision lâche varient en fonction des ressources et des fenêtres syntaxiques exploitées entre 90 et 40 %, avec une moyenne de 76 % sur l'ensemble des expériences et de 86 % pour les expériences sans l'utilisation de ressources de synonymes. Dans les travaux existants, la précision est de 67 % et 60 % (Deléger & Zweigenbaum, 2008), 66 % (Cartoni & Deléger, 2011), et 58 % (Elhadad & Sutaria, 2007).

Notons aussi dans les trois travaux cités, un seul (Elhadad & Sutaria, 2007) est basé sur l'exploitation de termes venant d'une terminologie existante. Les autres travaux exploitent le contenu des corpus et n'établissent pas de lien avec les terminologies existantes. De manière générale, notre travail va au-devant des travaux de l'état de l'art pour les paramètres discutés ici. Il est difficile de comparer nos résultats avec les travaux autour de la construction du CHV, car il s'agit d'une série de plusieurs travaux souvent faits de manière collaborative.

Par rapport aux gloses proposées par DériF (Namer, 2009), en (15), nous présentons les paraphrases pour quelques termes, qui sont à comparer avec les termes glosés en (12). Nous pensons que les paraphrases extraites offrent des informations exprimées plus naturellement et sont plus faciles à comprendre. Notons cependant que, grâce au langage formel de DériF, tous les termes décomposés et analysés morphologiquement reçoivent une glose, alors que la couverture de paraphrases que nous extrayons dépend du contenu des corpus et des ressources linguistiques exploitées.

7.6 Termes non paraphrasés

Plusieurs termes restent non paraphrasés, comme ceux présentés en (20). Une des raisons est que certains termes, comme *hémidesmosome* ou *hémohistioblaste*, contiennent plus de deux composants, ce qui rend la détection de leurs paraphrases plus difficile. Nous avons vu cependant qu'avec l'augmentation des fenêtres syntaxiques la taille des termes paraphrasés augmente également. D'autres termes non analysés contiennent des préfixes ou des composants qui apparaissent moins fréquemment dans les textes. Nous pensons que l'utilisation de corpus complémentaires permettra d'acquérir d'autres paraphrases. Un autre fait qui peut réduire le taux d'extraction de paraphrases concerne l'association de composants supplétifs avec les mots du français. En effet, plusieurs traductions sont parfois possibles mais ne peuvent pas être captées avec la méthode de traduction actuelle. D'autres méthodes, comme par exemple celle proposée dans (Claveau & Kijak, 2014), devraient être exploitées pour améliorer cet aspect.

- (20) *leptoméningé* : *affaibli, méningé*
hémipénis : *pénis, demi*
hémidesmosome : *corpuscule, demi, ligament*
hémohistioblaste : *cellule embryonnaire, tissu, sang*

8 Conclusion et travaux futurs

Nous avons proposé d’exploiter les articles de la Wikipédia pour détecter les paraphrases pour les termes techniques du domaine médical. Nous nous sommes concentrés sur les composés (e.g., *myocardiaque*, *cholecystectomie*, *galactose*, *acromégalie*). Les données traitées sont en français. La méthode s’appuie sur l’analyse morphologique de termes, la traduction des composants de termes vers le français moderne (e.g. {*card*; *cœur*}), et leur projection sur les syntagmes syntaxiques. La méthode permet d’extraire les paraphrases correctes et possiblement correctes pour 1 128 termes composés, tandis que les définitions fournissent des explications pour 1 028 termes. Mis ensemble, cela correspond à 2 089 termes. Un des avantages de la méthode est que les relations sémantiques entre les composants sont aussi extraites à partir des textes. Nous pensons que cette méthode peut en effet être utilisée pour la création d’un lexique nécessaire pour la simplification de termes médicaux. Notons aussi que la méthode proposée traite les composés néoclassiques qui en général ne sont pas traités par les méthodes existantes, car ils ne présentent pas de similarité formelle avec leurs paraphrases.

Une des difficultés actuelles est liée à la couverture des termes paraphrasés ou définis. Dans les travaux futurs, nous prévoyons d’utiliser d’autres méthodes, comme par exemple les méthodes distributionnelles (Claveau & Kijak, 2014), pour la segmentation de termes et leur association aux mots du français. Il est en effet possible qu’actuellement cette étape soit trop restrictive. Des corpus plus grands doivent aussi être exploités pour couvrir plus de matériel linguistique.

Nous voulons aussi traiter les termes complexes syntaxiquement (e.g. *vaporisateur hypodermique*, *fistule trachéo-œsophagienne*, *cardiopathie artérioscléreuse*), car ils peuvent aussi être difficiles à comprendre par les patients. La méthode proposée peut être appliquée à d’autres langues lorsque l’analyse morphologique et l’association aux mots de la langue peuvent être effectuées. L’objectif final de notre travail est d’exploiter la ressource, qui met en relation les termes spécialisés et leurs paraphrases grand public, pour la simplification de textes de spécialité.

Références

- AMA (1999). Health literacy : report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. *JAMA*, **281**(6), 552–7.
- AMIOT D. & DAL G. (2005). Integrating combining forms into a lexeme-based morphology. In *Mediterranean Morphology Meeting (MMM)*, p. 323–336.
- AMOIA M. & ROMANELLI M. (2012). SB : mmSystem - using compositional semantics for lexical simplification. In **SEM 2012*, p. 482–486, Montréal, Canada.
- BERLAND G., ELLIOTT M., MORALES L., ALGAZY J., KRAVITZ R., BRODER M., KANOUSE D., MUNOZ J., PUYOL J. & ET AL M. L. (2001). Health information on the internet. accessibility, quality, and readability in english and spanish. *JAMA*, **285**(20), 2612–2621.
- BOOIJ G. (2010). *Construction Morphology*. Oxford : Oxford University Press.
- CARTONI B. & DELÉGER L. (2011). Découverte de patrons paraphrastiques en corpus comparable : une approche basée sur les n-grammes. In *TALN*.
- CLAVEAU V. & KIJAK E. (2014). Generating and using probabilistic morphological resources for the biomedical domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, p. 3348–3354.
- CÔTÉ R. A., BROCHU L. & CABANA L. (1997). *SNOMED Internationale – Répertoire d’anatomie pathologique*. Secrétariat francophone international de nomenclature médicale, Sherbrooke, Québec.
- DAILLE B. (2003). Conceptual structuring through term variations. In *Proceedings of the ACL Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, p. 9–16.
- DELÉGER L. & ZWEIGENBAUM P. (2008). Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In *AMIA 2008*, p. 146–50.
- DUBAY W. H. (2004). The principles of readability. *Impact Information*. Available at <http://almacenplantillasweb.es/wp-content/uploads/2009/11/The-Principles-of-Readability.pdf>.
- DUJOLS P., AUBAS P., BAYLON C. & GRÉMY F. (1991). Morphosemantic analysis and translation of medical compound terms. *Methods in Informatics and Medicine (MIM)*, **30**, 30–35.
- ELHADAD N. & SUTARIA K. (2007). Mining a lexicon of technical terms and lay equivalents. In *BioNLP*, p. 49–56.

- EYSENBACH G. (2007). Poverty, human development, and the role of ehealth. *J Med Internet Res*, **9**(4), e34.
- FERNÁNDEZ-SILVA S., FREIXA J. & CABRÉ M. (2011). A proposed method for analysing the dynamics of cognition through term variation. *Terminology*, **17**(1), 49–73.
- FLESCH R. (1948). A new readability yardstick. *Journal of Applied Psychology*, **23**, 221–233.
- FRANÇOIS T. (2011). *Les apports du traitements automatique du langage à la lisibilité du français langue étrangère*. Phd thesis, Université Catholique de Louvain, Louvain.
- FRANÇOIS T. & FAIRON C. (2013). Les apports du TAL à la lisibilité du français langue étrangère. *TAL*, **54**(1), 171–202.
- GRABAR N. & HAMON T. (2006). Terminology structuring through the derivational morphology. In T. SALAKOSKI, F. GINTER, S. PYYSALO & T. PAHIKKALA, Eds., *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, number 4139 in LNAI, p. 652–663 : Springer.
- GRABAR N. & HAMON T. (2014a). Automatic extraction of layman names for technical medical terms. In *ICHI 2014*, Pavia, Italy.
- GRABAR N. & HAMON T. (2014b). Unsupervised method for the acquisition of general language paraphrases for medical compounds. In *Computerm 2014*, Dublin, Ireland.
- GRABAR N., HAMON T. & AMIOT D. (2014). Automatic diagnosis of understanding of medical words. In *Workshop on Predicting and Improving Text Readability for Target Reader Populations*, p. 11–20, Gothenburg, Sweden.
- GRABAR N., VAROUTAS P., RIZAND P., LIVARTOWSKI A. & HAMON T. (2009). Automatic acquisition of synonym ressources and assessment of their impact on the enhanced search in EHRs. *Methods of Information in Medicine*, **48**(2), 149–154. PMID 19283312.
- GRABAR N. & ZWEIGENBAUM P. (2000). A general method for sifting linguistic knowledge from structured terminologies. *JAMIASUP*, p. 310–314.
- GUNNING R. (1973). *The art of clear writing*. New York, NY : McGraw Hill.
- HAHN U., HONECK M., PIOTROWSKY M. & SCHULZ S. (2001). Subword segmentation - leveling out morphological variations for medical document retrieval. In *AMIA*, 229–233.
- HARGRAVE D., BARTELS U., LAU L., ESQUEMBRE C. & BOUFFET E. (2003). évaluation de la qualité de l'information médicale francophone accessible au public sur internet : application aux tumeurs cérébrales de l'enfant. *Bulletin du Cancer*, **90**(7), 650–5.
- IACOBINI C. (1997). Distinguishing derivational prefixes from initial combining forms. In *First mediterranean conference of morphology*, Mytilene, Island of Lesbos, Greece.
- JAUHAR S. & SPECIA L. (2012). UOW-SHEF : SimpLex – lexical simplicity ranking based on contextual and psycholinguistic features. In **SEM 2012*, p. 477–481, Montréal, Canada.
- JESSOP D., ADAMS S., WILLIGHAGEN E., HAWIZY L. & MURRAY-RUST P. (2011). Oscar4 : a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, **3**(41).
- JOHANNSEN A., MARTÍNEZ H., KLERKE S. & SØGAARD A. (2012). Emnlp@cph : Is frequency all there is to simplicity ? In **SEM 2012*, p. 408–412, Montréal, Canada.
- KLINGER R., KOLÁRIK C., FLUCK J., HOFMANN-APITIUS M. & FRIEDRICH C. (2008). Detection of iupac and iupac-like chemical names. In *ISMB 2008*, p. 268–276.
- KUSEC S. (2004). Les sites web relatifs au diabète, sont-ils lisibles ? *Dibète et société*, **49**(3), 46–48.
- LAURENT D., NÈGRE S. & SÉGUÉLA P. (2009). Apport des cooccurrences à la correction et à l'analyse syntaxique. In *TALN*.
- LEROY G., HELMREICH S., COWIE J., MILLER T. & ZHENG W. (2008). Evaluating online health information : Beyond readability formulas. In *AMIA 2008*, p. 394–8.
- LIGOZAT A., GROUIN C., GARCIA-FERNANDEZ A. & BERNHARD D. (2012). Annlor : A naïve notation-system for lexical outputs ranking. In **SEM 2012*, p. 487–492.
- LINDBERG D., HUMPHREYS B. & MCCRAY A. (1993). The unified medical language system. *Methods Inf Med*, **32**(4), 281–291.
- LOGINOVA-CLOUET E. & DAILLE B. (2013). Segmentation multilingue des mots composés. In *TALN 2013*, p. 564–571.
- LOVIS C., MICHEL P.-A., BAUD R. & SCHERRER J.-R. (1995). Word segmentation processing : a way to exponentially extend medical dictionaries. In *Medical Informatics in Europe (MIE)*, p. 28–32.

- MAX A., BOUAMOR H. & VILNAT A. (2012). Generalizing sub-sentential paraphrase acquisition across original signal type of text pairs. In *EMNLP*, p. 721–31.
- MCCRAY A. (2005). Promoting health literacy. *J of Am Med Infor Ass*, **12**, 152–163.
- MCCRAY A. T., BROWNE A. C. & MOORE D. (1988). The semantic structure of neo-classical compounds. In *Proceedings of the Annual SCAMC*, p. 165–168.
- NAMER F. (2003). Automatiser l’analyse morpho-sémantique non affixale : le système DériF. *Cahiers de Grammaire*, **28**, 31–48.
- NAMER F. (2009). *Morphologie, Lexique et TAL : l’analyseur DériF. TIC et Sciences cognitives*. London : Hermes Sciences Publishing.
- PACAK M. G., NORTON L. M. & DUNHAM G. S. (1980). Morphosemantic analysis of -itis forms in medical language. *Methods in Medical Informatics (MIM)*, **19**(2), 99–105.
- PATEL V., BRANCH T. & AROCHA J. (2002). Errors in interpreting quantities as procedures : The case of pharmaceutical labels. *International journal of medical informatics*, **65**(3), 193–211.
- PÉRY-WOODLEY M. & REBEYROLLE J. (1998). Domain and genre in sublanguage text : definitional microtexts in three corpora. In *First International Conference on Language Resources and Evaluation*, p. 987–992.
- SCHULZ S., ROMACKER M., FRANZ P., ZAISS A., KLAR R. & HAHN U. (1999). Towards a multilingual morpheme thesaurus for medical free-text retrieval. In *Medical Informatics in Europe (MIE)*, p. 891–894.
- SINHA R. (2012). Unt-simprank : Systems for lexical simplification ranking. In **SEM 2012*, p. 493–496.
- SPECIA L., JAUHAR S. & MIHALCEA R. (2012). Semeval-2012 task 1 : English lexical simplification. In **SEM 2012*, p. 347–355.
- TRAN T., CHEKROUD H., THIERY P. & JULIENNE A. (2009). Internet et soins : un tiers invisible dans la relation médecine/patient ? *Ethica Clinica*, **53**, 34–43.
- WANG Y. (2006). Automatic recognition of text difficulty from consumers health information. In IEEE, Ed., *Computer-Based Medical Systems*, p. 131–136.
- WILLIAMS M., PARKER R., BAKER D., PARIKH N., PITKIN K., COATES W. & NURSS J. (1995). Inadequate functional health literacy among patients at two public hospitals. *JAMA*, **274**(21), 1677–1682.
- WOLFF S. (1987). Automatic coding of medical vocabulary. In N. SAGER, C. FRIEDMAN & M. S. LYMAN, Eds., *Medical Language Processing. Computer Management of Narrative Data*, chapter 7, p. 145–162. New-York : Addison-Wesley.
- ZENG Q. & TSE T. (2006). Exploring and developing consumer health vocabularies. *JAMIA*, **13**, 24–29.
- ZENG-TREILER Q., KIM H., GORYACHEV S., KESELMAN A., SLAUGHTER L. & SMITH C. (2007). Text characteristics of clinical reports and their implications for the readability of personal health records. In *MEDINFO*, p. 1117–1121, Brisbane, Australia.
- ZWEIGENBAUM P. & GRABAR N. (2003). Corpus-based associations provide additional morphological variants to medical terminologies. In *AMIA*, p. 768–772.

Attribution d'Auteur : approche multilingue fondée sur les répétitions maximales

Romain Brixtel¹, Charlotte Lecluze², Gaël Lejeune³

(1) Université de Lausanne – HEC - Département de comportement organisationnel, Quartier Dorigny, 1015 Lausanne, Suisse

(2) GREYC, Campus Côte de Nacre, Boulevard du Maréchal Juin, 14032 CAEN cedex 5, France

(3) LINA, 2 rue de la Houssinière, 44322 Nantes, France

prenom.nom@unil.ch, unicaen.fr, univ-nantes.fr

Résumé. Cet article s'attaque à la tâche d'Attribution d'Auteur en contexte multilingue. Nous proposons une alternative aux méthodes supervisées fondées sur les n -grammes de caractères de longueurs variables : les *répétitions maximales*. Pour un texte donné, la liste de ses n -grammes de caractères contient des informations redondantes. A contrario, les *répétitions maximales* représentent l'ensemble des répétitions de ce texte de manière condensée. Nos expériences montrent que la redondance des n -grammes contribue à l'efficacité des techniques d'Attribution d'Auteur exploitant des sous-chaînes de caractères. Ce constat posé, nous proposons une fonction de pondération sur les traits donnés en entrée aux classifieurs, en introduisant les répétitions maximales du $n^{\text{ème}}$ ordre (c'est-à-dire des répétitions maximales détectées dans un ensemble de répétitions maximales). Les résultats expérimentaux montrent de meilleures performances avec des répétitions maximales, avec moins de données que pour les approches fondées sur les n -grammes.

Abstract.

Authorship Attribution through Character Substrings (and *vice versa*)

This article tackles the Authorship Attribution task according to the language independence issue. We propose an alternative of variable length character n -gram features in supervised methods : *maximal repeats* in strings. When character n -grams are by essence redundant, maximal repeats are a condensed way to represent any substring of a corpus. Our experiments show that the redundant aspect of character n -grams contributes to the efficiency of character-based Authorship Attribution techniques. Therefore, we introduce a new way to weight features in vector based classifier by introducing n -th order *maximal repeats* (maximal repeats detected in a set of maximal repeats). The experimental results show higher performance with maximal repeats, with less data than n -grams based approach.

Mots-clés : attribution d'auteur, multilinguisme, classification, chaînes de caractères, répétitions maximales.

Keywords: authorship attribution, multilinguism, classification, character substrings, maximal repeats.

1 Introduction

Internet donne la possibilité de partager facilement son opinion, de communiquer des informations ou publier ses productions. La mention de l'auteur n'y est pas alors systématiquement présente. La fouille de données textuelles permet de classer les auteurs par catégorie (par genre, âge ou par opinion politique) ou en tant qu'individu. Ce dernier cas de figure est appelé le problème d'Attribution d'Auteur (AA). Cela consiste à deviner l'auteur de textes à partir d'un ensemble de candidats. Ainsi, cette tâche peut être vue comme un sous-domaine de l'apprentissage automatique supervisé. Techniquement cela consiste à définir une nouvelle paire reliant un texte à un auteur. Ces méthodes peuvent aussi être utilisées pour savoir si un auteur est facilement détectable via ses productions dans un flux de textes. Ce domaine est aussi connu sous le nom de *writeprint*, en référence aux termes anglais « écriture » (*write*) et « empreinte digitale » (*fingerprint*). Pour un état de l'art complet, se référer aux travaux de Koppel *et al.* (2009), de Stamatatos (2009) et de El Bouanani & Kassou (2014).

La tâche d'AA est le plus souvent abordée sous l'angle de la stylométrie (ou étude du style). L'hypothèse sous-jacente est qu'un auteur laisse involontairement dans son message textuel des indices qui peuvent mener à son identification. El Bouanani & Kassou (2014) définissent un ensemble de traits (numériques) qui demeurent relativement constants pour un auteur donné et qui distinguent suffisamment son style d'écriture des autres auteurs. Dans de précédentes recherches, des données numériques – telles que la longueur des mots – et des données littérales – telles que des suites de mots ou de caractères – ont été utilisées pour capturer des traits stylistiques personnels (Koppel *et al.*, 2011). Si l'exploitation des

mots et des lemmes nécessite des ressources *a priori*, l'exploitation des chaînes de caractères d'un texte est indépendante de la langue de ce texte. Un profil d'auteur est alors construit à partir des n -grammes contenus dans les textes qui lui sont associés. Des techniques d'apprentissage automatique supervisé sont utilisées pour apprendre à partir de ces profils, en fonction d'un corpus d'entraînement où les paires (texte, auteur) sont connues. À l'issue, ces résultats sont utilisés pour attribuer de nouveaux textes au bon auteur. Il s'agit d'une classification multi-critères. SVM (*Support Vector Machine* ou Séparateur à Vaste Marge) est une méthode phare pour aborder de telles tâches en AA (Sun *et al.*, 2012). Nous adoptons la même approche dans cet article.

L'AA consiste à prédire l'auteur d'un texte à partir d'un ensemble de candidats. La difficulté augmente quand les objets d'étude proviennent du Web où se côtoient différents genres textuels, styles et langues. Dès lors, les recherches en AA peuvent se concentrer sur certains de ces problèmes : le passage à l'échelle quand un grand nombre d'auteurs candidats est considéré ou l'indépendance vis-à-vis de la langue lorsque les ressources linguistiques sont rares ou manquantes.

Dans ces travaux, l'indépendance vis-à-vis de la langue est abordée avec des méthodes fondées sur les caractères. Le calcul et l'exploitation de toutes les chaînes de caractères d'un texte est coûteux. La contribution principale de cet article consiste en l'utilisation d'un nouvel algorithme pour manipuler des chaînes de caractères, en vue de réduire les données et ainsi le temps et le coût d'entraînement, et ce sans perdre de précision lors de l'attribution des paires (texte, auteur). L'approche classique fondée sur les n -grammes de caractères de longueurs variables est comparée à une approche exploitant des *répétitions maximales* ainsi que des *répétitions maximales du 2^{ème} ordre*. Les expériences ont mené à la constitution de trois corpus : un en anglais, un en français et un correspondant à la concaténation des deux autres.

Les apports de cet article sont les suivants :

- nous présentons une alternative aux n -grammes de caractères en AA via les répétitions maximales ;
- nous montrons l'effet bénéfique de la redondance des n -grammes sur les méthodes utilisant une représentation vectorielle des textes ;
- en conséquence, nous proposons une nouvelle manière de prendre en compte l'interdépendance longueur-effectif pour la pondération d'une chaîne de caractères en fonction des sous-chaînes qu'elle encapsule.

Ces apports concernent l'AA, mais d'autres tâches manipulant des chaînes de caractères peuvent être considérées.

Cet article est organisé comme suit : la Section 2 présente l'état de l'art et les principaux traits utilisés dans cette tâche de classification. La Section 3 introduit le cadre expérimental, le corpus et ses caractéristiques ainsi que la chaîne de traitement. La Section 4 décrit les traits utilisés, en détaillant l'algorithme des répétitions maximales. La Section 5 expose les résultats expérimentaux et la Section 6 dresse les perspectives de cette nouvelle approche.

2 État de l'art

L'AA est une tâche de catégorisation multiclasse de textes à label unique. Comme détaillé dans Sun *et al.* (2012), trois caractéristiques principales doivent être définies : la nature des traits exploités, l'ensemble des traits représentant un texte et la façon de manipuler ces représentations pour relier un texte à un auteur.

2.1 Définitions des traits

Les traits utilisés en AA peuvent être séparés en différents groupes (Abbasi & Chen, 2008) :

- des valeurs numériques associées à des mots (nombre de mots dans les textes, nombre de caractères par mot, nombre de bi-grammes/tri-grammes de caractères au sein de ces mots) autrement dit des traits lexicaux ;
- des valeurs associées à la syntaxe des phrases (effectifs des mots outils, des mono-grammes/bi-grammes/tri-grammes de ces mots outils ou des séquences de parties du discours) ;
- des valeurs numériques associées à des unités plus grandes (nombre de paragraphes ou encore longueur moyenne des paragraphes), autrement dit des traits structurels ;
- des valeurs associées avec le contenu thématique (des sacs de mots, des n -grammes de mots clefs) ;
- des particularités en rapport avec les pratiques individuelles (telles que les fautes d'orthographe ou de frappe).

Parmi ces traits, certains sont spécifiques à des types de langue et de graphie. Si découper un texte en mots est aisé dans certains cas (en définissant un mot comme une chaîne de caractères entourée d'espaces), ce n'est pas une tâche triviale en chinois ou en japonais. Les approches exploitant les n -grammes de caractères apparaissent comme étant les plus simples pour traiter n'importe quelle langue (naturelle (Grieve, 2007; Stamatatos, 2006) ou non (Burrows *et al.*, 2014)), ainsi que les plus performantes.

Comme évoquée par Bender (2009), une méthode indépendante des langues ne doit pas forcément être dépourvue de considérations linguistiques. Si l'extraction de n -grammes est réalisée indépendamment de la langue traitée, le choix du paramètre n doit être fait en fonction des langues abordées. Étant donné les différences morphologiques des langues (flexionnelles, agglutinantes, *etc.*), ce paramètre ne pourra pas amener les mêmes résultats selon la langue.

Sun *et al.* (2012) défendent qu'utiliser une valeur fixe de n ne peut mener qu'à l'extraction d'informations lexicales (pour de petites valeurs de n), contextuelles ou thématiques (pour des plus grandes valeurs), mais n'expliquent pas pourquoi ou si cela est valide pour le chinois ou toutes les langues. Les auteurs soutiennent que cet inconvénient est évitable en exploitant des n -grammes de longueurs variables (des sous-chaînes de longueur entre 1 et n), donc en capturant des informations de types différents (lexicales, contextuelles et thématiques). Des sous-chaînes de longueurs variables sont également exploitées dans cette étude pour voir l'impact de ce paramètre sur les résultats en français et en anglais.

2.2 Représentation des textes et des auteurs fondée sur les traits

Un même trait peut être attribué à plusieurs paires (texte, auteur) mais chaque texte et auteur ne partagent pas pour autant un grand ensemble de traits. Différents ensembles de traits peuvent être définis pour représenter des textes (et par extension, pour représenter des auteurs). Considérant les méthodes d'AA existantes, deux catégories principales de traits peuvent être définies :

- les traits *hors-ligne* : traits *a priori* considérés pertinents pour cette tâche avec une connaissance préalable, comme ceux largement décrit par Chaski (2001). Ils peuvent être définis quand le corpus à traiter n'est pas encore collecté.
- les traits *en-ligne* : traits définis pendant le traitement (dans le cas de méthodes supervisées, en fonction des corpus d'entraînement et de test, comme le modèle de langue de caractères décrit par Peng *et al.* (2003)). Ils ne peuvent être définis que lorsque le corpus à traiter est complet.

Les traits *en-ligne* renvoient naturellement à la notion d'indépendance vis-à-vis des langues, aucun *a priori* n'est émis avant le traitement du corpus et aucune ressource linguistique extérieure n'est exploitée. La méthode décrite dans cet article suit ce principe.

2.3 Catégorisation de textes fondée sur les traits

Différentes techniques pour exploiter les traits extraits des textes ont été proposées. SVM (*Support Vector Machine* ou Séparateur à Vaste Marge) et les réseaux de neurones (*neural network*) sont des approches efficaces pour mener la tâche d'AA suivant le paradigme d'apprentissage automatique supervisé (Kacmarcik & Gamon, 2006; Tweedie *et al.*, 1996). Quand l'ensemble des auteurs candidats est extrêmement grand ou incomplet, d'autres approches comparent les textes comme des ensembles de traits avec des fonctions spécifiques pour calculer les similarités entre ces ensembles (Koppel *et al.*, 2011). D'autres approches utilisent des ensembles de traits individuels *via* apprentissage automatique pour construire un classifieur par auteur. Chaque classifieur agit tel un expert pour traiter un sous-ensemble de l'espace de recherche lors de la classification d'un corpus, chaque classifieur étant spécialisé dans la détection d'un auteur spécifique. Les expériences décrites dans cet article utilisent un unique classifieur SVM pour l'ensemble des auteurs en gardant les mêmes paramètres pour chaque expérience, en vue d'analyser finement l'influence du choix des traits sur le traitement. Cette analyse sur les traits est alors en principe valide, même pour d'autres méthodes se basant sur ces mêmes traits.

3 Chaîne de traitement expérimentale et description du corpus

Nous exploitons une chaîne de traitement classique pour la tâche d'AA (Figure 1). Cette chaîne est composée de deux principaux éléments : un extracteur de traits (des traits de même nature sont extraits des corpus d'entraînement et de test) et un classifieur (exploitant les traits extraits du corpus d'entraînement, chaque texte du corpus de test est alors classé).

Les expérimentations menées dans cet article soulignent les caractéristiques principales des méthodes d'AA fondées sur les chaînes de caractères (en opposition aux approches fondées sur les mots ainsi et l'exploitation de leurs étiquettes morphosyntaxiques). L'approche SVM est utilisée comme le classifieur dans cette chaîne de traitement, revendiquée comme approche la plus pertinente dans les travaux de Sun *et al.* (2012) et Brennan *et al.* (2012). L'étape de sélection des traits a pour but d'extraire les traits pertinents des corpus d'entraînement et de test sans *a priori* sur les langues traitées. Nous focalisons nos analyses sur l'influence de la sélection des traits en contexte multilingue. Pour ce faire, nous figeons les paramètres liés aux classifieurs afin de minimiser leurs influences sur l'interprétation des résultats liés aux changements des traits.

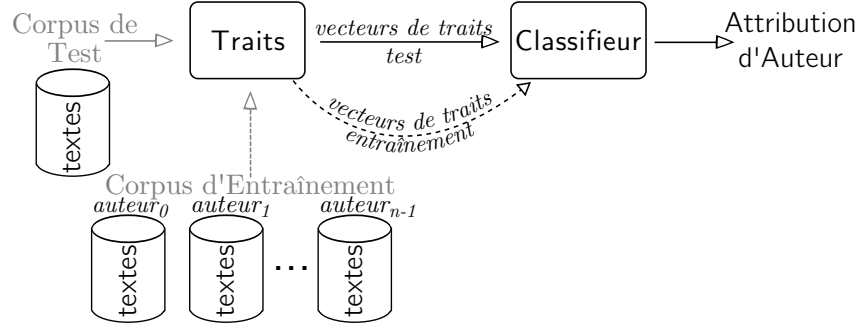


FIGURE 1: Chaîne de traitement utilisée pour l'Attribution d'Auteur.

3.1 Définitions

D est un ensemble de données pour l'analyse stylométrique constitué de I textes et de K auteurs. t_i est le $i^{\text{ème}}$ texte et a_k le $k^{\text{ème}}$ auteur. F désigne l'ensemble de traits calculables à partir de D et F_i l'ensemble des traits extraits du texte t_i . Chaque texte t_i est représenté sous la forme d'un vecteur de traits. Soit $o_{(i,j)}$ l'effectif du $j^{\text{ème}}$ trait dans le $i^{\text{ème}}$ texte t_i contenant n traits, $0 \leq j < n$. Nous représentons t_i sous la forme $\{o_{(i,0)}, o_{(i,1)}, \dots, o_{(i,n-1)}\}$. Une fonction de pondération w peut être appliquée sur chaque trait du texte t_i , $w(t_i) = \{w(f_0).o_{(i,0)}, w(f_1).o_{(i,1)}, \dots, w(f_{n-1}).o_{(i,n-1)}\}$. Un classifieur C est alors entraîné sur un sous-ensemble de textes écrits par des auteurs présélectionnés (corpus d'entraînement). L'ensemble des traits utilisés correspond à l'intersection des ensembles de traits du corpus de test et du corpus d'entraînement.

3.2 Corpus

Nous utilisons deux corpus différents, chacun constitué de textes écrits dans la même langue : un en anglais (corpus EBG), l'autre en français (corpus LIB). Ces deux langues ont été sélectionnées car elles partagent un alphabet et des origines en commun. Ceci rend la tâche plus difficile que lors du traitement de langues possédant plus de différences (anglais et chinois par exemple), les espaces de traits au grain caractère étant différents entre ces deux langues. Ainsi, une approche fondée sur SVM n'aurait aucune difficulté à séparer les textes analysés en deux sous-espaces contenant d'un coté les documents écrits en anglais, de l'autre les documents écrits en chinois (Figure 2).

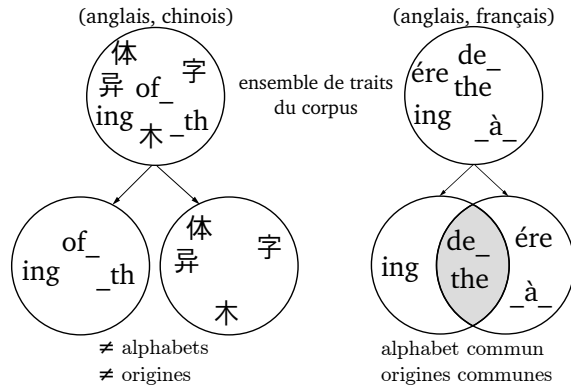


FIGURE 2: Influence de la différence de nature des traits sur l'apprentissage.

Un sous-corpus de textes écrits par 40 auteurs, EBG, a été sélectionné du EXTENDED BRENNAN GREENSTADT *adversarial corpus* (Brennan *et al.*, 2012). Ce corpus est exclusivement constitué de textes en anglais (Table 1). Les textes manipulés lors d'expériences menées par Brennan *et al.* (2012), plagats de styles ou de contenus, ont été exclus. La relation entre auteur et thème est tenue dans ce corpus, la plupart des auteurs ayant écrit des textes sur le même thème.

Le second corpus, LIB, est constitué d'articles de presse en provenance de la version en ligne du journal Libération, collectés pour les expériences décrites dans cet article. Il contient des textes écrits en français par 40 auteurs qui ont écrits dans plus d'une catégorie du journal (sport, santé, politique étrangère, *etc.*). En effet, les approches fondées sur des

	#caractères	#textes	#auteurs
corpus	$1.9 \cdot 10^6$	631	40
auteurs (moyenne \pm écart type)	$4.6 \cdot 10^4 \pm 8075$	15.8 ± 2.6	
textes (moyenne \pm écart type)	2945.1 ± 178.5		

TABLE 1: Caractéristiques du corpus EBG (anglais).

sous-chaînes de caractères sont souvent biaisées par le contenu thématique des textes analysés. Les auteurs qui écrivent exclusivement dans une seule de ces catégories ont donc été exclus afin d'éviter que le thème d'un article prime sur le style, ce qui rend ce corpus plus difficile à traiter. Les caractéristiques de ce corpus sont dressées Table 2.

	#caractères	#textes	#auteurs
corpus	$5.1 \cdot 10^6$	1247	40
auteurs (moyenne \pm écart type)	$1.3 \cdot 10^5 \pm 2.6 \cdot 10^4$	31.2 ± 4.2	
textes (moyenne \pm écart type)	4070.6 ± 1524.2		

TABLE 2: Caractéristiques du corpus LIB (français).

Le corpus LIB contient autant d'auteurs que le corpus EBG, mais le nombre de textes pour chaque auteur est plus important (31.2 ± 4.2 textes par auteur pour le corpus LIB, 15.8 ± 2.6 pour le corpus EBG). Chaque texte, dans ces deux corpus, contient plus de 250 mots (environ 1500 caractères), la longueur minimale nécessaire pour l'AA vue comme une tâche de classification (Forsyth & Holmes, 1996).

Le corpus MIXT est constitué à partir de la fusion des corpus EBG et LIB. Nous l'utilisons en vue d'éprouver le caractère multilingue des approches considérées. Des expériences sont aussi menées sur différents sous-corpus issus des corpus EBG, LIB et MIXT. Ainsi, EBG-10 (respectivement LIB-10 et MIXT-10) est un sous-ensemble de textes constitués de 10 auteurs du corpus EBG (LIB, MIXT). Aussi, les corpus MIXT-20, 40, 60 et 80 sont issus de la fusion des corpus LIB-10 + EBG-10, ... LIB-40 + EBG-40. Nous décrivons dans les sections suivantes les expérimentations menées sur ces corpus dans le but de souligner les différentes caractéristiques des traits utilisés et des éléments de la chaîne de traitement.

4 Définition des traits utilisés

Nous présentons dans cette section une alternative aux n -grammes de caractères pour les tâches d'AA. Les répétitions maximales (*maximal repeats* ou *motifs* dans les travaux de Ukkonen (2009)) sont calculées en se fondant sur les tableaux de suffixes (Kärkkäinen *et al.*, 2006). Les motifs représentent de manière condensée toutes les sous-chaînes d'un corpus. Pour la détection des chaînes *hapax* d'un corpus à partir de ses motifs, se référer aux travaux de Ilie & Smyth (2011).

4.1 Répétitions maximales

Les répétitions maximales (*motifs* dans les travaux de Ukkonen (2009)) sont des sous-chaînes de caractères avec les caractéristiques suivantes :

- répétition : les motifs apparaissent deux fois ou plus dans le corpus traité ;
- maximalité : étendre une occurrence d'un motif, ajouter à un motif le caractère se situant sur sa gauche ou sa droite, donne une chaîne de caractères avec un nombre d'occurrences moindre que le motif de base.

Les motifs se trouvant dans la chaîne $S = \text{HATTIVATTIA}$ sont T, A et ATTI. TT n'est pas maximal car il apparaît systématiquement à chaque occurrence de ATTI : son contexte droit est toujours I et son contexte gauche A. Les motifs d'un ensemble de chaînes peuvent être énumérés et leurs occurrences localisées, en utilisant un tableau de suffixes augmenté (*augmented suffix array* (Kärkkäinen *et al.*, 2006)). De par leurs caractéristiques de maximalité, ces motifs représentent toutes les sous-chaînes de caractères répétées d'un ensemble de chaînes de caractères de manière condensée.

Soit deux chaînes $S_0 = \text{HATTIV}$ et $S_1 = \text{ATTIAA}$, la Table 3 représente le tableau de suffixes augmenté calculé sur la concaténation de S_0 et S_1 , $S = S_0.\$1.S_1.\0 , avec $\$0$ et $\$1$ deux caractères lexicographiquement plus petits que ceux de l'alphabet Σ décrivant S_0 et S_1 et $\$0 < \1 . Le tableau de suffixes augmenté est composé du tableau de suffixes SA (*suffix array*) contenant les suffixes de S triés par ordre lexicographique, ainsi que de la table des plus longs préfixes communs LCP (*longest common prefix*) contenant la taille du préfixe commun entre les éléments de SA contigus deux à deux. Soit n le nombre de caractères de S , $S[i]$ est alors le $i^{\text{ème}}$ caractère de S , $S[k, l]$ est une sous-chaîne de S allant du $k^{\text{ème}}$

au $l^{\text{ème}}$ caractère, et $lpc(str_1, str_2)$ est le plus long préfixe commun entre deux chaînes str_1 et str_2 .

$$\begin{aligned} LCP_i &= lpc(\mathcal{S}[SA_i, n-1], \mathcal{S}[SA_{i+1}, n-1]) \\ LCP_{n-1} &= 0 \end{aligned}$$

La table LCP permet la détection de toutes les répétitions au sein d'un ensemble de textes. Le critère de maximalité n'est pas ici validé car le calcul des LCP permet seulement de vérifier la maximalité à gauche sur les préfixes répétés dans SA .

i	LCP_i	SA_i	$\mathcal{S}[SA_i] \dots \mathcal{S}[n]$
0	0	13	$\$0$
1	0	6	$\$1ATTIAA\0
2	1	12	$A\$0$
3	1	11	$AA\$0$
4	4	7	$ATTIAA\$0$
5	0	1	$ATTIV\$1ATTIAA\0
6	0	0	$HATTIV\$1ATTIAA\0
7	1	10	$IAA\$0$
8	0	4	$IV\$1ATTIAA\0
9	2	9	$TIAA\$0$
10	1	3	$TIV\$1ATTIAA\0
11	3	8	$TTIAA\$0$
12	0	2	$TTIV\$1ATTIAA\0
13	0	5	$V\$1ATTIAA\0

TABLE 3: Tableau des suffixes augmenté (SA et LCP) de $\mathcal{S} = \text{HATTIV\$1ATTIAA\$0}$.

Par exemple, la sous-chaîne $ATTI$ est présente dans \mathcal{S} aux offsets (1, 7) (voir LCP_4 dans la Table 3). Le processus d'énumération de tous les motifs s'effectue en parcourant la table des LCP . La détection de ses motifs est déclenchée en fonction de la différence de LCP entre un suffixe et le suivant en fonction de l'ordre établi sur SA .

TTI est équivalent à $ATTI$ car pour ces deux chaînes, leur dernier caractère se situe aux indices (4, 10). Ces deux chaînes sont en relation d'équivalence d'occurrences (*occurrence-equivalence*, (Ukkonen, 2009)). Pour cet exemple, $ATTI$ est considéré comme le motif *maximal* parce que cette chaîne est la plus grande de toutes celles qui sont en équivalence d'occurrences avec elle. Les autres motifs maximaux trouvés sont A et T car leur contexte gauche et leur contexte droit ne sont pas systématiquement les mêmes pour chacune de leurs occurrences. Toutes les occurrences de chaque motif, représentables par le couple ($id_{chaîne}, indice$), sont données en faisant l'équivalence entre les indices de \mathcal{S} et ceux de \mathcal{S}_0 et \mathcal{S}_1 . De cette manière, les motifs de \mathcal{S} peuvent être localisés dans chaque chaîne \mathcal{S}_i . Les tables SA et LCP sont construites en temps linéaire $O(n)$ (Kärkkäinen *et al.*, 2006), l'énumération de chacun des motifs est donnée en $O(k)$, avec k le nombre de motifs différents et $k < n$ (Ukkonen, 2009).

4.2 Répétitions maximales d'ordre n

Soit \mathcal{R} l'ensemble des répétitions maximales (ou *motifs*) détectées sur n chaînes de caractères $\mathcal{S} = \{\mathcal{S}_0, \dots, \mathcal{S}_{n-1}\}$, avec $|\mathcal{S}| = \sum_{i=1}^n size(\mathcal{S}_i)$. L'ensemble de motifs \mathcal{R} est calculé sur la concaténation de chaque chaîne $\mathcal{S}_i : c(\mathcal{S}) = \mathcal{S}_0\$_{n-1} \dots \mathcal{S}_{n-1}\0 . Les répétitions maximales du deuxième ordre \mathcal{R}^2 dans \mathcal{S} sont calculées sur la concaténation de l'ensemble des m motifs de \mathcal{R} , $c(\mathcal{R}) = \mathcal{R}_0\$_{m-1} \dots \mathcal{R}_{m-1}\0 avec $m < |\mathcal{S}|$, chaque \mathcal{R}_i étant un motif de \mathcal{S} . L'ensemble des motifs du $n^{\text{ème}}$ ordre est noté \mathcal{R}^n . Par exemple, soit $c(\mathcal{S}) = \text{HATTIV\$1ATTIAA\$0}$. L'ensemble de motifs \mathcal{R} sur $c(\mathcal{S})$ est constitué des motifs suivants : $\mathcal{R} = \{\text{ATTI}, A, T\}$. L'ensemble des répétitions du deuxième ordre \mathcal{R}^2 est composé des motifs T (deux fois dans $ATTI$ et une fois dans T) et A (une fois dans $ATTI$ et une fois dans A).

L'ensemble des motifs dans \mathcal{R}^n est un sous-ensemble de \mathcal{R}^{n-1} .

REDUCTIO AD ABSURDUM — Supposons que $\mathcal{R}^n \not\subset \mathcal{R}^{n-1}$. En d'autres termes, $\exists m$ un motif avec $m \in \mathcal{R}^n$ et $m \notin \mathcal{R}^{n-1}$. m a été extrait à partir de l'ensemble \mathcal{R}^{n-1} , donc m apparaît deux fois ou plus suivant deux configurations. m est un motif apparaissant (CAS 1) dans deux motifs différents et/ou apparaissant (CAS 2) deux fois (ou plus) dans un seul motif de \mathcal{R}^n . Les CAS 1 & 2 sont équivalents : m est un motif car il apparaît deux fois et est maximal dans $c(\mathcal{R}^{n-1})$ la concaténation de chaque élément de \mathcal{R}_i^{n-1} . Parce que m est maximal, ses contextes gauches (notés a et b) et droits (c et d) sont différents dans l'ensemble de ses occurrences, avec $a \neq b$, $c \neq d$ et a, b, c et d étant des caractères présents dans $c(\mathcal{R}^{n-1})$, dont les séparateurs $\$$ ou le caractère *vide* ϵ si m possède une occurrence au début de $c(\mathcal{R}^{n-1})$. \mathcal{R}^n a été calculé sur $c(\mathcal{R}^{n-1}) = \dots amc \dots bmd \dots$, donc $m \in \mathcal{R}^{n-1}$ — contradiction.

La Figure 3 représente le nombre de motifs différents en fonction de l'ordre des répétitions maximales. Parce que $\mathcal{R}^n \subset \mathcal{R}^{n-1} \iff |\mathcal{R}^n| < |\mathcal{R}^{n-1}|$, le nombre de motifs décroît plus l'ordre est important quelque soit le corpus. Le nombre de

motifs tombe à 0 pour $n = 26$ (EBG-40, LIB-40 et MIXT-80) et $n = 25$ (MIXT-40).

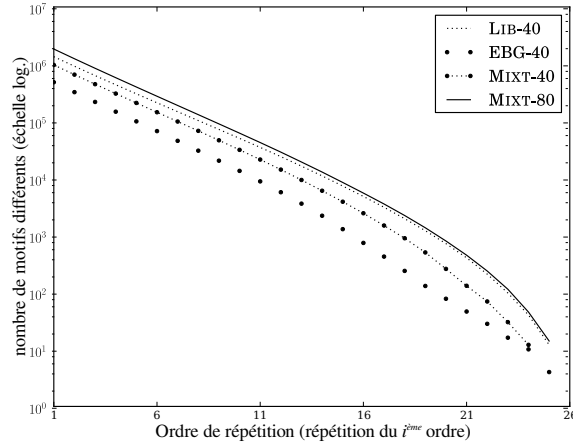


FIGURE 3: Évolution du nombre de motifs (échelle logarithmique) en fonction de l'ordre des répétitions maximales (LIB-40, EBG-40, MIXT-40 et MIXT-80).

Le calcul des répétitions maximales du deuxième ordre s'effectue avec le même algorithme que celui permettant le calcul des répétitions maximales, la complexité en temps pour l'énumération de ces motifs est donc faite en $O(n)$ car la taille des ensembles de motifs décroît plus l'ordre est important. Le calcul des répétitions maximales du deuxième ordre est utilisé pour détecter les motifs inclus dans d'autres motifs.

4.3 Exploitation des différences entre n -grammes de caractères et répétitions maximales

Cette section décrit les principales différences entre les n -grammes de caractères et les répétitions maximales, et comment exploiter cette différence sur les représentations de textes fondées sur les approches vectorielles. Comme décrit précédemment, les répétitions maximales sont une manière condensée de représenter toutes les sous-chaînes de caractères d'un corpus. En d'autres termes, pour une valeur donnée n , l'ensemble des répétitions maximales de taille n est un sous-ensemble des n -grammes de caractères d'un corpus (et de la même manière dans le cas de chaînes de caractères de longueurs variables : les répétitions maximales ayant une longueur comprise entre $[min, max]$ et les $[min, max]$ -grammes de caractères). Les sous-chaînes qui ne sont pas des répétitions maximales sont celles qui sont seulement maximales à gauche ou à droite (ou ni l'un ni l'autre, donc répétées mais non-maximales) ou des *hapax legomena*. Dans une tâche de classification supervisée, les *hapax legomena* du corpus complet n'ont alors pas d'impact sur les résultats car par définition, ces *hapax* apparaissent seulement une fois dans le corpus de test ou une fois dans le corpus d'entraînement.

Si les n -grammes de caractères peuvent capturer différentes caractéristiques sous-jacentes en fonction du choix du paramètre n (caractéristiques lexicales, contextuelles ou thématiques (Sun *et al.*, 2012)), ces n -grammes « capturent » des traits représentés par des sous-chaînes de taille supérieure à n . Par exemple, considérons *abcdef* un motif extrait d'un corpus et que ses caractères ne sont pas inclus dans d'autres sous-chaînes du corpus, parce que *abcdef* est maximale. Alors, chaque sous-chaîne de *abcdef* possède le même nombre d'occurrences que *abcdef* ($freq(abcdef) = k$). La Figure 4 représente comment l'usage de 3-grammes de caractères est affecté par le nombre d'occurrences du motif, et donc comment la représentation vectorielle du corpus contenant ce motif est elle aussi affectée.

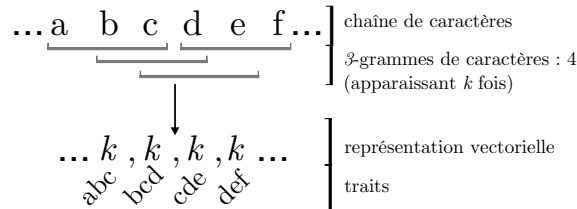


FIGURE 4: Sous-chaînes d'un motif et leur influence sur la représentation vectorielle du corpus.

Ainsi, exploiter seulement les répétitions maximales de taille 3 ne permettra pas dans cet exemple d'exploiter des sous-

chaînes ayant le même nombre d’occurrences que le motif `abcdef`. Seulement considérer certaines longueurs affectera la représentation vectorielle fondées sur les occurrences des chaînes, et inversement (interdépendance longueur-effectif telle que décrit dans les travaux de Zipf (1949)). Pour prendre en compte ces influences, nous définissons une fonction de pondération $w_{2nd}(trait)$ qui exploite les sous-chaînes qu’un trait encapsule. $w_{2nd}(trait) = pot(trait) - sub(trait)$, où $pot(trait)$ correspond au nombre de sous-chaînes potentielles à l’intérieur d’un trait et $sub(trait)$ correspond au nombre de motifs de deuxième ordre qui apparaissent à l’intérieur du trait et ailleurs dans le corpus :

- cette pondération est donc fonction de la longueur du trait ;
- pour deux traits de même longueur, le facteur de pondération peut être différent.

Si un trait varie d’un caractère par rapport à un autre, cette fonction de pondération minimisera cet ajout : les produits du facteur de pondération et de l’effectif des motifs utilisés comme traits seront proches. À l’inverse, un trait qui est plus qu’une légère variation d’un autre trait sera considéré comme « consistant » et donc aura plus d’importance. En d’autres termes, là où les n -grammes pondèrent naturellement les traits grâce à la redondance, cette fonction de pondération exploite la redondance au sein des traits.

Chaque \mathcal{R}_i est une répétition maximale utilisée comme trait et \mathcal{S} est l’ensemble des textes des auteurs analysés. Avec $\mathcal{S} = \{S_0, \dots, S_{n-1}\}$, \mathcal{R} l’ensemble des répétitions maximales calculées à partir de \mathcal{S} et \mathcal{R}^2 l’ensemble des répétitions maximales calculées à partir de \mathcal{R} , chaque répétition maximale dans \mathcal{R} est pondérée à partir de l’ensemble des répétitions maximales de \mathcal{R}^2 . Le nombre de sous-chaînes différentes d’un trait de taille n est donné par $pot(trait) = \frac{n(n+1)}{2}$ (c’est-à-dire le nombre triangulaire de taille n , la chaîne totale étant considérée comme une sous-chaîne potentielle). $sub(trait)$ consiste à calculer toutes les occurrences de \mathcal{R}^2 apparaissant à l’intérieur du trait (en excluant le trait lui-même). Si toutes les sous-chaînes potentielles d’un trait sont aussi des traits du corpus, alors $w_{2nd}(trait) = 1$. Dans les expériences de cet article, cette fonction de pondération est comparée avec celle prenant en compte seulement la longueur des traits $w_{len}(trait) = \frac{n(n+1)}{2}$ (avec n le nombre de caractères du trait).

5 Expériences

Cette section décrit les performances des approches proposées. Deux ensembles de traits de longueur variable sont envisagés : les n -grammes de caractères et les répétitions maximales (motifs). Les motifs sont considérés de trois façons différentes : pondérés par leur longueur (w_{len}), par les répétitions maximales du 2^{ème} ordre (w_{2nd}) et sans pondération.

Une validation croisée, *stratified 10-fold cross validation*, est utilisée pour valider les performances du système pour chaque trait. Les corpus sont échantillonnés en 10 sous-ensembles de taille égale, chaque échantillon contient la même proportion d’auteurs. Pour mesurer la performance des systèmes, le score d’attribution est calculé de la manière suivante : le nombre de textes correctement classés divisé par le nombre total de textes classés puis ramené à un pourcentage. SVM est utilisé avec un noyau linéaire, paramètre adapté quand le nombre de dimensions est beaucoup plus élevé que le nombre d’éléments à classer. Le paramètre de régularisation est fixé à $C = 1$ quelque soit le trait utilisé.

5.1 Impact de la longueur des sous-chaînes et des motifs

Le score d’AA est calculé sur trois corpus : EBG-40 (Figure 5), LIB-40 (Figure 6) et MIXT-80 (Figure 7). Chaque figure est constituée de quatre matrices pour chaque trait : les répétitions maximales (*motifs*), les n -grammes, les répétitions maximales pondérées par leur longueur ($motifs_{len}$) et les répétitions maximales pondérées par les répétitions maximales du 2^{ème} ordre ($motifs_{2nd}$). Le score écrit aux coordonnées (i, j) de chaque matrice est donné par l’exploitation de traits de longueur comprise entre i et j .

Les traits peuvent être ordonnés en fonction de leur performance sur les corpus : $motifs \leq motifs_{len} < n\text{-grammes} < motifs_{2nd}$. Le fait que $motifs < n\text{-grammes}$ montre l’effet positif de la redondance des traits. Les scores sur les diagonales des matrices (là où les traits sont de taille fixe) utilisant $motifs$ et $motifs_{len}$ sont identiques car chaque trait est pondéré par le même facteur avec la fonction de pondération w_{len} . Les scores d’attributions sur le corpus EBG sont élevés. Cela s’explique par le lien fort entre auteur et contenu thématique (pour un auteur, chacun de ses textes appartient à la même thématique comme économie, arts ou sport). La tâche est plus difficile sur le corpus LIB car contrairement au corpus EBG, chaque auteur a été sélectionné si son ensemble de textes est constitué de textes de différents thèmes.

Le score d’attribution sur les trois corpus a aussi été calculé en utilisant $motifs_{2nd}$ sans contrainte (tous les motifs sont considérés quelque soit leur longueur) avec les scores suivants : 66,40% sur le corpus EBG-40, 48,20% sur le corpus LIB-40 et 54,21% sur le corpus MIXT-80. Ceci souligne la nécessité de la présélection des traits parmi ceux disponibles.

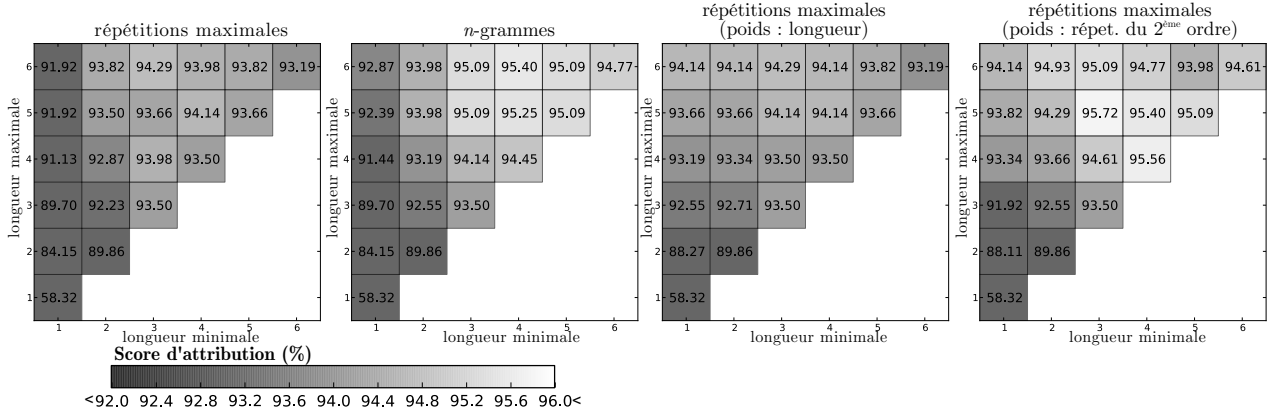


FIGURE 5: Score d'attribution sur le corpus EBG-40.

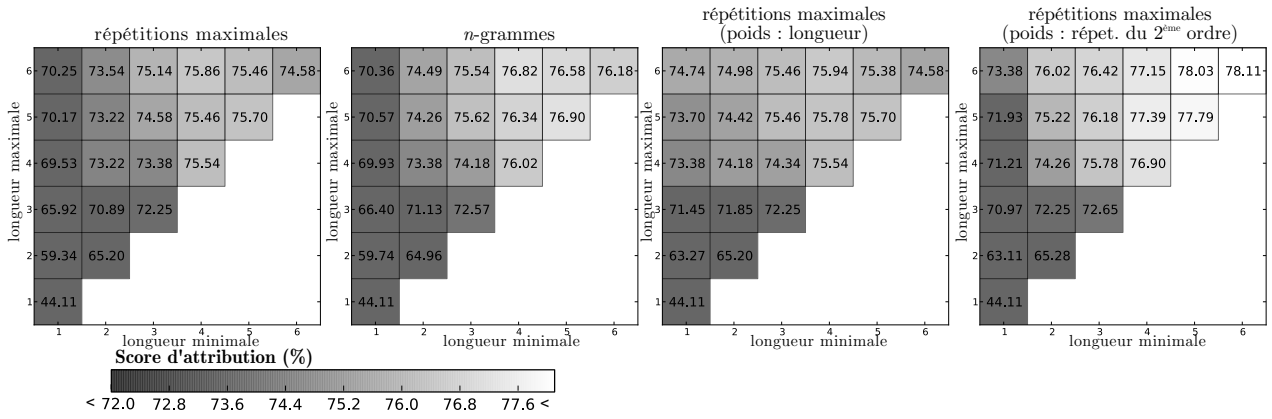


FIGURE 6: Score d'attribution sur le corpus LIB-40.

Les meilleurs paramètres de longueur sont sélectionnés en calculant la moyenne des scores d'attribution sur chacune des matrices pour chaque intervalle de longueurs $[min, max]$ (Table 4).

	meilleurs paramètres de longueur $[min, max]$	score moyen
n -grammes	[4, 6]	84,61%
$motifs$	[4, 6]	83,69%
$motifs_{len}$	[4, 6]	83,88%
$motifs_{2nd}$	[4, 5]	85,39%

TABLE 4: Meilleurs paramètres en fonction du score moyen sur les corpus LIB-40, EBG-40 et MIXT-80.

$motifs_{2nd}$ obtient les meilleurs résultats en utilisant le plus petit intervalle de longueurs. Les meilleurs paramètres de longueur calculés sur l'ensemble corpus ne constituent pas nécessairement le meilleur jeu de paramètres pour chaque corpus pris individuellement. Par exemple, $motifs_{2nd}$ obtient de meilleurs résultats avec les paramètres [6, 6] sur le corpus LIB-40 qu'avec les paramètres [4, 5]. Parce que les motifs sont une représentation condensée des n -grammes, l'espace de traits est naturellement moindre en utilisant des motifs. Les expériences montrent de meilleurs résultats avec l'utilisation de traits de longueur variable plutôt que fixe. Cependant, utiliser le plus grand intervalle de longueur lors de la sélection des traits n'est pas systématiquement un choix pertinent au regard des résultats. Par exemple, une différence de 4,01% est observable entre l'intervalle [1, 6] et l'intervalle optimale [4, 5] sur les résultats du corpus LIB-40 en utilisant $motifs_{2nd}$ (Figure 6). Considérer le plus grand ensemble de traits disponibles permet peut être de capturer des caractéristiques utiles à cette tâche, mais surtout des traits dégradant sensiblement les résultats.

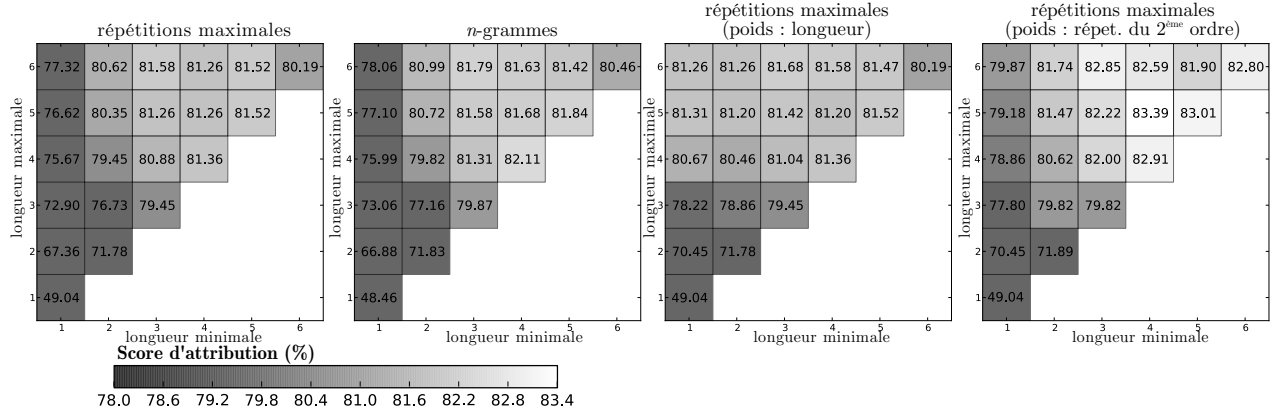


FIGURE 7: Score d'attribution sur le corpus MIXT-80.

5.2 Évolution de la qualité d'attribution en fonction des nombres de traits et d'auteurs

En choisissant les meilleurs paramètres pour chaque type de trait (Table 4), les expériences suivantes décrivent l'évolution du score d'attribution en fonction du nombre d'auteurs (Figure 8).

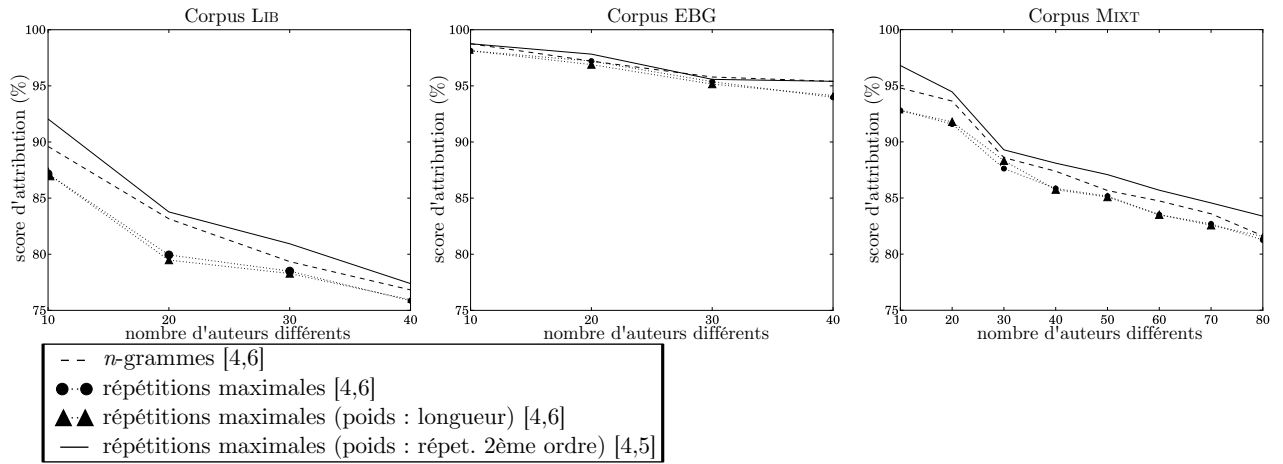


FIGURE 8: Évolution du score d'attribution en fonction du nombre d'auteurs.

Pour chaque corpus et chaque trait, le score d'attribution décroît quand le nombre d'auteurs augmente. Les résultats sont supérieurs en utilisant $motifs_{2nd}$, à l'exception des corpus EBG-30 et EBG-40. Les pires résultats sont obtenus sur le corpus LIB pour lequel le score décroît de 92,04% à 77,39% (de 89,60% à 76,82% en utilisant des n -grammes). Pondérer les motifs en fonction de la longueur ($motifs_{len}$) n'améliore pas le score de manière significative par rapport à l'utilisation des motifs sans pondération. Les évolutions des nombres de traits sont données sur la Figure 9. Le nombre de traits correspond à la moyenne des tailles des vecteurs représentant les textes sur chacun des dix échantillons de la validation croisée. Les résultats sont différents de ceux de la Figure 3 (Section 4.2). En effet nous montrons ici le nombre de traits utilisés pour la classification et non tous ceux qui sont calculables sur les corpus.

Utiliser des motifs de taille [4, 5] réduit significativement le nombre de traits par rapport à l'usage de n -grammes de taille [4, 6] et ou de motifs sans contrainte de longueur. Le nombre de motifs augmente linéairement en fonction du nombre d'auteurs (ou en fonction de la taille du corpus). Le nombre de n -grammes de taille [4, 6] est plus élevé que le nombre de motifs pour un faible nombre d'auteurs. Il est toutefois plus faible dès lors que le nombre d'auteurs augmente, du fait de la distribution sous-linéaire des n -grammes de taille [4, 6]. La taille de l'espace de recherche des motifs de taille [4, 5] est adaptée quand la taille des données croît.

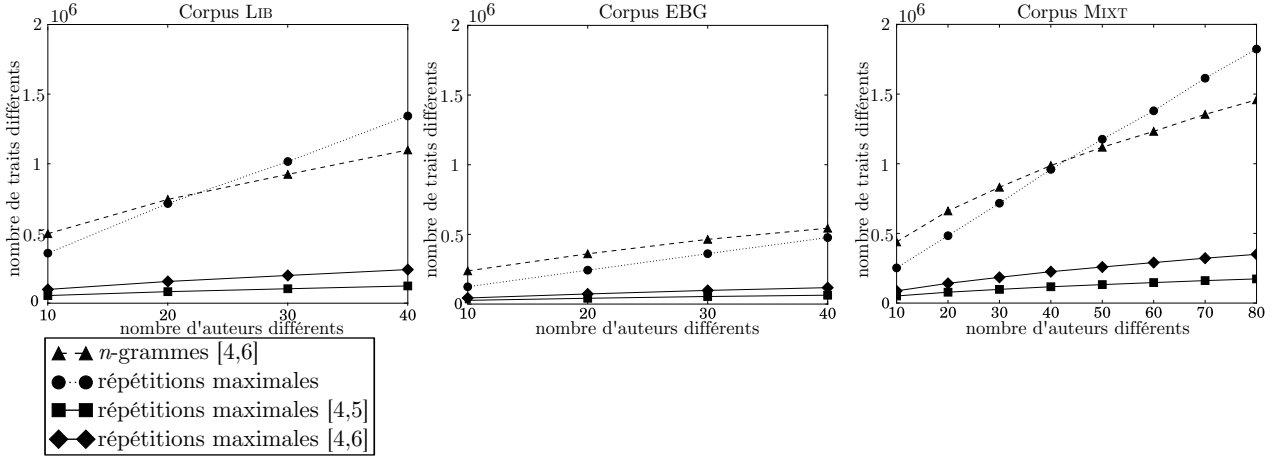


FIGURE 9: Évolution du nombre de traits en fonction du nombre d'auteurs.

5.3 Évaluation monolingue à partir de corpus multilingues

Le corpus MIXT est composé des corpus LIB en français et EBG en anglais, les deux langues partageant des traits en commun de par leurs origines communes. Dans le cadre d'une analyse multilingue, l'utilisation de deux langues proches est adaptée en AA. Cette expérience permet de vérifier si le trait choisi est efficace dans un corpus où chaque texte n'est pas écrit dans la même langue mais où les langues partagent des traits en commun. La Table 5 présente les scores d'attribution sur les deux corpus monolingues, LIB et EBG, pris indépendamment puis intriqués au sein du même corpus MIXT. Le but est d'analyser comment les traits influent sur le traitement lorsque plusieurs langues sont traitées en même temps.

<i>n</i> -grammes (longueur [4, 6])					<i>motifs</i> _{2nd} (longueur [4, 5])				
nb. d'auteurs	EBG	EBG issu de MIXT	LIB	LIB issu de MIXT	nb. d'auteurs	EBG	EBG issu de MIXT	LIB	LIB issu de MIXT
10	98,75%	98,75%	89,60%	91,13%	10	98,75%	98,75%	92,01%	92,35%
20	97,20%	96,89%	83,15%	82,69%	20	97,83%	97,52%	83,77%	83,46%
30	95,79%	94,85%	79,34%	78,65%	30	95,59%	96,84%	80,93%	80,08%
40	95,40%	94,10%	76,82%	75,03%	40	95,40%	95,09%	77,38%	77,47%

TABLE 5: Score d'attribution sur les corpus LIB et EBG indépendamment ou issus du traitement du corpus MIXT.

Les résultats sont proches en utilisant le corpus multilingue ou les corpus monolingues de manière indépendante. Quelques améliorations peuvent être notées en utilisant *motifs*_{2nd}, où les résultats sont plus souvent meilleurs quand les deux corpus EBG et LIB sont traités ensemble. En utilisant des *n*-grammes de taille variable sur les corpus multilingue et monolingue, la différence de résultats augmente avec le nombre d'auteurs : la différence de score d'attribution est de -1,30% sur le corpus LIB et de -1,79% sur le corpus EBG. À l'inverse, l'approche fondée sur les motifs est plus adaptée à la problématique multilingue (-0,31% sur le corpus LIB et +0,09% sur le corpus EBG).

6 Conclusion

Nous avons proposé une alternative efficiente aux approches fondées sur les *n*-grammes de tailles variables via les *répétitions maximales*. Ces répétitions surpassent les approches classiques fondées sur les sous-chaînes sur deux aspects. Premièrement, les *répétitions maximales* sont, par essence et à la différence des *n*-grammes, non-redondantes. En effet, leur caractère maximal évite la détection et l'utilisation de nombreuses occurrences de sous-chaînes équivalentes dans le corpus. Cela réduit considérablement le nombre de traits donc l'espace de recherche et nous préconisons leur usage conjointement à des méthodes de sélection de sous-espaces de recherche (Algorithme Génétique, Recuit Simulé, sélection de Corrélation Caractéristiques, Gain d'Information). Mais nous avons aussi souligné dans cet article que cette redondance avait un effet positif dans cette tâche de classification. Cette redondance a été exploitée pour proposer une méthode de pondération des traits en fonction de leur structuration interne. Deuxièmement, avec les répétitions maximales de deuxième ordre, nous réduisons davantage l'espace de recherche des traits et nous proposons une nouvelle façon d'améliorer la précision de la prédiction en AA. L'hypothèse qu'une longue chaîne répétée est plus importante si elle ne contient pas

trop de sous-répétitions, est validée. Nos premières expériences montrent que l’usage des répétitions maximales permet de traiter plus aisément des flux de textes dans des langues différentes. La dégradation des résultats est plus minime en contexte multilingue plus le flux de données à traiter grandi comparée aux approches fondées sur les n -grammes.

Références

- ABBASI A. & CHEN H. (2008). Writeprints : A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, **26**(2), 7.
- BENDER E. M. (2009). Linguistically naïve != language independent : Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction Between Linguistics and Computational Linguistics : Virtuous, Vicious or Vacuous ?*, ILCL '09, p. 26–32 : ACL.
- BRENNAN M., AFROZ S. & GREENSTADT R. (2012). Adversarial stylometry : Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, **15**(3), 12.
- BURROWS S., UITDENBOGERD A. & TURPIN A. (2014). Comparing techniques for authorship attribution of source code. *Software – Practice and Experience*, **44**(1), 1–32.
- CHASKI C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, **8**, 1–65.
- EL BOUANANI S. E. M. & KASSOU I. (2014). Authorship analysis studies : A survey. *International Journal of Computer Applications*, **86**, 22–29.
- FORSYTH R. S. & HOLMES D. I. (1996). Feature-finding for text classification. *Literary and Linguistic Computing*, **11**(4), 163–174.
- GRIEVE J. (2007). Quantitative authorship attribution : An evaluation of techniques. *Literary and linguistic computing*, **22**(3), 251–270.
- ILIE L. & SMYTH W. F. (2011). Minimum unique substrings and maximum repeats. *Fundamenta Informaticae*, **110**(1), 183–195.
- KACMARCIK G. & GAMON M. (2006). Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, p. 444–451 : Association for Computational Linguistics.
- KÄRKKÄINEN J., SANDERS P. & BURKHARDT S. (2006). Linear work suffix array construction. *Journal of the ACM*, **53**(6), 918–936.
- KOPPEL M., SCHLER J. & ARGAMON S. (2009). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, **60**(1), 9–26.
- KOPPEL M., SCHLER J. & ARGAMON S. (2011). Authorship attribution in the wild. *Language Resources and Evaluation*, **45**(1), 83–94.
- PENG F., SCHUURMANS D., WANG S. & KESELJ V. (2003). Language independent authorship attribution using character level language models. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, p. 267–274 : Association for Computational Linguistics.
- STAMATATOS E. (2006). Ensemble-based author identification using character n -grams. In *Proceedings of the 3rd International Workshop on Text-based Information Retrieval*, p. 41–46.
- STAMATATOS E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, **60**(3), 538–556.
- SUN J., YANG Z., LIU S. & WANG P. (2012). Applying stylometric analysis techniques to counter anonymity in cyberspace. *Journal of Networks*, **7**(2).
- TWEEDIE F. J., SINGH S. & HOLMES D. I. (1996). Neural network applications in stylometry : The Federalist papers. *Computers and the Humanities*, **30**(1), 1–10.
- UKKONEN E. (2009). Maximal and minimal representations of gapped and non-gapped motifs of a string. *Theoretical Computer Science*, **410**(43), 4341–4349.
- ZIPF G. K. (1949). *Human Behaviour and the Principle of Least-Effort : an Introduction to Human Ecology*. Addison-Wesley.

Mesurer la similarité entre phrases grâce à Wikipédia en utilisant une indexation aléatoire

Hai-Hieu Vu¹ Jeanne Villaneau¹ Farida Saïd² Pierre-François Marteau¹

(1) Université de Bretagne Sud, laboratoire IRISA

(2) Université de Bretagne Sud, laboratoire LMBA

hai-hieu.vu, jeanne.villaneau, farida.said, pierre-francois.marteau@univ-ubs.fr

Résumé. Cet article présente une méthode pour mesurer la similarité sémantique entre phrases qui utilise Wikipédia comme unique ressource linguistique et qui est, de ce fait, utilisable pour un grand nombre de langues. Basée sur une représentation vectorielle, elle utilise une indexation aléatoire pour réduire la dimension des espaces manipulés. En outre, elle inclut une technique de calcul des vecteurs de termes qui corrige les défauts engendrés par l'utilisation d'un corpus aussi général que Wikipédia. Le système a été évalué sur les données de SemEval 2014 en anglais avec des résultats très encourageants, au-dessus du niveau moyen des systèmes en compétition. Il a également été testé sur un ensemble de paires de phrases en français, à partir de ressources que nous avons construites et qui seront mises à la libre disposition de la communauté scientifique.

Abstract.

Semantic similarity between sentences based on Wikipedia and Random Indexing.

This paper proposes a semantic similarity measure for sentence comparison based on the exploitation of Wikipedia as the only language resource. Such similarity measure is therefore usable for a wide range of languages, basically those covered by Wikipedia. Random Indexing is used to cope with the great dimensionality and the sparseness of the data vectorial representations. Furthermore, a statistical weight function is used to reduce the noise generated by the use of a multi domain corpus such as Wikipedia. This semantic similarity measure has been evaluated on SemEval 2014 dataset for English language leading to very promising results, basically above the average level of the competing systems that exploit Wikipédia in conjunction with other sources of semantic information. It has been also evaluated on a set of pairs of sentences in French that we have build specifically for the task, and made freely available for the research community.

Mots-clés : Similarité sémantique, Indexation aléatoire, Wikipédia, Relation sémantique.

Keywords: Semantic Textual Similarity, Random indexing, Wikipédia, Semantic Relatedness.

1 Introduction

Mesurer la similarité entre deux phrases (ou textes courts) consiste à évaluer jusqu'à quel point le sens de ces phrases est proche. Cette tâche (STS : Semantic Textual Similarity) est souvent utilisée dans plusieurs domaines importants du Traitement Automatique des Langues (TAL), parmi lesquels on peut citer la recherche d'informations (Balasubramanian *et al.*, 2007), la catégorisation de textes (Ko *et al.*, 2002), le résumé de texte (Erkan & Radev, 2004), la traduction automatique, etc. Longtemps considérée comme une sous-tâche dans les domaines cités, la STS fait depuis quelques années l'objet d'un intérêt croissant. Depuis 2012, la tâche STS de SemEval confronte les résultats de différents systèmes, presque tous consacrés à la langue anglaise. La version 2014 de Semeval a cependant proposé une évaluation des systèmes sur des phrases en espagnol, à laquelle 9 équipes ont participé (Agirre *et al.*, 2014).

La similarité lexicale constitue une première approche pour mesurer la similarité entre deux textes (Hirao *et al.*, 2005; Lin, 2004). Cependant, elle ne tient compte, ni des relations sémantiques entre les mots ou groupes de mots d'un même texte, ni de la similarité sémantique entre les mots des deux textes (synonymie, paraphrase, etc.). Pour pallier ce manque et suivant le principe selon lequel les mots qui apparaissent dans un même contexte ont potentiellement une similarité sémantique importante, les systèmes récents se fondent sur des études statistiques de gros corpus de la langue qui permettent de prendre en compte ces contextes. Les meilleurs systèmes de la tâche STS de SemEval2014 utilisent des ressources

linguistiques qui ne sont disponibles que pour la langue anglaise en y incluant, outre des corpus de très grande taille, des corpus de paraphrases, le WordNet, etc. (Kashyap *et al.*, 2014; Sultan *et al.*, 2014). Il est également intéressant de constater que les systèmes qui sont arrivés en tête dans le challenge en langue espagnole de SemEval ont utilisé un système réalisé pour l'Anglais, en transformant les phrases données en espagnol en leur équivalent anglais (Chavez *et al.*, 2014; Kashyap *et al.*, 2014).

Pour les langues moins bien dotées en ressources linguistiques que ne l'est la langue anglaise, Wikipédia représente un corpus très intéressant en raison de sa taille croissante et de son caractère encyclopédique qui assure une couverture très générale de presque tous les domaines. Wikipédia représente donc une énorme ressource multilingue pour le traitement automatique de la langue naturelle (TAL), qui est exploitée de différentes façons, et en particulier pour définir des relations sémantiques entre termes et entre textes (cf. section 2).

Le système présenté dans cet article (WikiRI) repose sur un modèle vectoriel, ou Vector Space Models (VSM). Le principe consiste à construire un espace vectoriel de grande dimension, dans lequel un mot est représenté par un vecteur unique qui rend compte de ses contextes d'occurrence. Plus précisément, le modèle utilisé est celui des GVSM (Generalized Vector Space Model), où les documents sont utilisés comme base de l'espace. Les termes y sont représentés comme des vecteurs dans la base des concepts définis à partir des articles de Wikipédia. Pour remédier aux problèmes posés par le nombre d'articles présents dans Wikipédia et sa constante augmentation, nous proposons une représentation vectorielle de la sémantique des termes qui utilise le Random Indexing (RI) (cf. section 3). Par ailleurs, WikiRI introduit des modifications dans les calculs des vecteurs de termes pour corriger le bruit engendré par l'utilisation d'une ressource linguistique aussi encyclopédique que Wikipédia : elles sont détaillées dans la section 4.

Nous avons effectué les expérimentations et les évaluations sur des ensembles de données en français (Sensim-french¹ que nous avons construites et sur les données de SemEval 2014 pour l'anglais (SemEval-2014 Task 10²). Elles indiquent des résultats intéressants qui sont décrits dans la section 5.

2 Wikipédia en tant que ressource linguistique

Actuellement disponible dans 288 langues, Wikipédia est le plus grand référentiel de connaissances générales sur le Web. Les statistiques officielles de Wikipédia en date du 12/12/2014 font état d'un nombre d'articles en langue anglaise de 4 668 468 et de 1 569 491 articles pour la langue française.

- **Structure du réseau** : Si l'on ne tient pas compte de la direction des liens entre articles, le graphe de Wikipédia est presque entièrement connecté : 98.5% des articles sont liés les uns aux autres. En tenant compte de la direction des liens, on retrouve la structure en noeud papillon du Web : des composantes denses fortement connectées sont liées entre elles par des liens unidirectionnels. La zone centrale (SCC) - pour strongly connected component - est composée

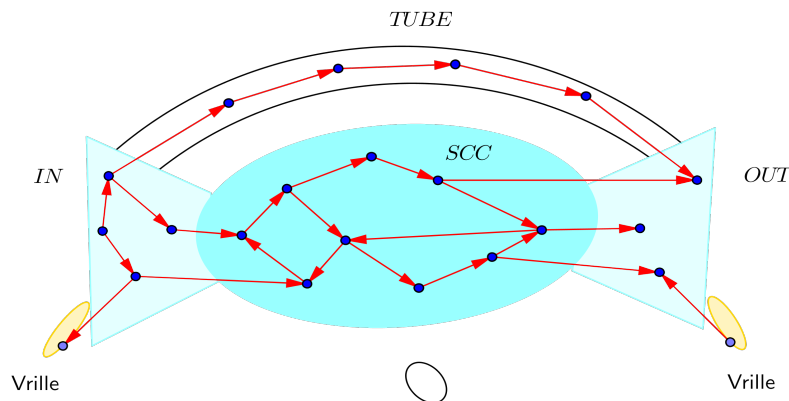


FIGURE 1 – Structure en noeud-papillon de Wikipédia

d'articles fortement liés entre eux : deux articles quelconques de cette zone peuvent toujours être liés par un chemin

1. <http://vuhaihiu-001-site1.smarterasp.net>

2. <http://alt.qcri.org/semeval2014/task10/>

direct ou indirect. La zone (IN), de taille plus réduite, est composée d'articles qui permettent d'accéder aux articles de la zone (SCC), mais qui ne sont pas accessibles depuis cette zone. La zone (OUT), de taille équivalente à (IN) est composée à l'inverse d'articles qui sont accessibles depuis la zone (SCC), mais qui n'y renvoient pas. Les tubes sont des zones de taille plus réduites, qui relient directement les articles de la zone (IN) aux articles de la zone (OUT), sans passer par la zone (SCC). Les vrilles sont des zones atypiques qui relient des articles isolés de l'ensemble, soit à la zone (OUT), soit à la zone (IN).

Plus de 2/3 des articles de Wikipédia appartiennent au large noyau (SCC) et un signe de maturité de Wikipédia est la bonne stabilité dans le temps des différentes composantes ; ce qui serait actuellement le cas du Wikipédia anglais.

- **Nature sémantique des liens** : alors que dans les documents Web, un auteur peut arbitrairement lier une page à une autre, les liens dans Wikipédia indiquent une pertinence par rapport à un contexte local : un lien de la page A vers la page B indique que la page B est sémantiquement reliée au contenu, ou une partie du contenu de la page A.
- **Structure des liens** : les liens entrants dans Wikipédia ont tendance à se comporter comme les liens sortants (Jaap & Marijn, 2009) ; ce qui est consistant avec la nature sémantique des liens dans Wikipédia : si un lien de la page A vers la page B souligne une certaine pertinence de B alors il est vraisemblable que A soit également pertinent pour B.
- **Domaines couverts et qualité** : Wikipédia couvre des domaines de connaissance très variés, Arts, Géographie, Histoire, Science, Sports, Jeux... Dans le domaine des Sciences, cette encyclopédie collaborative s'avère aussi précise que l'"Encyclopedia Britannica" (Giles, 2005).
- **Evolution dans le temps** : la structure de Wikipédia et son évolution dans le temps sont régulièrement analysés (Voss, 2005; Buriol *et al.*, 2006; Capocci *et al.*, 2006; Nakayama *et al.*, 2008) et il s'avère qu'à l'instar du Web, cette encyclopédie se densifie au fil du temps aussi bien dans son contenu (nombre d'articles, longueur des articles) que dans sa structure en liens (nombre de liens entrants et sortants par article).
- **Référencement des articles** : chaque article (ou concept) de Wikipédia est référencé de manière unique par une adresse URL ; ce qui élimine tout risque d'ambiguïté.

Les caractéristiques précédentes et son multilinguisme font de Wikipédia un outil de choix pour le TAL qui ont d'ores et déjà donné lieu à des résultats intéressants (Gabrilovich & Markovitch, 2007; Hadj Taieb *et al.*, 2013; Strube & Ponzetto, 2006; Chan *et al.*, 2013). Cependant sa généralité, sa taille et son évolution permanente posent des problèmes de mise en œuvre, particulièrement pour les méthodes basées sur la vectorisation, étant donné la taille des espaces manipulés. Le Random Indexing est la solution que nous avons retenue pour pallier cette difficulté.

3 Random Indexing

Dans la méthodologie des VSM, un espace vectoriel de grande dimension est généré par la construction d'une matrice de co-occurrences F , dans laquelle chaque ligne F_w représente un unique mot et chaque colonne F_c représente un contexte c , typiquement un segment de plusieurs mots tel qu'un document, ou un autre mot. Dans les GVSM, ce sont les documents qui sont utilisés comme base de l'espace, pour répondre à la critique selon laquelle les mots ne constituent pas une base de vecteurs libres (Carbonell *et al.*, 1997).

Le modèle construit souffre de deux problèmes majeurs : la dimensionnalité et les données éparées. Lorsque le vocabulaire et le nombre de documents du corpus augmentent, la matrice de co-occurrence F entre termes et documents devient numériquement lourde à exploiter. Par ailleurs, une très grande proportion des mots n'apparaissent que dans un ensemble de documents très limité. Ainsi, dans une matrice de co-occurrence typique, 99% des entrées sont des zéros.

Pour pallier ces problèmes, diverses techniques de réduction de dimension peuvent être mises en œuvre, comme la décomposition en valeurs singulières (SVD) de la matrice F (Kumar, 2009). La nécessité de construire préalablement la matrice de co-occurrence entre termes et documents est un gros inconvénient lorsque l'on utilise des corpus en évolution constante tels que Wikipédia.

Une alternative aux techniques de réduction de dimension est le Random Indexing, basé sur le travail de Pentti Kanerva sur les représentations de données éparées (Kanerva, 1988; Kanerva *et al.*, 2000). Le Random Indexing procède d'abord par la représentation de chaque concept par un vecteur index de taille réduite, et ensuite le vecteur concept de chaque mot est calculé par sommation des vecteurs index de tous les concepts auxquels il est associé. Ainsi, l'ajout de nouveaux contextes n'implique pas une reconstruction complète de la matrice : il suffit de créer de nouveaux vecteurs index et d'ajouter à la matrice les vecteurs colonnes correspondant aux nouveaux documents.

Les vecteurs index aléatoires sont choisis presque orthogonaux, ce qui conduit à une description approximative de l'espace contexte où les distances entre points sont approximativement préservées (William & Lindenstrauss, 1984). La description

qui suit du Random Indexing est faite à partir de celle qu'en a donnée Sahlgren (Sahlgren, 2005).

On alloue un vecteur index unique de longueur d à chaque contexte. Ces vecteurs sont constitués d'un grand nombre de 0 et d'un petit nombre de 1 et de -1. À chaque composante est allouée l'une de ces valeurs avec la probabilité suivante :

$$\begin{cases} +1 & \text{avec une probabilité } s/2 \\ 0 & \text{avec une probabilité } 1 - s \\ -1 & \text{avec une probabilité } s/2 \end{cases}$$

où s désigne le nombre d'éléments non nuls. Le choix de s et d se fait en fonction du nombre de contextes à représenter. Pour chaque nouveau concept, un vecteur index est produit. Le vecteur contexte d'un terme est la somme des vecteurs index de tous les contextes dans lesquels ce terme apparaît.

Le vecteur contexte d'un terme qui apparaît dans chacun des contextes $c_1 = [1, 0, 0, -1]$ et $c_2 = [0, 1, 0, -1]$ serait $[1, 1, 0, -2]$. Si le contexte c_1 est rencontré de nouveau, il n'y a pas création de nouveau vecteur index et la mise-à-jour du vecteur contexte de t se fait par addition du vecteur index de c_1 ; ce qui conduit au nouveau vecteur contexte de t : $[2, 1, 0, -3]$. La distance entre ces vecteurs contextes peut être évaluée au moyen de différentes mesures de distance. Sahlgren et Karlgren (2005) utilisent la mesure cosinus (Sahlgren & Karlgren, 2005).

Une version pondérée du Random Indexing a été proposée par (Gorman & Curran, 2006) et les auteurs l'utilisent pour mesurer la similarité sémantique entre phrases. Le vecteur contexte d'un mot y est calculé comme la somme pondérée des vecteurs index des contextes qui lui sont associés. Les auteurs comparent plusieurs fonctions de pondération dans une tâche d'extraction de synonymie : fréquence du mot dans le contexte, fréquence relative, $tf-idf$, $tf-idf^*$ (version log-pondérée du $tf-idf$), $DICE$, etc. Ils concluent à une nette amélioration des performances de RI en présence de grands corpus de données. Pour des ensembles de données réduits, RI est suffisamment robuste et la pondération n'a, au mieux, qu'un effet mineur. Ils constatent également une grande variabilité dans l'effet des fonctions poids utilisées et les bonnes performances de la fonction $tf-idf^*$.

4 Calcul de la similarité entre phrases

Le calcul de la similarité entre phrases a été mis en œuvre en effectuant les étapes suivantes.

- Un étiqueteur syntaxique (en l'occurrence TreeTagger³) traite l'ensemble des articles de Wikipédia et convertit chacun de leurs termes en lemmes ("*travaille*" → "*travailler*").
- Ensuite, le coefficient de pondération du $tf-icf$ (Term Frequency-Inverse Corpus Frequency) (Reed *et al.*, 2006) de chaque terme (lemme) est calculé pour chaque article :

$$tf-icf_{ij} = \log(1 + f_{ij}) \cdot \log\left(\frac{N + 1}{n_i + 1}\right) \quad (1)$$

où f_{ij} est le nombre d'occurrences du terme d'indice i dans le document d'indice j , N le nombre total de documents d'un sous-corpus choisi suffisamment large et diversifié et n_i le nombre de documents où apparaît le terme d'indice i . Le coefficient $tf-icf$ fournit une approximation du véritable $tf-idf$ construit sur le corpus entier et il permet de traiter à moindre coût, des corpus dynamiques ou de très grande taille. Dans les expérimentations que nous présentons, nous avons considéré une version complète et statique de Wikipedia.

- L'ensemble des concepts est identifié avec celui des articles, chaque article définissant un concept et un concept n'existant que s'il existe un article qui le définit. Les valeurs du $tf-icf$ d'un terme par rapport à l'ensemble des articles sont les composantes d'un vecteur appelé *vecteur sémantique de terme* dans la base des concepts.
- La valeur sémantique d'une phrase est calculée à partir des vecteurs sémantiques des termes qui la composent.

4.1 Calcul des vecteurs sémantiques

Un vecteur de terme est la représentation des liens entre ce terme et chacun des concepts, où l'ensemble des concepts est identifié à l'ensemble des articles de Wikipédia. Selon nos calculs, après avoir appliqué les étapes de prétraitement du corpus Wikipédia : filtrage du texte proprement dit, suppression des articles trop courts ou ayant un nombre trop faible

3. <http://www.cis.uni-muenchen.de/schmid/tools/TreeTagger>

de liens, suivant les étapes suivies dans (Bawakid, 2011, p. 129), il reste 1 015 879 articles avec le Wikipédia français du 20/11/2013 et 3 766 589 articles avec le Wikipédia anglais du 02/12/2013.

Pour résoudre le problème de la réduction de dimension des vecteurs de terme, nous avons utilisé la méthode d'indexation aléatoire du Random Indexing décrite dans la section 3, en suivant les étapes ci-après.

- **Définition des vecteurs index de concept** : à chaque concept Wikipédia est attribué un vecteur index unique \vec{c}_j dans un espace de dimension d fixée (cf. section 3). Étant donné le nombre de concepts des Wikipédia anglais et français, une dimension d de quelques milliers suffit pour assurer l'existence de vecteurs index presque orthogonaux.
- **Définition des vecteurs sémantiques de terme** : les vecteurs des termes présents dans le corpus Wikipédia sont calculés selon la formule (2).

$$\overrightarrow{terme_i} = \sum_{j=1}^N tf\text{-}icf_{ij} \cdot \vec{c}_j \quad (2)$$

où N est le nombre de concepts Wikipédia, \vec{c}_j est le vecteur index du concept j et $tf\text{-}icf_{ij}$, le *Term Frequency-Inverse Corpus Frequency* du terme d'indice i dans le document (concept) d'indice j calculé suivant la formule (1).

- **Similarité entre phrases** : pour calculer la similarité entre deux phrases, chacune d'elles doit d'abord être représentée comme un vecteur sémantique. On suppose que Wikipédia a une couverture des concepts et des mots suffisamment large pour contenir la plupart des termes sémantiquement significatifs utilisés dans les phrases en question. Le vecteur sémantique d'une phrase se calcule en faisant la somme des vecteurs sémantiques des termes qui la composent, suivant la formule (3).

$$\vec{S} = \sum_{i=1}^n \overrightarrow{terme_i}. \quad (3)$$

Toutefois, cette mesure ne prend pas en considération le poids interne des mots dans le texte ou dans l'ensemble de textes d'où la phrase est extraite. L'hypothèse est que, si un mot est très fréquent dans les documents concernés, il convient de minimiser son importance au niveau de la phrase. Pour cela et conformément aux travaux de Neto et al., nous utilisons la pondération par le *tf-isf* (term frequency \times inverse sentence frequency) (Neto et al., 2000, 2002). Le *tf* est ici le nombre d'occurrences du terme dans la phrase et l'*isf* est calculé d'après la proportion de phrases dans l'ensemble des documents qui contiennent le terme :

$$tf\text{-}isf_{is} = tf_{is} \cdot \log\left(\frac{|S|}{SF_i}\right) \quad (4)$$

où $|S|$ est le nombre de phrases et SF_i le nombre de phrases qui contiennent le terme d'indice i . Ainsi, l'importance d'un terme qui apparaît dans un grand nombre de phrases de l'ensemble des documents s'en trouve réduite. La sémantique d'une phrase est finalement représentée par une combinaison linéaire des vecteurs des termes qui la composent, pondérés par leurs *tf*s respectifs :

$$\vec{S}_i = \sum_{j=1}^n tf_{ij} \cdot \overrightarrow{terme_j}. \quad (5)$$

La similarité entre deux phrases S_i et S_j dans un document (ou multi-document) est ensuite définie comme le cosinus de leurs vecteurs sémantiques respectifs⁴ :

$$Sim_{WikiRI}(S_i, S_j) = \cos(\vec{S}_i, \vec{S}_j). \quad (6)$$

4.2 Nouveau calcul des vecteurs de termes

Nos premières expérimentations ayant donné des résultats décevants, nous avons analysé finement les mesures de similarités obtenues entre certains termes et groupements de termes pour mieux comprendre les insuffisances de la méthode. Des dysfonctionnements s'observent lorsque sont associés des termes qui diffèrent de par leur fréquence. Après avoir décrit le phénomène, nous proposons une modification dans le calcul des coordonnées des vecteurs de termes.

4. D'autres mesures ont été testées sans qu'une amélioration significative des résultats n'ait été constatée.

Les mots grammaticaux (*stop-words*) sont très fréquents dans les articles de Wikipédia, comme dans tous les textes écrits en langue française ou anglaise. Malgré leur importance pour la bonne compréhension d'un texte par ses lecteurs, ces termes ne sont pas pris en compte dans le calcul des vecteurs sémantiques.

Certains termes, que nous désignerons par *termes généraux*, ne sont pas des mots grammaticaux mais sont néanmoins très fréquents dans les articles de Wikipédia. La table 1 en donne quelques exemples pour la langue française, avec leur nombre d'occurrences dans Wikipédia, le pourcentage des articles dans lesquels ils apparaissent et la valeur de leur coefficient *icf*.

Terme	cf	Couverture	icf	Terme	cf	Couverture	icf
naître	298 963	29,60%	0,52	Lune	6 667	0,66%	2,18
pouvoir	293 035	29,01%	0,53	NASA	3 528	0,35%	2,45
grand	263 987	24,14%	0,58	peste	4 917	0,49%	2,31
nouveau	235 462	23,31%	0,63	sida	1 524	0,15%	2,82

TABLE 1 – Exemples de l'importance des termes généraux dans le Wikipédia français.

À l'inverse, un grand nombre de termes ont un nombre d'occurrences beaucoup plus faible. Il s'agit souvent de termes spécifiques à un domaine déterminé et qui sont essentiels pour une modélisation pertinente de la sémantique d'une phrase.

Ainsi, lorsque l'on évalue la similarité entre groupements de termes où sont associés un terme très fréquent avec un terme spécifique, on constate que l'influence du terme le plus fréquent écrase celui du terme spécifique. Par exemple, les lemmes *robot* et *infection* ont respectivement des *cf* relativement faibles, respectivement égaux à 5930 et 3593. À ce titre, ils peuvent être considérés comme des mots spécifiques. Par ailleurs, leur score de similarité (calculé comme le cosinus de leurs vecteurs de terme) est très faible (peu différent de 0,007). Or, les groupements de termes *petit robot*/*petite infection* obtiennent, avec le calcul de similarité défini précédemment, un score peu différent de 0,89, une valeur intuitivement beaucoup trop élevée, due à la prééminence du vecteur de termes *petit* sur les deux autres vecteurs de termes.

Autrement dit, bien que l'*icf* ait considérablement réduit le poids des termes généraux, la réduction qu'il opère n'est pas suffisante.

4.2.1 Modification des coordonnées des vecteurs de terme

L'objectif est donc de rééquilibrer le poids des termes très fréquents (mots généraux) par rapport à celui des termes plus rares, souvent spécifiques à un domaine donné, par rapport aux valeurs obtenues par le calcul classique du *tf-icf*. Pour ce, on introduit un paramètre $\alpha \geq 1$, destiné à renforcer le poids du *icf*, selon la formule (7).

$$tf\text{-}icf_{\alpha} = tf \cdot icf^{\alpha}, \quad (7)$$

Le paramètre α est estimé par apprentissage sur les ensembles de données SemEval-2012 TASK 6⁵, choisies comme données d'entraînement pour le système.

Plus précisément, pour chacun des cinq ensembles de données SemEval-2012, nous avons calculé les similarités pour chaque paire de phrases, puis les scores obtenus par le système ont été comparés avec les similarités du "gold standard" qui sont fournies par SemEval-2012 pour obtenir les scores d'évaluations. Après avoir examiné les résultats obtenus avec différentes valeurs du paramètre α comprises entre 1 et 7, nous avons constaté que la valeur $\alpha = 3$ correspondait au meilleur résultat d'évaluation pour chacun des cinq corpus de Semeval-2012 testés.

Avec la valeur $\alpha = 3$, le calcul de la similarité des groupes de termes *petit robot* et *petite infection*, qui combinent des mots très généraux avec des mots moins fréquents, donne un résultat intuitivement acceptable, avec une valeur égale à 0,091.

4.2.2 Modification des vecteurs sémantiques de phrase

Les résultats sont améliorés par l'introduction du paramètre α . Cependant, cette modification du calcul des coordonnées des vecteurs sémantiques des termes agit sur la partie *icf* du *tf-icf*: elle ne fait donc que modifier la norme des vecteurs de

5. <http://www.cs.york.ac.uk/semeval-2012/task6/>

termes. En particulier, elle ne résoud pas le caractère creux des vecteurs sémantiques des termes peu fréquents. En d'autres termes, ces derniers contiennent toujours principalement des coordonnées nulles. Conformément aux auteurs (Higgins & Burstein, 2007), les vecteurs des mots rares peuvent être enrichis en utilisant le vecteur centroïde du texte défini suivant la formule suivante.

$$\overrightarrow{centroid} = \frac{1}{n} \sum_{i=1}^n \overrightarrow{term}_i, \quad (8)$$

où n est le nombre de termes distincts dans le texte à calculer.

L'introduction dans le calcul du vecteur sémantique d'une phrase de son vecteur centroïde augmente l'apparition des coordonnées des vecteurs des termes rares et amoindrit le biais introduit par la fréquence des termes généraux. Le vecteur sémantique d'une phrase est finalement calculé en remplaçant la formule (3) par la formule (9).

$$\vec{S}_i = \sum_{j=1}^n tf_{ij} \cdot (\overrightarrow{term}_j - \overrightarrow{centroid}), \quad (9)$$

où \overrightarrow{term}_j est le vecteur du terme d'indice j et n le nombre de termes distincts dans la phrase d'indice i .

5 Expérimentations et résultats

Les expérimentations ont été effectuées sur deux langues, l'anglais et le français.

D'après (Kanerva *et al.*, 2000) et étant donné la taille des corpus obtenus après les opérations de prétraitement, les vecteurs index ont été représentés dans des espaces de dimension $d = 5\,000$ pour le Wikipédia français et $d = 10\,000$ pour le Wikipédia anglais. Suivant les indications des mêmes auteurs, le nombre de composantes non nulles est fixé à $s = 20$ dans le premier cas et à $s = 26$ dans le second.

Les résultats rendus par le système WIKIRI ont été évalués en utilisant le coefficient de corrélation de Pearson entre les scores de système et les scores des annotateurs humains, comme il est habituel pour ce type de tâche.

5.1 Évaluation pour l'anglais

L'évaluation a été réalisée sur les données de la tâche 10 de **SemEval-2014** (Agirre *et al.*, 2014) qui contient 6 types de corpus à évaluer pour l'anglais :

1. **Discussion de forum** (deft-forum) : 450 paires de phrases.
2. **Discussion de l'actualité** (deft-news) : 300 paires de phrases.
3. **Titres de l'actualité** (headlines) : 750 paires de phrases.
4. **Descriptions d'images** (image) : 750 paires de phrases.
5. **Définitions extraites de OntoNotes et de WordNet** (OnWN) : 750 paires de phrases
6. **Titres et commentaires de nouvelles sur tweeter** (tweet-news) : 750 paires de phrases.

La table 2 présente une analyse comparative des corpus de Semeval où figurent leur nombre de mots (non grammaticaux) par phrase, leurs pourcentages d'adverbes, d'adjectifs, de noms communs, de noms propres, de verbes, ainsi que le pourcentage moyen de mots (non grammaticaux) communs entre les phrases des paires testées. Le faible pourcentage de noms propres dans certains corpus correspond au fait que le choix y a été fait de supprimer les majuscules. Par ailleurs, on peut également noter le très important pourcentage de mots qu'ont en commun les phrases testées.

SemEval fournit les "gold standard" des 6 corpus et un outil pour évaluer les systèmes. En 2014, 15 équipes ont participé à cette évaluation et les résultats de 38 systèmes ont été comparés. En utilisant la valeur de $\alpha = 3$ déterminée avec les corpus de SemEval-2012, notre système a obtenu les scores suivants : 47,005% avec deft-forum, 63,820% avec deft-news, 56,584% avec headlines, 75,884% avec image et 73,995% avec OnWN. La Figure 2 compare les résultats du système (en rose) avec ceux des systèmes qui ont participé à SemEval2014. WikiRI se place au-dessus de la moyenne des systèmes pour tous les corpus, à l'exception de celui concernant les titres de l'actualité.

Or, les meilleurs systèmes utilisent des corpus qui sont soit plus grands soit plus élaborés que Wikipédia, tels que Stanford WebBase Project (Kashyap *et al.*, 2014) ou des corpus de paraphrases (Sultan *et al.*, 2014). WikiRI obtient donc des

	Nb_Mots/Ph	ADV	ADJ	NC	NP	V	Communs/Ph
deft-news	11,8	1,9%	11,2%	33,7%	0%	14,8%	32,6%
headlines	6,3	0,7%	7,5%	25,3%	21,1%	11,6%	22,4%
images	5,8	0,4%	10,4%	30,8%	0,7%	9,5%	25,1%
OnWN	5,25	2%	6,2%	24,9%	0,2%	14,8%	25,2%
deft-forum	6,6	6%	5,6%	16,8%	5,2%	19%	33%
tweet-news	7,4	2,2%	5,4%	18,7%	20,8%	11,1%	19%

TABLE 2 – Analyse comparative des différents corpus de tests de Semeval.

résultats tout à fait encourageants puisqu'il obtient des résultats au niveau de l'état de l'art en utilisant Wikipédia pour seule ressource.

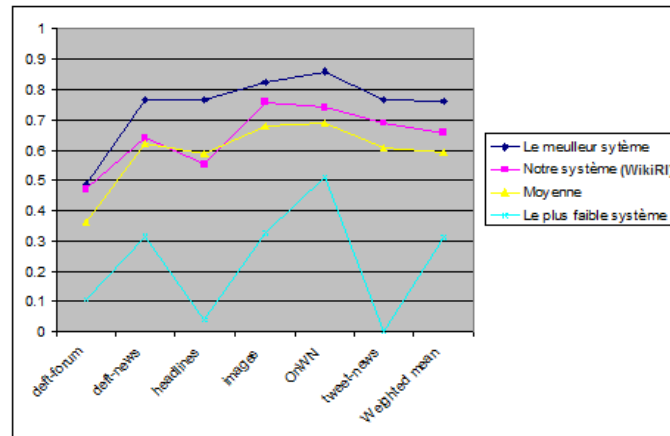


FIGURE 2 – Comparaison des résultats de WIKIRI avec ceux des systèmes proposés à SemEval-2014.

5.2 Évaluation pour le français

Si SemEval2014 contient des données pour l'anglais et pour l'espagnol, il n'existe pas de corpus annoté en français actuellement pour la tâche qui nous intéresse. Créer un tel corpus est un travail long et difficile : tester toutes les paires d'un ensemble de n phrases devient rapidement impraticable de par la croissance quadratique du nombre de paires en fonction de n . Nous avons extrait du Web deux corpus de textes français dans deux domaines différents définis respectivement par les mots-clefs "Épidémies" et "Conquête spatiale". Dans chaque corpus, nous avons sélectionné un ensemble de soixante-dix phrases, dont la longueur varie de 10 à 65 mots. Dix d'entre elles ont été choisies comme phrases de référence : elles contiennent diverses informations importantes concernant les domaines testés. Chacune de ces dix phrases a été associée à six autres phrases choisies de sorte que les différents niveaux de similarité entre phrases (sur une échelle de 0.0 à 4.0) soient représentés. La table 3 contient les mêmes indications que celles données pour le corpus Semeval : nombre de mots non grammaticaux par phrase, pourcentages d'adverbes, d'adjectifs de noms propres et de verbes, moyenne du nombre de mots non grammaticaux communs entre les phrases des paires testées. Ces données montrent que les phrases sont notablement plus longues que celles des corpus de Semeval, excepté celles du corpus *deft-news* ; par ailleurs, l'application visée étant le résumé multi-textes, le pourcentage de mots communs entre phrases est également beaucoup plus faible, notre échantillon se voulant représentatif de la tâche à laquelle devrait se confronter le système.

Sept volontaires humains, âgés de 18 à 60 ans, ont été impliqués dans la tâche d'annotation dont trois experts et quatre candides. Ils ont évalué la similitude des paires de phrases sur une échelle de 0,0 à 4,0 (les décimales étaient autorisées), selon les consignes indiquées dans la Table 4 et suivant la procédure d'annotation décrite dans (Li *et al.*, 2006).

La Table 5 donne l'une des phrases de référence (en gras) avec les phrases qui lui ont été associées. Les données du tableau correspondent à la moyenne des scores de similarité attribués par les sept annotateurs à chacune des six paires de phrases.

	Nb_Mots/Ph	ADV	ADJ	NC	NP	V	Communs/Ph
Epidémies	12,6	2,5%	10,9%	22,7%	3,7%	10%	9,7%
Conquête spatiale	16,1	2,2%	10,7%	21,4%	8,1%	11,4%	6,8%

TABLE 3 – Comparaison des corpus de tests *épidémies* et *conquête spatiale*.

4.0 : Les phrases sont complètement équivalentes ;
3.0 : Les phrases sont globalement équivalentes, mais elles diffèrent par quelques détails ;
2.0 : Les phrases ne sont pas équivalentes, mais elles partagent certaines parties de l'information ;
1.0 : Les phrases ne sont pas équivalentes, mais elles traitent du même sujet ;
0.0 : Les phrases ne sont pas liées.

TABLE 4 – Les instructions d'annotation pour le choix du score de similarité entre phrases

(1) *Mars est l'astre le plus étudié du système solaire, puisque 40 missions lui ont été consacrées, qui ont confirmé la suprématie américaine - des épopées Mariner et Viking aux petits robots Spirit et Opportunity (2003 et 2004).*
(2) *Le 28 novembre 1964, la sonde Mariner 4 est lancée vers Mars, 20 jours après l'échec de Mariner 3.*
(3) *Les robots Spirit et Opportunity, lancés respectivement le 10 juin 2003 et le 8 juillet 2003 par la NASA, représentent certainement la mission la plus avancée jamais réussie sur Mars.*
(4) *Le bilan de l'exploration de Mars est d'ailleurs plutôt mitigé : deux tiers des missions ont échoué et seulement cinq des quinze tentatives d'atterrissage ont réussi (Viking 1 et 2, Mars Pathfinder et les deux MER).*
(5) *Le 6 août 2012, le rover Curiosity a atterri sur Mars avec 80 kg de matériel à son bord.*
(6) *Arrivé sur Mars en janvier 2004 comme son jumeau Spirit, et prévu comme lui pour fonctionner au moins trois mois, Opportunity (alias MER-B) roule encore et plusieurs de ses instruments répondent présents.*
(7) *Mars est mille fois plus lointaine que la Lune et son champ d'attraction plus de deux fois plus intense : la technologie n'existe pas pour envoyer un équipage vers Mars et le ramener sur Terre.*

Paires des phrases	(1)-(2)	(1)-(3)	(1)-(4)	(1)-(5)	(1)-(6)	(1)-(7)
Score de similarité	0,49	2,06	1,86	1,19	1,57	1,1

TABLE 5 – Les scores de similarité d'une phrase de référence avec ses six phrases associées.

Les participants ont travaillé indépendamment et sans contrainte de temps sur une application Web⁶ conçue pour leur faciliter la tâche d'annotation. Pour chaque phrase de référence choisie au hasard, ses phrases associées ont été aléatoirement et successivement présentées à l'annotateur. Ce dernier disposait d'un historique des scores de similarité qu'il avait déjà attribués et il était libre de les modifier à tout moment. Pour estimer l'accord inter-annotateurs, nous avons comparé les scores de chaque annotateur à la moyenne des scores calculée sur le reste du groupe. Les coefficients de corrélation ainsi obtenus sont présentés dans la table 6⁷. Compris entre 0,8 et 0,941, ils indiquent que les évaluateurs humains sont largement d'accord sur les définitions utilisées dans l'échelle, même s'ils ont trouvé la tâche d'annotation particulièrement difficile.

Pour chacun des deux corpus, le système a été testé avec différentes valeurs du paramètre α . Les résultats ont été évalués à l'aide du coefficient de corrélation de Pearson, comme dans la tâche correspondante de SemEval. Ils sont donnés dans la première partie du tableau (lignes WikiRI) de la table 7. La deuxième partie du tableau contient les résultats obtenus avec un système précédemment implémenté (Vu *et al.*, 2014) inspiré de la méthode ESA (Gabrilovich & Markovitch, 2007), une variante du modèle GVSM. Chacun des corpus étant lié à un domaine spécifique, un choix des concepts les plus pertinents basé sur l'étude des liens Wikipédia précédait la construction de la matrice termes \times concepts. D'après Gotttron *et al.*, 2011), une réduction de dimension est d'autant plus efficace que l'on travaille dans un domaine spécifique (Gotttron *et al.*, 2011).

6. <http://vuhaihiu-001-site1.smarterasp.net>

7. Le choix de laisser les annotateurs utiliser des valeurs décimales ne permettait pas d'utiliser un kappa pour estimer l'accord.

Annotateurs	1	2	3	4	5	6	7
Corrélation (c. spatiale)	0,872	0,869	0,844	0,941	0,886	0,815	0,855
Standard Déviation (c. spatiale)	0,586	0,640	0,714	0,364	0,624	0,671	0,568
Corrélation (épidémies)	0,862	0,904	0,903	0,931	0,846	0,846	0,800
Standard Déviation (épidémies)	0,544	0,514	0,622	0,367	0,651	0,580	0,617

TABLE 6 – Les coefficients de corrélation entre les scores de chaque annotateur et la moyenne des scores des six autres.

WikiRI α	1	2	2,25	2,5	3	4	4,5	4,75	5
Epidémies	0,64788	0,79418	0,79955	0,79593	0,7754	0,72649	0,70075	0,68663	0,67148
Conquête spatiale	0,64811	0,74963	0,76075	0,77049	0,79191	0,83745	0,84846	0,84943	0,84696
ESA α	1	1,25	1,5	2	3,75	4	4,25	4,5	5
Epidémies	0,52541	0,53926	0,5388	0,51306	0,38146	0,36229	0,34119	0,31904	0,27683
Conquête Spatiale	0,55626	0,5611	0,56051	0,56563	0,61389	0,61692	0,61659	0,61241	0,59465

TABLE 7 – Les résultats du système pour les deux corpus en langue française suivant différentes valeurs du paramètre α .

Une première constatation est que les résultats obtenus par le système WikiRI, qui utilise l'ensemble de Wikipédia, sont très largement supérieurs à ceux obtenus par le système ESA pour des espaces de concepts limités à ceux des domaines considérés. Par ailleurs, l'introduction du paramètre α est plus efficace pour le système WikiRI que pour le système inspiré de la méthode ESA. Ces résultats sont conformes aux conclusions de Gordon et al. (Gorman & Curran, 2006) concernant l'influence des pondérations sur le système RI.

La seconde constatation est que, si la valeur optimale du paramètre α reste stable entre les différents corpus en langue anglaise de SemEval, il n'en est pas de même entre les deux corpus de domaine en langue française, puisque le meilleur résultat est obtenu avec $\alpha = 2,25$ pour le corpus *épidémies* et $\alpha = 4,75$ pour le corpus *conquêtes spatiales*. Néanmoins, l'introduction du paramètre s'avère très efficace : les résultats obtenus pour $\alpha = 1$, qui correspondent à l'utilisation du *tf-idf* classique, sont largement inférieurs à ceux obtenus pour les valeurs optimales (0,648 contre 0,800 et 0,648 contre 0,849). Par ailleurs, on constate la même variabilité de la valeur optimale de α pour le système inspiré de la méthode ESA que pour le système WikiRI.

Il est actuellement difficile de savoir si cette instabilité constatée du α optimal est imputable à la langue ou à la nature même des corpus que nous avons volontairement choisis très différents. D'après les données de la table 3, la principale différence concerne les noms propres (NP), presque trois fois plus fréquents dans le corpus *conquête spatiale* que dans le corpus *épidémies*. Dans ce second corpus en effet, les termes spécifiques au domaine sont souvent des noms communs : *peste*, *choléra*, *vaccin*, *bacille*, *virus*, etc. alors qu'ils concernent plus fréquemment des hommes ou des engins spatiaux dans le premier : *Gagarine*, *Curiosity*, *Spoutnik*, *Armstrong*, etc. Cependant, des expérimentations supplémentaires seront nécessaires pour pouvoir mieux comprendre la relation qui peut exister entre le choix du meilleur α et la nature du corpus.

6 Conclusion et perspectives

Nous avons présenté une méthode de modélisation de la sémantique d'un mot ou d'un texte basée sur l'utilisation de Wikipédia, qui utilise la technique d'indexation aléatoire RI pour réduire la dimension des espaces vectoriels de représentation. Par ailleurs, des modifications ont été introduites dans le calcul des vecteurs représentant les termes et les phrases pour réduire le bruit que peut engendrer la multiplicité des concepts dans une ressource linguistique aussi foisonnante. La technique d'indexation aléatoire a montré son efficacité dans la réduction de la complexité des calculs, mais elle semble très sensible au choix des pondérations utilisées. Les résultats obtenus sur les données de SemEval2014 pour l'anglais sont au niveau de l'état de l'Art, ce qui prouve l'efficacité de l'approche. Testée également sur la langue française, la méthode donne des résultats très encourageants, même si des expérimentations supplémentaires sont nécessaires pour mieux comprendre l'influence du paramètre α que nous avons introduit. Elle offre l'avantage d'être utilisable pour d'autres langues, à la condition d'y disposer de ressources Wikipédia suffisamment développées. Si le choix a été fait d'utiliser la totalité de Wikipédia, la question reste ouverte de savoir quel est le nombre minimal de documents qui pourrait assurer une qualité suffisante à la détermination des vecteurs de termes.

Nos travaux actuels cherchent à utiliser les similarités entre phrases pour implémenter une méthodologie de résumés multi-textes. Pour l'anglais, le système sera testé sur les données DUC. Pour le français, il utilisera les données du corpus rpm2 (de Loupy *et al.*, 2010).

Références

- AGIRRE E., BANECA C., CARDIE C., CER D., DIAB M., GONZALEZ-AGIRRE A., GUO W., MIHALCEA R., RIGAU G. & WIEBE J. (2014). Semeval-2014 task 10 : Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 81–91, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.
- BALASUBRAMANIAN N., ALLAN J. & CROFT W. B. (2007). A comparison of sentence retrieval techniques. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 813–814 : ACM.
- BAWAKID A. (2011). *Automatic Documents Summarization Using Ontology based Methodologies*. PhD thesis, University of Birmingham.
- BURIOL L. S., CASTILLO C., D. D., S. L. & S. M. (2006). Temporal analysis of the wikigraph. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference of Web Intelligence.*, p. 45–51.
- CAPOCCI A., SERVEDIO V., COLAIORI F. & BURIOL L. (2006). Preferential attachment in the growth of social networks : the case of wikipedia. *Arxiv preprint physics*.
- CARBONELL J. G., YANG Y., FREDERKING R. E., BROWN R., GENG Y. & D. L. (1997). Translingual information retrieval : a comparative evaluation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'97 Distinguished Paper Award)*.
- CHAN P., HIJIKATA Y. & NISHIDA S. (2013). Computing semantic relatedness using word frequency and layout information of wikipedia. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, p. 282–287 : ACM.
- CHAVEZ A., DÁVILA H., GUTIÉRREZ Y., FERNÁNDEZ-ORQUÍN A., MONTOYO A. & MUÑOZ R. (2014). Umcc_dlsi_semsim : Multilingual system for measuring semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 716–721, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.
- DE LOUPY C., GUÉGAN M., AYACHE C., SENG S. & MORENO J.-M. T. (2010). A french human reference corpus for multi-document summarization and sentence compression. In N. C. C. CHAIR, K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS, M. ROSNER & D. TAPIAS, Eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).
- ERKAN G. & RADEV D. R. (2004). Lexrank : Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.(JAIR)*, **22**(1), 457–479.
- GABRILOVICH E. & MARKOVITCH S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, p. 1606–1611.
- GILES J. (2005). Internet encyclopedias go head to head. *Nature*, **438**, 900–901.
- GORMAN J. & CURRAN J. R. (2006). Random indexing using statistical weight functions. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP '06*, p. 457–464, Stroudsburg, PA, USA : Association for Computational Linguistics.
- GOTTRON T., ANDERKA M. & STEIN B. (2011). Insights into explicit semantic analysis. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, p. 1961–1964 : ACM.
- HADJ TAIEB M. A., BEN AOUICHA M. & BEN HAMADOU A. (2013). Computing semantic relatedness using wikipedia features. *Knowledge-Based Systems*, **50**, 260–278.
- HIGGINS D. & BURSTEIN J. (2007). Sentence similarity measures for essay coherence. In *Proceedings of the 7th International Workshop on Computational Semantics*, p. 1–12.
- HIRAO T., OKUMURA M. & ISOZAKI H. (2005). Kernel-based approach for automatic evaluation of natural language generation technologies : Application to automatic summarization. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, p. 145–152 : Association for Computational Linguistics.

- JAAP K. & MARIJN K. (2009). Is wikipedia link structure different ? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, p. 232–241.
- KANERVA P. (1988). *Sparse distributed memory*. MIT Press.
- KANERVA P., KRISTOFERSSON J. & HOLST A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd annual conference of the cognitive science society*, volume 1036 : Erlbaum.
- KASHYAP A., HAN L., YUS R., SLEEMAN J., SATYAPANICH T., GANDHI S. & FININ T. (2014). Meerkat mafia : Multilingual and cross-level semantic textual similarity systems. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 416–423, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.
- KO Y., PARK J. & SEO J. (2002). Automatic text categorization using the importance of sentences. In *COLING*.
- KUMAR C. A. (2009). Analysis of unsupervised dimensionality reduction techniques. *Comput. Sci. Inf. Syst.*, **6**(2), 217–227.
- LI Y., MCLEAN D., BANDAR Z. A., O’SHEA J. D. & CROCKETT K. (2006). Sentence similarity based on semantic nets and corpus statistics. *Knowledge and Data Engineering, IEEE Transactions on*, **18**(8), 1138–1150.
- LIN C.-Y. (2004). Rouge : a package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out*, p. 25–26.
- NAKAYAMA K., HARA T. & NISHIO S. (2008). Wikipedia link structure and text mining for semantic relation extraction towards a huge scale global web ontology.
- NETO J. L., FREITAS A. A. & KAESTNER C. A. (2002). Automatic text summarization using a machine learning approach. In *Advances in Artificial Intelligence*, p. 205–215. Springer.
- NETO J. L., SANTOS A. D., KAESTNER C. A. & FREITAS A. A. (2000). Generating text summaries through the relative importance of topics. In *Advances in Artificial Intelligence*, p. 300–309. Springer.
- REED J. W., JIAO Y., POTOK T. E., KLUMP B. A., ELMORE M. T. & HURSON A. R. (2006). TF-ICF : A new term weighting scheme for clustering dynamic data streams. In M. A. WANI, T. LI, L. A. KURGAN, J. YE & Y. LIU, Eds., *The Fifth International Conference on Machine Learning and Applications, ICMLA 2006, Orlando, Florida, USA, 14-16 December 2006*, p. 258–263 : IEEE Computer Society.
- SAHLGREN M. (2005). An introduction to random indexing. In *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, TKE*, volume 5.
- SAHLGREN M. & KARLGREN J. (2005). Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Journal of Natural Language Engineering*, **11**(3). Special Issue on Parallel Texts.
- STRUBE M. & PONZETTO S. P. (2006). Wikirelate ! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, p. 1419–1424.
- SULTAN M. A., BETHARD S. & SUMNER T. (2014). Dls@cu : Sentence similarity from word alignment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 241–246, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.
- VOSS J. (2005). Measuring wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*.
- VU H. H., VILLANEAU J., SAÏD F. & MARTEAU P. (2014). Sentence similarity by combining explicit semantic analysis and overlapping n-grams. In P. SOJKA, A. HORÁK, I. KOPECEK & K. PALA, Eds., *Text, Speech and Dialogue - 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings*, volume 8655 of *Lecture Notes in Computer Science*, p. 201–208 : Springer.
- WILLIAM B. & LINDENSTRAUSS J. (1984). Extensions of lipschitz mappings into a hilbert space. In *Conference in Modern Analysis and Probability*.

Typologie automatique des langues à partir de treebanks

Philippe Blache¹, Grégoire de Montcheuil^{1,2}, Stéphane Rauzy¹
(1) Aix-Marseille Université & CNRS, LPL, Aix-en-Provence
(2) Equipex ORTOLANG

blache@lpl-aix.fr, stephane.rauzy@lpl-aix.fr, gregoire.montcheuil@lpl-aix.fr

Résumé. La typologie des langues repose sur l'étude de la réalisation de propriétés ou phénomènes linguistiques dans plusieurs langues ou familles de langues. Nous abordons dans cet article la question de la typologie syntaxique et proposons une méthode permettant d'extraire automatiquement ces propriétés à partir de treebanks, puis de les analyser en vue de dresser une telle typologie. Nous décrivons cette méthode ainsi que les outils développés pour la mettre en œuvre. Celle-ci a été appliquée à l'analyse de 10 langues décrites dans le *Universal Dependencies Treebank*. Nous validons ces résultats en montrant comment une technique de classification permet, sur la base des informations extraites, de reconstituer des familles de langues.

Abstract.

Automatic Linguistic Typology from Treebanks.

Linguistic typology studies different linguistic properties or phenomena in order to compare several languages or language families. We address in this paper the question of syntactic typology and propose a method for extracting automatically from treebanks syntactic properties, and bring them into a typology perspective. We present here the method and the different tools for inferring such information. The approach has been applied to 10 languages of the *Universal Dependencies Treebank*. We validate the results in showing how automatic classification correlates with language families.

Mots-clés : Typologie, syntaxe, treebank, inférence de grammaire, Grammaire de Propriétés.

Keywords: Typology, syntax, grammar inference, Property Grammars.

Introduction

Les treebanks sont des ressources désormais indispensables pour l'analyse syntaxique automatique. Ils constituent de plus une source d'information précieuse pour la description, au sens linguistique du terme, des propriétés syntaxiques des langues. En associant des informations syntaxiques à des données naturelles sur une grande échelle, ils permettent en effet d'extraire des régularités générales, mais offrent en même temps la possibilité de décrire des réalisations spécifiques de certaines tournures syntaxiques. De plus, et dans la mesure où un treebank repose sur un guide d'annotation précis, il est possible d'extraire automatiquement un grand nombre d'informations, en vue par exemple d'appliquer des techniques d'apprentissage automatique ou encore d'étudier la distribution de certains phénomènes.

Cependant, les formats utilisés restent à un niveau de généralité élevé. Par exemple, les treebanks de constituants reposent sur une grammaire syntagmatique implicite, à laquelle s'ajoute éventuellement l'indication des principales fonctions syntaxiques. Il est ainsi possible d'extraire automatiquement une grammaire d'un treebank et de l'illustrer en fournissant l'ensemble des réalisations des règles syntagmatiques de cette grammaire. Mais dans une perspective linguistique, il est nécessaire d'identifier également des informations plus fines, du type de celles associées aux phénomènes de rection (ordre linéaire, cooccurrence etc.) : une langue est en effet caractérisée par ce type d'indices plus que par une grammaire à proprement parler. De plus, une grammaire de constituants (ou de dépendants) extraite de treebank ne permet pas de fournir des informations globales sur la langue, concernant par exemple le type d'ordre utilisé (libre, fixe), qui constituent cependant une information essentielle pour caractériser une langue.

Ces questions se posent de façon cruciale lorsque nous adoptons une perspective typologique : comparer plusieurs langues en comparant les grammaires extraites des treebanks, n'a pas grand sens. En revanche, la typologie s'attache à comparer les langues au travers des phénomènes plus spécifiques qui la caractérisent. Nous trouvons ainsi, pour ce qui concerne la syntaxe, des typologies s'appuyant sur les relations verbes/arguments, sur l'ordre tête/modificateurs, etc.

Nous proposons dans cet article une approche s'appuyant sur une représentation particulière de l'information syntaxique (les Grammaires de Propriétés) et visant à permettre la caractérisation de langues dans une perspective typologique. Cette approche repose sur un ensemble d'outils permettant l'inférence automatique de l'information grammaticale à partir de treebanks et leur utilisation dans une perspective typologique. Nous décrivons tout d'abord la méthode développée en illustrant son application au français et proposons dans un second temps une approche comparative entre une dizaine de langues : tchèque, allemand, anglais, suédois, espagnol, français, italien, finnois, hongrois et irlandais.

1 Les treebanks, sources de l'inférence grammaticale

Nous nous appuyons pour cette étude sur le *Universal Dependencies Treebank* (Nivre, 2015). Il s'agit d'un ensemble de treebanks, utilisant le formalisme des Grammaires de Dépendance, pour 10 langues différentes. La principale caractéristique de cette ressource unique est de s'appuyer sur un même jeu d'étiquettes pour les catégories grammaticales, le *Universal POS Tags*, jeu de 17 étiquettes¹ (Petrov et al., 2012). De même, les relations de dépendance ont été standardisées et sont regroupées en un ensemble commun, le *Universal Dependency Relations*, jeu de 40 étiquettes² (de Marneffe et al., 2014).

La table suivante détaille la taille des treebanks respectifs des différentes langues, en précisant les familles des langues ainsi que leurs principales caractéristiques typologiques. Dans la suite, nous indiquerons par UD_xx le treebank correspondant à la langue xx.

Code	Langue	Famille	Genre	#Arbres	#Tokens	Caractéristiques typologiques
cs	Tchèque	Indo-Européenne	Slave	87.913	1.482.147	SVO ³ , accentuelle, ordre des mots libre
de	Allemand	Indo-Européenne	Germanique	15.918	297.985	V2 et SOV, flexionnelle, accusative, accentuelle, à accent d'intensité
en	Anglais	Indo-Européenne	Germanique	16.622	254.930	SVO, flexionnelle, accusative, accentuelle, à accent d'intensité
sv	Suédois	Indo-Européenne	Germanique	6.026	96.699	SVO, flexionnelle, accusative, accentuelle, à accent de hauteur
es	Espagnol	Indo-Européenne	Romane	16.006	430.764	SVO, syllabique
fr	Français	Indo-Européenne	Romane	16.418	398.964	SVO, flexionnelle, accusative, syllabique
it	Italien	Indo-Européenne	Romane	10.077	214.748	SVO, syllabique
fi	Finnois	Ouralienne	Fenique	13.581	181.022	SVO, ordre des mots libre
hu	Hongrois	Ouralienne	Ougrienne	1.299	25.064	SOV, ordre libre, agglutinante, accusative
ga	Irlandais	Indo-Européenne	Celte	1.020	23.686	VSO, flexionnelle, accusative, accentuelle, à accent d'intensité

Cette ressource, par l'effort de standardisation du jeu d'étiquettes et de relations ainsi que par la couverture multilingue, est unique en son genre. Elle s'avère parfaitement adaptée au projet de comparaison de langues sur la base de propriétés formelles acquises automatiquement.

La caractérisation des propriétés d'une langue de même que la comparaison des caractéristiques syntaxiques de plusieurs langues ne peut se faire en effet directement sur la base de la comparaison des structures syntaxiques. Elle s'avère également complexe à partir de grammaires complètes (qu'il s'agisse de grammaires syntagmatiques ou de grammaires de dépendance). En revanche, il est possible de comparer des propriétés spécifiques, comme il est d'usage en typologie. Par exemple, une typologie classique consiste à étudier les relations verbes/arguments et leur linéarité.

¹ <http://universaldependencies.github.io/docs/u/pos/all.html>

² <http://universaldependencies.github.io/docs/u/dep/all.html>

³ Les notations SVO, SOV ou VSO se réfèrent à l'ordre relatif entre sujet, verbe et objet ; et la notation V2 indique que le verbe est en seconde position.

Nous proposons d'extraire des treebanks ces propriétés élémentaires à partir desquelles il est possible d'établir des comparaisons entre les langues. Ces propriétés sont celles identifiées dans le cadre des *Grammaires de Propriétés*. Notre approche s'intéresse aux types de propriétés de base, telles que définies dans (Blache et al., 2012) :

- **linéarité** : deux composants (A,B) ont une relation de linéarité quand l'ordre d'apparition de ces composants est toujours le même. Nous noterons cette propriété *precede(A,B)* (i.e. A précède toujours B).
- **exigence** : deux composants (A,B) ont une relation d'exigence quand la présence de l'un exige la présence de l'autre. Nous noterons cette propriété *require(A,B)* (i.e. si A est présent, B est aussi présent, soit la formule logique $A \Rightarrow B$).
- **exclusion** : deux composants (A,B) ont une relation d'exclusion quand ils n'apparaissent jamais ensemble. Nous noterons cette propriété *exclude(A,B)* (i.e. si A est présent, B ne l'est pas). Contrairement aux 2 précédentes, cette dernière propriété est symétrique : $exclude(A,B) \Leftrightarrow exclude(B,A)$.
- **unicité** : un composant A répond à une propriété d'unicité s'il n'apparaît jamais plusieurs fois dans la partie droite des règles. Nous noterons cette propriété *unicity(A)*.

L'inférence des propriétés à partir des treebanks, s'appuie sur un processus en deux étapes :

1. Extraction de la grammaire hors-contexte implicite dans les treebanks
2. Génération des propriétés à partir de ces grammaires

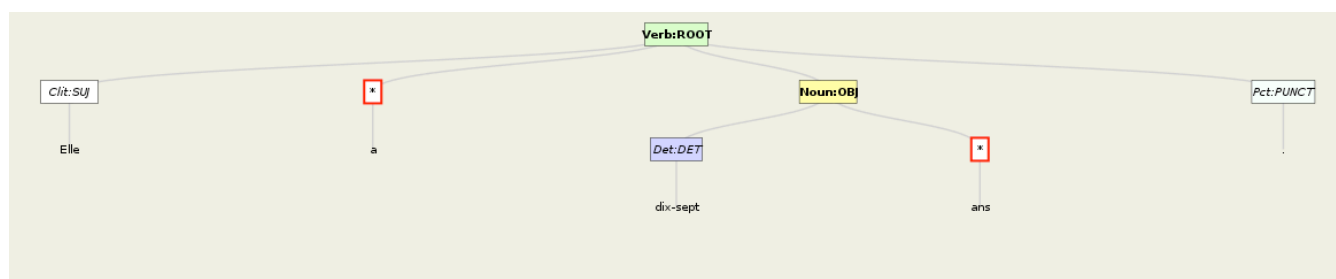
Le choix de s'appuyer sur un formalisme hors-contexte peut sembler paradoxal puisque nous utilisons en entrée des treebanks de dépendance. Il est cependant motivé par le fait qu'il est possible d'utiliser la même méthode pour tout type de treebank, quel que soit leur formalisme. Notre approche permet donc (même si cet aspect n'est pas décrit dans cet article) de traiter en entrée tout type de treebank.

1.1 Inférence de la grammaire CFG implicite

L'extraction d'une grammaire hors-contexte (CFG) à partir d'un treebank, dans le cas de formalismes syntagmatiques, repose sur méthode classique décrite dans (Charniak, 1996) : chaque nœud interne de l'arbre syntaxique correspond à une règle de réécriture dont la partie gauche est l'étiquette de ce nœud, et la partie droite la séquence d'étiquettes de ses descendants. La grammaire implicite est simplement formée de l'ensemble des différentes règles ainsi identifiées.

Dans le cas d'un treebank de dépendances, le principe est similaire : les arbres de dépendance ont une racine correspondant à la catégorie de la tête, celle-ci est indiquée de façon spécifique (par un étoile) parmi les descendants. L'application de la méthodologie décrite pour les constituants consiste à établir des règles dans lesquelles la partie gauche correspond à la tête de la relation de dépendance, la partie droite étant formée par les différents dépendants, dans leur ordre de projection, ainsi que de la projection de la tête indiquée par le symbole “*”.

L'exemple suivant illustre un arbre de dépendance du corpus français ainsi que les deux règles de la CFG implicite qui en ont été extraites :



Elle a dix-sept ans .

Verb:ROOT -> Clit:SUJ * Noun:OBJ Pct:PUNCT
 Noun:OBJ -> Det:DET *

1.2 Inférence des propriétés

À partir des CFG implicites, le calcul des propriétés est relativement aisé : pour chaque tête (partie gauche de règles), nous calculons la liste de ses différents dépendants (dans toutes les parties droites de règle correspondant à cette tête). Il est alors possible, à l'intérieur de chacun de ces sous-ensembles de catégories, d'identifier les propriétés en recherchant les patterns existants.

Concrètement, pour chaque paire de composants distincts (A,B), nous classons les règles de la CFG où A et/ou B apparaissent (i.e. ensemble ou non). Les parties droites de ces règles constituent donc des suites de catégories à partir desquelles il est possible d'inférer une propriété. Par exemple, à partir des règles suivantes extraites de la CFG implicite du UD_fr concernant le sujet nominal, nous pouvons inférer la propriété de linéarité entre le déterminant et le nom :

```
NOUN-nsubj -> DET-det *
NOUN-nsubj -> DET-det * NOUN-nmod      =>      Det < NOUN-nsubj
NOUN-nsubj -> DET-det * ADJ-amod
```

Ce même principe s'applique sur les autres types de propriétés comme l'exigence, l'exclusion ou l'unicité. Le principe consiste donc à prendre en considération, pour les propriétés binaires, tous les couples de catégories et vérifier dans quel ordre ils apparaissent, s'ils sont systématiquement en cooccurrence, ou encore systématiquement séparés. Dans l'exemple précédent, toutes les règles permettaient d'inférer la propriété considérée. Pour chacune des propriétés, les conditions de validation sont simplement exprimées, pour deux catégories A et B données :

- Linéarité : A et B sont réalisées et A apparaît avant B
- Exigence : A et B sont réalisées ensemble
- Exclusion A et B ne sont pas réalisées ensemble
- Unicité : A n'est réalisé qu'une fois

Nous avons développé un environnement permettant, à partir d'un treebank, d'inférer la grammaire CFG implicite et la grammaire de propriétés correspondante, et d'éditer le résultat sous la forme d'un navigateur HTML qui permet de visualiser les règles et les propriétés et d'explorer leurs occurrences dans le treebank. L'exemple suivant montre une page d'exploration dans laquelle la fenêtre se divise en 3 parties :

- Dans le cadre de gauche sont listés les différentes catégories du treebank avec les décomptes des occurrences, nombre de règles CFG associées et propriétés induites. Un lien permet de charger dans la partie principale la page décrivant le composant.
- La zone principale sert à présenter, pour chaque tête, les différents symboles qui le composent, les propriétés calculées et les règles CFG dont il est la partie gauche (dans la figure suivante, seule la description des catégories dépendantes est affichée)
- Le cadre du bas permet la visualisation des arbres de dépendance de la partie correspondante dans le treebank.

822 files, 16418 tree structures, 398964 tokens
20064 rules
2070 properties [CSV]
(all relations [CSV])

Symbols

18 symbols

symbol	nb_rules	properties	occurrences
ADJ	1168	172	22339
ADP	108	138	63695
ADV	150	178	13821
AUX	29	96	8920
CONJ	19	78	10050
DET	17	81	61421
INTJ	32	125	273
NOUN	5940	159	71709
NUM	275	165	9905
PART	7	11	909
PRON	391	155	17696
PROPN	2074	169	31497
PUNCT	9	45	44606
SCONJ	26	59	2898
SYM	91	142	411
VERB	9310	150	35980
X	418	147	2834

Head: VERB : 9310 rules, 30148 occurrences in 14628 trees, 17 distincts symbols

Symbols

17 distincts symbols

page size : 25 show page : 1 Disable Pager

symbol	nb_rules	occurrences	frequency
ADJ	570	768	2.55%
ADP	1304	5865	19.45%
ADV	3248	5840	19.37%
AUX	3788	7894	26.18%
CONJ	1944	2904	9.63%

VERB-root 461:1
VERB-acl 461:4
NOUN-dobj 461:7
DET-det 461:7

PROPN-nsubj Vilgax
ADP-case de
VERB-acl tente
NOUN-dobj récupérer
DET-det le
PUNCT-punct cristal

L'éditeur possède plusieurs options pour affiner le calcul des règles et des propriétés. Il est ainsi possible de choisir le niveau de finesse des symboles considérés, soit en ne conservant que la catégorie grammaticale, soit en utilisant la paire <catégorie, fonction>. Il est également possible de filtrer les règles (en fonction de leur nombre d'occurrences, ou de leur fréquence) pour considérer une propriété. L'outil est diffusé sur le site d'Ortolang/SLDR sous le nom de *MarsaGram*⁴.

1.3 Identifier l'importance des propriétés par leur distribution

Notre hypothèse est qu'il est possible de caractériser les langues en fonction de la répartition des propriétés. Pour cela, nous proposons de prendre en compte les occurrences de chacune d'entre elles ainsi que leur distribution au sein des propriétés caractérisant la construction étudiée. Nous définissons ainsi un certain nombre de critères à partir desquels il sera possible d'analyser et comparer les caractéristiques de différentes langues. Ces critères s'appuient sur la fréquence des occurrences des règles CFG à partir desquelles les propriétés sont inférées. Par exemple, dans le treebank UD_fr, nous aurons les données suivantes concernant la propriété de précédence entre la tête verbale et le complément d'objet (NOUN-dobj) :

precede	* NOUN-dobj	nb_rules	occurrences	frequency	rules
		2	434	77.78%	0 1

Ces données indiquent que la propriété de linéarité est inférée à partir de 2 règles (dont les indices 0 et 1, qui sont des hyperliens, pointent vers les règles VERB-root -> NOUN-nsubj * NOUN-dobj PUNCT-punct et VERB-root -> PRON-nsubj * NOUN-dobj PUNCT-punct). Ces deux règles apparaissent dans le treebank 434 fois, ce qui représente 77,78% des occurrences des règles décrivant VERB-root (elles sont au nombre de 558).

⁴ Accessible grâce à son identifiant pérenne : hdl:11041/ortolang-000917

Il est possible d'extraire plus d'informations de ces données. En particulier, une propriété sera d'autant plus importante qu'elle apparaît systématiquement. Nous avons un moyen direct d'évaluer cette importance : il suffit d'identifier les règles par lesquelles la propriété étudiée est activée, et de voir si cette propriété est vérifiée ou non. Ainsi, dans l'exemple précédent, toutes les règles dans lesquelles apparaissent les catégories `VERB` et `NOUN-dobj` comportent cet ordre linéaire. Nous considérerons donc que la propriété `VERB < NOUN-dobj` est importante et ne doit pas être transgressée. En revanche, l'exemple suivant illustre un cas pour lequel la propriété semble être moins stable. Il s'agit de la relation de précedence entre verbe et nom, tous deux dépendants d'un nom construit comme modifieur adverbial exprimé par la propriété : `VERB-cop < NOUN-nmod`. Cette propriété est validée dans la plupart des règles contenant ces deux catégories comme dans l'exemple suivant :

NOUN-advcl 7150:27							
ADP-mark	PRON-nsubj	VERB-cop	DET-det	*	NOUN-nmod 7150:33		
comme	c'	était	le	cas	ADP-case	DET-det	*
					dans	le	livre

Elle est cependant non satisfaite dans certaines constructions inversant le complément nominal par rapport à la copule, ces exemples étant beaucoup plus rares :

NOUN-advcl 6147:4									
		NOUN-nmod 6147:6							
SCONJ-mark	ADP-case	DET-det	*	ADJ-amod	PRON-nsubj	VERB-cop	ADV-neg	*	ADJ-amod
si	6147:7	la	convention	fiscale	vous	êtes	non	résident	français
	ADP-mwe								
	de	par							

Il est possible de ne conserver que les propriétés vérifiées dans tous les cas (i.e. sans aucune règle contradictoire) ou de relâcher cette contrainte, comme dans l'exemple précédent, pour avoir des propriétés pondérées, obtenant ainsi des propriétés plus *fortes* (fréquentes et vérifiées dans tous les cas) et d'autres plus *faibles* (moins fréquentes ou moins souvent vérifiées). Nous proposons donc de fournir une première indication du poids associé à une propriété en s'appuyant sur cet indice, pouvant être pondéré par la fréquence de la propriété.

Nous notons, pour une propriété p , $Validating(p)$ et $Violating(p)$ l'ensemble des règles validant ou contredisant la propriété. Il est alors possible de calculer un premier poids noté w_0 correspondant au ratio entre le nombres d'occurrences des règles validant p par rapport à l'ensemble des règles liées à la propriété. On note par ailleurs $Occ(p)$ la fonction retournant l'ensemble des occurrences de p , éventuellement contraintes par leurs satisfaction. Nous avons :

$$w_0 = Occ(Validating(p)) / Occ(Validating(p)) + Occ(Violating(p))$$

On peut noter que si les occurrences des règles validant p sont plus nombreuses que celles qui la contredisent, alors w_0 est supérieur à 0.5. Dans le cas où la propriété est toujours vérifiée, alors $w_0=1$. Il est possible d'affiner ce poids en prenant en compte la fréquence des règles. Il s'agit plus précisément de pondérer w_0 par le ratio entre le nombre d'occurrences des règles validant la propriété p et le nombre total d'occurrences des règles de la catégorie tête à laquelle p se réfère.

$$w_1 = w_0 * Occ(Validating(p)) / nb\ total\ d'occurrences$$

Le tableau suivant, extrait du navigateur, illustre cette répartition pour quelques propriétés de la catégorie `NOUN-nsubj` en français. Cette catégorie correspond dans le treebank à 5585 occurrences des 10 règles de la grammaire CFG qui la décrivent. Les deux premières propriétés se retrouvent en tant que pattern dans toutes les règles, avec donc une fréquence de 100%. Aucune règle ne correspondant à une violation, le poids w_0 est donc égal à 1, de même que le poids w_1 . En revanche, ce n'est pas le cas de la propriété d'exclusion entre le déterminant et le verbe modifieur du nom qui se trouve vérifiée dans la plupart des règles du corpus, à l'exception d'une, ayant une faible occurrence. L'indice w_0 , qui a dans ce cas la même valeur que la fréquence, conduit à un poids w_1 de plus faible valeur, indiquant la possibilité de relâchement de la propriété.

Property	1st cat	2nd cat	Freq	w0	w1				
Unicity	DET-det	-	100%	1.0000	1.0000		nb_rules	occ	freq rules

						Validating	10	5585	100.00%	0 1 2 3 4 5 6 7 8 9
Precede	DET-det *		100%	1.0000	1.0000		nb_rules	occ	freq	rules
						Validating	10	5585	100.00%	0 1 2 3 4 5 6 7 8 9
Exclude	DET-det	VERB-acl:relcl	98.57%	0.9857	0.9716		nb_rules	occ	freq	rules
						Validating	9	5505	98.57%	0 1 2 3 4 5 6 7 8
						Violating	1	80	1.43%	9

2 Caractériser les langues

Notre hypothèse est que la distribution des propriétés ainsi que leur importance relative permet d'établir une forme de caractérisation de la langue décrite dans le treebank. Cette caractérisation s'appuie sur un ensemble d'éléments qui, réunis, permettent de donner une image globale des caractéristiques syntaxiques.

Taille de la grammaire : une première indication porte sur le nombre de propriétés qu'il est possible d'identifier, en rapport avec la taille du *tagset*. Il s'agit d'un élément d'information régulièrement utilisé dans la description typologique des langues, notamment dans la perspective de l'étude de leur complexité (Dahl, 2004). Cependant, les propriétés fournissent d'autres types d'information. Par exemple, une langue comportant un très grand nombre de propriétés aura des formes de surface contraintes, avec une variabilité limitée. En effet, les propriétés réduisent par leur application l'espace de recherche définissant les formes possibles. Si les catégories utilisées sont soumises à un grand nombre de contraintes, leur combinatoire s'en trouvera donc limitée. Une conséquence directe sera alors la présence dans la langue de constructions nombreuses, mais peu variables. Le tableau suivant récapitule ces données pour les langues du corpus considéré :

CS	DE	EN	ES	FI	FR	GA	HU	IT	SV
598	683	755	708	523	716	547	448	750	547

Ces données confirment la prédiction émise : les langues considérées comme étant à ordre des mots libre (le tchèque, le finnois et le hongrois) sont décrites par des grammaires de propriétés de taille significativement plus réduite que les autres.

Caractérisation des têtes : toutes les catégories têtes sont caractérisées par l'ensemble des propriétés qui relient leurs dépendants. Il est donc possible pour chaque catégorie d'extraire plusieurs types d'information : la couverture des contraintes (le nombre de catégories qu'elles affectent) et leur importance. De la même façon que pour la langue, ces deux indications permettent de décrire ses réalisations possibles d'une construction. Nous proposons dans ce qui suit une analyse pour chaque type de propriété de l'influence de ces facteurs. Nous définissons pour cela un indice de cohésion qui est une fonction de ces facteurs :

On note C une catégorie, D_c l'ensemble des catégories dépendantes de C , et P_c l'ensemble des contraintes s'appliquant aux catégories de D_c . Nous appelons dans ce qui suit *construction* tout ensemble formé d'une catégorie tête et de ses dépendants.

- Taille (t_c) : c'est la taille du graphe (le nombre de sommets) formé par les contraintes de P_c . Il s'agit en d'autres termes du nombre de catégories de D_c affectées par les contraintes de P_c .

Densité ($Dens_c$) : c'est la densité du graphe formé par les contraintes de P_c . Un graphe dense contient un grand nombre de sommets (ici des catégories) connectés par des arrêtes (des contraintes). Un graphe dense peut être complet : tous les sommets sont connectés entre eux. Le nombre maximal de relation pour un type de propriété binaire donné étant de $t_c * (t_c - 1)$, nous avons donc l'indicateur de densité suivant :

$$Dens_c = nb_prop / t_c * (t_c - 1)$$

L'intuition est qu'un ensemble de dépendants est plus ou moins fortement contraint et donc plus ou moins variable. Si la taille du graphe de contrainte P_c s'approche du nombre total de catégories de D_c , si sa densité est élevée ainsi que la moyenne des poids des contraintes de P_c , alors la réalisation de la construction est très fortement limitée. L'interprétation de cette mesure est différente en fonction des types de contraintes. Par exemple, pour ce qui concerne la

linéarité, une densité indique le nombre de catégories affectées par un ordre. Dans ce cas, la densité est indicatrice de la liberté de l'ordre des mots dans le constituant. Une densité faible indique un ordre des mots libre. À l'opposé, un graphe de contraintes complet à l'ordre près (toutes les catégories sont contraintes 2 à 2), indique un ordre des mots fixe à l'intérieur de la construction.

Le tableau suivant illustre l'application de cette mesure au français et au finnois :

	ADJ	ADP	ADV	AUX	CONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN
FR	0,40	0,33	0,33	0,32	0,53	0,24	0,30	0,86	0,43	0,27	0,24	0,67
FI	0,04	0,30	0,08	0,36	0,42	0,00	0,46	0,06	0,07	0,00	0,06	0,14

	PUNCT	SCONJ	SYM	VERB	X	Moy
FR	0,35	0,75	0,60	0,46	0,40	0,42
FI	0,29	1,00	0,06	0,05	0,15	0,22

Ce tableau indique pour chaque catégorie tête la densité de linéarité calculée selon la formule précédente. Il révèle des différences importantes à la fois entre les catégories elles-mêmes, mais également au niveau global. La densité de linéarité moyenne du français est très supérieure à celle du finnois, prédisant ainsi correctement l'observation typologique selon laquelle le finnois a un ordre des mots libre, ce qui n'est pas le cas du français.

Caractérisation des langues : Nous proposons de synthétiser pour chaque langue les informations concernant toutes les catégories à l'intérieur d'un *ensemble de dépendances* (i.e. l'ensemble des dépendants d'une même tête). Nous regroupons ainsi, pour chaque propriété p affectant une catégorie c d'un tel ensemble (notée p_c) les informations suivantes :

- props : nombre de propriétés p_c
- %props : ratio entre le nombre de propriétés p_c et le nombre total de propriétés de type p pour la tête
- occ : nombre d'occurrences des règles (de la grammaire CFG) validant la propriété p_c
- %occ : ratio entre le nombre d'occurrences des règles validant p_c et le nombre total des règles validant une propriété de type p
- mean(w0) : moyenne des w0 pour p_c
- mean(w1) : moyenne des w1 pour p_c

L'exemple suivant propose une synthèse des informations concernant les propriétés de précédence affectant le déterminant pour le tchèque et l'allemand :

lang	head	prop_type	cat	props	%props	occ	%(occ)	mean(w0)	mean(w1)
cs	NOUN	precede	DET	6	15,38%	41151	7,34%	0,96	0,02
de	NOUN	precede	DET	9	23,08%	73508	53,99%	0,96	0,16

Ce tableau indique que pour le tchèque, parmi les dépendants du nom, le déterminant intervient dans 6 propriétés de linéarité, soit 15,38% des propriétés de linéarité des dépendants du nom. Ces propriétés correspondent à des règles dont la somme des occurrences est 41.151 (7,3% des occurrences des règles liées à une propriété de linéarité). La moyenne des w0 de ces 6 propriétés de linéarité est de 0,96, celle de w1 de 0,02. Une comparaison avec les valeurs comparables pour l'allemand montre des différences significatives : le déterminant a un ordre linéaire beaucoup plus contraint en allemand qu'en tchèque, ce qui est révélé d'une part par le ratio du nombre de propriétés de linéarité concernant le déterminant par rapport à toutes les propriétés de linéarité du nom, mais également (et surtout) par la distribution des occurrences des règles correspondantes (représentant 53,93% des règles validant une linéarité pour le nom).

Le même type de comparaison peut être opéré à un niveau un peu plus général, en prenant en compte simultanément tous les types de propriétés.

lang	head	prop_type	cat	props	%props	occ	%(occ)	mean(w0)	mean(w1)
cs	NOUN	ALL_	DET	22	13,41%	893059	12,26%	0,98	0,13
de	NOUN	ALL_	DET	19	13,97%	393535	31,83%	0,97	0,42

3 Classer les langues

Le projet d'approcher la typologie des langues par classification automatique sur la base de traits morpho-syntaxiques a été explorée dans différentes études, avec des perspectives différentes : traits lexicaux (Enright et al., 2011 ; Barbancon et al., 2007 ; Ellison et al. 2006), interférence langue maternelle/seconde (Nagata et al. 2013), mesures de distance (Batagelj et al., 1992; Kita, 1999). Certaines approches, plus rares, s'appuient sur des informations syntaxiques (Sidorov et al., 2013 ; Abramov et al., 2011). Nous proposons dans cet article d'appliquer des méthodes de classification à partir de paramètres syntaxiques précis, obtenus à grande échelle à partir du UDT.

Il est possible, grâce aux informations produites par les propriétés, de comparer les langues en vue de leur classification. Il s'agit ici de vérifier la pertinence de ces informations, mais également la possibilité d'établir un modèle prédictif de regroupement, voire de typologie. Cette classification repose sur l'identification des propriétés communes à plusieurs langues. Cette opération est rendue possible par le fait que le même jeu d'étiquettes est utilisé pour les différentes langues du *Universal Treebank*, rendant ainsi les propriétés directement comparables. L'hypothèse est que, à la différence des règles syntagmatiques ou de dépendance, les propriétés peuvent à la fois représenter des types d'information très spécifiques, entre deux sous-catégories particulières, mais également être regroupées par type.

Chaque propriété dans notre approche est représentée par un quadruplet $\langle C, tp, A, B \rangle$ où C est le contexte de la propriété (la partie gauche des règles sur laquelle celle-ci est calculée), tp est le type de propriété (*unicity*, *precede*, *require* ou *exclude*) et A et B sont les composants de la propriété. Étant donné que le jeu d'étiquettes auquel appartiennent A , B et C est le même pour toutes les langues, nous pouvons calculer si une propriété $p = \langle C, tp, A, B \rangle$ présente pour une langue l'est également dans une autre.

Le tableau suivant présente les occurrences des propriétés communes entre l'italien et d'autres langues :

properties	lang	cs	de	en	es	fi	fr	ga	hu	it	sv
ALL	it	706	740	1022	972	755	1023	566	505	-	630
precede	it	114	142	192	225	111	186	97	86	-	142

L'italien partage le plus de propriétés avec le français (1.023, dont 186 de linéarité), puis l'anglais (1.022, dont 192 de linéarité) et c'est avec le hongrois qu'il partage le moins de propriétés. Du point de vue de la linéarité, c'est avec l'espagnol qu'il y a le plus d'occurrences de propriétés communes (225).

Nous proposons d'affiner cette estimation par la définition d'une fonction de similarité. Soit $P(lg)$ un ensemble de propriétés calculées pour la langue lg , une mesure de similarité entre deux langues $simil(lg_1, lg_2)$ peut être obtenue par le rapport des propriétés communes aux deux langues relativement à l'ensemble des propriétés de l'une et l'autre de ces langues :

$$simil(lg_1, lg_2) = \text{card}(P(lg_1) \cap P(lg_2)) / \text{card}(P(lg_1) \cup P(lg_2))$$

	#properties	cs	de	en	es	fi	fr	ga	hu	it
cs	1665									
de	1695	0,645								
en	2267	0,656	0,634							
es	2001	0,675	0,562	0,588						
fi	1299	0,724	0,756	0,655	0,701					
fr	2070	0,652	0,621	0,517	0,528	0,662				
ga	1228	0,716	0,649	0,723	0,671	0,780	0,673			
hu	969	0,702	0,707	0,759	0,738	0,732	0,741	0,726		
it	1679	0,687	0,669	0,595	0,581	0,596	0,575	0,696	0,717	
sv	1250	0,626	0,637	0,690	0,663	0,747	0,654	0,652	0,655	0,681

La distance associée compte les propriétés présentes dans seulement une des deux langues (différence symétrique) :

$$\begin{aligned} \text{dist}(lg_1, lg_2) &= \text{card}(P(lg_1) \oplus P(lg_2)) / \text{card}(P(lg_1) \cup P(lg_2)) \\ &= 1 - \text{simil}(lg_1, lg_2). \end{aligned}$$

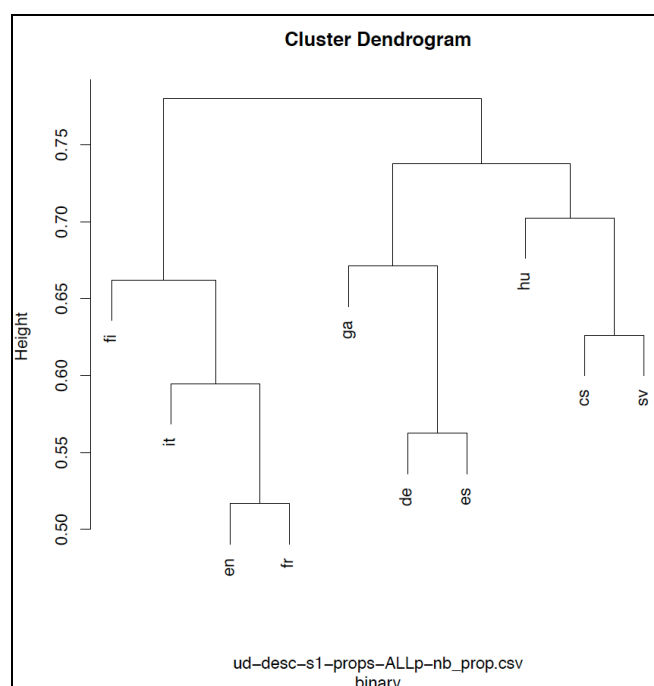
Nous obtenons grâce à ces mesures un calcul de distance entre les langues. La première table indique les distances entre langues sur la base de la prise en compte de toutes les propriétés. On remarque ainsi que l'anglais, l'espagnol et l'italien sont les langues les plus proches du français, ce qui correspond aux attentes, y compris pour l'anglais compte tenu de la situation très particulière de cette langue germanique, mais proche syntaxiquement des langues romanes. Un rapprochement peu attendu mais semblant robuste apparaît également entre l'allemand et l'espagnol.

Le même calcul peut être fait en prenant en compte un seul type de propriété. Nous présentons dans le tableau suivant ces informations calculées pour les propriétés de linéarité. On remarque que dans ce cas, l'espagnol et l'italien sont plus nettement proches du français. De même, un rapprochement entre l'allemand et le suédois peut-être observé.

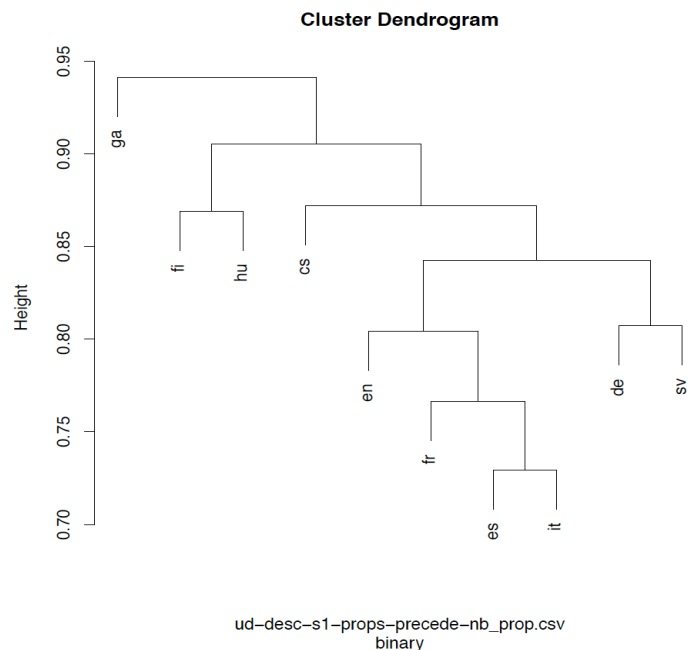
	#properties	cs	de	en	es	fi	fr	ga	hu	it
cs	375									
de	445	0,867								
en	562	0,850	0,811							
es	563	0,856	0,797	0,782						
fi	334	0,885	0,886	0,835	0,862					
fr	489	0,872	0,843	0,804	0,762	0,888				
ga	318	0,907	0,873	0,892	0,878	0,942	0,893			
hu	271	0,882	0,867	0,905	0,882	0,869	0,903	0,929		
it	493	0,849	0,822	0,778	0,729	0,845	0,766	0,864	0,873	
sv	341	0,830	0,807	0,833	0,795	0,895	0,841	0,862	0,860	0,795

À l'aide de ces mesures de distance entre toutes les langues, nous proposons de réaliser un clustering hiérarchique des différentes langues du *Universal Treebank*. Ce traitement a été effectué sous R à l'aide de la fonction *hclust*. Celle-ci réalise (par défaut) un groupement agglomératif à liens complets (*bottom-up, complete-linkage*).

Les figures suivantes présentent les dendrogrammes obtenus. La première figure prend en considération tous les types de propriétés (*unicity, precede, require* et *exclude*), les données étant celles de la table 9.



La classification dans ce cas ne permet pas de dégager des regroupements très nets, quelle que soit la typologie : ordre des mots fixe ou libre, position des arguments par rapport au verbe, etc. Ainsi que nous l'avions remarqué sur la base des données, des rapprochements ponctuels sont cependant significatifs entre le français, l'anglais et l'italien. À noter ici que l'anglais se retrouve très éloigné des autres langues germaniques. L'application du même type de clustering en se limitant aux propriétés de linéarité est en revanche très concluant, comme indiqué dans la figure suivante :



Les classes dégagées correspondent en effet à des attentes typologiques. Nous retrouvons en effet au sein d’une même classe les langues romanes (*es*, *it*, *fr*), desquelles se rapproche l’anglais (comme décrit précédemment). De même, nous retrouvons les langues germaniques au sein d’une même classe (*de*, *sv*). Le finnois et le hongrois se retrouvent également dans une classe, ce qui est également une attente typologique pour les langues finno-ougriennes. L’irlandais enfin, se retrouve dans une position éloignée étant seule représentante de sa classe dans ce corpus.

4 Conclusion

Nous avons présenté dans cet article une méthode ainsi qu’un ensemble d’outils pour acquérir automatiquement des informations en vue d’une description typologique des langues. La méthode présentée consiste à inférer automatiquement à partir de treebanks la grammaire hors-contexte implicite puis d’en extraire les propriétés telles que définies dans les Grammaires de Propriétés. Les outils développés pour cela constituent une plateforme de navigation dans le treebank en même temps qu’un outil de visualisation de données. Cet environnement, de même que les données produites sont disponibles via l’entrepôt de données « *anonyme* ».

Cette première opération permet donc d’inférer automatiquement deux types de grammaires, dans deux formalismes différents. La méthode proposée est générique et peut s’appliquer à des treebanks de dépendance (comme c’est le cas dans cet article), mais également à des treebanks en constituants. En termes de perspectives, nous appliquons actuellement cette méthode à trois grands treebanks en constituants : le Penn, l’Arabic et le Chinese treebank). À terme, nous serons ainsi en mesure d’inclure dans notre étude un très grand nombre de langues, proposant ainsi une technique automatique pour la typologie à grande échelle.

Nous avons montré dans cet article comment, à partir d’une représentation de l’information syntaxique sous la forme de propriétés, il était possible de dégager ou vérifier les grandes propriétés typologiques à partir desquelles nous avons proposé une technique de classification permettant de retrouver automatiquement, sur la base des propriétés de linéarité, les familles de langues sur la base de leur distance. Ce résultat permet de valider la pertinence de la représentation des informations syntaxiques sous la forme de propriétés en vue d’une description typologique. En particulier, plusieurs études portant sur la complexité linguistique ont montré l’intérêt d’une représentation de ce type. Il devient ainsi possible d’envisager le développement d’un outil de comparaison de la complexité des langues, s’appuyant sur des bases formelles.

Par ailleurs, cette méthode présente l’avantage de produire automatiquement des ressources de haut niveau (treebanks hybrides, ajoutant les propriétés syntaxiques explicites aux informations de dépendance et/ou de constituance ainsi que les informations dérivées comme la densité) à partir desquelles de nombreuses applications reposant sur des techniques

d'apprentissage automatique peuvent être appliquées. Ces ressources permettront le développement d'une plateforme d'analyse syntaxique automatique multilingue, reposant sur les Grammaires de Propriétés.

Remerciements

Ce travail réalisé dans le cadre du Labex BLRI (ANR-11-LABX-0036) et de l'Equipex ORTOLANG (ANR-11-EQPX-0032)), ayant ainsi bénéficié d'une aide de l'État au titre du projet A*MIDEX (ANR-11-IDEX-0001-02).

Références

- ABRAMOV O., MEHLER A. (2011) "Automatic Language Classification by means of Syntactic Dependency Networks", in *Journal of Quantitative Linguistics*, 4:291-336
- BATAGELJ V., PISANSKI T., AND KERZIC D. (1992), "Automatic clustering of languages", in *Computational Linguistics*, 18(3):339-352.
- BARBANCON F. WARNOW T., EVANS S., RINGE D., NAKHLEH L. (2007), "An experimental study comparing linguistic phylogenetic reconstruction methods", ms#732, *Department of Statistics, University of California, Berkeley*.
- BLACHE P., RAUZY S. (2012). « Enrichissement du FTB : un treebank hybride constituants/propriétés », in Actes de la conférence *JEP-TALN-RECITAL 2012*, volume 2, 307-320.
- BLACHE P. (2005). Property grammars: A fully constraint-based theory. In H. Christiansen et al., editor, *Constraint Solving and Language Processing*, volume LNAI 3438. Springer.
- CHARNIAK E. (1996). « Tree-bank Grammars ». In proceedings of 13th *National Conference on Artificial Intelligence*, 10311036.
- DAHL O. (2004) *The Growth and Maintenance of Linguistic Complexity*, John Benjamins
- DE MARNEFFE M.-C., DOZAT T., SILVEIRA N., HAVERINEN K., GINTER F., NIVRE J., MANNING C. (2014). Universal Stanford Dependencies: A cross-linguistic typology. In proceedings of 9th *International Conference on Language Resources and Evaluation (LREC'14)*.
- ELLISON M., KIRBY S. (2006), "Measuring Language Divergence by Intra-Lexical Comparison", in *Proceedings of COLING-ACL-2006*
- ENRIGHT J., KONDRAK G. (2011), "The application of chordal graphs to inferring phylogenetic trees of languages", in *Proceedings of 5th IJCNLP*
- FERRER I CANCHO R., SOLE R., KÖHLER R. (2004), "Patterns in syntactic dependency networks" in *Phys. Rev. E*, 69:5
- KITA K. (1999), "Automatic Clustering of Languages Based on Probabilistic Models", in *Journal of Quantitative Linguistics*, 6(2):167-171.
- NIVRE J., BOSCO C., CHOI J., DE MARNEFFE M.-C., DOZAT T., FARKAS R., FOSTER J., GINTER F., GOLDBERG Y., HAJIC J., KANERVA J., LAIPPALA V., LENCI A., LYNN T., MANNING C., McDONALD R., MISSILÄ A., MONTEMAGNI S., PETROV S., PYYSALO S., SILVEIRA N., SIMI M., SMITH A., TSARFATY R., VINCZE V., ZEMAN D. (2015). Universal Dependencies 1.0. <http://hdl.handle.net/11234/1-1464>.
- PETROV S., DAS D., McDONALD R. (2012). A Universal Part-of-Speech Tagset. In proceedings of the 8th *International Conference on Language Resources and Evaluation (LREC'12)*.
- Nagata R. and Whittaker E. (2013), "Reconstructing an Indo-European Family Tree from Non-native English Texts", in *Proceedings of the 51st Annual Meeting of the ACL*
- RAMA, T., SINGH K. (2009), "From Bag of Languages to Family Trees From Noisy Corpus", in *Proceedings of RANLP-2009*
- SIDOROV G., VELASQUEZ F., STAMATATOS E., GELBUKH A., CHANONA-HERNANDEZ L. (2013) "Syntactic Dependency-Based N-grams as Classification Features", in *Advances in Computational Intelligence*, LNCS-7630, Springer
- SINGH K. AND SURANA H. (2007), "Can Corpus Based Measures be Used for Comparative Study of Languages?", in *Proceedings of 9th Meeting of the ACL SIG in Computational Morphology and Phonology*

Utilisation de mesures de confiance pour améliorer le décodage en traduction de parole

Laurent Besacier¹ Benjamin Lecouteux¹ Ngoc Luong Quang¹

(1) LIG, Univ. Grenoble-Alpes, France

laurent.besacier@imag.fr, benjamin.lecouteux@imag.fr, quangngocluong@gmail.com

Résumé. Les mesures de confiance au niveau mot (*Word Confidence Estimation* - WCE) pour la traduction automatique (TA) ou pour la reconnaissance automatique de la parole (RAP) attribuent un score de confiance à chaque mot dans une hypothèse de transcription ou de traduction. Dans le passé, l'estimation de ces mesures a le plus souvent été traitée séparément dans des contextes RAP ou TA. Nous proposons ici une estimation conjointe de la confiance associée à un mot dans une hypothèse de traduction automatique de la parole (TAP). Cette estimation fait appel à des paramètres issus aussi bien des systèmes de transcription de la parole (RAP) que des systèmes de traduction automatique (TA). En plus de la construction de ces estimateurs de confiance robustes pour la TAP, nous utilisons les informations de confiance pour re-décoder nos graphes d'hypothèses de traduction. Les expérimentations réalisées montrent que l'utilisation de ces mesures de confiance au cours d'une seconde passe de décodage permettent d'obtenir une amélioration significative des performances de traduction (évaluées avec la métrique BLEU - gains de deux points par rapport à notre système de traduction de parole de référence). Ces expériences sont faites pour une tâche de TAP (français-anglais) pour laquelle un corpus a été spécialement conçu (ce corpus, mis à la disposition de la communauté TALN, est aussi décrit en détail dans l'article).

Abstract.

Word confidence estimation for re-decoding speech translation graphs

Word Confidence Estimation (WCE) for machine translation (MT) or automatic speech recognition (ASR) assigns a confidence score to each word in the MT or ASR hypothesis. In the past, this task has been treated separately in ASR or MT contexts and we propose here a joint estimation of word confidence for a spoken language translation (SLT) task involving both ASR and MT. We build robust word confidence estimators for SLT, based on joint ASR and MT features. Using these word confidence measures to re-decode the spoken language translation graph leads to a significant BLEU improvement (2 points) compared to the SLT baseline. These experiments are done for a French-English SLT task for which a corpus was specifically designed (this corpus being made available to the NLP community).

Mots-clés : Mesures de confiance, traduction automatique de la parole, paramètres joints, re-décodage de graphe.

Keywords: Word confidence estimation (WCE), spoken language translation (SLT), joint features, search graph re-decoding.

1 Introduction

L'estimation de mesures de confiance est un sujet important en reconnaissance automatique de la parole (RAP) ainsi qu'en traduction automatique (TA). En effet, tandis que ces systèmes produisent des sorties dont la qualité ne cesse de croître, une correction (ou post-édition) humaine de ces sorties est le plus souvent nécessaire pour produire des transcriptions ou des traductions parfaites. Ainsi, ces estimateurs de confiance nous permettent de répondre à des questions telles que : est-ce que ces transcriptions / traductions peuvent être publiées telles quelles ? La qualité est-elle suffisante pour qu'elles soient corrigées ou mieux vaut-il les retranscrire (re-traduire) à partir de zéro ? Quelles parties de la transcription / traduction doivent être corrigées en priorité ? Pour répondre à ces questions, il est nécessaire de construire un système automatique qui détecte les erreurs sur des segments d'une sortie de TA ou de RAP ; un tel système est appelé estimateur de confiance et génère des mesures de confiance au niveau de chaque segment de texte. Lorsque les segments considérés sont simplement les mots, on parle alors de *mesures de confiance au niveau des mots* (*Word Confidence Estimation* - WCE - en anglais). En plus d'être utiles pour des scénarios interactifs où l'humain participe à la tâche, les mesures de confiance permettent également de re-ordonner des hypothèses de type N-meilleures (Luong *et al.*, 2014b) ou de re-décoder un graphe de

recherche (Luong *et al.*, 2014a) en améliorant les performances.

Un estimateur de confiance au niveau des mots (WCE) assigne donc à chaque mot de l'hypothèse un score (typiquement entre 0 et 1). Plus spécifiquement, la détection d'erreurs consiste à seuiller ce score puis à étiqueter chaque mot comme correct ou incorrect. Pour cela, un système d'étiquetage séquentiel, entraîné sur un ensemble de paramètres, apprend à prédire des labels de type correct (*Good*) ou incorrect (*Bad*). Dans le passé, l'estimation de ces mesures a le plus souvent été traitée séparément dans des contextes RAP ou TA. Nous proposons ici une estimation conjointe de la confiance associée à un mot dans une hypothèse issue d'un système de traduction automatique de la parole (TAP). Cette estimation fait appel à des paramètres issus aussi bien des modules de transcription de la parole (RAP) que des modules de traduction automatique (TA), tous deux nécessaires à une tâche de TAP.

Cet article en français présente des résultats nouveaux (re-décodage d'un graphe de traduction de parole avec mesures de confiance) mais il s'appuie néanmoins sur deux publications récentes, en anglais, des mêmes auteurs qu'il convient ici de mentionner :

- une publication à IWSLT 2014 (Besacier *et al.*, 2014) qui présente en détail le corpus sur lequel s'appuie cette étude ; il nous semblait important de le présenter à la communauté TALN francophone et ce corpus est donc décrit à nouveau ici (25% de cette soumission), mais avec un peu moins de détails que dans l'article original en anglais,

- une publication à EAMT 2014 (Luong *et al.*, 2014a) qui présente un algorithme de re-décodage de graphes d'hypothèses de traduction automatique ; cet article ne concernait qu'une tâche de traduction de texte et nous reprenons ici l'algorithme (décrit de façon plus succincte - 25% de cette soumission) en l'adaptant à une tâche de traduction de parole.

Le reste de l'article décrit un résultat original et non encore publié ailleurs qui montre que l'utilisation de mesures de confiance jointes RAP+TA permet d'améliorer significativement les performances d'un système de traduction de la parole.

Cet article est organisé de la façon suivante : la section 2 résume rapidement les principaux travaux antérieurs concernant les estimateurs de confiance au niveau des mots (WCE). Les approches sont présentées séparément entre les tâches de RAP et de TA puisque, à notre connaissance, seule notre précédente publication (Besacier *et al.*, 2014) propose une estimation jointe. Ensuite, le corpus utilisé pour la partie expérimentale est décrit dans la section 3. Les parties 4 et 5 présentent nos systèmes de WCE pour des tâches de transcription et de traduction, respectivement. La section 6 présente, quant à elle, des résultats originaux et montre comment les estimateurs de confiance améliorent les performances sur une tâche de traduction de parole. Pour finir, nous concluons ce travail et donnons quelques perspectives dans la dernière partie.

2 Rapide aperçu des mesures de confiance pour la TA et la RAP

De nombreux travaux ont proposé d'estimer des mesures de confiance afin de détecter automatiquement les erreurs en sortie des systèmes de RAP. Dans ce domaine, ces mesures ont tout d'abord été introduites pour la détection des mots hors-vocabulaire (Asadi *et al.*, 1990). Ces travaux ont ensuite été exploités par (Young, 1994) qui a alors introduit l'utilisation des probabilités *a posteriori* comme mesures de confiance pour la RAP. Ces dernières sont estimées, le plus souvent, en utilisant le graphe (ou treillis) issu de la transcription automatique (Kemp & Schaaf, 1997). Plus récemment, d'autres paramètres sont venus enrichir les probabilités *a posteriori* (Lecouteux *et al.*, 2009) : nombre d'hypothèses concurrentes à un instant donné, paramètres linguistiques ou acoustiques (stabilité du signal, durée des phonèmes, etc.) ainsi que des paramètres sémantiques. Une liste exhaustive des différents paramètres qui peuvent être utilisés est présentée dans (Chase, 1997). Ces différents paramètres peuvent alors être classés selon diverses méthodes : des réseaux Bayésiens naïfs (Sanchis *et al.*, 2012), des *Support Vector Machines* (Zhang & Rudnicky, 2001), des réseaux de neurones (Weintraub *et al.*, 1997). Plus récemment, dans (Seigel *et al.*, 2011) et (Seigel & Woodland, 2012) les auteurs combinent les différents paramètres en utilisant des champs aléatoires conditionnels ((Conditionnal Random Fields (CRF) (Lafferty *et al.*, 2001)).

Par ailleurs, l'atelier WMT (*Workshop on Machine Translation*) a introduit en 2013 une nouvelle tâche d'évaluation dédiée aux mesures de confiance appliquées aux systèmes de TA. (Han *et al.*, 2013) (Luong *et al.*, 2013b) proposent d'utiliser des CRFs pour aborder le problème comme un étiquetage de séquence. En parallèle, (Bicici, 2013) a proposé un modèle permettant d'estimer la similarité sémantique entre phrases cibles et sources. Leur modèle, basé sur l'apprentissage global (*Global Learning Model*) est indépendant du moteur de traduction et utilise des paramètres liés à des informations syntaxiques, de contexte ou de forme. (Han *et al.*, 2013) proposent de s'attacher essentiellement aux combinaisons de n-grammes dans la langue cible. Finalement, dans les travaux de (Luong *et al.*, 2013b), l'ensemble des paramètres présentés précédemment sont intégrés en rajoutant des informations sur la topologie du graphe d'exploration, sur des pseudo-références ou encore des éléments liés à la polysémie et la complexité syntaxique.

A notre connaissance, les premiers travaux proposant des mesures de confiance pour la traduction orale utilisant des paramètres joints entre RAP et TA sont ceux présentés à IWSLT 2014 (Besacier *et al.*, 2014).

3 Notre corpus pour la construction d'estimateurs de confiance en traduction de la parole

Ce travail s'appuie sur un corpus, construit par nos soins, disponible en ligne¹ et déjà décrit dans (Besacier *et al.*, 2014). Nous en présentons les principales traits ci-dessous. Le corpus présente 2643 phrases prononcées en français (3 locuteurs * 881 phrases différentes - 5h au total environ - lectures de phrases de corpus journalistiques), et traduites vers l'anglais. Plus précisément, pour chaque phrase, un quintuplet est disponible : la sortie de transcription (*src-asr*), la transcription de référence ou verbatim (*src-ref*), la traduction automatique de ce verbatim (*tgt-mt*), la traduction automatique de la transcription automatique - c'est à dire la sortie d'un système de traduction de parole (*tgt-slt*) et la post-édition de la traduction (*tgt-pe*).

Concernant les systèmes automatiques, le système de RAP est construit à partir de la boîte à outils KALDI (Povey *et al.*, 2011). Le modèle de langue de type 3-gramme est appris à partir des corpus ESTER (Galliano *et al.*, 2006) et Gigaword français (taille du vocabulaire = 55k mots). Les modèles acoustiques, de type SGMM, sont appris sur le même corpus ESTER. Par ailleurs, un post-traitement est nécessaire sur les sorties de transcription pour les rendre compatibles avec le système de traduction automatique (conversion des nombres, restauration de la casse et de la ponctuation, etc.). Le système de TA français-anglais est construit à partir de la boîte à outils MOSES (Koehn *et al.*, 2007). C'est un système statistique à base de fragments (*statistical phrase-based*) et il est décrit plus en détails dans (Potet *et al.*, 2010).

Il est important ici de souligner que les post-éditions des sorties de traduction (*tgt-pe*) sont antérieures à l'enregistrement du corpus oral. Ainsi, le point de départ était un corpus de post-éditions, publié en 2012 dans (Potet *et al.*, 2012), puis les enregistrements à partir des phrases sources en français ont été réalisés. Les transcriptions automatiques issues du système de RAP ont donc été traduites par le système décrit ci-dessus pour obtenir *tgt-slt*. L'hypothèse forte, faite ici, est que nous avons supposé que les post-éditions (réellement obtenues à partir de *tgt-mt*) seraient aussi valables pour *tgt-slt*.

Référence	The	consequence	of	the	fundamentalist	movement		also	has	its importance	.
	T	S	T	T	S	Y	I	T	D	P	.
Hyp après Shift	The	result	of	the	hard-line	trend	is	also		important	.

TABLE 1 – Exemple de labels obtenus avec TERp-A.

L'étiquetage des sorties au niveau des mots (correct / incorrect) est réalisé avec l'outil TERp-A (Snover *et al.*, 2008). Comme illustré sur la table 1, chaque mot de l'hypothèse de traduction est aligné avec un mot ou un fragment de la post-édition selon différents types d'édition : "I" (insertions), "S" (substitutions), "T" (correspondance au niveau du lemme), "Y" (synonyme), "P" (substitution d'un segment) et "E" (identique). Ensuite, nous étiquetons l'hypothèse en groupant les labels E, T et Y selon la catégorie *Good* (G), tandis que les labels S, P et I correspondent à l'étiquette *Bad* (B).

Les principales statistiques sur ce corpus sont présentées dans la table 2, où nous montrons comment les étiquettes de confiance (G/B) sont obtenues. Pour l'ensemble de test, nous disposons donc de jeux de données pour construire des estimateurs de confiance pour trois tâches : RAP, TA et TAP.

- **RAP** : extraire les étiquettes G/B en calculant le taux d'erreur mots - WER - entre *src-asr* et *src-ref*,
- **TA** : extraire les étiquettes G/B en calculant le TERp-A entre *tgt-mt* et *tgt-pe*,
- **TAP** : extraire les étiquettes G/B en calculant le TERp-A entre *tgt-slt* et *tgt-pe*.

La table 3 donne un exemple de quintuplet disponible dans notre corpus. Une des transcriptions (*src-asr1*) a une erreur tandis que l'autre transcription (*src-asr2*) en a 4. Ceci donne lieu à, respectivement 2 et 4 étiquettes de type B pour (*tgt-slt1*) et (*tgt-slt2*) dans la sortie de TAP, alors que la sortie de TA *tgt-mt* ne présente qu'une seule étiquette de type B (incorrect).

Enfin, la table 4 résume les performances de nos systèmes de TA (traduction des références de transcription) et de TAP (traduction des sorties de transcription) obtenues sur notre corpus et évaluées en utilisant les post-éditions comme références. Nous donnons également la distribution des étiquettes *correct* (G) et *incorrect* (B) obtenues pour les deux tâches

1. <https://github.com/besacier/WCE-SLT-LIG>

Jeu de données	# train	# test	Méthode pour obtenir les étiquettes G/B
<i>src-ref</i> <i>src-asr</i>	10000	881 881*3	$wer(src-asr, src-ref)$
<i>tgt-mt</i> <i>tgt-slt</i> <i>tgt-pe</i>	10000 10000	881 881*3 881	$terpa(tgt-mt, tgt-pe)$ $terpa(tgt-slt, tgt-pe)$

TABLE 2 – Vue d’ensemble du corpus

<i>src-ref</i>	quand	notre	cerveau	chauffe
<i>src-asr1</i>	<i>comme</i>	notre	cerveau	chauffe
labels RAP	B	G	G	G
<i>src-asr2</i>	<i>qu’</i>	<i>entre</i>	<i>serbes</i>	<i>au chauffe</i>
labels RAP	B	B	B	B G
TA	when	our	brains	<i>chauffe</i>
labels TA	G	G	G	B
<i>tgt-slt1</i>	<i>as</i>	our	brains	<i>chauffe</i>
labels TAP	B	G	G	B
<i>tgt-slt2</i>	<i>between</i>	<i>serbs</i>	<i>in</i>	<i>chauffe</i>
labels TAP	B	B	B	B
<i>tgt-pe</i>	when	our	brain	heats up

TABLE 3 – Exemple de quintuplet avec étiquettes associées

(ces étiquettes seront considérées comme notre "vérité terrain" lorsque nous évaluerons la performance de nos estimateurs de confiance). Logiquement, le pourcentage d’étiquettes de type (B) augmente lorsqu’on passe d’une tâche de TA à une tâche de TAP. Ce corpus est téléchargeable en ligne sur *github* (le lien exact sera donné dans la version finale de cet article - si celui-ci est accepté).

4 Mesures de confiance pour la transcription de parole

Dans nos travaux, nous proposons d’extraire un certain nombre de traits issus du graphe du système de reconnaissance de la parole. Ces traits sont principalement issus des scores du modèle de langage et d’une analyse morphosyntaxique. Les traits utilisés sont les suivants (plus de détails sont donnés dans (Besacier *et al.*, 2014)) :

- Acoustiques : durée du mot (F-dur).
- Graphiques (extraits du réseau de confusion de la phrase courante) : nombre de chemins alternatifs (F-alt) entre deux noeuds et la probabilité *a posteriori* (F-post).
- Linguistiques (basés sur les probabilités du modèle de langue) : l’unigramme (F-word), probabilité du 3-gramme (F-3g) et utilisation du modèle de repli (F-back) telle que proposée dans (Fayolle *et al.*, 2010),
- Morpho-syntaxiques : les étiquettes morpho-syntaxiques (*Part-Of-Speech*) liées au mot (F-POS).

Nous utilisons un algorithme basé sur le *boosting* afin de combiner les différents traits. Le classifieur utilisé est *bonzaiboost* (Laurent *et al.*, 2014). Sa particularité est d’implémenter un algorithme de *boosting* (Adaboost.MH) sur des arbres de décision.

Pour chaque mot nous estimons les 7 traits (F-Word ; F-3g ; F-back ; F-alt ; F-post ; F-dur ; F-POS) décrits et l’apprentissage de l’estimateur est réalisé sur un corpus séparé mais de nature proche (parole lue - BREF 120 (Lamel *et al.*, 1991)). On notera que la préparation de ce corpus d’apprentissage aura nécessité le décodage de la totalité des signaux du corpus BREF 120, pour obtenir les transcriptions automatiques et leurs séquences d’étiquettes G/B associées.

5 Mesure de confiance pour la traduction automatique

Nous utilisons les CRFs (Conditional Random Fields (Lafferty *et al.*, 2001)) comme méthode d’apprentissage. En effet, la tâche est vue ici comme un étiquetage séquentiel d’une séquence de mots (avec labels G/B). Plus précisément, une implémentation du LIMSIS nommée WAPITI (Lavergne *et al.*, 2010), est utilisée pour entraîner notre estimateur de confiance. Les 10000 phrases du corpus d’apprentissage présenté dans la table 2 sont utilisées pour apprendre les modèles (corpus issu de (Potet *et al.*, 2012)).

La raison pour laquelle les méthodes d’apprentissage sont différentes pour les mesures de confiance en RAP (*boosting*) et

Tâche	RAP (WER)	TA (BLEU)	% G (cor- rect)	% B (in- correct)
TA	0%	52.8%	82.5%	17.5%
TAP	26.6%	30.6%	65.5%	34.5%

TABLE 4 – Performances de traduction de texte (TA) et de parole (TAP) sur notre corpus (2643 phrases)

en TA (CRFs) sont liées à la pre-existence de systèmes dans l'équipe avant ce travail. Une perspective est bien sûr d'avoir une approche unifiée pour ces estimateurs de confiance, en utilisant les CRFs par exemple.

Brièvement, un CRF permet de calculer la probabilité d'une séquence d'étiquettes $Y = (y_1, y_2, \dots, y_N)$ étant donnée une séquence de mots en entrée $X = (x_1, x_2, \dots, x_N)$ par :

$$p_\theta(Y|X) = \frac{1}{Z_\theta(X)} \exp \left\{ \sum_{k=1}^K \theta_k F_k(X, Y) \right\} \quad (1)$$

où $F_k(X, Y) = \sum_{t=1}^T f_k(y_{t-1}, y_t, x_t)$; $\{f_k\}$ ($k = \overline{1, K}$) est un ensemble de fonctions caractéristiques ; $\{\theta_k\}$ ($k = \overline{1, K}$) sont les paramètres du modèle ; et $Z_\theta(x)$ est une fonction de normalisation.

Concernant les traits extraits pour apprendre ces modèles, de multiples sources peuvent être utilisées. Nous utilisons plusieurs dizaines de traits qui sont décrits en détail dans (Luong *et al.*, 2013a). Ils sont présentés ici seulement brièvement ; il est cependant important de préciser qu'un vecteur de traits est extrait pour chaque mot de l'hypothèse de traduction en langue cible :

- Mots ou séquences de mots en langue cible : mot cible ; séquence bigramme (et trigramme) précédent le mot cible considéré,
- Mots en langue source : mots en langue source alignés avec le mot en langue cible considéré,
- Contexte de l'alignement (Bach *et al.*, 2011) : mots en langue source qui entourent (pour une fenêtre ± 2) le mot source aligné avec le mot en langue cible considéré,
- Probabilité *a posteriori* du mot cible considéré (WPP - *Word Posterior Probability* (Ueffing *et al.*, 2003)),
- Pseudo-référence (Google Translate) : en considérant un système de TA en ligne comme une pseudo référence, nous obtenons un trait binaire indiquant si chaque mot de la phrase cible se trouve (ou pas) dans cette (pseudo-) référence,
- Topologie du graphe d'hypothèse de TA (Luong *et al.*, 2013a) : à partir d'une listes des meilleurs hypothèses (*N-best list* en anglais) rassemblée dans un réseau de confusion, les traits suivants sont extraits pour chaque mot de l'hypothèse de TA : nombre de chemins alternatifs d'un noeud à l'autre du graphe (en d'autres termes, nombre de mots qui sont des alternatives au mot considéré), valeur minimale et maximale des probabilités *a posteriori* parmi les alternatives au mot considéré,
- Traits issus du modèle de langue : nous construisons tout d'abord deux modèles de langue 4-gramme pour les deux cotés (source et cible) ; ensuite, nous comptons la longueur la plus grande possible du n-gramme couvert par le mot courant et les mots précédents dans le modèle de langue du côté cible ainsi que du côté source (en utilisant les informations d'alignement) ; par exemple, avec le mot cible w_i : si la séquence $w_{i-2}w_{i-1}w_i$ existe dans le modèle de langue cible tandis que la séquence $w_{i-3}w_{i-2}w_{i-1}w_i$ n'existe pas, la valeur du trait calculée pour w_i est 3,
- Traits lexicaux : étiquette morpho-syntaxique du mot considéré (POS) ; étiquette du mot source aligné avec le mot considéré, séquences des POS en langue cible, traits binaires indiquant si on est en présence d'une marque de ponctuation, d'un nom propre ou d'une valeur numérique,
- Traits syntaxiques : obtenus à partir de l'arbre d'analyse de la phrase cible (nous utilisons ici la sortie de l'outil *Link Grammar Parser*) et calculons pour chaque mot : le label grammatical, la distance à la racine, et un trait binaire indiquant si un mot n'a pas de dépendance (*null link* (Xiong *et al.*, 2010)),
- Traits sémantiques : nombre de sens dans *WordNet* côté cible.

Un ensemble de traits similaires a été utilisé pour construire un estimateur de confiance pour un système de traduction anglais vers espagnol. L'estimateur, soumis à la campagne d'évaluation WMT 2013 (*Quality Estimation shared task*), s'est classé premier selon la métrique proposée par les organisateurs (Luong *et al.*, 2013a).

Tâche	Est. conf. RAP	Est. conf. TA	Est. conf. TAP	Est. conf. TAP	Est. conf. TAP
Traits issus de	RAP	TA	TA	RAP	RAP+TA
$F(G)$	87.85%	87.65%	77.17%	76.41%	77.54%
$F(B)$	37.28%	42.29%	39.34%	38.00%	43.96%

TABLE 5 – Résumé des performances de nos estimateurs de confiance (différentes tâches, différents traits utilisés)

6 Re-décodage d'un graphe de traduction de parole à l'aide de mesures de confiance

6.1 Estimateurs de confiance fusionnés RAP+TA

La table 5 présente tout d'abord les résultats obtenus par nos estimateurs de confiance sur des tâches de RAP et de TA (2 premières colonnes). Pour la traduction automatique (TA), l'estimateur peut être considéré comme "à l'état de l'art" puisqu'il présente des performances similaires à celle obtenues (même si c'est pour un autre couple de langues) au cours des campagnes d'évaluation WMT 2013 et 2014 auxquelles nous avons participé. Concernant la transcription de parole (RAP), il est plus difficile de se comparer car il n'existe pas, à notre connaissance, de corpus *étalon* sur lequel il est possible d'évaluer notre approche. Cependant, nous atteignons des résultats acceptables qui sont bien au delà d'une décision "au hasard". Les trois dernières colonnes de la table sont quant à elles directement comparables entre elles puisqu'elles rapportent des performances évaluées selon une même vérité terrain correspondant à des étiquettes de confiance pour une tâche de traduction de parole (TAP - 65.5% d'étiquettes G et 34.5% d'étiquettes B). Plus précisément, ces trois colonnes correspondent aux résultats des estimateurs suivants :

- Le premier système (estimateur de confiance TAP / traits issus de la TA) est celui décrit dans la section 5 et n'utilise que des traits issus du système de traduction automatique (TA). La seule différence est que, en entrée, la phrase source provient de la sortie de RAP (*src-asr*) plutôt que d'être une phrase texte sans erreur (*src-ref*).
- Le second système (estimateur de confiance TAP / traits issus de la RAP) est celui décrit dans la section 4 et n'utilise que des traits issus du système de transcription automatique (RAP). Ceci revient donc à prédire la confiance d'une sortie de traduction de parole en n'utilisant que des informations issues du module de transcription. L'alignement en mots, obtenu grâce au décodeur *moses* entre *src-asr* et *tgt-slt* est utilisé pour projeter les scores de confiance issus du système de RAP - qui sont donc sur la langue source - vers la sortie de traduction en langue cible.
- Le troisième système (estimateur de confiance TAP / traits issus de RAP+TA) combine les informations issues des deux estimateurs de confiance utilisés précédemment. Dans cette expérimentation, le score issu de l'estimateur RAP est projeté sur chaque mot cible comme pour le second système (présenté dans l'item précédent) puis combiné linéairement (simple moyenne pour cet expérimentation) avec le score issu de l'estimateur TA ($0.5score(MT) + 0.5score(ASR)$). Il est important de noter que les estimateurs de confiance ne sont pas re-entraînés ici, puisque nous réalisons une "fusion tardive" des scores de confiance issus des systèmes de RAP et de TA. Une perspective de ce travail consistera à entraîner un nouvel estimateur à partir de traits joints ASR+MT (en concaténant simplement les vecteurs de traits, par exemple).

Les résultats de ces trois systèmes sont donnés dans les trois dernières colonnes de la table 5. On voit clairement que la fusion d'informations RAP+TA permet d'améliorer les performances de l'estimateur de confiance pour une tâche de TAP². En effet, la performance (F-mesure) pour l'étiquette "B" passe de 39.34% (traits TA seulement) et 38% (traits RAP seulement) à 43.96% (traits RAP+TA fusionnées), tout en conservant un score similaire pour l'étiquette "G". Il est aussi intéressant de remarquer que l'utilisation des traits issus de la transcription seule donnent des performances très intéressantes. On peut d'ailleurs s'interroger sur les gains finalement limités obtenus en combinant des traits qui, séparément, donnent toutes les deux des résultats honorables. Une explication possible est que la fusion tardive proposée ici n'est sans doute pas la meilleure solution car l'observation des scores "G" et "B" obtenus par chacune des méthodes fait apparaître des distributions biaisées vers le label "G"; une normalisation de ces scores avant combinaison serait nécessaire, ainsi que des stratégies de fusion plus avancées qu'une simple moyenne des scores (arbres de décision par exemple). Il semble aussi que l'estimateur de confiance utilisant des traits uniquement TA (et appris sur des données dont la répartition des labels G/B est plutôt 80%/20%) n'est pas bien adapté aux données issues de TAP dont la répartition de labels G/B est plus équilibrée. Un re-entraînement et une optimisation de l'estimateur de confiance RAP+TA seraient donc une tâche à court terme importante.

2. Tous les résultats sont donnés avec un seuil de décision sur les scores p(G) et p(B) fixé à 0.7 - c'est à dire que l'étiquette est fixée à G si p(G)>0.7 et l'étiquette est B sinon - ce seuil de 0.7 est fixé empiriquement et permet de favoriser la détection d'erreurs

6.2 Re-décodage d'un graphe de traduction de parole

6.2.1 Quelques travaux antérieurs sur le sujet

Plusieurs travaux ont proposé une "seconde passe" de post-édition (Parton *et al.*, 2012) ou de re-ordonnancement des N-meilleures hypothèses (Duh & Kirchhoff, 2008; Bach *et al.*, 2011; Zhang *et al.*, 2006).

Concernant le re-décodage de graphes, (Zens & Ney, 2006) proposent un système de traduction en 2 passes qui utilise, au cours de la seconde passe, des paramètres de longueur de phrase et de probabilité *a posteriori* de séquences de mots. Les expérimentations sur un système de traduction Mandarin-Anglais (tâche NIST) montrent une amélioration significative des performances. Par ailleurs, (Tromble *et al.*, 2008) propose de re-décoder le graphe de traduction en trouvant l'hypothèse candidate qui correspond à la minimisation du risque de Bayes (décodage MBR - *Minimum Bayes Risk*). Les résultats expérimentaux sur des tâches de traduction Arabe-Anglais, Mandarin-Anglais et Anglais-Mandarin montrent que l'approche par décodage MBR de graphes surpasse le ré-ordonnancement d'hypothèses (fondé sur le même critère MBR). De son côté, (Venugopal *et al.*, 2007) utilise un modèle de langue (utilisant un historique de taille importante) pour re-décoder le graphe d'hypothèses de traduction d'un système fondé sur une grammaire probabiliste hors contexte.

Notre approche, qui consiste à utiliser des informations externes, rassemblées via un estimateur de confiance, au cours d'une seconde passe de traduction, est présentée dans la section suivante. Elle peut être comparée à celle de (Zens & Ney, 2006) mais avec un nombre de traits (ayant conduit à l'estimation de confiance) beaucoup plus important.

6.2.2 Notre approche

Maintenant que nous avons des estimateurs de confiance au niveau mot, pour une tâche de TAP, nous allons les intégrer dans une seconde passe de décodage pour la traduction. La technique proposée peut se résumer comme suit : les chemins dans le graphe de recherche passant par des mots étiquetés comme incorrects (B) seront pénalisés, tandis que des chemins passant par des mots étiquetés comme corrects (G) seront récompensés. Une fois ce principe énoncé, il convient cependant de préciser que nos estimateurs de confiance ne sont pas capables d'étiqueter directement un graphe (dans leur état actuel, ils doivent être appliqués sur des chaînes de mots). Ainsi, pour couvrir un maximum de mots présents dans le graphe de recherche, nos estimateurs de confiance sont appliqués sur une liste de N-meilleures hypothèses de traduction afin d'étiqueter un maximum de mots différents. Ensuite, pour chaque mot différent rencontré dans la liste des N-meilleures hypothèses, nous mettons à jour les scores des hypothèses du graphe de recherche qui contiennent ces mots. Enfin, nous recherchons à nouveau le meilleur chemin dans le graphe de recherche pour trouver la nouvelle traduction considérée comme "la meilleure". Cette approche est décrite en détail dans (Luong *et al.*, 2014a) où elle est appliquée sur une tâche de traduction de texte uniquement. Dans ce même article, plusieurs façons de mettre à jour les scores dans le graphe de recherche sont aussi présentées. Nous ne décrivons ici que la méthode utilisée dans les expérimentations de cette présente soumission, pour une tâche de TAP.

Si on formalise, notre décodeur génère N-meilleures hypothèses de traduction $T = \{T_1, T_2, \dots, T_N\}$ à la fin de la "première passe". Toutes ces hypothèses sont ensuite étiquetées par notre estimateur de confiance et nous obtenons alors, pour le j -ème mot dans l'hypothèse T_i , noté t_{ij} , une étiquette de qualité c_{ij} (e.g. "G" (correct, pas d'erreur), "B" (incorrect, doit être édité)). On remarquera que les scores $p(G)$ ou $p(B)$ auraient pu être utilisés plutôt que les étiquettes, mais des expériences conduites sur une tâche de TA (reportées dans (Luong *et al.*, 2014a)) montrent que ceci fait peu de différence au final. Ensuite, la seconde passe considère chaque mot t_{ij} et son étiquette c_{ij} . Si $c_{ij} = "G"$ (les mots sont pris en compte séquentiellement en parcourant la liste des N-best, de la meilleure à la moins bonne, et un mot et son étiquette sont "ignorés" si le mot a déjà été rencontré dans une hypothèse de meilleur rang - ainsi l'étiquette considérée pour un mot est celle correspondant à l'hypothèse placée la plus haut dans la liste N-Best), toutes les hypothèses H_k dans le graphe de recherche contenant ce mot t_{ij} vont être récompensées. En revanche, si $c_{ij} = "B"$, toutes les hypothèses H_k contenant ce mot seront pénalisées. Les autres hypothèses (ne contenant pas t_{ij}) ne seront, quant à elles, pas modifiées. Si on définit $reward(t_{ij})$ et $penalty(t_{ij})$ comme les récompenses (ou pénalités) pour t_{ij} , alors le nouveau score (de transition) de H_k , après mise à jour, sera défini par :

$$transition'(H_k) = transition(H_k) + \begin{cases} reward(t_{ij}) & \text{si } c_{ij} = G \\ penalty(t_{ij}) & \text{sinon} \end{cases} \quad (2)$$

La mise à jour des scores étant faite de la façon suivante :

$$penalty(t_{ij}) = -reward(t_{ij}) = \alpha * \frac{score(H^*)}{\#mots(H^*)} \quad (3)$$

Où $\#mots(H^*)$ est le nombre de mots cible dans H^* , le coefficient α (>0) pondère l'impact de la pénalité (ou de la récompense) sur le score final de l'hypothèse. Ce paramètre doit être optimisé : dans ce travail, en raison de la taille du corpus disponible, nous effectuons une validation croisée avec optimisation sur une moitié du corpus de test et validation sur l'autre moitié (et vice versa en inversant les données d'optimisation et de validation). Ainsi, $penalty(t_{ij})$ (négatif car $score(H^*) < 0$) sera ajouté au score de toutes les hypothèses contenant t_{ij} lorsque ce mot est étiqueté "B" ; tandis que $reward(t_{ij})$ (même valeur absolue mais signe opposé) est utilisé dans le cas contraire. Finalement, la mise à jour des scores s'arrête quand tous les mots différents de la liste des N-meilleures hypothèses ont été traités. Les scores des hypothèses complètes sont alors recalculés à partir du graphe de recherche modifié.

6.3 Résultats

Le graphe d'hypothèse de TAP est généré de la façon suivante : l'hypothèse de RAP (*src-asr*) est traduite par le système de TA (fondé sur l'outil mooses) qui génère un graphe d'hypothèses de traduction. Nous appelons ce graphe : *graphe de traduction de parole* ; cependant, une perspective de ces travaux consistera à construire un graphe plus riche en informations, où le système de RAP fournit lui-même un treillis d'hypothèses en entrée de la TA (ce qui n'est pas le cas ici).

Les performances de traduction de parole (TAP) obtenues avec ou sans re-décodage de graphe, et utilisant nos estimateurs de confiance, sont présentées dans la table 6. Il est important de noter ici qu'une seconde passe qui n'utiliserait aucun estimateur de confiance donnerait lieu à une hypothèse équivalente au système en une passe. La première colonne montre notre *baseline* de TAP (système à une seule passe) dont les résultats ont déjà été donnés dans la table 4. Les seconde, troisième et quatrième colonnes montrent l'amélioration de notre système en prenant en compte des mesures de confiance pour re-décoder le graphe de recherche. Si on compare la première colonne avec la dernière (système 1-passe vs système 2-passes avec le meilleur estimateur), les gains observés sont significatifs (p-valeur dans l'intervalle [0.00 ; 0.01]). On remarque également que l'estimateur de confiance obtenu à partir des traits joints RAP+TA donne une amélioration légèrement plus importante que l'estimateur obtenu à partir de traits TA uniquement (BLEU de 32.82% au lieu de 31.89%) ce qui peut paraître étonnant alors que l'amélioration de la qualité de l'estimateur RAP+TA donnée dans la table 5 était faible par rapport à un estimateur TA seul. Une première explication peut être liée au fait qu'ici, l'estimateur de confiance est appliqué non pas sur 2643 phrases, mais sur 2643*N phrases (avec dans ce cas N=100 meilleures hypothèses) et les différences de performances entre TA et RAP+TA sont peut être plus importantes dans ce cas. Une autre possibilité peut être aussi que même une faible amélioration de la détection d'erreurs (mots dont l'étiquette est B) peut conduire à un gain non négligeable du score BLEU (qui est, rappelons le, évalué ici avec comme référence la post-édition ayant servi également à générer la vérité terrain de nos étiquettes G/B). Enfin, l'estimateur de confiance obtenu à partir de traits RAP seuls permet d'améliorer le décodage par rapport à un système à une seule passe (31.12% au lieu de 30.60%) ; cependant, cette configuration est la plus faible en terme d'amélioration.

système	TAP baseline (BLEU)	TAP redécodage (BLEU)	TAP redécodage (BLEU)	TAP redécodage (BLEU)
estimateurs conf.	aucun	RAP	TA	RAP+TA
<i>Perf.</i>	30.60%	31.12%	31.89%	32.82%

TABLE 6 – Performance de TAP (BLEU) avec ou sans re-décodage de graphes - 2643 phrases

Des exemples de traduction de parole (TAP) obtenues avec ou sans re-décodage de graphe sont donnés dans la table 7 (sans chercher ici à analyser les différences fines entre les estimateurs TA et TA+RAP - ainsi, la ligne avec *re-décodage* indique l'un ou l'autre des estimateurs selon les cas). L'exemple 1 illustre un premier cas où le re-décodage du graphe de traduction permet une légère amélioration de l'hypothèse de traduction. L'analyse des labels issus de l'estimateur de confiance indique ici que les mots *a* (en début de phrase) et *penalty* étaient étiquetés comme incorrects ici ; le redécodage a permis de faire émerger une hypothèse très légèrement meilleure, même si l'erreur de reconnaissance n'a pas pu être rattrapée (puisque de toute façon, seule la meilleure hypothèse de RAP est traduite ici - et pas le graphe d'hypothèse de RAP complet). Pour l'exemple 2, l'estimateur de confiance a étiqueté comme incorrects les séquences *it has*, *speech that was* et *post route* ; à nouveau, une meilleure hypothèse de traduction a été trouvée via re-décodage (pronom correct, fin de phrase de meilleure qualité). Pour finir, l'exemple 3 indique un cas où cette fois-ci, la traduction issue de la première passe

s'est dégradée après redécodage ; l'analyse des sorties de l'estimateur de confiance montre que dans ce cas, la chaîne *to open* était bien étiquetée comme incorrecte, mais le re-décodage a fait émerger une hypothèse encore plus mauvaise. Cet exemple illustre, entre autres choses, les limites de notre approche actuelle qui, dans ce cas précis, aurait de toute façon été incapable de retrouver l'entité nommée *opel* puisque celle-ci n'était pas présente dans le graphe de traduction de parole. Nos travaux à venir, consistant à exploiter aussi le graphe issu de la transcription, nous laissent espérer qu'un tel cas aura une chance d'être résolu dans le futur.

<i>src-ref1</i>	une démobilisation des employés peut déboucher sur une démoralisation mortifère
<i>src-asr1</i>	une démobilisation des employés peut déboucher sur une démoralisation mort y faire
<i>tgt-slt1</i> base-line	a demobilisation employees can lead to a penalty demoralisation
<i>tgt-slt1</i> avec redécodage	a demobilisation of employees can lead to a demoralization death
<i>tgt-pe1</i>	demobilization of employees can lead to a deadly demoralization
<i>src-ref2</i>	celui-ci a indiqué que l'intervention s'était parfaitement bien déroulée et que les examens post- opératoires étaient normaux
<i>src-asr2</i>	celui-ci a indiqué que l' intervention c' était parfaitement bien déroulés , et que les examens post opéra-toire étaient normaux .
<i>tgt-slt2</i> base-line	it has indicated that the speech that was well conducted , and that the tests were normal post route
<i>tgt-slt2</i> avec redécodage	he indicated that the intervention is very well done , and that the tests after operating were normal
<i>tgt-pe2</i>	he indicated that the operation went perfectly well and the post-operative tests were normal
<i>src-ref3</i>	general motors repousse jusqu'en janvier le plan pour opel
<i>src-asr3</i>	general motors repousse jusqu' en janvier le plan pour open
<i>tgt-slt3</i> base-line	general motors postponed until january the plan to open
<i>tgt-slt3</i> avec redécodage	general motors puts until january terms to open
<i>tgt-pe3</i>	general motors postponed until january the plan for opel

TABLE 7 – Exemples de sortie des systèmes avec et sans redécodage de graphes

7 Conclusion

Cet article démontre que des mesures de confiance, issues d'estimateurs automatiques performants, sont utiles pour redécoder des graphes d'hypothèses de traduction du langage parlé. Par ailleurs, les estimateurs de confiances obtenus à partir de traits multiples (RAP et TA) sont plus performants que des estimateurs fondés sur l'une ou l'autre de ces modalités. Cette bonne intégration des informations des modules de transcription et traduction, conduit à des gains de performance (mesurés avec BLEU) encore plus importants. Enfin, ce travail a été possible en raison de la constitution d'un corpus spécifique (oral/écrit avec quintuplets *transcription/référence/TA(transcription)/TA(référence)/post-édition* pour 2643 phrases), également présenté ici, mis à disposition de la communauté TALN. Les perspectives de ce travail sont les suivantes : proposer une tâche d'estimation de confiance pour la TAP pour un workshop international tel que IWSLT, entraîner un estimateur de confiance à partir de véritables traits joints RAP+TA (au lieu de fusionner les sorties de deux estimateurs différents), utiliser les mesures de confiance dans un scénario interactif de traduction de parole (ce qui nécessiterait une étape indispensable d'optimisation de nos méthodes qui, à ce jour, ne sont pas en état de fonctionner "en ligne") ou dans un scénario interactif de transcription de discours (où une hypothèse de TAP serait recalculée "à la volée" en fonction des éditions de l'utilisateur).

Références

- ASADI A., SCHWARTZ R. & MAKHOUL J. (1990). Automatic detection of new words in a large vocabulary continuous speech recognition system. *Proc. of International Conference on Acoustics, Speech and Signal Processing*.
- BACH N., HUANG F. & AL-ONAIZAN Y. (2011). Goodness : A method for measuring machine translation confidence. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, p. 211–219, Portland, Oregon.
- BESACIER L., LECOUEUX B., LUONG N. Q., HOUR K. & HADJSALAH M. (2014). Word confidence estimation for speech translation. In *Proceedings of The International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, USA.

- BICICI E. (2013). Referential translation machines for quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, p. 343–351, Sofia, Bulgaria : Association for Computational Linguistics.
- CHASE L. (1997). *Error-responsive feedback mechanisms for speech recognizers*. PhD thesis, Carnegie Mellon University.
- DUH K. & KIRCHHOFF K. (2008). Beyond log-linear models : Boosted minimum error rate training for n-best re-ranking. In *Proc. of ACL, Short Papers*.
- FAYOLLE J., MOREAU F., RAYMOND C., GRAVIER G. & GROS P. (2010). Crf-based combination of contextual features to improve a posteriori word-level confidence measures. In *Interspeech*.
- GALLIANO S., GEOFFROIS E., GRAVIER G., BONASTRE J.-F., MOSTEFA D. & CHOUKRI K. (2006). Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006)*, p. 315–320.
- HAN A. L.-F., LU Y., WONG D. F., CHAO L. S., HE L. & XING J. (2013). Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, p. 365–372, Sofia, Bulgaria : Association for Computational Linguistics.
- KEMP T. & SCHAAF T. (1997). Estimating confidence using word lattices. *Proc. of European Conference on Speech Communication Technology*, p. 827–830.
- KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses : Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, p. 177–180, Prague, Czech Republic.
- LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting et labeling sequence data. In *Proceedings of ICML-01*, p. 282–289.
- LAMEL L. F., GAUVAIN J.-L., ESKÉNAZI M. *et al.* (1991). Bref, a large vocabulary spoken corpus for french1. *training*, **22**(28), 50.
- LAURENT A., CAMELIN N. & RAYMOND C. (2014). Boosting bonsai trees for efficient features combination : application to speaker role identification. In *Interspeech*.
- LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale crfs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 504–513.
- LECOUTEUX B., LINARÈS G. & FAVRE B. (2009). Combined low level and high level features for out-of-vocabulary word detection. *INTERSPEECH*.
- LUONG N. Q., BESACIER L. & LECOUTEUX B. (2013a). Word confidence estimation and its integration in sentence quality estimation for machine translation. In *Proceedings of The Fifth International Conference on Knowledge and Systems Engineering (KSE 2013)*, Hanoi, Vietnam.
- LUONG N. Q., BESACIER L. & LECOUTEUX B. (2014a). An Efficient Two-Pass Decoder for SMT Using Word Confidence Estimation. In *European Association for Machine Translation (EAMT)*, Dubrovnik, Croatie.
- LUONG N.-Q., BESACIER L. & LECOUTEUX B. (2014b). Word Confidence Estimation for SMT N-best List Re-ranking. In *Proceedings of the Workshop on Humans and Computer-assisted Translation (HaCaT) during EACL*, Gothenburg, Suède.
- LUONG N. Q., LECOUTEUX B. & BESACIER L. (2013b). LIG system for WMT13 QE task : Investigating the usefulness of features in word confidence estimation for MT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, p. 396–391, Sofia, Bulgaria : Association for Computational Linguistics.
- PARTON K., HABASH N., MCKEOWN K., IGLESIAS G. & DE GISPERT A. (2012). Can automatic post-editing make mt more meaningful ? In *Proceedings of the 16th EAMT*, p. 111–118, Trento, Italy.
- POTET M., BESACIER L. & BLANCHON H. (2010). The lig machine translation system for wmt 2010. In A. WORKSHOP, Ed., *Proceedings of the joint fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT2010)*, Uppsala, Sweden.
- POTET M., EMMANUELLE E R., BESACIER L. & BLANCHON H. (2012). Collection of a large database of french-english smt output corrections. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

- POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* : IEEE Signal Processing Society. IEEE Catalog No. : CFP11SRW-USB.
- SANCHIS A., JUAN A. & VIDAL E. (2012). A word-based naïve Bayes classifier for confidence estimation in speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**(2), 565–574.
- SEIGEL M. S. & WOODLAND P. C. (2012). Using sub-word-level information for confidence estimation with conditional random field models. In *INTERSPEECH*.
- SEIGEL M. S., WOODLAND P. C. *et al.* (2011). Combining information sources for confidence estimation with crf models. In *INTERSPEECH*, p. 905–908.
- SNOVER M., MADNANI N., DORR B. & SCHWARTZ R. (2008). Terp system description. In *MetricsMATR workshop at AMTA*.
- TROMBLE R., KUMAR S., OCH F. J. & MACHEREY W. (2008). Lattice minimum bayes risk decoding for statistical machine translation. In *Lattice minimum bayesrisk Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 620–629.
- UEFFING N., MACHEREY K. & NEY H. (2003). Confidence measures for statistical machine translation. In *Proceedings of the MT Summit IX*, p. 394–401, New Orleans, LA.
- VENUGOPAL A., ZOLLMANN A. & VOGEL S. (2007). An efficient two-pass approach to synchronous-cfg driven statistical mt. In *Proceedings of Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics*.
- WEINTRAUB M., BEAUFAYS F., RIVLIN Z., KONIG Y. & STOLCKE A. (1997). Neural-network based measures of confidence for word recognition. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, p. 887–890.
- XIONG D., ZHANG M. & LI H. (2010). Error detection for statistical machine translation using linguistic features. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, p. 604–611, Uppsala, Sweden.
- YOUNG S. R. (1994). Recognition confidence measures : Detection of misrecognitions and out-of-vocabulary words. *Proc. of International Conference on Acoustics, Speech and Signal Processing*, p. 21–24.
- ZENS R. & NEY H. (2006). N-gram posterior probabilities for statistical machine translation. In *Workshop on Statistical Machine Translation - StatMT*, Stroudsburg, PA, USA.
- ZHANG R. & RUDNICKY A. I. (2001). Word level confidence annotation using combinations of features.
- ZHANG Y., HILDEBRAND A. S. & VOGEL S. (2006). Distributed language modeling for n-best list re-ranking. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, p. 216–223, Sydney.

Multi-alignement vs bi-alignement : à plusieurs, c'est mieux !

Olivier Kraif

LIDILEM, Univ. Grenoble Alpes, BP 25, 38040 Grenoble cedex 9

Résumé. Dans cet article, nous proposons une méthode originale destinée à effectuer l'alignement d'un corpus multi-parallèle, i.e. comportant plus de deux langues, en prenant en compte toutes les langues simultanément (et non en composant une série de bi-alignements indépendants). Pour ce faire, nous nous appuyons sur les réseaux de correspondances lexicales constitués par les transfuges (chaînes identiques) et cognats (mots apparentés), et nous montrons comment divers tuilages des couples de langues permettent d'exploiter au mieux les ressemblances superficielles liées aux relations génétiques interlinguistiques. Nous évaluons notre méthode par rapport à une méthode de bi-alignement classique, et montrons en quoi le multi-alignement permet d'obtenir des résultats à la fois plus précis et plus robustes.

Abstract.

Multi-alignment vs bi-alignment: the more languages the better

In this paper, we propose an original method for performing the alignment of a multi-parallel corpus, i.e. a parallel corpus involving more than two languages, taking into account all the languages simultaneously (and not by merging a series of independent bi-alignments). To do this, we rely on the networks of lexical correspondences formed by identical chains and cognates (related words), and we show how various tiling of language pairs allow to exploit the surface similarities due to genetic relationships between languages. We evaluate our method compared to a conventional method of bi-alignment, and show how the multi-alignment achieves both more accurate and robust results.

Mots-clés : Alignement multilingue, corpus parallèles, cognats.

Keywords: Multilingual alignment, parallel corpora, cognates.

1 Introduction

Dans une étude pionnière dans le domaine de la désambiguïsation lexicale, Dagan *et al.* (1991) avaient intitulé leur article « *Two Languages Are More Informative Than One* ». Généralisant le même mot d'ordre (« *Five language are better than one* ») pour la désambiguïsation cross-lingue, Lefever *et al.* (2013) ont récemment illustré cette idée sur un corpus multi-parallèle, en bâtissant un système de classification automatique qui incorpore les contextes de traduction de 5 autres langues que la langue cible (l'anglais) : leur expérience montre que pour 4 langues sur 5, les résultats sont meilleurs quand on fait intervenir les contextes de traduction de plusieurs langues. Dans le domaine de l'alignement de corpus parallèle, cependant, il existe une piste qui a encore été assez peu explorée : celle du *multi-alignement*, à savoir l'alignement de plus de deux langues en même temps. Il paraît pourtant légitime de se poser la question suivante : le fait de prendre en compte plusieurs langues *simultanément* permet-il de mieux aligner ? Ce surcroît d'information permet-il de gagner en robustesse ? en précision ?

Notons que les corpus multi-parallèles ne sont pas rares : l'*Acquis communautaire*, qui constitue le socle législatif et réglementaire de l'Union Européenne en est un des exemples les plus représentatifs : il est actuellement distribué en version 3.0, sous le nom de JRC-Acquis Corpus, et concerne 22 langues européennes¹ – pour un total d'environ 636 millions de mots toutes langues confondues. A l'instar de l'UE, de nombreuses organisations internationales (ONU, GIEC, OMC, OMS, OIT, OCDE, etc.) sont pourvoyeuses de rapports et textes multi-parallèles touchant à des domaines très variés (économie, environnement, médecine, diplomatie, droit, technique, etc.). Dans le cadre du projet

¹ cf. <http://ipsc.jrc.ec.europa.eu/index.php?id=198>

Carmel (Chen et al., 2005), des œuvres littéraires (des récits de voyages) ont été réunies en 4 langues : anglais, espagnol, français, italien. On trouve par ailleurs sur le site de l'OPUS - Open parallel corpus (Tiedemann, 2012) une grande variété de corpus multi-parallèles déjà alignés. Certains de ces corpus sont massivement multilingues, comme le corpus *OpenSubtitles2013* qui compte 59 langues et intègre la plupart des paires de langues impliquées (1 211 paires de langues sur 1 711).

Il est frappant de constater que pour ces corpus massivement multi-parallèles, on utilise toujours des techniques d'alignement bilingue. Pour le JRC-Acquis, tous les alignements ont été effectués 2 à 2, en utilisant l'aligneur Vanilla (Danielsson & Riding, 1997), qui implémente la méthode de Gale & Church (1991), ainsi que l'aligneur Hunalign (Varga et al. 2005). De même les textes du corpus OPUS ont été alignés 2 à 2 grâce à une méthode similaire. Le principal inconvénient de l'alignement 2 à 2 d'un corpus de textes multi-parallèles réside dans le grand nombre de couples à considérer. Par exemple, pour les 22 langues du JRC Acquis, il faut considérer $22 \times 21 / 2 = 231$ couples différents. D'un point de vue général, pour n langues, le nombre de couples impliqués est quadratique : $n \times (n-1) / 2$. Cette complexité peut être pénalisante à la fois du point de vue du temps de calcul et de l'espace de stockage des résultats. Quand on a 22 textes parallèles, pourquoi ne pas aligner les 22 langues en même temps, et représenter l'alignement résultant dans une seule structure de données, par exemple un seul fichier au format TMX contenant tous les groupes de phrases équivalents, plutôt que 231 fichiers différents ? La question du multi-alignement constitue donc également un enjeu en termes de complexité : le problème est-il abordable sur un plan calculatoire, et permet-il de faire mieux qu'une complexité quadratique ?

Cet article se propose de donner un début de réponse à ces questions, dans le cadre restreint de l'alignement phrastique. Dans la section 2, nous évoquerons les rares travaux qui ont cherché à sortir du carcan de l'alignement 2 à 2. Nous décrirons ensuite l'architecture d'un aligneur baptisé JAM (pour *Just A Multialigner*) qui s'appuie sur des réseaux de correspondances multilingues. Dans la quatrième section, nous chercherons à identifier ces réseaux de correspondance à travers un petit corpus en 11 langues tiré du corpus Europarl. Nous en déduirons une méthode de tuilage permettant de tirer le meilleur parti de ces correspondances. Enfin, dans la section 5, nous comparerons les performances de JAM avec celles de deux autres aligneurs bilingues, Vanilla et Yasa (Lamraoui & Langlais, 2013).

2 Etat de l'art

En ce qui concerne l'alignement phrastique, force est de constater que depuis les travaux pionniers du début des années 1990, peu de choses ont bougé : les modèles superficiels faisant intervenir les longueurs de phrases (Gale & Church, 1991, Brown et al. 1991) ont largement fait leur preuve, complétés par des méthodes intégrant le repérage de ressemblances de surface, n-grammes, transfuges (i.e. les chaînes identiques, souvent des nombres et des entités nommées) ou cognats (Simard, Foster & Isabelle, 1992 ; Mc Enery & Oakes, 1995 ; Kraif, 2001). Nombre de ces algorithmes s'appuient sur le cercle vertueux énoncé par Kay et Röschensein dès 1988 (Kay & Röschensein, 1993) : aligner au niveau des phrases pour aligner au niveau des mots, et aligner au niveau des mots pour aligner au niveau des phrases. Davis et al. (1995), dans le souci de tenir compte des ruptures de parallélisme fréquentes dans les traductions réelles, ont proposé de combiner ces différents types d'indices pour les intégrer dans un même cadre algorithmique. La campagne d'évaluation Arcade 2 (Chiao et al. 2006), a montré qu'il était possible d'appliquer ces méthodes, avec une certaine robustesse, à des couples de langues éloignées ne partageant pas le même alphabet (le français avec le russe, le chinois, le japonais ou l'arabe).

En complément des modèles de surface basés sur les longueurs, certaines méthodes s'appuient sur les alignements au niveau lexical, ce qui peut s'avérer nécessaire dans le cas de traductions « bruitées » s'écartant du parallélisme. Ce type de modèle requiert dans certains cas un lexique bilingue externe, comme dans Li et al. (2010), qui utilisent ce lexique à la fois pour extraire des points d'ancrage fiables destinés à réduire l'espace de recherche, et pour calculer l'alignement final. Moore (2002) obtient des résultats à la fois robustes et précis pour des alignements 1-1, en combinant les longueurs de phrases et l'alignement lexical, les alignements obtenus dans une première passe étant utilisés pour entraîner le modèle 1 d'IBM (Brown et al., 1993) en vue d'affiner l'alignement des phrases à partir de l'alignement des mots. Braune & Frazer (2010) améliorent cette dernière méthode en proposant de regrouper les phrases non alignées avec les alignements 1-1, afin de pouvoir constituer des alignements 1-n et augmenter le rappel. Plus récemment, Lamraoui & Langlais (2013), insistant sur le fait que l'alignement phrastique était un champ de recherche encore ouvert et méritant de nouvelles investigations, ont pourtant montré que leur aligneur Yasa, dont l'architecture très simple s'appuie sur un préalignement basé sur les cognats et un alignement phrastique « classique » incorporant longueurs de phrases et densité de cognat, était capable de dépasser les systèmes état de l'art plus complexes tels que BMA (Moore, 2002) et Hunalign (Varga et al., 2005), et ceci sur différents genres de textes.

Enfin, s'affranchissant de la contrainte de parallélisme, notons que de nombreux travaux ont porté sur l'alignement à travers des corpus comparables, au niveau phrastique (Munteanu et al., 2004) ou sub-phrastique (Hewavitharana & Vogel, 2011).

Le succès de ces méthodes explique peut-être le fait que le multi-alignement ait été si peu exploré. Dans le cas de trois langues, une étude assez complète a été effectuée par Simard (1999), avec un article dont le titre fait écho à l'article précédemment cité : « *Text-Translation Alignment: Three Languages Are Better Than Two* ». Il y présente une méthode d'alignement ternaire, nommée *Trial*, basée sur la réitération de la méthode bilingue. Étant donné 3 textes A, B, C, on aligne d'abord A avec B, puis C avec le bi-texte AB (le calcul du coût d'un appariement entre une phrase c et une bi-phrase (a,b) étant une simple combinaison linéaire des coûts d'appariement entre c et a , et c et b). La méthode présentée par Simard n'a pas pour but d'économiser le temps de calcul, puisque tous les alignements bilingues AB, BC et AC sont calculés préalablement, afin de choisir la paire de langue optimale, qui sera ensuite réalignée avec la langue restante. En outre, les trois alignements bilingues permettent de dégager des points d'ancrage pour l'alignement ternaire, lorsqu'ils sont concordants (i.e. quand pour trois phrase a,b et c on a les appariements (a,b) (b,c) et (c,a)). Ce que montre Simard, ce n'est donc pas une réduction du calcul, mais une amélioration (certes modeste, avec 1% de F-mesure en plus) de la qualité de l'alignement final. Ainsi, quand un couple de langues est défaillant (p.ex. parce qu'on a trop peu de mots apparentés pour guider l'alignement des phrases), une troisième langue peut apporter une information complémentaire et suppléer à cette défaillance.

Dans la perspective de l'alignement sous-phrastique, la méthode d'alignement par échantillonnage proposée par Lardilleux (2010) présente l'originalité d'aligner simultanément plus de deux langues, toutes les unités des phrases alignées pouvant être fusionnées dans un même contexte d'occurrence sans langue définie – la méthode étant dite « alingue ». Des tirages aléatoires d'échantillon du corpus permettent de regrouper les unités qui partagent les mêmes distributions, ces regroupements pouvant aussi bien réunir des unités dans une même langue (i.e. des expressions polylexicales) que des équivalents traductionnels. Cette méthode est intéressante du point de vue de l'économie des traitements, l'alignement simultané du corpus multi-parallèle étant moins coûteux que l'alignement des langues deux à deux, mais elle n'est pas directement transférable à la problématique de l'alignement phrastique, et ne s'appuie pas, comme dans *Trial*, sur un renforcement des alignements par triangulation.

3 Cadre algorithmique pour un multi-aligneur

Les méthodes bilingues telles que celles de Gale & Church sont difficilement généralisables au cas de n langues, la complexité des algorithmes de programmation dynamique mis en œuvre étant exponentielle en $O(t^n)$, pour des textes de taille t . Le système *trial*, dans la mesure où il implique de pré-calculer tous les alignements 2 par 2, nous semble également assez lourd sur le plan algorithmique lorsqu'un grand nombre de langues est mis en jeu. Il n'a d'ailleurs jamais été étendu au-delà de trois langues, à notre connaissance. D'autres méthodes peu coûteuses sont envisageables, comme l'alignement par transitivité : si A est aligné avec B et B est aligné avec C, alors on peut calculer rapidement, par transitivité, un alignement entre A et C. Mais lorsque l'on prend la clôture transitive des alignements, on obtient en général des alignements plus grossiers, regroupant plusieurs phrases, ce qui aboutit à une baisse de la précision. Par ailleurs cette méthode ne tire pas parti du principe de triangulation : tout repose sur une seule langue pivot, et si l'alignement au niveau d'un couple est faible, cette faiblesse sera propagée vers la troisième langue par le jeu de la transitivité, au lieu d'être éventuellement compensée par la prise en considération d'un autre couple plus solide.

3.1 Cognats et multi-alignement

Pour des corpus multi-parallèles tels que ceux de l'Union Européenne, il apparaît que la parenté linguistique entre les différents groupes de langues impliqués (langues romanes, langues germaniques, langues slaves, langues baltes, langues finno-ougriennes, pour ne citer que les principaux groupes) doit pouvoir jouer un rôle prépondérant dans le multi-alignement. Afin d'explorer cette hypothèse, nous avons téléchargé la transcription de la session du 17 janvier 2000 du parlement européen, tiré du corpus Europarl3 (cf. <http://opus.lingfil.uu.se/Europarl3.php>), qui contient des versions alignées dans les langues suivantes : allemand, anglais, danois, espagnol, français, finnois, grec, italien, portugais, néerlandais, suédois (on utilisera désormais les codes ISO, par ordre alphabétique : DA, DE, EN, ES, EL, FI, FR, IT, NL, PT, SV). La partie française comporte environ 30 000 tokens (mots graphiques, ponctuations, nombres, etc.). Nous avons manuellement révisé les alignements fournis pour tous les couples impliquant le français afin d'avoir une référence fiable (la plupart des alignements fournis étaient de bonne qualité à part pour le couple FR-NL qui a nécessité un peu plus de révisions).

La première tâche a consisté à mesurer le degré de proximité graphique des formes alignées entre toutes les langues prises deux à deux, afin d'évaluer jusqu'à quel point la parenté génétique peut se traduire en un critère automatiquement exploitable (l'identification des candidats cognats). Pour chaque couple de phrases, nous avons compté les candidats cognats en retenant toutes les paires de mots d'au moins 7 caractères pour laquelle la sous-chaîne commune maximale (SCM, cf. Kraif, 2001) correspond à au moins 80 % des caractères de la chaîne la plus courte des deux chaînes comparées. Avec de tels critères, plutôt sélectifs, on trouve de très nombreux cognats avec un minimum de bruit. Par exemple, pour les langues DA, DE, EN, on trouve les paires suivantes : *periodiske*↔*Periodischen*, *Schroedter*↔*Schroedterin*, *Parlaments*↔*Parliament*, *Regionalpolitik*↔*Regional*, *regionaler*↔*regional*, *Europa-Parlamentets*↔*Europaparlamentets*, *Europæiske*↔*Europäischen*, *Kommission*↔*Commission*, etc.

Les résultats, cumulant le nombre de chaînes identiques (hormis les nombres et les noms commençant par une majuscule) et le nombre de cognats identifiés avec les critères précédents, sont présentés dans le tableau 1 :

	DA	DE	EN	ES	FI	FR	IT	NL	PT	SV	Total
DA		1 114	1 202	705	458	1 984	1 041	1 019	479	2 325	14 327
DE	1 114		863	448	397	735	747	722	376	925	10 327
EN	1 202	863		1 968	527	2 367	2 225	1 174	1 493	1 256	17 075
ES	705	448	1 968		222	1 829	2 234	638	3 750	764	16 558
FI	458	397	527	222		292	481	197	174	617	7 365
FR	1 984	735	2 367	1 829	292		2 120	936	1 350	851	16 464
IT	1 041	747	2 225	2 234	481	2 120		978	1 935	354	16 115
NL	1 019	722	1 174	638	197	936	978		489	893	11 046
PT	479	376	1 493	3 750	174	1 350	1 935	489		579	14 625
SV	2 325	925	1 256	764	617	851	354	893	579		12 564
Total	14 327	10 327	17 075	16 558	7 365	16 464	16 115	11 046	14 625	12 564	136 466

TABEAU 1 : Nombre de transfuges et cognats identifiés dans les bi-phrases par couples de langues (grec exclu)

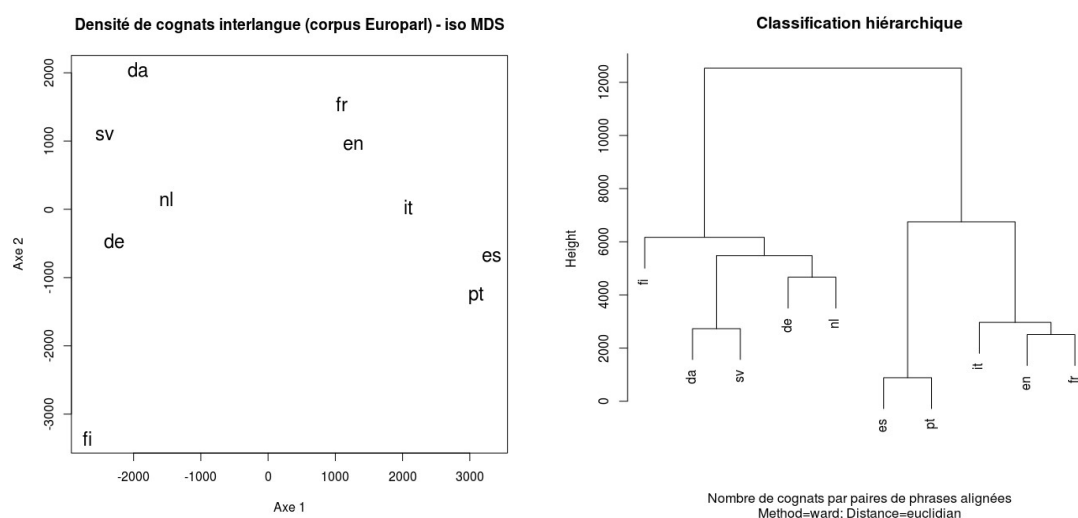


FIGURE 1 : Analyse multivariée du nombre de cognats par paires de phrases alignées (iso MDS et hclust)

Quand on considère les valeurs marginales, on constate que certaines langues cumulent beaucoup plus de cognats que d'autres : elles occupent en quelque sorte une position plus centrale au sein de ces différentes familles linguistiques, position qui leur confère en moyenne une plus grande ressemblance avec un plus grand nombre de langues – c'est notamment le cas du français et de l'anglais. Par ailleurs, ces densités de mots ressemblants font apparaître les affinités entre langue et groupe de langues. La figure ci-dessous montre les résultats de deux méthodes multivariées (classification hiérarchique ascendante obtenue avec la fonction *hclust()* de 'R', et échelonnement multidimensionnel avec iso MDS) en interprétant la matrice du tableau 1 comme une matrice de similarité (arbitrairement, on remplit la diagonale de la matrice par la valeur 4000 qui correspond approximativement à la borne supérieure des similarités observées). On note que les aspects génétiques sont étroitement corrélés aux similarités qui ressortent de ces données.

3.2 L'aligneur JAM

Pour utiliser ces cognats comme point d'appui pour le multi-alignement, on se dote d'une structure de données complexe que l'on nomme treillis de multi-alignement (cf. le tableau 2, qui contient les numéros de phrases appariées : chaque ligne correspond à un nœud du treillis, et les arcs sont les liens de successions propres à chaque langue). Cette table permet de mettre en correspondance les appariements simples entre phrases des différents textes (ici représentées par leur numéro), ces appariements pouvant être fragmentaires (lignes incomplètes) et l'ordre des lignes étant fixé par un chaînage propre à chaque langue : p.ex. pour NL le successeur de la ligne (1;2;2;2;2;2;2;2;2;2) est la ligne (1;7;1;6;1;7;1;9;7;1). Par construction, tout appariement d'une phrase avec une autre est transitif : à l'intérieur d'une même ligne, toutes les phrases sont alignées entre elle. Pour un groupe de langues donné (p.ex. S=EN-NL-SV) une étape de l'alignement consiste à : 1/ considérer tous les candidats cognats définis pour ce groupe de langues Co(S) 2/ identifier les intervalles à aligner, entre deux points consécutifs issus des étapes précédentes (pour les langues considérées p.ex. ici $P_i=(2;2;2)$ et $P_{i+1}=(7;10;9)$) 3/ à l'intérieur de chaque intervalle, sélectionner les cognats qui apparaissent avec la même fréquence pour chaque langue et ajouter les nouveaux points correspondant aux appariements de ces cognats 4/ Revenir à 2 sauf si stabilité 5/ Enfin, à l'issue de ces itérations, effectuer une étape de complétion : pour tous les couples de langues, lancer l'algorithme d'alignement de Gale & Church en ne calculant que les chemins qui passent par les points précédemment obtenus. Dans l'exécution de cette étape, les couples de langues sont ordonnés par complétude décroissante, c'est-à-dire que l'on traite en priorité les couples qui ont produit de nombreux points. Au fur et à mesure des alignements successifs, tous les appariements 1:1 qui correspondent à des points jugés *équilibrés*² et *cohérents* (i.e. sans croisement avec des points existants), sont ajoutés dans le treillis – et pourront être utilisés comme point d'ancrage pour les couples de langues non encore alignés.

Lors de chacune de ces étapes, pour tout ajout d'un nouveau point au treillis, trois cas de figure sont à considérer : soit il existe plusieurs lignes du treillis de multi-alignement liées à des phrases du nouveau point, et ces lignes doivent alors être fusionnées (si c'est possible) ; soit il existe une seule ligne liée à ces phrases, et celle-ci doit être complétée ; soit une nouvelle ligne doit être créée.

Ces étapes peuvent être exécutées pour n'importe quelle combinaison de groupes de langues, p.ex. Comb= {DA-DE-FI-NL-SV-EN, EN-ES-EL-FR-GR-IT-PT}. Notons qu'il est important que ces groupes possèdent des intersections non vides afin de relier toutes les langues au final : comme nous le verrons dans la section 3.3, un *tuilage* approprié est déterminant pour les résultats.

DA	DE	EL	GR	EN	ES	FI	FR	IT	NL	PT	SV
		1	1	1			1	1		1	
	2	2	2	2	2	2	2	2	2	2	2
	3	3	3		3	3	3	3		3	
	4				4	4	4	4		4	
	5			4	5	5	5	5		5	
	6			5	6	6	6	6		6	
	7			6		7	7		9	7	
	8			7	8	8	8	8	10	8	9
	10			8	9	10	9	9	11	9	
11	11			9		11	10	10	12		13
...

TABLEAU 2 : Extrait du treillis de multi-alignement obtenu pour le début de notre corpus, contenant les numéros de phrases appariées (dans ce multi-alignement, ainsi que dans tous les suivants, nous incluons le grec EL, ainsi que sa variante translittérée, qui sera désormais notée GR)

Calculer l'ensemble Co(S) des candidats cognats propres à une combinaison donnée peut se révéler assez coûteux : pour n textes avec un vocabulaire moyen de v formes, on aura environ $v^2 * n * (n-1) / 2$ couples de formes à comparer. Pour éviter cela, on commence par amorcer l'alignement avec les transfuges (les chaînes identiques) : calculer l'ensemble de transfuges Tr(S) est assez simple, en construisant initialement un hachage associant pour chaque forme

² Par point « équilibré », nous entendons que toutes les longueurs des phrases, prises deux à deux, sont comparables : on impose que le rapport de la plus courte sur la plus longue, en longueur relative, doit être compris entre 0.75 et 1.

la liste des différentes langues comportant cette forme. Si les combinaisons de Comb sont connues à l'avance, alors il est aisé d'alimenter $Tr(S)$ dès la construction de ce hachage. À chaque ajout d'un nouveau point dans le treillis de multi-alignement, on compare alors les phrases alignées 2-à-2, et l'on en extrait les candidats cognats, cette fois en se basant sur la recherche de SCM (cf. section 3.1).

L'algorithme complet de JAM est décrit dans la figure ci-dessous :

```

 $T \leftarrow \{Premier, Dernier\}$ ; # on initialise le treillis avec le premier et le dernier point du multi-texte
 $Comb \leftarrow \{\text{ensemble des combinaisons de langues considérées}\}$ 
Pour chaque combinaison de langues  $S = \{L_1 - L_2 - \dots - L_K\}$  de l'ensemble  $Comb$ 
   $CO(S) \leftarrow \{\text{ensemble des transfuges identifiés pour } S\}$ 
  # 1 - itérations
  Pour chaque cognat ou transfuge  $C$  de  $CO(S)$ 
    Pour chaque couple de points  $(P_i, P_{i+1})$  résultant de la suite ordonnée des points de  $T$  définis pour les langues de  $S$ 
      Si  $C$  a  $n$  occurrences correspondant aux phrases  $occ_{L_1}, occ_{L_2}, \dots, occ_{L_n}$  dans l'intervalle  $[P_i, P_{i+1}]$  pour chaque langue  $L$  de  $S$ 
        Pour  $j = 1 \dots n$ 
          Si le nouveau point  $(occ_{L_1,j}, occ_{L_2,j}, \dots, occ_{L_K,j})$  est équilibré et cohérent
             $T \leftarrow T \cup (occ_{L_1,j}, occ_{L_2,j}, \dots, occ_{L_K,j})$ 
            Pour toutes les paires de mots  $(M_{L_{x,j}}, M_{L_{y,j}})$  des phrases  $occ_{L_{x,j}}, occ_{L_{y,j}}$  alignées du nouveau point
              Si  $longueur(M_{L_{x,j}}) > 6$  et  $SCM(M_{L_{x,j}}, M_{L_{y,j}}) \geq 0.8 * \min(longueur(M_{L_{x,j}}), longueur(M_{L_{y,j}}))$ 
                associer le même identifiant de cognat à  $M_{L_{x,j}}$  et à  $M_{L_{y,j}}$ 
              Fin Si
            Fin Pour
          Fin Si
        Fin Pour
      Fin Si
    Fin Pour
  Fin Si
  # 2 - complétion
  Ordonner tous les couples de langue  $(L_i, L_j)$  par ordre décroissant en fonction du nombre de bi-points obtenus.
  Pour chaque couple de langue  $(L_i, L_j)$  de cette liste ordonnée
    Appliquer l'algorithme de Gale & Church entre chaque point existant dans  $T$ 
    Pour tout bi-point  $(Num_i, Num_j)$  de type 1 : 1
      Si  $(Num_i, Num_j)$  est équilibré et cohérent
         $T \leftarrow T \cup (Num_i, Num_j)$ 
      Fin Si
    Fin Pour
  Fin Pour

```

FIGURE 2 : Algorithme itératif d'appariement des transfuges et cognats

Afin de garantir le maximum de précision, nous appliquons les deux critères suivants dans la sélection des points :

- *redondance* : dans un premier cycle d'itérations, on ne tient compte que des points contenant au moins *minMatchNumber* appariements de cognats (ou transfuges). Après stabilité, on réitère en décrémentant cette valeur. Dans les résultats qui suivent, on a testé *minMatchNumber*=2 puis 1.
- *parallélisme* : à chaque ajout d'un nouveau point P , on considère les deux points existants P_{inf} et P_{sup} qui encadrent ce point (pour les langues considérées dans ce point). On peut alors calculer la longueur des intervalles entre P_{inf} et P (notons Inf_L) et entre P et P_{sup} (notons Sup_L) pour chaque langue L . La vérification de parallélisme se fait alors pour chaque couple de langues L_1, L_2 : si $déviations(Inf_{L_1}, Inf_{L_2})^3 > maxDiffInterval$ et/ou $déviations(Sup_{L_1}, Sup_{L_2}) > maxDiffInterval$ alors les deux coordonnées du point correspondant à L_1 et L_2 sont rejetées. On peut appliquer cette contrainte sur les deux intervalles en même temps (« parallélisme fort »), ce qui n'autorise aucun décrochement dans l'alignement des points, ou bien d'un côté seulement (demi-parallélisme), ce qui peut permettre des sauts.

Nous avons effectué un premier test en utilisant un jeu de combinaisons de langues simple, prenant le français comme pivot (on notera FR-pivot) : $Comb_{FR-pivot} = \{EN-FR, FR-IT, ES-FR, FR-PT, DA-FR, FR-NL, FR-SV, DE-FR, FR-FI, FR-GR, FR-EL\}$. Comme le montre la première colonne du tableau 3, on obtient une précision excellente mais un rappel assez faible. En examinant les points obtenus à ce stade, on obtient encore de nombreux « trous », comme l'illustre l'exemple du tableau 2 : l'alignement n'est pas complet car certaines langues se trouvent isolées, comme DA, EL, NL ou SV. Pour tenter d'éliminer ces trous, nous appliquons alors l'algorithme de complétion finale en lançant l'algorithme d'alignement de Gale & Church (1991) pour toutes les langues prises deux à deux. Notons que cet algorithme livre des appariements groupés de type 1:2, 2:1, 2:2 tandis que notre multi-alignement n'enregistre que des

³ Pour deux intervalles Int_{L_1} et Int_{L_2} , on a : $déviations(Int_{L_1}, Int_{L_2}) = 2x \mid Int_{L_1} - Int_{L_2} \mid / (Int_{L_1} + Int_{L_2})$. Le seuil *maxDiffInterval* prend des valeurs entre 0.5 et 0.25 suivant la taille de l'intervalle.

correspondances 1:1 sans fusion ni croisement. Pour rester dans le cadre imposé par notre structure de données, seuls les appariements 1:1 sont retenus (ce qui explique que le rappel ne puisse être optimal).

On obtient alors les résultats de la deuxième colonne du tableau 3. Notons que le coût de cet algorithme est modéré, étant donné l'étroitesse de l'espace de recherche : pour obtenir les résultats précédents, l'algorithme de Gale & Church a été lancé 6 662 fois sur des intervalles d'une longueur moyenne de 4 phrases environ, l'intervalle le plus grand ayant une longueur de 75 phrases⁴.

Couple de langues	Sans complétion finale		Avec complétion finale	
	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>
DA-FR	96,84	24,35	97,47	80,42
DE-FR	98,36	30,36	97,81	81,52
EL-FR	97,33	15,07	98,18	83,59
EN-FR	98,53	64,39	98,56	87,31
ES-FR	99,40	50,67	99,53	87,10
FI-FR	99,13	12,46	98,82	91,80
FR-GR	97,74	17,89	98,06	83,45
FR-IT	98,51	49,89	98,60	83,91
FR-NL	96,27	32,32	96,46	65,80
FR-PT	98,99	53,51	99,16	90,68
FR-SV	98,53	26,72	99,01	79,66

TABEAU 3 : Résultats de JAM pour les combinaisons FR-pivot

3.3 Tuilage des couples de langues

Cherchons maintenant une combinaison de langues qui soit optimale, sans s'appuyer *a priori* sur le français. Une piste consiste à chercher le meilleur *tuilage* des alignements. Par *tuilage* on entend un ensemble de combinaisons de langues tel que chaque langue apparaît dans au moins une combinaison et que chaque combinaison possède au moins une langue en commun avec une autre combinaison. Parmi les *tuilages* possibles, on cherchera le *tuilage* qui met en jeu les groupes des langues plus fortement associées (d'après les données du tableau 1), afin qu'il puisse s'appuyer les uns sur les autres, de manière complémentaire, pour former un tout plus solide.

L'anglais étant la langue obtenant la plus importante valeur marginale (cf. tableau 1), on peut supposer qu'un *tuilage* basé sur l'anglais comme pivot donnera de bons résultats : on notera cette combinaison $Comb_{EN-pivot} = \{EN-FR, EN-IT, EN-ES, EN-PT, EN-SV, DA-EN, EN-NL, DE-EN, EN-FI, EN-GR, EN-EL\}$. Notons que l'ordre d'application des couples peut jouer un rôle : on traitera d'abord les meilleurs couples, qui forment un réseau d'associations plus denses et plus sûres.

Une autre stratégie consiste à s'appuyer sur des triplets plutôt que sur des couples, la triangulation des langues permettant peut-être d'améliorer la précision des résultats. Nous avons cherché à identifier les meilleurs triplets potentiels en retenant, pour chaque langue, ses deux langues les plus fortement associées en termes de densité de cognats. Comme on peut supposer que la validité d'une combinaison n'est pas dépendante d'un texte en particulier, nous avons pris un autre corpus, tiré du JRC Acquis⁵, pour calculer les densités de cognats à travers les phrases alignées. En ordonnant les triplets de façon décroissante en fonction du nombre global de cognats obtenus pour la première langue de chaque triplet, on obtient : $Comb_{TRI} = \{EN-FR-ES, FR-IT-EN, ES-PT-EN, PT-ES-IT, NL-EN-FR, SV-DA-NL, DE-DA-NL, FI-EN-SV, EL-GR-EN\}$. Enfin, à titre de *baseline*, nous avons testé également un *tuilage* « aléatoire » basé sur l'ordre alphabétique des codes de langue : $Comb_A = \{DA-DE, DE-EL, EL-EN, EN-ES, ES-FR, FR-FI, FI-GR, GR-IT, IT-NL, NL-PL, PT-SV\}$.

Les résultats du tableau 4 montrent qu'au final les différences sont peu significatives entre ces différents *tuilages* : si on observe des différences marquées avant complétion au niveau du rappel (p.ex. $R=16,75\%$ pour la *baseline* contre $34,33\%$ pour *FR-Pivot*), la complétion permet un certain rattrapage qui uniformise les résultats à 2 point de F-mesure

⁴ Si on utilisait directement l'algorithme de Gale & Church (1991) sur l'intégralité des textes pour les 66 couples en présence, chaque texte faisant environ 1 000 phrases, on aurait une complexité bien supérieure.

⁵ Décision du 27 février 2002, ref. Celex=42002D0234, cf. http://optima.jrc.it/Acquis/index_2.2.html.

prêt. À ce stade, *FR-Pivot* semble être la meilleure combinaison – ce résultat étant toutefois à prendre avec précaution et peut être dû à des spécificités de notre corpus : il faudrait des études sur un corpus plus grand et plus varié pour tirer des conclusions sur ce plan.

Couple de langues	<i>Comb_{FR-pivot}</i>		<i>Comb_{EN-PIVOT}</i>		<i>Comb_A</i>		<i>Comb_{TRI}</i>	
	<i>P</i> %	<i>R</i> %	<i>P</i> %	<i>R</i> %	<i>P</i> %	<i>R</i> %	<i>P</i> %	<i>R</i> %
DA-FR	97,47	80,42	97,60	80,72	97,53	78,43	96,88	77,14
DE-FR	97,81	81,52	97,51	79,59	98,24	79,19	95,04	75,84
EL-FR	98,18	83,59	98,16	82,56	98,62	81,32	98,38	81,63
EN-FR	98,56	87,31	98,44	87,42	99,13	85,29	99,14	86,14
ES-FR	99,53	87,10	97,08	81,68	98,81	85,26	97,47	82,91
FI-FR	98,82	91,80	94,03	80,87	98,80	90,27	99,39	89,73
FR-GR	98,06	83,45	98,15	82,42	97,76	81,39	98,25	81,49
FR-IT	98,60	83,91	97,94	81,97	98,28	80,35	98,84	82,61
FR-NL	96,46	65,80	95,51	64,82	95,18	61,86	96,88	66,61
FR-PT	99,16	90,68	97,80	87,61	97,44	83,55	98,76	87,06
FR-SV	99,01	79,66	97,74	77,57	95,51	72,18	98,46	76,47
Moyenne	98,33	83,21	97,27	80,66	97,76	79,92	97,95	80,69

TABLEAU 4 : Résultats comparés pour différents tuilages (après complétion finale)

4 Comparaison avec des méthodes de bi-alignement

Reste à déterminer, à l'issue de ces différentes observations, si le recours au multi-alignement présente vraiment un intérêt par rapport à l'alignement binaire : c'est la question centrale à laquelle il nous faut maintenant tenter de donner une réponse. Pour ce faire, nous avons comparé JAM avec l'aligneur Vanilla (Danielsson & Riding, 1997), basé sur l'algorithme de Gale & Church (1991), et l'aligneur Yasa (Lamraoui & Langlais, 2013) qui a obtenu des résultats au niveau de l'état de l'art en conjuguant longueurs de phrases et densité de cognats. Pour Jam, nous avons utilisé *Comb_{FR-PIVOT}* en appliquant la contrainte de « parallélisme fort » dans le filtrage des points. Dans le tableau 5, nous ajoutons les résultats de JAM+GC, pour les bi-alignements obtenus par l'application de l'algorithme de Gale & Church à l'issue de l'étape finale de complétion :

Couple de langues	Vanilla		JAM <i>Comb_{FR-PIVOT}</i>		JAM <i>Comb_{FR-PIVOT}</i> + GC		YASA	
	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>	<i>P</i>	<i>R</i>
DA-FR	0,94	0,93	0,99	0,82	0,96	0,95	0,97	0,94
DE-FR	0,94	0,95	0,98	0,80	0,95	0,95	0,95	0,94
EL-FR	0,09	0,12	0,98	0,82	0,94	0,96	0,96	0,94
EN-FR	0,98	0,98	0,99	0,88	0,98	0,98	0,98	0,98
ES-FR	0,90	0,92	0,99	0,85	0,97	0,98	0,98	0,96
FI-FR	0,96	0,97	0,98	0,90	0,94	0,96	0,99	0,99
FR-GR	0,94	0,95	0,98	0,81	0,93	0,95	0,96	0,94
FR-IT	0,95	0,96	0,98	0,82	0,96	0,97	0,97	0,97
FR-NL	0,80	0,80	0,97	0,63	0,90	0,90	0,95	0,90
FR-PT	0,96	0,97	1,00	0,91	0,98	0,99	0,99	0,98
FR-SV	0,87	0,89	1,00	0,76	0,91	0,92	0,97	0,94
Moyenne (hors EL)	0,92	0,93	0,99	0,82	0,94	0,95	0,97	0,95

TABLEAU 5 : Résultats comparés de Vanilla, de JAM, de JAM+GC (avec l'application a posteriori de l'algorithme de Gale & Church entre les points d'ancrage), et de YASA

Les résultats de Vanilla pour le grec (EL) sont mauvais, mais nous pensons qu'il s'agit d'un problème de prise en compte du codage UTF-8 par Vanilla, et nous n'avons pas intégré ces résultats dans la moyenne (à priori, l'algorithme

de Gale & Church ne s'appuyant que sur les longueurs de phrases, EL et GR devraient être identiques). Nous avons donc calculé les moyennes sans cette ligne (en gris).

Vanilla obtient une F-mesure globale de 92,9 %, tandis que JAM obtient 95,07%, contre 96,3% pour YASA. Les résultats sont serrés, mais YASA s'en sort mieux globalement, et semble plus robuste, notamment pour les alignements avec le suédois et le néerlandais qui ont posé des problèmes à JAM. Pour NL, notamment, un point déviant issu de la phase 1 n'a pas été éliminé par les contraintes de parallélisme, et ce point a localement dégradé une partie de l'alignement. L'architecture multilingue permet donc d'améliorer l'alignement basé sur les longueurs (Gale & Church), mais sans toutefois surpasser un algorithme bilingue qui combine de façon optimale densité de cognats et longueurs. Un paramétrage plus fin de JAM permet d'améliorer les résultats, notamment en utilisant un tuilage basé sur les meilleures paires de langue d'abord : avec $Comb_{Max} = \{ES-PT, EN-FR, DA-SV, ES-IT, EN-IT, SV-EN, NL-EN, DE-DA, FI-SV\}$, on obtient une F-mesure de 96,2%. Cependant, la recherche des paramètres optimaux ayant été effectuée sur ce même corpus, nous ne pouvons en tenir compte, d'autant que YASA n'a pas bénéficié d'un tel réglage. Par ailleurs, au plan du temps d'exécution, JAM (qui est implémenté en Perl) est encore assez lent : il prend 230 s. pour aligner tous les couples, tandis que YASA prend seulement 82 s. pour la même opération.

Il faut noter que les résultats de JAM (deuxième colonne) ne sont pas vraiment comparables avec ceux des autres colonnes, puisqu'il s'agit d'un multi-alignement ne comportant que des correspondances 1:1. Un multi-alignement construit à partir de regroupements de type 1:2, 2:1, 2:2 etc. serait *ipso facto* beaucoup moins précis. En effet, si on applique la propriété de transitivité sur des alignements binaires complets, on peut obtenir des regroupements très larges : il suffit qu'un alignement pour un couple de langues chevauche deux groupes de phrases différents pour d'autres couples pour que ceux-ci fusionnent, et ainsi de suite. Nous en avons fait l'essai en prenant la clôture transitive de nos alignements de référence avec le français, et nous obtenons des groupes élargis qui peuvent compter jusqu'à 13 phrases pour un seul groupe. Le tableau 6 en donne un échantillon pour le début du corpus :

DA	DE	EL	EN	ES	FI	FR	GR	IT	NL	PT	SV
.....
7.3	7.3	7.3	7.3	7.3	7.3	7.3	7.3	7.3	7.3	7.3	7.4
8.1 8.2	8.1 8.2	8.1 8.2	8.1 8.2	8.1 8.2	8.1 8.2	8.1 8.2	8.1 8.2	8.1	8.1 8.2 8.3	8.1 8.2	8.1 8.2
8.3 8.4	8.3 8.4	8.3 8.4	8.3 8.4	8.3 8.4	8.3 8.4	8.3 8.4	8.3 8.4	8.2	8.4 8.5	8.3 8.4	8.3 8.4
9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1
10.1 11.1 11.2	10.1 10.2 11.1 11.2	10.1 11.1 11.2	10.1 10.2 11.1 11.2	10.1 11.1 11.2 11.3	10.1 11.1	10.1 11.1 11.2	10.1 11.1 11.2	10.1 11.1	10.1 10.2 11.1 11.2	10.1 11.1	10.1 10.2 11.1
12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1	12.1
13.1 13.2	13.1	13.1	13.1	13.1	13.1	13.1	13.1	13.1	13.1	13.1	13.1
13.3	13.2	13.2	13.2	13.2	13.2	13.2	13.2	13.2	13.2	13.2	13.2
14.1 14.2	14.1 14.2 14.3	14.1 14.2	14.1 14.2	14.1 14.2	14.1 14.2	14.1 14.2	14.1 14.2	14.1	14.1	14.1 14.2	14.1 14.2 14.3
.....

TABLEAU 6 : Groupes obtenus par clôture transitive des 11 alignements de référence avec le français

4.1 Robustesse vis-à-vis des ruptures de parallélisme

Nous terminerons cette comparaison par une évaluation de la robustesse comparée de ces approches vis-à-vis des ruptures dans le parallélisme des traductions. Pour ce faire, nous avons créé artificiellement des « trous » dans la version française du corpus, en éliminant de façon aléatoire des blocs de phrases. Dans une première expérimentation, nous avons supprimé aléatoirement un bloc d'une seule phrase, en réitérant respectivement 10 fois, 50 fois et 100 fois. Nous avons alors lancé JAM et JAM+GC avec les combinaisons $Comb_{EN-PIVOT}$, $Comb_{FR-PIVOT}$ et $Comb_{TRI}$, en relâchant la contrainte de « parallélisme fort » (pour tenir compte des ruptures dans le chemin d'alignement). Comme nous nous y attendions, c'est la combinaison $Comb_{TRI}$ avec les triplets qui est la plus robuste, l'alignement à 3 langues étant moins sensible aux écarts locaux que l'alignement 2-à-2.

Nous avons également paramétré YASA afin d'éviter une dégradation brutale des résultats, en élargissant l'espace de recherche⁶. On obtient les résultats suivants :

Nb. blocs supprimés	Vanilla P	JAM P	JAM+GC P	YASA P	Vanilla R	JAM R	JAM+GC R	YASA R	Vanilla F	JAM F	JAM+GC F	YASA F
10	0,89	0,97	0,93	0,95	0,91	0,81	0,95	0,95	0,90	0,88	0,90	0,95
50	0,77	0,97	0,89	0,92	0,84	0,80	0,95	0,94	0,80	0,88	0,92	0,93
100	0,62	0,93	0,80	0,88	0,80	0,74	0,90	0,92	0,66	0,83	0,84	0,90

TABLEAU 7 : Résultats comparés de Vanilla, JAM ($Comb_{TRI} + GC$) et YASA pour le corpus français dégradé (blocs de taille 1)

Ces résultats montrent que la précision de JAM se maintient à un niveau élevé et se dégrade légèrement pour 100 phrases supprimées – ceci étant dû à l'étape de complétion, basée sur les alignements de GC, qui gèrent mal ce cas de figure (les probabilités de suppression étant a priori très faibles). En effet, avant l'étape de complétion, JAM obtient une précision de 98% pour un rappel de 24 %. Globalement, pour le bi-alignement complet, si JAM+GC résiste mieux à la dégradation du corpus que Vanilla, YASA obtient à nouveau de meilleures performances, avec une F-mesure au-dessus de 90 %, malgré une diminution de 5 points.

Dans une deuxième expérimentation, nous avons étudié l'effet de la taille des blocs supprimés : cette fois nous ne supprimons qu'un seul bloc, comportant respectivement 10, 50, 100, 200 et 300 phrases.

Taille du bloc supprimé	Vanilla P	JAM+GC P	YASA P	Vanilla R	JAM+GC R	YASA R	Vanilla F	JAM+GC F	YASA F
10	0,82	0,93	0,94	0,85	0,95	0,93	0,83	0,94	0,94
50	0,45	0,94	0,87	0,50	0,95	0,82	0,47	0,95	0,84
100	0,54	0,94	0,94	0,62	0,93	0,94	0,57	0,94	0,94
200	0,25	0,94	0,92	0,33	0,94	0,93	0,28	0,94	0,93
300	0,06	0,93	0,84	0,08	0,93	0,87	0,07	0,93	0,85

TABLEAU 8 : Evolution des résultats en fonction de la taille des blocs supprimés

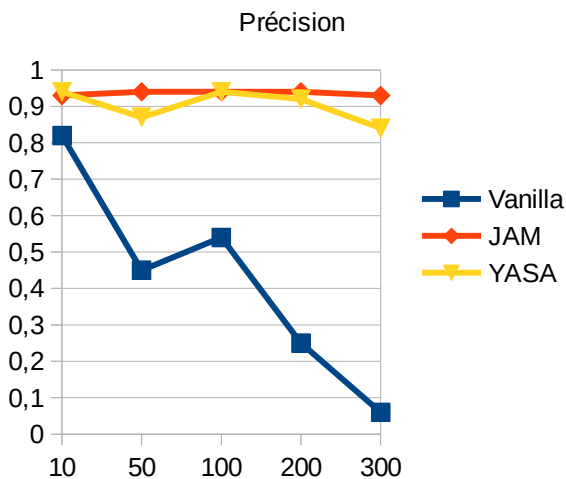


FIGURE 3 : Evolution de la précision en fonction de la taille des blocs supprimés

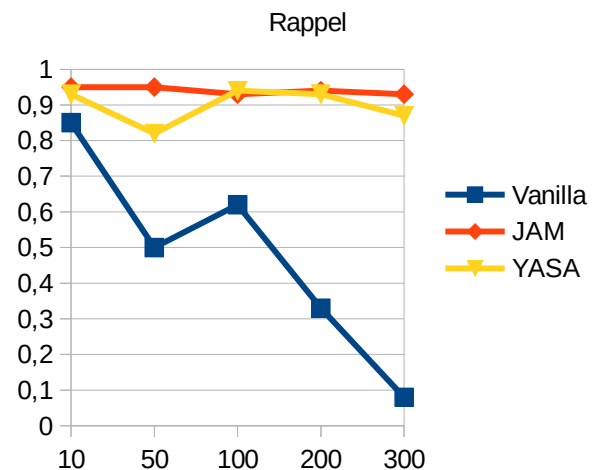


FIGURE 4 : Evolution du rappel en fonction de la taille des blocs supprimés

Cette fois c'est JAM+GC qui se maintient à un niveau élevé, avec une F-mesure presque inchangée. YASA semble plus instable par rapport à ce type de rupture, étant donné la dégradation assez nette observée pour la suppression de 50

⁶ Avec l'option -b 300, pour fixer à 300 phrases le rayon de recherche autour des points d'ancrage.

phrases (à l'instar de Vanilla). Par ailleurs, notons que le paramétrage spécifique de YASA a un coût algorithmique assez important : cette fois l'alignement de tous les couples est effectué en 600 s., alors que rien n'a changé du côté de JAM sur ce plan, qui prend un peu plus de 200 s.

5 Conclusion

Dans cette étude, nous avons proposé un cadre algorithmique pour la mise en œuvre d'une véritable méthode de multi-alignement, destinée à aligner simultanément toutes les versions d'un corpus multi-parallèle (i.e. comportant plus de deux langues). Notre algorithme repose sur une structure de données spécifique : un treillis de multi-alignement, où l'on définit pour chaque langue un chaînage spécifique des appariements partiels avec d'autres langues. L'algorithme itératif proposé permet de tirer parti dans un premier temps des transfuges, puis des cognats, pour apparier des phrases de combinaisons de langues déterminées. Une étape de complétion basée sur l'algorithme de Gale & Church (1991) est enfin appliquée, tant pour remplir le treillis de façon optimale que pour produire, si besoin est, des alignements complets deux à deux.

Dans une expérimentation sur un multi-texte de 11 langues, nous montrons qu'un tuilage des couples de langues suffit à obtenir des résultats satisfaisants, même s'il est possible d'optimiser ce tuilage en se basant sur les densités de cognats propres à chaque couple de langues. Nos résultats montrent par ailleurs que le multi-alignement produit de meilleurs résultats qu'une simple méthode bilingue telle que celle de Gale & Church (1991), et paraît adapté pour fournir des pré-alignements de qualité – avec seulement des appariements 1:1 – pour guider dans un second temps des méthodes d'alignement bilingues destinées à extraire des alignements complets.

Par la suite, nous avons montré que la prise en compte de plusieurs langues dans la même structure de données rend l'alignement, grâce au tuilage, très robuste face aux ruptures de parallélisme : que l'on supprime de nombreuses phrases disséminées çà et là, ou des blocs de grande taille, le pré-alignement stocké dans le treillis de multi-alignement reste d'une grande précision. Les comparaisons avec l'aligneur YASA montrent que celui-ci reste supérieur dans la tâche d'alignement bilingue proprement dite, sauf dans les cas de ruptures importantes de parallélisme, où JAM se révèle plus robuste et plus rapide. Dans des cas de figure où de nombreuses langues sont disponibles, et où les ruptures de parallélisme sont fréquentes, le multi-alignement peut donc constituer une alternative crédible, en termes de robustesse et de fiabilité.

Notons enfin que le multi-alignement présente l'avantage de fournir en sortie une structure de données compacte renfermant un grand nombre de couples – avec une complexité en espace pour le stockage des résultats bien meilleure (en $O(n)$ pour n langues, contre $O(n^2)$ dans le cas bilingue) : un seul fichier au format TMX ou CesAlign peut renfermer les appariements de n langues - avec un rappel oscillant entre 80 % et 90 % vu qu'on n'y retient que les alignements 1:1.

Dans des travaux futurs, nous envisageons d'intégrer l'approche de YASA, qui combine densité de cognats et longueurs de phrases dans la phase de programmation dynamique, en sortie de JAM, pour améliorer l'extraction d'alignements bilingues autour des points d'ancrage. Par ailleurs, le corpus sur lequel nous avons mené cette étude est de taille modeste : il serait intéressant de l'élargir à un corpus plus volumineux et diversifié. Il serait utile, également, de proposer un cadre algorithmique pour étendre l'idée du multi-alignement à l'alignement sous-phrastique – en tirant réellement partie du principe de transitivité, contrairement aux méthodes multilingues proposées jusqu'ici (cf. Lardilleux, 2010). On pourra alors pleinement confirmer que pour l'alignement, tout comme pour la désambiguïsation multilingue, plusieurs langues valent mieux que deux.

Remerciements

Merci à Bettina Schrader, avec qui j'avais commencé à explorer cette piste de recherche il y a quelques années dans le cadre du projet Carmel.

Références

- BROWN P., LAI J., MERCER R. (1991). Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, ACL-91*, Morristown, NJ, 169-176.
- BRAUNE, FABIENNE AND ALEXANDER FRASER. (2010). Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of 23rd COLING*, 81-89.

- CHEN, B., EL-BÈZE, M., HADDARA, M., KRAIF, O., MOREAU DE MONTCHEUIL, G. (2005). Contextes multilingues alignés pour la désambiguïsation sémantique : une étude expérimentale, *Actes de TALN-RECITAL 2005*, 6-10 juin 2005, Dourdan, vol. 1, 415-420.
- CHIAO Y.-C., KRAIF O., LAURENT D., NGUYEN T. M. H., SEMMAR N., STUCK F., VÉRONIS J., ZAGHOUANI W. (2006). Evaluation of multilingual text alignment systems: the ARCADE II project, *Proceedings of the fifth international conference on Language Resources and Evaluation, LREC 2006*, Genova, May 2006.
- DAGAN I., ITAI A., SHWALL U. (1991). Two Languages Are More Informative Than One. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, ACL-91*, Morristown, NJ, 130-137.
- DANIELSSON P., RIDINGS D. (1997). Practical presentation of a "vanilla" aligner. Presented at the *TELRI Workshop on Alignment and Exploitation of Texts*. Institute Jožef Stefan, Ljubljana.
- GALE W., CHURCH K. (1991). "A Program for Aligning Sentences in Bilingual Corpora," *Association for Computational Linguistics*, 177-184.
- HEWAVITHARANA S., VOGEL S. (2011). Extracting Parallel Phrases from Comparable Data, *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*.
- KAY M., RÖSCHEISEN M. (1993). Text-translation alignment. *Computational Linguistics* 19, 1 (March 1993), 121-142.
- KRAIF O. (2001). Exploitation des cognats dans les systèmes d'alignement bi-textuel : architecture et évaluation, *TAL* 42 :3, ATALA, Paris, 833-867.
- LAMRAOUI F., LANGLAIS P. (2013). Yet Another Fast and Robust and Open Source Sentence Aligner. Time to Reconsider Sentence Alignment ?, *Proceedings of the XIV Machine Translation Summit (Nice, September 2-6, 2013)*, 77-84.
- LARDILLEUX A. (2010). Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle. Thèse de doctorat, sous la direction d'Yves Lepage. Université de Caen, 2010.
- LI, P., SUN M., XUE P. (2010). Fast-champollion: a fast and robust sentence alignment algorithm. In *Proceedings of 23rd COLING*, 710-718.
- LEFEVER E., HOSTE V., DE COCK M. (2013) Five languages are better than one: an attempt to bypass the data acquisition bottleneck for WSD, *CICLing 2013, Part I, LNCS 7816*, Springer-Verlag: Berlin, 343-354.
- MOORE, R.-C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *5th AMTA*, 135-144.
- MUNTEANU D.S., FRASER A., MARCU D. (2004). Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora, *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*.
- SIMARD M. (1999). Text-Translation Alignment: Three Languages Are Better Than Two. *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 2-11.
- SIMARD M., FOSTER G., ISABELLE P. (1992). Using Cognates to Align Sentences in Bilingual Corpora. *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, TMI-92*, Montréal, CCRIT, 67-81.
- TIEDEMANN, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*.
- VARGA D., NÉMETH L., HALÁCSY P., KORNAI A., TRÓN V., NAGY V. (2005). Parallel corpora for medium density languages. In *Proceedings of 3rd RANLP*, 590-596.

Apprentissage discriminant des modèles continus de traduction

Quoc-Khanh Do^{1,2} Alexandre Allauzen^{1,2} François Yvon¹

(1) LIMSI/CNRS, rue John von Neumann, Campus Universitaire Orsay 91 403 Orsay

(2) Université Paris Sud, 91 403 Orsay
prenom.nom@limsi.fr

Résumé. Alors que les réseaux neuronaux occupent une place de plus en plus importante dans le traitement automatique des langues, les méthodes d'apprentissage actuelles utilisent pour la plupart des critères qui sont décorrélés de l'application. Cet article propose un nouveau cadre d'apprentissage discriminant pour l'estimation des modèles continus de traduction. Ce cadre s'appuie sur la définition d'un critère d'optimisation permettant de prendre en compte d'une part la métrique utilisée pour l'évaluation de la traduction et d'autre part l'intégration de ces modèles au sein des systèmes de traduction automatique. De plus, cette méthode d'apprentissage est comparée aux critères existants d'estimation que sont le maximum de vraisemblance et l'estimation contrastive bruitée. Les expériences menées sur la tâche de traduction des séminaires TED Talks de l'anglais vers le français montrent la pertinence d'un cadre discriminant d'apprentissage, dont les performances restent toutefois très dépendantes du choix d'une stratégie d'initialisation adéquate. Nous montrons qu'avec une initialisation judicieuse des gains significatifs en termes de scores BLEU peuvent être obtenus.

Abstract.

Discriminative Learning of Continuous Translation Models.

This paper proposes a new discriminative framework to train translation models based on neural network. This framework relies on the definition of a new objective function that allows us to introduce the evaluation metric in the learning process as well as to consider how the model interacts with the translation system. Moreover, this approach is compared with the state of the art estimation methods, such as the maximum likelihood criterion and the noise contrastive estimation. Experiments are carried out on the English to French translation task of TED Talks. The results show the efficiency of the proposed approach, whereas the initialization has a strong impact. We show that with a tailored initialization scheme significant improvements can be obtained in terms of BLEU scores.

Mots-clés : Modèle neuronal de traduction, traduction automatique par approche statistique, apprentissage discriminant.

Keywords: Neural network based translation model, statistical machine translation, discriminative learning.

1 Introduction

Les modèles neuronaux occupent aujourd'hui dans le traitement automatique des langues (TAL) une place importante car ils permettent grâce à leur caractère continu des avancées significatives dans de nombreux domaines applicatifs. Historiquement, les modèles de langue neuronaux ont été une des premières réalisations marquantes, avec des applications en reconnaissance automatique de la parole (RAP), depuis les travaux pionniers de (Nakamura *et al.*, 1990) jusqu'aux développements ultérieurs de (Bengio *et al.*, 2003; Schwenk, 2007; Mnih & Hinton, 2007; Le *et al.*, 2011; Mikolov *et al.*, 2011). Les modèles neuronaux ont été également appliqués à d'autres tâches complexes de modélisation linguistique, comme par exemple l'analyse syntaxique (Socher *et al.*, 2013), l'estimation de similarité sémantique (Huang *et al.*, 2012), les modèles d'alignement de mots (Yang *et al.*, 2013) ou encore en traduction automatique statistique (TAS) (Le *et al.*, 2012; Kalchbrenner & Blunsom, 2013; Devlin *et al.*, 2014; Cho *et al.*, 2014).

Une des caractéristiques importantes des modèles neuronaux pour le TAL est leur caractère continu. En effet les modèles d'apprentissage probabilistes usuels reposent sur une représentation discrète des unités linguistiques considérées (mots, syntagmes, etc.). Typiquement, pour un modèle de traduction à base de segments, l'occurrence d'un segment est considérée comme la réalisation d'une variable aléatoire discrète, dont l'espace de réalisation est l'ensemble des segments

observés dans les données d'apprentissage. Au sein de cet espace, il n'existe aucune relation entre les éléments permettant de modéliser une notion de similarité, par exemple sémantique ou syntaxique. Le caractère très inégal des distributions d'occurrences dans les textes implique que les modèles résultants sont souvent estimés à partir de petits nombres d'occurrences, qu'ils possèdent une faible capacité de généralisation et que la modélisation du contexte est très coûteuse et donc souvent à horizon très limité.

Par opposition, les modèles neuronaux (Bengio *et al.*, 2003) se caractérisent par une méthode d'estimation alternative qui se fonde sur une représentation *continue* des unités qu'ils modélisent et en particulier des mots¹. Dans le cas par exemple d'un modèle de langue, chaque mot du vocabulaire est représenté comme un point dans un espace métrique. La probabilité n -gramme d'un mot est alors une fonction des représentations continues des mots qui composent son contexte. Ces représentations, ainsi que les paramètres de la fonction d'estimation, sont apprises conjointement par un réseau de neurones multi-couches ; une stratégie d'estimation qui permet que les mots partageant des similarités distributionnelles aient des représentations proches. Ainsi, ce type de modèle introduit la notion de similarité entre mots et son exploitation permet une meilleure exploitation des données textuelles. L'intégration de ce type de modèle a permis des améliorations systématiques et significatives des performances en RAP et en TAS (Schwenk, 2007; Le *et al.*, 2011, 2012). Les représentations continues peuvent de plus servir à de nombreuses tâches, comme par exemple l'étiquetage en parties du discours et en rôle sémantique (voir (Turian *et al.*, 2010; Collobert *et al.*, 2011) pour une vue d'ensemble).

De nombreux travaux récents proposent différents types de modèles de traduction. Une part importante est dédiée aux modèles n -grammes de traduction (Schwenk *et al.*, 2007; Le *et al.*, 2012; Devlin *et al.*, 2014). Néanmoins ces travaux ont en commun d'apprendre les modèles de manière à maximiser la vraisemblance mesurée sur les données d'apprentissage. Or ce critère est peu corrélé avec d'une part les métriques utilisées pour évaluer la traduction et d'autre part l'intégration de ces modèles au sein des systèmes de TAS. De plus, cet estimateur oblige le modèle à être normalisé ce qui représente un coût computationnel prohibitif étant donné les espaces de réalisation utilisés. Il est alors nécessaire d'avoir recours à des solutions permettant d'alléger ce coût, comme l'utilisation d'une couche de sortie structurée ou l'usage d'un critère alternatif permettant de contourner cette contrainte.

Les contributions de cet article sont d'une part de proposer un cadre discriminant pour l'apprentissage des modèles continus de traduction permettant d'orienter l'optimisation du modèle vers les difficultés du système de TAS et donc d'apprendre à discriminer les hypothèses considérées selon la métrique utilisée lors de l'évaluation. D'autre part, cette approche est comparée à deux méthodes d'estimation compétitives : le maximum de vraisemblance, et l'estimation contrastive bruitée. Les résultats expérimentaux montrent des gains significatifs en termes de scores BLEU, et donc l'intérêt d'un tel cadre d'apprentissage pour la TAS. Le reste de l'article est organisé de la manière suivante : la section 2 introduit les modèles continus de traduction qui seront utilisés dans ces travaux ; puis les différentes méthodes d'apprentissage étudiées sont décrites à la section 3, avec en particulier la méthode discriminante ; les résultats expérimentaux sont enfin présentés à la section 4.

2 Modèles neuronaux pour la traduction automatique

Cette section propose une vue d'ensemble des modèles continus de traduction tels que nous allons les utiliser dans ces travaux. Si ce type de modèle s'intègre naturellement dans l'approche n -gramme en traduction automatique, il peut également être utilisé avec les approches usuelles à base de segments (Do *et al.*, 2014b). Pour plus de détails sur ces modèles et leur intégration, le lecteur peut se reporter à (Le *et al.*, 2012; Schwenk, 2012).

2.1 Approche n -gramme en traduction automatique

L'approche n -gramme en traduction automatique est une variante de l'approche à base de segments (ou *phrase-based*) (Zens *et al.*, 2002; Koehn, 2010). Décrite dans (Casacuberta & Vidal, 2004) puis (Mariño *et al.*, 2006; Crego & Mariño, 2006), elle s'en distingue par une décomposition spécifique de la probabilité jointe d'une paire de phrases parallèles où l'on suppose que la phrase source a été réordonnée au préalable. Ainsi, notons $P(s, t)$ cette probabilité jointe, où s est une phrase source de I mots (s_1, \dots, s_I) réordonnés, et t la phrase cible associée et composée de J mots cibles (t_1, \dots, t_J) . Cette paire de phrases est décomposée en L unités bilingues appelées *tuples*, $(s, t) = (u_1, \dots, u_L)$. Une illustration de cette décomposition est donnée Figure 1.

1. Les modèles neuronaux sont souvent qualifiés de modèles continus.

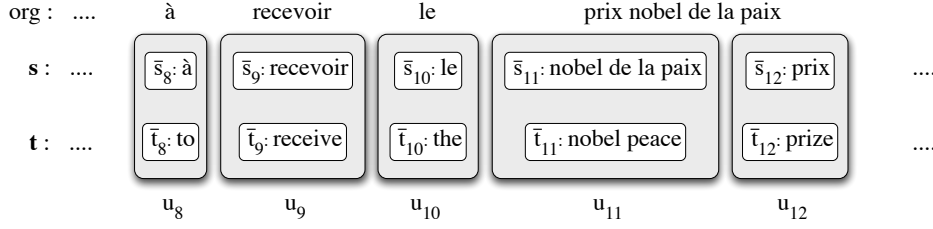


FIGURE 1: Extrait d'une paire de phrases parallèles segmentée. La phrase source originale (*org*) est indiquée au dessus de la phrase source réordonnée *s* et de la phrase cible *t*. La paire de phrases (*s*, *t*) est décomposée en une séquence de *L* unités bilingues (*tuples*) u_1, \dots, u_L . Chaque tuple u_i associe un segment source à un segment cible : \bar{s}_i et \bar{t}_i .

$$P \left(\begin{array}{|c|c|c|} \hline \bar{s}_{11}: \text{nobel de la paix} & \bar{s}_9: \text{recevoir} & \bar{s}_{10}: \text{le} \\ \hline \bar{t}_{11}: \text{nobel peace} & \bar{t}_9: \text{receive} & \bar{t}_{10}: \text{the} \\ \hline \end{array} \right) = \frac{P \left(\begin{array}{|c|} \hline \bar{t}_{11}: \text{nobel peace} \\ \hline \end{array} \middle| \begin{array}{|c|c|c|c|c|} \hline \bar{s}_{11}: \text{nobel de la paix} & \bar{s}_9: \text{recevoir} & \bar{s}_{10}: \text{le} & \bar{t}_9: \text{receive} & \bar{t}_{10}: \text{the} \\ \hline \end{array} \right)}{P \left(\begin{array}{|c|} \hline \bar{s}_{11}: \text{nobel de la paix} \\ \hline \end{array} \middle| \begin{array}{|c|c|c|c|c|} \hline \bar{s}_9: \text{recevoir} & \bar{s}_{10}: \text{le} & \bar{t}_9: \text{receive} & \bar{t}_{10}: \text{the} \\ \hline \end{array} \right)}$$

FIGURE 2: Exemple de décomposition en segment source et cible d'une paire de phrases parallèles sous l'hypothèse 3-gramme. Reprenant l'exemple de la figure 1, il s'agit de prédire le u_{11} connaissant u_9 et u_{10} .

Dans cette modélisation, les *tuples* sont les unités élémentaires de traduction², représentant une correspondance $u = (\bar{s}, \bar{t})$ entre une séquence \bar{s} de mots sources et une séquence de mots cibles \bar{t} . En utilisant l'hypothèse markovienne, la probabilité jointe peut être factorisée de la manière suivante :

$$P(\mathbf{s}, \mathbf{t}) = \prod_{i=1}^L P(u_i | u_{i-n+1}^{i-1}), \quad (1)$$

où u_{i-n+1}^{i-1} représente la séquence de tuples $u_{i-n+1}, \dots, u_{i-1}$. Le modèle complet d'une paire de phrases parallèles contient donc les variables latentes précisant d'une part le réordonnancement de la phrase source, ainsi que la segmentation en unités bilingues. Ces variables latentes définissent la dérivation de la phrase source qui génère la phrase cible. Elles sont ignorées par la suite afin d'alléger les notations. Comme détaillé dans (Mariño *et al.*, 2006; Crego & Mariño, 2006), ces variables latentes sont inférées lors de la phase d'apprentissage à partir des données parallèles alignées automatiquement et ce en deux étapes : pour chaque paire de phrases parallèles, la phrase source est d'abord réordonnée de manière à suivre l'ordre des mots de la phrase cible, puis la segmentation en unités bilingues est effectuée.

Le modèle de traduction ainsi défini est un modèle de séquences utilisant l'hypothèse de *n*-gramme. La différence avec les modèles de langue monolingues est que les unités manipulées ne sont plus les mots mais les tuples. L'espace de réalisation considéré est alors bien plus grand qu'un inventaire monolingue de mots, alors que les données d'apprentissage disponibles se réduisent aux données parallèles. Ainsi, le caractère parcimonieux des données textuelles en général et des données parallèles en particulier rend difficile une estimation directe de ce type de modèle. Une solution est de décomposer les tuples en unités plus petites, comme, par exemple, en distinguant la partie source de la partie cible. L'équation (1) peut ainsi être décomposée de deux manières différentes :

$$\begin{aligned} P(u_i | u_{i-n+1}^{i-1}) &= P(\bar{t}_i | \bar{s}_{i-n+1}^i, \bar{t}_{i-n+1}^{i-1}) P(\bar{s}_i | \bar{s}_{i-n+1}^{i-1}, \bar{t}_{i-n+1}^{i-1}) \\ &= P(\bar{s}_i | \bar{t}_{i-n+1}^i, \bar{t}_{i-n+1}^{i-1}) P(\bar{t}_i | \bar{s}_{i-n+1}^{i-1}, \bar{t}_{i-n+1}^{i-1}) \end{aligned} \quad (2)$$

Considérons, par exemple, la première décomposition. Une illustration en est donnée à la figure 2. Désormais, deux espaces de réalisation sont impliqués, un par langue, qui recensent l'ensemble des segments. Il est encore possible de réduire ces espaces de réalisation en décomposant les segments en séquence de mots. Le modèle obtenu considère alors

2. Les tuples sont assimilables aux paires de segments ou bissegments (*phrase pairs*) utilisés dans l'approche plus classique à base de segments.

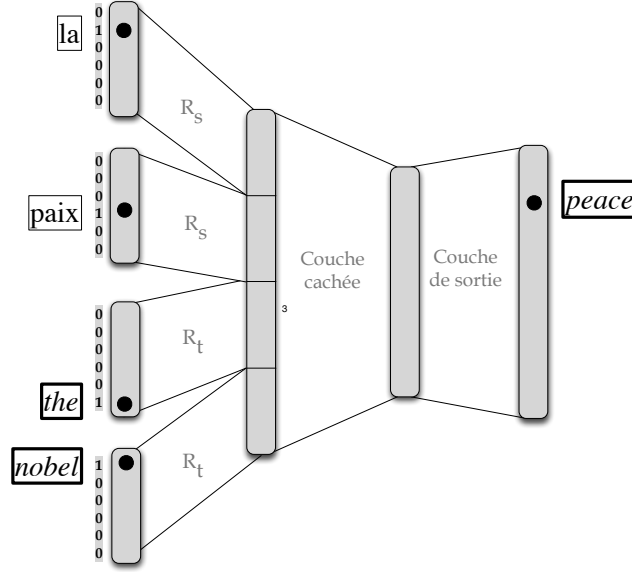


FIGURE 3: Architecture neuronale pour l’estimation des distributions n -grammes bilingues (ici $n = 3$). Cette figure illustre l’estimation de la distribution $P(\bar{t}_i | \bar{s}_{i-n+1}^i, \bar{t}_{i-n+1}^{i-1})$ relativement à l’exemple de la figure 1.

deux séquences de mots, l’une cible et l’autre source, synchronisées sur la segmentation en unités de traduction et dont la partie source a été réordonnée au préalable. Cela correspond à un modèle n -gramme bilingue de mots tel qu’il est initialement décrit dans (Le *et al.*, 2012) et étendu dans (Devlin *et al.*, 2014).

2.2 Architecture neuronale des modèles de traduction

L’estimation des distributions n -grammes peut être réalisée par des réseaux de neurones multi-couches, comme proposé dans (Bengio *et al.*, 2003; Schwenk, 2007) pour une application monolingue. Une architecture couramment utilisée est l’architecture *feed-forward*, illustrée sur la figure 3. Nous en résumons ici l’idée principale, les détails peuvent être trouvés dans (Le *et al.*, 2012) : les mots du contexte sont d’abord projetés dans un espace continu de représentation, chaque langue ayant sa matrice de projection (R_s et R_t respectivement pour les langues source et cible) ; leur concaténation permet d’obtenir une représentation continue du contexte bilingue ; puis une transformation non-linéaire est appliquée afin de prédire le mot cible grâce à la couche de sortie.

Avec ce type d’architecture, la taille du vocabulaire de sortie est la principale limitation³. Ainsi des solutions ont été proposées concernant spécifiquement la couche de sortie afin de réduire le coût d’inférence et d’apprentissage tout en permettant l’usage de vocabulaire de taille réaliste. La première consiste à structurer la couche de sortie comme proposé par (Mnih & Hinton, 2008). Dans cet article nous utilisons la structure SOUL décrite dans (Le *et al.*, 2011, 2013). Le modèle n -gramme peut être alors considéré comme un modèle neuronal de classe de mots. Une autre solution propose un cadre d’apprentissage pour des modèles non-normalisés, permettant de garder une couche de sortie de forme conventionnelle. Cette approche nommée estimation contrastive bruitée (ou NCE pour *noise contrastive estimation*) a été introduite par (Gutmann & Hyvärinen, 2010), puis appliquée aux modèles de langues par (Mnih & Teh, 2012), enfin intégrée à un système de traduction dans (Vaswani *et al.*, 2013). Cette approche sera détaillée à la section 3.1. Dans les deux cas, le modèle neuronal attribue un score positif à un mot w dans son contexte \mathbf{c} noté $\mathbf{b}_\theta(w, \mathbf{c})$, où θ représente l’ensemble des paramètres du modèle à estimer. Concrètement, ce score positif est l’exponentiel de l’activation de la dernière couche linéaire du modèle, $\mathbf{b}_\theta(w, \mathbf{c}) = \exp(\mathbf{a}_\theta(w, \mathbf{c}))$. Les scores $\mathbf{b}_\theta(\cdot)$ peuvent ensuite être normalisés de manière efficace dans le cas du modèle SOUL, ou utilisés tels quels dans le cas d’un modèle NCE.

3. La majeure partie du coût computationnel se situe en effet au niveau de la couche de sortie, où il est nécessaire de normaliser la distribution en effectuant la somme sur l’ensemble du vocabulaire.

3 Méthodes d'apprentissage des modèles de traduction

Les modèles de traduction neuronaux décrits dans la section 2 sont habituellement appris en maximisant la log-vraisemblance, ou plus récemment en utilisant l'estimation contrastive bruitée. Or ces critères d'apprentissage n'ont qu'un lien lointain avec, d'une part, leur utilisation usuelle en traduction automatique et, d'autre part, avec les métriques d'évaluation. En effet, à cause du coût computationnel qu'ils impliquent, les modèles neuronaux sont le plus souvent utilisés en post-traitement d'un système de traduction conventionnel. Leur rôle est alors d'aider le système à trier un ensemble d'hypothèses en se basant sur une mesure automatique de qualité de la traduction, la plupart du temps le score BLEU (Papineni *et al.*, 2002). Cette étape est en général nommée *N-best reranking*, soit la réévaluation des N -meilleures hypothèses.

Dans cette section, nous commençons par décrire les deux critères d'apprentissage habituels des modèles de traduction neuronaux, puis nous formalisons (§ 3.2) un algorithme d'apprentissage discriminant visant à estimer directement les paramètres du modèle de traduction, de manière à optimiser l'étape de réévaluation des N -meilleures hypothèses. Cette méthode s'appuie sur la définition d'une fonction objectif que nous présentons dans un troisième temps (§ 3.3).

3.1 Maximum de vraisemblance et estimation contrastive bruitée

Traditionnellement, les modèles neuronaux de traduction sont entraînés de manière à maximiser la vraisemblance. En pratique, les données d'apprentissage sont présentées comme un ensemble de n -grammes \mathcal{S}_n , et la fonction objectif à minimiser⁴ est la suivante :

$$\mathcal{L}_{cll}(\theta, \mathcal{S}_n) = \sum_{(w, \mathbf{c}) \in \mathcal{S}_n} -\log \mathbf{p}_\theta(w|\mathbf{c}) + \mathcal{R}(\theta), \quad (3)$$

où $\mathcal{R}(\theta)$ est le terme de régularisation L_2 défini par $\mathcal{R}(\theta) = \gamma \times \frac{\|\theta\|^2}{2}$ et γ est l'hyper-paramètre associé. Ce critère $\mathcal{L}_{cll}(\theta, \mathcal{S}_n)$ correspond en fait à la somme négative des log-probabilités conditionnelles des n -grammes contenus dans les données d'apprentissage. Pour calculer cette fonction objectif, il est nécessaire de normaliser la sortie du réseau de neurones sur tous les mots du vocabulaire \mathcal{V} , selon :

$$\mathbf{p}_\theta(w|\mathbf{c}) = \frac{\mathbf{b}_\theta(w, \mathbf{c})}{\sum_{w' \in \mathcal{V}} \mathbf{b}_\theta(w', \mathbf{c})},$$

où $\mathbf{b}_\theta(w, \mathbf{c}) = \exp(\mathbf{a}_\theta(w, \mathbf{c}))$ et $\mathbf{a}_\theta(w, \mathbf{c})$ désigne l'activité du neurone de sortie associé au mot w . La minimisation de $\mathcal{L}_{cll}(\theta, \mathcal{S}_n)$ se fait par descente de gradient stochastique. Néanmoins, le coût de la normalisation peut être prohibitif pour les tailles de vocabulaires typiquement utilisées en traduction automatique, qui contiennent des dizaines, voire des centaines de milliers d'entrées. Dans cet article, nous utilisons le modèle SOUL proposé par (Le *et al.*, 2011) qui, grâce à une couche de sortie structurée en arbre, permet de ramener le temps de calcul à des niveaux raisonnables.

Une approche différente permet de contourner le calcul induit par la normalisation : l'estimation contrastive bruitée ou *Noise Contrastive Estimation* (Gutmann & Hyvärinen, 2010). L'idée principale est de reformuler le problème comme une tâche de classification binaire entre, d'une part, les exemples positifs rencontrés dans les données d'apprentissage et, d'autre part, des exemples négatifs générés artificiellement selon une distribution de bruit $\mathbf{p}_N(\cdot)$. Soit \mathcal{X}^w la variable aléatoire binaire indiquant si le mot w est un exemple positif ou négatif. Nous faisons de plus l'hypothèse (justifiée ci-dessous) que les échantillons négatifs sont *a priori* K fois plus fréquents que les positifs, alors les probabilités *a priori* des deux événements sont données par :

$$\mathbf{p}(\mathcal{X}^{w'} = 1) = \frac{1}{K+1}; \mathbf{p}(\mathcal{X}^{w'} = 0) = \frac{K}{K+1}$$

En supposant de plus que \mathbf{p}_θ est une bonne approximation de la distribution empirique des exemples positifs, il est possible d'écrire :

$$\begin{aligned} \mathbf{p}(w'|\mathbf{c}, \mathcal{X}^{w'} = 1) &= \mathbf{p}_\theta(w'|\mathbf{c}) \\ \mathbf{p}(w'|\mathbf{c}, \mathcal{X}^{w'} = 0) &= \mathbf{p}_N(w'), \end{aligned}$$

4. Dans la suite de cet article nous adoptons la convention habituelle en apprentissage automatique qui consiste à formuler l'apprentissage comme la minimisation d'une fonction objectif. Ainsi maximiser la vraisemblance est équivalent à minimiser le critère défini par l'équation (3).

puis de déduire, en appliquant le théorème de Bayes, les probabilités à posteriori suivantes :

$$\begin{aligned} \mathbf{p}(\mathcal{X}^{w'} = 1 | w', \mathbf{c}) &= \frac{\mathbf{p}_\theta(w' | \mathbf{c})}{\mathbf{p}_\theta(w' | \mathbf{c}) + K \mathbf{p}_N(w')} \\ \mathbf{p}(\mathcal{X}^{w'} = 0 | w', \mathbf{c}) &= \frac{K \mathbf{p}_N(w')}{\mathbf{p}_\theta(w' | \mathbf{c}) + K \mathbf{p}_N(w')}. \end{aligned} \quad (4)$$

La fonction objectif à minimiser devient alors l'espérance de $-\log(\mathbf{p}(\mathcal{X}^{w'} | w', \mathbf{c}))$ sur l'ensemble d'exemples constitué d'un unique exemple positif (w), auxquels sont associés K exemples négatifs $\{w_1^*, \dots, w_K^*\}$. La fonction objectif s'écrit alors de la manière suivante (en ré-intégrant tous les mots observés) :

$$\mathcal{L}_{nce}(\theta, \mathcal{S}_n) = \sum_{(w, \mathbf{c}) \in \mathcal{S}_n} \left[-\log \frac{\mathbf{p}_\theta(w | \mathbf{c})}{\mathbf{p}_\theta(w | \mathbf{c}) + K \mathbf{p}_N(w)} - \sum_{i=1}^K \log \frac{K \mathbf{p}_N(w_i^*)}{\mathbf{p}_\theta(w_i^* | \mathbf{c}) + K \mathbf{p}_N(w_i^*)} \right] + \mathcal{R}(\theta), \quad (5)$$

où (w, \mathbf{c}) est un n -gramme issu des données d'apprentissage et $(w_i^*)_{i=1}^K$ l'ensemble des K exemples négatifs qui lui sont associés. Ces exemples négatifs sont tirés aléatoirement de la distribution de bruit. Dans (Gutmann & Hyvärinen, 2010; Mnih & Teh, 2012), les auteurs insistent sur l'importance du choix de cette distribution de bruit. Il semble en effet nécessaire qu'elle soit proche de la distribution empirique, tout en permettant un échantillonnage efficace. Ainsi, le choix le plus répandu est d'utiliser la distribution unigramme sur les données d'apprentissage. Dans notre cas, il s'agit d'une distribution unigramme sur les mots cibles.

Notons que, dans l'équation (4), le terme $\mathbf{p}_\theta(w | \mathbf{c})$ apparaît au numérateur et au dénominateur. En faisant l'approximation que \mathbf{p}_θ et \mathbf{p}_N ont des normalisations proches, il est possible de se débarrasser du terme de normalisation, et donc de remplacer (avantageusement) $\mathbf{p}_\theta(w | \mathbf{c})$ par $\mathbf{b}_\theta(w, \mathbf{c})$. De plus, il est possible de montrer que lorsque K tend vers l'infini, cette fonction objectif tend vers \mathcal{L}_{cll} . Ainsi l'estimation contrastive bruitée est une méthode d'optimisation formalisant le calcul approché de la constante de normalisation en échantillonnant K exemples négatifs au lieu d'effectuer la somme sur l'ensemble du vocabulaire.

3.2 Méthode discriminante d'apprentissage pour la traduction automatique

Malgré les progrès récents, l'inférence avec un réseau de neurones reste trop coûteuse pour que ce type de modèle puisse être intégré au décodage aussi facilement que les modèles de langue discrets utilisés dans les systèmes de traduction automatique⁵. L'usage est donc d'utiliser ces modèles lors d'une seconde étape de réévaluation des N -meilleures hypothèses.

Afin de définir ce cadre, supposons que pour chaque phrase source \mathbf{s} à traduire, le décodeur génère une liste des N meilleures hypothèses $\{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N\}$. Chaque hypothèse \mathbf{h}_i est constituée d'une phrase cible \mathbf{t}_i et de la dérivation \mathbf{a}_i qui l'a engendrée⁶. Elle est évaluée par le système de traduction grâce à la fonction suivante :

$$F_\lambda(\mathbf{s}, \mathbf{h}) = \sum_{k=1}^M \lambda_k f_k(\mathbf{s}, \mathbf{h}), \quad (6)$$

où M fonctions caractéristiques (f_k) sont pondérées par un jeu de poids λ . Les fonctions caractéristiques utilisées dans cet article sont similaires à celles que l'on peut trouver dans les systèmes usuels à base de segments (voir (Crego *et al.*, 2011) pour plus de précisions).

L'introduction d'un modèle continu lors de l'étape de réévaluation des hypothèses se traduit par l'ajout à $F_\lambda(\cdot)$ d'une fonction caractéristique supplémentaire $f_\theta(\mathbf{s}, \mathbf{h})$, qui varie selon le modèle utilisé :

$$f_\theta(\mathbf{s}, \mathbf{h}) = \begin{cases} \log \mathbf{p}_\theta(\mathbf{s}, \mathbf{h}) = \sum_{(w, \mathbf{c}) \in (\mathbf{s}, \mathbf{h})} \log \mathbf{p}_\theta(w | \mathbf{c}) & \text{pour le modèle SOUL,} \\ \sum_{(w, \mathbf{c}) \in (\mathbf{s}, \mathbf{h})} \log \mathbf{b}_\theta(w, \mathbf{c}) = \sum_{(w, \mathbf{c}) \in (\mathbf{s}, \mathbf{h})} \mathbf{a}_\theta(w, \mathbf{c}) & \text{pour le modèle NCE.} \end{cases} \quad (7)$$

5. Notons néanmoins les tentatives récentes d'intégration des modèles neuronaux de traduction dans le décodeur (Niehues & Waibel, 2012; Vaswani *et al.*, 2013; Devlin *et al.*, 2014).

6. \mathbf{a}_i regroupe l'ensemble des variables latentes du processus de traduction. Dans le cas d'un système de traduction n -gramme, il s'agit du réordonnement de la phrase source et du choix des unités bilingues (cf. § 2.1).

Algorithm 1 Procédure d'optimisation jointe de θ et λ

```

1: Init. de  $\theta$  et  $\lambda$ 
2: Pour chaque itération faire
3:   Pour  $M$  paquets faire ▷  $\lambda$  fixé
4:     Calcul du sous-gradient de  $\mathcal{L}(\theta)$  pour chaque phrase  $s$  du paquet
5:     Mise à jour de  $\theta$ 
6:   Fin Pour
7:   Mise à jour de  $\lambda$  en utilisant le dev. ▷  $\theta$  fixé
8: Fin Pour

```

Dans les deux cas, il est nécessaire de prendre la somme sur tous les n -grammes extraits de la dérivation considérée. Comme précédemment, θ désigne le vecteur de paramètres définissant le modèle continu de traduction. Ainsi la fonction d'évaluation devient :

$$G_{\lambda, \theta}(s, h) = F_{\lambda}(s, h) + \lambda_{K+1} f_{\theta}(s, h) \quad (8)$$

Cette fonction dépend à la fois des paramètres θ du modèle continu de traduction et des paramètres de mélange λ de la fonction d'évaluation. Ainsi, dans l'approche que nous proposons, l'optimisation nécessite *d'alterner l'estimation des poids de mélange λ et l'apprentissage des paramètres θ du modèle continu* : la première étape utilise classiquement les données de développement alors que la deuxième utilise les données parallèles d'apprentissage.

Cette procédure d'optimisation est décrite par l'algorithme 1. Les données d'apprentissage sont découpées en paquets de 128 phrases successives. Chacun de ces paquets sert à la mise à jour de θ à λ constant et ces derniers sont réestimés tous les P paquets. Notons que cet algorithme nécessite la définition d'une fonction objectif $\mathcal{L}(\theta)$ pour le modèle continu de traduction qui sera décrite à la section 3.3. Dans cet article, l'optimisation de λ utilise les outils standards, en l'occurrence l'algorithme *K-Best Mira* décrit dans (Cherry & Foster, 2012) et tel qu'il est implémenté dans MOSES⁷.

3.3 Une fonction objectif discriminante

Le critère discriminant d'apprentissage proposé dans cet article s'inspire à la fois des méthodes à vaste marge et des approches de *ranking*. Comme expliqué précédemment, chaque hypothèse de traduction h_i engendrée par le système de traduction est évaluée selon l'équation (8). Mais sa qualité peut également être évaluée selon un critère de qualité de traduction, ici le score BLEU, ou plus précisément selon une approximation du score BLEU au niveau de la phrase, que l'on note $sBLEU(h_i)$. Si h^* désigne l'hypothèse ayant le meilleur score, il est possible de définir un critère visant à maximiser la marge (Freund & Schapire, 1999; McDonald *et al.*, 2005; Watanabe *et al.*, 2007) de la manière suivante :

$$\mathcal{L}_{mm}(\theta, s) = -G_{\lambda, \theta}(s, h^*) + \max_{1 \leq j \leq N} (G_{\lambda, \theta}(s, h_j) + \text{cost}_{\alpha}(h_j)), \quad (9)$$

où $\text{cost}_{\alpha}(h_j) = \alpha(sBLEU(h^*) - sBLEU(h_j))$ représente la fonction de coût et le paramètre α pondère sa contribution. Lorsque $\alpha = 0$, nous retrouvons la fonction objectif du perceptron structuré (Collins, 2002). Si ce critère introduit une marge entre h^* et les autres hypothèses, il existe parmi les autres hypothèses des traductions qui pourraient être acceptables et qu'il conviendrait de considérer autrement qu'en les jugeant mauvaises. Ainsi, une alternative est de s'inspirer du classement par paire (ou *pairwise ranking*) comme le propose le système PRO de Hopkins & May (2011). Supposons que r_i désigne le rang de l'hypothèse h_i lorsque la liste des hypothèses est triée avec comme critère $sBLEU$, il est alors possible de définir la fonction objectif suivante :

$$\mathcal{L}_{pro}(\theta, s) = \sum_{1 \leq i, k \leq N} \mathbb{I}_{\{r_i + \delta \leq r_k, G_{\lambda, \theta}(s, h_i) < G_{\lambda, \theta}(s, h_k)\}} (-G_{\lambda, \theta}(s, h_i) + G_{\lambda, \theta}(s, h_k)). \quad (10)$$

Notons que cette fonction objectif implique un sous-ensemble de $N(N-1)/2$ paires d'hypothèses. En effet, une paire d'hypothèses n'est prise en compte que si les hypothèses qui la composent sont suffisamment éloignées en terme de rang : formellement la différence absolue des rangs doit excéder un seuil prédéfini δ .

Le critère que nous proposons est une combinaison des deux critères précédents. Ce choix s'appuie sur les résultats expérimentaux de (Do *et al.*, 2014a), qui a introduit ces critères dans le cadre de l'adaptation de modèle. Considérons que

7. <http://www.statmt.org/moses/>

pour une paire d'hypothèses $(\mathbf{h}_i, \mathbf{h}_k)$ telle que $r_i + \delta < r_k$, la différence de score $G_{\lambda, \theta}(\mathbf{s}, \mathbf{h}_i) - G_{\lambda, \theta}(\mathbf{s}, \mathbf{h}_k)$ doit être au-delà d'une certaine marge. Comme précédemment, la marge s'exprime grâce à l'approximation du score BLEU au niveau de la phrase et donc via la fonction de coût cost_α . Nous pouvons alors définir le sous-ensembles des hypothèses critiques comme :

$$\mathcal{C}_\delta^\alpha = \{(i, k) : 1 \leq i, k \leq N, r_i + \delta \leq r_k, G_{\lambda, \theta}(\mathbf{s}, \mathbf{h}_i) - G_{\lambda, \theta}(\mathbf{s}, \mathbf{h}_k) < \text{cost}_\alpha(\mathbf{h}_k) - \text{cost}_\alpha(\mathbf{h}_i)\}. \quad (11)$$

La fonction objectif que nous allons utiliser pour apprendre les modèles de traduction continus se définit de la manière suivante :

$$\mathcal{L}_{pro-mm}(\theta, \mathbf{s}) = \sum_{(i, k) \in \mathcal{C}_\delta^\alpha} \text{cost}_\alpha(\mathbf{h}_k) - \text{cost}_\alpha(\mathbf{h}_i) - G_{\lambda, \theta}(\mathbf{s}, \mathbf{h}_i) + G_{\lambda, \theta}(\mathbf{s}, \mathbf{h}_k). \quad (12)$$

Cette fonction objectif ne requiert pas, tout comme le NCE, que le score du modèle neuronal $f_\theta(\mathbf{s}, \mathbf{h})$ inclus dans $G_{\lambda, \theta}(\mathbf{s}, \mathbf{h}_i)$ soit normalisé. Il est donc possible d'apprendre un modèle de type SOUL selon ce critère mais également un modèle non-normalisé de type NCE. Remarquons enfin que si $\alpha = 0$, cette fonction se ramène à celle du classement par paire de l'équation (10).

4 Expériences

Afin de comparer les différentes méthodes d'apprentissage présentées à la section 3, une série d'expériences est menée sur la tâche de traduction automatique anglais vers français des séminaires TED Talks . Cette tâche fait partie de la campagne d'évaluation internationale sur la traduction de la parole organisée dans le cadre des ateliers IWSLT⁸.

4.1 Cadre expérimental

La tâche considérée est la traduction des séminaires TED Talks (Federico *et al.*, 2012) dans leur version transcrite manuellement. Les données parallèles d'apprentissage servant à l'apprentissage des modèles neuronaux contiennent 107058 paires de phrases. Les données de développement et de test contiennent respectivement 934 et 1664 paires de phrases. En suivant (Le *et al.*, 2012), ces données sont échangées ; les coefficients λ sont estimés à partir des 1664 paires de phrases et l'autre corpus sert à l'évaluation. Le critère d'évaluation de la traduction est le score BLEU (Papineni *et al.*, 2002).

Le système de traduction utilise une implémentation libre de l'approche n -gramme⁹ et ses modèles ont été appris à partir de vaste quantité de données bilingues et monolingues dans le cadre de la campagne d'évaluation WMT. Le système est décrit plus précisément dans l'article (Allauzen *et al.*, 2013).

Les modèles continus de traduction sont appris uniquement sur les données TED Talks . Chaque modèle, SOUL et NCE, est initialisé à partir de modèles n -grammes monolingues estimés respectivement sur la partie source et cible du corpus bilingue. Tous les modèles n -grammes continus sont des 10-grammes. Pour l'apprentissage discriminant, le système de traduction est d'abord utilisé pour générer une liste des 300 meilleures hypothèses pour chaque phrase source. Le seuil δ (cf. l'équation (11)) a été empiriquement fixé à 250 en fonction des scores BLEU sur les données de développement. Ces scores servent aussi de choisir la meilleure itération qui correspond au modèle qui est ensuite évalué sur les données de test.

Comme décrit à la section 2.1, il existe deux manières de décomposer la probabilité jointe d'une paire de phrases (voir l'équation (2)), et il est donc possible de définir 4 modèles continus de traduction. Par souci de clarté, seul le modèle $P(\bar{t}_i | \bar{s}_{i-n+1}^i, \bar{t}_{i-n+1}^{i-1})$ est utilisé par la suite. Néanmoins des tendances similaires ont été observées avec les autres modèles.

4.2 Résultats expérimentaux

Afin de comparer les différents critères décrits à la section 3, la première série d'expériences concerne les modèles non-normalisés selon qu'ils soient appris avec le NCE, le critère discriminant que nous proposons ou les deux. Dans tous les

8. International Workshop on Spoken Language Translation : <http://workshop2014.iwslt.org/>

9. perso.limsi.fr/Individu/jmcrego/bincoder

	dev	test
Système de traduction	33,9	27,6
Ajout d'un modèle continu standard		
NCE	35,0	28,8
Ajout d'un modèle continu discriminant		
initialisation aléatoire	34,3	28,4
initialisation monolingue NCE	35,3	29,0
NCE + discriminant	35,4	29,7
Oracle	46,1	39,0

TABLE 1: Comparaisons des résultats obtenus en terme de score BLEU avec différents modèles de traduction continus non normalisés.

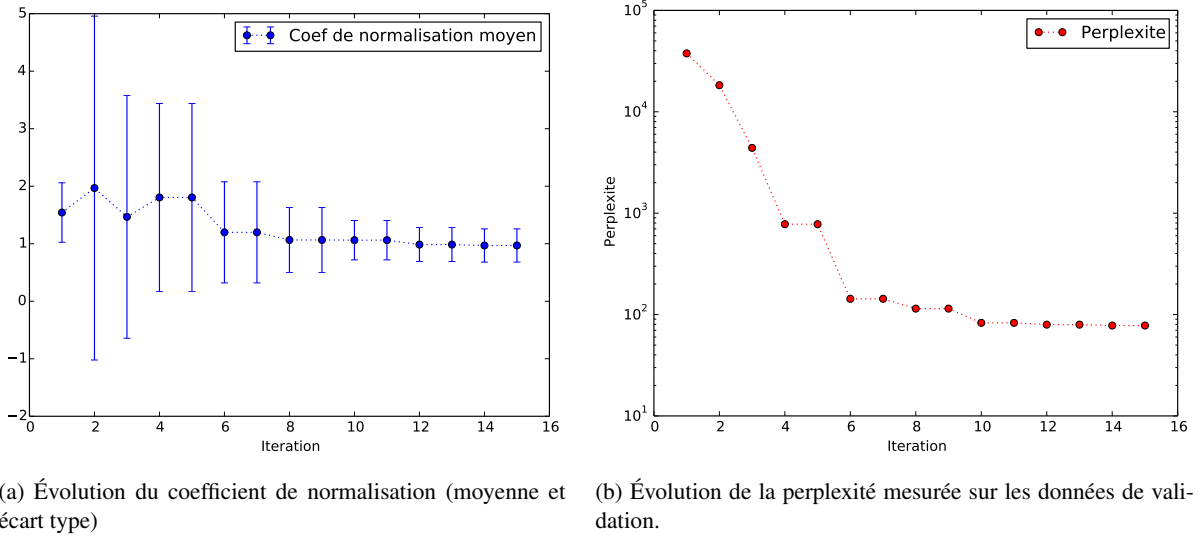


FIGURE 4: Évolution de l'apprentissage d'un modèle NCE au cours du temps.

cas, le modèle neuronal est intégré via son score non-normalisé $f_{\theta}(s, h) = \sum_{(w, c) \in (s, h)} \mathbf{a}_{\theta}(w, c)$. Le tableau 1 rassemble les résultats obtenus avec différents critères d'apprentissage et différentes configurations. Observons d'abord que le modèle continu de traduction appris avec le NCE comme critère permet d'obtenir une amélioration de 1,2 point BLEU et ce malgré l'utilisation d'une distribution de bruit relativement éloignée de la distribution empirique puisqu'il s'agit d'une distribution monolingue.

La seconde partie du tableau regroupe les différentes expériences utilisant le critère discriminant d'apprentissage et montre l'importance de l'initialisation pour ce critère d'apprentissage¹⁰. En partant d'une initialisation aléatoire, le modèle discriminant permet d'obtenir un gain significatif de 0,8 point BLEU mais qui est moindre que celui obtenu avec un modèle NCE. Par contre, si on utilise la même initialisation que celle du modèle NCE (voir section 4.1), le modèle discriminant apporte un gain supplémentaire par rapport au modèle NCE de 0,2 points BLEU. Enfin, le meilleur résultat est obtenu en combinant les deux critères : le modèle discriminant est appris avec en guise d'initialisation le modèle NCE. Avec cette configuration le modèle de traduction permet d'obtenir un gain de 2,1 points BLEU.

Afin de mieux comprendre comment fonctionne l'apprentissage NCE, la figure 4a représente l'évolution des coefficients de normalisation (moyenne et écart type) du modèle au cours des itérations. Un tel coefficient est calculé pour chaque contexte présent dans les données de validation¹¹. Nous observons que la valeur moyenne, ainsi que l'écart type de ce coefficient, convergent rapidement. Néanmoins, l'écart type reste élevé, montrant ainsi que le modèle NCE n'est pas

10. Les modèles neuronaux, par leurs couches cachées donnent lieu à des fonctions objectifs non-convexes. Puisqu'une descente de gradient est utilisée lors de l'optimisation, le point de convergence dépend de l'initialisation.

11. Le coefficient de normalisation se calcule pour un contexte donné par $\sum_{w \in \mathcal{V}} \mathbf{b}_{\theta}(w, c)$.

	dev	test
Système de traduction	33, 9	27, 6
Ajout d'un modèle continu standard		
SOUL	35, 1	28, 9
Ajout d'un modèle continu discriminant		
initialisation aléatoire	33, 8	27, 7
initialisation monolingue SOUL	35, 0	28, 9
SOUL + discriminant	35, 7	29, 3
Oracle	46, 1	39, 0

TABLE 2: Comparaison et utilisation des modèles SOUL dans le cadre discriminant d'apprentissage.

	SOUL	NCE	DISCRIM
Vitesse d'entraînement (mots/second)	1000	1000	
Nombre d'itérations	3	14	
Temps d'entraînement total, init incl. (heures)	9	9	15
Vitesse d'inférence (mots/second)	20000	25000	dépend du modèle

TABLE 3: Vitesse de traitement lors de l'apprentissage et de l'inférence, ainsi que le temps total d'apprentissage (comprenant la phase d'initialisation) des modèles décrits à la section 3. Si les vitesses d'entraînement des modèles SOUL et NCE sont équivalentes, l'inférence avec le modèle NCE est légèrement plus rapide. On note également que même si l'entraînement NCE demande plus d'itérations pour converger, son initialisation est bien plus simple par rapport à celle d'un modèle SOUL qui nécessite de construire une structure d'arbre pour tous les mots du vocabulaire.

normalisé. De plus, la figure 4b représente l'évolution de la perplexité¹² mesurée sur les mêmes données de validation. On constate également une convergence rapide et un comportement similaire à celui d'un modèle estimé selon le maximum de vraisemblance.

La seconde série d'expériences permet de comparer le critère usuel d'apprentissage qui est le maximum de vraisemblance et donc du modèle SOUL. Les résultats sont rassemblés dans le tableau 2. Remarquons tout d'abord qu'il n'y a qu'une différence de 0,1 points BLEU en faveur du modèle SOUL par rapport au modèle NCE. À ce stade, il est important de noter que les deux méthodes, SOUL et NCE, induisent des temps d'apprentissage équivalents pour les modèles de traduction. Ainsi, le caractère normalisé ne semble pas indispensable. Par contre, introduire le score normalisé du modèle SOUL semble moins favorable au cadre discriminant¹³. En effet, en partant d'une initialisation aléatoire, on notera qu'une très faible amélioration par rapport au résultat du système de traduction. De même, en partant d'une initialisation utilisant les modèles SOUL monolingues, l'apprentissage discriminant n'apporte rien par rapport au modèle SOUL bilingue. Le seul véritable gain est obtenu en combinant les deux critères, mais là encore, l'utilisation du NCE permet d'obtenir un meilleur résultat. Enfin, le tableau 3 rassemble les vitesses et temps de calcul liés à l'apprentissage et à l'inférence des différentes méthodes d'apprentissage décrites dans cet article.

5 Conclusions

Dans cet article nous avons proposé un cadre discriminant pour l'apprentissage des modèles neuronaux de traduction. Ce cadre s'appuie sur la définition d'un critère d'optimisation qui permet d'une part d'introduire la mesure servant à évaluer la traduction, et d'autre part de prendre en compte l'état courant du système de base pendant l'entraînement, contrairement aux autres méthodes existantes comme l'estimation au maximum de vraisemblance et l'estimation contrastive bruitée. Ces trois critères sont décrits puis comparés expérimentalement dans le cadre d'une tâche de traduction automatique de l'anglais vers le français des séminaires TED Talks. Les résultats montrent d'une part qu'il est possible d'apprendre un modèle continu de traduction de manière discriminante et d'autre part que le choix de l'initialisation revêt une grande importance. Nous proposons d'ailleurs à ce sujet des solutions efficaces permettant d'obtenir des gains significatifs en termes de scores BLEU. Notamment, la meilleure configuration consiste à enchaîner l'entraînement discriminant sur un

12. Le modèle original n'étant pas normalisé, il est nécessaire d'effectuer cette normalisation pour le calcul de la perplexité.

13. ici $f_{\theta}(\mathbf{s}, \mathbf{h}) = \sum_{(w, \mathbf{c}) \in (\mathbf{s}, \mathbf{h})} \log \mathbf{p}_{\theta}(w|\mathbf{c})$

modèle NCE à scores non-normalisés, ce qui dispense de normaliser les scores sur l'ensemble du vocabulaire. En guise de futurs travaux, il semble intéressant d'explorer ce cadre d'apprentissage pour des paires de langues peu dotées en données parallèles. En effet ce cadre semble permettre une meilleure exploitation des données d'apprentissage.

Références

- ALLAUZEN A., PÉCHEUX N., DO Q. K., DINARELLI M., LAVERGNE T., MAX A., LE H.-S. & YVON F. (2013). LIMSIS @ WMT13. In *Proceedings of the Workshop on Statistical Machine Translation*, p. 62–69, Sofia, Bulgaria.
- BENGIO Y., DUCHARME R., VINCENT P. & JAUVIN C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, **3**, 1137–1155.
- CASACUBERTA F. & VIDAL E. (2004). Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, **30**(3), 205–225.
- CHERRY C. & FOSTER G. (2012). Batch tuning strategies for statistical machine translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*, p. 427–436.
- CHO K., VAN MERRIENBOER B., GULCEHRE C., BAHDANAU D., BOUGARES F., SCHWENK H. & BENGIO Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1724–1734, Doha, Qatar : Association for Computational Linguistics.
- COLLINS M. (2002). Discriminative training methods for hidden Markov models : theory and experiments with perceptron algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1–8.
- COLLOBERT R., WESTON J., BOTTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, **12**, 2493–2537.
- CREGO J. M. & MARIÑO J. B. (2006). Improving statistical MT by coupling reordering and decoding. *Machine Translation*, **20**(3), 199–215.
- CREGO J. M., YVON F. & MARIÑO J. B. (2011). N-code : an open-source bilingual N-gram SMT toolkit. *Prague Bulletin of Mathematical Linguistics*, **96**, 49–58.
- DEVLIN J., ZBIB R., HUANG Z., LAMAR T., SCHWARTZ R. & MAKHOUL J. (2014). Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1370–1380, Baltimore, Maryland.
- DO Q. K., ALLAUZEN A. & YVON F. (2014a). Discriminative adaptation of continuous space translation models. In *International Workshop on Spoken Language Translation (IWSLT 2014)*, Lake Tahoe, USA.
- DO Q. K., HERRMANN T., NIEHUES J., ALLAUZEN A., YVON F. & WAIBEL A. (2014b). The KIT-LIMSIS Translation System for WMT 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, p. 84–89, Baltimore, Maryland, USA : Association for Computational Linguistics.
- FEDERICO M., STÜKER S., BENTIVOGLI L., PAUL M., CETTOLO M., HERRMANN T., NIEHUES J. & MORETTI G. (2012). The IWSLT 2011 evaluation campaign on automatic talk translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) : European Language Resources Association (ELRA)*.
- FREUND Y. & SCHAPIRE R. E. (1999). Large margin classification using the perceptron algorithm. *Machine learning*, **37**(3), 277–296.
- GUTMANN M. & HYVÄRINEN A. (2010). Noise-contrastive estimation : A new estimation principle for unnormalized statistical models. In Y. TEH & M. TITTERINGTON, Eds., *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9, p. 297–304.
- HOPKINS M. & MAY J. (2011). Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, p. 1352–1362, Edinburgh, Scotland, UK.
- HUANG E., SOCHER R., MANNING C. & NG A. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 873–882, Jeju Island, Korea : Association for Computational Linguistics.
- KALCHBRENNER N. & BLUNSOM P. (2013). Recurrent continuous translation models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1700–1709, Seattle, Washington, USA.

- KOEHN P. (2010). *Statistical Machine Translation*. New York, NY, USA : Cambridge University Press, 1st edition.
- LE H.-S., ALLAUZEN A. & YVON F. (2012). Continuous space translation models with neural networks. In *Proceedings of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*, p. 39–48, Montréal, Canada.
- LE H.-S., OPARIN I., ALLAUZEN A., GAUVAIN J.-L. & YVON F. (2011). Structured output layer neural network language model. In *Proceedings of ICASSP*, p. 5524–5527.
- LE H.-S., OPARIN I., ALLAUZEN A., GAUVAIN J.-L. & YVON F. (2013). Structured output layer neural network language models for speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, **21**(1), 197–206.
- MARIÑO J. B., BANCHS R. E., CREGO J. M., DE GISPERS A., LAMBERT P., FONOLLOSA J. A. & COSTA-JUSSÀ M. R. (2006). N-gram-based machine translation. *Computational Linguistics*, **32**(4), 527–549.
- MCDONALD R., CRAMMER K. & PEREIRA F. (2005). Online large-margin training of dependency parsers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 91–98.
- MIKOLOV T., KOMBRINK S., BURGET L., CERNOCKÝ J. & KHUDANPUR S. (2011). Extensions of recurrent neural network language model. In *Proceedings of ICASSP*, p. 5528–5531.
- MNIH A. & HINTON G. E. (2007). Three new graphical models for statistical language modelling. In *ICML*, p. 641–648.
- MNIH A. & HINTON G. E. (2008). A scalable hierarchical distributed language model. In D. KOLLER, D. SCHUURMANS, Y. BENGIO & L. BOTTOU, Eds., *Advances in Neural Information Processing Systems 21*, volume 21, p. 1081–1088.
- MNIH A. & TEH Y. W. (2012). A fast and simple algorithm for training neural probabilistic language models. In *ICML*.
- NAKAMURA M., MARUYAMA K., KAWABATA T. & KIYOHIRO S. (1990). Neural network approach to word category prediction for english texts. In *Proceedings of the 13th conference on Computational linguistics (COLING)*, volume 3, p. 213–218.
- NIEHUES J. & WAIBEL A. (2012). Continuous space language models using restricted Boltzmann machines. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, p. 164–170, Hong-Kong, China.
- PAPINENI K., ROUKOS S., WARD T. & ZHU W. J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics*, p. 311–318.
- SCHWENK H. (2007). Continuous space language models. *Computer Speech and Language*, **21**(3), 492–518.
- SCHWENK H. (2012). Continuous space translation models for phrase-based statistical machine translation. In *Proceedings of COLING 2012 : Posters*, p. 1071–1080, Mumbai, India : The COLING 2012 Organizing Committee.
- SCHWENK H., R. COSTA-JUSSÀ M. & R. FONOLLOSA J. A. (2007). Smooth bilingual n -gram translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 430–438, Prague, Czech Republic.
- SOCHER R., BAUER J., MANNING C. D. & ANDREW Y. N. (2013). Parsing with compositional vector grammars. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 455–465, Sofia, Bulgaria.
- TURIAN J., RATINOV L.-A. & BENGIO Y. (2010). Word representations : A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 384–394, Uppsala, Sweden : Association for Computational Linguistics.
- VASWANI A., ZHAO Y., FOSSUM V. & CHIANG D. (2013). Decoding with large-scale neural language models improves translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1387–1392, Seattle, Washington, USA.
- WATANABE T., SUZUKI J., TSUKADA H. & ISOZAKI H. (2007). Online large-margin training for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) : Cite-seer*.
- YANG N., LIU S., LI M., ZHOU M. & YU N. (2013). Word alignment modeling with context dependent deep neural network. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 166–175, Sofia, Bulgaria.
- ZENS R., OCH F. J. & NEY H. (2002). Phrase-based statistical machine translation. In *KI '02 : Proceedings of the 25th Annual German Conference on AI*, p. 18–32, London, UK : Springer-Verlag.

Utiliser les interjections pour détecter les émotions

Amel Fraisse Patrick Paroubek

LIMSI-CNRS, Bât. 508 Université Paris-Sud, 91403 Orsay Cedex, France
fraise@limsi.fr, pap@limsi.fr

Résumé. Bien que les interjections soient un phénomène linguistique connu, elles ont été peu étudiées et cela continue d'être le cas pour les travaux sur les microblogs. Des travaux en analyse de sentiments ont montré l'intérêt des émoticônes et récemment des mots-dièses, qui s'avèrent être très utiles pour la classification en polarité. Mais malgré leur statut grammatical et leur richesse sémantique, les interjections sont restées marginalisées par les systèmes d'analyse de sentiments. Nous montrons dans cet article l'apport majeur des interjections pour la détection des émotions. Nous détaillons la production automatique, basée sur les interjections, d'un corpus étiqueté avec les émotions. Nous expliquons ensuite comment nous avons utilisé ce corpus pour en déduire, automatiquement, un lexique affectif pour le français. Ce lexique a été évalué sur une tâche de détection des émotions, qui a montré un gain en mesure F1 allant, selon les émotions, de +0,04 à +0,21.

Abstract.

Using interjections for emotion detection

Although interjections have been recognized as linguistic phenomena for a long time, they have somehow been rarely studied and continue to be left aside in works dealing with microblogs. Users of this new kind of communication platforms have popularized widely the use of linguistic constructions, like emoticons or interjections. In spite of their grammatical status and semantic richness for describing emotional states, interjections have been mostly ignored. In this article we show the importance of the role that interjections can play for detecting emotions. We detail how using interjections we have tagged automatically a French microblog corpus with emotion labels. Then we describe how we did deduce automatically from this corpus a fine-grained affective lexicon. The usefulness of the lexicon was evaluated in an emotion recognition task where, depending on the emotion, the F1-measure improvement ranged from +0.04 to +0.21.

Mots-clés : Interjections, détection des émotions, lexique affectif, analyse de sentiments, fouille d'opinions .

Keywords: Interjections, emotion recognition, affective lexicon, sentiment analysis, opinion mining .

1 Introduction

Le microbloggage est devenu aujourd'hui parmi les moyens de communication les plus populaires. En effet, en limitant la longueur de messages à quelques caractères, ce service permet de partager rapidement et facilement une information ou une opinion. Twitter¹ est la plus célèbre plateforme de microbloggage. Elle permet d'envoyer et de recevoir ce que l'on nomme des *tweets*, c'est-à-dire des messages courts dont la longueur est limitée à 140 caractères. D'après les derniers chiffres avancés par la compagnie de Twitter², la plateforme dénombre 284 millions d'utilisateurs actifs mensuels produisant au total environ 500 millions de tweets par jour en moyenne. Le public utilisant ce service est très varié et il croît de jour en jour. Dans la plupart des cas, ce sont des utilisateurs lambda qui souhaitent communiquer une information factuelle, ponctuelle, anecdotique, politique, people, marketing, etc. qui exploitent ce réseau. Mais, ce site communautaire qui a connu une très forte croissance ces dernières années est également utilisé par des professionnels liés pour la plupart à l'industrie des médias (CNN, BBC envoient des flashs d'information sur Twitter), des partis ou des hommes politiques (John Edwards et Barack Obama ont exploité twitter comme outil médiatique lors de la campagne pour l'élection présidentielle américaine de 2008), des personnalités tous domaines confondus à l'intention de leurs fans ou d'un public concerné.

En ce qui concerne, les langues utilisées sur Twitter, même si l'anglais demeure la langue la plus utilisée (34 % des mes-

1. <https://twitter.com/>

2. <https://about.twitter.com/company>

@AnneDuBois hihi ! ça y est ! on est arrivé !
@isa#TeamOlympe Le coup de chaud que j'ai eu. Mon ordi voulait plus démarrer. c'était ma barette de RAM installé récemment qui fonctionnait plus #ouf
@Flavien 2000 : communes, départements, régions. 2014 : intercommunalités en plus ; et bientôt regroupements de communes ! argh .
@Capu ce projet de loi est vraiment inutile, et pas de sens pfff

TABLE 1 – Des exemples de messages de Twitter avec des interjections onomatopéiques exprimant l'état émotionnel des utilisateurs.

sages en septembre 2013 soit 170 millions de Tweets quotidiens), il n'en demeure pas moins qu'aujourd'hui, +60 % des tweets sont rédigés dans d'autres langues. Ce flux incessant de données représente une source d'information multilingue intéressante pour analyser les sentiments et les opinions des utilisateurs à l'égard de différentes questions.

Du fait, de la taille réduite de messages de Twitter les utilisateurs utilisent très souvent des formes langagières qui, permettent à la fois une réelle concision de l'écriture et d'exprimer de façon très explicite leurs états émotionnels et affectifs tels que les émoticônes ou les interjections.

En effet, ces deux pratiques discursives permettent d'exprimer rapidement des émotions, d'indiquer l'attitude subjective du locuteur, toutes choses qui se font très facilement dans une interaction orale en face à face en recourant au système sémiotique mimo-gestuel, ou encore à l'intonation, mais qui sont moins évidentes à faire passer à l'écrit. Le tableau 1 montre des exemples de tweets contenant des interjections.

Nous nous intéressons dans ce travail, en particulier, aux interjections, utilisées sur le réseau Twitter comme marqueurs de subjectivité et particulièrement d'émotion. En effet, l'interjection est classée par les grammaires parmi les parties du discours, au même titre que le verbe, l'adverbe, l'adjectif, le substantif ou le déterminant. Mais, si elle a bien une étiquette terminologique qui semble lui accorder un statut grammatical, elle est souvent marginalisée dans les analyses malgré sa richesse sémantique.

Il s'agira d'abord d'étudier les caractéristiques des interjections, puis d'évaluer leur apport pour les systèmes d'analyse de sentiments et de fouille d'opinions, en particulier pour la tâche de détection des émotions. Nous montrons, dans la section 4, qu'il est possible d'utiliser les interjections pour produire automatiquement un corpus étiqueté avec des annotations décrivant les émotions exprimés dans les blogs. Ensuite, dans la section 5 nous évaluons la pertinence du corpus collecté pour la détection automatique des émotions. Nous présentons, dans la section 6 notre méthode pour la création automatique de lexique affectif fin à partir du corpus collecté. Dans la section 7 nous évaluons la qualité du lexique produit.

2 Les interjections

2.1 Caractéristiques formelles

Parmi les interjections nous distinguons les deux sous-classes suivantes :

- **Les interjections onomatopéiques** sont celles qui imitent un bruit naturel qui peut être d'origine humaine (*ouille !*, *ha !*, *ouf !*, *hihi !*, etc.) ou non-humaine (*baoum !*, *tac tac !*, *flic floc !*, etc.) (Barberis, 1992). En s'appuyant sur les travaux de (Tesnière, 1959), (Kleiber, 2006) a proposé de les nommer "*interjections primaires émotives*". Les grammaires proposent parfois des listes d'interjections les plus fréquentes (Grevisse, 1969; Bonnard, 1971), mais aucune ne peut prétendre en donner une liste exhaustive et stable, du fait qu'il est possible d'introduire sans cesse de nouvelles créations imitatives dans le discours.

@Candy mouah .
@cntdpie héhé c'est vraiment marrant ça :)
@eelv pff ! mais elle sert à quoi cette loi ! !
@eelv ouf je l'ai trouvé, je pensais l'avoir perdu dans le bus !
@sam Oh ! c'est quoi ce match !

TABLE 2 – Exemples de tweets contenant des interjections onomatopéiques.

- **Les interjections non onomatopéiques** (Gonçalves, 2008) sont un ensemble de mots ou expressions figées empruntant

leur formulation à d'autres classes de mots : noms (*pardon, flûte, sans blague, etc.*), verbes (*allez !, vive !, etc.*), adjectifs (*hardi !*), adverbes (*là, comment, eh bien*). Dans ce cas, les morphèmes lexicaux perdent leur relation symbolique à l'objet du monde qu'ils représentaient, pour devenir des indices de subjectivité ou de l'émotion du locuteur. Ainsi les expressions "*Mon dieu !*" et "*Diable !*" sont marquées par un déplacement sémantique et pragmatique. Elles ne désignent plus le *le diable et le bon Dieu* mais l'état émotionnel du locuteur. Dans ses travaux de thèse (Halté, 2013), a qualifié cette catégorie d'*interjections secondaires*.

@xsof Purée ça ne marche toujours pas !
@elao non mdr il n'y a rien d'intéressant dans ce programme !
@eelv tiens voilà une bonne mesure
@sam voyons ! t'as bien d'autres solutions ?

TABLE 3 – Exemples de tweets contenant des interjections non onomatopéiques.

2.2 Valeurs subjectives et fonctionnement textuel

Les interjections ont ceci de particulier qu'elles sont nécessairement les indices linguistiques d'une émotion. Même si exprimer une émotion n'est pas toujours leur fonction principale (notamment les interjections vocatives comme "*euh !*" ou "*pst !*", ou celles dont la fonction consiste à donner un ordre comme "*Stop !*" ou "*ouste !*"), c'est toujours un mode d'expression linguistique d'une émotion ou plus généralement d'une attitude subjective. Selon (Guillaume, 1973), *l'expressivité* (sens d'intention subjectif et momentané visé par le locuteur dans l'instant de parole) croît aux dépens de *l'expression* (recours au langage institué, avec des sens stabilisés et des propositions grammaticalement formées). L'interjection constitue le cas limite d'expressivité. D'après (Barberis, 1992), si l'on pose que tout message linguistique repose sur l'équation $expressivité + expression = 1$, l'interjection fait tendre l'expressivité vers 1 et l'expression vers 0. Les interjections sont donc des marques de subjectivité. Les valeurs subjectives des interjections sont conjoncturelles, par exemple, l'interjection *Ah* peut exprimer dans certains contextes le *dégoût* et dans d'autres *la peur, l'étonnement, le soulagement, la colère, etc.*

En ce qui concerne son fonctionnement textuel, l'interjection se caractérise par son détachement du reste de l'énoncé. Aucun lien syntaxique ne la relie à son contexte discursif, ce qui lui permet d'y occuper des places variables. Par exemple, à l'intérieur de l'énoncé "*je n'ai pas réussi mon examen*", l'interjection "*Hélas*" peut s'insérer soit au début, après le verbe ou à la fin du texte. Cependant, les interjections qui soulignent l'orientation affective du locuteur (surprise, peur, joie, soulagement, colère, etc.), interviennent souvent au début de l'énoncé : "*Ouf ! il est arrivé à l'heure.*" ou "*Ah ! il ne tient jamais sa promesse*".

3 Travaux antérieurs

Plusieurs travaux de recherche en traitement automatique de la langue écrite et orale, se sont intéressés à l'étude des interjections dans le discours. (Garcia-Fernandez *et al.*, 2010) ont procédé à une analyse contextuelle de l'interjection onomatopéique d'hésitation "*Euh*" et ont montré qu'elle peut être utilisée comme indicateur de reformulation ou de niveau de confiance d'une réponse utilisateur dans un système de question-réponse.

Concernant, leur utilisation dans les conversations en ligne, (Falaise, 2005) a montré que le recours aux interjections est assez courant. En effet, sur un corpus de 77 messages choisi aléatoirement par l'auteur, les interjections représentent à elles seules 10 % du corpus. Dans un travail plus récent, (Halté, 2013), a démontré que les interjections relèvent des marques modales qui permettent aux locuteurs de faire porter une émotion ou une attitude subjective sur l'énonciation d'un contenu.

Dans le domaine de l'analyse de sentiments et de la fouille d'opinions, plusieurs travaux de recherche utilisent d'autres formes discursives comme marqueur de subjectivité. (Read, 2005; Pak & Paroubek, 2010) ont utilisé les émoticônes comme marqueur de polarité pour distinguer les textes positifs et négatifs. Dans un premier temps, ils ont identifié une liste d'émoticônes positives (:), :-), :-D, etc.) et une liste d'émoticônes négatives (:(), :-(, etc.). Ensuite, les deux listes ont été utilisées comme critère de recherche pour récupérer des messages positifs et négatifs depuis Twitter

Dans des travaux plus récents (Mohammad, 2012; Qadir & Riloff, 2013; Fraisse & Paroubek, 2014b) ont utilisé une liste de mot-dièses (*hashtag* en anglais) (*#sad, #happy, #angry, #fear, #anxious, #disappointed, #unhappy, etc.*) pour collecter des corpus émotionnels et construire de façon automatique des lexiques affectifs. Les lexiques ont été ensuite utilisés dans

des tâches de détection automatique des émotions.

4 Création du corpus émotionnel annoté à partir de Twitter

Les utilisateurs de Twitter font souvent recours aux interjections, qui allient la concision d'écriture à la richesse sémantique, pour exprimer leurs états émotionnels et plus généralement leurs attitudes subjectives. Le tableau 4 montre quelques exemples de tweets subjectifs contenant des interjections. Pour les trois premiers messages, il est possible de détecter les émotions exprimées par les trois locuteurs sans lire les interjections employées (*colère* dans le premier message, *Joie* dans le deuxième et *surprise positive* dans le troisième). Cependant, il est quasiment impossible d'identifier avec exactitude les émotions exprimées dans les messages 4 (*Colère*), 5 (*Dégoût*) et 6 (*Peur*) sans prendre en compte l'interjection employée dans chacun de ces messages. En effet, dans ce cas, l'interjection fournit une information importante sur l'état émotionnel du locuteur et qui n'est pas présente (de façon implicite ou explicite) dans le reste du message.

D'autant plus, du fait de la taille réduite du tweet, des fautes d'orthographe, des abréviations et de tous les autres phénomènes linguistiques et extra-linguistiques qu'il peut contenir, il est difficile d'effectuer une analyse correcte de l'intégralité du tweet. Il est donc important de repérer et de concentrer l'analyse sur les parties pertinentes du tweet. Dans notre cas, il s'agit de parties qui contiennent des informations utiles sur l'état émotionnel du locuteur et les interjections en font partie intégrante. L'objectif de ce papier est d'étudier l'apport des interjections pour les systèmes d'analyse de sentiments et en

1.	Argh! train en grève ! c'est parti pour des heures d'attentes !
2.	hihi! j'ai gagné mon pari :)
3.	Ah! je ne m'y attendais vraiment pas super !
4.	pff ... je pars seul
5.	Beurk Fast-food. McDonald's dévoile les 19 ingrédients de ses frites
6.	aaah j'ai entendu un grand bruit en bas !

TABLE 4 – Exemples de tweets exprimant des émotions et contenant des interjections.

particulier pour la tâche de détection des émotions dans les textes courts. Nous nous posons alors les questions suivantes :

- Est-il possible d'utiliser les interjections pour créer et étiqueter un large corpus émotionnel ?
- Est-il possible de considérer les interjections comme une annotation émotionnelle fiable malgré le grand nombre d'annotateurs (les utilisateurs) et leurs différences culturelles et sociales ?

Dans un premier temps, nous nous sommes intéressés au nombre ainsi qu'aux types de classes émotionnelles à considérer. En effet, à ce jour, il est impossible de dénombrer les émotions avec exactitude. Plusieurs travaux (Mohammad, 2012; Qadir & Riloff, 2013) se basent sur les six émotions de base définies par (Ekman, 1970) (*tristesse, colère, peur, surprise, joie, dégoût*). D'autres travaux ont défini leur propres modèles et classes d'émotions (Matsumoto, 2009; Levenson, 2011; Cambria *et al.*, 2012). Cependant ces modèles sont souvent adaptés à une langue, à une culture, à un domaine ou à une application donnée et ils ne couvrent que partiellement le spectre de la subjectivité.

Après une étude complète de l'état de l'art sur ce sujet, nous avons sélectionné le modèle proposé par (Fraisse & Paroubek, 2014a). Ce choix a été motivé par la généralité du modèle proposé (indépendant du domaine) ainsi par sa couverture totale du spectre de la subjectivité (opinion, sentiment, émotion). Les auteurs proposent 18 classes subjectives qui couvrent les différents types d'expressions subjectives (émotion, sentiment, jugement, opinion) (Tableau 5). Les auteurs répartissent les informations subjectives en trois grandes catégories : une première catégorie nommée *Opinion* qui inclut les 4 classes suivantes : *accord, désaccord, valorisation et dévalorisation*, une deuxième catégorie *Sentiment* qui inclut les 2 classes : *satisfaction et insatisfaction* et enfin la catégorie *Émotion* avec les 12 classes affectives : *colère, peur, tristesse, ennui, dérangement, surprise négative, déplaisir, mépris, surprise positive, apaisement, plaisir, amour*. Afin d'identifier les classes d'émotions ayant des interjections connues et utilisées par un grand nombre d'utilisateur. Nous avons, d'abord, extrait manuellement une liste d'interjections depuis différentes sources (Grevisse, 1969; Bonnard, 1971). Ensuite, nous avons utilisé cette liste comme requête de recherche pour collecter les tweets contenant au moins une interjection parmi la liste fournie dans la requête. Enfin, nous avons retenu uniquement les interjections qui sont à la fois fréquentes (ayant un nombre d'occurrence élevé dans le corpus) et souvent utilisées pour exprimer la même émotion par différents utilisateurs. Le tableau 6 décrit l'ensemble des interjections que nous avons retenu ainsi que les classes d'émotions associées.

#	Classe	Catégorie	Spécification
1	SURPRISE NÉGATIVE	e-	surprise négative / étonnement négatif
2	DÉRANGEMENT	e-	dérangement / embarras
3	PEUR	e-	peur / terreur / inquiétude
4	ENNUI	e-	ennui
5	DÉPLAISIR	e-	déplaisir/ déception
6	TRISTESSE	e-	tristesse / chagrin / souffrance / désespoir / résignation
7	COLÈRE	e-	colère / rage / agacement / exaspération / énervement / impatience
8	MÉPRIS	e-	mépris / dédain / dégoût / haine/
9	INSATISFACTION	s-	insatisfaction / mécontentement
10	DÉVALORISATION	o-	dévalorisation / désintérêt / dépréciation
11	DÉSACCORD	o-	désaccord / désapprobation
12	VALORISATION	o+	valorisation / intérêt / appréciation
13	ACCORD	o+	accord / compréhension / approbation
14	SATISFACTION	s+	satisfaction / contentement / fierté
15	SURPRISE POSITIVE	e+	surprise positive / étonnement positif
16	APAISEMENT	e+	apaisement / soulagement / reconnaissance / pardon / sérénité
17	PLAISIR	e+	plaisir / divertissement / joie /euphorie / bonheur /extase
18	AMOUR	e+	amour / tendresse / affection / dévouement / passion / envie

TABLE 5 – Les 18 classes subjectives proposées par (Fraisie & Paroubek, 2014a), e=émotion, s=sentiment, o=opinion, += valence positive, -=valence négative.

#	Classe	Interjections onomatopéiques	Interjections non onomatopéiques
1	COLÈRE	Argh , pff	Voyons !
2	PLAISIR	hihi , haha	lol , youpi
3	PEUR	aah	
4	APAISEMENT	ouf	
5	TRISTESSE	Aïe, ouille	zut, hélas
6	INSATISFACTION	Bof	
7	MÉPRIS	Beurk	
8	SURPRISE NÉGATIVE	oups	

TABLE 6 – Les classes émotionnelles exprimées ayant des interjections clairement définies sur Twitter.

4.1 Recherche basée sur les interjections

Nous avons utilisé l'API Search³ de Twitter pour collecter et filtrer les messages. L'API permet de spécifier la langue de messages et une requête de recherche par mot clé. Ainsi, pour chaque classe émotionnelle du Tableau 6, nous utilisons la liste d'interjections, qui lui est associée, comme mots clés de la requête. Nous attribuons par la suite l'étiquette de la classe correspondante à tous les messages qui ont été retourné par la requête. Par exemple, tous les messages qui ont été collectés en utilisant la liste d'interjections "*hihi*", "*haha*", "*lol*", "*youpi*" sont étiquetés *PLAISIR*.

Certains utilisateurs n'utilisent pas souvent les formes normalisées des interjections, par exemple l'interjection "*argh*" peut être écrite "*arggh*", "*arrrrgh*", "*arghhhh*", *etc.*. En plus, certaines interjections n'ont pas vraiment de forme d'écriture normalisée. Ainsi, pour collecter aussi les messages contenant des interjections orthographiées de façon non normalisée, nous avons inclus, moyennant des expressions régulières, dans les requêtes de recherche, la plupart des variations orthographiques des interjections par exemple : "*arghh*", "*arrrrgh*", "*arrrrgh*", "*arghhhh*", *etc.* pour l'interjection "*argh*". En total, pour les 8 classes d'émotions nous avons collecté un corpus de 30,123 tweets. Après suppression des *retweets* (messages commençant par l'abréviation "RT" et qui consistent à re-publier tel quel un message d'un autre twitteur) le corpus final était composé de 19,061 tweets.

4.2 Distribution des interjections dans le corpus

Le tableau 7 décrit la distribution des interjections dans le corpus émotionnel collecté depuis Twitter. Certaines interjections sont plus fréquentes que d'autres comme "*pff*", "*hihi*", "*haha*". Nous observons aussi que pour certaines émotions

3. <https://dev.twitter.com/docs/api/1/get/search>

comme par exemple la *COLÈRE*, les twitteurs utilisent plus les interjections onomatopéiques que les non-onomatopéiques. Par exemple, parmi les 3896 messages étiquetés *COLÈRE* 3440 contiennent les interjections "*argh*" et "*pff*" et seulement 456 messages avec l'interjection non-onomatopéique "*Voyons !*". Ce phénomène peut être expliqué par le fait que les interjections onomatopéiques sont plus explicites pour exprimer certaines émotions intenses.

Interjection	Nombre de messages	% de messages
Argh	1078	5,65
pff	2362	12,39
Voyons !	456	2,39
hihi	3350	17,57
haha	3987	19,14
lol	1456	7,63
youpi	1211	6,35
aah	844	4,42
ouf	1300	6,82
Aïe	320	1,67
ouille	337	1,76
zut	158	0,82
hélas	987	5,17
Bof	650	3,41
beurk	435	2,82
oups	467	2,45
Total tweets	19061	100
Total utilisateurs	19032	

TABLE 7 – Distribution des différentes interjections dans le corpus émotionnel de Twitter.

5 Pertinence et utilité du corpus collecté depuis Twitter

Un des problèmes majeurs inhérent à la classification supervisée des émotions est d'obtenir un corpus d'apprentissage annoté manuellement. L'annotation est complexe, elle coûte cher et nécessite beaucoup de temps. En outre, du fait de l'ambiguïté des textes émotionnels, il est souvent difficile d'avoir des scores inter-annotateur élevés lors de l'annotation. Dans notre cas, nous considérons que le corpus collecté est un corpus annoté par les utilisateurs eux mêmes, puisque nous attribuons le label émotionnel de l'interjection présente dans le texte à tout le message. Par exemple si le message contient l'une des interjections suivantes : "*argh*", "*pff*", "*voyons !*" alors il est étiqueté *COLÈRE*. Comme le montre le tableau 7, le corpus contient 19061 messages créés par 19032 utilisateurs différents. Ainsi, nous posons les deux hypothèses suivantes :

1. Les annotations données par ce grand nombre d'annotateurs sont cohérentes et pertinentes.
2. Ce corpus peut être utilisé par un classifieur pour apprendre à détecter les émotions.

Pour vérifier ces deux hypothèses, nous avons mené une expérience de classification automatique des émotions en utilisant ce corpus comme corpus d'apprentissage. Nous avons entraîné un système de classement à vecteurs supports en utilisant des traits à base de n-grammes extraits du corpus d'apprentissage. Nous avons utilisé la bibliothèque LIBLINEAR (Fan *et al.*, 2008) avec un noyau linéaire et un paramétrage par défaut. Pour pondérer chaque terme nous avons eu recours aux deux fonctions de poids suivantes :

- **Binaire** : 1 si le terme apparaît dans le message et 0 sinon.
- **TF-IDF normalisée** : nous avons utilisé une forme normalisée de la fonction *TF-IDF* (Pak *et al.*, 2014). Cette normalisation consiste à diviser la fonction de poids *TF-IDF* par la fréquence moyenne d'un terme (*avg.tf*). La normalisation à base de fréquence moyenne d'un terme est basée sur l'observation que les utilisateurs ont tendance à utiliser un vocabulaire riche quand ils expriment leur attitude subjective. Ainsi, les termes subjectifs et émotionnels (par exemple *adorable*, *satisfaisant*) ont une fréquence moyenne proche de 1, tandis que les termes non subjectifs ont une fréquence moyenne plus élevée. Ainsi, afin de normaliser le poids de chaque n-gramme (unigramme dans notre cas) du document, nous le divisons par sa fréquence moyenne (Équation 1).

$$w(t_i) = \frac{tfidf(t_i)}{avg.tf(t_i)} \quad (1)$$

$$\text{tfidf}(t_i) = \text{tf}(t_i) \cdot \log \frac{|D|}{\text{df}(t_i)} \quad (2)$$

$$\text{avg.tf}(t_i) = \frac{\sum_{\{d \in D | t_i \in d\}} \text{tf}(t_i)}{|\{d \in D | t_i \in d\}|} \quad (3)$$

où $\{d \in D | t_i \in d\}$ est l'ensemble des messages qui contiennent le terme t_i .

Pour la classification multi-étiquettes, nous avons entraîné indépendamment un système de classement pour chaque émotion. Le tableau 8 montre la répartition des données en apprentissage et test par classe d'émotion. Afin de diminuer le bruit dans les tweets, nous avons pré-traité le corpus en supprimant les mentions utilisateurs (les mots qui commencent par "@" et qui font référence aux utilisateurs) et les URL. Enfin, pour ne pas biaiser le classement, nous avons supprimé des tweets toutes les interjections qui ont servi pour collecter et étiqueter le corpus. Les résultats de classification illustrés dans le

Classes	Apprentissage	test
COLÈRE	2598	1298
PLAISIR	6445	3222
PEUR	563	281
APAISEMENT	867	433
TRISTESSE	1202	600
INSATISFACTION	434	216
MÉPRIS	290	145
SURPRISE NÉGATIVE	312	155
Total	12711	6350

TABLE 8 – Caractéristiques du corpus Twitter utilisé pour l'apprentissage.

Classes	unigr.+binaire			unigr.+TF-IDF/avg.tf		
	Précision	Rappel	F1	Précision	Rappel	F1
COLÈRE	0,72	0,61	0,66	0,43	0,37	0,39
PLAISIR	0,42	0,51	0,46	0,71	0,76	0,73
PEUR	0,21	0,13	0,16	0,11	0,09	0,09
APAISEMENT	0,13	0,23	0,16	0,25	0,31	0,27
TRISTESSE	0,31	0,42	0,35	0,34	0,37	0,35
INSATISFACTION	0,12	0,22	0,15	0,27	0,12	0,16
MÉPRIS	0,09	0,12	0,10	0,19	0,12	0,14
SURPRISE NÉGATIVE	0,07	0,11	0,08	0,04	0,09	0,05

TABLE 9 – Résultats de la classification des émotions, les meilleurs score en mesure F1 sont montrés en gras.

tableau 9, montrent que certains classifieurs atteignent des scores acceptables en mesure F1 : 0,66 pour le classifieur *COLÈRE* et 0,73 pour le classifieur *PLAISIR*. Ces deux bons scores s'expliquent par le fait que les deux classifieurs disposent de plus de données d'entraînement que les autres classifieurs. Nous notons aussi que les systèmes utilisant les vecteurs de traits *unigrammes+TF-IDF/avg.tf* ont de meilleures performances en classification. En effet, la fonction *TF-IDF/avg.tf* permet de favoriser les termes qui caractérisent au mieux une émotion et donc de mieux apprendre les traits représentatifs d'une classe donnée. Nos résultats sont comparables à d'autres travaux similaires comme ceux de (Roberts *et al.*, 2012), qui annoncent un score de 0.74 en mesure F1 atteint par leur classifieur supervisé et entraîné sur un corpus de tweets annoté manuellement par des experts.

6 Création du lexique d'émotions

Un lexique d'émotions est une liste de mots associés à une ou plusieurs émotions. Par exemple, l'adjectif "*épanoui*" peut être associé aux deux émotions *joie* ou *amour* en fonction du contexte. Ce type de lexique est très utile et peut être utilisé dans plusieurs applications et tâches telles que la détection des émotions dans les textes, l'identification de passages émotionnels et affectifs dans les documents ou encore l'analyse de personnalité. Nous présentons dans cette section notre approche pour la construction automatique d'un lexique affectif fin à partir du corpus émotionnel de Twitter.

6.1 Méthode

En se basant sur une approche statistique, notre méthode consiste à extraire, à partir du corpus émotionnel collecté (décrit dans la section 4), et pour chaque classe d'émotion (du Tableau 6), l'ensemble de mots qui lui est associé. Afin de mesurer l'association entre un mot m du corpus et une émotion e , nous nous sommes basés sur l'information mutuelle introduite par (Fano, 1961; Church & Hanks, 1990), qui pour chaque couple de variables aléatoires (X, Y) mesure leur degré de dépendance au sens probabiliste. L'information mutuelle est donnée par la formule suivante :

$$IM(X, Y) = \log_2 \left(\frac{P(X, Y)}{P(X) \cdot P(Y)} \right) \quad (4)$$

Ainsi, dans notre cas, il s'agit de mesurer le degré de dépendance entre un mot m et une émotion e .

$$IM(m, e) = \log_2 \left(\frac{freq(m, e)}{freq(m) \cdot freq(e)} \right) \quad (5)$$

Avec $freq(m, e)$ est le rapport entre le nombre de messages contenant le mot m et étiqueté e ($|T_{m,e}|$) et le nombre total de messages ($|T|$).

$$freq(m, e) = \frac{|T_{m,e}|}{|T|} \quad (6)$$

$freq(m)$ est le rapport entre le nombre total de messages contenant le mot m ($|T_m|$) et le nombre total de messages.

$$freq(m) = \frac{|T_m|}{|T|} \quad (7)$$

et $freq(e)$ est le rapport entre le nombre total de messages étiquetés e ($|T_e|$) et le nombre total de messages.

$$freq(e) = \frac{|T_e|}{|T|} \quad (8)$$

Certains mots peuvent être associés à plusieurs émotions. Nous calculons donc comme dans (Mohammad, 2012), le degré de dépendance entre les deux variables m et $\neg e$ ($\neg e$ représente toutes les classes d'émotions sauf e).

$$IM(m, \neg e) = \log_2 \left(\frac{freq(m, \neg e)}{freq(m) \cdot freq(\neg e)} \right) \quad (9)$$

Ainsi, le degré d'association entre un m et une e est donné par l'équation 10. Nous considérons qu'un mot m est fortement associé à une émotion e , si son degré d'association $Asso(m, e)$ est supérieur à 0.

$$Asso(m, e) = IM(m, e) - IM(m, \neg e) \quad (10)$$

6.2 Expérimentation et résultats

Les messages Twitter peuvent contenir : des URLs, des mentions utilisateur (par exemple @oXc11), des retweets commençant par la mention "RT", des caractères spéciaux, etc. Ainsi, avant de procéder à l'extraction des mots affectifs, depuis le corpus collecté, nous avons tout d'abord effectué certaines opérations de pré-traitement : (1) suppression des liens URL, retweets ainsi que les mentions utilisateurs (2) segmentation et (3) suppressions des mots outils. Ensuite, nous avons calculé pour chaque m du corpus son degré d'association $Asso(m, e)$ selon la méthode présentée dans la section précédente. En fonction de son degré d'association, un mot peut être attribué à une ou plusieurs classes d'émotions.

La taille du lexique affectif produit est de 1530 mots répartis sur les 8 classes d'émotions (Tableau 10).

7 Évaluation du lexique

7.1 Expérimentations

Afin, d'évaluer le lexique produit, nous l'avons utilisé dans une tâche de détection des émotions à partir de textes dont le corpus de référence est issu du projet DOXA (Paroubek *et al.*, 2010). À notre connaissance, c'est le seul corpus en

#	Classe émotionnelle	Nombre de termes	Exemples
1	COLÈRE	234	colère, furieux, fâché
2	PLAISIR	423	super, enjoy, cool
3	PEUR	184	peur, tremble, inquiet
4	APAISEMENT	178	ouf, enfin, fini
5	TRISTESSE	168	triste, seul, mal
6	INSATISFACTION	58	bof, insatisfait, moyen
7	MÉPRIS	198	déteste, dégoûtant, dégueulasse
8	SURPRISE NÉGATIVE	87	non, stupéfait, impossible

TABLE 10 – Nombre de termes par classe d’émotion dans le lexique.

français qui est annoté avec des classes d’émotions fines et auquel nous avons accès. Il s’agit d’un corpus de critiques de films annoté manuellement avec les classes d’émotion correspondantes (18 classes en total). Les annotations ont été faites sur deux niveaux :

- le niveau *macro* : annotations au niveau du document ;
- et le niveau *meso* : annotation au niveau du paragraphes.

Les paragraphes peuvent avoir de 1 à 5 étiquettes d’émotions. Le corpus total compte 7126 paragraphes dont 612 annotées avec plus d’une émotion (609 annotées avec 2 émotions et 3 paragraphes avec 3 émotions). Pour nos expérimentations, nous considérons les paragraphes comme des documents. Ainsi, dans un premier temps, nous avons constitué un corpus avec l’ensemble de paragraphes regroupés par classe d’émotion. Ensuite, nous avons écarté les paragraphes qui sont étiquetées avec une classe d’émotion non-présente dans notre lexique. Au total, nous avons formé un corpus de 1,000 documents repartis sur les 8 classes d’émotions de notre lexique. Le Tableau 11 décrit la répartition des données, en apprentissage et en test, à travers les différentes classes d’émotions du corpus. Nous avons entraîné un système de classement

#	Classe émotionnelle	Apprentissage	test	Total
1	COLÈRE	86	42	128
2	PLAISIR	288	144	432
3	PEUR	16	7	23
4	APAISEMENT	82	40	122
5	TRISTESSE	51	25	76
6	INSATISFACTION	9	4	13
7	MÉPRIS	117	58	175
8	SURPRISE NÉGATIVE	21	10	31
Total		670	330	1000

TABLE 11 – Caractéristiques du corpus d’apprentissage DOXA.

à vecteurs supports utilisant différents traits extraits du corpus d’apprentissage. Nous avons utilisé la bibliothèque LIBLINEAR (Fan *et al.*, 2008) avec un noyau linéaire et un paramétrage par défaut. Pour la classification multi-étiquettes, nous avons utilisé une stratégie en parallèle, c’est-à-dire que nous avons entraîné indépendamment un système de classement pour chaque émotion. Chaque système de classement fournit pour chaque message une indication de présence ou d’absence de l’émotion qu’il a été entraîné à détecter. Ainsi nous pouvons obtenir pour chaque document, de 0 à 8 étiquettes d’émotions. La liste des traits utilisés pour l’apprentissage comprenait :

- **n-grammes** : nous avons utilisé des unigrammes et des bigrammes avec la fonction de poids TF-IDF normalisée (décrite dans la section 5).
- **Lexique** : nous avons créé un vecteur de traits avec les noms de classes d’émotions présents dans le lexique (8 classes en total) et dont la valeur est initialisée à 0. Ensuite pour chaque occurrence de mot présent dans un tweet, nous vérifions s’il est présent dans le lexique. Si c’est le cas alors nous incrémentons de 1 la valeur numérique de classes d’émotions correspondantes dans le vecteur de traits. Par exemple si le tweet contient 2 occurrences de mots qui sont présents dans le lexique et étiquetées *PLAISIR* alors l’élément *PLAISIR* du vecteur de traits aura comme valeur 2.

Afin, d’évaluer l’apport de notre lexique pour la tâche de détection des émotions, nous avons utilisé l’algorithme de classification suivant :

1. D’abord nous avons entraîné le système uniquement avec les deux vecteurs *unigr.TF-IDF/avg.tf* et *bigr.TF-IDF/avg.tf*.
2. Ensuite, nous avons rajouté les traits *Lexique* aux deux vecteurs *unigr.TF-IDF/avg.tf* et *bigr.TF-IDF/avg.tf* et évalué

la différence de performance.

7.1.1 Discussion de résultats

Afin d'affiner le paramétrage de notre système de classement, nous avons effectué une validation croisée à 10 replis sur le corpus d'apprentissage. Nous avons utilisé *la précision, le rappel et la mesure F1* pour évaluer la performance du système.

D'abord, nous avons analysé la performance de différents vecteurs de traits utilisés pour la détection des émotions : *unigr.TF-IDF/avg.tf.*, *unigr.TF-IDF/avg.tf+lexique*, *bigr.TF-IDF/avg.tf*, *bigr.TF-IDF/avg.tf+lexique*. La figure 1 montre les performances du système en mesure F1 pour chaque émotion détectée et pour chaque vecteur de traits utilisé par le classifieur. La performance de classification de classes fréquentes comme *PLAISIR*, *COLÈRE*, *MÉPRIS*, *TRISTESSE* et

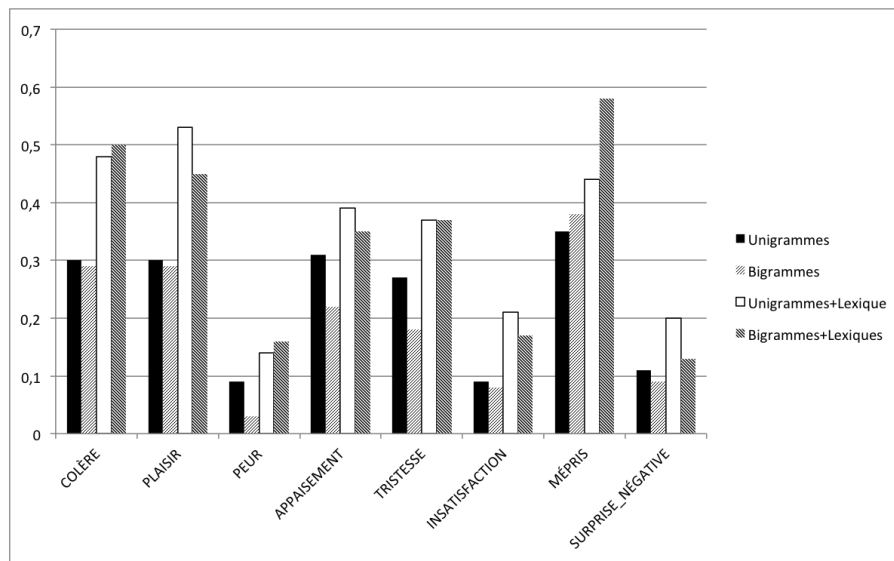


FIGURE 1 – Performance de différents types de traits utilisés pour la détection des émotions.

DÉGOÛT est meilleure que celles qui sont moins fréquentes dans les données d'apprentissage comme *PEUR*, *SURPRISE NÉGATIVE* et *INSATISFACTION*. Cependant dans les deux cas, les vecteurs de traits ayant obtenus les meilleurs scores en mesure F1 sont les *unigr.TF-IDF/avg.tf+lexique* et *bigr.TF-IDF/avg.tf+lexique*. D'après les résultats illustrés dans le tableau 12, le vecteur de traits *unigr.TF-IDF/avg.tf+lexique* permet une augmentation de +0,25 en mesure F1 pour la classe *PLAISIR* et le vecteur *bigr.TF-IDF/avg.tf+lexique* une augmentation +0,21 pour la classe *COLÈRE*. Cela prouve que le lexique que nous avons construit apporte un gain significatif pour la tâche de détection des émotions par rapport aux systèmes qui utilisent uniquement les N-grammes. En effet, les meilleurs scores en mesure F1 obtenu par le système à base de N-grammes sont : 0,38 en utilisant le vecteur *bigr.TF-IDF/avg.tf* et 0,35 en utilisant le vecteur de traits *unigr.TF-IDF/avg.tf* contre 0,53 avec le vecteur de traits *unigrammes+lexique* et 0,5 avec le vecteur *bigrammes+lexique*.

Nous avons comparé aussi nos résultats à ceux obtenus par (Roberts *et al.*, 2012) pour la détection des émotions dans les Tweets. Malgré le fait qu'il ne s'agit pas du même corpus de Tweets, cela nous permet de se situer par rapport à d'autres méthodes dans l'état de l'art. (Roberts *et al.*, 2012) utilisent un corpus de Tweets annoté automatiquement avec les sept émotions : *PEUR*, *COLÈRE*, *SURPRISE*, *DÉGOÛT*, *JOIE*, *AMOUR* et *TRISTESSE*. Afin de détecter les émotions présentes dans les tweets, les auteurs utilisent un classifieur par émotion en se basant sur différents types de traits : *unigrammes*, *bigrammes*, *trigrammes*, *synsets*, *contains !* (qui indique si le tweet contient un point d'exclamation), *contains ?* (qui indique si le tweet contient un point d'interrogation) et *Topic* (qui indique le sujet du tweet). Malgré la différence de taille de corpus d'entraînement et la simplicité des traits utilisés pendant la phase d'apprentissage, nous obtenons, pour certaines émotions (*MÉPRIS* et *PLAISIR*) des scores comparables à ceux obtenus par (Roberts *et al.*, 2012).

Classes	unigrammes			unigr.+lexique			Δ	bigrammes			bigr.+lexique			Δ
	P.	R.	F1	P.	R.	F1		P.	R.	F1	P.	R.	F1	
COLÈRE	0,32	0,29	0,3	0,49	0,48	0,48	+0,18	0,29	0,31	0,29	0,56	0,46	0,5	+0,21
PLAISIR	0,39	0,27	0,28	0,56	0,51	0,53	+0,25	0,28	0,31	0,29	0,51	0,41	0,45	+0,19
PEUR	0,11	0,09	0,09	0,17	0,12	0,14	+0,05	0,07	0,02	0,03	0,13	0,21	0,16	+0,13
APAISEMENT	0,35	0,28	0,31	0,41	0,39	0,39	+0,08	0,29	0,18	0,2	0,35	0,37	0,35	+0,15
TRISTESSE	0,29	0,27	0,27	0,39	0,36	0,37	+0,1	0,17	0,21	0,18	0,41	0,34	0,37	+0,19
INSATISFACTION	0,11	0,09	0,09	0,17	0,29	0,21	+0,12	0,1	0,07	0,08	0,13	0,27	0,17	+0,09
MÉPRIS	0,41	0,32	0,35	0,47	0,42	0,44	+0,09	0,39	0,38	0,38	0,61	0,56	0,58	+0,2
SURPRISE NÉGAT.	0,13	0,11	0,11	0,23	0,19	0,20	+0,09	0,12	0,08	0,09	0,18	0,11	0,13	+0,04

TABLE 12 – Résultats de la classification des émotions avec les différents types de traits (unigr.,bigr.,lexi.)

8 Conclusion

Dan ce papier, nous avons étudié et évalué l'apport des interjections pour les systèmes d'analyse de sentiments et de fouille d'opinions. Malgré leur statut grammatical et leur richesse sémantique pour expliciter un état émotionnel, les interjections sont restées marginalisées par les systèmes d'analyse de sentiments. Nous avons proposé une méthode automatique pour collecter et étiqueter les messages de Twitter en considérant les interjections présentes dans les messages comme des annotations bruitées. Nous avons ainsi créé un corpus émotionnel de 19061 tweets repartis sur 8 classes d'émotion. Afin, d'évaluer la pertinence du corpus collecté, nous l'avons utilisé comme corpus d'apprentissage pour la tâche de détection des émotions. Nos résultats ont montré que nous obtenons des performances comparables à d'autres systèmes entraînés sur des corpus annotés manuellement par des experts. Ensuite, nous avons proposé une méthode permettant d'extraire, de façon automatique, à partir du corpus collecté un lexique affectif fin. Enfin, nous avons évalué la qualité du lexique produit sur une tâche de détection des émotions, qui a montré un gain en mesure F1 allant selon les émotions de +0,04 à +0,21. L'approche proposée dans cet article, affine et complète les méthodes existantes basées sur l'utilisation des émoticônes et des mots-dièses pour la classification des textes subjectifs. Pour la suite de nos travaux, nous souhaitons évaluer la généralité de notre approche en l'appliquant à d'autres langues.

Remerciements. Les auteurs remercient les membres du comité scientifique de TALN pour leurs remarques constructives. Les travaux présentés ici ont été réalisés dans le cadre du projet CHIST-ERA uComp, convention ANR-12-CHRI-0003-03.

Références

- BARBERIS J.-M. (1992). Onomatopée, interjection : un défi pour la grammaire. *L'information grammaticale*, **53**, p.52–57.
- BONNARD H. (1971). Interjections. *Grand Larousse de la langue française*, tome IV, p. p.2758.
- CAMBRIA E., LIVINGSTONE A. & HUSSAIN A. (2012). The hourglass of emotions. *Cognitive behavioural systems*.
- CHURCH K. W. & HANKS P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**, p.22–29.
- EKMAN P. (1970). Universal facial expressions of emotions. In *Digest, California mental Health Research*.
- FALAISE A. (2005). Constitution d'un corpus de français tchaté. In *actes de RÉCITAL*, Dourdan.
- FAN R.-E., CHANG K.-W., HSIEH C.-J., WANG X.-R. & CHIH-JEN (2008). Liblinear : A library for large linear classification. *Journal of Machine Learning Research*, **9**, p.1871–1874.
- FANO R. (1961). Transmission of information : A statistical theory of communications. *MIT Press, Cambridge*.
- FRAISSE A. & PAROUBEK P. (2014a). Toward a unifying model for opinion, sentiment and emotion information extraction. In *the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.
- FRAISSE A. & PAROUBEK P. (2014b). Twitter as a comparable corpus to build multilingual affective lexicons. In *the 7th Workshop on Building and Using Comparable Corpora*.
- GARCIA-FERNANDEZ A., VASILESCU I. & ROSSET S. (2010). Euh as cue for speaker confidence and word searching in human spoken answers in french. In *DISS-LPSS Joint Workshop/Disfluency*.

- GONÇALVES M. (2008). Sur le statut linguistique de l'interjection. In *Actas del VIII congreso de Lingüística General*, p. p.14, Universidad Autónoma de Madrid.
- GREVISSE M. (1969). *Le bon usage*. Duculot et Gembloux, Haltier.
- GUILLAUME G. (1973). Principes de linguistiques théoriques. *Presses del'université Laval, Québec et Klincksieck, Paris*, p. p.146.
- HALTÉ P. (2013). *Les marques modales dans les chats : étude sémiotique et pragmatique des interjections et des émoticônes dans un corpus de conversations synchrones en ligne*. Universités de Luxembourg et de Lorraine.
- KLEIBER G. (2006). Sémiotique de l'interjection. *Langages*, **161**.
- LEVENSON R. W. (2011). Basic emotion questions. *Emotion Review*, **3**(4), p.379–386.
- MATSUMOTO D. (2009). Spontaneous facial expressions of emotion of blind individuals. *Journal of Personality and Social Psychology*, **96**(1), p.1–10.
- MOHAMMAD S. M. (2012). Emotional tweets. In *proceedings of First Joint Conference on Lexical and Computational Semantics*, p. p.246–255.
- PAK A. & PAROUBEK P. (2010). Construction d'un lexique affectif pour le français à partir de twitter. In *proceedings of TALN (Traitement Automatique des Langues Naturelles) 2010*, Montréal, Canada.
- PAK A., PAROUBEK P., FRAISSE A. & FRANCOPOULO G. (2014). Normalization of term weighting scheme for sentiment analysis. *Human Language technology Challenges for Computer Science and Linguistics. Series : Lecture Notes in Artificial Intelligence*, **8387**.
- PAROUBEK P., PAK A. & MOSTEFA D. (2010). Annotations for opinion mining evaluation in the industrial context of the doxa project. In *the Seventh International Conference on Language Resources and Evaluation*.
- QADIR A. & RILOFF E. (2013). Bootstrapped learning of emotion hashtags hashtags4you. In *the 4th Workshop on Computational Approaches to Subjectivity Sentiment and Social Media Analysis*, Atlanta.
- READ J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *proceedings of The 43rd Annual Meeting of the Association for Computational Linguistics*, p. p.43–48.
- ROBERTS K., ROACH M. A., JOHNSON J., GUTHRIE J. & HARABAGIU S. M. (2012). Empatweet : Annotating and detecting emotions on twitter. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.
- TESNIÈRE L. (1959). Éléments de syntaxe structurale. *Langue française*, **1/1**, p.36–40.

Comparaison d'architectures neuronales pour l'analyse syntaxique en constituants

Maximin Coavoux¹ Benoît Crabbé^{1,2}

(1) ALPAGE, INRIA, Université Paris Diderot, Place Paul Ricœur, 75013 Paris

(2) Institut Universitaire de France

maximin.coavoux@inria.fr, bcrabbe@linguist.univ-paris-diderot.fr

Résumé. L'article traite de l'analyse syntaxique lexicalisée pour les grammaires de constituants. On se place dans le cadre de l'analyse par transitions. Les modèles statistiques généralement utilisés pour cette tâche s'appuient sur une représentation non structurée du lexique. Les mots du vocabulaire sont représentés par des symboles discrets sans liens entre eux. À la place, nous proposons d'utiliser des représentations denses du type plongements (*embeddings*) qui permettent de modéliser la similarité entre symboles, c'est-à-dire entre mots, entre parties du discours et entre catégories syntagmatiques. Nous proposons d'adapter le modèle statistique sous-jacent à ces nouvelles représentations. L'article propose une étude de 3 architectures neuronales de complexité croissante et montre que l'utilisation d'une couche cachée non-linéaire permet de tirer parti des informations données par les plongements.

Abstract.

A Comparison of Neural Network Architectures for Constituent Parsing

The article deals with lexicalized constituent parsing in a transition-based framework. Typical statistical approaches for this task are based on an unstructured representation of the lexicon. Words are represented by discrete unrelated symbols. Instead, our proposal relies on dense vector representations (embeddings) that are able to encode similarity between symbols : words, part-of-speech tags and phrase structure symbols. The article studies and compares 3 increasingly complex neural network architectures, which are fed symbol embeddings. The experiments suggest that the information given by embeddings is best captured by a deep architecture with a non-linear layer.

Mots-clés : Analyse syntaxique en constituants lexicalisée, plongements, réseaux de neurones.

Keywords: Lexicalized constituent parsing, embeddings, neural networks.

1 Introduction

On s'intéresse dans cet article à l'analyse syntaxique lexicalisée pour des grammaires de constituants. Dans le cas robuste, pour gérer l'ambiguïté inhérente à la langue naturelle, on augmente l'algorithme d'analyse d'une méthode de pondération. Ici on propose d'étudier plus spécifiquement le procédé de pondération basé sur un modèle statistique discriminant de type neuronal dans le cas de l'analyse par décalage réduction.

Pour ce problème, les propositions existantes en termes de modèle statistique s'appuient principalement sur la régression logistique multinomiale (dorénavant RLM) (Sagae & Lavie, 2006; Zhao *et al.*, 2013) ou sur le perceptron global (Zhang & Clark, 2009; Zhu *et al.*, 2013; Crabbé, 2014b). Ces architectures s'appuient sur un codage des données par des vecteurs creux de très haute dimensionnalité par l'intermédiaire de fonctions de trait. En première analyse, pour un vocabulaire Σ , chaque mot est codé sur une dimension spécifique d'un vecteur de dimension $|\Sigma|$. Autrement dit, aucune structure n'est donnée à la représentation du lexique. Dans le cas de l'analyse syntaxique lexicalisée, et dans le cas où les paramètres du modèle statistique sont estimés sur un corpus arboré, qui reste de petite taille, ce type de codage a comme inconvénient principal que les paramètres liés aux mots sont très mal estimés. Chaque mot ou couple de mots, dans un modèle d'analyse syntaxique lexicalisé, est en général vu très rarement dans les données.

Dans ce qui suit, on propose de changer la représentation des mots par des représentations vectorielles denses du type plongements lexicaux (*word embeddings*). Dans ce cas, un mot est représenté par un vecteur réel de faible dimension, ce qui permet d'espérer que les propriétés apprises pour un mot par un modèle d'analyse syntaxique puisse se transférer à des mots dont les vecteurs sont similaires. L'introduction de représentations vectorielles pour les symboles discrets nous mène

également à adapter les modèles statistiques habituels à ces nouvelles représentations, ce qui nous pousse à utiliser des modèles neuronaux. De plus, pour homogénéiser et généraliser les représentations, nous codons l'ensemble des symboles discrets de la grammaire (mots, catégories syntagmatiques, parties de discours) par des vecteurs.

L'article est organisé comme suit. Après avoir revu les résultats récents en termes d'analyse syntaxique et de représentations vectorielles de mots (section 2), on rappelle le fonctionnement de l'algorithme d'analyse syntaxique qui sert de support à notre étude (section 3). On propose en section 4 une légère reformulation des méthodes de pondération bien connues pour le cas de représentations creuses, comme la RLM, au cas des représentations vectorielles denses. Nous suggérons ensuite différentes extensions qui permettent d'acquérir à la fois des représentations vectorielles de symboles discrets et une méthode de pondération des analyses. La section 5 nous permet de comparer expérimentalement les différents modèles et de discuter leur formulation dans un cas de prédiction structurée comme l'analyse syntaxique lexicalisée en constituants.

2 Contexte scientifique et état de l'art

L'usage d'analyseurs syntaxiques avec modèle de pondération discriminant n'est en lui-même pas nouveau. L'algorithme que l'on utilise ici est inspiré des travaux de Sagae & Lavie (2006). Nous présentons un algorithme glouton de recherche de la meilleure solution, ce qui le distingue des méthodes de recherche en faisceau telles que présentées par ailleurs par (Zhu *et al.*, 2013; Crabbé, 2014a).

Ce que nous proposons de nouveau, c'est de construire un analyseur dont le modèle de pondération est fondamentalement basé sur des vecteurs denses de réels. De telles représentations, appelées plongements lexicaux ou *word embeddings* sont connues depuis longtemps en traitement automatique des langues et ont été particulièrement popularisées par Mikolov *et al.* (2013). L'usage de représentations vectorielles pour modéliser le langage n'est pas nouveau non plus. On peut d'ailleurs voir d'une certaine manière notre proposition comme une transposition du travail de Bengio *et al.* (2003) pour les modèles de langue au cas de l'analyse syntaxique en constituants. L'usage de représentations vectorielles pour la prédiction structurée en traitement automatique des langues dérive généralement du travail de Collobert & Weston (2008).

L'idée sous-jacente à l'utilisation de représentations vectorielles en analyse syntaxique lexicalisée vient du fait que les corpus arborés sont petits et que les statistiques liées aux mots – considérées comme cruciales – sont en général mal estimées. Un nombre important d'auteurs a tenté de combattre ce problème de dispersion en ajoutant des traits additionnels à des algorithmes d'analyse traditionnels. Ces traits prennent la forme de clusters ou de plongements lexicaux (Bansal *et al.*, 2014; Turian *et al.*, 2010; Candito & Crabbé, 2009; Koo *et al.*, 2008). Mais si des représentations continues ou issues de clusters permettent d'extraire des informations syntaxiques pertinentes, ces informations sont largement redondantes avec celles apportées par des traits classiques, comme les parties du discours (Andreas & Klein, 2014), de sorte que les traits basés sur les représentations continues n'apportent qu'une très faible amélioration pour l'analyse en constituants. C'est une des raisons pour lesquelles on propose ici un modèle d'analyse entièrement fondé sur des représentations vectorielles de symboles discrets.

Deux propositions principales pour l'analyse syntaxique en constituants basée sur des vecteurs ont émergé récemment. D'une part, Henderson (2004); Titov & Henderson (2007) proposent une architecture fondée sur des réseaux de neurones récurrents. Celle-ci prédate toutefois l'arrivée du *deep learning* et cet algorithme utilise des représentations lexicales qui ne correspondent pas à des plongements lexicaux. Plus récemment, Socher *et al.* (2013) proposent l'utilisation de réseaux de neurones comme algorithme de réordonnancement d'analyses pour une PCFG.

À notre connaissance, nous sommes donc les premiers à présenter un modèle de type *deep learning* pour l'analyse syntaxique en constituants. On remarquera toutefois que notre proposition est analogue à celle de Chen & Manning (2014) pour l'analyse en dépendances.

3 Algorithme d'analyse

Cette section introduit rapidement aux propriétés de l'algorithme d'analyse syntaxique par décalage réduction sous-jacent à cette étude. On commence par caractériser la grammaire utilisée par l'analyseur avant d'introduire l'algorithme proprement dit.

Pour l'analyse lexicalisée en constituants, on utilise une grammaire analogue à celle de Crabbé (2014b). On contraint

la grammaire à être en forme binaire lexicalisée (2-LCFG) qui est une forme binarisée dont les règles prennent nécessairement une des formes données en Table 1. Les symboles h, x dénotent des symboles terminaux. Les symboles de la

$$\begin{aligned} A[h] &\rightarrow B[h] C[x] \\ A[h] &\rightarrow B[x] C[h] \\ A[h] &\rightarrow h \end{aligned}$$

TABLE 1 – Formes des règles 2-LCFG

forme $A[h]$ dénotent des symboles non terminaux lexicalisés. Un tel symbole est composé d'un non terminal délexicalisé noté A, B, C et d'un terminal h ou x . Une règle 2-LCFG sera par exemple de la forme $NP[chat] \rightarrow D[le] N[chat]$. Le symbole de la forme $X[h]$ situé en partie droite de la règle est appelé tête de la règle. Une telle grammaire est extraite d'un corpus arboré binarisé. Dans cet article nous avons utilisé la construction décrite par Crabbé (2014b) de manière à garantir que, pour toute analyse d'une phrase de longueur n par un analyseur à décalage réduction, la dérivation a exactement une longueur de $3n - 1$ pas.

L'algorithme d'analyse par décalage-réduction (*shift-reduce*) se fonde sur deux structures de données : une pile S et une file \mathcal{W} , et deux familles d'actions : le décalage et la réduction. On appelle *configuration* ou *état*, le couple $\langle S, \mathcal{W} \rangle$ à une étape t de l'analyse. À l'état initial, la pile est vide et la file contient la séquence de mots à analyser. L'action de décalage consiste à déplacer le sommet de \mathcal{W} et à l'empiler sur S . Une action de réduction est paramétrée par le symbole X . Celle-ci dépile 1 (réduction unaire) ou 2 éléments au sommet de S , et empile le symbole non-terminal X à la place. De plus, comme les éléments de la pile sont lexicalisés, on distingue des réductions gauches et des réductions droites. Une réduction gauche empile un élément $X[h]$ si le sommet de la pile est $B[h] C[x]$ alors qu'une réduction droite empile un élément $X[h]$ si le sommet de la pile est $B[x] C[h]$. En exécutant une action a à partir d'une configuration C_i , on dérive une nouvelle configuration C_{i+1} , ce que l'on note : $C_i \xrightarrow{a} C_{i+1}$. On appelle dérivation une séquence de configurations $C_0 \xrightarrow{a_0} C_1 \xrightarrow{a_1} \dots \xrightarrow{a_{k-1}} C_k$, telle que C_0 est l'état initial. Au terme de l'analyse, si la pile contient seulement l'axiome et si la file est vide, le mot est reconnu. On présente en figure 1 l'algorithme sous forme de règles d'inférence.

ITEM	$\langle S, \mathcal{W} \rangle : w$	
ÉTAT INITIAL	$\langle \emptyset, [w_1 \ w_2 \ \dots \ w_n] \rangle : 0$	
ÉTAT FINAL	$\langle [S], \emptyset \rangle : w$	
DÉCALAGE	$\frac{\langle S, [w_i \ w_{i+1} \ \dots \ w_n] \rangle : w}{\langle S + w_i, [w_{i+1} \ \dots \ w_n] \rangle : w + f(\text{DÉCALAGE}, \langle S, \mathcal{W} \rangle)}$	
RÉDUCTION-GAUCHE-X	$\frac{\langle [\ \alpha \ B[h] \ C[x]], \mathcal{W} \rangle : w}{\langle [\ \alpha \ X[h]], \mathcal{W} \rangle : w + f(\text{RG-X}, \langle S, \mathcal{W} \rangle)}$	$X \rightarrow B \ C \in R$
RÉDUCTION-DROITE-X	$\frac{\langle [\ \alpha \ B[x] \ C[h]], \mathcal{W} \rangle : w}{\langle [\ \alpha \ X[h]], \mathcal{W} \rangle : w + f(\text{RD-X}, \langle S, \mathcal{W} \rangle)}$	$X \rightarrow B \ C \in R$
RÉDUCTION-UNAIRE-X	$\frac{\langle [\ \alpha \ A[h]], \mathcal{W} \rangle : w}{\langle [\ \alpha \ X[h]], \mathcal{W} \rangle : w + f(\text{RU-X}, \langle S, \mathcal{W} \rangle)}$	$X \rightarrow A \in R$

FIGURE 1 – Algorithme pondéré pour une grammaire bilinguale. S est l'axiome de la grammaire, R est l'ensemble des règles de grammaire. Chaque état de l'analyse possède un poids, calculé comme la somme du poids d'une action et de celui de l'état dont il est issu. Chaque action a est pondérée par une fonction $f(a, \langle S, \mathcal{W} \rangle)$, qui dépend de la configuration courante.

L'algorithme n'est pas déterministe. Comme les actions de réduction sont paramétrées par un symbole X qui est un symbole non terminal, en posant Σ l'ensemble de non terminaux, il y a $3|\Sigma| + 1$ actions au total. Il y a en effet trois types de réductions et l'action de décalage. Pour choisir une dérivation, on s'appuie sur une fonction de pondération qui donne un poids à chaque dérivation possible $C_0 \Rightarrow C_k = C_0 \xrightarrow{a_0} \dots \xrightarrow{a_{k-1}} C_k$. Par convention, la dérivation à choisir est celle de poids maximal. La méthode de pondération s'appuie sur une fonction de scorage de la forme :

$$f(C_0 \Rightarrow C_k) = \sum_{i=0}^{k-1} f(a_i, C_i) \quad (1)$$

où f est une fonction qui dépend de l'action a_i à effectuer et d'informations C_i accessibles localement à la configuration (Figure 2). Autrement dit, on suppose que le score d'une dérivation se décompose comme la somme des scores de chacun de ses pas. Dans ce contexte, le problème de l'analyse syntaxique consiste à trouver la meilleure dérivation possible de la phrase c'est-à-dire à donner la solution de :

$$C^* = \underset{C_{0 \Rightarrow 3n-1} \in \text{GEN}(w_1^n)}{\text{argmax}} f(C_{0 \Rightarrow 3n-1})$$

où $\text{GEN}(w_1^n)$ est l'ensemble des analyses possibles pour une phrase de longueur n . L'algorithme que nous présentons ici ne construit pas l'ensemble des dérivations possibles pour une phrase donnée mais procède par recherche gloutonne. À chaque point de choix, l'action qui a le meilleur score est sélectionnée en fonction de l'information localement disponible sans possibilité de remise en question à une étape ultérieure. Cela le distingue des méthodes de recherche par faisceau comme celle de Crabbé (2014a).

4 Représentations denses pour l'analyse syntaxique déterministe

Cette section introduit notre proposition principale : celle-ci porte sur la conception de la fonction locale de scorage $f(a, C)$. On commence par reformuler le système de représentation creux traditionnel, basé sur des fonctions de traits, par un système dense dans lequel chacun des symboles (mots, parties du discours et catégories syntagmatiques) est codé par un vecteur de réels de basse dimensionnalité. On montre que ce changement de représentation ne modifie ni le comportement de la fonction de scorage ni le fonctionnement de l'algorithme d'analyse. La seconde étape du développement introduit une famille d'architectures neuronales qui d'une part permet d'obtenir par apprentissage un dictionnaire de représentations denses pour les symboles discrets et d'autre part qui vise à introduire des interactions entre les variables du modèle.

4.1 Des représentations creuses aux représentations denses

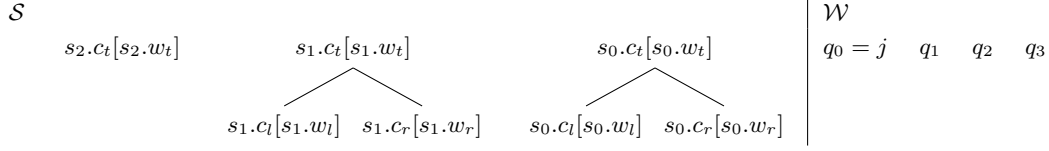
Représentations creuses Le passage de représentations creuses à des représentations denses impacte la fonction de scorage $f(a, C)$ locale à chaque pas d'analyse. Les modèles statistiques discriminants bien connus, comme le perceptron définissent typiquement $f(a, C)$ comme un produit scalaire $f(a, C) = \mathbf{w} \cdot \Phi(a, C)$ où \mathbf{w} dénote un vecteur de poids et $\Phi(a, C)$ dénote un vecteur booléen valué par d fonctions indicatrices $\phi_j(a, C)$ appelées fonction de trait ($1 \leq j \leq d$). Dans le cas de la RLM qui nous intéresse plus particulièrement pour la suite, on normalise le produit scalaire à l'aide de la fonction softmax, et on interprète $f(a, C)$ comme la probabilité de l'action a sachant la configuration C , ce qui donne ¹ :

$$f(a, C) = \frac{\exp(\mathbf{w} \cdot \Phi(a, C))}{\sum_{a' \in A} \exp(\mathbf{w} \cdot \Phi(a', C))} = P(a | C; \mathbf{w}) \quad (2)$$

Les fonctions de trait ont pour paramètres l'action a à évaluer et la configuration C de l'analyseur. Pour une configuration C , les fonctions de trait ont accès au sommet de la pile d'analyse et à la file d'attente. On donne en Figure 2 un exemple de représentation de l'information locale accessible pour valuer les fonctions de traits. Dans cet exemple, les 3 premiers éléments de la pile, les descendants directs des 2 premiers éléments de la pile et les 4 premiers éléments de la file sont accessibles. s_i et q_i représentent respectivement les i^e éléments de la pile et de la file. c_t , c_l et c_r codent respectivement les symboles non-terminaux d'un élément de la pile, de son fils gauche et de son fils droit. w_t , w_l et w_r codent les têtes de ces non-terminaux. On peut adresser les éléments de la file uniquement à l'aide des indices correspondant à leur position dans la phrase : j pour le sommet de la file, $j+1$, $j+2 \dots$ pour les suivants. Les éléments de la file comme les éléments lexicaux de la pile – notés entre crochets – sont eux-même structurés. Il s'agit de tokens lexicaux. On notera par exemple $q_i.w$ pour dénoter la forme lexicale du token et $q_i.tag$ pour dénoter le tag de ce token.

Dans un tel paradigme les fonctions de trait codent effectivement un couple (a, C) sur un vecteur creux de très haute

1. On remarque que la formule de scorage des analyses donnée en (1) reste inchangée si on se place, comme c'est le cas en pratique, dans un espace log probabilisé et en s'appuyant sur l'égalité : $\log \left(\prod_{i=1}^{k-1} P(a_i | C_i; \mathbf{W}) \right) = \sum_{i=1}^{k-1} \log P(a_i | C_i; \mathbf{W})$, c'est-à-dire en posant que $f(a, C) = \log P(a | C; \mathbf{W})$

FIGURE 2 – Information accessible aux fonctions de trait depuis une configuration C

dimensionnalité (d est très grand). Celles-ci ont par exemple une forme du type :

$$\phi_j(a, C) = \begin{cases} 1 & \text{si l'action est } a_i \text{ et le mot tête en sommet de pile } (s_0.w_t) \text{ est le mot } \textit{chat} \\ 0 & \text{sinon} \end{cases}$$

$$\phi_k(a, C) = \begin{cases} 1 & \text{si l'action est } a_i \text{ et le mot tête en sommet de pile } (s_0.w_t) \text{ est le mot } \textit{chat} \\ & \text{et le symbole grammatical } (s_1.c_t) \text{ en sous-sommet de pile est } \textit{Déterminant} \\ 0 & \text{sinon} \end{cases}$$

Si les représentations creuses sont couramment utilisées avec succès en TAL, observons toutefois que les fonctions de trait codent sur des dimensions distinctes des informations liées à la connaissance du lexique. Par exemple on codera indépendamment sur une dimension j un trait lié au mot *chat* et sur une dimension i un trait lié au mot *chien*. Vu la petite taille des corpus arborés utilisés pour l'entraînement, on souhaiterait idéalement qu'un modèle soit capable de tirer parti du fait que *chat* et *chien* sont des mots sémantiquement proches ou encore que des catégories syntagmatiques comme 'phrase principale' et 'phrase subordonnée' sont plus proches que 'nom' et 'verbe'. Autrement dit, on souhaiterait pouvoir améliorer l'estimation des poids en intégrant dans le modèle statistique une notion de similarité entre symboles, qui sera inférée à partir des seules données du corpus lors de l'apprentissage.

Représentations denses Partant de ces deux observations, on propose ici un système de codage qui s'appuie sur des représentations denses de symboles discrets. Un tel codage peut se concevoir, en première approximation, comme un dictionnaire $\delta : \Sigma \rightarrow \mathbb{R}^d$ qui envoie un ensemble Σ de symboles discrets sur des vecteurs de réels de faible dimension (d est petit). Dans un tel contexte on peut redéfinir une fonction de codage $\Phi(C)$ comme une concaténation de vecteurs $\delta(\cdot)$ à valeurs réelles qui code l'ensemble des symboles discrets $x_1 \dots x_c$ accessibles à l'analyseur depuis une configuration donnée :

$$\Phi(C) = [\delta(x_1)\delta(x_2) \dots \delta(x_c)]^T$$

Chaque x_i se voit attribuer une valeur accessible depuis la configuration, comme par exemple $q_0.w$ ou $s_0.w_t$ ou encore $s_1.c_t \dots$ (voir Figure 2). En posant une liste d'actions $A = a^{(1)} \dots a^{(m)}$ de dimension $out = |A|$ et un vecteur $\Phi(C)$ de dimension $in = |\Phi(C)|$, on représente les poids par une matrice $\mathbf{W}^{out \times in}$ dont chaque rangée r représente le vecteur de poids \mathbf{w}_r correspondant à l'action $a^{(r)}$. Dans cette nouvelle représentation on redéfinit par conséquent $f(a, C)$ par $f(a^{(r)}, C) = \mathbf{w}_r \cdot \Phi(C)$. On observe plus généralement qu'en normalisant $f(a^{(r)}, C)$ à l'aide de la fonction softmax, on peut réexprimer $f(a^{(r)}, C)$ par un réseau de neurones élémentaire de type RLM :

$$f(a^{(r)}, C) = \frac{\exp(\mathbf{w}_r \cdot \Phi(C))}{\sum_{j=1}^{|A|} \exp(\mathbf{w}_j \cdot \Phi(C))} \quad (3)$$

Celui-ci ne diffère pas fondamentalement de l'équation (2). La seule différence est que (3) s'interprète plus directement comme un réseau de neurone à une couche. On peut d'ailleurs exprimer ce modèle dans une notation graphique inspirée de Bengio *et al.* (2003) pour décrire les architectures de réseaux de neurones, comme illustré en Figure 3.

On tire de (3) une interprétation probabilisée : $P(a^{(r)} | C; \mathbf{W}) = f(a^{(r)}, C)$. On a montré dans cette section que le passage au modèle dense ne change pas fondamentalement le fonctionnement de l'algorithme d'analyse² : il s'agit essentiellement d'une représentation alternative de la fonction de scorage.

2. En particulier, on remarquera que la représentation adoptée ici, par codage dynamique des symboles discrets se généralise en théorie aux cas d'algorithmes d'analyse qui structurent l'espace de recherche par programmation dynamique.

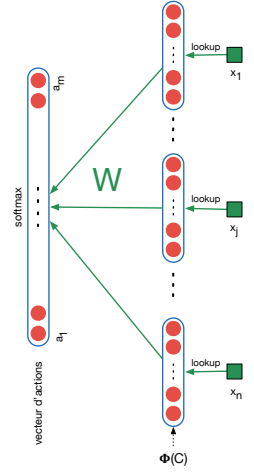


FIGURE 3 – La régression logistique multinomiale vue comme un réseau de neurones à une couche

Par contre, il nous reste à traiter les questions suivantes. D’où vient la fonction δ (ou dictionnaire) et quelles propriétés cette fonction est-elle supposée satisfaire ? Comment, pour le cas dense, exprimer une contrepartie des interactions entre variables, exprimées naturellement dans les modèles creux, et illustrées ci-dessus par une fonction de trait telle que $\phi_k(a, C)$? Comment estimer la matrice de poids ? On propose un ensemble de réponses à ces questions dans la section suivante, en réexprimant plus généralement $f(a^{(r)}, C)$ par une famille de réseaux de neurones artificiels à propagation avant (*feedforward*).

4.2 Architectures de réseaux

De manière à répondre aux trois questions soulevées dans la section précédente, nousinstancions trois architectures de réseaux de neurones, de complexité croissante, potentiellement utilisables pour l’analyse syntaxique. On peut voir ces différentes architectures comme des extensions progressives de modèles de la famille Maximum Entropy Markov Model (MEMM).

Les deux premiers modèles que nous proposons cherchent à illustrer l’introduction du dictionnaire dense dans le modèle statistique. On suppose que dans une configuration donnée l’analyseur a accès aux informations représentées en Figure 2. Il s’agit de valeurs discrètes, que nous notons $x_1 \dots x_c$ et qui regroupent des formes de mots, des catégories syntagmatiques et des catégories morphosyntaxiques. Les formes de mots accessibles sont les formes lexicales représentant la tête lexicale d’un syntagme dans la pile, $s_0.w_t, s_1.w_t, s_2.w_t$. Il s’agit également des formes de mots accessibles dans la file et que nous notons $q_0.w, q_1.w, q_2.w, q_3.w$. Les catégories syntagmatiques sont accessibles dans la pile : $s_0.c_t, s_1.c_t, s_2.c_t$ et finalement les catégories morphosyntaxiques sont accessibles à la fois dans la pile $s_0.tag_t, s_0.tag_l, s_0.tag_r, s_1.tag_t, s_1.tag_l, s_1.tag_r, s_2.tag_t$ et dans la file : $q_0.tag, q_1.tag, q_2.tag, q_3.tag$.

Pour cette raison, tout symbole d’une séquence $x_1, x_2 \dots x_c$ ainsi observée à partir d’une configuration, est systématiquement typé par un type t indiquant si il s’agit d’un mot (W), d’une catégorie syntagmatique (S) ou d’une catégorie morphosyntaxique (T). Pour tirer parti du typage, on distingue dorénavant trois fonctions de dictionnaire $\delta_t : \Sigma_t \mapsto \mathbb{R}^{d_t}$ de telle sorte que chaque type de symbole est potentiellement codé sur des vecteurs de dimension propre à ce type : intuitivement, on peut par exemple envisager de coder les tags sur des vecteurs de dimensionnalité plus petite que les mots. Le dictionnaire de chaque type, est représenté par une table de correspondance (*lookup table*) qui prend la forme d’une matrice $\mathbf{E}^{(t)} \in \mathbb{R}^{|t| \times d_t}$. Chaque symbole x de type t , est assigné à un unique vecteur creux $\mathbf{e}_x \in \mathbb{R}^{1 \times |t|}$ dont une valeur unique est mise à 1. On obtient sa représentation vectorielle dense $\mathbf{e}_x^{(t)} \in \mathbb{R}^{1 \times d_t}$ en le multipliant avec la matrice du type correspondant : $\mathbf{e}_x^{(t)} = \mathbf{e}_x \cdot \mathbf{E}^{(t)}$. Il doit être clair, dans la suite de cet article, que nous ne supposons pas que les représentations denses nous sont données. Au contraire, nous proposons ci-dessous des modèles statistiques qui vont inférer les plongements $\mathbf{e}_x^{(t)}$ par effet de bord de la procédure d’apprentissage.

La représentation du dictionnaire dense étant posée, on propose par la suite de construire d’abord un modèle trivial de type RLM qui ne fait pas appel à la représentation lexicale dense, on propose ensuite un second modèle qui ajoute la représentation lexicale dense au premier modèle et finalement un troisième modèle « profond » qui incorpore une couche cachée supplémentaire.

Régression logistique multinomiale comme modèle de base Ce premier modèle, utilisé comme *baseline* dans nos expériences, représente chaque symbole discret x par un vecteur creux \mathbf{e}_x . Celui-ci représente un modèle naïf de type RLM. On peut l’exprimer comme un réseau de neurones *feedforward* (figure 4, partie gauche). Il se compose essentiellement d’une couche de sortie softmax et s’appuie sur des représentations vectorielles creuses des symboles. Ce modèle, de paramètres $\theta = (\mathbf{W}, \mathbf{b})$, se formule comme suit :

$$\mathbf{h}^{(0)} = \text{softmax} \left(\mathbf{W} [\mathbf{e}_{x_1} \mathbf{e}_{x_2} \dots \mathbf{e}_{x_c}]^T + \mathbf{b} \right)$$

$$P(a|C, \theta) = h_a^{(0)}$$

Réseau de neurones superficiel Le second modèle testé cherche à introduire explicitement une représentation dense de type *embeddings* pour les symboles (mots, catégories, tags). On procède par extension du modèle de RLM en ajoutant une couche intermédiaire entre la couche d’entrée et la couche de sortie (figure 4, partie centrale). Cette couche intermédiaire est obtenue par une opération de *look-up* dans les matrices $\mathbf{E}^{(w)}, \mathbf{E}^{(s)}, \mathbf{E}^{(t)}$, ou de manière équivalente par multiplication

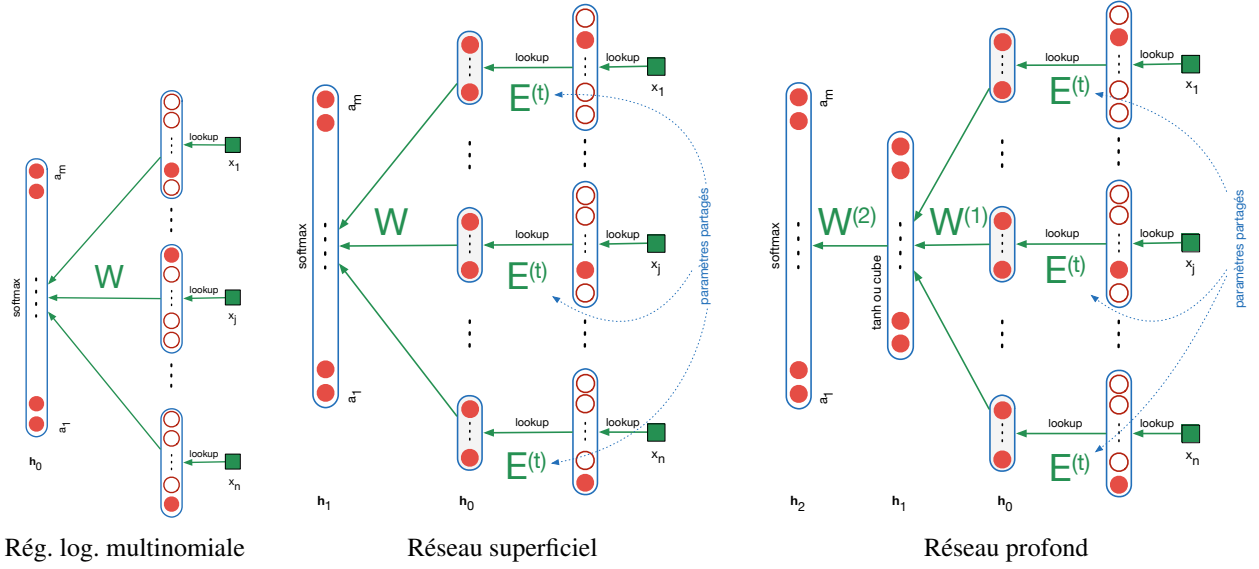


FIGURE 4 – Architectures de réseaux pour la prédiction locale. Les biais ne sont pas représentés sur les schémas.

matricielle entre la matrice $\mathbf{E}^{(t)}$ de type approprié et le vecteur \mathbf{e}_x . La probabilité conditionnelle d'une action prend alors la forme suivante :

$$\begin{aligned} \mathbf{h}^{(0)} &= [\mathbf{e}_{x_1}^{(t_1)} \mathbf{e}_{x_2}^{(t_2)} \dots \mathbf{e}_{x_n}^{(t_n)}]^T & (\text{opération de look-up}) \\ \mathbf{h}^{(1)} &= \text{softmax}(\mathbf{W} \mathbf{h}^{(0)} + \mathbf{b}) \\ P(a|C, \theta) &= h_a^{(1)} \end{aligned}$$

Les paramètres à estimer sont $\theta = (\mathbf{E}^{(w)}, \mathbf{E}^{(s)}, \mathbf{E}^{(t)}, \mathbf{W}, \mathbf{b})$. Il s'agit d'un réseau *feedforward*. Ce second modèle fait apparaître les plongements $\mathbf{e}_{x_i}^{(t_i)}$ qu'ils soient lexicaux, syntagmatiques ou morphosyntaxiques. Autrement dit, cette architecture cherche à tirer parti de similarités distributionnelles entre les symboles et à améliorer l'estimation des paramètres.

Notons toutefois qu'il ne prend pas en compte les interactions entre les variables. Par ailleurs, ce modèle ne permet pas d'apprendre de fonction de décision plus complexe que le modèle de RLM. On peut montrer que sa fonction de décision est équivalente à celle d'un tel modèle en réécrivant le produit $\mathbf{W} \mathbf{h}^{(0)}$:

$$\begin{aligned} \mathbf{W} \mathbf{h}^{(0)} &= \mathbf{W} [\mathbf{e}_{x_1}^{(t_1)} \mathbf{e}_{x_2}^{(t_2)} \dots \mathbf{e}_{x_n}^{(t_n)}]^T \\ &= \mathbf{W} \left(\begin{bmatrix} \mathbf{E}^{(t_1)} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{E}^{(t_1)} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{E}^{(t_n)} \end{bmatrix} [\mathbf{e}_{x_1} \mathbf{e}_{x_2} \dots \mathbf{e}_{x_n}]^T \right) \\ &= \left(\mathbf{W} \begin{bmatrix} \mathbf{E}^{(t_1)} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{E}^{(t_1)} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{E}^{(t_n)} \end{bmatrix} \right) [\mathbf{e}_{x_1} \mathbf{e}_{x_2} \dots \mathbf{e}_{x_n}]^T \\ &= \mathbf{W}' [\mathbf{e}_{x_1} \mathbf{e}_{x_2} \dots \mathbf{e}_{x_n}]^T \end{aligned}$$

où \mathbf{W}' est la matrice de poids d'un modèle de RLM. Au lieu d'estimer directement une matrice de poids, comme dans le modèle de RLM, cette architecture tente d'apprendre une factorisation de cette matrice.

Réseau de neurones profond Dans un troisième temps, nous cherchons à modéliser des interactions entre les représentations denses. Dans le cas de représentations creuses, ces interactions sont capturées à l'aide de conjonctions de fonctions

indicatrices comme celle présentée en section 4.1. Néanmoins, le nombre de fonction de trait à utiliser pour capturer toutes les interactions entre des groupes de 2 voire 3 variables est très grand et les considérer toutes est en général redondant. En ajoutant une couche cachée non-linéaire, on peut extraire automatiquement les informations de la couche précédente pertinentes pour l'analyse (figure 4, partie droite). Nous obtenons alors un réseau *feedforward* profond avec deux couches cachées dont une non-linéaire. La première couche cachée ($\mathbf{h}^{(0)}$) récupère les représentations denses correspondant à chaque symbole de l'entrée x (plongements). La deuxième couche cachée ($\mathbf{h}^{(1)}$) extrait des traits latents à partir de la couche précédente, en lui appliquant une transformation affine suivie d'une fonction d'activation non-linéaire g . La fonction d'activation la plus couramment utilisée est la tangente hyperbolique. Enfin, la couche de sortie ($\mathbf{h}^{(2)}$) est obtenue en appliquant la fonction softmax à une transformation affine de la couche cachée. On obtient ainsi une distribution de probabilités sur l'ensemble des actions étant donné la configuration.

$P(a|C; \theta)$ se calcule de la manière suivante :

$$\begin{aligned} \mathbf{h}^{(0)} &= \left[\mathbf{e}_{x_1}^{(t_1)} \mathbf{e}_{x_2}^{(t_2)} \dots \mathbf{e}_{x_n}^{(t_n)} \right]^T && \text{(opération de look-up)} \\ \mathbf{h}^{(1)} &= g \left(\mathbf{W}^{(1)} \mathbf{h}^{(0)} + \mathbf{b}^{(1)} \right) \\ \mathbf{h}^{(2)} &= \text{softmax} \left(\mathbf{W}^{(2)} \mathbf{h}^{(1)} + \mathbf{b}^{(2)} \right) \\ P(a|C; \theta) &= h_a^{(2)} \end{aligned}$$

Les paramètres du modèles sont alors $\theta = (\mathbf{E}^{(w)}, \mathbf{E}^{(s)}, \mathbf{E}^{(r)}, \mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{b}^{(1)}, \mathbf{b}^{(2)})$. On s'autorisera par la suite à utiliser une fonction d'activation g cubique introduite par Chen & Manning (2014) comme alternative à la fonction tangente hyperbolique.

4.3 Estimation des paramètres

Le corpus d'apprentissage est vu comme un ensemble $\mathcal{D} = \{(C^{(i)}, a^{(i)})\}_{i=1}^N$. Chaque $(C^{(i)}, a^{(i)})$ est un couple, configuration – action de référence, extrait à partir des séquences de dérivation de référence du corpus arboré.

Les modèles présentés modélisent la distribution $P(a^{(r)}|C; \theta)$. Chaque action est alors pondérée par sa log probabilité $f(a, C) = -\log P(a|C; \theta)$. Pour estimer les paramètres, nous minimisons la log vraisemblance négative des données en supposant les données i.i.d. La fonction objective est la suivante :

$$\mathcal{L}(\mathcal{D}; \theta) = -\frac{1}{N} \sum_{i=1}^N \log P(a^{(i)}|C^{(i)}; \theta)$$

où θ est l'ensemble des paramètres du modèle choisi. Les trois modèles sont optimisés en utilisant une descente de gradient stochastique dans son incarnation ADAGRAD (Duchi *et al.*, 2011). Cette méthode est moins sensible que la descente de gradient stochastique standard à la valeur du pas d'apprentissage. Elle permet également une convergence plus rapide. Pour chaque exemple, la valeur du gradient est calculée par l'algorithme de rétropropagation du gradient.

Il faut d'abord remarquer que le vecteur de paramètres θ est initialisé par tirage au sort. Dans le cas du réseau profond, la fonction objective n'est pas convexe, et nous n'avons pas de garantie théorique de convergence vers un minimum global. On remarque finalement qu'on obtient un produit dérivé de l'estimation des paramètres. La couche cachée $\mathbf{h}^{(0)}$ des modèles superficiel et profond code des vecteurs de réels qui sont associés aux symboles d'entrée x_i et qui s'interprètent comme des plongements lexicaux, syntagmatiques ou morphosyntaxiques selon le type du symbole x_i .

5 Expériences

Les expériences s'appuient sur le jeu de données français SPMLR décrit dans (Abeillé *et al.*, 2003; Seddah *et al.*, 2013). Le jeu de données SPMLR instancie les données French Treebank dans deux scénarios : le scénario 'tags prédits' comporte un jeu de test où les tags de référence sont remplacés par des tags prédits par un tagger (exactitude d'étiquetage = 97.35%) et un scénario 'tags donnés' où le jeu de test comporte les tags de référence.

Époque	Corpus d'entraînement			Corpus de développement	
	F	Exactitude	Obj.	F	Exactitude
1	84.40	96.93	0.506	79.33	95.99
2	88.60	97.93	0.478	79.90	96.13
3	90.94	98.47	0.463	79.86	96.07
4	92.72	98.81	0.453	80.16	96.05
5	93.38	98.94	0.450	80.10	96.03
6	93.85	99.00	0.447	80.13	96.02

TABLE 2 – Données d'apprentissage pour le réseau profond dans le scénario 'tags donnés'. Les colonnes 'Exactitude' évaluent la qualité de la prédiction locale sur l'ensemble des couples configuration/action extraits des arbres de référence.

		tags donnés			tags prédits		
	apprentissage	F ≤ 40	F	Cov.	F ≤ 40	F	Cov.
Réseau profond (P+I)	local	83.5	80.7	99.96	81.3	78.3	99.96
Réseau superficiel (P)	local	80.6	77.3	99.96	79.1	75.6	99.96
Rég. log. multinomiale (RLM)	local	79.1	75.6	99.96	77.29	74.2	99.96
Perceptron moyenné	local	82.0	78.8	99.96	80.1	76.9	99.96
Perceptron moyenné (beam=4)	global	83.28	80.04	99.96	81.16	77.60	99.96
Berkeley (Petrov <i>et al.</i> , 2006)	global	86.4	84.0	99.96	83.2	80.7	99.96
Perc. moy. (beam=4, I + morpho.)	global	90.35	87.84	99.99	85.14	82.38	99.69

TABLE 3 – Résultats sur le corpus de test du FRENCH TREEBANK-SPMRL (P)longements, (I)interactions.

Les expériences sont menées avec une implémentation de l'algorithme décrit dans l'article, écrite en C++/Eigen. Les arbres sont binarisés par une markovisation par la tête d'ordre 0. Les expériences sont réalisées sur les données de développement. Pour comparaison, nous donnons également les résultats sur les données de test pour un perceptron moyenné local et un perceptron moyenné structuré (beam = 4) qui utilisent des traits identiques aux modèles neuronaux (et ne prennent pas en compte d'interactions entre variables). On propose également deux derniers modèles qui donnent une idée de l'état de l'art : il s'agit de l'analyseur de Petrov *et al.* (2006), et enfin un modèle global – fondé sur des fonctions de traits riches incluant des interactions entre variables et des informations morphologiques supplémentaires notamment pour les mots composés – qui est une extension de (Crabbé, 2014b). Nous mesurons le F-Score et la couverture sur les données débinaisées à l'aide du logiciel *evalb*³.

Les 3 modèles neuronaux ont été entraînés pour minimiser la fonction objective. L'apprentissage des réseaux de neurones est réputé sensible à de nombreux hyperparamètres. Pour calibrer les différents hyperparamètres et initialiser les matrices de poids, nous avons suivi les recommandations proposées par la communauté (Bengio, 2012; Do *et al.*, 2014). Le pas d'apprentissage initial est 0.04. Il est divisé par deux à la fin d'une itération si la fonction objective croît par rapport à l'itération précédente. Nous avons utilisé 300 unités dans la couche cachée pour le réseau profond. La dimension des plongements lexicaux d_w est fixée à 50. Celles des plongements de catégories syntagmatiques et de tags sont fixées à $d_s = d_w = 20$. Les coefficients initiaux des plongements sont tous initialisés aléatoirement dans l'intervalle $[-0.01, 0.01]$. Les matrices de poids des réseaux de neurones sont également initialisées aléatoirement. Nous avons comparé deux fonctions non linéaires pour le réseau profond : la tangente hyperbolique et la fonction cube proposée par Chen & Manning (2014). Nous présentons les résultats pour la fonction cube qui produit des résultats légèrement meilleurs que la tangente hyperbolique. Pour chaque modèle, le nombre d'itérations choisi est celui qui maximise le F-score sur le corpus de développement dans le scénario 'tags donnés' : 4 pour le réseau profond, 6 pour le réseau superficiel, 8 pour le modèle de RLM. On donne en table 2 les résultats de l'apprentissage sur le corpus de développement pour le réseau profond. Les 5 premiers modèles du tableau ont accès aux mêmes informations (traits) qui sont données au début de la section 4.2.

Les résultats sont présentés en table 3. Tout d'abord, on observe que la substitution de représentations denses à des représentations creuses entraîne un gain sur le modèle de RLM, sans augmenter l'expressivité de ce modèle. Cela suppose qu'une représentation structurée du lexique — et plus généralement des informations pertinentes pour désambiguïser — permet un meilleur apprentissage du comportement des mots et des relations entre les mots. L'utilisation d'une couche cachée non linéaire améliore encore ce résultat. Cela suggère que les classifieurs habituellement utilisés pour l'analyse syntaxique (perceptron, RLM) ne sont pas suffisamment puissants pour tirer pleinement parti des informations contenues dans les plongements lexicaux. La couche cachée permet d'apprendre des interactions entre les différents symboles et de sélectionner les informations pertinentes pour la désambiguïstation.

3. Nous utilisons la version standard du logiciel telle que distribuée sur le site <http://nlp.cs.nyu.edu/evalb>.

Le réseau de neurones profond utilisé apprend des représentations denses pour toutes les informations qu’il utilise lors de l’analyse. On donne en figure 5 une visualisation des plongements obtenus pour les symboles non-terminaux après une réduction de dimensionalité à l’aide de t-SNE (van der Maaten & Hinton, 2008). Notons que les représentations de symboles obtenus pourraient être en principe utilisées comme traits semi-supervisés pour toute autre tâche de traitement automatique des langues.

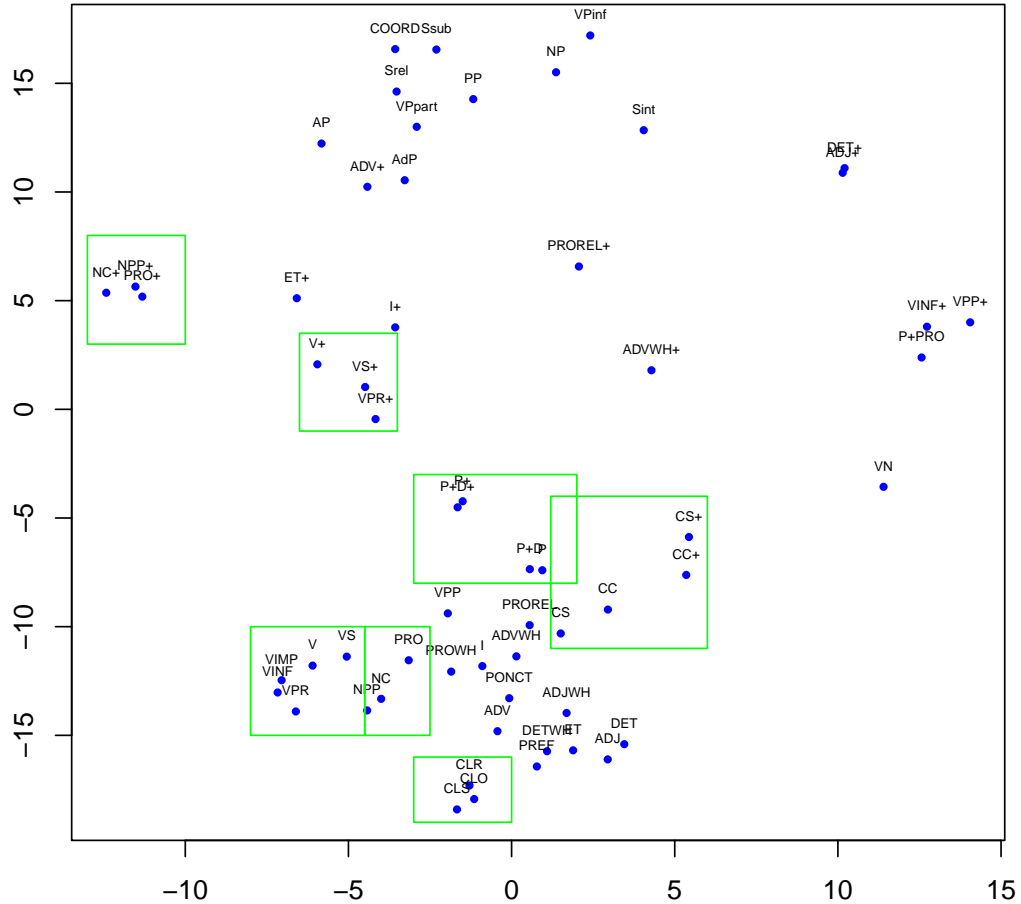


FIGURE 5 – Représentations denses de symboles non-terminaux apprises par le réseau de neurones profond. Pour leur visualisation, ces représentations ont été projetées sur 2 dimensions à l’aide de t-SNE. Le modèle parvient à capturer certaines similarités entre les différents symboles. Par exemple, les représentations pour différents tags de verbes (V, VS, VIMP, VINP, ...) sont proches, de même que celles pour les clitiques (CLO, CLS, CLR). Les symboles représentant des mots composés (suffixés par +) se regroupent entre eux ou avec leur contrepartie non composée.

L’architecture que nous proposons reste encore en deçà des performances d’un analyseur génératif comme celui de Petrov *et al.* (2006), ou de perceptrons globaux comme celui de Crabbé (2014b). Nous envisageons plusieurs suites possibles pour améliorer ce résultat.

Tout d’abord, le modèle utilisé est un modèle local appliqué à un problème de classification structuré. Les modèles neuronaux présentés dans cet article utilisent tous une recherche gloutonne, de sorte que des erreurs en début d’analyse sont irrécupérables et peuvent compromettre la suite des opérations. Traditionnellement, étant donné la taille de l’espace de recherche ($\mathcal{O}(|A|^{3n-1})$), les analyseurs par transition s’appuient sur des algorithmes de recherche approximative pour rester efficace, le plus souvent une recherche par faisceau (Zhang & Clark, 2009; Zhu *et al.*, 2013; Crabbé, 2014a) ou une recherche de type meilleur d’abord (Zhao *et al.*, 2013). Or, des données empiriques (Zhang & Nivre, 2012) ont suggéré

Réseau profond	tags prédits		
	F <= 40	F	Cov.
beam=1	81.3	78.3	99.96
beam=2	81.63	78.5	99.96
beam=4	81.9	78.9	99.96
beam=8	81.9	78.9	99.96

TABLE 4 – Résultats pour le réseau profond avec une recherche par faisceau.

que, pour l'analyse syntaxique, l'utilisation d'un faisceau de recherche n'apporte un gain que dans le cas d'un algorithme d'apprentissage global. Cette observation se confirme pour l'architecture proposée (table 4), le gain est limité et plafonne avec une petite taille de faisceau. Ce gain est plus important si l'on passe du perceptron local au perceptron structuré.

De plus, la généralisation à la classification structurée pose problème. Reformuler un modèle neuronal global qui serait une contrepartie de CRF ou du perceptron structuré est rendu délicat par l'introduction de la couche cachée. Des méthodes hybrides ont été proposées pour la prédiction structurée à l'aide d'un réseau profond et de méthodes globales de type CRF (Wang & Manning, 2013). Mais celles-ci sont plus coûteuses en terme de temps de calcul. En revanche, un paradigme pour l'apprentissage structuré comme SEARN (Daumé III *et al.*, 2009), basé sur un classifieur local pourrait se révéler intéressant. De même, les méthodes plus traditionnelles (réseaux récurrents) qui cherchent à améliorer la modélisation de l'historique d'analyse ont donné de bons résultats également (Henderson, 2004; Titov & Henderson, 2007).

En second lieu, les architectures présentées apprennent elles-mêmes les représentations denses de symboles qu'elles utilisent. Au lieu de les initialiser aléatoirement, il est possible d'initialiser ces représentations pour les mots par des plongements lexicaux appris sur de gros volumes de données non annotées, par exemple à l'aide du modèle SKIP-GRAM (Mikolov *et al.*, 2013), ou d'un modèle de langue comme celui de Bengio *et al.* (2003). Cela permet généralement d'améliorer l'apprentissage (Chen & Manning, 2014).

Finalement, les sources d'informations utilisées par le classifieur sont encore relativement limitées si on regarde en comparaison les patrons de traits classiquement utilisés par les modèles d'analyse discriminants. Il serait en fait assez facile de donner au modèle accès à d'autres sources d'information : notons que si nous nous sommes limités à 3 types d'informations, le modèle proposé est assez général pour utiliser des représentations denses pour un nombre arbitraire de types de symboles, par exemple pour des informations morphologiques sous la forme de suffixes (voir également Collobert & Weston 2008).

6 Conclusion

L'article propose plusieurs architectures neuronales pour l'analyse automatique en constituants. Leur comparaison suggère 2 choses. En premier lieu, le remplacement de représentations creuses des symboles de la grammaire par des représentations denses permet d'améliorer l'estimation du modèle statistique. En second lieu, une architecture comprenant au moins une couche cachée non-linéaire tire mieux parti des informations données par ces plongements qu'un classifieur linéaire. La suite des travaux portera principalement sur le problème de la prédiction structurée et sur l'intégration de représentations pré-entraînées.

Références

- ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a Treebank for French. In *Treebanks : Building and Using Parsed Corpora*, p. 165–188. Springer.
- ANDREAS J. & KLEIN D. (2014). How much do word embeddings encode about syntax ? *Proceedings of ACL*.
- BANSAL M., GIMPEL K. & LIVESCU K. (2014). Tailoring continuous word representations for dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 809–815, Baltimore, Maryland : Association for Computational Linguistics.
- BENGIO Y. (2012). *Practical recommendations for gradient-based training of deep architectures*. Rapport interne arXiv :1206.5533, U. Montreal, Lecture Notes in Computer Science Volume 7700, Neural Networks : Tricks of the Trade Second Edition, Editors : Grégoire Montavon, Geneviève B. Orr, Klaus-Robert Müller.
- BENGIO Y., DUCHARME R. & VINCENT P. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3, 1137–1155.

- CANDITO M. & CRABBÉ B. (2009). Improving generative statistical parsing with semi-supervised word clustering. In *IWPT*, p. 138–141 : The Association for Computational Linguistics.
- CHEN D. & MANNING C. D. (2014). A fast and accurate dependency parser using neural networks. In *Empirical Methods in Natural Language Processing (EMNLP)*.
- COLLOBERT R. & WESTON J. (2008). A unified architecture for natural language processing : Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning, ICML '08*, p. 160–167, New York, NY, USA : ACM.
- CRABBÉ B. (2014a). An LR-inspired generalized lexicalized phrase structure parser. In *Proceedings of the twenty-fifth International Conference on Computational Linguistics*, Dublin, Ireland.
- CRABBÉ B. (2014b). Un analyseur discriminant de la famille lr pour l'analyse en constituants. In *TALN*.
- DAUMÉ III H., LANGFORD J. & MARCU D. (2009). Search-based structured prediction.
- DO Q.-K., ALLAUZEN A. & YVON F. (2014). Modèles de langue neuronaux : une comparaison de plusieurs stratégies d'apprentissage. In *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, p. 256–267, Marseille, France : Association pour le Traitement Automatique des Langues.
- DUCHI J., HAZAN E. & SINGER Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, **12**, 2121–2159.
- HENDERSON J. (2004). Discriminative training of a neural network statistical parser. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, p. 95–102, Barcelona, Spain.
- KOO T., CARRERAS X. & COLLINS M. (2008). Simple semi-supervised dependency parsing. In *Proceedings of ACL-08 : HLT*, p. 595–603, Columbus, Ohio : Association for Computational Linguistics.
- MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. *CoRR*, **abs/1301.3781**.
- PETROV S., BARRETT L., THIBAU R. & KLEIN D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, p. 433–440, Sydney, Australia : Association for Computational Linguistics.
- SAGAE K. & LAVIE A. (2006). A best-first probabilistic shift-reduce parser. In *Proceedings of the COLING/ACL on Main conference poster sessions*, p. 691–698 : Association for Computational Linguistics.
- SEDDAH D., TSARFATY R., KÜBLER S., CANDITO M., CHOI J. D., FARKAS R., FOSTER J., GOENAGA I., GOJE-NOLA GALLETEBEITIA K., GOLDBERG Y., GREEN S., HABASH N., KUHLMANN M., MAIER W., NIVRE J., PRZEPIÓRKOWSKI A., ROTH R., SEEKER W., VERSLEY Y., VINCZE V., WOLIŃSKI M., WRÓBLEWSKA A. & DE LA CLERGERIE E. V. (2013). Overview of the SPMRL 2013 shared task : A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, p. 146–182, Seattle, Washington, USA : Association for Computational Linguistics.
- SOCHER R., BAUER J., MANNING C. D. & NG A. Y. (2013). Parsing With Compositional Vector Grammars. In *ACL*.
- TITOV I. & HENDERSON J. (2007). Constituent parsing with incremental sigmoid belief networks. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, p. 632–639, Prague, Czech Republic : Association for Computational Linguistics.
- TURIAN J., RATINOV L. & BENGIO Y. (2010). Word representations : A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, p. 384–394, Stroudsburg, PA, USA : Association for Computational Linguistics.
- VAN DER MAATEN L. & HINTON G. (2008). Visualizing high-dimensional data using t-sne.
- WANG M. & MANNING C. D. (2013). Effect of non-linear deep architecture in sequence labeling. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*.
- ZHANG Y. & CLARK S. (2009). Transition-based parsing of the chinese treebank using a global discriminative model. In *Proceedings of the 11th International Conference on Parsing Technologies, IWPT '09*, p. 162–171, Stroudsburg, PA, USA : Association for Computational Linguistics.
- ZHANG Y. & NIVRE J. (2012). Analyzing the effect of global learning and beam-search on transition-based dependency parsing. In *COLING (Posters)*, p. 1391–1400.
- ZHAO K., CROSS J. & HUANG L. (2013). Optimal incremental parsing via best-first dynamic programming. In *EMNLP*, p. 758–768 : ACL.
- ZHU M., ZHANG Y., CHEN W., ZHANG M. & ZHU J. (2013). Fast and accurate shift-reduce constituent parsing. In *ACL (1)*, p. 434–443 : The Association for Computer Linguistics.

...des conférences enfin disons des causeries...

Détection automatique de segments en relation de paraphrase dans les reformulations de corpus oraux

Natalia Grabar¹ Iris Eshkol-Taravella²

(1) CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France

natalia.grabar@univ-lille3.fr

(2) CNRS UMR 7270 LLL, Université d'Orléans, 45100 Orléans, France

iris.eshkol@univ-orleans.fr

Résumé. Notre travail porte sur la détection automatique des segments en relation de reformulation paraphrastique dans les corpus oraux. L'approche proposée est une approche syntagmatique qui tient compte des marqueurs de reformulation paraphrastique et des spécificités de l'oral. Les données de référence sont consensuelles. Une méthode automatique fondée sur l'apprentissage avec les CRF est proposée afin de détecter les segments paraphrasés. Différents descripteurs sont exploités dans une fenêtre de taille variable. Les tests effectués montrent que les segments en relation de paraphrase sont assez difficiles à détecter, surtout avec leurs frontières correctes. Les meilleures moyennes atteignent 0,65 de F-mesure, 0,75 de précision et 0,63 de rappel. Nous avons plusieurs perspectives à ce travail pour améliorer la détection des segments en relation de paraphrase et pour étudier les données depuis d'autres points de vue.

Abstract.

...des conférences enfin disons des causeries... **Automatic detection of segments with paraphrase relation in spoken corpora rephrasings.**

Our work addresses automatic detection of segments with paraphrastic rephrasing relation in spoken corpus. The proposed approach is syntagmatic. It is based on paraphrastic rephrasing markers and the specificities of the spoken language. The reference data used are consensual. Automatic method based on machine learning using CRFs is proposed in order to detect the segments that are paraphrased. Different descriptors are exploited within a window with various sizes. The tests performed indicate that the segments that are in paraphrastic relation are quite difficult to detect. Our best average reaches up to 0.65 F-measure, 0.75 precision, and 0.63 recall. We have several perspectives to this work for improving the detection of segments that are in paraphrastic relation and for studying the data from other points of view.

Mots-clés : Corpus oraux, Paraphrase, Reformulation, Marqueur de reformulation paraphrastique, Apprentissage supervisé.

Keywords: Spoken Corpora, Paraphrase, Reformulation, Paraphrastic Reformulation Marker, Supervised Learning.

1 Introduction

La paraphrase joue un rôle important dans la langue. Ceci justifie en particulier la conception de la langue comme d'un système de paraphrasage par certains linguistes (Melčuk, 1988). Voilà quelques exemples de contextes où la paraphrase se relève importante :

- Dans les cours de langues, on demande souvent aux élèves de paraphraser des expressions ou des phrases afin de contrôler leur maîtrise de la langue, qu'elle soit maternelle ou étrangère ;
- De la même manière, il est possible de contrôler la compréhension d'une idée. Les premiers exercices de paraphrasage aurait ainsi apparus en effectuant l'exégèse des textes anciens : des textes sacrés (Bible, Coran, Tora) d'abord, et ensuite des textes théologiques, philosophiques ou scientifiques. Notons que la production d'explications ou de commentaires sur ces textes occupe toujours une place importante en philosophie, théologie et philologie des langues anciennes ;
- De manière plus naturelle, les locuteurs recourent à la paraphrase pour préciser leurs pensées et les transmettre au mieux à leurs interlocuteurs. Dans ce cas, la paraphrase découle de l'activité de reformulation. Notons que l'écriture

relève également du paraphrasage : entre les différentes versions d'une oeuvre littéraire (Fuchs, 1982), d'un article de Wikipédia (Vila *et al.*, 2014) ou d'un article scientifique, les auteurs peuvent réécrire plusieurs fois leur texte avant de produire celui qui les satisfait enfin mieux.

En dehors des situations communes de la vie, la paraphrase joue aussi un rôle important pour différentes applications de TAL (Androutsopoulos & Malakasiotis, 2010; Madnani & Dorr, 2010; Bouamor *et al.*, 2012). L'objectif est alors de détecter les expressions linguistiques formellement différentes mais véhiculant une sémantique similaire ou proche :

- En recherche et extraction d'information, et dans les systèmes de questions-réponses, les paraphrases permettent d'augmenter la couverture des résultats grâce aux expressions équivalentes entre les requêtes ou les questions et les textes dans lesquelles les réponses doivent être trouvées. Par exemple, les paires {*infarctus du myocarde*; *crise cardiaque*} et {*maladie d'Alzheimer*; *maladie neurodégénérative*} contiennent des expressions différentes qui véhiculent cependant une sémantique identique ou proche. Si le système automatique dispose de telles connaissances, la couverture et la qualité de ses résultats peuvent être améliorées ;
- En traduction automatique, les paraphrases permettent d'éviter des répétitions lexicales et peuvent introduire ainsi une légèreté du texte cible. Par exemple, le segment original en anglais *Figure 10.2 shows money growth and output growth. There is a strong, but not absolute, link between money growth and output growth* est traduit en français de manière à éviter la répétition *money growth and output growth* : *Le graphique 10.2 montre une augmentation des fonds et de la production. Il existe entre ces éléments un lien étroit, bien que non absolu* (Scarpa, 2010). Différentes langues ont en effet une tolérance variable vis-à-vis des répétitions (Hatim & Mason, 1990; Baker, 1992) ;
- L'inférence textuelle (Dagan *et al.*, 2013) consiste à établir une relation entre deux segments textuels, appelés Texte et Hypothèse. L'inférence textuelle est une relation directionnelle, dans laquelle la vérité de l'Hypothèse peut être inférée à partir du sens du Texte, ou, en d'autres mots, il est possible de vérifier si l'Hypothèse est subsumée par le Texte. Par exemple, le Texte *The drugs that slow down or halt Alzheimer's disease work best the earlier you administer them* permet d'inférer que l'Hypothèse *Alzheimer's disease is treated by drugs* est vraie ; par contre, l'Hypothèse *Alzheimer's disease is cured by drugs* ne peut pas être inférée à partir de ce Texte. Dans cet exemple, le lien de paraphrase entre les paires {*administer drugs*; *treated by drugs*} et {*slow down or halt*; *cured by drugs*} permet justement d'établir ce lien d'inférence entre le Texte et l'Hypothèse.

Comme ces quelques exemples montrent, en fonction des objectifs poursuivis, la paraphrase requiert des phénomènes linguistiques plus ou moins nombreux. L'étendue des classifications de paraphrases proposées peut être ainsi plus ou moins couvrante, et aller de 25 catégories de paraphrases (Bhagat & Hovy, 2013) à 67 fonctions lexicales pour le paraphrasage (Melčuk, 1988). Le plus souvent, ces classifications focalisent sur un aspect donné, comme par exemple les moyens linguistiques mis en oeuvre (Melčuk, 1988; Vila *et al.*, 2011; Bhagat & Hovy, 2013), la taille de l'unité paraphrasée (Flottum, 1995; Fujita, 2010; Bouamor, 2012), les connaissances requises (Milicevic, 2007), le registre de langue. À notre connaissance, la seule classification multidimensionnelle est celle de (Milicevic, 2007) : elle couvre plusieurs des dimensions mentionnées. Dans notre travail précédent, nous avons également proposé de travailler sur une classification à plusieurs dimensions (Eshkol-Taravella & Grabar, 2014). Elle prend en compte les dimensions suivantes :

- la catégorie syntaxique des segments paraphrasés,
- le type de la relation lexicale entre ces segments (*e.g.* hyperonyme, synonyme, antonyme, instance, méronyme),
- le type de la modification lexicale (*e.g.* remplacement, suppression, ajout),
- le type de la modification morphologique (*i.e.* flexion, dérivation, composition),
- le type de la modification syntaxique (*e.g.* passif/actif),
- le type de la relation pragmatique liée aux fonctionnalités de la paraphrase et de la reformulation (*e.g.* définition, explication, précision, résultat, correction linguistique, correction référentielle, équivalence).

Dans notre travail, nous adoptons donc une acception large de la paraphrase.

2 Travaux existants en acquisition automatique de paraphrases

Plusieurs approches sont proposées pour la détection automatique de la paraphrase. De manière générale, ces approches reposent sur les propriétés paradigmatiques des mots et sur leur capacité de se substituer mutuellement dans un contexte donné. Ces approches dépendent du type de corpus exploités. Quatre types de corpus sont généralement distingués : corpus monolingues, corpus monolingues parallèles, corpus monolingues comparables, corpus bilingues parallèles.

Corpus monolingues. En corpus monolingues, deux types d'approches sont à noter :

- la similarité des chaînes d'édition permet de détecter les unités linguistiques (mots, syntagmes, etc) qui montrent des traits communs de surface et permettent de rapprocher deux chaînes comme {*When did Charle de Gaulle die ?*; *Charles de Gaulle died in 1970*} (Malakasiotis & Androutsopoulos, 2007),

- les méthodes distributionnelles permettent de détecter les unités qui apparaissent dans des contextes similaires, auquel cas ces unités montrent aussi des vecteurs contextuels ou syntaxiques similaires, et sont alors de bons candidats pour la paraphrase (e.g. {*Y is solved by X; Y is resolved in X*}) (Lin & Pantel, 2001; Pasca & Dienes, 2005).

Corpus monolingues parallèles. Lorsqu'un texte dans une langue est traduit plus d'une fois dans une autre langue, les traductions de ce texte permettent de constituer un corpus monolingue parallèle. Un des corpus les plus utilisés est constitué des traductions en anglais de *20 000 lieux sous la mer* de Jules Verne. Lorsque les phrases de tels corpus sont alignées, il devient alors possible de les exploiter grâce aux méthodes d'alignement de mots (Och & Ney, 2000). Différentes méthodes ont été proposées pour l'exploitation de tels corpus (Barzilay & McKeown, 2001; Ibrahim *et al.*, 2003; Quirk *et al.*, 2004). Elles permettent d'extraire les paraphrases comme {*countless; lots of*}, {*undertone; low voice*}, {*shrubs; bushes*}, {*refuse; say no*}, {*dull tone; gloom*}, {*sudden appearance; apparition*} (Barzilay & McKeown, 2001).

Corpus monolingues comparables. Les corpus monolingues comparables contiennent typiquement des textes produits indépendamment sur un même événement, comme par exemple les articles de presse qui couvrent l'actualité. La cohérence thématique de ces textes d'un côté et les méthodes distributionnelles ou bien l'alignement de phrases comparables de l'autre côté permettent d'induire les relations de paraphrase entre les segments de texte (Shinyama *et al.*, 2002; Sekine, 2005; Shen *et al.*, 2006). En particulier, les entités nommées et les nombres font partie des repères pour l'extraction de paraphrases comme {*PERS1 killed PERS2; PERS1 let PERS2 die from loss of blood*} ou {*PERS1 shadowed PERS2; PERS1 kept his eyes on PERS2*} (Shinyama *et al.*, 2002).

Corpus bilingues parallèles. Les corpus bilingues parallèles, qui contiennent typiquement la traduction d'un texte dans une autre langue, peuvent aussi être utilisés pour la détection de la paraphrase. Les traductions multiples d'une expression ou d'un mot peuvent alors correspondre aux paraphrases (Bannard & Callison-Burch, 2005; Callison-Burch *et al.*, 2008; Kok & Brockett, 2010). Par exemple, les paraphrases {*under control; in check*} peuvent être extraites parce qu'elles sont utilisées pour la traduction de *unter Kontrolle* dans différents contextes (Bannard & Callison-Burch, 2005).

3 Objectifs

Dans notre précédent travail, nous avons commencé à étudier les reformulations paraphrastiques de l'oral, formées autour de trois marqueurs de reformulation paraphrastique : *c'est-à-dire*, *je veux dire* et *disons* dans l'exemple (1)). Nous avons en particulier proposé un guide pour l'annotation des reformulations paraphrastiques et l'avons testé lors de l'annotation des tours de parole de corpus *ESLO1* et *ESLO2*. Nous avons également défini et testé une méthode à base de règles pour distinguer, au sein d'un ensemble de tours de parole comportant un des trois marqueurs étudiés, les tours de parole qui comportent des reformulations paraphrastiques. Dans le travail actuel, nous poursuivons nos objectifs. D'une part, nos données de référence sont consensuelles et comportent plus d'exemples annotés. D'autres part, nous nous focalisons maintenant sur la détection automatique de segments faisant partie des reformulations paraphrastiques. Nous mettons en place pour ceci un système par apprentissage supervisé. La structure étudiée est de type : *segment1 marqueur-de-reformulation-paraphrastique segment2*. Les marqueurs de reformulation paraphrastique établissent alors le lien sémantique entre les deux segments en relation de paraphrase. Dans l'exemple (1), il s'agit du segment source *des conférences* et du segment cible *des causeries*. Il a été observé que, dans plusieurs cas, les segments impliqués n'ont aucun lien sémantique évident, mais, grâce au marqueur de reformulation et à la relation de paraphrase établie, ce lien peut apparaître (Gulich & Kotschi, 1983; Rossari, 1993). Notre objectif principal est donc de trouver les critères formalisables qui permettent de détecter automatiquement les segments en relation de paraphrase et d'établir ce lien sémantique entre ces segments. Il est possible que les corpus étudiés contiennent d'autres reformulations paraphrastiques introduites par d'autres phénomènes (d'autres marqueurs, l'intonation, les pauses...). Cependant, nous n'étudions pas ces autres reformulations paraphrastiques.

- (1) *des conférences y en a assez souvent sur France culture enfin disons des causeries* [ESLO1_ENT_121_C]

Nous présentons d'abord les données linguistiques que nous traitons (section 4). Nous présentons ensuite la méthode (section 5), et les résultats obtenus (section 6). Nous terminons avec les orientations pour les travaux futurs (section 7).

4 Données linguistiques

Nous utilisons plusieurs types de données linguistiques : les marqueurs de reformulation paraphrastique (section 4.1), les corpus traités (section 4.2), les marqueurs de disfluence (section 4.3) et les données de référence grâce auxquelles le système automatique est créé et évalué (section 4.4).

4.1 Marqueurs de reformulation paraphrastique (MRP)

Nous exploitons trois marqueurs de reformulation paraphrastique (MRP) : *c'est-à-dire* (Gulich & Kotschi, 1983; Hölker, 1988; Beeching, 2007), *je veux dire* (Teston-Bonnard, 2008) et *disons* (Hwang, 1993; Petit, 2009; Saunier, 2012). Le point commun entre eux est qu'ils sont formés à partir du même verbe *dire*. Le marqueur *c'est-à-dire* est le plus lexicalisé et étudié des trois. Ces trois marqueurs sont reconnus pour leur capacité d'introduire les reformulations paraphrastiques, mais ils peuvent également jouer d'autres rôles dans le discours, comme par exemple l'argumentation ou les disfluences.

4.2 Provenance et composition de corpus

Nous travaillons avec les corpus ESLO (Enquêtes Sociolinguistiques à Orléans) (Eshkol-Taravella *et al.*, 2012) : *ESLO1* et *ESLO2*. *ESLO1*, la première enquête sociolinguistique à Orléans, a été réalisée en 1968-1971 par des professeurs de français de l'University of Essex, Language Centre, Colchester (Royaume-Uni), en collaboration avec des membres du B.E.L.C. (Bureau pour l'étude de l'enseignement de la langue et de la civilisation françaises de Paris). Le corpus *ESLO1*, constitué à Orléans mais archivé ensuite de manière fragmentaire ailleurs, est revenu dans les années 1990 au LLL (Laboratoire Ligérien de Linguistique). Le laboratoire a mis au format standard ce corpus d'enquêtes sociolinguistiques. Il comprend 300 heures de parole (4 500 000 mots environ) et inclue une gamme d'enregistrements variés. En prenant en compte l'expérience d'*ESLO1* et l'évolution des cadres théoriques et méthodologiques de la constitution et de l'exploitation de grands corpus oraux à visée variationniste, une nouvelle enquête *ESLO2* a été entamée en 2008. À terme, *ESLO2* comprendra plus de 350 heures d'enregistrements afin de former avec *ESLO1* un corpus de plus de 700 heures et d'atteindre les dix millions de mots. Les corpus *ESLO1* et *ESLO2* sont accessibles en ligne (<http://eslo.tge-adonis.fr/>).

4.3 Marqueurs de disfluence

Nous utilisons un ensemble de marqueurs de disfluence : *allez, allons, alors, là, enfin, euh, heu, bah, ben, hm, hum, hein, quoi, ah, oh, donc, bon, bè, eh*.

4.4 Données de référence

Comme observé dans la littérature, les marqueurs de reformulation paraphrastique peuvent apparaître dans des emplois qui ne sont pas toujours paraphrastiques (Gulich & Kotschi, 1983; Flottum, 1995; Rossari, 1993). L'objectif de l'annotation manuelle est de faire cette distinction et de marquer les segments en relation de paraphrase. Ainsi, l'exemple (1) contient une reformulation paraphrastique, alors que les exemples (2) et (3) ne contiennent pas de relations de paraphrase. Dans ce dernier cas, les marqueurs de reformulation paraphrastique peuvent ainsi être associés aux marqueurs discursifs ou aux disfluences. Lorsqu'un tour de parole contient une reformulation paraphrastique, les annotations plus précises, comme indiquées à la fin de la section 1, sont établies. Les segments en relation de reformulation paraphrastique peuvent être de différentes tailles et avoir différentes fonctions syntaxiques : noms, adjectifs, verbes, adverbes et pronoms ; groupes nominaux (exemple en (1)), verbaux (exemple en (12)), adjectivaux ou adverbiaux ; propositions (exemple en (10)) ; présentateur...

- (2) *est-ce que vous remarquez une différence sensible entre vos différents clients dans leur façon de choisir la viande dans ce qu'ils achètent et caetera indépendamment <MRP>disons</MRP> de leurs oui origines de classe* [ESLO1_ENT_001_C]

- (3) *mais il y a des termes qui sont plus ou moins euh euh <MRP>disons</MRP> euh grossiers qui sont employés plus ou- plus facilement euh par certaines couches de la société en fonction des fréquentations des uns ou des autres quoi* [ESLO1_ENT_003_C]
- (4) *euh <VP1>démocratiser l'enseignement</VP1> <MRP>c'est-à-dire</MRP> <VP2 rel-lex="syn(démocratiser/ permettre à tout le monde) syn(enseignement/faculté)" modif-lex="ajout(rentre à)" rel-pragm="explic"> per-
mettre à tout le monde de rentrer en faculté</VP2>* [ESLO1_ENT_121_C]

Trois paires d'annotateurs ont participé dans la création des données de référence. Après une annotation indépendante, des séances de consensus ont été tenues. Le consensus porte sur la présence de la reformulation paraphrastique et sur les segments source et cible.

Tout marqueur de reformulation paraphrastique confondu, 611 tours de parole du corpus *ESLO1* et 498 tours de parole du corpus *ESLO2* de la partie *entretiens* sont analysés. Ces tours de parole contiennent 168 paraphrases du corpus *ESLO1* et 186 paraphrases du corpus *ESLO2*. Les tours de parole étudiés proviennent de 59 et 37 entretiens des corpus *ESLO1* et *ESLO2*, respectivement. Ces données de référence sont utilisées pour entraîner le système automatique à la détection des segments en relation de paraphrase et pour évaluer les résultats.

5 Méthode

La figure 1 présente le schéma général de la méthode, composée de plusieurs étapes : le pré-traitement des données (section 5.1), la détection de tours de parole avec les reformulations paraphrastiques (section 5.2). Le traitement continue avec les tours de paroles qui contiennent les reformulations paraphrastiques. Nous effectuons alors la détection des segments en relation de paraphrase (section 5.3) et l'évaluation de résultats (section 5.4). Dans le travail proposé, l'accent principal est mis sur la détection des segments en relation de paraphrase et l'évaluation de ces résultats avec les données de référence.

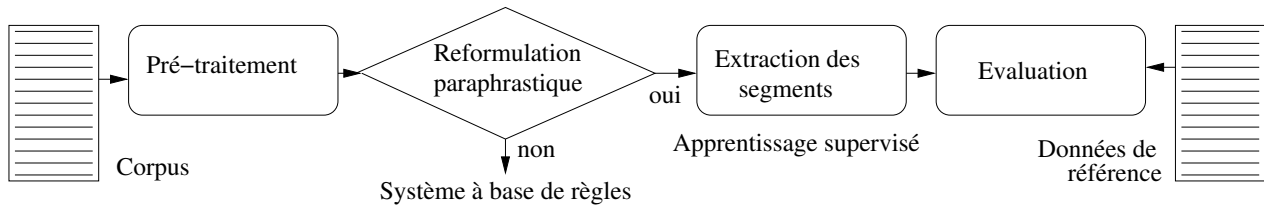


FIGURE 1 – Schéma général de la méthode.

5.1 Pré-traitement des données

Les fichiers transcrits d'ESLO respectent deux principes : l'adoption de l'orthographe standard et le non-recours à la ponctuation de l'écrit. La segmentation d'origine est faite soit sur une unité intuitive de type "groupe de souffle" repérée par le transcripteur humain, soit sur un *tour de parole*, défini uniquement par le changement de locuteurs. Nous avons utilisé les versions C de transcription, qui ont bénéficié de la correction orthographique supplémentaire par rapport aux versions A et B. Pour avoir des données comparables, nous avons sélectionné des entretiens dans les deux corpus. En l'absence de la ponctuation, nous avons reconstitué les tours de paroles en utilisant comme segmenteur :

- les tours de parole marqués dans la transcription par un changement de locuteur,
- et en traitant les chevauchements où deux locuteurs ou plus parlent en même temps.

Dans le cas des chevauchements, les segments correspondants sont associés aux énoncés de chacun des locuteurs impliqués et lorsqu'un locuteur continue de parler après un chevauchement, son tour de parole continue. Les tours de parole sont notre unité de travail. Lorsque les tours de parole se retrouvent sur plus d'un groupe de souffle, ces groupes sont séparés par les virgules dans nos sorties (exemple en (8)).

Les corpus avec les tours de parole reconstitués sont traités avec le chunker SEM (Tellier *et al.*, 2014) adapté à la langue orale. SEM détecte les chunks minimaux, comme présenté dans l'exemple (5) (même exemple qu'en (2)).

- (5) *(est/V)VN (-ce/CLS)NP (que/CS)CONJ (vous/CLS)NP (remarquez/V)VN (une/DET différence/NC sensible/ADJ)NP (entre/P vos/DET différents/ADJ clients/NC dans/P leur/DET façon/NC de/P choisir/VINF)PP (la/DET viande/NC)NP (dans/P ce/PRO)PP (qu'/PROREL ils/CLS)NP (achètent/V)VN (et/CC)CONJ (caetera/V)VN (./CLS)NP (indépendamment/V disons/VPP)VN (de/P leurs/NC)PP (oui/I)IntP (origines/NC)NP (de/P classe/NC ./ADJ)PP*

5.2 Distinction automatique entre les reformulations paraphrastiques et non paraphrastiques

Cette étape n'a pas été modifiée par rapport au travail précédent. Par soucis de clarté, nous en rappelons les principaux moments. L'objectif de cette étape est d'analyser les tours de parole qui contiennent les marqueurs de reformulation paraphrastique pour décider si, dans ces tours de parole, les marqueurs de reformulation paraphrastique introduisent la relation de reformulation paraphrastique ou non. Actuellement, cette étape est fondée sur des filtres communs à l'ensemble des marqueurs de reformulation paraphrastique :

1. Si le marqueur de reformulation paraphrastique est placé en début ou en fin de tours de parole, alors il est considéré que le contexte n'est pas suffisant et que ce tour de parole ne comporte pas de paraphrase ;
2. Si le marqueur de reformulation paraphrastique est entouré des marqueurs discursifs (comme ceux présentés dans la section 4.3), d'euh d'hésitation, d'interjections (*ben hm ouais*), d'amorces (*s-*), etc. répétés, il est considéré que le marqueur de reformulation paraphrastique fait partie des disfluences de l'oral (une accumulation d'éléments qui brisent le déroulement syntagmatique (Blanche-Benveniste *et al.*, 1991)) et n'introduit pas la paraphrase ;
3. Si le marqueur de reformulation paraphrastique apparaît dans un contexte lexical spécifique (emploi de *nous* devant *disons*), ou si le marqueur de reformulation paraphrastique apparaît dans des suites argumentatives (*e.g. par contre, mais, en revanche, au contraire*), ce tour de parole ne comporte pas de paraphrase ;
4. Si le marqueur de reformulation paraphrastique apparaît à l'intérieur d'une locution, comme *indépendamment de* ou *plus ou moins grossiers* (exemples (2) et (3)), alors il est considéré que ce contexte ne comporte pas de paraphrase. Ce test est effectué sur les sorties du chunker (exemple (5)). Pour vérifier que la locution existe dans la langue, nous interrogeons un moteur de recherche généraliste et analysons les fréquences attestées sur la Toile. Ces fréquences donnent une indication générale et relative sur les segments testés. Chaque segment est testé de trois manières (exemple (5)) : avec un *((caetera)VN (indépendamment)VN (de leurs)PP)*, deux *((et)CONJ (caetera)VN (indépendamment)VN (de leurs)PP (origines)NP)* ou trois chunks *(achètent)VN (caetera)VN (indépendamment)VN (de leurs)PP (origines)NP (de classe)PP* à droite et à gauche du marqueur de reformulation paraphrastique, excepté les disfluences et la ponctuation. La taille maximale du segment est empiriquement limitée à sept mots. La fréquence moyenne de ces segments doit être inférieure à des seuils donnés pour qu'il soit considéré que ce segment comporte une paraphrase. Dans le cas où les fréquences sont plus élevées que le seuil, ce test indique que la locution existe dans la langue et qu'il s'agit en effet de disfluence.

L'application et le test de ces filtres ont montré que la meilleure combinaison consiste à donner la priorité aux filtres 1, 3 et 4, et d'attribuer moins d'importance au filtre 2 car les disfluences peuvent apparaître autour des paraphrases, comme *disons* dans l'exemple (1). Dans ce cas, la précision atteint jusqu'à 66,4 %.

5.3 Extraction automatique de segments en relation de reformulation paraphrastique

Cette étape correspond à l'apport principal de notre travail. Les segments en relation de paraphrase sont détectés avec un système d'apprentissage supervisé. En disposant des données de référence et en travaillant sur le paraphrasage syntagmatique, nous exploitons les fonctionnalités des CRF telles qu'elles sont implémentées dans Wapiti (Lavergne *et al.*, 2010) pour créer notre système. Les données de référence sont divisées aléatoirement en deux ensembles : un ensemble d'entraînement et un ensemble de test, composés de 60 % et 40 % des tours de parole, respectivement.

5.3.1 Les catégories à détecter

L'objectif principal est de détecter deux segments en relation paraphrastique : le segment source, qui est repris ultérieurement dans le texte, et le segment cible, qui propose une nouvelle expression linguistique pour l'idée déjà exprimée par le segment source. Les catégories à détecter sont donc les suivantes :

1. **M** : marqueur de reformulation paraphrastique,
2. **SEG1** : segment source, qui apparaît avant le marqueur de reformulation paraphrastique,

3. **SEG2** : segment cible, qui apparaît après le marqueur de reformulation paraphrastique,
4. **O** : tout autre token dans les tours de parole (out).

<i>form</i>	<i>POS</i>	<i>chunkBI</i>	<i>chunk</i>	<i>heu</i>	<i>num</i>	<i>dmf</i>	<i>stem</i>	<i>MRP</i>	<i>ref.</i>
...									
la	DET	I-PP	PP	N	28	MIL	la	O	O
cuisson	NC	I-PP	PP	N	29	MIL	cuisson	O	O
rapide	ADJ	I-PP	PP	N	30	MIL	rapid	O	O
quoi	PROWH	I-PP	PP	EUH	31	MIL	quoi	O	O
des	DET	B-NP	NP	N	32	MIL	de	O	SEG1
morceaux	NC	I-NP	NP	N	33	MIL	morceau	O	SEG1
nobles	ADJ	I-NP	NP	N	34	MIL	nobl	O	SEG1
ce	PRO	I-NP	NP	N	35	MIL	ce	O	O
qu'	PROREL	B-NP	NP	N	36	MIL	qu'	O	O
ils	CLS	B-NP	NP	N	37	MIL	il	O	O
appellent	V	B-VN	VN	N	38	MIL	appellent	O	O
quoi	PROWH	B-NP	NP	EUH	39	MIL	quoi	O	O
c'	CLS	B-NP	NP	N	40	MIL	c'	M	M
est	V	B-VN	VN	N	41	MIL	est	M	M
à	P	B-PP	PP	N	42	MIL	à	M	M
dire	VINF	I-PP	PP	N	43	MIL	dir	M	M
les	DET	B-NP	NP	N	44	MIL	le	O	SEG2
rosbifs	ADJ	I-NP	NP	N	45	MIL	rosbif	O	SEG2
les	DET	I-NP	NP	N	46	MIL	le	O	SEG2
biftecks	NC	I-NP	NP	N	47	MIL	bifteck	O	SEG2
et	CC	B-CONJ	CONJ	N	48	MIL	et	O	SEG2
tout	PRO	B-NP	NP	N	49	MIL	tout	O	SEG2
ça	PRO	B-NP	NP	N	50	MIL	ça	O	SEG2
...									

TABLE 1 – Un extrait de données annotées, avec une reformulation paraphrastique.

5.3.2 Les unités et les descripteurs

Les unités traitées sont les tours de parole. Chaque mot des tours de parole est décrit avec un ensemble de descripteurs, comme exemplifié dans le tableau 1. Nous présentons les descripteurs et justifions leur choix :

- *form* : forme de chaque mot graphique tel qu'il apparaît dans le texte. La forme correspond à l'information de base directement disponible dans le texte ;
- *POS* : étiquette morpho-syntaxique calculée par SEM. L'étiquette morpho-syntaxique peut être indicative d'un lien syntaxique entre les segments à mettre en relation, et peut ainsi faciliter la détection de la relation de paraphrase ;
- *chunksBI* : chunk de SEM avec le marquage de début et de fin. Les chunks marquent la segmentation des tours de parole en segments syntaxiques. Cette information peut aider à délimiter les frontières des segments en relation de paraphrase ;
- *chunks* : chunk de SEM sans le marquage de début et de fin. Par rapport à *chunksBI*, il s'agit d'un descripteur simplifié car il ne s'agit plus de détecter le début et la suite des segments syntaxiques, mais juste de faire la distinction entre les différents segments syntaxiques. Nous pensons que ce descripteur peut être aussi intéressant pour la tâche visée ;
- *heu* : marquage des disfluences exprimées par des marqueurs spécifiques considérés dans notre travail (section 4.3). Nous pensons que les marqueurs de disfluences peuvent également aider à détecter les segments en relation de paraphrase car, dans ce cas, étant employés avec les marqueurs de reformulation, les marqueurs de disfluences peuvent souligner l'acte de reformulation ;
- *num* : numéro du mot depuis le début de chaque tour de parole. Ce descripteur permet surtout de marquer les mots apparaissant au début des tours de parole et donc de limiter potentiellement la taille des segments à détecter ;
- *début/milieu/fin* : position relative de chaque mot graphique où les positions de début et de fin correspondent à 20 % des tokens du début et de la fin des tours de parole, respectivement. Le reste des mots fait partie de la position du milieu. Ce descripteur est similaire au descripteur précédent *num* ;

- *stem* : mot racinisé avec le raciniseur `Snowball` (Porter, 1980) implémenté dans le module `Perl Lingua::Stem`. Nous pensons que les mots racinisés peuvent aider à établir un lien entre les segments en relation de paraphrase au cas où ces segments comportent des mots formellement similaires. Bien qu'étant assez brutal avec les mots, `Snowball` permet de les réduire à des chaînes de caractères plus facilement comparables ;
- *MRP* : marquage du marqueur de reformulation paraphrastique. Ce marqueur correspond également à l'information disponible directement dans le texte. De plus, comme c'est lui qui permet d'établir la relation de paraphrase, son rôle pour la détection des segments peut être important.

5.3.3 Les patrons pour la gestion des descripteurs et des contextes

Les patrons sous `Wapiti` indiquent quels descripteurs il faut exploiter, comment les combiner, dans quel contexte les étudier, etc. Nous appliquons trois séries de patrons pour tester plusieurs hypothèses :

1. utilisation de la forme seule : dans une fenêtre de 3 à 12 mots avant et après le token courant la forme est considérée, de même que la combinaison *forme/MRP*. Il s'agit des informations directement disponibles dans le texte ;
2. la taille de la fenêtre autour du mot courant a une influence : pour l'ensemble de descripteurs, nous faisons varier la largeur de la fenêtre allant de 3 à 12 mots avant et après un token donné ;
3. les combinaisons de différents descripteurs ont une influence sur les résultats : dans la fenêtre de 7 mots avant et après le token courant, nous faisons varier les combinaisons de descripteurs.

5.4 Évaluation

L'évaluation est effectuée par rapport aux données de référence. Nous calculons la macro-précision, le macro-rappel et la macro-F-mesure au niveau des catégories (Sebastiani, 2002) et les moyennes globales. La baseline correspond à l'exploitation de la forme seule dans une fenêtre de 7 mots avant et après le token courant, avec la combinaison *forme/MRP*. Il s'agit des informations disponibles dans le texte directement et de la largeur de fenêtre moyenne dans nos expériences.

6 Résultats et Discussion

6.1 Trois séries d'expériences

Dans l'ensemble de test, les catégories *SEG1* et *SEG2* représentent 8,9 % et 11,2 %, respectivement, au niveau des tokens.

Les tableaux 2 à 4 présentent les résultats obtenus pour les trois séries d'expériences. Pour chaque expérience, nous indiquons les valeurs de la précision, du rappel et de la F-mesure pour les catégories visées : *O*, *SEG1* et *SEG2*. De manière générale, le marqueur est la catégorie la plus facile à reconnaître (toujours proche de 1,0), tandis que les segments en relation de paraphrase restent plus difficiles à détecter. Nous calculons aussi la moyenne des performances.

Taille de la fenêtre	<i>O</i>			<i>SEG1</i>			<i>SEG2</i>		
	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>
<i>E1 (3 mots)</i>	0,81	0,97	0,88	0,61	0,13	0,21	0,50	0,13	0,20
<i>E2 (4 mots)</i>	0,81	0,95	0,88	0,51	0,18	0,27	0,39	0,13	0,20
<i>E3 (5 mots)</i>	0,81	0,97	0,88	0,57	0,13	0,21	0,45	0,12	0,19
<i>E4 (6 mots)</i>	0,81	0,97	0,88	0,59	0,11	0,19	0,46	0,12	0,19
<i>E5 (7 mots) - Baseline</i>	0,82	0,94	0,88	0,47	0,13	0,21	0,37	0,18	0,24
<i>E6 (8 mots)</i>	0,81	0,98	0,88	0,62	0,12	0,20	0,51	0,11	0,18
<i>E7 (9 mots)</i>	0,81	0,97	0,88	0,59	0,11	0,19	0,49	0,14	0,22
<i>E8 (10 mots)</i>	0,81	0,97	0,88	0,61	0,13	0,21	0,48	0,15	0,22
<i>E9 (11 mots)</i>	0,82	0,96	0,88	0,61	0,20	0,30	0,51	0,16	0,24
<i>E10 (12 mots)</i>	0,81	0,97	0,88	0,61	0,13	0,21	0,46	0,15	0,22

TABLE 2 – Utilisation de la forme seule, dans une fenêtre allant de 3 à 12 mots.

Le tableau 2 montre les résultats lorsque seules les formes et la combinaison *forme/MRP* sont utilisées. La différence entre les expériences concerne la largeur de la fenêtre qui varie de 3 à 12 mots à gauche et à droite du token donné. Cette configuration montre une moyenne autour de 0,60 : la précision moyenne est autour de 0,70 tandis que le rappel est autour de 0,58. Les segments en relation de paraphrase sont reconnus avec une précision assez élevée mais un rappel très bas (le plus souvent, moins de 0,20). La différence entre les expériences n'est pas très importante. La baseline est bien placée. Les meilleures moyennes de F-mesure sont observées avec les fenêtres de 11 et 4 mots.

Taille de la fenêtre	<i>O</i>			<i>SEG1</i>			<i>SEG2</i>		
	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>
<i>Baseline</i>	0,82	0,94	0,88	0,47	0,13	0,21	0,37	0,18	0,24
<i>E1 (3 mots)</i>	0,84	0,77	0,81	0,27	0,29	0,28	0,34	0,51	0,40
<i>E2 (4 mots)</i>	0,82	0,58	0,68	0,24	0,17	0,20	0,20	0,64	0,30
<i>E3 (5 mots)</i>	0,84	0,37	0,52	0,40	0,23	0,30	0,16	0,81	0,26
<i>E4 (6 mots)</i>	0,84	0,40	0,54	0,20	0,61	0,31	0,22	0,68	0,34
<i>E5 (7 mots)</i>	0,83	0,46	0,60	0,27	0,50	0,35	0,16	0,55	0,25
<i>E6 (8 mots)</i>	0,86	0,45	0,59	0,13	0,70	0,21	0,42	0,29	0,35
<i>E7 (9 mots)</i>	0,85	0,61	0,71	0,16	0,63	0,25	0,40	0,23	0,29
<i>E8 (10 mots)</i>	0,87	0,29	0,43	0,18	0,70	0,29	0,21	0,72	0,33
<i>E9 (11 mots)</i>	0,87	0,35	0,50	0,26	0,64	0,37	0,19	0,74	0,30
<i>E10 (12 mots)</i>	0,88	0,51	0,65	0,15	0,70	0,25	0,45	0,48	0,46

TABLE 3 – Utilisation de l'ensemble des descripteurs dans une fenêtre allant de 3 à 12 mots.

Le tableau 3 montre les résultats lorsque l'ensemble des descripteurs est utilisé dans une fenêtre variant de 3 à 12 mots. Par rapport au tableau 2, nous observons plus de différences dans les résultats. La reconnaissance des segments en relation de paraphrase est meilleure grâce à l'amélioration du rappel. Par contre, les performances de la catégorie *O* diminuent, ce qui mène aussi à la diminution de moyennes globales. Vu la différence entre les expériences, nous supposons que la portée optimale de chaque descripteur est différente et la largeur optimale de sa fenêtre doit être adaptée. Pour ceci, des tests complémentaires doivent être effectués.

Différentes combinaisons	<i>O</i>			<i>SEG1</i>			<i>SEG2</i>		
	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>	<i>P</i>	<i>R</i>	<i>F</i>
<i>Baseline</i>	0,82	0,94	0,88	0,47	0,13	0,21	0,37	0,18	0,24
<i>E1</i>	0,84	0,94	0,89	0,47	0,42	0,44	0,70	0,17	0,27
<i>E2 (E1 + 7 mots(chunk/MRP))</i>	0,83	0,36	0,50	0,12	0,59	0,20	0,27	0,52	0,36
<i>E3 (E2 + 7 mots(POS/MRP))</i>	0,85	0,19	0,32	0,36	0,43	0,39	0,15	0,93	0,26
<i>E4 (E3 + 7 mots(stem/MRP))</i>	0,87	0,36	0,51	0,20	0,65	0,31	0,21	0,69	0,32
<i>E5 (E4 + 7 mots(début))</i>	0,83	0,46	0,60	0,27	0,50	0,35	0,16	0,55	0,25
<i>E6 (E5 - 7 mots(POS))</i>	0,84	0,48	0,61	0,23	0,52	0,32	0,22	0,66	0,33
<i>E7 (E5 - 7 mots(chunk))</i>	0,84	0,26	0,40	0,20	0,60	0,30	0,18	0,78	0,30
<i>E8 (E5 - 7 mots(heu))</i>	0,85	0,39	0,53	0,24	0,45	0,32	0,20	0,81	0,32
<i>E9 (E5 - 7 mots(stem))</i>	0,85	0,32	0,47	0,31	0,28	0,29	0,15	0,82	0,26
<i>E10 (E5 - 7 mots(MRP))</i>	0,83	0,42	0,56	0,17	0,47	0,25	0,22	0,69	0,34

TABLE 4 – Différentes combinaisons de descripteurs dans une fenêtre de 7 mots. L'expérience *E1* correspond à la combinaison de descripteurs : dans une fenêtre de 7 mots (POS, chunk, heu, stem, MRP), dans une fenêtre de 1 mot (chunk/MRP, POS/MRP, stem/MRP). Les autres combinaisons dérivent de *E1*.

Le tableau 4 montre les résultats obtenus avec l'exploitation de différentes combinaisons de descripteurs dans une fenêtre de 7 mots. Ces différentes expériences dépassent la baseline. L'expérience *E1* correspond à la combinaison de descripteurs : dans une fenêtre de 7 mots (POS, chunk, heu, stem, MRP), dans une fenêtre de 1 mot (chunk/MRP, POS/MRP, stem/MRP). Les autres combinaisons dérivent de *E1*. Les expériences qui montrent la meilleure reconnaissance des segments en relation de paraphrase (F-mesure supérieure à 0,30) sont *E4*, *E6* et *E8*. Il est difficile de faire des généralisations, mais l'ajout de nouveaux descripteurs, de même que la suppression de POS et heu, semblent être bénéfiques. Le marqueur de reformulation paraphrastique est considéré comme le token pivot car, dans notre approche, c'est ce marqueur

qui a le potentiel d'établir la relation de paraphrase. Avec ces expériences, la précision et le rappel sont assez équivalents, ce qui permet d'obtenir de meilleures performances globales par rapport à la baseline. La reconnaissance des segments paraphrasés restent cependant difficile.

En résumé, nous pouvons faire plusieurs observations :

- les marqueurs de reformulation sont toujours bien reconnus,
- les positions O out sont aussi assez bien reconnues,
- la détection des segments cible et source restent difficile et montre des performances variables selon les expériences,
- parmi les meilleures expériences, nous avons la baseline (utilisation des formes dans une fenêtre de 7 mots et de la combinaison *forme/MRP*) et les expériences basées sur différentes combinaisons du tableau 4,
- parmi les meilleures fenêtres, nous avons 4, 7 et 11 mots.

6.2 Discussion

Concernant les 3 séries d'expériences proposées, nous pouvons remarquer que :

- l'exploitation de la forme seule est assez efficace du point de vue des valeurs moyennes. Cette série d'expériences est surtout intéressante car la précision est élevée par comparaison aux autres expériences ;
- les combinaisons ciblées de l'expérience présentée dans le tableau 4 montrent l'intérêt d'effectuer de telles combinaisons. Avec certains tests, nous obtenons des valeurs de précision et de rappel assez élevées. Nous devons cependant effectuer plus de tests pour optimiser les résultats ;
- le rappel reste le point faible des expériences actuelles. L'amélioration principale de la méthode concerne cet aspect.

Dans d'autres expériences, non présentées ici, nous avons traité séparément les corpus *ESLO1* et *ESLO2*, et les annotations de deux annotateurs *A1* et *A2*. Il s'agit des données de référence sans le consensus. Les ensembles d'entraînement et de test étaient également indépendants. Les modèles créés sur chacun des ensembles de données (*ESLO1/A1*, *ESLO1/A2*, *ESLO2/A1*, *ESLO2/A2*) ont été appliqués sur d'autres ensembles. L'objectif était de voir quelle est la portabilité de ces modèles. Ces expériences montrent que :

- il est plus facile de détecter les segments paraphrasés dans le corpus *ESLO1*, quel que soit le modèle appliqué (corpus ou annotateur). En effet, le corpus *ESLO2* comporte des tours de parole beaucoup plus longs, ce qui rend la détection des segments paraphrasés plus compliquée dans ce corpus. Cette situation peut s'expliquer par la méthodologie mise en œuvre dans la constitution du corpus *ESLO2* où l'entretien mené est beaucoup moins formel et l'intervieweur laisse plus de liberté à son interlocuteur ;
- les modèles liés aux annotateurs montrent aussi des performances variables, mais qui se ressentent moins par comparaison avec l'influence des corpus ;
- assez systématiquement, la détection des segments cible *SEG2* est plus aisée que la détection des segments source *SEG1*.

En relation avec cette observation, une hypothèse est que dans le flux des énoncés, après l'apparition du marqueur de reformulation, il est plus évident de se rendre compte que le *SEG2* doit être retrouvé.

De manière générale, le travail avec les données de référence consensuelles entre les annotateurs et la fusion des annotations provenant des corpus *ESLO1* et *ESLO2* montrent un effet légèrement bénéfique sur les résultats. Par ailleurs, nous avons aussi observé que la prise en compte de la catégorie *M*, même si sa détection est évidente, permet d'améliorer la détection des segments en relation de paraphrase de quelques points.

6.3 Analyse des erreurs

L'analyse des sorties indique que très souvent les segments sont détectés, mais avec des frontières différentes que celles définies par les données de référence. En (6) et (7), nous présentons deux exemples différents : en *A* se trouvent les tours de parole tels qu'annotés dans les données de référence, en *B* se trouvent les résultats de la détection automatique. Les segments en relation de paraphrase sont en bleu et surlignés. Nous pouvons voir que les segments détectés automatiquement sont plus larges. Les segments proposés par le modèle sont aussi acceptables et des hésitations quant à la taille de ces segments étaient aussi présentes lors de l'annotation manuelle. Dans de nombreux cas, les sorties sont intéressantes et utilisables au moins comme une base d'annotation. Cependant, comme l'évaluation des résultats de CRF dépend de la détection des frontières, cela diminue les chiffres de l'évaluation automatique des résultats. Notons aussi que la taille des segments détectés augmente avec l'agrandissement de la fenêtre (la série d'expériences du tableau 3).

- (6) A. et cinq kilomètres c'est-à-dire j'avais quatre kilomètres à faire quatre et quatre huit je faisais huit kilomètres tous les jours et à pieds ah oui [ESLO1_ENT_011_C]
 B. et cinq kilomètres c'est-à-dire j'avais quatre kilomètres à faire quatre et quatre huit je faisais huit kilomètres tous les jours et à pieds ah oui [ESLO1_ENT_011_C]
- (7) A. et et vous par exemple approximativement vous combien de fois euh quelle est la fréquence avec laquelle vous regardez le dictionnaire c'est à dire une fois par mois une fois par an une fois par euh , oh [ESLO1_ENT_047_C]
 B. et et vous par exemple approximativement vous combien de fois euh quelle est la fréquence avec laquelle vous regardez le dictionnaire c'est à dire une fois par mois une fois par an une fois par euh , oh [ESLO1_ENT_047_C]

Dans notre modèle, nous cherchons à établir une relation de paraphrase dans la séquence *SEG1 MRP SEG2*, où le marqueur de reformulation paraphrastique fait le lien entre les deux segments. Pourtant, la détection des segments *SEG1* et *SEG2* peut être dissociée et seulement un des deux segments trouvé, comme dans les exemples (8) et (9). Il semblerait donc que le marqueur de reformulation paraphrastique n'est pas le seul élément qui marque la paraphrase. D'autres descripteurs utilisés peuvent donc aussi participer dans l'établissement de cette relation. Par ailleurs, comme noté plus haut, le segment cible est détecté plus facilement que le segment source.

- (8) A. n'est-ce pas , alors quand toutes les pièces sont coupées on le on le met en plomb c'est-à-dire qu'on prend un tout petit plomb et on met on rassemble toutes les pièces [ESLO1_ENT_002_C]
 B. n'est-ce pas , alors quand toutes les pièces sont coupées on le on le met en plomb c'est-à-dire qu'on prend un tout petit plomb et on met on rassemble toutes les pièces [ESLO1_ENT_002_C]
- (9) A. oui enfin par industriel je veux dire euh j'ai le côté commercial [ESLO1_ENT_002_C]
 B. oui enfin par industriel je veux dire euh j'ai le côté commercial [ESLO1_ENT_002_C]

Un autre type d'erreurs peut être observé avec les marqueurs, qui sont fusionnés avec les segments paraphrasés. Notons que cette erreur apparaît avec le marqueur *disons*, qui est le moins grammaticalisé dans le rôle de marqueur de reformulation paraphrastique. Notons aussi que cette fusion s'accroît avec l'agrandissement de la fenêtre dans laquelle les tokens et les descripteurs sont analysés (la série d'expériences du tableau 3).

Finalement, les énoncés avec plusieurs reformulations paraphrastiques (comme en (10)) sont mal gérées actuellement.

- (10) s- y en a sûrement mais si vous voulez euh je s- sincèrement je crois que *<P1>il faut pas mettre les inconvénients en exergue</P1>* enfin *<MRP>je veux dire</MRP>* *<P2>on raisonne pas comme ça</P2>* *<MRP>je veux dire</MRP>* euh *<P3>on est heureux là où on est</P3>* et je pense qu'on serait oui on essaierait de trouver euh des moyens d'être heureux donc euh y a sûrement des inconvénients mais sincèrement on les cherche pas hein et donc on les trouve pas

7 Conclusion et Travaux futurs

Nous avons proposé un travail sur la détection automatique de segments en relation de paraphrase. Ce travail a deux points originaux principaux :

- la paraphrase étudiée est formée de manière syntagmatique dans une structure particulière de type : *SEG1 MRP SEG2*,
- le travail est effectué sur un corpus de l'oral, où les traces de reformulations sont fréquentes et peuvent être observées grâce à l'emploi de marqueurs spécifiques.

Nous avons mis en place un système d'apprentissage qui repose sur les CRF et une série de descripteurs, et de leurs combinaisons. Les descripteurs sont étudiés dans une fenêtre plus au moins grande. Parmi les meilleures expériences, nous trouvons la baseline, qui consiste en l'utilisation des formes dans une fenêtre de 7 mots et de la combinaison *forme/MRP*, et les expériences basées sur différentes combinaisons (e.g. *chunk/MRP*, *pos/MRP*, *stem/MRP*) dans une fenêtre de 7 mots. Les meilleurs résultats atteignent une moyenne allant jusqu'à 0,65 de F-mesure, 0,75 de précision et 0,63 de rappel. Nous observons qu'il est aisé de reconnaître les marqueurs de reformulation paraphrastique. Par contre, les segments en relation de paraphrase restent plus difficiles à détecter. Il s'agit surtout de la difficulté à détecter correctement leurs frontières. Le défi principal consiste en amélioration du rappel.

Nous avons plusieurs perspectives à ce travail. Tout d'abord, le travail actuel peut être amélioré de plusieurs points de vue. Nous pouvons ainsi tester d'autres classifieurs, comme les réseaux à longues mémoires court-terme (LSTM) (Schmidhuber, 1997) indiqués par un des relecteurs, et améliorer le traitement des énoncés avec plusieurs reformulations paraphrastiques. Nous pouvons aussi tester d'autres descripteurs et leurs combinaisons pour améliorer la détection des segments en relation de paraphrase. Cela concerne surtout le rappel, mais les performances globales devraient être améliorées également. Une fois stabilisé, le modèle d'apprentissage peut être utilisé pour pré-annoter d'autres tours de paroles et préparer ainsi les données à valider et corriger par les annotateurs humains. Nous pensons que cela peut faciliter l'annotation humaine et la création des données de référence plus importantes. Nous voulons aussi tester les modèles générés avec les trois marqueurs traités ici sur les tours de parole qui contiennent d'autres marqueurs (*e.g. ça veut dire, j'allais dire, notamment, autrement dit, en d'autres termes, en d'autres mots*). De la même manière, les modèles générés peuvent être testés sur d'autres corpus. Ainsi, pouvons-nous étudier si les reformulations paraphrastiques introduites par différents marqueurs montrent des régularités communes. Lorsque les données de référence le permettent, nous prévoyons de tester les marqueurs de reformulation séparément. Cela permettra d'étudier la généralité des modèles et de la reformulation paraphrastique d'une autre manière.

D'autres perspectives consistent à appréhender ces données d'autres points de vue. Par exemple, nous pouvons prendre en compte et analyser conjointement les éléments prosodiques et acoustiques associés aux différents tours de parole. Notre hypothèse est que les tours de parole avec les reformulations paraphrastiques montrent des différences par rapport aux tours de parole avec les marqueurs étudiés mais n'introduisant pas les reformulations paraphrastiques. De cette manière, le filtrage entre ces deux types de tours de parole peut reposer sur ce critère aussi. L'étape de distinction entre les tours de parole paraphrastiques et non devrait évoluer et reposer sur des mécanismes plus robustes qu'un système à base de règles.

Nous avons aussi commencé à explorer les reformulations paraphrastiques, introduites par les mêmes marqueurs ou bien par des marqueurs différents, dans les corpus écrits (les forums de discussion et la presse journalistique). Une comparaison entre ces types de corpus et les modes de reformulation correspond à une perspective intéressante. Par ailleurs, dans les corpus oraux d'autres paraphrases peuvent être introduites par d'autres moyens que les marqueurs de reformulation paraphrastique.

Parmi d'autres perspectives, se trouvent par exemple :

- l'exploitation de *Distagger* (Constant & Dister, 2010) pour améliorer la détection des disfluences et des répétitions,
- la création des annotations consensuelles concernant les autres paramètres (relation lexicale, morphologique, syntaxique et pragmatique),
- l'exploitation des propriétés inhérentes des segments annotés (et détectés automatiquement) pour induire ces autres paramètres. Par exemple, en (11), le segment source est beaucoup plus long que le segment cible, ce qui semble être typique de la relation pragmatique *résultat*. Tandis que dans l'exemple (12), la situation est contraire : le segment cible est plus développé que le segment source. Nous pensons que cette situation peut être typique de relation pragmatique comme *définition* ou *explication*.
- le traitement des relations de paraphrase entre différents tours de parole, alors qu'actuellement nous le faisons au sein d'un même tour de parole uniquement,
- l'étude de l'emploi des reformulations paraphrastiques en croisant les annotations avec les critères sociologiques des locuteurs.

- (11) *voilà <P1>le côté très bétonné voilà c'est pas ils ont pas développé les les logements étudiants suffisamment ils ont pas développé l'off- l'offre culturelle euh en même temps</P1> donc enfin <MRP>je veux dire</MRP> voilà <P2 rel-pragm="res"> c'est mort</P2> [ESLO2_ENT_1012_C]*
- (12) *euh <VP1>démocratiser l'enseignement</VP1> <MRP>c'est-à-dire</MRP> <VP2 rel-lex="syn(démocratiser/ permettre à tout le monde) syn(enseignement/faculté)" modif-lex="ajout(rentre à)" rel-pragm="explic"> permettre à tout le monde de rentrer en faculté</VP2> [ESLO1_ENT_121_C]*

Finalement, nous voudrions exploiter les données acquises dans notre travail pour les applications TAL, comme la recherche d'information ou l'inférence textuelle. Lorsque les entretiens sont effectués avec des spécialistes de domaines (*e.g. boulangers, métiers médicaux, peintres, imprimeurs, artistes, commerçants*), les personnes peuvent proposer les paraphrases ou définitions pour les termes techniques, comme dans les exemples en (13). De telles connaissances peuvent être exploitées pour effectuer la simplification lexicale des textes (Carroll *et al.*, 1998; Candido, Jr. *et al.*, 2009).

- (13) {on le met en plomb; on prend un tout petit plomb et on met on rassemble toutes les pièces}

{le ballottage; il reste deux personnes en présence}
 {l'appareil digestif; foie vésicule pancréas et cætera}
 {la médecine interne; la médecine spécialisée mais de toutes les maladies générales}
 {le français; d'employer des mots comme snack bar}

Remerciements

Nous remercions les relecteurs pour leurs remarques, qui nous ont permis de prendre plus de recul, d'améliorer la qualité du papier et de prévoir d'autres pistes dans les futurs travaux.

Références

- ANDROUTSOPOULOS I. & MALAKASIOTIS P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, **38**, 135–187.
- BAKER M. (1992). In *Other Words : A Coursebook on Translation*. London, UK : Routledge.
- BANNARD C. & CALLISON-BURCH C. (2005). Paraphrasing with bilingual parallel corpora. In *ACL*, p. 597–604.
- BARZILAY R. & MCKEOWN L. (2001). Extracting paraphrases from a parallel corpus. In *ACL*, p. 50–57.
- BEECHING K. (2007). La co-variation des marqueurs discursifs bon, c'est-à-dire, enfin, hein, quand même, quoi et si vous voulez : une question d'identité ? *Langue française*, **154**(2), 78–93.
- BHAGAT R. & HOVY E. (2013). What is a paraphrase ? *Computational Linguistics*, **39**(3), 463–472.
- BLANCHE-BENVENISTE C., BILGER M., ROUGET C. & VAN DEN EYNDE K. (1991). *Le français parlé. Études grammaticales*. Paris : CNRS Éditions.
- BOUAMOR H. (2012). *Étude de la paraphrase sous-phrastique en traitement automatique des langues*. Thèse de doctorat, Université Paris Sud, Paris.
- BOUAMOR H., MAX A. & VILNAT A. (2012). Étude bilingue de l'acquisition et de la validation automatiques de paraphrases sous-phrastiques. *TAL*, **53**(1), 11–37.
- CALLISON-BURCH C., COHN T. & LAPATA M. (2008). Parametric : An automatic evaluation metric for paraphrasing. In *COLING*, p. 97–104.
- CANDIDO, JR. A., MAZIERO E., GASPERIN C., PARDO T. A. S., SPECIA L. & ALUISIO S. M. (2009). Supporting the adaptation of texts for poor literacy readers : a text simplification editor for Brazilian Portuguese. In *EdAppsNLP '09 Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 34–42.
- CARROLL J., MINNEN G., CANNING Y., DEVLIN S. & TAIT J. (1998). Practical simplification of english newspaper text to assist aphasic readers. In *AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, p. 7–10.
- CONSTANT M. & DISTER A. (2010). *Automatic detection of disfluencies in speech transcriptions*, In C. S. PUBLISHING, Ed., *Spoken Communication*, p. 259–272.
- DAGAN I., ROTH D., SAMMONS M. & ZANZOTTO F. (2013). *Recognizing Textual Entailment*. Milton Keynes, UK : Morgan & Claypool Publishers.
- ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER I. (2012). Un grand corpus oral disponible : le corpus d'Orléans 1968-2012. *Traitement Automatique de Langues*, **52**(3), 17–46.
- ESHKOL-TARAVELLA I. & GRABAR N. (2014). Repérage et analyse de la reformulation paraphrastique dans les corpus oraux. In *TALN 2014*.
- FLOTTUM K. (1995). *Dire et redire. La reformulation introduite par "c'est-à-dire"*. Thèse de doctorat, Hogskolen i Stavanger, Stavanger.
- FUCHS C. (1982). *La paraphrase*. Paris : PUF.
- FUJITA A. (2010). Typology of paraphrases and approaches to compute them. In *CBA to Paraphrasing & Nominalization*, Barcelona, Spain. Invited talk.

- GULICH E. & KOTSCHI T. (1983). Les marqueurs de la reformulation paraphrastique. *Cahiers de linguistique française*, **5**, 305–351.
- HATIM B. & MASON I. (1990). *Discourse and the Translator*. London, UK : Longman.
- HÖLKER K. (1988). *Zur Analyse von Markern*. Stuttgart : Franz Steiner.
- HWANG Y. (1993). Eh bien, alors, enfin et disons en français parlé contemporain. *L'Information Grammaticale*, **57**, 46–48.
- IBRAHIM A., KATZ B. & LIN J. (2003). Extracting structural paraphrases from aligned monolingual corpora. In *International Workshop on Paraphrasing*, p. 57–64.
- KOK S. & BROCKETT C. (2010). Hitting the right paraphrases in good time. In *NAACL*, p. 145–153.
- LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *ACL*, p. 504–513.
- LIN D. & PANTEL L. (2001). Dirt - discovery of inference rules from text. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, p. 323–328.
- MADNANI N. & DORR B. J. (2010). Generating phrasal and sentential paraphrases : A survey of data-driven methods. *Computational Linguistics*, **36**, 341–387.
- MALAKASIOTIS P. & ANDROUTSOPOULOS I. (2007). Learning textual entailment using SVMs and string similarity measures. In *ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, p. 42–47.
- MELČUK I. (1988). Paraphrase et lexique dans la théorie linguistique sens-texte. In *Lexique et paraphrase*. *Lexique*, **6**, 13–54.
- MILICEVIC J. (2007). *La paraphrase : Modélisation de la paraphrase langagière*. Peter Lang.
- OCH F. & NEY H. (2000). Improved statistical alignment models. In *ACL*, p. 440–447.
- PASÇA M. & DIENES P. (2005). Aligning needles in a haystack : Paraphrase acquisition across the Web. In *IJCNLP*, p. 119–130.
- PETIT M. (2009). *Discrimination prosodique et représentation du lexique : application aux emplois des connecteurs discursifs*. Thèse de doctorat, Université d'Orléans, Orléans.
- PORTER M. (1980). An algorithm for suffix stripping. *Program*, **14**(3), 130–137.
- QUIRK C., BROCKETT C. & DOLAN W. (2004). Monolingual machine translation for paraphrase generation. In *EMNLP*, p. 142–149.
- ROSSARI C. (1993). *Les opérations de reformulation. Analyse du processus et des marques dans une perspective contrastive français-italien*, In P. LANG, Ed., *Sciences pour la communication*.
- SAUNIER E. (2012). Disons : un impératif de dire ? Remarques sur les propriétés du marqueur et son comportement dans les reformulations. *L'Information Grammaticale*, **132**, 25–34.
- SCARPA F. (2010). *La Traduction spécialisée. Une approche professionnelle à l'enseignement de la traduction*. Ottawa, Canada : University of Ottawa Press. Language Arts & Disciplines.
- SCHMIDHUBER S. H. J. (1997). Long short-term memory. *Neural computation*, **7**(8), 1735–1780.
- SEBASTIANI F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1), 1–47.
- SEKINE S. (2005). Automatic paraphrase discovery based on context and keywords between NE pairs. In *International Workshop on Paraphrasing*, p. 80–87.
- SHEN S., RADEV D., PATEL A. & ERKAN G. (2006). Adding syntax to dynamic programming for aligning comparable texts for the generation of paraphrases. In *ACL-COLING*, p. 747–754.
- SHINYAMA Y., SEKINE S., SUDO K. & GRISHMAN R. (2002). Automatic paraphrase acquisition from news articles. In *Proceedings of HLT*, p. 313–318.
- TELLIER I., ESHKOL I., DUPONT Y. & WANG I. (2014). Peut-on bien chunker avec de mauvaises étiquettes pos ? In *TALN 2014*.
- TESTON-BONNARD S. (2008). Je veux dire est-il toujours une marque de reformulation ? In M. L. BOT, M. SCHUWER & E. RICHARD, Eds., *Rivages linguistiques. La Reformulation. Marqueurs linguistiques. Stratégies énonciatives*, p. 51–69. Rennes : PUR.
- VILA M., ANTÒNIA MART M. & RODRÍGUEZ H. (2011). Paraphrase concept and typology. a linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, **46**, 83–90.
- VILA M., RODRÍGUEZ H. & MARTÍ M. (2014). Relational paraphrase acquisition from wikipedia : The WRPA method and corpus. *Natural Language Engineering*, p. 1–35.

Une méthode discriminant formation simple pour la traduction automatique avec Grands Caractéristiques

Tian Xia Shaodan Zhai Zhongliang Li Shaojun Wang
Wright State University
3640 Colonel Glenn Hwy, Dayton, OH 45435
{Xia.7, Zhai.6, Li.141, Shaojun.Wang}@wright.edu

Résumé. Marge infusé algorithmes détendus (MIRAS) dominant modèle de tuning dans la traduction automatique statistique dans le cas des grandes caractéristiques de l'échelle, mais ils sont également célèbres pour la complexité de mise en œuvre. Nous introduisons une nouvelle méthode, qui concerne une liste des N meilleures comme une permutation et minimise la perte Plackett-Luce de permutations rez-de-vérité. Des expériences avec des caractéristiques à grande échelle démontrent que, la nouvelle méthode est plus robuste que MERT ; si ce est seulement à rattacher avec Miras, il a un avantage comparativement, plus facile à mettre en œuvre.

Abstract.

A Simple Discriminative Training Method for Machine Translation with Large-Scale Features

The margin infused relaxed algorithm (MIRAS) dominates model tuning in statistical machine translation in the case of large scale features, but also they are famous for the complexity in implementation. We introduce a new method, which regards an N-best list as a permutation and minimizes the Plackett-Luce loss of ground-truth permutations. Experiments with large-scale features demonstrate that, the new method is more robust than MERT ; though it is only matchable with MIRAS, it has a comparatively advantage, easier to implement.

Mots-clés : Traduction automatique ; ajustement du modèle ; caractéristiques à grande échelle.

Keywords: machine translation ; model tuning ; large-scale features .

1 Introduction

Since Och (Och, 2003) proposed minimum error rate training (MERT) to exactly optimize objective evaluation measures, MERT has become a standard model tuning technique in statistical machine translation (SMT). Though MERT performs better by improving its searching algorithm (Macherey *et al.*, 2008; Cer *et al.*, 2008; Galley & Quirk, 2011; Moore & Quirk, 2008; Kumar *et al.*, 2009), it does not work reasonably when there are lots of features¹. As a result, the margin infused relaxed algorithm (MIRA) dominates in this case (McDonald *et al.*, 2005; Watanabe *et al.*, 2007; Chiang *et al.*, 2008; Tan *et al.*, 2013; Cherry & Foster, 2012).

In SMT, MIRAS consider margin losses related to sentence-level BLEUs. However, since the BLEU is not decomposable into each sentence, these MIRA algorithms use some heuristics to compute the exact losses, e.g., pseudo-document (Chiang *et al.*, 2008), and document-level loss (Tan *et al.*, 2013).

Recently, another successful work in large-scale feature tuning include force decoding based (Yu *et al.*, 2013), classification based (Hopkins & May, 2011).

We aim to provide a simpler tuning method for large-scale features than MIRAS. Our motivation derives from an observation on MERT. As MERT considers the quality of only top1 hypothesis set, there might have more-than-one set of parameters, which have similar top1 performances in tuning, but have very different topN hypotheses. Empirically, we expect an ideal model to benefit the total N-best list. That is, better hypotheses should be assigned with higher ranks, and this might decrease the error risk of top1 result on unseen data.

Plackett (Plackett, 1975) offered an easy-to-understand theory of modeling a permutation. An N-best list is assumedly generated by sampling without replacement. The *i*th hypothesis to sample relies on those ranked after it, instead of on the

1. The regularized MERT seems promising from Galley *et al.* (Galley *et al.*, 2013) at the cost of model complexity.

whole list. This model also supports a partial permutation which accounts for top k positions in a list, regardless of the remaining. When taking k as 1, this model reduces to a standard conditional probabilistic training, whose dual problem is actual the maximum entropy approach (Och & Ney, 2002; Berger *et al.*, 1996). Although Och (Och, 2003) substituted direct error optimization for a maximum entropy based training, probabilistic models correlate with BLEU well when features are rich enough. The similar claim also appears in (Zhu & Hastie, 2001). This also make the new method be applicable in large-scale features.

2 Plackett-Luce Model

Plackett-Luce was firstly proposed to predict ranks of horses in gambling (Plackett, 1975). Let $\mathbf{r} = (r_1, r_2 \dots r_N)$ be N horses with a probability distribution \mathcal{P} on their abilities to win a game, and a rank $\pi = (\pi(1), \pi(2) \dots \pi(|\pi|))$ of horses can be understood as a generative procedure, where $\pi(j)$ denotes the index of the horse in the j th position.

In the 1st position, there are N horses as candidates, each of which r_j has a probability $p(r_j)$ to be selected. Regarding the rank π , the probability of generating the champion is $p(r_{\pi(1)})$. Then the horse $r_{\pi(1)}$ is removed from the candidate pool.

In the 2nd position, there are only $N - 1$ horses, and their probabilities to be selected become $p(r_j)/Z_2$, where $Z_2 = 1 - p(r_{\pi(1)})$ is the normalization. Then the runner-up in the rank π , the $\pi(2)$ th horse, is chosen at the probability $p(r_{\pi(2)})/Z_2$. We use a consistent terminology Z_1 in selecting the champion, though Z_1 equals 1 trivially.

This procedure iterates to the last rank in π . The key idea for the Plackett-Luce model is the choice in the i th position in a rank π only depends on the candidates not chosen at previous stages. The probability of generating a rank π is given as follows

$$p(\pi) = \prod_{j=1}^{|\pi|} \frac{p(r_{\pi(j)})}{Z_j} \quad (1)$$

where $Z_j = 1 - \sum_{t=1}^{j-1} p(r_{\pi(t)})$.

We offer a toy example (Table 1) to demonstrate this procedure.

\mathbf{r}	r_1	r_2	r_3
π	2	3	1
Z	1	$1-p(r_2)$	$1-(p(r_2) + p(r_3))$
$p(\pi)$	$\frac{p(r_2)}{Z_1}$	$\frac{p(r_3)}{Z_2}$	$\frac{p(r_1)}{Z_3}$

TABLE 1 – The probability of the rank $\pi = (2, 3, 1)$ is $p(r_2) \cdot p(r_3)/(1 - p(r_2))$ in a simplified form, as $\frac{p(r_2)}{Z_1} = p(r_2)$ and $\frac{p(r_1)}{Z_3} = 1$.

Theorem 1 *The permutation probabilities $p(\pi)$ form a probability distribution over a set of permutations Ω_π . For example, for each $\pi \in \Omega_\pi$, we have $p(\pi) > 0$, and $\sum_{\pi \in \Omega_\pi} p(\pi) = 1$.*

We have to note that, Ω_π is not necessarily required to be completely ranked permutations in theory and in practice, since gamblers might be interested in only the champion and runner-up, and thus $|\pi| \leq N$. In experiments, we would examine the effects on different length of permutations, systems being termed $PL(|\pi|)$.

Theorem 2 *Given any two permutations π and π' , and they are different only in two positions p and q , $p < q$, with $\pi(p) = \pi'(q)$ and $\pi(q) = \pi'(p)$. If $p(\pi(p)) > p(\pi(q))$, then $p(\pi) > p(\pi')$.*

In other words, exchanging two positions in a permutation where the horse more likely to win is not ranked before the other would lead to an increase of the permutation probability.

This suggests the ground-truth permutation, ranked decreasingly by their probabilities, owns the maximum permutation probability on a given distribution. In SMT, we are motivated to optimize parameters to maximize the likelihood of ground-truth permutation of an N-best hypotheses.

Due to the limitation of space, see (Plackett, 1975; Cao *et al.*, 2007) for the proofs of the theorems.

3 Plackett-Luce Model in Statistical Machine Translation

In SMT, let $\mathbf{f} = (f_1, f_2 \dots)$ denote source sentences, and $\mathbf{e} = (\{e_{1,1}, \dots\}, \{e_{2,1}, \dots\} \dots)$ denote target hypotheses. A set of features are defined on both source and target side. We refer to $h(e_{i,*})$ as a feature vector of a hypothesis from the i th source sentence, and its score from a ranking function is defined as the inner product $h(e_{i,*})^T \mathbf{w}$ of the weight vector \mathbf{w} and the feature vector.

We first follow the popular exponential style to define a parameterized probability distribution over a list of hypotheses.

$$p(e_{i,j}) = \frac{\exp\{h(e_{i,j})^T \mathbf{w}\}}{\sum_k \exp\{h(e_{i,k})^T \mathbf{w}\}} \quad (2)$$

The ground-truth permutation of an n best list is simply obtained after ranking by their sentence-level BLEUs. Here we only concentrate on their relative ranks which are straightforward to compute in practice, e.g. add 1 smoothing. Let π_i^* be the ground-truth permutation of hypotheses from the i th source sentences, and our optimization objective is maximizing the log-likelihood of the ground-truth permutations and penalized using a zero-mean and unit-variance Gaussian prior. This results in the following objective and gradient :

$$\mathcal{L} = \log\{\prod_i p(\pi_i^*, \mathbf{w})\} - \frac{1}{2} \mathbf{w}^T \mathbf{w} \quad (3)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \sum_i \sum_j \{h(e_{i,\pi_i^*(j)}) - \sum_{t=j} (h(e_{i,\pi_i^*(t)}) \cdot \frac{p(e_{i,\pi_i^*(t)})}{Z_{i,j}})\} - \mathbf{w} \quad (4)$$

where $Z_{i,j}$ is defined as the Z_j in Formula (1) of the i th source sentence.

The log-likelihood function is smooth, differentiable, and concave with the weight vector \mathbf{w} , and its local maximal solution is also a global maximum. Iteratively selecting one parameter in α for tuning in a line search style (or MERT style) could also converge into the global maximum (Bertsekas, 1999). In practice, we use the faster limited-memory BFGS (L-BFGS) algorithm (Byrd *et al.*, 1995).

N-best Hypotheses Resample

The log-likelihood of a Plackett-Luce model is not a strict upper bound of the BLEU score, however, it correlates with BLEU well in the case of rich features. The concept of “rich” is actually qualitative, and obscure to define in different applications. We empirically provide a formula to measure the richness in the scenario of machine translation.

$$r = \frac{\text{the size of features}}{\text{the average size of N-best lists}} \quad (5)$$

The greater, the richer. In practice, we find a rough threshold of r is 5.

In engineering, the size of an N-best list with unique hypotheses is usually less than several thousands. This suggests that, if features are up to thousands or more, the Plackett-Luce model is quite suitable here. Otherwise, we could reduce the size of N-best lists by sampling to make r beyond the threshold.

There may be other efficient sampling methods, and here we adopt a simple one. If we want to m samples from a list of hypotheses \mathbf{e} , first, the $\frac{m}{3}$ best hypotheses and the $\frac{m}{3}$ worst hypotheses are taken by their sentence-level BLEUs. Second, we sample the remaining hypotheses on distribution $p(e_i) \propto \exp(h(e_i)^T \mathbf{w})$, where \mathbf{w} is an initial weight from last iteration.

4 Evaluation

	MT02(dev)	MT04(%)	MT05(%)
MERT	34.61	31.76	28.85
MIRA	35.31	32.25	29.37
PL(1)	34.20	31.70	28.90
PL(2)	34.31	31.83	29.10
PL(3)	34.39	32.05	29.20
PL(4)	34.40	32.13	29.46+
PL(5)	34.46	32.19+	29.42+
PL(6)	34.37	32.16	29.30
PL(7)	34.39	32.20+	29.32
PL(8)	34.70	32.19+	29.10
PL(9)	34.30	32.07	29.22
PL(10)	34.30	32.14	29.19

TABLE 2 – $PL(k)$: Plackett-Luce model optimizing the ground-truth permutation with length k . The significant symbols (+ at 0.05 level) are compared with MERT.

We compare our method with MERT and MIRA² in two tasks, iterative training, and N-best list reranking. We do not list PRO (Hopkins & May, 2011) as our baseline, as Cherry et al. (Cherry & Foster, 2012) have compared PRO with MIRA and MERT massively.

In the first task, we align the FBIS data (about 230K sentence pairs) with GIZA++, and train a 4-gram language model on the Xinhua portion of Gigaword corpus. A hierarchical phrase-based (HPB) model (Chiang, 2007) is tuned on NIST MT 2002, and tested on MT 2004 and 2005. All features are eight basic ones (Chiang, 2007) and extra 220 group features. We design such feature templates to group grammar rules by the length of source side and target side, ($feat\text{-}type, a \leq src\text{-}side \leq b, c \leq tgt\text{-}side \leq d$), where the $feat\text{-}type$ denotes any of the relative frequency, reversed relative frequency, lexical probability and reversed lexical probability, and $[a, b]$, $[c, d]$ enumerate all possible subranges of $[1, 10]$, as the maximum length on both sides of a hierarchical grammar is limited to 10. There are 4×55 extra group features.

In the second task, we rerank an N-best list from a HPB system with 7491 features from a third party. The system uses six million parallel sentence pairs available to the DARPA BOLT Chinese-English task. This system includes 51 dense features (translation probabilities, provenance features, etc.) and up to 7440 sparse features (mostly lexical and fertility-based). The language model is a 6-gram model trained on a 10 billion words, including the English side of our parallel corpora plus other corpora such as Gigaword (LDC2011T07) and Google News. For the tuning and test sets, we use 1275 and 1239 sentences respectively from the LDC2010E30 corpus.

4.1 Plackett-Luce Model for SMT Tuning

We conduct a full training of machine translation models. By default, a decoder is invoked for at most 40 times, and each time it outputs 200 hypotheses to be combined with those from previous iterations and sent into tuning algorithms.

In getting the ground-truth permutations, there are many ties with the same sentence-level BLEU, and we just take one randomly. In this section, all systems have only around two hundred features, hence in Plackett-Luce based training, we sample 30 hypotheses in an accumulative n best list in each round of training.

All results are shown in Table 2, we can see that all $PL(k)$ systems do not perform well as MERT or MIRA in the development data, this maybe due to that $PL(k)$ systems do not optimize BLEU and the features here are relatively not enough compared to the size of N-best lists (empirical Formula 5). However, $PL(k)$ systems are better than MERT in testing. $PL(k)$ systems consider the quality of hypotheses from the 2th to the k th, which is guessed to act the role of the margin like SVM in classification. Interestingly, MIRA wins first in training, and still performs quite well in testing.

The $PL(1)$ system is equivalent to a max-entropy based algorithm (Och & Ney, 2002) whose dual problem is actually maximizing the conditional probability of the best BLEU hypothesis. When we increase the k , performances improve at first. After reaching a maximum around $k = 5$, they decrease slowly. We explain this phenomenon as this, when features

2. MIRA is from the open-source Moses (Koehn *et al.*, 2007)

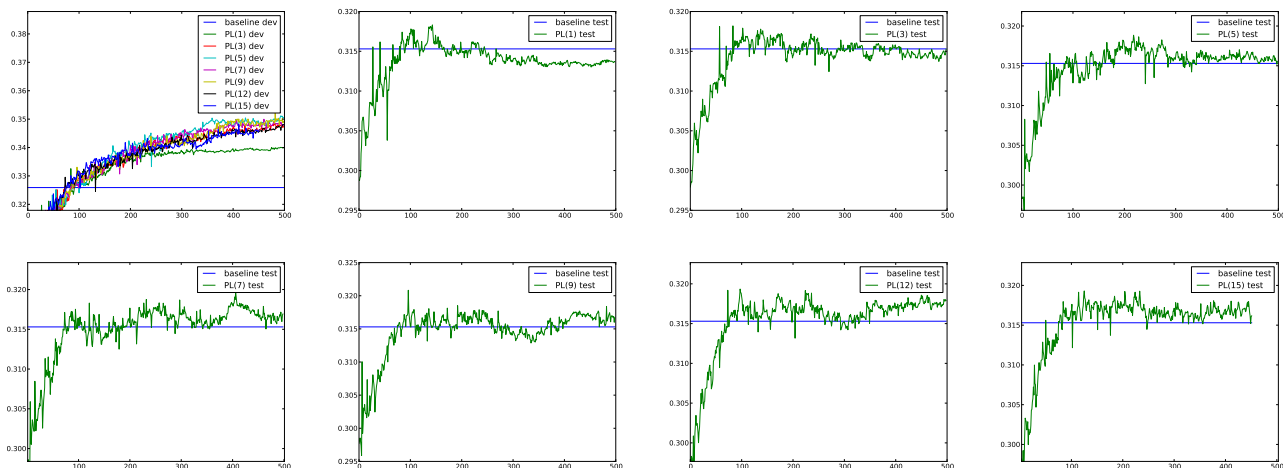


FIGURE 1 – $PL(k)$ with 500 L-BFGS iterations, $k=1,3,5,7,9,12,15$ compared with MIRA in reranking.

are rich enough, higher BLEU scores could be easily fitted, then longer ground-truth permutations include more useful information.

4.2 Plackett-Luce Model for SMT Reranking

After being de-duplicated, the N-best list has an average size of around 300, and with 7491 features. Referring to Formula 5, this is ideal to use the Plackett-Luce model. Results are shown in Figure 1. We observe some interesting phenomena.

First, the Plackett-Luce models boost the training BLEU very greatly, even up to 2.5 points higher than MIRA. This verifies our assumption, richer features benefit BLEU, though they are optimized towards a different objective.

Second, the over-fitting problem of the Plackett-Luce models $PL(k)$ is alleviated with moderately large k . In $PL(1)$, the over-fitting is quite obvious, the portion in which the curve overpasses MIRA is the smallest compared to other k , and its convergent performance is below the baseline. When k is not smaller than 5, the curves are almost above the MIRA line. After 500 L-BFGS iterations, their performances are no less than the baseline, though only by a small margin.

This experiment displays, in large-scale features, the Plackett-Luce model correlates with BLEU score very well, and alleviates overfitting in some degree.

5 Conclusion

This work has successfully brought the Plackett-Luce loss into statistical machine translation for model tuning. When in the case of large scale features, our method works better than MERT, and similarly with MIRA. This method, compared with MIRA, has a merit in practice, that is simpler to implement. Moreover, in the N-best list reranking experiment, our method also shows interesting property, resilient to overfitting. In the future work, we plan to expand the Plackett-Luce loss on packed translation forests to acquire better performance.

Références

- BERGER A. L., PIETRA V. J. D. & PIETRA S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, **22**(1).
- BERTSEKAS D. P. (1999). Nonlinear programming.
- BYRD R. H., LU P., NOCEDAL J. & ZHU C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, **16**(5).

- CAO Z., QIN T., LIU T.-Y., TSAI M.-F. & LI H. (2007). Learning to rank : from pairwise approach to listwise approach. In *Proc. of ICML*.
- CER D., JURAFSKY D. & MANNING C. D. (2008). Regularization and search for minimum error rate training. In *Proc. of WMT*.
- CHERRY C. & FOSTER G. (2012). Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 427–436 : Association for Computational Linguistics.
- CHIANG D. (2007). Hierarchical phrase-based translation. *computational linguistics*, **33**(2).
- CHIANG D., MARTON Y. & RESNIK P. (2008). Online large-margin training of syntactic and structural translation features. In *Proc. of EMNLP*.
- GALLEY M. & QUIRK C. (2011). Optimal search for minimum error rate training. In *Proc. of EMNLP*.
- GALLEY M., QUIRK C., CHERRY C. & TOUTANOVA K. (2013). Regularized minimum error rate training. In *EMNLP*, p. 1948–1959.
- HOPKINS M. & MAY J. (2011). Tuning as ranking. In *Proc. of EMNLP*.
- KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R. *et al.* (2007). Moses : Open source toolkit for statistical machine translation. In *Proc. of ACL : Poster*.
- KUMAR S., MACHEREY W., DYER C. & OCH F. (2009). Efficient minimum error rate training and minimum bayes-risk decoding for translation hypergraphs and lattices. In *Proc. of Joint ACL and AFNLP*.
- MACHEREY W., OCH F. J., THAYER I. & USZKOREIT J. (2008). Lattice-based minimum error rate training for statistical machine translation. In *Proc. of EMNLP*.
- MCDONALD R., CRAMMER K. & PEREIRA F. (2005). Online large-margin training of dependency parsers. In *Proc. of ACL*.
- MOORE R. C. & QUIRK C. (2008). Random restarts in minimum error rate training for statistical machine translation. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, p. 585–592 : Association for Computational Linguistics.
- OCH F. J. (2003). Minimum error rate training in statistical machine translation. In *Proc. of ACL*.
- OCH F. J. & NEY H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of ACL*.
- PLACKETT R. L. (1975). The analysis of permutations. *Applied Statistics*.
- TAN M., XIA T., WANG S. & ZHOU B. (2013). A corpus level mira tuning strategy for machine translation. In *EMNLP*, p. 851–856.
- WATANABE T., SUZUKI J., TSUKADA H. & ISOZAKI H. (2007). Online large-margin training for statistical machine translation. In *Proc. of EMNLP*.
- YU H., HUANG L., MI H. & ZHAO K. (2013). Max-violation perceptron and forced decoding for scalable mt training. In *EMNLP*, p. 1112–1123.
- ZHU J. & HASTIE T. (2001). Kernel logistic regression and the import vector machine. In *Proc. of NIPS*.

Natural Language Reasoning using Coq: Interaction and Automation

Stergios Chatzikyriakidis
LIRMM, University of Montpellier 2
stergios.chatzikyriakidis@lirmm.fr

Résumé. Dans cet article, nous présentons une utilisation des assistants des preuves pour traiter l'inférence en Langage Naturel (NLI). D'abord, nous proposons d'utiliser les théories des types modernes comme langue dans laquelle traduire la sémantique du langage naturel. Ensuite, nous implémentons cette sémantique dans l'assistant de preuve Coq pour raisonner sur ceux-ci. En particulier, nous évaluons notre proposition sur un sous-ensemble de la suite de tests FraCas, et nous montrons que 95.2% des exemples peuvent être correctement prédits. Nous discutons ensuite la question de l'automatisation et il est démontré que le langage de tactiques de Coq permet de construire des tactiques qui peuvent automatiser entièrement les preuves, au moins pour les cas qui nous intéressent.

Abstract. In this paper, we present the use of proof-assistant technology in order to deal with Natural Language Inference. We first propose the use of modern type theories as the language in which we translate natural language semantics to. Then, we implement these semantics in the proof-assistant Coq in order to reason about them. In particular we evaluate against a subset of the FraCas test suite and show a 95.2% accuracy and also precision levels that outperform existing approaches at least for the comparable parts. We then discuss the issue of automation, showing that Coq's tactical language allows one to build tactics that can fully automate proofs, at least for the cases we have looked at.

Mots-clés : Inference en Langage Naturel, Théorie des Types, Sémantique Formelle, FraCas, Coq, Automatisation des Preuves.

Keywords: Natural Language Inference, Type Theory, Formal Semantics, FraCas, Coq, Proof automation.

1 Introduction

1.1 Natural Language Inference

Central within a theory of formal semantics for Natural Language (NL) is the study of Natural Language Inference (NLI). Roughly put, NLI is the task of determining whether an NL hypothesis can be inferred from an NL premise. Human beings do not only have the ability to understand infinite many NL sentences but can further reason about these. In effect, understanding a NL sentence amounts (among others) to knowing what can be inferred or not from such a sentence.

Natural Language Inference has been also central in the field of computational semantics. As Cooper et al. aptly put it 'inferential ability is not only a central manifestation of semantic competence but is in fact centrally constitutive of it' (Cooper & Ginzburg, 1996). Inferential ability according to Cooper et al. (Cooper & Ginzburg, 1996) is the best way to test the semantic adequacy of NLP systems.¹

A number of NLI platforms have been proposed over the years in order to evaluate NLI systems. The two most important ones are : a) the FraCas test suite, and b) the Recognizing Textual Entailment (RTE) challenges. For the needs of this paper, we concentrate on the FraCas test suite (Cooper & Ginzburg, 1996)

1. At this point, it is necessary to distinguish between deep and shallow approaches to inference. In a nutshell, shallow approaches refers to approaches where no translation to an intermediate language is done (Romano *et al.*, 2006; Glickmann *et al.*, 2005; Hickl *et al.*, 2005; MacCartney *et al.*, 2008) among many others, where deep approaches concern approaches that perform a translation to a logical language prior to inference (Bos & Markert, 2005; Pulman, 2013). There are also hybrid approaches like (MacCartney, 2009). Obviously, the approach pursued here is a deep approach.

1.2 The FraCas test suite

The FraCas Test Suite (Cooper & Ginzburg, 1996) arose out of the FraCas Consortium, a huge collaboration with the aim to develop a range of resources related to computational semantics. The FraCas test suite is specifically designed to reflect what an adequate theory of NLI should be able to capture. It comprises NLI examples formulated in the form of a premise (or premises) followed by a question and an answer. It involves 345 examples classified according to the semantic phenomenon involved, e.g. quantifiers, adjectives, tense etc. Here are some illustrative examples from the suite :

- (1) Some irish delegates finished the survey on time.
Did any delegate finish the survey on time ? [Yes, FraCas 3.55]
- (2) All mice are small animals.
Mickey is a large mouse.
Is Mickey a large animal ? [No, FraCas 3.210]

In this paper, we discuss inference against a subset of the FraCas test suite, approximately 1/4 of the suite.

1.3 Modern Type Theories

The term Modern Type Theories (MTTs), refers to type theories within the tradition of Martin L  f (Martin-L  f, 1971; Martin-L  f, 1984).² In linguistics, this tradition has been initiated with the pioneering work of Ranta (Ranta, 1994).³ In this paper, we use one such TT, specifically Luo’s Type Theory with Coercive Subtyping (Luo, 2011, 2012a) among many others. One of the advantages of MTTs compared to traditional Montagovian approaches is that MTTs can be seen as being both model-theoretic and proof-theoretic. NL semantics can first be represented in an MTT in a model-theoretic way and then these semantic representations can be understood inferentially in a proof-theoretic way. (Luo, 2014). Besides this advantage, MTTs offer a number of features that allow for semantic fine-grainedness and expresiveness. Some of the most important ones are briefly mentioned below :

1. Type structures in MTTs are very rich. Types in MTTs can be used to represent collections of objects (or constructive sets, informally) in a model-theoretic sense, although they are syntactic entities in MTTs.
2. The notion of signature in an MTT, as introduced in (Luo, 2014; Chatzikyriakidis & Luo, 2014b), can be used to represent situations or (incomplete) possible worlds.

As regards the second point, see (Luo, 2014) for more examples. Now, elaborating on the expressiveness of typing structures of MTTs, we briefly mention the following type structures :

- Dependent sum types (Σ -types $\Sigma(A, B)$ which have product types $A \times B$ as a special case). Σ -types have been used to interpret intersective and subjective adjectives without the need of resorting to meaning postulates. The inferences follow directly from typing (Ranta, 1994; Chatzikyriakidis & Luo, 2013). Note that subtyping is essential for the Σ -type to work (Luo, 2012b).
- Dependent product types (Π -types $\Pi(A, B)$, which have arrow-types $A \rightarrow B$ as a special case). These are basic dependent types that, together with universes (see below), provide polymorphism among other things. To give an example, verb modifying adverbs are typed by means of dependent Π -types (together with the universe CN of common nouns) (Luo, 2012b; Chatzikyriakidis, 2014).
- Disjoint union types ($A + B$). Disjoint union types have been proposed to give interpretations of privative adjectives (Chatzikyriakidis & Luo, 2013).
- Universes. These are types of types, basically collections of types. Typical examples of universes in MTT-semantics include, among others, the universe *Prop* of logical propositions as found in impredicative type theories and the universe CN of (the interpretations of) common nouns (Luo, 2012b) Further uses of universes can be seen in (Chatzikyriakidis & Luo, 2012) where the universe *LType* of all linguistic types is used in order to deal with coordination.

2. Note that this is a term introduced by (Luo, 2011, 2012a) in order to distinguish type theories with the tradition of Martin-L  f and simple type theories as these are generally used within the Montagovian tradition. A similar term, ‘rich’ type theories has been used by other people like (Cooper *et al.*, 2014).

3. Potentially, even further back, with the work of Sundholm (Sundholm, 1986, 1989)

- Dot-types ($A \bullet B$). These are special types introduced to study co-predication (Luo, 2012b). It is worth mentioning that coercive subtyping is essentially employed in the formulation of dot-types.⁴

Besides the above, we should also emphasise that *subtyping* is crucial for an MTT to be a viable language for formal semantics. Furthermore, also very importantly, subtyping is needed when considering many linguistic features such as copredication (Asher, 2012; Pustejovsky, 1995).

2 Using Coq as a Natural Language Reasoner

Given MTT's proof-theoretic aspect, it is not surprising that many proof assistants implement MTTs. Starting from the early AUTOMATH system and all the way to the state of the art proof-assistants like Coq (Coq, 2004) or Agda (Agda 2008, 2008), MTTs have been shown to be a good language for interactive theorem proving. Perhaps, the most advanced of these provers is the Coq proof-assistant (Coq, 2004), a powerful proof assistant that has been used successfully to derive a number of impressive results. Some of these include a complete mechanized proof of the four colour theorem (Gonthier, 2005), the odd order theorem (Gonthier *et al.*, 2013) as well as CompCert, a formally verified compiler for C (Leroy, 2013).

Now, given : a) Coq's powerful reasoning ability and b) that it implements a MTT, a new avenue of research is opened up. To use Coq as a NL reasoner. This has been attempted in (Chatzikyriakidis & Luo, 2014a) with a number of promising results. In this paper, I present an extension of this approach that improves on accuracy and precision over previous deep approaches to NLI. We then discuss, how we can use proof automation in order to fully automate NL reasoning.

2.1 How the system works

As already said, Coq implements a MTT, more specifically the Calculus of Inductive Constructions, a type theory which is very close to the MTT we are using for representing NL semantics. It is thus straightforward to provide MTT NL semantics in Coq since Coq 'speaks' so to say the language already. Now, given that Coq is a powerful theorem prover, it can further reason about the implemented semantics. In fact, we can use Coq's proof mechanisms to prove valid NL inferences, in the same sense we use Coq to prove valid mathematical or logical theorems. We now move to exemplify how this idea can be evaluated against the FraCas test suite (Cooper *et al.*, 1996). The FraCas test suite involves three categories of NLIs : positive (YES), negative (NO) and unknown (UNK). For positive NLIs, we construct the example as a declarative hypothesis in the form of a conditional and try to prove it as a theorem. A correct account should be able to prove all YES NLIs as theorems. For negative NLIs, we formulate the example but instead we try to prove the negation of the consequent. For UNK NLIs, we should not be able to find a proof for either case of the consequent (both positive and negative). We use the modified Grammatical Framework parser (GF, (Ljunglof & Siverbo, 2011)) in order to parse the FraCas examples and then we translate the examples to the syntax of Coq.⁵ But let us see how this works by looking at example (1) repeated below as (3) :

- (3) Some irish delegates finished the survey on time.
Did any delegate finish the survey on time ? [Yes]

We assume a Σ type approach to modification, where the first projection is a coercion ($\Sigma(delegate, Irish) < delegate$ with $delegate:CN$). In Coq this is done by using dependent record types :

- (4) Record Irishdelegate : CN := mkIrishdelegate [d : > delegate ; _ : Irish d]

Adverbs and quantifiers are given the following types respectively :

- (5) $\Pi A: CN. (A \rightarrow Prop) \rightarrow (A \rightarrow Prop)$

4. See (Bassac *et al.*, 2010) for another proposal of using coercions to deal with co-predication. See also (Chatzikyriakidis & Luo, 2015) for further elaboration on the existing dot-type account.

5. Note that for the moment, we do not have an automatic translation procedure between the two. This is something that we are currently working on. Given GF's ability to translate accurately between languages (natural or formal), this task seems feasible.

(6) $\Pi A: \text{CN}. (A \rightarrow \text{Prop}) \rightarrow \text{Prop}$

These assumptions put together are enough to prove the example. The start of the proof is shown below :

```
Theorem IRISH:(some Irishdelegate(on_time(finish(the survey)))->
(some delegate)(on_time (finish(the survey)))).
```

The command *theorem* puts Coq into proof mode. We unfold the definitions using *cbv delta* which replaces the occurrences of a defined notion by the definition itself in the current goal (or in any of the hypotheses). We then apply *intro*, which introduces the antecedent as an assumption :

```
H:exists x:Irishdelegate,(let (a, _) := ADV (finish (the survey)) in a)
=====
exists x : delegate, (let (a, _):=ADV (finish (the survey))
```

Then we apply *induction* which performs *elim H* (it applies the correct destructor to an inductive type, in our case to the hypothesis *H*) and *intro* :

```
x:Irishdelegate
H:(let (a, _):= ADV (finish(the survey))in a) (let (c, _):=x in c)
=====
exists x0:delegate,(let (a, _):=ADV(finish(the survey))in a)x0
```

At this point we can substitute *x0* with *x*. The treatment of adjectival modification allows this substitution, and thus a proof can be found :

```
IRISH2 < exists x.
1 subgoal
x : Irishdelegate
H : (let (a, _):= ADV(finish (thesurvey))in a)(let (c, _)
=====
(let (a, _):=ADV(finish(the survey))in a)x
IRISH2 < assumption.
Proof completed.
```

We have tested against almost 30% of the FraCas test suite (a total of 102 plus 3 examples outside the suite). The evaluation involved examples from the following sections :

- Quantifiers and monotonicity (41 examples).
- Conjoined noun phrases (15 examples).
- Adjectives (18 examples).
- Dependent plurals (2 examples)
- Comparatives (10).
- Epistemic, intensional and reportive attitudes (11 examples).
- Collective predication (6 examples).
- Quantificational adverbs (2 examples)

100/105 examples were correctly captured, giving an overall accuracy of 95.2%. The state of the art in precision on the FraCas test suite is MacCartney (MacCartney, 2009) with an overall accuracy of 70.5%, evaluated however against 53.3% of the suite (183 examples) and against semantic phenomena that we have not tested against (e.g. ellipsis). In order to look at more direct comparisons, MacCartney offers an accuracy of 97.7% in the quantifier section, evaluating against 44 examples, while we offer an accuracy of 100% on an evaluation against 41 examples. We offer an accuracy of 87.5% while (MacCartney, 2009) 80% for the same number of examples, while for plurals we offer an accuracy of 82.5% for 17 examples while (MacCartney, 2009) offers 75% for 25 examples. The system also achieves higher accuracy than the earlier similar system proposed by (Chatzikyriakidis & Luo, 2014a), raising accuracy from 93.5% to 95.2%. However, the

current system lacks an automatic translation between the parser and the syntax of Coq and this work is done manually in this respect, as we have already pointed out.⁶ The nature of the automatic translation can significantly reduce these results if it is not efficient (basically because of low recall problems). The system as it stands can however guarantee great precision. To put things in perspective, it offers a precision of 100% for the YES section in the quantifier section while (MacCartney, 2009) offers 95.5% precision, while for the adjectival case it offers a precision of 81.8%, compared to 71.4% of (MacCartney, 2009). The following table summarizes the results from four sections of the FraCas test suite and a comparison between the approach presented here, (MacCartney, 2009) as well as (Angeli & Manning, 2014) is shown :

	Category	Count	Precision AM MC N	Recall AM MC N	Accuracy AM MC N
1	Quantifiers	AM-MC :44 N :41	91 95 100	100 100 100	95 97 100
2	Plurals	AM-MC :25 N :17	80 90 100	29 64 69.2	38 75 82.5
3	Adjectives	AM-MC :15 N :15	80 71 81.8	66 83 100	73 80 87.5
4	Attitudes	AM-MC :9 N :11	- 100 100	0 83 100	22 89 100

We conclude that the approach as presented here offers a number of welcome results compared to previous approaches, most importantly a very high precision and accuracy. It is however worth pointing out that in order to have a more realistic evaluation of the system, this should be developed into an automatic system, where a) translation into Coq is not done manually but automatically and b) the system is further evaluated against a bigger fragment of the FraCas test suite as well as against inferences using natural text like the RTE challenges.

2.2 Automation

Given that Coq is an interactive rather than an automated theorem prover, in order to prove a given theorem the user has to guide the prover to the proof. The way to do this is via Coq’s predefined tactics or any other tactic libraries that have been defined for different purposes. Coq itself involves a number of tactics that are designed in order to automate part of proofs. For example, the tactic *intuition* solves all first-order intuitionistic tautologies. For a number of examples and once the definitions have been unfolded, these can be solved with *intuition* only. Given that Coq allows the construction of new tactics, one can define a new tactic that will just unfold definitions followed by the application of *intuition*. This tactic, can then fully automate a number of the examples we have dealt in this paper. Another variant of this auto tactic, much more effective can be achieved by further adding *jauto*, a tactic like Coq’s pre-defined *auto* tactic but much more powerful since it can break up existentials as well, and congruence which can reason with equalities and inequalities. This new tactic, let us call it *AUTO* is shown below :

(7) Ltac AUTO :=cbv delta ;intuition ;try repeat congruence ;jauto ;intuition.

Most of the examples we have looked at, can be solved with this tactic (for example 3 is such a case). Other examples need more powerful tactics. For example, comparatives and collective predication examples need more powerful tactics to be solved. However, these tactics can be provided and then they can be gathered into one general auto tactic that can effectively solve all the examples. This is indeed what we have done, we defined 3 different auto tactics and then combined these three tactics (let us call them *a*, *b* and *c*) into a general auto tactic (*d*) :

(8) Ltac d := solve[a|b|c].

With this tactic full automation can be achieved. The tactic succeeds if one of the tactics can solve the theorem. Otherwise, it fails.

3 Conclusions

In this paper, we have presented the use of Coq as an NL reasoner. Given Coq’s ‘understanding’ of MTT semantics, we straightforwardly implemented MTT semantics for NL. We evaluate the approach against almost 30% of the FraCas test

⁶. The details of such a translation are currently under development and we hope that an efficient translation that will maintain the impressive results as regards accuracy will be available.

suite, where an accuracy of 95.2% was achieved and a better overall performance in the three comparable sections of the FraCas with earlier approaches was shown. We then discussed ways of using Coq’s tactical language in order to fully automate the proof process. It was shown that at least for the examples we have looked at, this is feasible.

Références

- Agda 2008 (2008). Agda proof assistant. <http://appserv.cs.chalmers.se/users/ulfn/wiki/agda.php>
- ANGELI G. & MANNING C. D. (2014). Naturalli : Natural logic inference for common sense reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- ASHER N. (2012). *Lexical Meaning in Context : a Web of Words*. Cambridge University Press.
- BASSAC C., MERY B. & RETORÉ C. (2010). Towards a type-theoretical account of lexical semantics. *Journal of Logic, Language and Information*, **19**(2).
- BOS J. & MARKERT K. (2005). Recognising textual entailment with logical inference. In *Proc. of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 98–103.
- CHATZIKYRIAKIDIS S. (2014). Adverbs in a modern type theory. In N. ASHER & S. SOLOVIEV, Eds., *Proceedings of LACL2014, LNCS 8535*, p. 44–56.
- CHATZIKYRIAKIDIS S. & LUO Z. (2012). An account of natural language coordination in type theory with coercive subtyping. In Y. PARMENTIER & D. DUCHIER, Eds., *Proc. of Constraint Solving and Language Processing (CSLP12)*. LNCS 8114, p. 31–51, Orleans.
- CHATZIKYRIAKIDIS S. & LUO Z. (2013). Adjectives in a modern type-theoretical setting. In G. MORRILL & J. NEDERHOF, Eds., *Proceedings of Formal Grammar 2013*. LNCS 8036, p. 159–174.
- CHATZIKYRIAKIDIS S. & LUO Z. (2014a). Natural language inference in Coq. *J. of Logic, Language and Information*, **23**(4), 441–480.
- CHATZIKYRIAKIDIS S. & LUO Z. (2014b). Using signatures in type theory to represent situations. *Logic and Engineering of Natural Language Semantics 11*. Tokyo.
- CHATZIKYRIAKIDIS S. & LUO Z. (2015). Individuation criteria, dot-types and copredication : A view from modern type theories. In *Under Review*.
- COOPER R., CROUCH D., VAN EIJCK J., FOX C., VAN GENABITH J., JASPARS J., KAMP H., MILWARD D., PINKAL M., POESIO M. & PULMAN S. (1996). Using the framework. *Technical Report LRE 62-051r*. <http://www.cogsci.ed.ac.uk/fracas/>.
- COOPER R., DOBNIK S., LAPPIN S. & LARSSON S. (2014). A probabilistic rich type theory for semantic interpretation. In *Proceedings of the European Association of Computational Linguistics*.
- COOPER R. & GINZBURG J. (1996). A compositional situation semantics for attitude reports. In J. SELIGNMANN & D. WESTERSTAHL, Eds., *Logic, language and computation*, CSLI.
- Coq (2004). *The Coq Proof Assistant Reference Manual (Version 8.0)*, INRIA. The Coq Development Team.
- GLICKMANN O., DAGAN I. & KOPPEL M. (2005). Web based probabilistic textual entailment. In *Proceedings of the PASCAL Challenges Workshop on Recognizing Textual Entailment*.
- GONTHIER G. (2005). A computer checked proof of the four colour theorem.
- GONTHIER G., ASPERTI A., AVIGAD J., BERTOT Y., COHEN C., GARILLOT F., LE ROUX S., MAHBOUBI A., O’CONNOR R., BIHA S. O. *et al.* (2013). A machine-checked proof of the odd order theorem. In *Interactive Theorem Proving*, p. 163–179. Springer.
- HICKL A., WILLIAMS J., BENSLEY J., ROBERTS K., RINK B. & SHI Y. (2005). Recognizing textual entailment with ICC’s groundhog system. In *Proc. of Second PASCAL Challenges Workshop on Recognizing Textual Entailment*, p. 80–85.
- LEROY X. (2013). The compcert c verified compiler : Documentation and user’s manual. <http://compcert.inria.fr/man/manual.pdf>.
- LJUNGLOF P. & SIVERBO M. (2011). *A bilingual treebank for the FraCaS test suite*. Clt project report, University of Gothenburg.

- LUO Z. (2011). Contextual analysis of word meanings in type-theoretical semantics. *Logical Aspects of Computational Linguistics (LACL'2011)*. LNAI 6736.
- LUO Z. (2012a). Common nouns as types. In D. BECHET & A. DIKOVSKY, Eds., *Logical Aspects of Computational Linguistics (LACL'2012)*. LNCS 7351.
- LUO Z. (2012b). Formal semantics in modern type theories with coercive subtyping. *Linguistics and Philosophy*, **35**(6), 491–513.
- LUO Z. (2014). Formal Semantics in Modern Type Theories : Is It Model-theoretic, Proof-theoretic, or Both ? *Invited talk at Logical Aspects of Computational Linguistics 2014 (LACL 2014), Toulouse*. LNCS 8535, p. 177–188.
- MACCARTNEY B. (2009). *Natural Language Inference*. PhD thesis, Stanford University.
- MACCARTNEY B., GALLEY M. & C.D. MANNING I. (2008). A phrase-based alignment model for natural language inference. In *Proceedings of EMNLP-08*, p. 802–811.
- MARTIN-LÖF P. (1971). An intuitionistic theory of types. manuscript.
- MARTIN-LÖF P. (1984). *Intuitionistic Type Theory*. Bibliopolis.
- PULMAN S. (2013). Second order inference in NL semantics. Talk given at the KCL Language and Cognition seminar, London.
- PUSTEJOVSKY J. (1995). *The Generative Lexicon*. MIT.
- RANTA A. (1994). *Type-Theoretical Grammar*. Oxford University Press.
- ROMANO L., KUYLEKOV M., SZPEKTOR I., DAGAN I. & LAVELLI A. (2006). Investigating a generic paraphrase-based approach for relation extraction. In *Proceedings of EACL 2006*, p. 409–416.
- SUNDHOLM G. (1986). Proof theory and meaning. In D. GABBAY & F. GUENTHNER, Eds., *Handbook of Philosophical Logic III : Alternatives to Classical Logic*, p. 471–506. Reidel.
- SUNDHOLM G. (1989). Constructive generalized quantifiers. *Synthese*, **79**(1), 1–12.

Vous aimez ?...ou pas ? LikeIt, un jeu pour construire une ressource lexicale de polarité

Mathieu Lafourcade¹, Nathalie Le Brun², Alain Joubert¹

(1) Lirimm, Université Montpellier, France

(2) Imagin@t, 34400 Lunel, France

mathieu.lafourcade@lirimm.fr, imaginat@imaginat.name, alain.joubert@lirimm.fr

Résumé. En analyse de discours ou d'opinion, savoir caractériser la connotation générale d'un texte, les sentiments qu'il véhicule, est une aptitude recherchée, qui suppose la constitution préalable d'une ressource lexicale de polarité. Au sein du réseau lexical JeuxDeMots, nous avons mis au point LikeIt, un jeu qui permet d'affecter une valeur positive, négative, ou neutre à un terme, et de constituer ainsi pour chaque terme, à partir des votes, une polarité résultante. Nous présentons ici l'analyse quantitative des données de polarité obtenues, ainsi que la méthode pour les valider qualitativement.

Abstract.

Do you like it? or not? LikeIt, a game to build a polarity lexical resource

The ability to analyze the feelings that emerge from a text requires having a polarity lexical resource. In the lexical network JeuxDeMots we designed LikeIt, a GWAP that allows attributing a positive, negative or neutral value to a term, and thus obtaining for each term a resulting polarity. We present a quantitative analysis of polarity data obtained, together with the comparison method we developed to validate them qualitatively.

Mots-clés : polarité, sentiment, réseau lexical, crowdsourcing, GWAP

Keywords: polarity, feelings, lexical network, crowdsourcing, GWAP

1 Introduction

Être capable de caractériser les sentiments est devenu incontournable dans le cadre d'applications dédiées à l'analyse de discours politiques, ou d'opinions relatives à la fourniture de services touristiques, culturels, ou de biens de grande consommation. La constitution d'une ressource lexicale de polarité (associer à un terme une connotation positive, négative, neutre, et éventuellement objective ou subjective) est un préalable à ce type de recherche, que les approches soient statistiques supervisées ou plus linguistiques (Brun, 2011). Une telle polarité peut être exprimée avec une valeur numérique (Taboada *et al.*, 2011) ou plusieurs valeurs. Par exemple, (Saif et Turney, 2013) ont utilisé deux valeurs (positif/négatif) pour leur ressource (EmoLex) de polarité/sentiment pour l'anglais obtenue par crowdsourcing à l'aide d'Amazon Mechanical Turk (ce qui peut poser problème, voir (Fort *et al.*, 2014)). SentiWordNet (Esuli et Sebastiani, 2006) tout comme WordNet Affect (Strapparava et Valitutti, 2004) sont des extensions de WordNet où les termes sont polarisés sur trois valeurs (positif, négatif et objectif) dont la dernière s'oppose aux deux premières. Des approches par propagation à partir d'un noyau manuel (voir, par exemple (Gala et Brun, 2012) et (Lafourcade et Fort, 2014)) ont également été utilisées, mais peuvent ne pas refléter exactement l'opinion des locuteurs. Les approches compositionnelles ou par apprentissage sont généralement celles adoptées pour l'utilisation de ce type de données (voir, par exemple (Kim et Hovy, 2004) et (Turney, 2002)).

Un réseau lexical, tel celui obtenu grâce au jeu en ligne JeuxDeMots (Lafourcade, 2007), comporte des termes associés par des relations lexico-sémantiques. Le projet JeuxDeMots (JDM) a permis non seulement la constitution d'un réseau lexical en perpétuelle extension et libre d'accès, mais également la validation/vérification des relations qui le constituent via un certain nombre de jeux et contre-jeux (Lafourcade *et al.*, 2015). L'expérimentation de méthodes d'acquisition d'informations de polarité peut être réalisée avec une telle ressource.

Il peut être intéressant d'associer à des termes des informations correspondant à des ensembles fermés de valeurs. Par exemple, la polarité peut être définie sur la base de trois valeurs : positif, négatif et neutre. On remarquera que de nombreuses caractéristiques sémantiques peuvent être associées à de tels ensembles de taille variable : les sentiments/émotions (colère, peur, joie, amour, tristesse, ...), ou encore les couleurs (rouge, bleu, jaune, vert, orange, violet, noir, blanc, ...). Comme ce type d'association ne peut être obtenu via le jeu principal du réseau JDM, nous avons développé plusieurs jeux permettant de tagger les termes selon des critères (j'aime / je n'aime pas, émotion associée, couleur associée ...). Les données collectées ont de nombreuses applications, que ce soit en analyse du discours ou en

désambiguïsation. Une telle annotation est complexe, car souvent subjective et/ou contextuelle : par exemple, une même remarque adressée à un individu peut être considérée comme un trait d'humour, un conseil, une critique, une réprimande ... selon l'énonciateur, l'interlocuteur, et le contexte.

Dans cet article, nous commencerons par présenter les principes de la détermination de cette polarisation via le jeu LikeIt, ainsi que les caractéristiques de sa propagation dans le réseau. Nous indiquerons ensuite les résultats obtenus via une analyse quantitative et qualitative. La méthodologie d'évaluation qualitative, qui se fonde sur un croisement entre données de polarité et données de sentiments (c'est-à-dire sentiments que les joueurs associent spontanément à un terme donné) sera approfondie. Nous concluons en explicitant quelques-unes des perspectives envisagées.

2 LikeIt, un jeu de polarité

Dans LikeIt, la polarisation consiste, comme sur les réseaux sociaux, à affecter la mention « J'aime » / « Je n'aime pas » (ou encore « neutre ») à un terme. Bien que cette polarisation soit très subjective, en plus d'être étroitement liée au contexte, notre hypothèse de travail est que de nombreux termes et usages de termes possèdent une polarité intrinsèque, qu'il est possible de capturer globalement, en interrogeant un grand nombre de locuteurs : on vérifie ainsi si une polarité majoritaire se dégage, et si oui, laquelle.

2.1 Principe de fonctionnement

LikeIt procède par l'*appréciation* simplifiée, avec des questions simples posées au joueur, sur le modèle : *Est-ce que vous aimez l'idée de* suivi d'un terme. Les seules réponses possibles sont *oui*, *non* et *neutre*. Ce jeu enrichit ainsi le réseau lexical avec trois types de polarités, ce qui nous a semblé être la manière à la fois la plus souple et la plus exhaustive pour caractériser les termes. Cela permet notamment de distinguer les termes laissant majoritairement indifférent (majorité de *neutre*) de ceux qui suscitent des avis très partagés (valeurs à peu près égales pour *positif* et *néglatif*). L'exploitation préliminaire de ces données seules, dans le cadre d'une désambiguïsation lexicale, semble montrer que la polarité prise isolément permet de sélectionner le bon usage d'un terme en contexte dans approximativement 50% des cas. Ce type de données peut aussi être utilisé en analyse d'opinions, par composition des polarités des termes fortement polarisés (i.e. ceux dont la polarité majoritaire représente plus de 50% des valeurs cumulées des trois possibles). La figure 1 est constituée de deux copies d'écran d'une partie de LikeIt.



FIGURE 1. Deux écrans consécutifs de LikeIt. Suite à la réponse donnée dans l'écran de gauche (*examen scolaire*), le joueur voit immédiatement en haut de l'écran suivant (image de droite et zoom du bas), le pourcentage de joueurs qui partagent son avis : le jeu fournit ainsi un retour direct au joueur avec une relance immédiate (nouvelle question). La mention "...et curieusement vous n'avez répondu cela que 33% des fois" indique au joueur qu'il a déjà été confronté à ce mot et qu'il n'a pas été constant dans son choix.

LikeIt est un jeu de consensus à votes dont certaines propriétés sont développées ci-dessous. Concernant son intérêt ludique et ses qualités en tant que GWAP (Lafourcade, 2015), on relève parmi quelques caractéristiques notables qui en font un jeu addictif :

- **la simplicité** : les réponses positive/négative/neutre, s'apparentent à celles demandées lors d'un sondage ; cependant, d'après les retours, la diversité du vocabulaire et des sujets qu'il évoque est telle, que les joueurs n'ont pas l'impression de répondre à une enquête d'opinion. De plus, la réponse par simple clic rend possible de jouer depuis un smartphone sur des plages temporelles relativement réduites (tram, salles d'attente, etc.). LikeIt est donc un jeu à

parties très courtes et relance immédiate, ce qui, sur un plan quantitatif, le rend déjà très efficace pour collecter des données.

– **diversité du vocabulaire et variabilité de la réponse** : certains termes suscitent des sentiments très partagés (par exemple, le terme *bloc opératoire* est positif dans l'absolu : c'est une structure médicale indispensable, qui sert à soigner, mais qui peut être connotée négativement lorsqu'on est personnellement concerné) et ce faisant, l'avis d'un même joueur peut évoluer dans le temps, et en fonction des circonstances. Ainsi, le terme *baccalauréat* ou *examen scolaire* peut susciter un sentiment négatif quand on est lycéen, mais il est toujours perçu positivement lorsqu'on l'a obtenu. Le fait de jouer sur du vocabulaire à très large couverture rend le jeu intéressant et varié.

2.2 Nature des données obtenues

Les réponses issues de LikeIt génèrent pour chaque terme un triplet de nombres de votes pour chacune des trois polarités possibles. De ce triplet, on déduit une répartition en pourcentages que l'on nomme *polarisation* ainsi que des valeurs d'intensité (le nombre de votes total pour le terme, et la *norme* de la polarisation, alors vue comme un vecteur à trois composantes). Plus l'intensité, et la norme en particulier, sont élevées, plus la polarisation est fiable. Il est difficile de définir un seuil moyen à partir duquel on pourrait considérer que le nombre de votes est suffisant pour générer des pourcentages de polarité fiables. En effet ce nombre est étroitement lié à la nature et aux caractéristiques de chaque terme (polarité unique forte ou très partagée, terme polysémique, etc.). Toutefois, on peut faire empiriquement une approximation et évaluer le nombre minimal de votes garantissant une répartition fiable des polarités à 10 fois le nombre de polarités (soit au moins 10 votes pour une monopolarité, 20 pour une bipolarité, 30 pour une tripolarité). Un terme sera considéré comme étant fortement *orienté* si une des polarités est majoritaire (au moins supérieure à 50%). La table 1 présente quelques exemples de résultats où le nombre de votes paraît suffisant pour juger fiables les polarisations obtenues.

Terme	Distribution de la polarité en %	Intensité
cadeau	POS: 82 NEUT: 14 NEG: 4	Nb votes : 280 norme : 232.73
retraite	POS: 48 NEUT: 18 NEG: 34	Nb votes : 303 norme : 190.6
CRS	POS: 29 NEUT: 15 NEG: 56	Nb votes : 274 norme : 177.45
automne	POS: 37 NEUT: 44 NEG: 18	Nb votes : 277 norme : 168.34

TABLE 1. Exemples des polarisations avec LikeIt : le terme *cadeau* présente une quasi-monopolarité, alors que les trois autres termes sont bi, voire tripolarisés. Le terme *cadeau* est fortement positif, *CRS* fortement négatif.

2.3 Mécanisme de sélection des termes

Il serait contre-productif de sélectionner un terme à proposer au joueur de façon totalement équiprobable sur l'ensemble des termes du réseau lexical de JDM. En effet, on peut imaginer qu'une grande proportion de ces termes a une polarité globalement neutre. De plus, le réseau comporte des termes très spécialisés, ce qui est à double-tranchant : intéressant si on connaît le terme, mais décourageant sinon, d'où la nécessité de minimiser la fréquence de proposition de ce type de terme. Pour ces raisons, la sélection d'un terme à proposer au joueur est effectuée selon une approche par propagation dans le réseau lexical de JDM. Le principe de l'algorithme est le suivant :

- un terme T ayant une valeur de polarité positive et/ou négative est choisi aléatoirement dans le graphe (nous ignorons délibérément les polarités neutres) ;
- on propose T au joueur selon une probabilité p (empiriquement fixée à 0.5), ou un voisin V de T selon une probabilité $1-p$. V est choisi de façon équiprobable parmi tous les voisins possibles ;
- si le nombre de votes toutes polarités confondues pour V dépasse un seuil s (empiriquement fixé à 1000) alors p vaut 0.9. Nous avons donc une probabilité de 90% de proposer T , et non pas V lui-même.
- L'amorçage a été effectué uniquement en polarisant à la main les termes *bon* (1 vote de polarité positif), et *mauvais* (1 vote de polarité négatif).

Ainsi, cet algorithme très simple (une pseudo marche aléatoire à un pas) effectue une propagation au sein du graphe entre termes potentiellement intéressants pour l'acquisition de valeurs de polarité, à savoir ceux qui ne sont pas neutres, en évitant partiellement ceux qui sont déjà très renseignés. Un terme neutre ne peut être sélectionné que par le voisinage. Un terme ayant un grand nombre de voisins sera plus rapidement polarisé que les autres, car plus souvent atteint par voisinage (tant que le nombre de votes reste sous le seuil s). Une sélection strictement aléatoire au sein du réseau lexical donnerait une proportion trop importante de termes neutres, ce qui diminuerait l'intérêt du jeu. L'échantillonnage des termes est donc réalisé par cet algorithme de propagation en fonction de la topologie du réseau lexical et des données de polarité déjà disponibles (qui viennent des joueurs). En ce qui concerne la polysémie des

termes, notons que des mots raffinés, donc des sens différents d'un même terme, peuvent être sélectionnés, ainsi que certains termes en contexte existant déjà dans le réseau, tels que *chat* [*carac*] *pelé*.

2.4 Biais observés

Le jeu LikeIt présente un premier biais : pour un terme polysémique, il est possible que la réponse du joueur subisse une *contamination* par un sens anecdotique fortement polarisé. Par exemple, *vache* dont le sens premier, l'animal, est globalement neutre (voire légèrement positif) peut être contaminé par le sens de *vache* (*méchant*), qui a une forte polarité négative. Ainsi, les joueurs, se trouvant *de facto* par le jeu dans un contexte de polarité, vont voter en pensant au sens le plus polarisé, et donc choisir une polarité négative pour le terme *vache*. Il en est de même pour les termes : *fumier* (sens premier engrais, contaminant insulte), *cellule* (sens premier biologie, contaminant prison), etc. Cependant, les raffinements de chacun de ces termes, eux, ne subissent aucune contamination et disposent bien d'une polarité conforme à celle attendue. Certains termes ont une polarité double (positive et négative) car les joueurs peuvent les interpréter de façon aussi bien diégétique, c'est-à-dire en s'identifiant au personnage ou à l'action, qu'extradiégétique, c'est-à-dire en adoptant un point de vue extérieur. Ainsi, les termes *dragon*, *orc*, *sorcière*, *vampire*, etc. sont tout à la fois négatifs (quand on les appréhende au niveau diégétique) et positifs (si on les considère de façon extradiégétique). Cette divergence de perception est un second biais.

Lors de l'évaluation quantitative des données obtenues, nous avons constaté un troisième biais qui a tendance à favoriser la polarité positive. Ce troisième biais est explicité à la section suivante.

3 Evaluation des données de polarité obtenues avec LikeIt

3.1 Evaluation quantitative

Durant les trois premiers mois, plus de 25 000 termes ont été *polarisés* (c'est-à-dire dotés d'une information de polarité) pour un total dépassant 150 000 votes. Depuis 2008, plus de 360 000 termes ont été *polarisés* pour un total supérieur à 75 millions de votes (<http://www.jeuxdemots.org/likeit.php?action=list>) pour plusieurs centaines de milliers de joueurs (nombre d'adresses IP différentes mesuré en janvier 2013). L'ensemble du réseau JDM contient environ 490 000 termes, c'est-à-dire qu'environ 70% des termes du réseau ont été atteints par l'algorithme de propagation.

323 292 polarités positives (38.3 %)	44 762 101 votes positifs (57.5 %)
350 536 polarités neutres (41.6 %)	21 767 725 votes neutres (28 %)
169 768 polarités négatives (20.1 %)	11 276 692 votes négatifs (14.5 %)
Total = 843 596 polarités (100 %)	Total = 77 806 518 votes (100 %)

51 151 polarités < 10 votes	792 445 polarités >= 10 votes
151 847 polarités < 20 votes	691 749 polarités >= 20 votes
289 886 polarités < 40 votes	553 710 polarités >= 40 votes
518 624 polarités < 80 votes	324 972 polarités >= 80 votes

10 515 termes à polarité positive uniquement	(2.9 %)
11 193 termes à polarité neutre uniquement	(3 %)
6 498 termes à polarité négative uniquement	(1.8 %)
177 221 termes à polarité positive et neutre uniquement	(48.1 %)
1 150 termes à polarité positive et négative uniquement	(0.3 %)
27 715 termes à polarité neutre et négative uniquement	(7.5 %)
134 405 termes ayant les trois polarités	(36.5 %)
Total = 368 697 termes ayant au moins une polarité	(100 %)

TABLES 2A, 2B ET 2C. Données quantitatives des polarisations obtenues avec LikeIt.

Les données, présentées aux tables 2a et 2c, semblent présenter un biais vers la polarité positive qui représente plus de la moitié des votes. En effet, en interrogeant les joueurs, on s'aperçoit que beaucoup de termes perçus comme relativement neutres (comme par exemple *odonate* - www.jeuxdemots.org/diko.php?gotermrel=odonate), peuvent souvent être néanmoins l'objet d'un vote positif. Il semblerait que le biais soit la conséquence d'un adage qui serait « je peux aimer ce que je ne déteste pas ». Dans quelle mesure ce second biais influence-t-il les résultats ? Il est difficile de l'évaluer car on ne connaît pas *a priori* les termes du lexique qui seraient ou positifs ou neutres.

On peut aussi expliquer partiellement ce biais positif en considérant qu'une majorité des termes proposés sont des entités nommées soit relèvent de domaines suscitant l'adhésion : par exemple, l'immense majorité des personnes célèbres ont plutôt une polarité positive (les acteurs et actrices notamment). Pour les hommes politiques, cela peut être

d'avantage partagé. De même, les entités nommées d'œuvres (films, tableaux, etc.) ont très majoritairement une polarité positive, ainsi que la majeure partie du vocabulaire lié à la gastronomie (et particulièrement les noms de plats, de boissons).

3.2 Evaluation qualitative et méthodologie

La question de la méthode à adopter pour une évaluation qualitative se pose dans la mesure où il n'existe aucune ressource relative à la polarité à laquelle les données issues de LikeIt pourraient être confrontées. Il est possible d'entreprendre une évaluation manuelle, qui consisterait à regarder une par une un certain nombre d'entrées et à se prononcer sur la pertinence des polarités. Cette perspective est déraisonnable vu la taille des données (plus de 360 000 termes polarisés) et la question de la constitution de l'échantillonnage reste posée (comment sélectionner les termes à inspecter ?).

Nous disposons, avec la mise à disposition des données de JDM, (via le jeu principal, et un jeu spécifique Emot - <http://www.jeuxdemots.org/emot.php>), d'associations entre termes et termes désignant des sentiments ou des émotions. Emot est un jeu de choix semi-ouvert (choix fermé et réponses libres en mode avancé) qui demande au joueur d'associer des sentiments ou des émotions à un terme donné. Jouer à JDM sur la relation *sentiment* est complètement ouvert (réponses libres). Les données obtenues pour un terme prennent la forme d'une liste de termes (sentiments) pondérés, par exemple :

- cadeau : joie (+) 1712 ; surprise (+):1142 ; bonheur (+) 980 ; amour (+) 780 ; plaisir (+) 741 ; amitié (+) 660 ; reconnaissance (+) 310 ; déception (-) 260 ; étonnement (+) 222 ; gratitude (+) 210 ; générosité (+) 200 ; satisfaction (+) 160 ; content (+) 140 ; contentement (+) 120 ; envie (+) 100 ; gêne (-) 90 ; émotion (0) 81 ; émerveillement (+) 80 ; impatience (-) 70 ; jalousie (-) 70 ; heureux (+) 60 ; présent (+) 59 ; fête (+) 56 ; sympathie (+) 55 ; frustration(-) 50 ; confusion (-) 50 ;
- CRS : sécurité (+) 1027 ; peur (-) 1007 ; violence (-) 817 ; haine (-) 357 ; crainte (-) 297 ; colère (-) 186 ; force (0) 137 ; protection (+) 127 ; répression (-) 127 ; insécurité (-) 127 ; angoisse (-) 117 ; révolte (-):117 ; injustice (-) 99 ; brutalité (-) 97 ; panique (-) 97 ; respect (+) 93 ; terreur (-) 87 ; agressivité (-) 87 ; rage (-) 87 ; méfiance (-) 87 ; inquiétude (-) 77 ; douleur (-) 77 ; rejet (-) 77 ; trouille (-):67 ; aveuglement (-) 66 ; défiance (-) 65 ; honte (-) 63 ; incompréhension (-) 57 ; détresse (-) 57 ; soulagement (+) 57 ; frayeur (-) 32 ; appréhension (-) 32 ;
- bras : force (0) 110 ; protection (+) 100 ; soutien (+) 80 ; union (+) 5 ; indifférence (*) 4 ;

Dans les exemples ci-dessus nous avons indiqué à la suite de chaque terme sa polarité absolument majoritaire (plus de 50% des votes) au moyen d'un symbole entre parenthèses : (+) pour *positif*, (-) pour *négatif* et (*) pour *neutre*. Le symbole (0) caractérise un terme qui n'a pas de polarité absolument majoritaire ; c'est le cas du terme *force* associé à *bras* par exemple.

Comme les exemples présentés ci-dessus le suggèrent, il est possible de calculer une polarisation pour un terme auquel des termes de sentiments ont été associés. Ceci n'est évidemment possible que si les termes associés disposent eux-mêmes d'une polarisation. Mais on constatera qu'en pratique, c'est forcément le cas : les termes désignant des sentiments ont été les premiers à être polarisés car ils ont été très rapidement atteints par l'algorithme de propagation. Pour un terme, on calcule donc la polarisation en faisant la somme des vecteurs de polarité de chaque sentiment associé.

La polarisation calculée via les sentiments associés peut ainsi être comparée à celle issue du jeu LikeIt. On compare donc une polarité inférée à une polarité directement établie par les joueurs. Il est possible de comparer les deux polarités via une mesure cosinus, ou une mesure du max (1 si les deux polarités maximum coïncident). L'intérêt de cette approche est qu'elle est directement automatisable, et nous permet de réserver l'effort d'inspection manuelle aux cas divergents. Nous avons considéré les mesures cos et max, en ordonnant les 5 000 premiers termes par poids décroissants pour la relation sentiment (donc ceux qui ont été le plus alimentés en termes de sentiments en premier).

n premiers termes	Moyenne du Cos	Moyenne du Max	Moyenne des polarités max
1 000	0.80	0.76	85.65 %
2 000	0.83	0.79	86.55 %
3 000	0.80	0.75	87.40 %
4 000	0.82	0.77	87.49 %
5 000	0.83	0.79	87.63 %

TABLE 3. Evaluation qualitative des données de polarisation issues de LikeIt par comparaison avec celles calculées à partir des associations de sentiments produites par Emot et JDM.

D'après la table 3, on constate une concordance assez élevée entre la polarité définie par les réponses des joueurs et celle induite par les sentiments associés aux termes. La moyenne des polarités maximales issues des données de LikeIt, *mpm*, représente le taux d'accord maximum que peut atteindre en moyenne l'opinion générale, pour les *n* termes les plus joués. Nous constatons une variabilité de 15 à 12% ($100 - mpm$) entre les 1000 et les 5 000 termes les plus renseignés. La variabilité a tendance à augmenter quand le nombre de votes augmente, ce qui semble logique car il y a plus d'opinions divergentes quand le nombre de votes est élevé.

Nous avons examiné manuellement les cas de divergence (valeur de la mesure Max égale à 0). Il s'avère que la quasi-totalité des termes pour lesquels il y a peu d'accord sont ceux qui sont diégétiquement contrastés, à savoir ceux qui sont négatifs selon une perspective diégétique et positifs selon une perspective extradiégétique. Quelques termes contrastés :

classe prépa, thèse, pince-oreille, analyses, murène, micropénis, exprimer [sujet] femme, dragon

Pour le dire autrement, il semblerait que le processus d'association des sentiments tende à mettre le joueur en situation (configuration diégétique) alors que la demande de polarité (LikeIt) est typiquement extradiégétique (le joueur adopte un point de vue externe). Par exemple, si l'on demande à un joueur d'énumérer les sentiments qu'il associe au terme *dieu vengeur*, il va avoir tendance à le faire en s'impliquant personnellement dans le contexte que lui suggère le mot, et à répondre *peur, crainte, soumission...* Inversement si il doit dire, via LikeIt, s'il aime cette expression ou pas, il va répondre *oui*, car *dieu vengeur* évoque un récit mythologique, donc quelque chose de plutôt distrayant. C'est ainsi que *dieu vengeur* a une polarité négative lorsqu'on la calcule via les sentiments associés, et positive lorsqu'elle est issue de LikeIt. On notera que tous les cas de divergence observés portent sur un terme polarisé négativement via les sentiments associés, et positivement via LikeIt. Les termes à polarité forte sont insensibles à l'aspect diégétique. Soulignons également que les termes les plus soumis à la subjectivité du jugement montrent des polarités multiples, mais dont la répartition est concordante quel que soit le mode d'évaluation (polarité directe ou polarité induite).

Perspectives et conclusion

Au vu de notre expérience sur les différents jeux présentés ci-dessus, nous pouvons envisager un certain nombre de perspectives, à commencer par la poursuite de cette double approche comme moyen d'accroître encore notre ressource lexicale de polarité (librement accessible à <http://www.jeuxdemots.org/likeit.php?action=list>). Toutes sortes de caractéristiques (dimension, température, importance/prépondérance, temporalité, localisation ...) peuvent être soumises à l'appréciation des joueurs, et donc donner lieu à des jeux de type LikeIt, visant à les quantifier. Mais une étude préliminaire pour identifier les plus utiles, les plus informatives doit impérativement être réalisée, afin d'éviter, en multipliant ce type de jeu, de provoquer une lassitude, donc de la démotivation chez les joueurs. Notons que les données produites via ce type de jeux, qui ne nécessitent aucune autre connaissance qu'une relative maîtrise de la langue, sont de qualité, ce qui légitime ce type d'approche.

On peut se poser la question de l'évolution possible de la polarité d'un terme, en particulier dans le temps : un même terme peut inspirer de manière globale des sentiments différents en fonction du temps, du contexte ; par exemple, le terme *volcan* suscite plutôt de la curiosité ou de l'indifférence, sauf lorsqu'une éruption imminente menace des populations ou le trafic aérien, l'inquiétude ou la peur prenant alors le dessus, dans la diversité des sentiments exprimés. Les sentiments à l'égard d'une personne publique, d'une œuvre (entités nommées) peuvent être très fluctuants dans le temps, et si des sentiments contradictoires apparaissent dans le réseau, il pourrait être intéressant pour les représenter d'y associer une notion de contexte, par exemple *Depardieu [contexte] cinéma*, et *Depardieu [contexte] fiscalité*. Dans le cadre d'une extension de LikeIt restant dans l'esprit du crowdsourcing, les joueurs pourraient être sollicités pour fournir un ou plusieurs contextes associés à des polarités contrastées.

Certains termes pourraient être polarisables automatiquement, en fonction des relations qui les concernent. Par exemple, la relation *caractéristique* est très porteuse de polarité, *veuve [caractéristique] triste* permet de polariser le terme *veuve* négativement. Cependant, l'approche par crowdsourcing est globalement plus fiable et plus rapide, aussi bien pour les termes à forte polarité que pour les termes complexes. L'approche et les outils présentés dans cet article sont relativement récents, et le nombre de termes polarisés représente une proportion non anecdotique (70%) de l'ensemble du réseau lexical. Compte tenu des résultats obtenus présentés ici, nous estimons avoir montré la faisabilité, l'intérêt, et les perspectives d'un tel projet, et très largement commencé à construire la ressource correspondante.

Références

- BRUN C. (2011). Detecting opinions using Deep Syntactic Analysis. Proceedings of *Recent Advances in Natural Language Processing* (RANLP 2011), Hissar, Bulgaria, pp. 392-398.
- ESULI A. AND SEBASTIANI F. (2006). SentiWordNet: a publicly available lexical resource for opinion mining. Proceedings of *LREC-06*, Gêne, Italie, 6 p.
- FORT K, ADDA G., SAGOT B., MARIANI J. ET COUILLAUT A. (2014) Crowdsourcing for Language Resource Development : Criticisms about Amazon Mechanical Turk Overpowering Use. *Lecture Notes in Artificial Intelligence*. Springer, pp. 303-314, 2014, 978-3-319-08957-7.
- GALA N., ET BRUN C. (2012). Propagation de polarités dans des familles de mots : impact de la morphologie dans la construction d'un lexique pour l'analyse d'opinions. Actes de *Traitement Automatique des Langues Naturelles* (TALN 12), Grenoble, juin 2012, pp. 495-502.
- KIM S. M., AND HOVY E. (2004). Determining the sentiment of opinions. Proceedings of *COLING- 04*, Barcelone, Espagne, pp. 1367-1373.
- LAFOURCADE M. (2007). Making people play for Lexical Acquisition. Proceedings of *7th Symposium on Natural Language Processing*, Pattaya, Thailand, 13-15 December 2007, 8 p.
- LAFOURCADE M., LE BRUN N., ET JOUBERT A. (2015). *Jeux et intelligence collective – résolution de problèmes et acquisition de données sur le Web*. Collection Science cognitive et management des connaissances (sous la direction de Joseph Mariani et Patrick Paroubek), ISTE éditions, 2015, 156 p.
- LAFOURCADE M., AND FORT K. (2014). Propa-L: a semantic filtering service from a lexical network created using Games With A Purpose. Proceedings of the Ninth International *Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Islande, 26-31 May 2014, pp. 1676-1681.
- SAIF, M., AND TURNEY, P. (2013). Crowdsourcing a Word-Emotion Association Lexicon. In *Computational Intelligence*, 29 (3), pp. 436-465, 2013.
- STRAPPARAVA C., AND VALITUTTI A. (2004). WordNet Affect : an affective extension of WordNet. Proceedings 4th International *Conference on Language Resources and Evaluation (LREC-04)*, Lisbon, Portugal, pp. 1083-1086.
- TABOADA M., BROOKE J., TOFILOSKI M., VOLL K., AND STEDE M. (2011). Lexicon-based methods for sentiment analysis. In *Computational Linguistics*, Volume 37 (2), pp. 267-307.
- TURNEY P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of *ACL-02*, Philadelphia, USA, pp. 417-424.

Étude des verbes introducteurs de noms de médicaments dans les forums de santé

François Morlane-Hondère¹ Cyril Grouin¹ Pierre Zweigenbaum¹

(1) LIMSI-CNRS, UPR 3251, rue John von Neumann, 91400 Orsay

{prenom.nom}@limsi.fr

Résumé. Dans cet article, nous combinons annotations manuelle et automatique pour identifier les verbes utilisés pour introduire un médicament dans les messages sur les forums de santé. Cette information est notamment utile pour identifier la relation entre un médicament et un effet secondaire. La mention d'un médicament dans un message ne garantit pas que l'utilisateur a pris ce traitement mais qu'il effectue un retour. Nous montrons ensuite que ces verbes peuvent servir pour extraire automatiquement des variantes de noms de médicaments. Nous estimons que l'analyse de ces variantes pourrait permettre de modéliser les erreurs faites par les usagers des forums lorsqu'ils écrivent les noms de médicaments, et améliorer en conséquence les systèmes de recherche d'information.

Abstract.

Study of Drug-Introducing Verbs on Health Forums

In this paper, we combine manual/automatic annotation to identify the verbs used by the users of a health forum to say that they are taking a drug. This information is important in many aspects, one of them being the identification of the relation between drugs and side effects : the mere mention of a drug in a message is not enough to assess that the user is taking this drug, and is thus likely to provide a feedback on it. In a second part of the study, we show how the set of verbs that we identified can be used to automatically extract variants of drug names. We assume that the analysis of the variants could shed light on patterns of mistakes that users make when spelling drug names and thus, improve medical information retrieval systems.

Mots-clés : contenu généré par l'utilisateur, forum, verbes, noms de médicaments.

Keywords: user-generated content, forum, verbs, drug names.

1 Introduction et état de l'art

Cette étude s'inscrit dans le cadre d'un projet visant l'identification d'effets secondaires de médicaments dans des textes produits par des utilisateurs sur des forums de discussion. Il s'agit d'une problématique récente née du besoin de surveiller les médicaments après leur mise sur le marché et du développement des forums de santé en ligne.

L'autorisation de mise sur le marché d'un médicament est soumise à une batterie d'essais cliniques qui permettent d'évaluer le rapport entre effets bénéfiques et les éventuels effets néfastes. Si ce rapport est jugé acceptable, le médicament est commercialisé et les effets indésirables sont documentés dans sa notice. Toutefois, certains effets indésirables échappent aux tests cliniques et, dans les cas les plus graves, peuvent conduire à des crises sanitaires comme l'*affaire du Mediator* en France, qui n'était pas qu'une affaire politique.

En 2007, l'étude EMIR (Effets indésirables des Médicaments : Incidence et Risque) menée par le réseau des centres régionaux de pharmacovigilance a estimé que 3,6 % des admissions dans des hôpitaux français étaient dues à des effets indésirables de médicaments, soit 1 480 885 journées d'hospitalisation par an¹. Ces chiffres témoignent des enjeux humains et financiers considérables liés au processus de pharmacovigilance, qui consiste à détecter, identifier et gérer les effets indésirables de médicaments après leur commercialisation.

Les dispositifs mis en œuvre par les organismes de pharmacovigilance et les laboratoires pour permettre aux praticiens et aux patients de rapporter les effets indésirables se composent principalement de centres d'appel et de formulaires en

1. <http://www.sante.gouv.fr/IMG/pdf/EMIR.pdf>

ligne². Un rapport de 2014 de l'Académie nationale de Pharmacie montre toutefois que seuls 4 à 5 % des effets indésirables sont signalés de façon spontanée³. Cela est principalement dû à la méconnaissance des dispositifs de signalement de la part des utilisateurs de médicaments et professionnels de santé (Alshakka *et al.*, 2013).

Avec la mise au format électronique des données de santé, l'utilisation de méthodes d'extraction d'information apparaît alors comme une approche intéressante pour identifier les effets indésirables de médicaments par l'analyse de dossiers de patients (Trifirò *et al.*, 2009; Wang *et al.*, 2009; Gurulingappa *et al.*, 2012). Les résultats encourageants fournis par ces méthodes ont conduit à les appliquer sur les données produites par les utilisateurs de médicaments eux-mêmes, principalement sur les réseaux sociaux et forums de discussion (voir Sarker *et al.* (2015) pour un état de l'art). Ces données présentent plusieurs avantages par rapport à celles produites par des professionnels de santé dans le milieu médical : (i) elles sont facilement accessibles, (ii) elles sont massives, (iii) elles sont produites en continu et peuvent être soumises à un système de pharmacovigilance dès leur mise en ligne, ce qui permet une réactivité inédite (Egberts *et al.* (1996) rapportent un délai de 229 jours entre le signalement par téléphone et la production d'un rapport sur les effets indésirables sur la paroxétine par la Pharmacovigilance Foundation), et (iv) elles sont volontairement mises à disposition par les utilisateurs à destination de la communauté, ce qui facilite la question de leur confidentialité. Le principal inconvénient est lié au caractère non contrôlé des données, ce qui entraîne une variabilité orthographique et stylistique. Ceci complique l'utilisation d'outils d'analyse morpho-syntaxique traditionnels et la projection de lexiques d'entités (Nikfarjam & Gonzalez, 2011). Nous relevons par exemple dans nos données plus de neuf orthographes différentes pour le mot *anxiolytique* ainsi que de nombreuses classes d'équivalences comme *être agité/survolé/comme une pile électrique...*

Cette étude prospective, fondée sur une analyse manuelle, est envisagée comme une première étape dans l'identification automatique d'effets indésirables. Elle vise à donner des éléments de description linguistique de la façon dont les utilisateurs de forums de santé expriment le fait qu'ils prennent un médicament par l'analyse des verbes employés. Comme mis en évidence par Leaman *et al.* (2010), les verbes constituent des indices forts pour le repérage d'informations médicales (indications, effets bénéfiques ou indésirables...). Les verbes qui ont pour rôle d'introduire des noms de médicaments présentent donc un intérêt pour identifier des effets indésirables dans les textes : en plus d'indiquer la prise d'un médicament, ils donnent des indices sur la nature de ce médicament. Nous montrons également que ces verbes peuvent servir à identifier des variantes de noms de médicaments.

2 Méthode et données

La mention de l'utilisation d'un médicament peut être modélisée comme une construction composée de trois éléments : (i) un nom de médicament, (ii) un verbe introducteur de médicament (VIM) ; ce type de verbe est indispensable pour distinguer les cas où un utilisateur témoigne de son expérience avec un médicament ou demande des renseignements en vue de le prendre (Wu *et al.*, 2013), et (iii) un pronom de première personne singulier. Bien que seuls le verbe et l'objet soient suffisants pour identifier la mention d'une prise de médicament, nous restreignons le sujet au scripteur lui-même dans le but d'écarter les cas de témoignages indirects. Des modalités comme le dosage ou le rythme de prise du médicament peuvent se greffer à cette construction mais ne font pas partie de ses éléments essentiels. Puisqu'il existe des listes des noms de médicaments et que les pronoms de première personne singuliers se limitent à *je* et *me*, les VIMs sont les seuls éléments inconnus de cette construction. Cette section présente les approches manuelle puis automatique que nous avons adoptées pour les identifier, puis une expérience visant à montrer leur pertinence pour identifier les variantes de noms de médicaments.

2.1 Annotation manuelle des VIMs

Dans une approche préliminaire, nous avons manuellement annoté les VIMs dans un corpus de 11 735 mots constitué de messages extraits du forum de santé francophone Doctissimo⁴. Nous avons identifié douze verbes, dont dix ont une fréquence inférieure à 5. Afin d'estimer la productivité des verbes identifiés sur un plus gros volume de données, nous avons utilisé le moteur de recherche Google pour accéder à l'ensemble des données du forum Doctissimo. Nous avons construit 91 requêtes constituées d'un sous-ensemble de sept verbes parmi les douze identifiés précédemment et de treize noms de médicaments choisis parmi les plus prescrits en France. Ces requêtes sont construites selon les modalités suivantes :

2. https://www.formulaires.modernisation.gouv.fr/gf/cerfa_15031.do

3. http://www.acadpharm.org/dos_public/GTNotif_Patients_Rap_VF__2015.01.22.pdf

4. <http://forum.doctissimo.fr>

- le verbe est à la première personne du singulier du présent de l'indicatif, sauf dans le cas de *prescrire* et *donner*, qui sont au participe passé (l'utilisateur est celui à qui l'on donne un médicament) ;
- le verbe et le nom de médicament sont séparés par un astérisque, qui remplace un ou plusieurs mots dans la syntaxe de requêtes de Google ;
- la requête est encadrée de guillemets pour que l'ordre des éléments se retrouve à l'identique dans les textes.

Ces 91 requêtes ont été soumises manuellement à Google avec le paramètre `site:forum.doctissimo.fr` pour restreindre la recherche à tous les forums de Doctissimo. Bien que cette expérience ait fourni des résultats intéressants, le protocole mis en place reste limité (*i*) par le nombre de contextes et de médicaments testés, (*ii*) par l'utilisation de requêtes Google, qui ne permettent pas de prendre en compte les formes fléchies des verbes et (*iii*) par l'utilisation du nombre de pages retournées comme un substitut de la fréquence.

2.2 Extraction semi-automatique des VIMs

Cette deuxième approche vise à étendre la liste des douze VIMs identifiés manuellement par la projection de patrons sur un corpus de 17,5 millions de mots constitué de messages extraits du forum médical Atoute.org⁵. Elle consiste à chercher les VIMs dans les contextes dont nous faisons l'hypothèse qu'ils sont les plus susceptibles d'apparaître, à savoir entre un pronom de première personne singulier et un nom de médicament.

Cette approche se divise en trois étapes :

1. une liste de noms de médicaments est utilisée pour annoter automatiquement les noms de médicaments qui apparaissent dans les messages. Cette liste est composée de 4 ressources distinctes. La construction de cette liste prend en compte les versions accentuées et désaccentuées et ne tient pas compte de la casse. Cette liste contient :
 - les 8691 entités en français de l'UMLS appartenant au type sémantique *Pharmacologic substance* ;
 - 9064 noms de médicaments génériques fournis par l'Agence nationale de sécurité du médicament et des produits de santé (ANSM)⁶ ;
 - 10 870 noms de médicaments extraits du dictionnaire de médicaments en ligne EurekaSanté⁷ ;
 - la liste des 100 médicaments les plus prescrits en France⁸.
2. les séquences de n ($n \leq 6$) mots qui figurent entre un pronom de première personne singulier et un nom de médicament sont extraites ;
3. les verbes figurant dans ces séquences sont identifiés et lemmatisés par la projection de Glàff⁹, un lexique généraliste du français. Bien que fruste, cette approche se justifie par la nature de nos données, qui complique l'utilisation d'un étiqueteur morpho-syntaxique. Cette méthode reste néanmoins très bruitée.

La pertinence des verbes candidats identifiés a été évaluée manuellement.

2.3 Utiliser les VIMs pour extraire les variantes de noms de médicaments

Parce que les textes produits sur les forums sont non contrôlés et que les noms de traitements sont complexes, il est fréquent que les noms de médicaments soient mal orthographiés ou abrégés (Pimpalkhute *et al.*, 2014). Ces orthographes alternatives ne figurent pas dans les lexiques et posent un problème pour les systèmes d'extraction d'information.

Dans cette expérience, nous illustrons une utilisation possible des VIMs pour extraire des variantes de noms de médicaments. Le protocole utilisé est une déclinaison de celui décrit en section 2.2 : au lieu de chercher les verbes situés entre un pronom et un nom de médicament, nous ciblons les mots inconnus – définis comme n'apparaissant ni dans notre lexique de médicaments ni dans Glàff – situés après une séquence pronom+VIM. Nous faisons l'hypothèse que les VIMs sont quasi-systématiquement suivis de noms de médicaments. En conséquence, un mot inconnu qui apparaît dans une séquence de 1 à 6 mots après un VIM a plus de chances d'être la variante d'un nom de médicament que les mots inconnus qui apparaissent ailleurs dans le texte.

5. <http://www.atoute.org/n/forum/>

6. http://ansm.sante.fr/var/ansm_site/storage/original/text/97b3c42da571c69da1e837f759076675.txt

7. <http://www.eurekasante.fr/medicaments/alphabetique.html>

8. <http://www.doctissimo.fr/asp/medicaments/les-medicaments-les-plus-prescrits.htm>

9. <http://redac.univ-tlse2.fr/lexiques/glafl.html>

3 Résultats

L'étape d'annotation manuelle nous a permis d'identifier les douze VIMs suivants (la fréquence du verbe apparaît entre parenthèses) : *prendre* (37), *prescrire* (19), *être sous* (4), *passer* (4), *donner* (3), *commencer* (3), *aval* (1), *absorber* (1), *entamer* (1), *suivre* (1), *tester* (1) et *utiliser* (1). Le nombre de pages Web indiqué par Google après la soumission des requêtes contenant 7 de ces VIMs et 13 noms de médicaments est rapporté au tableau 1. La proportion de pages rapportées pour chaque VIM et chaque contexte est fournie dans le tableau 2

	Doliprane	Levothyrox	Kardégic	Spasfon	Tahor	Voltaire	Forlax	Subutex	Gaviscon	Lexomil	Lysanxia	Atarax	Xanax	total
<i>prendre</i>	6350	1550	71	35700	55	36	319	56	4570	806	576	365	1340	51 794
<i>prescrire</i>	2070	334	27	7270	23	64	510	35	5040	750	426	824	1570	18 943
<i>donner</i>	3730	86	17	8570	12	34	619	8	1560	414	74	1190	673	16 987
<i>être sous</i>	80	1340	202	6390	24	30	47	75	61	536	177	93	811	9866
<i>aval</i>	1787	0	0	35	0	1	0	0	12	9	2	5	20	1871
<i>commencer</i>	8	89	5	28	1	0	18	10	63	12	6	4	26	270
<i>passer</i>	35	21	5	33	4	3	15	6	37	1	7	1	18	186
total	16 600	5080	327	58 026	119	168	1528	190	11 343	2528	1268	2482	4458	

TABLE 1 – Nombre de pages rapportées par Google pour chaque requête.

	Doliprane	Levothyrox	Kardégic	Spasfon	Tahor	Voltaire	Forlax	Subutex	Gaviscon	Lexomil	Lysanxia	Atarax	Xanax	moyenne
<i>prendre</i>	38.3	30.5	20.7	58.4	40.4	20.1	20.6	22.3	40.1	31.1	43	14.6	25.5	31.2
<i>prescrire</i>	12.5	6.6	7.9	11.9	16.9	35.8	32.9	13.9	44.3	28.9	31.8	32.9	29.9	23.6
<i>être sous</i>	0.5	26.4	58.9	10.4	17.6	16.8	3	29.9	0.5	20.7	13.2	3.7	15.4	16.7
<i>donner</i>	22.5	1.7	5	14	8.8	19	40	3.2	13.7	16	5.5	47.5	12.8	16.1
<i>aval</i>	10.8	0	0	0.1	0	0.6	0	0	0.1	0.3	0.1	0.2	0.4	1.0
<i>passer</i>	0.2	0.4	1.5	0.1	2.9	1.7	1	2.4	0.3	0	0.5	0	0.3	0.9
<i>commencer</i>	0	1.8	1.5	0	0.7	0	1.2	4	0.6	0.5	0.4	0.2	0.5	0.9

TABLE 2 – Proportion de pages rapportées pour chaque VIM et chaque nom de médicament.

La méthode d'extraction de VIMs (cf. section 2.2) nous permet d'identifier 28 934 occurrences de 1053 verbes candidats parmi lesquels 44 ont été manuellement identifiés comme des VIMs. Le tableau 3 présente ces verbes et leur fréquence.

freq.	verb	freq.	verb	freq.	verb	freq.	verb
2953	<i>prendre</i>	37	<i>passer</i>	13	<i>manger</i>	5	<i>repasser</i>
1570	<i>prescrire</i>	33	<i>continuer</i>	12	<i>aval</i>	4	<i>vacciner</i>
842	<i>être sous</i>	31	<i>appliquer</i>	11	<i>(se) soigner</i>	4	<i>refiler</i>
764	<i>donner</i>	25	<i>tester</i>	10	<i>consommer</i>	4	<i>recommencer</i>
400	<i>avoir</i>	24	<i>injecter</i>	9	<i>rajouter</i>	2	<i>sniffer</i>
296	<i>mettre</i>	20	<i>remettre</i>	9	<i>administrer</i>	2	<i>(se) droguer</i>
181	<i>commencer</i>	16	<i>tenter</i>	8	<i>(se) gaver</i>	2	<i>effectuer</i>
140	<i>essayer</i>	15	<i>représcrire</i>	7	<i>(se) badigeonner</i>	2	<i>absorber</i>
133	<i>utiliser</i>	15	<i>filer</i>	7	<i>baisser</i>	1	<i>bouffer</i>
122	<i>reprandre</i>	15	<i>diminuer</i>	6	<i>boire</i>	1	<i>bénéficier</i>
115	<i>suivre</i>	15	<i>augmenter</i>	5	<i>sucer</i>	1	<i>abuser</i>

TABLE 3 – Fréquence des 44 VIMs identifiés dans le corpus Atoute.org.

4 Discussion

Emploi des verbes La première approche consistant à annoter manuellement les VIMs dans un petit corpus a montré que 61 % des mentions de prise d'un médicament se font en employant les verbes *prendre* et *prescrire* (bien que ce verbe n'exprime pas la prise d'un médicament, les données montrent que cette dernière est très souvent sous-entendue). Les résultats fournis par l'approche basée sur la construction de requêtes Google montrent que les verbes *prendre*, *prescrire*, *donner* et *être sous* prévalent aussi bien en terme de nombre de pages rapportées que de proportion : sur l'ensemble des pages rapportées par le moteur de recherche, 98 % l'ont été par une requête contenant l'un de ces quatre VIMs. On constate également que le type de verbe employé pour évoquer la prise d'un médicament varie en fonction du médicament :

- Lexomil, Lysanxia, Atarax et Xanax, qui appartiennent à la classe des benzodiazépines, sont souvent employés avec le verbe *prescrire*. Ce n'est pas le cas de Levothyrox et Kardégic, ce qui peut s'expliquer par le fait que ce sont des traitements à vie : on peut alors supposer que l'action de prescrire ce type de traitement est moins fréquente que pour d'autres médicaments, et que cela se répercute dans les messages ;
- Doliprane, Gaviscon et Forlax ne s'emploient que très rarement avec *être sous*. On peut supposer que cela est dû au fait que ces médicaments s'utilisent de façon ponctuelle pour traiter des affections passagères. Le fait que Levothyrox et – surtout – Kardégic sont souvent introduits par cette locution verbale va dans le sens de cette hypothèse ;
- le verbe *avalier* s'emploie quasi-exclusivement avec Doliprane. L'emploi de ce verbe relativement familier est peut-être à mettre en relation avec le fait que le Doliprane est un médicament très répandu et considéré comme inoffensif, quand bien même aucun médicament n'est inoffensif. Ce VIM est également particulier en cela qu'il contient une information sur la façon dont le médicament est pris, donc sur sa forme galénique : il paraît improbable qu'un médicament sous forme de crème soit introduit par ce VIM.

Ces résultats mettent ainsi en lumière certaines restrictions distributionnelles imposées par les VIMs, qui sélectionnent certains types de médicaments en fonction de propriétés liées à leur fréquence de prise ou leur forme galénique. Comme le montre le tableau 3, la prévalence des 4 VIMs identifiés manuellement est confirmée par les données obtenues sur le corpus Atoute.org. De nouveaux VIMs particulièrement fréquents ont pu être identifiés, comme *mettre*, qui s'emploie soit avec un nom de médicament – principalement une lotion ou un crème – en COD, soit avec la préposition *sous*. La haute fréquence du verbe *avoir* est une erreur due à la méthode d'identification des verbes employée. Du fait de son caractère hautement polysémique, il apparaît dans une grande variété de contextes, mais n'est que rarement un VIM.

Informations complémentaires Une des caractéristiques des VIMs les plus fréquents est qu'ils sont relativement *neutres* : le fait qu'ils imposent peu de restrictions sélectionnelles sur le type de médicament qu'ils prennent comme objets entraîne leur utilisation pour l'introduction d'une grande variété de médicaments, ce qui explique leur fréquence. À l'inverse, d'autres VIMs véhiculent une connotation qui réduit le spectre des médicaments avec lesquels ils peuvent s'employer, et donc leur fréquence générale dans les textes. On peut distinguer 4 groupes parmi ces VIMs : (i) *commencer*, *essayer*, *passer*, *continuer*, *tester*, *tenter*, *repasser*, *recommencer* fournissent des informations temporelles sur le traitement ; (ii) *diminuer*, *augmenter*, *baisser* indiquent l'évolution du dosage ; (iii) *manger*, *(se) gaver*, *(se) droguer*, *bouffer*, *abuser* apportent des précisions sur la quantité et la fréquence de prise du médicament. Ils renseignent également du point de vue du scripteur qui porte un jugement négatif sur la quantité de médicaments prise, jugée excessive ; (iv) *injecter*, *avalier*, *(se) badigeonner*, *boire*, *sucer*, *sniffer* informent sur la forme galénique du médicament. En plus d'indiquer la prise d'un médicament, ces VIMs apportent des informations complémentaires utiles pour identifier des effets indésirables.

Identification de variantes Un des problèmes liés aux données générées par des utilisateurs est la fréquence d'entités nommées mal orthographiées. Il est possible d'identifier ces variantes en rapprochant leur différentes occurrences sur la base de leur structure phonétique ou à l'aide de la distance d'édition. De manière similaire aux techniques de clustering non supervisées qui se fondent sur l'analyse du contexte, nous proposons ici d'utiliser les VIMs en les projetant sur le corpus Atoute.org dans le but de recueillir des variantes de noms de médicaments. Cette méthode nous a permis de recueillir 5 638 occurrences de 2 769 candidats. Ces derniers peuvent être classés dans cinq catégories :

- nom de médicament
 1. orthographe officielle : bien que correctement orthographiés, ces noms n'apparaissent pas dans notre liste. Nos lexiques ne contiennent aucun nom de produits paramédicaux (tels que les compléments alimentaires) et se limitent aux médicaments prescrits en France (certains médicaments sont évoqués sous leur nom commercial canadien par des utilisateurs québécois) ;
 - orthographe non officielle
 2. variante intentionnelle : nous considérons comme des variantes intentionnelles les mots dérivés des noms offi-

- ciels par des procédés morphologiques tels que l'apocope (*lévo* pour *Lévothyrox*), la reduplication (*dudu* pour *Duphaston*) ou encore la siglaison (*pdl* pour *pilule du lendemain*) ;
3. variante fautive : nous considérons comme des fautes les variantes de noms de médicaments qui ne semblent pas résulter d'un procédé de formation morphologique volontaire comme ceux évoqués précédemment. Nous avons distingué ces variantes selon les quatre opérations utilisées pour mesurer la distance de Damerau-Levenshtein (cf. tableau 4). Les erreurs d'orthographe n'ont pas été distinguées des éventuelles fautes de frappe ;
 - autre type de nom
 4. domaine médical : mots du lexique médical mal orthographiés ou n'apparaissant pas dans nos lexiques (variantes *amniosynthese* et *amnio* pour *amniocentèse*) ;
 5. domaine non médical : mots du vocabulaire général mal orthographiés ou n'apparaissant pas dans Gläff.

effacement	substitution	insertion	inversion	combinaison
alocardyl (Avlocardyl)	allupirinol (Allopurinol)	corguard (Corgard)	pantesa (Pentasa)	calements (calmants)
lévotyrox (Lévothyrox)	anxiolitique (anxiolytique)	cortisonne (cortisone)	procolaran (Procoralan)	cylbatant (Cymbalta ?)
luthenyl (Lutényl)	anxyolitique (anxiolytique)	dacktarin (Daktarin)	steridil (Stediril)	rhénomicine (Rinomicine)
stomectol (Stromectol)	celccept (Cellcept)	duphastion (Duphaston)		doxycilline (Doxycycline)
utrogestant (Utrogestan)	cortencil (Cortancyl)	endoxant (Endoxan)		methojet (Metoject)

TABLE 4 – Les différentes catégories de variantes fautives (les noms officiels sont entre parenthèses). La colonne combinaison contient des variations produites par la combinaisons de plusieurs opérations.

La distinction des variantes extraites permet d'observer que les variantes intentionnelles relèvent de procédés de création réguliers, contrairement aux variantes fautives. Parmi les variantes fautives, nous observons des tendances telles que les substitutions récurrentes *i/y* ou *s/z*, le rajout d'un *t* aux noms de médicaments terminés par la séquence *-an*, ou encore des confusions dans les noms contenant des doubles lettres (*Xyzaal* pour *Xyzall*). Bien qu'une analyse phonétique permettrait d'identifier la majorité de ces variantes, il reste des cas pour lesquels une analyse plus poussée demeure nécessaire (*steridil* vs *Stediril*). En conséquence, nous estimons que l'analyse des procédés de formation des variantes intentionnelles et fautives pourrait permettre de prédire les erreurs faites par les utilisateurs et ainsi d'améliorer les performances de systèmes d'identification des effets indésirables de médicaments.

5 Conclusion

Dans cet article, nous avons présenté l'étude réalisée pour identifier un ensemble de verbes introduisant des noms de médicaments dans un corpus de messages postés sur un forum de santé. Nous avons mis en évidence que les verbes sont utilisés différemment selon le médicament qu'ils introduisent. Nous avons montré que cet ensemble de verbes peut être utilisé dans des règles pour extraire automatiquement des variantes de noms de médicaments. Enfin, nous estimons que ces propriétés sont utiles pour identifier certaines catégories de médicaments, ou pour extraire des relations entre médicaments et effets secondaires.

Remerciements

Ce travail a été réalisé dans le cadre du projet Vigi4MED (ANSM-2013-S-060), financé par l'ANSM (Agence Nationale de Sécurité du Médicament).

Références

- ALSHAKKA M. A., IBRAHIM M. I. M. & HASSALI M. A. A. (2013). Do health professionals have positive perception towards consumer reporting of adverse drug reactions ? *J Clin Diagn Res*, 7(10), 2181–5.
- EGBERTS T. C., SMULDERS M., DE KONING F. H., MEYBOOM R. H. & LEUFLENS H. G. (1996). Can adverse drug reactions be detected earlier ? a comparison of reports by patients and professionals. *Br Med J*, 313(7056), 530–1.
- GURULINGAPPA H., MATEEN-RAJPUT A. & TOLDO L. (2012). Extraction of potential adverse drug events from medical case reports. *Journal of Biomedical Semantics*, 3(1).

- LEAMAN R., WOJTULEWICZ L., SULLIVAN R., SKARIAH A., YANG J. & GONZALEZ G. (2010). Towards internet-age pharmacovigilance : Extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, BioNLP '10, p. 117–125, Stroudsburg, PA, USA : Association for Computational Linguistics.
- NIKFARJAM A. & GONZALEZ G. H. (2011). Pattern mining for extraction of mentions of adverse drug reactions from user comments. *AMIA Annual Symposium Proceedings*, **2011**, 1019.
- PIMPALKHUTE P., PATKI A., NIKFARJAM A. & GONZALEZ G. (2014). Phonetic spelling filter for keyword selection in drug mention mining from social media. *AMIA Summits on Translational Science Proceedings*, **2014**, 90.
- SARKER A., NIKFARJAM A., O'CONNOR K., GINN R., GONZALEZ G., UPADHAYA T., JAYARAMAN S. & SMITH K. (2015). Utilizing social media data for pharmacovigilance : A review. *Journal of Biomedical Informatics*, (0), –.
- TRIFIRÒ G., PARIENTE A., COLOMA P. M., KORS J. A., POLIMENI G., MIREMONT-SALAMÉ G., CATANIA M. A. A., SALVO F., DAVID A., MOORE N., CAPUTI A. P. P., STURKENBOOM M., MOLOKHIA M., HIPPISELEY-COX J., ACEDO C. D. D., VAN DER LEI J., FOURRIER-REGLAT A. & EU-ADR GROUP (2009). Data mining on electronic health record databases for signal detection in pharmacovigilance : which events to monitor ? *Pharmacoepidemiology and drug safety*, **18**(12), 1176–1184.
- WANG X., HRIPCSAK G., MARKATOU M. & FRIEDMAN C. (2009). Research paper : Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records : A feasibility study. *JAMIA*, **16**(3), 328–337.
- WU H., FANG H. & STANHOPE S. J. (2013). Exploiting online discussions to discover unrecognized drug side effects. *Methods Inf Med*, **52**(2), 152–9.

Initialisation de Réseaux de Neurons à l'aide d'un Espace Thématique

Mohamed Morchid, Richard Dufour, Georges Linarès
Laboratoire Informatique d'Avignon
prénom.nom@univ-avignon.fr

Résumé. Ce papier présente une méthode de traitement de documents parlés intégrant une représentation fondée sur un espace thématique dans un réseau de neurones artificiels (ANN) employé comme classifieur de document. La méthode proposée consiste à configurer la topologie d'un ANN ainsi que d'initialiser les connexions de celui-ci à l'aide des espaces thématiques appris précédemment. Il est attendu que l'initialisation fondée sur les probabilités thématiques permette d'optimiser le processus d'optimisation des poids du réseau ainsi qu'à accélérer la phase d'apprentissage tout en améliorant la précision de la classification d'un document de test. Cette méthode est évaluée lors d'une tâche de catégorisation de dialogues parlés entre des utilisateurs et des agents du service d'appels de la Régie Autonome Des Transports Parisiens (RATP). Les résultats montrent l'intérêt de la méthode proposée d'initialisation d'un réseau, avec un gain observé de plus de 4 points en termes de bonne classification comparativement à l'initialisation aléatoire. De plus, les expérimentations soulignent que les performances sont faiblement dépendantes de la topologie du ANN lorsque les poids de la couche cachée sont initialisés au moyen des espaces de thèmes issus d'une allocation latente de Dirichlet ou *latent Dirichlet Allocation* (LDA) en comparaison à une initialisation empirique.

Abstract.

Neural Network Initialization using a Topic Space

This paper presents a method for speech analytics that integrates topic-space based representation into an artificial neural network (ANN), working as a document classifier. The proposed method consists in configuring the ANN's topology and in initializing the weights according to a previously estimated topic-space. Setup based on thematic priors is expected to improve the efficiency of the ANN's weight optimization process, while speeding-up the training process and improving the classification accuracy. This method is evaluated on a spoken dialogue categorization task which is composed of customer-agent dialogues from the call-centre of Paris Public Transportation Company. Results show the interest of the proposed setup method, with a gain of more than 4 points in terms of classification accuracy, compared to the baseline. Moreover, experiments highlight that performance is weakly dependent to ANN's topology with the LDA-based configuration, in comparison to classical empirical setup.

Mots-clés : Réseau de neurones artificiels, Allocation latente de Dirichlet, Initialisation de poids.

Keywords: Artificial neural network, Latent Dirichlet allocation, Weights initialization.

1 Introduction

Plusieurs méthodes d'analyse de documents parlés projettent leurs transcriptions automatiques dans un espace de thèmes¹ obtenu par le biais d'un apprentissage non-supervisé sur des corpus de documents de grande taille. L'objectif de cette projection est d'abstraire la représentation de surface des termes composant le document transcrit car ceux-ci peuvent rendre l'analyse directe des documents difficile (erreurs de transcription, disfluences...). Le module d'analyse de documents opère alors dans ces espaces thématiques lors de tâches de classification ou d'identification. Souvent, la représentation du contenu du document et le module d'analyse sont traités ou optimisés indépendamment : l'espace de représentation est conçu pour être le plus expressif et le plus compact possible, alors que le module d'analyse est optimisé par la fonction objective de la tâche finale. Dans ce travail, nous proposons une approche holistique où les espaces de thèmes et le système d'analyse de contenu sont optimisés conjointement. Les réseaux de neurones artificiels (ANN) sont maintenant une approche standard pour le traitement de la langue mais nécessitent un processus lourd et coûteux pour l'estimation des paramètres lors de la phase d'apprentissage. Cette difficulté d'élaboration de l'architecture du ANN est essentiellement

1. Il sera par la suite appelé "thème" un ensemble de termes regroupés dans une même classe dans l'espace LDA.

due au fait que la phase d'apprentissage est un processus d'optimisation stochastique dépendant de plusieurs facteurs comme la distribution des termes dans le corpus d'apprentissage ou les conditions d'initialisation du réseau. Le choix de la configuration initiale du réseau est un point crucial lors de la phase d'apprentissage (poids des couches cachées, topologie...). Ce choix peut considérablement affecter le temps d'apprentissage (Adhikari & Joshi, 1956) ainsi que les performances du ANN (optimum local). Plusieurs études adaptent le momentum et le taux d'apprentissage pour accélérer l'apprentissage (Beale, 1972; Møller, 1993; Powell, 1977; Nguyen & Widrow, 1990; Drago & Ridella, 1992; Thimm & Fiesler, 1997) ou se concentrent sur l'initialisation des poids ou du biais en appliquant un pré-traitement fondé sur l'analyse de données ou des méthodes de classification (Breukelen & Duin, 1998; Kathirvalavakumar & Subavathi, 2011; Dahl *et al.*, 2012).

Dans ce papier, nous proposons une méthode d'initialisation d'un ANN évaluée lors d'une tâche d'identification de catégories de conversations transcrites automatiquement et issues de la RATP (Bechet *et al.*, 2012) potentiellement bruitées. Afin de gérer cette difficulté, un espace de thèmes permettant une abstraction des transcriptions en sortie du système de reconnaissance automatique de la parole (SRAP) est utilisé. Dans un schéma classique, la classification devrait s'opérer dans ces espaces thématiques. Ici, nous étudions l'impact de notre méthode d'initialisation d'un réseau de neurones (ANN) s'appuyant sur un espace de thèmes issu d'une allocation latente de Dirichlet (LDA). Dans un premier temps, nous comparons différentes entrées du ANN en utilisant la fréquence des termes de la conversation puis les probabilités issues d'espaces de thèmes. Nous proposons ensuite d'évaluer différentes initialisations des poids de la couche cachée d'un ANN : une initialisation aléatoire suivant une loi uniforme classique et notre initialisation originale au moyen des probabilités estimées avec une LDA.

La partie 2 présente les études précédentes liées à la représentation de documents ainsi qu'aux méthodes d'initialisation des ANN. Les concepts de base d'un ANN ainsi que les caractéristiques thématiques sont décrits dans la partie 3. La partie 4 présente les expériences ainsi que les résultats avant de conclure dans la partie 5.

2 Travaux antérieurs

L'approche classique à base de fréquences de mots TF-IDF (Robertson, 2004), a été très largement utilisée afin d'extraire les mots discriminants contenus dans des textes. D'autres approches ont proposé de considérer le document comme un mélange de thèmes cachés telles que *Latent Semantic Analysis* (LSA) (Deerwester *et al.*, 1990) ou encore *Latent Dirichlet Allocation* (LDA) (Blei *et al.*, 2003), permettant un niveau de représentation plus élevé. Les performances de ces méthodes ont pu être démontrées sur de nombreuses tâches. En particulier, dans l'approche LDA, un thème est associé à chacune des occurrences d'un terme contenu dans le document (et non un thème pour l'ensemble du document). Il en résulte que les thèmes appartenant à un document peuvent changer d'un terme à l'autre. Dans ce papier, les probabilités des thèmes cachés, estimées avec une LDA, capturent les dépendances possibles entre les termes pour permettre de modéliser le contenu sémantique d'une conversation donnée. Les réseaux de neurones (ANN) constituent un environnement standard aujourd'hui pour des tâches de classification ou de prédiction. L'un des modèles les plus populaires est le *feed-forward multilayer perceptron*, habituellement entraîné à l'aide de l'algorithme de rétro-propagation du gradient ou une de ses nombreuses variantes (Cazorla & Escolano, 2003; Hagan *et al.*, 1996). La rétro-propagation est une technique d'optimisation de descente du gradient offrant des propriétés de convergence rapide mais qui est fortement dépendante des conditions d'initialisation souvent choisies empiriquement. Ce problème est d'actualité et est abordé par plusieurs chercheurs. (Feuring, 1996) propose d'employer l'algorithme de rétro-propagation pour calculer les bornes de l'intervalle des valeurs potentiellement optimales pour les poids d'un ANN pour une tâche donnée (Draghici, 2002). Ensuite, le ANN doit résoudre ce problème avec des poids dont la valeur est un entier dans cet intervalle. Généralement, la plupart des méthodes proposées reposent sur l'analyse de données, des méthodes d'apprentissage automatique ou sur des connaissances *a priori* (Kathirvalavakumar & Subavathi, 2011; Dahl *et al.*, 2012).

3 Approche proposée

Un réseau de neurones (ANN) ou *feed-forward neural network* est composé de trois couches comme présenté dans la figure 1 : une couche d'entrée (x), une ou plusieurs couche(s) cachée(s) (θ) et une couche de sortie (y). Un ANN contient une couche cachée totalement connectée aux couches d'entrée et de sortie dans ce papier.

La première expérimentation consiste à évaluer l'impact de différents jeux de caractéristiques d'entrée d'un ANN issus de la fréquence classique des termes et issus d'espaces thématiques (voir partie 3.2). Le nombre de neurones contenus dans la couche d'entrée (x) correspond au nombre de caractéristiques (*i.e.* nombre de termes ou nombre de thèmes LDA). La seconde expérimentation cherche à évaluer l'impact, dans une tâche de catégorisation, de l'initialisation des poids de la

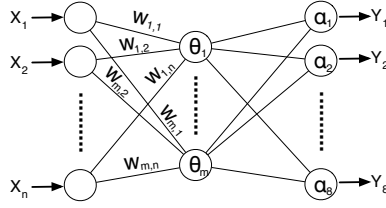


FIGURE 1 – Exemple d’architecture d’un ANN.

couche cachée soit aléatoirement, soit au moyen de probabilités estimées durant une LDA. Dans ce schéma, les neurones de la couche d’entrée représentent le vocabulaire et chacun des neurones de la couche cachée représente un thème LDA, les entrées de la couche cachée étant initialisées avec les probabilités de ces thèmes sachant les termes représentés par les neurones de la couche d’entrée. Ensuite, l’algorithme d’apprentissage reposant sur la rétro-propagation du gradient est réalisé. Cette dernière étape peut être vue comme un optimisation conjointe de la représentation de la couche thématique et de discrimination de la catégorie (*i.e.* classe) à associer à la conversation.

3.1 Concepts de base d’un réseau de neurones artificiels (ANN)

3.1.1 Fonction d’activation

La fonction d’activation utilisée durant les expérimentations est la fonction classique de *tangente hyperbolique* :

$$\alpha(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (1)$$

Plus d’informations concernant les fonctions de transfert en général sont disponibles dans (Duch & Jankowski, 1999).

3.1.2 Algorithme d’apprentissage du *feed-forward*

Trois étapes sont nécessaires : calcul de la sortie, rétro-propagation de l’erreur et mise-à-jour des poids et biais.

Phase de calcul des sorties

Soit N_l le nombre de neurones contenus dans la couche l ($1 \leq l \leq M$) et M le nombre de couches du ANN. $\theta_{n,l}$ est le biais du neurone n ($1 \leq n \leq N_l$) de la couche l . Soit un ensemble de p exemples d’entrée x_i ($1 \leq i \leq p$) et un ensemble de classes y_i associées à chacun des x_i . La sortie $\gamma_{n,l}$ du neurone n de la couche l (voir figure 1) est donnée par :

$$\gamma_{n,l} = \alpha_{n,l} = \alpha\left(\sum_{m=0}^{N_{l-1}} w_{nm}^l \times \gamma_{m,l-1}\right) + \theta_{n,l} \quad (2)$$

Phase d’apprentissage

L’erreur e observée entre la sortie attendue y et le résultat de la phase de calcul des sorties γ est évaluée comme suit :

$$e_n^l = y_n - \gamma_{n,M} \quad (3)$$

$$e_n^l = \sum_{m=1}^{N_{l+1}} w_{m,n} \times \delta_{m,l+1} \quad (4)$$

pour respectivement la couche de sortie M (3) et les couches cachées (4). Le gradient δ est calculé ainsi : $\delta_{n,l} = e_n^l \times \alpha_{n,l}$

Phase de mise-à-jour

Lorsque les erreurs entre la sortie attendue et le résultat sont calculées, les poids $w_{n,m}^l$ et les biais $\theta_{n,l}$ doivent alors être respectivement mis-à-jour comme $w_{n,m}^{l*}$ et $\theta_{n,l}^*$:

$$w_{n,m}^{l*} = w_{n,m}^l + \epsilon \delta_{n,l} \times \alpha_{n,l} \quad (5)$$

$$\theta_{n,l}^* = \theta_{n,l} + \epsilon \delta_{n,l} \quad (6)$$

3.2 Caractéristiques issues du document pour l'ANN

La méthode proposée d'initialisation du ANN est évaluée lors d'une tâche de catégorisation de conversations issues du corpus du projet DECODA (Bechet *et al.*, 2012). Un ANN a besoin d'un ensemble de caractéristiques en entrée x_i et de classes (*i.e.* catégories) y_i associées à un dialogue en sortie. Deux représentations différentes du document fondées respectivement sur la fréquence classique des termes discriminants contenus dans le document, et une représentation plus abstraite issue d'un espace de thèmes LDA, sont présentées dans les parties suivantes.

Fréquence de termes discriminants

Un ensemble de mots discriminants \mathbf{V} de taille 166 est composé avec le critère de TF-IDF-Gini. Pour chacun des dialogues d , un ensemble de caractéristiques x^d est déterminé. La $k^{\text{ème}}$ caractéristique x_k^d est composée du nombre d'occurrences du mot t_k ($|t_k|$) dans d et le score Δ de t_n dans la liste de termes discriminants \mathbf{V} : $x_k^d = |t_k| \times \Delta(t_k)$

Espace de thèmes LDA

L'échantillonnage de Gibbs, présenté dans (Griffiths & Steyvers, 2004), est utilisé pour estimer les paramètres LDA et pour représenter un nouveau document dans l'espace des thèmes r de taille T . Ce modèle extrait un ensemble de caractéristiques de d depuis l'espace de représentation en thèmes. La $k^{\text{ème}}$ caractéristique est composée comme suit : $x_k^d = \theta_{k,d}^r$, où $\theta_{k,d}^r = P(z_k^r | d)$ est la probabilité du thème z_k^r ($1 \leq k \leq T$) soit généré par le dialogue d dans l'espace de thème r^{th} .

4 Expériences

4.1 Protocole expérimental

Les expériences sur l'identification du thème d'une conversation sont menées sur le corpus du projet DECODA (Bechet *et al.*, 2012). Ce corpus est composé de 1 242 conversations téléphoniques (environ 74 heures de signal) découpées en un corpus d'apprentissage (740 dialogues), un corpus de développement (175 dialogues) et un corpus de test (327 dialogues). Ces dialogues ont été manuellement annotés selon 8 thèmes : *problème d'itinéraire*, *objet perdu et trouvé*, *horaire*, *carte de transport*, *état du trafic*, *prix du ticket*, *infraction* et *offre spéciale*. Les expérimentations conduites utilisent des transcriptions automatiques issues d'un système de reconnaissance automatique de la parole (SRAP) décrit dans (Morchid *et al.*, 2014) et obtenues avec le système Speeral (Linarès *et al.*, 2007). Enfin, le processus de validation croisée (apprentissage sur le corpus d'entraînement et validation à chacune des itérations avec l'ensemble de développement) est employé pour trouver la meilleure configuration (*i.e.* nombre d'itérations).

4.2 Résultats

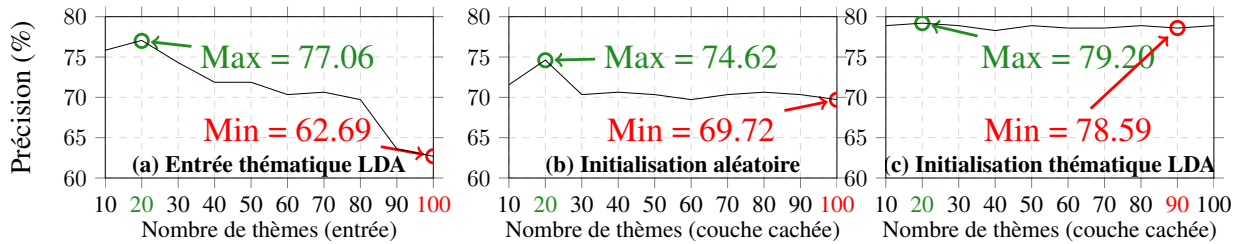


FIGURE 2 – Précision de la classification (%) en faisant varier le nombre de thèmes LDA en entrée du ANN (a), avec la couche cachée initialisée aléatoirement (b) et avec les poids de l'espace thématique (c).

Les premières expérimentations comparent deux ensembles de caractéristiques d'un document utilisant la représentation classique en fréquence de termes et utilisant un espace de thèmes LDA (voir parties suivantes). Ces représentations sont utilisées comme les entrées du ANN permettant l'apprentissage de celui-ci avec des documents textuels. Le ANN considéré est composé de trois couches : entrée (x issue de fréquences de termes ou de l'espace de thèmes LDA), cachée (1 couche de 8 neurones) et sortie (nombre de thèmes du corpus DECODA = 8). Les poids du ANN sont initialisés aléatoirement durant ces expérimentations initiales. Dans un second temps, nous comparons l'initialisation classique des poids w de la couche cachée à l'aide d'une variable aléatoire et notre méthode issue d'espaces de thèmes LDA. Les réseaux de neurones sont appris en utilisant, dans ce cas, la représentation thématique des conversations téléphoniques.

Entrée	# Neurones	#n	Précision
Fréquence de termes	8	X	75,84 %
LDA	20	8	77,06 %
Fréqu. de termes + initialisation LDA	20	20	79,20 %

TABLE 2 – Meilleures précisions lors de l’identification de thèmes.

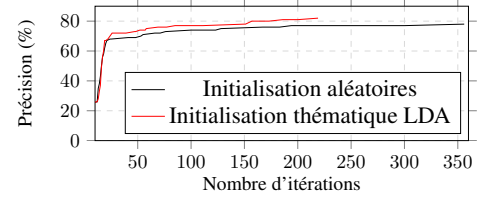


FIGURE 3 – Précision lors de la validation croisée (%) avec les deux approches d’initialisation de la couche cachée.

Comparaison des caractéristiques d’entrée du ANN

La première expérimentation utilisant les caractéristiques fondées sur la fréquence des termes permet d’obtenir une précision lors de la tâche de classification de 75.84 % (8 neurones dans la couche cachée). La seconde expérimentation utilise comme ensemble d’entrée les probabilités issues d’espaces de thèmes LDA. Sachant que la configuration des modèles LDA peut agir sur les performances de classification (Morchid *et al.*, 2014), nous proposons d’évaluer la performance du ANN en faisant varier le nombre de thèmes dans l’espace abstrait. La figure 2-(a) présente la précision obtenue avec différentes configurations de l’espace de thèmes (10 à 100 thèmes) au moyen de l’algorithme LDA, toujours avec 8 neurones dans la couche cachée. La première remarque est que la meilleure précision obtenue est de 77,06 % avec un gain possible de 1,22 points comparativement à la représentation fondée sur la fréquence de termes. Cependant, les résultats obtenus avec les probabilités issues de modèles LDA comme entrée sont instables et dépendent du nombre de thèmes dans l’espace (différence de 77,06 – 62,69 \simeq 15 points). La partie suivante tire un avantage des deux représentations, en utilisant les caractéristiques fondées sur la fréquence des termes (plus stable) comme entrée du ANN tout en initialisant le ANN à l’aide des probabilités issues de LDA.

Initialisation des poids

La partie précédente a comparé un ensemble de caractéristiques en entrée du ANN en considérant que les poids de la couche cachée sont initialisés aléatoirement. L’objectif des expérimentations suivantes est de résoudre le choix difficile des poids initiaux du ANN en utilisant comme entrée la représentation en fréquence de termes tout en initialisant les poids à l’aide des probabilités thématiques issues de LDA. La méthode originale d’initialisation est comparée à une initialisation aléatoire des poids. Pour ce faire, deux ANNs sont construits avec la même architecture : couche d’entrée composée de x_n neurones avec la fréquence de termes discriminants t_n (vocabulaire $V = 166$ mots discriminants, *i.e.* 166 neurones), couche cachée composée de $|T|$ neurones ($|T|$ = nombre de classes contenues dans l’espace de thèmes LDA $10 \leq |T| \leq 100$) et couche de sortie contenant 8 neurones (8 catégories dans le corpus DECODA). L’initialisation des poids de la couche cachée, fondée sur les thèmes LDA, consiste à considérer chacun des neurones de la couche cachée comme un thème LDA z_m . Ainsi, les poids sont considérés comme les probabilités thématiques du terme discriminant t_n sachant le thème représenté par le neurone dans la couche cachée z_m :

$$w_{m,n} = P(t_n|z_m) \quad (7)$$

La figure 2-(b) montre la précision obtenue avec une initialisation aléatoire alors que la figure 2-(c) présente les précisions obtenues avec une initialisation fondée sur les espaces de thèmes. En comparant ces deux courbes, nous constatons clairement que les résultats obtenus en initialisant des poids à l’aide de l’espace de thèmes LDA, sont meilleurs que ceux obtenus en initialisant les poids aléatoires, quel que soit le nombre de neurones contenus dans la couche cachée. En effet, la meilleure précision est de 74,62 %, obtenue pour une initialisation aléatoire, alors que l’initialisation à l’aide des probabilités thématiques LDA atteint une précision maximum de 79,20 % (gain de 4.58 points). Finalement, la méthode proposée (entrée=TF-IDF-Gini et poids=LDA) permet d’améliorer les performances du ANN en utilisant des caractéristiques d’entrée fondées sur les espaces de thèmes (entrée=LDA et poids=aléatoire) avec un gain de $79,20 - 77,06 = 2,14\%$ comme le montre le tableau 2. Les résultats présentés dans la figure 2-(b) sont également plus consistants (la différence entre la valeur minimum et maximum atteinte en termes de précision atteint 0,6 point) en comparaison avec la robustesse du réseau de neurones initialisé aléatoirement présenté dans la figure 2-(a) (différence de 4,9 points). Cette approche permet donc d’atteindre de meilleurs résultats qu’une initialisation classique aléatoire, mais plus important, élimine le choix difficile du nombre de neurones dans la couche cachée (résultats équivalents lorsque le nombre de neurones dans la couche cachée varie). La figure 3 présente la précision lors de la phase de validation croisée (ensemble de développement) avec une initialisation aléatoire et une initialisation fondée sur les espaces de thèmes LDA. Nous pouvons aisément constater que l’initialisation des poids à l’aide des probabilités thématiques permet d’obtenir une meilleure précision (78 % et 82 % pour respectivement une initialisation aléatoire et fondée sur l’espace LDA) avec un nombre d’itérations plus faible (356 et 219 itérations pour respectivement une initialisation aléatoire et fondée sur l’espace LDA). Un gain de 137 itérations

est alors observé, ce qui correspond à un gain en termes de temps de traitement (apprentissage du ANN) de 38.5% .

5 Conclusion

Ce papier présente une configuration originale de poids initiaux d'un réseau de neurones artificiels (ANN) au moyen des probabilités issues d'un espace thématique LDA. Les expérimentations ont montré l'intérêt de l'utilisation de variables latentes (probabilités thématiques) pour initialiser les poids du réseau de neurones durant une tâche de classification. LDA fournit ainsi une représentation robuste et pertinente de contenus bruités durant la phase d'apprentissage, optimisée selon la fonction objective liée à la tâche de classification. Cette méthode obtient de meilleurs résultats qu'un schéma classique fondé sur une représentation thématique du document à l'aide de LDA suivie par une classification à l'aide d'un ANN. Le gain est d'environ 4 points en termes de précision alors que le temps d'apprentissage est considérablement réduit (ce gain est d'environ 38%). Nous envisageons, dans des travaux futurs, d'évaluer cette approche en utilisant des réseaux de neurones profonds ainsi que des espaces de représentation hiérarchiques.

Références

- ADHIKARI B. & JOSHI D. (1956). Distance discrimination et resume exhaustif. *Publ. Inst. Statist. Univ. Paris*, **5**, 57–74.
- BEALE E. (1972). A derivation of conjugate gradients. *Numerical methods for nonlinear optimization*, p. 39–43.
- BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BEZE M., DE MORI R. & ARBILLOT E. (2012). Decoda : a call-centre human-human spoken conversation corpus. In *LREC'12*.
- BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, **3**, 993–1022.
- BREUKELLEN M. V. & DUIN R. P. W. (1998). Neural network initialization by combined classifiers. In *Proceedings of the 14th International Conference on Pattern Recognition, ICPR'98*, p. 215–.
- CAZORLA M. A. & ESCOLANO F. (2003). Two bayesian methods for junction classification. *Image Processing, IEEE Transactions on*, **12**(3), 317–327.
- DAHL G. E., YU D., DENG L. & ACERO A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, **20**(1).
- DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K. & HARSHMAN R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, **41**(6), 391–407.
- DRAGHICI S. (2002). On the capabilities of neural networks using limited precision weights. *Neural networks*, **15**(3).
- DRAGO G. P. & RIDELLA S. (1992). Statistically controlled activation weight initialization (scawi). *Neural Networks, IEEE Transactions on*, **3**(4), 627–631.
- DUCH W. & JANKOWSKI N. (1999). Survey of neural transfer functions. *Neural Computing Surveys*, **2**(1), 163–212.
- FEURING T. (1996). Learning in fuzzy neural networks. In *Neural Networks, 1996., IEEE International Conference on*, volume 2, p. 1061–1066 : IEEE.
- GRIFFITHS T. L. & STEYVERS M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, **101**(Suppl 1), 5228–5235.
- HAGAN M. T., DEMUTH H. B., BEALE M. H. *et al.* (1996). *Neural network design*, volume 1. Pws Boston.
- KATHIRVALAVAKUMAR H. & SUBAVATHI S. J. (2011). A new weight initialization method using cauchy's inequality based on sensitivity analysis. *Journal of Intelligent Learning Systems and Applications*, **3**(1), 242–248.
- LINARÈS G., NOCÈRA P., MASSONIE D. & MATROUF D. (2007). The lia speech recognition system : from 10xrt to 1xrt. In *Text, Speech and Dialogue*, p. 302–308 : Springer.
- MØLLER M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, **6**(4).
- MORCHID M., DUFOUR R., BOUSQUET P.-M., BOUALLEGUE M., LINARÈS G. & DE MORI R. (2014). Improving dialogue classification using a topic space representation and a gaussian classifier based on the decision rule. In *ICASSP*.
- NGUYEN D. & WIDROW B. (1990). Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*, p. 21–26 : IEEE.
- POWELL M. J. D. (1977). Restart procedures for the conjugate gradient method. *Mathematical programming*, **12**(1).
- ROBERTSON S. (2004). Understanding inverse document frequency : on theoretical arguments for idf. *Journal of Documentation*, **60**(5), 503–520.
- THIMM G. & FIESLER E. (1997). High-order and multilayer perceptron initialization. *Neural Networks, IEEE Transactions on*, **8**(2), 349–359.

FDTB1: Repérage des connecteurs de discours en corpus

Jacques Steinlin¹ Margot Colinet¹ Laurence Danlos^{1, 2}

(1) ALPAGE, INRIA et Université Paris Diderot, 75013 Paris

(2) IUF

jacques.steinlin@gmail.com, margotcolinet@gmail.com, Laurence.Danlos@inria.fr

Résumé. Cet article présente le repérage manuel des connecteurs de discours dans le corpus FTB (French Treebank) déjà annoté pour la morpho-syntaxe. C'est la première étape de l'annotation discursive complète de ce corpus. Il s'agit de projeter sur le corpus les éléments répertoriés dans LexConn, lexique des connecteurs du français, et de filtrer les occurrences de ces éléments qui n'ont pas un emploi discursif mais par exemple un emploi d'adverbe de manière ou de préposition introduisant un complément sous-catégorisé. Plus de 10 000 connecteurs ont ainsi été repérés.

Abstract.

FDTB1 : Identification of discourse connectives in a French corpus

This paper presents the manual identification of discourse connectives in the corpus FTB (French Treebank) already annotated for morpho-syntax. This is the first step in the full discursive annotation of this corpus. The method consists in projecting on the corpus the items that are listed in LexConn, a lexicon of French connectives, and then filtering the occurrences of these elements that do not have a discursive use. More than 10K connectives have been identified.

Mots-clés : connecteurs de discours, annotation discursive de corpus, grammaire et discours.

Keywords: discourse connectives, discourse annotation, grammar and discourse.

1 Introduction

Le projet FDTB (French Discourse Treebank) s'inscrit dans la lignée du projet PDTB, Penn Discourse Treebank (Prasad *et al.*, 2008) qui a consisté à ajouter manuellement une couche d'annotation discursive sur le PTB (Penn Treebank), corpus composé d'articles du *Wall Street Journal*, déjà annoté en morpho-syntaxe. De même, le projet FDTB consiste à ajouter manuellement une couche d'annotation discursive sur le FTB, French Treebank (Abeillé *et al.*, 2003), corpus composé d'articles du journal *Le Monde* annoté en morpho-syntaxe. L'annotation complète du PDTB ou FDTB consiste *grosso modo* à repérer les connecteurs (« explicites » et « implicites »¹), et à annoter leurs sens et leurs arguments. Des expériences préliminaires d'annotation du FDTB (Danlos *et al.*, 2012) ont montré qu'il était difficile d'effectuer toutes ces opérations en une seule passe, entre autres du fait que de nombreux items lexicaux (e.g. *et*, *en gros*, *ainsi*, *alors*) sont ambigus entre un emploi comme connecteur de discours et un emploi non discursif. A titre d'illustration, la conjonction de coordination *et* est connecteur en (1-a) et non-connecteur en (1-b). De même, l'adverbial *en gros* est connecteur en (2-a) et non-connecteur en (2-b).

- (1) a. Fred a fini d'écrire son article **et** il est parti en vacances.
b. Fred **et** Marie sont de très bons amis.
- (2) a. Paul s'est cassé le bras et a attrapé la grippe. **En gros**, il ne va pas bien du tout.
b. Ce film traite **en gros** du réchauffement climatique.

1. Un connecteur implicite n'est pas réalisé : c'est le connecteur vide entre deux phrases simplement juxtaposées dans une parataxe. A l'inverse, un connecteur explicite est un item lexical non vide.

La détermination du statut discursif de *et* dans les exemples en (1) est triviale, mais ceci est loin d'être toujours le cas, comme le montre la littérature sur *ainsi* (Molinier, 2013; Karssenbergh & Lahousse, 2014) ou *alors* (Bras, 2008; Degand & Fagard, 2011). De ce fait, il est apparu qu'il valait mieux effectuer l'annotation du FDTB en commençant par une première étape, appelée FDTB1, qui consiste **uniquement** à repérer tous les connecteurs de discours du corpus. C'est cette étape que nous présentons ici. Signalons que l'annotation du PDTB n'est pas passée par cette première étape : seuls les 100 connecteurs anglais considérés comme les plus fréquents ont été annotés. Il n'est pas clair de savoir comment la fréquence des connecteurs anglais a été déterminée vu l'ambiguïté dont nous venons de parler. Seule une étude telle que celle menée dans le FDTB1 permet de déterminer la fréquence des connecteurs et d'identifier les 100 connecteurs français les plus fréquents (au moins dans un corpus journalistique).

Ce travail repose donc crucialement sur la notion de connecteur de discours qui est définie de manière fonctionnelle : les connecteurs de discours sont des items lexicaux qui permettent d'exprimer explicitement les relations discursives (sémantiques ou rhétoriques) entre deux segments de discours, « élémentaires » ou « complexes »². Les connecteurs de discours du français ont été répertoriés dans LexConn (Roze *et al.*, 2012), un lexique qui recense de la manière la plus exhaustive possible les connecteurs avec leur catégorie syntaxique et la ou les relations de discours qu'ils lexicalisent. Les catégories syntaxiques sont : conjonction de coordination, conjonction de subordination, préposition (introduisant un VP à l'infinitif ou au participe présent) et adverbial (catégorie qui regroupe principalement des adverbes simples et des syntagmes prépositionnels).

Le travail effectué dans le FDTB1 s'appuie sur LexConn tant sur le plan théorique que méthodologique. Sur le plan théorique, les principes qui ont guidé l'élaboration de LexConn ont tous été retenus dans le FDTB1. Un de ces principes est qu'un segment de discours élémentaire doit comporter un syntagme verbal VP (à temps fini ou non). Ce principe a éliminé de LexConn des prépositions comme *à cause de* ou *en raison de* qui ne peuvent introduire que des syntagmes nominaux (SN). Ce principe a aussi été appliqué dans le FDTB1 : les occurrences d'éléments de LexConn qui n'ont pas porté sur un VP dans le corpus ont été éliminées automatiquement. A titre d'illustration, seules les occurrences de la préposition *pour* introduisant un VP à l'infinitif ont été projetées sur le FTB, en excluant celles introduisant un SN³.

Sur le plan méthodologique, nous avons projeté automatiquement sur le FTB les éléments de LexConn respectant le principe ci-dessus, puis effectué des tâches de désambiguïsation pour savoir si ces occurrences étaient effectivement employées comme connecteurs. Les tâches de désambiguïsation sont les suivantes :

- désambiguïsation morpho-syntaxique (Section 3), par exemple pour les homonymes comme *bref* qui peut être un adjectif ou un adverbe connecteur,
- désambiguïsation entre grammaire et discours (Section 4) pour les adverbiaux (comme *ainsi* et *alors*) qui peuvent avoir un emploi comme connecteur et un emploi d'ajout à l'intérieur de leur phrase hôte,
- désambiguïsation entre grammaire et discours (Section 5) pour les prépositions et conjonctions de subordination (comme *pour* et *pour que*) qui peuvent avoir un emploi comme connecteur et un emploi d'introducteur de complément sous-catégorisé par un élément (verbal, nominal, adjectival ou adverbial) de la phrase où ils apparaissent.

Le corpus FDTB1 est librement disponible à l'adresse https://gforge.inria.fr/frs/?group_id=6145 où se trouve un manuel d'annotation complet (Danlos *et al.*, à paraître). Les résultats quantitatifs de l'annotation sont donnés à la Section 6 qui décrit aussi les utilisations potentielles du corpus. Avant d'expliquer les tâches de désambiguïsation, nous allons préciser la notion de connecteur de discours qui est au cœur du FDTB1.

2 La notion de connecteur de discours

Nous allons préciser la notion de connecteurs de discours (explicites) en l'opposant à celle de connecteur implicite et de « AltLex ».

Lorsqu'une phrase (typographique) ne contient aucun connecteur explicite, il est souvent considéré qu'elle est reliée à son contexte gauche par un connecteur implicite (voir note 1). Toutefois, il a été souligné dans

2. Un segment de discours est complexe s'il couvre plusieurs segments élémentaires contigus reliés eux-mêmes par des relations discursives.

3. Un tel filtrage bénéficie de l'annotation morpho-syntaxique du FTB et s'effectue automatiquement avec l'outil Tregex (Levy & Andrew, 2006).

divers travaux qu'une phrase sans connecteur explicite peut se voir relier à son contexte gauche par une relation discursive lexicalisée par des items lexicaux n'appartenant pas à la catégorie des connecteurs de discours, qui ont été baptisés AltLex (Alternative Lexicalization) dans le PDTB. Illustrons sur des exemples : en (3-a), le connecteur explicite *parce que* lie les deux propositions avec un sens causal. En (3-c), le lecteur doit inférer que les deux phrases sont reliées par une relation causale : on doit positionner un « connecteur implicite », noté ϕ . A l'intermédiaire, (3-b) ne comporte pas de connecteur explicite mais le lecteur ne doit faire aucune inférence : le fait que le contenu de la proposition *Fred a mal dormi* explique le contenu de *Fred est de mauvaise humeur* est explicitement indiqué par la séquence *Ceci est dû au fait que* qui se voit attribuer le statut d'AltLex.

- (3) a. Fred est de mauvaise humeur *parce qu'* il a mal dormi.
 b. Fred est de mauvaise humeur. *Ceci est dû au fait qu'* il a mal dormi.
 c. Fred est de mauvaise humeur. ϕ Il a mal dormi.

Le PDTB décrit quelques cas d'AltLex pour l'anglais et les définit par le fait qu'on ne peut pas leur ajouter de connecteur sans produire un effet de redondance. Pour le français, c'est un vaste champ d'étude non exploré (à l'exception des « verbes de discours » (Danlos, 2006)), mais il nous semble qu'une définition reposant sur une absence d'inférence par le lecteur soit préférable à une définition reposant sur un effet de redondance, la redondance n'étant pas exclue de la langue et éventuellement non perçue⁴.

La séquence *Ceci est dû au fait que* en (3-b) est compositionnelle et à ce titre ne saurait en aucun cas être considérée comme un connecteur de discours. En effet, un des premiers critères pour déterminer qu'une séquence composée de plusieurs mots est un connecteur est le fait qu'elle ne soit pas compositionnelle (Roze *et al.*, 2012). Considérons l'adverbial *à ce moment-là*. En (4-a), cet adverbial est compositionnel avec un sens de concomitance temporelle où le déterminant *ce ... là* est anaphorique comme le montre la paraphrase en *au moment où il a commencé à pleuvoir*. A l'inverse, cet adverbial est non compositionnel en (5-a) où *ce ... là* est non anaphorique. Ceci indique que seul *à ce moment-là* en (5-a) peut prétendre au statut de connecteur (lexicalisant une relation de conséquence entre les deux phrases). Ce statut est confirmé par deux autres faits : d'abord *à ce moment-là* en (5-a) ne peut pas être clivé — (5-b) est inacceptable contrairement à (4-b) — ce qui va de pair avec le fait qu'un connecteur adverbial n'est pas intégré au contenu propositionnel de sa phrase hôte contrairement à un adverbial temporel. Ensuite, *moment* en (5-a) ne peut pas être modifié — (5-c) est inacceptable contrairement à (4-c) — ce qui va de pair avec le figement versus la compositionnalité de la séquence composée.

- (4) a. Il a commencé à pleuvoir. *A ce moment-là*, Pierre est arrivé.
 b. Il a commencé à pleuvoir. C'est *à ce moment-là* que Pierre est arrivé.
 c. Il a commencé à pleuvoir. *A ce moment-là précis*, Pierre est arrivé.
- (5) a. Tu as l'air de penser qu'elle n'est pas honnête. *A ce moment-là*, tu devrais ne rien lui raconter.
 b. Tu as l'air de penser qu'elle n'est pas honnête. #C'est *à ce moment-là* que tu devrais ne rien lui raconter.
 c. Tu as l'air de penser qu'elle n'est pas honnête. #*A ce moment-là précis*, tu devrais ne rien lui raconter. (Roze *et al.*, 2012)

Tous les critères convergent donc pour indiquer que *à ce moment-là* en (4) est un AltLex tandis que c'est un connecteur en (5-a). Toutefois, la situation n'est pas toujours aussi tranchée : il semble exister un continuum entre AltLex et connecteur de discours, continuum qui reflète un processus de grammaticalisation (une étude dans ce sens a été menée par (Rysová & Rysová, 2014) sur le Tchèque).

En résumé, les connecteurs de discours ont pour fonction de lexicaliser les relations discursives entre deux segments de discours. Ils sont non intégrés au contenu propositionnel de leur phrase hôte, et non compositionnels et figés lorsqu'ils sont composés de plusieurs mots⁵.

4. Ainsi, la requête sur Google « Cela a ensuite été suivi » avec deux marqueurs (redondants) de la relation de précédence temporelle, à savoir le connecteur *ensuite* et le verbe de discours *suivre*, ramène aux alentours de 22 800 résultats, comme le texte suivant qui n'est pas perçu comme redondant : *L'excitation a commencé vendredi après un très laconique annonce de quatre lignes par la FINMA. Cela a ensuite été suivi de certains reportages à la fois par ...*

5. Néanmoins, certaines conjonctions de subordination comportant le complémenteur *que* (*pour que, avant que*) peuvent accepter l'insertion d'un adverbial ou d'une incise : *pour, dit-on, que*.

3 Ambiguïtés morpho-syntaxiques

Le premier aspect de la désambiguation dans le FDTB1 consiste, pour chaque occurrence d’item qui peut être connecteur, à décider si elle correspond morpho-syntaxiquement à la catégorie du connecteur recherché. Le premier cas d’ambiguïté est celui des homonymes, par exemple le mot *car* qui peut-être une conjonction de coordination (répertoriée dans LexConn) ou un nom commun. Le second cas correspond à une suite de mots qui a été répertoriée comme connecteur dans LexConn mais qui peut correspondre à d’autres catégories morpho-syntaxiques. Par exemple, la suite de mots *en fait* est répertoriée dans LexConn comme adverbial (de type syntagme prépositionnel), (6-a), mais elle peut correspondre à un pronom suivi d’un verbe comme en (6-b).

- (6) a. Fred avait l’air sûr de lui. **En fait**, il était mort de trouille.
 b. La Grand-Place était piétonne. Le maire **en fait** un parking.

Ces deux cas d’ambiguïté peuvent être levés automatiquement grâce à l’annotation morpho-syntaxique du corpus initial. Le manuel d’annotation du FDTB1 dresse la liste d’une trentaine d’éléments de LexConn qui présentent une ambiguïté morpho-syntaxique.

4 Les adverbiaux entre grammaire et discours

Le deuxième aspect de la désambiguïssation consiste à distinguer les occurrences des adverbiaux de LexConn qui ont une fonction discursive de ceux qui ont un rôle sémantique à l’intérieur de leur phrase hôte (avec la fonction syntaxique d’ajout et plus rarement de complément). Dans les termes de (Molinier & Lévrier, 2000), ceci s’approche de la distinction entre « adverbe de phrase » et autre adverbe. Cette désambiguïssation s’appuie sur les critères utilisés dans LexConn (brièvement rappelés à la Section 2). Au cas par cas, pour aider à déterminer si un adverbial potentiellement connecteur est effectivement employé comme connecteur en contexte, il a paru nécessaire de lister un emploi comme connecteur en donnant un aperçu de la ou les relations de discours qu’il lexicalise, et un emploi comme non connecteur en précisant le rôle sémantique à l’intérieur de la phrase hôte. Ce travail prolonge à large échelle celui qui a été effectué par les linguistes sur quelques connecteurs adverbiaux ; il est illustré pour *au contraire* (jamais étudié) et *ainsi* (largement étudié).

Au contraire a été annoté comme connecteur lorsqu’il lexicalise un contraste, (7-a), ou une sorte de reformulation du contexte gauche perçu comme une litote, (7-b) ; 38 occurrences. *Au contraire* n’est pas retenu comme connecteur lorsqu’il renforce une assertion négative, (8) ; 8 occurrences.

- (7) a. Selon cette enquête, 15% se prononcent pour un arrêt rapide du programme nucléaire français, 22% sont **au contraire** favorables à sa poursuite et à la construction de nouvelles centrales.
 b. Qu’il y ait aujourd’hui, ou qu’il y ait encore après le prochain comité directeur, plusieurs textes d’orientation en présence n’est pas en soi nuisible. Cela peut être **au contraire** une preuve de la vitalité du seul parti véritablement démocratique en France [...]
 (8) La nouvelle diminution du taux d’escompte de la Banque du Japon n’a nullement déprimé la monnaie japonaise, **au contraire**.

Ainsi a été annoté comme connecteur lorsqu’il lexicalise une relation de résultat ou d’exemplification, comme en (9-a) sans inversion de l’ordre canonique sujet-verbe ou en (9-b) avec inversion (Molinier, 2013; Karssenberg & Lahousse, 2014) ; 291 occurrences. *Ainsi* n’est pas connecteur lorsqu’il est utilisé comme anaphore de manière, (10-a), ou comme anaphore ou cataphore d’un discours rapporté, (10-b) ; 32 occurrences.

- (9) a. La Commission nationale [...] se limite à vérifier si les obligations comptables et financières sont remplies. **Ainsi**, il n’existe à ce jour aucun contrôle des dépenses des partis .
 b. M. Hockey ne mâche pas ses mots. **Ainsi** a-t-il invité les pays émergents à « se sevrer de la morphine de l’argent facile et à engager des réformes ».
 (10) a. Luc s’est comporté **ainsi** parce qu’il était fatigué.

- b. M. Michel Charasse, ministre du budget, a **ainsi** déclaré au micro de RMC : « C'est une affaire privée, et je ne vois pas pourquoi les pouvoirs publics seraient impliqués là-dedans ».

Dans notre corpus, il y a une centaine d'adverbiaux qui sont ambigus entre grammaire et discours. Le manuel d'annotation du FDTB1 dresse la liste des adverbiaux de LexConn qui sont toujours employés comme connecteurs (au moins dans ce corpus) : ils sont une cinquantaine.

5 Les prépositions et conjonctions entre grammaire et discours

Le troisième aspect de la désambiguïsation du FDTB1 consiste à distinguer les occurrences des prépositions et conjonctions de subordination qui ont une fonction discursive de celles qui sont sous-catégorisées par un élément verbal, nominal, adjectival ou adverbial. Cette désambiguïsation concerne cinq prépositions qui introduisent des infinitives — *pour*, *afin de*, *plutôt que de*, *jusqu'à* et *avant de* — et trois conjonctions de subordination reliées morphologiquement à trois de ces prépositions, à savoir *pour que*, *afin que* et *plutôt que*. Le cas le plus complexe et le plus fréquent est celui de la préposition *pour* suivie d'une infinitive qui a fait l'objet d'une publication, (Colinet *et al.*, 2014), résumée ici dans les grandes lignes. La préposition *pour* peut être connecteur, avec une valeur finale, causale ou temporelle, (11).

- (11) a. Côté alliances, DEC, qui s'est associé à Olivetti **pour** développer notamment des machines Risc - un microprocesseur à jeu d'instructions réduit...
 b. L' an dernier, le correspondant du quotidien britannique Financial Times s'est fait expulser **pour** avoir fait état de « l'évaporation » des énormes bénéfices tirés des exportations de pétrole pendant la guerre du Golfe.
 c. De son côté, la construction de logements reprend effectivement, après une forte baisse en 1991, **pour** remonter à un rythme annuel de 1,3 million de mises en chantier contre 1 million l'année précédente.

La préposition *pour* peut également introduire un complément sous-catégorisé par un verbe (12-a), un nom (12-b), un adjectif (12-c) ou encore un adverbe (12-d) (l'élément sous-catégorisant est souligné dans ces exemples).

- (12) a. Le gouvernement n'a pas profité de l'occasion **pour** trancher.
 b. Olivetti a toutes les qualités **pour** profiter de la nouvelle phase de croissance.
 c. 280000 tonnes de céréales seront nécessaires, chaque année, **pour** nourrir les poules.
 d. Ceci est trop rapide **pour** être durable.

Enfin, *pour* peut introduire une « relative sans mot QU » (Huddleston & Pullum, 2002), (13-a), et des emplois méta-discursifs, (13-b).

- (13) a. Un pont **pour** franchir l'Amazone a été construit en 1745.
 b. pour conclure, pour ne citer que lui, pour le dire autrement, ...

Si les « relatives sans mot QU » et les *pour* introducteurs d'expressions métadiscursives sont faciles à identifier, la distinction entre *pour* connecteur de discours et *pour* introduisant un argument sous-catégorisé n'est pas aisée. Il s'agit en effet d'une instance particulièrement délicate du problème général de la distinction entre arguments et modificateurs, pour laquelle une batterie de critères a été mise au point (Colinet *et al.*, 2014). Ces critères ont permis d'annoter manuellement les 1161 occurrences de *pour* introduisant une infinitive dans le FTB : 518 sont des connecteurs de discours (44%), 558 introduisent des compléments sous-catégorisés, 52 introduisent des relatives sans mot QU, et 33 introduisent des expressions métadiscursives. Ce travail a aussi permis de compléter les lexiques syntaxiques dans lesquels la préposition *pour* est largement ignorée comme introducteur de complément sous-catégorisé (Sagot *et al.*, 2014).

6 Conclusion

Les données chiffrées concernant la taille du FTB et le nombre de connecteurs annotés dans le FDTB1 avec leurs catégories sont données dans la Table 1. Les 536 occurrences de *en V-ant* correspondent à des gérondifs, voir *Fred a réconforté Marie en la complimentant sur son travail*, qui ont toutes été considérées comme connecteurs avec la particularité que le connecteur est en fait *en ... -ant*, c'est-à-dire la préposition *en* et le suffixe *-ant*.

FTB		FDTB1	
		adverbiaux	3221
		conj coord	3653
articles	1005	conj sub	1949
phrases	18535	prép V-inf	1070
mots	535 000	en V-ant	536
		TOTAL	10429

TABLE 1 – Taille du FTB et nombre de connecteurs dans le FDTB1

Cette annotation a mis en évidence 29 connecteurs non répertoriés dans LexConn dont une nouvelle version (comptant 353 éléments) est disponible sur le site du FDTB1. Près de 70% des éléments de LexConn ont au moins une occurrence dans le FDTB1. Le manuel donne la liste des 100 connecteurs les plus fréquents du corpus. L'accord entre deux annotateurs experts (deux auteurs de cet article) sur un échantillon de 13 articles donne un kappa de 0,70.

Le seul autre corpus du français écrit qui a été annoté pour le discours est le corpus Annodis (Péry-Woodley *et al.*, 2011) qui est 20 fois plus petit que le FTB. Ce corpus a reçu deux annotations : annotation en relations rhétoriques et annotation en structures multi-échelles. La première correspond à l'étude de l'organisation discursive qui est étudiée dans le FDTB, même si les approches sont différentes : l'annotation en relation rhétoriques d'Annodis s'inspire de la SDRT, Segmented Discourse Representation Theory, (Asher & Lascarides, 2003), tandis que, rappelons-le, l'annotation du FDTB1 et dans le futur du FDTB s'inspire du PDTB avec un focus sur les marques lexicales (connecteurs et AltLex) des relations discursives.

Le FDTB1 est donc le premier corpus écrit où les connecteurs du discours du français sont repérés systématiquement. Ce corpus peut être utilisé par les linguistes intéressés par les connecteurs. Il peut aussi être utilisé pour développer des méthodes d'apprentissage afin de repérer automatiquement les connecteurs dans un autre corpus, et ce d'autant plus aisément qu'il repose sur une annotation morpho-syntaxique.

Pour arriver à une annotation discursive complète à partir du FDTB1, trois tâches seront à effectuer :

1. annotation du sens et des arguments des connecteurs explicites repérés dans le FDTB1,
2. identification des AltLex et des connecteurs implicites (éléments définis à la Section 2),
3. annotation du sens et des arguments des éléments identifiés à l'étape 2.

La première et la troisième tâche seront effectuées dans l'esprit du PDTB, avec quelques modifications mineures concernant la hiérarchie des sens de connecteurs et l'annotation de leurs arguments (Danlos *et al.*, 2012). La première tâche est en cours.

Références

- ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In A. ABEILLÉ, Ed., *Treebanks*. Dordrecht : Kluwer Academic Publishers.
- ASHER N. & LASCARIDES A. (2003). *Logics of Conversation*. Cambridge : Cambridge University Press.
- BRAS M. (2008). *Entre relations temporelles et relations de discours*. Université de Toulouse le Mirail : HDR.
- COLINET M., DANLOS L., DARGNAT M. & WINTERSTEIN G. (2014). Emplois de la préposition *pour* suivie d'une infinitive : description, critères formels et annotation en corpus. In *Actes du Congrès Mondial de Linguistique Française (CMLF, 2014)*, Berlin, Allemagne.

- DANLOS L. (2006). Discourse verbs and discourse periphrastic links. In *Proceedings of the second workshop on Constraints in Discourse (CID 2006)*, Maynooth, Ireland.
- DANLOS L., ANTOLINOS-BASSOS D., BRAUD C. & ROZE C. (2012). Vers le FDTB : French Discourse Tree Bank. In *Actes de TALN 2012*, Grenoble, France.
- DANLOS L., COLINET M. & JACQUES STEINLIN (à paraître). FDTB1 : Repérage des connecteurs de discours dans un corpus français. *Revue Discours*.
- DEGAND L. & FAGARD B. (2011). *Alors* between discourse and grammar : The role of syntactic position. *Functions of Language*, **18(1)**, 29–56.
- HUDDLESTON R. & PULLUM G. (2002). *The Cambridge Grammar of the English Language*. Cambridge : Cambridge University Press.
- KARSENBERG L. & LAHOUSSE K. (2014). *Ainsi* en tête de phrase + inversion : une analyse de corpus. *SHS Web of Conferences*, **8**, 2413–2427.
- LEVY R. & ANDREW G. (2006). Tregex and Tsurgeon : tools for querying and manipulating tree data structures. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Gènes, Italie.
- MOLINIER C. (2013). *Ainsi* : Deux emplois complémentaires d’un adverbe type. *Linguisticae Investigationes*, **36-2**, 311–327.
- MOLINIER C. & LÉVRIER F. (2000). *Grammaire des adverbes*. Genève : Droz.
- PRASAD R., DINESH N., LEE A., MILTSAKAKI E., ROBALDO L., JOSHI A. & WEBBER B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Maroc.
- PÉRY-WOODLEY M.-P., AFANTENOS S. D., HO-DAC L.-M. & ASHER N. (2011). La ressource Annodis, un corpus enrichi d’annotations discursives. *Revue TAL*, **52(3)**, 71–101.
- ROZE C., DANLOS L. & MULLER P. (2012). LexConn : a French Lexicon of Discourse connectives. *Revue Discours*.
- RYSOVÁ M. & RYSOVÁ K. (2014). The centre and periphery of discourse connectives. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation*, p. 452–459, Phuket, Thailand.
- SAGOT B., DANLOS L. & COLINET M. (2014). Sous-catégorisation en *pour* et syntaxe lexicale. In *Traitement Automatique du Langage Naturel 2014*, Marseille, France.

ROBO : Une Mesure d'édition pour la comparaison de phrases Application au résumé automatique

Aurélien Bossard¹ Christophe Rodrigues²

(1) Université Paris 8, Laboratoire d'Informatique Avancée de Saint-Denis

(2) CNRS, UMR 7030, Laboratoire d'Informatique de Paris Nord

(1) aurelien.bossard@iut.univ-paris8.fr, (2) christophe.rodrigues@lipn.fr

Résumé. Dans cet article, nous proposons une mesure de distance entre phrases fondée sur la distance de Levenshtein, doublement pondérée par la fréquence des mots et par le type d'opération réalisée. Nous l'évaluons au sein d'un système de résumé automatique dont la méthode de calcul est volontairement limitée à une approche fondée sur la similarité entre phrases. Nous sommes donc ainsi en mesure d'évaluer indirectement la performance de cette nouvelle mesure de distance.

Abstract.

ROBO, an edit distance for sentence comparison – Application to automatic summarization

We here propose a sentence edit distance metric, ROBO, based on Levenshtein distance. This metric distance is weighted by words frequency and operation type. We apply ROBO on an automatic summarization system whose sentence selection metrics are on purpose restricted to sentence similarity approaches. ROBO performance can then be evaluated indirectly.

Mots-clés : résumé automatique, similarité sémantique, distance d'édition.

Keywords: automatic summarization, semantic similarity, edit distance.

1 Introduction

Dans cet article, nous nous intéressons à une nouvelle mesure de similarité entre phrases qui considère leur structure en utilisant un indice de surface – l'ordre des mots – et l'importance des mots. Le calcul de similarité entre phrases est un composant clé dans beaucoup de domaines du TAL, notamment la détection de plagiat et de paraphrase ou le résumé automatique. Cependant, dans ce dernier cas, la similarité entre phrases est souvent calculée à l'aide de méthodes dites de « sacs de mots », ou à l'aide de méthodes utilisant des traitements linguistiques complexes, syntaxiques et sémantiques. Alors que le premier type d'approches ne prend pas en compte la structure des phrases, la seconde est dépendante de la langue, requiert des ressources importantes et est coûteuse en temps de calcul. Nous proposons ici une mesure de similarité qui se situe à la croisée des mesures purement statistiques et de celles fondées sur une analyse linguistique profonde.

Étudier des mesures de similarité entre phrases pose le problème de leur évaluation. En effet, une mesure de similarité répond à un besoin particulier. Elles sont conçues pour mettre en avant des traits spécifiques des objets traitées pour un problème spécifique : Damerau-Levenshtein (Damerau, 1964) pour la vérification orthographique, (Smith & Waterman, 1981) pour l'alignement de séquences génétiques, la similarité cosinus (Salton & McGill, 1986) pour la classification de document ou la recherche d'information... L'évaluation d'une mesure de similarité ne peut donc pas être décorrélée de son application. Dans cet article, nous proposons d'appliquer et d'évaluer notre mesure dans le cadre du résumé automatique à base de graphe. Ce type de méthodes de résumé automatique a été introduite par (Salton *et al.*, 1997) et est très répandue parmi les résumeurs automatiques. Le calcul de la similarité entre phrases est au cœur de ces méthodes, qui sont donc un bon moyen d'évaluer notre mesure de similarité.

2 État de l'art

Afin de pouvoir comparer deux phrases entre elles, ou plus généralement deux séquences de symboles, différentes méthodes sont disponibles. Parmi les mesures les plus utilisées, nous pouvons distinguer deux principaux groupes : les méthodes dites de « sac de mots » ainsi que les distances d'édition. Dans cette section, pour étayer nos travaux, nous restreignons la présentation des distances ou mesures de similarité à leur application entre phrases, bien que celles-ci permettent de façon plus générale de comparer toute séquence de symboles. La liste des distances exposées ci-dessous n'est pas exhaustive et présente les mesures les plus utilisées. Ces mesures peuvent toutes être adaptées pour prendre en compte des n-grammes, moyennant un coût computationnel plus élevé (Damashek *et al.*, 1995).

2.1 Sacs de mots

Indice de Jaccard Cet indice permet de dénombrer les mots communs aux deux phrases et de les pondérer par le nombre total de mots différents des deux phrases. Soient les phrases p_1 et p_2 , l'indice de Jaccard est défini par :

$$J(p_1, p_2) = \frac{|p_1 \cap p_2|}{|p_1 \cup p_2|}$$

Cosinus Alors que l'indice de Jaccard repose sur une représentation ensembliste des phrases, la méthode cosinus utilise une représentation vectorielle. Une phrase devient alors un vecteur dont chaque dimension contient le poids d'un mot, généralement sa fréquence ou un *tf.idf*. Comparer deux phrases revient alors à comparer l'angle entre deux vecteurs.

$$Cosinus(p_1, p_2) = \sum_{w \in p_1 \cap p_2} V(w, p_1) \cdot V(w, p_2)$$

WordOrder (Li *et al.*, 2006) a développé une mesure fondée sur un modèle vectoriel. Chaque dimension des vecteurs qui représentent les phrases reçoit la position d'un mot dans la phrase. La norme de la différence entre deux vecteurs ainsi formés, normalisée par leur somme donne la similarité entre deux phrases. Des informations sémantiques issues de corpus et de ressources sémantiques affinent le calcul.

2.2 Distances d'édition

Jaro La distance de Jaro (Jaro, 1989) permet de prendre en compte dans une certaine mesure la position respective des mots identiques aux deux phrases (éloignement). Conceptuellement, un mot commun aux deux phrases mais situé en début d'une phrase et à la fin de l'autre sera fortement pénalisé (premier et deuxième terme de l'équation). De plus, les mots communs identiques se trouvant aux mêmes positions dans les deux phrases augmentent la similarité (troisième terme de l'équation). Plus la distance de Jaro est élevée (au maximum égale à un), plus les phrases sont similaires.

$$Jaro(p_1, p_2) = \frac{1}{3} \left(\frac{|m|}{|p_1|} + \frac{|m|}{|p_2|} + \frac{|m| - t}{|m|} \right) \quad \begin{array}{l} - |p_i|, \text{ le nombre de mots de la phrase } i ; \\ - |m|, \text{ le nombre de mots identiques (dont l'éloignement n'atteint pas la moitié de la longueur de la plus longue phrase)} ; \\ - t, \text{ le nombre de transpositions : le nombre de fois où des mots différents sont présents à des positions identiques.} \end{array}$$

La distance de Jaro-Winkler (Winkler, 1999) est une adaptation de la distance de Jaro permettant de favoriser les phrases commençant par des mots identiques.

$$JW(p_1, p_2) = Jaro(p_1, p_2) + (lp(1 - Jaro(p_1, p_2))) \quad \begin{array}{l} - l, \text{ la longueur du plus grand préfixe commun aux deux phrases} \\ - p, \text{ un coefficient fixé à la valeur } 0.1. \end{array}$$

Ainsi, plus les phrases sont similaires à leur début (l grand), plus la similarité obtenue est grande.

Levenshtein La distance de (Levenshtein, 1966) repose sur le calcul du nombre minimal d'opérations d'édition nécessaires pour transformer une phrase p_1 en une phrase p_2 : insérer ou supprimer un mot, substituer un mot par un autre. Le nombre minimal d'opérations nécessaires pour changer p_1 en p_2 représente alors la distance entre les deux phrases. Cette distance utilise une technique de programmation dynamique. Une solution optimale d'un problème est trouvée à partir de solutions optimales de sous-problèmes qui le composent.

Une matrice Lev de taille $(|p_1| + 1) \times (|p_2| + 1)$ est d'abord initialisée : chaque case i de la première ligne de Lev représente le coût nécessaire pour transformer les i premiers mots de p_1 en une phrase vide. De même, chaque case j de la première colonne de Lev représente le coût nécessaire pour transformer la phrase vide en les j premiers mots de p_2 .

Une fois la matrice Lev initialisée, on reporte dans l'ordre des indices i et j pour chaque case vide de la matrice, la valeur minimale parmi les cases adjacentes antérieures :

- la valeur de la case de gauche, de coordonnées $(i - 1, j)$ incrémentée de un, correspondant à une suppression ;
- la valeur de la case du dessus, de coordonnées $(i, j - 1)$ incrémentée de un, correspondant à une insertion ;
- et la valeur de la case en diagonale, de coordonnée $(i - 1, j - 1)$ correspondant à une substitution, incrémentée de un seulement si le $(i - 1)^{\text{ème}}$ mot de p_1 est différent du $(j - 1)^{\text{ème}}$ mot de p_2 .

Le processus est réitéré pour chaque case restante du tableau jusqu'à parvenir à la dernière case contenant le nombre d'opérations minimal pour changer p_1 en p_2 . Comme toute la matrice doit être parcourue entièrement mais seulement une fois, l'algorithme introduit par (Levenshtein, 1966) possède une complexité quadratique relative à la taille des phrases.

Des variantes de la distance de Levenshtein comme Smith-Waterman (Smith & Waterman, 1981) ou Smith-Waterman-Gotoh (Gotoh, 1982) existent, issues de travaux sur l'alignement de séquences biologiques. Elles ont pour objectif principal d'identifier efficacement les sous-séquences communes.

3 La mesure ROBO

Distance générale, la distance de Levenshtein peut être utilisée sur différents types de problèmes. Celle-ci permet de prendre en compte des éléments de structure des séquences comme par exemple l'ordre des symboles, contrairement à des méthodes « sac de mots ». De plus, elle possède l'avantage de distinguer explicitement les différentes opérations permettant de passer d'une séquence à une autre. Enfin, la fonction de coût associée à la substitution de deux symboles peut être directement adaptée au problème à traiter. Néanmoins, dans certains contextes, il peut être judicieux d'adapter son comportement afin de prendre en compte des caractéristiques particulières au problème à traiter.

Par exemple, pour le système de résumé automatique sur lequel repose l'évaluation de notre distance et que nous présentons en §4.1, il est important de rendre compte le plus fidèlement possible à la fois des similarités structurelles de deux phrases et de l'importance, relative au contexte des documents à résumer, des mots qui les composent. C'est pourquoi nous proposons d'adapter la distance de Levenshtein afin de prendre en compte une pondération des mots qui reflète leur importance au sein des textes, ainsi que la pondération des opérations d'édition afin de rendre compte le plus fidèlement possible de la structure des phrases.

L'idée d'adapter la distance de Levenshtein à des problèmes spécifiques a déjà été abordée dans la littérature comme par exemple pour la reconnaissance vocale en prenant en compte des poids pour chaque transcription phonétique (Ziolko *et al.*, 2010) ou encore (Barat *et al.*, 2010) pour la classification d'images.

L'algorithme 1 illustre les modifications apportées à la distance de Levenshtein. Dès l'initialisation, le poids de chaque mot est pris en compte en place d'une valeur unitaire constante. La fonction de coût utilisée demeure une fonction affine usuelle ; ici, la fonction moyenne permet lors de la substitution de deux mots de prendre en compte à part égale leur poids dans leur phrase respective. Enfin un facteur k est ajouté aux substitutions et permet de modifier l'importance de la fonction de coût des substitutions. Intuitivement, il peut être considéré qu'une substitution est l'équivalent d'une suppression suivie d'une insertion, ce qui peut être pris en compte en fixant $k = 2$. Si tous les poids sont affectés à une valeur unitaire, alors la distance entre deux phrases est identique à celle obtenue par la distance de Levenshtein d'origine. De même, la complexité de la méthode reste quadratique en le nombre de mots des phrases, dans la mesure où toute la matrice Lev doit être parcourue une fois.

4 Évaluation

Nous avons intégré la mesure ROBO à un système de résumé automatique afin de pouvoir l'évaluer indirectement, grâce à son impact sur les résultats du système de résumé par rapport à d'autres mesures de similarité. Le système de résumé a été simplifié au maximum, afin d'isoler les effets de la mesure ROBO sur les résultats obtenus. Nous utilisons pour toutes les mesures évaluées le *tf.idf* comme poids des mots.

Algorithm 1 retourne la distance de Levenshtein pondérée entre les phrases p_1 et p_2

```

1: procedure LEVENSSTEIN PONDÉRÉ( $p_1, p_2$ )
2:    $Lev[0][0] \leftarrow 0$ 
3:   for  $i \leftarrow 1$  to  $|p_1|$  do
4:      $Lev[i, 0] \leftarrow Lev[i - 1, 0] + w_{i-1}$ 
5:   end for
6:   for  $j \leftarrow 1$  to  $|p_2|$  do
7:      $Lev[0, j] \leftarrow Lev[0, j - 1] + w_{j-1}$ 
8:   end for
9:   for  $i \leftarrow 1$  to  $|p_1|$  do
10:    for  $j \leftarrow 1$  to  $|p_2|$  do
11:      if  $p_1[i - 1] = p_2[j - 1]$  then coût  $\leftarrow 0$ 
12:      else coût  $\leftarrow \frac{w_{i-1} + w_{j-1}}{2}$ 
13:      end if
14:       $suppr \leftarrow Lev[i - 1, j] + w_{i-1}$ 
15:       $insert \leftarrow Lev[i, j - 1] + w_{j-1}$ 
16:       $subst \leftarrow Lev[i - 1, j - 1] + k \times \text{coût}$ 
17:       $Lev[i, j] \leftarrow \text{minimum}(suppr, insert, subst)$ 
18:    end for
19:  end for
20:  Return  $Lev[|p_1|, |p_2|]$ 
21: end procedure

```

4.1 Système de résumé pour l'évaluation

Un résumé doit contenir les informations centrales des documents, mais aussi éviter la redondance et préserver la diversité informationnelle. On obtient souvent cela en appliquant un algorithme de sélection des phrases centrales suivi d'un algorithme d'élimination de la redondance. Nous appliquons ici les algorithmes bien connus LexRank (Erkan & Radev, 2004) puis MMR (Carbonell & Goldstein, 1998). L'algorithme LexRank consiste à simuler une marche aléatoire au sein d'un graphe où les nœuds sont les phrases et les arêtes les similarités entre elles. Il affecte itérativement à chaque nœud un score qui dépend du score des nœuds adjacents et de leur similarité. Le calcul des scores suivant est appliqué jusqu'à convergence :

$$s(u) = \frac{d}{N} + (1 - d) \sum_{v \in \text{adj}(u)} \frac{s(v)}{\text{sim}(u, v)} \quad \text{où } s \text{ est le score LexRank, } u \text{ un nœud, } \text{adj}(u) \text{ les nœuds adjacents à } u, \text{sim}(u, v) \text{ la similarité du nœud } u \text{ avec le nœud } v, \text{ et } N \text{ le nombre total de nœuds.}$$

LexRank nécessite donc deux paramètres : le *damping* (probabilité de saut aléatoire), et ϵ (seuil de convergence).

L'algorithme MMR consiste à donner un score à chaque phrase itérativement, en fonction d'un score de centralité et de la similarité aux phrases sélectionnées dans les étapes précédentes. La phrase qui maximise le score MMR est ajoutée au résumé, tant qu'elle n'amène pas le résumé à dépasser la taille maximum autorisée. L'algorithme est répété jusqu'à ce qu'aucune phrase ne vérifie cette dernière contrainte.

$$MMR = \underset{P_i \in D \setminus S}{\operatorname{argmax}} \left[\lambda \text{score}(P_i) - (1 - \lambda) \underset{P_j \in S}{\operatorname{argmax}} \text{sim}(P_i, P_j) \right] \quad \begin{array}{l} \text{où } D \text{ l'ensemble des phrases à résumer, } S \text{ l'ensemble des phrases déjà sélectionnées,} \\ \lambda \text{ le facteur de nouveauté, } \text{score} \text{ l'évaluation de la pertinence d'une phrase et } \text{sim} \\ \text{une similarité entre deux phrases.} \end{array}$$

A chaque lancement du système de résumé, les algorithmes LexRank et MMR se fondent sur la même mesure de similarité entre phrases. La mesure ROBO est une mesure de distance, et est comprise entre 0 et la somme des poids des mots de chacune des deux phrases. Nous avons donc normalisé la distance ROBO d'après la formule suivante :

$$\text{sim}_{ROBO}(P_1, P_2) = 1 - (\text{ROBO}(P_1, P_2) / \sum_{m \in P_1, P_2} \text{poids}(m))$$

4.2 Comparaison avec d'autres mesures de similarité

Nous comparons ici l'impact des mesures de similarités suivantes, toutes testées en pondérant les mots par leur *tfidf* : Cosinus, Jaccard, Jaro, Levenshtein, WordOrder (pas d'ajout d'informations sémantiques dans notre cas).

Corpus d'évaluation Nous avons évalué nos mesures sur le corpus RPM2, qui est à notre connaissance le seul corpus libre (licence GPL) en français disposant de résumés de références réalisés manuellement (de Loupy *et al.*, 2010). Il est constitué, à l'instar des corpus des campagnes DUC 2007 et TAC de 2008 à 2010, d'une partie dédiée au résumé classique et d'une partie consacrée au résumé de mise à jour, tâche plus complexe qui nécessite notamment de gérer la temporalité. Nous nous sommes intéressés uniquement à la partie dédiée au résumé classique afin d'évaluer le plus précisément possible l'apport de notre mesure de distance. Cette partie est constituée de 200 dépêches de presse regroupées en 20 thèmes de

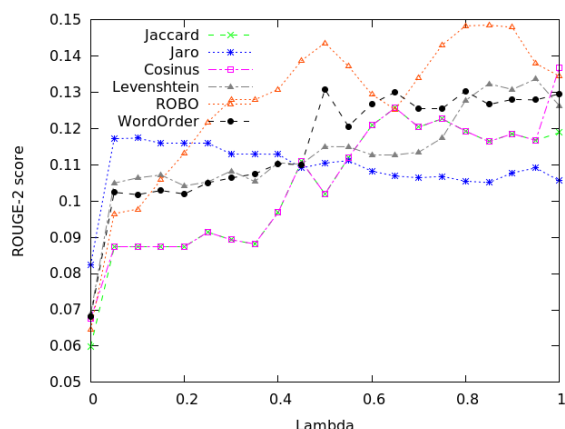
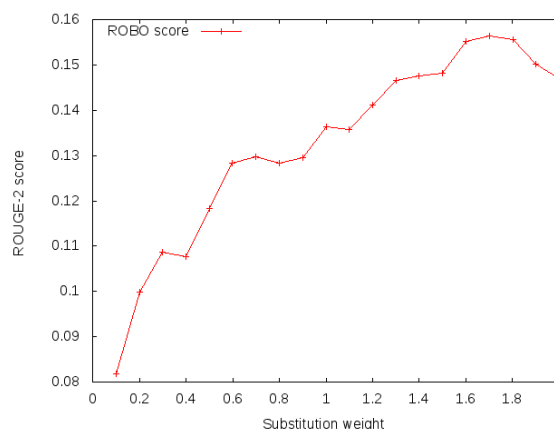


FIGURE 1: Résultats pour différentes mesures de similarités


 FIGURE 2: Scores ROBO selon le paramètre k , λ fixé à 0.8

dix documents. Chaque thème est pourvu de quatre résumés de référence de 100 mots maximum écrits manuellement. Les documents sont composés en moyenne de 230 mots. Nous avons utilisé les cinq premiers groupes de documents comme corpus de test, et les quinze derniers pour l'évaluation.

Paramètres expérimentaux Nous avons employé les paramètres suivants, conseillés dans l'article d'origine (Erkan & Radev, 2004). *dampling* : 0.15 ; ϵ : 0.000001 ; poids des substitutions (ROBO) : 1.5. La taille maximum des résumés a été fixée à 100 mots afin de les évaluer par rapport aux résumés de référence de RPM2, qui sont également de 100 mots maximum.

Mesures d'évaluation Les mesures ROUGE, entièrement automatiques, permettent de discriminer efficacement des systèmes procédant uniquement à de l'extraction de phrases, sans traitement supplémentaire risquant d'altérer la lisibilité et la cohérence des résumés générés (Lin, 2004)¹. Nous utilisons donc ici ROUGE qui répondent bien à nos besoins de rapidité d'évaluation et de corrélation aux jugements humains de pertinence des résumés (corrélation de Pearson de 0.96 sur une tâche similaire (Lin, 2004)). Les mesures ROUGE procèdent par comparaison de n-grammes entre résumés de référence et résumés automatiques. Nous évaluons les résumés d'après la variante ROUGE-2 (comparaison de bigrammes), la plus corrélée aux évaluations humaines pour notre cas (génération de résumés de 100 mots) d'après (Lin, 2004). L'outil d'évaluation ROUGE est « orienté rappel », puisque d'une part les résumés sont limités en nombre de mots et d'autre part, malgré la multiplicité des références, une information n'apparaissant pas dans une des références n'est pas forcément incorrecte.

5 Résultats

La figure 2 montre l'évolution des scores de résumé avec la mesure ROBO selon le poids donné aux opérations de substitution. Pour cette expérience, le paramètre λ de MMR a été fixé à 0.8. Passée la valeur 1, qui correspond à la mesure Levenshtein_{tf.idf}, les scores des résumés montrent une amélioration notable jusqu'à atteindre un palier aux alentours de la valeur 1.5. Cela correspond à l'idée intuitive suivante : les substitutions doivent être pénalisées moins qu'un couple (insertion, suppression), mais plus qu'une insertion ou suppression simple.

La figure 1 présente les scores ROUGE obtenus par toutes les mesures de similarité citées en §4.2 en faisant varier le paramètre λ de MMR. Ce paramètre gère l'élimination de la redondance ; plus il est proche de 1, moins la redondance compte dans la constitution du résumé. A 0, le score LexRank n'est plus pris en compte et la génération d'un résumé dépend uniquement de l'élimination de la redondance. Les résultats montrent que la mesure Cosinus permet d'obtenir de meilleurs résultats pour λ égal à 1. En revanche, et Levenshtein et ROBO améliorent les résultats lorsque le système procède à l'élimination de la redondance. Le score ROUGE pour λ fixé à 0.8 (paramètre conseillé dans (Carbonell & Goldstein, 1998)) montre une amélioration de 14% par rapport au score obtenu avec la mesure Cosinus, et de 9% par

1. Il est actuellement impossible d'évaluer la qualité linguistique et la cohérence d'un résumé sans expertise humaine.

ROBO, score ROUGE-2 : 0.22581

" Ca arrive tous les jours " Jérôme Kerviel, qui serait à l'origine de cette fraude, est aujourd'hui introuvable.

Après des perquisitions au domicile de Jérôme Kerviel et à la Société générale, le jeune homme a été mis en garde à vue.

La Société générale a révélé jeudi avoir été victime d'une fraude colossale, portant sur 40 à 50 milliards d'euros de positions, entraînant une perte de 4,9 milliards.

A la question Yahoo " Que pensez-vous de Jérôme Kerviel, le trader qui a fait perdre des milliards à la Société Générale ?

La Société Générale a déposé plainte contre Jérôme Kerviel.

Cosinus_{tfidf}, score ROUGE-2 : 0.19523

Une perquisition a été effectuée mercredi au domicile parisien du frère de Jérôme Kerviel, trader de la Société générale accusé d'être à l'origine d'une perte record de la banque de près de cinq milliards d'euros, a constaté un journaliste de l'AFP.

Des perquisitions ont eu lieu vendredi soir au siège de la Société Générale à La Défense, suite à la "fraude" massive qui a fait perdre à la banque près de 5 milliards d'euros, a indiqué samedi à l'AFP une porte-parole de la banque.

La Société Générale a déposé plainte contre Jérôme Kerviel.

FIGURE 3: Résumés comparés du Cluster 7 de RPM2 avec les mesures ROBO et Cosinus_{tfidf}

rapport au score obtenu avec la mesure Levenshtein pondérée par le *tfidf*, et ce dans les mêmes conditions ($\lambda = 0.8$). La figure 3 donne en exemple deux résumés, l'un généré d'après la mesure de similarité Cosinus_{tfidf}, l'autre d'après la mesure ROBO, du cluster 7 du corpus RPM2 relatif aux débuts de l'affaire Kerviel.

6 Discussion et perspectives

Dans cet article, nous avons montré l'intérêt de l'intégration d'une double pondération dans une mesure de distance d'édition pour une application de résumé automatique. Testée sur un ensemble reconnu de données libres (le corpus RPM2), la mesure ROBO a permis d'améliorer nettement la qualité des résumés générés, par rapport aux mesures « sacs de mots » ou à des distances d'édition non pondérées ou pondérées par l'importance des mots uniquement. Contrairement à des mesures de similarité qui exploiteraient des analyses syntaxiques lourdes, la mesure ROBO conserve un coût quadratique pour rendre compte uniquement de la structure de surface des phrases. Elle est générique et peut donc être directement adaptée à n'importe quelle langue ou corpus.

Etant donné les performances de la mesure ROBO sur l'évaluation de similarité inter-phrastique pour du résumé automatique, surtout en ce qui concerne l'élimination de la redondance, celle-ci pourrait être évaluée sur d'autres domaines du TAL, notamment la détection de paraphrases, ou l'évaluation automatique de résumé. Elle pourrait également être intégrée à des systèmes de résumé automatique plus complexes car ce système dépendant uniquement des similarités entre phrases nous a montré la pertinence de ROBO dans ce cadre.

Les travaux présentés dans cet article sont préliminaires, et de nombreuses pistes d'amélioration et d'évaluations complémentaires restent à explorer.

La mesure ROBO s'est avérée performante sur un corpus d'évaluation en français ; une évaluation sur un corpus multilingue permettra de valider l'approche sur d'autres langues qui arborent des structures différentes. Les comparaisons proposées dans l'article ne sont pas exhaustives et peuvent être étendues.

La mesure ROBO prend en compte la structure de surface des phrases et l'importance des mots, mais pas leur sémantique. Intégrer à la mesure ROBO des similarités sémantiques entre mots est une piste d'évolution qui nous semble naturelle.

Références

- BARAT C., DUCOTTET C., FROMONT E., LEGRAND A.-C. & SEBBAN M. (2010). Weighted symbols-based edit distance for string-structured image classification. In *Machine Learning and Knowledge Discovery in Databases*, volume 6321 of *Lecture Notes in Computer Science*, p. 72–86. Springer Berlin Heidelberg.
- CARBONELL J. & GOLDSTEIN J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98 : Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, p. 335–336, New York, NY, USA : ACM.
- DAMASHEK M. *et al.* (1995). Gauging similarity with n-grams : Language-independent categorization of text. *Science*, **267**(5199), 843–848.
- DAMERAU F. J. (1964). A technique for computer detection and correction of spelling errors. *Commun. ACM*, **7**(3), 171–176.

- DE LOUPY C., GUÉGAN M., AYACHE C., SENG S. & TORRES MORENO J.-M. (2010). A french human reference corpus for multi-document summarization and sentence compression. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.
- ERKAN G. & RADEV D. R. (2004). Lexrank : Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*.
- GOTOH O. (1982). An improved algorithm for matching biological sequences. *Journal of molecular biology*, **162**(3), 705–708.
- JARO M. A. (1989). Advances in record linking methodology as applied to the 1985 census of tampa florida. In *Journal of the American Statistical Society*, volume 64, p. 1183–1210.
- LEVENSHTEIN V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, **10**, 707.
- LI Y., MCLEAN D., BANDAR Z., O'SHEA J. & CROCKETT K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, **18**(8), 1138–1150.
- LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, p.10.
- SALTON G. & MCGILL M. J. (1986). *Introduction to Modern Information Retrieval*. New York, NY, USA : McGraw-Hill, Inc.
- SALTON G., SINGHAL A., MITRA M. & BUCKLEY C. (1997). Automatic text structuring and summarization. *Inf. Process. Manage.*, **33**(2), 193–207.
- SMITH T. F. & WATERMAN M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, **147**(1), 195–197.
- WINKLER W. E. (1999). *The state of record linkage and current research problems*. Rapport interne RR/1999/04, Statistics Research Division, U.S. Bureau of the Census.
- ZIOLKO B., GALK A. J. & SKURZOK D. (2010). Speech modelling using phoneme segmentation and modified weighted levenshtein distance. In *Audio Language and Image Processing (ICALIP), 2010 International Conference on*, p. 743–746.

Classification d'entités nommées de type « film »

Olivier Collin Aleksandra Guerraz

Orange Labs

2, avenue Pierre Marzin, 22307 Lannion Cedex, France

{olivier.collin,aleksandra.guerraz}@orange.com

Résumé. Dans cet article, nous nous intéressons à la classification contextuelle d'entités nommées de type « film ». Notre travail s'inscrit dans un cadre applicatif dont le but est de repérer, dans un texte, un titre de film contenu dans un catalogue (par exemple catalogue de films disponibles en VoD). Pour ce faire, nous combinons deux approches : nous partons d'un système à base de règles, qui présente une bonne précision, que nous couplons avec un modèle de langage permettant d'augmenter le rappel. La génération peu coûteuse de données d'apprentissage pour le modèle de langage à partir de Wikipedia est au coeur de ce travail. Nous montrons, à travers l'évaluation de notre système, la difficulté de classification des entités nommées de type « film » ainsi que la complémentarité des approches que nous utilisons pour cette tâche.

Abstract.

Named Entity Classification for Movie Titles

In this article, we focus on contextual classification of named entities for « movie » type. Our work is part of an application framework which aims to identify, in a text, a movie title contained in a catalog (e.g. VoD catalog). To do this, we combine two approaches : we use a rule-based system, which has good accuracy. To increase recall we couple our system with a language model. The generation of training data for the language model from Wikipedia is a crucial part of this work. We show, through the evaluation of our system, the complementarity of approaches we use.

Mots-clés : reconnaissance d'entités nommées, films, classification, règles, modèle de langage, Wikipedia.

Keywords: named entity recognition, movies, classification, rules, language model, Wikipedia.

1 Introduction

La détection de « film » est un cas particulier de détection d'entités nommées. La segmentation de ce type d'entités nommées, comme d'ailleurs celle des titres d'oeuvre en général, semble plus difficile à réaliser que celle des entités de type personne, lieu ou organisation. Cette tâche qui consiste à déterminer le début et la fin du titre est rendue plus complexe car de nombreux titres sont aussi des groupes de mots (noms communs, verbes, locutions adverbiales ou adjectivales) du lexique général de la langue : *A fleur de peau*, *A bout de souffle*, *Minuit à Paris*, *Vivre vite*. Dans le contexte de la phrase, il est donc difficile de déterminer l'étendue de ce type d'entité : *A bout de souffle* a été tourné...

Dans un cadre opérationnel, nous avons été confrontés à un problème différent. Les bases de données de titres étant de plus en plus nombreuses et le but applicatif étant de repérer un titre de film dans un texte et proposer au client de le voir en VoD, l'ensemble des titres est connu à l'avance. La détection est donc simplifiée puisqu'elle se résume essentiellement à repérer dans les documents des groupes de mots correspondant aux titres de la base, ce qui constitue techniquement un accès à un lexique de titres. Ce repérage entraîne cependant deux types de problème :

1. La forme d'un titre dans le document peut être différente de la forme du titre de la base. Cette différence peut être faible car liée à des variantes morphologiques (accentuation, ponctuation, conjugaison) ou plus importante car liée à des versions de l'oeuvre (*Mission impossible*, *Mission impossible 2*, *Mission impossible : Protocole Fantôme...*) ou à des versions non traduites (*Mission : Impossible – Ghost Protocol*)
2. Cette détection entraîne des annotations non désirées des entités qui sont ambiguës avec des entités de type différent (lieu, personne...) ou des noms communs (*Elle*, *Vie privée...*).

Le travail présenté se focalise uniquement sur ce problème de classification (film/pas film). La technique est relativement simple : nous avons réalisé un modèle de classification statistique qui s'appuie (apprentissage et test) sur la représentation fournie par notre système initial de détection d'entités nommées (basé sur les règles). La décision issue du classifieur statistique qui complète celle du premier système n'est utilisée que pour les données non étiquetées « film » de manière sûre par notre système initial. Il existe différentes manières de coupler un système symbolique à un système probabiliste telles que dans (Béchet *et al.*, 2011) et (Nouvel *et al.*, 2012). Nous présentons une hybridation très simple qui consiste à utiliser un classifieur basé sur des règles offrant une très bonne précision en amont d'un classifieur statistique qui ne remet pas en cause la décision prise par le système symbolique mais la complète sur les cas ambigus. Pour réaliser notre classifieur statistique, nous avons utilisé une technique standard de désambiguïsation statistique : les modèles de langage. Leur mise en oeuvre a été faite par l'un des outils les plus connus et techniquement aboutis dans ce domaine : la librairie SRILM¹.

Le corpus d'apprentissage pour le modèle de langage a été constitué à partir de Wikipedia. En effet, cette ressource a suscité un grand intérêt pour la tâche de Reconnaissance d'Entités Nommées et son utilisation a été largement approuvée. (Nothman *et al.*, 2008), (Charton & Torres-Moreno, 2009) montrent qu'il est possible d'obtenir à partir de Wikipedia une ressource annotée en entités nommées de large couverture et de très bonne qualité. Les systèmes de détection d'entités nommées entraînés avec des ressources issues de Wikipedia peuvent obtenir de très bons résultats sur d'autres types de corpus (Balasuriya *et al.*, 2009). Les ressources de Wikipedia sont utilisées pour compléter les ressources des systèmes de détection d'entités par (Stern & Sagot, 2010) et (Okinina *et al.*, 2013).

Dans cet article, nous présentons d'abord comment les données d'apprentissage pour le modèle de langage ont été produites (section 2). Le couplage de notre système de détection d'entités nommées à base de règles avec un modèle de langage est décrit dans la section 3. Enfin, nous présentons une évaluation de notre approche (section 4).

2 Production de données d'apprentissage pour le modèle de langage

La production de données d'apprentissage ou l'annotation consiste à ajouter des métadonnées (étiquettes sémantiques) telles que *<film> Hollywood </film>*, en regard des entités à classer, de manière à lever contextuellement l'ambiguïté portant sur l'entité. Dans le cas présent *Hollywood* est potentiellement un lieu ou un film mais ne possède qu'une catégorie dans un contexte donné. Ce travail est généralement réalisé manuellement en parcourant visuellement les textes à annoter, ce qui est long et fastidieux et peut nécessiter des ressources humaines importantes pour obtenir des données annotées en quantité suffisante.

Les données textuelles des pages de Wikipedia possèdent des liens internes qui sont un type d'annotation sémantique « gratuite ». Ce sont les rédacteurs des pages qui ajoutent manuellement ces annotations. Nous avons réalisé un travail de récupération et de filtrage des pages de Wikipedia (français). Le résultat de ce travail est un gros fichier texte qui associe à chaque page de Wikipedia une partie de son texte (hors tableaux, citations ...) en conservant les liens internes. Ces liens sont des références à des pages de Wikipedia qui associent à la forme du texte le nom de la page correspondante. Par exemple :

```
[[Jean Renoir]] : texte = Jean Renoir, page = Jean Renoir
[[Jacques Martin (auteur)|Jacques Martin]] : texte = Jacques Martin, page = Jacques Martin (auteur)
[[Jacques Martin (animateur)|Jacques Martin]] : texte = Jacques Martin, page = Jacques Martin (animateur)
```

Généralement, ces liens internes sont désambiguïsés par les rédacteurs des pages. Cela signifie que dans le cas où plusieurs formes existent (comme dans le cas de *Jacques Martin*), le nom de la page associée spécifie quel est le « bon » *Jacques Martin* dans le contexte de la phrase. Sans précision particulière (cas de *Jean Renoir*), c'est la page unique de *Jean Renoir* qui est référencée. Ceci étant, les références ne possèdent pas souvent un type sémantique tel que « auteur » ou « animateur ». Dans le cas de *Jean Renoir*, il faut donc trouver un moyen de connaître quel est son type.

Pour récupérer une étiquette sémantique nous avons utilisé une stratégie automatique qui est très simple mais qui ne permet pas d'étiqueter tous les liens de Wikipedia. En ce qui concerne les films et un certain nombre d'autres types d'entités, cela permet toutefois de constituer un corpus partiellement annoté. L'hypothèse est donc que ce corpus partiellement annoté, et probablement un peu bruité, nous permette de réaliser un modèle de langage utile. Nous avons en quelque sorte inversé la problématique. Plutôt que de compléter l'annotation de tous les liens internes de Wikipedia, nous avons

1. <http://www.speech.sri.com/projects/srilm/manpages/>

essayé de qualifier les pages dont nous sommes sûrs (ou presque !) de la catégorie sémantique. Une fois ces pages qualifiées, nous avons complété les liens où ces pages apparaissent avec la catégorie récupérée. Par exemple : *Jean Renoir* sera complété avec le type « réalisateur » et son lien interne pourra être transformé en une annotation de type XML `<realisateur>Jean Renoir</realisateur>`. La récupération des types sémantiques a été réalisée à partir de deux sources d'information suivantes :

- le type fourni dans le nom de la page Wikipedia, lorsqu'il existe, tel que dans l'exemple précédent de *Jacques Martin*
- la relation « est un » entre le nom de la page et le premier lien à droite qui apparaît généralement dans le premier paragraphe de la description de la page.

Pour la page de *Jean Renoir*² on a :

```
Jean Renoir est un [[réalisateur]] et [[scénariste]] [[français]], né à [[Paris]]...
```

Le type associé directement à la page n'est pas très fréquent et n'est pas toujours un type sémantique mais peut être un type thématique ou une date. La relation « est un » peut induire des erreurs mais est généralement précise. En fusionnant ces deux types d'informations, nous avons pu étiqueter automatiquement près de 200 000 pages de Wikipedia français et notamment près de 6 000 titres de film. Chaque titre apparaissant en moyenne plusieurs fois dans Wikipedia, nous avons obtenu environ 26 000 phrases contenant au moins un titre de film étiqueté par phrase. Ce type de corpus contenant un étiquetage contextuel de films, avec une telle quantité d'annotation est unique. D'autre part, cette stratégie n'est pas limitée aux films et peut être appliquée sur tous les types d'entités rencontrés lors de l'étiquetage initial des pages de Wikipedia. Elle n'est pas non plus limitée au français, elle devrait être reproductible à faible coût pour toutes les autres langues traitées par Wikipedia. Nous avons ensuite projeté les étiquettes extraites dans une taxonomie simple permettant de catégoriser les principaux types d'entités : film, evt (événement), jeu, time, livre, loc (lieu), org (organisation), album, pers(personne), amount. Cette projection a été réalisée automatiquement. Nous avons utilisé ce corpus annoté pour réaliser une désambiguïsation statistique avec un modèle de langage que nous couplons avec notre système de détection d'entités nommées à base de règles.

3 Couplage du système à base de règles avec un modèle de langage

Nous avons choisi un couplage relativement simple qui consiste à utiliser notre système à base de règles pour générer une représentation qui est ensuite utilisée par le modèle de langage statistique, aussi bien en apprentissage qu'en test. L'apprentissage se fait « à la suite » du système à base de règles ce qui permet d'utiliser la segmentation produite par notre système (notamment multi-mots et entités nommées) ainsi que les étiquettes sémantiques des entités nommées, notamment des personnes, des lieux et des organisations. Le modèle de langage est ensuite généré et utilisé en reprenant cette segmentation des données. Il ne sert donc pas à détecter l'étendue des entités nommées (réalisée par notre système) mais uniquement à classer le type des entités en fonction du contexte. Le modèle généré est donc dépendant de la segmentation et de la précision de notre système en ce qui concerne le typage des entités nommées.

3.1 Configuration initiale du système de détection d'entités nommées à base de règles

Notre système de détection d'entités nommées est basé sur la plate-forme de traitement linguistique de textes TiLT décrite dans (Heinecke *et al.*, 2008). Ce système permet de manière générale de :

- repérer des entités nommées déjà connues grâce à des informations lexicales,
- découvrir des entités nommées sur la base de déclencheurs lexicaux et de contraintes de bonne formation.

Le repérage d'entités nommées s'appuie sur des indices lexicaux, typographiques et contextuels. En l'absence d'indices contextuels, la confiance que l'on peut avoir dans le lexique varie en fonction des caractéristiques de l'entité :

- une entité nommée doit comporter une majuscule,
- une entité nommée mono-mot ne sera pas repérée en cas d'ambiguïté avec un autre mot du lexique et en l'absence d'un contexte syntaxique fiable (*Puma* lieu ne sera pas étiqueté lieu en l'absence de contexte fiable, par exemple *la ville de Puma*),
- une entité nommée multi-mots ne sera pas repérée si elle est de type mot outil + mot du lexique (*La Flèche* ne sera pas étiqueté lieu en l'absence de contexte fiable, par exemple *la piscine de La Flèche*).

2. http://fr.wikipedia.org/wiki/Jean_Renoir

La découverte d'entités nommées se fait au moyen de déclencheurs typographiques et lexicaux. La présence de majuscule est un élément majeur permettant de soupçonner la présence d'une entité nommée et de la segmenter. Les déclencheurs lexicaux sont internes ou externes selon qu'ils font (ou non) partie de l'entité nommée elle-même. La chaîne de caractères constituée par le(s) déclencheur(s) interne(s) et les noms propres est identifiée via les règles de grammaire en dépendance et est visible sous forme d'une locution : *Jean-Marcel Dupont* devient une locution de type nom de personne. Les déclencheurs sont principalement des noms qui introduisent de façon directe ou indirecte (c'est-à-dire à l'intermédiaire d'une préposition) les entités nommées.

Le repérage des titres de film consiste alors à valider un titre de film (existant dans le lexique) dans un contexte donné. La fonctionnalité de découverte n'est pas utilisée pour la détection de films. Le repérage des titres de film exploite des informations lexicales, typographiques (guillemets) et stylistiques (énumération), contextuelles modélisées par une grammaire locale sous forme de règles de « chunking » et des informations sur la longueur des titres de film (on part de l'hypothèse que plus le titre est long, moins il peut être confondu avec un autre élément de la phrase). Le lexique contient 70 054 entités nommées de type film. 41% des titres de film sont ambigus avec un autre mot du lexique (par exemple, *Vampires*, *Pirates*, *A fleur de peau*). Un contexte déclencheur est nécessaire pour leur identification dans les phrases. 2% de titres de film sont ambigus avec une autre entité nommée (parmi les types personne, lieu, organisation). Il est à noter que le lexique ne contient pas de titres d'autres oeuvres artistiques tels que les titres de livre, les titres de musique ou les titres de spectacle, etc.

3.2 Apprentissage du modèle de langage

Afin de coupler le système décrit dans la section précédente avec un modèle de langage, une mesure de confiance binaire est attribuée à chaque titre de film pour distinguer les films en contexte fiable. Les films en contexte fiable sont étiquetés 1, les autres titres de film potentiels sont étiquetés 0.

Le modèle de langage va attribuer une probabilité à tous les titres de film étiquetés 0 (indice de confiance). Le but est d'augmenter le rappel sans trop diminuer la précision. Nous considérons comme résultat positif soit les films initialement étiquetés 1 par notre système soit les films initialement étiquetés 0 mais dont le résultat du modèle de langage donne pour l'hypothèse du film (pour l'étiquette *NAM:film*) une probabilité forte. Nous pouvons donc définir une valeur de seuil pour cette probabilité. Dans un premier temps, les résultats ont été calculés avec une valeur de seuil de 0,8. Si $P(NAM:film) \geq 0,8$ alors le résultat est positif (on considère que l'entité est un film). Cette valeur de seuil permet d'effectuer un réglage précision/rappel, ce qui peut être utile dans un cadre applicatif de manière à favoriser le rappel ou la précision. Le seuil est relatif aux probabilités conditionnelles de l'étiquette *NAM:film* versus toutes autres catégories (personne, lieu, organisation...). Nous considérons donc implicitement que notre modèle de classification des « non-films » est constitué de la somme des probabilités des étiquettes autres que *NAM:film*.

Lors de l'apprentissage du modèle de langage nous avons utilisé des données de supervision (étiquetage des entités à classer) issues des étiquettes supposées fiables par notre système initial. Mais en nous limitant à cet étiquetage, le modèle ne pourrait qu'apprendre implicitement des règles statistiques conduisant au fonctionnement actuel de notre système à base de règles. Nous avons donc remplacé ou complété les étiquettes générées par des étiquettes apportant une information complémentaire. Cette information est essentiellement contenue dans de nouveaux contextes. De cette manière, nous espérons obtenir un système conservant la précision de notre système initial mais qui étend ses performances à d'autres contextes non gérés par notre système, et par conséquent augmente le rappel. L'exemple suivant illustre ce processus :

En 1928, pour le bimillénaire de la cité de Carcassonne, Jean Renoir réalise
Le Tournoi dans la cité.

1. Segmentation et étiquetage produit par notre système à base de règles (*NAM:* indique la catégorie de l'entité nommée, les « _ » relient les mots d'un même segment) :

En 1928/*NAM:time* , pour le bimillénaire de la cité de Carcassonne/*NAM:loc* ,
Jean_Renoir/*NAM:pers* réalise Le Tournoi dans la cité .

Notre système étiquette *1928*, *Carcassonne*, *Jean_Renoir* mais pas *Le_Tournoi_dans_la_cité*

2. Etiquetage externe à partir de Wikipedia (cf. section 2) :

En 1928, pour le bimillénaire de la cité de Carcassonne,
<*pers*>Jean Renoir</*pers*> réalise <*film*>Le Tournoi dans la cité</*film*>.

Les données produites étiquettent *Jean Renoir* et *Le Tournoi dans la cité*

3. Etiquetage cumulé et projeté dans le même formalisme d'annotation :

En 1928/NAM:time , pour le bimillénaire de la cité de Carcassonne/NAM:loc ,
Jean_Renoir/NAM:pers réalise Le_Tournoi_dans_la_cité/NAM:film .

4. Etiquetage utilisé pour l'apprentissage du modèle de langage :

En NAM:time , pour le bimillénaire de la cité de NAM:loc ,
NAM:pers réalise NAM:film .

Chaque segment annoté est remplacé par sa catégorie sémantique.

En cas de double étiquetage effectué par notre système et par l'étiquetage externe à partir de Wikipedia (cf. section 2), ce sont ces dernières étiquettes qui sont utilisées. Les étiquettes externes peuvent donc soit remplacer celles de notre système soit les compléter.

Pour réaliser le calcul des probabilités du modèle de langage, nous avons constitué un ensemble d'exemples représentatifs des contextes d'apparition des films, mais aussi des contextes d'apparition d'entités qui ne sont pas des films, de manière à en « peser » les probabilités respectives. En phase d'utilisation du modèle de langage, ces probabilités calculées à partir des données annotées vont être réutilisées (trigrammes, bigrammes, unigrammes) pour calculer la probabilité maximale d'émission de chaque catégorie sémantique pour l'ensemble de la phrase. Le modèle statistique ne s'applique que sur les décisions portant sur les titres de film jugés « non fiables » (étiquetés 0) par le système initial.

Au final, nous avons donc un système qui réalise un apprentissage statistique à partir d'une représentation des données filtrées par notre système et annotées automatiquement par Wikipedia.

4 Evaluation

4.1 Corpus d'évaluation

Nous avons constitué un corpus d'évaluation. L'annotation de ce corpus en entités nommées a été effectuée manuellement par un seul annotateur, selon les conventions d'annotations internes largement inspirées de celles de la campagne ESTER2 et de celles de la campagne QUAERO³. Aucun accord inter-annotateurs n'a pu être mesuré. Le corpus se compose de 287 documents textuels répartis comme suit :

- 257 textes courts de type dépêche AFP,
- 7 textes issus de Wikipedia décrivant des personnalités du monde cinématographique et musical,
- 5 dépêches AFP issues du portail Orange,
- 18 textes issus des Inrockuptibles, du portail Orange portant sur des chanteurs ou des groupes de musique.

4.2 Résultats

Le tableau 1 donne les résultats obtenus pour la détection de film avec le système initial et avec le même système auquel un modèle de langage a été ajouté a posteriori.

	Précision	Rappel	F-mesure
Système initial	0,84	0,33	0,48
Système initial+ML (Seuil $\geq 0,8$)	0,76	0,53	0,63

TABLE 1 – Résultats

La configuration hybride (système initial + modèle de langage) apporte une amélioration du système. La F-mesure augmente de 31% (elle passe de 0,48 à 0,63). Ceci se traduit par une augmentation du rappel (de 60%) et une baisse de la précision (de 9%) qui reste, toutefois, acceptable.

3. <https://perso.limsi.fr/rosset/quaero-guide-annotation-2011.pdf>

4.2.1 Bruit

42 % du bruit est lié à la confusion de catégorie. On y trouve principalement (18% du bruit) des confusions avec d'autres sous-types de la catégorie oeuvre artistique, par exemple avec :

- un titre de musique (*Carla Bruni enregistre **Douce France***)
- un titre d'album (*Le premier single extrait de l'album, **Falling Down**, s'accompagne d'un clip*)
- un titre de livre (*tout en reprenant le thème du **Désert des Tartares** de Dino Buzzati*)

Une sous-catégorisation de la classe personne en acteurs, écrivains, réalisateurs, chanteurs pourrait limiter ce genre de bruit. Les confusions avec le sous-type personnage de la catégorie personne sont également fréquentes et représentent 17% du bruit. Par exemple, une confusion avec un personnage éponyme du film dans « 1993 Fanfan de Alexandre Jardin : **Fanfan** ». Le reste de bruit est dû notamment à des erreurs de segmentation. Le titre de film détecté n'est pas complet (par exemple dans le lexique on a *Die Hard* alors que le film à détecter est *Die Hard 4*). Ce type d'erreur est fréquent pour les différents volets de films (par exemple *Mission Impossible 2*) et pourrait être limité par un repérage par *approximate string matching* (technique fréquemment utilisée en correction). On trouve également ce type d'erreur pour des films avec des titres contenant un sous-titre (par exemple dans le lexique on a *Belphégor* pour le titre de film *Belphégor, le fantôme du Louvre*). Le titre de film détecté fait partie d'un titre d'une autre oeuvre artistique (titre de livre, titre d'album ou titre de chanson) Ce type d'erreurs (de segmentation et de catégorisation) est fréquent dans les textes qui comportent beaucoup de mots en anglais, par exemple : sur l'album *This Is The Sea* où *The Sea* est détecté comme titre de film.

4.2.2 Silence

41% du silence est dû à l'incomplétude du lexique : le titre de film est inconnu de la base de données, source du lexique de notre système. En se ramenant aux données présentes uniquement dans le lexique, le rappel serait donc nettement meilleur et par conséquent donc la F-mesure. Voici un exemple de non-détection (exemple en gras) : *Si **Manon** n'est pas une grande réussite, c'est sans doute parce que le metteur en scène n'était pas en forme.*

Les variations relatives des résultats sont toutefois indicatives des propriétés du système. Un lexique exhaustif devrait produire un accroissement relatif des résultats analogues à celui que nous constatons. En terme opérationnel, cela veut dire que les films qui ne sont pas dans la base ne seront pas détectés.

5 Conclusions

Ce travail nous a permis d'améliorer les résultats initiaux en classification des titres de film obtenus par notre système basé sur des règles. Cette amélioration globale de la F-mesure porte surtout sur le rappel, au prix d'une légère dégradation de la précision. Les avantages du processus sont :

- l'intégration simple du modèle de langage au système initial basé sur des règles par un couplage a posteriori,
- l'utilisation d'une mesure de probabilité permettant de régler les performances suivant les axes précision/rappel,
- la possibilité d'ordonner les solutions par probabilités décroissantes pour un filtrage utilisateur,
- une génération automatique du corpus d'apprentissage annoté, donc aucun coût humain d'annotation,
- la possibilité de reproduire cette technique sur d'autres langues.

Références

- BALASURIYA D., RIGLAND N., NOTHMAN J., MURPHY T. & CURRAN J.-R. (2009). Named entity recognition in wikipedia. In *People's Web 2009*, p. 10–18, Morristown, NJ, USA : Association for Computational Linguistics.
- BÉCHET F., SAGOT B. & STERN R. (2011). Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées. In *Actes de TALN 2011 (Traitement automatique des langues naturelles)*, Montpellier : ATALA LIRMM.
- CHARTON E. & TORRES-MORENO J.-M. (2009). Classification d'un contenu encyclopédique en vue d'un étiquetage par entités nommées. In *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

- HEINECKE J., SMITS G., CHARDENON C., GUIMIER DE NEEF E., MAILLEBUAU E. & BOUALEM M. (2008). Tilt : plate-forme pour le traitement automatique des langues naturelles. In *TAL 2008 (Traitement automatique des langues)*, p. 17–41 : ATALA.
- NOTHMAN J., CURRAN J.-R. & MURPHY T. (2008). Transforming wikipedia into named entity trainig data. In *Actes de ALTA 2008*, p. 124–132, Tasmania : ACL Australian Language Technology Workshop.
- NOUVEL D., ANTOINE J.-Y., FRIBURGER N. & SOULET A. (2012). Coupling knowledge-based and data-driven systems for named entity recognition. In *Actes de EACL 2012*, Avignon : ACL.
- OKININA N., NOUVEL D., FRIBURGER N. & ANTOINE J.-Y. (2013). Apprentissage supervisé sur ressources encyclopédiques pour l’enrichissement d’un lexique de noms propres destiné à la reconnaissance des entités nommées. In *Actes de TALN 2013 (Traitement automatique des langues naturelles)*, Sables d’Olonne : ATALA LINA.
- STERN R. & SAGOT B. (2010). Détection et résolution d’entités nommées dans des dépêches d’agence. In *Actes de TALN 2010 (Traitement automatique des langues naturelles)*, Montréal : ATALA RALI.

A critical survey on measuring success in rank-based keyword assignment to documents

Natalie Schluter

Center for Language Technology, University of Copenhagen, Copenhagen, Denmark
natschluter@hum.ku.dk

Abstract. Evaluation approaches for unsupervised rank-based keyword assignment are nearly as numerous as are the existing systems. The prolific production of each newly used metric (or metric twist) seems to stem from general dissatisfaction with the previous one and the source of that dissatisfaction has not previously been discussed in the literature. The difficulty may stem from a poor specification of the keyword assignment task in view of the rank-based approach. With a more complete specification of this task, we aim to show why the previous evaluation metrics fail to satisfy researchers' goals to distinguish and detect good rank-based keyword assignment systems. We put forward a characterisation of an ideal evaluation metric, and discuss the consistency of the evaluation metrics with this ideal, finding that the average standard normalised cumulative gain metric is most consistent with this ideal.

1 Introduction

Automatic keyword assignment is concerned with assigning documents their representative keywords, either through extracting directly from the text (keyword extraction), or some other process (keyword generation). The task plays an important role in important approaches to, for example, document indexing (for search engines), summarisation, clustering, and classification. Research in data-driven methods for automatically finding keywords for documents has been both classifier-based and rank-based. In recent years, the rank-based systems have come to dominate and appear to form the state-of-the-art under all forms of the keyword assignment task. However, with the move towards rank-based systems, there is a general sense that previously adopted techniques of system evaluation (especially set-based) are inadequate, as testified by the fact that virtually every new publication on the topic introduces some new evaluation metric or metric *twist*. In a nutshell, the problem stems from the added dimension of *ranking*, where the order of items in the list is meaningful, unlike classification systems.

Following the definition of rank-based keyword assignment systems and a discussion of the specification of the keyword assignment task in view of this definition (Section 2), we present a critical survey of the evaluation approaches to this task adopted in the past and attempt to highlight some crucial weaknesses with respect to the rank-based system approach (Section 3). We finish by arguing that the currently most consistent approach to evaluation in this context makes use of the standard Normalised Discounted Cumulative Gain (NDCG) metric, which is mathematically proven to distinguish between ranking systems where one system is substantially better than another (Section 3.3).

2 The definition of rank-based keyword assignment systems

We define rank-based keyword assignment as the following (Cf. (Wang *et al.*, 2013)).

Definition 1. Given a set of candidate keywords $\mathcal{X} = \{x_1, \dots, x_m\}$ for a document D and a set \mathcal{Y} of degrees of relevancy, a **rank-based keyword assignment system** $f : \mathcal{X} \rightarrow \mathcal{Y}$ generates a score $f(x) \in \mathcal{Y}$, according to which the m keywords in \mathcal{X} can be organised, resulting in the returned list $x_{i_1}^f, \dots, x_{i_m}^f$ ($i_j \in [m]$ for all $j \in [m]$ and $i_{j_1} \neq i_{j_2}$ if $j_1 \neq j_2$), which satisfies $f(x_{i_1}^f) \geq \dots \geq f(x_{i_m}^f)$.

The set \mathcal{Y} can, for instance, be the set of real numbers or the interval $[0, 1]$. It can also be finite. If we set $\mathcal{Y} = \{0, 1\}$, then we see that a classification-based system can be viewed as a simple type of rank-based keyword assignment system. Thus any ranking-system specific evaluation metric can be used for their evaluation also (though indeed this introduces excessive complexity if there is no specific comparison with more complex ranking-based systems).

The original keyword assignment problem does not ask for a *ranking* of keyword candidates, but a set of *correct* keywords. Keywords are short representations of documents ; and as a set they come in small numbers. One could think of this small set size as part of their definition. So, researchers of rank-based systems have generally resorted to returning the top n items of the ranked list. The question of how many keywords to return then arises, since rank-based systems have only organised the candidate keywords into a list (in which all possibilities appear).

Cutting the returned keyword list from a rank-based system off at n , simply because n is the number of positives (correct keyword candidates), seems to be too harsh : a cut-off at $n = 5$, when, say, the sixth and seventh keywords in the list are correct is not the full story. At the same time, a cut-off at $n = 78$ seems far too generous (to at least the recall score) and contrary to the definition of a keyword. Indeed, there seems to be some small upper limit on the size of a keyword set, though it is not clear what this is ; call this observation (O1).

Moreover, we need to be conscious of how much tolerance for error a user or down-stream application receiving the keyword set has. Perhaps precision of at least $\frac{1}{3}$ is tolerable, but $\frac{1}{5}$ becomes fairly useless. Obviously, the higher density and quantity of true positives, the better, but this bound is better determined by the down-stream application. There exists some reasonably low bound on the error tolerance of keyword assignment systems for down-stream applications, though it is not clear what this is ; call this observation (O2).

With the ranking approach to keyword assignment, as Liu et al. Liu *et al.* (2010) note, the ranking order of extracted keyphrases is an important indicator for method preference. We argue that this ranking order in Definition 1 completely specifies the task and accounts for (O1) and (O2) if we observe that producing correct keywords lower down the list in the ranking should be not as important as producing correct keywords high up in the list, as some sort of quality control. We are not sure what the keyword set size is, but we are aware that it should be small by (O1), therefore a decaying “reward” for finding a correct keyword as we move down the ranked list can account for this. Moreover, by (O2), we are aware that down-stream applications could probably afford some density of errors, but that this density should be small ; since the set of correct (gold) keywords is small, as we move down the ranked list, the density of errors probably increases. Once again a decaying “reward” for finding a correct keyword as we move down the ranked list can account for this. Call this observation (O3).

3 Previous system evaluation and their adequacy for rank-based system evaluation

Four general categories of approaches to rank-based keyword assignment system evaluation have been adopted in the past, each category differing in its selection of the parameter n : (1) low choice(s) of n , (2) choice of n as a function of document length, (3) considering all values of n as equal, and (4) oracle choice of n (i.e., the choice of n which maximises the evaluation metric). We discuss these now, in turn.

3.1 Selecting a strong list cut-off n : precision, recall and f-score

The majority of systems have been evaluated using the popular measures of precision (P), recall (R) and f-score (F_1), where $P := \frac{\text{correctly returned keywords}}{\text{returned keywords}}$, $R := \frac{\text{correctly returned keywords}}{\text{all correct keywords}}$, and $F_1 := 2 \cdot \frac{P \cdot R}{P + R}$.

3.2 Keyword set size at a constant cut-off

When using these set-based evaluation measures, researchers typically choose a set size n , and true to the definition of a keyword (keyword set), this parameter is usually chosen to be small.

(Wan & Xiao, 2008) and (Wan *et al.*, 2007) evaluate systems using precision, recall and f-score at $n = 10$ explaining that 10 is the limit, because the guidelines they set for the manual annotation of keywords of the DUC2001 documents gave a limit of 10 keywords. Semeval 2010 task 5 organisers evaluated submitted systems using precision, recall, and f-score, with $n \in \{5, 10, 15\}$ (Kim *et al.*, 2010). (Liu *et al.*, 2010) present precision, recall and f-scores for specific n values selected with respect to the dataset : $n = 5$ for the Inspec dataset and $n = 10$ for the DUC2001 dataset.¹ (Litvak

1. We note that the mean number of keywords in the Inspec training set documents is 9.788 with standard deviation of 4.877. Also, this number was found to be normally distributed with high probability ($K^2 = 127.384, p = 0.0$). Therefore, (Liu *et al.*, 2010)’s value for $n = 5$ cannot be motivated

& Last, 2008) are generous with the smallness boundary (perhaps unrealistically), reporting precision, recall and f-scores for $n \in \{10, 20, 30, 40\}$.

The decision of (Wan & Xiao, 2008) and (Wan *et al.*, 2007) to evaluate with n as the number of positives may be too harsh for a rank-based system, because it allows no error tolerance, which goes against (O2). Moreover, f-scores can be highly chaotic when n is so low.

Consider the hypothetical systems in Table 1, which shows the ranked keyword lists of each of the systems, where 0 is a false positive and 1 is a true positive, and there are a total of 7 positives in the data. The highest f-score is achieved at $n = 8$; however evaluating, as in the Semeval 2010 task 5 at $n \in \{5, 10, 15\}$ doesn't provided any evidence of this. In fact, at $n = 5$, Systems 1 and 2 are tied, and at $n \in \{10, 15\}$, all three systems have the same f-score. However, we can observe some behaviour of the three systems which clearly demarcates System 2 as generally superior given these results, since it finds the positives *earlier* than the two other systems. Unfortunately, f-score at arbitrary cut-offs cannot account for this.

n	System 1		System 2		System 3	
	kw	f-score	kw	f-score	kw	f-score
1	0	0	1	0.29	0	0
2	0	0	1	0.5	0	0
3	1	0.22	1	0.67	0	0
4	1	0.4	0	0.6	0	0
5	1	0.55	0	0.55	1	0.18
6	0	0.5	1	0.67	1	0.33
7	0	0.46	1	0.77	1	0.46
8	1	0.57	1	0.85	1	0.57
9	1	0.67	0	0.8	1	0.67
10	1	0.75	0	0.75	1	0.75
11	0	0.71	1	0.82	0	0.71
12	0	0.67	0	0.78	0	0.67
13	0	0.63	0	0.74	0	0.63
14	1	0.7	0	0.7	0	0.6
15	0	0.67	0	0.67	1	0.67

TABLE 1 – F-scores of hypothetical systems at various levels of n . kw stands for keyword. A value 0 in this column indicates a false positive at the corresponding rank level n , and a value 1 indicates a true positive. The total number of positives is set at 7.

3.2.1 Keyword set size as some fraction of document size

One approach to the manner in which n is chosen for evaluation using precision, recall and f-score is to let n be some fraction of the length of the document. (Mihalcea & Tarau, 2004) used the knowledge of the relatively short length N of the documents (which were abstracts from the Inspec corpus), and set $n := \frac{1}{3} \cdot N$. Clearly the approach of taking first one-third of the returned ranked list does not work for longer sized documents, where n could end up in the 1000s. But, that does not mean that N could not be used to guide the selection of n .

Abstraction made of the problem of the harsh cut-off outlined in the previous section, an additional problem remains. This approach assumes that there is some correlation between n and document size N . We tested this assumption for the training set of Inspec corpus. This training set consists of 1000 scientific abstracts, which form the document set, and as in all previous uses of the corpus (to our knowledge) we considered the set of keywords for abstracts, which were designated as *uncontrolled* (Hulth, 2003). We carried out the D'Agostino-Pearson test for normality on document length, which showed that N is highly likely to be normally distributed ($K^2 = 114.565, p = 0.0$). The number of keywords for document n was also found to be normally distributed by the same test for normality ($K^2 = 127.384, p = 0.0$). We then calculated the Pearson correlation of N and n , however, to discover that there is in fact no correlation between these variables ($R = 0.060$). As such, this method for choosing some appropriate n does not replace the previous one.

On a related note, (Liu *et al.*, 2009) adopt a similar approach to “parameter n problem” in keyword assignment system, reporting precision, recall and f-score for different top fractions of the returned keyword list size : $n := r \cdot m$, where, we recall, m is size of the complete returned ranked list of candidate keywords, and $r \in \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{4}{5}\}$. It is not possible to carry out regression and significance tests for such an approach, since different systems return different candidate lists. However, r is also a choice.

by the Inspec data.

3.2.2 Oracle selection of n (best parameter evaluation)

(Hasan & Ng, 2010) consider several different datasets (DUC2001, Inspec, NUS, and ICSI) comparing their re-implementations of four previously published systems with a simple baseline, giving oracle parameter settings— n for the specific system and dataset that maximises the f-score; n varies widely from one system to then next (as well as from one dataset to the next), and from $n = 9$ through to $n = 190$, to the length of the candidate list. This type of evaluation may seem like cheating in the unsupervised setting. And though we do not advocate solely reporting results for oracle parameter settings, we do note that such results are very informative in the following sense. They reveal how fast systems reach their own optimality. If n is large at a systems best f-score, then this is generally a failure of the system. However, if n is fairly small, this can be seen as a success of the system. Moreover, if a system outperforms others at a reasonably low n , it seem fair to say that the system performs best. We believe this is what was intended to be shown by the ROC curves and precision-recall curves of the following section (Section 3.2.3); though these attempts, we hope to show, are problematic.

3.2.3 Curves and summarising over all n

With arbitrary or informed selection of n becoming increasingly unsatisfactory, researchers have attempted to avoid this selection altogether, by sketching the curve over all values of n , and drawing conclusions from this curve by means of either taking the area under the curve (AUC) to obtain a single numerical representation, or discussing how one curve *dominates* another.

We hope to show that neither of these strategies is really appropriate for the evaluation of keyword assignment systems. In fact, the former strategy can already be refuted by use of the definition of a keyword (keyword set). The strategy of taking the area under a curve over all values of n treats these values as equal. However, values of metrics at large n should at least be less important than values of metrics at small n (by (O3)). Still, we discuss further faults of the two types of curves previously adopted for rank-based keyword assignment system evaluation : ROC curves and precision-recall curves.²

ROC curves. (Litvak & Last, 2008) calculate the AUC of the (average) ROC curve as a means of evaluation.

The ROC curve of a binary classifier plots the true positive rate (TPR) (on the y-axis) against the false positive rate (FPR) (on the x-axis) at incremental levels of n , where $TPR := R$ and $FPR := \frac{\text{false positives}}{\text{negatives}}$.

When using such an evaluation approach for rank-based systems, an immediate problem is therefore the question of true negatives. It is not obvious what can be used as a true negative of a rank-based system for keyword assignment systems. However, ignoring this fact, other problems with the AUC ROC metric persist.

There are some important weaknesses about this measure that are vital to understand for classifiers evaluation in general.³ Specifically for the case of keyword assignment systems, in addition to the weakness mentioned above for all curve-based metrics, a critical short-coming is the metrics proven inability to always determine the best system when ROC curves cross (which is likely to happen when systems have performances worth comparing) (Hand, 2009; Lobo *et al.*, 2008).

Precision-recall curves. A precision-recall curve plots recall on the x-axis and precision on the y-axis. (Hasan & Ng, 2010) and (Liu *et al.*, 2010) plot precision-recall curves and discuss curve dominance. However, (Davis & Goadrich, 2006) mathematically prove that an algorithm dominates in precision-recall space if and only if it dominates in ROC space. Therefore such an evaluation method is problematic for the reasons already outlined above.

3.3 Standard Normalised Discounted Cumulative Gain

The metric we propose in this paper, (standard) Normalised Discounted Cumulative Gain (NDCG), is already widely used in information retrieval and machine learning research on ranking. It is defined as follows for the keyword assignment task.

2. In fact there is a third curve in the literature. (Litvak & Last, 2008) provide a graph of cumulative AUC for the average precision, with respect to $n \in [1, 589]$. But this seems simply to be a way to plot precision as a smooth curve, which can really only be used to detect an optimum precision point. Therefore, we do not discuss this type of curve here, but refer to Section 3.1.

3. See (Hand, 2009; Lobo *et al.*, 2008) for a complete discussion.

Definition 2. Let f be a keyword ranking function and $S_i = \{x_{i,1}, \dots, x_{i,m_i}\}$ be the dataset of keyword candidates for the document D_i , with $|S_i| = m_i$. The **Discounted Cumulative Gain (DCG)** of f on S_i (yielding the ranked list $x_{j_1}^f, \dots, x_{j_{m_i}}^f$) is defined as

$$DCG(f, S_i) := \sum_{t=1}^{m_i} \frac{\mathbb{I}\{x_{j_t}^f \text{ is a correct keyword}\}}{\log(1+t)}.$$

The **Ideal DCG** of f on S_i is defined as

$$IDCG(f, S_i) = \max_{f'} DCG(f', S_i).$$

The **NDCG** of f on S_i is defined as

$$NDCG(f, S_i) := \frac{DCG(f, S_i)}{IDCG(f, S_i)}.$$

For system evaluation, one would present the average NDCG over all documents.

We observe from the definition that the evaluation metric is in line with (O3), having a decaying reward of success with respect to rank : $\frac{1}{\log(1+\text{rank})}$. Moreover, an important result on standard NDCG is that every two substantially different⁴ ranking functions are *consistently distinguishable*⁵ by standard NDCG (Wang *et al.*, 2013).⁶ This makes the metric attractive in and of itself.

As illustration, let us consider again our hypothetical systems from Table 1, which were not always distinguishable using the precision, recall and f-score evaluations at $n \in \{5, 10, 15\}$. Their NDCG scores are given in Table 2. We see that the NDCG metric is reflective of our observations on the system performance : System 2 is best, followed by System 1, and last System 3. For further analysis, we can look at the best-parameter f-scores. We see that System 2 achieves its optimal with $n = 8$ (and precision 6/8 and recall 6/7), which further assures us that the system is also performing at its optimal with a good size n . We admit that this is a toy example, but it is only meant for illustration of the concepts and discussion of this paper and not as their proof. For a proof, the reader is referred to, for example, (Wang *et al.*, 2013).

System 1	System 2	System 3
0.681	0.939	0.613

TABLE 2 – NDCG scores for the three hypothetical systems.

On a side note, Liu *et al.* (Liu *et al.*, 2010) also introduce two new metrics for keyword evaluations : *mean reciprocal rank* and *binary preference measure*. These latter two metrics are meant to account for the ranking order of extracted keyphrases. Unfortunately, for the binary preference measure, the same n parameter must be chosen by the evaluator and for the mean reciprocal rank, only the rank of the first positive keyword in the ranked lists is accounted for. Therefore, we do not consider these as appropriate measures for keyword assignment.

4 Conclusion

Evaluation metrics should fit the task at hand. We hope to have shed some light on how the keyword assignment task should be re-specified under the rank-based approach. In doing so, we have been able explain some important weaknesses of the numerous pre-existing approaches to keyword assignment system evaluation, and motivate (and illustrate) an ideal evaluation metric : average standard NDCG.

Références

DAVIS J. & GOADRICH M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, p. 233–240, New York, NY, USA : ACM.

4. Roughly “substantially different” means “not almost always the same” (Wang *et al.*, 2013).

5. A function $neg : \mathbb{R} \rightarrow \mathbb{R}$ is negligible if for all c , $neg(N) < N^{-c}$, for sufficiently large N . So, roughly, two ranking systems are “consistently distinguishable” by some metric if there exists some negligible function $neg(N)$, that shows preference for one system over the other, with probability $1 - neg(N)$ when keyword candidate lists are at least as big as N (Wang *et al.*, 2013).

6. See Wang *et al.* (Wang *et al.*, 2013) for details.

- HAND D. J. (2009). Measuring classifier performance : A coherent alternative to the area under the roc curve. *Machine Learning*, **77**, 145–151.
- HASAN K. S. & NG V. (2010). Conundrums in unsupervised keyphrase extraction : making sense of the state-of-the-art. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, COLING '10, p. 365–373, Stroudsburg, PA, USA : Association for Computational Linguistics.
- HULTH A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, p. 216–223, Stroudsburg, PA, USA : Association for Computational Linguistics.
- KIM S. N., MEDELYAN O., KAN M.-Y. & BALDWIN T. (2010). Semeval-2010 task 5 : Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, p. 21–26, Uppsala, Sweden.
- LITVAK M. & LAST M. (2008). Graph-based keyword extraction for single-document summarization. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, MMIES '08, p. 17–24, Stroudsburg, PA, USA : Association for Computational Linguistics.
- LIU Z., HUANG W., ZHENG Y. & SUN M. (2010). Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, p. 366–376, Stroudsburg, PA, USA : Association for Computational Linguistics.
- LIU Z., LI P., ZHENG Y. & SUN M. (2009). Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1 - Volume 1*, EMNLP '09, p. 257–266, Stroudsburg, PA, USA : Association for Computational Linguistics.
- LOBO J. M., JIMENEZ-VALVERDE A. & R. R. (2008). Auc : a misleading measure of the performance of predictive distributive models. *Global Ecol. Biogeogr.*, **17**, 145–151.
- MIHALCEA R. & TARAU P. (2004). Textrank : Bringing order into text. In *Proceedings of EMNLP*, p. 404–411.
- WAN X. & XIAO J. (2008). Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2*, AAAI'08, p. 855–860 : AAAI Press.
- WAN X., YANG J. & XIAO J. (2007). Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *ACL*.
- WANG Y., WANG L., LI Y., HE D. & LIU T.-Y. (2013). A theoretical analysis of ndcg type ranking measures. In *Proceedings of COLT*, p. 25–54.

Effects of Graph Generation for Unsupervised Non-Contextual Single Document Keyword Extraction

Natalie Schluter

Center for Language Technology, University of Copenhagen, Copenhagen, Denmark
natschluter@hum.ku.dk

Abstract. This paper presents an exhaustive study on the generation of graph input to unsupervised graph-based non-contextual single document keyword extraction systems. A concrete hypothesis on concept coordination for documents that are scientific articles is put forward, consistent with two separate graph models : one which is based on word adjacency in the linear text—an approach forming the foundation of all previous graph-based keyword extraction methods, and a novel one that is based on word adjacency modulo their modifiers. In doing so, we achieve a best reported NDCG score to date of 0.431 for any system on the same data. In terms of a *best parameter* f-score, we achieve the highest reported to date (0.714) at a reasonable ranked list cut-off of $n = 6$, which is also the best reported f-score for any keyword extraction or generation system in the literature on the same data. The best-parameter f-score corresponds to a reduction in error of 12.6% conservatively.

1 Introduction

Recent and state-of-the-art approaches to unsupervised non-contextual single document keyword extraction typically work on some sort of graph of the input text, formed with respect to word order. The graphs generally pose word-forms as nodes and place edges between words so long as their proximity in the linear text is within some threshold d ; however the characteristics of the edges, and the deletion or association of certain nodes (in both pre- and post-processing) vary from one approach to the other and no exhaustive study of these choices has been motivated or exhaustively tested using a single *relevant* measure. Similarly, the use of the linear order of words in the text as the basis for the edge relations in text graphs has been loosely associated with a linguistic syntax motivation, however no more sophisticated accounting of syntactic relations has been attempted to date. The work presented in this paper partially bridges this gap.

We present an exhaustive study on the generation of graph input to unsupervised graph-based non-contextual single document keyword extraction systems. We consider the question of graph motivation, and put forward a concrete hypothesis on concept coordination for documents that are scientific articles. Corresponding to the requirements of the graph model, we consider two types of relations between words for such a graph, one which is based on word adjacency in the linear text—an approach forming the foundation of all previous graph-based keyword extraction methods, and a novel one that is based on word adjacency modulo their modifiers. In doing so, we achieve a best reported NDCG score to date of 0.431 for any system on the same data. In terms of a *best parameter* f-score, we achieve the highest reported to date (0.714) at a reasonable ranked list cut-off of $n = 6$, which is also the best reported f-score for any keyword extraction or generation system in the literature on the same data. The f-score corresponds to a reduction in error of 12.6%, or even more if we set both systems to a ranked list cut-off of $n = 6$ (since the previous best f-score was achieved at a best parameter of $n = 9$). (This latter score is also reproduced in Table 1.)

Following some preliminaries on the definition of the task, we discuss previous work in unsupervised non-contextual single document keyword extraction (Section 2). We present the graph models we investigate, in Section 3, which is the main contribution of this paper. Section 4 reviews the centrality measures used. Finally, we present the evaluation of the resulting systems (Section 5), followed by a brief discussion of conclusions and open problems (Section 6).

2 Preliminaries

We identify two broad types of *single document keyword extraction* (SDKE). *Contextual SDKE* makes use of the document set to which the relevant document belongs, and in which there are similar documents; other information outside of the

document set may also be used in some types of contextual SDKE. *Non-contextual SDKE* makes use of only the relevant document with no other information. The latter does not necessarily make the assumption of independence of documents in general. In fact, non-contextual SDKE is important for the case of isolated documents (not part of a document set), as well as for documents for which relevant supplementary information may be non-existent or unreliable.

Undirected Graphs A simple *graph* G is a pair (V, E) , where V is the set of vertices and $E \subseteq V \times V$ is the set of edges (where the pairs of vertices are unordered). The edge uv is said to be *incident* with the vertices u and v . A *multi-graph* is a graph where there may be more than one edge between two vertices (the edge set is a multi-set). The *degree* $\deg(v)$ of a vertex v is then the number of distinct edges with which it is incident. A *walk* of length k from vertex u to vertex v is a sequence of k edges, $v_1v_2, v_2v_3, \dots, v_{k-1}v_k, v_kv_{k+1}$ and a *path* from u to v is a walk from u to v where no edge is repeated. Finally, a *complete graph* is a graph with all $|V|(|V| - 1)/2$ possible edges, and a *clique* is a subgraph that is complete.

2.1 Previous Work

Published work including some discussion of text graph representations for non-contextual SDKE has considered only (1) the directed-/undirected-ness of edges on stop-word filtered graphs and (2) different proximity thresholds d for placing these edges between word nodes (Mihalcea & Tarau, 2004; Litvak & Last, 2008; Litvak *et al.*, 2013; Rose *et al.*, 2010; Schluter, 2014; Boudin, 2013).¹ The proximity threshold of $d = 1$ was found by (Mihalcea & Tarau, 2004) to perform best, and all other research on the task has consistently maintained this threshold; the present work follows suit in that respect.

In terms of the unsupervised non-contextual single document keyword extraction task itself, (Mihalcea & Tarau, 2004) had the pioneering work, also introducing graph-based techniques for this task with the application of PageRank (Page *et al.*, 1999). (Litvak & Last, 2008) follow this approach, but apply HITS instead (on a different dataset). Finally, (Rose *et al.*, 2010) observed that using the simple degree of a vertex in the network produced what were at the time state-of-the-art results (with precision 0.337, recall 0.415 and f-score 0.372, at a ranked list cut-off of $n = \frac{1}{3}N$, where N is the number of words in the document, on the Inspec corpus); the technique was later re-discovered by (Litvak *et al.*, 2013) and (Boudin, 2013).

3 Graph generation and other pre-processing

In this section we present the graph model of the document text as well as two consistent instances of this model : adjacency graphs and parse graphs.

3.1 Graph Model

We follow the document model proposed in (Schluter, 2014) for keyword extraction, which proposes a graph model from the point of view of document *synthesis* (as opposed to the document *analysis* model proposed by (Mihalcea & Tarau, 2004)). In generating scientific text on a given topic (or given related topics), the “author” may require other concepts to regularly support the discussion (for example, definitions or explanations); this is a sort of concept coordination. Two basic assumptions are adopted about this concept coordination in the model. The first assumption is that the author is communicating in the most efficient manner possible, and that supporting concepts are named only when absolutely necessary. The second assumption is that in supporting or defining a concept, textual mention of a topic concept and supporting concepts should occur rather “close” to each other, in terms of the linear order of concepts (words) in the texts. These concept support relations are approximated therefore by co-occurrence relations—relations that are essentially symmetric (undirected): there is no clear order that should be observed between topic concepts and supporting concepts within a single sentence (or over several sentences for that matter). We note that the network is *not* the meaning of the documentation; rather it is a *representation of its construction*. Flow through the concept network is seen as *communicative*—concept-building on the part of the author for the reader.

1. (Litvak & Last, 2008) motivate their choice of directed graphs by the extensive clustering and classification results-driven graph study presented in (Schenker *et al.*, 2005), but do not motivate the choice with respect to the keyword extraction task they undertake.

As such, we model text as an *undirected* graph, where vertices are words appearing in the text and edges model the concept coordination relationships discussed above. There are many methods of producing (undirected) edges for our graph that are compatible with the model described above. We consider two plausible ones for this paper. In Section 3.3 we describe a graph model similar to that of (Mihalcea & Tarau, 2004; Litvak & Last, 2008; Rose *et al.*, 2010; Litvak *et al.*, 2013) and in Section 3.4 we propose a novel graph model created out of parse graphs of document sentences. First we discuss the pre-processing of the text carried out prior to graph construction, as well as graph parameters that common to both types of graphs.

3.2 Pre-processing and common graph parameters

Preprocessing. For both main types of models, we first carry out sentence detection, tokenisation and part-of-speech tagging on the corpus, using the Stanford POS Tagger (Toutanova *et al.*, 2003). We remove all punctuation from individual sentences.

Filtering out stop-words. For both main models, we construct reduced and full graphs. Their exact manner of construction is specific to the graph type (adjacency or parse) (Cf. Sections 3.3 and 3.4).

1. The **reduced** graph contains the text stripped of stop-words, in order to have edges reflect relationships between semantically full words more directly, resulting in a denser graph. For the remaining non-stop-words, words of the same form and part-of-speech are merged into a single node.
2. The **full** graph is constructed from the full text. However, it contains a special type of node for stop-words. Nodes decorated with distinct non-stop-words of the same form and part-of-speech are still merged into a single node, but nodes decorated with stop-words are never merged, resulting in a sparser graph. The centrality measure, rather than the pre-processing is left with the full burden of ranking the important words. Stop-words are generally words that occur frequently in text, so by not merging identical ones into single nodes, we hope to prevent the centrality measure of choice from finding these units important.

Edge multiplicity. For both adjacency graphs and parse graphs we test their **simple** and **multi**-graph versions. Edges from both graphs reflect of course relations between words, but the multi-graph versions are meant to also reflect frequency of concept coordination.

Note however that for some centrality measures, there is no difference between multi- and simple graphs (Cf. Section 4).

3.3 Document adjacency graphs

Document adjacency graphs model text linear relationships between words (i.e., that they are beside or close to each other in the text). As such, for full (reduced) document adjacency graphs, an edge between two words is added to the graph if these two words are adjacent in the (stop-word filtered) text. There are therefore four different document adjacency graph models that we investigate, considering all common graph parameters combinations.

The generated reduced and full adjacency text graphs for Ex 1 below are given in the top of Figure 1.

- (Ex 1) Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types.²

3.4 Document parse graphs

It is straightforward that a better approximation of concepts can be achieved by first organising sub-strings of a sentence into units observing the communicative flow between units and sub-units. For English, the dependency tree syntactic re-

2. This is abstract 1939 from the test files in the Inspec corpus, first used as an example in (Mihalcea & Tarau, 2004).

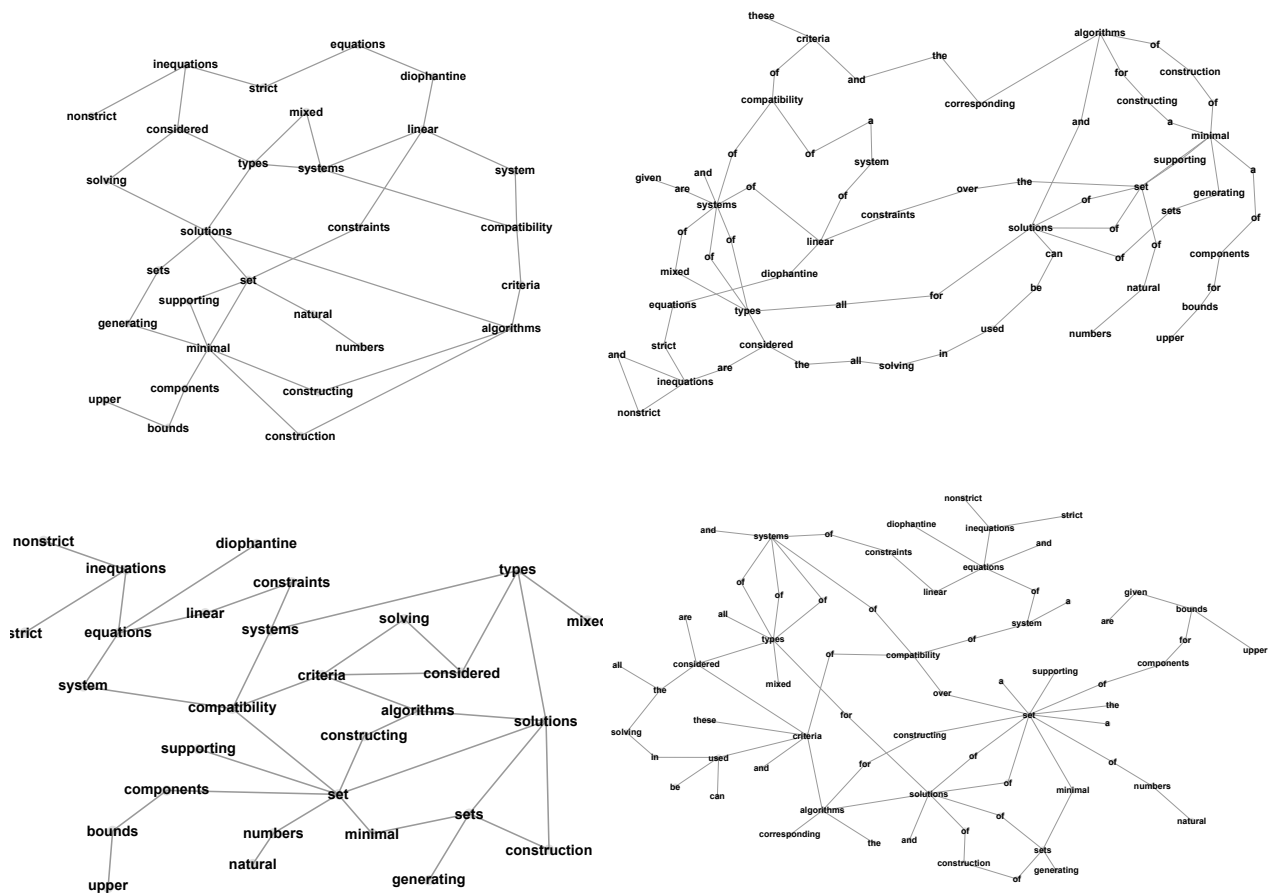


FIGURE 1 – Simple adjacency graphs (top), basic parse graphs (bottom) for Ex 1. Reduced graphs are on the left and the full graphs on the right.

presentation of sentences provides exactly the organisation of a sentence that we need to model this. The set of dependency trees of a document are organised into what we refer to as the *document parse graph*.

Following preprocessing (and before any stop-word filtering), the text is sent through the Stanford parser to obtain the associated dependency parses (basic dependency conversion) (de Marneffe *et al.*, 2006). This yields two additional graph models that we investigate (considering all other common graph parameter combinations).

The construction of full parse graphs is similar to that of adjacency graphs. The construction of reduced parse graphs is slightly more involved. By contracting stop-word vertices some information can potentially be lost : the children of stop-words become closer to their original grandparent than they are to each other, which is not the intended model. In an attempt at circumventing this effect, we first create a clique out of the children of stop-words before contraction.

For Ex 1, the generated reduced and full parse are given at the bottom of Figure 1.

3.5 Post-processing.

We carry out similar post-processing to (Mihalcea & Tarau, 2004). That is, sequences of adjacent keywords from the text are possibly collapsed into a multi-word keyword, depending on their scores. We score a multi-unit keyword by the maximum score of words they are composed with ; we also tried the using the average score of word components, but with worst performance and so do not report these scores here. This yields a candidate list where there may be unit overlaps in keywords. We therefore test an extra post-processing step which keeps only the keyword with the highest score among two overlapping keywords (this corresponds to **excl** (as opposed to **incl**) in results Tables 1-3). Ties are broken with a

preference for longer keywords ; moreover, the proposed keywords must not start or end with a stop-word, and must be grounded in a noun (i.e., the rightmost word of a multi-word keyword must be a noun). Keywords consisting of at most three words are considered.

4 Centrality measures

(Schluter, 2014) investigates seven different centrality measures for ranking nodes in undirected graphs, two of which corresponded to previously published state-of-the-art approaches to non-contextual SDKE among the centrality measure classes of degree-like centrality, closeness centrality and betweenness centrality, outlined by (Borgatti & Everett, 2006). This work showed that closeness centrality measures, which rank nodes according to their general (minimum) distance to other nodes in the graph, did not perform as well as degree-like centrality measures or betweenness centrality measures. In this paper we consider the performance of systems based on different graphs across the three betweenness centrality measures as well as the degree-like centralities.

The *degree centrality* of a vertex v in a graph G is simply its degree $\deg(v)$. Within the context of text graphs, this is a measure of how much of a first-hand support a text vertex (concept) is for other text vertices (concepts).

The *eigenvector centrality* is essentially the deterministic version of the PageRank algorithm (Page *et al.*, 1999) for *undirected* graphs, as well as the output of the HITS algorithm (for *directed* graphs) upon convergence (provided all eigenvalues are distinct) (Kleinberg, 1999). The eigenvector centrality of a node $v_i \in V(G)$, $C_{EI}(v_i)$ is found by calculating the principal eigenvector of the adjacency matrix for the graph. The i th entry in this vector is $C_{EI}(v_i)$. To bypass connectivity issues, we use the PageRank “teleportation trick”, transforming the input graph into a complete graph, simply incrementing the weight of all possible edges by 1.

The *betweenness centrality* of a vertex quantifies how often a node acts as a bridge along the shortest path between two other nodes. In the context of our text graph, the betweenness centrality can be seen as a measure of how the presentation of a scientific subject must employ a given word (concept) as support when moving the discussion between two different concepts. We consider three different betweenness centrality measures.

The (*normalised*) *betweenness centrality* $C_B(x)$ for vertex x is defined as $C_B(x) := \sum_{s \in V(G)} \sum_{t \in V(G)} \frac{\sigma_{st}(x)}{\sigma_{st}}$, where σ_{st} is the number of shortest paths between nodes s and t .³

$C_B(x)$ gives more weight to pairs of vertices at a larger distance from each other. If one wishes to consider all shortest paths to contribute the same weight, one approach is to normalise by the shortest distance between s and t , which yields *length-scaled betweenness centrality*, $C_{LSB}(x) : C_{LSB}(x) := \sum_{s \in V(G)} \sum_{t \in V(G)} \frac{\sigma_{st}(x)}{d(s,t)\sigma_{st}}$.

Finally, the *distance-weighted fragmentation* $C_{DWF}(x)$ of vertex x measures the fragmentation of a graph if we took x out of it. It is defined as $C_{DWF}(x) := C_{DWF}(G - x) - C_{DWF}(G)$, where $C_{DWF}(G) := 1 - \frac{2 \sum_{i \neq j} \frac{1}{d(i,j)}}{n(n-1)}$. Note that $G - x$ (the graph obtained from G by removing vertex x and any edges incident to x) should be more fragmented than G . (We also shift all scores, so that they are positive.)

Note that by these definitions, there will be no difference in betweenness centrality measure results on simple graphs versus multi-graphs.

5 Experiments and Evaluation

We carry out our experiments on the test set from the Inspec abstract corpus (Hulth, 2003) consisting of 500 abstracts for scientific articles, along with the *uncontrolled* corresponding keywords.

We evaluate the systems across the variety of graph inputs in terms of average standard Normalised Discounted Cumulative Gain (NDCG). We also provide best parameter (for ranked list cut-off n) precision, recall and f-score, which informs us of when a system reaches its optimality, rather than providing any general system evaluation. Document keyword sets are relatively small, but not too small, so if n is too small or too large when it reaches a good optimal f-score, the system cannot necessarily be considered successful. On the other hand, if for example $n \in \{5, \dots, 10\}$ and it reaches a global optimum among all systems, it is easier to argue this to be a success.

The results are reported in Tables 1 through 3. We observe that best scores according to both metrics are generally achieved by parse graphs, suggesting that the parse graph model is superior to the simple adjacency graph model.

3. In fact, the normalised version of betweenness centrality normalises $C_B(x)$ by the number of pairs of nodes in the graph. However, since we are not comparing two different graphs, but only two different nodes of the same graph, this expression of betweenness centrality has the same power as its normalised version.

We observe that degree centrality has the best NDCG score of 0.431, for the parse full graph. We note that these NDCG scores differ from those of (Schluter, 2014) which were erroneous due to a bug in the evaluation software. The best f-score of 0.714 among all models is achieved by the parse graph under the distance-weighted fragmentation measure, at a cut-off of $n = 6$, which is very reasonable for this task ; however, curiously this model (and measure) achieves a relatively poor NDCG score, which indicates that after this ideal cut-off, the ranking system fails.

With these degree-centrality results, we can observe differences between simple and multi-graphs, and more so for reduced graphs than for full graphs. This makes sense since we never merge stop-word nodes in full graphs and thereby account for a type of co-occurrence frequency via stop-words. Still the differences in simple and multi-graph scores are relatively small, perhaps contrary to intuition. We hypothesis this to be the result of the nature of the document set in question ; the documents are very short and therefore contain less repeat co-occurrences.

graph	pre-p	post-p	n	prec	rec	f1	NDCG
adj	reduced	incl	14	0.357	0.625	0.455	0.418
		excl	13	0.308	0.5	0.381	0.390
	full	incl	18	0.333	0.75	0.462	0.410
		excl	12	0.333	0.5	0.4	0.383
	parse	reduced/	6	0.833	0.625	0.714	0.406
		full	5	0.6	0.375	0.462	0.374

graph	pre-p	post-p	n	prec	rec	f1	NDCG
adj	reduced	incl	8	0.5	0.5	0.5	0.399
		excl	2	1.0	0.25	0.4	0.386
	full	incl	8	0.5	0.5	0.5	0.398
		excl	2	1.0	0.25	0.4	0.385
parse	reduced	incl	8	0.5	0.5	0.5	0.414
		excl	2	1.0	0.25	0.4	0.403
	full	incl	8	0.5	0.5	0.5	0.412
		excl	2	1.0	0.25	0.4	0.401

TABLE 1 – Distance-weighted fragmentation results (left). Betweenness centrality results (right). The scores for multi-graph are precisely the same.

graph	pre-p	post-p	n	prec	rec	f-score	NDCG
adj	reduced	incl	14	0.429	0.75	0.545	0.397
		excl	11	0.364	0.5	0.421	0.378
	full	incl	11	0.455	0.625	0.526	0.395
		excl	2	1.0	0.25	0.4	0.375
	parse	reduced	12	0.5	0.75	0.6	0.417
		excl	10	0.4	0.5	0.444	0.406
parse	full	incl	11	0.455	0.625	0.526	0.417
		excl	11	0.364	0.5	0.421	0.408

graph	pre-p	post-p	n	prec	rec	f1	NDCG
adj	reduced	incl	20	0.3	0.75	0.429	0.419
		excl	15	0.267	0.5	0.348	0.395
	full	incl	19	0.316	0.75	0.444	0.413
		excl	14	0.286	0.5	0.363	0.388
parse	reduced	incl	13	0.385	0.625	0.476	0.426
		excl	17	0.235	0.5	0.32	0.376
	full	incl	17	0.353	0.75	0.48	0.419
		excl	14	0.286	0.5	0.364	0.367

TABLE 2 – Length scaled betweenness centrality results (left). The scores for simple and multi-graphs are precisely the same. Eigenvector centrality results (right).

edge	pre-p	post-p	n	prec	rec	f1	NDCG
mult							
simple	reduced	incl	5	0.6	0.375	0.462	0.423
		excl	14	0.286	0.5	0.363	0.401
	full	incl	5	0.6	0.375	0.462	0.415
		excl	14	0.286	0.5	0.363	0.392
multi	reduced	incl	19	0.316	0.75	0.444	0.423
		excl	14	0.286	0.5	0.363	0.402
	full	incl	19	0.316	0.75	0.444	0.418
		excl	14	0.286	0.5	0.363	0.396

edge	pre-p	post-p	n	prec	rec	f1	NDCG
mult							
simple	reduced	incl	11	0.455	0.625	0.526	0.431
		excl	9	0.333	0.375	0.353	0.381
	full	incl	5	0.6	0.375	0.462	0.421
		excl	16	0.25	0.5	0.333	0.370
multi	reduced	incl	11	0.455	0.625	0.526	0.427
		excl	9	0.333	0.375	0.353	0.382
	full	incl	5	0.6	0.375	0.462	0.419
		excl	16	0.25	0.5	0.333	0.372

TABLE 3 – Degree centrality results for adjacency graphs (left) and parse graphs (right).

6 Conclusions and open questions

We have introduced a novel parse text graph for the representation of documents that is shown to perform better in non-contextual single document keyword extraction, producing the highest reported NDCG score to date, as well as the highest best parameter f-score. In our opinion this model is more language independent than the adjacency graph document model, as it relies slightly less on sentential linear order ; this is an open question for future investigation. In addition, the question as to whether the multi-graph version of document graph models helps systems when the input are larger documents remains open ; for smaller documents the answer seems to be negative.

Références

- BORGATTI S. P. & EVERETT M. G. (2006). A graph-theoretic perspective on centrality. *Social Networks*, **28**(4), 466 – 484.
- BOUDIN F. (2013). A comparison of centrality measures for graph-based keyphrase extraction. In *Proc. of IJCNLP 2013*, Nagoya, Japan.
- DE MARNEFFE M.-C., MACCARTNEY B. & MANNING C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, p. 449–454.
- HULTH A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, p. 216–223, Stroudsburg, PA, USA : Association for Computational Linguistics.
- KLEINBERG J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, **46**(5), 604–632.
- LITVAK M. & LAST M. (2008). Graph-based keyword extraction for single-document summarization. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, MMIES '08, p. 17–24, Stroudsburg, PA, USA : Association for Computational Linguistics.
- LITVAK M., LAST M. & KANDEL A. (2013). Degext : a language-independent keyphrase extractor. *J. Ambient Intelligence and Humanized Computing*, **4**(3), 377–387.
- MIHALCEA R. & TARAU P. (2004). Textrank : Bringing order into text. In *Proceedings of EMNLP*, p. 404–411.
- PAGE L., BRIN S., MOTWANI R. & WINOGRAD T. (1999). *The PageRank Citation Ranking : Bringing Order to the Web*. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.
- ROSE S., ENGEL D., CRAMER N. & COWLEY W. (2010). *Automatic Keyword Extraction from Individual Documents*, In *Text Mining. Applications and Theory*, p. 1–20. John Wiley and Sons, Ltd.
- SCHENKER A., LAST H. & KANDEL M. (2005). *Graph-Theoretic Techniques for Web Content Mining*, volume 62 of *Series in Machine Perception and Artificial Intelligence*. World Scientific.
- SCHLUTER N. (2014). Centrality measures for non-contextual graph-based unsupervised single document keyword extraction. In *Proc. of TALN 2014*.
- TOUTANOVA K., KLEIN D., MANNING C. D. & SINGER Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, Stroudsburg, PA, USA : Association for Computational Linguistics.

Adaptation par enrichissement terminologique en traduction automatique statistique fondée sur la génération et le filtrage de bi-segments virtuels

Christophe Servan^{1,2} Marc Dymetman²

(1) LIG équipe GETALP, 41 rue des mathématiques, BP 53 38041 Grenoble Cedex 9

(2) Xerox Research Centre Europe, 6 chemin de Maupertuis, 38240 Meylan

christophe.servan@imag.fr, marc.dymetman@xrce.xerox.com

Résumé. Nous présentons des travaux préliminaires sur une approche permettant d'ajouter des termes bilingues à un système de Traduction Automatique Statistique (TAS) à base de segments. Les termes sont non seulement inclus individuellement, mais aussi avec des contextes les englobant. Tout d'abord nous générons ces contextes en généralisant des motifs (ou patrons) observés pour des mots de même nature syntaxique dans un corpus bilingue. Enfin, nous filtrons les contextes qui n'atteignent pas un certain seuil de confiance, à l'aide d'une méthode de sélection de bi-segments inspirée d'une approche de sélection de données, précédemment appliquée à des textes bilingues alignés.

Abstract.

Statistical machine translation adaptation through terminological enrichment based on virtual phrase generation and filtering

We propose a technique for adding bilingual terms to a phrase-based SMT system which includes not only individual words, but also induces phrasal contexts around these words. We first generate these contexts by generalizing patterns observed for similar words in a bilingual corpus, but then filter out those contexts that fall below a certain confidence threshold, based on an original phrase-pair selection process inspired by existing sentence selection techniques.

Mots-clés : Traduction Automatique Statistique, Génération Automatique de Texte, contexte phrastique, terminologie bilingue.

Keywords: Statistical Machine Translation, Natural Language Generation, phrasal context, bilingual terminology.

1 Introduction

La plupart des recherches concernant l'*adaptation* au domaine en Traduction Automatique Statistique (TAS) se basent sur un entraînement à partir d'un corpus bilingue de ce domaine (le plus souvent conjointement avec un corpus hors-domaine plus important, voir section 5) mais n'utilisent pas directement de terminologie spécifique à ce domaine. Par exemple, la traduction en français du mot anglais « layer » sera « calque » dans le domaine graphique, mais « couche » dans le domaine informatique ou encore « niveau » si on est dans un contexte administratif.

Une solution est d'ajouter un dictionnaire bilingue spécifique à la table de bi-segments du modèle de traduction. Mais le fait d'ajouter des mots individuels sans aucun contexte (c'est-à-dire des mots individuels par opposition à des segments contenant ces mots) ne représente pas une utilisation optimale des capacités d'un système de TAS à base de segments [TAS-BS = PB-SMT en anglais] : un système limité à des bi-segments unigramme/unigramme est typiquement inférieur à un système utilisant des bi-segments plus étendus.

Ce type d'approche fondée sur l'enrichissement lié à une terminologie n'a que peu d'influence sur les résultats, de manière générale. Cependant, dans l'optique de fournir une traduction à post-éditer à un traducteur, ce type d'enrichissement est précieux et permet de gagner du temps de post-édition. Cette approche est donc appropriée dans le cadre d'un processus de traduction assistée par ordinateur.

La proposition principale de cet article consiste à ajouter des entités nommées provenant d'un dictionnaire bilingue à la table de bi-segments, non directement, mais en reconstituant leurs contextes potentiels. Ces contextes potentiels sont obtenus

nus par copie de contextes entourant des entités nommées similaires représentées dans la table originelle de bi-segments. Les bi-segments “virtuels” ainsi produits peuvent être sur-générés, c’est pourquoi nous proposons de leur appliquer certaines techniques de filtrage. Les travaux préliminaires présentés dans cet article proposent d’évaluer l’approche proposée sur une terminologie associée aux entités nommées, dans l’optique de l’étendre dans de futurs travaux à d’autres types de mots (noms, adjectifs...)

2 Approche

Notre approche pour générer les contextes peut se décrire à l’aide de la « règle de déduction » suivante :

$$\frac{\alpha\beta\gamma \leftrightarrow_{pt} \alpha'\beta'\gamma'; \delta \leftrightarrow_{lex} \delta'; \beta, \delta : T; \beta', \delta' : T'}{\alpha\delta\gamma \leftrightarrow_{tbg} \alpha'\delta'\gamma'} \quad (1)$$

Dans cette règle, chaque lettre grecque dénote une suite de mots, et T, T' sont des types ; pt est la table de bi-segments (standard) originelle extraite du corpus bilingue. lex est une table de traduction fournie par l’utilisateur qui définit les correspondances terminologiques du domaine. Quant à tbg , c’est la table de bi-segments « généralisés » obtenue par application des déductions.

La règle dit que si $\alpha\beta\gamma \leftrightarrow_{pt} \alpha'\beta'\gamma'$ est une entrée dans la table originelle et si $\delta \leftrightarrow_{lex} \delta'$ est une correspondance terminologique, où β et δ sont du même type T (resp. β', δ' et T'), alors nous pouvons générer une nouvelle entrée $\alpha\delta\gamma \leftrightarrow \alpha'\delta'\gamma'$ où δ remplace β et respectivement δ' remplace β' . L’entrée nouvellement générée à l’aide de cette règle de déduction est alors ajoutée à la table de bi-segments généralisée ($\alpha\delta\gamma \leftrightarrow_{tbg} \alpha'\delta'\gamma'$).

2.1 Exemple : Entités Nommées

Nous illustrons notre approche par un exemple de type terminologique : les noms de pays (c.-à-d. une classe d’entités nommées). Supposons que certains pays soient rarement ou pas du tout mentionnés dans notre corpus d’entraînement anglais-français, mais que nous ayons un dictionnaire qui nous donne leurs traductions lexicales. Par exemple, « Ecuador » apparaît environ seulement 100 fois dans le corpus Europarl (Koehn, 2005), alors que « Germany » apparaît environ 60 fois plus souvent. Nous pouvons considérer que les contextes linguistiques observés autour des pays peu représentés sont trop peu nombreux pour être fiables, et notre méthode consiste à tenter de transposer les contextes concernant les pays bien représentés aux pays peu représentés.

La première étape de notre approche consiste à identifier les noms de pays dans le corpus d’apprentissage. Une fois que les entités nommées de type pays sont identifiées, nous les remplaçons par un « marqueur » comme indiqué dans la Table 1. Pour plus d’efficacité, le processus d’extraction de patrons est effectué sur les bi-segments déjà extraits du corpus d’entraînement. Avec notre exemple « Ecuador », nous générons un nouveau bi-segment en remplaçant dans le patron $@COUNTRY@$ is ||| L’ $@COUNTRY@$ est le marqueur source avec le terme « Ecuador » et le marqueur cible par sa traduction « Équateur ».

England is		L’ Angleterre est	@COUNTRY@ is		L’ @COUNTRY@ est
Spain is		L’ Espagne est	Ecuador is		L’ Équateur est
Italy is		L’ Italie est			
@COUNTRY@ is		L’ @COUNTRY@ est			

TABLE 1 – Exemple d’extraction du patron $@COUNTRY@$ is ||| L’ $@COUNTRY@$ est (tableau de gauche) et de son application pour un même couple terminologique « Ecuador : Équateur ». (tableau de droite)

Le processus de génération de nouvelles entrées ($\alpha\delta\gamma \leftrightarrow \alpha'\delta'\gamma'$) peut générer des erreurs, au cas où les contextes virtuels générés ne sont pas compatibles avec les termes considérés.

Ainsi, plusieurs des segments virtuels générés pour le terme « Ecuador », illustrés dans le tableau 2 sont erronés : en français, l’article « Le » doit être éliminé en « L’ » devant une voyelle. Un grand nombre de problèmes de ce type peuvent apparaître et c’est pourquoi un processus de filtrage doit être appliqué. Ce processus de filtrage est une contribution centrale de cet article décrit dans la section 3.

@COUNTRY@ is		L' @COUNTRY@ est	from @COUNTRY@ ,		des @COUNTRY@ ,
Ecuador is		L' Équateur est	from Ecuador is		des Équateur ,
@COUNTRY@ is		Les @COUNTRY@ sont	from @COUNTRY@ ,		de la @COUNTRY@ ,
Ecuador is		Les Équateur sont	from Ecuador ,		de la Équateur ,

TABLE 2 – Exemples de bi-segments générés qui peuvent contenir des erreurs (e.g. « from Ecuador » → « de la Équateur »).

3 Le processus de filtrage

Pour réaliser le filtrage, nous proposons d'utiliser une technique basée sur la différence de scores d'entropie croisée, inspirée par des approches récentes en sélection de données pour l'adaptation au domaine en TAS. Ces techniques de sélection sont appliquées soit au corpus cible uniquement (Moore & Lewis, 2010) (filtrage monolingue), soit conjointement aux corpus source et cible (Axelrod *et al.*, 2011) (filtrage bilingue).

Filtrage monolingue Tout d'abord nous entraînons deux modèles de langue (ML), l'un correspondant à des données « en-domaine » (ED), l'autre à un sous-ensemble des données « hors-domaine » (HD). Nous donnons ensuite un score \hat{H}_{Pp} à chaque segment c de la partie cible de la table de traduction augmentée (à savoir tbg) avec ces deux modèles de langue (ML_{ED} et ML_{HD}) :

$$\hat{H}_{Pp}(c) = H_{ED}(c) - H_{HD}(c) \quad (2)$$

Ici, $H_{ED}(c)$ est l'entropie croisée (c-à-d le \log_2 de la perplexité) de c par rapport à ML_{ED} . $H_{HD}(c)$ est l'entropie croisée de c par rapport à ML_{HD} . Ensuite nous trions l'ensemble des bi-segments d'après leur score (\hat{H}_{Pp}) appliqué au côté cible (c). Le score $\hat{H}_{Pp}(c)$, qui peut prendre des valeurs positives ou négatives, est une indication de la « proximité » de la phrase cible c relativement au corpus en-domaine : un score plus bas indique une proximité plus grande¹. La phase suivante propose de choisir le point de coupure du corpus ainsi trié, grâce à la mesure de perplexité. Pour cela, nous considérons des tranches incrémentales de notre liste triée de bi-segments sur chacune desquelles nous entraînons un modèle de langue (ML) sur les segments cibles de la tranche. Enfin, nous calculons la perplexité de chacun de ces modèles de langue sur un corpus de développement en-domaine (ED). Le point de coupure correspond à la perplexité la plus faible ainsi calculée.

Filtrage bilingue On peut aussi appliquer la procédure précédente de façon bilingue. Pour un bi-segment (s, c) , on calcule un score $\hat{H}_{Pp}(s, c)$ de la façon suivante :

$$\hat{H}_{Pp}(s, c) = [H_{ED}(s) - H_{HD}(s)] + [H_{ED}(c) - H_{HD}(c)] \quad (3)$$

Maintenant, le processus de tri est effectué sur le score $\hat{H}_{Pp}(s, c)$, qui dépend des segments source et cible, mais le processus d'identification du point de coupure est effectué seulement sur le côté cible, en nous servant uniquement d'un corpus de développement en-domaine pour la langue cible, comme dans le cas précédent.

4 Expériences

Pour ces expériences, les systèmes de TAS à base de segments ont été entraînés en utilisant l'outil open-source MT Moses (Koehn *et al.*, 2007). Les modèles de langue utilisés sont des n -gram (avec $n = 5$), en appliquant un lissage Kneser-Ney (Chen & Goodman, 1999) grâce aux outils du SRI (Stolcke, 2002). Nous avons utilisé les scores de BLEU [*BiLingual Evaluation Understudy*] (Papineni *et al.*, 2002) et TER [*Translation Edit Rate*] (Snover *et al.*, 2006) comme mesures de performances des modèles pour les expériences.

4.1 Données

Le système de traduction a été entraîné avec les corpus Europarl V.7 (*ep7*) et News-Commentary V.8 (*nc8*), détaillés dans le tableau 3.

1. Ce score est inspiré de (Moore & Lewis, 2010), mais nous l'appliquons pour effectuer un filtrage sur des éléments de la table de bi-segments, alors qu'à l'origine il est appliqué sur des éléments du corpus pour effectuer une sélection au niveau des phrases.

Type	Corpus	# lignes	# mots src (en)	# mots cible (fr)
Apprentissage	<i>ep7</i>	2 007 K	56 192 K	61 811 K
	<i>nc8</i>	157 K	4 105 K	4 815 K
Développement	<i>ntst11</i>	3 003	75 K	86 K
Évaluation	<i>tstTerm</i>	2 577	87 K	103 K

TABLE 3 – Détail des données bilingues utilisées pour les expériences.

Modèle de traduction			<i>tstTerm</i> (ML référentiel)			<i>tstTerm</i> (ML dégradé)	
Appellation	taille	Nbr. de bi-seg. ajoutés	MHV	BLEU	TER	BLEU	TER
Référentiel	77 203 175	N/A	2 565 (2,9%)	30,7	56,1	27,4	59,0
Base	77 138 148	N/A	5 237 (6,0%)	27,2	58,9	27,2	58,5
Unigrammes	77 138 149	1	2 565 (2,9%)	30,6	56,1	28,4	57,0
Contextes générés (sans filtrage)	78 611 118	1 472 970	2 565 (2,9%)	31,1	55,6	28,8	56,6
Contextes générés ($\hat{H}_{Pp}(t)$)	77 193 106	54 958	2 565 (2,9%)	30,5	56,1	28,6	57,0
Contextes générés ($\hat{H}_{Pp}(s, t)$)	77 754 665	616 517	2 565 (2,9%)	31,0	55,6	29,2	56,7

TABLE 4 – Tableau de statistiques et de résultats pour les différentes configurations appliquées sur le modèle de traduction.

Le corpus de développement est issu de la campagne d'évaluation WMT 2014 (*ntst11*). Le corpus de test spécifique consiste en un ensemble de 2 500 phrases récupérées du corpus MultiUN (*Nations Unies*) (Eisele & Chen, 2010) et noté « *tstTerm* ». Ce dernier est donc un bitexte qui contient au moins une fois la traduction de « Germany » vers « Allemagne » par phrase, soit 2 672 occurrences (environ 3% des mots sources).

4.2 Résultats

Le tableau 4 présente les résultats et les statistiques suivant plusieurs configurations :

- « Référentiel » : un système appris sur les données telles quelles ;
- « Base » : les bi-segments extraits du corpus d'apprentissage sont filtrés pour retirer toute mention de « Germany » et « Allemagne », la table de traduction est entraînée sur les bi-segments restant ;
- « Unigrammes » : la configuration « base » est enrichie par le bi-segment « Germany » → « Allemagne » ;
- « Contextes générés (sans filtrage) » : enrichissement de la configuration « base » par notre approche de génération de contexte (voir section 2) ;
- « Contextes générés ($\hat{H}_{Pp}(s)$) » et « Contextes générés ($\hat{H}_{Pp}(s, t)$) » : respectivement filtrage monolingue et bilingue des contextes générés (description section 3) et ajout des bi-segments du filtrage à la configuration « base » ;

Nous indiquons également la quantité de données générées et les mots hors-vocabulaire (*MHV*). Avec ces configurations, s'ajoutent deux modèles de langues cibles possibles : un modèle appelé « référentiel », appris sur les données cibles telles quelles, ou alors, un modèle « dégradé » appris sur ces mêmes données mais en supprimant « Allemagne » du vocabulaire. Ceci pour simuler l'absence de ce mot dans le modèle de langue et permettre de mieux voir l'influence du contexte de la table de traduction, indépendamment du modèle de langue.

Les résultats montrent sans surprise une amélioration très significative entre les configurations « Base » et « unigrammes », principalement due à la diminution du nombre de mots-hors-vocabulaire (*MHV*). On constate que le système « unigrammes » est du même niveau que le système « Référentiel ». Le système « Contextes générés (sans filtrage) » donne des résultats très encourageants en surpassant le système « Unigrammes » mais également le système « Référentiel ». La contre-performance de ce dernier semble être uniquement lié à des ambiguïtés sur les bi-segments retirés, à savoir toutes les traductions de « Germany » vers « Allemagne ».

Enfin, les deux dernières approches utilisent la sélection monolingue des paires de segments générés avec du contexte (« contexte généré ($\hat{H}_{pp}(t)$) ») et la sélection bilingue de paires de segments (« contexte généré ($\hat{H}_{Pp}(s, t)$) »). La première sélection semble être trop forte : nous observons une diminution des scores BLEU et TER. Cependant, la sélection bilingue de paires de segments nous permet d'être aussi efficace que l'approche « contextes générés (sans filtrage) », mais en ne conservant que 45% des paires de segments générés, et ce, dans les deux tableaux de résultats. Ces dernières expériences semblent valider l'utilisation de la sélection de données bilingues associée à notre approche de génération de contextes. De plus, cette approche permet d'améliorer significativement les résultats par rapport au système « Référentiel ».

5 Etat de l'Art

5.1 Adaptation au domaine en TAS

En Traduction automatique statistique, l'un des sujets les plus étudiés concerne l'adaptation au domaine des systèmes de TAS. Il y a différentes façons d'effectuer cette adaptation. L'une des plus courantes consiste à appliquer une sélection sur les données d'apprentissage. Plusieurs travaux ont été réalisés en utilisant des approches fondées sur la recherche d'information afin d'extraire les parties du corpus qui sont les plus pertinentes (Eck *et al.*, 2004). Des travaux plus récents sont fondés sur l'entropie croisée pour sélectionner les parties les plus pertinentes des données d'apprentissage (Moore & Lewis, 2010; Axelrod *et al.*, 2011).

Dans notre cas, nous nous concentrons sur la traduction en enrichissant le vocabulaire grâce une terminologie spécifique. Or, il n'existe pas de grande quantité de données d'entraînement pour chaque domaine spécifique. C'est pourquoi ces approches ne sont pas adaptées à notre problème. Cependant, ce type d'approche est tout indiqué pour filtrer les bi-segments virtuels générés.

5.2 Enrichissement d'informations lexicales pour la TAS

La plupart des approches dans ce domaine proposent un moyen d'extraire la terminologie spécifique à partir de corpus bilingues (qu'ils soient parallèles ou comparables). Ces approches visent à construire le même genre de dictionnaires que ceux que nous voulons utiliser.

Des travaux antérieurs ont été proposés dans le but de réduire le nombre de mots hors-vocabulaire (MHV) comme (Habash, 2008). Ces approches visent, d'une certaine façon, à ajouter des MHV en les ajoutant dans le corpus d'apprentissage, en utilisant le dictionnaire comme une mémoire de traduction en plus du modèle de traduction. Certains utilisent un pré- ou post-traitement pour éviter le problème MHV (Banerjee *et al.*, 2012; Nikoulina *et al.*, 2012; Tsvetkov *et al.*, 2013). En ce sens, notre approche permet d'éviter tout pré- ou post-traitement lors du processus de traduction et d'utiliser les outils de traduction classique (Koehn *et al.*, 2007) sans modification.

L'approche la plus proche de la nôtre est proposée par (Skadiņš *et al.*, 2013). Dans leur article, ils proposent une technique de sélection de données d'entraînement selon une terminologie spécifique. Cela signifie qu'ils sélectionnent les bi-phrases qui ne contiennent que la terminologie spécifique recherchée. Ensuite, l'ensemble du processus d'entraînement n'est pas modifié (trouver l'alignement de texte, extraire les bi-segments et enfin estimer les paramètres des modèles de traduction). Enfin, les approches de génération de bi-segments par analogie (Chen *et al.*, 2011; Luo *et al.*, 2013) ne s'intéressent généralement assez peu à la problématique de la terminologie d'un domaine. Or notre approche offre comme principal intérêt d'ajouter la terminologie d'un domaine avec son contexte phrastique. Dans cette catégorie, (Langlais *et al.*, 2009) proposent une approche analogique traduire de la terminologie, cependant, ces hypothèses de traduction ne sont pas intégrées à un système de TAS à base de segments (TAS-BS), contrairement à notre approche.

6 Conclusion et discussion

Cet article présente les premiers résultats d'une approche d'enrichissement terminologique automatique pour la traduction automatique statistique. Cette approche propose l'ajout de contextes phrastiques autour des termes nouvellement introduits. Notre méthode conduit à des résultats encourageants en terme de scores de BLEU et TER sur notre corpus de test spécifique. D'un point de vue linguistique, les apports sont principalement situés sur le contexte gauche de la terminologie ainsi insérée. L'un des cas les plus courants est l'ajout d'un déterminant ou d'une préposition correctement traduite.

Nous prévoyons d'étendre cette approche à d'autres entités nommées et à d'autres types mots comme les adjectifs, les noms et verbes. Ces derniers sont souvent accompagnés de flexions liées principalement aux accords en genre et en nombre. Enfin, nous souhaitons également évaluer notre approche dans le cadre d'une application réelle de post-édition afin de mesurer son impact.

Remerciements

Les auteurs souhaitent remercier les différents relecteurs pour leur remarques et conseils qui ont permis d'améliorer considérablement cet article.

Ce travail a été partiellement financé par la Commission Européenne à travers le projet TransLectures (FP7/2007-2013 convention de subvention N° 287755) et par l'Agence Nationale de la Recherche dans le cadre du projet KEHATH.

Références

- AXELROD A., HE X. & GAO J. (2011). Domain Adaptation via Pseudo In-Domain Data Selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Edimbourg, Ecosse, Royaume-Uni.
- BANERJEE P., NASR S. K., ROTURIER J., WAY A. & VAN GENABITH J. (2012). Domain adaptation in SMT of user-generated forum content guided by OOV word reduction : Normalization and/or supplementary data ? In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT)*, Trento, Italie.
- CHEN B., KUHN R. & FOSTER G. (2011). Semantic smoothing and fabrication of phrase pairs for SMT. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT-2011)*, San Francisco, Etats-Unis.
- CHEN S. F. & GOODMAN J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 4(13), 359–393.
- ECK M., VOGEL S. & WAIBEL A. (2004). Language model adaptation for statistical machine translation based on information retrieval. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Lisbonne, Portugal.
- EISELE A. & CHEN Y. (2010). MultiUN : A Multilingual Corpus from United Nation Documents. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malte.
- HABASH N. (2008). Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL-HLT)*, Columbus, Etats-Unis.
- KOEHN P. (2005). Europarl : A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, Phuket, Thaïlande.
- KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses : Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, République Tchèque.
- LANGLAIS P., YVON F. & ZWEIGENBAUM P. (2009). Improvements in analogical learning : Application to translating multi-terms of the medical domain. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, Athenes, Grèce.
- LUO J., MAX A. & LEPAGE Y. (2013). Using the productivity of language is rewarding for small data : Populating smt phrase table by analogy. In *Proceedings of the 6th Language & Technology Conference (LTC'13)*, Poznan, Pologne.
- MOORE R. C. & LEWIS W. (2010). Intelligent selection of language model training data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL-HLT)*, Uppsala, Suède.
- NIKOULINA V., SANDOR A. & DYMETMAN M. (2012). Hybrid adaptation of named entity recognition for statistical machine translation. In *Proceedings of the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT (ML4HMT - 2012)*, Mumbai, Inde.
- PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL-HLT)*, Philadelphie, Etats-Unis.
- SKADINŠ R., PINNIS M., GORNOSTAY T. & VASIŁJEVS A. (2013). Application of online terminology services in statistical machine translation. In *Proceedings of the XIV Machine Translation Summit (MT Summit XIV)*, Nice, France.
- SNOVER M., DORR B., SCHWARTZ R., MICCIULLA L. & MAKHOUL J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, Massachusetts.

STOLCKE A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado.

TSVETKOV Y., DYER C., LEVIN L. & BHATIA A. (2013). Generating english determiners in phrase-based translation with synthetic translation options. In *Proceedings of the 8th Workshop on Statistical Machine Translation (WMT'13)*, Sofia, Bulgarie.

Une mesure d'intérêt à base de surreprésentation pour l'extraction des motifs syntaxiques stylistiques

Mohamed-Amine Boukhaled, Francesca Frontini, Jean-Gabriel Ganascia

LIP6 (Laboratoire d'Informatique de Paris 6), Université Pierre et Marie Curie and CNRS (UMR7606),
ACASA Team, 4, place Jussieu,
75252-PARIS Cedex 05 (France)
{mohamed.boukhaled, francesca.frontini, jean-gabriel.ganascia}@lip6.fr

Résumé. Dans cette contribution, nous présentons une étude sur la stylistique computationnelle des textes de la littérature classique française fondée sur une approche conduite par données, où la découverte des motifs linguistiques intéressants se fait sans aucune connaissance préalable. Nous proposons une mesure objective capable de capturer et d'extraire des motifs syntaxiques stylistiques significatifs à partir d'un œuvre d'un auteur donné. Notre hypothèse de travail est fondée sur le fait que les motifs syntaxiques les plus pertinents devraient refléter de manière significative le choix stylistique de l'auteur, et donc ils doivent présenter une sorte de comportement de surreprésentation contrôlé par les objectifs de l'auteur. Les résultats analysés montrent l'efficacité dans l'extraction de motifs syntaxiques intéressants dans le texte littéraire français classique, et semblent particulièrement prometteurs pour les analyses de ce type particulier de texte.

Abstract.

An Overrepresentation-based Interestingness Measure for Syntactic Stylistic Pattern Extraction

In this contribution, we present a computational stylistic study of the French classic literature texts based on a data-driven approach where discovering interesting linguistic patterns is done without any prior knowledge. We propose an objective measure capable of capturing and extracting meaningful stylistic syntactic patterns from a given author's work. Our hypothesis is based on the fact that the most relevant syntactic patterns should significantly reflect the author's stylistic choice and thus they should exhibit some kind of overrepresentation behavior controlled by the author's purpose. The analysed results show the effectiveness in extracting interesting syntactic patterns from classic French literary text, and seem particularly promising for the analyses of such particular text.

Mots-clés : Stylistique computationnelle, fouille de texte, motifs syntaxiques, mesure d'intérêt

Keywords: Computational stylistic, text mining, syntactic patterns, interestingness measure

1 Introduction

La stylistique computationnelle est un sous-domaine de la linguistique informatique qui se situe à l'intersection de plusieurs domaines de recherche comme le traitement automatique du texte, l'analyse littéraire et la fouille de données statistique. L'objectif de la stylistique computationnelle est d'extraire des motifs de style caractérisant un type particulier de textes à l'aide des méthodes statistiques et automatiques. En prenant le cas de l'étude du style d'écriture d'un auteur particulier, la tâche sera d'explorer automatiquement les formes linguistiques de son style qui ne sont pas seulement caractéristiques mais aussi volontairement surutilisées par cet auteur par rapport à une norme linguistique. Cependant, la notion de style dans le contexte de la stylistique computationnelle se révèle être assez large vu qu'elle se manifeste sur plusieurs niveaux linguistiques : lexical, syntaxique, sémantique et pragmatique. Chaque niveau possède ses propres marqueurs de styles et ses propres unités linguistiques qui le caractérisent.

Grace à la notion du style, la stylistique computationnelle interfère avec de nombreuses autres tâches connexes telles que l'attribution d'auteur (Stamatatos 2009), la classification stylistique de texte (Kessler et al. 1997), génération de texte basée sur style (Hovy 1990), l'évaluation automatique de la lisibilité et de la complexité du texte (Pitler & Nenkova 2008).

Les techniques de la stylistique computationnelle ont été utilisées pendant plusieurs années pour étudier les questions relatives au style dans le contexte de l'analyse littéraire, voir (Siemens & Schreibman 2013) pour un aperçu et une discussion. D'un point de vue méthodologique, deux types différents d'approches ont émergé:

- L'approche conduite par classification, qui peut être simplifiée comme suit: une classification connue a priori se trouve dans la littérature (comme les comédies vs tragédies de Shakespeare); certaines caractéristiques linguistiques sont identifiées sur la base de leur pertinence et de leur capacité à reproduire cette classification. Ces caractéristiques linguistiques sont utilisées par la suite pour voir si la distinction a priori se maintient ou non quand on se base sur des techniques de classification automatique comme le clustering par exemple (Craig 2004).
- L'approche herméneutique, dans laquelle les textes littéraires sont analysés afin d'en extraire automatiquement, sans aucune connaissance préalable, les caractéristiques linguistiques intéressantes qui peuvent ensuite être utilisées par les experts du domaine pour produire une analyse critique mieux informée (Mahlberg 2012, Ramsay 2011).

Dans cette contribution, nous présentons une étude stylistique computationnelle des textes classiques de la littérature française basée sur une approche herméneutique conduite par données où la découverte des formes linguistiques intéressantes se fait sans aucune connaissance préalable. Plus précisément, la méthode proposée est fondée sur l'évaluation de la surreprésentation des motifs syntaxiques dans un texte par rapport à un corpus de norme. Cette méthode est destinée à soutenir une analyse textuelle en focalisant sur :

- 1) La vérification du degré d'importance de chaque motif syntaxique (segments syntagmatiques avec d'éventuel trous).
- 2) L'extraction automatique d'une liste de motifs syntaxiques caractérisant le style syntaxique d'une œuvre d'un auteur donné.

2 Approche pour l'extraction des motifs syntaxiques pertinents

Notre méthode est composée de deux étapes. D'abord, un algorithme d'extraction de motif séquentiel est appliqué sur les textes pour en extraire des motifs syntaxiques récurrents. Deuxièmement, une mesure d'intérêt basée sur l'évaluation de la surreprésentation (en termes de fréquence d'apparition) par rapport à un corpus de norme est appliquée à l'ensemble des motifs syntaxiques extraits. Ainsi, à chaque motif syntaxique sera affecté un poids en fonction de sa surreprésentation indiquant son importance et sa pertinence dans la caractérisation du style syntaxique du texte en question. Dans ce qui suit, nous présentons le corpus à traiter et le protocole de son découpage en deux éléments : texte à analyser et texte de norme dans la sous-section 2.1. Ensuite, la sous-section 2.2 introduit quelques éléments nécessaires pour la compréhension du processus d'extraction de motifs syntaxiques séquentiels. Enfin, la formulation et les détails statistiques de la mesure d'intérêt proposée sont présentés à la section 2.3.

2.1 Corpus analysé

Dans notre étude, nous avons utilisé quatre romans écrits par quatre célèbres auteurs français: *Eugénie Grandet* de Balzac, *Madame Bovary* de Flaubert, *Notre Dame de Paris* de Hugo et le *Ventre de Paris* de Zola. Ce choix est motivé par notre intérêt particulier pour la littérature française classique du 19^{ème} siècle. Le fait que tous ces textes soit du même genre littéraire et écrits par des auteurs appartenant à la même époque permet de réduire l'effet qu'a la variation du genre et de l'époque sur le style d'écriture. Ce qui permet à son tour d'avoir une étude moins biaisée et bien focalisée sur le style d'écriture propres à ces auteurs. Au moment de l'analyse des motifs syntaxiques chaque texte écrit par un de ces quatre auteurs est mis en contraste avec les textes écrits par les trois autres auteurs. C'est à dire que ces trois textes seront considérés comme corpus de norme à partir duquel on va évaluer l'hypothèse de la surreprésentation des motifs syntaxiques dans le quatrième texte restant, comme expliqué dans la suite de cette section.

2.2 Extraction des motifs syntaxiques

Dans notre étude nous considérons une approche syntagmatique. Le texte est d'abord segmenté en un ensemble de phrases, puis chaque phrase est représentée par une séquence d'étiquettes syntaxiques (POS-tag) ¹correspondantes aux mots de la phrase. Ce qui permet de produire à la fin un ensemble de séquences syntaxique pour chaque texte. Par exemple, la phrase «*Le silence profond régnait nuit et jour dans la maison .* » sera représenté par la séquence :

< DET , NOM , ADJ , VER , NOM , KON , NOM , PRP , DET , NOM , SENT >

Puis, des motifs séquentiels d'une longueur déterminée avec leurs fréquences d'apparition (comptées par nombre de phrases et décrit plus souvent par le terme « support ») sont extraits de cette base de données séquentielle syntaxique en utilisant un algorithme d'extraction de motifs séquentiels (Viger et al. 2014). Un motif syntaxique consiste en un segment syntagmatique séquentiel (avec d'éventuels trous) présent dans la séquence syntaxique. Il peut être considéré

¹ Liste complète des étiquettes syntaxiques sur : <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html>

comme une sorte de généralisation de la notion des n-gram très utilisée dans le domaine du traitement automatique de la langue. Voici quelques exemples de motifs syntaxiques présents dans la séquence de l'exemple cité ci-dessus ²:

- $\langle DET \rangle \langle NOM \rangle \langle ADJ \rangle$
- $\langle NOM \rangle \langle ADJ \rangle \langle VER \rangle \langle NOM \rangle$
- $\langle KON \rangle \langle NOM \rangle \langle * \rangle \langle DET \rangle \langle NOM \rangle$

Pour éviter l'effet de la fluctuation statistique sur l'analyse des motifs avec basses fréquences, nous avons considéré une contrainte de seuil minimum de fréquence de 1%. C'est-à-dire que nous nous concentrons uniquement sur des motifs qui sont présents dans au moins 1% des phrases du texte analysé.

Cependant, comme le processus d'extraction de motifs séquentiels est connu par sa propriété de produire une grande quantité de motifs, et cela même dans des échantillons de textes relativement petits, une mesure d'intérêt doit être appliquée afin d'identifier les motifs les plus importants et pertinents pour la caractérisation du style syntaxique du texte en question. Cette mesure d'intérêt est expliquée dans la sous-section suivante.

2.3 Evaluation de la pertinence des motifs syntaxiques

Notre hypothèse est basée sur le fait que les motifs syntaxiques les plus pertinents devraient refléter de manière significative le choix stylistique de l'auteur et doivent ainsi se caractériser par une considérable surreprésentation dans ses textes. Cependant, pour capturer cette surreprésentation on ne peut pas se référer seulement à la fréquence brute, ou même relative, des motifs syntaxiques. En effet, une utilisation plus fréquente d'un motif syntaxique par un auteur (ce qui se traduit par une fréquence relative très élevée) n'indique pas nécessairement un choix ou un trait stylistique puisque ça peut être très bien une propriété imposée par la grammaire de la langue ou par les caractéristiques syntaxiques du genre du texte.

Ainsi, pour évaluer la surreprésentation des motifs dans un texte, on utilisera une approche empirique basée sur la comparaison de la fréquence d'apparition d'un motif syntaxique dans un texte par rapport à sa fréquence d'apparition dans un corpus de norme. Un ratio α entre ces deux quantités est calculé :

$$\alpha = \frac{\text{Fréquence du motif dans le corpus de norme}}{\text{Fréquence du motif dans le texte}}$$

Dans notre expérimentation nous avons constaté empiriquement que la distribution du ratio α suit un comportement Gaussien. En effet, les valeurs du ratio α sont réparties autour d'une valeur centrale (voir Fig. 1). Cela est dû au fait que la fréquence d'apparition d'un motif syntaxique dans un texte est fortement corrélée à sa fréquence d'apparition dans un corpus de norme avec quelques cas particuliers qui présentent une certaine aberrance (voir Fig. 2). Ce sont ces cas aberrants qui représentent un intérêt particulier pour notre étude parce qu'ils représentent une certaine déviation linguistique (propres au style de l'auteur) par rapport à ce qu'on s'attend de voir dans un corpus de norme.

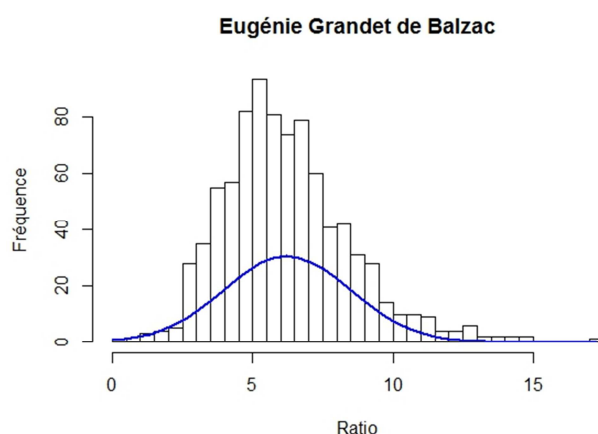


Fig. 1. Illustration du comportement gaussien du ratio α dans le roman *Eugénie Grandet* de Balzac

² Le symbole $\langle * \rangle$ correspond à un trou qui peut être remplacé par n'importe quelle étiquette syntaxique

Cela nous permet d'utiliser la méthode de détection des cas aberrants basée sur la distribution Gaussienne (Chandola et al. 2009). En effet, la surreprésentation d'un motif dans ce cas se traduira par un comportement aberrant négatif plus grand par rapport aux autres motifs. Les motifs les plus surreprésentés dans un texte seront ceux associés aux valeurs de z-score standard Z les moins élevées. Les valeurs de z-score sont calculées comme suit :

$$Z_i = \frac{\alpha_i - \hat{\alpha}}{s}$$

Où Z_i et α_i sont respectivement le ratio α et le z-score Z associé au i -ème motif syntaxique, α_i et s sont respectivement la moyenne et l'écart-type du ratio α

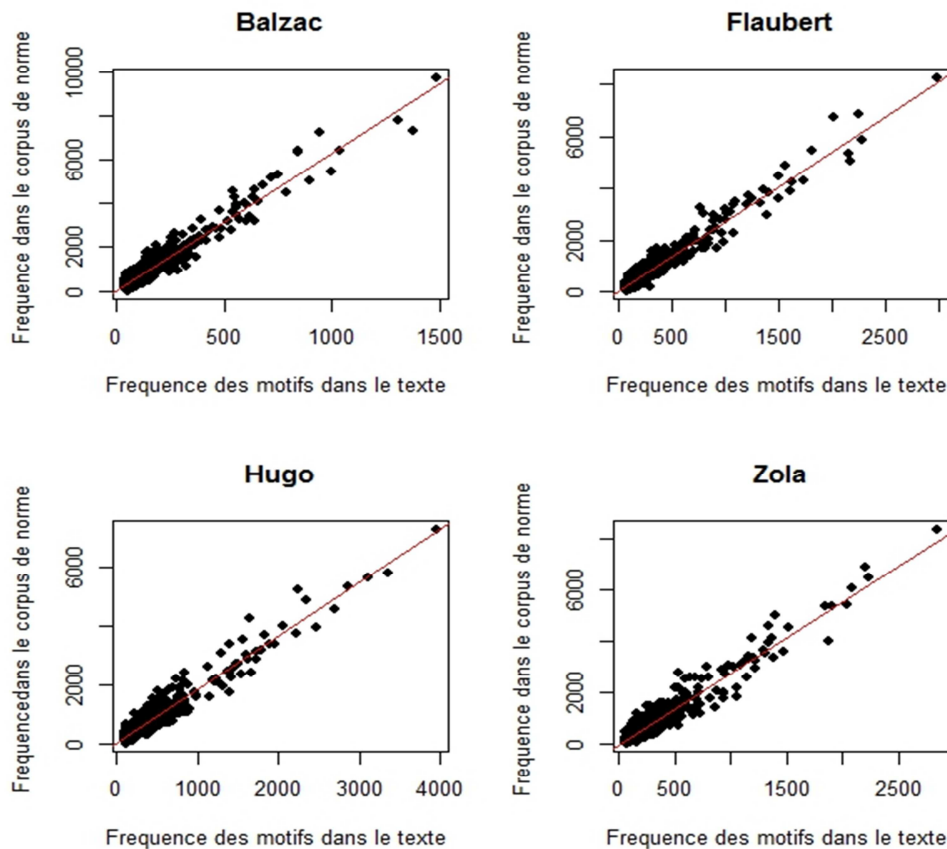


Fig. 2. Fréquence d'apparition d'un motif syntaxique dans un texte par rapport sa fréquence d'apparition dans un corpus de norme pour les romans étudiés, chaque point dans un graphe représente un motif syntaxique. Les lignes tracées représentent les courbes de régression linéaire capturant le comportement attendu du ratio α

3 Résultats et Discussion

Dans cette section, nous présenterons quelques exemples de motifs syntaxiques extraits et classés comme motifs caractéristiques du style des textes inclus dans notre corpus. En utilisant la méthode proposée, les motifs extraits semblent avoir une forte pertinence pour caractériser le style de l'auteur du texte en question, mais aussi pour le scénario du roman et le genre littéraire dans lequel il s'inscrit. Dans le roman *Madame Bovary* de Flaubert, les motifs extraits représentent bien la fonction rythmique plutôt que fonctionnelle que donne Flaubert à la virgule et au point-virgule (Mangiapane 2012). Par exemple dans le cas du motif (1) nous voyons la virgule précédant la conjonction, suivie par une clause imbriquée.

Motif (1) <PUN> <KON>< PUN> <PRP>, avec support= 113, exemples d'instances dans le texte :

- , et , à
- , mais , avant
- ; et , à

Dans le *Ventre de Paris* de Zola, et dans le même sens, les motifs classés comme pertinent qualifient clairement l'utilisation des clauses imbriquées pour décrire des situations ou des attitudes comme dans le motif (2) ou bien pour les descriptions des lieux publics et des objets en affiche comme dans le motif (3) :

Motif (2) : <PUN> <PRP> <PRP> <NOM>, support= 104, exemple (en gras) d'instances dans le texte :

« Florent se heurtait à mille obstacles , **à des porteurs** qui se chargeaient , **à des marchandes** qui discutaient de leurs voix rudes ; il glissait sur le lit épais d' épluchures et de trognons qui couvrait la chaussée , il étouffait dans l' odeur puissante des feuilles écrasées . »

Motif (3): <NOM> <PUN> <PRP> <NOM> <ADJ>, support= 68, exemples d'instances dans le texte :

- angles , à fenêtres étroites
- très-jolies , des légendes miraculeuses
- écrevisses , des nappes mouvantes

Dans *Eugénie Grandet* de Balzac, nous pouvions constater d'autres différentes fonctions communicatives accomplies par les motifs syntaxiques et leurs instances textuelles, par exemples :

Le motif (4): <PUN> <VER> <NAM> <PRP>, avec support= 49, est utilisé pour faire une post introduction d'un discours directe sous forme d'une clause imbriquée. Exemples d'instances dans le texte :

- , dit Grandet en
- , reprit Charles en
- , dit Cruchot en

Le motif (5): <NUM> <NUM> <NOM>, avec support= 54, est un motif utilisé pour parler des sommes d'argent, ce qui est typique pour le scénario du roman où l'argent joue un rôle très important . Exemples d'instances dans le texte :

- vingt mille francs
- deux mille louis
- sept mille livres

Le motif (6) : <ADV> <VER> <PRO> <ADV>, avec support= 59, est utilisé pour exprimer des questions négatives :

- n' avait -il pas
- ne disait -on pas
- ne serait -il pas

Le motif (7) : <PUN> <NOM> <PUN> <VER>, avec support= 44, représente la ponctuation largement utilisée pour imiter l'intonation orale et même de reproduire les phénomènes de performance tels que bégayer :

- , messieurs , cria
- , madame , répondi
- , mademoiselle , disait

Enfin pour le dernier texte, *Notre Dame de Paris* de Hugo en l'occurrence, nous avons remarqué que les motifs syntaxiques extraits comme étant caractéristique sont beaucoup plus pertinents pour décrire le contenu de ce roman et la manière dont l'auteur a utilisé pour introduire le lecteur dans l'histoire et le familiariser avec l'endroit où se déroulera la plupart des événements.

Par exemple, dans le motif (8) le nom propre est souvent un endroit, surtout au début du roman où les pièces descriptives sont plus fréquentes dans le but de guider le lecteur dans la topographie de Paris médiéval.

Motif (8) : <NOM> <PRP> <NAM> <PUN>, support= 340, exemples d'instances dans le texte :

- hôtel de Bourbon ,
- murailles de Paris ,
- dauphin de Vienne ;

Par ailleurs, le motif (9) est souvent utilisé pour présenter les personnages en indiquant d'abord leur nom et leur titre. Il convient de noter que le roman *Notre Dame de Paris* présente une pléthore de personnages secondaires.

Motif (9): <NAM> <PRP> <NAM> <PUN>, support = 118, exemples d'instances dans le texte :

- Marguerite de Flandre ,
- Jehan de Troyes ,

Les quelques exemples analysés indiquent d'une part que la technique présentée est efficace pour extraire des motifs syntaxiques intéressants dans des textes littéraires, et cela semble particulièrement prometteur pour les analyses de ce type de texte. D'autre part, la technique proposée, ainsi que d'autres techniques semblables, nous invite à poser plus de questions sur l'interprétation linguistique et la significativité de ce qui est réellement capturé par ces motifs syntaxiques. Certaines structures syntaxiques peuvent être importantes car elles sont typiques du style de l'auteur (son empreinte stylistique), mais elles peuvent être aussi très bien dictées par les besoins fonctionnels, en raison de la question particulière de l'œuvre, ou les conventions du genre choisi. Ceci est particulièrement vrai pour l'analyse syntaxique, où les contraintes fonctionnelles imposées, qui limitent la liberté d'auteur, sont plus évidents.

4 Conclusion

Dans cette contribution, nous avons présenté une étude sur la stylistique computationnelle des textes de la littérature classique française basée sur une approche conduite par données, où la découverte des motifs linguistiques intéressants se fait sans aucune connaissance préalable. Nous avons proposé une mesure objective capable de capturer et d'extraire des motifs syntaxiques stylistiques significatifs à partir d'une œuvre d'un auteur donné. Pour évaluer l'efficacité de la méthode proposée, nous avons mené une expérience sur quatre romans classiques français très célèbres. Les résultats analysés montrent l'efficacité dans l'extraction de motifs syntaxiques très intéressants d'un point de vue stylistique à partir de ce type particulier de texte.

Sur la base de la présente étude, nous avons déduit plusieurs perspectives et futures directions de recherches. Premièrement, nous allons explorer l'utilité d'utilisation d'autres mesures statistiques pour évaluer l'intérêt stylistique d'un motif syntaxique donné. Deuxièmement, cette étude sera élargie pour inclure les motifs morphosyntaxiques (forme et lemme des mots). Troisièmement, nous avons l'intention d'expérimenter avec d'autres différentes langues en utilisant d'autres corpus largement employés dans le domaine de la stylistique computationnelle en général.

Remerciement

Ce travail a bénéficié d'une aide d'État gérée par l'Agence Nationale de la Recherche dans le cadre des Investissements d'Avenir portant la référence ANR-11-IDEX-0004-02

Références

- Chandola, V., Banerjee, A. & Kumar, V., 2009. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3), p.15.
- Craig, H., 2004. Stylistic analysis and authorship studies. *A companion to digital humanities*, 3, pp.233–334.
- Hovy, E.H., 1990. Pragmatics and natural language generation. *Artificial Intelligence*, 43(2), pp.153–197.
- Kessler, B., Numberg, G. & Schütze, H., 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. pp. 32–38.
- Mahlberg, M., 2012. *Corpus stylistics and Dickens's fiction*, Routledge.
- Mangiapane, S., 2012. Ponctuation et mise en page dans *Madame Bovary*: les interventions de Flaubert sur le manuscrit du copiste. Flaubert. *Revue critique et génétique*, (8).
- Pitler, E. & Nenkova, A., 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pp. 186–195.
- Ramsay, S., 2011. *Reading machines: Toward an algorithmic criticism*, University of Illinois Press.
- Siemens, R. & Schreibman, S., 2013. *A companion to digital literary studies*, John Wiley & Sons.
- Stamatatos, E., 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3), pp.538–556.
- Viger, P.F. et al., 2014. SPMF: A Java Open-Source Pattern Mining Library. *Journal of Machine Learning Research*, 15, pp.3389–3393.

Une Approche évolutionnaire pour le résumé automatique

Aurélien Bossard¹ Christophe Rodrigues²

(1) Université Paris 8, Laboratoire d'Informatique Avancée de Saint-Denis

(2) CNRS, UMR 7030, Laboratoire d'Informatique de Paris Nord

(1) aurelien.bossard@iut.univ-paris8.fr, (2) christophe.rodrigues@lipn.fr

Résumé. Dans cet article, nous proposons une méthode de résumé automatique fondée sur l'utilisation d'un algorithme génétique pour parcourir l'espace des résumés candidats couplé à un calcul de divergence de distribution de probabilités de n-grammes entre résumés candidats et documents source. Cette méthode permet de considérer un résumé non plus comme une accumulation de phrases indépendantes les unes des autres, mais comme un texte vu dans sa globalité. Nous la comparons à une des meilleures méthodes existantes fondée sur la programmation linéaire en nombre entier, et montrons son efficacité sur le corpus TAC 2009.

Abstract.

Automatic Summarization Using a Genetic Algorithm

This paper proposes a novel method for automatic summarization based on a genetic algorithm that explores candidate summaries space following an objective function computed over ngrams probability distributions of the candidate summary and the source documents. This method does not consider a summary as a stack of independant sentences but as a whole text. We compare this method to one of the best existing methods which is based on integer linear programming, and show its efficiency on TAC 2009 corpus.

Mots-clés : Résumé automatique, algorithme génétique, modèles probabilistes.

Keywords: automatic summarization, genetic algorithm, probabilistic models.

1 Introduction

Les systèmes de résumé automatique (RA) sont des constituants essentiels des systèmes d'information. En effet, La multiplication des sources d'information numérique rend parfois difficile la lecture et l'assimilation d'un contenu en ligne, même dans le cas où celui-ci est issu d'une recherche précise. Résumer automatiquement ces contenus peut alors proposer une nouvelle approche d'un contenu informationnel. Le RA est donc naturellement devenu une des premières thématiques de recherche en traitement automatique du langage (Luhn, 1958), et reste encore aujourd'hui un domaine largement étudié.

Afin de pouvoir valider les améliorations apportées par des changements de méthode ou de paramètres dans un système de RA, il est nécessaire de disposer de méthodes d'évaluation robustes, si possible automatisées. Jusqu'au début des années 2000, deux types d'évaluation existaient : les évaluations entièrement manuelles avec grille de lecture et les évaluations semi-automatiques qui comparent les résumés automatiques avec des résumés de référence écrits par des humains. Depuis, des approches qui permettent d'évaluer un RA sans référence humaine ont été mises au point, mais ne sont devenues réellement performantes que très récemment, avec l'utilisation de modèles probabilistes (Louis & Nenkova, 2009).

Les résumés automatiques se créent majoritairement par extraction itérative de fragments textuels pertinents. La pertinence d'un fragment est établie d'après son importance au sein des documents source (centralité) et d'après sa similarité aux fragments précédemment sélectionnés pour éviter la redondance (diversité). Les récentes avancées dans l'évaluation entièrement automatique de résumés permettent de penser le résumé différemment. Plutôt que d'extraire itérativement des fragments textuels selon un critère de pertinence sur chacun des fragments, on peut voir l'acte de résumer comme la construction d'un texte guidée par une fonction d'objectif calculée sur le résumé dans intégralité : les mesures d'évaluation automatiques précédemment citées.

Dans cet article, nous proposons une nouvelle méthode de résumé, qui explore l'espace des résumés candidats grâce à un

algorithme génétique pour y trouver une solution approchée de la maximisation d’une fonction d’objectif calculée d’après une vision globale d’un résumé. Dans une première section, nous présentons les méthodes itératives et d’exploration d’espace pour le RA. Nous présentons ensuite notre méthode et l’expérience pour l’évaluer. Enfin, nous présentons nos conclusions et exposons nos perspectives.

2 État de l’art

Les systèmes de RA combinent généralement un score de pertinence pour l’extraction de fragments textuels et une méthode d’extraction des fragments. Les premiers systèmes (Luhn, 1958) extrayaient simplement les fragments les plus pertinents. La méthode MMR (Carbonell & Goldstein, 1998) permet, elle, d’extraire itérativement des phrases selon un score de pertinence et un score de non redondance. La méthode fondée sur CSIS, présentée dans (Radev, 2000), permet d’éliminer, depuis une liste de phrases triée selon la pertinence, toute phrase trop similaire à une autre mieux classée. Ces méthodes possèdent un inconvénient majeur : les résumés générés dépendent pour beaucoup de la sélection de la première phrase. Ainsi, ces méthodes risquent d’omettre des résumés issus de l’assemblage de phrases moyennement classées mais qui combinées ensemble reflètent mieux le contenu des documents à résumer.

D’autres méthodes ont vu le jour récemment, pour pallier ce problème. Il s’agit d’explorer l’espace des résumés possibles et d’en tirer la solution qui maximise une fonction d’objectif. Ce problème est np-complet : choisir 10 phrases parmi 200 conduit à 10^{25} résumés possibles. L’ajout de contraintes sur la sélection de phrases et l’utilisation de la programmation linéaire en nombre entier – *ILP* – (McDonald, 2007; Gillick & Favre, 2009) permet de limiter l’espace de recherche et d’y trouver une solution avec un très faible coût computationnel. L’espace de recherche est limité par des contraintes sur la taille des phrases et d’autres qui empêchent l’inclusion de phrases qui n’apportent aucune information supplémentaire. (Liu *et al.*, 2006) ont proposé une méthode de parcours de l’espace des résumés candidats par un algorithme génétique. Cela permet de s’affranchir des contraintes de la programmation linéaire en nombre entier (fonctions à maximiser et fonctions de contraintes limitées à des fonctions linéaires). Cependant, les méthodes issues de cette dernière famille continuent de considérer un résumé comme un ensemble de phrases indépendantes, et ne profitent pas de la possibilité offerte par les algorithmes de parcours de l’espace de calculer un score d’après une vision globale d’un résumé candidat. Dans l’état de l’art, les algorithmes génétiques sont généralement utilisés en RA non pas pour générer directement un résumé (Litvak *et al.*, 2010), mais pour apprendre de manière supervisée les meilleurs paramètres d’un système.

Nous proposons ici d’utiliser des fonctions de score fondées sur la comparaison de distributions de probabilités construites sur le document source et les résumés candidats. Les lissages utilisés dans la construction des distributions de probabilités considèrent un résumé candidat dans son ensemble et non plus comme un assemblage de phrases indépendantes. Ainsi, les fonctions de score proposées permettent de mieux tirer parti des possibilités des algorithmes génétiques.

3 Notre méthode

Notre méthode est fondée sur l’utilisation d’un algorithme génétique afin d’explorer l’espace des résumés candidats possibles et d’y trouver une solution approchée du meilleur résumé vis-à-vis d’une fonction d’objectif. Les résumés générés sont contraints en nombre de mots. Nous présentons d’abord la fonction d’objectif que nous utilisons avant de détailler l’algorithme génétique.

3.1 Fonction d’objectif

Notre fonction d’objectif consiste à comparer la distribution de probabilités des mots dans les documents source avec la distribution de probabilité des mots dans les résumés. Nous nous sommes fondés sur le travail de (Louis & Nenkova, 2009), dont l’efficacité de l’approche a été confirmée par (Torres-Moreno *et al.*, 2010) et qui utilise la divergence de Jensen-Shannon :

$$J(P||Q) = \frac{1}{2}[D(P||A) + D(Q||A)]$$

avec P et Q deux distributions, $A = \frac{P+Q}{2}$ la distribution moyenne de P et Q et $D(P||A)$ la divergence de Kullback-Leibler définie comme suit :

$$D(P||Q) = \sum_w p_P(w) \log_2 \frac{p_P(w)}{p_Q(w)}$$

Dans l'article de (Louis & Nenkova, 2009), la probabilité d'un mot est lissée d'après un lissage de Laplace modifié :

$$p(w) = \frac{C(w) + \delta}{N + \delta \times 1.5 \times |V|}$$

avec δ réglé à 0.0005, $C(w)$ le nombre d'occurrences de w , N la somme des occurrences sur tous les mots, et V le vocabulaire.

(Lin, 2004) a montré que les mesures d'évaluation semi-automatiques ROUGE étaient plus corrélées aux évaluations manuelles lorsqu'elles utilisent les bigrammes comme modèle pour évaluer des résumés de taille standard : 50 mots ou plus. Aussi faisons-nous l'hypothèse que l'utilisation d'un modèle probabiliste fondé sur les bigrammes et non sur les unigrammes améliorera les résultats de notre fonction d'objectif. De plus, nous lissons les probabilités avec un lissage *Dirichlet*, qui ajoute à tout comptage d'un *token* (dans notre, des unigrammes ou des bigrammes) dans un résumé à évaluer sa probabilité d'apparition dans les documents source :

$$p_d(t|R) = \frac{C_R(t) + \mu p_{ML}(t|S)}{N_R + \mu}$$

où t est un token, R un résumé, S l'ensemble des documents source, $C_R(t)$ le nombre d'occurrences de t dans R , $p_{ML}(t|S)$ l'estimation du maximum de vraisemblance de t dans S , N_R le nombre de *tokens* et μ un paramètre fixe (pseudo-fréquence). Les résumés candidats sont des sous-ensembles des documents source. Les lissages sont donc effectués uniquement pour les résumés

3.2 Algorithme génétique

Définition d'un individu L'algorithme génétique, dans notre cas, ne peut pas être vu au sens traditionnel. En effet, un résumé (un individu) n'est pas constitué d'un nombre fixe de phrases considérées comme des chromosomes. De fait, les résumés sont limités en nombre de mots. Le nombre de phrases extraites dépend donc du nombre de mots dans chacune. Chaque individu est donc composé d'un nombre variable de chromosome, codant chacun pour l'indice d'une phrase dans les documents source. Un résumé pourrait aussi être vu comme un vecteur de variables booléennes codant chacune pour l'extraction ou non d'une phrase. Cependant, cela fait perdre la notion d'ordre dans le résumé dont nous comptons nous servir dans les modèles futurs afin de prendre en compte des scores de cohésion textuelle. De plus, tant que la taille en nombre de mots d'un résumé n'est pas excessive, notre technique tend à restreindre l'espace de recherche.

Déroulement de l'algorithme Une population de départ est créée aléatoirement. Elle contient un nombre d'individus égal à la somme du nombre de parents, du nombre d'individus mutés et du nombre d'individus croisés, des paramètres choisis par l'utilisateur. N_p parents sont alors sélectionnés, qui engendrent par mutation puis par croisement N_m et N_c individus supplémentaires. Les parents, et les individus qu'ils ont contribué à générer forment une nouvelle génération. La sélection d'une nouvelle génération est répétée N_g fois. À la fin de l'algorithme, le meilleur individu selon la fonction d'objectif est sélectionné.

Sélection d'une population de départ $N_p + N_m + N_c$ individus sont créés aléatoirement. Une nouvelle phrase est ajoutée aléatoirement à chaque individu parmi les phrases qui vérifient la contrainte : $\sum_{p_i \in I} Taille(p_i) + Taille(p) <$

$TailleMax$ où I est un individu et p la phrase à tester.

Sélection des parents Il existe différentes méthodes de sélection des parents. Chacune privilégie soit l'exploration de l'espace, soit l'exploitation en sélectionnant les meilleurs individus. Nous avons choisi une forme de sélection qui constitue un compromis entre exploration et exploitation : la sélection par tournoi. N_t tournois de N/N_t individus – N étant la taille totale de la population – sont organisés aléatoirement. Le meilleur individu de chaque tournoi selon la fonction d'objectif est sélectionné comme parent de la génération précédente. Ainsi, le meilleur individu d'une génération est systématiquement sauvegardé, tandis que les autres individus ont une chance de faire partie des parents qui ne dépend pas uniquement de leur score, mais également du tournoi dans lequel ils ont été placés aléatoirement.

Opérateur de mutation Notre opérateur de mutation est défini comme suit : une phrase est supprimée aléatoirement d'un individu. L'individu est complété aléatoirement avec d'autres phrases tant que la somme de la taille de ses phrases n'excède pas la taille maximum autorisée.

Opérateur de croisement Les phrases de deux individus sont mises en commun dans un ensemble. Des phrases sont sélectionnées aléatoirement depuis cet ensemble pour constituer un nouvel individu, toujours en vérifiant la contrainte de taille. Cela correspond à un opérateur de croisement standard mais avec un nombre variable de chromosomes.

Un individu ainsi constitué peut alors avoir une taille en nombre de mots assez inférieure à limite de taille pour se voir ajouter une autre phrase absente des parents. Un tel individu en concurrence avec des individus dont la taille en nombre de mots est maximale serait alors potentiellement pénalisé. L'individu est donc, à la suite du croisement, complété aléatoirement par des phrases issues des documents source et absentes des parents.

4 Expériences

4.1 Protocole

Nous comparons notre méthode de résumé à quatre *baselines* sur le corpus de la campagne d'évaluation internationale TAC 2009. Ce corpus est disponible sur demande à l'adresse <http://www.nist.gov/tac/data/index.html>.

Corpus Le corpus de TAC 2009 comprend deux parties : l'une, qui nous intéresse, est dédiée au résumé standard. L'autre est dédiée au résumé de mise à jour, c'est-à-dire au résumé des informations nouvelles d'un groupe de documents si l'on considère qu'un utilisateur a déjà lu les informations des documents qui ont servi au résumé standard. Nous avons choisi ce corpus car c'est le dernier à proposer une tâche de résumé qui n'est pas guidée par un sujet particulier. Les années suivantes, chaque résumé doit être adapté à un sujet particulier parmi les cinq suivants : accidents et catastrophes naturelles, attaques, santé et sécurité, ressources menacées, et enquêtes et jugements. Réaliser des résumés adaptés à une telle tâche nécessite de prendre en compte les spécificités de chaque tâche. Les dernières campagnes TAC n'entrent donc pas dans le cadre de notre étude.

La partie du corpus de TAC 2009 dédiée au résumé standard est composée de 44 jeux de 10 documents issus d'organismes de presse et rédigés en anglais. Pour chaque jeu de documents, la tâche consiste à générer un résumé en 100 mots maximum. Les documents ont une longueur moyenne de 610 mots.

Notre système Notre système consiste en deux phases : lemmatisation et annotation morpho-syntaxique du corpus avec *tree-tagger*¹ de manière à travailler uniquement avec les formes canoniques des mots, élimination des méta-données des documents à résumer, puis sélection d'un résumé à l'aide de l'algorithme génétique décrit en §3.2. Nous avons implémenté les fonctions d'objectif suivantes :

- divergence Jensen-Shannon, unigrammes, lissage de Laplace modifié (unilap) ;
- divergence Jensen-Shannon, bigrammes, lissage de Laplace modifié (bilap) ;
- divergence Jensen-Shannon, unigrammes, lissage Dirichlet (unidir) ;
- divergence Jensen-Shannon, bigrammes, lissage Dirichlet (bidir) ;
- somme des poids des bigrammes : nombre de documents dans lesquels un bigramme apparaît (bisimple).

Le corpus TAC 2009 propose un corpus de test composé de 3 documents. Nous nous en sommes servis pour choisir manuellement et de manière empirique les paramètres de l'algorithme génétique. Ceux-ci ont été simplement définis de manière à obtenir une convergence en un temps raisonnable. L'algorithme génétique est paramétré comme suit : $N_p = 80$, $N_m = 160$, $N_c = 80$, $N_t = N_p$, et $N_g = 150$.

Baselines Nous avons implémenté deux *baselines*. La première est la méthode largement utilisée de *scoring* LexRank (Erkan & Radev, 2004), suivie d'une étape d'élimination de la redondance par MMR (Carbonell & Goldstein, 1998) (*lexmmr*). La deuxième est la méthode de (Gillick *et al.*, 2009) décrite en §2 ; par souci de reproductibilité, nous avons

1. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

	unilap	bilap	unidir	bidir	bisimple	lexmmr	ilp	ilpheur	hextac
ROUGE-1	0.35490	0.36638	0.35482	0.37625	0.37619	0.33877	0.37009	0.39292	0.37946
ROUGE-2	0.08599	0.09985	0.08493	0.10264	0.10251	0.08428	0.09914	0.12163	0.10655
ROUGE-L	0.17248	0.18752	0.17060	0.19229	0.18811	0.18392	0.18914	0.25416	0.20313

TABLE 1: Résultats moyens des différents systèmes de résumé sur le corpus TAC 2009

implémenté ces méthodes avec le minimum de pré-traitements possibles (communs aux autres systèmes). Elle a été la mieux classée en scores ROUGE-2 sur la campagne d'évaluation TAC 2009.

La *baseline ilpheur* est le système de (Gillick & Favre, 2009) qui a participé à TAC 2009. Ce système utilise des pré-traitements supplémentaires : segmentation en phrases efficace, élimination de parties de textes inutiles dans le corpus (TAC est constitué de dépêches de presse issues de différents organismes avec un formatage), et élimination des phrases qui contiennent des pronoms dont la référence est hors de la phrase. Il double également le poids des concepts qui apparaissent dans la première phrase d'un document. La dernière *baseline* est la *baseline 3* de TAC 2009 : des résumés par extraction générés par des humains (*hextac*).

Importance des premières phrases Le corpus TAC2009 est un corpus d'actualités constitué majoritairement de dépêches et d'articles. Les premières phrases, qui constituent « l'accroche » d'une dépêche, sont généralement considérées comme plus importantes que les autres. (Gillick *et al.*, 2009) doublent le poids des mots qui apparaissent dans la première phrase d'un document, et ce faisant augmentent leur score de 16% sur le corpus TAC2009. Afin de nous comparer plus précisément, nous testons également l'influence de cette modification sur les systèmes *bidir* et *bisimple* et la *baseline ilp* en faisant varier la pondération des mots des premières phrases (cf figure 1).

4.2 Résultats

Le tableau 1 présente les résultats obtenus par les différents systèmes et *baselines* présentés en §4.1 sur le corpus TAC2009. Le meilleur système selon toutes les évaluations est le système *bidir*, qui utilise comme fonction d'objectif la divergence de distribution de bigrammes entre documents source et résumés candidats. Il devance la *baseline ilp*, l'implémentation du système de résumé fondé sur la programmation linéaire en nombre entier. Le même système utilisé avec des pré-traitements différents, qui favorise les bigrammes présents dans les premières phrases (*ilpheur*) et présenté officiellement à TAC2009, lui reste cependant 15% supérieur. Il est toutefois intéressant de constater que l'écart qui sépare le système *bidir* de la *baseline hextac*, générée manuellement, est très faible.

La figure 1 présente les résultats obtenus par les systèmes *bidir*, *bisimple* et la *baseline ilp* en faisant varier le facteur multiplicateur des poids des bigrammes associés aux premières phrases. Le système *bidir* atteint rapidement un pallier puis reste constant. Le système *bisimple* atteint de moins bons score que *bidir*, mais augmenter le poids des bigrammes présents dans les premières phrases améliore globalement ses performances. Quant à la *baseline ilp*, ses résultats croissent avec le poids des bigrammes des premières phrases mais restent en dessous de *bidir*. Ses résultats n'atteignent toutefois pas ceux d'*ilpheur*. Une hypothèse est que la différence avec cette dernière réside dans les pré-traitements supplémentaires qu'elle effectue. Le système *bidir* ainsi que la *baseline ilp* se placent au-dessus de la *baseline hextac*, générée par extraction de phrases manuelle. Cela signifie que le système réussit à extraire autant d'informations essentielles qu'un humain ; cependant, les résumés de la *baseline hextac* ont sûrement une qualité linguistique supérieure.

5 Discussion

Nous avons proposé une méthode qui obtient des résumés de bonne qualité sur le corpus TAC 2009. Grâce à l'utilisation d'un algorithme génétique plutôt qu'un système fondé sur l'ILP, l'expression des scores et des contraintes n'est pas limitée. Cependant, cette méthode ne gère pas spécifiquement la redondance. Un résumé sera jugé bon si sa distribution de probabilités colle au mieux à celle des documents source. Ainsi, si les documents en entrée sont très redondants, ce qui n'est visiblement pas le cas du corpus TAC 2009, la méthode générera des résumés également redondants. Il est toutefois possible de pourvoir notre système de scores qui visent à pénaliser la redondance. La question de la sensibilité au bruit de notre système reste à évaluer.

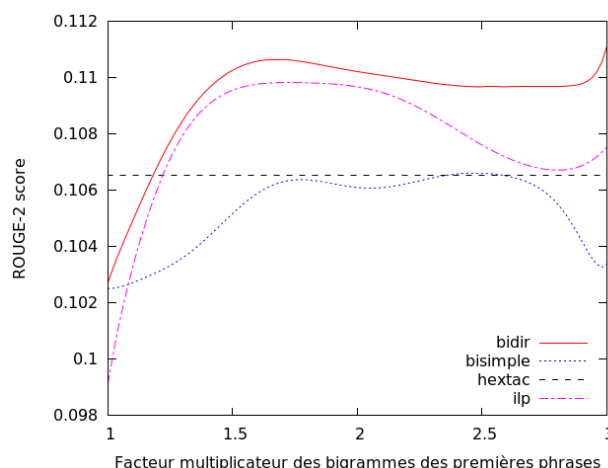


FIGURE 1: Résultats des systèmes *bidir* et *bisimple* et des *baselines* base3 et base4 en fonction du facteur multiplicateur des poids des bigrammes associés aux premières phrases

Le système proposé n'effectue aucun traitement pour obtenir des résumés de meilleure qualité linguistique : le problème de la cohésion par l'articulation des phrases n'est pas géré. Des méthodes existent, comme les chaînes lexicales (Barzilay & Elhadad, 1999) qui peuvent être intégrées à la fonction d'objectif ou utilisées en post-traitement.

Enfin, un algorithme génétique possède des paramètres : taille de la population, pourcentage de mutants et de croisés. Il convient d'étudier leur influence sur la convergence de l'algorithme en utilisant divers jeux de données.

Il est à souligner que des systèmes de RA par extraction réussissent à dépasser (en scores ROUGE) des résumés par extraction générés manuellement. Cela pose naturellement la question de la limite supérieure que l'on peut atteindre avec des méthodes purement extractives, et des méthodes à envisager dorénavant : compression de phrases, utilisation de paradigmes génératifs pour des résumés spécialisés, reformulation de phrases pour une meilleure cohésion textuelle...

Références

- BARZILAY R. & ELHADAD M. (1999). Using lexical chains for text summarization. *Advances in automatic text summarization*, p. 111–121.
- CARBONELL J. & GOLDSTEIN J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR'98 : Proceedings of the 21st ACM SIGIR Conference*, p. 335–336.
- ERKAN G. & RADEV D. R. (2004). Lexrank : Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*.
- GILICK D. & FAVRE B. (2009). A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, p. 10–18 : Association for Computational Linguistics.
- GILICK D., FAVRE B., HAKKANI-TÜR D., BOHNET B., LIU Y. & XIE S. (2009). The ICSI/UTD summarization system at TAC 2009. In *Proceedings of Workshop on Summarization task at TAC 2009 conference*.
- LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, p. 10.
- LITVAK M., LAST M. & FRIEDMAN M. (2010). A new approach to improving multilingual summarization using a genetic algorithm. In *Proceedings of the 48th Annual Meeting of ACL, ACL '10*, p. 927–936.
- LIU D., HE Y., JI D. & YANG H. (2006). Genetic algorithm based multi-document summarization. In Q. YANG & G. WEBB, Eds., *PRICAI 2006 : Trends in Artificial Intelligence*, volume 4099 of *Lecture Notes in Computer Science*, p. 1140–1144. Springer Berlin Heidelberg.
- LOUIS A. & NENKOVA A. (2009). Automatically evaluating content selection in summarization without human models. In *Proc. of the 2009 EMNLP Conference : Volume 1*, p. 306–314 : ACL.
- LUHN H. (1958). The automatic creation of literature abstracts. *IBM Journal*, 2(2), 159–165.

- MCDONALD R. (2007). *A study of global inference algorithms in multi-document summarization*. Springer.
- RADEV D. R. (2000). A common theory of information fusion from multiple text sources step one : cross-document structure. In *Proceedings of the 1st SIGdial workshop*, p. 74–83 : Association for Computational Linguistics.
- TORRES-MORENO J.-M., SAGGION H., DA CUNHA I., SANJUAN E. & VELÁZQUEZ-MORALES P. (2010). Summary evaluation with and without references. *Polibits Research Journal on Computer Science and Computer Engineering and Applications*, **42**.

Identification des unités de mesure dans les textes scientifiques

Soumia Lilia Berrahou^{1, 2} Patrice Buche^{1, 2} Juliette Dibie³ Mathieu Roche^{1, 4}

(1) LIRMM, 161 rue Ada, Montpellier, France

(2) IATE, 2 place Viala, Montpellier, France

(3) MIA, 16 rue Claude Bernard, Paris, France

(4) TETIS, 500 rue Jean-François Breton, Montpellier, France

berrahou@lirmm.fr, Patrice.Buche@supagro.inra.fr

Juliette.Dibie@agroparistech.fr, mathieu.roche@cirad.fr

Résumé. Le travail présenté dans cet article se situe dans le cadre de l'identification de termes spécialisés (unités de mesure) à partir de données textuelles pour enrichir une Ressource Termino-Ontologique (RTO). La première étape de notre méthode consiste à prédire la localisation des variants d'unités de mesure dans les documents. Nous avons utilisé une méthode reposant sur l'apprentissage supervisé. Cette méthode permet de réduire sensiblement l'espace de recherche des variants tout en restant dans un contexte optimal de recherche (réduction de 86% de l'espace de recherché sur le corpus étudié). La deuxième étape du processus, une fois l'espace de recherche réduit aux variants d'unités, utilise une nouvelle mesure de similarité permettant d'identifier automatiquement les variants découverts par rapport à un terme d'unité déjà référencé dans la RTO avec un taux de précision de 82% pour un seuil au dessus de 0.6 sur le corpus étudié.

Abstract.

Identification of units of measures in scientific texts.

The work presented in this paper consists in identifying specialized terms (units of measures) in textual documents in order to enrich a onto-terminological resource (OTR). The first step permits to predict the localization of unit of measure variants in the documents. We have used a method based on supervised learning. This method permits to reduce significantly the variant search space staying in an optimal search context (reduction of 86% of the search space on the studied set of documents). The second step uses a new similarity measure identifying automatically variants associated with term denoting a unit of measure already present in the OTR with a precision rate of 82% for a threshold above 0.6 on the studied corpus .

Mots-clés : ressource termino-ontologique, apprentissage, similarité.

Keywords: onto-terminological resource, learning, similarity.

1 Introduction

Le travail présenté dans cet article se situe dans le cadre de l'identification de termes spécialisés à partir de données textuelles pour enrichir une Ressource Termino-Ontologique (RTO). Les travaux de (McCrae *et al.*, 2011; Cimiano *et al.*, 2011) proposent d'associer une partie terminologique et/ou linguistique aux ontologies afin d'établir une distinction claire entre la manifestation linguistique (le terme) et la notion qu'elle dénote (le concept). Nous nous intéressons à l'enrichissement d'une RTO permettant de modéliser des relations n-aires entre des données quantitatives expérimentales (Touhami *et al.*, 2011), où les arguments peuvent être des concepts symboliques ou des quantités caractérisées par des unités de mesure. En effet, l'extraction des données quantitatives est un enjeu majeur pour de nombreux domaines scientifiques dont l'objectif concerne la capitalisation et la pérennisation des connaissances du domaine. Cependant, la forte variation d'écriture des unités de mesure dans les documents engendre des problèmes d'identification des instances numériques dans les textes.

Dans une démarche consensuelle, le Systeme International (SI) (Thompson & Taylor, 2008) organise, en posant plusieurs définitions formelles, le système des quantités et des unités de mesure. Il définit ainsi des unités de base, i.e. unités simples comme *kilogram*, et des unités dérivées, i.e. unités plus complexes comme $kg.m^{-1}$. Ce standard pose les règles d'écriture de l'ensemble des unités de mesure mais n'intègre pas la notion de variants d'unités. Ces principes sont repris dans des

travaux récents (Rijgersberg *et al.*, 2013) afin de modéliser formellement cette connaissance dans une ontologie dédiée à la représentation des données quantitatives et des unités de mesure. Les auteurs ont ainsi modélisé *OM* (Ontology of Units of Measure and Related Concepts). Les travaux de (Van Assem *et al.*, 2010) posent la problématique d'identification des données quantitatives présentes dans les cellules des tableaux représentés dans les documents. La localisation des variants d'unités n'est pas problématique dans ces travaux car la méthode repose sur le format structuré des tableaux. Les travaux de (Grau *et al.*, 2009) proposent des méthodes d'extraction des données expérimentales dans le domaine biomédical. L'identification des unités de mesure repose sur les unités référencées dans le Systeme International (Thompson & Taylor, 2008), la problématique de l'identification des variants d'unité référencée n'y est pas abordée.

Ainsi, à notre connaissance, les méthodes de l'état de l'art partageant l'objectif d'extraction de données quantitatives, ne permettent pas de résoudre la problématique d'extraction et d'identification des variants d'unités de mesure dispersés dans les documents scientifiques au format textuel non structuré. Dans cet article, nous présentons notre proposition qui tente de répondre à deux questions concernant l'identification des variants d'unités de mesure dans les documents textuels non structurés :

- La question concernant la localisation des variants dans le document. Sachant que nous travaillons sur l'intégralité des documents, nous préférons l'apprentissage afin de prédire la localisation des variants sans poser d'hypothèses préalables.
- La question de l'identification du variant une fois qu'il est localisé. A quel autre terme d'unité de mesure référencée dans la RTO peut-on le rapprocher, en sachant que les termes d'unités répondent à leurs propres règles syntaxiques ? Les méthodes existantes doivent être adaptées à ces nouvelles règles.

Les deux questions de recherche précédentes sont respectivement traitées en sections 2 et 3. Ces propositions sont alors expérimentées en section 4, avant la conclusion et les perspectives décrites en section 5.

2 Localisation d'unité de mesure dans les textes

La première étape de notre méthode illustrée dans la figure 1 consiste à prédire la localisation des variants d'unités de mesure dans les documents. Nous avons utilisé une méthode reposant sur l'apprentissage supervisé. Les pré-traitements choisis pour préparer nos documents reposent sur les étapes ci-contre : segmentation des documents en phrases, tokenisation des phrases, suppression des mots vides de la phrase, annotation automatique des phrases. Cette dernière étape, illustrée sur les figures 2 et 3, consiste à identifier les phrases contenant des concepts d'unités de mesure présents dans une RTO pour constituer un corpus d'apprentissage avec un ensemble d'exemples positifs (cf. phase 1 de la figure 1). Les figures montrent deux cas d'annotation automatique selon que l'unité est considérée comme un ou plusieurs tokens. Pour isoler la fraction du variant \tilde{A} identifier dans la deuxième étape, décrite dans la section 3, nous balayons la phrase de part et d'autre de l'unité identifiée jusqu'à trouver un terme du dictionnaire. Dans nos travaux, nous définissons une nouvelle représentation des documents sous forme de fenêtrages textuelles, représentant un contexte phrastique précis (en considérant une à deux phrases autour de chaque phrase courante – cf. phase 2 de la figure 1). Dans notre contexte d'étude, pour chacune des fenêtrages, nous avons sélectionné uniquement les termes apparaissant plus d'une fois dans le corpus dédié à l'apprentissage pour constituer un *sac de mots* (descripteurs). Cela réduit sensiblement l'espace de représentation des textes sans avoir pour autant un impact sur les résultats de la classification. Les termes ainsi pré-sélectionnés sont projetés sur chaque document. Hormis la représentation booléenne des descripteurs (présence/absence), les descripteurs propres à la représentation vectorielle peuvent être pondérés selon différentes approches statistiques comme TF, TF.IDF et OKAPI qui sont expérimentés en section 4.1 (cf. phase 3 de la figure 1). L'ensemble des vecteurs constitue la nouvelle représentation des documents pour les algorithmes d'apprentissage supervisé. Le but des modèles appris étant de prédire si une phrase d'un ensemble de test est susceptible de contenir une unité de mesure.

La section suivante décrit la deuxième étape du processus : une fois l'espace de recherche réduit aux variants d'unités, elle propose une nouvelle mesure de similarité permettant d'identifier automatiquement les variants découverts par rapport à un terme d'unité déjà référencé dans la RTO.

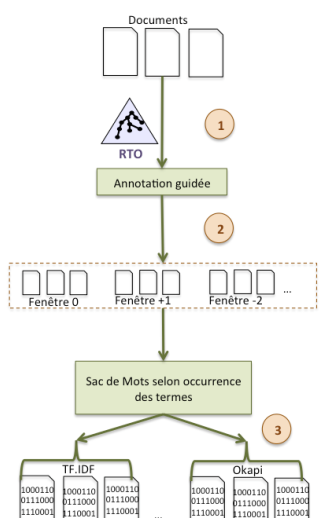


FIGURE 1: Représentation textuelle adaptée au contexte

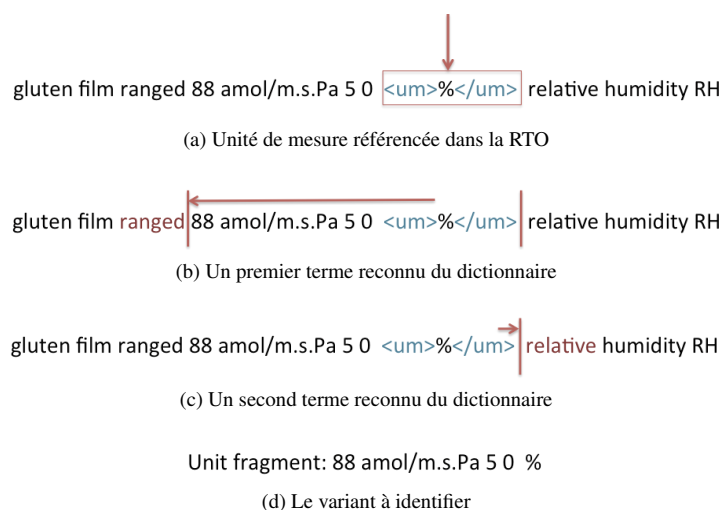


FIGURE 2: Isoler le variant considéré comme un token

3 Identification de nouvelles unités de mesures

Les unités de mesures subissent de fortes variations terminologiques. Un même concept d'unités de l'ontologie peut donc être représenté par des termes d'unités très différents dans les documents, nous les nommons les variants d'unités. Contrairement aux variations terminologiques considérées pour évaluer la similarité entre deux chaînes de caractères, les unités de mesure possèdent leurs propres règles d'écriture établies librement par l'auteur du document.

Par exemple, l'unité de mesure *amol/(m.s.Pa)* définie dans la RTO peut être écrite à l'aide de différents variants dans les documents scientifiques selon les cas :

- d'insertions de caractères comme dans *amol/m.sec.Pa* ou *amol.m-1.s-1.Pa-1*
- de suppressions de caractères comme dans *mol/m.s.Pa*
- d'inversions de certains blocs dans l'unité comme dans *amol.s-1.m-1.Pa-1*
- d'écriture non plus ponctuée mais comme un ensemble composé de blocs indépendants comme dans *amol m-1 s-1 Pa-1*

Nos travaux proposent d'extraire de telles variations terminologiques dans les documents afin d'enrichir la RTO de ces

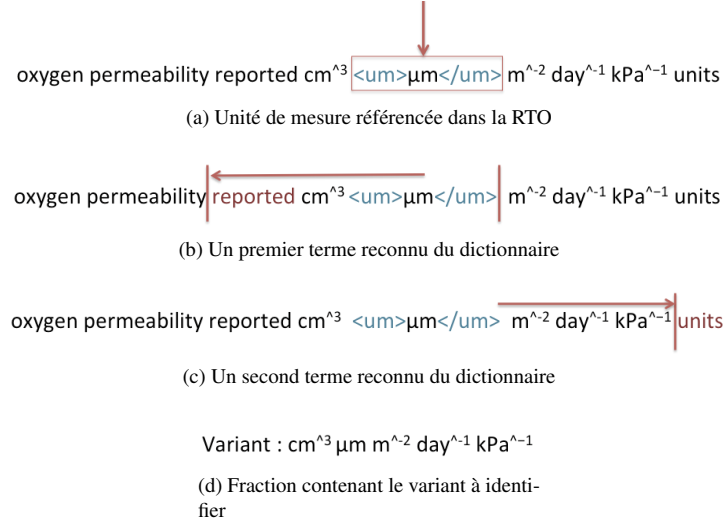


FIGURE 3: Isoler le variant considéré comme plusieurs tokens

variants d'unités de mesure. Une telle extraction ne peut pas reposer sur des méthodes utilisant des expressions régulières car il n'existe aucune règle précise pour les établir du fait de la grande variation typographique des unités de mesure. Nous proposons d'utiliser et adapter les mesures de proximité aux spécificités des unités de mesure afin de répondre à la problématique d'identification des variants d'unités.

Dans notre approche, il est fondamental de pouvoir prendre en considération les particularités d'écriture des unités de mesure, en notant que chaque bloc est indépendant dans l'écriture de l'unité. De ce fait, l'ordre des blocs n'est pas important à prendre en compte, en revanche, la comparaison des blocs entre eux nous semble plus pertinente et plus adaptée dans le calcul de la similarité. Il est alors intéressant de proposer une mesure qui calcule la similarité en deux temps, qui s'appuie à la fois sur les unités déjà référencées dans la RTO et sur les caractères spécifiques (*/*, *(*, *)*, *.*, *×*, *^*...) utilisés comme séparateurs de blocs.

Dans un premier temps, **les candidats sont préselectionnés** selon la mesure de Jaccard. Le principe ci-dessous est alors mis en œuvre :

- Soit un couple composé du variant candidat u_i et d'une unité référence dans la RTO u_j . $J(u_i, u_j)$ (cf. formule (1)) calcule dans un premier temps le score de similarité entre l'ensemble u_i et l'ensemble u_j par rapport aux blocs communs sans tenir compte de leur ordre.

$$J(u_i, u_j) = \frac{|u_i \cap u_j|}{|u_i \cup u_j|} \quad (1)$$

- On sélectionne le couple (u_i, u_j) comme étant pertinent à être comparé si $J(u_i, u_j) > K'$, K' étant le seuil minimal défini préalablement par l'utilisateur.

Prenons l'exemple du couple composé d'un variant candidat localisé et extrait à partir d'un document $kg \text{ Pa}^{-1} \text{s}^{-1} \text{m}^{-2}$ et son référent dans la RTO $lb.m.m^{-2}.s^{-1}.Pa^{-1}$. Dans ce contexte, le calcul de la mesure de Jaccard donne le résultat suivant :

$$J(kg \text{ m Pa}^{-1} \text{s}^{-1} \text{m}^{-2}, lb.m.m^{-2}.s^{-1}.Pa^{-1}) = \frac{4}{6} = 0.7$$

Dans un deuxième temps, après cette phase de présélection, **les candidats sont sélectionnés** selon une mesure étendue de Damereau-Levenshtein.

La mesure de similarité de Levenshtein (Levenshtein, 1966) calcule le coût minimal pour transformer une première chaîne de caractères en une deuxième chaîne de caractères en considérant les opérations de remplacement de caractères entre les deux chaînes, d'ajout d'un caractère ou de suppression d'un caractère. Le coût est ensuite normalisé pour obtenir une valeur de la distance entre les deux chaînes entre 0 et 1. Cette mesure est étendue par Damerau (Damerau, 1964) qui

inclut dans celle de Levenshtein la notion de transposition de caractères d'une chaîne à une autre, i.e. dans *litre* et *liter*, il y a transposition entre les caractères "e" et "r".

La mesure adaptée à notre contexte (cf. formule (2)) ne considère plus la comparaison des caractères mais des blocs de caractères, correspondant à des unités simples. Le variant candidat et l'unité de référence, composant le couple présélectionné lors de la première phase, sont, dans cette seconde phase, comparés bloc à bloc pour déterminer leur similarité finale.

$$SM_{D_b}(u_i, u_j) = \max[0; \frac{\min(|u_i|, |u_j|) - D_b(u_i, u_j)}{\min(|u_i|, |u_j|)}]; SM_{D_b}(u_i, u_j) \in [0; 1] \quad (2)$$

- (u_i, u_j) représente le couple sélectionné à partir de la mesure de Jaccard ;
- Chaque bloc de u_i est comparé aux blocs de u_j pour calculer la nouvelle distance D_b ;
- u_i est validée comme un variant de l'unité u_j si $SM_{D_b} > K$, avec K un seuil de similarité défini préalablement.

En posant $K = 0.5$, le couple $kg\ m\ Pa^{-1}s^{-1}m^{-2}, lb.m.m^{-2}.s^{-1}.Pa^{-1}$, SM_{D_b} calcule la similarité du couple en comparant chaque bloc dans les unités :

$$SM_{D_b}(kg\ m\ Pa^{-1}s^{-1}m^{-2}, lb.m.m^{-2}.s^{-1}.Pa^{-1}) = \max[0; \frac{5-1}{5}] = 0.8$$

Un tel processus fondé sur ces deux phases consécutives de présélection et de sélection finale permet la découverte de nouveaux variants à intégrer comme détaillé dans la section suivante.

4 Expérimentations

Les expérimentations ont été menées à partir d'un corpus de 115 articles scientifiques en anglais issus du domaine des emballages alimentaires. Elles s'appuient également sur une liste de 211 termes dénotant les différents concepts d'unités de mesure pour le domaine des emballages alimentaires. Ces différentes fenêtrures correspondent à des sous-ensembles du corpus représentant 5000 phrases (e.g. f_0) à 15000 phrases (e.g. f_{+2}). Le corpus complet comportant plus de 35000 phrases. Le sac de mots représente un ensemble de 3000 à 4800 descripteurs selon les différentes représentations.

4.1 Évaluation de la méthode de localisation des unités de mesure

Notre objectif, au cours de cette première étape (cf. section 2), est de produire un modèle d'apprentissage appris à partir des données représentées sous forme de fenêtrures textuelles, qui permette de réduire l'espace de recherche des variants d'unités. Nous avons testé plusieurs fenêtrures textuelles. Nous restituons en résultats des expérimentations uniquement ceux révélant les fenêtrures d'étude les plus pertinentes dans le tableau 1. Par souci de lisibilité, les fenêtrures textuelles sont exprimées de la manière suivante :

- f_0 : représente la fenêtrure comportant la phrase où au moins un terme d'unité dénotant un concept de la RTO est identifié.
- f_{+2} : représente la fenêtrure comportant la phrase où au moins un terme d'unité dénotant un concept de la RTO est identifié ainsi que les deux phrases suivantes.
- f_{-2} : représente la fenêtrure comportant la phrase où au moins un terme d'unité dénotant un concept de la RTO est identifié ainsi que les deux phrases précédentes.

Les tableaux 1 et 2 restituent les résultats obtenus sur le corpus des emballages réalisé avec quatre algorithmes d'apprentissage (Naïves Bayes, C4.5, DMNB (Discriminative Multinomial Naïve Bayes), SMO (Sequential minimal optimization qui est une variante de SVM)) et une 10-validation croisée. Le tableau 1 restitue les résultats, toutes mesures confondues. L'analyse des résultats montrent que Naïves Bayes produit une F-mesure allant de 0.85 à 0.88, l'arbre de décision établi sur l'algorithme C4.5 (J48) produit de meilleurs résultats autour de 0.93 à 0.96. DMNB et SMO produisent les meilleurs

résultats, conformément à ce qui est souligné dans la littérature du domaine (0.95 à 0.99). Outre ces résultats analytiques, nous remarquons qu'un plus large contexte, à partir des fenêtres f_{+2} et f_{-2} , n'améliorent pas les résultats d'apprentissage. Nous pouvons donc en déduire que la plus petite fenêtre textuelle, c'est-à-dire celle où au moins un terme d'unité référencé dans la RTO apparaît, est le contexte le plus favorable à la découverte de variants d'unités. Cette conclusion permet de réduire sensiblement l'espace de recherche des variants, i.e. 5000 phrases à considérer plutôt que 35000 initialement dénombrées, tout en restant dans un contexte optimal de recherche (réduction de 86% de l'espace de recherche).

Le tableau 2 synthétise les résultats selon les différentes mesures de pondération et la matrice booléenne pour la fenêtre optimale f_0 . Notre objectif étant d'évaluer quel algorithme produit le modèle restituant des valeurs de F-mesure stables sur les différentes mesures de pondération et la matrice booléenne. La F-mesure, ainsi que les valeurs de précision et de rappel restent stables et élevées avec le modèle DMNB, en restituant une valeur constante autour de 0.95.

	Dec. Tree J48			Naive Bayes			DMNB			SMO		
	P	R	F	P	R	F	P	R	F	P	R	F
f_0	0.99	0.87	0.93	0.83	0.93	0.88	0.95	0.96	0.95	0.99	0.99	0.99
f_{+2}	0.99	0.92	0.96	0.95	0.77	0.85	0.93	0.96	0.95	0.99	0.97	0.99
f_{-2}	0.99	0.92	0.95	0.77	0.98	0.86	0.94	0.96	0.95	0.99	0.97	0.98

TABLE 1: Résultats des instances de la classe "Unit" : Précision (P), Rappel (R), F-mesure (F) restitués pour chaque fenêtre textuelle.

	Dec. Tree J48			Naive Bayes			DMNB			SMO		
	P	R	F	P	R	F	P	R	F	P	R	F
Boolean	0.99	0.87	0.93	0.83	0.93	0.88	0.95	0.96	0.95	0.99	0.99	0.99
TF	0.99	0.86	0.92	0.69	0.85	0.76	0.95	0.96	0.95	0.84	0.90	0.87
TF.IDF	0.99	0.86	0.92	0.69	0.85	0.76	0.95	0.96	0.95	0.84	0.90	0.87
Okapi	0.99	0.86	0.92	0.69	0.86	0.76	0.95	0.96	0.95	0.77	0.88	0.82

TABLE 2: Résultats des instances de la classe "Unit" sur f_0 : Précision (P), Rappel (R), F-mesure (F) restitués pour chaque mesure de pondération et le modèle booléen.

4.2 Évaluation de la méthode d'identification des unités de mesure

Dans la deuxième étape de notre processus (cf. section 3), nous nous appuyons sur les résultats obtenus précédemment afin d'identifier les variants d'unités. Notons que dans notre contexte, nos mesures doivent sélectionner les variants les plus pertinents à présenter aux experts, ce qui revient à minimiser le bruit. Ainsi, nous privilégions la mesure de précision pour évaluer nos propositions.

Le tableau 3 restitue les résultats obtenus avec la nouvelle mesure et permet de les comparer par seuil de similarité :

1. La précision est globalement plus élevée avec le processus complet comparativement à l'application de la seule mesure de présélection (Jaccard), ceci confirme donc l'intérêt d'utiliser notre mesure SMD_b .
2. Les seuils les plus intéressants à exploiter sont au-dessus de 0.6 avec un taux de précision supérieur à 82% après application des deux étapes successives ; l'essentiel des variants sont identifiés.
3. En dessous de 0.5, les résultats se tassent largement. En choisissant de ne considérer que les seuils au-delà de 0.5, nous créons forcément du silence mais un silence "contrôlé". En effet, le processus d'extraction et d'identification des variants étant un processus itératif, les nouvelles unités intégrées dans la RTO favorisent la découverte d'autres variants qui s'expriment dans cette plage de silence.

Le tableau 3 montre la validation des couples variants et unités de référence pertinents (avec K et K' ayant les valeurs 0.5). Un même variant peut former un couple avec plusieurs unités de référence dans la RTO. En effet, prenons l'exemple du variant $amol / m s Pa$, sa comparaison avec les unités de référence $amol/m/s/Pa$, $amol/(m.s.Pa)$, $amol/(m s Pa)$... est considérée comme pertinente. Pour tous les couples pertinents validés, nous n'intégrons qu'une seule fois le variant $amol / m s Pa$. Considérant cette remarque, sur les 267 couples cumulés dans le tableau 3, validés par SMD_b (260 couples sélectionnés), nous obtenons 121 variants d'unités uniques à intégrer dans notre RTO.

Seuil de similarité	Présélection par Jaccard (étape 1)	Précision (étape 1)	Sélection par SMD_b (étape 2)	Précision (étape 2)
[0.9-1]	64	0.87	54	0.84
[0.8-1]	121	0.79	102	0.84
[0.7-1]	238	0.73	209	0.88
[0.6-1]	266	0.75	220	0.82
[0.5-1]	317	0.77	249	0.78
[0.4-1]	479	0.50	267	0.56

TABLE 3: Résultats obtenus avec la nouvelle mesure combinée

5 Conclusion

La méthode proposée, guidée par la connaissance de la RTO, permet à partir d'un processus complet de localisation et d'identification des variants d'unités de mesure, d'enrichir la RTO de nouveaux termes d'unités. Cet enrichissement est une étape fondamentale car les problématiques d'identification des unités de mesure complexes sont une des causes des problématiques d'identification et d'extraction des instances d'arguments quantitatifs.

Nous avons montré que la première étape de notre méthode, s'appuyant sur l'apprentissage supervisé, permet de localiser automatiquement les variants d'unité dans un contexte phrastique optimal de recherche. La méthode repose sur une nouvelle représentation des données, sous forme de fenêtrages textuelles d'étude, que nous obtenons en étant guidé par la RTO. Par la suite, nous avons proposé une nouvelle mesure de similarité adaptée aux spécificités des unités de mesure. Le choix de la mesure de Damereau-Levenshtein est appropriée à notre contexte car elle prend en charge toutes les variations constatées pour les unités de mesure. De plus, associée à l'indice de Jaccard, la nouvelle mesure permet de rapprocher les couples variant-unité référencée de manière plus pertinente en octroyant un premier score global de similarité qui ne tient pas compte de l'ordre des blocs dans la construction de l'unité complexe. Dans un second temps, la nouvelle mesure SMD_b affine ce rapprochement en comparant chaque bloc du variant et de l'unité référencée sélectionnée. Le processus d'enrichissement de la RTO étant un processus itératif, une nouvelle phase d'extraction et d'identification permettrait alors de comparer d'autres variants avec ces nouvelles unités intégrées dans la RTO, qui deviennent des référents.

Références

- CIMIANO P., BUITELAAR P., MCCRAE J. & SINTEK M. (2011). LexInfo : A declarative model for the lexicon-ontology interface. *J. Web Sem.*, **9**(1), 29–51.
- DAMERAU F. (1964). A technique for computer detection and correction of spelling errors. *Commun. ACM*, **7**(3), 171–176.
- GRAU B., LIGOZAT A.-L. & MINARD A.-L. (2009). Corpus study of kidney-related experimental data in scientific papers. In *Proceedings of the Workshop on Biomedical Information Extraction*, p. 21–26.
- LEVENSHTEIN V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, **10**, 707.
- MCCRAE J., SPOHR D. & CIMIANO P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th extended semantic web conference on The semantic web : research and applications - Volume Part I*, ESWC'11, p. 245–259, Berlin, Heidelberg : Springer-Verlag.
- RIJGERSBERG H., VAN ASSEM M. & TOP J. (2013). Ontology of units of measure and related concepts. *Semantic Web*.
- THOMPSON A. & TAYLOR B. N. (2008). Guide for the use of the international system of units (SI).
- TOUHAMI R., BUCHE P., DIBIE-BARTHÉLEMY J. & IBANESCU L. (2011). An Ontological and Terminological Resource for n-ary Relation Annotation in Web Data Tables. In *International Conference ODBASE, OTM Workshops 2011*, volume 7045, p. 662–679 : Lecture Notes in Computer Science series.
- VAN ASSEM M., RIJGERSBERG H., WIGHAM M. & TOP J. (2010). Converting and annotating quantitative data tables. *The Semantic Web–ISWC 2010*, p. 16–31.

Évaluation intrinsèque et extrinsèque du nettoyage de pages Web

Gaël Lejeune¹, Romain Brixtel², Charlotte Lecluze³

(1) LINA, Université de Nantes, 2 rue de la Houssinière, 44322 Nantes, France

(2) Université de Lausanne – HEC - Département de comportement organisationnel, Quartier Dorigny, 1015 Lausanne, Suisse

(3) GREYC, Campus Côte de Nacre, Boulevard du Maréchal Juin, 14032 CAEN cedex 5, France
prenom.nom@univ-nantes.fr, unil.ch, unicaen.fr

Résumé. Le nettoyage de documents issus du web est une tâche importante pour le TAL en général et pour la constitution de corpus en particulier. Cette phase est peu traitée dans la littérature, pourtant elle n'est pas sans influence sur la qualité des informations extraites des corpus. Nous proposons deux types d'évaluation de cette tâche de *détourage* : (I) une évaluation intrinsèque fondée sur le contenu en mots, balises et caractères ; (II) une évaluation extrinsèque fondée sur la tâche, en examinant l'effet du détourage des documents sur le système placé en aval de la chaîne de traitement. Nous montrons que les résultats ne sont pas cohérents entre ces deux évaluations ainsi qu'entre les différentes langues. Ainsi, le choix d'un outil de détourage devrait être guidé par la tâche visée plutôt que par la simple évaluation intrinsèque.

Abstract.

Intrinsic and extrinsic evaluation of boilerplate removal tools

In this article, we tackle the problem of evaluation of web page cleaning tools. This task is seldom studied in the literature although it has consequences on the linguistic processing performed on web-based corpora. We propose two types of evaluation : (I) an intrinsic (content-based) evaluation with measures on words, tags and characters ; (II) an extrinsic (task-based) evaluation on the same corpus by studying the effects of the cleaning step on the performances of an NLP pipeline. We show that the results are not consistent in both evaluations. We show as well that there are important differences in the results between the studied languages. We conclude that the choice of a web page cleaning tool should be made in view of the aimed task rather than on the performances of the tools in an intrinsic evaluation.

Mots-clés : Nettoyage de pages Web, collecte de corpus, évaluation intrinsèque, évaluation extrinsèque, détourage.

Keywords: Web page cleaning, corpus collecting, intrinsic evaluation, extrinsic evaluation, web scraping.

1 Introduction

La quantité grandissante de documents numériques disponibles permet de disposer de corpus pour différentes tâches de Traitement Automatique des Langues (TAL). Cependant, le format des documents n'est pas toujours adapté aux modules placés en aval de la chaîne de traitement de TAL. Ces documents contiennent des éléments non-informatifs qu'il convient de détecter pour faciliter les analyses ultérieures. Aussi, un même rendu peut être généré par des codes sources différents : il n'y a pas de biunivocité source-rendu. L'opération d'extraction du contenu des documents HTML peut être nommée suppression du squelette de page (*boilerplate removal*), détection de modèle (*Web Page Template Detection*) ou plus généralement nettoyage (*Web Page Cleaning*) de page Web. Toutefois ces termes, et en particulier « nettoyage », sont réducteurs vis-à-vis de l'importance de cette étape dans la chaîne de traitement. Nous proposons d'utiliser le terme de *détourage*. Issu de la photographie, ce terme désigne le fait de n'extraire d'une illustration que les parties utiles. Le détourage de pages Web consiste à extraire le texte recherché à partir des données brutes et du rendu tout en conservant certaines données de structure (titraisons, paragraphes...). Cette opération est effectuée par un *détoureur*. Nous proposons dans cet article deux types d'évaluation pour le détourage :

Évaluation intrinsèque Nous utilisons une évaluation fondée sur le contenu détourné. Cette modalité d'évaluation est la plus fréquente dans la littérature (Endrédý & Novák, 2013). Nous exploitons les métriques de la compétition CLEANVAL (Baroni *et al.*, 2008) : distance d'édition au niveau des mots avec ou sans balise(s). Nous y ajoutons une distance d'édition sur les caractères pour permettre une meilleure évaluation sur le chinois.

Évaluation extrinsèque Nous proposons une évaluation par la tâche qui consiste à mesurer la qualité d'un détoureur en fonction des résultats obtenus par un système placé en aval de la chaîne de traitement. Nous utilisons pour ce faire DANIEL (Brixtel *et al.*, 2013), un système de veille multilingue, ainsi que le corpus de référence associé.

Le corpus que nous utilisons est composé d'articles de presse uniquement, cible principale de la campagne CLEANVAL. Dans la Section 2, nous exposons la problématique du détournement. Puis nous détaillons dans la Section 3 les caractéristiques des différents détournements comparés. La Section 4 est consacrée à la description du corpus et à l'évaluation. La Section 5 propose une conclusion de notre travail et des perspectives d'évolution.

2 La problématique du détournement des pages Web

Pour l'humain, la détection du contenu purement textuel ne semble pas poser de difficulté. Le contenu apparaît la plupart du temps au centre de la page et le titre permet de fixer rapidement un point de départ pour la lecture. Toutefois, automatiser ce processus de reconnaissance du contenu textuel reste un défi à l'heure actuelle. Ceci est illustré par la Figure 1 qui présente deux exemples de détournements erronés sur deux sources différentes dans les premiers résultats de *Google News*¹.

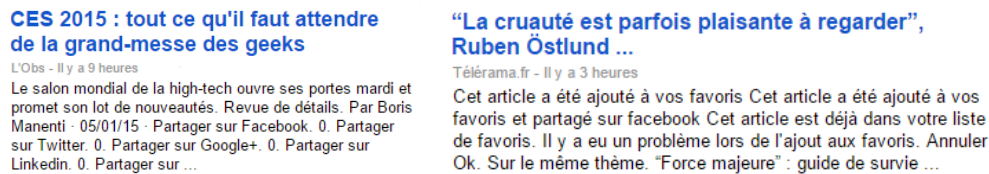


FIGURE 1: Deux exemples d'erreurs de détournement sur des chapeaux d'articles tirés de *Google News*.

Le processus de détournement peut être décomposé en deux sous-tâches : le **nettoyage** (suppression du code JAVASCRIPT, du style, des menus, entêtes et pieds de page) et la **structuration** (récupération des titres, des paragraphes, des listes...). L'approche la plus intuitive pour détourner les pages Web est l'exploitation du *Document Object Model* (DOM), tel que dans les travaux de Chakrabarti *et al.* (2008) et Vieira *et al.* (2006). Dans ces derniers, différentes pages issues du même site Web sont exploitées pour identifier les similarités du DOM. Ce qui est structurellement commun correspond au contenu non-informatif (publicités, liens de navigation) et ce qui est structurellement différent constitue le contenu informatif. D'autres approches représentent la spécificité des pages Web d'un même site sous la forme d'un arbre. La position relative des nœuds permet alors de classer les différents segments dans la catégorie « informatif » ou « non-informatif » (Das *et al.*, 2012). La densité en balises HTML est une autre façon d'exploiter le code source pour classer les segments (Ferraresi *et al.*, 2008). D'autres approches exploitent la distribution des n-grammes de caractères comme dans NCLEANER (Evert, 2008) ou VICTOR (Spousta *et al.*, 2008). L'utilisation combinée de la densité en balises et de la distribution de n-grammes de caractères a été proposée par Pasternack & Roth (2009).

3 Caractéristiques des détournements utilisés

Les détournements font appel à quatre niveaux d'analyse : le site Web (caractéristiques inhérentes à différentes pages d'un même site), le rendu (simulation du rendu de la page donné par un navigateur), la structure HTML (informations de hiérarchie entre les blocs) ou le contenu textuel (phrases, mots, caractères...). Nous nous concentrons ici sur trois outils librement disponibles : Boilerpipe, NCleaner et Justext.

3.1 Boilerpipe

Boilerpipe² est probablement le détournement le plus utilisé dans la communauté TAL. Il est basé sur une combinaison de critères locaux (internes aux blocs) et contextuels (relatifs aux blocs voisins). Les balises considérées comme les plus communes dans les zones textuelles sont utilisées (balises de titres <h1> à <h6>, de paragraphes <p> et de conteneurs <div>). Les balises de liens <a>, les mots capitalisés, les liens hypertextes ou le caractère « | » sont des indicateurs

1. Consultés respectivement le 5 janvier 2015 et le 28 janvier 2015.

2. <http://code.google.com/p/boilerpipe/> (consulté le 1er juin 2015)

de contenu non-informatif ; les points ou les virgules sont la marque de segments informatifs. Les indices contextuels se fondent sur une hypothèse de position relative des blocs de texte et de squelette : les blocs informatifs sont souvent consécutifs. L'étiquette « informatif » ou « non-informatif » d'un bloc est donc fortement dépendante de l'étiquette du bloc précédent. C'est une comparaison des densités en *tokens* (mots graphiques) par ligne dans l'affichage (largeur estimée à 80 caractères) qui permet de juger si l'étiquette du bloc doit changer.

3.2 NCleaner

Le détoureur NCleaner³ (Evert, 2008), présenté lors de la compétition CLEANEVAL, utilise des modèles de langue en n -grammes de caractères. NCleaner mesure la probabilité qu'un caractère appartienne à la langue sachant les caractères qui le précèdent. NCleaner cherche à identifier les n -grammes (avec $1 \leq n \leq 3$) qui maximisent la probabilité d'appartenance d'un bloc au contenu informatif et ceci pour chaque langue. Trois configurations de base sont possibles :

Par défaut (NC) : Utilise un modèle n -gramme « indépendant » de la langue ;

Non-lexical (NCNL) : Transforme les lettres en a ($[: \alpha :] \rightarrow a$) et les chiffres en 0 ($[: \text{digit} :] \rightarrow 0$) ;

Avec entraînement (NCT x) : Se base sur un échantillon de x paires de documents ($d_{\text{brut}}, d_{\text{détouré}}$) fournies au système, d_{brut} étant un document non détourné et $d_{\text{détouré}}$ étant un attendu de document détourné à partir de d_{brut} .

3.3 Justext

Justext⁴ est un détoureur plus récent qui dépasse les résultats de BOILERPIPE selon les évaluations menées par son auteur (Pomikálek, 2011). La méthode utilisée comporte deux étapes. La première étape (dite *context-free*) consiste à collecter trois traits pour chaque bloc de texte : sa longueur en *tokens*, le nombre de liens qu'il contient et, optionnellement, la quantité de mots outils en faisant appel à une ressource externe qui existe pour 100 langues. Nous utilisons dans cet article deux configurations : avec et sans ressource(s) externe(s). En fonction de ces indices, chaque bloc reçoit une première étiquette :

Bad : Bloc de squelette.

Near good : Bloc probablement informatif.

Good : Bloc informatif.

Short : Bloc trop court pour être étiqueté.

La seconde étape (dite *context-sensitive*) consiste à étiqueter les blocs sans étiquette en fonction des étiquettes de leurs voisins. Un bloc de type *short* devient informatif s'il est entouré de blocs de la classe *good* ou *near-good*. Un bloc de type *near-good* est considéré comme informatif si le bloc qui le suit ou le bloc qui le précède est lui-même un bloc informatif.

4 Corpus et modalités d'évaluation

Nous présentons dans la Section 4.1 le format des données et les modalités d'évaluation intrinsèque proposés lors de CLEANEVAL (Baroni *et al.*, 2008). Le corpus d'évaluation ainsi que l'outil choisi pour l'évaluation extrinsèque sont présentés dans la Section 4.2. Enfin, nous décrivons les résultats obtenus dans la Section 4.3.

4.1 Format des textes et modalités d'évaluation de CLEANEVAL

Le format de référence a été obtenu par le travail d'annotateurs humains munis d'instructions précises⁵. La tâche consistait à enlever les traces du squelette de page, les codes HTML et JAVASCRIPT et ne conserver qu'une structure de texte simplifiée utilisant trois balises : `<h>` pour les titres, `<p>` pour les paragraphes et `` pour les éléments de listes.

Le format de texte amène à évaluer deux aspects : la séparation du contenu et du squelette d'une part et la conservation de la structure du texte d'autre part. Le script d'évaluation de la campagne CLEANEVAL considère deux grains : les mots seuls (TO : *text only*) et les mots avec les balises (TM : *text and markup*). Pour le second cas, deux modalités sont proposées : *labelled* qui tient compte du type de balise et *unlabelled* qui n'en tient pas compte (la séquence `<p><p>` est alors équivalente à `<p><p><p>`). Pour chaque fichier, la version détournée automatiquement est comparée avec la

3. http://webasrcorpus.sourceforge.net/PHITE.php?page=FILES_10_Software (consulté le 1er juin 2015)

4. <http://nlp.fi.muni.cz/projects/justext/> (consulté le 1er juin 2015)

5. http://cleaneval.sigwac.org.uk/annotation_guidelines.html (consulté le 1er juin 2015)

version de référence. Chaque version est normalisée en deux étapes : remplacement des caractères de contrôle (sauts de lignes, tabulations ...) par des espaces et normalisation des espaces (les espaces consécutifs sont remplacés par un seul).

Chaque version est découpée à chaque signe de ponctuation ou espace en une séquence de *tokens*. Deux séquences de *tokens* sont comparées en utilisant l'algorithme de Ratcliff (Ratcliff & Metzner, 1988). Calculer la similarité entre elles revient à diviser le nombre de *tokens* en commun par le nombre de *tokens* dans les deux séquences. Les *tokens* en commun sont extraits de la plus longue sous-séquence commune, puis récursivement sur les *tokens* en commun autour de cette sous-séquence. L'algorithme de Ratcliff permet d'obtenir la liste des opérations permettant de passer d'une séquence à l'autre (Baroni *et al.*, 2008). Le fait que les métriques sur les mots et sur les balises soient intriquées gêne l'interprétation des résultats : un détoureur qui n'extraierait que les mots du texte sans structure aurait un score convenable dans la configuration TM. Deux détoueurs proches suivant l'évaluation TO peuvent présenter des résultats différents dans la configuration TM. L'utilisation du grain mot dans les deux modalités d'évaluation de CLEANVAL pose problème pour des langues telles que le chinois. Nous introduisons donc une évaluation par caractère destinée à mieux évaluer les performances des détoueurs, sur le chinois notamment. Pour cette modalité d'évaluation, un *token* correspond à un caractère.

4.2 Corpus de référence et outil pour l'évaluation extrinsèque

L'outil DANIEL et son corpus associé ont été proposés par Brixtel *et al.* (2013) pour la classification de documents pertinents ou non-pertinents pour la veille épidémiologique. L'outil utilise à la fois des éléments de contenu et des éléments de structure, ce qui permet de mesurer la conservation de la structure à l'issue du nettoyage. Nous comparons les résultats obtenus sur les documents détournés manuellement (détournage supposé « idéal ») avec ceux obtenus sur des documents détournés automatiquement. Le corpus associé contient originellement plus de 2000 documents en cinq langues (anglais, chinois, grec, polonais et russe). Le corpus permet d'évaluer la variation en langue et en système d'écriture. Les fichiers n'étant pas directement disponibles, nous avons donc utilisé les *urls* fournies pour les télécharger⁶. Nous avons pu collecter 80% des documents bruts utilisés par DANIEL (Table 1) avec une répartition comparable entre les deux classes.

Langues	Anglais	Chinois	Grec	Polonais	Russe	Toutes
#documents du corpus d'origine	475	446	390	352	426	2089
#documents pertinents	31	16	26	30	41	144
#documents retrouvés	475 (100%)	405 (90,8%)	273 (70%)	274 (77,8%)	267 (62,7%)	1694 (81,1%)
#documents pertinents retrouvés	31 (100%)	16 (100%)	17 (65,4%)	27 (90%)	29 (70,7%)	120 (83,3%)

TABLE 1: Corpus DANIEL, documents présents dans le corpus d'origine et documents retrouvés.

4.3 Évaluation intrinsèque vs. évaluation extrinsèque

La Table 2 présente les résultats de différents détoueurs et de leurs combinaisons en utilisant les métriques de *Cleaneval* auxquelles nous avons ajouté une évaluation par caractère. Pour les Tables suivantes, TO désigne l'évaluation sur le texte seul (*Text Only*), TM l'évaluation sur le texte avec les balises (*Text and Markup*) et CAR désigne l'évaluation au grain caractère.

Les détoueurs utilisés sont également désignés par des abréviations : BP correspond à BOILERPIPE, JTA et JTS à JUSTEXT dans ses deux configurations (respectivement avec (JTA) et sans ressource externe (JTS)), NC est la configuration par défaut de NCLEANER, NCT5 et NCT25 sont les configurations utilisant respectivement 5 et 25 paires de textes pour l'apprentissage. Cinq paires ont été constituées manuellement pour chaque langue. Pour NCT5, les cinq paires sont utilisées pour chaque langue et une paire par langue pour le corpus complet (la paire où le document détourné est le plus long en caractères). Pour NCT25, les 25 paires sont utilisées pour chaque langue de même que pour le corpus complet. Les mesures utilisées sont le rappel (*R*), la précision (*P*) et la F_1 -mesure (F_1). Les combinaisons de détoueurs consistent à exploiter le résultat d'un premier détoueur en entrée d'un second détoueur. Il s'agit d'analyser la complémentarité des détoueurs. La Table 2 présente les résultats de l'évaluation intrinsèque pour l'ensemble du corpus⁷.

Sur l'évaluation intrinsèque JT est surclassé par BP dans les 3 catégories de *tokens* évalués. NC dans sa configuration par défaut présente des résultats intéressants en terme de précision mais au prix d'un rappel moindre. Les autres configurations de NC offrent des résultats peu significatifs.

La Table 3 présente les résultats de l'évaluation extrinsèque, pour chacune des langues ainsi que pour le corpus complet. La

6. https://daniel.greyc.fr/public/api_daniel.php (consulté le 1er juin 2015)

7. Les résultats des chaînages impliquant NCLEANER ont été omis car les performances sont faibles et n'apportent aucune information intéressante.

Mesures	TO			CAR			TM		
	P	R	F ₁	P	R	F ₁	P	R	F ₁
BP	81.80	88.89	85.20	76.93	81.12	78.97	64.47	85.42	73.48
BP-JTA	85.01	80.20	82.54	76.94	63.86	69.79	73.30	58.74	65.22
BP-JTS	83.23	82.87	83.05	75.12	66.03	70.28	69.22	62.07	65.45
JTA	68.75	83.41	75.37	63.79	67.03	65.37	61.94	63.23	62.58
JTA-BP	72.54	85.86	78.64	69.43	73.28	71.31	66.76	69.34	68.02
JTS	62.68	86.30	72.62	56.93	68.63	62.23	54.24	66.57	59.78
JTS-BP	66.31	88.74	75.90	62.95	75.76	68.77	59.42	72.70	65.39
NC	98.53	39.38	56.27	96.65	23.15	37.36	89.01	30.82	45.78
NCNL	01.28	02.73	01.74	01.30	02.97	01.81	02.01	04.14	02.71
NCT5	60.43	23.83	34.18	53.81	16.03	24.70	48.41	19.89	28.19
NCT25	56.14	25.70	35.26	53.25	18.73	27.72	45.11	21.77	29.36

TABLE 2: Résultats de l'évaluation intrinsèque pour le corpus complet en cinq langues (Précision, Rappel et F₁-mesure exprimés en %) sur les grains *Text Only* (TO), CARactère (CAR) et *Text and Markup* (TM).

ligne « Référence » indique le résultat attendu à partir du détournage manuel⁸. Nous observons que l'évaluation extrinsèque établit une hiérarchie différente de celle issue de l'évaluation intrinsèque. JT, dans sa version avec ou sans ressource(s), est meilleur que BP en terme de F₁-mesure hormis sur le sous-corpus chinois. La meilleure option est cependant le chaînage BP-JT. Les résultats obtenus sur le corpus chinois sont identiques pour JTA et JTS, ceci est dû à l'absence de ressource externe pour cette langue. Ces résultats sont par ailleurs strictement équivalents à ceux obtenus pour les deux chaînages JTA-BP et JTS-BP. NC obtient des résultats intéressants sur l'anglais et dans une moindre mesure sur le polonais. Nous pouvons remarquer une grande variabilité des résultats selon les langues sur l'évaluation extrinsèque.

Mesures	Anglais			Chinois			Grec			Polonais			Russe			Corpus complet		
	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁	P	R	F ₁
BP	60.00	25.71	36.00	78.95	93.75	85.71	85.71	35.94	50.00	76.47	48.15	59.09	76.19	55.17	64.00	74.68	47.58	58.13
BP-JTA	61.54	45.71	52.46	71.43	31.25	43.48	66.67	70.59	68.57	65.63	77.78	71.19	76.67	79.31	77.97	68.14	62.10	64.98
BP-JTS	65.22	42.86	51.72	71.43	31.25	43.48	63.16	70.59	66.67	64.71	81.48	72.13	74.19	79.31	76.67	67.54	62.10	64.71
JTA	55.17	45.71	50.00	66.67	37.50	48.00	59.09	76.47	66.67	59.26	59.26	59.26	82.14	79.31	80.70	64.35	59.68	61.92
JTA-BP	59.09	37.14	45.61	66.67	37.50	48.00	62.50	58.82	60.61	65.38	62.96	64.15	76.00	65.52	70.37	66.33	52.42	58.56
JTS	55.56	42.86	48.39	66.67	37.50	48.00	66.67	58.82	62.50	56.67	62.96	59.65	82.61	65.52	73.08	64.42	54.03	58.77
JTS-BP	60.87	40.00	48.28	66.67	37.50	48.00	66.67	47.06	55.17	62.96	62.96	62.96	78.26	62.07	69.23	67.02	50.81	57.80
NC	58.33	60.00	59.15	N/A	0.00	N/A	N/A	0.00	N/A	80.00	14.81	25.00	N/A	0.00	N/A	60.98	20.16	30.30
NCNL	50.00	14.29	22.22	N/A	0.00	N/A	N/A	0.00	N/A	23.81	18.52	20.83	100	6.90	12.90	26.09	9.68	14.12
NCT5	52.94	25.71	34.62	83.33	31.25	45.45	N/A	0.00	N/A	82.35	51.85	63.64	60.00	20.69	30.77	62.96	27.42	38.20
NCT25	50.00	25.71	33.96	83.33	31.25	45.45	20.00	5.88	9.09	82.35	51.85	63.64	61.54	27.59	38.09	62.71	29.84	40.44
Référence	68.89	88.57	77.50	80.00	100	88.89	68.42	76.47	72.22	61.76	77.78	68.85	72.73	82.76	77.42	69.54	84.68	76.36

TABLE 3: Résultats de l'évaluation extrinsèque (anglais, chinois, grec, polonais, russe et corpus complet) avec N/A représentant les valeurs non-calculables (nombre nul de vrais positifs).

La Table 4 récapitule pour chaque langue et chaque mesure, l'outil ou le chaînage le plus performant et le score associé. BP est souvent l'outil le plus efficace (première place dans un cas sur deux). Toutefois, il existe de nombreuses configurations où d'autres outils ou chaînages (notamment BP-JTA et BP-JTS) offrent de meilleurs résultats.

Si BP donne les meilleurs résultats sur l'évaluation intrinsèque, le choix de ce détournement semble moins évident lorsque l'on s'intéresse à la tâche. De plus, la qualité d'extraction du contenu informatif est fortement dépendante de la langue traitée. Le corpus polonais est le seul pour lequel on observe des résultats comparables sur les deux évaluations ($F_1 = 72.75$ pour l'évaluation intrinsèque TM, $F_1 = 72.13$ pour l'évaluation extrinsèque). Le contenu est correctement extrait et les résultats de l'évaluation extrinsèque sont très bons et même meilleurs dans certains cas que ceux obtenus sur la version détournée manuellement. Ceci semble dû au fait que les résultats pour le polonais du système DANIEL sont les plus faibles des cinq langues étudiées ($F_1 = 68.85$), la segmentation opérée par le système est plus sujette à caution sur ce sous-corpus. Les résultats obtenus sur le russe sont assez différents de ceux obtenus sur le polonais, bien que les deux langues soient apparentées. L'évaluation intrinsèque et l'évaluation extrinsèque donnent de mauvais résultats. Par ailleurs, le russe est la seule langue pour laquelle BP ne présente pas les meilleures performances pour aucun des traits évalués.

BP donne les meilleurs résultats sur le chinois, et dans une moindre mesure sur le grec et le polonais. Dès lors que

8. Les résultats sont différents de ceux présentés par Brixstel *et al.* (2013) car seuls les documents dont les sources HTML ont été retrouvées sont comptabilisés.

	TO			CAR		
	P	R	F_1	P	R	F_1
anglais	BP-JTA (86.98)	BP (92.02)	BP (88.89)	BP-JTA (87.60)	BP (91.03)	BP (87.87)
chinois	BP (61.32)	BP (52.90)	BP (56.80)	BP (77.12)	BP (63.55)	BP (69.68)
grec	BP-JTA (93.62)	BP (96.48)	BP (94.10)	BP (87.58)	BP (91.59)	BP (89.54)
polonais	BP-JTA (85.24)	JTS-BP (87.54)	BP (84.26)	BP-JTA (82.95)	JTS-BP (82.50)	BP (81.42)
russe	BP-JTA (67.77)	JTS-BP (86.81)	BP-JTA (69.92)	BP-JTA (53.65)	JTS-BP (79.96)	JTA-BP (59.56)
toutes	NC (98.53)	BP (88.89)	BP (85.20)	NC (96.65)	BP (81.12)	BP (78.97)

	TM			Tâche		
	P	R	F_1	P	R	F_1
anglais	BP-JTA (81.09)	BP (93.59)	BP-JTS (79.79)	BP-JTS (65.22)	NC (60.00)	NC (59.15)
chinois	JT-BP (89.91)	BP (67.99)	BP (75.34)	NCT (83.33)	BP (93.75)	BP (85.71)
grec	BP-JTA (86.18)	BP (91.67)	BP-JTS (82.90)	BP (85.71)	JTA (76.47)	BP-JTA (68.57)
polonais	BP-JTA (76.27)	BP (85.57)	BP (72.75)	NCT (82.35)	BP-JTS (81.48)	BP-JTS (72.13)
russe	BP-JTA (48.63)	JTS-BP (86.68)	JTA-BP (58.74)	NCNL (100)	BP-JT, JTA (79.31)	JTA (80.70)
toutes	NC (89.01)	BP (85.42)	BP (73.48)	BP (74.68)	BP-JT (62.10)	BP-JTA (64.98)

TABLE 4: Outils et chaînages offrant les meilleurs scores pour chaque métrique d'évaluation et sous-corpus.

l'on intègre le balisage dans l'évaluation (TM), l'écart entre BP et JT se réduit, exception faite du chinois. NC affiche des résultats variables (en langue et en type d'évaluation) pour être considéré comme fiable. Les meilleurs résultats sur l'évaluation extrinsèque sont obtenus par le chaînage BP-JTA : BP détecte plus de contenu que nécessaire, contenu qui est ensuite rogné par JTA. Les résultats sur l'évaluation intrinsèque TM sont les plus corrélés avec ceux de l'évaluation par la tâche. L'évaluation TM apporte donc le plus d'indices sur le choix du détoureur. Ceci est dû au fait que DANIEL, l'outil utilisé pour l'évaluation extrinsèque, requiert des informations de contenu et de structure.

5 Discussion

Nous avons procédé à une évaluation intrinsèque et extrinsèque d'outils de détournement. Notre questionnaire a été le suivant : quelle influence a le détournement sur les résultats d'un système placé en aval de la chaîne de traitement. Nous avons montré que les détourneurs offrant les meilleurs résultats dans l'évaluation intrinsèque ne garantissent nullement de bons résultats pour l'évaluation extrinsèque. Enfin, les résultats ne sont pas constants entre les langues. De manière générale, nous voyons le détournement comme l'illustration d'une problématique rarement abordée qui est celle des erreurs en cascade dans une chaîne de traitement de TAL. L'évaluation de l'influence d'un outil sur les résultats d'un autre n'est pas suffisamment explorée, surtout dans des processus impliquant le chaînage de nombreux outils (détournement, lemmatisation, étiquetage...). Le choix d'un outil de détournement ne devrait se faire qu'en fonction de la tâche visée car les résultats obtenus sur l'évaluation intrinsèque ne donnent que de maigres indications sur la pertinence de l'outil en conditions réelles.

Pour une analyse plus globale des résultats présentés dans cet article, nous pouvons reprendre les termes utilisés par les organisateurs de CLEANEVAL. Le détournement est une tâche peu gratifiante et il est sans doute aussi peu gratifiant d'évaluer un système de traitement automatique conjointement avec les modules de nettoyage dont il dépend. Cela conduit à présenter des résultats détériorés, parfois de manière très significative, ce qui pourrait inciter à ne présenter que des résultats obtenus sur des corpus « idéaux ». Il nous semble au contraire tout à fait justifié d'évaluer l'intégralité du processus de traitement en plus des différentes étapes qui le constituent. Évaluer comment les *conditions de laboratoire*, des textes parfaitement détournés, influent sur les résultats devrait alors être un souci plus constant des travaux de TAL. S'il n'existe pas de méthode permettant de détourner efficacement les pages Web autrement que *ad hoc* à un site, alors il convient d'admettre que le détournement n'appartient pas au domaine de l'ingénierie mais reste un champ de recherche que le TAL devrait investiguer.

Références

- BARONI M., CHANTREE F., KILGARRIFF A. & SHAROFF S. (2008). Cleaneval : a competition for cleaning web pages. In *Actes du 4ème Workshop Web as Corpus, LREC 2008* : European Language Resources Association.
- BRIXTEL R., LEJEUNE G., DOUCET A. & LUCAS N. (2013). Any Language Early Detection of Epidemic Diseases from Web News Streams. In *International Conference on Healthcare Informatics (ICHI)*, p. 159–168.

- CHAKRABARTI D., KUMAR R. & PUNERA K. (2008). A graph-theoretic approach to webpage segmentation. In *Proceedings of the 17th international conference on World Wide Web, WWW '08*, p. 377–386, New York, NY, USA : ACM.
- DAS S. N., VIJAYARAGHAVAN P. K. & MATHEW M. (2012). Article : Eliminating noisy information in web pages using featured dom tree. *International Journal of Applied Information Systems*, **2**(2), 27–34. Published by Foundation of Computer Science, New York, USA.
- ENDRÉDY I. & NOVÁK A. (2013). More effective boilerplate removal – the goldminer algorithm. *Polibits*, **48**, 79–83.
- EVERT S. (2008). A lightweight and efficient tool for cleaning web pages. In *Actes du 4ème Workshop Web as Corpus, LREC 2008*.
- FERRARESI A., ZANCHETTA E., BARONI M. & BERNARDINI S. (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Actes du 4ème Workshop Web as Corpus, LREC 2008*.
- PASTERNAK J. & ROTH D. (2009). Extracting article text from the web with maximum subsequence segmentation. In *WWW*, p. 971–980.
- POMIKÁLEK J. (2011). Removing boilerplate and duplicate content from web corpora. *Disertacni práce, Masarykova univerzita, Fakulta informatiky*.
- RATCLIFF J. W. & METZENER D. E. (1988). Pattern matching : The gestalt approach. *Dr. Dobbs Journal*, **13**(7), 46, 47, 59–51, 68–72.
- SPOUSTA M., MAREK M. & PECINA P. (2008). Victor : the Web-Page Cleaning Tool. In *Actes du 4ème Workshop Web as Corpus, LREC 2008*.
- VIEIRA K., DA SILVA A. S., PINTO N., DE MOURA E. S., CAVALCANTI J. A. M. B. & FREIRE J. (2006). A fast and robust method for web page template detection and removal. In *ACM international conference on Information and knowledge management, CIKM '06*, p. 258–267, New York, NY, USA : ACM.

CANÉPHORE : un corpus français pour la fouille d’opinion ciblée

Joseph Lark^{1,2} Emmanuel Morin¹ Sebastián Peña Saldarriaga²

(1) LINA, UMR CNRS 4261, 2 chemin de la Houssinière, Nantes, France

(2) Dictanova, 2 chemin de la Houssinière, Nantes, France

joseph@dictanova.com, emmanuel.morin@univ-nantes.fr, sebastian@dictanova.com

Résumé. La fouille d’opinion ciblée (*aspect-based sentiment analysis*) fait l’objet ces dernières années d’un intérêt particulier, visible dans les sujets des récentes campagnes d’évaluation comme *SemEval* 2014 et 2015 ou bien DEFT 2015. Cependant les corpus annotés et publiquement disponibles permettant l’évaluation de cette tâche sont rares. Dans ce travail nous présentons en premier lieu un corpus français librement accessible de 10 000 tweets manuellement annotés. Nous accompagnons ce corpus de résultats de référence pour l’extraction de marqueurs d’opinion non supervisée. Nous présentons ensuite une méthode améliorant les résultats de cette extraction, en suivant une approche semi-supervisée.

Abstract.

CANÉPHORE : a French corpus for aspect-based sentiment analysis evaluation

Aspect-based sentiment analysis knows a renewed interest these last years, according to recent opinion mining evaluation series (*SemEval* 2014 and 2015, DEFT 2015). However, publicly available evaluation resources are scarce. This work firstly introduces a publicly available annotated French Twitter corpus for sentiment analysis evaluation on aspect, subject and opinion word levels (10 000 documents). We present baseline results on this corpus for the task of opinion word extraction and then show that these results can be improved with simple semi-supervised methods.

Mots-clés : Fouille d’opinion, web social, corpus annoté, extraction d’information semi-supervisée.

Keywords: Opinion mining, social web, annotated corpus, semi-supervised information extraction.

1 Introduction

Ce travail s’inscrit dans le cadre de la fouille d’opinion sur le web social en français, et plus particulièrement la fouille d’opinion ciblée (*aspect-based sentiment analysis*). Nous entendons par fouille d’opinion ciblée l’analyse des arguments avancés par les internautes au niveau des sujets ou de leurs aspects, par opposition à l’analyse d’un document dans son ensemble (Pang & Lee, 2004; Peña Saldarriaga *et al.*, 2013). Ce domaine de recherche fait dernièrement l’objet d’un fort intérêt d’après les sujets de récents ateliers tels que *SemEval* 2014 (*Semantic Evaluation*) (Pontiki *et al.*, 2014) ainsi que DEFT 2015 (Défi Fouille de Texte). D’un point de vue applicatif, ce type de fouille est une promesse très intéressante.

En dehors des corpus annotés fournis pour les campagnes d’évaluation citées précédemment, peu de ressources sont disponibles pour l’évaluation de ces systèmes, et donc pour leur bon développement. Nous proposons ici un corpus annoté d’interactions sur Twitter, CANÉPHORE¹ (Corpus Annoté pour l’Évaluation de Fouille d’Opinion pour la Recherche), disponible publiquement. L’annotation réalisée manuellement permet d’évaluer un système de fouille d’opinion au niveau des sujets abordés, de leurs aspects (partie ou caractéristique jugée du sujet), ou des marqueurs d’opinion, c’est-à-dire des mots portant une orientation sémantique (ou polarité) et qualifiant une cible. Nous associons à ce corpus les résultats d’une méthode de référence non supervisée pour l’extraction de ces marqueurs. Dans un deuxième temps nous expérimentons l’identification faiblement supervisée des patrons morpho-syntaxiques caractéristiques des marqueurs d’opinion, à partir d’un lexique affectif restreint. Les résultats obtenus confortent cette approche du problème de l’adaptation au domaine.

Après une brève mise en contexte de nos travaux (*cf.* section 2), nous décrivons dans cet article le processus d’annotation ainsi que le contenu du corpus d’évaluation proposé (*cf.* section 3). Enfin nous montrons les résultats de nos expériences en extraction de marqueurs d’opinion (*cf.* section 4), dont nous commentons les points forts et les limites (*cf.* section 5).

1. Le corpus est disponible à <https://github.com/ressources-tal/canephore>. Les tweets doivent être acquis à partir de leur identifiant, qui est associé à un fichier unique au format d’annotation de Brat (<http://brat.nlplab.org/standoff.html>).

2 Travaux connexes

Le corpus que nous proposons permet d'évaluer dans un cadre formel un système de fouille d'opinion ciblée. Après avoir présenté en quoi cela diffère d'un corpus pour la fouille d'opinion au niveau document, nous définissons ce qu'est ce cadre formel, puis nous dressons un bref état de l'art sur les problèmes soulevés dans le domaine de la fouille d'opinion ciblée.

2.1 Corpus pour la fouille d'opinion

En ce qui concerne la classification de documents en polarité positive ou négative, plusieurs méthodes efficaces permettent d'obtenir automatiquement un corpus d'évaluation à partir du web social. Ainsi parmi les travaux français, Harb *et al.* (2008) effectuent des requêtes incluant des marqueurs d'opinion sur le moteur de recherche de Google Blog, et Pak & Paroubek (2010) interrogent l'API de Twitter avec des émoticônes positives ou négatives. Une autre méthode consiste à utiliser les notes des internautes pour déduire la polarité d'une critique, c'est le cas par exemple de Boubel & Bestgen (2011) pour des critiques de films, ou de Vincent & Winterstein (2013) sur des critiques d'hôtels, de films et de romans.

Malheureusement il n'existe à ce jour aucune méthode permettant de construire de la même façon un corpus d'évaluation pour la fouille d'opinion ciblée – une annotation manuelle est donc nécessaire. À notre connaissance les données disponibles en anglais sont celles produites lors de la campagne d'évaluation *SemEval* 2014, proposant des critiques annotées sur des restaurants et ordinateurs portables. En français, nous n'avons connaissance d'aucun corpus libre de ce type.

2.2 Fouille d'opinion ciblée

À notre sens, le travail fondateur de la fouille d'opinion ciblée est celui de Hu & Liu (2004), dans lequel les auteurs recherchent les adjectifs cooccurents des sujets abordés parmi des critiques sur le web social afin d'extraire les opinions portées sur les caractéristiques d'appareils numériques. Le choix des adjectifs s'explique par la corrélation forte entre la subjectivité du discours et la présence d'adjectifs, comme ont pu le montrer Hatzivassiloglou & Wiebe (2000). L'orientation sémantique (ou polarité) des adjectifs retrouvés est soit directement attribuée si l'adjectif est présent dans un lexique affectif, soit inférée en recherchant un synonyme ou un antonyme dans la ressource WordNet. Liu (2010) propose par la suite une définition formelle d'une opinion ciblée, modélisée par un quintuplet (e, a, s, h, t) représentant respectivement la cible de l'opinion (*entity*), l'aspect visé (*aspect*), le marqueur d'opinion (*sentiment*), l'émetteur de l'opinion (*holder*) et le moment d'émission de cette opinion (*time*). Le problème de la fouille d'opinion ciblée est alors décomposé en sous-tâches visant à détecter chaque élément de ce tuple. Nous ne considérons dans cet article qu'un triplet (e, a, s) soit (*cible*, *aspect*, *marqueur*). Les sous-tâches dans notre cas sont donc l'identification des sujets, des aspects et des marqueurs d'opinion ainsi que la détection de la polarité de ces marqueurs.

Pour ce qui est de l'identification des sujets et aspects cibles, Brun *et al.* (2014) ont recours à la création d'un arbre syntaxique, indiquant le lien entre un marqueur et une cible. Cependant Kiritchenko *et al.* (2014) montrent que des résultats satisfaisants peuvent être atteints par apprentissage sur une fenêtre de n mots incluant un marqueur d'opinion et sa cible. Enfin, Qiu *et al.* (2009) proposent une méthode dite de double propagation par laquelle les sujets sont détectés à l'aide des marqueurs déjà identifiés et *vice versa*, par le moyen d'un modèle de champ aléatoire conditionnel (CRF).

L'identification des marqueurs d'opinion et l'inférence de leur polarité peuvent être réalisées conjointement, c'est notamment le cas si un lexique affectif est utilisé. Plusieurs travaux visant à construire ce type de ressource ont été entrepris, du fait de la difficulté de retrouver l'orientation sémantique d'un mot sans ce point de repère. Parmi ceux-ci nous pouvons citer le lexique de subjectivité MPQA (Wiebe *et al.*, 2005) ainsi que le lexique SentiWordNet (Baccianella *et al.*, 2010). Plusieurs projets de traduction de SentiWordNet ont vu le jour, cependant aucun équivalent français n'est encore disponible. En revanche, Pak & Paroubek (2010) décrivent la construction d'un tel lexique à partir de Twitter.

Il est important de rappeler que ces lexiques affectifs indépendants du domaine peuvent toutefois induire des ambiguïtés, puisque de nombreux mots présentent une orientation sémantique voire une valence subjective différente selon leur contexte d'apparition. Garcia-Fernandez & Ferret (2012) réalisent une étude sur les différentes stratégies applicables pour la construction de ressources spécifiques au domaine, permettant de réduire ces ambiguïtés. Parmi ces stratégies nous notons l'adaptation d'un lexique affectif générique à un domaine spécifique (Jijkoun *et al.*, 2010) et l'adaptation de ressources pour la fouille d'opinion d'un domaine spécifique vers un autre (Marchand, 2013).

3 Présentation du corpus CANÉPHORE

Notre objectif est d'évaluer progressivement un système de fouille d'opinion ciblée, depuis la détection des sujets abordés jusqu'à la qualification en polarité binaire des marqueurs qualifiant les aspects de ces sujets. Nous expliquons ici en quoi le corpus annoté que nous proposons est adapté à cette évaluation.

3.1 Description du corpus

Le corpus provient d'un ensemble de tweets échangés pendant l'événement "Miss France" en 2012. Les doublons, les retweets, les citations ainsi que les tweets considérés trop courts (moins de 3 mots) ont été retirés afin d'éviter le biais que peuvent apporter les répétitions. Ces suppressions réduisent le corpus à 10 000 tweets, soit environ 127 000 mots.

Marqueurs d'opinion CANÉPHORE est analogue à un corpus de critiques comparatives, tels que ceux proposés pour les campagnes d'évaluation récentes, au sens où les internautes s'expriment sur les différents aspects de quelques entités. D'après notre analyse sur plusieurs corpus de ce type, une conséquence de cette configuration est que les internautes emploient fréquemment les mêmes mots porteurs d'opinion pour qualifier la plupart des sujets abordés. Toutefois ce constat peut être nuancé dans le cas de ce corpus puisque bon nombre de ces mots, que nous appelons marqueurs d'opinion, sont uniques (tableau 2) car provenant d'un vocabulaire argotique, et bien souvent incorrectement orthographiés.

Éléments annotés du corpus	#	Marqueurs d'opinion	#
Tweets contenant une annotation	5372	Positifs	687
Sujets (toutes variantes)	708	Négatifs	1290
Marqueurs	1967	Mots (unigrammes)	1075
Aspects	292	Dont uniques	740
Triplets cible-aspect-marqueur	955	Expressions (n-grammes, n > 1)	892
Couples cible-marqueur	5220	Dont uniques	800

Tableau 1: Informations sur les éléments du corpus annotés

Tableau 2: Informations sur les marqueurs d'opinion

Sujets Les principaux sujets du corpus sont bien évidemment les "Miss", dont les caractéristiques sont jugées par les internautes. Cependant l'identification des cibles précises des opinions émises reste un défi car dans bien des cas plusieurs sujets sont jugés dans un même tweet. De plus, les erreurs orthographiques et l'emploi fréquent d'anaphores nominales (par exemple "Alsace" ou "Miss Rousse" pour "Miss Alsace") complexifient la consolidation des opinions qualifiant chaque sujet. Nous avons ainsi retrouvé 708 variantes ou erreurs orthographiques pour ces sujets (tableau 1), alors qu'il n'est question que d'une centaine de cibles uniques au cours de ces discussions.

3.2 Annotation du corpus

Nous décrivons la méthode utilisée pour annoter les éléments du corpus indiquant une opinion, ainsi que les difficultés que nous avons rencontrées lors de cette étape.

Protocole L'annotation du corpus a été réalisée grâce à l'outil libre Brat², avec lequel il est possible d'étiqueter pour chaque tweet les entités (sujets, aspects ou marqueurs) et les relations entre ces entités. Nous modélisons donc par ces annotations les tuples (*sujet*, *aspect*, *marqueur*) définis précédemment. Pour chaque tweet exprimant une opinion directe sur un sujet explicitement mentionné, nous relevons les segments de texte les plus courts permettant à un humain sans connaissance du sujet de disposer d'une information non ambiguë. Si l'opinion est indirecte, ou fait référence à un sujet par une anaphore pronominale (sans mention du sujet dans le tweet), aucune annotation n'est retenue. Les informations ajoutées à chaque tweet sont donc : le sujet jugé, l'aspect du sujet le cas échéant, le marqueur d'opinion, sa polarité et éventuellement la marque de négation inversant cette polarité. À ces informations s'ajoutent les relations possibles entre les entités : un lien "est un aspect de" entre un aspect et un sujet, un lien "est une opinion positive (ou négative) sur" entre un marqueur d'opinion et un sujet ou un aspect et enfin un lien "inverse" entre une marque de négation et un marqueur d'opinion. L'exemple d'annotation en figure 1 présente ces différents éléments dans un cas non ambigu.

2. <http://brat.nlplab.org/>

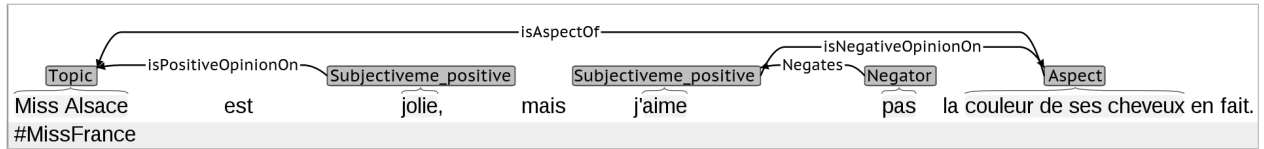
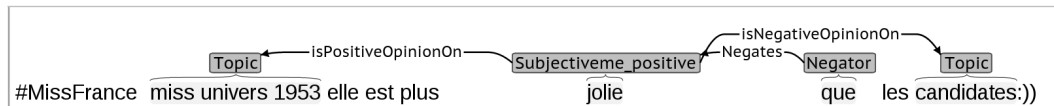


FIGURE 1: Capture d'écran d'annotation à l'aide de Brat, dans un cas non ambigu

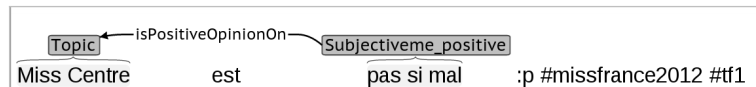
Difficultés Les annotations effectuées décrivent des opinions directes, qu'elles soient explicites ou implicites. Notre objectif est que l'information annotée désigne ce qu'un humain lisant le tweet pourrait comprendre sans ambiguïté. Cependant, une opinion peut être exprimée au moyen d'une construction complexe ou de connaissances implicites, ce qui peut impliquer des difficultés d'annotation. La figure 2 répertorie quelques exemples parmi ces difficultés. Le cas de la construction comparative (figure 2.a) est fréquent. Nous avons choisi dans ce cas de considérer le pronom relatif "que" comme l'élément inversant la polarité du marqueur d'opinion, car cela représente un pivot indissociable d'une comparaison entre deux sujets. Un autre annotation pourrait considérer la construction "*plus...que*" (ou "*moins...que*") en entier, cependant l'outil que nous avons utilisé ne permet pas d'annoter des empanns de texte non connexes.



(a) Annotation d'une opinion sous la forme d'une comparaison



(b) Annotation d'opinions implicites, soit imagées (à gauche), soit demandant des connaissances extérieures (à droite)



(c) Annotation d'une opinion positive incluant une négation

FIGURE 2: Captures d'écran d'annotation, dans le cas d'opinion énoncées de façon ambiguë ou complexe

Une autre forme d'opinion complexe est celle reposant sur des connaissances externes, ce que nous appelons opinion implicite. En réalité, toute forme d'opinion repose sur des connaissances extérieures à son expression, et l'utilisation même de ressources linguistiques pour la fouille d'opinion pose la question de la limite de cet apport. En effet, de la même façon que nous disposons dans un lexique de la polarité "positif" pour le mot "*belle*", nous pourrions disposer d'une telle polarité *a priori* pour l'expression "*réincarnation réussie de Julia Channel*" (figure 2.b, à droite). La spécificité des expressions de ce type nous encourage cependant à les considérer comme trop ambiguës – aucune annotation n'est alors effectuée. Ce n'est pas le cas des expressions imagées (figure 2.b, à gauche), dont la polarité peut être inférée dans leur contexte, ou des expressions idiomatiques, dont le sens est connu. Enfin, nous avons fait le choix d'annoter comme un tout certaines expressions composées courantes, comme "*pas si mal*" (figure 2.c), "*pas mal*" ou "*pas trop mal*". Compte tenu de la granularité de l'analyse visée, et du grand nombre de ces cas particuliers, nous avons préféré effectuer plusieurs passes de vérification, en particulier pour valider la cohérence entre les informations, à une annotation multiple. Nous n'avons par conséquent pas d'accord inter-annotateur pour ce corpus.

3.3 Extraction des marqueurs d'opinion du corpus

Le système que nous souhaitons évaluer à l'aide de ce corpus doit en définitive produire un récapitulatif des éléments positifs ou négatifs exprimés par les internautes à propos des sujets discutés. Cette tâche complexe est décomposée en plusieurs sous-tâches, dont la détection des sujets et de leurs aspects, l'extraction des marqueurs d'opinion et l'inférence de la polarité de ces marqueurs. La première des tâches à réaliser pour la désambiguïsation des opinions émises est l'identification des marqueurs d'opinion. La subjectivité et la polarité de ces mots sont bien souvent spécifiques au contexte

d'apparition, c'est pourquoi nous pensons qu'une approche peu supervisée est adaptée pour ce problème. Nous montrons dans ce travail les résultats de l'extraction des marqueurs (mots simples uniquement) sans tenir compte de leur polarité.

Sujet	Précision	Rappel	F1
Miss Alsace	30,75	33,16	31,91
Miss Bretagne	19,07	19,07	19,07
Miss Réunion	40,00	37,36	38,64
Miss Languedoc	56,82	55,56	56,18
Miss Martinique	35,23	35,23	35,23
Miss Guyane	46,60	47,06	46,83
Miss Provence	28,89	36,62	32,30

Tableau 3: Évaluation de la méthode de référence pour l'extraction des marqueurs d'opinion (%)

Sujet	Précision	Rappel	F1
Miss Alsace	29,10	9,87	14,74
Miss Bretagne	55,56	25,77	35,21
Miss Réunion	45,10	12,64	19,74
Miss Languedoc	12,90	2,96	4,82
Miss Martinique	59,00	22,73	32,79
Miss Guyane	25,71	8,82	13,14
Miss Provence	40,00	8,45	13,95

Tableau 4: Évaluation de l'extraction par projection du lexique affectif restreint (%)

La méthode de référence dont nous montrons les performances en tableau 1 est celle proposée par Hu & Liu (2004), qui consiste à extraire les adjectifs présents dans la même phrase qu'un sujet identifié. Nous avons choisi sept sujets parmi les plus discutés dans le corpus afin d'évaluer cette extraction, ce qui correspond à notre cadre d'application. L'étiquetage grammatical a été réalisé en suivant le travail de Dejean *et al.* (2010), car nous utilisons UIMA³ pour cette analyse. Nous observons que cette méthode fournit des résultats très variables, et dans le meilleur des cas assez moyens. Plusieurs facteurs peuvent expliquer cela. Premièrement la précision ne peut être maximale dans la mesure où nous ne réalisons pas de détection du lien entre un marqueur et sa cible. Deuxièmement, le rappel est limité puisque seuls les adjectifs sont extraits, alors que les marqueurs d'opinion peuvent être des verbes ou des substantifs. Enfin des erreurs d'étiquetage grammatical peuvent expliquer ces performances, car nous savons notre prétraitement imparfait sur ce corpus de tweets.

4 Expériences

Au vu des performances moyennes de la méthode de référence sur notre corpus, nous avons souhaité mener quelques expériences sur l'extraction de marqueurs d'opinion peu supervisée, dont nous montrons ici les résultats.

4.1 Méthodes proposées

Les méthodes proposées réalisent l'extraction en deux étapes : nous identifions dans un premier temps des patrons morpho-syntaxiques – c'est-à-dire les séquences d'étiquettes grammaticales ou de lemmes de n mots consécutifs – caractérisant la présence d'un marqueur d'opinion à partir d'un lexique de mots d'opinion restreint (que nous décrivons par la suite). Ces patrons sont ensuite projetés sur le corpus pour découvrir de nouveaux marqueurs.

Patrons fixes Notre première méthode consiste à identifier les patrons syntaxiques fixes les plus fréquents entourant un marqueur connu. Nous retenons les patrons syntaxiques dans une fenêtre centrée sur le marqueur de 7 mots au maximum et dont la fréquence figure parmi le top 3. Cela représente en tout quatre patrons : *nom-verbe-candidat adjectif* (24 occurrences), *pronom-verbe-candidat verbe* (24 occurrences), *verbe-adverbe-candidat adjectif* (22 occurrences) et *déterminant-adverbe-candidat adjectif* (20 occurrences).

Modèle SVM Notre seconde méthode consiste à réaliser un apprentissage par machine à vecteurs de support⁴ (SVM) sur les patrons morpho-syntaxiques, en reprenant une méthode que nous avions expérimentée précédemment (Lark *et al.*, 2014). Les traits de classification pour ces patrons sont les étiquettes grammaticales et les lemmes des mots entourant un mot candidat dans une phrase, en tenant compte de leur position. Les exemples positifs pour ce modèle sont les patrons entourant les marqueurs du lexique. Le modèle est ensuite utilisé pour classer tous les adjectifs et verbes du corpus, considérés comme exemples négatifs lors de la phase d'apprentissage s'ils sont inconnus du lexique.

4.2 Lexique de marqueurs stables

L'extraction est réalisée de manière semi-supervisée au sens où aucun exemple du corpus n'est fourni, cependant nous l'amorçons à partir d'un lexique affectif stable, et indépendant du domaine. Nous entendons par stable le fait que les mots

3. Unstructured Information Management, <https://uima.apache.org/>

4. Nous utilisons la librairie LIBLINEAR (<http://liblinear.bwaldvogel.de/>)

indexés dans cette ressource ont été choisis pour leur faible variation de sens et de polarité en fonction de la thématique abordée ou de leur contexte d'apparition. Afin d'obtenir un tel lexique nous avons retenu les éléments les plus fréquents dans un large corpus du web social depuis un lexique affectif indépendant du domaine dont nous disposons⁵. Le lexique final, vérifié manuellement, contient 189 marqueurs d'opinion, dont 70 positifs et 119 négatifs.

4.3 Résultats

Nous reprenons le cadre d'expérimentation de la méthode de référence pour évaluer l'extraction des marqueurs par les méthodes proposées, dont les résultats sont indiqués dans le tableau 5. Afin de mettre en perspective ces résultats, nous montrons les performances d'une simple projection de notre lexique restreint sur le corpus, en tableau 4. Enfin nous évaluons l'ensemble des marqueurs d'opinion obtenus par l'une et l'autre des deux méthodes proposées (colonne de droite dans le tableau 5). Les marqueurs extraits dans les deux cas n'étant pas les mêmes, nous montrons que le rappel peut être amélioré en utilisant cette combinaison.

Sujet	Patrons fixes			SVM			Union		
	Précision	Rappel	F1	Précision	Rappel	F1	Précision	Rappel	F1
Miss Alsace	96,47	20,76	34,17	83,08	13,67	23,48	88,89	26,33	40,62
Miss Bretagne	80,95	8,76	15,81	80,36	23,20	36,00	80,56	29,90	43,61
Miss Réunion	98,08	28,02	43,59	96,77	16,48	28,17	96,83	33,52	49,80
Miss Languedoc	94,59	51,85	66,99	85,19	17,04	28,40	90,12	54,07	67,59
Miss Martinique	92,31	13,64	23,76	95,45	23,86	38,18	93,55	32,95	48,74
Miss Guyane	95,00	37,25	53,52	77,27	16,67	27,42	89,80	43,14	58,28
Miss Provence	88,89	11,27	20,00	83,33	7,04	12,99	84,62	15,49	26,19

Tableau 5: Évaluation de l'extraction par occurrence (%)

4.4 Discussion

Les résultats que nous obtenons par ces méthodes simples sont globalement satisfaisants en matière de précision. Cela corrobore notre approche par amorçage dans la mesure où des marqueurs inconnus du lexique sont retrouvés, tandis que les éléments faisant partie de notre ressource mais non pertinents pour ce corpus sont filtrés. Toutefois, c'est sur la notion de couverture que les méthodes testées trouvent leurs limites. D'une part le choix des patrons les plus présents peut éliminer des candidats les mots peu fréquents. D'autre part, d'après notre analyse des patrons retenus, l'agrégation des marqueurs candidats dont le patron morpho-syntaxique est similaire à ceux déjà reconnus ne permet pas d'acquérir les marqueurs dont le patron est très différent, quand bien même leur fréquence serait élevée.

5 Conclusions et perspectives

Cet article présente la diffusion d'un corpus français annoté pour l'évaluation de méthodes de fouille d'opinion ciblée. Une telle ressource n'existait pas à notre connaissance, et nous espérons que d'autres seront créées selon le même type d'annotation que nous décrivons ici. La diffusion de ce corpus offre la possibilité à chacun de l'exploiter, mais également de l'enrichir. Certaines informations ne sont pour le moment pas présentes car elles ne correspondent pas à notre cadre d'analyse : l'opinion générale de chaque tweet n'est pour le moment pas annotée, et les marqueurs d'opinion sont simplement catégorisés selon une polarité binaire et non en classes d'opinion fines. Enfin, pour que ce corpus puisse être exploité par tout un chacun, il serait pertinent d'évaluer son annotation dans un cadre plus large que celui de ce travail.

Nous évaluons sur ce corpus des méthodes faiblement dépendantes du domaine et de la langue. Les résultats indiquent qu'elles permettent d'extraire avec précision les éléments clés qualifiant l'opinion émise sur les sujets abordés. Cette extraction non supervisée montre cependant certaines limites quant à la capacité de détection des marqueurs peu fréquents ou peu similaires aux marqueurs déjà reconnus. Afin de détecter ces éléments nous prévoyons d'étudier l'extraction par double propagation au moyen d'une classification par séquence comme ont pu le tester Qiu *et al.* (2009).

5. Ressource établie à partir du lexique Apopsis, disponible sur demande : <http://taln.lina.univ-nantes.fr/apopsis/>

Références

- BACCIANELLA S., ESULI A. & SEBASTIANI F. (2010). Sentiwordnet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC 2010*, Valletta, Malta.
- BOUBEL N. & BESTGEN Y. (2011). Une procédure pour identifier les modifieurs de la valence affective d'un mot dans des textes. In *Actes de TALN 2011*, p. 137–142, Montpellier, France.
- BRUN C., POPA D. N. & ROUX C. (2014). XRCE : Hybrid classification for aspect-based sentiment analysis. In *Proceedings of SemEval 2014*, p. 838–842, Dublin, Ireland.
- DEJEAN C., FORTUN M., MASSOT C., POTTIER V., POULARD F. & VERNIER M. (2010). Un étiqueteur de rôles grammaticaux libre pour le français intégré à Apache UIMA. In *Actes de TALN 2010*, Montréal, Canada.
- GARCIA-FERNANDEZ A. & FERRET O. (2012). Etude de différentes stratégies d'adaptation à un nouveau domaine en fouille d'opinion. In *Proceedings of JEP-TALN-RECITAL 2012*, p. 391–398, Grenoble, France.
- HARB A., DRAY G., PLANTIÉ M., PONCELET P., ROCHE M. & TROUSSET F. (2008). Détection d'opinion : Apprenons les bons adjectifs ! In *Actes de INFORSID 2008 - Atelier FODOP*, p. 59–66, Fontainebleau, France.
- HATZIVASSILOGLOU V. & WIEBE J. M. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of COLING 2000*, p. 299–305, Saarbrücken, Germany.
- HU M. & LIU B. (2004). Mining and summarizing customer reviews. In *Proceedings of KDD 2004*, p. 168–177, Seattle, WA, USA.
- JIJKOUN V., DE RIJKE M. & WEERKAMP W. (2010). Generating focused topic-specific sentiment lexicons. In *Proceedings of ACL'10*, p. 585–594, Stroudsburg, PA, USA.
- KIRITCHENKO S., ZHU X., CHERRY C. & MOHAMMAD S. (2014). Nrc-canada-2014 : Detecting aspects and sentiment in customer reviews. In *Proceedings of SemEval 2014*, p. 437–442, Dublin, Ireland.
- LARK J., PEÑA SALDARRIAGA S. & MORIN E. (2014). Consumer concern extraction in social web reviews. In *Proceedings of Digital Intelligence 2014*, Nantes, France.
- LIU B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*.
- MARCHAND M. (2013). Fouille d'opinion : ces mots qui changent de polarité selon le domaine. In *Actes de CORIA 2013*, p. 347–352, Neuchâtel, Switzerland.
- PAK A. & PAROUBEK P. (2010). Construction d'un lexique affectif pour le français à partir de twitter. In *Actes de TALN 2010*, Montréal, Canada.
- PANG B. & LEE L. (2004). A sentimental education : Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL'04*, Barcelona, Spain.
- PEÑA SALDARRIAGA S., VINTACHE D. & DAILLE B. (2013). Démonstrateur Apopsis pour l'analyse des tweets. In *Actes de TALN 2013*, p. 807–808, Les Sables d'Olonne, France.
- PONTIKI M., GALANIS D., PAVLOPOULOS J., PAPAGEORGIOU H., ANDROUTSOPOULOS I. & MANANDHAR S. (2014). Semeval-2014 task 4 : Aspect based sentiment analysis. In *Proceedings of SemEval 2014*, Dublin, Ireland.
- QIU G., LIU B., BU J. & CHEN C. (2009). Expanding domain sentiment lexicon through double propagation. In *Proceedings of IJCAI 2009*, p. 1199–1204, Pasadena, CA, USA.
- VINCENT M. & WINTERSTEIN G. (2013). Construction et exploitation d'un corpus français pour l'analyse de sentiment. In *Actes de TALN 2013*, p. 764–771, Les Sables d'Olonne, France.
- WIEBE J., WILSON T. & CARDIE C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, Volume 39, p. 165–210.

Extraction de Contextes Riches en Connaissances en corpus spécialisés

Firas Hmida Emmanuel Morin Béatrice Daille

Université de Nantes, LINA UMR 6241, 2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 3
{firas.hmida, emmanuel.morin, beatrice.daille}@univ-nantes.fr

Résumé. Les banques terminologiques et les dictionnaires sont des ressources précieuses qui facilitent l'accès aux connaissances des domaines spécialisés. Ces ressources sont souvent assez pauvres et ne proposent pas toujours pour un terme à illustrer des exemples permettant d'appréhender le sens et l'usage de ce terme. Dans ce contexte, nous proposons de mettre en œuvre la notion de Contextes Riches en Connaissances (CRC) pour extraire directement de corpus spécialisés des exemples de contextes illustrant son usage. Nous définissons un cadre unifié pour exploiter tout à la fois des patrons de connaissances et des collocations avec une qualité acceptable pour une révision humaine.

Abstract.

Knowledge-Rich Contexts Extraction in Specialized Corpora

The term banks and dictionaries are valuable resources that improve access to knowledge in specialized domains. These resources are often relatively poor and do not always provide, for a given term, examples of its typical use. In this context, we implement Knowledge-Rich Contexts (KRCs) to extract examples of contexts providing illustration of terms in specialized domain. We propose a unified framework to apply at the same time knowledge pattern and collocations with acceptable quality for human review.

Mots-clés : corpus spécialisé, CRC, patrons de connaissances, collocations.

Keywords: specialized corpus, KRC, knowledge patterns, collocations.

1 Introduction

Les banques terminologiques et les dictionnaires sont des ressources linguistiques précieuses qui facilitent l'accès aux connaissances des domaines spécialisés. Ces ressources proposent généralement pour un terme à illustrer une définition, des exemples d'utilisation et d'autres termes en relation. Habituellement, le terme à illustrer est considéré comme « terme favori » et les autres termes reflètent des relations paradigmatiques comme la synonymie ou l'hyponymie. La définition associée au terme favori est souvent de nature encyclopédique et les quelques exemples de contextes proposés, lorsqu'ils existent, ne couvrent qu'une partie des usages de ce terme favori (Bowker, 2011). Des travaux récents laissent entendre que les banques terminologiques actuelles n'ont pas connu d'améliorations significatives depuis les années 60 (Bowker, 2011). Elles contiennent trop peu d'informations contextuelles et les connaissances sur l'usage du terme sont assez limitées (p. ex. des exemples de collocations peuvent être indiqués mais cela n'est pas systématique). En outre, les définitions fournies par les dictionnaires comme les banques terminologiques sont généralement insuffisantes pour permettre la compréhension du terme. Les corpus spécialisés représentent un réservoir important d'informations contextuelles pour analyser le fonctionnement des termes. Cependant, tous les contextes dans lesquels les termes apparaissent ne sont pas utiles à leur compréhension. Dans ce contexte, Meyer (2001) a introduit la notion de Contextes Riches en Connaissances (CRC) désignant les contextes qui jouent un rôle prépondérant dans la compréhension des termes et renseignent sur leur fonctionnement linguistique.

Dans ce travail, nous postulons que les corpus monolingues spécialisés contiennent des contextes conceptuels et linguistiques qui peuvent être extraits automatiquement. Nous nous limitons aux corpus étudiés sans solliciter de ressources externes et mettons en œuvre deux méthodes pour identifier des CRC. La première méthode s'appuie sur la présence du terme à illustrer et l'exploitation des patrons lexicaux afin d'extraire des contextes riches en connaissances conceptuelles. Ces CRC permettent l'accès à la dimension conceptuelle du terme. Les patrons lexicaux exploités ont pour but notamment d'accéder en corpus spécialisé à la définition du terme. La seconde méthode exploite des mesures d'association pour identifier des contextes riches en connaissances linguistiques aidant ainsi à comprendre l'usage du terme. Elle repose sur

le repérage en corpus des collocations. Nous focalisons notre travail sur deux corpus spécialisés de discours scientifiques relevant du domaine de la vulcanologie en français et en anglais.

2 État de l'art

De nombreuses recherches ont été menées sur les CRC dans différentes perspectives. Dans cette section, nous présentons succinctement les principaux travaux exploitables dans une perspective d'aide à la compréhension. Nous présentons tout d'abord la notion de CRC, ensuite les méthodes permettant leur extraction.

Meyer (2001) propose la notion de CRC pour désigner les contextes qui illustrent des relations entre les termes d'un domaine spécialisé. Ces relations sont souvent représentées par des unités lexico-syntaxiques appelées « *patrons de connaissances* » (PC). La phrase « *L'Olympus ci-contre est le volcan géant du système solaire* » est définie comme un CRC pour le terme *Olympus*. Dans cet exemple, le patron de connaissances *est le* explicite une relation hiérarchique entre les termes *Olympus* et *volcan*. Schumann (2012) a entrepris d'extraire des CRC à partir du Web dans le but d'enrichir une banque terminologique en langue russe. Les contextes ont tout d'abord été repérés au moyen de PC, puis ordonnés grâce à une méthode supervisée. Ce travail est similaire au nôtre. Néanmoins, dans le cadre de notre problématique, nous étudions l'identification des CRC dans un corpus spécialisé de taille modeste. Marshman (2014) a étudié la nécessité d'utiliser des ressources terminologiques mettant en évidence des CRC extraits par des PC, telles que CREATerminal¹. Ces recherches ont également montré l'utilité des ressources enrichies par des CRC, particulièrement pour des traducteurs étudiants. Une des difficultés majeures dans le domaine des PC tient au fait qu'il n'existe aucune bibliothèque de PC qui manifesterait l'aspect cumulatif de ces travaux. Pour chaque nouvelle étude, il faut refaire une synthèse des études existantes pour établir des listes de PC. D'une part, les travaux concernant la variation dans le fonctionnement des PC sont encore récents. En effet, selon Marshman *et al.* (2008), bien que l'intérêt des PC pour repérer les CRC est indéniable, leur identification est coûteuse et l'on doit chercher à les réutiliser pour d'autres études, dans d'autres types de corpus. D'autre part, il est primordial de chercher à mesurer leur portabilité, c'est-à-dire leur degré de variabilité d'un corpus à l'autre, et donc d'un domaine à l'autre.

La notion de CRC fait écho à d'autres types de contextes tels que les définitions étudiées par Saggion (2004) et les collocations extraites par Kilgarriff *et al.* (2008). Saggion (2004) a eu recours à deux stratégies afin de repérer des définitions à partir de textes disponibles sur le Web. Il a utilisé des PC modélisant des relations de définitions, ainsi que des « termes secondaires » fournissant des connaissances spécifiques au terme en question. Saggion (2004) introduit ces termes secondaires comme les termes qui co-occurrent significativement avec le terme à illustrer dans des définitions. Un terme secondaire peut être un nom, un adjectif ou un verbe. Cette notion fait référence à celle de collocation identifiée dans un corpus de définitions. La maîtrise des collocations est une composante essentielle de la maîtrise de la langue ou d'un discours spécifique. Ceci explique l'importance accordée à cette notion pour l'aide à la compréhension. Au sens restreint, les collocations représentent des associations lexicales transparentes du point de vue de la compréhension mais qu'un locuteur non natif doit tout particulièrement apprendre à maîtriser. C'est le cas des exemples suivants : *prescrire une ordonnance*, *tenir debout*, *nuît blanche*. La notion de collocation reçoit des définitions variables selon le contexte de recherche dans lequel elle est employée. Sinclair *et al.* (1970), par exemple, définissent la collocation par la co-occurrence significative de deux items dans un contexte spécifié. En s'appuyant sur les collocations, Kilgarriff *et al.* (2008) propose GDEX (Good Dictionary Examples), un outil permettant de produire automatiquement des exemples pour des lexicographes. Cet outil a pour but de sélectionner l'exemple lexicographique le plus pertinent à partir d'un corpus de données massives. De ce point de vue, GDEX peut être considéré comme un système de filtres permettant de retenir les « bons » exemples respectant un ensemble de critères. Son fonctionnement consiste à identifier les collocations du terme en question pour associer ensuite des exemples à chaque collocation. Kilgarriff *et al.* (2008) se sont basés sur les travaux d'Atkins & Rundell (2008) pour qualifier un « bon » exemple en s'appuyant sur différents critères tels que i) la lisibilité, c'est-à-dire intelligible aux lecteurs aussi bien lexicalement que structurellement et ii) l'informativité, qui illustre des contextes typique (par exemple une collocation) et qui aide à comprendre le terme à exemplifier. Ces critères ont été mis en œuvre comme des traits positifs tels que la présence de troisième collocatif, et négatifs comme la présence de mots rares dans le contexte. Les exemples de Kilgarriff sont incontestablement des contextes riches illustrant l'usage du terme. Bien que notre objectif soit similaire à celui de Kilgarriff, nous cherchons les contextes qui fournissent des connaissances spécifiques au domaine étudié. Ces contextes permettront également au lecteur de positionner le terme par rapport à la terminologie du domaine. Ces deux types de connaissances linguistiques (collocations) et conceptuelles (PC) sont nécessaires.

1. Interface fournissant des contextes aidant à la traduction terminologique (anglais-français) dans le domaine du cancer du sein.

3 Contribution

Dans cette section nous étudions les PC et les collocations dans une perspective d'aide à la compréhension d'un terme donné. Pour ce faire, nous mettons en œuvre deux méthodes permettant d'exploiter ces notions pour identifier des CRC à partir de corpus monolingues spécialisés.

3.1 Patrons de Connaissances pour CRC

Dans la littérature, trois relations considérées comme universelles ont été majoritairement étudiées. Il s'agit des relations d'hyponymie, de méronymie et de cause. La relation d'hyponymie est connue comme la plus structurante, du fait de son exploitation dans les définitions et de sa propriété de transitivité. Dans ce travail, nous mettons en œuvre l'automatisation du repérage et la portabilité des CRC à l'aide de PC d'hyponymie. Nous exprimons ces PC sous la forme d'un triplet ($terme_1$, PC d'hyponymie, $terme_2$) avec $terme_1$ et $terme_2$ deux termes distincts du corpus étudié. La démarche suivie consiste à utiliser tout d'abord un outil d'extraction terminologique pour identifier les termes du corpus ; ensuite à retenir les contextes phrastiques contenant le patron ($terme$ à illustrer, PC d'hyponymie, $terme$ du domaine). Ces contextes potentiellement riches en connaissances sont considérés comme des CRC candidats. La table 1 illustre des CRC candidats associés aux termes *cendre* et *volcan*. Ces CRC candidats sont repérés grâce aux PC présentés dans la même table.

Terme à illustrer (X)	Terme du domaine (Y)	PC	CRC candidat
Cendre	produit volcanique	X_ÊTRE_LE_PRINCIPAL_Y	Les cendres sont les principaux produits volcaniques émis par les volcans explosifs de la ceinture de feu du Pacifique.
Volcan	Actif	X_ÊTRE_Y_LE_PLUS	Le groupe Klyvcheskoy, dont les volcans sont les <u>plus actifs</u> de l'arc des Kouriles...

TABLE 1 – Exemples de CRC candidats identifiés par des PC pour le français

3.2 Collocations pour CRC

Plusieurs mesures d'association ont été appliquées pour extraire automatiquement des collocations. Si l'Information Mutuelle spécifique (Fano, 1961) permet d'identifier des unités lexicales qui apparaissent plus souvent ensembles que séparément, le Z-score (Berry-Rogghe, 1973) est souvent privilégié pour déterminer les collocatifs candidats d'un terme donné. Dans ce travail, nous associons à une liste de termes donnés leurs meilleurs collocatifs en nous appuyant sur la mesure Z-score puisque nous connaissons *a priori* les termes que nous souhaitons illustrer. Ces collocations serviront, par la suite, à sélectionner des CRC candidats.

Les mesures d'association peuvent également être combinées à des analyses linguistiques telles que l'analyse syntaxique (Fellbaum, 1998). Ces analyses, jouant un rôle de filtre, permettent d'affiner la qualité des collocations obtenues et de les classer selon leurs catégories grammaticales. Evert & Krenn (2005) montrent qu'il faut distinguer les catégories syntaxiques des collocations avant d'appliquer une mesure d'association. Nous retenons alors deux catégories de collocations nominales dans lesquelles la base est un terme à illustrer : ($terme$, nom) et ($terme$, $adjectif$). Ces collocations ont été identifiées dans Josselin-Leray *et al.* (2014) comme pertinentes dans un exercice d'aide à la compréhension.

Après avoir filtré les mots outils dans le corpus, nous avons repéré les collocations constituées de deux mots pleins dans une fenêtre bigramme : un mot avant ou un mot après la base (sans compter les mots vides) en respectant les catégories syntaxiques étudiées. Afin d'extraire les CRC candidats nous avons suivi les deux étapes suivantes :

1. Identifier pour un terme à illustrer ses collocatifs en fonction de sa catégorie syntaxique et les ordonner selon le Z-score ;
2. Parcourir les collocatifs de chaque terme à illustrer et retenir le premier (selon le Z-score) qui procure au moins un contexte phrastique lisible, tel que défini par Kilgarrieff *et al.* (2008). Ici, nous nous sommes limités à un seul collocatif en vue d'obtenir un nombre acceptable de CRC candidats pour une évaluation humaine.

Les CRC candidats de la table 2 sont identifiés par des collocations dans lesquelles la base est le terme à illustrer.

Terme à illustrer	Collocatif	CRC candidat
Gaz	carbonique	<i>Ce gaz carbonique qui, transformé par les plantes, a donné de l'oxygène, indispensable à la vie.</i>
Gas	dissolved	<i>Gas dissolved in the molten rock expanded and literally blew the volcano apart...</i>
Cendre	retombée	<i>Les explosions phréatiques se font plus violentes qu'en 1792, et deux ou trois d'entre elles provoquent des retombées de cendres sur les villes du prêcheur.</i>
Cendre	retombée	<i>Veaucoup d'habitants du prêcheur et de ses environs viennent se réfugier à Saint-Pierre, épargnée par les retombées de cendres.</i>

TABLE 2 – Exemples de CRC candidats identifiés par des collocations

4 Expériences et résultats

Dans cette section nous décrivons les différentes ressources mobilisées pour nos expériences et présentons ensuite les résultats obtenus en appliquant les deux méthodes précédentes.

4.1 Ressources linguistiques

Les expériences ont été réalisées sur deux corpus français et anglais relevant du domaine de la vulcanologie. Il s'agit d'un corpus comparable constitué par Amélie Josselin-Leray de CLLE-ERSS. Il est composé de documents scientifiques contenant environ 400 000 mots par langue, obtenus grâce à une recherche thématique à partir de journaux et magazines tels que *Le Monde*, *Sciences et avenir*, *Sciences et Vie*... L'ensemble des documents ont été nettoyés et normalisés à travers les traitements suivants réalisés par la plateforme TermSuite² : segmentation en occurrences de formes, étiquetage morphosyntaxique, lemmatisation et extraction terminologique. En ce qui concerne l'extraction terminologique, nous nous sommes limités aux termes simples et complexes qui apparaissent au moins 5 fois dans chaque corpus. Ces termes sont nécessaires à l'exploitation des PC. À ce niveau, nous nous appuyons sur une liste de PC relatifs à la relation d'hyponymie : 33 en français (Rebeyrolle & Tanguy, 2000) (cf. table 3 pour un extrait) et 34 en anglais (Séguéla, 2001; Marshman *et al.*, 2012). Ces PC permettent d'extraire des CRC comportant éventuellement des connaissances définitoires qui illustrent les termes en question.

Patrons de connaissances (FR)
X_ÊTRE_UN_Y
X_ÊTRE_UNE_SORTE_DE_Y
X_ÊTRE_LE_Y_LE_PLUS
X_ÊTRE_AUTRES_Y
Y_ET_ADVERBE_DE_SPECIFICATION_X

TABLE 3 – Exemples de PC exploités pour le français (X est un terme et Y son hyperonyme)

Enfin, les termes que nous cherchons à illustrer avec nos deux approches et qui sont caractéristiques du domaine de la vulcanologie sont présentés dans la table 4. Ces termes ont été sélectionnés par des linguistes pour des expériences portant sur l'aide à la traduction.

4.2 Résultats des patrons de connaissances

La table 5 présente les résultats obtenus après projection des PC d'hyponymie sur les corpus de vulcanologie en français et en anglais afin d'illustrer les termes de la table 4 rappelés en colonne # *Termes à illustrer*. Nous désignons par # *Termes*

2. <https://logiciels.lina.univ-nantes.fr/redmine/projects/termsuite>

Corpus	Termes à illustrer
Français	<i>basalte, cendre, cratère, cône, débris, dégazage, dôme, fontaine, gaz, jaillir, lave, magma, phase, roche, scorie, téphra, vacuole, volcan, vésicule, éruption</i>
Anglais	<i>basalt, blobs, cinder, cone, eruption, fountaining, gas, layers, scoria, softball, spongelike, vesicles</i>

TABLE 4 – Liste des termes à illustrer

extraits le nombre de termes intégrant des PC avec au moins un hyperonyme, et par # *CRC candidats* le nombre de CRC associés aux termes extraits. # *CRCC* indique le nombre de CRC Conceptuels. Ce sont des CRC candidats contenant le terme à illustrer ainsi que d'autres termes du domaine étudié. Des relations conceptuelles doivent être explicitées à travers un lien sémantique entre le terme visé et les autres termes présents dans le CRC candidat.

Le CRC candidat « *Les **cendres** sont les principaux **produits volcaniques** émis par les volcans explosifs de la ceinture de feu du Pacifique* » (cf. table 1) explicite une relation d'hyperonymie définissant le terme *cendre*. Il s'agit alors d'un CRCC. En ce qui concerne le deuxième cas « *Le groupe Klyvcheskoy, dont les **volcans** sont les plus **actifs** de l'arc des Kouriles* », la relation d'hyperonymie est invalide. En effet, le terme *actif* n'est pas un hyperonyme de *volcan*, d'où nous considérons ce CRC candidat comme non intéressant. Même s'ils peuvent contenir d'autres relations sémantiques intéressantes, les CRC candidats ne sont retenus que lorsque la relation d'hyperonymie est valide.

La colonne # *Termes extraits* montre que les PC sont présents dans les deux corpus spécialisés même si uniquement 33 à 40 % des termes à illustrer sont retrouvés. Cependant, la qualité des CRC obtenus après validation est relativement bonne. Ces résultats sont finalement assez conformes à l'état de l'art (Morin, 1999), à savoir un faible rappel des patrons de connaissances au bénéfice de la précision.

Corpus	# Termes à illustrer	# Termes extraits	# CRC candidats	# CRCC (P.)
Français	20	8	21	17 (80,95 %)
Anglais	12	4	14	10 (71,42 %)

TABLE 5 – Résultats de la projection des PC après une validation manuelle

4.3 Résultats des collocations

En ce qui concerne l'évaluation de la méthode basée sur les collocations, chaque contexte identifié a été évalué par un annotateur natif. Nous désignons par un contexte intéressant, un contexte révélant des connaissances conceptuelles ou linguistiques et aidant à comprendre le terme à illustrer. Nous distinguons, en plus des CRCC, les contextes riches en connaissances linguistiques (CRCL). Un CRCL représente un contexte contenant seulement le terme à illustrer et son collocatif, à l'exception d'autres termes du domaine auxquels le terme à illustrer est conceptuellement lié. Un CRCL doit être grammaticalement bien formé.

Les tables 6 et 7 présentent les résultats obtenus pour les collocations respectivement de type (*terme, adjectif*) et (*terme, nom*). Dans ces tables, # *Termes extraits* est le nombre de termes ayant une collocation fournissant des contextes, *CRCC* et *CRCL* représentent quant à eux les pourcentages des connaissances respectivement conceptuelles et linguistiques illustrées par les contextes repérés grâce à des collocations. *Non CRC* est le pourcentage de contextes qui ne sont pas intéressants par rapport à la compréhension des termes visés. Nous focalisons notre analyse des résultats sur le type des connaissances illustrées ainsi que la cohérence entre les résultats des corpus étudiés. Nous pouvons néanmoins constater que la qualité des CRC obtenue par cette approche est en dessous de la seule utilisation des patrons de connaissances.

Dans le cas du corpus français de vulcanologie, les contextes identifiés par les collocations de type (*terme, adjectif*) contiennent plus souvent des connaissances conceptuelles que linguistiques. Par exemple, la phrase *Ce **gaz carbonique** qui, transformé par les plantes, a donné de l'oxygène, indispensable à la vie* (cf. table 2) contient une relation sémantique qui peut être traitée comme une relation de cause. Dans d'autres cas, les collocatifs de cette catégorie peuvent révéler des connaissances conceptuelles quand il s'agit de participe passé ou présent. En effet, ce type de collocatif, mettant en jeu un verbe conjugué, traduit éventuellement un lien sémantique entre le terme en question et d'autres termes du domaine. Nous parlons alors d'un contexte conceptuel. Dans le corpus anglais de vulcanologie, le contexte ***Gas dissolved** in the molten rock expanded and literally blew the volcano apart* (cf. table 2), considéré comme CRCC, le terme *gas* est illustré par son collocatif *dissolved*. Ces résultats semblent être cohérents avec ceux des corpus anglais de vulcanologie dont les contextes ont été évalués par des linguistes.

Les collocations de type (*terme, nom*) favorisent l'illustration des connaissances conceptuelles et linguistiques de façon mitigée dans les deux corpus étudiés : dans le corpus vulcanologie français 31,33 % des contextes sont conceptuels et 27,71 % sont linguistiques. En effet, ces collocations peuvent informer sur l'usage du terme, notamment en présence de prépositions comme dans le cas de *retombée de cendre* (base : *cendre* et collocatif : *retombée*). Les résultats sont mis en évidence dans les deux corpus vulcanologies.

Corpus	# Termes à illustrer	# Termes extraits	# CRC candidats	CRCC	CRCL	Non CRC
Français	20	18	74	45,95 %	12,16 %	41,81 %
Anglais	12	10	41	41,46 %	14,63 %	43,90 %

TABLE 6 – Évaluation manuelle des contextes extraits par la collocation de type (*terme adjectif*)

Corpus	# Termes à illustrer	# Termes extraits	# CRC candidats	CRCC	CRCL	Non CRC
Français	20	18	83	31,33 %	27,71 %	40,96 %
Anglais	12	10	43	41,86 %	30,23 %	27,91 %

TABLE 7 – Évaluation manuelle des contextes extraits par la collocation de type (*terme, nom*)

4.4 Synthèse

Si nous combinons maintenant les résultats obtenus par les deux précédentes approches, nous pouvons constater à la lecture de la table 8 que nous sommes en mesure de proposer des CRC pour l'ensemble des termes de la liste de référence. La qualité de ces CRC est autour de 70 %, ce qui reste acceptable. Il n'y a que deux termes anglais *spongelike* et *softball* pour lesquels nous ne proposons pas de CRC pertinents.

Il serait en outre intéressant de pouvoir ordonner ces derniers pour ne conserver que les plus intéressants. Dans un premier temps, il semble pertinent de proposer en premier lieu les CRC issus des patrons de connaissances puis ceux issus des collocations.

Corpus	# Termes à illustrer	# Termes extraits	# CRC candidats (sans doublons)	# CRC valides (P.)
Français	20	20	143	100 (69,93 %)
Anglais	12	10	97	67 (69,07 %)

TABLE 8 – Tableau récapitulatif de la combinaison des deux méthodes

5 Conclusion et perspectives

Dans ce travail, nous avons proposé de mettre en œuvre la notion de Contextes Riches en Connaissances pour extraire directement de corpus des exemples illustrant le fonctionnement des termes. Ces CRC, qui sont extraits de corpus en s'appuyant sur des patrons de connaissances et des collocations, permettent d'accéder tout à la fois aux connaissances linguistiques et conceptuelles. L'originalité de notre approche est de considérer l'ensemble des CRC disponibles à la différences des travaux existants qui se restreignent soit à des patrons de connaissances pour extraire des définitions (Marshman, 2014) soit à des collocations pour extraire des exemples (Kilgarriff *et al.*, 2008). La complémentarité des deux approches mises en œuvre permet d'obtenir des CRC variés avec une qualité acceptable pour une révision humaine. Néanmoins, il serait intéressant de pouvoir réduire le nombre proposé de CRC en cherchant à proposer systématiquement pour un terme à illustrer un exemple de CRC linguistique et un autre de CRC conceptuel. Pour ce faire, il sera nécessaire de pouvoir associer à chacun de ces CRC un score de confiance qui pourrait être fonction du patron de connaissances déclenché dans un cas et de l'ordonnement global des collocations dans l'autre cas.

Remerciements

Ce travail qui s'inscrit dans le cadre du projet CRISTAL www.projet-cristal.org a bénéficié d'une aide de l'Agence National de la Recherche portant la référence ANR-12-CORD-0020.

Références

- ATKINS B. S. & RUNDELL M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.
- BERRY-ROGGHE G. (1973). The computation of collocations and their relevance in lexical studies. *The Computer and Literary Studies*, p. 103–112.
- BOWKER L. (2011). Off the record and on the fly : Examining the impact of corpora on terminographic practice in the context of translation. *Corpus-based Translation Studies : Research and Applications*. London/New York : Continuum, p. 211–236.
- EVERT S. & KRENN B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language*, **19**(4), 450–466.
- FANO R. M. (1961). *Transmission of Information : A Statistical Theory of Communication*. MIT Press.
- FELLBAUM C. (1998). *WordNet : An electronic lexical database*. MIT Press.
- JOSSELIN-LERAY A., FABRE C., REBEYROLLE J., PICTON A. & PLANAS E. (2014). Good Contexts for Translators - A First Account of the Cristal Project. In *Proceedings of the XVI EURALEX International Congress*, p. 631–645, Bolzano, Italy.
- KILGARRIFF A., RYCHLÝ P., HUSÁK M., RUNDELL M. & MCADAM K. (2008). GDEX : Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress*, p. 425–432, Barcelona.
- MARSHMAN E. (2014). Enriching terminology resources with knowledge-rich contexts : A case study. *Terminology*, **20**(2), 225–249.
- MARSHMAN E., GARIÉPY J. L. & HARMS C. (2012). Helping language professionals relate to terms : Terminological relations and termbases. *JoSTrans*, **18**.
- MARSHMAN E., L'HOMME M.-C. & SURTEES V. (2008). Portability of cause-effect relation markers across specialised domains and text genres : a comparative evaluation. *Corpora*, **3**(2), 141–172.
- MEYER I. (2001). Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework. In B. DIDIER, J. CHRISTIAN & M.-C. L'HOMME, Eds., *Recent Advances in Computational Terminology*, p. 279–302.
- MORIN E. (1999). Des patrons lexico-syntaxiques pour aider au dépouillement terminologique. *Traitement Automatique des Langues (TAL)*, **40**(1), 143–166.
- REBEYROLLE J. & TANGUY L. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de grammaire*, **25**, 153–174.
- SAGGION H. (2004). Identifying Definitions in Text Collections for Question Answering. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, p. 1927–1930.
- SCHUMANN A.-K. (2012). Towards the Automated Enrichment of Multilingual Terminology Databases with Knowledge-Rich Contexts—Experiments with Russian EuroTermBank Data. In *Proceedings of the 2nd Workshop on the Creation, Harmonization and Application of Terminology Resources (CHAT'12)*, p. 27–34.
- SÉGUÉLA P. (2001). Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques. *Thèse en Informatique, Université Toulouse 3*.
- SINCLAIR J. M., JONES S. & DALEY R. (1970). *English Lexical Studies. Final Report of O.S.T.I. Programme C/LP/08*.

Traitement automatique des formes métriques des textes versifiés

Éliane Delente & Richard Renault

CRISCO EA 4255

Université de Caen Basse-Normandie

eliane.delente@unicaen.fr, richard.renault@unicaen.fr,

Résumé. L'objectif de cet article est de présenter tout d'abord dans ses grandes lignes le projet *Anamètre*¹ qui a pour objet le traitement automatique des formes métriques de la poésie et du théâtre français du début du XVII^e au début du XX^e siècle. Nous présenterons ensuite un programme de calcul automatique des mètres appliqué à notre corpus dans le cadre d'une approche déterministe en nous appuyant sur la méthode métricométrique de B. de Cornulier ainsi que la procédure d'appariement des rimes et la détermination des schémas de strophes dans les suites périodiques et les formes fixes.

Abstract.

Automatic Processing of Metrical Forms in Verse Texts.

The purpose of this paper is to present the project *Anamètre*. The project proposes automatic processing of metrical forms in French poetry and drama from the early seventeenth to the early twentieth century. Then we present the calculation program of meters on our corpus using a deterministic approach relying on the "métricométric" method of B. de Cornulier. Finally, we present the procedure of matching rimes and determination of rhyme schemes in periodic sequences and specific forms.

Mots-clés : *métrique française, corpus poétique, calcul du mètre, appariement des rimes, détermination des schémas de strophes*

Keywords: *French metrics, poetic corpus, calculation of meter, matching rhymes, determining of rhyme schemes*

1 Introduction

Le projet en question, mis en place il y a huit ans, a pour objet le traitement automatique des formes métriques dans la poésie littéraire française et dans le théâtre, du début du XVII^e au début du XX^e siècle. Nous présenterons le projet dans sa globalité ainsi que les différentes étapes de l'analyse automatique qui conduisent à calculer de manière satisfaisante les différents mètres des vers, à repérer les rimes, à déterminer les schémas strophiques, ainsi qu'à restituer la forme globale des textes (formes fixes : sonnet, ballade, terza rima, rondeau... et suites périodiques : suites de distiques, de quatrains, de sizains...).

À ce jour, le corpus analysé est constitué de 8900 poèmes et 75 pièces de théâtre pour 56 auteurs, ce qui représente un total de 540 000 vers intégrés dans une base de données qui regroupe les différents résultats de l'analyse métrique, rimique et strophique.

2 Métrique et informatique

En tant qu'objets construits, les textes versifiés sont intégralement structurés par des équivalences métriques portant sur les vers et les schémas de strophes au moyen des rimes. Pourtant, la métrique a très peu été investie par le traitement automatique. Notons cependant les travaux pionniers à cet égard de Roubaud (1986 et 1988) et Beaudouin & Yvon (1994) et Beaudouin (2002).

Le renouvellement de la métrique française dans les années 80 par B. de Cornulier fournit un cadre théorique et méthodologique au développement de procédures et programmes d'analyse des régularités métriques. Pourtant, cette méthode quantitative, fiable et reproductible est parfois perçue comme trop formelle pour un objet littéraire, alors qu'elle est à même de fournir aux études stylistiques des analyses métriques rigoureuses et argumentées.

Par ailleurs, ce type de traitement est aujourd'hui plus abordable. Les linguistes peuvent rapidement s'approprier des langages de scripts tels que Perl ou Python et l'émergence du langage de description de données XML ainsi que la normalisation des pratiques d'encodage encouragée par la TEI ont considérablement facilité la façon de traiter les données linguistiques.

¹ Le site web *Métrique en ligne* est associé au projet *Anamètre* : <http://www.crisco.unicaen.fr/verlaine/>

3 Objectifs et résultats attendus

3.1 Les exploitations strictement métriques

La grande nouveauté est que chaque unité rythmique est désormais balisée. On peut donc rapidement et aisément les extraire pour diverses exploitations :

- 1- Établir des dictionnaires de diérèses par auteur notamment et ainsi constituer une histoire de la versification à cet égard.
- 2- Observer systématiquement la composition des vers complexes (composé de 2 hémistiches), notamment par l'examen des propriétés des mots à la césure.
- 3- Étudier systématiquement les relations entre unités rythmiques : l'hémistiche, le vers, le module de strophe et la strophe.

Ces analyses, qui relèvent de l'organisation micro-textuelle, conduisent au second grand type d'exploitation qu'on peut appeler macro-textuelle.

3.2 Les exploitations textuelles

Le balisage des unités rythmiques permettra d'examiner l'organisation syntaxique et sémantique des textes versifiés :

- 1- les études statistiques en lexicométrie et textométrie gagneront en pertinence puisque les recherches lexicales se feront non plus sur le texte poétique traité en bloc mais selon son organisation rythmique – strophe, module de strophe, vers et hémistiches- jusqu'alors négligée faute d'outils.
- 2- les études sur le degré de convergence entre les expressions linguistiques et les unités rythmiques pourront se développer (propriétés sémantiques du second hémistiche par rapport au premier, du second module de strophe par rapport au premier, de la strophe)

L'enjeu est le développement d'une linguistique textuelle caractéristique du texte versifié.

4 Le projet dans ses grandes lignes

Le projet comporte trois volets :

- 1- la constitution d'une base de données de textes poétiques et théâtraux annotés, du début du XVIIIe au début du XXe siècle
- 2- la conception et la mise au point de programmes d'analyse métrique pour l'élaboration d'une base de données de relevés métriques générés automatiquement (un relevé métrique est une série d'analyses métriques portant sur un objet métriquement cohérent et autonome)
- 3- la constitution d'une base de données de relevés métriques faits manuellement par divers métriciens depuis les années 80 qu'il convient de normaliser et d'optimiser pour les rendre pérennes

Le développement qui suit porte uniquement sur le second volet.

5 Les étapes du traitement du corpus

5.1 Formatage initial du texte

Le texte fourni en entrée du traitement est un fichier au format XML-TEI qui contient un formatage minimal en sections et sous-sections <div>, strophes typographiques <lg>, et vers <l> pour un poème ; actes et scènes <div>, tours de parole <sp> et vers <l> pour les pièces de théâtre. La gestion des retraits des vers en cas de polymétrie est réalisée automatiquement après une première passe calculant la longueur métrique des vers. Exemple extrait d'un poème de Théodore de Banville (*Le Budget*, Occidentales, 1875, vers 5) :

(1) Formatage initial

```
<l>Malgré ses ailes d'aigle et son corps de lion,</l>
```

5.2 Découpage du texte en mots

La notion de mot est celle définie par les expressions régulières. Bien que le découpage en mots ne soit pas nécessaire à l'identification des voyelles de l'étape suivante, il est indispensable dès lors qu'il s'agit de mettre en rapport les mots du texte avec des ressources lexicales. Un algorithme déterministe ne peut pas décider si un mot doit être traité en diérèse ou en synérèse, ou si un mot commençant par un "h" doit être traité comme jonctif ou disjonctif. Le découpage en mots

est également utile lors du calcul du mètre pour l'identification de prépositions ou de clitiques et pour l'identification de voyelles masculines ou féminines (une voyelle masculine ou féminine ne peut se définir que par rapport au mot qui la contient). Enfin, les mots sont également pertinents pour la description des rimes.

(2) Découpage du vers en mots

[Malgré][ses][ailes][d]'[aigle][et][son][corps][de][lion],

5.3 Insertion des balises de noyaux syllabiques

Les voyelles des noyaux syllabiques sont identifiées au moyen de règles de transposition de graphèmes en phonèmes, écrites sous forme d'expressions régulières qui prennent en compte les contextes gauche et droit du graphème. Quatre attributs sont associés à la balise <seg> qui délimite la voyelle : 1) le type (voyelle stable, voyelle instable, semi-consonne ou voyelle ambiguë), 2) le numéro de la règle appliquée, 3) le symbole phonétique (API) du phonème, 4) la valeur par défaut pour le traitement métrique (0 ou 1).

(3) Identification des noyaux syllabiques (voyelles)

Malgré ses ailes d'aigle et son corps de lion,
 vs vs vs vs vi vs vi vs vs vs vi scvs

vs = voyelle stable
 vi = voyelle instable
 sc = semi-consonne

5.4 Statut métrique ou non des "e" instables

Selon le contexte, les "e" instables sont transformés en voyelles métriques ("e" masculin ou "e" féminin) ou non ("e" élidé, "e" ignoré, "e" écarté). Par exemple, un "e" de fin de mot sera traité comme un "e" féminin s'il est suivi d'un mot commençant par une consonne, et traité comme "e" élidé s'il est suivi d'un mot commençant par une voyelle. Les mots commençant par un "h" ou une semi-consonne sont traités au moyen d'une ressource lexicale qui précise s'ils sont jonctifs ou disjonctifs. Le mot est jonctif s'il commence phonologiquement par une voyelle, et disjonctif, s'il commence phonologiquement par une consonne. Les modifications portent sur les attributs de l'élément <seg> : type, règle et valeur.

(4) Traitement des "e" instables

Malgré ses ailes d'aigle et son corps de lion,
 ef ee em

em = e masculin
 ef = e féminin
 ee = e élidé

5.5 Conversion des semi-consonnes en voyelles ou en consonnes

Les semi-consonnes sont, soit assimilées à des consonnes, soit traitées comme des voyelles (diérèse). Par défaut, le statut des semi-consonnes est donné par un dictionnaire de dièses dont les entrées ont été établies à partir d'une version fléchée des mots du dictionnaire Littré dont le découpage syllabique manifeste une diérèse (ex : [li-on] pour *lion*). Les modifications portent sur les attributs : type, règle, phonème et valeur. Dans l'exemple suivant, la semi-consonne du mot *lion* est convertie en voyelle stable (diérèse) conduisant ainsi à une interprétation rythmique conforme au système classique :

(5) Traitement des semi-consonnes

Malgré ses ailes d'aigle et son corps de lion,
 vs

vs = voyelle stable

5.6 Calcul de la longueur métrique des vers

Les voyelles reçoivent un attribut de place (numéro d'ordre) si elles ont une valeur métrique. La longueur métrique est obtenue à partir du décompte de toutes les voyelles métriques jusqu'à la dernière voyelle stable ; le "e" féminin de fin

de vers en est exclu puisqu'il n'appartient pas au domaine du mètre mais à celui de la rime.

(6) Calcul de la longueur métrique des vers

Malgré ses ailes d'aigle et son corps de lion,
 1 1 1 1 1 1 0 1 1 1 1 1 = 12

5.7 Calcul du mètre des vers

5.7.1 Distribution de propriétés pertinentes pour le calcul du mètre

Pour tous les vers dont la longueur métrique est ≤ 8 , le mètre correspond à la longueur métrique. Pour tous les vers dont la longueur métrique est > 8 , le calcul du mètre met en œuvre une procédure en trois temps : 1) Repérage de propriétés phonologiques (place des voyelles masculines et féminines dans le mot), morphologiques (distribution des proclitiques dans le vers), et syntaxiques (distribution des prépositions monosyllabiques). Ces propriétés (propriétés métricométriques²) sont pertinentes pour déterminer le profil métrique du texte puis le mètre des vers. Précisons que ce traitement n'utilise pas d'analyseur syntaxique mais seulement un algorithme fondé sur des règles d'identification contextuelle des proclitiques, des prépositions monosyllabiques, et de mots ambigus tels que *entre* et *contre* qui sont des prépositions monosyllabiques en contexte d'élision. Cet algorithme utilise une ressource lexicale constituée uniquement de mots grammaticaux (prépositions, déterminants, pronoms...). Une voyelle métrique reçoit un attribut de propriété dans les cas suivants : 1) si elle appartient à une préposition monosyllabique (propriété P), si elle appartient à un proclitique (propriété C), si la voyelle masculine est prétonique de mot (propriété M), si la voyelle féminine est post-tonique de mot (propriété F).

(7) Distribution des propriétés métricométriques

Malgré [ses] ailes d'aigle et [son] corps [de] lion,
 M [C] F [C] [P]

propriétés phonologiques :
 M = voyelle masculine prétonique
 F = voyelle féminine
 propriétés morphosyntaxiques :
 [P] = préposition monosyllabique
 [C] = proclitique

5.7.2 Détermination d'un profil métrique

La distribution régulière des propriétés repérées sur l'ensemble du texte permet d'en évaluer le profil métrique (6+6 par exemple pour un poème constitué uniquement de vers classiques de 12 voyelles). Ce profil métrique est nécessaire notamment au calcul du mètre des vers afin de départager le mètre 4+6 du 5+5 pour une même longueur métrique de 10 voyelles.

5.7.3 Calcul du mètre des vers complexes

Le profil métrique correspondant à chaque longueur métrique permet ensuite de préciser le mètre de chaque vers en tenant compte de la présence éventuelle des propriétés métricométriques dans les places pertinentes pour le calcul du mètre (places 6, 4 et 8, par exemple, pour les vers de 12 voyelles). À l'issue de cette procédure, une balise de césure <caesura /> est introduite entre les hémistiches.

(8) Calcul du mètre des vers (après détermination du profil métrique du poème)

Malgré ses ailes d'aigle | et son corps de lion, 6+6

| = césure
 6+6 = mètre du vers

² Cornulier, Benoît de, 1982, *Théorie du vers*, Éditions du Seuil, Paris
 Cornulier, Benoît de, 1982 1995, *Art Poétique*, Presses universitaires de Lyon.

5.8 Traitement des rimes

5.8.1 Appariement des vers en rimes

L'appariement (appel-écho) se fait à partir de la dernière voyelle tonique du vers. La relation rimique entre deux vers est établie s'il y a : 1) une équivalence phonétique de la voyelle tonique, 2) une équivalence graphique de la consonne finale, 3) une équivalence graphique des consonnes subséquentes (consonnes placées entre la voyelle tonique et le "e" des terminaisons féminines). L'appariement des vers est soumis à une contrainte de localité³ qui fait qu'un appel trouvera son écho dans un domaine proche. Les schémas de rimes permettent l'identification des strophes qui sont ensuite comparées avec des schémas de formes fixes (sonnet, ballade, rondeau...) ou des schémas de formes périodiques (distique, quatrain, quintil...) afin de déterminer la forme globale du poème.

(9) Appariement des vers en rimes

Malgré ses ailes d'aigle | et son corps de lion, 3
 ...
 Et je pense qu'avec un petit million 3

■ = segment phonologique de la rime
 3 = n° d'ordre d'appariement de la rime

5.8.2 Détermination de la forme globale

Quatre procédures sont utilisées pour déterminer la forme globale : 1) Reconnaissance d'une forme fixe à partir des données quantitatives (nombre total de vers, nombre de strophes, nombre de vers par strophe), des schémas de rimes et des éventuelles répétitions totales ou partielles de vers. Cette procédure permet d'identifier ainsi les sonnets, les ballades, les rondeaux, les triolets, les terza rima, les pantoums, les sextines... 2) Reconnaissance d'une forme périodique à partir de la distribution régulière des schémas de rimes relativement au découpage du poème en strophes typographiques. Cette procédure permet d'identifier les suites régulières de distiques, quatrains, quintils, sizains... ainsi que les différentes combinaisons de schémas rimiques. 3) Reconnaissance d'une suite de distiques tels qu'on les trouve dans les pièces de théâtre. 4) Reconnaissance d'une suite de schémas de rimes distincts au moyen d'expressions régulières construites à partir d'un répertoire de base de schémas de rimes. Cette dernière procédure permet d'identifier des séquences de schémas rimiques dans les poèmes sans découpage typographique en strophes.

(10) Construction des strophes au moyen des schémas de rimes

Malgré ses ailes d'aigle | et son corps de lion, a
 Il n'a pas du tout l'air farouche b
 Et je pense qu'avec un petit million a
 Nous pourrons lui fermer la bouche b

■ = segment phonologique de la rime
 a = étiquette de la rime dans le schéma de strophe

6. Choix théoriques et méthodologiques

Notre approche diffère très nettement de celle de V. Beaudouin, tout d'abord, par l'ampleur et la nature du corpus traité. Le *Métromètre*, appliqué essentiellement au théâtre classique, n'épuise donc pas, loin s'en faut, la variété métrique et strophique de la poésie littéraire française puisqu'il ne traite guère que du seul mètre 6+6 et du seul schéma strophique (aa) en rimes suivies. *Anamètre* n'a pas de limitation de ce type puisqu'il couvre près de trois siècles d'écriture théâtrale et poétique dans le cadre du système métrique classique, offrant ainsi toutes les variétés métriques et strophiques caractéristiques de ce système. Mais il traite également les productions de la deuxième moitié du XIX^e siècle et du début du XX^e siècle qui obéissent à des régularités assouplies conduisant progressivement à l'abandon de la notion même de régularité métrique. Le corpus offre ainsi une perspective historique indispensable pour l'exploitation et l'interprétation des données.

³ Ruwet, Nicolas, 1981, "Linguistique et poétique : une brève introduction", *Le Français moderne*, 49:1, p. 1-19
 Cornulier, Benoît de, 1982, *Théorie du vers*, Éditions du Seuil, Paris

Le type d'approche est également très différent. Nos programmes ne procèdent pas à un découpage du vers en syllabes puisque les consonnes ne sont pas pertinentes pour le traitement du mètre. Ils ne phonétisent pas non plus les vers. Ce traitement se révèle en effet inutile ; puisque les régularités métriques reposent pour une large part sur la graphie, la problématique des régularités métriques et celle de la déclamation sont en grande partie indépendantes. Enfin, nous n'avons pas recours à des patrons métriques extérieurs au poème ; notre approche consiste à intégrer le principe d'équivalence contextuelle au moyen de la notion de profil métrique.

Afin d'évaluer la fiabilité des résultats du traitement, une comparaison systématique, vers par vers, avec le corpus analysé manuellement par B. de Cornulier est en cours. Mais un exemple peut d'ores et déjà nous servir d'indice de fiabilité. *Les Méfaits de la lune* est un poème appartenant à la production tardive de Verlaine et présentant une complexité métrique telle qu'elle rend problématique l'analyse humaine du poème et malgré cela, l'analyse automatique du poème est en accord avec celle de J. L. Aroui⁴.

7. Conclusion

Comme on peut s'y attendre, l'état actuel du traitement présente des limites, touchant notamment à l'interprétation de nouveaux mètres, dits de substitution, apparaissant à partir de 1860. En effet, certains de ces vers présentent une coupe possible en 4e et 8e positions. Ces vers sont tous potentiellement 4.4.4. mais pour certains, l'interprétation 8.4 ou bien 4.8 paraît plus évidente que l'interprétation 4.4.4. Seule une analyse syntaxique du vers, croisée à l'analyse métrique effectuée, pourrait permettre de décider quelle est l'interprétation rythmique la plus naturelle entre ces trois possibilités. À ce jour, l'analyse syntaxique du vers n'a pas encore été intégrée dans nos programmes qui proposent pour l'instant une interprétation correcte mais sous-déterminée.

⁴ Répertoire métrique des *Œuvres Poétiques complètes* de Paul Verlaine, Nantes, M.S.H. Ange Guépin, Centre d'Études Métriques, 1993.

Bibliographie

- Beaudouin V. (2000), *Rythme et rime de l'alexandrin classique, étude empirique des 80 000 vers du théâtre de Corneille et Racine*. Thèse de Doctorat. EHESS.
- Beaudouin V. (2002), *Mètre et rythmes du vers classique : Corneille et Racine*, Paris, Honoré Champion.
- Bobillot, J. P. (1991) : Recherches sur la crise d'identité du vers dans la poésie française _ 1873-1913 [4 vol.], Université Sorbonne Nouvelle, Paris 3.
- Bobillot, J. P. (1993) : « Entre mètre & non-mètre : le « décasyllabe chez Verlaine », *Revue Verlaine* n° 1, pp. 179-200, Musée Bibliothèque Rimbaud, Charleville-Mézières.
- Bobillot, J. P. (1994) : « De l'anti-nombre au quasi-mètre : le « hendécasyllabe » chez Verlaine », *Revue Verlaine* n° 1, pp. 66-86, Musée Bibliothèque Rimbaud, Charleville-Mézières.
- Cornulier, B. de (1979) : *Problèmes de métrique française*, thèse d'État, Université de Provence.
- Cornulier B. de (1982), *Théorie du vers : Rimbaud, Verlaine, Mallarmé*, Paris, Seuil.
- Cornulier B. de (1995), *Art poétique*, IUFM, Presses universitaires de Lyon.
- Cornulier B. de (1999), *Petit dictionnaire de métrique*, Université de Nantes.
- Cornulier B. de (2000a), "La place de l'accent, ou l'accent à sa place : position, longueur, concordance", dans *Le Vers Français : Histoire, théorie, esthétique*, textes recueillis par Michel Murat, Champion, Paris, p. 57-91.
- Cornulier B. de (2000b), « L'invention du « décasyllabe » chez Verlaine décadent. Le 4-6, le 5+5, le mixte, et le n'importe quoi », Actes du colloque *Verlaine à la loupe*, 11-18 juillet 1996, dir. J. M. Gouvard et S. Murphy, Paris, Honoré Champion.
- Cornulier B. de (2003), « Problèmes d'analyse rythmique du non-métrique » Semen.
- Cornulier, B. de (2009) : « Types de césures ou plutôt manières de rythmer le vers composé », *L'Information grammaticale* n° 121, Peters.
- Cornulier, B. de (2010) : *Notions d'analyse métrique*.
<http://www.normalesup.org/~bdecornulier/notmet.pdf>
- Delente, E. (à paraître) : « La dimension textuelle du rythme. Étude des *Poèmes Saturniens* jusqu'à *Bonheur de Verlaine* », dans *Cahiers du Centre d'Études Métriques* n° 7, université de Nantes.
- Delente E. (2003), "Distribution du syntagme adjectival épithète dans l'alexandrin verlainien", in *Le sens et la mesure. De la pragmatique à la métrique*. Hommages à Benoît de Cornulier, Textes réunis et édités par Jean-Louis Aroui, Paris, Ed. Honoré Champion, p. 399-414.
- Delente E. (2010), « Métrique et traitement automatique. Une question difficile : l'étude de la concordance », Actes du colloque *Linguistique et Littérature : Cluny, 40 ans après*, Besançon, 29-31 octobre 2007.
- Delente E. & Renault R. (2011) : « Annotation automatique des textes versifiés », Colloque « Patrimoine à l'ère du numérique » – 10-11 décembre 2009 à Caen, Schedae, prépublication n° 5, fascicule n° 1, p. 39-52.
- Dominicy, M. (1984) : « Sur la notion d'e féminin ou masculin en métrique et en phonologie », *Recherches linguistiques* n° 12, pp. 7-45.
- Gouvard J.-M. (2000), *Critique du vers*, Paris, Honoré Champion.
- Holowacz, W. (1997) : « Le décasyllabe dans l'œuvre de Verlaine », *Cahiers du Centre d'Études Métriques* de Nantes, Université de Nantes.
- Murat M. (éd.) (2000), *Le vers français : histoire, théorie, esthétique*, Paris, Honoré Champion.
- Roubaud J. (1986) et (1988) : « Dynastie : études sur le vers français, sur l'alexandrin classique » *Cahiers de poétique comparée* n°13 et n°16, publications des langues'O.
- Ruwet N. (1981), « Typography, Rhymes, and Linguistic Structures in Poetry », in W. Steiner (éd.) *The Sign in Music and Literature*, University of Texas Press, Austin, 242 p.
- Tobler, A. (1885) : *Le vers français ancien et moderne*, Slatkine, Genève, éd. 1972.

Apprentissage automatique d'un modèle de résolution de la coréférence à partir de données orales transcrites du français : le système CROC

Adèle Désoyer^{1, 2} Frédéric Landragin¹ Isabelle Tellier¹

(1) Lattice, CNRS, ENS et Université de Paris 3 - Sorbonne Nouvelle

(2) MoDyCo, CNRS, Université Paris Ouest - Nanterre La Défense

adele.desoyer@gmail.com, frederic.landragin@ens.fr, isabelle.tellier@univ-paris3.fr

Résumé. Cet article présente CROC¹ (*Coreference Resolution for Oral Corpus*), un premier système de résolution des coréférences en français reposant sur des techniques d'apprentissage automatique. Une des spécificités du système réside dans son apprentissage sur des données exclusivement orales, à savoir ANCOR (anaphore et coréférence dans les corpus oraux), le premier corpus de français oral transcrit annoté en relations anaphoriques. En l'état actuel, le système CROC nécessite un repérage préalable des mentions. Nous détaillons les choix des traits – issus du corpus ou calculés – utilisés par l'apprentissage, et nous présentons un ensemble d'expérimentations avec ces traits. Les scores obtenus sont très proches de ceux de l'état de l'art des systèmes conçus pour l'écrit. Nous concluons alors en donnant des perspectives sur la réalisation d'un système *end-to-end* valable à la fois pour l'oral transcrit et l'écrit.

Abstract.

Machine Learning for Coreference Resolution of Transcribed Oral French Data : the CROC System

We present CROC (Coreference Resolution for Oral Corpus), the first machine learning system for coreference resolution in French. One specific aspect of the system is that it has been trained on data that are exclusively oral, namely ANCOR (ANaphora and Coreference in ORal corpus), the first corpus in oral French with anaphorical relations annotations. In its current state, the CROC system requires pre-annotated mentions. We detail the features that we chose to be used by the learning algorithms, and we present a set of experiments with these features. The scores we obtain are close to those of state-of-the-art systems for written English. Then we give future works on the design of an end-to-end system for oral and written French.

Mots-clés : corpus de dialogues, détection de coréférences, apprentissage, paires de mentions.

Keywords: Dialogue corpus, Coreference resolution, Machine learning, Mention-pair model.

1 Introduction

Depuis les vingt dernières années, la reconnaissance automatique des chaînes de coréférence représente un objet d'étude à part entière du TAL, au cœur de grandes campagnes d'évaluation telles que celles proposées par MUC (*Message Understanding Conference*²), ACE (*Automatic Content Extraction*³), SemEval (*Semantic Evaluation*⁴) ou CoNLL (*Computational Natural Language Learning*⁵). Ces chaînes constituent une unité discursive complexe qui contribue à la cohésion du discours. Les identifier automatiquement oblige à prendre en compte la séquence des phrases qui le composent, et leurs relations. Ce domaine a donné lieu à de nombreux travaux, mais les données sur lesquelles ils se sont fondés (issues des campagnes précédemment citées) étaient jusqu'à présent essentiellement de l'anglais écrit. Les travaux présentés dans cet article ont la particularité de se concentrer sur la reconnaissance automatique de chaînes de coréférence présentes dans de

1. Présenté plus en détails dans (Désoyer *et al.*, 2015)

2. Voir notamment la tâche sur la coréférence dans MUC-7 en 1998, cf. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.

3. Cf. <http://www.itl.nist.gov/iad/mig//tests/ace/>.

4. Voir notamment la tâche sur la coréférence dans SemEval-2 en 2010, cf. [http://semeval2.fbk.eu/semeval2.php?location=](http://semeval2.fbk.eu/semeval2.php?location=tasks)tasks.

5. Voir notamment la tâche sur la coréférence en 2011 et 2012, cf. <http://conll.cemantix.org/2011/> et <http://conll.cemantix.org/2012/>.

l’oral transcrit français. Cette modalité rend-elle les coréférences plus ou moins fréquentes, explicites et faciles à repérer ? Ce sera tout l’enjeu des expériences que nous avons menées.

Commençons par définir précisément notre objet d’étude. Les chaînes de coréférence s’appuient sur la notion plus restreinte d’anaphore. Cette dernière décrit une procédure référentielle regroupant les phénomènes de renvoi à un antécédent du discours immédiat. Les anaphores sont des relations asymétriques entre un antécédent et une expression anaphorique qui ne peut être interprétée qu’à partir de son antécédent. Dans l’exemple de la figure 1, extrait d’un dialogue du corpus OTG d’ANCOR, l’anaphorique nominal « le nom » ne peut être interprété qu’à partir de son antécédent « une grande libraire » ; il en va de même pour le pronom « elle ».

- on m’a parlé d’une **grande libraire** mais on se rappelle plus **le nom**
- **Arthaud**
- **Arthaud** peut-être
- **elle** se trouve dans le centre ville

FIGURE 1 – Mise en évidence d’une chaîne de coréférence dans un dialogue

Le phénomène plus large de la coréférence se décrit, quant à lui, comme la relation existant entre plusieurs expressions référant à une même entité. Contrairement à l’anaphore qui distinguait strictement ses deux parties, la relation de coréférence est symétrique. Dans l’exemple de la figure 1, l’ensemble des expressions en gras composent une chaîne de coréférence.

Les systèmes de résolution automatique de la référence sont encore aujourd’hui extrêmement rares s’agissant du français, et même, à notre connaissance, inexistant – à l’exception de systèmes à base de règles tels que celui décrit dans (Trouilleux, 2001) ou RefGen présenté dans (Longo, 2013). Sans recourir intégralement aux modèles statistiques, la méthode hybride de RefGen allie la performance de ceux-ci à la pertinence des modèles symboliques avec un premier module de segmentation probabiliste, puis un module linguistique repérant les marqueurs de cohésion. L’apprentissage automatique n’avait pu encore être mis en œuvre dans ce contexte, faute jusqu’à présent de corpus annotés et disponibles librement. Notre travail repose sur les données du corpus ANCOR⁶, réalisé dans le cadre du projet du même nom financé par la région Centre, qui annote différentes formes de reprise dans le discours : la reprise fidèle (représentant 40 % des reprises annotées) est la relation qui unit deux groupes nominaux référant à la même entité du discours, et qui ont la même tête nominale ; la reprise infidèle est plus rare (7 % des reprises) et se distingue de la précédente par le fait que les deux mentions qu’elle relie ont des têtes nominales distinctes souvent proches sémantiquement (par synonymie, hyperonymie ou hyponymie). La relation la plus représentée dans le corpus (42 % des reprises) est l’anaphore pronominale, c’est-à-dire la reprise d’un groupe nominal par un pronom. Les deux dernières relations représentées relèvent de l’anaphore associative, c’est-à-dire que les deux expressions en relation réfèrent à des entités distinctes, qui entrent par exemple dans une relation méronymique. Parmi ces phénomènes, l’annotation d’ANCOR distingue l’anaphore associative nominale, reliant deux syntagmes nominaux (10 % des reprises) et l’anaphore associative pronominale, reliant un syntagme nominal à un pronom (1 % des reprises). Ce corpus global de 488 000 mots se décompose en fait en quatre sous-corpus issus de précédents projets de recherche : deux d’entre eux sont extraits du projet ESLO et sont essentiellement composés d’interviews, faiblement interactives (CO2, qui représente 8 % des données, et ESLO qui correspond à 85 % des données) ; les deux autres ensembles sont composés de dialogues interactifs, avec d’une part l’office du tourisme de Grenoble (corpus OTG, représentant 5 % des données), d’autre part le standard téléphonique de l’Université de Basse-Normandie (corpus UBS, représentant 2 % des données).

2 Résolution de la coréférence comme tâche de classification

2.1 État de l’art

Les premiers systèmes de résolution automatique de la coréférence traitent la tâche de façon symbolique, avec des règles écrites à la main. Dans les années 1970, la problématique est limitée à la résolution des anaphores pronominales, avec une mise en avant et des pistes pour calculer la saillance (Lappin & Leass, 1994). Elle débouche sur des approches dites *knowledge-poor*, à l’instar de celle de (Mitkov, 2002) qui, pour seul prétraitement, nécessite une analyse morphosyntaxique et un découpage en *chunks*. Ces approches ont rapidement trouvé leurs limites et l’essor de la linguistique de

6. http://tln.li.univ-tours.fr/Tln_Ancor.html.

corpus a encouragé l'exploitation de données attestées. Les efforts de recherche actuels se concentrent désormais sur des approches fondées sur l'apprentissage supervisé, ce qui nécessite deux prérequis : reformuler l'identification d'une chaîne de coréférence comme une tâche que l'on sait aborder par de telles méthodes (par exemple une tâche de classification ou d'annotation) ; disposer d'un corpus annoté servant à la fois pour l'apprentissage et le test, afin d'évaluer les performances du système ainsi construit. Différents types d'approches s'opposent quant à la façon de formuler la tâche confiée aux algorithmes d'apprentissage, parmi lesquels :

- les modèles *mention-pair* ou *pairwise* qui sont fondés sur une classification binaire comparant une anaphore à des antécédents potentiels situés dans les phrases précédentes. Concrètement, les exemples fournis au programme sont des paires de mentions (une anaphore et un antécédent potentiel) pour lesquelles l'objectif est de déterminer si elles sont coréférentes ou non. Une anaphore ne pouvant avoir qu'un unique antécédent, une deuxième phase doit déterminer quel est le véritable antécédent de l'anaphore parmi tous ceux qui sont possibles (à l'intérieur des paires de mentions classées comme coréférentes). Différents systèmes ont été implémentés pour cela, parmi lesquels ceux de (Soon *et al.*, 2001) et (Ng & Cardie, 2002) et plus récemment ceux de (Bengtson & Roth, 2008) et (Stoyanov *et al.*, 2010), régulièrement utilisés comme systèmes de référence à partir desquels les nouveaux systèmes comparent leur performance. Pour les premiers, l'antécédent sélectionné parmi un ensemble pour une anaphore donnée est celui qui en est le plus proche. Il s'agit d'un regroupement dit *Closest-First* qui, pour chaque anaphore, parcourt l'ensemble du texte vers la gauche, jusqu'à trouver un antécédent ou atteindre le début du texte. Les seconds proposent une alternative à cette approche, dit regroupement *Best-First*, qui sélectionne comme antécédent celui ayant le plus haut score de « probabilité coréférentielle » parmi l'ensemble des précédentes mentions. L'inconvénient de ce type de méthode est la réduction du problème à une série de classifications binaires indépendantes, qui ne prend pas en compte l'ensemble des différents maillons d'une même chaîne de coréférence ;
- les modèles *twin-candidate*, proposés dans (Yang *et al.*, 2003) considèrent également le problème comme une tâche de classification, mais dont les instances sont cette fois composées de trois éléments (x, y_i, y_j) où x est une anaphore et y_i et y_j deux antécédents candidats (y_i étant le plus proche de x en termes de distance). L'objectif du modèle est d'établir des critères de comparaison des deux antécédents pour cette anaphore, et de classer l'instance en *FIRST* si le bon antécédent est y_i et en *SECOND* si le bon antécédent est y_j . Cette classification alternative est intéressante car elle ne considère plus la résolution de la coréférence comme l'addition de résolutions anaphoriques indépendantes, mais prend en compte l'aspect « concurrentiel » des différents antécédents possibles pour une anaphore ;
- les modèles *mention-ranking*, tels celui décrit dans (Denis, 2007), envisagent non plus d'étiqueter chaque paire de mentions mais de classer l'ensemble des antécédents possibles pour une anaphore donnée selon un processus itératif qui compare successivement cette anaphore à deux antécédents potentiels : à chaque itération, on conserve le meilleur candidat, puis on forme une nouvelle paire de candidats avec ce « gagnant » et un nouveau candidat. L'itération s'arrête lorsqu'il n'y a plus de candidat possible. Une alternative à cette méthode propose de comparer simultanément tous les antécédents possibles pour une anaphore donnée ;
- les modèles *entity-mention* (Yang *et al.*, 2008) déterminent quant à eux la probabilité qu'une expression réfère à une entité ou à une classe d'entités précédemment considérées comme coréférentes (*i.e.* un candidat est comparé à un unique antécédent ou à un cluster contenant toutes les références à une même entité).

Enfin, parmi les travaux les plus récents, certains cherchent à concilier les avantages de chaque méthode, y compris celles à base de règles, en distinguant plusieurs strates de résolution de manière à optimiser les performances en fonction des phénomènes ciblés par chaque strate (Lee *et al.*, 2013).

2.2 Reformulation du problème en tâche de classification

CROC est fondé sur une classification binaire opérant sur des paires d'unités référentielles, pour les ranger soit dans la classe des mentions coréférentes, soit dans celle des mentions non coréférentes. La qualité d'un tel système repose sur les données qui lui sont fournies en apprentissage, et particulièrement sur les traits linguistiques (ou *attributs*) décrivant les unités à classer.

Les systèmes de l'état de l'art procédant de la sorte s'inspirent tous de l'ensemble des traits définis dans (Soon *et al.*, 2001), composé de douze attributs décrivant les propriétés d'un antécédent i et d'une reprise potentielle j : 1) Distance en nombre de phrases entre i et j ; 2) i est-il un pronom ? ; 3) j est-il un pronom ? ; 4) Les chaînes de caractères de i et j sont-elles égales ? ; 5) j est-il un SN défini ? ; 6) j est-il un SN démonstratif ? ; 7) i et j s'accordent-ils en nombre ? ; 8) i et j s'accordent-ils en genre ? ; 9) i et j appartiennent-ils à la même classe sémantique ? ; 10) i et j sont-ils tous deux des noms propres ? ; 11) i et j sont-ils alias l'un de l'autre ? ; 12) i et j sont-ils au sein d'une structure appositive ?

(Ng & Cardie, 2002) ajoutent à ces douze traits de référence quarante et un nouveaux, également repris dans les travaux plus récents. Ces attributs se répartissent dans deux familles : les traits non relationnels, qui décrivent une mention d'entité, et les traits relationnels, qui caractérisent la relation unissant les deux mentions d'une paire. Notre propre ensemble reprend ceux-ci et en ajoute certains, en s'adaptant aux spécificités des données dont nous disposons. Ainsi, les traits non relationnels que nous intégrons sont de différents types :

- morphosyntaxiques (correspondant aux traits 2, 3, 5 et 6 de l'ensemble de (Soon *et al.*, 2001)) ;
- énonciatifs (une mention initie-t-elle une chaîne de coréférence ou non ?) ;
- sémantiques (une mention est-elle une entité nommée ? Si oui, de quel type ?).

Quant aux traits relationnels, ils nous permettent de caractériser différentes distances entre un antécédent m_1 et une reprise m_2 potentielle :

- distances lexicales (m_1 et m_2 sont-elles strictement égales ? partiellement égales ?) ;
- distances morphosyntaxiques (m_1 et m_2 s'accordent-elles en genre ? en nombre ?) ;
- distances spatiales (m_1 et m_2 sont séparées de combien de caractères ? de mots ? de mentions ? de tours de paroles ?) ;
- distance syntaxique (l'une des deux mentions est-elle une partie de l'autre ? Autrement dit, si l'une des deux est un syntagme nominal complexe l'autre est-elle une partie de ce syntagme ?) ;
- distances contextuelles (m_1 et m_2 sont-elles précédées du même token ? suivies du même token ?) ;
- distance énonciative (m_1 et m_2 sont-elles produites par le même locuteur ?).

Un total de 30 traits décrivant différents niveaux linguistiques constituent l'ensemble que nous utilisons dans nos séries de test.

3 Expérimentations d'apprentissage et résultats

3.1 Plan d'expérimentations

Différents paramètres entrent en jeu dans la construction du système de détection de chaînes de coréférence ; c'est l'optimisation de la combinaison de ces paramètres qui permettra au système d'améliorer ses performances. Les expérimentations que nous menons ici consistent à générer différents modèles de classification en faisant varier d'une part la représentation des données, d'autre part l'algorithme de calcul du modèle. Afin de mesurer la qualité de chacun des modèles générés sans introduire de biais lié à la dépendance aux données d'apprentissage, nous en distinguons trois ensembles :

- un ensemble d'apprentissage (60 % des données initiales) utile au calcul des différents modèles
- un ensemble de développement (20 % des données initiales) fourni à chacun des modèles générés afin de déterminer celui qui optimise au mieux les paramètres
- un ensemble de test (20 % des données initiales) fourni au système final pour en évaluer ses performances sur de nouvelles données

Afin d'évaluer l'influence de la taille de l'ensemble d'apprentissage sur les résultats de classification, un premier paramètre consiste en la variation du nombre de données utiles au calcul du modèle (trois ensembles sont donc sélectionnés : un réduit dit *small_trainingSet*, un moyen dit *medium_trainingSet* et un grand dit *big_trainingSet*)⁷. Un second paramètre concerne l'ensemble d'attributs décrivant ces données, et de nouveau, différents sont testés : un premier ensemble contient tous les attributs, un second uniquement les attributs relationnels, et un troisième exclut les traits spécifiques de l'oral tels que la correspondance des locuteurs ou la distance en tours de parole. Enfin, nous nous inspirons de l'état de l'art pour tester trois algorithmes distincts : les arbres de décision, les SVM et Naïve Bayes tels qu'implémentés dans la plate-forme Weka, avec leurs paramètres par défaut.

Le plan d'expérimentations mis en place consiste à combiner les données des différents corpus d'apprentissage avec les trois ensembles d'attributs produits, puis à fournir chacune de ces représentations de données aux trois algorithmes d'apprentissage. Chacun des modèles ainsi généré permet alors de classer de nouvelles paires (ensemble de développement) qui sont ensuite filtrées pour ne conserver qu'un unique antécédent pour une anaphore (les résultats de classification peuvent en effet associer une même reprise à différents antécédents). Cette sélection s'appuie sur celle décrite dans les travaux de (Soon *et al.*, 2001) : il s'agit de la stratégie *Closest-First* qui, lorsqu'une mention a plusieurs antécédents possibles, sélectionne le plus proche à gauche pour former une chaîne. Tous les systèmes ainsi générés seront comparés quantitativement à partir de leurs résultats chiffrés.

7. Les paires de mentions sont sélectionnées aléatoirement au sein du corpus, sans distinguer les sous-corpus dans lesquels elles apparaissent, et sont réparties telles que : 71 881 instances dont 11 908 paires coréférentes et 59 973 non coréférentes dans *small_trainingSet* ; 101 919 instances dont 17 844 paires coréférentes et 84 075 non coréférentes dans *medium_trainingSet* ; 142 498 instances dont 24 620 paires coréférentes et 117 878 non coréférentes dans *big_trainingSet*.

System	Langue	Corpus	MUC	B ³	CEAF	BLANC
<i>Systèmes end-to-end</i>						
(Soon <i>et al.</i> , 2001)	ANGLAIS	MUC-7	60.4	-	-	-
(Ng & Cardie, 2002)	ANGLAIS	MUC-7	63.4	-	-	-
(Stoyanov <i>et al.</i> , 2009)	ANGLAIS	ACE-2003	67.9	65.9	-	-
(Stoyanov <i>et al.</i> , 2010)	ANGLAIS	MUC-7	62.8	79.4	-	-
(Haghighi & Klein, 2010)	ANGLAIS	ACE-2004	67.0	77.0	-	-
(Lassalle, 2015)	ANGLAIS	CoNNL-2012	68.8	54.56	50.20	-
(Longo, 2013)	FRANCAIS	MULTI-GENRES	36	69.7	55	59.5
<i>Systèmes pré-annotés</i>						
(Yang <i>et al.</i> , 2003)	ANGLAIS	MUC-7	60.2	-	-	-
(Luo <i>et al.</i> , 2004)	ANGLAIS	ACE-2	80.7	77.0	73.2	77.2
(Denis & Baldridge, 2008)	ANGLAIS	ACE-2	71.6	72.7	67.0	-
(Bengtson & Roth, 2008)	ANGLAIS	ACE-2004	75.1	80.8	75.0	75.6
CROC	FRANCAIS	ANCOR	63.45	83.76	79.14	67.43

TABLE 1 – Résultats de systèmes de résolution de la coréférence

3.2 Évaluations

La tâche de la résolution de la coréférence est traditionnellement évaluée selon quatre métriques :

- MUC (dont le nom est issu de la campagne d'évaluation *Message Understanding Conference*) se concentre sur l'évaluation des liens de coréférence qui sont communs à l'ensemble des chaînes avérées et à celui des chaînes prédites par le système.
- B³ : (Bagga & Baldwin, 1998) considère comme unité de base la mention plutôt que le lien.
- CEAF, développé dans les travaux de (Luo, 2005), est fondée sur l'entité, c'est-à-dire la référence commune à tous les maillons d'une chaîne de coréférence (le nom complet de la mesure est *Constrained Entity Aligned F-Measure*).
- BLANC (pour *BiLateral Assessment of Noun-phrase Coreference*) est la métrique la plus récente mise au point dans les travaux de (Recasens, 2010), dont la vocation est de considérer conjointement les liens de coréférence et de non-coréférence.

À titre de repères, le haut du tableau 1 présente les résultats de certains des systèmes *end-to-end* (c'est-à-dire sans connaître *a priori* les positions des mentions dans le corpus de test) les plus connus à ce jour pour la langue anglaise. Les résultats du système RefGen de (Longo, 2013), correspondant à la moyenne des scores obtenus pour différents genres textuels, sont également présentés dans ce tableau 1. Par ailleurs, les résultats des systèmes non *end-to-end* qui, comme le nôtre, basent leurs expérimentations sur de *vraies* mentions, sont présentés en bas de tableau 1 (à noter que tous sont conçus pour l'anglais écrit standard).

Les résultats obtenus sur l'ensemble de développement démontrent nettement que NaiveBayes n'est pas idéal pour la résolution de la coréférence. C'est dans la majorité des cas les systèmes appris par l'algorithme SVM sur les données du plus petit ensemble d'apprentissage qui présentent les meilleurs résultats sur de nouvelles données (*i.e.* l'ensemble de développement). Concernant l'ensemble de traits, le plus pertinent semble être le plus exhaustif puisque c'est en l'incluant que les modèles sont meilleurs dans la grande majorité des systèmes. Le fait de ne pas prendre en compte les deux traits spécifiques de l'oral est apparemment assez peu pénalisant. Mais ces résultats seraient certainement à nuancer si nous avions distingué les différents sous-corpus, qui varient quant à leur degré d'interactivité, lors de l'apprentissage. Notamment, la distance en tours de parole entre deux mentions successives est sans doute différente d'un sous-corpus à un autre, et utiliser ce trait en mélangeant les sous-corpus restreint son impact.

Le modèle finalement intégré au système de résolution est donc celui calculé par SVM sur l'ensemble complet d'attributs et le corpus d'apprentissage *small_trainingSet*, puisque la moyenne de ses quatre métriques obtenues lors de la phase de développement est la plus haute (71,9). Les résultats de ce système sur l'ensemble de test forment une moyenne de 73,4 et sont présentés plus en détails en fin de tableau 1.

De manière générale, il est extrêmement délicat de comparer nos résultats à ceux déjà connus, tant les données varient : la langue des corpus, la modalité (oral vs écrit), le codage des coréférences spécifiques au corpus, les traits disponibles, etc., sont différents. CROC se veut plutôt une base nouvelle, à laquelle pourront se comparer les futurs autres systèmes

de reconnaissance de la coréférence dédiés au français. A titre d'exemple, sur l'énoncé extrait du sous-corpus OTG d'ANCOR présenté en figure 2, l'annotation manuelle relève deux chaînes de coréférence : une première composée des maillons (2, 4, 5, 7, 8, 9) et une seconde composée des maillons (3, 6). Le système CROC détecte quant à lui trois chaînes de coréférence : une première composée des maillons (1, 8, 9), une seconde composée des maillons (2, 4, 5) et une dernière composée des maillons (3, 6). On constate que l'annotation automatique produit des erreurs de plusieurs types : elle intègre le maillon *le numéro de l'office de tourisme d'Espagne* qui fonctionne en fait comme un singleton, et le considère comme référent des mentions anaphoriques *les* et *ils*. Cette erreur implique la suivante, puisque la chaîne dont le référent est *l'office de tourisme d'Espagne* perd deux de ses maillons, qui ont été précédemment intégrés à une fausse chaîne. De plus, on observe que le maillon *leur* n'est pas considéré par le système, il s'agit donc ici d'une omission.

— je peux donner [*le numéro de [l'office de tourisme d'Espagne à Paris*₃]₂]₁ **qui**₄ vous enverront tout ce que vous désirez **c'**₅ est à **Paris**₆ vous **leur**₇ écrivez vous **les**₈ appelez **ils**₉ vous envoient tout

FIGURE 2 – Exemple d'énoncé avec annotation des maillons

Ces premières observations nous amènent à envisager de mettre en place une métrique supplémentaire pour évaluer la qualité de l'annotation en chaînes de coréférence, qui prendrait en compte le nombre d'opérations à effectuer sur les chaînes automatiques pour parvenir à la véritable annotation : il s'agirait d'une sorte de calcul de distance d'édition qui relèverait les différentes substitutions, insertions et délétions de maillons dans les différentes chaînes de coréférence.

4 Conclusion et perspectives

Depuis que la tâche de résolution de la coréférence occupe une place importante dans les problématiques de traitement automatique des langues, beaucoup de travaux se sont attachés à développer des systèmes de détection de chaînes de coréférence pour l'anglais. Très peu cependant, mis à part celui décrit dans (Longo, 2013), ont étudié le phénomène et ses pistes de résolution automatique sur le français. De fait ce travail présente l'un de ces premiers systèmes appris automatiquement sur un corpus annoté. En plus de cette évolution de langue, c'est une distinction de canal qui caractérise ce travail, puisqu'à ce jour aucun modèle de résolution fondé sur l'apprentissage supervisé n'avait été développé spécifiquement pour l'oral transcrit.

Les résultats d'évaluation obtenus par notre modèle de résolution, proches de ceux de certains systèmes de l'état de l'art, suppose toutefois que le texte sur lequel on l'applique est déjà annoté en mentions et que certains attributs de ces mentions (genre, nombre, caractère nouveau ou non de l'entité référencée...) sont disponibles. Ces résultats expérimentaux nous ont néanmoins permis d'observer certaines propriétés du phénomène étudié et d'envisager des pistes de travail pour améliorer les performances du modèle de classification, étendre ses capacités à celles d'un système *end-to-end*, et en compléter l'évaluation. Pour obtenir un tel système *end-to-end* en français, il faudrait coupler CROC avec un étiqueteur POS, un reconnaiseur d'entités nommées et divers autres outils capables d'identifier les genres et nombres des mentions, notamment.

Il est difficile, en l'état actuel de nos expériences, de mesurer l'impact spécifique des différents traits utilisés. Une perspective de ce travail serait de s'attacher précisément à cette sélection d'attributs, par exemple *via* une méthode de sélection ascendante qui évaluerait un modèle appris sur un ensemble ne contenant qu'un trait, puis ajouterait de manière incrémentale un nouveau trait à l'ensemble, en ne le conservant que si les résultats de classification sont meilleurs que pour l'ensemble précédent.

Dans ce travail, la tâche de classification ne permet de ranger les instances que sous deux classes : soit coréférentes, soit non-coréférentes. En procédant ainsi, nous ne distinguons pas les différentes formes de reprises telles qu'annotées dans le corpus ANCOR, et ne prenons pas en compte le fait que chacune d'entre elles est susceptible d'avoir des propriétés qui lui sont propres. Pour l'anaphore pronominale, par exemple, on pourrait supposer que la distance entre les deux mentions d'une paire ne doit pas excéder un certain seuil. Mais ce typage, dans le corpus ANCOR, est toujours réalisé *relativement à la première mention de l'entité*, alors que CROC cherche à identifier les *mentions coréférentes successives*. Une fois la classification des paires de mentions adaptée, on pourra s'inspirer des recherches de (Denis, 2007), qui propose d'apprendre des modèles spécifiques pour chaque type de reprises (anaphore fidèle, infidèle, pronominale et associative).

Références

- BAGGA A. & BALDWIN B. (1998). Entity-based cross-document coreferencing using the vector space model. In *Proceedings of ACL'98*, p. 79–85.
- BENGTSON E. & ROTH D. (2008). Understanding the Value of Features for Coreference Resolution. In *Proceedings of EMNLP 2010*, p. 236–243.
- DENIS P. (2007). *New Learning Models for Robust Reference Resolution*. PhD thesis, University of Texas at Austin.
- DENIS P. & BALDRIDGE J. (2008). Specialized models and ranking for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08*, p. 660–669, Stroudsburg, PA, USA : Association for Computational Linguistics.
- DÉSOYER A., LANDRAGIN F., TELLIER I., LEFEUVRE A. & ANTOINE J.-Y. (2015). Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ancor. *Traitement Automatique des Langues*, **55**(2), 97–121.
- HAGHIGHI A. & KLEIN D. (2010). Coreference resolution in a modular, entity-centered model. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10*, p. 385–393, Stroudsburg, PA, USA : Association for Computational Linguistics.
- LAPPIN S. & LEASS H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, **20**, 535–561.
- LASSALLE E. (2015). *Structured learning with latent trees : A joint approach to coreference resolution*. PhD thesis, Université Paris Diderot.
- LEE H., CHANG A., PEIRSMAN Y., CHAMBERS N., SURDEANU M. & JURAFSKY D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, **39**(4), 885–916.
- LONGO L. (2013). *Vers des moteurs de recherche intelligents : un outil de détection automatique de thèmes*. PhD thesis, Université de Strasbourg.
- LUO X. (2005). On Coreference Resolution Performance Metrics. In *Proceedings of Human Language Technology - Empirical Methods in Natural Language Processing (EMNLP 2005)*.
- LUO X., ITTYCHERIAH A., JING H., KAMBHATLA N. & ROUKOS S. (2004). A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, ACL '04*, Stroudsburg, PA, USA : Association for Computational Linguistics.
- MITKOV R. (2002). *Anaphora resolution*. Longman.
- NG V. & CARDIE C. (2002). Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of ACL'02*, p. 104–111.
- RECASENS M. (2010). *Coreference : Theory, Resolution, Annotation and Evaluation*. PhD thesis, University of Barcelona.
- SOON W. M., NG H. T. & LIM D. C. Y. (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, **27**(4), 521–544.
- STOYANOV V., CARDIE C., GILBERT N., RILOFF E., BUTTLER D. & HYSOM D. (2010). *Reconcile : A Coreference Resolution Research Platform*. Rapport interne.
- STOYANOV V., GILBERT N., CARDIE C. & RILOFF E. (2009). Conundrums in noun phrase coreference resolution : Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2 - Volume 2, ACL '09*, p. 656–664, Stroudsburg, PA, USA : Association for Computational Linguistics.
- TROUILLEUX F. (2001). *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français*. PhD thesis, Université Blaise Pascal.
- YANG X., SU J., LANG J., TAN C. L., LIU T. & LI S. (2008). An entity-mention model for coreference resolution with inductive logic programming. In *Proc. of ACL'08*, p. 843–851.
- YANG X., ZHOU G., SU J. & TAN C. L. (2003). Coreference resolution using competition learning approach. In *Proceedings of ACL'03*, p. 176–183.

Vers un diagnostic d'ambiguïté des termes candidats d'un texte

Gaël Lejeune, Béatrice Daille

LINA, Université de Nantes, 2 rue de la Houssinière, 44322 Nantes, France
prenom.nom@univ-nantes.fr

Résumé. Les recherches autour de la désambiguïssation sémantique traitent de la question du sens à accorder à différentes occurrences d'un mot ou plus largement d'une unité lexicale. Dans cet article, nous nous intéressons à l'ambiguïté d'un terme en domaine de spécialité. Nous posons les premiers jalons de nos recherches sur une question connexe que nous nommons le diagnostic d'ambiguïté. Cette tâche consiste à décider si une occurrence d'un terme est ou n'est pas ambiguë. Nous mettons en œuvre une approche d'apprentissage supervisée qui exploite un corpus d'articles de sciences humaines rédigés en français dans lequel les termes ambigus ont été détectés par des experts. Le diagnostic s'appuie sur deux types de traits : syntaxiques et positionnels. Nous montrons l'intérêt de la structuration du texte pour établir le diagnostic d'ambiguïté.

Abstract.

Towards diagnosing ambiguity of candidate terms

Researches in the field of Word Sense Disambiguation focus on identifying the precise meaning of a lexical unit found in a text. This article tackles another kind of problem : assessing the ambiguity of a lexical unit. In other words, we try to identify if a particular unit is ambiguous or not, we define this task as ambiguity diagnosis. Our evaluation dataset contains scientific articles where ambiguous words have been tagged by experts. In order to give an ambiguity diagnosis for each term, we use two types of features : POS tags and positions in the text. We show that the position of an occurrence in the text is a strong hint for such a task.

Mots-clés : diagnostic d'ambiguïté, extraction de mot-clés, terminologie.

Keywords: ambiguity diagnosis, keyword extraction, terminology.

1 Introduction

La désambiguïssation sémantique est un verrou important pour le Traitement Automatique des Langues. Ce problème a souvent été abordé dans une perspective de résolution. Étant donné les sens possibles d'une unité lexicale (mot ou groupe de mots) en contexte, il s'agit de déterminer lequel de ces sens est activé pour une occurrence particulière. Ce champ de recherches a été principalement investi à la suite des travaux de Yarowski (1992,1995) bien que les recherches dans le domaine soient bien plus anciennes avec notamment les travaux de Lesk et l'algorithme éponyme (Lesk, 1986). Dans cet article, nous abordons l'ambiguïté sémantique d'un terme, simple ou complexe, en domaine de spécialité. Nous nous intéressons au diagnostic d'ambiguïté, c'est à dire que nous cherchons à déterminer si, dans un contexte particulier, le sens d'un terme est difficile à appréhender. Par exemples, si l'on a le mot « classe » dans un texte relevant de la linguistique, il s'agit de savoir si l'on a un emploi terminologique (biunivoque) ou non (susceptible d'être ambigu). Il peut revêtir son sens général ou servir d'équivalent référentiel pour un terme plus complexe qui serait son expansion (Jacques, 2003).

Pour l'unité lexicale non-ambiguë, e.g. dont le sens est clair, le choix du sens à activer est trivial : si son emploi relève d'un domaine de spécialité, il s'agit d'un cas de monosémie. Autrement dit, le nombre d'inférences à effectuer pour déterminer le sens est minimal pour le récepteur du texte (Sperber & Wilson, 1998; Coursil, 2000; Wilson & Sperber, 2004). Si nous nous replaçons dans le domaine de la désambiguïssation sémantique, cela signifie que parmi tous les sens possibles de l'unité lexicale considérée, c'est le plus terminologique qui doit être activé. Détecter les cas d'emploi terminologique permet donc de guider le processus d'analyse. Nous pensons que le diagnostic d'ambiguïté permet de limiter la combinatoire des sens à explorer. Identifier s'il y a une réelle ambiguïté favorise alors la résolution de cette ambiguïté en permettant de savoir si l'on peut ou non se référer au domaine de spécialité concerné. Ce peut aussi être

un indice pour déterminer quels sont les mots-clés pertinents pour décrire un document. En effet, du point de vue de la terminologie en tant que discipline, les termes d'un document sont non-ambigus.

Pour poser les premiers jalons de nos recherches sur le diagnostic d'ambiguïté, nous exploitons ici un corpus de textes en sciences humaines dont les termes candidats ont été classés selon leur degré d'ambiguïté. Dans ce corpus, nous utiliserons des indices syntaxiques et positionnels pour donner pour chaque terme un diagnostic d'ambiguïté. Nous comparons ce diagnostic automatique avec le jugement humain de manière à évaluer la pertinence des indices choisis. Nous détaillerons la problématique de l'ambiguïté dans la section 2 puis nous décrirons le corpus utilisé pour nos expériences ainsi que notre méthodologie dans la section 3.2. Dans la section 4 nous montrerons nos premiers résultats de nos recherches avant de proposer quelques conclusions et pistes pour des recherches futures (section 5).

2 Problématique de l'ambiguïté

La désambiguïsation sémantique (*Word Sense Disambiguation*) est un champ de recherches très actif dans le domaine du TAL. Cette tâche relève de la classification : il s'agit pour chaque occurrence d'un terme de déterminer le sens le plus approprié parmi tous ceux que ce terme peut revêtir. Ce sens de l'occurrence est une étiquette qui selon les ressources exploitées peut revêtir différentes formes : une définition exprimée en langue naturelle, la position dans une ressource de type ontologie ou encore les traductions possibles de ce terme dans différentes langues. Résoudre l'ambiguïté des termes candidats d'un texte permet par exemple d'améliorer les performances des systèmes de traduction automatique. Pour mesurer l'intérêt de cette classification, nous pouvons également donner en exemple le service *Linguee*¹ qui permet de voir en contexte les différentes acceptions d'une unité lexicale.

D'un point de vue méthodologique, la désambiguïsation sémantique a suivi l'évolution du TAL en général. Les travaux répertoriés les plus anciens (Bar-Hillel, 1960; Wilks, 1975; Small & Rieger, 1982) ont traité la désambiguïsation sémantique comme un problème de sélection que l'on pourrait résoudre à l'aide de systèmes experts. L'approche la plus emblématique du domaine est due à (Lesk, 1986) qui a exploité les premiers dictionnaires électroniques à large couverture pour utiliser les relations entre les définitions pour raffiner les connaissances sémantiques sur chaque mot. Puis, c'est l'apprentissage automatique qui est devenu en vogue (Gale *et al.*, 1992) ce qui a permis d'améliorer considérablement les résultats et d'autoriser l'extension vers de nouveaux domaines et des langues autres que l'anglais. D'autre part, d'autres recherches ont amené de nouvelles problématiques pour le domaine comme la limitation des ressources impliquées (de Loupy & El-Bèze, 2000; Jin *et al.*, 2009) ou l'interprétabilité des modèles générés (Navigli & Velardi, 2005). Les traits exploités dans ces travaux et leurs successeurs sont principalement de deux ordres : classes sémantiques et étiquettes morpho-syntaxiques. Sont considérés les termes à désambiguïser ainsi que leurs voisins selon une certaine fenêtre (n termes avant et/ou après). Une des principales contraintes rencontrées est la largeur de cette fenêtre, plus elle est grande et plus la complexité de calcul est élevée. Avec une fenêtre de taille n et en moyenne m sens par terme à observer, on a une complexité exponentielle en la largeur de la fenêtre. Le choix de cette largeur ne peut donc être qu'un compromis entre efficacité et temps de calcul.

Donner un diagnostic d'ambiguïté permet, par exemple, de limiter le nombre de combinaisons à envisager. Chaque mot non-ambigu permet de réduire la combinatoire pour le calcul du sens de ses voisins ou encore d'élargir à moindre coût le contexte exploré pour améliorer les résultats. Par ailleurs, le diagnostic d'ambiguïté permet d'identifier plus finement les candidats qui sont véritablement des termes pour le document considéré. Pour le terminologue, le mot terminologique est par définition non-ambigu. Diagnostiquer l'ambiguïté revient alors à distinguer en contexte les emplois terminologiques des emplois non-terminologiques. Nous décrivons dans la section suivante, le corpus et la méthode déployée pour aboutir à un diagnostic d'ambiguïté.

3 Description du corpus et de la méthodologie

3.1 Le corpus

Le corpus que nous avons utilisé est constitué de textes scientifiques (articles et communications) relevant des sciences humaines collectés dans le cadre du projet SCIENTEXT². La portion de SCIENTEXT utilisée est composée uniquement

1. <http://www.linguee.org> (consulté le 1er juin 2015)

2. <http://scientext.msg-alpes.fr> (consulté le 1er juin 2015)

de textes en français relevant de la linguistique, de la psychologie des sciences de l'éducation et du traitement automatique des langues. Dans ces textes, des candidats termes ont été identifiés automatiquement en utilisant l'extracteur de termes TERMSUITE, outil librement disponible et *Open Source*³. Chacun des candidats a été évalué par un annotateur humain ce qui a permis d'obtenir 4 classes de candidats (DM signifiant Désambiguïsation Manuelle, les modalités précises d'annotation sont disponibles en ligne⁴) :

DM0 Candidat terme rejeté au niveau syntaxique.

DM1 Candidat terme validé au niveau syntaxique. La validation repose sur des critères propres à chaque discipline.

DM3 Candidat terme validé au niveau disciplinaire. La validation repose sur l'appartenance effective du terme au champ scientifique dont relève les textes.

DM4 Candidat terme validé au niveau terminologique. La validation repose sur un emploi véritablement terminologique dans le contexte du document où l'on retrouve le candidat.

La classe DM4 correspond à un usage purement terminologique, et par conséquent non-ambigu, du terme. Pour chaque document, chaque candidat terme est identifié par une classe parmi les quatre décrites ci-dessus. Les documents utilisés sont disponibles au format XML, les données structurales (sections, paragraphes, listes et légendes) y sont identifiées. Par contre, les informations sur la mise en forme matérielle (graisse, italique...) sont absentes.

3.2 Notre méthode : exploiter la mise en saillance

Nous faisons l'hypothèse que la position des candidats est un indicateur fort de leur ambiguïté. L'idée est que le texte forme un écosystème dans lequel certains candidats termes sont plus mis en valeur que d'autres. C'est une manière pour l'émetteur du texte de faciliter le travail de son lecteur en plaçant ce qui est pertinent pour la compréhension à des positions remarquables. Le nombre de configurations permettant de mettre en valeur les termes est limité (e.g. tout ne peut pas être pertinent). Ceci permet de limiter le nombre d'inférences que doit faire le lecteur pour discriminer ce qui est important de ce qui est secondaire. Ce qui est important doit alors être non-ambigu (Wilson & Sperber, 2004). Nous faisons de plus l'hypothèse que les vrais termes sont globalement « grégaires », c'est à dire que nous les retrouvons souvent ensemble. À l'opposé, ce qui est ambigu est distribué plus uniformément au sein de l'écosystème texte.

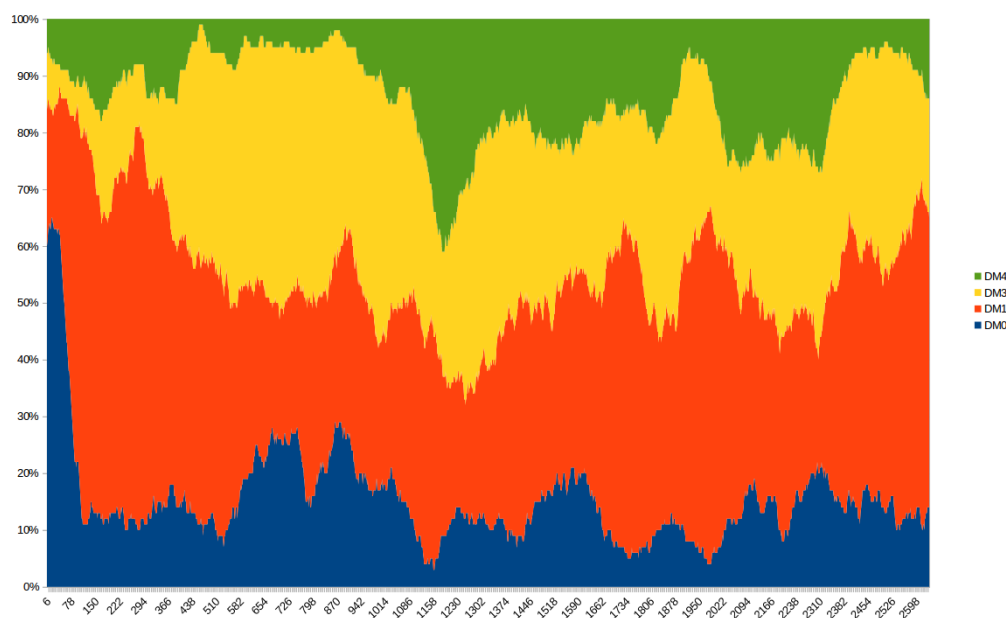


FIGURE 1 – Exemple de distribution des quatre classes au sein d'un texte du corpus

3. <https://logiciels.lina.univ-nantes.fr/redmine/projects> (consulté le 1er juin 2015)

4. <https://apps.atilf.fr/smarties/GuideSmarties.pdf> (consulté le 1er juin 2015)

La figure 1 présente la distribution des candidats termes dans chaque classe au fil d'un des textes du corpus. En ordonnée figure le pourcentage de candidats affecté à chaque classe sur chaque série de 100 candidats. En abscisse figure le début de la série considérée. Ainsi, au point d'abscisse zéro figurent la proportion d'appartenance à chaque classe parmi les 100 premières occurrences dans l'ordre du document. Ici, nous pouvons observer que la proportion de candidats appartenant à la classe DM4 (validés au niveau terminologique) est en augmentation dans différentes zones. Si nous alignons dans cet exemple, les parties avec leurs correspondants dans le modèle Introduction-Matériel-Résultats-Discussion ou IMRAD (Sollaci & Pereira, 2004; Bertin & Atanassova, 2014), cela correspond à :

La fin de l'introduction : de 150 à 366 en abscisse ;

La partie méthode : de 870 à 1878 en abscisse ;

Le cœur des résultats : de 1878 à 2310 en abscisse ;

La fin de la discussion : à partir de 2526 en abscisse.

Ce sont donc des zones où sans être majoritaires, les termes véritables du document sont plus fréquemment présents. Autrement dit, un candidat dont les occurrences seraient régulièrement placées dans ces zones aurait une plus forte « propension terminologique ». Les documents que nous étudions sont de deux types différents : articles et communications. Nous décrivons dans la table 1 un certain nombre de statistiques sur le corpus. Nous observons que dans chacun des deux types de textes scientifiques présents dans le corpus, il y a une relative proximité structurelle ainsi qu'une densité proche en termes non-ambigus (DM4). Parmi les candidats extraits par l'extracteur de termes, les occurrences validées au niveau terminologiques ne représentent qu'une minorité : dans les articles moins d'un candidat sur trois est terminologique. Cette proportion est toutefois plus forte dans les communications. Nous proposons dans la section 4 une première série d'expérience visant à identifier au sein de ce corpus des indices positionnels pour identifier automatiquement ces termes non-ambigus.

	Articles	Communications	Corpus combiné
#textes	11	42	53
#parties	251	509	760
parties /texte, moy.(écart-type)	22,82 (± 10,7)	12,12 (± 4,47)	14,34 (± 7,64)
#paragraphes	1350	2318	3668
paragraphes/texte, moy.(écart-type)	122,73 (± 49,93)	55,19 (± 29,14)	69,21 (± 44,05)
paragraphes/parties, moy.(écart-type)	5,66 (± 2,01)	5,08 (± 3,18)	5,2 (± 2,99)
#mots	99942	171119	271061
mots/texte, moy.(écart-type)	9085,64 (± 3116,82)	4074,26 (± 688,25)	5114,36 (± 2553,84)
mots/parties, moy.(écart-type)	420,03 (± 91,35)	378,95 (± 154,33)	387,44 (± 144,51)
mots/paragraphes, moy.(écart-type)	79,16 (± 18,73)	84,23 (± 26,49)	83,18 (± 25,16)
#DM4	1938	5785	7723
DM4/texte, moy.(écart-type)	176,18 (± 23,63)	137,74 (± 36,0)	145,72 (± 37,23)
#occurrences DM4	6384	13415	19799
occurrences DM4/texte, moy.(écart-type)	580,36 (± 187,0)	319,4 (± 115,6)	373,57 (± 170,43)
#candidats	6080	12447	18527
candidats/texte, moy.(écart-type)	552,73 (± 127,3)	296,36 (± 43,91)	349,57 (± 125,3)
candidats/DM4, moy.(écart-type)	3,14 (± 0,67)	2,23 (± 0,39)	2,42 (± 0,59)
#occurrences	22535	31709	54244
occurrences/texte, moy.(écart-type)	2048,64 (± 733,82)	754,98 (± 153,9)	1023,47 (± 637,01)
occurrences/candidat, moy.(écart-type)	3,64 (± 1,0)	2,54 (± 0,34)	2,77 (± 0,7)

TABLE 1 – Statistiques sur différents grains d'analyses disponibles dans le corpus.

4 Résultats

Dans un but exploratoire, nous utilisons ici les indices positionnels de manière brute. L'objectif est de ne pas utiliser de connaissance *a priori* sur le balisage XML exploité. Pour chaque occurrence d'un candidat nous cherchons pour chaque type de balise (en distinguant les fermantes et les ouvrantes), celle qui est la plus proche de celui-ci. Cette proximité est

	Sans étiquettes morpho-syntaxiques								Avec étiquettes morpho-syntaxiques							
	4 classes				2 classes				4 classes				2 classes			
Classifieur	P	R	F ₁	F _{0,5}	P	R	F ₁	F _{0,5}	P	R	F ₁	F _{0,5}	P	R	F ₁	F _{0,5}
Bayésien naïf	0,50	0,52	0,51	0,50	0,50	0,52	0,51	0,50	0,50	0,52	0,51	0,50	0,51	0,52	0,51	0,51
Régression logistique	0,47	0,63	0,54	0,49	0,75	0,42	0,54	0,65	0,57	0,66	0,61	0,59	0,67	0,48	0,56	0,62
Decision Stump	0,99	0,41	0,58	0,77	0,99	0,41	0,58	0,77	0,99	0,41	0,58	0,77	1	0,41	0,58	0,78
Arbre C4.5 (J48)	0,93	0,40	0,56	0,74	1	0,40	0,57	0,77	0,74	0,60	0,66	0,71	0,94	0,54	0,68	0,82
Baseline	0,34	1	0,51	0,39	0,34	1	0,51	0,39	0,34	1	0,51	0,39	0,34	1	0,51	0,39

TABLE 2 – Résultats de la classification pour la classe DM4 (34,44% des candidats) sur notre fichier initial, la *baseline* classe chaque candidat en DM4.

mesurée en caractères et normalisée en la taille du document ⁵. Les principales balises existant dans le corpus exploité sont les suivantes :

text l'intégralité du texte avec **title** son titre et **body** l'ensemble de ses sections (hors résumé) ;

div une section (ou sous-section) avec **head** son titre et **p** ses paragraphes ;

list les listes à puces dont les items sont signalés par **item** ;

keywords la liste des mots-clés attribués par les auteurs ;

ref les appels à références.

Les autres balises rencontrées (encodingDesc, addrLine, editor...) présentent un certain nombre de méta-données, donc de « l'extra-texte ». Ces balises ne sont pas exclues du processus d'apprentissage de manière à rester fidèle à l'objectif de ne pas introduire de connaissance *a priori*.

La stratégie que nous employons vise à situer chaque occurrence candidat dans sa position relative avec les éléments qui structurent l'écosystème texte (e.g. les balises ouvrantes et fermantes). Nous avons utilisé les implémentations de classifieurs disponibles dans l'outil WEKA ⁶. Nous présentons ici des résultats sur la détection des occurrences appartenant à la classe DM4 (cf. Section 3.2) en termes de Rappel, Précision, F₁-mesure et F_{0,5}-mesure (de manière à pénaliser les faux positifs). Les vrais positifs sont ici les candidats étiquetés comme non-ambigus (DM4) par les annotateurs et effectivement classés comme tels par notre méthode.

La table 2 ⁷ présente les résultats des premières expériences menées sur le texte mentionné dans la figure 1. Ce premier test a été mené en effectuant une validation croisée en dix strates. Il s'agit à partir de l'observation d'un seul texte de mesurer si les indices positionnels constituaient de bons traits pour la classification. Nous comparons ces résultats avec ceux obtenus en ajoutant les étiquettes morpho-syntaxiques obtenues à l'issue de la phase de détermination des candidats. Pour les termes complexes, l'étiquette utilisée est la concaténation des étiquettes des éléments lemmatisés qui le composent. Par exemple, le terme complexe « Science du Langage » sera étiqueté « NOM-PRP-NOM ». Nous ne présentons pas ici les résultats obtenus avec les étiquettes seules car ceux-ci sont faibles. Enfin, nous testons deux cas : celui où les 4 classes sont concernées par la phase d'apprentissage et celui où il n'y a que deux classes (DM4 VS le reste). Nous pouvons remarquer d'une part que les traits que nous avons identifiés sont bien adaptés aux arbres de décisions (y compris *Decision Stump* qui n'utilise qu'un règle par classe) et ce d'autant plus que les informations morpho-syntaxiques sont également présentes.

Nous avons réparti les 10 articles scientifiques restants de manière à disposer d'un corpus d'apprentissage de 9 articles et nous avons gardé le dernier pour constituer le jeu de test. La table 3 présente les résultats obtenus sur ce jeu de test à partir du modèle appris. Nous pouvons observer que la régression logistique et le classifieur bayésien ont une nouvelle fois des résultats équilibrés entre rappel et précision. Nous pouvons voir que les arbres de décisions offrent toujours les meilleurs résultats notamment pour ce qui est de la précision. Toutefois l'arbre *Decision Stump* obtient une précision excellente au prix d'un rappel très faible. En comparant avec les résultats présentés dans le tableau 2, nous pouvons remarquer un cas patent de surapprentissage. À l'opposé, l'arbre de décision J48 est plus robuste et offre un équilibre plus intéressant. Ainsi, nous obtenons un premier résultat intéressant pour l'identification des emplois terminologiques. De nombreuses règles utilisées par ces arbres valident notre hypothèse initiale : la position dans la structure du document permet de détecter les termes non ambigus avec une forte précision. Pour donner un exemple de règle extraite : la proximité du candidat avec un début de partie et avec un indicateur de référence bibliographique détermine à plus de 90% la non-ambiguïté du candidat. Autrement dit, la probabilité d'emploi terminologique est d'autant plus forte que l'occurrence est proche du début ou de la fin d'une section et d'une référence bibliographique. L'utilisation des étiquettes morpho-syntaxiques accroît de manière significative les résultats.

5. Une distance en mots aurait également pu être exploitée.

6. www.cs.waikato.ac.nz/ml/weka (consulté le 1er juin 2015)

7. Nous avons choisi d'écarter ici les classifieurs dont les résultats étaient les moins significatifs, par exemple les séparateurs à vaste marge (SVM).

	Sans étiquettes morpho-syntactiques								Avec étiquettes morpho-syntactiques							
	4 classes				2 classes				4 classes				2 classes			
Classifieur	P	R	F ₁	F _{0,5}	P	R	F ₁	F _{0,5}	P	R	F ₁	F _{0,5}	P	R	F ₁	F _{0,5}
Bayésien naïf	0,43	0,46	0,45	0,44	0,44	0,45	0,44	0,44	0,49	0,49	0,49	0,49	0,47	0,47	0,47	0,47
Régression Logistique	0,49	0,37	0,42	0,46	0,83	0,24	0,37	0,56	0,50	0,62	0,55	0,52	0,73	0,23	0,35	0,51
Decision Stump	1	0,14	0,25	0,45	1	0,14	0,25	0,45	1	0,14	0,25	0,45	0,82	0,19	0,30	0,49
Arbre C4.5 (J48)	0,55	0,27	0,36	0,45	1	0,16	0,28	0,49	0,55	0,53	0,53	0,55	0,75	0,74	0,71	0,75
Baseline	0,32	1	0,48	0,37	0,32	1	0,48	0,37	0,32	1	0,48	0,37	0,32	1	0,48	0,37

TABLE 3 – Résultats de la classification pour la classe DM4 (32,3% des candidats) sur le corpus de test, la *baseline* classe chaque candidat en DM4.

Nous avons appliqué les modèles appris sur les articles scientifiques à la seconde partie de notre corpus, composée uniquement de communications. Les résultats obtenus figurent dans la table 4. Nous observons que le changement de sous-genre scientifique affecte fortement les performances des arbres de décisions. À l’opposé, le classifieur bayésien naïf obtient des résultats plus réguliers. Par ailleurs, la plus-value obtenue en ajoutant les traits syntaxiques est moins nette sur ce sous-corpus. Les indices positionnels pertinents diffèrent entre les deux sous-corpus, les conditions de l’emploi terminologique ne sont pas tout à fait les mêmes. Par exemple, les règles exploitant les références bibliographiques sont moins efficaces, à l’opposé, la proximité avec des items de liste est plus significative dans les communications que dans les articles.

	Sans étiquettes morpho-syntactiques								Avec étiquettes morpho-syntactiques							
	4 classes				2 classes				4 classes				2 classes			
Classifieur	P	R	F ₁	F _{0,5}	P	R	F ₁	F _{0,5}	P	R	F ₁	F _{0,5}	P	R	F ₁	F _{0,5}
Bayésien naïf	0,56	0,50	0,53	0,55	0,55	0,52	0,54	0,54	0,62	0,50	0,56	0,59	0,58	0,52	0,55	0,57
Régression logistique	0,51	0,15	0,23	0,35	0,43	0,49	0,46	0,44	0,55	0,24	0,33	0,44	0,43	0,49	0,46	0,44
Decision Stump	1	0,18	0,31	0,52	1	0,17	0,28	0,51	0,80	0,30	0,44	0,60	0,86	0,18	0,30	0,49
Arbre C4.5 (J48)	0,67	0,23	0,34	0,48	0,98	0,15	0,26	0,47	0,68	0,26	0,38	0,51	0,85	0,21	0,33	0,53
Baseline	0,39	1	0,56	0,44	0,39	1	0,56	0,44	0,39	1	0,56	0,44	0,39	1	0,56	0,44

TABLE 4 – Résultats de la classification pour la classe DM4 (39,37% des candidats) sur le corpus de communications scientifiques, la *baseline* classe chaque candidat en DM4.

5 Conclusion et perspectives

Nous avons présenté dans cet article quelques pistes pour diagnostiquer l’ambiguïté de candidats termes dans des textes scientifiques. Nous avons fait l’hypothèse que les emplois terminologiques étaient repérables par leur présence à des positions remarquables (ou saillantes) dans les documents. Nous avons défini la saillance comme une mesure proximité vis-à-vis des balises de structure présentes dans les documents XML que nous avons étudié. Nous avons ainsi obtenu un profil des termes non-ambigus que nous avons pu projeter sur de nouveaux documents. En combinaison avec les patrons syntaxiques, ceci nous a permis d’obtenir des premiers diagnostics assez prometteurs en particulier pour détecter avec confiance une certaine proportion des emplois terminologiques. Nous avons identifié des différences significatives entre les règles efficaces pour les articles et les règles efficaces pour les communications. Certains phénomènes observés (proximité avec des débuts et des fins de section par exemple) sont réguliers dans les deux genres tandis que d’autres traduisent de véritables différences entre les deux sous-genres. Il s’agit par exemple de la récurrence des emplois terminologiques dans les items de liste qui est plus marquée dans le sous-genre des communications. Cette première étude devra être approfondie pour mieux combiner les indices positionnels et les indices concernant les candidats termes en eux-mêmes. L’utilisation des lemmes ou des formes prises par chaque terme pourrait ainsi permettre d’améliorer le diagnostic.

Pour approfondir ce diagnostic, il serait intéressant de l’évaluer en fonction d’une tâche particulière. Ce peut-être la désambiguïsation sémantique, l’extraction de termes ou encore de l’aide à l’écriture (e.g. les termes clés sont ils bien placés dans le texte). Pour aller plus loin, une piste serait de traiter des documents utilisant d’autres jeux de balises. Il s’agirait d’identifier automatiquement en contexte quelles balises déterminent des positions remarquables et quelles balises sont à écarter. Le modèle serait de ce fait plus robuste à la variation en genre que les modèles obtenus dans nos expériences. Ainsi, nous pourrions mettre la méthode décrite ici en liaison avec les travaux portant sur l’utilisation du modèle IMRAD dans les textes scientifiques, notamment dans les communications qui est un sous-genre que l’on pourrait considérer comme moins normé. Enfin, si le genre est une variable importante pour l’efficacité de la méthode, il serait intéressant d’étudier cette fois le même genre mais dans différentes langues.

Références

- BAR-HILLEL Y. (1960). *Automatic Translation of Languages*. Academic press, New York.
- BERTIN M. & ATANASSOVA I. (2014). A Study of Lexical Distribution in Citation Contexts through the IMRaD Standard. In *Bibliometric-enhanced Information Retrieval Workshop at the 36th European Conference on Information Retrieval (ECIR-2014)*, Amsterdam, Netherlands.
- COURSIL J. (2000). *La fonction muette du langage*. Ibis Rouge.
- DE LOUPY C. & EL-BÈZE M. (2000). Using few clues can compensate the small amount of resources available for word sense disambiguation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)* : European Language Resources Association (ELRA).
- GALE W. A., CHURCH K. W. & YAROWSKY D. (1992). Using bilingual materials to develop word sense disambiguation methods.
- JACQUES M.-P. (2003). *Approche en discours de la réduction des termes complexes dans les textes spécialisés*. Doctorat Nouveau Régime, Université Toulouse II Le Mirail, Toulouse. 3.
- JIN P., MCCARTHY D., KOELING R. & CARROLL J. (2009). Estimating and exploiting the entropy of sense distributions. In *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume : Short Papers*, NAACL-Short '09, p. 233–236, Stroudsburg, PA, USA : Association for Computational Linguistics.
- LESK M. (1986). Automatic sense disambiguation using machine readable dictionaries : How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC '86*, p. 24–26, New York, NY, USA : ACM.
- NAVIGLI R. & VELARDI P. (2005). Structural semantic interconnections : A knowledge-based approach to word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**(7), 1075–1086.
- SMALL S. & RIEGER C. (1982). Parsing and comprehending with word experts (a theory and its realization). In W. G. LEHNERT & M. H. RINGLE, Eds., *Strategies for Natural Language Processing*, p. 89–147. Hillsdale, NJ : Erlbaum.
- SOLLACI L. & PEREIRA M. (2004). The introduction, methods, results, and discussion (imrad) structure : a fifty-year survey. *Journal of the Medical Library Association : JMLA*, **92**(3), 364–367.
- SPERBER D. & WILSON D. (1998). *Relevance : Communication and cognition*. Blackwell press, Oxford U.K.
- WILKS Y. (1975). A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence*, p. 53–74.
- WILSON D. & SPERBER D. (2004). *Relevance theory*. Blackwell press, Oxford U.K.
- YAROWSKY D. (1992). Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the International Conference on Computational Linguistics, COLING 1992*, p. 454–460.
- YAROWSKY D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, p. 189–196.

Augmentation d'index par propagation sur un réseau lexical Application aux comptes rendus de radiologie

Lionel Ramadier^{1,2}, Mathieu Lafourcade²,
(2) Imaios, 34000 MONTPELLIER - France
(1) Lirrm, Université Montpellier, France

lionel.ramadier@imaio.com, mathieu.lafourcade@lirrm.fr

Résumé. Les données médicales étant de plus en plus informatisées, le traitement sémantiquement efficace des rapports médicaux est devenu une nécessité. La recherche d'images radiologiques peut être grandement facilitée grâce à l'indexation textuelle des comptes rendus associés. Nous présentons un algorithme d'augmentation d'index de comptes rendus fondé sur la propagation d'activation sur un réseau lexico-sémantique généraliste.

Abstract.

Index augmentation through propagation over a lexical network – application to radiological reports

Medical data being increasingly computerized, semantically effective treatment of medical reports has become a necessity. The search of radiological images can be greatly facilitated through textual indexing of the associated reports. We present here an index enlargement algorithm based on spreading activations over a general lexical-semantic network.

Mots-clés : réseau lexico-sémantique, propagation, indexation, recherche d'information, imagerie médicale

Keywords: lexico-semantic network, propagation, indexation, information retrieval, medical imaging

1 Introduction

L'informatisation des professions de santé et le développement du dossier médical personnalisé (DMP) entraînent une progression rapide du volume d'information numérique. Les systèmes informatiques médicaux permettent de stocker de grandes masses d'informations (dossier médical, résultats d'examens complémentaires, images et comptes rendus radiologiques, par exemple), d'y accéder en vue d'améliorer la prise en charge des patients, de collecter de nouvelles informations ou encore de fournir une aide à la décision pour l'amélioration de la qualité des soins. Un accès simple et efficace à ces documents médicaux est devenu un objectif primordial pour les établissements de santé. La recherche d'information dans le domaine médical fait à l'heure actuelle l'objet de nombreux travaux de recherche ainsi que des campagnes d'évaluation (Voorhees et al, 2011, Diaz-Galiano et al, 2009, L.Goeuriot, 2014). L'indexation efficace des divers compte rendus (opératoires, radiologiques, etc.) est une tâche nécessaire qui permet d'améliorer la recherche d'informations dans un but clinique mais aussi pédagogique. Par exemple, Dinh *et al.*, (2010) ont réalisé une indexation sémantique des dossiers médicaux des patients afin qu'elle serve de support à des processus de recherche d'information. Leur système d'indexation repose sur l'utilisation de MeSH (Medical Subject Headings), implique un traitement de la désambiguïsation, de l'extraction de valeurs cliniques et de la pondération de concepts. Pouliquen, (2002) a aussi réalisé une indexation automatique par reconnaissance et extraction des concepts médicaux. Il a exploité les mots composés et les associations de mots pour convertir une phrase en mots de référence avec l'aide d'un thésaurus médical.

Dans le domaine de l'imagerie médicale, la quantité d'images et de comptes rendus devient de plus en plus importante, ce qui fait de leur accessibilité pour le corps médical un enjeu majeur. En effet, tirer le meilleur parti d'une telle collection d'images radiologiques en identifiant rapidement l'information pertinente suppose qu'elles soient correctement indexées à partir de leurs comptes rendus. Cette tâche d'indexation, afin d'être utile aux praticiens, doit tenir compte des requêtes qu'ils effectuent. Plusieurs auteurs, notamment Hersh *et al.*, (2001) et Huang *et al.*, (2003) ont réalisé une indexation automatique des comptes rendus radiologiques en se fondant sur le métathésaurus UMLS. Pour améliorer leurs résultats (surtout en ce qui concerne la précision) ils ont utilisé une sous-section de la terminologie UMLS. Toujours dans l'optique d'augmenter la précision, Hersh *et al.*, (2001) ont délibérément écarté certaines parties des comptes rendus, en particulier la section *indications* ; ils ont ainsi obtenu un index ne contenant que les termes strictement médicaux. Or, en pratique, dans leurs requêtes, les utilisateurs d'un système de recherche dédié à la radiologie ont besoin de rechercher non seulement des termes médicaux précis (*perforation digestive*, *glioblastome*) mais aussi des termes composés ou des périphrases de sens général (*accident de ski*, *femmes jeunes*, *coups de couteau*,

coup de sabot). Nous appelons *termes médicaux précis* les termes qui dans le réseau utilisé pour ce travail sont liés par la relation *domaine* à la *médecine*. Les termes généraux ne sont pas directement liés par cette relation à la médecine.

Une grande partie de la complexité de l'extraction automatique d'informations pertinentes à partir de corpus médicaux provient de la forme non structurée de la plupart des textes, de l'écriture informelle (beaucoup d'abréviations, de raccourcis, d'incorrections, etc.), de la quantité d'informations à analyser et de l'identification de leur pertinence. La difficulté d'analyse automatique du sens (en particulier la gestion précise des négations (Huang *et al.*, 2007) et des apocopes, d'identification de termes inconnus (non présents dans la base de connaissances), d'analyse syntaxique de phrases agrammaticales, de reconnaissance d'entités médicales figurant souvent sous une forme d'écriture très dégradée, d'extraction de relations sémantiques présentes dans le texte (Bundschuh *et al.*, 2008), sont autant d'obstacles à une indexation fine de ce genre de documents. Pour réaliser une telle analyse, il est donc crucial de disposer d'un support sémantique non seulement de grande couverture, c'est-à-dire une base de connaissances non réduite aux formes normées, mais également dynamique (i.e. capable d'évoluer et de s'enrichir par apprentissage permanent).

A notre connaissance, l'indexation automatique de comptes rendus radiologiques a jusqu'alors essentiellement porté sur les termes strictement médicaux sans tenir compte des informations d'ordre général. Cependant, Xu *et al.*, (2014) ont réalisé une reconnaissance d'entités nommées de termes anatomiques avec l'aide de ressources externes générales comme Wikipedia et WordNet, en supplément des ressources médicales usuelles, à savoir UMLS, RadLex, MeSH et BodyPart3D (<http://lifesciencedb.jp/bp3d/>). Un autre type de ressource, qui n'avait encore jamais été utilisé dans le cadre médical ou biomédical, permet de prendre en compte non seulement les mots et les concepts du domaine de spécialité, mais également le langage commun couramment utilisé dans les comptes rendus (notamment dans la section *indications*) : il s'agit du réseau lexico-sémantique JeuxDeMots (<http://www.jeuxdemots.org>) que nous utilisons comme base de connaissances support pour l'indexation automatique des comptes rendus radiologiques.

Dans le projet IMAIOS (en collaboration avec des médecins radiologues de Montpellier), afin d'être en mesure d'indexer correctement des comptes rendus de radiologie, nous réalisons non seulement une description des termes et concepts du domaine, mais nous visons également à déterminer les sens (ou les usages) des termes ou des abréviations très fréquentes en médecine. McInnes et Stevenson (2014) ont souligné la difficulté de réaliser cette tâche dans le domaine biomédical, et Ramadier (*et al.*, 2014) cherche à le faciliter à l'aide d'annotations et d'inférences de relations sémantiques. Nous décrivons dans cet article comment à partir de ses informations sémantiques, il est possible de définir une *augmentation des index bruts construits pour chaque compte rendu* afin d'améliorer le rappel de la recherche documentaire. En effet, les médecins radiologues peuvent exprimer leurs requêtes en utilisant des génériques (par exemple, *tumeur bénigne du cerveau*, *tumeur du cerveau*, *tumeur bénigne*, *tumeur*), des conséquences, des circonstances, etc. sans que ces termes soient pour autant explicitement présents dans les comptes rendus. L'article est organisé comme suit : nous présentons en premier lieu le support utilisé pour réaliser cette indexation, c'est-à-dire le réseau lexical JeuxDeMots, puis décrivons la forme précise que peut prendre un index augmenté ainsi que l'algorithme d'augmentation basé sur une propagation au sein du réseau lexical. Enfin nous discutons des expérimentations et analysons les résultats.

2 Augmentation d'index et propagation

La base de connaissances sur laquelle s'appuie notre stratégie d'indexation des comptes rendus médicaux est le réseau lexical JeuxDeMots (Lafourcade 2007). Bien que généraliste, le réseau JDM contient un grand nombre de données de spécialités, notamment des données de médecine/radiologie introduites dans le cadre du projet IMAIOS. Ce réseau sert de base à un algorithme de propagation visant à augmenter un index brut obtenu par des moyens classiques en recherche d'informations.

2.1 Le réseau JeuxDeMots

Le réseau JDM est un graphe lexico-sémantique pour le français obtenu via des jeux - des GWAP, voir (Lafourcade *et al.* 2015) - et un outil contributif nommé Diko. Au moment de l'écriture de cet article, le réseau JDM contient près de 20 millions de relations entre 490 000 termes. Les propriétés de ce réseau qui sont d'intérêt ici sont les suivantes :

- il existe environ 80 types de relations différentes. Celles qui nous intéressent sont les relations essentiellement sémantiques, comme l'hyponymie, les caractéristiques typiques, les lieux typiques, les constituants typiques, le domaine (de spécialité), etc. ;
- les termes polysémiques sont associés (via la relation raffinement) à leurs différents usages. Environ 9 000 termes sont ainsi raffinés en plus de 25 000 usages, comme par exemple, fracture → fracture (lésion), fracture (rupture), fracture (sociologie). Le terme entre parenthèse est une *glose* qui permet de savoir (ou de deviner) de quel sens (*raffinement*) il s'agit ;
- les relations sont pondérées, le poids traduisant la force d'association entre les termes. Environ 70 000 relations ont des poids négatifs, qui indiquent une relation fautive (mais intéressante à conserver, car pouvant aider à la désambiguïsation lexicale) comme par exemple : *fracture du tibia *hyperonyme* (< 0) fracture (sociologie) ;

- il existe une relation d'inhibition qui à un terme t , associe un usage d'un autre terme dont un sens frère est en rapport avec t . Par exemple : fracture → fracture talus (pente), talus (imprimerie), talus (remblai), astragale (architecture), astragale (botanique), ... Au moins un des autres sens des termes associés est en rapport avec *fracture* : ou talus (os) ou astragale (os).

fracture du tibia Nom, Nom féminin singulier Informations diverses wiki polarité

Associations d'idées fracture ► tibia ► fracture (lésion) ► jambe ► plâtre ► traumatisme ► fracture spiroïde ► lésion ► médecine ► os ► lésion physique ► lésion osseuse ► chute ► os (squelette) ► ostéosynthèse ► accident ► traumatisme (physique) ► avoir mal ► plâtre (médecine) ► blessure ► jambe (membre) ► cassé ► fissure (médecine) ► blessé ► douleur (physique) ► douleur ► clou centro-médullaire ► fracture de Segond ► blessure sportive ► blessure (lésion physique) ► fracture du plateau tibial ► fracture ouverte ► Médecine ► médecine (science) ► orthopédie ► traumatologie ► radiologie ► fracture du tibia ► fracture ► tibia ► médecine ► médecine (science) ► Médecine ► radiologie ► orthopédie ► traumatologie ► lésion physique ► fracture (lésion) ► lésion ► lésion osseuse

Est souvent accompagné par ► fracture du péroné ► fracture de la fibula

Thèmes/domaines ► médecine (science) ► médecine ► traumatologie ► radiologie ► orthopédie ► Médecine

Génériques H ► fracture (lésion) ► fracture ► lésion osseuse [1] ► lésion physique ► lésion [1] ► * fracture (sociologie)

Symptôme(s) ► déformation (médecine) ► déformation ► douleur (physique) ► douleur ► **Diagnostic(s)** ► scanner (médecine, technique) ► scanner (médecine) ► radiographie (cliché) ► radiographie ►

Plus intense que fracture du tibia ► fracture double ► double fracture **Moins intense que fracture du tibia** ► foulure ► entorse ►

Locutions/termes composés ► tibia ► fracture ► fracture du ► fracture (lésion)

Caractéristiques de fracture du tibia ► fermée ► ouverte [1] ► spiroïde ► nette ► non déplacée [1] ► invalidante ► plâtrée ► grave ► complexe (complicé) ► complexe ► [1] ► diaphysaire ► douloureuse ► [1] ► comminutive [1] ► douloureuse (souffrance) ► * hépatique ► **A quoi fracture du tibia peut-il s'opposer/combattre ?** ► marche (mouvement) ► marche ►

Lieux incluant/contenant fracture du tibia ? ► tibia ► jambe (membre) ► membre inférieur ► jambe ► [1] ► corps ► [1] ► * genou ► * bras ►

Que peut faire fracture du tibia ? (agent) ► faire souffrir ► faire mal ► **Que peut-on faire à/de fracture du tibia ? (patient)** ► réduire ►* ► visualiser ► radiographier ► plâtrer ► opérer ► opérer (chirurgie) ► diagnostiquer

Causes associées à fracture du tibia ► ski (sport) ► ski ► se blesser ► se battre ► sport ► sport (activité physique) ► traumatisme (physique) ► traumatisme ► tomber ► glisser ► coup (choc) ► accident de ski ► accident de moto ► accident de la route ► accident ► activité physique ► blessure sportive ► coup ► chute ► choc ► Sport

Conséquences associées à fracture du tibia ► radio ► radiographie ► soin ► soin (acte médical) ► plâtre (médecine) ► plâtre ► broche (médecine) ► douleur (physique) ► immobilité ► marcher avec des béquilles ► broche ►

Sentiments/émotions associés à fracture du tibia ► colère ► fatalité ► amertume (tristesse) ► contrariété ► malchance ► ennui (contrariété) ► ennui ► mécontentement ► rage ► triste (malheureux) ► tracas ► souffrance ► dépit ► peur ► découragement ► consternation ► calamité ► angoisse (médecine) ► amertume ► culpabilité ► douleur ► déception ► dépendance (assujettissement) ► douleur (physique) ► abattement ► horrible

Rôles agents fracture du tibia ► se faire ► provoquer ► occasionner

Figure 1. Capture d'écran de la page Diko pour l'entrée « fracture du tibia ». Diko est un outil de visualisation et de contribution en ligne pour le réseau lexical JeuxDeMots. On remarquera dans cet exemple que l'entrée contient à la fois des informations médicales précises (voir symptômes, diagnostics etc.) et des associations d'ordre plus général (voir causes, conséquences, etc.).

L'indexation de mots clés dans le domaine médical concerne souvent certains aspects d'une maladie (Andrade, 2000) ou une partie de l'anatomie. Comme le but de cette indexation est la recherche de documents dans un objectif de pratique quotidienne, nous indexons non seulement les termes anatomiques au sens large (*genou, paroi antérieure du côlon, genou du corps calleux...*), les signes cliniques et les maladies, mais aussi des termes du langage commun susceptibles de faire l'objet de requêtes par le radiologue.

Concernant les domaines (de spécialité) pouvant nous intéresser spécifiquement pour le projet IMAIOS, le tableau suivant donne une idée de l'ordre de grandeur relatif à la quantité d'informations dont nous disposons :

terme	nb de liens sortants	nb de liens entrants
médecine	21408	22666
anatomie	10477	11453
radiologie	382	502
accident	741	956
imagerie médicale	541	556

Tableau 1 : Nombres de liens entre termes dans JeuxDeMots pour certains termes clés.

2.2 Indexation standard de comptes rendus

Notre corpus de travail est constitué d'environ 40 000 comptes rendus de radiologie (Exemple 1) englobant les différentes modalités d'imagerie médicale (imagerie par résonance magnétique, scanner, échographie, radiologie conventionnelle, radiologie vasculaire). Ces comptes rendus sont écrits de manière semi-structurée, c'est-à-dire qu'ils sont généralement divisés en quatre parties distinctes (*indications, technique, résultats*, et une *conclusion* optionnelle). Chaque partie est rédigée par le médecin radiologue sous une forme très libre, avec souvent une profusion d'acronymes (*ATCD* pour *antécédent*, *ACR* pour *american college of radiologie*, *tt* pour *traitement* etc.) d'élisions (par exemple, la *communicante antérieure* au lieu de *artère communicante antérieure*), et toutes sortes d'incorrections diverses. Les comptes rendus contiennent une grande quantité d'informations implicites, devant être explicitées si nous souhaitons obtenir une indexation répondant aux besoins des praticiens.

La création de l'index à partir des comptes rendus reste relativement simple. Nous utilisons les méthodes classiques de la recherche documentaire, soit la fréquence des termes (TF) et la fréquence documentaire (DF) pour calculer l>IDF (Inverse Document Frequency). La reconnaissance des termes composés est effectuée en amont par comparaison au contenu de JDM. Un tiret bas remplacera l'espace entre chaque élément d'un terme composé afin de les conserver comme unité autonome lors de l'extraction (*fracture_du_tibia*).

Malgré le filtrage fréquentiel, nous conservons les termes situés au voisinage de la médecine même pour de faibles valeurs du TFR-IDF. Si un mot simple ou composé du texte est présent dans le réseau JDM, et qu'il est lié par la relation domaine au terme *médecine* (voisinage à distance 1) alors il est ajouté à l'index. De même, des termes non médicaux (*accident de moto, prise de drogue, boule de pétanque*) sont également capturés et ajoutés à l'index dès lors qu'ils possèdent une relation avec un terme lui-même lié à *médecine* (voisinage à distance 2) : ainsi *accident de moto* est ajouté car lié à *polytraumatisé* par la relation conséquence et *polytraumatisé* est lui-même lié à *médecine* par la relation domaine.

<p>indications : fracture du tibia droit, chute de ski</p> <p>technique : une série de coupes axiales transverses sur l'ensemble de la cheville droite sans injection de produit de contraste</p> <p>étude : en fenêtres parties molles et osseuses.</p> <p>résultats : fractures diaphysaires spiroïdes à trois fragments principaux du 1/3 distal du tibia et de la fibula avec discret déplacement vers l'avant, sans retrait de refend articulaire. Fractures de la base de M2 et de M3 non articulaire et non déplacée. Fracture articulaire de la partie interne de la base de M1 non déplacée. Atrophie avec dégénérescence marquée des corps musculaires de l'ensemble des loges.</p>	<p>atrophie • cheville • chute • corps musculaire • coupe axiale transverse • dégénérescence • déplacement • fibula • fracture • fracture du tibia • loge • non articulaire • non déplacée • ski • spiroïde • tibia</p>
---	---

Exemple 1 : compte rendu de radiologie typique (à gauche) et index brut (à droite) : liste de termes présents extraits, ordonnée par ordre alphabétique (les pondérations ne sont pas représentées et la liste est simplifiée). Les termes composés sont identifiés pour peu qu'ils soient présents dans le réseau JDM sous une forme connexe.

Par ailleurs, pour chaque terme de l'index brut, il est intéressant d'essayer de déterminer le ou les bons raffinements sémantiques (s'il en possède). Par exemple, dans le compte rendu ci-dessus, les termes *fracture*, *cheville*, *chute* et *loge* sont polysémiques. Les résultats montreront que la détermination des bons raffinements a de l'importance. Ainsi, *l'augmentation est un processus visant à rajouter dans l'index des termes pertinents, mais non présents dans le texte*.

accident de ski • accident de sports d'hiver • atrophie • cheville • cheville>anatomie • chute • chute>tomber • corps musculaire • coupe axiale transverse • dégénérescence • dégénérescence musculaire • déplacement • fibula • fracture • fracture articulaire • fracture des membres inférieurs • fracture multiple • fracture diphysaire • fracture du tibia • fracture non articulaire • fracture non déplacée • fracture spiroïde • fracture avec déplacement • fracture>lésion • imagerie médicale • jambe • lésion • lésion osseuse • loge • loge>anatomie • médecine • non articulaire • non déplacée • péroné • radiologie • ski • spiroïde • sports d'hiver • tibia • traumatisme des membres inférieurs • ...

Exemple 2 : Index augmenté correspondant à l'exemple présenté ci-dessus (termes triés par ordre alphabétique avec en gras les termes ajoutés). Les thématiques générales du texte sont bien identifiées (médecine, imagerie médicale, radiologie). Les termes polysémiques ont été raffinés avec leur usage correct en contexte.

2.3 Algorithme d'augmentation par propagation

Pour constituer l'index augmenté à partir de l'index brut, nous adoptons une stratégie consistant à propager des signaux sur le réseau JDM à partir des termes de l'index brut. L'idée principale est *d'allumer* les termes de l'index brut et de récupérer à leur suite les termes du réseau qui s'allument également.

A chaque cycle, les termes déchargent en parallèle leur activation courante vers leurs voisins. L'activation totale n'est que la mémoire des décharges reçues par un terme sur l'ensemble du processus. Pour les relations à poids négatif ou inhibitrices, l'activation n'est pas ajoutée mais soustraite. Un terme avec une $AC < 0$ ne décharge pas. La séquence itérée est réalisée en parallèle pour tous les termes. On remarquera que la distribution du signal se fait proportionnellement aux logarithmes des poids (et non pas proportionnellement aux poids eux-mêmes).

A l'issue de l'itération (lignes 5 à 7), nous obtenons une liste de termes pondérés que nous ordonnons par poids décroissants. Nous retenons (filtrage) les N termes de poids les plus forts tels que la somme de leur poids représente S% du poids total des termes de cette liste.

Plus précisément, l'algorithme que nous avons mis au point s'énonce informellement comme suit :

```

1  Init : les termes T du réseau sont associés à un couple de valeurs (AC, AT), activation courante et activation totale.
2      pour les termes T appartenant à l'index brut, nous fixons AC = AT = 1. // les T sont les sources d'activation
3      pour tous les autres termes, AC = AT = 0.
4      nous fixons un nombre d'itérations NBI
5      nous répétons NBI fois l'opération suivante :
6          pour chaque terme T du réseau ayant des voisins {t1, ..., tn}
              via une relation de type r de T vers ti de pondération positive wi, nous modifions les AC et AT des ti :
                  
$$AC(t_i) = AC(t_i) + AC(T) \times \frac{\log(w_i)}{\sum_{k=0}^n \log(w_k)}$$

                  
$$AT(t_i) = AT(t_i) + AC(t_i) \quad // \text{ on mémorise ce que reçoit } t_i \text{ dans } AT(t_i)$$

7      AT(T) = 1 // tous les T ont déchargé leur activation, on recharge les T
8  filtrage des termes activés avec pourcentage de surface S ; nous retournons les termes activés restants.
```

Algorithme 1 : calcul d'un index augmenté à partir d'un index brut, à l'aide d'une propagation sur le réseau lexical JDM. Les deux principaux paramètres sont NBI (nombre d'itérations) et S (% de surface retenu pour le filtrage).

Nous n'exploitons pas tous les types de relations disponibles dans JDM, certaines, très lexicales, auraient tendance, dans le cadre de notre application, à dégrader la précision. Les relations que nous utilisons sont les suivantes (avec leur pondération éventuelle, sinon le poids par défaut est 1) : idées associées (poids 1/2), hyperonymes (poids 2), synonymes, caractéristiques typiques, symptômes, diagnostiques, parties/tout, lieux typiques, causes, conséquences, domaine, et fréquemment associé. Dans l'algorithme 1 ci-dessus, par souci de simplification, tous les types de relations ont un poids identique (le poids serait rajouté de part et d'autre de la fraction).

3 Evaluation des index augmentés

Nous avons évalué l'algorithme de propagation de façon statistique par une sélection aléatoire de 200 index augmentés (sur 30 000 calculés). Chaque terme de l'index augmenté a été manuellement évalué comme pertinent ou non. L'évaluation manuelle a été réalisée par des experts en radiologie. Un terme (ou multi termes) est considéré pertinent par un spécialiste lorsqu'il est susceptible de faire l'objet de requêtes. Les couples de valeurs du Tableau 2 sont donc (a) le nombre moyen de termes de l'index augmenté qui ne sont pas dans l'index brut (valeur *nouv*) et (b) le pourcentage moyen de termes pertinents de l'index augmenté (valeur *pert*).

NBI \ S	10 %	20 %	30 %	40 %	50 %
1	22 / 82 %	45 / 80 %	67 / 78 %	93 / 53 %	127 / 38 %
2	31 / 95 %	55 / 92 %	83 / 89 %	211 / 57 %	439 / 41 %
3	48 / 99 %	90 / 97 %	139 / 95 %	356 / 53 %	755 / 34 %
4	111 / 97 %	223 / 92 %	335 / 87 %	747 / 45 %	1259 / 23 %
5	387 / 96 %	774 / 87 %	1161 / 76 %	1671 / 26 %	2089 / 15 %

Tableau 2 : Présentation des valeurs *nouv* (à gauche de chaque colonne) et *pert* (à droite) en fonction des paramètres NBI et S. NBI est le nombre d'itérations effectuées dans le réseau lexical. S est la part retenue de la surface sous la courbe des poids cumulés des termes atteints par l'algorithme de propagation.

En pratique, l'évaluation manuelle de la valeur *pert* n'a besoin d'être réalisée qu'une fois indépendamment des paramètres NBI et S. En effet, il suffit pour un compte rendu de considérer l'ensemble des termes obtenus pour toutes les valeurs possibles des paramètres, puis d'évaluer la pertinence de chaque terme dans l'ordre de leurs poids décroissants. Au bout de 5 termes consécutifs non pertinents, on considère que tout ce qui suit est également non pertinent. La valeur *nouv*, elle, peut être calculée automatiquement. Pour un même nombre d'itérations, plus la surface retenue est grande plus le nombre de termes atteints est important (filtrage faible). C'est-à-dire que le rappel est d'autant plus important, mais en contrepartie la précision a tendance à diminuer (voire s'écroule au-delà de 30%), les termes ajoutés à l'index brut ayant tendance être de moins en moins pertinents. A l'inverse, plus le nombre d'itérations augmente, plus les termes pertinents sont renforcés (ce sont les voisins mutuels des termes de l'index brut). Le réseau

lexical contient des boucles (directes et indirectes) qui agissent comme autant d'auto-renforcements. Le temps de calcul croît considérablement à chaque nouvelle itération, le nombre de termes déchargeant leur activation augmentant très fortement. Pour NBI = 5, la quasi-totalité du réseau est atteinte (si on exclut le filtrage par S), le diamètre étant d'environ 6 (le réseau JDM est de type petit monde). Globalement, la zone qui semble la plus intéressante pour un temps de calcul raisonnable (quelques secondes) correspond à 3 à 4 itérations pour une surface inférieure à 30%.

La totalité des termes ambigus ont été correctement désambiguïsés. Cela signifie que l'index augmenté a systématiquement inclus le bon raffinement quand un raffinement était proposé (ce n'est pas forcément le cas pour des valeurs faibles de NBI et de S). Nous avons recalculé les index augmentés en interdisant les accès aux termes raffinés, et avons constaté une baisse globale d'environ 10% de la valeur de *pert* quelle que soit la configuration (NBI et S). Chercher à sélectionner les sens corrects des termes polysémiques peut donc être réalisé conjointement à la sélection de termes pertinents et aurait même tendance à la favoriser. Enfin, tous les domaines identifiés ont été pertinents. Rajouter les domaines pertinents dans l'index brut avant l'augmentation n'améliore pas significativement les résultats (ni ne les dégrade). Enfin, nous avons également recalculé les index (bruts et augmentés) en n'autorisant l'accès qu'aux termes directement liés au terme médecine (quelle que soit la relation) durant la propagation. Cela s'est traduit par une diminution moyenne de 12% de la valeur *pert*. Il semblerait donc que l'utilisation d'une base de connaissances non limitée au domaine de spécialité améliore grandement la pertinence de l'index produit.

On remarquera que l'ensemble du processus présenté ci-dessus fonctionne de façon thématique sur le texte et sémantique sur le réseau lexical. Il n'y a pas d'analyse sémantique fine, qui impliquerait selon toute vraisemblance une analyse en constituants et en dépendance des comptes rendus. Les cas d'erreurs manifestes (23 termes pour les 200 index soit pour environ 10 000 termes) que nous avons relevés peuvent avoir plusieurs causes :

- défaut d'information dans la base de connaissances (20% des cas d'erreur) ;
- défaut de rôle sémantique, impliquant la nécessité d'une analyse fine (55%) ;
- chimérisme – deux parties distinctes du compte rendu ont fait émerger un terme non pertinent (25%).

Perspectives et conclusion

Notre objectif est d'indexer automatiquement des comptes rendus radiologiques, non seulement avec les termes médicaux mais aussi avec des termes du langage courant susceptibles d'être utilisés dans des requêtes d'utilisateurs, notamment de praticiens hospitaliers. Pour augmenter le rappel sans notablement dégrader la précision, nous ajoutons à l'index brut des termes implicites des comptes rendus, en utilisant comme support la base de connaissances qu'offre le réseau lexico-sémantique JeuxDeMots. A notre connaissance, très peu de travaux prennent en compte des éléments non médicaux présents dans le compte rendu, ou encore effectuent de l'inférence implicite afin de trouver des termes pertinents non présents. Les approches classiques d'augmentation du rappel consistent essentiellement à inclure des termes plus généraux (hyperonymes ou synonymes) à partir d'une ontologie médicale. La présence d'informations de sens commun bonifie les résultats : l'hypothèse selon laquelle la *non séparation des connaissances* (spécialisées et générales) est plus intéressante que l'usage exclusif de celles de spécialité semble se confirmer, au moins dans nos travaux.

Le travail présenté ici reste préliminaire et il manque une évaluation de fond des index sur l'ensemble du corpus. Nos premiers résultats semblent prometteurs, mais pour être réalisée à grande échelle, l'évaluation doit pouvoir être mécanisée. Il serait intéressant de pouvoir comparer nos résultats avec une indexation obtenus à partir d'un métathésaurus comme l'UMLS. Nous pourrions ensuite aller plus loin dans l'analyse de comptes rendus via l'extraction des relations entre les termes du texte à l'aide de celles présentes dans le réseau JDM. L'indexation portera alors non seulement sur les termes mais également sur les relations entre ces termes. A ce moment là, on pourra réaliser une évaluation plus détaillée concernant la recherche d'information.

Un des buts poursuivis dans le projet IMAIOS est aussi de découvrir dans les comptes rendus de nouvelles connaissances permettant d'alimenter le réseau lexical. Nous envisageons également de déduire à partir du corpus des règles d'inférence et de faire ainsi un raisonnement authentique, c'est-à-dire de proposer par déduction et induction de nouvelles informations médicales, voire des diagnostics.

Références

- ANDRADE M. A. & BORK, P. (2000). *Automated extraction of information in molecular biology*. FEBS letters, Elsevier, 476/1, pp. 12–17.
- BUNDSCHUS M., DEJORI M., STETTER M., TRESP V. & KRIEGEL H.-P. (2008). *Extraction of semantic biomedical relations from text using conditional random fields*. BMC bioinformatics, 9:207, 14 p.
- DINH D., TAMINE L. *et al.* (2010). *Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients*. In Conférence francophone en Recherche d'Information et Applications, CORIA 2010, pp. 325–336.
- DÍAZ-GALIANO, Manuel Carlos, MARTÍN-VALDIVIA, Maite Teresa, et UREÑA-LÓPEZ, L.A. Query expansion with a medical ontology to improve a multimodal information retrieval system. *Computers in biology and medicine*, 2009, vol. 39, no 4, p. 396-403.
- GOEURIOT, Lorraine, KELLY, Liadh, et LEVELING, Johannes. An analysis of query difficulty for information retrieval in the medical domain. In : *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014. p. 1007-1010.
- HERSH W., MAILHOT M., ARNOTT-SMITH C. & LOWE H. (2001). *Selective automated indexing of findings and diagnoses in radiology reports*. Journal of biomedical informatics, 34(4), pp. 262–273.
- HUANG Y. & LOWE H. J. (2007). *A novel hybrid approach to automated negation detection in clinical radiology reports*. Journal of the American Medical Informatics Association, 14(3), pp. 304–311.
- HUANG Y., LOWE H. J. & HERSH W. R. (2003). *A pilot study of contextual UMLS indexing to improve the precision of concept-based representation in xml-structured clinical radiology reports*. Journal of the American Medical Informatics Association, 10(6), pp. 580–587.
- LAFOURCADE M. (2007). *Making people play for lexical acquisition with the JeuxDeMots prototype*. In SNLP'07 : 7th international symposium on natural language processing.
- LANGLOTZ C. P. (2006). *Radlex : A new method for indexing online educational material*. Radiographics, 26(6), pp. 1595–1597.
- MCINNES B. T. & STEVENSON M. (2014). *Determining the difficulty of word sense disambiguation*. Journal of biomedical informatics, 47, pp. 83–90.
- POULIQUEN B. (2002). *Indexation de textes médicaux par extraction de concepts, et ses utilisations*. Thèse de doctorat, Faculté de Médecine, Université Rennes 1, juin 2002, 163 p.
- RAMADIER L., ZARROUK M., LAFOURCADE M. & MICHEAU A. (2014). *Annotations et inférences de relations dans un réseau lexico-sémantique : application à la radiologie*. TALN 2014, Marseille, juillet 2014, pp. 103-112.
- RAMOS J. (2003). *Using TF-IDF to Determine Word Relevance in Document Queries*. In Proceedings of the first instructional conference on machine learning., 4 p.
- ROBERTSON S. E. & JONES K. S. (1976). *Relevance weighting of search terms*. Journal of the American Society for Information science, 27(3), pp. 129–146.
- ROBERTSON, S. (2004). "Understanding inverse document frequency: On theoretical arguments for IDF". *Journal of Documentation* 60 (5): pp. 503–520.
- VOORHEES, E. et TONG, R. Overview of the TREC 2011 medical records track. In : *Proc. of TREC*. 2011.
- XU Y., HUA J., NI Z., CHEN Q., FAN Y., ANANIADOU S., ERIC I., CHANG C. & TSUJII J. (2014). *Anatomical entity recognition with a hierarchical framework augmented by external resources*. PloS one, 9(10), e108396.
- ZARROUK M., LAFOURCADE M. & JOUBERT A. (2013). *Inference and reconciliation in a crowdsourced lexical semantic network*. Computación y Sistemas, 17(2), pp. 147–159.

Détection automatique de l'ironie dans les tweets en français

Jihen Karoui^{1,3} Farah Benamara Zitoune¹ Véronique Moriceau² Nathalie Aussenac-Gilles¹
Lamia Hadrach Belguith³

(1) IRIT, CNRS, Université Paul Sabatier, 118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9

(2) LIMSI, CNRS, Université Paris Sud, Rue John von Neumann, 91403 ORSAY CEDEX

(3) MIRACL, Pôle technologique de Sfax, Route de Tunis Km 10 B.P. 242, 3021 SFAX

jihen.karoui@irit.fr, benamara@irit.fr, moriceau@limsi.fr, Nathalie.Aussenac-Gilles@irit.fr,

l.belguith@fsegs.rnu.tn

Résumé. Cet article présente une méthode par apprentissage supervisé pour la détection de l'ironie dans les tweets en français. Un classifieur binaire utilise des traits de l'état de l'art dont les performances sont reconnues, ainsi que de nouveaux traits issus de notre étude de corpus. En particulier, nous nous sommes intéressés à la négation et aux oppositions explicites/implicites entre des expressions d'opinion ayant des polarités différentes. Les résultats obtenus sont encourageants.

Abstract.

Automatic Irony Detection in French tweets.

This paper presents a supervised learning method for irony detection in tweets in French. A binary classifier uses both state of the art features whose efficiency has been empirically proved and new groups of features observed in our corpus. We focused on negation and explicit/implicit oppositions of opinions with different polarities. Results are encouraging.

Mots-clés : Analyse d'opinion, détection de l'ironie, apprentissage supervisé.

Keywords: Opinion analysis, irony detection, supervised learning.

1 Introduction

L'extraction d'opinion dans les textes s'est beaucoup développée pendant la dernière décennie (Liu, 2012), et surtout depuis l'expansion du web social qui permet aux internautes d'émettre des opinions, des émotions ou des évaluations (critiques, etc.). Il existe plusieurs approches pour l'extraction d'opinions allant de représentations par sac de mots à des modèles plus complexes qui traitent de phénomènes dépendant du contexte ou du niveau discursif. Bien que les systèmes actuels obtiennent des résultats relativement bons sur des tâches de classification objective/subjective, l'analyse de polarité (positif ou négatif) doit encore être améliorée pour pouvoir prendre en compte notamment des formes figuratives telles que l'ironie.

L'ironie est un phénomène linguistique complexe largement étudié en philosophie et en linguistique (Grice *et al.*, 1975; Sperber & Wilson, 1981; Utsumi, 1996). Même si les théories diffèrent sur la définition, elles s'accordent sur le fait que l'ironie implique une incongruité entre ce qui est dit et la réalité. Par exemple, dans les tweets ironiques, l'incongruité consiste souvent en l'opposition d'au moins deux propositions P_1 et P_2 qui s'opposent. Elles peuvent être dans le même énoncé (c'est-à-dire explicitement lexicalisées), ou bien l'une est présente et l'autre est implicite. Nous avons défini deux types d'opposition. L'*opposition explicite* peut impliquer une contradiction entre des mots de P_1 et des mots de P_2 qui ont des polarités opposées comme dans (1), ou qui n'ont pas de relation sémantique comme dans (2). L'*opposition explicite* peut aussi provenir d'un contraste positif/négatif explicite entre une proposition subjective P_1 et une situation P_2 qui décrit une activité ou un état indésirable. L'ironie est alors inférée grâce aux connaissances partagées ou aux normes sociales et culturelles : par exemple, (3) suppose que tout le monde s'attend à ce qu'un téléphone sonne assez fort pour être entendu.

(1) J'**adore** quand mon téléphone **tombe en panne** quand j'en ai besoin.

(2) **The Voice** est plus important que **Fukushima** ce soir.

- (3) J'adore quand mon téléphone baisse le son automatiquement.

L'opposition implicite quant à elle, se produit quand il y a une opposition entre une proposition lexicalisée P_1 décrivant un événement ou un état et un contexte pragmatique externe à l'énoncé qui souvent nie P_1 ou son existence. La proposition P_1 peut être soit subjective (cf. (4)) ou objective. Par exemple, dans (5), l'ironie vient du fait que les lecteurs savent que la proposition P_1 (en italique) n'est pas vraie au moment de l'émission du tweet.

- (4) #Hollande est vraiment un bon diplomate #Algérie.

- (5) #Valls a appris la mise sur écoute de #Sarkozy en lisant le journal. *Heureusement qu'il n'est pas Ministre de l'Intérieur.*

Pour détecter l'ironie dans les oppositions explicites et implicites, la plupart des approches existantes utilisent un classifieur binaire (i.e. apprendre si un texte est ironique ou non) avec une variété de traits allant de simples traits de surface (ponctuations, émoticônes, etc.) à des traits comme la polarité, la synonymie ou le contexte émotionnel (Reyes & Rosso, 2014; Barbieri & Saggion, 2014). Une analyse de notre corpus de tweets en français (nous le présentons dans la section suivante) montre que plus de 62 % des tweets contiennent des opérateurs de négation ("ne...pas") ou des quantificateurs de négation ("jamais", "personne"). Ainsi, la négation nous semble être un indice important dans les énoncés ironiques. Nous faisons donc l'hypothèse que la présence d'une négation ou d'une opposition implicite ou explicite peut aider à la détection automatique de l'ironie.

Dans les sections suivantes, nous présentons notre corpus de tweets puis nous présentons le classifieur binaire utilisé pour la détection de l'ironie dans les tweets. Enfin, nous présentons les expériences menées et les résultats. Finalement, nous concluons en présentant quelques pistes de travaux futurs.

2 Corpus

Un de nos objectifs est de tester si la négation est un bon indice pour détecter les tweets ironiques. Pour cela, nous avons constitué un corpus de tweets ironiques et non ironiques contenant ou non des mots de négation tels que *ne*, *n'*, *pas*, *non*, *ni*, *sans*, *plus*, *jamais*, *rien*, *aucun(e)*, *personne*. Nous considérons comme ironiques les tweets contenant les hashtags *#ironie* ou *#sarcasme*, les autres sont considérés comme non ironiques.

Pour collecter les tweets, nous avons dans un premier temps sélectionné un ensemble de thèmes discutés dans les médias au printemps 2014. Nous avons choisi 184 thèmes répartis en 9 catégories (politique, sport, musique, etc.). Pour chaque thème, nous avons sélectionné un ensemble de mots-clés avec et sans hashtag, par exemple : politique (Sarkozy, Hollande, UMP, ...), santé (cancer, grippe), sport (#Zlatan, #FIFAworldcup, ...), médias sociaux (#Facebook, Skype, MSN), artistes (Rihanna, Beyoncé, ...), télévision (TheVoice, XFactor), pays ou villes (Cordée du Nord, Brésil, ...), Printemps Arabe (Marzouki, Ben Ali, ...) et d'autres thèmes plus génériques (pollution, racisme). Nous avons ensuite sélectionné des tweets ironiques contenant les mots-clés, le hashtag *#ironie* ou *#sarcasme* et un mot de négation ainsi que des tweets ne contenant pas de négation. De la même manière, nous avons aussi sélectionné des tweets non ironiques (i.e. ne contenant pas *#ironie* or *#sarcasme*). Une fois les tweets collectés, nous avons supprimé les doublons, les retweets et les tweets contenant des liens vers du contenu extérieur. Pour les expériences décrites par la suite, les hashtags *#ironie* et *#sarcasme* sont supprimés des tweets. Pour identifier automatiquement les vrais usages de négation (par exemple, *pas* peut être un nom, *une personne* n'est pas une négation), nous avons utilisé l'analyseur syntaxique MELt¹ ainsi que des règles manuelles pour corriger les sorties de l'analyseur si nécessaire.

Au total, nous avons un ensemble de 6742 tweets. Pour mesurer l'effet de la négation sur la tâche de détection de l'ironie, nous avons constitué 3 corpus : les tweets avec négation (*NegOnly*), les tweets sans négation (*NoNeg*), et un corpus regroupant l'ensemble des tweets (*All*). Le tableau 1 montre la répartition des tweets.

Pour s'assurer que les hashtags indiquant l'ironie sont fiables, deux annotateurs ont annoté 3 sous-ensembles : 50 tweets ironiques et 50 non ironiques pour chacun des corpus *All*, *NoNeg* et *NegOnly*. L'accord inter-annotateur (kappa de Cohen) par rapport à la référence (i.e. par rapport aux hashtags) est $\kappa = 0.78$ pour *All*, $\kappa = 0.73$ pour *NoNeg* et $\kappa = 0.43$ pour *NegOnly*. Ces scores montrent que les hashtags *#ironie* and *#sarcasme* sont relativement fiables mais que la présence d'une négation est une cause d'ambiguïté pour la détection de l'ironie par des humains.

1. http://alpage.inria.fr/statgram/frdep/fr_stat_dep_malt.html

Corpus	Ironique	Non ironique	TOTAL
<i>NegOnly</i>	470	3761	4231
<i>NoNeg</i>	1075	1436	2511
<i>All</i>	1545	5197	6742

TABLE 1 – Répartition des tweets dans le corpus.

3 Un classifieur pour la détection de l’ironie

3.1 Traits utilisés

Un tweet est représenté par un vecteur composé de 6 groupes de traits que nous présentons ici. Certains d’entre eux ont été utilisés avec succès pour la détection de l’ironie (dans ce cas, nous citons les références), d’autres sont nouveaux.

Traits de surface : ce sont principalement les traits utilisés dans l’état de l’art. Le premier est la longueur du tweet en nombre de mots (Tsur *et al.*, 2010). Les autres sont tous binaires et indiquent la présence ou non de : ponctuation (Kreuz & Caucci, 2007; Gonzalez-Ibanez *et al.*, 2011), mots en lettres majuscules (Tsur *et al.*, 2010; Reyes *et al.*, 2013), interjections (Gonzalez-Ibanez *et al.*, 2011; Buschmeier *et al.*, 2014), émoticônes (Gonzalez-Ibanez *et al.*, 2011; Buschmeier *et al.*, 2014), citation (Tsur *et al.*, 2010; Reyes *et al.*, 2013), argot (Burfoot & Baldwin, 2009), mots d’opposition tels que “mais” et “bien que” (Utsumi, 2004), séquence de points d’exclamation ou d’interrogation (Carvalho *et al.*, 2009), combinaison de points d’exclamation et d’interrogation (Buschmeier *et al.*, 2014). Nous avons ajouté un nouveau trait qui indique la présence de connecteurs discursifs qui ne déclenchent pas d’opposition (“ainsi, donc, ...”) car nous faisons l’hypothèse que les tweets non ironiques sont susceptibles d’être plus verbeux. Pour implémenter ces traits, nous avons utilisé trois lexiques : un pour les connecteurs discursifs (Roze *et al.*, 2012), un pour l’argot (389 entrées), construit manuellement à partir de diverses sources trouvées sur le web² et le lexique CASOAR (Benamara *et al.*, 2014) pour les interjections (236 entrées) et les émoticônes (595 entrées).

Traits de sentiment : ce sont les traits qui indiquent la présence de mots ou d’expressions d’opinion positive ou négative (Reyes & Rosso, 2011, 2012), leur nombre (Barbieri & Saggion, 2014). Nous avons ajouté 3 nouveaux traits : la présence de mots ou expressions de surprise ou d’étonnement, la présence et le nombre d’opinions neutres. Pour obtenir ces traits, nous avons utilisé deux lexiques :

- CASOAR (Benamara *et al.*, 2014), un lexique pour le français de 2732 mots ou expressions d’opinion catégorisés en 4 catégories sémantiques (REPORTAGE, JUGEMENT, SENTIMENT-APPRECIATION et CONSEIL comme définies dans (Asher *et al.*, 2009)), ainsi que 184 entrées correspondant aux adverbes de doute, affirmation, intensifieur, et adverbes de négation,
- EMOTAIX³, un lexique émotionnel et affectif disponible publiquement de 4921 entrées regroupées en 9 catégories : malveillance, mal-être, anxiété, bienveillance, bien-être, sang-froid, surprise, impassibilité, émotion non spécifique. Il contient 1308 entrées positives, 3078 négatives et 535 neutres.

Traits pour les modifieurs de sentiment : ils regroupent deux nouveaux traits qui indiquent si un tweet contient un mot d’opinion dans la portée d’une modalité ou d’un adverbe d’intensité. Les **traits pour les modifieurs** vérifient aussi si un tweet contient : un intensifieur (Liebrecht *et al.*, 2013; Barbieri & Saggion, 2014), une modalité, un mot de négation ou un verbe de discours rapporté.

Traits d’opposition : ils sont nouveaux par rapport à ceux traditionnellement utilisés. Ils indiquent la présence d’opposition explicite grâce à des patrons lexico-syntaxiques spécifiques. Ces traits ont été partiellement inspirés de (Riloff *et al.*, 2013) qui a proposé une méthode par bootstrapping pour détecter les tweets sarcastiques correspondant à une opposition entre un sentiment/opinion positif et une situation négative. Nous avons donc étendu ce patron afin de traiter d’autres types d’opposition. Par exemple, nos patrons indiquent si un tweet contient (a) une opposition de sentiment/opinion, ou (b) une opposition explicite positive/négative entre une proposition subjective et une proposition objective. Soit P_+ (resp. P_-) une proposition subjective contenant une expression positive (resp. négative), soit P_{obj} une proposition objective ne contenant pas d’expression d’opinion (P_{obj} peut contenir une négation ou non), et soit Neg un opérateur qui change la polarité des mots subjectifs dans P_+ (resp. P_-). Les patrons pour (a) sont de la forme $[Neg(P_+)].[P'_+]$, $[Neg(P_-)].[P'_-]$, $[P_+].[Neg(P'_+)]$, $[P_-].[Neg(P'_-)]$, $[P_-].[P'_+]$, et $[P_+].[P'_-]$;

2. <http://www.linternaute.com/dictionnaire/fr/usage/argot/1/>

3. http://www.tropes.fr/download/EMOTAIX_2012_FR_V1_0.zip

ceux pour (b) sont de la forme : $[Neg(P_+)].[P'_{obj}]$, $[Neg(P_-)].[P'_{obj}]$, $[P_{obj}].[Neg(P'_+)]$, $[P_{obj}].[Neg(P'_-)]$, $[P_+].[P'_{obj}]$, $[P_-].[P'_{obj}]$, $[P'_{obj}].[P_+]$, et $[P'_{obj}].[P_-]$.

Nous considérons qu'un mot d'opinion est dans la portée d'une négation, s'ils sont séparés par au maximum deux tokens (puisque les tweets sont des messages courts limités à 140 caractères).

Traits de contexte : le contexte d'énonciation est important pour comprendre l'ironie d'un énoncé. Ces traits indiquent donc la présence/absence d'éléments de contexte tels que les pronoms personnels, les mots-clés d'un thème donné et les entités nommées donnés par l'analyseur syntaxique. Par exemple, l'ironie dans le tweet *Elle nous avait manqué !* est difficile à détecter car il ne contient pas d'élément contextuel.

3.2 Expériences and résultats

Nous avons testé plusieurs classifieurs sous Weka et avons obtenu les meilleurs résultats avec SMO. Comme nous avons 3 corpus (*NegOnly*, *NoNeg* et *All*), nous avons entraîné 3 classifieurs, un par corpus, notés $C_{NegOnly}$, C_{NoNeg} , et C_{All} . Comme le nombre d'instances ironiques dans *NegOnly* est relativement petit (470 tweets), le classifieur $C_{NegOnly}$ a été entraîné sur un sous-ensemble équilibré de 940 tweets avec une validation croisée sur 10 échantillons. Pour C_{NoNeg} et C_{All} , nous avons utilisé 80% du corpus pour l'apprentissage et 20% pour le test, avec une distribution égale entre les instances ironiques (notées IR) et non ironiques (notées NIR)⁴. Le nombre de tweets non ironiques étant plus grand que le nombre d'ironiques (cf. Tableau 1), nous avons entraîné les classifieurs en fixant l'ensemble de tweets ironiques tout en faisant varier l'ensemble de tweets non ironiques. Les résultats pour ces différentes combinaisons sont relativement similaires. Les résultats présentés ici ont été obtenus en entraînant C_{NoNeg} sur 1720 tweets et en testant sur 430 tweets. C_{All} a été entraîné sur 2472 tweets (1432 contenant une négation –404 IR et 1028 NIR) et testé sur 618 tweets (360 contenant une négation –66 IR et 294 NIR).

Pour chaque classifieur, nous avons étudié l'apport de chaque groupe de traits (cf. Section 3.1) au processus d'apprentissage. Nous avons appliqué à chaque ensemble d'apprentissage un algorithme de sélection de traits (Chi2 et GainRatio), puis avons mesuré l'impact des groupes de traits les plus pertinents sur la tâche de détection de l'ironie. Pour toutes les expériences, nous avons utilisé les traits de surface comme baseline. Pour C_{NoNeg} et $C_{NegOnly}$, le trait qui indique la présence d'une négation a été désactivé. Les résultats en terme d'exactitude sont présentés dans le tableau 2.

	<i>NegOnly</i>	<i>NoNeg</i>	<i>All</i>
Baseline (traits de surface)	73.08	63.25	55.50
Meilleurs traits de surface	73.08	64.65	56.31
Meilleurs traits de sentiment	57.02	67.90	58.25
Modifieurs de sentiment	53.51	56.51	51.94
Modifieurs	53.72	55.81	86.89
Opposition	55.31	63.02	79.77
Contexte interne	55.53	53.25	53.55

TABLE 2 – Résultats des 3 expériences en terme d'exactitude.

Comparé aux autres traits, la baseline obtient de bons résultats sur *NegOnly* alors que les résultats sont beaucoup moins bons que les 2 autres corpus. Pour *NoNeg*, les meilleurs résultats sont obtenus en utilisant les traits {longueur du tweet, interjections, connecteurs discursifs, ponctuations, citations} alors que pour *All*, la meilleure combinaison correspond à {présence de ponctuation, mots en lettres majuscules}. Les principales conclusions que l'on peut tirer du tableau 2 sont : (1) Dans *NegOnly*, les traits sémantiques pris séparément (sentiment, modifieurs, opposition, etc.) ne sont pas suffisants pour classer les tweets NIR et IR. On note en particulier que les résultats de la baseline pour la classe NIR sont meilleurs que ceux pour IR (respectivement 77,60 et 66,40 en F-mesure). (2) Les traits de sentiment sont les plus fiables pour *NoNeg* en utilisant le trait de surprise/étonnement associé aux traits de fréquence des mots d'opinion. Ici aussi, NIR obtient 12,7 points de plus que IR avec une F-mesure de 73,30. (3) Les traits pour les modifieurs et oppositions sont les meilleurs pour *All*. Comme pour les autres corpus, on remarque que les prédictions du classifieur sont meilleures pour la classe NIR que pour IR mais avec un écart moindre (2,2 en utilisant les modifieurs et 7,4 en utilisant les oppositions).

4. Pour C_{NoNeg} et C_{All} , nous avons testé une validation croisée sur 10 échantillons avec une distribution équilibrée entre les instances ironiques et non ironiques mais les résultats sont beaucoup moins bons

Le tableau 3 détaille les résultats globaux quand les classifieurs sont entraînés sur tous les traits pertinents de chaque groupe. Les résultats sont donnés en termes de précision (P), rappel (R), F-mesure (F, macro-moyenne) et exactitude. Les résultats sont meilleurs pour *All* que pour *NegOnly* et *NoNeg*. Ces résultats sont obtenus en utilisant les 3 traits de surface {mots en lettres majuscules, connecteurs d'opposition, longueur du tweet}, les modificateurs {présence d'intensificateurs et négations} et les traits d'opposition {présence d'opposition explicite et implicite}. La meilleure combinaison pour *NegOnly* est composée de 2 traits de surface {mots en lettres majuscules, citation} et du trait d'opposition. Finalement, si on ne considère pas les tweets contenant des négations (i.e. *NoNeg*), les performances tombent à 69,30% d'exactitude. La meilleure combinaison est la suivante : traits de surface {ponctuation, mots en lettres majuscules, interjection, citation, connecteurs discursifs, connecteurs d'opposition, longueur du tweet}, sentiment {(présence de mots d'opinion positifs/négatifs/neutres) et modificateurs de sentiment {mots d'opinion modifiés par un intensificateur ou une modalité}}. Nous pouvons ainsi tirer 4 conclusions : (1) Les traits de surface sont essentiels pour la détection de l'ironie, surtout pour les tweets sans négation, (2) La négation est un trait important pour cette tâche mais ne suffit pas : en effet, parmi les 76 tweets mal classés par *C_{All}*, 60% contiennent des négations (37 IR et 9 NIR), (3) Pour les tweets contenant une négation, les traits d'opposition sont les plus efficaces, (4) Les mots d'opinion sont plus susceptibles d'être utilisés dans les tweets sans négation.

	Ironique (IR)			Non ironique (NIR)		
	P	R	F	P	R	F
<i>C_{NegOnly}</i>	0.889	0.56	0.687	0.679	0.933	0.785
<i>C_{NoNeg}</i>	0.711	0.651	0.68	0.678	0.735	0.705
<i>C_{All}</i>	0.93	0.816	0.869	0.836	0.939	0.884
Résultats (meilleure combinaison)						
	F-score (macro-moyenne)			Exactitude		
<i>C_{NegOnly}</i>	73.60			74.46		
<i>C_{NoNeg}</i>	69.25			69.30		
<i>C_{All}</i>	87.65			87.70		

TABLE 3 – Résultats pour les meilleures combinaisons de traits.

Pour les 3 classifieurs, une analyse d'erreur montre que les erreurs de classification sont principalement dues à 4 facteurs : la présence de comparaison, l'absence de contexte, l'humour ou de mauvais hashtags *#ironie* ou *#sarcasme*. La comparaison est une forme d'ironie par laquelle on attribue des caractéristiques à un élément en le comparant à un élément complètement différent (e.g. "*Benzema en équipe de France c'est comme le dimanche. Il sert à rien*"). Ce type d'ironie utilise souvent des marqueurs de comparaison. Nous ne traitons pas ce phénomène pour le moment mais une approche par similarité sémantique pourrait être utilisée (Veale & Hao, 2010). L'absence de contexte est responsable de la majorité des erreurs. En effet, l'interprétation des tweets mal classés nécessite des connaissances contextuelles extérieures aux tweets. Cette absence de contexte peut se manifester de plusieurs façons : (1) Le thème du tweet n'est pas mentionné (e.g. "*Elle nous avait manqué !*" ou bien l'ironie doit être inférée des hashtags (e.g. *#poissonnavril*) ; (2) L'ironie porte sur une situation spécifique, par exemple un épisode d'une série télé ou une situation géographique ; (3) de fausses assertions comme dans "*Ne vous inquiétez pas. Le Sénégal sera champion du monde de Football*"; (4) Des oppositions qui impliquent une contradiction entre 2 mots qui ne sont pas sémantiquement reliés (e.g. "*ONU*" et "*organization terroriste*", "*Tchad*" et "*élection démocratique*"). Ce cas est plus fréquent dans les tweets sans négation alors que les cas (2) et (3) le sont plus dans les tweets avec négation. Ces résultats sont très encourageants car les travaux qui se sont intéressés à cette même tâche ont atteint des scores de précision de 30% pour le néerlandais (Liebrecht *et al.*, 2013) et 79% (Reyes *et al.*, 2013) pour l'anglais par exemple.

4 Conclusion

Dans cet article, nous avons présenté une approche par apprentissage automatique pour la détection de l'ironie dans les tweets. Nous avons vu que les traits de surface traditionnellement utilisés pour cette tâche dans d'autres langues sont aussi efficaces pour le français. Nous avons introduit de nouveaux traits qui nous ont permis de tester deux hypothèses : la présence de négation et celle d'opposition explicite ou implicite peut aider à détecter l'ironie. Les résultats obtenus sont très encourageants. A court terme, nous prévoyons d'améliorer la classification des tweets pour lesquels le contexte est absent, en exploitant par exemple l'information extra-linguistique.

Remerciements

Ce travail a été financé par le projet ANR ASFALDA ANR-12-CORD-023.

Références

- ASHER N., BENAMARA F. & MATHIEU Y. (2009). Appraisal of Opinion Expressions in Discourse. *Linguisticae Investigationes* 32 :2.
- BARBIERI F. & SAGGION H. (2014). Modelling Irony in Twitter : Feature Analysis and Evaluation. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, p. 4258–4264.
- BENAMARA F., MORICEAU V. & MATHIEU Y. Y. (2014). *TALN-RECITAL 2014 Workshop DEFT 2014 : DÉfi Fouille de Textes (DEFT 2014 Workshop : Text Mining Challenge)*, chapter Catégorisation sémantique fine des expressions d’opinion pour la détection de consensus, p. 36–44. Association pour le Traitement Automatique des Langues.
- BURFOOT C. & BALDWIN C. (2009). Automatic satire detection : Are you having a laugh ? In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, p. 161–164 : Association for Computational Linguistics.
- BUSCHMEIER K., CIMIANO P. & KLINGER R. (2014). An impact analysis of features in a classification approach to irony detection in product reviews. *ACL 2014*, p.42.
- CARVALHO P., SARMENTO L., SILVA M. J. & OLIVEIRA E. D. (2009). Clues for detecting irony in user-generated contents : oh...!! it’s so easy ;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, p. 53–56 : ACM.
- GONZALEZ-IBANEZ R., MURESAN S. & WACHOLDE N. (2011). Identifying sarcasm in Twitter : a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers-Volume 2*, p. 581–586 : Association for Computational Linguistics.
- GRICE H. P., COLE P. & MORGAN J. L. (1975). Syntax and semantics. *Logic and conversation*, **3**, 41–58.
- KREUZ R. J. & CAUCCI G. M. (2007). Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on computational approaches to Figurative Language*, p. 1–4 : Association for Computational Linguistics.
- LIEBRECHT C., KUNNEMAN F. & VAN DEN B. A. (2013). The perfect solution for detecting sarcasm in tweets# not. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* : New Brunswick, NJ : ACL.
- LIU B. (2012). Sentiment Analysis and Opinion Mining (Introduction and Survey). In M. . C. PUBLISHERS, Ed., *Synthesis Lectures on Human Language Technologies*.
- REYES A. & ROSSO P. (2011). Mining subjective knowledge from customer reviews : a specific case of irony detection. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, p. 118–124 : Association for Computational Linguistics.
- REYES A. & ROSSO P. (2012). Making objective decisions from subjective data : Detecting irony in customer reviews. *Decision Support Systems*, **53**(4), 754–760.
- REYES A. & ROSSO P. (2014). On the difficulty of automatically detecting irony : beyond a simple case of negation. *Knowledge and Information Systems*, **40**(3), 595–614.
- REYES A., ROSSO P. & VEALE T. (2013). A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, **47**(1), 239–268.
- RILOFF E., QADIR A., SURVE P., SILVA L. D., GILBERT N. & HUANG R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, p. 704–714.
- ROZE C., DANLOS L. & MULLER P. (2012). Lexconn : A French lexicon of discourse connectives. *Discours, Multi-disciplinary Perspectives on Signalling Text Organisation*, **10**, (on line).
- SPERBER D. & WILSON D. (1981). Irony and the use-mention distinction. *Radical pragmatics*, **49**, 295–318.
- TSUR O., DAVIDOV D. & RAPPOPORT A. (2010). ICWSM-A Great Catchy Name : Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. In *Proceedings of ICWSM*.
- UTSUMI A. (1996). A unified theory of irony and its computational formalization. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, p. 962–967 : Association for Computational Linguistics.
- UTSUMI A. (2004). Stylistic and contextual effects in irony processing. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, p. 1369–1374.
- VEALE T. & HAO Y. (2010). Detecting ironic intent in creative comparisons. In *Proceedings of ECAI*, volume 215, p. 765–770.

Dictionnaires morphologiques du français contemporain : présentation de Morfetik, éléments d'un modèle pour le TAL

Aude Grezka¹ Emmanuel Cartier² Michel Mathieu-Colas¹

(1) LDI UMR 7187, Université Paris 13 Sorbonne Paris Cité

(2) LIPN-RCLN UMR 7030, Université Paris 13 Sorbonne Paris Cité

aude.grezka@ldi.univ-paris13.fr, emmanuel.cartier@lipn.univ-paris13.fr, michel.mathieu-colas@univ-paris13.fr

Résumé. Dans cet article, nous présentons une ressource linguistique, Morfetik, développée au LDI. Après avoir présenté le modèle sous-jacent et spécifié les modalités de sa construction, nous comparons cette ressource avec d'autres ressources du français : le GLAFF, le LEFF, Morphalou et Dicolecte. Nous étudions ensuite la couverture lexicale de ces dictionnaires sur trois corpus, le *Wikipedia* français, la version française de *Wacky* et les dix ans du *Monde*. Nous concluons par un programme de travail permettant de mettre à jour de façon continue la ressource lexicographique du point de vue des formes linguistiques, en connectant la ressource à un corpus continu.

Abstract.

French Contemporary Morphological Dictionaries : Morfetik Database, Elements of a Model for Computational Linguistics

In this article, we present a morphological linguistic resource for Contemporary French called Morfetik. We first detail its composition, features and coverage. We compare it to other available morphological dictionaries for French (GLAFF, LEFF, Morphalou and Dicolecte). We then study its coverage on big corpora (French *Wikipedia*, French version of *Wacky* and *Le Monde* 10 years). We conclude with a proposition for updating the dictionary by connecting the resource with a continuously live corpus.

Mots-clés: dictionnaire, morphologie, français, ressource linguistique, corpus

Keywords: dictionary, morphology, French language, linguistic resource, corpus

1 Morfetik, une ressource morphologique pour le TAL

La ressource lexicale Morfetik, développée au laboratoire LDI, est un dictionnaire morphologique des mots simples du français¹. Nous présentons ici une mise à jour importante de la ressource présentée en 2009 (Buvet *et al.*, 2009 ; Mathieu-Colas *et al.*, 2009).

Le recensement lexical a fait appel à de nombreuses sources lexicographiques. Pour ce qui est de la langue générale, les dictionnaires les plus courants ont été pris en compte, y compris les dictionnaires bilingues : le *DELAS* (Dictionnaire électronique du LADL, cf. B. Courtois, 1990) ; le *Petit Robert* et le *Grand Robert* ; le *Petit Larousse illustré*, le *Lexis*, le *Grand Larousse encyclopédique* et le *Grand Dictionnaire encyclopédique Larousse* (GDEL) ; le *Trésor de la langue française* ; le *Harrap's* et le *Robert & Collins* ; des dictionnaires d'argot ; des tables de conjugaison (dont le *Bescherelle* et les *Verbes logiques* de A. Dugas) ; *Le Bon Usage* de Grevisse et des dictionnaires de « difficultés » pour le traitement des cas problématiques. Pour les termes spécialisés, l'exploration a été largement étendue. Des dictionnaires encyclopédiques ont été consultés : c'est ainsi qu'une partie non négligeable de la nomenclature du GDEL a été intégrée.

1.1 Mises à jour

Au total, 102 962 lemmes (noms, adjectifs, déterminants, pronoms, verbes, adverbes, prépositions, conjonctions, interjections) et 758 035 formes ont ainsi été identifiés. L'inventaire n'est pas clos puisque, actuellement, nous rentrons dans la ressource lexicale :

1/ L'ensemble des propositions des Rectifications orthographiques du français de 1990 (http://www.academie-francaise.fr/sites/academie-francaise.fr/files/rectifications_1990.pdf). Celles-ci ont pour objectif de rectifier l'orthographe de certains mots, sans pour autant constituer une réforme. Elles permettent notamment de lever

¹ Cette ressource est le résultat du travail d'une vingtaine d'années de collecte et de description, sous la direction de Michel Mathieu-Colas.

l'ambiguïté de l'orthographe de certains mots. Ces rectifications touchent entre 2 000 mots d'un dictionnaire d'usage courant qui en contient de 50 000 à 60 000 et plus de 5 000 mots si on prend en compte ceux qui sont rares et techniques. Nous ne prenons en compte pour le moment que les règles orthographiques relatives aux mots simples :

RÈGLES	EXEMPLES	
	ORTHOGRAPHE TRADITIONNELLE	ORTHOGRAPHE RÉFORMÉE
Un certain nombre de mots remplaceront le trait d'union par la soudure, notamment : - les mots composés de <i>contr(e)-</i> et <i>entr(e)-</i> - les mots composés de <i>extra-</i> , <i>infra-</i> , <i>intra-</i> , <i>ultra-</i> - les onomatopées - les mots d'origine étrangère - les mots composés avec des éléments « savants »	<i>contre-appel</i> <i>extra-terrestre</i> <i>tic-tac</i> <i>week-end</i> <i>agro-alimentaire</i>	<i>contrappel</i> <i>extraterrestre</i> <i>tictac</i> <i>weekend</i> <i>agroalimentaire</i>
Pour montrer la prononciation du <i>u</i> , le tréma est, dans les mots comportant : - <i>guë-</i> et - <i>guï-</i> , déplacé sur cette lettre - <i>geure-</i> , ainsi qu'avec le verbe <i>arguer</i> , rajouté à cette lettre	<i>aiguë</i> , <i>ambiguë</i> <i>ambiguïté</i> <i>gageure</i> , <i>arguer</i>	<i>aigüe</i> , <i>ambigüe</i> <i>ambigüité</i> <i>gagüe</i> , <i>argüer</i>
Au lieu de l'accent aigu, emploi de l'accent grave dans un certain nombre de mots et au futur et au conditionnel des verbes qui se conjuguent comme <i>céder</i> .	<i>événement</i> <i>je céderai</i>	<i>évènement</i> <i>je cèderai</i>
L'accent circonflexe disparaît sur <i>i</i> et <i>u</i> , mais est maintenu dans les terminaisons verbales du passé simple (1 ^{ère} et 2 ^e personnes du pluriel), l'imparfait et le plus-que-parfait du subjonctif (3 ^e personne du singulier) et en cas d'homonymie.	<i>coût</i> <i>entraîner</i> , <i>nous entraînons</i> <i>paraître</i> , <i>il paraît</i>	<i>cout</i> <i>entraîner</i> , <i>nous entraînons</i> <i>paraître</i> , <i>il paraît</i>
Les verbes en - <i>eler</i> ou - <i>eter</i> se conjuguent comme <i>peler</i> ou <i>acheter</i> . Les dérivés en - <i>ment</i> suivent les verbes correspondants. Exceptions : <i>appeler</i> , <i>jeter</i> et leurs composés.	<i>j'amoncelle</i> , <i>amoncellement</i> <i>tu époussetteras</i>	<i>j'amoncèle</i> , <i>amoncèlement</i> <i>tu époussèteras</i>
Les mots en - <i>olle</i> et les verbes en - <i>otter</i> (et leurs dérivés) s'écrivent respectivement - <i>ole</i> et - <i>oter</i> . Exceptions : <i>colle</i> , <i>folle</i> , <i>molle</i> et les mots de la même famille qu'un nom en - <i>otte</i> (comme <i>botter</i> , de <i>botte</i>).	<i>corolle</i> <i>frisotter</i> , <i>frisottis</i>	<i>corole</i> <i>frisoter</i> , <i>frisotis</i>
Les mots empruntés forment leur pluriel comme les mots français et sont accentués conformément aux règles qui s'y appliquent. Exceptions : les mots ayant conservé une valeur de citation (comme <i>des mea culpa</i>).	<i>des länder</i> <i>des sandwiches</i> <i>revolver</i>	<i>des lands</i> <i>des sandwiches</i> <i>révolver</i>

TABLEAU 1 : SYNTHÈSE DE LA RÉFORME DE L'ORTHOGRAPHE DE 1990 (MOTS SIMPLES)

Il y a, en outre, plus d'une soixantaine de modifications orthographiques isolées. Ce sont des modifications sur des mots divers : par exemple *charriot* sur le modèle de *charrue*, *boursoufflement* (au lieu de *boursoufflement*), *boursouffler* (au lieu de *boursouffler*), *boursoufflure* (au lieu de *boursoufflure*), *cahutte* (au lieu de *cahute*), etc.

2/ Les mots de la base France Terme (<http://www.culture.fr/franceterme>). Cette base est consacrée aux termes recommandés au *Journal officiel de la République française*. Il regroupe un ensemble de termes de différents domaines scientifiques et techniques mais ne constitue en aucun cas un dictionnaire de langue générale : *édumétrie*, *psychométrie*, *innumérisme*, etc.

3/ Les correspondances masculin-féminin, notamment la féminisation des noms de métier (un ou une *pilote*, un *professeur*/une *professeure*).

4/ Les pluriels sémantiques (une *assise*, les *assises* ; un *ciseau*, des *ciseaux* ; un *échec*, les *échecs* ; un *papier*, les *papiers* ; la *vacance*, les *vacances*).

5/ Le vocabulaire spécialisé : médecine, minéralogie, etc. (*abstension*, *acanthite*...).

Par la suite, nous souhaitons également enrichir la base par les formes verbales composées (choix de l’auxiliaire, identification des verbes pronominaux) et les mots composés.

1.2 Structuration

La structure des tables étant différente selon les catégories morphosyntaxiques, nous avons mis en place cinq groupes distincts. Pour certains types de mots (comme les adverbes), un simple listage suffit. En revanche, pour d’autres catégories (noms, adjectifs et verbes), il convient d’élaborer deux tables complémentaires : (i) des tables de flexion pour identifier et coder tous les types flexionnels ; (ii) des tables attribuant à chaque lemme le code flexionnel correspondant. Ce sont ces tables qui seront ensuite utilisées par le moteur de flexion pour produire l’ensemble de toutes les formes fléchies. Au total, 226 codes de flexion pour les verbes ont été définis, 59 pour les adjectifs et 63 pour les noms.

A titre d’exemple, nous présentons ici les encodages retenus pour les adjectifs. Dans ce cadre, la table des lemmes va comprendre, pour chaque lemme, un identifiant vers son code de flexion. Dans la table des flexions, on trouvera les différentes informations liées à chaque flexion, ainsi que la forme à ajouter. Les 59 codes à genre variable ont été définis, sur le modèle suivant :

Code	Rad	Masculin sing.	Masculin plur.	Féminin sing.	Féminin plur.	Exemples
30	0					albinos, ocre
31	0			e	es	gris
32	1	s	s	ce	ces	tiers
33	1	x	x	ce	ces	doux
34	1	x	x	se	ses	heureux
35	0			se	ses	gros
36	2	ès	ès	esse	esses	exprès
37	1	x	x	sse	sses	faux
38	1	s	s	te	tes	dissous
39	2	is	is	îche	îches	frais
3C	2	ux	ux	ille	illes	vieux
40	0		s		s	démocrate
42	0		S	e	es	petit

TABEAU 2 : EXTRAIT DE LA TABLE DES FLEXIONS POUR LES ADJECTIFS

Ils sont précédés par quelques codes conçus plus spécialement pour les adjectifs à genre fixe :

Code	Rad	Masculin sing.	Masculin plur.	Féminin sing.	Féminin plur.	Exemples
00F	0	NULL	NULL			azygos
00M	0			NULL	NULL	preux
01F	0	NULL	NULL		s	enceinte
01M	0		s	NULL	NULL	extenseur
02M	0		x	NULL	NULL	bijumeau
03M	1	l	ux	NULL	NULL	multicanal
20M	2	an	en	NULL	NULL	gentleman

TABEAU 3 : EXTRAIT DE LA TABLE DES FLEXIONS POUR LES ADJECTIFS

Le champ « Rad » indique le nombre de caractères à enlever pour construire un radical artificiel utilisé par le fléchisseur pour générer les formes fléchies.

2 Couverture lexicale de la ressource

La ressource produite a été comparée avec les ressources lexicales analogues en français. La couverture lexicale a également été validée par comparaison avec trois corpus du français, les 10 ans du *Monde*, le *Wikipedia* français et la version française de *Wacky*.

2.1 Comparaison avec les ressources lexicales en français contemporain

Quatre autres ressources sont disponibles aujourd'hui² : le GLAFF, le Lefff, Morphalou et le Dicolecte. Nous donnons dans le tableau 4 les données principales pour ces différents dictionnaires³.

	MORFETIK		GLAFF		MORPHALOU		LEFFF		DICOLECTE	
	lemmes	formes	lemmes	formes	lemmes	formes	lemmes	formes	lemmes	formes
ADJ	24 391	96 964	42 204	125 409	15 208	47 392	17 416	60 044	11 403	31 859
ADV	1 897	1 897	2 648	2 649	1 579	1 597	3 119	3 143	2 097	2 098
FCTW ⁴	351	483	142	542	352	478	220	459	3 727	3 783
NC	66 393	138 963	104 218	192 386	41 000	80 261	40 109	84 276	44 139	98 532
PREP	57	60	50	56	(FCTW)		128	159	62	62
V	10 223	519 668	21 402	1 085 422	7 207	278 944	7 795	341 528	7 990	334 681
totaux	102 962	758 035	170 664	1 406 464	65 346	408 672	68 787	489 609	69 418	471 015

TABEAU 4 : COMPOSITION DES DIFFÉRENTS DICTIONNAIRES MORPHOLOGIQUES

Le GLAFF (Hathout *et al.*, 2014 ; Sajous *et al.*, 2013, 2014) est un dictionnaire extrait automatiquement à partir du Wiktionnaire français. Il comprend, pour chaque forme, les informations suivantes : la forme graphique, la description morphosyntaxique au format GRACE, le lemme, la ou les prononciation(s) en API et les prononciations équivalentes dans le format SAMPA. Il est à noter que le Wiktionnaire comprend un très grand nombre de gentilés et de lexies spécialisées, ce qui explique le très grand nombre de lemmes et d'entrées. Chaque entrée comprend également sa fréquence dans différents corpus (Wikipedia, LM10 et FrWac).

Le Lefff (Clément *et al.*, 2004 ; Sagot, 2010) se place dans le modèle lexical Alexina, avec pour objectif d'être indépendant des langues spécifiques ainsi que des formalismes syntaxiques ; le format est compatible avec LMF (Francopoulo *et al.*, 2006) qui couvre les niveaux morphologique et syntaxique. Au niveau morphologique, chaque forme comprend son lemme, sa partie du discours et sa classe flexionnelle. Les classes flexionnelles sont définies dans le même esprit que celles de Morfetik. Les données elles-mêmes proviennent de plusieurs sources : récupération automatique (avec validation manuelle) à partir de techniques statistiques sur gros corpus, ainsi que récupération de données provenant d'autres ressources (essentiellement Multext, Veronis, 1998). Les noms propres, initialement intégrés à Lefff, ont ensuite été retirés. C'est la version révisée que nous prenons en compte ici.

Morphalou, développé par l'ATILF, est un lexique des formes fléchies du français construit à partir de la nomenclature du *Trésor de la Langue Française* (539 413 formes fléchies, pour 68 075 lemmes). Le dictionnaire résultant comprend un grand nombre de champs répondant à la norme LMF.

Dicolecte est un dictionnaire construit collaborativement pour les applications Open Office. Il comprend les informations suivantes : forme fléchie, lemme, étiquette grammaticale, métagraphe et métaphone, ainsi que des informations de fréquence dans trois corpus (Google 1-grams, Wikipedia, corpus de littérature issue du site gutenberg.org).

Le tableau 4 montre que, globalement, la couverture lexicale de Morfetik, du point de vue des lemmes comme des formes, est plus importante que celle des trois dictionnaires Morphalou, Lefff et Dicolecte, mais bien moindre que celle de GLAFF. Mais cette différence doit être affinée pour deux raisons principales : d'une part, le GLAFF comprend un grand nombre de formes dont la seule variation est la casse (première lettre en majuscule ou non, exemple : Aïd, aïd) ; d'autre part, un très grand nombre (1 071 327 lexies) ont une fréquence nulle dans les trois corpus que nous étudions, ce qui laisse 335 530 lexies « utiles » et 235 388 formes uniques. Les lexies à valeur nulle sont essentiellement des dérivés de noms propres (gentilés). Notons également que le GLAFF, pour les prépositions, adverbes et autres mots-outils, ne propose pas les listes les plus complètes.

Une autre différence entre les dictionnaires concerne le mode de description des variantes morphologiques : en effet, seuls Morfetik et le Lefff proposent des matrices morphologiques, les autres (Morphalou, DicoLecte, Glaff) se contentant de décrire les différentes formes liées à un lemme. Ces matrices sont particulièrement utiles car elles permettent d'étendre la couverture des dictionnaires de manière dynamique, notamment pour les parties du discours lexicales qui constituent des classes ouvertes. Nous allons voir dans la comparaison des dictionnaires sur corpus que ces matrices permettent une reconnaissance dynamique de formes inconnues, sans avoir à décrire les formes effectives.

Recouvrement lexicographique : les différents dictionnaires ont chacun des spécificités, et il convient à ce point d'étudier le recouvrement des dictionnaires, pour chacune des parties du discours, en partant du principe que les entrées des dictionnaires sont toutes valides. Le tableau 5 compare les trois dictionnaires les plus couvrants et explicites : les entrées communes (intersection), la combinaison des entrées (union), les entrées spécifiques à chaque dictionnaire, et les entrées présentes dans l'un des dictionnaires sauf Morfetik pour les verbes, noms et adjectifs. On constate que : 1/ l'intersection est faible (inférieure à 50% par rapport au dictionnaire le plus couvrant), et corrélativement l'union

² Le Delas fait aussi partie de cette liste, mais la comparaison a déjà été faite dans (Buvet *et al.*, 2009). On consultera (Cougnon et Fairon, 2009) pour une mise à jour de cette ressource.

³ En gras les volumétries les plus importantes.

⁴ (FunctionWord) Correspond à : conjonction, pronom, déterminant, interjection.

améliore significativement la couverture ; 2/ les lexies spécifiques à chaque dictionnaire sont en nombre conséquent, notamment pour les noms et les adjectifs ; 3/ Morfetik : le nombre de lemmes manquants, présents dans au moins l'un des deux autres dictionnaires, est très important, mais il faut analyser ces « manques » ; en effet, parmi les 66 686 noms manquants, 64 153 sont des dérivés par affixation⁵, ce qui laisse 2 545 lemmes manquants ; parmi les 11 877 lemmes verbaux, seul 2 – *voilà* – n'est pas un dérivé ; enfin, parmi les 35 491 lemmes adjectivaux, 6 178 sont des participes passés considérés comme adjectifs (GLAFF) et 29 027 sont des dérivés, ce qui laisse 286 lemmes manquants. Parmi les lemmes manquants, la totalité des lemmes proviennent du GLAFF, dont une très grande majorité sont des emprunts récents, et, de fait, néologiques (exemples : *cokney, sabaoth, mamelouk, glamour...*), des termes techniques ou populaires (exemples : *sextil, tapuscrit, cornecul, pignouf, feu, capout...*), ou encore comportent des erreurs typographiques (*succint, ...*). Somme toute, ces résultats nous semblent d'une part montrer l'intérêt de combiner les différents dictionnaires, et surtout de prévoir, en complément d'un dictionnaire des lexies usuelles, des matrices morphologiques permettant de reconnaître dans les textes des entrées liées à la productivité dérivationnelle. Cela apparaîtra encore plus clairement dans la confrontation des dictionnaires avec des corpus contemporains.

	NOMS	%	VERBES	%	ADJECTIFS	%
Entrées Morfetik	66 393		10 223		24 391	
Entrées Glaff	104 218		21 402		42 204	
Entrées Lefff	40 108		7 795		17 416	
Intersection entre les 3 diction.	31 473		6 856		8 614	
Union des entrées	133 079		22 100		59 882	
Lexies spécifiques Morfetik	21 610	32,55%	380	3,72%	9 106	37,33%
Lexies spécifiques Glaff	62 123	59,61%	10 966	51,24%	27 756	65,77%
Lexies spécifiques Lefff	3 179	7,93%	290	3,72%	7 505	43,09%
Autre dico sauf Morfetik	66 686		11 877		35 491	

TABLEAU 5 : COMPARAISON DES ENTRÉES (LEMES) DES TROIS DICTIONNAIRES LES PLUS COUVRANTS

2.2 Couverture des dictionnaires sur corpus

Pour vérifier la couverture sur corpus, nous avons repris la méthodologie de Sajous (2014), en utilisant trois corpus suffisamment volumineux et représentatifs de la langue générale : version française de Wikipedia (août 2014, 100 millions de mots), version française Wacky⁶ (1 milliard de mots), corpus des 10 ans du *Monde*, 1992-2002 (126 millions de mots). Nous avons étudié la couverture des trois dictionnaires les plus couvrants de la phase précédente (GLAFF, Morfetik, Lefff). Nous avons effectué un prétraitement des corpus afin d'éliminer l'effet « noms propres » (qui représentent près de 50% des lexies et constituent une classe ouverte non couverte par les dictionnaires morphologiques) en remplaçant dans les corpus toute entrée commençant par une majuscule, sauf premier mot de phrase n'ayant aucune autre occurrence commençant par majuscule, par la mention NP, sans en tenir compte dans les comptages. Nous avons également centré l'analyse sur les seules lexies simples.

Le tableau 6 présente les résultats, en effectuant la comparaison, d'une part, en ne considérant que les formes uniques (total formes uniques non reconnues), puis en considérant les occurrences et les fréquences, en faisant varier ce dernier paramètre : Fréquence >0 (toutes les occurrences du corpus), >1 (les formes ayant une fréquence supérieure à 1), etc. Pour chaque paramètre, nous notons le nombre d'occurrences non reconnues (exemple : FR Wikipedia – GLAFF – Fréquence > 100 : 2193 occurrences non reconnues) et le pourcentage par rapport à la totalité des occurrences du corpus. Les lignes COMBI correspondent à un dictionnaire construit par combinaison des entrées des trois dictionnaires.

Corpus	Dictionnaire	Total formes uniques non reconnues	Total occurrences non reconnues (% des formes uniques)					
			Fréquence >0	Fréquence >1	Fréquence >4	Fréquence >9	Fréquence >100	Fréquence >1000
FR Wikipedia (567 429 formes uniques, 99 731 049 occurrences)	Morfetik	419 000 (73,84%)	2 984 594 (2,99%)	166 765 (0,16%)	62 268	30 522	2 506	204
	GLAFF	415 637 (73,24%)	3 147 019 (3,15%)	164 492 (0,16%)	60 636	29 145	2 193	174
	LEFFF	470 496 (82,91%)	19 060 668 (19,11%)	206 639 (0,20%)	92 840	55 523	11 986	2 497
	COMBI	377 862 (66,59%)	1 910 637 (1,91%)	137 653 (0,13%)	45 803	20 596	1 134	88
FrWac (1 606 069 formes uniques, 1 031 810 340 occurrences)	Morfetik	1 378 805 (85,84%)	26 309 234 (2,55%)	657 668 (0,06%)	274 464	152 077	21 606	2 409
	GLAFF	1 367 236 (85,12%)	26 698 724 (2,58%)	651 172 (0,06%)	271 165	149 684	20 025	1 944
	LEFFF	1 474 666 (91,81%)	182 580 835 (17,69%)	739 433 (0,07%)	340 428	207 376	48 482	12 084

⁵ Nous avons calculé le nombre des dérivés en utilisant des listes de préfixes et suffixes productifs en français, en considérant que si un lemme inconnu débutait, se terminait ou débutait et se terminait par l'un d'eux, il s'agissait d'un dérivé.

⁶ <http://wacky.sslmit.unibo.it/doku.php?id=corpora>

	COMBI	1 320 418 (82,21%)	16 582 686 (1,60%)	614 572 (0,05%)	246 326	132 013	16 153	1 432
LM10 (227 262 formes uniques, 126 729 329 occurrences)	Morfetik	91 761 (40,37%)	721 952 (5,70%)	30 511 (0,02%)	10 393	4 924	533	63
	GLAFF	89 051 (39,18%)	6 205 818 (4,89%)	29 259 (0,02%)	9 815	4 557	496	85
	LEFFF	147 411 (64,86%)	26 220 353 (20,69%)	75 429 (0,05%)	44 928	32 917	12 118	2 918
	COMBI	78 761 (34,65%)	352 674 (0,27%)	23 301 (0,01%)	6 866	2 901	248	30

TABLEAU 6 : COUVERTURE DE GLAFF, MORFETIK, LEFF ET COMBINAISON DES TROIS SUR GROS CORPUS

Plusieurs enseignements en découlent :

1/ Le nombre de formes uniques inconnues se situe entre 82% (Lefff) et 73% (Glaff et Morfetik) du vocabulaire dans Wikipedia, entre 91% (Lefff) et 85% (Glaff et Morfetik) pour FrWac et entre 64% (Lefff) et 40% (Glaff et Morfetik) pour LM10 : la couverture du Lefff apparaît donc bien moindre que les deux autres, et Morfetik faisant jeu égal avec le Glaff malgré un lexique bien plus important pour Glaff. La disparité selon les corpus s'explique d'une part par les propriétés des deux premiers corpus, qui comprennent un très grand nombre de termes très spécialisés (Wikipedia essentiellement : *polyolefin*, *furocémide*, etc.), de noms propres sans majuscule initiale, d'erreurs typographiques (*techonopoles*, *accompagnéede*, *respectivement*, etc.), de mots d'origine étrangère (*organizatsiya*, *roommate*,...). Les dix ans du *Monde* sont le corpus le plus « propre » de ce point de vue, mais révèlent également un nombre conséquent de néologismes, principalement par affixation (*supercentres*, *irremplaçabilité*, *autocommémore*, *juridictionnalisation*...), ainsi que des lexies composées dont les composants n'ont pas de valeur autonome (*statu quo*, *stricto sensu*...).

2/ Le nombre d'occurrences inconnues est dès le départ très faible (en dehors du Lefff), entre 3% (FrWikipedia, FrWac) et 5% (LM10), preuve d'une très bonne couverture lexicographique du Glaff et de Morfetik ; on notera que la moins bonne couverture concerne LM10, pourtant réputé comme corpus le plus proche d'un langage courant ; on notera également que si l'on ne considère que les formes ayant une fréquence supérieure à 1, les taux de couverture s'équilibrent (à environ 0,02%) pour tous les dictionnaires. Enfin, si l'on considère les formes inconnues ayant une fréquence supérieure à 10, il s'agit de lexies manquantes qui peuvent être utilement ajoutées aux dictionnaires ; ainsi, par exemple, Morfetik n'a pas pris en compte les abréviations des différentes unités de mesure (*km*, *cl*, ...), les monnaies (*euro*, *yen*...).

3/ Pour chacun des corpus, la combinaison des lexiques conduit à une couverture plus grande, mais cet effet n'est plus visible dès que l'on considère les formes d'une fréquence supérieures à 1.

3 Conclusions et perspectives

Les dictionnaires morphologiques sont utiles pour l'analyse automatique des textes. Nous avons montré que Morfetik est la ressource la plus couvrante parmi les dictionnaires existants et qu'elle soutenait la comparaison avec le Glaff, la ressource collaborative, malgré une couverture certes moins grande, mais qui n'a qu'un effet limité si l'on considère l'exploitation de la ressource dans un système d'analyse des textes. Morfetik et ses mises à jour seront disponibles sous licence LGPL-LR à l'adresse suivante : <http://extranet-ldi.univ-paris13.fr/Morfetik/>

Enfin, pour être utilisé dans un système de TAL, un dictionnaire de formes n'est jamais suffisant, en raison de différents phénomènes discursifs et de la productivité continuelle des langues : un correcteur orthographique, un générateur de formes liées notamment à des matrices d'affixation permettant de rendre compte des dérivés, un traitement spécifique des noms propres et des termes sont ainsi indispensables. De ce point de vue, une étude complémentaire doit être menée afin d'exploiter les matrices morphologiques dont dispose Morfetik.

Les dictionnaires, comme toutes les ressources linguistiques, nécessitent également une confrontation continuelle avec des corpus. Du point de vue des formes linguistiques, cela revient à mettre en regard la ressource linguistique et un corpus continu, afin de suivre l'évolution fréquentielle des lexies, d'une part, de repérer les lexies qui sortent de l'usage (fréquence nulle sur une période), et celles qui semblent s'implanter (néologismes qui atteignent une fréquence suffisante, sur une période donnée). Un dictionnaire morphologique est donc l'un des composants d'un système plus large impliquant un corpus continu et un module néologismes, ainsi que différents outils pour suivre la fréquence d'usage des lexies du dictionnaire.

La combinaison des dictionnaires est également une piste intéressante pour améliorer la couverture sur corpus : nous avons montré l'intérêt d'une telle combinaison, chaque dictionnaire apportant des entrées spécifiques utiles à l'analyse automatique. La consolidation des ressources présentées ici sera prochainement proposée.

Enfin, un dictionnaire des unités linguistiques, pour être efficace en TAL, doit décrire un maximum d'unités polylexicales, même si cela nécessite des mécanismes de description incluant notamment les possibilités d'insertion entre composants, ainsi que des variations morphologiques des composants. Nous passons ainsi du dictionnaire au *construction*.

Références

- BUVET P.-A., CARTIER E., ISSAC F., MATHIEU-COLAS M., MEJRI S., MADIOUNI Y. (2009). Morfetik, ressource lexicale pour le TAL, *TALN 2009*, Senlis, 24-26 juin 2009. <hal.archives-ouvertes.fr/halshs-00739036/>
- CLÉMENT, L., LANG, B., SAGOT, B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 1841–1844, Lisboa, Portugal.
- COUGNON, L.-A., FAIRON, C. (2009). La mise à jour d'un dictionnaire électronique : Une expérience pédagogique liée à la mise à jour du Delaf, *Arena Romanistica*, 28th Conference on Lexis and Grammar, Bergen (29/09/2009-03/10/2009) - Vol. 1, no. 4, p. 58-71.
- COURTOIS B. (1990). Un système de dictionnaires électroniques pour les mots simples du français, *Langue française*, 87, Paris, Larousse, p. 11-22.
- FRANCOPOULO G., MONTE G. (2006). *Lexical Markup Framework* (LMF aka ISO-24613), CD revision 9 : 15 mars 2006.
- HATHOUT N., SAJOUS F., CALDERONE B. (2014). GLÀFF, a Large Versatile French Lexicon. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 1007-1012, Reykjavik, Iceland.
- MATHIEU-COLAS M. (2009). *Morfetik*, une ressource lexicale pour le TAL, *Cahiers de Lexicologie*, Paris, pp. 137-146.
- SAGOT B. (2010). The *Lefff*, a freely available and large-coverage morphological and syntactic lexicon for French, In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 2744-2751, Istanbul, Turkey.
- SAJOUS F., HATHOUT N., CALDERONE B. (2013). GLÀFF, un Gros Lexique À tout Faire du Français. *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN 2013)*, pp. 285-298, Les Sables d'Olonne, France.
- SAJOUS F., HATHOUT N., CALDERONE B. (2014). Ne jetons pas le Wiktionnaire avec l'oripeau du Web ! Études et réalisations fondées sur le dictionnaire collaboratif. *Actes du 4^e Congrès Mondial de Linguistique Française (CMLF 2014)*, pp. 663-680, Berlin, Allemagne.
- VÉRONIS J. (1998). *Multext-Lexicons. A set of Electronic Lexicons for European Languages*. [CD-ROM]: Distributed by ELRA/ELDA.

Une métagrammaire de l'interface morpho-sémantique dans les verbes en arabe

Simon Petitjean, Younes Samih, Timm Lichte
Heinrich-Heine-Universität Düsseldorf, Allemagne
simon.petitjean@hhu.de, samih@phil.hhu.de, lichte@phil.hhu.de

Résumé. Dans cet article, nous présentons une modélisation de la morphologie dérivationnelle de l'arabe utilisant le cadre métagrammatical offert par XMG. Nous démontrons que l'utilisation de racines et patrons abstraits comme morphèmes atomiques sous-spécifiés offre une manière élégante de traiter l'interaction entre morphologie et sémantique.

Abstract.

A metagrammar of the morphology-semantics interface in Arabic verbs

In this article we propose to model the derivational morphology of Arabic using the metagrammatical framework of XMG. We demonstrate that treating abstract roots and patterns as semantically underspecified atomic morphemes offers an elegant way to account for the interaction between morphology and semantics.

Mots-clés : Morphologie, arabe, métagrammaire, frame semantics.

Keywords: Morphology, Arabic, metagrammar, frame semantics.

1 Introduction

Dans cet article, nous présentons une implémentation de la morphologie arabe utilisant le formalisme XMG (pour eX-tensible MetaGrammar). Dans le même temps, nous proposons et décrivons des stratégies générales pour exploiter les capacités des mécanismes de résolution de contraintes pour représenter les propriétés de la morphologie verbale de type 'racine-et-patron' de l'arabe. Bien que nous privilégions ici une approche basée sur le concept de métagrammaire, nous reconnaissons que les approches basées sur les automates finis pour l'analyse et la génération des langues sémitiques, comme celles de (Beesley, 1998) et (Yona & Wintner, 2005), (Shaalán *et al.*, 2012), (Habash & Rambow, 2006), (Kiraz, 2001), ont été fructueuses. Cependant, nous pensons que notre approche a un important avantage sur les méthodes utilisant les automates finis : de par sa modularité, elle permet d'exprimer de manière simple l'interface morphologie-sémantique, ou plus brièvement le problème de l'interface. Dans les automates finis, le travail du transducteur est de traduire des analyses en formes de surface et inversement, sans permettre de façon directe de fournir de l'information sémantique indiquant quel segment contribue à quelles parties ou comment. De nombreux linguistes ont tendance à trouver cette méthode de description lourde, en particulier s'ils considèrent généralement que les systèmes morphologique et sémantique sont interconnectés. Bien que les morphologies non concaténatives constituent probablement le meilleur cas d'étude pour passer des méthodes basées sur les automates finis aux métagrammaires pour traiter l'interface morphologie-sémantique, nous montrerons que cette approche n'est pas sans mérite dans le cas des morphologies hautement concaténatives.

Dans la section 2, nous donnons les principes de la morphologie verbale en arabe. Puis, dans la section 3, nous présentons le cadre de développement que nous utilisons pour sa description. La section 4 donne les détails de la métagrammaire développée pour générer le lexique de formes verbales. Dans la section 5, nous comparons notre approche avec des travaux similaires. Enfin, la dernière section présente la conclusion de cet article et annonce les étapes suivantes de ces travaux.

2 La morphologie verbale en arabe

L'interprétation linguistique standard du processus de formation des mots dans les langages sémitiques décrit les mots comme la combinaison de deux morphèmes : une racine et un patron (parfois appelé schème), pour lesquels le premier effort de génération formelle a été réalisé par (McCarthy, 1981). Les racines sont généralement composées de trois consonnes, parfois quatre, et leur nombre est estimé à 7502, dont 2903 sont fréquemment utilisées (Altabbaa *et al.*, 2010). Un patron peut se présenter comme une séquence de lettres¹, définissant les positions des voyelles relativement aux consonnes de la racine.

Par exemple, les verbes *ma\$YaY*, 'marcha', *ma\$~Y*, 'fit marcher', partagent tous le même morphème racine *m\$y*, 'lié à la marche'.

- | | | | | | |
|-----|------------------------------------|-----------------|-----|-----------------------------|--------|
| (1) | ma\$~Y | Al>abu Alwalada | (2) | ma\$YaY | Al>abu |
| | marcher.CAUSE-PAST le_père le_fils | | | marcher.SIMPLE-PAST le_père | |
| | 'Le père fit marcher l'enfant.' | | | 'Le père marcha.' | |

La deuxième et la troisième colonne du tableau 1 présentent 9 patrons compatibles avec les racines verbales composées de trois consonnes, à l'actif et au passif (il faut également noter que toutes les racines ne sont pas compatibles avec tous les patrons). C_1 , C_2 et C_3 représentent les trois consonnes de la racine.

Patron	Actif	Passif	Sémantique
1	$C_1aC_2aC_3$	$C_1uC_2iC_3$	
2	$C_1aC_2C_2aC_3$	$C_1uC_2C_2iC_3$	Causatif du transitif 1
3	$C_1aaC_2aC_3$	$C_1uuC_2iC_3$	Associatif
4	$\text{ʔ}aC_1C_2aC_3$	$\text{ʔ}uC_1C_2iC_3$	Causatif de 1
5	$taC_1aC_2C_2aC_3$	$tuC_1uC_2C_2ib$	Reflexif de 2 (médiopassif)
6	$taC_1aaC_2aC_3$	$tuC_1uuC_2iC_3$	Reciproque de 3
7	$nC_1aC_2aC_3$	$nC_1uC_2iC_3$	Reflexif / resultatif / passif / médiopassif
8	$C_1taC_2aC_3$	$C_1tuC_2iC_3$	Reflexif / médiopassif
10	$staC_1C_2aC_3$	$stuC_1C_2iC_3$	Requestatif

TABLE 1 – Patrons pour les racines composées de trois consonnes, et sémantiques des patrons verbaux proposées par (Ryding, 2005), extraites de (Danks, 2011)

Nous supposons, comme Doron (2003, 2013); Schneider (2010) parmi d'autres, qu'au moins une partie des patrons est associée à une contribution sémantique sous spécifiée (ces contributions sont présentées dans le tableau 1). Sous cette hypothèse, la combinaison d'une racine et d'un patron mène à la composition de leurs sémantiques. Des incompatibilités entre racines et patrons peuvent en conséquence être motivées par l'incompatibilité de leurs sémantiques respectives. Dans ce travail, nous implémentons cette idée en utilisant des représentations basées sur la théorie des *frames*, dans la tradition de Fillmore (1977) et Barsalou (1992). Dans la mesure où nous les traitons comme des structures de traits typées étendues (Petersen, 2007; Kallmeyer & Osswald, 2013; Lichte & Petitjean, to appear), la composition est vue comme l'unification. Dans la figure 1 nous présentons quelques frames préliminaires pour le deuxième patron, qui introduisent la causalité, et pour la racine *m\$y*. Les frames de type *causation* et *locomotion* sont empruntées à (Kallmeyer & Osswald, 2013).

Notons que dans la figure 1 la frame de la racine est unifiée avec la valeur du trait *EFFECT* de la frame du patron (toutes les deux étiquetées $\boxed{\text{IN}}$). L'unification est donc parfois effectuée entre des sous parties des frames. L'unification des types est déterminée par une hiérarchie de types comme celle de la figure 2. C'est pourquoi l'unification des types *activity* et *locomotion* dans la figure 1 produits *locomotion*, qui est leur sous-type commun le plus spécifique.

3 eXtensible MetaGrammar

eXtensible MetaGrammar, ou XMG (Crabbé *et al.*, 2013)², désigne à la fois un formalisme métagrammatical et l'outil utilisé pour traiter les descriptions reposant sur ce formalisme. L'outil en question, appelé compilateur, permet de générer une ressource linguistique à partir d'une description abstraite et plus compacte de celle-ci (la métagrammaire). Les

1. Nous utilisons dans cet article la méthode de translittération de Buckwalter (Buckwalter, 2004) : www.qamus.org/transliteration.htm.

2. La nouvelle implémentation du compilateur, XMG-2, est disponible librement à l'adresse suivante : <https://sourcesup.cru.fr/xmg>

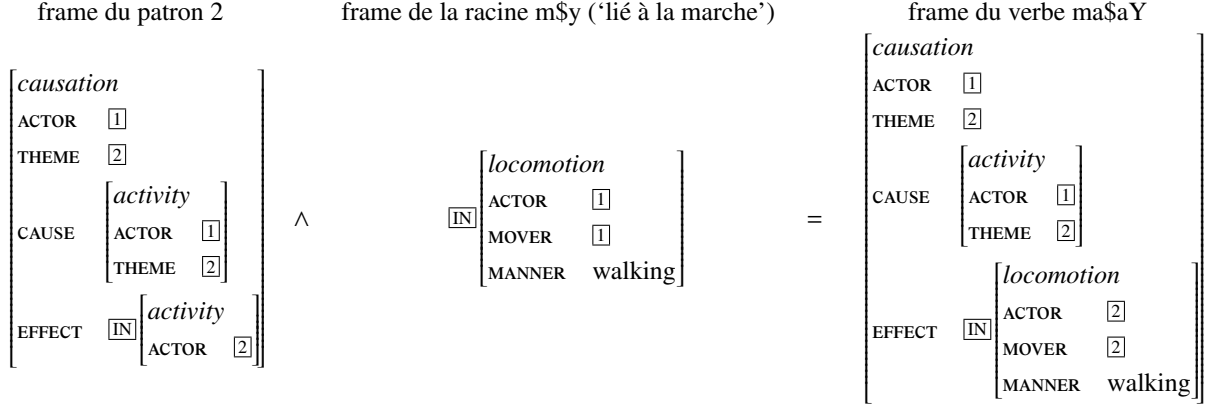


FIGURE 1 – Représentations sémantiques et composition du patron 2 et de la racine m\$y ('lié à la marche')

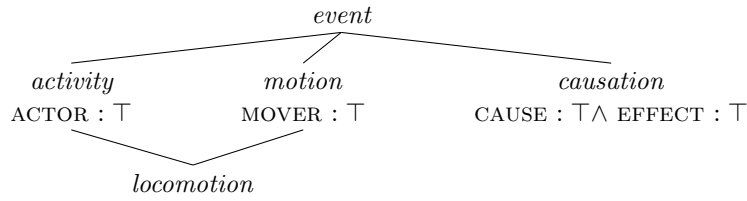


FIGURE 2 – Hiérarchie de types avec contraintes de type extraite de Kallmeyer & Osswald (2013)

descriptions métagrammaticales utilisant ce formalisme sont totalement déclaratives, et peuvent être formalisées de la manière suivante :

$$\begin{aligned}
 \text{Classe} &:= \text{Nom} \rightarrow \text{Contenu} \\
 \text{Contenu} &:= \langle \text{dim} \rangle \{ \text{Contribution} \} \mid \text{Nom} \mid \\
 &\quad \text{Contenu} \vee \text{Contenu} \mid \text{Contenu} \wedge \text{Contenu}
 \end{aligned}$$

La première règle correspond à la notion d'abstraction : une classe permet d'associer un contenu à un identifiant. La seconde règle fait intervenir les trois autres concepts centraux de XMG, ceux de dimension, de conjonction et de disjonction. Ainsi, le contenu d'une abstraction peut être la contribution d'une description, l'utilisation d'une autre abstraction, la combinaison de contenus, ou bien l'expression d'une alternative entre contenus. Les contributions sont effectuées en spécifiant une dimension cible, une dimension étant un accumulateur correspondant à un niveau de description linguistique. Chaque accumulateur étant indépendant des autres, le concept de dimension permet aux métagrammaires utilisant le formalisme XMG de modéliser l'interface entre les niveaux de description linguistique (par exemple l'interface entre syntaxe et sémantique), puisqu'il est possible de partager de l'information explicitement entre dimensions au moyen de variables d'unification.

Si XMG a principalement été utilisé pour décrire la syntaxe des langues, la modularité du compilateur lui permet d'être facilement adapté à de nouvelles tâches de description, par la création de nouveaux modules. La création d'un nouveau compilateur, pour un nouveau langage métagrammatical, est réalisée par un assemblage de ces modules, appelés briques (voir Petitjean 2014). Chaque module définit un langage de description, dédié à de nouvelles structures.

Certains travaux utilisant les nouvelles extensions de XMG ont prouvé que l'approche métagrammaticale pouvait se révéler utile pour la description de différents autres niveaux de description linguistique. Les travaux en question s'intéressent notamment à la représentation de la sémantique au moyen de structures de traits typées (Lichte *et al.*, 2013) et à celle de la morphologie verbale de l'ikota, langue bantoue (Duchier *et al.*, 2012). C'est la dimension morphologique de XMG créée pour cette dernière tâche que nous utilisons pour modéliser la morphologie verbale de l'arabe. Nous formulons donc l'hypothèse que l'outil n'est pas seulement adapté à la description de langues agglutinantes (telles que l'ikota), mais également à celle de langues sémitiques.

4 Métagrammaire de la morphologie verbale de l'arabe

La méthode utilisée pour la description de l'ikota dans (Duchier *et al.*, 2012) consiste à contribuer des morphèmes dans des champs topologiques ordonnés. Le fait que l'ordre de ces champs soit fixe différencie cette tâche de la nôtre. Le nombre et les contraintes sur l'ordre des contributions à la dimension morphologique diffèrent selon la racine et le patron utilisés.

Nos descriptions contiennent donc trois types d'information :

- (i) des contraintes sur le nombre et l'ordre des champs,
- (ii) des instructions affectant un contenu (soit une chaîne de caractères) dans un champ,
- (iii) des informations morphosyntaxiques sous la forme de structures de traits,
- (iv) des descriptions de frames.

En comparaison, la métagrammaire de l'ikota ne contient que les types d'information (ii) et (iii). La métagrammaire peut être vue comme un assemblage de blocs élémentaires (notation empruntée à (Duchier *et al.*, 2012)), chacun de ces blocs pouvant contenir ces quatre types d'information. L'exemple de bloc élémentaire de la figure 3 définit deux champs topologiques nommés C1 et C2, et contraignant leur ordre grâce à l'opérateur de précedence linéaire >>. La deuxième partie du bloc indique la contribution de la lettre /k/ dans le champ C1, et la troisième ajoute un trait morphosyntaxique précisant le patron utilisé dans la dérivation. La quatrième partie contient la description d'une frame très générale de type *activity*. Le langage de description utilisé est celui développé pour XMG dans (Lichte & Petitjean, to appear).

champ C1
champ C2
C1 >> C2
C1 <- k
patron = p1
[activity, actor:?X1]

FIGURE 3 – Exemple de description morpho-sémantique dans XMG

La figure 4 présente la métagrammaire que nous proposons. La classe *Forme* est l'unique axiome de cette métagrammaire, ce qui signifie que ce sont les modèles de cette classe que le compilateur doit calculer. Une *Forme* est obtenue par l'assemblage de quatre abstractions. La première, *Consonnes* est utilisée pour déclarer les trois champs qui contiennent les consonnes de la racine, ainsi que pour insérer ces dernières dans les champs. Les consonnes en question sont contenues dans des variables, les valeurs de ces variables étant obtenues par unification.

Un *Patron* est obtenu en combinant deux blocs élémentaires. Le premier réalise le même travail que le bloc *Consonnes* pour deux voyelles (utilisées dans chaque patron), soit la déclaration des champs, et l'insertion du contenu obtenu par unification. Le second bloc élémentaire est spécifique au patron, et est donc choisi parmi un ensemble de blocs (un par patron). La deuxième de ces alternatives, correspondant au patron 2 ($C_1aC_2C_2aC_3$), déclare dans un premier temps un nouveau champ topologique (C21), pour recevoir la consonne géminée, et ordonne la totalité des champs. Dans un second temps, on place dans le nouveau champ la deuxième consonne de la racine. Enfin, on ajoute l'information que le patron utilisé pour construire cette forme est le deuxième.

L'abstraction *Voix* permet d'exprimer l'alternative entre les blocs élémentaires associés aux voix active et passive. Chacun d'entre eux donne simplement la valeur des deux voyelles utilisées dans les patrons (/a/ et /a/ pour l'actif, /u/ et /i/ pour le passif). Enfin, l'abstraction *Racine* exprime le choix de la racine verbale (ici, par exemple, écrire, étudier et marcher, respectivement /ktb/, /drs/, et /m\$y/).

Le résultat de la combinaison des blocs pour le deuxième patron, la voix active et la racine /m\$y/ est montré dans la figure 5. Cette accumulation, après résolution (c'est à dire ordonnement des champs et concaténation de leurs contenus) produit une forme intermédiaire de l'entrée lexicale *ma\$\$ay*. Pour créer la forme finale *ma\$~Y*, des règles morphophonémiques ultérieures doivent être appliquées, qui ne sont pas montrées ici.

5 Autres travaux

Comme annoncé précédemment, la motivation initiale pour notre travail était de fournir une approche pour la morphologie en arabe plus modulaire que celles utilisant des méthodes basées sur les automates finis. XMG offre un cadre à la fois déclaratif, flexible et multi-dimensionnel. Par conséquent il semble particulièrement adapté à la modélisation de morphologies non-concaténatives et de l'interface morpho-sémantique. Cependant, une limitation cruciale est qu'XMG ne peut jusqu'ici être utilisé que pour la génération.

(Bhuyan & Ahmed, 2008) réalisent l'une des rares propositions pour intégrer une morphologie basée sur les racines et les patrons à une grammaire de précision. Ils proposent d'étendre l'architecture basée sur les traites de HPSG avec un trait

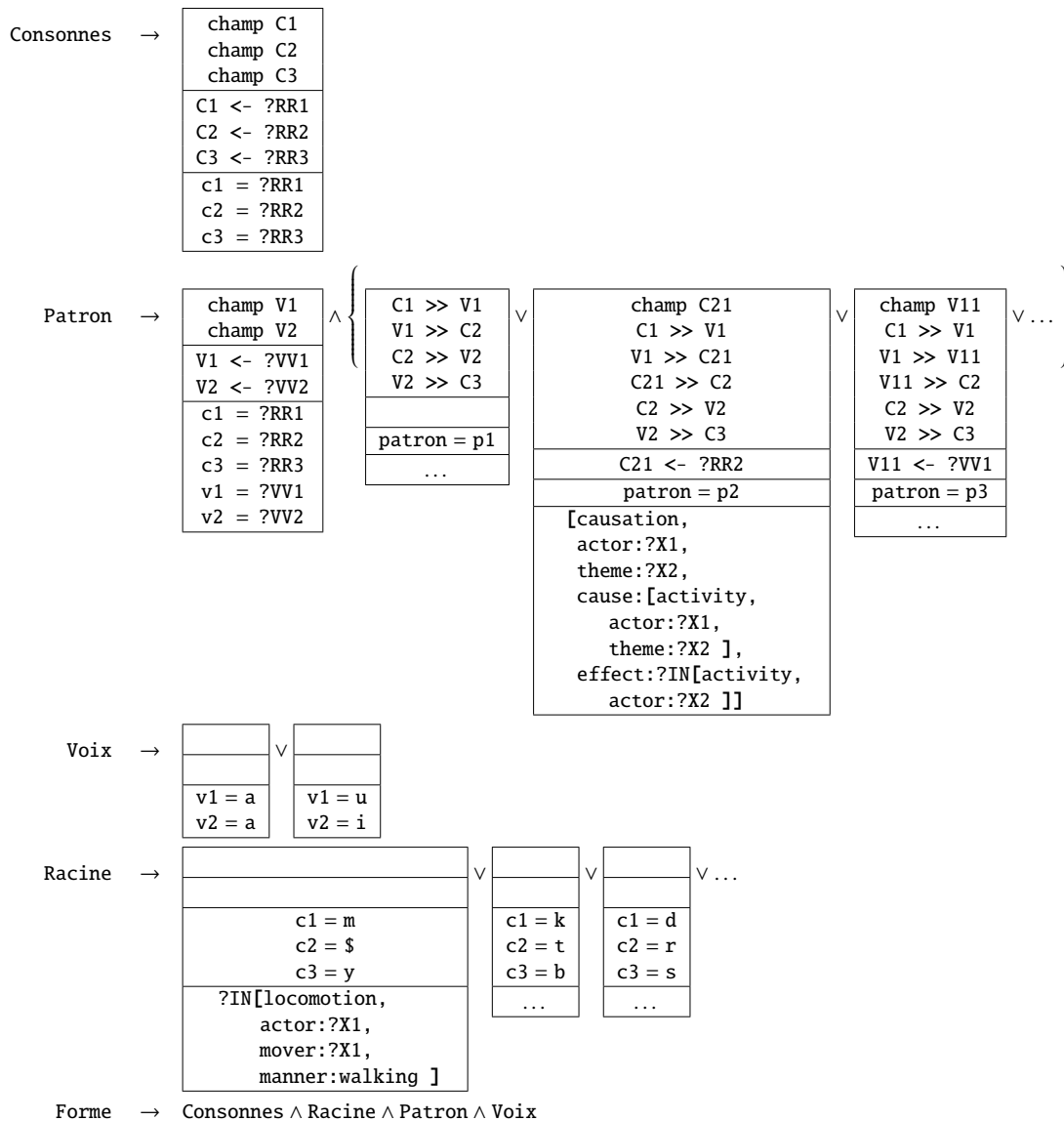


FIGURE 4 – Métagrammaire de la morphologie verbale de l'arabe

MORPH contenant une représentation richement structurée des racines et des patrons. Toutefois, Bhuyan et Ahmed abordent les combinaisons de racines et de patrons sans les décomposer sémantiquement.

Notre approche partage d'avantage de similarités avec des travaux relativement récents sur la morphologie constructionnelle reportés dans (Schneider, 2010). Dans le cadre de l'étude de l'hébreu moderne, Schneider propose d'augmenter la morphologie racine-patron de l'hébreu avec une sémantique compositionnelle, en mettant de côté les problèmes phonologiques. Néanmoins, il utilise les conventions de notation propres à la *Embodied Construction Grammar* (Bergen & Chang, 2005), qui sont très différentes des nôtres, au moins en surface. Un outil d'implémentation et un parser pour les grammaires ECG est disponible,³ mais aucun outil pour la génération ne semble exister.

3. Voir <http://www1.icsi.berkeley.edu/~lucag/>.

champ C1	C1 >> V1
champ C2	V1 >> C21
champ C3	C21 >> C2
champ C21	C2 >> V2
champ V1	V2 >> C3
champ V2	
C1 <- m	V1 <- a
C2 <- \$	V2 <- a
C3 <- y	C21 <- \$
patron = p2	c1 = m
v1 = a	c2 = \$
v2 = a	c3 = y
[causation, actor:?X1, theme:?X2, cause:[activity, actor:?X1, theme:?X2] effect:?IN[activity, actor:?X2]	
?IN[locomotion, actor:?X1, mover:?X1, manner:walking]	

FIGURE 5 – Une accumulation pour la classe *Forme* de la figure 4 incluant le patron 2, la voix active et la racine /m\$/y/, menant à la solution *ma\$\$ay*

6 Conclusion et perspectives

Nous avons présenté une formalisation de la morphologie verbale de l’arabe sous la forme d’une métagrammaire. À partir de cette description, le compilateur XMG génère un lexique de formes verbales non fléchies. Les travaux présentés dans cet article sont la première étape d’un projet plus ambitieux : l’objectif est d’enrichir ce lexique en intégrant l’interface morpho-sémantique. Nous utiliserons pour ceci des structures de traits typées, que nous décrirons au moyen de la dimension sémantique proposée dans (Lichte & Petitjean, to appear). La dimension morphologique de XMG utilise des séquences ordonnées de champs topologiques, l’ordre de ces champs étant défini dans la métagrammaire, ce qui constitue une extension de la dimension utilisée dans (Duchier *et al.*, 2012), apportant la flexibilité nécessaire au traitement de l’arabe. Nous prévoyons à terme d’intégrer des lexiques générés de cette manière à des chaînes de traitement plus importantes, par exemple dans le cadre d’une analyse syntaxique.

Ce travail pourrait évoluer par la suite dans différentes directions. D’un point de vue technique, la dimension morphologique de XMG pourrait être enrichie pour permettre des descriptions plus compactes (par exemple une notation concaténative telle que $C1 \gg V1 \gg C2 \gg V2 \gg C3$ pourrait remplacer un ensemble de contraintes binaires de précedence linéaire) ainsi qu’un opérateur de précedence non immédiate. De plus, une dimension phonologique pourrait être ajoutée pour la prise en compte de règles morphophonémiques. En ce qui concerne la couverture, d’autres racines et patrons doivent être pris en compte, ainsi que l’affixation et l’attachement de clitiques. Enfin, nous pourrions étudier plus en détail les analyses obtenues en utilisant la morphologie constructionnelle, et éventuellement réimplémentées.

Remerciements

Les travaux présentés dans cet article ont été financés par la fondation allemande pour la recherche (Deutsche Forschungsgemeinschaft, DFG), par l’intermédiaire du SFB 991. Nous remercions également les trois relecteurs de TALN pour leurs précieux commentaires.

Références

ALTABBA M., AL-ZARAE A. & ARIF SHUKAIRY M. (2010). *An Arabic Morphological Analyzer and Part-Of-Speech Tagger*. PhD thesis, Arab International University, Damascus, Syria. Thèse.

- BARSALOU L. (1992). Frames, concepts, and conceptual fields. In A. LEHRER & E. F. KITTEY, Eds., *Frames, fields, and contrasts : New essays in semantic and lexical organization*, p. 21–74. Hillsdale : Lawrence Erlbaum Associates.
- BEESELEY K. (1998). Arabic morphological analysis on the internet. In *Proceedings of the International Conference on Multi-Lingual Computing*.
- BERGEN B. & CHANG N. (2005). Embodied Construction Grammar in simulation-based language understanding. In J.-O. ÖSTMAN & M. FRIED, Eds., *Construction Grammars : Cognitive grounding and theoretical extensions*, p. 147–190. Amsterdam : John Benjamins.
- BHUYAN M. S. I. & AHMED R. (2008). An HPSG analysis of Arabic verb. In *The International Arab Conference on Information Technology (ACIT 2008)*.
- BUCKWALTER T. (2004). Issues in arabic orthography and morphology analysis. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*, Semitic '04, p. 31–34, Stroudsburg, PA, USA : Association for Computational Linguistics.
- CRABBÉ B., DUCHIER D., GARDENT C., LE ROUX J. & PARMENTIER Y. (2013). XMG : eXtensible MetaGrammar. *Computational Linguistics*, **39**(3), 1–66.
- DANKS W. (2011). *The Arabic Verb : Form and Meaning in the Vowel-lengthening Patterns*. Number 63 in Studies in functional and structural linguistics. Amsterdam : John Benjamins.
- DORON E. (2003). Agency and voice : The semantics of the Semitic templates. *Natural Language Semantics*, **11**(1), 1–67.
- DORON E. (2013). Binyanim : Modern Hebrew. In G. KHAN, Ed., *Encyclopedia of Hebrew Language and Linguistics*. Brill Online. Available online at http://referenceworks.brillonline.com/entries/encyclopedia-of-hebrew-language-and-linguistics/binyanim-modern-hebrew-EHLL_COM_000000247.
- DUCHIER D., MAGNANA EKOUKOU B., PARMENTIER Y., PETITJEAN S. & SCHANG E. (2012). Décrire la morphologie des verbes en ikota au moyen d'une métagrammaire. In *19e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2012) – Atelier sur le traitement automatique des langues africaines (TALAf 2012)*, p. 97–106, Grenoble, France.
- FILLMORE C. J. (1977). The case for case reopened. In P. COLE & J. M. SADOCK, Eds., *Grammatical Relations*, volume 8 of *Syntax and Semantics*, p. 59–81. New York : Academic Press.
- HABASH N. & RAMBOW O. (2006). MAGEAD : A morphological analyzer and generator for the arabic dialects. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*.
- KALLMEYER L. & OSSWALD R. (2013). Syntax-driven semantic frame composition in Lexicalized Tree Adjoining Grammar. *Journal of Language Modelling*, **1**, 267–330.
- KIRAZ G. A. (2001). *Computational Nonlinear Morphology : With Emphasis on Semitic Languages*. New York, NY, USA : Cambridge University Press.
- LICHTE T., DIEZ A. & PETITJEAN S. (2013). Coupling trees, words and frames through XMG. In *Proceedings of the ESSLLI 2013 workshop on High-level Methodologies for Grammar Engineering*.
- LICHTE T. & PETITJEAN S. (to appear). Implementing semantic frames as typed feature structures with XMG. *Journal of Language Modelling*.
- MCCARTHY J. J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic inquiry*, p. 373–418.
- PETERSEN W. (2007). Representation of concepts as frames. *The Baltic International Yearbook of Cognition, Logic and Communication*, **2**, 151–170.
- PETITJEAN S. (2014). *Génération Modulaire de Grammaires Formelles*. PhD thesis, Université d'Orléans. Thèse de Doctorat.
- RYDING K. (2005). *A Reference Grammar of Modern Standard Arabic*. A Reference Grammar of Modern Standard Arabic. Cambridge University Press.
- SCHNEIDER N. (2010). Computational cognitive morphosemantics : Modeling morphological compositionality in Hebrew verbs with Embodied Construction Grammar. In *Proceedings of the 36th Annual Meeting of the Berkeley Linguistics Society (BLS)*. Available online at <http://www.cs.cmu.edu/~nschneid/bls36.pdf>.
- SHAALAN K. F., SAMIH Y., ATTIA M., PECINA P. & VAN GENABITH J. (2012). Arabic word generation and modelling for spell checking. In *LREC*, p. 719–725.
- YONA S. & WINTNER S. (2005). A finite-state morphological grammar of Hebrew. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, p. 9–16, Ann Arbor, Michigan : Association for Computational Linguistics.

Création d'un nouveau treebank à partir de quatrièmes de couverture

Philippe Blache¹ Grégoire Montcheuil² Stéphane Rauzy¹ Marie-Laure Guénot²

(1) Aix Marseille Université, CNRS, Laboratoire Parole et Langage UMR 7309

(2) Equipex ORTOLANG, CNRS

{prenom.nom}@lpl-aix.fr

Résumé. Nous présentons ici 4-couv, un nouveau corpus arboré d'environ 3 500 phrases, constitué d'un ensemble de quatrièmes de couverture, étiqueté et analysé automatiquement puis corrigé et validé à la main. Il répond à des besoins spécifiques pour des projets de linguistique expérimentale, et vise à rester compatible avec les autres treebanks existants pour le français. Nous présentons ici le corpus lui-même ainsi que les outils utilisés pour les différentes étapes de son élaboration : choix des textes, étiquetage, parsing, correction manuelle.

Abstract.

Creation of a new treebank with backcovers

We introduce 4-couv, a treebank of approximatively 3 500 trees, built from a set of literacy backcovers. It has been automatically tagged and parsed, then manually corrected and validated. It was developed in the perspective of linguistic experiment projects, and aims to be compatible with other standard treebanks for french. We present in the following the corpus itself, then the tools we used or developed for the different stages of its elaboration : texts' selection, tagging, parsing, and manual correction.

Mots-clés : Corpus arboré, Étiquetage automatique, Analyse syntaxique automatique, Parsing stochastique, Conventions d'annotation, Outils d'annotation, Linguistique expérimentale.

Keywords: Treebank, Tagging, Parsing, Stochastic parsing, Annotation scheme, Annotation tools, Experimental linguistics.

1 Introduction

Les treebanks constituent une ressource indispensable non seulement pour la description de la syntaxe d'une langue, mais également pour l'entraînement ou la validation des systèmes d'analyse automatique. Pour le français, les premières ressources véritablement disponibles restent bien entendu le *French Treebank* (Abeillé & Crabbé, 2013) et toutes ses déclinaisons. Il existe également d'autres ressources, notamment la partie française du *Universal Dependencies Treebank*¹. Cependant, si l'on élargit le champs d'application au cadre de la **linguistique expérimentale**, il devient nécessaire de disposer de ressources de haut niveau pouvant être utilisées dans des expériences avec des sujets humains. Il est dans ce cas crucial de prendre en compte des éléments tels que la nature des textes, leur style, leur lisibilité. Le type d'expérience typique pour l'étude du traitement du langage par l'homme consiste à analyser les données associées à la lecture (mouvement oculaire, électro-encéphalographie). Or, l'intérêt de disposer d'un treebank pour faire passer ce type d'expérience est très important : il devient en effet possible d'établir des modèles prédictifs (par exemple de difficulté) à partir d'informations syntaxiques. Toutefois la plupart des études conduites à ce jour ne prennent en compte que le niveau morphosyntaxique (Demberg & Keller, 2008). Ces expériences ont consisté à faire lire des textes à un certain nombre de sujets. Une expérience similaire a été conduite pour le français (Rauzy & Blache, 2012) en utilisant des extraits du FTB. À cette occasion, nous avons pu constater un biais important dans l'acquisition des données : la **nature des textes** eux-mêmes. Il s'agit en effet d'articles du *Monde*, anciens, et dont l'intérêt est souvent très limité. Un effet de lecture superficielle suscité par le manque d'intérêt est alors important, entraînant une chute d'attention importante, en même temps qu'un déficit de compréhension.

Nous avons donc décidé, à la fois pour enrichir le patrimoine existant, mais également dans la perspective de pouvoir les

1. <https://code.google.com/p/uni-dep-tb/>

utiliser dans un environnement expérimental, de constituer un nouveau treebank, sur la base de textes courts, sémantiquement consistants (en d’autres termes, auto-suffisants pour leur interprétation), et suscitant l’intérêt de façon à maintenir l’attention pendant la lecture. Nous avons pour cela choisi de constituer un corpus de “quatrièmes de couverture”, permettant de respecter les contraintes indiquées : il s’agit du **corpus 4-Couv**, en cours de développement et dont une première livraison pourra être effectuée fin 2015. Ce corpus est constitué d’un ensemble de textes provenant de différents éditeurs (Pocket, Gallimard) ayant donné leur accord pour un usage à fins de recherche. Nous avons ainsi pu récupérer environ 8 000 textes. Un premier corpus de 500 textes a été constitué, représentant environ 3 500 phrases, formant la première livraison du treebank 4-Couv.

Nous proposons dans cet article de décrire la méthodologie et les outils mis au point pour la constitution de 4-Couv. Au-delà des problèmes classiques soulevés pour la constitution de ce type de treebank (analyse automatique, correction manuelle) s’ajoute l’étape de **sélection de textes** parmi un ensemble volumineux : nous avons pour cela développé un système d’aide à la sélection, décrit dans la première partie. Il s’agit d’un outil permettant d’évaluer (manuellement) certains critères tout en effectuant au passage un certain nombre de vérifications ou corrections (segmentations, mots inconnus, etc.). Il est ensuite nécessaire de traiter les problèmes posés par l’analyse syntaxique. Nous décrivons ainsi dans la seconde partie la question du **jeu d’étiquettes et des annotations syntaxiques** ; afin d’assurer une certaine interopérabilité, nous proposons pour cela de rester dans le cadre proposé par le FTB, en introduisant quelques modifications mineures. Nous avons ainsi développé un **analyseur syntaxique**, décrit dans la 3ème partie, entraîné sur le FTB ainsi modifié, et qui nous permet de produire les arbres d’entrée du treebank. Le résultat est enfin corrigé manuellement. Nous décrivons dans la dernière partie deux systèmes que nous avons développés dans cette perspective : un **système d’aide à la correction morphosyntaxique** et un **éditeur d’arbres syntaxiques**.

2 Constitution du corpus et outil d’aide à la sélection

2.1 Description des textes

Comme nous l’avons dit précédemment, nous construisons le corpus à partir des descriptions qui se trouvent sur la quatrième de couverture des livres. Une petite étude statistique réalisée sur 1 000 textes pris au hasard avant sélection nous a confirmé que ce sont des textes assez courts : 136 742 tokens² pour 6 838 phrases, 80% des textes ont entre 80 et 200 tokens et entre 4 et 10 phrases ; la taille moyenne des phrases est de 20 tokens (80% des phrases ont moins de 30 tokens et moins de 10% en ont plus de 40). Les textes correspondent généralement à (a) un extrait du livre, (b) un résumé ou synopsis de l’histoire, (c) la genèse du texte, (d) un commentaire à propos de l’œuvre, ou encore (e) une combinaison de deux ou trois de ces éléments. Chacun de ces textes courts est sémantiquement autonome et, élément crucial pour notre corpus, est censé entretenir l’intérêt à la lecture, en minimisant autant que possible la chute de l’attention et de la recherche de compréhension.

2.2 Outil d’aide à la sélection

Afin de choisir les textes les plus pertinents pour le corpus, nous avons mis au point un système d’aide à la sélection basé sur des fichiers HTML constituant de véritables petits wikis autonomes présentant une dizaine de textes à évaluer. Cette stratégie de fichiers HTML autonomes permet de répartir facilement le travail de relecture à diverses personnes, sans la nécessité d’installer un logiciel particulier (les fichiers étant utilisables directement par la plupart des navigateurs web modernes³), ni de se connecter à un serveur central (ce qui permet un travail hors-ligne). Nous nous sommes basés pour cela sur l’outil TiddlyWiki⁴ qui nous fournit le squelette des fichiers de wikis autonomes que nous avons configuré pour nos besoins. Nous avons ensuite utilisé un script Perl pour “remplir” chaque fichier avec les informations des quatrièmes de couverture d’une dizaine de livres.

Comme le montre la figure 1, pour chaque texte un “tiddler”⁵ est créé présentant les sections suivantes : (a) les métadonnées du livre (auteur, titre, éditeur, ISBN,...), (b) le texte dans sa présentation initiale⁶, (c) la découpe en phrases du

2. Les tokens incluent les mots et les signes de ponctuation.

3. Seule l’installation d’un petit complément étant parfois nécessaire pour la sauvegarde des fichiers.

4. <http://classic.tiddlywiki.com/> (version 2.8.1)

5. L’unité d’information de base dans TiddlyWiki qui, dans notre cas, correspond à un onglet.

6. En réalité seules quelques informations typographiques sont préservées : paragraphe, italiques, gras,...

The screenshot shows the '4Couv selector' web application. The main header is blue with the text '4Couv selector' and a subtitle 'Un TiddlyWiki pour sélectionner les quatrièmes de couvertures'. Below the header, there's a sidebar on the left with a list of books under 'Lunes :'. The main content area is titled '[01] Le cycle d'Eric' and shows a last modification date of 'Wednesday 03 September 2014 at 14:00:00'. It contains a form with fields for 'Auteur', 'Titre', 'Tomer', 'Série', 'Sous-titre', 'Langue originale', 'Traduction', 'ISBN', 'Éditeur', 'Collection(s)', 'Format', and 'Parution'. Below the form is a 'Description' section with a text area and a 'Phrases' section with a list of phrases. The interface is clean and functional, with a focus on text input and selection.

FIGURE 1 – Outil de sélection : vue générale

texte, (d) un cadre d'évaluation, (e) une liste des mots non-reconnus par le POS tagger. La section d'évaluation se compose essentiellement de cases à cocher et de champs à sélectionner pour simplifier celle-ci. D'autre part, grâce à la syntaxe wiki, il est relativement aisé de corriger d'éventuelles erreurs dans la découpe en phrases (celles-ci sont les lignes d'une table) ou d'introduire des limites de sections dans la composition de texte (en ajoutant des lignes blanches). Enfin, au cas où ce serait nécessaire, les champs de la section méta-données constituent un formulaire qui permet des rectifications.

3 Les annotations syntaxiques

3.1 L'étiquetage lexical

Pour l'annotation des unités syntaxiques minimales on se base sur un lexique⁷ qui associe à chaque forme une étiquette lexicale (partie du discours) et un vecteur de traits de sous-catégorisation. Le découpage en tokens est maximal, dans le sens où l'on découpera en unités lexicales distinctes même les formes très contraintes dès lors qu'elles obéissent aux règles de construction syntaxique standard ; p.ex. on étiquettera séparément les constituants d'expressions semi-figées telles que *il était une fois* ou bien *mettre à nu*, mais pas ceux de formes telles que *d'autant plus* ou *tant mieux* parce que celles-ci ne répondent plus à des contraintes syntaxiques générales.

À chaque catégorie lexicale correspond un jeu de traits spécifique, bien que de nombreux traits se retrouvent sur plusieurs catégories (typiquement le genre, le nombre, la personne). Les catégories ainsi que les jeux de traits utilisés sont somme toute assez standard, compatibles avec la plupart des corpus étiquetés automatiquement, et permettent d'indiquer un ensemble d'informations lexicales, morphologiques, syntaxiques ou parfois sémantiques qui auront une incidence sur la construction syntaxique des unités aux niveaux supérieurs, p.ex. le nombre d'un déterminant, les compléments attendus par un verbe ou le cas d'un pronom clitique.

Nous n'avons pas de constituants lexicaux discontinus, ni ne conservons d'ambiguïté dans l'étiquetage (i.e., tous les éléments reçoivent une catégorie lexicale, dont les traits de sous-catégorisation peuvent être sous-spécifiés le cas échéant). On ne modifie pas la catégorie des unités qui changent de paradigme (*une tarte maison*, *il est très zen*).

7. MarsaLex, <https://www.ortolang.fr/#/market/item/02f75cf8-8fcd-4305-alec-b34d516e716c>

3.2 L'annotation syntaxique

Les unités lexicales et syntaxiques entretiennent des relations syntagmatiques que l'on représente sous forme d'arbres. Ici aussi, afin de veiller à la compatibilité avec les autres treebanks existants, nous observons les contraintes de forme suivantes :

- On n'introduit pas de catégories vides dans les arbres (p.ex. dans le cas d'une construction elliptique) : chaque noeud est instancié par une unité lexicale ou syntagmatique.
- On fait une distinction entre niveau lexical et niveau syntagmatique, qui fait que l'on pourra avoir des syntagmes unaires, p.ex. *Simone* sera le constituant unique d'un NP dans (1).
(1) *Simone m'en donne trois*
- L'annotation ne comporte pas de constituants discontinus. Il s'agit d'une contrainte forte qui s'applique sur les choix linguistiques d'analyse que l'on peut faire, p.ex. pour des constructions présentant des discontinuités dans la structure syntagmatique, comme on trouve dans (1) ou (2), pour lesquelles notre liberté d'analyse se trouve limitée par cette obligation de forme.
(2) *Ce film, Paul et moi on a adoré*
- Pour annoter les syntagmes, on veille à n'attribuer que des étiquettes correspondant à des constructions proprement syntagmatiques (typiquement syntagmes nominaux, verbaux, adjectivaux, etc.) ; cela a pour conséquence notable que notre annotation des constructions coordonnées, phénomène canoniquement problématique, est différente de celle utilisée par le FTB et ses déclinaisons.
- On utilise le même type d'annotation des fonctions syntaxiques que celui introduit pour le FTB⁸.

Notre approche de l'annotation syntaxique est guidée par les usages (*corpus driven*, cf. p.ex. Lacheret *et al.* (2014) parmi les projets récents), c'est-à-dire que la correction manuelle des arbres obtenus automatiquement peut mener, rétroactivement, à des modifications du fonctionnement du parseur (qui réapprend sur les sorties corrigées) afin d'en améliorer les résultats de manière dynamique.

4 Génération automatique de treebank

Le treebank est généré à partir de l'analyseur stochastique du LPL (Rauzy & Blache, 2009). La chaîne de traitement suit un schéma classique. Dans un premier temps, le texte brut est segmenté en tokens par un segmenteur à base de règles. Un lexique permet ensuite d'associer à la forme de chaque token la distribution des catégories morphosyntaxiques correspondantes. Le processus de désambiguïsation est réalisé par un étiqueteur stochastique utilisant la technologie HMM (Rabiner, 1989) pour identifier la séquence de catégories morphosyntaxiques la plus probable. Dans un dernier temps, un analyseur stochastique permet de générer les structures d'arbre aptes à décrire chaque énoncé et de sélectionner la structure d'arbre la plus probable décrivant l'énoncé. Cette chaîne de traitement est implémentée dans *MarsaTag* (Rauzy *et al.*, 2014).

Le modèle probabiliste pour la phase d'étiquetage est entraîné sur le corpus GraceLPL, une version du corpus Grace/Multi-tag (Paroubek & Rajman, 2000) contenant 700 000 tokens, que nous corrigeons et enrichissons régulièrement. L'information morphosyntaxique est dans notre modèle organisée sous la forme de 48 étiquettes distinctes (version 2013). Sur ce jeu d'étiquettes, l'évaluation de notre étiqueteur atteint un score de 0.974 (F-mesure).

Le modèle probabiliste pour l'analyseur est obtenu à partir du corpus FTLPL (Blache & Rauzy, 2012), une version du MFT (Schluter & van Genabith, 2007) extrait du FTB (Abeillé *et al.*, 2001). Le corpus FTLPL compte actuellement 1 500 phrases validées (soit environ 26 000 tokens), pour lesquelles la structure en constituants (*AP*, *NP*, *VP*,...) et leurs fonctions syntaxiques (*SUBJ*, *OBJ*, *ATR*, ...) sont disponibles. L'information syntaxique est retenue dans le modèle sous forme de 36 constituants distincts (en tenant compte des fonctions différentes associées aux constituants). L'algorithme de génération des arbres et de la sélection de l'arbre le plus probable s'inscrit dans l'approche *Augmented Transition Network* (Woods, 1970). L'évaluation complète de notre analyseur stochastique reste à venir. Sur les 1 500 phrases du corpus d'apprentissage, l'analyseur associe la même analyse que la référence pour 500 phrases environ (même structure d'arbre et mêmes étiquettes de constituants et de fonction), et la même structure d'arbre (mais avec des étiquettes différentes) pour 900 d'entre elles.

8. Ce type d'annotation est malheureusement moins précis et plus *ad hoc* que le système qui avait p.ex. été utilisé dans Gendner *et al.* (2009) où informations constructionnelles et fonctionnelles étaient mentionnées indépendamment.

5 Les outils de correction utilisés

La correction des annotations automatiques est réalisée en deux étapes. La première concerne la **correction de l'étiquetage morphosyntaxique** ; l'interface de correction est illustrée figure 2. La seconde étape consiste à **réviser les arbres syntagmatiques** produits à l'aide de l'outil *MarsaTag* (Rauzy *et al.*, 2014).

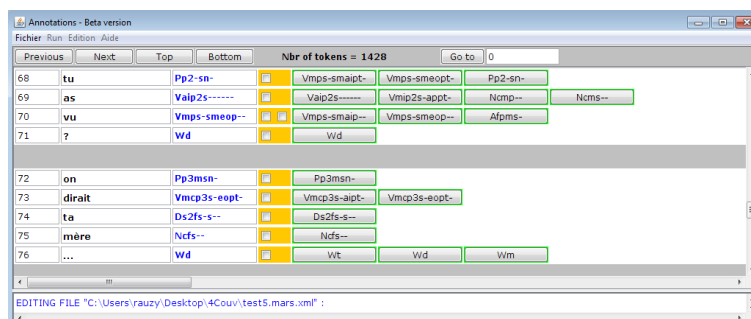


FIGURE 2 – L'interface utilisée pour corriger l'étiquetage. Le texte est présenté horizontalement, une ligne par token. Chaque ligne contient la forme, la solution retenue par l'étiqueteur (deuxième colonne), et la liste des étiquettes possibles associées à chaque forme. Pour corriger une étiquette, l'annotateur clique sur l'étiquette désirée dans la liste ou saisira les traits de l'étiquette si celle-ci n'est pas proposée.

Rares sont les éditeurs d'arbres syntaxiques adaptés au formalisme de constituants. Parmi les applications "installées" nous pouvons citer WordFreak (Morton & LaCivita, 2003) utilisé dans de nombreux projets d'annotation ou encore TrED 2.0 (Pajas & Štěpánek, 2008) qui nous a semblé l'un des plus complets pour ses grandes possibilités de personnalisation. En outre, ces dernières années ont vu l'émergence de nouvelles plateformes d'annotation linguistique totalement "en ligne", c'est-à-dire basées sur une architecture client-serveur, permettant souvent une annotation intuitive et rapide de textes (brat (Stenetorp *et al.*, 2012), TextAE⁹) et intégrant parfois une gestion du processus d'annotation et des différents "rôles" comme ceux d'annotateur, de curateur ou de chef de projet (GATE Teamware (Bontcheva *et al.*, 2013), WebAnno (Yimam *et al.*, 2013)). Cependant, si ces derniers outils sont assez bien adaptés au formalisme de dépendance, qui ne nécessite que deux "étages" d'annotation : les tokens et les liens de dépendance entre ceux-ci, ils ne sont pas vraiment adaptés au formalisme de constituants, où de nombreux "étages" de syntagmes peuvent se superposer. Pour conserver la simplicité d'un outil accessible avec n'importe quel navigateur web, nous avons donc créé une librairie JavaScript d'édition d'arbre syntaxique qui peut-être intégrée à une simple page html (figure 3) — et pour laquelle nous travaillons également à son intégration dans une plateforme d'annotation. Nous avons utilisé la librairie d3.js¹⁰ qui permet de générer des images SVG dynamiques, auxquelles il est possible d'appliquer des styles CSS pour personnaliser l'affichage. L'édition de l'arbre se fait à l'aide de déplacement de nœuds (*drag & drop*) et de fonctions d'édition accessibles via le menu contextuel.

6 Conclusion

Cet article a une double vocation : présenter un ensemble d'outils pour le treebanking et présenter un nouveau type de treebank, adapté à l'acquisition de données comportementales, physiologiques et cérébrales pour l'étude du traitement du langage.

Du point de vue des **outils** (disponibles via la plateforme de ressources SLDR@ORTOLANG¹¹), nous disposons désormais d'un environnement complet pour la constitution de treebank, permettant la sélection de textes et leur pré-traitement, l'analyse syntaxique automatique ainsi que deux outils d'aide à la correction : éditeur morpho-syntaxique et éditeur d'arbres.

Le **treebank** ainsi construit est composé d'un ensemble de textes courts, consistants du point de vue discursif (*i.e.* formant une unité discursive complète) et adaptés à l'expérimentation (par exemple pour l'acquisition de mouvements oculaires).

9. <http://textae.pubannotation.org/>

10. <http://d3js.org/>

11. <http://sldr.org/>

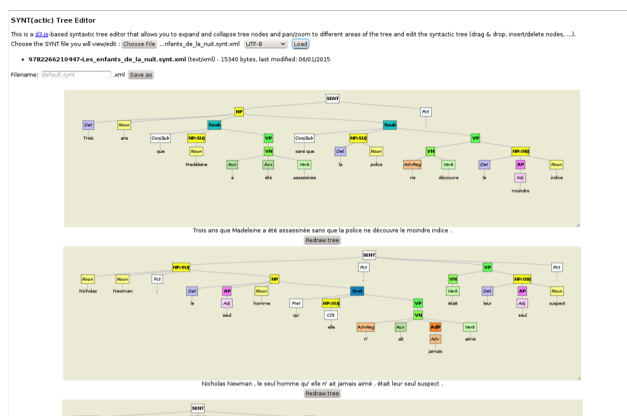


FIGURE 3 – Éditeur d’arbre : vue générale

Les outils présentés ici sont actuellement utilisés pour la constitution d’un premier treebank de 500 textes (3 500 arbres) qui sera opérationnel fin 2015, et lui-même également distribué via SLDR@ORTOLANG à des fins de recherche scientifique.

Nous sommes actuellement engagés dans le développement d’un treebank comparable en mandarin avec la collaboration de Hong-Kong Polytechnic University), dont une partie est formée de quatrièmes de couvertures d’ouvrages existants dans les deux langues.

Références

- ABEILLÉ A., CLÉMENT C., KINYON A. & TOUSSENEL F. (2001). Un corpus français arboré : quelques interrogations. In *Actes de Traitement Automatique des Langues Naturelles*, volume 1, p. 33–42, Tours, France.
- ABEILLÉ A. & CRABBÉ B. (2013). Vers un treebank du français parlé. In *Actes de TALN*.
- BLACHE P. & RAUZY S. (2012). Enrichissement du FTB : un treebank hybride constituants/propriétés. In *Actes de TALN*.
- BONTCHEVA K., CUNNINGHAM H., ROBERTS I., ROBERTS A., TABLAN V., ASWANI N. & GORRELL G. (2013). Gate teamware : a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, **47**(4), 1007–1029.
- DEMBERG V. & KELLER F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, **109**(2), 193–210.
- GENDNER V., VILNAT A., MONCEAUX L., PAROUBEK P., ROBBA I., FRANCOPOULO G. & GUÉNOT M.-L. (2009). *Les annotation syntaxiques de référence PEAS*. Rapport interne, version 2.2.
- HERNANDEZ N. & BOUDIN F. (2013). Construction automatique d’un large corpus libre annoté morpho-syntaxiquement en français. In *Actes de la conférence TALN-RECITAL 2013*.
- LAGHERET A., KAHANE S., BELIAO J., DISTER A., GERDES K., GOLDMAN J.-P., OBIN N., PIETRANDREA P. & TCHOBANOV A. (2014). Rhapsodie : Un treebank annoté pour l’étude de l’interface syntaxe-prosodie en français parlé. In *4e congrès mondial de linguistique française*, volume 8, p. 2675–2689.
- MORTON T. & LACIVITA J. (2003). Wordfreak : An open tool for linguistic annotation. In *Proceedings of NAACL-Demonstrations ’03*, p. 17–18, Stroudsburg, PA, USA : Association for Computational Linguistics.
- PAJAS P. & ŠTĚPÁNEK J. (2008). Recent advances in a feature-rich framework for treebank annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, p. 673–680, Manchester, UK : Coling 2008 Organizing Committee.
- PAROUBEK P. & RAJMAN M. (2000). Multitag, une ressource linguistique produit du paradigme d’évaluation. In *Actes de Traitement Automatique des Langues Naturelles*, p. 297–306, Lausanne, Suisse.

- RABINER L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**, 257–286.
- RAUZY S. & BLACHE P. (2009). Un point sur les outils du lpl pour l'analyse syntaxique du français. In *Actes du workshop ATALA 'Quels analyseurs syntaxiques pour le français ?'*, p. 1–6, Paris, France.
- RAUZY S. & BLACHE P. (2012). Robustness and processing difficulty models. a pilot study for eye-tracking data on the french treebank. In *Proceedings of Workshop on Eye-tracking and Natural Language Processing at The 24th International Conference on Computational Linguistics (COLING)*.
- RAUZY S., MONTCHEUIL G. & BLACHE P. (2014). MarsaTag, a tagger for French written texts and speech transcriptions. In *Second Asia Pacific Corpus Linguistics Conference*, Hong Kong.
- SCHLUTER N. & VAN GENABITH J. (2007). Preparing, restructuring, and augmenting a french treebank : Lexicalised parsers or coherent treebanks ? In *Proceedings of PACLING 07*, p. 200–209.
- STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). brat : a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 102–107, Avignon, France : Association for Computational Linguistics.
- WOODS W. A. (1970). Transition network grammars for natural language analysis. *Communications of the ACM*, **13**(10), 591–602.
- YIMAM S. M., GUREVYCH I., DE CASTILHO R. E. & BIEMANN C. (2013). Webanno : A flexible,web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, p. 1–6, Stroudsburg, PA, USA : Association for Computational Linguistics.

Entre écrit et oral ? Analyse comparée de conversations de type tchat et de conversations téléphoniques dans un centre de contact client

Géraldine Damnati, Aleksandra Guerraz, Delphine Charlet

Orange Labs, OPENSERV/CONTENT/FAST, 2 av. Pierre Marzin, 22307 LANNION cedex
{geraldine.damnati,aleksandra.guerraz,delphine.charlet}@orange.com

Résumé. Dans cet article nous proposons une première étude descriptive d'un corpus de conversations de type tchat issues d'un centre de contact d'assistance. Les dimensions lexicales, syntaxiques et interactionnelles sont analysées. L'étude parallèle de transcriptions de conversations téléphoniques issues d'un centre d'appel dans le même domaine de l'assistance permet d'établir des comparaisons entre ces deux modes d'interaction. L'analyse révèle des différences marquées en termes de déroulement de la conversation, avec une plus grande efficacité pour les conversations de type tchat malgré un plus grand étalement temporel. L'analyse lexicale et syntaxique révèle également des différences de niveaux de langage avec une plus grande proximité entre le client et le téléconseiller à l'oral que pour les tchats où le décalage entre le style adopté par le téléconseiller et l'expression du client est plus important.

Abstract.

Comparing Written and Spoken Languages: a Descriptive Study of Chat and Phone Conversations from an Assistance Contact Center

In this article we propose a first descriptive study of a chat conversations corpus from an assistance contact center. The analysis includes lexical, syntactic and interactional dimensions. Transcriptions of telephone conversations from an assistance call center are also studied, allowing comparisons between these two interaction modes to be drawn. The study reveals significant differences in terms of conversation flow, with an increased efficiency for chat conversations in spite of longer temporal span. The lexical and syntactic analyses also reveal differences in terms of language level with a tighter proximity between agent and customers on the phone than for chats where the style adopted by agents is different from the style adopted by customers.

Mots-clés : Centre de contact, conversations tchat, interaction, analyse lexicale et syntaxique.

Keywords: Contact center, chat conversations, interaction, lexical and syntactic analysis.

1 Introduction

Même si le canal téléphonique reste prépondérant dans l'interaction entre une entreprise et ses clients, les modes d'interaction se sont nettement diversifiés et la Gestion de la Relation Client se doit aujourd'hui d'intégrer la multi-canalité dans ses études analytiques. En particulier, les conversations en ligne par Messagerie Instantanée (ou tchat) se développent très rapidement et sont plébiscitées à la fois par les clients et par les téléconseillers qui y voient un moyen de garder un contact direct avec les clients tout en préservant une certaine distance. Du point de vue d'une entreprise, ces conversations, disponibles aisément en grande quantité, constituent une formidable mine d'information pour mieux comprendre les besoins des clients. Du point de vue du traitement du langage, elles constituent un nouveau terrain d'étude encore très peu exploré. Nous présentons dans cet article une analyse descriptive d'un corpus de conversations tchat en prenant le parti de conduire une étude comparative avec un corpus de conversations téléphoniques afin d'en faire apparaître les points communs et les différences.

Des travaux ont été menés sur des conversations de type tchat issues de *chatroom* (salons de conversations) (Falaise, 2005) a ainsi constitué un corpus de français tchaté. (Martel & al., 2007) décrivent un corpus similaire en anglais. (Cadilhac et al., 2013) étudient la structure relationnelle des conversations à travers une analyse discursive profonde de sessions tchat multi-utilisateurs dans un jeu vidéo en ligne. On trouve également quelques travaux sur les tchats en centre de contact. (Dickey et al., 2007) étudient un corpus de tchats client-téléconseiller du point de vue des stratégies employées pour favoriser la compréhension mutuelle entre les interlocuteurs, avec un focus sur les phénomènes de discontinuité dans l'interaction en cherchant à analyser les raisons pour lesquelles une mauvaise communication peut

s'installer. (Wu et al., 2012) proposent une typologie des modes de communications entre client et téléconseillers dans le cadre d'une étude sur l'interface de conversation. Le média ne peut pas être le seul prisme à travers lequel étudier les conversations. Le domaine et le degré de connaissance mutuelle des interlocuteurs est une dimension à prendre en compte pour caractériser ce type d'interaction. L'adaptation du mode d'expression à l'interlocuteur est décrite dans la littérature en termes de *conscience sociolinguistique*. C'est pourquoi nous proposons d'étudier spécifiquement les conversations de type tchat en centre de contact client, qui peuvent présenter des caractéristiques communes avec les conversations issues de *chatroom* mais qui ont la particularité de se placer sur un niveau plus formel et institutionnel. En choisissant l'angle du domaine, nous proposons de mener cette étude en relatif par rapport aux propriétés d'un corpus de conversations téléphoniques issues du même périmètre. Les conversations téléphoniques ont pour leur part fait l'objet de plus nombreuses études avec des travaux plus avancés d'extraction d'information à partir des conversations en français issues de centres d'appel d'EDF (Garnier-Rizet et al., 2008) ou de la RATP (Béchet et al., 2012), avec comme principale difficulté la nécessité de prendre en compte les phénomènes liés à la parole spontanée (phénomènes intrinsèques ou bruit induit dans la transcription automatique).

Nous proposons ici de conduire une analyse descriptive d'un corpus de conversations tchat issues d'un centre de contact d'assistance technique d'Orange. Le contexte d'assistance a une influence sur la nature des données étudiées. En effet les clients peuvent contacter l'assistance pour des renseignements mais également pour résoudre des problèmes techniques auquel cas la conversation peut être émaillée de manipulations diverses de la part du téléconseiller ou du client. L'analyse est conduite en parallèle sur un corpus de conversations téléphoniques issues d'un centre d'appel sur un périmètre d'assistance comparable. Nous nous intéressons donc dans cette étude à faire apparaître les différences entre ces deux types de conversations, décrits à la section 2, en nous concentrant dans la section 3 sur la dimension de l'interaction puis à la section 4 sur les dimensions lexicales et syntaxiques.

2 Description des données

2.1 Définitions et analogies

Afin de mener l'étude comparative des deux types de conversations, nous proposons en avant-propos de préciser le vocabulaire employé. Il s'agit d'une proposition visant à mettre en parallèle les éléments d'une conversation téléphonique et les éléments d'une conversation tchat. Le vocabulaire employé dans le cadre des conversations tchat hérite du vocabulaire de la messagerie instantanée où chaque élément envoyé correspond à un message, comme pour les messageries asynchrones. Les définitions posées dans la *TABLE 1* font apparaître de nombreuses analogies, à l'exception notable de la superposition.

	Tchat	Téléphone
Participants	Scripteurs	Locuteurs
Unité élémentaire de l'interaction	Message : séquence de mots tapés par le scripteur, suivie d'un « envoi ». La segmentation volontaire en messages peut être interprétée comme un retour à la ligne.	Groupe de souffle : séquence de mots prononcés par le locuteur, suivie d'une pause. La pause n'est pas nécessairement volontaire.
Unité de l'interaction	Tour de clavier : concaténation des messages consécutifs provenant d'un même scripteur.	Tour de parole : concaténation des GS consécutifs provenant d'un même locuteur.
Informations temporelles de l'interaction	A chaque message est associé l'instant de l'envoi du message. L'information précise du début du message n'est pas disponible.	L'annotation manuelle produit des marqueurs temporels de début et fin des tours de parole et des groupes de souffle.
Superposition dans l'interaction	Pas de superposition d'édition des messages qui apparaissent séquentiellement mais il se peut que les scripteurs écrivent en même temps et qu'un message soit écrit sans connaître le précédent.	Parole superposée : un locuteur parle tandis que son interlocuteur est encore en train de parler. Les locuteurs en question sont conscients de la superposition.

TABLE 1 : Définitions

2.2 Corpus de conversations tchat

Le corpus étudié est issu de l'assistance en ligne sur la TV d'Orange. Les clients contactent l'assistance pour un problème technique ou des renseignements sur leur offre. Dans certains cas, la conversation se déroule de façon linéaire comme dans l'exemple de la *FIGURE 1* et dans d'autres, le téléconseiller peut effectuer des tests à distance sur la ligne qui peuvent prendre du temps ou le client est amené à faire des manipulations sur son installation (débrancher, rebrancher, réinitialiser, ...) qui induisent également des temps de latence dans la conversation. Dans tous les cas le corpus a la forme suivante (*FIGURE 1*) où les instants renseignés en début de ligne correspondent à l'instant où le participant (client ou téléconseiller) presse la touche entrée de son clavier et donc à l'instant où le message présent sur la ligne devient visible à l'autre participant.

```

[12:04:20] Vous êtes en relation avec _TC_.
[12:04:29] _TC_: Bonjour, je suis _TC_, que puis-je pour vous ?
[12:05:05] _CLIENT_: mes enfant ont perdu la carte dans le modem et je nai plus de tele comment dois je faire?
[12:05:27] _TC_: Pouvez vous me confirmer votre numéro ligne fixe afin que je sois sûr d'avoir le bon dossier ?
[12:05:56] _CLIENT_: _NUMTEL_
[12:07:04] _TC_: Si je comprend bien vous avez perdu la carte d'accès de votre décodeur.
[12:07:27] _CLIENT_: oui ces bien sa
[12:07:47] _CLIENT_: code erreur S03
[12:09:09] _TC_: Pas de souci, je vais vous envoyer une autre carte par voie postale à votre domicile.
[12:09:38] _CLIENT_: est ce que je peux venir chez orange la chercher aujourd'hui
[12:10:36] _TC_: Vous ne pouvez pas récupérer une carte depuis une boutique Orange puisque vous n'avez pas une.
[12:11:02] _TC_: Car dans une boutique Orange, ils peuvent seulement faire un échange.
[12:11:33] _CLIENT_: ok merci de me l'envoyer au plus vite vous avez bien mes coordonnées
[12:11:57] _TC_: Oui je les bien sur votre dossier.
[12:12:51] _CLIENT_: ok tres bien dici 48h au plus tard 72h pour la carte
[12:14:06] _TC_: Vous la recevrez selon les délais postaux à l'adresse figurant sur votre dossier (entre 3 et 5 jours).
[12:14:25] _CLIENT_: ok tres bien en vous remerciant a bientot
[12:15:20] _TC_: Je vous en prie.
[12:15:29] _TC_: Avant de nous quitter avez-vous d'autres questions ?
[12:17:23] _CLIENT_: non merci

```

FIGURE 1 : Exemple de conversation tchat

2.3 Corpus de conversations téléphoniques

Le corpus de conversations téléphoniques est issu du centre d'appel d'assistance technique 3901, réservé aux clients professionnels. Le périmètre est plus large que pour les conversations tchat car il couvre l'assistance sur la TV, la connexion Internet et le téléphone. On retrouve néanmoins des problématiques similaires avec un vocabulaire général propre aux services Orange et des conversations qui peuvent se dérouler de façon linéaire ou qui impliquent des manipulations (tests de ligne, etc...). La transcription manuelle des conversations a été réalisée à l'aide de l'outil Transcriber (Barras et al., 2001) qui permet d'obtenir des informations temporelles en plus de la simple retranscription du contenu de la conversation. Pour cette transcription, réalisée antérieurement à la collecte des tchats, les consignes données à l'annotateur¹ n'ont donc pas été orientées pour produire des unités comparables à celles des tchats. Cependant, l'annotateur pouvait insérer, à l'intérieur d'un tour de parole, des points de synchronisation temporelle lorsque le tour de parole lui semblait trop long pour être considéré d'un seul bloc et ces points de synchronisation ont été insérés à l'endroit où des pauses étaient perceptibles. La différence notable est que dans le cas des tchats, c'est le scripteur lui-même qui décide de la segmentation en message (en appuyant volontairement sur la touche envoi), tandis que dans le cas de la conversation orale annotée par un tiers, c'est le transcripateur qui fait cette segmentation. Avec toutes ces réserves, nous conservons cependant cette similarité entre message tchat et groupe de souffle dans la suite.

2.4 Description générale des données

Les données tchats sont disponibles en grande quantité car elles sont directement sauvegardées dans les logs du système à l'issue de la conversation. En revanche, la transcription manuelle de conversations téléphoniques est un processus long et coûteux. De façon à disposer de corpora de tailles comparables, nous avons pris le parti de sélectionner un corpus de tchats contenant un nombre total de mots équivalent au nombre de mots présents dans le corpus de conversations orales disponible. Le corpus de tchats regroupe ainsi 230 conversations, pour un total de 6879 messages et 76839 mots et le corpus téléphonique est constitué de 56 conversations pour un total de 6870 groupes de souffles et 76463 mots.

Les données ont été anonymisées préalablement à l'étude. Les noms des clients et téléconseillers sont remplacés par un seul symbole (_CLIENT_ et _TC_ respectivement) ainsi que les numéros de téléphone, de contrat, les adresses et adresses mail. La variabilité lexicale exclue donc tout ce qui a trait aux données personnelles. Les transcriptions manuelles des conversations téléphoniques ont été réalisées en respectant les conventions classiques de transcription de l'oral. Le corpus de tchat quant à lui présente une forme non normalisée, telle que saisie par les scripteurs. De façon à unifier dans la mesure du possible les deux corpora, des pré-traitements ont été réalisés. Les données sont systématiquement passées en minuscule et les signes de ponctuation sont supprimés dans le corpus de tchats. Les mots finissant par une apostrophe sont séparés du mot suivant par un espace, et les traits d'union ne sont pas conservés dans le cas où ils font office de liaison entre deux mots (comme dans le cas le plus typique de l'inversion du sujet dans les phrases interrogatives).

¹ <http://trans.sourceforge.net/en/transguidFR.php>

3 Analyse de l'interaction

Dans le *TABLE 2*, nous proposons une analyse des interactions, selon leur durée, leur longueur en tour de clavier/parole et en messages/GS. Les messages/GS sont subdivisés en 2 catégories :

- début de tour de clavier/parole (B): le message/GS précédent provient d'un autre scripteur/locuteur
- interne au tour de clavier/parole (I) : le message/GS précédent provient du même scripteur/locuteur

		Tchat (230 conversations)			Téléphone (56 conversations)		
		Total	CLIENT	TC	Total	CLIENT	TC
durée moyenne de la conversation (seconde)		1185.7	549.3	636.4	594.5	162.0	221.8
#tours par conversation		21.2	10.3	10.9	83.3	41.5	41.8
#messages par tour		1.41	1.27	1.54	1.47	1.33	1.62
#mots par message	tous	11.2	8.6	13.2	11.1	10.0	12.1
	B-	11.1	8.7	13.3	9.6	8.4	10.8
	I-	11.4	8.1	13.0	14.3	14.9	14.1

TABLE 2 : Analyse de l'interaction

- Les conversations tchat sont 2 fois plus longues en durée que les conversations téléphoniques. Il faut noter que pour les conversations téléphoniques le temps de parole total est de 383.8s en moyenne par conversion soit 64.5% de la durée totale de la conversation. Il ne nous est pas possible en l'état d'établir le ratio équivalent pour les conversations tchat. Par ailleurs, les conversations de type tchat présentent une plus grande diversité avec un écart type de 901s sur la durée contre un écart type de 316s pour les conversations téléphoniques.
- Les conversations téléphoniques sont 4 fois plus longues en termes d'interactions que les conversations tchat (83.3 tours de parole contre 21.2 tours de clavier en moyenne).
- Le nombre de messages par tour de parole est sensiblement le même que le nombre de messages par tour de clavier (1.47 contre 1.41), et l'on observe la même différence entre TC et CLIENT en tchat comme à l'oral : dans les 2 cas, le nombre de messages par tour de parole/clavier est plus important pour le téléconseiller que pour le client.
- Si le nombre de mots par message est globalement comparable entre tchat et oral, on observe cependant des différences entre CLIENT et TC. Pour les tchats, le nombre de mots par message est beaucoup plus important pour les téléconseillers que pour les clients, cet écart est plus réduit à l'oral. Ceci peut s'expliquer par le fait que les TC ont accès à des bibliothèques de réponses qu'ils peuvent insérer sans avoir à les saisir. Ils fournissent en particulier des explications détaillées sur la marche à suivre en cas de dysfonctionnement.
- Dans les tchats, on n'observe pas de différences entre le nombre de mots par messages B- et le nombre de mots par message I-. En revanche, les messages I- sont significativement plus longs que les B- dans le corpus téléphonique.

Les deux premières observations peuvent sembler paradoxales. Ceci peut sans doute s'expliquer, d'une part parce qu'il est plus rapide de parler que d'écrire, et d'autre part par le fait qu'une conversation orale est une activité exclusive (c'est-à-dire que le client et le téléconseiller ne font que cela quand ils sont en conversation ensemble), tandis que le tchat est une activité potentiellement menée en parallèle d'autres activités (le TC peut avoir deux sessions tchat en parallèle, le CLIENT peut avoir d'autres activités en parallèle). Concernant la dernière observation, nous pouvons formuler l'hypothèse qu'un locuteur ne fournit pas trop d'information au début d'un tour de parole, et développe son propos progressivement, alors que la rémanence du message à l'écran peut conduire le scripteur à saisir dès le début un message complet et circonstancié, quitte à ce que son interlocuteur le relise plusieurs fois pour bien le comprendre.

4 Analyse lexicale et syntaxique

4.1 Diversité lexicale

Le tableau ci-dessous illustre la composition des deux corpora en termes de nombre de mots et de nombre de lemmes. Nous rappelons que le processus de lemmatisation inclue une correction automatique des erreurs d'accentuation ainsi qu'un regroupement de locutions. Par ailleurs, pour le corpus oral, les marques d'hésitation (ou pauses remplies *eah* et *hum*) sont supprimées avant de lancer l'analyse alors qu'elles sont comptabilisées au niveau des mots. Les pauses remplies correspondent à 1034 occurrences pour les clients et 873 pour les téléconseillers. Le calcul des lemmes différents est fait à partir du lemme associé à sa catégorie grammaticale. Ainsi deux homographes correspondant à des sens différents sont comptabilisés comme deux lemmes différents.

	Tchat			Téléphone		
	Total	CLIENT	TC	GLOBAL	CLIENT	TC
# total de mots	76839	25867	50972	76463	30751	45712
# mots différents	4446	3088	2641	3726	2369	2821

(% d'occurrence 1)	(46.4%)	(52.3%)	(39.1%)	(42.7%)	(46.4%)	(43.1%)
# total de lemmes	74245	25370	48875	70053	28197	41856
# lemmes différents	4192	2968	2575	3772	2426	2908
(% d'occurrence 1)	(38.5%)	(43.9%)	(32.3%)	(31.4%)	(36.1%)	(31.8%)

TABLE 3 : Analyse de la diversité lexicale

On observe une plus grande proportion de mots n'apparaissant qu'une fois chez le client, et ce de façon nettement plus significative pour les tchats. Ceci s'explique en partie par la plus grande proportion de formes incorrectes. Nous ne pouvons relater ici l'étude exhaustive des formes incorrectes recensées dans les conversations tchat, mais mentionner cependant quelques éléments. Les formes incorrectes s'entendent au sens du dictionnaire de notre analyseur linguistique (Heinecke et al. 2008) et n'englobent pas les erreurs grammaticales qui conduisent à remplacer un mot par un autre mot présent dans le dictionnaire. L'analyse des messages du client révèle 1559 occurrences de formes inconnues (6,14% des occurrences) pour 648 formes différentes. Comme l'on pouvait s'y attendre ce nombre est plus réduit pour les téléconseillers où 535 occurrences ont été recensées (soit 1,09% des occurrences totales) pour 229 formes différentes. Pour le client, parmi les 648 formes inconnues, on recense 289 mots *mal accentués* (soit 44,6% des formes incorrectes); 257 *erreurs typographiques* (39,7%) dont 120 *suppressions* (par exemple accueilment, maintenat, essage pour message) ainsi que des cas de suppression de la lettre finale que l'on peut observer dans le langage SMS (Véronis et al. 2006), 70 *ajouts* (parabolle, voptre), 41 *substitutions* (commercial pour commercial, instrinctions pour instruction) et 16 *inversions* (inetrnet, besion); 53 formes (8,2%) sont des *agglutinations* (par exemple explicationbonne, dela, derien, ainsi que des élisions cest, jespere, lécran, daccord, narrive, nai); et 33 formes (5,1%) sont des *abréviations non standard* (par exemple msg pour message, teleph pour téléphone) ou des troncations de mots (manip, dispo) ainsi que des abréviations fréquemment utilisées dans les SMS (bjr, pb, qd, tt). On retrouve chez le TC la même typologie de formes incorrectes à l'exception de la dernière catégorie. En effet, contrairement aux autres qui sont des erreurs, les abréviations sont volontaires et le téléconseiller ne se permet pas ce niveau de langage. Les erreurs typographiques (du CLIENT et du TC) sont principalement des fautes de saisie et sont dues notamment à la rapidité d'écriture. Chez le CLIENT, des fautes d'orthographe « classiques » sont plus fréquentes que chez le TC, par exemple : *êteind*, *rappelera*, *parabolle*. Le CLIENT utilise une « écriture SMS », alors que ce phénomène n'apparaît pas chez le TC. On trouve, par exemple, chez le CLIENT 22 occurrences de « c » (au sein de 16 conversations différentes) qui apparaissent 19 fois pour « c'est », 2 fois pour « ça » et 1 fois pour « ce ».

En observant maintenant les mots qui sont employés à la fois par le client et le téléconseiller, nous voyons dans la figure ci-dessous seuls 29% des mots se retrouvent à la fois chez le client et chez le conseiller ($CLIENT \cap TC$) dans le corpus de tchats (contre 39% dans le corpus téléphonique). Ces mots communs aux deux scripteurs représentent 88% des occurrences de mots du client dans les tchats et 95% des occurrences de mots du client dans les conversations téléphonique. Hormis les différences en termes de formes incorrectes, cette observation nous permet d'observer que le niveau de langue employé par le téléconseiller est plus éloigné de celui du client pour les tchats que pour l'oral. Ceci se vérifie de façon plus nette avec l'analyse syntaxique développée à la section suivante.

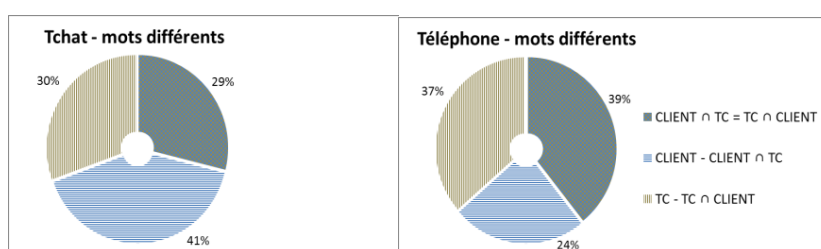


FIGURE 2 : Répartition des mots employés par le client et/ou le téléconseiller

4.2 Distribution des catégories syntaxiques

Les statistiques présentées ici ne sont pas le résultat d'une annotation manuelle en *Part of Speech* mais le résultat de l'analyseur syntaxique TiLT d'Orange Labs (Heinecke et al., 2008). Les statistiques sont évaluées relativement au nombre total d'occurrences de lemmes (après regroupement des locutions).

La proportion de **verbes** est similaire entre les deux corpus et relativement équilibrée entre les clients et les téléconseillers. En dehors des verbes modaux qui sont les plus fréquents pour le client comme pour le téléconseiller, les verbes les plus fréquents à l'oral pour le client sont *appeler* et *voir* et pour le téléconseiller *dire* et *regarder*; pour les tchats, les verbes les plus fréquents pour le client sont *fonctionner*, *changer*, *venir*, *dire* alors que pour le téléconseiller ce sont *patienter*, *remercier*, *souhaiter*, *inviter*. Le téléconseiller dans les tchats a plus recours à des formules de politesse explicites qu'à l'oral.

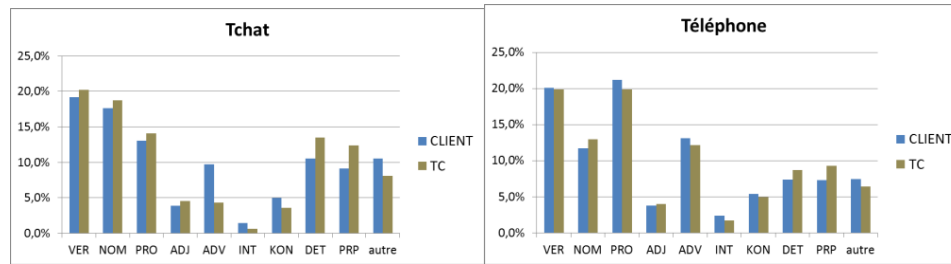


FIGURE 3 : Distribution des catégories syntaxiques

Nous observons en revanche une inversion dans la proportion du ratio de **noms** et de **pronoms**. Les trois pronoms les plus fréquents pour le client dans le corpus téléphonique sont *je*, *c'* et *j'*. Ces deux derniers sont bien souvent mal saisis dans le corpus tchat où l'élision de l'apostrophe est un phénomène très fréquent. Or dans une première approche nos statistiques portent sur le résultat de l'analyseur du Français, sans correction du texte ni adaptation de l'analyseur et les lettre *j* et *c* ne sont pas répertoriées comme pronoms. C'est une des raisons pour laquelle la proportion de pronoms dans le corpus tchat est moins large que dans le corpus audio. Ces « monolettres » étant principalement étiquetés comme des noms, cela explique également pour partie la plus forte proportion de noms pour les tchats que pour le corpus téléphonique.

La différence entre la proportion d'**adverbes** à l'oral par rapport au tchat est surtout marquée pour les téléconseillers et s'explique par la présence à l'oral d'adverbes qui sont principalement employés comme marqueurs discursifs (*donc*, *ben*, *alors*). Ce type de marqueurs ne se retrouve pas dans les conversations écrites. De même, on retrouve parmi les locutions adverbiales les plus fréquentes dans le corpus téléphonique *en fait*, *de toutes façons*, *quand même*, qui ne se retrouvent pas de façon aussi prononcée dans le corpus tchat. Si ces locutions aident à articuler la conversation orale, la conversation écrite se fait dans un style beaucoup plus direct et efficace dans lequel les locutions adverbiales les plus fréquentes sont utilisées pour des descriptions factuelles (*à distance*, *à la demande*, *par la suite*, ...).

De façon similaire, on retrouve dans les **interjections** des marques propres à l'oral comme (*ouais*, *hein*, *bah*, *hop*) qui ne se retrouvent pas à l'écrit. En revanche, le terme de type INT le plus fréquent chez le téléconseiller dans le corpus de tchats est « *s'il vous plaît* », qui apparaît dans 4% des messages alors qu'on le retrouve dans seulement 0,8% des groupes de souffle à l'oral. Une hypothèse est qu'en l'absence de marques non verbales de respect dans l'intonation, les téléconseillers ont plus souvent recours aux marqueurs explicites de politesse dans les conversations écrites.

5 Conclusion

Nous avons mené une analyse comparative de deux corpora de conversations issues d'un centre de contact d'assistance, l'un via le canal téléphonique et l'autre via une interface de messagerie instantanée. L'étude révèle des différences marquées en termes de déroulement de la conversation avec une durée deux fois plus longue pour les conversations tchat mais cependant quatre fois moins d'échanges. Cette première observation reflète une notion d'*efficacité* accrue à travers les tchats où la rémanence de l'information à l'écran favorise vraisemblablement la compréhension mutuelle des intervenants, malgré une durée plus longues liées au fait que le tchat n'est pas nécessairement une activité exclusive. L'analyse lexicale et syntaxique révèle également que les niveaux de langage employés par le client et les téléconseiller sont plus similaires à l'oral que dans les tchats où le téléconseiller adopte un style plus formel. A court terme, nous envisageons d'approfondir l'analyse de l'interaction à travers l'étude de la durée de chaque message. En effet, un des enjeux importants pour mieux comprendre le déroulement de la conversation serait de pouvoir détecter les désynchronisations dans l'interaction. Une analyse des marques typographiques d'expressivité a également été menée mais n'est pas relatée dans cet article. Enfin l'objectif de nos travaux est de pouvoir proposer des méthodes d'extraction d'information performantes de façon à alimenter les outils d'analyse dans le cadre de la gestion de la Relation Client.

Références

- BARRAS, C., GEOFFROIS, E., WU, Z., & LIBERMAN, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1), 5-22.
- BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BEZE M., DE MORI R., ARBILLOT E., (2012). DECODA: a call-centre human-human spoken conversation corpus. Actes de *LREC*.
- CADILHAC A., ASHER N., BENAMARA F., & LASCARIDES A. (2013). Grounding Strategic Conversation: Using Negotiation Dialogues to Predict Trades in a Win-Lose Game. Actes de *EMNLP*.
- DICKEY M., BURNETT G., CHUDоба K., & KAZMER, M. (2007). Do you read me? Perspective making and perspective taking in chat communities. *Journal of the Association for Information Systems*, 8(1), 3.

FALAISE A. (2005). Constitution d'un corpus de français tchaté. Actes de *RECITAL*.

GARNIER-RIZET M., ADDA G., CAILLIAU F., GAUVAIN J. L., GUILLEMIN-LANNE S., LAMEL L., ... & WAAST-RICHARD, C. (2008). CallSurf: Automatic Transcription, Indexing and Structuration of Call Center Conversational Speech for Knowledge Extraction and Query by Content. Actes de *LREC*.

HEINECKE, J., SMITS, G., CHARDENON, C., GUIMIER DE NEEF, E., MAILLEBUAU, E., BOUALEM, M. (2008). TiLT : plate-forme pour le traitement automatique des langues naturelles. *Traitement automatique des langues* , 49(2):17-41.

MARTELL E., FORSYTH N., & CRAIG H. (2007). Lexical and Discourse Analysis of Online Chat Dialog. Actes de *International Conference on Semantic Computing (ICSC)*.

VERONIS J. AND GUIMIER DE NEEF É. (2006). Le traitement des nouvelles formes de communication écrite. In Sabah, Gérard, editor, *Compréhension automatique des langues et interaction*, pages 227– 248. Paris: Hermès Science.

WU M., BHOWMICK A., & GOLDBERG J. (2012). Adding structured data in unstructured web chat conversation. Actes de *ACM symposium on User interface software and technology*.

Construction et maintenance d'une ressource lexicale basées sur l'usage

Laurie Planes¹,

(1) Inbenta France, 164 route de Revel, 31400 TOULOUSE

lplanes@inbenta.com

Résumé. Notre société développe un moteur de recherche (MR) sémantique basé sur la reformulation de requête. Notre MR s'appuie sur un lexique que nous avons construit en nous inspirant de la Théorie Sens-Texte (TST). Nous présentons ici notre ressource lexicale et indiquons comment nous l'enrichissons et la maintenons en fonction des besoins détectés à l'usage. Nous abordons également la question de l'adaptation de la TST à nos besoins.

Abstract.

Lexical resource building and maintenance based on the use

Our company develops a semantic search engine based on queries rephrasing. Our search engine relies on a lexicon we built on the basis of the Meaning-Text theory. We introduce our lexical resource and explain how we enrich and update it according to the needs we detect. We also mention the customization of the Meaning Text Theory to our needs.

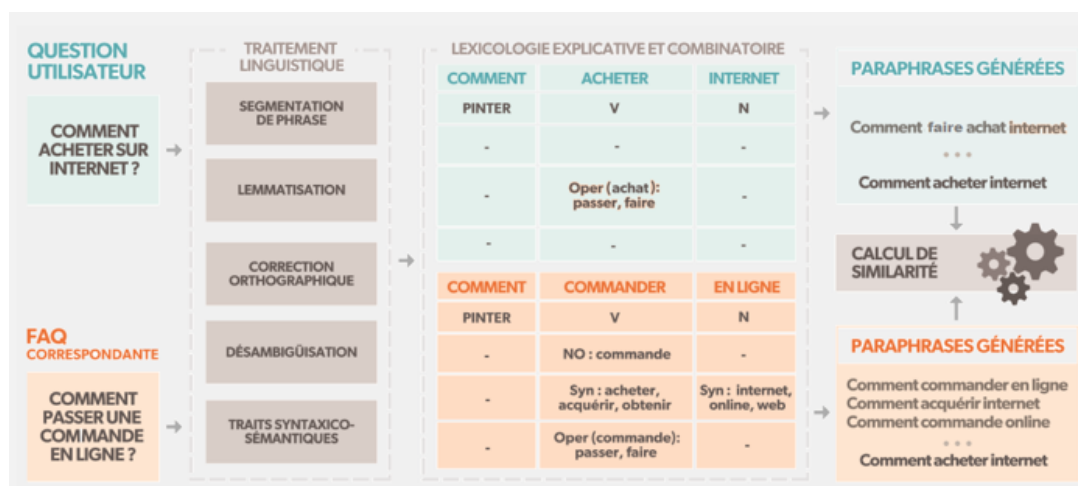
Mots-clés : Ressources lexicales, Théorie Sens-Texte, Recherche d'Information.

Keywords: Lexical Ressources, Meaning-Text Theory, Information Retrieval.

1 Contexte

Notre société a développé son propre Moteur de Recherche Sémantique. Nous utilisons notamment ce moteur dans une Foire Aux Questions (FAQ) dynamique. Une FAQ dynamique est un outil permettant d'accéder à une liste de questions-réponses, notamment en posant une question en langage naturel dans un moteur de recherche.

Pour cela, le moteur procède à l'analyse de la question utilisateur afin d'en produire une représentation sémantique. Dans l'exemple ci-dessous la question « Comment acheter sur internet ? » est lemmatisée et la préposition est supprimée. Ensuite une paraphrase de la question est générée grâce à la fonction lexicale *Oper* (présente dans notre ressource linguistique) qui indique que « faire un achat » a le même sens qu'« acheter ». Parallèlement, nous générons un ensemble de reformulations des titres des FAQs (questions-réponses) à l'aide des relations stockées dans notre ressource : le lexique. Ci-dessous, en orange, des paraphrases du titre de la FAQ sont générées à partir des Fonctions Lexicales associées aux lemmes (formes canoniques) de la question. Enfin, nous comparons ces différentes reformulations à la représentation sémantique de la question utilisateur, ce qui nous permet de retrouver les FAQs dont les titres sont les plus similaires à la question utilisateur. Il s'agit du calcul de similarité.



Comme on le voit dans ce schéma, le cœur du fonctionnement du moteur de recherche repose sur le lexique. Il est donc essentiel de disposer de ressources de « qualité ». Comme le souligne (Saint-Dizier, 2002) la qualité d'une ressource lexicale dépend de son adéquation aux besoins qui sous-tendent son utilisation. *« A défaut d'une analyse fine et précise des besoins, il revient donc au concepteur de ressources de faire lui-même les choix, peut-être de façon arbitraire, et ensuite de tenter de les faire correspondre aux intuitions et aux besoins des utilisateurs. Une attitude prudente consiste à développer des ressources cohérentes, par rapport à sa propre vision théorique et technique, puis de mettre en œuvre des ponts, dits d'insertion lexicale (Strigin, 1998), qui sont autant de canaux qui vont filtrer et reformater ces ressources selon les besoins des applications. ».*

Notre besoin est donc de disposer d'une ressource nous permettant de formaliser le sens d'un énoncé et de générer des reformulations de ce sens. C'est pourquoi nous avons choisi de construire un lexique inspiré de la Théorie Sens-Texte (TST) puisque cette théorie fournit un modèle de traduction du sens en énoncés (textes).

Nous aborderons ici les questions de construction et maintenance de nos ressources lexicales avant de présenter les principes généraux de la TST. Nous verrons ensuite dans quelle mesure nous avons adapté cette dernière.

2 Les méthodes de constitution du lexique

Afin de disposer d'une base de départ nous avons importé un lexique existant pour le français que nous avons retravaillé par la suite. Nous avons limité notre lexique aux domaines qui nous concernent, à savoir la banque et l'assurance. Nous ne visons pas la description exhaustive du français, c'est pourquoi nous avons éliminé certains termes non pertinents pour nos domaines afin de limiter les ambiguïtés.

2.1 Enrichissement du lexique

Pour chaque nouveau site sur lequel nous intégrons une FAQ dynamique nous créons un « projet ». Chaque projet a sa propre base de questions-réponses et son lexique. Ce dernier est constitué d'une base commune à l'ensemble des projets ainsi que d'éléments spécifiques au projet. En général, l'enrichissement du lexique coïncide avec l'arrivée de nouveaux projets et de nouvelles thématiques. Ajouter les termes au fur et à mesure nécessite une réactivité certaine pour maintenir les performances de notre moteur mais nous permet en contrepartie de démarrer rapidement un projet et de constituer un lexique adapté, répondant exactement à nos besoins.

Lorsque de nouvelles questions font ressortir des termes absents du lexique, nous nous appuyons sur l'expertise de nos clients pour intégrer ces termes et définir les relations qu'ils entretiennent avec les termes déjà présents. En effet, nos clients sont capables de nous indiquer si la question X fait référence à la FAQ Y. Les linguistes enrichissent donc le lexique sur la base de cette expertise ainsi que de leurs propres intuitions, suivant une démarche introspective qui sera par la suite validée et complétée par une analyse de corpus.

De plus, nous restons informés des actualités du domaine banque-assurance pour ajouter les termes et relations en rapport avec les nouveautés de ce domaine. Nous avons par exemple enrichi notre lexique de termes relatifs à l'ANI (L'Accord National Interprofessionnel) du 11 janvier 2013, qui généralise la complémentaire santé pour tous les salariés.

Nous repérons les failles du lexique à travers des tâches de maintenance, comme nous le verrons dans le 3.2, ou bien ponctuellement aux cours de nos analyses des performances du moteur basées sur les notions de bruit et de silence.

Pour les questions restées sans réponse (silence), soit il s'agit d'un manque dans nos FAQs (aucune FAQ ne traite de cette thématique), soit le moteur n'a pas été capable de ramener la FAQ pertinente. Dans le second cas nous pouvons être amenés à modifier ou ajouter des relations lexicales. Par exemple, lorsque nous repérons une formulation utilisant des synonymes des mots de la FAQ, nous devons alors ajouter ces synonymes ou bien seulement ajouter une relation de synonymie entre ces mots s'ils sont déjà présents dans le lexique.

En l'absence de gold standard, nous définissons la notion de bruit comme suit : lorsque des FAQs ramenées pour une question ne sont pas cliquées (et donc pas consultées), nous en déduisons qu'elles n'étaient probablement pas pertinentes. Il faudra alors effectuer des ajustements dans le lexique, par exemple en réduisant l'importance (« poids sémantique ») d'un mot qui génère du bruit. Dans ce sens, nous avons minoré l'importance du verbe « vouloir » qui a, par conséquent, un poids faible dans le calcul de similarité. Nous évitons ainsi qu'une FAQ comme « Je *souhaite* changer de formule. Comment faire ? » remonte pour les questions du type « je *veux* résilier mon assurance » (souhaiter et vouloir étant synonymes).

L'enrichissement du lexique peut également se faire à travers des analyses de corpus. Par exemple, nous avons procédé à l'analyse des cooccurrences au sein des questions utilisateurs. Cette analyse, visant à repérer les termes ou groupes de termes qui apparaissent au sein des mêmes contextes, nous a permis d'identifier qu'une des relations présente dans la TST serait pertinente au sein de notre lexique. Il s'agit de la relation entre « faire un paiement » et « payer », intitulée « *Oper* ». Ainsi, nous faisons appel à des scripts automatiques pour enrichir notre lexique. Cependant, une étape de vérification manuelle est ajoutée afin de garantir la qualité de la ressource.

2.2 Maintenance

Nous avons mis en place un système de maintenance reposant sur des tâches effectuées régulièrement par les linguistes de la société. Nous utilisons des outils spécifiques pour repérer automatiquement les anomalies au sein du lexique. Le lexique est stocké sous forme de base de données SQL. Nous pouvons extraire tous les lemmes qui n'ont pas de valeur associée pour un type de relation. Cela nous permet de repérer notamment les verbes qui n'auraient pas de participe passé associé via la relation *PartVerb*.

Dans le même esprit, un autre outil contrôle que le nombre de formes associées aux lemmes des verbes est supérieur à minimum défini et repère ainsi d'éventuels manques.

Dans le but de travailler sur les ambiguïtés catégorielles, nous extrayons tous les mots reconnus comme identiques après normalisation des caractères. Ainsi, les formes identiques associées à des lemmes différents sont listées avec les catégories grammaticales de leurs différents lemmes. Nous disposons de ce fait de la liste de tous les mots grammaticalement ambigus. Cette réserve de cas spécifiques nous sert à rechercher des exemples pour tester et améliorer nos règles de désambiguïsation. Nous pouvons, par exemple, extraire tous les mots ayant l'ambiguïté Verbe vs Nom vs Participe passé. En lançant une recherche sur ces mots parmi les questions utilisateurs, nous obtenons une liste des contextes d'apparitions de l'ambiguïté.

D'autres outils utilisés à l'échelle du projet permettent de repérer les mots de la base de questions-réponses du projet qui sont absents du lexique. Nous détectons ainsi de nouveaux mots à ajouter à notre lexique ainsi que des relations à ajouter avec les lemmes existants.

A l'inverse, nous pouvons voir parmi les questions utilisateurs quels sont les mots absents du lexique. Dans ce cas, cela signifie, soit qu'il faut ajouter un mot, soit qu'un mot a été mal écrit et n'a pas été géré par le correcteur orthographique. Nous améliorons alors notre système de correction orthographique.

3 La lexicologie explicative et combinatoire en théorie vs. en pratique

3.1 Introduction à la TST

Comme nous l'avons expliqué précédemment, le moteur de recherche génère des reformulations sur la base de relations entre les mots. La théorie Sens-Texte (TST) de I. Mel'čuk (Mel'čuk, 1995) propose un modèle dit « traductif » qui permet de mettre en correspondance des représentations sémantiques avec toutes les représentations phoniques qui peuvent les exprimer dans une langue donnée. Il s'agit donc d'un modèle reposant sur le paraphrasage.

Pour permettre ces mises en correspondance, Mel'čuk a produit un dictionnaire explicatif et combinatoire du français contemporain (DEC) dans lequel il décrit les lexies (lemmes) à différents niveaux. Ces niveaux, appelés « zones » permettent la description exhaustive des unités : zone phonologique, syntaxique, sémantique, de combinatoire lexicale, de combinatoire syntaxique, d'exemple et de phraséologie.

La zone de combinatoire lexicale permet le choix du mot juste, notamment à travers l'usage des Fonctions Lexicales (FL) « *La vocation des fonctions lexicales est de fournir au locuteur la totalité des moyens lexicaux nécessaires à l'expression la plus riche, la plus variée et la plus complète de la pensée et, en même temps, de garantir le choix le plus précis de la formulation appropriée. En d'autres termes, les FL [...] alimentent un système puissant de paraphrasage, qui est à la fois une raison d'être des FL et un outil fondamental de leur vérification.* » (Mel'čuk, 1995)

Exemple : 'synonyme'(accuser) = inculper

Les FL syntagmatiques relient les lexies apparaissant en relation de cooccurrence alors que les FL paradigmatiques relient les lexies qui sont liées par des relations sémantiques communes. En d'autres termes, les FL syntagmatiques visent la

combinatoire lexicale (un élément de la valeur de la FL est utilisé **à côté de** son mot-clé) tandis que les FL paradigmatiques visent la sélection lexicale (un élément de la valeur de la FL est utilisé **à la place** de son mot-clé). (Mel'čuk, 2003)

3.2 Notre utilisation de la TST

Notre société a fait le choix de s'intéresser uniquement à la zone de combinatoire lexicale, notre but n'étant pas de capitaliser un nombre maximal d'informations sur les lexies contrairement à l'objectif de Mel'čuk lors de la création de son D.E.C. Cette zone permet le choix du mot juste et la production de la combinaison lexicale adéquate dans un paradigme sémantique donné.

Notre lexique s'inspire de la TST puisqu'il reprend le concept des Fonctions Lexicales (FL) pour le paraphrasage. Il réutilise 5 FL telles qu'elles sont décrites dans le DEC (*Phrase*, *Loc*, *Neg*, *Oper*, *Real*) et en adapte d'autres telles que *Syn 0*, *Syn 1* et *Syn 2*. Nous utilisons également *OperInv*, adaptation de la FL *Oper* pour les formes passives. Les relations *Nominalisation* et *Verbalisation* ont été regroupées en une seule relation réciproque « *NounVerb* ».

5 FL supplémentaires ont été créées pour nos besoins :

- *AddSic* permet de contracter les expressions multitermes : retard + paiement = impayé.
- *ClearSic* permet de réduire une expression en ignorant les mots non porteurs de sens : merci + d'avance → merci.
- *Origen Final* permet de matérialiser un lien de cause à effet ; lorsqu'un utilisateur utilise le lemme « trompé » (ex : « je me suis trompé dans ma date de naissance ») la finalité de sa question est de savoir comment rectifier son erreur. Il faut donc lui répondre par une FAQ concernant la modification des informations personnelles. Par conséquent, nous associons « modifier » à « trompé » via la fonction *Origen Final*.
- *Local FL* est une FL « joker » permettant d'utiliser, au sein du projet seulement, un FL ne répondant pas aux propriétés des FL existantes.
- *PartVerb* : matérialise le lien réciproque entre un Verbe et son participe passé. En effet, il était important pour nous de pouvoir distinguer les verbes de leurs participes passés : la question « comment bloquer ma carte » est relative au fait de faire opposition sur sa carte, alors que « ma carte est bloquée » fait référence à un problème de fonctionnement de la carte. Nous avons donc créé deux catégories distinctes. Nous souhaitons tout de même conserver un lien entre le verbe et le participe car dans certaines situations ils sont utilisés indifféremment. C'est pourquoi nous avons créé la relation *PartVerb*.

Comme évoqué précédemment, notre lexique vise la description limitée des domaines qui nous intéressent (principalement la banque et l'assurance) dans la limite des mots effectivement employés dans nos projets afin d'éviter de générer des ambiguïtés qui n'ont pas lieu d'être au sein d'un domaine spécifique. Par exemple, lorsque nous rencontrons le mot « franchise », nous savons que dans le cadre des assurances il ne s'agit pas de « sincérité » mais bien de la somme restant à la charge de l'assuré dans le cas où survient un sinistre. Inutile donc d'ajouter deux versions du lemme « franchise », qu'il faudrait ensuite désambigüiser.

Cependant, certaines ambiguïtés persistent même en se limitant au domaine banque-assurance. Elles sont alors gérées par des règles de désambigüisation par mot. Ces règles, sous forme de patrons, utilisent les lemmes du contexte précédant ou suivant le mot ambigu. Par exemple, la forme « suis » existe pour les verbes « être » et « suivre ». Il ne s'agit pas d'une ambiguïté catégorielle, il faut donc créer une règle spécifique pour ce mot.

A l'issue d'une analyse de corpus effectuée dans les questions utilisateurs (requêtes) de nos projets, nous avons constaté que « suis » est presque toujours utilisé pour le verbe « être ». Peu d'occurrences de « suivre » ont été constatées pour la forme « suis » : « suivre une procédure » et « suivre des instructions ». Notre règle de désambigüisation par mot indique que « suis » devant un déterminant puis le lemme « procédure » sera identifié comme « suivre », idem pour « suis » devant un déterminant puis le lemme « instruction ». Par défaut, « suis » sera identifié comme « être ».

1. [Ambigüité] + [*|DET] + [procédure|N] = [suivre|V]

2. [Ambigüité] + [*|DET] + [instruction|N] = [suivre|V]

3. Conditionnel par défaut = [etre|V]

Cette règle, bien trop simpliste pour un corpus de textes tout-venant, fonctionne bien dans le cadre limité de nos projets. Une fois de plus, notre lexique est adapté aux besoins spécifiques de nos projets et non à l'ensemble des textes du français.

De plus, nous adaptons certaines FL à nos besoins. Par exemple, la FL « synonyme » proposée par Mel'čuk distingue les synonymes absolus des synonymes plus spécifiques (notion d'hyponymie), des synonymes moins spécifiques (hyperonymie) et des synonymes à intersection. Alors que nous distinguons nos synonymes par degré de proximité (niveaux 0, 1 et 2). Les *syn 0* sont exactement substituables, les *syn 1* ont un sens proche, les *syn 2* ont des sèmes communs mais entretiennent une relation plus lointaine que les *syn 1*. Ces degrés de proximité subissent un traitement distinct lors du calcul de similarité entre un titre de FAQ et une question utilisateur. En effet, le score de similarité sera plus élevé pour un *syn 0* que pour un *syn 1*, et plus élevé pour un *syn 1* que pour un *syn 2*. Cela traduit une confiance plus grande dans le score de similarité pour le *syn 0* que pour le *syn 2*.

Afin de déterminer ce degré de proximité, nous testons la pertinence de la substitution d'un lemme par son synonyme dans nos différents projets.

- Si la substitution est pertinente dans tous les cas nous utilisons la relation *syn 0*. Par ex. : « deuxième » est un *syn 0* de « second ».
- Si elle fonctionne dans la majorité des cas mais peut interférer avec d'autres FAQs, nous utilisons la relation *syn 1*. Par ex. : « annuler » est un *syn 1* de « résilier » car nous rencontrons des formulations du type « Comment annuler mon assurance ? ». Cependant, pour les questions contenant « annuler », il n'est pas souhaitable que les FAQs sur la résiliation obtiennent un score de similarité plus important que les FAQs du type « Comment annuler la modification de mon assurance ? ».
- Si le cas est plutôt incertain nous utilisons la relation *syn 2* qui permettra de réduire le silence sans trop risquer de produire du bruit, étant donné son score plus faible lors du calcul de similarité. C'est ce que nous faisons pour « devis » avec son *syn 2* « tarif ». Nous voulons en effet ramener la FAQ « Comment obtenir un devis d'assurance auto ? » pour les questions du type « Quels sont les tarifs de l'assurance auto ? ». Cette relation relève de la connaissance du monde mais ne se vérifie pas d'un point de vue strictement linguistique, d'où l'usage de la relation *syn 2*.

4 Structure de notre lexique

4.1 Le lexique en quelques chiffres

Nous distinguons les lemmes (formes canonique considérées comme des concepts), des formes qui leurs sont associées, que nous appelons « mots ». Au lemme « payer » sont associés les mots « payer, payons, payez, payent » etc.

Dans notre lexique français nous avons actuellement :

- 23013 lemmes et 163337 mots
- 52170 relations lexicales.
- Une moyenne de 30 mots locaux (mots spécifiques à un projet) par projet
- Une moyenne de 85 particularisations/projet (mots particularisés et relations locales, cf. 4.2)

Pour chaque mot (forme) du lexique nous disposons du lemme correspondant associé à sa catégorie grammaticale ainsi que de la catégorie sémantique du mot. La catégorie sémantique permet de graduer le poids du mot dans le projet, c'est-à-dire l'importance qui lui sera accordée lors du calcul de similarité entre question utilisateur et titre de FAQ.

Pour chaque lemme nous disposons de sa catégorie grammaticale ainsi que des valeurs qui lui sont associées pour l'ensemble des Fonctions Lexicales utilisées dans le lexique.

Notre société gère plus de 20 langues dont le catalan, le castillan, le galicien, le basque, l'anglais, le français, le portugais, l'italien, l'allemand, le néerlandais, le russe, le turque. Chacune de ces langues possède un lexique dédié. Une détection de la langue est faite en entrée, ce qui permet de déterminer le lexique à utiliser. A ce jour, les lexiques sont indépendants les uns des autres.

4.2 Spécificités

Comme évoqué précédemment, nous devons adapter notre lexique aux spécificités des domaines considérés ainsi qu'à celles du projet concerné. Ainsi, le lexique est organisé à différents niveaux : nous distinguons un dictionnaire général partagé par l'ensemble des projets, des dictionnaires locaux, qui sont propres à chaque projet. Un lemme peut être ajouté spécifiquement dans un projet et rester absent des autres projets (éventuellement pour ne pas produire d'ambiguïté inutile). Il s'agit alors d'un lemme « local ». C'est le cas notamment des noms de produits spécifiques à un projet.

Il est également possible d'ajouter des particularisations. Il s'agit d'attribuer des caractéristiques et des relations spécifiques à un mot au sein d'un projet alors que ce terme est utilisé dans l'ensemble des projets avec des caractéristiques différentes.

Par exemple, dans un projet dédié exclusivement aux assurances nous pouvons créer une relation de synonymie *syn I* entre « relevé d'information » (attestation de bonus/malus) et « relevé » car dans la majorité des cas, les utilisateurs emploient « relevé » pour « relevé d'information », alors que dans un projet banque ou banque/assurance, cette relation risque de poser problème car « relevé » peut également faire référence au « relevé de compte ». « Relevé d'information » aura donc un *syn I* « relevé » uniquement dans le projet d'assurance. Dans le lexique général, partagé par les autres projets, il n'aura pas ce synonyme.

Des particularisations sont également effectuées sur la notion de « poids sémantique » traduisant l'importance d'un lemme dans le calcul de similarité entre FAQ et Question utilisateur. Dans le projet du groupe « Banque Accord » la plupart des noms de produit contiennent le mot « accord », comme par exemple « garantie hospitalisation Accord », « compte carte Accord ». Le mot « accord » revient donc très souvent et n'apporte pas de sens aux questions. Nous avons donc diminué son importance (« poids sémantique ») au sein du projet. Ainsi, il est peu pris en compte dans le calcul de similarité de ce projet mais conserve une importance normale dans les autres projets.

Cette structure par niveaux nous permet une grande flexibilité. Nous sommes ainsi en mesure de modéliser le lexique précisément en fonction des besoins du projet considéré sans perturber d'autres projets, tout en construisant une base commune qui sert de socle à l'ensemble des projets.

En plus de ces niveaux (local vs général), nous avons choisi une organisation en conformité avec la structure du langage telle que décrite dans la linguistique structuraliste. Nous reprenons la distinction Saussurienne entre axe paradigmatique et syntagmatique pour qualifier nos différents types de relations entre lemmes. Nous sommes ainsi en mesure d'opérer un traitement différent au moment de la reformulation des questions.

Les fonctions paradigmatiques, qui concernent la sélection des mots, incluent la synonymie (divisée en degré de proximité), la relation *PartVerb* entre un Verbe et son participe passé, la relation *Nominalisation/Verbalisation* entre un nom et son déverbal. Au niveau du calcul de similarité entre questions utilisateurs et FAQ, ces FL permettent de générer une paraphrase par substitution d'un lemme par un autre.

Les relations syntagmatiques sont celles entretenues par les termes dans leur enchaînement au sein de la phrase. Lors de la reformulation des FAQs, ces FL permettent soit, de contracter une expression : remplacer plusieurs lemmes par un seul ou au contraire développer une forme contractée : remplacer un lemme par plusieurs lemmes. Nous avons, par exemple, la relation *Oper* : « faire un paiement » est substituable par « payer », ou bien la relation *Neg* : « rappelle plus » est substituable par « oublié » (ex : « j'ai oublié mon mot de passe » sera reformulé en « je ne me rappelle plus de mon mot de passe »). On remplace donc un lemme par plusieurs ou inversement.

5 Conclusions et perspectives

Nous venons d'évoquer un cas de construction de ressource lexicale adaptée à un usage spécifique. Ici, nous mettons en pratique une formalisation théorique du langage. Le passage à la pratique nécessite certaines adaptations et des ajustements permanents.

Nous avons notamment insisté sur la différence fondamentale entre une ressource visant la description exhaustive du langage et une ressource visant la description limitée des domaines de la banque et de l'assurance, dans la limite des mots effectivement employés dans nos projets. Ainsi, la maintenance et l'enrichissement de la ressource sont continus et nous permettent d'avoir une ressource de plus en plus performante.

Grâce à une structuration à différents niveaux, nous avons doté notre ressource d'une grande flexibilité. Elle s'adapte ainsi aux spécificités de chaque projet dans lequel elle intervient.

Notre ressource continue aujourd'hui de s'enrichir et nous sommes en recherche perpétuelle de moyens de l'affiner et de la compléter.

Références

MEL'CUK, I., POLGUERE, A. (1995) *Introduction à la lexicologie explicative et combinatoire*. Bruxelles : Duculot.

MEL'CUK, I. (2003) Collocations dans le dictionnaire, *Les écarts culturels dans les Dictionnaires bilingues*, Paris: Honoré Champion, p.19-64,

SAINT-DIZIER, P. (2002) Quelques défis et éléments de méthode pour la construction de ressources lexicales sémantiques, *Revue française de linguistique appliquée* I (Vol. VII), p 39-51

STRIGIN, A. (1998) : Lexical Rules as Hypothesis Generators. *Journal of Semantics* 15.

Utilisation d'annotations sémantiques pour la validation automatique d'hypothèses dans des conversations téléphoniques

Carole Lailler¹ Yannick Estève¹ Renato de Mori^{2,3} Mohamed Bouallègue¹ Mohamed Morchid²
(1) LIUM — Université du Maine, France
(2) LIA — Université d'Avignon et des Pays de Vaucluse, France
(3) McGill University, Montréal, Canada

Résumé. Les travaux présentés portent sur l'extraction automatique d'unités sémantiques et l'évaluation de leur pertinence pour des conversations téléphoniques. Le corpus utilisé est le corpus français DECODA. L'objectif de la tâche est de permettre l'étiquetage automatique en thème de chaque conversation. Compte tenu du caractère spontané de ce type de conversations et de la taille du corpus, nous proposons de recourir à une stratégie semi-supervisée fondée sur la construction d'une ontologie et d'un apprentissage actif simple : un annotateur humain analyse non seulement les listes d'unités sémantiques candidates menant au thème mais étudie également une petite quantité de conversations. La pertinence de la relation unissant les unités sémantiques conservées, le sous-thème issu de l'ontologie et le thème annoté est évaluée par un DNN, prenant en compte une représentation vectorielle du document. L'intégration des unités sémantiques retenues dans le processus de classification en thème améliore les performances.

Abstract.

Use of Semantic Annotations for Validating Mentions of Semantic Hypotheses in Telephone Conversations.

The presented work focuses on the automatic extraction of semantic units and evaluation of their relevance to telephone conversations. The corpus used is DECODA corpus. The objective of the task is to enable automatic labeling theme of each conversation. Given the spontaneous nature of this type of conversations and the size of the corpus, we propose to use a semi-supervised strategy based on the construction of an ontology and a simple active learning : a human annotator analyses not only the lists of semantic units leading to the theme, but also studying a small amount of conversations. The relevance of the relationship between the conserved semantic units, sub-theme from the ontology and annotated theme is assessed by DNN, taking into account a vector representation of the document. The integration of semantic units included in the theme classification process improves performance.

Mots-clés : analyse de conversation humain/humain, extraction automatique d'unités sémantiques pertinentes, validation d'une ontologie.

Keywords: human/human conversation analysis, automatic extraction of relevant semantic units, ontology validation.

1 Introduction

Les travaux présentés concernent une analyse sémantique du corpus DECODA (<http://decoda.univ-avignon.fr/>). Il s'agit d'un corpus issu d'une application de conseil à la clientèle par téléphone mise en place en interne par la Régie Autonome des Transports Parisiens (dorénavant RATP). L'objectif est de permettre l'extraction automatique du thème de chaque conversation en utilisant un DNN (DNN pour « *deep neural networks* »). Toutefois, deux difficultés majeures demeurent avec ce corpus : d'une part, le nombre de conversations reste faible au regard des besoins d'apprentissage d'un DNN. D'autre part, les Word Error Rate (dorénavant WER) obtenus par les systèmes de reconnaissance de la parole sur ce type d'oral demeurent élevés. Ces deux constats nous ont conduits à utiliser une méthode semi-supervisée avec annotations pour faciliter la détection en thème. Dans ce cadre, l'article propose l'utilisation de réseaux de neurones profonds pour valider l'extraction des mentions sémantiques pertinentes selon les relations sémantiques établies dans l'ontologie construite après extraction automatique de ngrams. Il s'agit de valider la relation tripartite [ngram ; sous-thème ; thème] mise au point dans l'ontologie. Les DNN utilisés ont été inspirés par des travaux récemment publiés et portent sur l'extraction de relations dans le discours (Bost *et al.*, 2015) et (Ji & Eisenstein, 2014).

Il s'agit tout d'abord, d'obtenir une ontologie suffisamment générique et exhaustive pour faciliter l'extraction automatique d'éléments sémantiques pertinents permettant de conduire au thème de la conversation. Par ailleurs, chaque conversation doit faire l'objet d'un rapport final construit automatiquement consignait les éléments les plus informatifs des interactions Conseiller-Usager de la RATP. La documentation fournie par la société gestionnaire de l'application téléphonique contient des informations suffisantes pour en extraire une ontologie. Elle fait suite à l'extraction automatique de ngrams jugés pertinents, selon le critère de Gini (Breiman *et al.*, 1984). Elle permet de pallier le manque de données et d'organiser dans une structure suffisamment générique les ngrams extraits automatiquement. Elle étend également la recherche de ces ngrams selon des principes sémantiques et discursifs logiques. Elle est établie selon plusieurs niveaux de granularité (5 au total) en fonction du type d'informations recherchées. Elle doit permettre l'extraction d'éléments sémantiques pertinents conduisant non seulement au thème mais levant aussi les ambiguïtés de la conversation. Nous n'aborderons ici que les deux principaux niveaux sémantiques de l'ontologie, à savoir le niveau thématique et le niveau syntagmatique qui organise, selon des sous-thèmes, les ngrams les plus discriminants pour chaque thème. Le niveau dialogique fera l'objet d'une présentation ultérieure. Il s'agit pour l'instant de ne s'intéresser qu'aux éléments sémantiques permettant de conduire au thème de la conversation, indépendamment des tours de parole et de la fonction des locuteurs.

Le premier niveau est celui qui contient le thème de la conversation. Le second rassemble, en une série de vingt sous-thèmes, les éléments sémantiques les plus discriminants, ceux qui ont pour but d'inférer le thème. Ces sous-thèmes contiennent les éléments les plus signifiants au regard de l'application. Ils permettent de relier un thème aux unités sémantiques pertinentes, c'est-à-dire aux ngrams porteurs de l'information. Constitués d'un ngram, d'un sous-thème indiquant le champ sémantique abordé et d'un thème, cet ensemble tripartite, exprimé en langage naturel, est de taille variable : il va du mot au syntagme (jusqu'à six entités lexicales sans blanc) et peut intégrer des valeurs numériques (numéro de bus, de rue, etc.) et des entités nommées. Nous l'appellerons dorénavant TRSC pour « *theme-specific report component* ». Ainsi, le TRSC [mouvement social ; Grève ; ETFC] appartient au thème « ETFC » (État du trafic). Il a pour sous-thème « Grève » et pour mention sémantique pertinente « mouvement social ». La relation entre le sous-thème et la mention est le prédicat « perturber ». Ainsi, les deux arguments du prédicat sont des éléments issus de l'ontologie qui doivent être insérés dans le rapport.

Néanmoins, les ngrams porteurs de l'information conduisant au thème sont disséminés dans l'ensemble des énoncés d'une conversation. Compte tenu notamment du caractère spontané des échanges (importance des disfluences et des répétitions), il peut être difficile de les retrouver ou d'appliquer des bornes pour permettre leur identification. Pour cette raison, les méthodes généralement proposées pour l'identification en thème et en particulier pour l'identification des traits sémantiques ne sont pas appropriées. La relation exprimée par le prédicat n'est pas immédiatement appréhendable morphosyntaxiquement. Il est souvent difficile pour un agent de suivre le protocole fixé à l'avance. Le client adopte rarement un comportement prévisible et normé. Il tend à s'écarter du champ d'application se rapportant au domaine des transports. Toutes ces digressions perturbent la bonne marche de l'échange et retardent sa conclusion. Ainsi, les modèles issus de l'ontologie décrivant l'expression d'un TRSC sont prévus pour tenir compte non seulement des scénarii-protocoles mais également des aléas d'une conversation téléphonique. En se fondant sur la documentation de l'application et sur le protocole devant être suivi par l'agent, il est possible de détecter des TRSC pertinents qui conduiront au thème de la conversation et dont les sous-thèmes pourront servir de fils conducteurs dans l'élaboration du rapport.

2 Travaux connexes

Des résultats prometteurs ont récemment été obtenus sur l'extraction de relations unissant des entités exprimées dans une phrase (Mesquita *et al.*, 2013), selon une analyse de dépendance entre la phrase et les relations définies. Plus récemment, de nouvelles solutions ont permis l'évaluation de la cohérence dans des successions de phrases au sein d'un même texte (Li & Hovy, 2014). Ces solutions sont inspirées par la Théorie de la Structure Rhétorique (RST pour « *Rhetorical Structure Theory* ») (Mann & Thompson, 1988). La RST considère un texte comme cohérent s'il est composé d'unités de discours élémentaires (EDU), portées par des phrases et un petit ensemble de relations de discours typiques les unissant.

Toutefois les relations entre les unités sémantiques et les thèmes considérés dans cette étude diffèrent de celles prises en considération dans d'autres types d'extraction de relation. En effet, les ngrams candidats à l'élection d'un thème peuvent être utilisés dans des conversations renvoyant à un autre thème. Il peut arriver, par exemple, qu'un ngram se rattachant à une notion temporelle, déclarée comme TRSC pour le thème HORR (Horaires) apparaisse dans une conversation dont le thème principal est ITNR (Itinéraire). C'est le cas notamment en fin de conversation quand le client souhaite s'assurer de son temps de trajet. Dans ce cas, la mention sémantique exprimée par le modèle n'a pas à être déclarée comme un TRSC pertinent pour le thème principal de la conversation. De plus, les ngrams pertinents sont généralement disséminés dans l'ensemble des segments de conversation. Une réflexion menée sur l'utilisation de mots avec des liens de dépendance

à longue distance qui ne peuvent donc être retrouvées avec de simples analyseurs de phrases est présentée dans (Ji & Eisenstein, 2014) et (Prasad *et al.*, 2014).

Dans notre cas, le problème relève d'une difficulté supplémentaire, puisqu'au lieu d'utiliser un document sous forme d'un texte cohérent, nous analysons des conversations téléphoniques en langue spontanée au sein desquelles les interventions du client sont souvent imprévisibles, avec une grande variabilité morphosyntaxique et de nombreux "bruits" (disfluences, répétitions, etc.). Par ailleurs, il nous faut essayer de trouver une solution pour établir des relations pertinentes entre une mention sémantique locale exprimée par un ngram et un thème caractéristique qui reste commun à l'ensemble de la conversation. Les sous-thèmes ont justement pour but de permettre l'établissement de cette relation : en offrant la possibilité de diviser une conservation selon des unités sémantiques plus petites, ils soulignent la progression argumentative et thématique d'un échange en tenant compte des éléments de variabilité du discours. Les sous-thèmes permettent d'établir une connexion sémantique unique et nécessaire entre un ngram structuré et le thème de la conversation. La détection des TRSC est une étape qui va au-delà de la description d'un thème (Hazen, 2011), (Morchid *et al.*, 2014a), (Morchid *et al.*, 2014b). Le réseau de neurones, décrit dans (Estève *et al.*, 2015), est considéré ici comme le système de référence. Il sert de point de comparaison à notre étude. Nous le nommerons SystOne.

3 Domaine d'application

Les segments de dialogues considérés dans l'application de la RATP sont constitués d'un problème rencontré par le client, d'une phase de reformulation puis d'une réponse (ou de bribes de réponse) apportée par l'agent. Ce dernier tente de résoudre le problème posé tout en suivant un protocole prédéfini. Le problème est formulé de telle façon que les mentions sémantiques révélant le thème sont livrées tout au long de l'échange selon un principe de pertinence. Ainsi, l'énoncé du problème et sa réponse sont reliés par des relations de continuité discursive. Le discours se veut généralement collaboratif entre les deux parties. Après avoir reformulé en des éléments clairs et concis le problème de l'utilisateur, le conseiller cherche à rapidement apporter une réponse. Suivre les protocoles imposés par la RATP induit également des types de réponse. L'ontologie créée prend en compte ces contingences discursives et les utilise. Elle se fonde sur la documentation fournie par le service et son site internet mais aussi sur les relations qui se créent au cours des échanges entre les locuteurs. Elle prend la forme d'un graphe dont les nœuds représentent les unités sémantiques les plus discriminantes et essentielles. Ces nœuds sont unis aux thèmes par des liens qui représentent les relations sémantiques prédictives les plus couramment utilisées dans l'application. Le contenu du rapport est guidé par une stratégie de composition reposant sur une planification résultant du DNN. Cette stratégie permet alors la formulation d'hypothèses dévoilant les relations sémantiques menant au thème de l'échange.

L'ensemble des thèmes du domaine liés à l'application et leurs abréviations sont au nombre de douze : *ITNR* = itinéraire, *OBJT* = objets trouvés et perdus, *HORR* = horaires, *NVGO* = cartes de transport et abonnement, *VGC* = Vente Grand Compte (il s'agit d'un thème rassemblant les ventes aux entreprises et collectivités ainsi que les cartes des ayants-droits), *ETFC* = état du trafic, *TARF* = les tarifs, *PV* = les infractions, *OFTP* = l'offre de transport palliatif (en cas de travaux de réfection par exemple), *CPAG* = problème avec un agent, *JSTF* = les justificatifs de retard, *RETT* = les remboursements. Un treizième thème, *AAPL*, concerne les conversations qui font référence à un problème concernant non pas la RATP mais la SNCF (le réseau francilien étant partagé par ces deux compagnies). Enfin, les conversations hors domaine sont rassemblées sous le thème *NULL*. Des exemples sur la segmentation du dialogue et les détails de la détection des huit premiers thèmes peuvent être retrouvés dans (Morchid *et al.*, 2014a) et (Morchid *et al.*, 2014b).

Initialement, seuls les thèmes ont été annotés au sein des transcriptions manuelles. Concernant les conversations qui présentent dans leur déroulé plusieurs thèmes, un unique thème majoritaire a été retenu en se référant aux règles contenues dans la documentation de service. Toutefois, ces conversations multithèmes restent un frein à l'analyse automatique. Elles déclenchent de fausses alertes : des TRSC appartenant à d'autres thèmes sont ainsi détectés et conduisent à un mauvais étiquetage. Afin de réduire l'effort d'annotation des TRSC, une procédure semi-automatique, minimisant l'effort humain, est proposée. Elle consiste à utiliser l'ensemble des conversations du corpus d'apprentissage pour trouver automatiquement une liste de ngrams selon leur indice de pureté dans chaque thème. L'indice de pureté est calculé selon le critère de Gini (Breiman *et al.*, 1984). Cette liste est ensuite analysée par un expert humain qui sélectionne et étend dans les modèles de l'ontologie les ngrams les plus pertinents en les associant à un sous-thème afin de construire un TRSC efficace et unique. Il s'agit pour l'expert de nettoyer cette première liste automatique de ses scories et d'étendre ses éléments pour la rendre plus robuste. Ainsi, l'effort humain dépend principalement de la taille de la liste initiale. L'expert se sert également de ses connaissances sur l'application et de la documentation de service pour enrichir cette première liste automatique de mentions sémantiques à conserver. Les TRSC pourront être enrichis au fur et à mesure de l'acquisition de nouvelles données

afin de ne pas voir diminuer leur pureté : les nouveaux TRSC ainsi injectés permettent d'accroître la reconnaissance en thème et de réduire les confusions inter-thèmes.

Nous considérons $\Gamma_t = \{\gamma_{t,1}, \dots, \gamma_{t,i}, \dots, \gamma_{t,I_t}\}$ comme un ensemble de TRSC du thème t . Chaque TRSC $\gamma_{t,i}$ est exprimé par un ensemble de mentions $M_{t,i} = \{m_1^{t,i}, \dots, m_j^{t,i}, \dots, m_{J_{t,i}}^{t,i}\}$. Une mention $m_j^{t,i}$ est exprimée par un modèle de mentions qui peut contenir des mots, des déclinaisons de syntagmes, ou encore des mots entrecoupés par des éléments spécifiques et identifiables (entités nommées ou valeurs numériques). L'objectif poursuivi est d'obtenir un nombre suffisant d'exemples positifs et négatifs pour entraîner les DNN. Pour une conversation donnée, tous les ngrams issus de TRSC conduisant à un autre thème sont considérés comme des exemples négatifs. En revanche, tous les ngrams appartenant à des TRSC du thème annotés manuellement constituent potentiellement des exemples positifs. Cette hypothèse de travail a été validée en utilisant un échantillonnage aléatoire du corpus d'apprentissage contenant une proportion d'exemples négatifs suffisants par rapport aux exemples positifs. Une vérification en OUI/NON pour chacun des TRSC est ensuite effectuée : le OUI correspond à une adéquation entre le TRSC et le thème annoté, le NON est signe de fausse alarme.

Par ailleurs, il faut noter que les conversations qui ne contenaient aucun TRSC ont été isolées. Il s'agit de constituer un sous-ensemble suffisant et cohérent permettant d'effectuer un apprentissage actif simple : en l'occurrence, une analyse minutieuse des conversations qui ne contiennent aucun TRSC pour essayer de capturer de nouveaux ngrams porteurs d'information et les insérer ensuite dans une relation tripartite [ngram ; sous-thème ; thème]. En s'assurant de la robustesse des ngrams déjà engagés dans les niveaux de l'ontologie par une vérification manuelle binaire et en y ajoutant des ngrams issus des conversations étudiées selon leur thème principal, la stratégie d'apprentissage actif mise en place n'a nécessité que peu d'effort.

Une mention $m_j^{t,i}$ est une expression de $\gamma_{t,i}$ quand elle est pertinente avec le thème t dans une conversation annotée avec t . Sinon, il ne s'agit pas d'une expression de $\gamma_{t,i}$ et cela entraîne une ambiguïté dans les hypothèses formulées concernant le thème t . Comme cela a été observé empiriquement, la cooccurrence d'une mention $m_j^{t,i}$ et du thème correspondant est fréquente, une mesure d'ambiguïté peut être estimée en suivant l'entropie conditionnelle suivante :

$$H[t|m_j^{t,i}] = -P[t|m_j^{t,i}] * \log P[t|m_j^{t,i}] \quad (1)$$

Une mention $m_j^{t,i}$ provoque peu d'ambiguïté si elle a un haut degré de pureté $P[t|m_j^{t,i}]$. Les mentions sémantiques pertinentes sont sélectionnées manuellement au sein d'une liste L_t de candidats issus d'une procédure de sélection automatique, qui classe des ngrams de mots en fonction de leur pureté par rapport au thème t . La sélection est effectuée et complétée par un expert humain selon une méthode semi-supervisée simple, en utilisant les connaissances recueillies auprès de la documentation de service. L'expert humain introduit également des déclinaisons de syntagmes, des mots sélectionnés pour leur unicité voire des séquences de mots exprimant des domaines sémantiques identifiés, par exemple, l'heure, la date, la localisation, le type d'objets perdus, etc. Un ngram $m_j^{t,i}$ est exprimé dans une conversation d par une instance $m_j^{t,i}(n, d)$ débutant avec le ngram qui doit conduire à un unique thème de la transcription de d .

4 Détection des relations de discours

Récemment, des architectures de réseaux de neurones ont été utilisées pour extraire des relations de discours en ayant recours à la distribution du texte et à son évolution paragraphe après paragraphe (Li & Hovy, 2014). Par ailleurs, (Ji & Eisenstein, 2014) et (Prasad *et al.*, 2014) ont démontré que les caractéristiques syntaxiques ne peuvent suffire à elles-seules pour capturer le contenu sémantique d'un document, notamment en raison de "réalisations lexicales alternatives" (Prasad *et al.*, 2014). C'est la raison pour laquelle il a été proposé d'utiliser une représentation des documents par sacs de mots, censée permettre de révéler des caractéristiques cachées idoines qui identifient les relations de discours. Compte-tenu de l'application utilisée et des conversations humain/humain obtenues, l'objectif est ici de détecter la relation de pertinence entre un TRSC issu de la liste obtenue de manière semi-supervisée et le thème annoté en amont.

Un DNN est proposé pour évaluer la relation de pertinence

$\mathcal{R}_{\text{pertinence}}[m_j^{t,i}(n, d), t]$ entre une mention $m_j^{t,i}(n, d)$ et un thème t . Une telle relation de discours valide la relation de pertinence, représentée telle que $m_j^{t,i}(n, d) \rightarrow \gamma_{t,i}$, conduisant de la mention au thème. Ceci valide alors l'inférence suivante :

$$\mathcal{R}_{\text{pertinence}}[m_j^{t,i}(n, d), t] \Rightarrow [m_j^{t,i}(n, d) \rightarrow \gamma_{t,i}] \quad (2)$$

Les entrées du DNN sont des représentations continues de TRSC de $m_j^{t,i}(n, d)$ obtenues en sommant les vecteurs de mots

les composant, un vecteur, appelé C-vector, représentant la totalité de la conversation (dont le calcul est décrit dans (Morchid *et al.*, 2014a)) et les scores de pureté de $m_j^{t,i}(n, d)$. Ces scores sont obtenus sur les données du corpus d'apprentissage pour chaque thème en ajoutant le vecteur des scores calculés avec d pour chaque thème (en utilisant le système SystOne). Le vecteur présenté en entrée du réseau résulte d'une concaténation de ces éléments. Pour cette étude, des représentations continues de mots sur 100 dimensions ont été calculées à partir d'un corpus conséquent, d'environ 2 milliards de mots. Ce corpus a été élaboré à partir de l'outil word2vec, décrit dans (Mikolov *et al.*, 2013), en utilisant les articles du quotidien français "Le Monde", le corpus Gigaword en français, des articles de Google News et des transcriptions manuelles de journaux télévisuels français, à hauteur de 400 heures d'enregistrement.

La sortie du DNN est un score mesurant la validité de l'inférence $\mathcal{R}_{pertinence}[m_j^{t,i}(n, d), t] \Rightarrow [m_j^{t,i}(n, d) \rightarrow \gamma_{t,i}]$. Le DNN est entraîné avec les données du corpus d'apprentissage en réglant la sortie pour s'approcher au plus près de l'annotation de référence de la conversation d sur le thème t , pour lequel $[m_j^{t,i}(n, d) \rightarrow \gamma_{t,i}]$. La sortie doit s'approcher de zéro quand $[m_j^{t,i}(n, d) \rightarrow \gamma_{\bar{t},i}]$ où \bar{t} ne constitue pas le thème avec lequel la conversation d a été annotée. L'évaluation sur les transcriptions manuelles du corpus de développement a montré que 91,3% des décisions prises par le DNN sont pertinentes. Ce pourcentage descend à 88,2% sur les sorties issues d'un système de reconnaissance de la parole. Compte tenu du taux de WER relativement élevé pour ce type de conversations, l'écart entre les résultats obtenus sur des transcriptions manuelles ou des sorties de système est faible : les décisions du DNN restent fiables. Ainsi, sur les 636 occurrences de ngrams appartenant à la première liste de 365 éléments détectés dans le corpus de développement, 389 sont conservées et utilisées par le DNN. Une autre évaluation a été effectuée pour établir la pertinence de la conversation d dans sa totalité en prenant en compte les mentions annotées avec le même TRSC $\gamma_{t,i}$ pour la liste de thème t . Ce type d'évaluation peut s'avérer utile pour lever les erreurs de thème et les hypothèses de TRSC erronées, ainsi que pour révéler la présence éventuelle de plusieurs thèmes dans une conversation. Plusieurs TRSC conduisant à des thèmes différents peuvent en effet entrer en conflit au sein d'une même conversation.

Considérons

$$S_{t,i}^d(n_1, q_{t,i}) = [m_1^{t,i}(n_1, d), \dots, m_j^{t,i}(n_j, d), \dots, m_{Q_{t,i}}^{t,i}(n_{Q_{t,i}}, d)] \quad (3)$$

comme étant la séquence pleine et entière de toutes les instances des mentions de $\gamma_{t,i}$ détectées dans une conversation d , en commençant avec une mention à la position n_1 et contenant $q_{t,i}$ mentions. La pureté de $S_{t,i}^d(n_1, q_{t,i})$ correspond à la probabilité de $P[t|S_{t,i}^d(n_1, q_{t,i})]$. L'approximation suivante est utilisée pour son calcul :

$$P[t|S_{t,i}^d(n_1, q_{t,i})] \approx P[t|S_{t,i}^d(n_1, q_{t,i})] \quad (4)$$

La stratégie conduit à construire une séquence de mentions $S_{t,i}^d(n_1, q_{t,i})$ si au moins un ngram possède un haut degré de pureté avec t . Une heuristique destinée à opérer des corrections est utilisée. Elle se fonde sur l'inférence logique suivante : SI {la pureté de $S_{t,i}^d(n_1, q_{t,i})$ est élevée et que t n'est pas annoté automatiquement pour d ET qu'il n'existe aucun ngram pertinent appartenant à un TRSC pour le thème annoté automatiquement pour d } ALORS { t remplace le thème annoté automatiquement pour la conversation d }. Il s'agit, ce faisant, de lever les erreurs d'annotations en thème en utilisant un haut degré de précision dans la détection des ngrams pertinents.

5 Expériences

L'objectif de ces expériences est d'évaluer la robustesse de la méthode semi-supervisée proposée. Le corpus utilisé pour les expériences comprend 2109 conversations.

Un sous-ensemble du corpus d'apprentissage contenant 600 conversations sur les 1 489 disponibles a été annoté automatiquement en utilisant 345 ngrams sélectionnés avec la procédure semi-automatique décrite en section 3. Ces ngrams sont inclus dans un TRSC, selon un jeu de 20 sous-thèmes possibles connectés à 13 thèmes appartenant à l'application, auxquels s'ajoute le thème NULL. Chaque ensemble tripartite est unique et ne permet de conduire qu'à un seul thème. Les sous-thèmes dévoilent les arguments les plus couramment utilisés dans une conversation. Toutes les conversations qui ne contiennent pas de TRSC sont étiquetées NULL. Sur ces 600 conversations, plus de 60% lèvent une alarme pour au moins un TRSC appartenant au thème annoté, 11% constituent des exemples négatifs sur lesquels entraîner un DNN. Les 29% restants sont des conversations qui ne contiennent aucune mention pertinente pour le thème annoté. Ces conversations ont donc formé un sous-ensemble destiné à être analysé par un expert humain selon un processus d'apprentissage actif, qui a permis d'ajouter 435 TRSC et d'obtenir une liste finale de 780 occurrences tripartites. Cette méthode présente l'intérêt d'être économe dans la mesure où, à partir d'un nombre relativement restreint de données et d'un apprentissage

semi-supervisé simple, on aboutit à un résultat. Toutefois, l'évaluation de ce résultat nécessiterait des éléments de comparaison voire une métrique ajustée. Cette dernière permettrait d'associer au volume de données, le temps de construction de l'ontologie et celui d'annotation pour mieux mesurer le gain par rapport à une seule extraction automatique de ngrams selon des critères de pureté. Ainsi, la portabilité de ce type de méthode pourrait être évaluée.

Toutes les conversations ont été transcrites manuellement et annotées en thème. Parmi elles, 322 conversations ont été utilisées pour constituer le corpus de TEST et 298 pour le corpus de développement. Les ngrams composant un TRSC tripartite des corpus de test et de développement pour lesquels $m_j^{t,i}(n, d) \rightarrow \gamma_{t,i}$ ont été vérifiés manuellement. Toutes les conversations ont été transcrites automatiquement avec un système de reconnaissance automatique de la parole (ASR). Ce système, décrit dans (Rousseau *et al.*, 2014), génère, pour chaque conversation, les hypothèses de séquence de mots les plus probables. Le modèle de langage du système a été adapté au domaine et un procédé du Leave-One-Out a été utilisé. Les taux de « Word Error Rate » (WER) du système sont respectivement de 33,8% pour le DEV et de 34,5% pour le TEST. Le tableau 1 présente les résultats de l'annotation automatique en thème (en utilisant les 13 thèmes à notre disposition plus le thème NULL) avant et après l'utilisation d'heuristiques de correction décrites à la fin de la section précédente. La dernière ligne du tableau est consacrée aux résultats en termes de pertinence de détection des ngrams en utilisant le DNN décrit plus haut.

	DEV	TEST
Classification en thèmes	83.2%	81.4%
Classification en thèmes après correction	84.6%	83.2%
Pertinence de la détection en mentions	89.4%	87.0%

TABLE 1 – Résultats obtenus en termes de détection d'un thème

Le haut degré de précision quant à la classification en thème, observé avec les résultats de sorties du système ASR, est encourageant. Il est de 96% sur le Dev et de 89% sur Test. Le rappel est, quant à lui, de 87% sur le DEV et de 85% sur le Test. En se fondant sur un effort de modélisation des TRSC, on peut thématiser des conversations téléphoniques mais aussi concevoir une stratégie de correction d'erreurs. La présence de ngrams inclus dans des TRSC non compatibles avec le thème annoté est souvent dû au fait que certaines conversations sont multithématiques. Or, seul le thème dominant a été annoté et évalué.

Enfin, on doit noter que si l'on ne tenait compte que de la détection des TRSC dans les conversations sans appliquer la méthode de correction décrite plus haut (fondée sur l'utilisation d'un DNN), aucune amélioration ne serait introduite. Au contraire, nos expérimentations montrent que, sur le corpus de Test, le taux de bonne classification en thème passerait de 81,4% à 79,9%. Dans certaines applications de dialogue sur des tâches courantes, comme c'est le cas ici avec le corpus DECODA, on s'aperçoit que l'utilisation d'un DNN accompagnée d'une approche semi-supervisée simple permet de mieux percevoir l'objet de l'échange et la détection en thème alors même que le WER montre que la tâche de reconnaissance de la parole reste délicate.

6 Conclusion et Perspectives

Nous avons proposé une méthode d'extraction automatique d'hypothèses pour la détection de thèmes dans des conversations téléphoniques. Le but est de permettre l'élaboration automatique de rapports indiquant, entre autres, le thème des échanges. Les TRSC, qui constituent autant de déclencheurs permettant de remonter directement au thème, sont annotés à partir du corpus d'apprentissage en utilisant une approche semi-supervisée et une stratégie d'apprentissage actif simple, ne nécessitant qu'un effort humain limité. Les résultats expérimentaux montrent une bonne capacité de détection en thème lorsque seul le thème principal de la conversation est annoté. Certes, les conversations multi-thématiques viennent brouiller les pistes et constituent une difficulté majeure dans ce type de travail. Toutefois, la mesure de précision indique de fortes valeurs pour les mentions validées par le DNN. Les taux de rappel sont eux-aussi corrects. Actuellement, de nouvelles annotations sont menées pour faciliter la formulation d'hypothèses y compris pour les thèmes abordés dans un second temps au sein d'une conversation. Les bons résultats obtenus par l'annotation manuelle sur l'ensemble du TEST indiquent qu'une approche fondée sur un effort humain limité mais guidée par une ontologie respectant l'application peut être efficace. En sélectionnant et en généralisant le contenu des listes de ngrams générées et évaluées automatiquement, il est possible d'entraîner des DNN et de valider les hypothèses ainsi émises. Parallèlement, des travaux de recherche sont actuellement en cours pour étendre cette approche à la détection des actes de dialogue.

Remerciements

Ce travail a été partiellement financé par la Commission Européenne à travers le projet EUMSSI, sous le numéro de contrat 611057, selon le numéro d'identification FP7-ICT-2013-10. Ce travail a également fait l'objet d'un financement de la part de l'Agence Nationale pour la Recherche (ANR) à travers le projet VERA qui porte le numéro ANR-12-BS02-006-01.

Références

- BOST X., DENAY G., EL-BEZE M. & MORI R. D. (2015). Multiple topic identification in human/human conversations. In *Actes de In Computer Speech and Language Journal2015(ICSLJ2015)*.
- BREIMAN L., FRIEDMAN J. H., OLSHEN R. & STONE C. (1984). Classification and regression trees. In *Technical report, Wadsworth international*, Monterey, CA.
- ESTÈVE Y., BOUALLEGUE M., LAILLER C., MORCHID M., DUFOUR R., LINARÈS G. & MORI R. D. (2015). Integration of word and smeantic features for theme identification in telephone conversations. In *Actes de International Worshop on Spoken Dialogue System2015(IWSDS2015)*, Buzan, South Korea.
- HAZEN T. J. (2011). Topic identification. In G. TUR & J. R. DE MORI, Eds., *Spoken language understanding : Systems for extracting semantic information from speech*. John Wiley and Sons, p. 319–356.
- JI Y. & EISENSTEIN J. (2014). Representation learning for text-level discourse parsing. In *Actes de the 23rd International Conference on Computational Linguistics2014*, p. 595–603.
- LI J. & HOVY E. (2014). A model of coherence based on distributed sentence representation. In *Actes de Conference on Empirical Methods in Natural Language Processing*, p. 2039–2048, Doha, Qatar.
- MANN W. C. & THOMPSON S. A. (1988). Rhetorical structure theory : Toward a functional theory of text organization. In *Text*, 8(3), p. 243–281.
- MESQUITA F., SCHMIDEK J. & BARBOSSA D. (2013). Effectiveness an efficiency of open relation extraction. In *Actes de Conference on Empirical Methods in Natural Language Processing*, p. 447–457, Seattle, Washington, USA.
- MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In *Actes de The Second International Conference on Learning Representations*.
- MORCHID M., DUFOUR R., BOUALLEGUE M., LINARÈS G., MATROUF D. & MORI R. D. (2014a). An i-vector based approach to compact multi-granularity topic spaces representation of textual documents. In *Actes de Conference on Empirical Methods in Natural Language Processing*, p. 443–454, Doha, Qatar.
- MORCHID M., DUFOUR R., BOUSQUET P.-M., LINARÈS G. & MORI R. D. (2014b). Improving dialogue classification using a topic space representation and a gaussian classifier based on the decision rule. In *Actes de the IEEE International Conference on Acoustic, Speech and Signal Processing*, p. 126–130, Florence, Italy.
- PRASAD R., JOSHI A. & WEBBER B. (2014). Realization of discourse relations by other means : alternative lexicalizations. In *Actes de the 23rd International Conference on Computational Linguistics*, p. 1023–1031 : Association for Computational Linguistics.
- ROUSSEAU A., BOULIANNE G., DELÉGLISE P., ESTÈVE Y., GUPTA V. & MEIGNIER S. (2014). Lium and crim asr system combination for the repere evaluation campaign. In *Actes de 17th Internation Conference on Text, Speech and Dialogue*, Brno, Czech republic.

Étiquetage morpho-syntaxique en domaine de spécialité: le domaine médical

Christelle Tiana Rabary¹ Thomas Lavergne^{1,2} Aurélie Névéal¹

(1) LIMSI-CNRS, Campus Universitaire d'Orsay, bât 508, 91405 ORSAY, FRANCE

(2) Université Paris-Sud, 91403 ORSAY, FRANCE

prenom.nom@limsi.fr

Résumé. L'étiquetage morpho-syntaxique est une tâche fondamentale du Traitement Automatique de la Langue, sur laquelle reposent souvent des traitements plus complexes tels que l'extraction d'information ou la traduction automatique. L'étiquetage en domaine de spécialité est limité par la disponibilité d'outils et de corpus annotés spécifiques au domaine. Dans cet article, nous présentons le développement d'un corpus clinique du français annoté morpho-syntaxiquement à l'aide d'un jeu d'étiquettes issus des guides d'annotation French Treebank et Multitag. L'analyse de ce corpus nous permet de caractériser le domaine clinique et de dégager les points clés pour l'adaptation d'outils d'analyse morpho-syntaxique à ce domaine. Nous montrons également les limites d'un outil entraîné sur un corpus journalistique appliqué au domaine clinique. En perspective de ce travail, nous envisageons une application du corpus clinique annoté pour améliorer l'étiquetage morpho-syntaxique des documents cliniques en français.

Abstract.

Part of Speech tagging for specialized domains : a case study with clinical documents in French.

Part-of-Speech (PoS) tagging is a core task in Natural Language Processing, often used as a stepping stone to perform more complex tasks such as information extraction or machine translation. PoS tagging of specialized documents is often challenging due to the limited availability of tools and annotated corpora dedicated to specialized domains. Herein, we present the development of a PoS annotated corpus of clinical documents in French, using annotation guidelines from the FrenchTree Bank and Multitag datasets. Through analysis of the annotated corpus, we characterize the clinical domain, including specific targets for domain adaptation. We also show the limitations of a PoS tagger trained on news documents when applied to clinical text. We expect that the domain-specific resource presented in this paper will contribute to improve PoS tagging for clinical documents in French.

Mots-clés : adaptation ; analyse morpho-syntaxique ; langue de spécialité ; dossier électronique patient.

Keywords: domain adaptation ; part-of-speech tagging ; specialized domain ; EHR.

1 Introduction

1.1 Contexte et motivation

L'étiquetage morpho-syntaxique est une tâche fondamentale du Traitement Automatique de la Langue, sur laquelle reposent souvent des traitements plus complexes tels que l'extraction d'information ou la traduction automatique. Les méthodes d'apprentissage statistiques ont fortement progressé ces dernières années, mais restent limitées en domaine de spécialité par la disponibilité de corpus annotés spécifiques au domaine. Sur l'anglais, des travaux utilisant des corpus cliniques (Pakhomov *et al.*, 2006; Liu *et al.*, 2007; Ferraro *et al.*, 2013) montrent le potentiel d'adaptation à ce domaine lorsque des corpus spécialisés, même de taille modeste, sont disponibles. Une étude avec l'outil statistique MedPOST caractérise plus spécifiquement l'apport des données annotées et des ressources lexicales spécialisées riches pour l'adaptation d'un étiqueteur au domaine biomédical (Smith *et al.*, 2005). A partir de ces résultats, Pécheux *et al.* (Pécheux *et al.*, 2014) montrent qu'un gain de performance significatif peut être obtenu dans un système de traduction automatique de documents biomédicaux utilisant un étiqueteur morpho-syntaxique adapté au domaine. Pour le français, le développement du corpus Sequoia (Candito & Seddah, 2012) a abordé l'adaptation en domaine pour des outils d'analyse syntaxique. Le corpus comporte notamment deux rapports publics européens discutant la mise sur le marché d'un médicament. Cependant, la principale ressource annotée morpho-syntaxiquement pour le français reste le corpus journalistique French Tree Bank (Abeillé *et al.*, 2003), à la fois en terme de taille du corpus et de complexité du jeu d'étiquettes utilisé.

1.2 Objectif et contribution

Dans cet article nous nous intéressons à l'étiquetage morpho-syntaxique de textes en domaine de spécialité, et plus particulièrement les documents cliniques issus des dossiers électroniques de patients en français. A travers le développement d'un corpus clinique du français annoté morpho-syntaxiquement, nous présentons une analyse du corpus permettant de caractériser le domaine clinique et de dégager les points-clés pour l'adaptation d'outils d'analyse morpho-syntaxique à ce domaine. Le corpus annoté constitue une contribution importante de notre travail, et a vocation par la suite d'entraîner et d'évaluer un outil d'étiquetage spécialisé.

2 Méthodes

2.1 Présentation du corpus de travail

Pour cette étude, nous avons utilisé des documents d'un corpus de textes cliniques issus d'un groupe d'institutions hospitalières françaises. Ce corpus contient plusieurs types de documents (principalement des courriers, des comptes-rendus de séjour, des comptes-rendus d'actes et des ordonnances). Pour ce travail, nous avons sélectionné des documents issus du service d'hépatogastro-nutrition faisant déjà l'objet d'une dé-identification (Grouin & Névéol, 2014) et d'autres études (Deléger & Névéol, 2014) afin d'enrichir le corpus au niveau morpho-syntaxique.

2.2 Schéma d'annotation morpho-syntaxiques

Notre choix d'étiquettes a été guidé par l'état de l'art en annotation morpho-syntaxique pour le français, exposé dans le guide d'annotation French Tree Bank (Abeillé *et al.*, 2003) dénommé "FTB" dans la suite de cet article) et le guide d'annotation Multitag élaboré dans le cadre de la campagne PASSAGE (Villemonte De La Clergerie *et al.*, 2008).

Le schéma d'annotation du FTB est très riche et détaillé comme on peut le voir sur l'exemple suivant :

FTB :	CL-suj-1mp	V-P-1p	D-ind-ms	N-m-s
	Nous	administrons	un	traitement

Ce schéma est composé de 15 catégories lexicales et 38 sous-catégories ainsi que d'un grand nombre de traits morphologiques pour toutes les formes fléchies. Au contraire, le schéma d'annotation Multitag est principalement restreint à la syntaxe et se compose de 11 catégories lexicales et 33 sous-catégories. Il encode peu de traits morphologiques comme on peut le voir sur l'exemple suivant :

Passage :	Pp	Vm	Da	Nc
	Nous	administrons	un	traitement

La complexité du schéma du FTB demande de plus grandes compétences pour l'annotateur et augmente à la fois le temps d'annotation et le risque d'erreurs. Il semble donc plus pertinent d'utiliser ce second schéma.

Le corpus FTB a été converti au schéma Multitag ainsi que la partie médicale (notices de médicaments éditées par l'Agence Européenne du Médicament, EMEA) du corpus Sequoia (Candito & Seddah, 2012). L'ensemble de ces corpus ont été utilisés pour entraîner un modèle CRF à l'aide de l'outil Wapiti tel que décrit dans (Lavergne *et al.*, 2010) (les caractéristiques utilisées y sont décrites à la section 5.1.2) afin de pré-annoter les documents médicaux. Le tokenizer utilisé lors de la première phase d'annotation est un outil maison suivant une segmentation proche de celui du FTB. Il n'est pas adapté au domaine médical, il a été important de modifier la segmentation manuellement afin d'obtenir une annotation morpho-syntaxique de qualité optimale. Les corpus FTB et PASSAGE sont constitués de textes journalistiques (articles du journal Le Monde) et littéraires. Le corpus EMEA est en revanche plus proche du domaine médical mais est de taille relativement réduite. Le tableau 2 présente l'ensemble du jeu d'étiquette que nous utilisons.

2.3 Pré-traitement du corpus de travail

Des travaux sur le développement de corpus clinique en anglais annoté morpho-syntaxiquement ont montré qu'il était possible de minimiser la taille du corpus annoté grâce à des heuristiques de fréquence simple (Liu *et al.*, 2007). Par ailleurs, les études précédentes menées sur notre corpus ont montré que certaines parties des documents étaient redondantes et

Etiquette complète	Etiquette réduite	Etiquette complète	Etiquette réduite
Adjectif qualificatif	Aq	Adjectif ordinal	Ao
Adjectif cardinal	Ak	Adjectif indéfini	Ai
Adjectif interrogatif	At	Adjectif possessif	Ap
Conjonction de coordination	CC	Conjonction de subordination	CS
Article	Da	Déterminant démonstratif	Dd
Déterminant indéfini	Di	Déterminant cardinal	Dk
Déterminant relatif	Dr	Déterminant possessif	Dp
Déterminant exclamatif ou interrogatif	Dt	Mot-phrase	I
Nom commun	NC	Nom propre	NP
Nom cardinal	Nk	Pronom démonstratif	Pd
Pronom indéfini	Pi	Pronom cardinal	Pk
Pronom personnel	Pp	Pronom relatif	Pr
Pronom possessif	Ps	Pronom interrogatif	Pt
Pronom réfléchi	Pf	Adverbe	Qg
Adverbe exclamatif ou interrogatif	Qx	Particule négative	Qn
Préposition	Sp	Introduceur	Sd
Verbe plein	Vm	Verbe auxiliaire	Va
Résidu	X	Ponctuations	F

TABLE 1 – Jeu d’étiquette pour l’annotation morpho-syntaxique

pouvaient présenter un intérêt limité pour l’analyse clinique et morpho-syntaxique (par exemple, en-têtes des documents). Ainsi, afin d’optimiser les efforts d’annotation dans notre étude, nous avons procédé à une sélection de phrases à l’intérieur des documents afin de concentrer le travail sur des phrases pertinentes et non-redondantes entre elles. La sélection a été opérée à l’aide d’un outil libre développé par Cohen et al. (Cohen *et al.*, 2013) en choisissant 20% comme taux de similarité maximum acceptable (valeur défaut recommandée par les auteurs) et 10 caractères comme taille des segments sur lesquels se fonde la comparaison. La sélection conduit à retenir moins de 50% de phrases ; cela s’explique en partie par le fait que les phrases "courtes" (typiquement contenues dans les en-têtes et pieds de page) de longueur inférieure à la taille des segments sont systématiquement exclues de la sélection. Les paires de phrases dont la similarité est au-dessus du seuil comprennent par exemple des variantes des phrases décrivant des techniques d’examen reprises plus ou moins à l’identique dans plusieurs documents ; ou alors des phrases comme "compte rendu d’hospitalisation de Nom Prénom, Né(e) le Date."

Un jeu de 60 documents a été préannoté avec le modèle décrit à la section précédente. Seules les phrases sélectionnées ont été pré-annotées, et présentées à un linguiste pour correction à l’aide de l’outil BRAT (Stenetorp *et al.*, 2012).

3 Annotation morphosyntaxique du corpus

3.1 Déroulement du travail d’annotation

Le travail d’annotation est fait par un seul et unique linguiste : ce choix a été imposé par le fait que nous n’avions pas d’autres linguistes disponibles à cette période. Il est prévu par la suite d’intégrer un deuxième annotateur qui viendra étayer le travail effectué en amont. Nous sommes tout à fait conscient du biais modéré que cette configuration induit, cependant il est à préciser que le linguiste ne disposait d’aucune information concernant le corpus sur lequel il a travaillé (préparation et taille des données).

Le travail d’annotation effectué sur le corpus s’est déroulé en deux temps : dans un premier temps, le linguiste a travaillé sur le corpus pré-annoté automatiquement, afin de valider ou corriger l’étiquetage proposé par l’outil statistique. Dans cette partie du travail, seule une sélection de phrases sont découpées en tokens de manière similaire au corpus FTB et pré-annotées à l’aide du modèle CRF. Les phrases sont cependant présentées à l’intérieur du document d’origine afin de fournir tous les éléments de contexte nécessaires à l’annotateur, qui peut modifier les étiquettes, sans changer le découpage en tokens proposé par l’outil. Suite à cette première phase d’annotation, une première analyse d’erreurs a permis d’identifier notamment des problèmes liés à la tokenisation du corpus.

En conséquence, dans un deuxième temps, un nouveau jeu de 30 documents vierges (sans pré-annotation) est présenté au linguiste qui peut alors effectuer l’annotation morpho-syntaxique en choisissant librement le découpage en token. Lors de cette phase, le temps moyen d’annotation a augmenté de 50% de par l’absence de pré-annotation et la nécessité de corriger les erreurs de tokenisation. Une partie des phrases annotées lors de cette étape est commune avec l’étape précédente afin de pouvoir évaluer l’impact de la pré-annotation.

Enfin, une partie des documents (10 documents) étant commune aux deux phases d’annotation, un consensus entre les deux annotations a été réalisé afin de déterminer l’étiquetage final de ces documents. Afin de limiter le biais de l’annotateur dans cette phase du travail, les documents communs ont été présentés sans distinction particulière. De plus, l’annotateur ne savait pas a priori que certains documents avaient déjà été utilisés dans la phase précédente. Interrogé sur ce point, il a indiqué ne pas s’en être rendu compte. Lors de ce consensus, les deux tokenisations et étiquetages sont présentés au linguiste qui peut choisir l’une ou l’autre pour chaque segment de phrase. L’accord intra-annotateur sur cette sous-partie du corpus est de 0.84 (précision), malgré les divergences de tokenisation.

Le tableau 2 présente la taille du corpus annoté final. A titre indicatif, nous indiquons également les chiffres équivalents pour les corpus Sequoia-EMEA et FTB. Il est intéressant d’observer que le taux de redondance des tokens dans notre corpus est relativement faible : en moyenne, 4,31 occurrences par mot de vocabulaire contre 7,96 dans EMEA et 19,99 dans FTB. Cela atteste du succès de notre méthode de sélection de phrases non-redondantes.

	Corpus Clinique	Sequoia-EMEA	FTB
Nombre de fichiers	80	2	44
Nombre de phrases annotées	722	1 108	11 116
Nombre de tokens annotés	13 721	22 275	679 730
Taille du vocabulaire	3 181	2 797	33 988

TABLE 2 – Statistiques descriptives du corpus clinique

3.2 Performances du système de pré-annotation

La motivation de cet étiquetage étant de caractériser les particularités du domaine médical et de fournir un corpus permettant l’adaptation des modèles existants, la première étape d’analyse a consisté à évaluer les performances du système de pré-annotation.

Le système CRF atteint 0.80 de taux d’erreur sur l’ensemble des étiquettes ce qui est relativement faible pour une tâche d’analyse morpho-syntaxique générale mais est raisonnable pour un système non adapté. Pour comparaison, le système CRF atteint 0.964 de taux d’erreur sur des documents de test issus du FTB et un système entraîné uniquement sur les corpus FTB et passage obtient quant à lui 0.968 sur ce même test ce qui est comparable à l’état-de-l’art (Constant *et al.*, 2011; Denis & Sagot, 2012). Une chute de performance pouvant atteindre jusqu’à 15% est également constatée sur l’anglais, lorsque des étiqueteurs génériques sont appliqués sur des textes cliniques (Ferraro *et al.*, 2013). Une analyse plus fine des erreurs montre que la majorité des erreurs sont dues soit à des erreurs de tokenisation, soit à des particularités du domaine médical. Parmi ces dernières on retrouve notamment des confusions entre noms communs et noms propres dans les noms de procédures ou dispositifs médicaux. Par exemple, dans le syntagme *confection d’un Hartmann*, le dernier token doit être annoté comme un nom commun contrairement à ce que la majuscule suggère car il s’agit d’une marque de dispositifs médicaux (cas similaire à l’emploi du terme *Frigidaire* dans la langue générale). On retrouve aussi des mots composés néoclassiques : *thoraco-abdomino-pelvien*, des noms d’appareils : *Endoscope Olympus XQ30*) ou des noms de médicaments : *Interféron pégylé*, composés de tokens spécifiques au domaine médical.

Au final, nous calculons qu’environ 41% des erreurs de pré-annotation sont dues à des particularités de tokenisation et 22% sont liées au vocabulaire médical. En dehors de ces deux classes d’erreurs, qui sont analysées plus finement dans les sections suivantes, la pré-annotation montre des résultats de bonne qualité. Ces observations indiquent qu’avec un tokeniseur adapté, l’utilisation d’une pré-annotation permet un gain de temps significatif. Cela rejoint les résultats de travaux antérieurs sur l’annotation de corpus (Névél *et al.*, 2011).

4 Analyse d'erreurs

4.1 Difficultés liées à la tokenisation

Parmi les phrases à étiqueter qui ont été préalablement sélectionnées, nous avons rencontré une quantité importante d'abréviations, de termes contenant des éléments de ponctuation et d'autres termes qui à première vue ressembleraient à des sigles. Ces éléments sont des formes d'expressions propres au domaine médical : ce sont des abréviations de noms de procédures médicales, des descriptions de stades d'évolution de maladies, des modalités d'administration de la prise de médicaments, et des mesures. Ce qui fait la singularité de ces termes, c'est qu'ils suivent une structure orthographique étrangère à ce qui est connu du CRF (rappelons qu'il a été exclusivement entraîné sur du corpus journalistique) ; de surcroît, la quantité de sigles est ici beaucoup plus élevée que d'ordinaire (toujours en comparaison avec le contenu d'un corpus générique).

Abréviation des termes médicaux. Les textes du domaine médical et en particulier les comptes rendus cliniques étudiés ici comportent une grande part d'abréviations. Ce phénomène, largement étudié du point de vue de la désambiguïsation lexicale (Stevenson *et al.*, 2009), est dû à l'abondance de termes spécialisés ainsi qu'à la rédaction de type prise de notes.

On trouve deux types d'abréviations. D'une part des abréviations partielles, telles que *chir.* pour *chirurgie*, qui ne font pas partie des listes classiques d'abréviations et sont donc inconnues à la fois du tokeniseur et du modèle CRF. La ponctuation qui doit être considérée comme faisant partie du token est généralement séparée car reconnue comme une ponctuation finale. Le token se trouve donc incorrectement étiqueté et le marqueur de fin de phrase a tendance à propager cette erreur aux mots suivants. D'autre part, de nombreux termes médicaux tels que *anesthésie générale* ou *sérum glutamoxaloacétate transférase* sont complètement abrégés en *A.G.* et *SGOT*. Même si la morphologie de ces tokens permet de les reconnaître plus simplement, leur regroupement sous une seule étiquette « abréviation » est ici peu approprié, ces termes étant souvent porteurs d'une information sémantique importante pour l'analyse des documents médicaux. Le choix le plus approprié est de réaliser une tokenisation assurant un découpage en tokens similaire à ceux du terme non abrégé ainsi qu'un étiquetage complet de la séquence. L'abréviation *A.G.* est donc annotée *A. :NC G. :ADJc* au contraire de *A.G. :NP* suggéré par le système de pré-annotation.

Mesures et abréviations complexes. Une deuxième particularité du domaine médical est l'abondance de quantités et de mesures. Si certaines sont simples, comme *3mm*, et sont correctement analysées par un système non-adapté au domaine, ce n'est pas le cas pour les plus complexes telles que *3x/j* ou *5,4 mmoles/l*. De plus, une même mesure peut-être abrégée de manière différentes, pour *3 fois par jour* par exemple, nous avons observé les abréviations suivantes dans notre corpus : *3 fois/jours*, *3 fois/j*, *3x/j*, *3/j*. . .

On trouve aussi des termes ressemblant à la fois à des abréviations et des mesures tels que *T2N+* dans *lésion du rectum T2N+* qui indique le degré d'évolution d'un cancer. Il s'agit de la classification TNM (« tumor, nodes, metastasis » en anglais) qui prend en compte la taille et la localisation de la tumeur primitive (notée parfois pT), le nombre et le site des ganglions lymphatiques régionaux qui contiennent des cellules cancéreuses, et la propagation du cancer, ou métastases, vers une autre partie du corps.

Ces deux types de termes : *3x/j* et *T2N+*, ont tendance à être considérés comme un seul token par la chaîne de traitement non-adaptée au domaine. Comme pour les abréviations simples, il est pourtant pertinent ici de les décomposer afin de les étiqueter de manière similaire à leur écriture non-abrégée.

On annotera donc *3 :Det x :NC / :Prep j : :NC* et *T :NC 2 :ADJc N :NC + :ADV*. Cette annotation complète bien que plus coûteuse et demandant des connaissances médicales dans certains cas permet de faciliter les étapes suivantes de l'analyse automatique de ces documents.

4.2 Difficultés liées au vocabulaire

Le domaine médical a un vocabulaire très riche qui met facilement le système de pré-annotation en difficulté. On a pu noter principalement deux formes de problèmes : ceux liés à des mots spécifiques au domaine et donc inconnus du système, et ceux liés à une utilisation différente dans le domaine médical de mots classiques et donc connus.

Termes médicaux et mots hors vocabulaire. Le système de pré-annotation a été entraîné sur des corpus de textes journalistiques. Pour ce type d'étiquetage, si le corpus d'entraînement est suffisamment important, la très grande majorité des mots hors-vocabulaires seront des noms propres. On peut aussi trouver des variantes morphologiques de mots connus,

comme une forme conjuguée non vue à l'apprentissage d'un verbe connu, mais le système peut facilement gérer ces cas en exploitant le lemme du terme. La régularisation ℓ_1 utilisée à l'apprentissage du modèle CRF induit une sélection de caractéristiques. Cela a notamment pour effet de pénaliser l'utilisation des caractéristiques lexicales pour les mots les moins fréquents au profit des caractéristiques non-lexicales. Au décodage, cet effet se traduit par un fort biais du modèle vers l'étiquette NP pour les mots inconnus qui n'ont aucune caractéristique lexicale active.

Ces mots sont très fréquents dans les documents médicaux et les erreurs qu'ils engendrent, de par la structure du modèle CRF, se propagent facilement aux mots voisins. Pour le syntagme *microfilaments de type actine-myosine* par exemple, seul le token *de* sera correctement étiqueté car il appartient à une catégorie grammaticale fermée. Les deux tokens extrêmes sont par contre hors-vocabulaires et donc mal étiquetés et ces erreurs se propagent sur le token *type* pourtant connu du système.

Il est à noter que ces mots hors-vocabulaires ne peuvent être traités uniquement grâce à l'ajout de données médicales étiquetées. Il est en effet impossible d'en avoir une couverture suffisante et l'utilisation de lexiques se trouve ici indispensable.

Homographies. L'homographie de certains termes existant en dehors du domaine médical a aussi été un facteur d'erreurs d'étiquetage. De nombreux termes sont utilisés différemment en contexte médical et leur catégorie grammaticale change, par exemple dans le syntagme *veine porte*, le terme *porte* est un adjectif qualificatif, alors qu'il est un nom commun ou une forme conjuguée du verbe *porter* dans les textes du corpus d'entraînement.

De même, le terme *scanner* est un nom commun dans le domaine médical. S'il peut aussi être un nom commun dans le domaine général, il est plus fréquemment étiqueté comme verbe, ce qui conduit à des erreurs. Une étude sur notre corpus montre que parmi ces homographies, dans 23% des cas, la catégorie grammaticale utilisée en domaine médicale est inobservée dans le corpus général et donc impossible à prédire pour le système.

Concernant l'homographie, nous avons aussi pu observer des erreurs plus complexes. Par exemple, le syntagme *le toucher rectal* peut être étiqueté PROP VRB ADV ou DET NC ADJ. En dehors de tout contexte, les deux choix d'annotations sont plausibles, mais dans le vocabulaire médical, *toucher* désigne une pratique. Ce terme est utilisé en tant que nom commun et seule la deuxième possibilité est donc acceptable.

Ambiguïté des noms de médicaments. Les noms de médicaments tels que *Dafalgan effervescent*, *Fudicine pommade* et *Néoral 50* présentent une difficulté supplémentaire. Deux étiquetages sont possibles pour chacun d'eux : soit les deux tokens forment un nom propre et reçoivent tous les deux le tag NP, soit seul le premier token est considéré comme constituant le nom du médicament, le deuxième étant un adjectif qui en précise la forme. Dans ce deuxième cas, le deuxième token recevra le tag ADJ_q ou ADJ_c suivant les cas.

Ces deux schémas d'annotation sont justifiables mais il est nécessaire d'assurer la cohérence du choix sur tout le corpus. Le système de pré-annotation est incohérent ici car, lorsque l'adjectif ne fait pas partie de son vocabulaire il choisit la première forme, mais dans le cas contraire ou pour les numéraux, c'est la deuxième forme qui l'emporte.

Médicalement, le deuxième token correspond à une caractéristique associée au médicament, telle que la concentration, la forme ou la voie d'administration. Ainsi, il semble plus pertinent de ce point de vue de choisir la deuxième solution d'étiquetage.

5 Conclusion et perspectives

L'analyse d'un corpus clinique du français annoté morpho-syntaxiquement et en particulier des erreurs d'étiquetage faites par un outil générique a mis en évidence deux principales caractéristiques du domaine clinique. D'une part, un *vocabulaire spécialisé* qui dénote des connaissances à apporter à l'étiqueteur grâce à des lexiques spécialisés et des données étiquetées. D'autre part, un besoin d'une *tokénisation particulière* pour des phénomènes linguistiques particulièrement prévalents dans les textes cliniques, tels que les posologies, mesures et abréviations. Pour permettre à un outil statistique d'apprendre ce type de construction, il est nécessaire de disposer d'une quantité satisfaisante de données annotées. Une contribution importante du travail présenté dans cet article est le développement d'un corpus du domaine clinique annoté morpho-syntaxiquement, dans le but d'entraîner et d'évaluer un outil d'étiquetage spécialisé. Le développement d'un tel outil est en cours. Une perspective à moyen terme est d'enrichir le corpus existant avec l'intervention d'un deuxième annotateur linguiste afin de faire de ce corpus une référence de qualité destinée à évaluer différents outils. Cet objectif reste conditionné à l'obtention d'une autorisation de la CNIL en raison de la nature sensible des documents.

Remerciements

Nous remercions le Service d'Informatique Biomédicale (SIBM) ainsi que l'équipe CISMeF du CHU de Rouen qui nous ont permis d'utiliser le corpus LERUDI pour cette étude. Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence CAbReNeT ANR-13-JS02-0009-01.

Références

- ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In A. ABEILLÉ, Ed., *Treebanks*. Kluwer, Dordrecht.
- CANDITO M.-H. & SEDDAH D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de TALN 2012*, p. 321–334.
- COHEN R., ELHADAD M. & ELHADAD N. (2013). Redundancy in electronic health record corpora : analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics*, **14**, 10.
- CONSTANT M., TELLIER I., DUCHIER D., DUPONT Y., SIGOGNE A. & BILLOT S. (2011). Intégrer des connaissances linguistiques dans un crf : application à l'apprentissage d'un segmenteur-étiqueteur du français. In *Actes de TALN 2011*.
- DELÉGER L. & NÉVÉOL A. (2014). Identification automatique de zones dans des documents pour la constitution d'un corpus médical en français. In *Actes de TALN 2014*, p. 568–573.
- DENIS P. & SAGOT B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Lang. Resour. Eval.*, **46**(4), 721–736.
- FERRARO J., DAUMÉ H., DUVALL S., CHAPMAN W., HARKEMA H. & HAUG P. (2013). Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. *J Am Med Inform Assoc*, **20**(5), 931–939.
- GROUIN C. & NÉVÉOL A. (2014). De-identification of clinical notes in french : towards a protocol for reference corpus developpement. *J Biomed Inform*, **50**, 151–61.
- LAVERGNE T., CAPPÉ O. & YVON F. C. (2010). Practical very large scale crfs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 504–513, Uppsala, Sweden.
- LIU K., CHAPMAN W., HWA R. & CROWLEY R. (2007). Heuristic sample selection to minimize reference standard training set for a part-of-speech tagger. *J Am Med Inform Assoc*, **14**(5), 641–650.
- NÉVÉOL A., DOĞAN R. I. & LU Z. (2011). Semi-automatic semantic annotation of PubMed queries : a study on quality, efficiency, satisfaction. *J Biomed Inform*, **44**(2), 310–8.
- PAKHOMOV S., CODEN A. & CHUTE C. (2006). Developing a corpus of clinical notes manually annotated for part-of-speech. *International Journal of Medical Informatics*, **75**(6), 418–429.
- PÉCHEUX N., GONG L., DO Q. K., MARIE B., IVANISHCHEVA Y., ALLAUZEN A., LAVERGNE T., NIEHUES J., MAX A. & YVON F. (2014). Limsi @ wmt'14 medical translation task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, p. 246–253, Baltimore, Maryland, USA : Association for Computational Linguistics.
- SMITH L., RINDFLESH T. & WILBUR W. (2005). The importance of the lexicon in tagging biological text. *Natural Language Engineering*, **12**(2), 1–17.
- STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). Brat : A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, p. 102–107, Stroudsburg, PA, USA : Association for Computational Linguistics.
- STEVENSON M., GUO Y., AL AMRI A. & GAIZAUSKAS R. (2009). Disambiguation of biomedical abbreviations. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, p. 71–79, Stroudsburg, PA, USA : Association for Computational Linguistics.
- VILLEMONT DE LA CLERGERIE E., HAMON O., MOSTEFA D., AYACHE C., PAROUBEK P. & VILNAT A. (2008). Passage : from french parser evaluation to large sized treebank. In *Proceedings of the 6th International Conference on Languages Resources and Evaluation, LREC 2008*, Marrakech, Morocco.

Vers une typologie de liens entre contenus journalistiques

Rémi BOIS¹ Guillaume GRAVIER¹ Pascale SÉBILLOT² Emmanuel MORIN³

(1) CNRS, IRISA & INRIA, Campus de Beaulieu, 35000 Rennes

(2) INSA, IRISA & INRIA, Campus de Beaulieu, 35000 Rennes

(3) Université de Nantes, LINA, 2 Rue de la Houssinière, 44300 Nantes

(1, 2) prenom.nom@irisa.fr (3) emmanuel.morin@univ-nantes.fr

Résumé. Nous présentons une typologie de liens pour un corpus multimédia ancré dans le domaine journalistique. Bien que plusieurs typologies aient été créées et utilisées par la communauté, aucune ne permet de répondre aux enjeux de taille et de variété soulevés par l'utilisation d'un corpus large comprenant des textes, des vidéos, ou des émissions radiophoniques. Nous proposons donc une nouvelle typologie, première étape visant à la création et la catégorisation automatique de liens entre des fragments de documents afin de proposer de nouveaux modes de navigation au sein d'un grand corpus. Plusieurs exemples d'instanciation de la typologie sont présentés afin d'illustrer son intérêt.

Abstract.

Towards a typology for linking newswire contents

In this paper, we introduce a typology of possible links between contents of a multimedia news corpus. While several typologies have been proposed and used by the community, we argue that they are not adapted to rich and large corpora which can contain texts, videos, or radio stations recordings. We propose a new typology, as a first step towards automatically creating and categorizing links between documents' fragments in order to create new ways to navigate, explore, and extract knowledge from large collections. Several examples of links in a large corpus are given.

Mots-clés : typologie, liens inter-documents, hypertexte, multimédia, presse.

Keywords: typology, linking documents, hypertext, multimedia, newswire.

1 Introduction

L'explosion de la quantité de sources d'informations disponible sur le web a rendu nécessaire l'utilisation d'outils permettant de guider la navigation des internautes, notamment dans le cas pratique de l'extraction d'information liée au domaine journalistique. Or, les outils existants aujourd'hui peinent à proposer à leurs utilisateurs une navigation qui soit à la fois intuitive et diversifiée. Une personne cherchant à étendre ses connaissances sur un événement lié à l'actualité sera le plus souvent amenée à utiliser un moteur de recherche, outil certes performant mais laissant à son utilisateur la charge cognitive de faire le lien entre les différentes pages qu'il aura consultées. Certains professionnels, comme les attachés de presse, ont également besoin de compiler rapidement toute une série d'informations autour d'un même sujet. Une fois encore, les moteurs de recherche sont la seule solution à leur disposition, faute d'outils plus performants.

Le projet LIMAH¹, dans lequel cette étude se déroule, vise à répondre à cette problématique en créant de façon automatisée des liens explicites, fondés sur une similarité sémantique, entre des fragments de documents issus du domaine journalistique. Un grand corpus multimédia, contenant un mois de données journalistiques récupérées sur le web sous forme de vidéos (*e.g.* journal télévisé), de podcasts (*e.g.* chroniques radio), ou de pages HTML (*e.g.* lemonde.fr) a été constitué pour le projet. L'objectif est de permettre une navigation éclairée au sein de ce corpus, dans lequel une personne peut par exemple choisir de suivre l'évolution d'une actualité, ou au contraire d'organiser son parcours de l'information autour de résumés afin d'acquérir une connaissance rapide des principaux enjeux la concernant. Organisée sous forme d'hypergraphe, cette collection de fragments liés doit permettre de proposer une telle navigation, lors de laquelle l'utilisateur dispose de plusieurs choix, décrits de façon explicite, pour orienter son parcours.

1. Linking Media in Acceptable Hypergraphs <http://limah.irisa.fr/>

La première étape à considérer pour la création de tels hypergraphes est la caractérisation des liens à construire entre les fragments de documents. Si plusieurs typologies de liens ont déjà été proposées (*cf.* section 3.1), elles ne nous semblent pas suffisamment riches pour représenter la diversité des relations existantes. Dans cet article, nous proposons donc une typologie des liens dans un corpus multimédia ancré dans le domaine journalistique avec comme objectif la création et la catégorisation automatiques de ces liens.

Dans un premier temps, nous décrivons le type de corpus sur lequel se fonde cette étude (*cf.* section 2). Puis, nous exposons la typologie proposée et la façon dont elle a été construite (*cf.* section 3). Nous montrons ensuite que cette typologie est adaptée au corpus étudié au travers d'exemples issus de ces données (*cf.* section 3.3). Nous concluons ce travail en exposant les possibilités offertes par cette typologie (*cf.* section 4).

2 Corpus

La recherche d'informations liées à l'actualité sur le web met en œuvre de très nombreux types de documents. La plupart des journaux papiers sont en effet disponibles dans une version électronique, le plus souvent accessible gratuitement, et parfois enrichie de contenus multimédias (*e.g.* vidéos explicatives, illustrations, graphiques). Les stations radio et chaînes de télévision mettent également à disposition des internautes leurs émissions, sous forme de podcasts pour les premiers et de vidéos à la demande pour les seconds. Les utilisateurs se tournent également de plus en plus vers une information communautaire, où les réactions à un événement deviennent aussi importantes que la description de l'événement lui-même. Les titres des articles de la presse écrite sont tweetés, leurs en-têtes publiés sur Facebook, et les internautes interagissent directement pour commenter non seulement l'information, mais également la façon dont celle-ci est transmise. Le réseau social Twitter est notamment largement utilisé pour s'informer, comme le montrent plusieurs études (Kwak *et al.*, 2010).

Pour appréhender ce domaine, disposer d'un corpus représentant cette masse et cette diversité semble nécessaire. Nous nous fondons donc sur un mois de données journalistiques, mélangeant des ressources issues des chaînes de télévision (France Télévision) ou des stations de radio (Radio France), des articles de blogs ou de presse écrite (Le Monde, Le Figaro), ainsi que les commentaires qui leur sont associés directement sur leur site ou via les réseaux sociaux (Twitter et Facebook). Parmi ces sources se trouvent des documents engagés (articles de blogs, chroniques radios) ainsi que des documents plus neutres (articles de presse). Sont également présentes des parodies diffusées dans des émissions (Le petit journal, Les guignols de l'info) ou transmises via des réseaux sociaux.

Dans le cadre du projet LIMAH, l'ensemble des données vidéo ou audio sont transcrites automatiquement et seules ces transcriptions sont utilisées pour la création de liens. Les documents utilisés ont donc une qualité variable selon leur modalité (*e.g.* article de presse écrite *vs* transcription d'une émission radio) ou leur provenance (*e.g.* abréviations dans des textes issus de Twitter). Ce corpus de travail totalisant plusieurs centaines de gigaoctets de données n'a pas vocation à être diffusé.

3 Construction d'une typologie de liens adaptée au domaine journalistique

Dans cette section, nous décrivons un ensemble de typologies et expliquons leurs faiblesses pour notre cas d'usage, avant de proposer une nouvelle typologie plus adaptée. Notre description des typologies existantes s'articule autour des différentes communautés qui les ont proposées. Ces communautés ont souvent des buts différents, et choisissent donc de lier différents éléments (*e.g.* événements, thèmes, documents) avec différents types de liens.

3.1 Typologies existantes

La création de liens entre documents, ou fragments de document, a été explorée par différentes communautés. La communauté du traitement automatique des langues s'est principalement intéressée, au travers de corpus journalistiques, à lier des événements entre eux, le plus souvent via deux types de liens : un lien de similarité (les deux documents présentent le même événement) et un lien de causalité temporelle (un événement en provoque un second) (Nallapati *et al.*, 2004; Renison, 1994; Muller & Tannier, 2004). Ces relations temporelles facilitent le parcours d'une collection de documents en proposant une navigation chronologique permettant de recomposer l'évolution d'une série d'événements. Il nous semble néanmoins que l'utilisation de ces deux seuls types de liens est peu adaptée à un corpus multimédia, dans lequel de nom-

breux événements sont décrits simultanément par différentes sources, commentés sur les réseaux sociaux et repris par des blogueurs. Un typage plus fin nous paraît donc nécessaire. Une seconde approche explorée par cette communauté consiste à relier des thèmes (*topics*) émergeant des documents. Une fois encore, le but principal consiste à regrouper ces thèmes et à proposer un parcours chronologique de ceux-ci (Ide *et al.*, 2004), avec les mêmes limites que celles décrites précédemment.

La communauté du multimédia s'est également intéressée à la création automatique de liens entre documents. Le plus souvent, ces liens ne sont pas typés et ne servent qu'à mettre en lumière une relation non explicite entre deux documents (Eskevich *et al.*, 2012). Ces liens non explicites se révèlent particulièrement utiles pour de petites collections, ou pour le développement de moteurs de recommandation. Cette absence de typage limite néanmoins les usages possibles et rend difficile une navigation éclairée. D'autres travaux dans ce même domaine mettent d'ailleurs en avant le besoin d'un typage pour faciliter le parcours de grandes collections (Cleary & Bareiss, 1996). Cette dernière étude expose une partie de la typologie qu'elle utilise où les 8 types les plus fréquents sont décrits sous forme de questions. Nous y trouvons des types classiques du domaine journalistique tels que les relations de causalité, mais aussi des liens de type « conseils : comment puis-je capitaliser sur cette situation ? » qui sont difficilement instanciables sur un corpus d'actualités.

La communauté des sciences de l'information et de la communication s'est aussi penchée sur le rôle des liens hypertextes. L'un de ces travaux a notamment influencé la typologie que nous proposons (Ertzscheid, 2002). Cette étude, bien que très générique et se plaçant dans un contexte éloigné du domaine journalistique, met en avant plusieurs grandes catégories de liens dont nous nous inspirons (*e.g.* une relation de récurrence, peu abordée dans les autres travaux, mais qui nous paraît pertinente dès lors que le corpus considéré est volumineux).

3.2 Typologie proposée

Dans le cadre du projet LIMAH, il nous paraît intéressant de lier des informations entre elles, plutôt que des événements trop fermés, ou des thèmes trop larges. Une information est définie par le Larousse comme « tout événement, tout fait, tout jugement porté à la connaissance d'un public plus ou moins large, sous forme d'images, de textes, de discours, de sons ». Nous choisissons d'étendre cette définition en constatant qu'une information peut également correspondre à une série d'événements ou de faits. Un exemple concret est un sujet d'un journal télévisé. Ce sujet peut durer plusieurs minutes et réunir différents événements présentés en un tout cohérent, qu'on appelle l'information. Cette notion de liens entre informations a déjà été exploitée dans d'autres travaux avec succès (Shahaf & Guestrin, 2010). Il s'agit donc de lier entre eux des fragments de documents, chacun de ces fragments représentant une information.

Nous proposons trois grandes catégories de liens, dont deux sont divisées en sous-catégories. Pour chacun des liens présentés, un lien inverse existe de telle sorte que tout fragment de document lié à un autre est à la fois source et cible d'un lien. Cette relation double peut se caractériser par des liens non orientés (*e.g.* un lien de type quasi duplicat est non orienté) ou par des liens duaux (*e.g.* le lien dual du développement est le résumé). Les types proposés ne sont pas exclusifs, un lien entre deux documents pouvant disposer de plusieurs types (*cf.* section 3.3). Les trois catégories retenues sont :

- la récurrence** : répétition d'une information. Le contenu est similaire mais peut être présenté de diverses manières, indépendamment de la modalité utilisée ;
- l'extension** : enrichissement d'une information. L'extension peut correspondre à un enrichissement en volume, avec un contenu plus large, ou bien à une extension temporelle correspondant à un suivi d'information ;
- la réaction** : l'information est commentée par un nouvel intervenant.

La récurrence est la relation la plus fréquemment rencontrée. Elle peut être envisagée sous trois formes :

- le quasi duplicat** : l'information est répétée, de façon similaire, sans ajout ou suppression notable ;
- la citation** : une référence à une information délivrée précédemment est incluse ;
- la parodie** : une information est reprise et détournée.

Nous considérons la parodie comme une forme de récurrence car elle reprend une information identique et change son traitement à des fins de divertissement. L'information traitée reste néanmoins la même.

L'extension enrichit une information en la développant ou en exhibant un lien temporel avec une autre information. Elle peut donc se préciser selon les deux sous-catégories suivantes :

- le développement** : l'information est développée, son contenu est plus important ;
- la postériorité** : une relation de suivi temporel est exhibée entre les deux informations.

La réaction concerne l'ensemble des commentaires qui peuvent être apportés sur une information, que ceux-ci aient lieu dans un milieu contrôlé (*e.g.* diffusion de la réponse d'un homme politique à une critique adverse) ou libre (*e.g.* réaction d'un internaute sur Twitter). Nous choisissons de ne pas offrir de sous-catégories à la réaction, bien qu'il soit possible d'utiliser les typologies existantes en analyse d'opinion pour affiner ce type (Ekman, 1992).

La typologie proposée est donc issue à la fois d'un réagencement de types couramment utilisés par la communauté, ainsi que de types rarement utilisés, mais pertinents dans le cadre d'un corpus multimodal diversifié. Elle reprend donc les relations classiques d'antériorité/postériorité, qui permettent de suivre une information d'un point de vue temporel, ou bien de source/citation, largement étudiées dans le cadre de corpus scientifiques (Nanba *et al.*, 2011; Thelwall, 2003), mais aussi des relations moins souvent exploitées telles que la parodie ou le quasi duplicat.

La figure 1 présente la typologie proposée par cette étude. Elle indique la nature des relations duales lorsqu'elles existent. Lorsqu'il n'y a pas de dualité, le lien est considéré comme non orienté. Ainsi, un lien d'antériorité entraîne nécessairement un lien inverse de postériorité, ou un lien de développement correspond toujours à un lien de résumé. Cette typologie nous paraît couvrir une très large majorité des liens possibles et permet d'envisager de nouveaux moyens de parcourir une grande collection de documents.

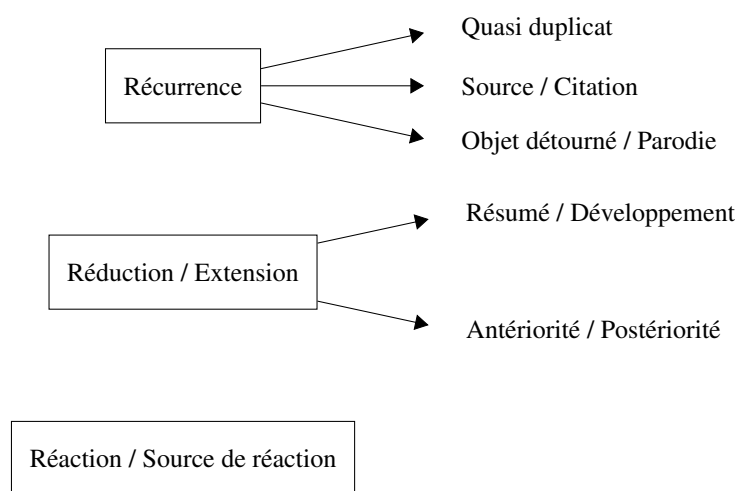


FIGURE 1: Typologie des liens entre informations

Cette typologie paraît adaptée au corpus décrit précédemment, comme le montrent les exemples développés dans la section suivante.

3.3 Exemples extraits du corpus

Nous exposons ici deux exemples extraits du corpus. Trois fragments de documents sont présentés. Le premier est un article du Figaro daté du 27 février 2015 et rapportant une allocution de Monsieur Manuel Valls lors d'un meeting électoral. Lors de cette allocution, M. Manuel Valls désigne l'extrême droite comme « l'adversaire principal ». Le deuxième est une partie d'une interview de Monsieur Florian Philippot. Durant cet entretien qui se déroule le 28 février 2015, M. Florian Philippot critique l'allocution de M. Manuel Valls et ses propos à l'encontre de son parti. Le troisième est un article du Point qui reprend par écrit l'interview de M. Florian Philippot. Également daté du 28 février 2015, l'article cite sa source et rapporte les paroles de M. Florian Philippot. La figure 2 montre les liens existant entre ces fragments de documents, en accord avec la typologie décrite précédemment.

La figure 3 reprend trois documents illustrant le dépôt de plainte de la ville de Paris après que la chaîne américaine Fox News ait qualifié certains quartiers parisiens de « no-go zones ». L'article du Point présente l'affaire tandis qu'un tweet résume l'article en reprenant l'en-tête tout en citant sa source. L'émission Le petit journal parodie l'information en décrivant le dépôt de plainte comme « une bataille entre Madame Anne Hidalgo, maire de Paris, et le premier amendement de la constitution américaine ».

Les liens duaux ne sont pas représentés sur les figures 2 et 3 pour des soucis de lisibilité.

Valls: l'extrême droite, "adversaire principal"

🏠 > ACTUALITE > FLASH ACTU Par LeFigaro.fr avec AFP | Mis à jour le 27/02/2015 à 07:47 | Publié le 26/02/2015 à 21:52

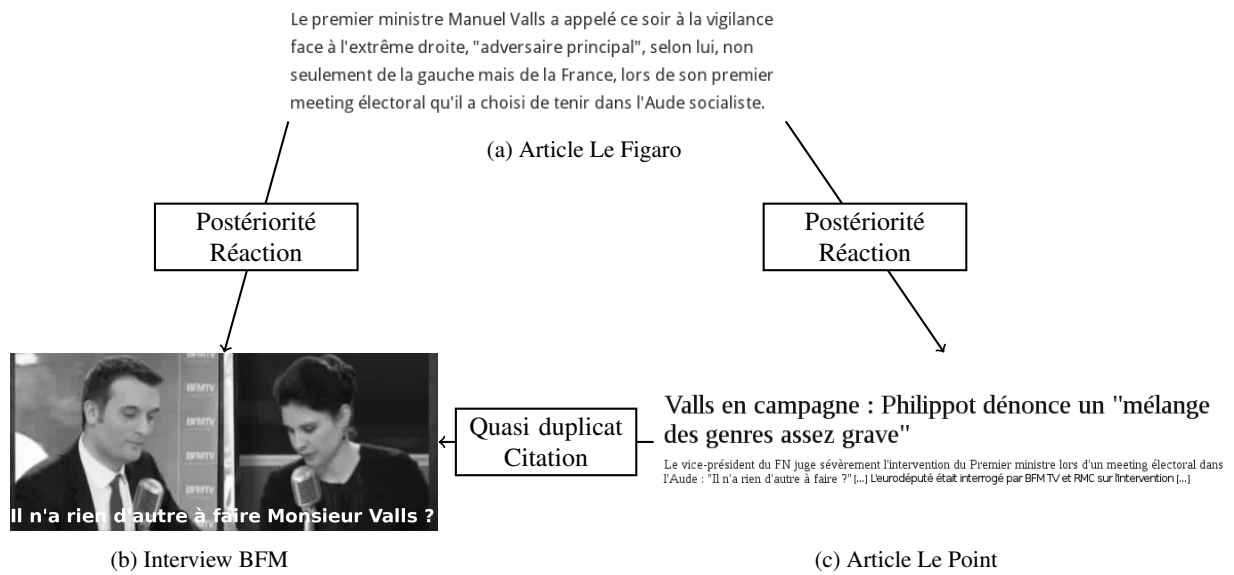


FIGURE 2: Divers liens entre trois informations

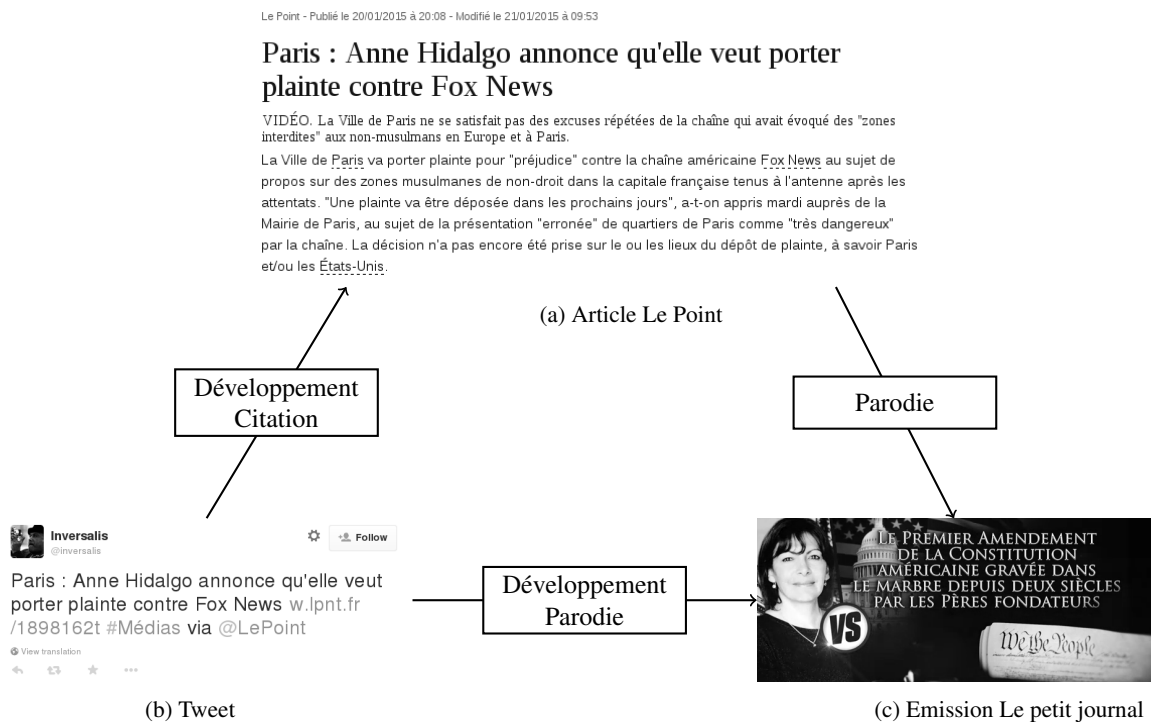


FIGURE 3: Liens de parodie et de développement

Les liens créés entre ces documents ou fragments de documents (*e.g.* dans la figure 3 on trouve des documents entiers tels le tweet ou l'article du Point ainsi qu'un fragment de l'émission télévisuelle « Le petit journal ») sont indépendants de la modalité de ces derniers. La figure 2 montre ainsi un article reprenant les propos tenus par M. Florian Philippot lors d'une interview. L'article n'apporte pas davantage d'informations que sa source, et un lien de quasi duplicat est donc créé entre les deux fragments. On peut néanmoins envisager que certains liens soient plus fréquents pour certaines modalités. Ainsi, les tweets, de par leur limite de 140 caractères, ont peu de chances de développer une information présente dans un autre média.

4 Conclusion et travaux futurs

Dans cet article, nous avons proposé une typologie pour catégoriser les liens entre des informations de type journalistique. Cette typologie offre la possibilité d'enrichir le parcours des utilisateurs en s'écartant de la logique des moteurs de recherche pour aller vers une navigation éclairée dans un ensemble de documents dont les différentes informations sont reliées entre elles.

Notre prochain objectif consiste à développer les algorithmes qui permettront de créer et de catégoriser ces liens de façon automatique. L'un des enjeux consiste à rendre ces algorithmes efficaces sur des gros volumes de données, mais aussi, à terme, d'envisager les moyens de mettre en place un système dynamique, permettant de créer de nouveaux liens pertinents au fur et à mesure que des documents lui sont présentés. Nous souhaitons également confronter notre typologie et les systèmes qui la mettront en œuvre à une réalité terrain, et préparons donc une campagne de retours d'utilisateurs afin de mieux cerner les avantages et les limites de tels liens.

La création automatique de liens typés pour enrichir un corpus multimodal est une tâche complexe et la typologie proposée est un premier pas pour encourager le développement de tels algorithmes.

Remerciements

Ce travail a bénéficié d'une aide de l'État attribuée au labex COMIN LABS et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01.

Références

- CLEARY C. & BAREISS R. (1996). Practical methods for automatically generating typed links. In *Proceedings of the the seventh ACM conference on Hypertext*, p. 31–41 : ACM.
- EKMAN P. (1992). An argument for basic emotions. *Cognition & emotion*, **6**(3-4), 169–200.
- ERTZSCHEID O. (2002). *Le lieu, le lien, le livre : les enjeux cognitifs et stylistiques de l'organisation hypertextuelle*. PhD thesis, Université de Toulouse 2.
- ESKEVICH M., JONES G. J., CHEN S., ALY R., ORDELMAN R. & LARSON M. (2012). Search and hyperlinking task at MediaEval 2012. *CEUR Workshop Proceedings*, **927**.
- IDE I., MO H., KATAYAMA N. & SATOH S. (2004). Topic threading for structuring a large-scale news video archive. In *Image and Video Retrieval*, p. 123–131. Springer.
- KWAK H., LEE C., PARK H. & MOON S. (2010). What is twitter, a social network or a news media ? In *Proceedings of the 19th international conference on World wide web*, p. 591–600 : ACM.
- MULLER P. & TANNIER X. (2004). Annotating and measuring temporal relations in texts. In *Proceedings of the 20th international conference on Computational Linguistics*, p. 50–56 : ACL.
- NALLAPATI R., FENG A., PENG F. & ALLAN J. (2004). Event threading within news topics. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, p. 446–453 : ACM.
- NANBA H., KANDO N. & OKUMURA M. (2011). Classification of research papers using citation links and citation types : Towards automatic review article generation. *Advances in Classification Research Online*, **11**(1), 117–134.

- RENNISON E. (1994). Galaxy of news : An approach to visualizing and understanding expansive news landscapes. In *Proceedings of the 7th annual ACM symposium on User interface software and technology*, p. 3–12 : ACM.
- SHAHAF D. & GUESTRIN C. (2010). Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 623–632 : ACM.
- THELWALL M. (2003). What is this link doing here ? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information research*, **8**(3).

CDGFr, un corpus en dépendances non-projectives pour le français

Denis Béchet¹ Ophélie Lacroix²

(1) LINA, 44322 Nantes Cedex 3, France

(2) LIMSI-CNRS, 91403 Orsay Cedex, France

denis.bechet@univ-nantes.fr, ophelie.lacroix@limsi.fr

Résumé. Dans le cadre de l'analyse en dépendances du français, le phénomène de la non-projectivité est peu pris en compte, en majeure partie car les données sur lesquelles sont entraînés les analyseurs représentent peu ou pas ces cas particuliers. Nous présentons, dans cet article, un nouveau corpus en dépendances pour le français, librement disponible, contenant un nombre substantiel de dépendances non-projectives. Ce corpus permettra d'étudier et de mieux prendre en compte les cas de non-projectivité dans l'analyse du français.

Abstract.

CDGFr, a Non-projective Dependency Corpus for French.

The non-projective cases, as a part of the dependency parsing of French, are often disregarded, mainly because the treebanks on which parsers are trained contain little or no non-projective dependencies. In this paper, we present a new freely available dependency treebank for French that includes a substantial number of non-projective dependencies. This corpus can be used to study and process non-projectivity more effectively within the context of French dependency parsing.

Mots-clés : Corpus français, annotation en dépendances, dépendances non-projectives.

Keywords: Treebank for French, dependency annotation, non-projective dependencies.

1 Introduction

Les développements actuels d'analyseurs syntaxiques en TALN reposent principalement sur l'utilisation de corpus arborés. Ces données se substituent à la notion plus traditionnelle de grammaire que l'on utilise maintenant principalement pour modéliser la syntaxe des langages artificiels comme les langages informatiques. De nombreuses raisons peuvent expliquer ce phénomène comme la difficulté de développer une grammaire pour une langue qui soit précise, robuste, à large couverture, qui puisse évoluer au cours du temps, en fonction du domaine, etc. De l'autre côté, l'approche *dirigée par les données* est simple à mettre en œuvre, rapide (si l'on ne tient pas compte de la difficulté de disposer des données d'apprentissage) et donne de bons résultats à la fois en termes de vitesse d'analyse et de robustesse.

Il semble important de noter que même si les analyseurs n'ont plus besoin de grammaire pour fonctionner, les corpus qu'ils utilisent lors de la phase d'apprentissage reposent souvent sur un modèle linguistique qui a permis de créer directement ce corpus ou bien de l'obtenir par des transformations depuis d'autres ressources qui elles ont un modèle linguistique. Suivant le point de vue de Dikovsky (2011), le lien entre les corpus et les modèles sur lesquels ils reposent et en tenant compte des transformations utilisées est essentiel à la compréhension des corpus. Par exemple, les corpus FTB (*French Treebank*) (Abeille *et al.*, 2003) et Sequoia (Candito & Seddah, 2012) permettent d'entraîner des analyseurs syntaxiques en dépendances après transformation en arbres de dépendances (Candito *et al.*, 2010). Le modèle en constituant de départ ne permet en général pas d'obtenir beaucoup de dépendances non-projectives (i.e. des dépendances qui peuvent croiser les autres dépendances en raison d'une discontinuité dans la langue comme la dépendance entre « en » et « directives » de la figure 1). Par conséquent, un analyseur basé sur ces corpus va peu produire ce type de structures même s'il est capable de traiter ces dépendances efficacement.

D'ailleurs, les corpus en dépendances librement accessibles pour le français et comportant un nombre conséquent de structures de dépendances non-projectives ne sont pas très nombreux. Le FTB convertit en dépendances ne contient pas de dépendances non-projectives. Le corpus Sequoia comprend 1,2 % de structures de dépendances non-projectives et, plus récent, le corpus UDT (*Universal Dependency Treebank*) (McDonald *et al.*, 2013) en comprend 12,4 % pour le français,

ce qui correspond à, respectivement, 0,2 % et 2,4 % de dépendances non-projectives¹ sur l'ensemble des dépendances de chacun de ces corpus. En conséquence, les cas de non-projectivité dans la langue française sont difficiles à traiter et sont donc peu ou pas pris en compte lors de l'analyse de ces données. Notons également qu'il existe d'autres corpus en dépendances pour le français avec lesquels nous ne nous comparerons pas directement puisque ces corpus proposent des représentations en dépendances différentes de celles que nous proposons et donc ciblent des usages différents. Par exemple, il existe une version en dépendances profondes du Sequoia (Candito *et al.*, 2014). Citons également le corpus Rhapsodie (Lacheret *et al.*, 2014) traitant du français parlé.

Pour ces raisons, nous avons entrepris le développement d'un corpus de structures de dépendances centré sur la notion de dépendances non-projectives comprenant un noyau ayant servi, dans un premier temps, à développer en parallèle une grammaire catégorielle pour le français et son corpus de développement puis, dans un deuxième temps, à ajouter des corpus supplémentaires. Et tandis que les corpus en dépendances existants pour le français rassemblent des phrases provenant majoritairement de textes journalistiques, nous avons choisi ici d'annoter des extraits de textes divers dont des périodiques mais également des textes littéraires variés en terme de style et de genre. Les structures de dépendances du corpus sont disponibles² sous la licence LGPL-LR (Lesser General Public Licence For Linguistic Resources)³, les œuvres d'origines restant la propriété des auteurs ou de leurs ayants droit. Le corpus a été développé en suivant la méthode proposée dans (Dikovsky, 2011) basée sur la construction incrémentale d'une grammaire catégorielle de dépendance (CDG) et d'un corpus de développement. Nous avons utilisé pour cela l'environnement de développement des CDG proposé par (Béchet *et al.*, 2014). Nous pensons que ce corpus permettra de montrer l'intérêt d'étudier et de prendre en compte les structures non-projectives dans les langues naturelles en particulier pour le français.

Dans la suite de l'article, la structure générale du corpus et la provenance des textes sont abordées. Nous présentons ensuite la structure et le schéma d'annotation des unités lexicales (mots ou groupe de mots), de leurs classes grammaticales et de leurs traits puis la structure des types de dépendances présents dans les arbres de dépendances. La section révèle également la méthodologie employée lors de la création du corpus et de ses annotations et comporte une analyse statistique du contenu. Nous évoquons pour finir des travaux d'analyse en dépendances dirigés par les données présentant des résultats préalables sur le corpus CDGFr.

2 Corpus

2.1 Origines

Le corpus CDGFr comporte des phrases de trois origines différentes : un noyau de développement du modèle linguistique, des extraits de textes littéraires de la fin du XIX^{ème} siècle et du XX^{ème} siècle et des extraits de périodiques contemporains. La première partie du corpus, nommée **CDGFr-devel** est particulièrement importante. Elle est constituée de phrases souvent assez courtes provenant de multiples sources et représentant les différentes structures syntaxiques du français. Les autres sources (littérature et périodiques) ont permis de valider le modèle sur des extraits réels de textes.

Littérature

Les œuvres littéraires ont été choisies en fonction de leur style assez différents les uns des autres.

- **Zola** : extrait du chapitre 1 de « *Germinal* » de E. Zola publié en 1885.
- **Céline** : extrait du chapitre 1 du « *Voyage au bout de la nuit* » de L.F. Céline publié en 1932.
- **Camus** : extrait du chapitre 1 de la première partie de « *L'étranger* » de A. Camus publié en 1942.
- **Le Clézio** : extrait de « *L'échappé* » de « *La ronde et autres faits divers* » de J.M.G. Le Clézio publié en 1982.

Dans *Germinal*, on trouve des phrases descriptives souvent longues avec des constructions apposées donnant lieu à des dépendances parfois très longues. Le texte est ponctué de dialogues avec leur syntaxe propre, enchâssés dans le texte. Le style de Céline tient plus du langage parlé parfois pas très grammatical (au sens académique du terme). Les phrases y sont plutôt courtes même s'il a été choisi de parfois les regrouper en une seule analyse comme par exemple la phrase suivante :

1. Formellement, une dépendance est non-projective s'il existe au moins un mot situé entre sa tête et son dépendant direct qui n'est pas dominé par la tête, i.e. qui ne dépend pas directement ou indirectement de la tête.

2. La version actuelle est consultable à l'adresse suivante : <http://pagesperso.lina.univ-nantes.fr/~bechet-d/CDGFr>

3. <http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/lgpllr.html>

Quand il fait très froid, non plus, il n'y a personne dans les rues ; c'est lui, même que je m'en souviens, qui m'avait dit à ce propos : « Les gens de Paris ont l'air toujours d'être occupés, mais en fait, ils se promènent du matin au soir ; la preuve, c'est que lorsqu'il ne fait pas bon à se promener, trop froid ou trop chaud, on ne les voit plus ; ils sont tous dedans à prendre des cafés-crème et des bocks. C'est ainsi ! Siècle de vitesse ! qu'ils disent. Où ça ? Grands changements ! qu'ils racontent. Comment ça ? Rien n'est changé en vérité. Ils continuent à s'admirer et c'est tout. Et ça n'est pas nouveau non plus. Des mots, et encore pas beaucoup, même parmi les mots, qui sont changés ! Deux ou trois par-ci, par-là, des petits ... »

Les phrases de Camus portent aussi sur un texte à la première personne mais dans un style beaucoup plus neutre. Finalement, le texte de Le Clézio est au présent et comporte peu de dialogues. Sa forme est moderne et neutre.

Ainsi, à travers ses quatre textes, nous avons un échantillon assez large de formes : description au passé, au présent ; dialogue ordinaire, dialogue argotique, narration à la première personne. Nous pensons qu'ils représentent une partie importante des structures syntaxiques utilisées dans la littérature française.

Périodiques

Les deux extraits choisis comportent un texte de journalisme scientifique paru dans le journal Le Monde et un texte de journalisme historique paru dans la revue mensuelle de la ville de Nantes distribuée gratuitement dans l'agglomération nantaise et disponible sur Internet⁴. Ce ne sont donc pas des dépêches mais plutôt des textes déjà bien construits exprimant des faits réels actuels ou passés.

- **Nantes Passion** : extrait de l'article « Il y a 70 ans, le procès des « 42 » » de L. Abed-Denesle pour le magazine municipale mensuel Nantes Passion (numéro 230, janvier 2013)
- **Univers** : extrait de l'article « L'enfance de l'univers dévoilée » de J.L. Puget pour Le Monde (22 mars 2013)

2.2 Caractéristiques

Les caractéristiques des différents corpus présentés précédemment sont exposées dans la table 1, ainsi que les caractéristiques du corpus total correspondant à l'union de ces corpus. Il est intéressant d'observer les différences effectives entre les données provenant de ces différentes sources. Les corpus dont les phrases proviennent d'oeuvres littéraires rassemblent les phrases les plus longues. On remarquera en particulier le corpus Zola, dont la longueur des phrases varie fortement et qui comprennent en moyenne 30 mots. Par ailleurs, le corpus CGDFr-devel concentre le plus grand nombre de phrases mais celles-ci sont relativement courtes en moyenne par rapport à celles des autres sources.

Corpus	Phrases			Mots (ponctuations comprises)	
	total	longueur moyenne	écart-type	total	formes fléchies diff.
CDGFr-devel	1 995	11,1	7,0	22 195	3 526
Zola	100	30,0	25,0	3 004	1 029
Céline	91	19,8	25,1	1 801	603
Camus	319	16,5	12,4	5 253	1 268
Le Clézio	528	18,7	10,2	9 894	1 730
Nantes Passion	42	22,8	13,4	957	436
Univers	64	25,3	13,5	1 619	643
Total	3 139	14,3	11,5	44 723	5 892

TABLE 1 – Caractéristiques des corpus

4. <http://www.nantes.fr/nantes-passion>

3 Annotations

3.1 Schéma d’annotation morpho-syntaxique (en classes grammaticales)

La grammaire catégorielle servant de modèle linguistique regroupe les unités lexicales en classes ayant des propriétés syntaxiques proches. Ces classes sont elles-mêmes regroupées en classes grammaticales générales listées dans la table 2. Une classe se distingue d’une autre classe, dans la même classe générale, par sa fonction, sa forme ou le type de ses arguments. Par exemple, pour la classe générale *N* des noms, *N(Lex=proper)* correspond aux noms propres, *N(Lex=common)* aux noms communs, *N(Lex=time)* aux noms des dates ou du temps ainsi qu’aux adverbes pouvant être une réponse à une question comme « quand venez-vous ? » : « aujourd’hui », « bientôt », « jamais », « heure », « avril », etc. Dans la classe générale *Vi* des verbes intransitifs, on trouvera *Vi(F=fin)* pour les formes finies (conjuguées), *Vi(F=inf)* pour les infinitifs, *Vi(F=pz, T=past)* pour les participes passés et *Vi(F=pz, T=pres)* pour les participes présents. Les verbes avec deux arguments de *V2t* sont regroupés suivant le type des deux compléments : *V2t(F=fin, C1=a, C2=d)* correspond aux verbes avec un complément d’objet direct (accusatif) et un complément second introduit par la préposition *à* (datif), etc. En plus de la classe grammaticale de chaque unité de la phrase, le corpus précise sa forme dans le lexique (par exemple, sans majuscule pour le premier mot d’une phrase), sa forme normalisée (infinitif pour un verbe ou forme au masculin singulier) et la liste des traits (genre, nombre, mode, temps, personne). Les deux derniers attributs indiquent la provenance de la forme dans le lexique (principalement basée sur le Lefff (Sagot, 2010)) et la manière avec laquelle la forme a été associée à la classe grammaticale (basée principalement sur les informations disponibles sur les formes et leurs arguments dans le Lefff). Quelques classes supplémentaires ont été introduites pour les termes qui ne sont pas reconnus par le lexique : les formes pouvant correspondre à des noms propres (avec des majuscules), à des nombres (composés de chiffres), des nombres complexes en toutes lettres et sinon des termes inconnus de classe *UT(Lex=V|N|Adj|Adv)* traitée comme une agrégation des propriétés syntaxiques des verbes, des noms, des adjectifs et des adverbes. Des exemples de phrases comportant cette classe se trouvent dans le corpus *CDGFr-devel* : « Adam va y xxx bientôt » où « xxx » porte cette classe. Dans ce cas, la pseudo-forme associée dans le lexique est *\$UnknownTerm*. La classe des termes inconnus est utilisée dans le corpus pour signifier qu’un lexique ne pourra jamais être complet (terme ancien, trop récent, peu utilisé, utilisé localement, faute d’orthographe, etc) et qu’il peut être intéressant de les traiter de cette manière.

Adjectifs	Adj	Prépositions	PP	Ponctuations	Dash
Adverbes	Adv	Verbes auxiliaires	Vaux		Parentheses
Collocations	Colloc	Verbes copules	Vcopul		QuestMark
Conjonctions	Conj	Verbes intransitifs	Vi		Quotes
Déterminants	Det	Verbes substituts	Vlight		SemiColon
Interjections	Expletives	Verbes transitifs	Vt		Chevrans
Noms	N	Verbes ditransitifs	V2t		Colon
Nombres	Num	Unités inconnues	UT		FullStop
Partitifs	Part				EmphatMark
Pronoms	PN				Comma

TABLE 2 – Liste des classes grammaticales générales de la grammaire catégorielle de dépendances du français

3.2 Schéma d’annotation en dépendances

Les structures de dépendances du corpus sont techniquement des graphes acycliques orientés (*DAG*) superposant deux arbres dont les dépendances sont des arcs de différents types. Ceux-ci sont de 3 types : les dépendances projectives (*projective*), les dépendances discontinues (*discontinuous*) et les ancrs (*anchor*). La majeure partie des dépendances sont des dépendances projectives (les arcs pleins en noir dans la figure 1) tandis que les dépendances discontinues (les arcs en pointillé) sont des dépendances qui peuvent croiser les autres dépendances dans la structure. Les ancrs sont, en outre, des pseudo-dépendances de deux sortes : d’une part, la plupart des pseudo-dépendances arrivant sur une ponctuation sont des ancrs (les arcs dont l’étiquette est préfixée par « @ »), d’autre part, chaque mot recevant une dépendance discontinue reçoit également une ancre (les arcs dont l’étiquette est préfixée par « # ») de même rôle syntaxique, e.g. le mot « en » dans la figure 1 possède donc deux têtes, une vraie tête discontinue « données » et une pseudo-tête projective « sont ». De la sorte, l’union des dépendances projectives et des dépendances discontinues, pour une phrase donnée, forme

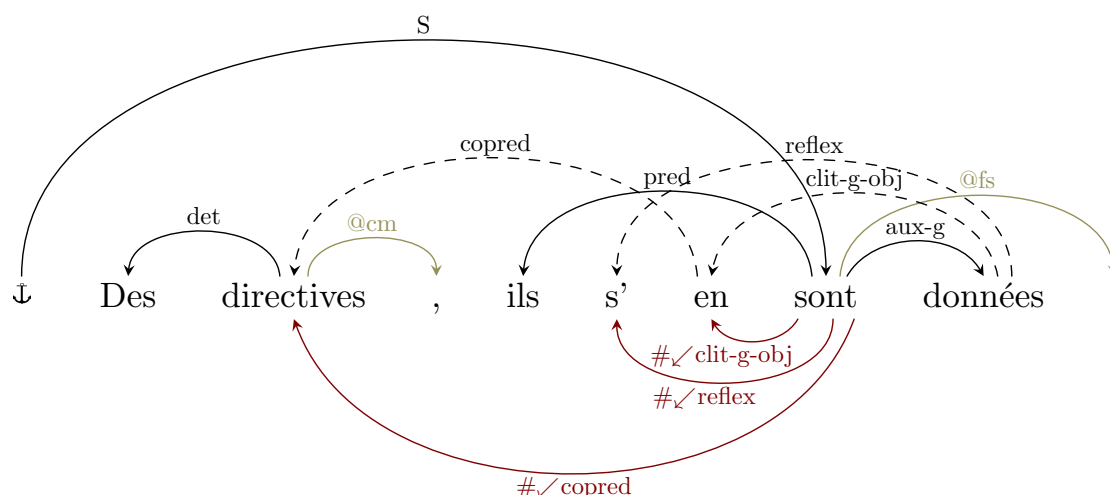


FIGURE 1 – Structure de dépendances du corpus CDGFr-devel pour la phrase « Des directives, ils s'en sont données »

un arbre éventuellement non-projectif⁵ tandis que l'union des dépendances projectives et des ancrs forme toujours un arbre projectif.

De plus, les dépendances se distinguent également par le rôle syntaxique (nom de la dépendance) qui leur est associé. La terminologie des noms de dépendances provient de la théorie sens-texte de Mel'čuk & Pertsov (1986); Mel'čuk (1988). Une présentation succincte des dépendances utilisées ici se trouve dans (Dikovsky, 2011). La présentation exhaustive est disponible dans la documentation associée au corpus. On y trouve notamment les informations sur le site d'ancrage des mots sur lesquels une dépendance discontinue arrive. Les noms des dépendances (figurants en majuscules dans le corpus) sont regroupés en groupes (figurants en minuscules). Les dépendances d'un même groupe se distinguent par le cas de l'argument (nominatif, accusatif, datif, etc), son type (nom, adjectif, complément verbal), ou des propriétés du contexte. Par exemple, dans le groupe *OBJ* des arguments des verbes (autre que le sujet), on trouve *a-obj* pour le complément d'objet direct, *d-obj* pour un complément indirect au datif (introduit par « à »), etc. Le groupe *COPUL* des copules comporte la dépendance *a_copul* pour les adjectifs attributs (comme « il est jeune »), *n_copul* pour les noms attributs (comme « il s'appelle Pierre ») ou *c_copul* pour des compléments introduits par une préposition (comme « il était au commencement »). Dans le groupe *OBJ*, la dépendance *a-obj-g* lie un verbe et son complément d'objet direct mais indique aussi que le complément doit posséder une dépendance vers un clitique placé avant le verbe comme la dépendance entre « avait » et « besoin » dans la phrase « il en avait besoin ».

AGENT	(0,21 % / 0,02 %)	COPUL	(2,67 % / 0,04 %)	PRED	(9,94 % / 0,00 %)
AGGR	(1,46 % / 0,05 %)	CORREL	(0,04 % / 0,00 %)	PREFIXA	(0,02 % / 0,00 %)
APPOS	(0,92 % / 0,15 %)	DEICT	(0,03 % / 0,00 %)	PREPOS	(8,29 % / 0,00 %)
APPROX	(0,06 % / 0,00 %)	DET	(10,64 % / 0,00 %)	PUNCT	(12,74 % / 0,00 %)
ATTR	(2,95 % / 0,02 %)	EMPHAT	(0,71 % / 0,00 %)	QUANT	(0,92 % / 0,08 %)
AUX	(2,35 % / 0,00 %)	EXPLET	(0,11 % / 0,02 %)	QUANTIF	(0,29 % / 0,00 %)
CIRC	(6,38 % / 0,00 %)	GER	(0,15 % / 0,00 %)	REFLEX	(0,97 % / 0,50 %)
CLAUS	(2,83 % / 0,00 %)	INF	(2,96 % / 0,00 %)	REL	(1,03 % / 0,26 %)
CLIT	(1,82 % / 0,85 %)	INTERROG	(0,06 % / 0,00 %)	RESTRICT	(1,14 % / 0,02 %)
COMPAR	(0,44 % / 0,00 %)	JUNC	(3,49 % / 0,00 %)	SELECT	(0,12 % / 0,02 %)
CONJ	(0,47 % / 0,00 %)	MODIF	(3,14 % / 0,08 %)	SENT	(7,53 % / 0,00 %)
COORDV	(1,36 % / 0,00 %)	NEG	(1,51 % / 1,15 %)	VOCATIVE	(0,19 % / 0,00 %)
COPRED	(0,08 % / 0,33 %)	OBJ	(6,26 % / 0,12 %)		

TABLE 3 – Liste des groupes de dépendances de la grammaire catégorielle de dépendances du français et pourcentage de dépendances (projectives / discontinues) parmi les dépendances associées aux groupes dans l'ensemble du corpus

5. Une structure de dépendances ne peut être non-projective que si elle contient au moins une dépendance discontinue.

3.3 Méthodologie d’annotation

Le corpus comporte deux parties bien distinctes. Le corpus *CDGFr-devel* a été développé en même temps que le modèle grammatical associé qui se présente sous la forme d’une grammaire catégorielle de dépendances. Le développement a consisté, à partir d’une grammaire catégorielle noyau du français, en l’ajout progressif au corpus de phrases introduisant de nouvelles structures syntaxiques et de la modification correspondante de la grammaire catégorielle.

La seconde partie comportant les corpus sur la littérature et les périodiques a été créée en utilisant les outils d’analyse syntaxiques dérivés du modèle défini par la première phase, la documentation dérivée portant sur les dépendances syntaxiques et les exemples d’analyses fournis par le corpus de développement. Ce processus d’annotation semi-automatique pouvait également être renforcé, suivant le choix de l’annotateur, par une étape de pré-annotation manuelle (combinant segmentation et étiquetage grammatical) permettant de réduire le temps de la phase d’analyse avec la grammaire. Après analyse, le processus a requis une étape de validation manuelle des structures de dépendances et des annotations morphosyntaxiques. En dernier ressort, les analyses sélectionnées ont été vérifiées par l’équipe qui a créé le modèle. Les problèmes éventuels ont été détectés pour que la grammaire initiale et les corpus déjà créés puissent être modifiés (par exemple si un mot n’appartenait pas à une classe grammaticale ou plus rarement, si une structure syntaxique a été oubliée ou n’était pas utilisée de manière cohérente).

En outre, le corpus a d’ores et déjà été exploité pour l’élaboration d’outils de pré-annotation automatique supervisé (Lacroix *et al.*, 2014) dans le but d’améliorer la rapidité et le confort d’annotation pour le développement futur du corpus.

3.4 Statistiques

La table 4 rassemble les statistiques sur les structures de dépendances et les dépendances elles-mêmes pour les différents corpus annotés. En particulier, sont présentés les taux de dépendances discontinues et de structures de dépendances non-projectives. Les statistiques sur les dépendances ne prennent pas en compte les ancrs (i.e. les ancrs liées aux ponctuations et les ancrs qui vont de pair avec les dépendances discontinues).

Corpus	Structures de dépendances			Dépendances		
	total	non-projectives (%)		total	discontinues (%)	
CDGFr-devel	1 995	864	(43,3 %)	19 340	1 095	(5,7 %)
Zola	100	41	(41,0 %)	2 646	65	(2,7 %)
Céline	91	36	(39,6 %)	1 560	68	(4,4 %)
Camus	319	158	(49,5 %)	4 707	216	(4,6 %)
Le Clézio	528	151	(28,6 %)	8 791	187	(2,1 %)
Nantes Passion	42	9	(21,4 %)	868	9	(1,0 %)
Univers	64	21	(32,8 %)	1 460	21	(1,4 %)
Total	3 139	1 280	(40,8 %)	39 372	1 661	(4,2 %)

TABLE 4 – Statistiques sur les dépendances et les structures

4 Analyse en dépendances

À partir du schéma d’annotation en dépendances présenté en section 3.2 il est possible d’extraire, depuis le corpus CDGFr, des arbres standards possiblement non-projectifs ou uniquement projectifs (en conservant les ancrs à la place des dépendances discontinues). Ces arbres sont donc adaptables aux systèmes standards d’analyse en dépendances dirigés par les données tels que les analyseurs par transition ou les analyseurs basés sur les graphes. Inversement, tout corpus standard non-projectif peut également être adapté à la représentation en dépendances des CDG par l’emploi d’une méthode automatique de projectivisation des dépendances non-projectives.

Des résultats d’analyses en dépendances, effectuées sur le corpus CDGFr à l’aide de différents algorithmes d’analyse par transition, peuvent être trouvés dans (Lacroix & Béchet, 2014), ainsi que la présentation d’un algorithme spécialement adapté à la représentation en dépendances exploitées dans nos corpus.

5 Conclusion

Nous avons présenté un nouveau corpus arboré en dépendances pour le français contenant un nombre substantiel de dépendances non-projectives. Les différents corpus formant le corpus intégral sont visualisables à l'adresse <http://pagesperso.lina.univ-nantes.fr/~bechet-d/CDGFr> et téléchargeable au format XML. Cet ensemble de phrases annotées en dépendances servira à l'étude, notamment, des phénomènes non-projectifs (discontinuités) dans la langue française et ainsi amènera à prendre en compte et à mieux traiter ces cas particuliers dans le cadre de l'analyse en dépendances globale de données sur le français. Certaines dénominations comme le nom de la dépendance *pred* (predicative) proviennent de (Mel'čuk & Pertsov, 1986; Mel'čuk, 1988). Par respect pour le travail d'Alexandre Dikovsky présenté dans (Dikovsky, 2011), nous n'avons pas cherché à les renommer dans cette version du corpus (version 3.4).

Remerciements

La grammaire de CDG sur laquelle repose le CDGFr a été développée par Alexandre Dikovsky qui est aussi l'auteur du rapport technique sur les dépendances que l'on peut consulter sur le site de téléchargement du corpus. Le lexique associé a été développé par Denis Béchet, Alexandre Dikovsky et Ramadan Alfared. Nous voudrions remercier particulièrement Danièle Bauquier qui a analysé une partie conséquente du corpus. Nous dédions cette article à Alexandre Dikovsky qui nous a malheureusement quitté en 2014. Sans lui ce projet n'aurait jamais existé.

Références

- ABEILLE A., CLEMENT L. & TOUSSENEL F. (2003). *Building a Treebank for French*, In A. ABEILLÉ, Ed., *Treebanks*, volume 20 of *Text, Speech and Language Technology*, p. 165–187. Springer Netherlands.
- BÉCHET D., DIKOVSKY A. & LACROIX O. (2014). “CDG Lab” : an Integrated Environment for Categorical Dependency Grammar and Dependency Treebank Development. In *Computational Dependency Theory*, volume 258 of *Frontiers in Artificial Intelligence and Applications*, p. 153–169. IOS Press.
- CANDITO M., CRABBÉ B. & DENIS P. (2010). Statistical French Dependency Parsing : Treebank Conversion and First Results. In *Proceedings of the Language Resources and Evaluation Conference*, LREC 2010, Valletta, Malta.
- CANDITO M., PERRIER G., GUILLAUME B., RIBEYRE C., FORT K., SEDDAH D. & DE LA CLERGERIE É. (2014). Deep Syntax Annotation of the Sequoia French Treebank. In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC 2014, Reykjavik, Iceland.
- CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, Grenoble, France.
- DIKOVSKY A. (2011). Categorical Dependency Grammars : from Theory to Large Scale Grammars. In *Proceedings of the International Conference on Dependency Linguistics*, DEPLING 2011, Barcelona, Spain.
- LACHERET A., KAHANE S., BELIAO J., DISTER A., GERDES K., GOLDMAN J.-P., OBIN N., PIETRANDREA P., TCHOBANOV A. *et al.* (2014). Rhapsodie : a Prosodic-Syntactic Treebank for Spoken French. In *Proceedings of the International Conference on Language Resources and Evaluation*, LREC 2014, Reykjavik, Iceland.
- LACROIX O. & BÉCHET D. (2014). A three-step transition-based system for non-projective dependency parsing. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING 2014, Dublin, Irlande.
- LACROIX O., BÉCHET D. & BOUDIN F. (2014). Label pre-annotation for building non-projective dependency treebanks for french. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics*, CICLing 2014, Kathmandu, Népal.
- MCDONALD R., NIVRE J., QUIRMBACH-BRUNDAGE Y., GOLDBERG Y., DAS D., GANCHEV K., HALL K., PETROV S., ZHANG H., TÄCKSTRÖM O., BEDINI C., BERTOMEU CASTELLÓ N. & LEE J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL'13, Sofia, Bulgaria.
- MEL'ČUK I. (1988). *Dependency syntax : Theory and Practice*. State University of New York Press.
- MEL'ČUK I. A. & PERTSOV N. V. (1986). *Surface syntax of English : A formal model within the Meaning-Text framework*. John Benjamins Publishing Company.
- SAGOT B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, LREC'10, Valletta, Malte.

Utilisation des réseaux de neurones récurrents pour la projection interlingue d'étiquettes morpho-syntaxiques à partir d'un corpus parallèle

Othman Zennaki^{1,2} Nasredine Semmar¹ Laurent Besacier²

(1) CEA, LIST, Laboratoire Vision et Ingénierie de Contenus, F-91191, Gif-sur-Yvette, France

(2) Laboratoire d'Informatique de Grenoble, Univ. Grenoble-Alpes, Grenoble, France

othman.zennaki, nasredine.semmar@cea.fr, laurent.besacier@imag.fr

Résumé. La construction d'outils d'analyse linguistique pour les langues faiblement dotées est limitée, entre autres, par le manque de corpus annotés. Dans cet article, nous proposons une méthode pour construire automatiquement des outils d'analyse via une projection interlingue d'annotations linguistiques en utilisant des corpus parallèles. Notre approche n'utilise pas d'autres sources d'information, ce qui la rend applicable à un large éventail de langues peu dotées. Nous proposons d'utiliser les réseaux de neurones récurrents pour projeter les annotations d'une langue à une autre (sans utiliser d'information d'alignement des mots). Dans un premier temps, nous explorons la tâche d'annotation morpho-syntaxique. Notre méthode combinée avec une méthode de projection d'annotation basique (utilisant l'alignement mot à mot), donne des résultats comparables à ceux de l'état de l'art sur une tâche similaire.

Abstract.

Use of Recurrent Neural Network for Part-Of-Speech tags projection from a parallel corpus.

In this paper, we propose a method to automatically induce linguistic analysis tools for languages that have no labeled training data. This method is based on cross-language projection of linguistic annotations from parallel corpora. Our method does not assume any knowledge about foreign languages, making it applicable to a wide range of resource-poor languages. No word alignment information is needed in our approach. We use Recurrent Neural Networks (RNNs) as cross-lingual analysis tool. To illustrate the potential of our approach, we firstly investigate Part-Of-Speech (POS) tagging. Combined with a simple projection method (using word alignment information), it achieves performance comparable to the one of recently published approaches for cross-lingual projection.

Mots-clés : Multilinguisme, transfert crosslingue, étiquetage morpho-syntaxique, réseaux de neurones récurrents.

Keywords: Multilingualism, cross-Lingual transfer, part-of-speech tagging, recurrent neural network.

1 Introduction

L'annotation linguistique de ressources consiste à ajouter des informations de nature interprétative aux données brutes originales (Garside *et al.*, 1997). Ces informations peuvent être d'ordre terminologique, lexical, morphologique, syntaxique ou sémantique et les ressources linguistiques peuvent être des lexiques, des dictionnaires, des transcriptions de dialogue ou des corpus de textes (Véronis, 2000). Ces ressources linguistiques sont annotées par des outils d'analyse linguistique et utilisées dans de nombreuses applications : recherche d'information translingue, fouille de textes, extraction d'informations, traduction automatique, etc.

Dans la littérature, il a été montré que les outils d'analyse linguistique les plus performants sont ceux construits pour les quelques langues (richement dotées) disposant des ressources linguistiques manuellement annotées nécessaires aux algorithmes d'apprentissage supervisé. Cependant, la plus grande majorité des langues (faiblement dotées) ne disposent pas de telles ressources annotées.

La construction manuelle de ces ressources est lente et coûteuse, rendant ainsi l'utilisation des approches supervisées difficile voire impossible. Dans cet article, nous nous intéressons à l'induction de ressources linguistiques adéquates à moindre coût pour les langues faiblement dotées, et aussi à la construction automatique d'outils d'analyse linguistique pour ces langues. Pour cela, nous proposons d'utiliser des approches fondées sur la *projection interlingue d'annotations*.

Celles-ci s'articulent autour de l'exploitation des corpus parallèles multilingues entre une langue source richement dotée (disposant d'outils d'analyse linguistique) et une langue cible faiblement dotée. En partant d'un corpus parallèle dont les

textes en langue *source* sont déjà annotés, les textes en langue *cible* sont annotés par projection des annotations à l'aide de techniques d'alignement automatique au niveau des mots.

Bien que prometteuses, ces approches non supervisées ont des performances assez éloignées de celles des méthodes supervisées. Par exemple, pour une tâche d'analyse morpho-syntaxique supervisée, (Petrov *et al.*, 2012) obtient une précision moyenne de 95.2% pour 22 langues richement dotées, tandis que les analyseurs morpho-syntaxiques non supervisés construits par (Das & Petrov, 2011; Duong *et al.*, 2013) donnent une précision moyenne de 83.4% pour 8 langues Européennes.

Dans cet article, nous explorons la possibilité d'employer les réseaux de neurones récurrents (RNN) pour induire des outils multilingues d'analyse linguistique. Dans un premier temps, nous abordons la possibilité de les utiliser comme analyseurs morpho-syntaxiques. Pour cela, nous utilisons un corpus parallèle entre une langue bien dotée et une autre langue moins bien dotée, pour assigner aux mots du corpus parallèle (appartenant aux vocabulaires des langues source et cible) une représentation commune, obtenue à partir d'un alignement au niveau des phrases. Cette représentation commune permet d'apprendre — à partir d'une seule langue étiquetée parmi N — un seul analyseur multilingue capable de traiter N langues.

Après un bref état de l'art présenté dans la section 2, notre modèle est décrit dans la partie 3 et son évaluation est présentée dans la partie 4, la partie 5 conclut notre étude et présente nos travaux futurs.

2 État de l'art

La projection interlingue d'annotations a été introduite par (Yarowsky *et al.*, 2001), en utilisant un corpus parallèle pour adapter des outils monolingues (analyseurs morpho-syntaxiques, analyseurs syntaxiques de surface et analyseurs morphologiques) à de nouvelles langues. Le transfert entre les langues a été rendu effectif en utilisant les alignements au niveau des mots entre les phrases d'un corpus parallèle. Plusieurs outils permettent d'obtenir automatiquement de tels alignements, dont GIZA++ (Och & Ney, 2000). Cet outil implémente divers modèles de traduction (IBM 1, 2, 3, 4, 5 et HMM). Ces modèles utilisent l'algorithme EM (Dempster *et al.*, 1977) pour l'apprentissage à partir de corpus bilingues. L'alignement des mots est réalisé à l'aide d'un algorithme de recherche de type Viterbi. GIZA++ est un outil efficace pour aligner les mots simples, mais il est moins performant, d'une part, lorsque les langues source et cible ont des morphologies et des structures syntaxiques différentes, et d'autre part, pour aligner les expressions multi-mots (Allauzen & Wisniewski, 2009; Abdulhay, 2012).

Cette méthode a été ensuite utilisée avec succès dans plusieurs autres travaux. Ainsi, (Das & Petrov, 2011; Duong *et al.*, 2013) ont montré qu'il était possible d'apprendre des analyseurs morpho-syntaxiques de bonne qualité de cette manière. Dans cette lignée, (Wisniewski *et al.*, 2014; Täckström *et al.*, 2013) ont obtenu de meilleures performances encore, en combinant les informations obtenues par projection avec les informations extraites d'un dictionnaire qui associe à chaque mot (de la langue cible) l'ensemble des étiquettes morpho-syntaxiques autorisées, puis en utilisant des méthodes d'apprentissage faiblement supervisées.

La projection interlingue a été aussi adaptée avec succès pour transférer d'autres types d'annotations. Par exemple, la projection d'annotations en sens réalisée par (Bentivogli *et al.*, 2004; Van der Plas & Apidianaki, 2014), l'annotation en rôles sémantiques sur l'allemand par projection interlingue à partir de la paire de langues anglais-allemand (Padó & Lapata, 2005, 2006), dont la généricité a plus spécifiquement été évaluée dans (Padó & Pitel, 2007). De plus cette méthode permet la portabilité multilingue des applications utilisant les annotations linguistiques, (Jabaian *et al.*, 2013) l'ont utilisé pour la portabilité d'un système de compréhension de la parole pour des langues ou domaines différents.

Dans ces approches, les annotations du côté source sont projetées vers le côté cible, à travers les alignements automatiques du corpus parallèle obtenus au niveau des mots. Cette annotation partielle et bruitée des textes cibles est ensuite utilisée par des méthodes d'apprentissage robustes. Cependant, les performances des algorithmes d'alignement au niveau des mots ne sont pas toujours satisfaisantes (du point de vue de la qualité des alignements prédits) et l'étape d'alignement au niveau des mots (un alignement n'est pas toujours 1-1, il peut être 1-N, N-N, etc.) constitue aujourd'hui un facteur limitant la projection d'annotations linguistiques (Fraser & Marcu, 2007). Pour cette raison, notre approche utilise un corpus parallèle aligné au niveau des phrases seulement et n'applique aucun pré-traitement du type *alignement automatique en mots* qui est source d'erreurs et de bruit.

3 Méthode proposée

Pour faire face aux limitations relatives à l'étape d'alignement mot à mot des phrases du corpus parallèle, nous proposons de ne pas prendre en compte les informations bruitées issues de cet alignement, mais de représenter ces informations

de façon intrinsèque dans l'architecture du réseau de neurones. Dans ce travail initial, nous implémentons un analyseur morpho-syntaxique multilingue basé sur les réseaux de neurones récurrents, et nous montrons que ses performances sont proches de l'état de l'art des autres analyseurs morpho-syntaxiques non supervisés.

Avant de décrire notre étiqueteur morpho-syntaxique multilingue basé sur les réseaux de neurones récurrents (RNN), nous décrivons tout d'abord l'approche par projection simple à laquelle nous allons nous comparer (et qui sera aussi combinée — au cours des expériences qui vont suivre — avec la méthode que nous proposons).

3.1 Annotateur morpho-syntaxique non supervisé par projection simple

L'approche pour construire notre étiqueteur morpho-syntaxique non supervisé par projection simple (décrit par l'algorithme 1) est très proche de celle introduite par (Yarowsky *et al.*, 2001). Cette approche, qui a été réutilisée plus récemment par (Duong *et al.*, 2013), correspond à l'état de l'art des annotateurs morpho-syntaxiques non supervisés. Ces auteurs utilisent l'alignement automatique en mots (obtenu à partir d'un corpus parallèle) pour projeter les annotations de la langue source vers la langue cible, en vue de construire des annotateurs morpho-syntaxiques pour la langue cible.

L'algorithme 1 est décrit dans l'encadré ci-dessous :

Algorithme 1 : Méthode de référence par projection d'annotations selon un alignement automatique en mots

- 1 : Annoter le côté source du corpus parallèle.
 - 2 : Aligner automatiquement le corpus parallèle en utilisant GIZA++ ou un autre outil d'alignement en mots.
 - 3 : Projeter les annotations directement pour les alignements 1-1.
 - 4 : Pour les correspondances N-1, projeter l'annotation du mot se trouvant à la position $N/2$ arrondi à l'entier supérieur.
 - 5 : Annoter les mots non-alignés avec l'étiquette la plus fréquente qui leur est associée dans le corpus.
 - 6 : Apprendre un analyseur morpho-syntaxique à partir de la partie cible du corpus désormais annotée (par exemple, dans notre cas, nous utilisons TNT tagger (Brants, 2000)).
-

3.2 Annotateur morpho-syntaxique non supervisé fondé sur les réseaux de neurones récurrents

Les réseaux de neurones sont généralement classés dans deux grandes catégories : les réseaux de neurones *Feed-forward* (Bengio *et al.*, 2006) et les réseaux de neurones Récurrents (on utilise l'acronyme RNN en anglais - pour *Recurrent Neural Networks*) (Mikolov *et al.*, 2010). (Sundermeyer *et al.*, 2013) ont montré que les modèles de langue statistiques basés sur une architecture récurrente présentent de meilleures performances que les modèles basés sur une architecture *Feed-forward*. Cela vient du fait que les réseaux de neurones récurrents utilisent un contexte de taille non limitée, contrairement aux réseaux *Feed-forward* dont la topologie limite la taille du contexte pris en compte. Cette propriété a motivé notre choix d'utiliser, dans nos expériences, un réseau de neurones de type récurrent (Elman, 1990).

Dans cette section, nous décrivons en détail l'approche proposée pour la construction d'un étiqueteur morpho-syntaxique multilingue basé sur les RNNs. L'approche, qui ne nécessite aucune ressource externe, requiert simplement un corpus parallèle et un annotateur morpho-syntaxique pré-existant dans la langue source.

3.2.1 Description du modèle

Un RNN est au minimum composé d'une succession de trois couches de neurones : une couche d'entrée au temps t notée $x(t)$, une couche cachée $h(t)$ (aussi appelée couche de contexte), et une couche de sortie $y(t)$. Chaque neurone de la couche d'entrée est relié à tous les neurones de la couche cachée par les matrices des poids U et W . La matrice des poids V connecte tout neurone de la couche cachée à chaque neurone de la couche de sortie, cf. (Figure 1).

Dans notre modèle, la couche d'entrée est formée par la concaténation de la représentation vectorielle $w(t)$ du mot courant, et de la couche cachée au temps précédent $h(t-1)$ (information de la première des couches cachées, dans le cas où on utilise plusieurs). La première étape de notre modèle est donc d'associer à chaque mot w (appartenant aux vocabulaires des langues source et cible) une représentation vectorielle spécifique.

Notre idée est la suivante : si on arrive à construire un espace de représentation commun, où un mot source et sa traduction cible possèdent des représentations vectorielles proches, nous pourrons — à partir de cette représentation commune —

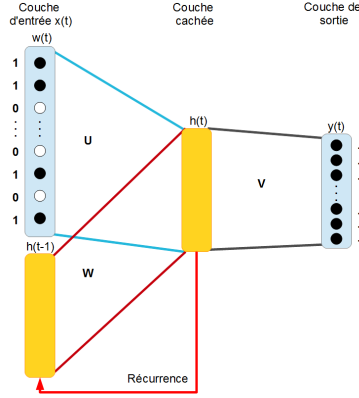


FIGURE 1 – Exemple de réseau de neurones récurrent

utiliser l’annotateur morpho-syntaxique de type RNN (appris initialement sur le coté source) pour annoter un texte en langue cible.

En général, un mot source et sa traduction cible apparaissent le plus souvent ensemble dans les mêmes bi-phrases, et donc leurs empreintes distributionnelles sont proches. Nous faisons le choix de construire notre espace de représentation commun, en associant à chaque mot (source, cible) son empreinte distributionnelle V_w de dimension N (nombre de bi-phrases dans le corpus parallèle) indiquant si le mot apparaît ou pas dans chaque bi-phrase $Phi \{i = 1, \dots, N\}$ du corpus parallèle :

$$V_w = \begin{cases} V_{wi} = 1 & \text{si } w \in Phi_i \\ V_{wi} = 0 & \text{sinon} \end{cases} \quad (1)$$

Par ailleurs, nous utilisons deux couches cachées (des expériences préliminaires ont montré que ceci permet d’obtenir de meilleures performances), avec des tailles variables (de 80 à 1024 neurones). Pour la fonction d’activation, nous utilisons la fonction *sigmoïde*. Nous pensons que ces couches cachées devraient permettre de capturer intrinsèquement des informations d’alignement au niveau des mots.

De plus, pour pouvoir transférer les annotations morpho-syntaxiques d’une langue à une autre, il est nécessaire que ces annotations soient décrites de la même manière dans les deux langues. (Petrov *et al.*, 2012) définissent un ensemble de 12 étiquettes morpho-syntaxiques universelles à gros grain, communes au plus grand nombre de langues (*Universal Tagset*). Ces étiquettes universelles sont les suivantes : NOUN (noms), VERB (verbes), ADJ (adjectifs), ADV (adverbes), PRON (pronoms), DET (déterminants et articles), ADP (prépositions et postpositions), NUM (numéraux), CONJ (conjonctions), PRT (particules), « . » (symboles de ponctuations) et X (pour tout ce qui échappe aux autres catégories). Dans nos travaux, nous adoptons ces étiquettes morpho-syntaxiques universelles. Par conséquent, la couche de sortie de notre modèle comporte 12 neurones, chaque neurone correspondant à une étiquette morpho-syntaxique universelle. On utilise la fonction d’activation *softmax* sur la couche de sortie afin d’obtenir des scores assimilables à des probabilités, le mot w en entrée du réseau est annoté par l’étiquette la plus probable en sortie du réseau.

3.2.2 Construction du modèle

Tout d’abord, avant l’apprentissage du modèle, quelques étapes de pré-traitement sont nécessaires. Celles-ci sont appliquées sur notre corpus d’apprentissage (corpus parallèle source / cible) et / ou sur notre corpus de validation en langue source :

- Étiqueter le côté source du corpus parallèle et le corpus de validation (avec l’étiqueteur supervisé disponible) :
- Construire les représentations vectorielles communes (empreintes distributionnelles) des mots source et cible, à partir du corpus parallèle initial ¹.

Ensuite, le réseau de neurones est entraîné sur plusieurs itérations (épques). L’algorithme 2 présenté ci-dessous décrit une époque d’entraînement du réseau.

1. Il est important de noter que si ce corpus parallèle change - par exemple si de nouvelles données sont disponibles - les représentations vectorielles pourront être soit conservées à l’identique soit mises à jour (en augmentant la taille du vecteur) avant le ré-apprentissage du RNN.

Algorithme 2 : Apprentissage d'un analyseur morpho-syntaxique multilingue basé sur un RNN

- 1 : Initialiser les matrices des poids du réseau avec une distribution normale.
 - 2 : Initialiser le compteur du temps $t=0$, et initialiser l'état des neurones de la couche cachée $h(t)$ à 1.
 - 3 : Incrémenter le compteur du temps t de 1.
 - 4 : Présenter le vecteur représentant le mot $w(t)$ dans la couche d'entrée.
 - 5 : Recopier l'état de la couche cachée $h(t-1)$ dans la couche d'entrée.
 - 6 : Calculer la valeur de la couche cachée $h(t)$ et de la couche de sortie $y(t)$.
 - 7 : Calculer l'erreur de prédiction $e_0(t) = d(t) - y(t)$ (différence entre la sortie prédite et la sortie attendue).
 - 8 : Mettre à jour les matrices des poids V et U avec l'algorithme de rétropropagation (RP) du gradient de l'erreur (Rumelhart *et al.*, 1985).
 - 9 : Mettre à jour la matrice des poids de récurrence W avec l'algorithme de la rétro-propagation du gradient de l'erreur à travers le temps (RPTT) (Rumelhart *et al.*, 1985).
 - 10 : Si le corpus d'apprentissage comporte encore des exemples, alors revenir à 3.
-

Les matrices des poids du réseau sont mises à jour en utilisant l'erreur de prédiction pondérée par un pas d'apprentissage α , initialement fixé à 0.1.

Après chaque époque, le corpus de validation est annoté en utilisant le réseau de neurones appris jusque-là. Les sorties sont comparées aux sorties de l'annotateur supervisé, pour calculer le taux d'erreur du réseau. Si le taux d'erreur diminue d'une époque à une autre, le pas d'apprentissage reste inchangé et l'apprentissage continue durant une nouvelle époque. Sinon, le pas d'apprentissage est diminué de moitié au début de la nouvelle époque. Pour éviter un sur-apprentissage des poids du réseau, l'algorithme d'apprentissage est arrêté si le taux d'erreur ne diminue plus durant deux époques successives. Généralement le réseau converge en 5 à 10 époques.

La deuxième étape de notre approche consiste simplement à utiliser le modèle entraîné sur le côté source comme annotateur morpho-syntaxique pour la langue cible, via l'utilisation de la représentation vectorielle commune. Il est important de noter que si l'on dispose d'un corpus parallèle en N langues (au lieu de 2), un même réseau RNN pourra étiqueter toutes ces langues sans être re-entraîné. On dispose donc d'un véritable étiqueteur multilingue.

4 Expérimentations et Résultats

4.1 Corpus et outils

Initialement, nous avons expérimenté notre approche sur le couple de langues anglais-français, où le français est considéré comme langue cible. Le français n'est certainement pas une langue faiblement dotée, mais le fait qu'il dispose d'un annotateur morpho-syntaxique supervisé (TreeTagger (Schmid, 1995)), nous a permis de construire une *pseudo vérité terrain* (sur le corpus de test) pour évaluer notre approche. Nous avons utilisé un corpus d'apprentissage de 10000 bi-phrases, extrait du corpus parallèle (anglais-français) ARCADEII (Véronis *et al.*, 2008), dont le côté source a été annoté par l'outil *TreeTagger* (Schmid, 1995) pour l'anglais. Notre corpus de validation (en anglais - pour le réglage du RNN) contient 1000 phrases (non présentes dans le corpus d'apprentissage), et est aussi extrait du corpus ARCADEII puis annoté par le toolkit *TreeTagger* pour l'anglais. Nous avons construit notre corpus test (français) à partir de 1000 phrases extraites du corpus ARCADEII, et annoté par le toolkit *TreeTagger* pour le français, puis corrigées manuellement.

Ayant obtenu des résultats intéressants sur le couple de langues anglais-français, nous nous sommes ensuite intéressés à la généralisation de notre approche sur d'autres langues : l'allemand, le grec et l'espagnol. Afin de pouvoir rendre nos résultats comparables avec ceux de (Das & Petrov, 2011) et (Duong *et al.*, 2013), nous suivons leur protocole : nous partons de l'anglais comme langue source et utilisons un corpus parallèle et un corpus de validation (anglais) extraits d'Europarl (Koehn, 2005). Nous évaluons les résultats de nos approches sur les mêmes corpus de test, qui sont ceux des campagnes d'évaluation d'analyse en dépendances CoNLL (Buchholz & Marsi, 2006). Ces corpus ont été annotés manuellement par des experts linguistes. Nous utilisons aussi la même métrique d'évaluation (le taux d'erreur d'étiquetage) et le même jeu d'étiquettes (*Universal Tagset* (Petrov *et al.*, 2012)).

Afin de construire nos modèles par projection simple (Algorithme 1), la partie cible des corpus d'apprentissage est étiquetée par projection des annotations du côté source (annoté par le toolkit *TreeTagger* pour l'anglais) en utilisant les alignements obtenus par GIZA++. Par souci d'uniformité, nous avons aussi transformé les étiquettes morpho-syntaxiques

finies (de TreeTagger et de CoNLL) en leurs équivalents dans le jeu étiquettes universelles via les règles de (Petrov *et al.*, 2012).

Pour implémenter notre approche (décrite dans l’Algorithme 2), nous avons adapté l’outil *Recurrent Neural Network Language Modeling Toolkit* (RNNLM) fourni par (Mikolov *et al.*, 2011), pour apprendre et tester notre annotateur morpho-syntaxique neuronal ².

Pour tirer parti des avantages de chacun de ces deux modèles *M1* (Projection Simple) et *M2* (RNN), il est intéressant d’étudier un moyen de les combiner. Le mot w est annoté avec l’étiquette t_w la plus probable, en utilisant la fonction f donnée par l’équation ci-dessous :

$$f(w) = \arg \max_t (\mu P_{M1}(t|w, C_{M1}) + (1 - \mu) P_{M2}(t|w, C_{M2})) \quad (2)$$

Où, C_{M1} et C_{M2} sont, respectivement les contextes de w considérés par *M1* et *M2*. Le paramètre d’interpolation μ (importance de chaque modèle) est ajusté par validation croisée sur le corpus de test.

4.2 Résultats et discussion

Les résultats obtenus par notre approche sont résumés dans le tableau 1. Les scores obtenus par (Das & Petrov, 2011) et (Duong *et al.*, 2013) sont également inclus lorsque ceux-ci sont disponibles sur le même corpus de test

Modèle	français		allemand		grec		espagnol	
	Tous mots	OOV	Tous mots	OOV	Tous mots	OOV	Tous mots	OOV
Projection Simple	80.3 %	77.1 %	78.9%	73%	77.5%	72.8%	80%	79.7%
RNN-640-160	78.5 %	70 %	76.1%	76.4%	75.7%	70.7%	78.8%	72.6%
Projection+RNN	84.5%	78.8%	81.5 %	77%	78.3%	74.6%	83.6%	81.2%
(Das, 2011)	na	na	82.8%	na	82.5%	na	84.2%	na
(Duong, 2013)	na	na	85.4%	na	80.4%	na	83.3%	na

TABLE 1 – Performances en taux d’erreur d’étiquetage (Projection Simple, RNN et Projection+RNN) - et comparaison avec Das & Petrov (2011) et Duong *et al* (2013).

Nous avons évalué plusieurs topologies de réseaux de neurones récurrents, avec une ou deux couches cachées, et avec différentes tailles. Les meilleures performances obtenues sont celles des annotateurs basés sur des réseaux de neurones à deux couches cachées, contenant respectivement 640 et 160 neurones (RNN-640-160). Ces performances sont proches des annotateurs par projection simple, les différences proviennent de la gestion des mots inconnus (OOV) qui est pour l’instant quasi-inexistante dans notre approche. En effet, les représentations vectorielles des OOV sont nulles, et pour les annoter, le réseau de neurones n’utilise que l’information de récurrence (étiquette précédente prédite) ce qui est une information insuffisante pour une bonne annotation. Une perspective à très court terme consistera à traiter le cas des mots inconnus dans notre RNN.

En attendant, nous avons également combiné l’approche classique avec notre méthode par réseaux récurrents. Dans l’ensemble, les résultats expérimentaux de notre combinaison (Projection+RNN) sont proches de ceux de l’état de l’art des annotateurs morpho-syntaxiques non supervisés (Das & Petrov, 2011; Duong *et al.*, 2013) et montre une bonne complémentarité entre projection simple et RNN. Toutefois, nos résultats sont légèrement inférieurs aux résultats de référence, et une meilleure gestion des OOV, ainsi que l’utilisation d’une faible quantité de données cible annotée pour *adapter* le réseau, semblent des perspectives intéressantes à court terme.

5 Conclusion

Dans cet article, nous avons présenté une approche utilisant les réseaux de neurones récurrents comme annotateurs morpho-syntaxiques multilingues (non supervisés pour les langues cibles). Cette approche n’a besoin que d’un corpus parallèle et d’un annotateur morpho-syntaxique pré-existant en langue source. Bien que nos résultats initiaux soient positifs, ils doivent être améliorés. Dans nos futurs travaux, nous envisageons donc d’utiliser une meilleure représentation pour les OOV. Par ailleurs, nous envisageons d’utiliser une technique similaire pour des tâches plus complexes du TALN (par exemple annotation en sens, en entités nommées et en rôles sémantiques).

2. L’adaptation du RNNLM est disponible à l’url https://github.com/othman-zennaki/RNN_POS_Tagger.git

Références

- ABDULHAY A. (2012). *Constitution d'une ressource sémantique arabe à partir d'un corpus multilingue aligné*. PhD thesis, Université de Grenoble.
- ALLAUZEN A. & WISNIEWSKI G. (2009). Modèles discriminants pour l'alignement mot à mot. *Traitement Automatique des Langues*, **50**(3), 173–203.
- BENGIO Y., SCHWENK H., SENÉCAL J.-S., MORIN F. & GAUVAIN J.-L. (2006). Neural probabilistic language models. In *Innovations in Machine Learning*, p. 137–186. Springer.
- BENTIVOGLI L., FORNER P. & PIANTA E. (2004). Evaluating cross-language annotation transfer in the multisetcor corpus. In *Proceedings of the 20th international conference on Computational Linguistics*, p. 364 : Association for Computational Linguistics.
- BRANTS T. (2000). Tnt : a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, p. 224–231 : Association for Computational Linguistics.
- BUCHHOLZ S. & MARSI E. (2006). Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, p. 149–164 : Association for Computational Linguistics.
- DAS D. & PETROV S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, p. 600–609 : Association for Computational Linguistics.
- DEMPSTER A. P., LAIRD N. M. & RUBIN D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, p. 1–38.
- DUONG L., COOK P., BIRD S. & PECINA P. (2013). Simpler unsupervised pos tagging with bilingual projections. In *ACL (2)*, p. 634–639.
- ELMAN J. L. (1990). Finding structure in time. *Cognitive science*, **14**(2), 179–211.
- FRASER A. & MARCU D. (2007). Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, **33**(3), 293–303.
- GARSDIE R., LEECH G. N. & MCENERY T. (1997). *Corpus annotation : linguistic information from computer text corpora*. Taylor & Francis.
- JABAIAN B., BESACIER L. & LEFEVRE F. (2013). Comparison and combination of lightly supervised approaches for language portability of a spoken language understanding system. *IEEE Transactions on Audio, Speech & Language Processing*, **21**(3), 636–648. (Impact-F 1.67 estim. in 2012).
- KOEHN P. (2005). Europarl : A parallel corpus for statistical machine translation. In *MT summit*, volume 5, p. 79–86.
- MIKOLOV T., KARAFIÁT M., BURGET L., CERNOCKÝ J. & KHUDANPUR S. (2010). Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, p. 1045–1048.
- MIKOLOV T., KOMBRINK S., DEORAS A., BURGET L. & CERNOCKÝ J. (2011). Rnnlm-recurrent neural network language modeling toolkit. In *Proc. of the 2011 ASRU Workshop*, p. 196–201.
- OCH F. J. & NEY H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, p. 440–447 : Association for Computational Linguistics.
- PADÓ S. & LAPATA M. (2005). Cross-linguistic projection of role-semantic information. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, p. 859–866 : Association for Computational Linguistics.
- PADÓ S. & LAPATA M. (2006). Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, p. 1161–1168 : Association for Computational Linguistics.
- PADO S. & PITEL G. (2007). Annotation précise du français en sémantique de rôles par projection cross-linguistique. *Proceedings of TALN-07, Toulouse, France*.
- PETROV S., DAS D. & McDONALD R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.
- RUMELHART D. E., HINTON G. E. & WILLIAMS R. J. (1985). *Learning internal representations by error propagation*. Rapport interne, DTIC Document.

- SCHMID H. (1995). Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop* : Citeseer.
- SUNDERMEYER M., OPARIN I., GAUVAIN J.-L., FREIBERG B., SCHLUTER R. & NEY H. (2013). Comparison of feedforward and recurrent neural network language models. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 8430–8434 : IEEE.
- TÄCKSTRÖM O., DAS D., PETROV S., McDONALD R. & NIVRE J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, **1**, 1–12.
- VAN DER PLAS L. & APIDIANAKI M. (2014). Cross-lingual word sense disambiguation for predicate labelling of french. *Proceedings of TALN-14, Marseille, France*, p.46.
- VÉRONIS J. (2000). *Chapitre 4*, In *Annotation automatique de corpus : panorama et état de la technique*. Editions Hermès.
- VÉRONIS J., HAMON O., AYACHE C., BELMOUHOU B., KRAIF O., LAURENT D., NGUYEN T. M. H., SEMMAR N., STUCK F. & WAJDI Z. (2008). *Chapitre 2*, In *ArcadeII Action de recherche concertée sur l'alignement de documents et son évaluation*. Editions Hermès.
- WISNIEWSKI G., PÉCHEUX N., KNYAZEVA E., ALLAUZEN A. & YVON F. (2014). Apprentissage partiellement supervisé d'un étiqueteur morpho-syntaxique par transfert cross-lingue. *Proceedings of TALN-14, Marseille, France*, p. 173.
- YAROWSKY D., NGAI G. & WICENTOWSKI R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, p. 1–8 : Association for Computational Linguistics.

Segmentation et Titrage Automatique de Journaux Télévisés

Abdessalam Boucekif^{1,2} Géraldine Damnat¹ Nathalie Camelin² Delphine Charlet¹ Yannick Estève²

⁽¹⁾Orange Labs, 2 avenue Pierre Marzin 22300, Lannion, France.

⁽²⁾Laboratoire d'Informatique de l'Université du Maine, LIUM - France.

Résumé. Dans cet article, nous nous intéressons au titrage automatique des segments issus de la segmentation thématique de journaux télévisés. Nous proposons d'associer un segment à un article de presse écrite collecté le jour même de la diffusion du journal. La tâche consiste à apparier un segment à un article de presse à l'aide d'une mesure de similarité. Cette approche soulève plusieurs problèmes, comme la sélection des articles candidats, une bonne représentation du segment et des articles, le choix d'une mesure de similarité robuste aux imprécisions de la segmentation. Des expériences sont menées sur un corpus varié de journaux télévisés français collectés pendant une semaine, conjointement avec des articles aspirés à partir de la page d'accueil de Google Actualités. Nous introduisons une métrique d'évaluation reflétant la qualité de la segmentation, du titrage ainsi que la qualité conjointe de la segmentation et du titrage. L'approche donne de bonnes performances et se révèle robuste à la segmentation thématique.

Abstract.

Automatic Topic Segmentation and Title Assignment in TV Broadcast News

This paper addresses the task of assigning a title to topic segments automatically extracted from TV Broadcast News video recordings. We propose to associate to a topic segment the title of a newspaper article collected on the web at the same date. The task implies pairing newspaper articles and topic segments by maximising a given similarity measure. This approach raises several issues, such as the selection of candidate newspaper articles, the vectorial representation of both the segment and the articles, the choice of a suitable similarity measure, and the robustness to automatic segmentation errors. Experiments were made on various French TV Broadcast News shows recorded during one week, in conjunction with text articles collected through the Google News homepage at the same period. We introduce a full evaluation framework allowing to measure the quality of topic segment retrieval, topic title assignment and also joint retrieval and titling. The approach yields good titling performance and reveals to be robust to automatic segmentation.

Mots-clés : Segmentation thématique, Titrage automatique, Pondération Okapi, Mesures de similarité.

Keywords: Topic segmentation, Title assignation, Okapi weighting, Similarity measures.

1 Introduction

Avec les dernières innovations technologiques, les sites des chaînes TV proposent gratuitement à leurs téléspectateurs des services de TV de rattrapage (*Replay/catch up TV*) via Internet. Ce service donne à l'utilisateur la possibilité de voir les émissions des chaînes TV à travers les podcasts. Ces derniers intéressent aussi les fournisseurs d'actualités comme Google Actualités, Yahoo Actualités, Orange Actualités, *etc.* Ils collectent des informations de différentes sources (site web des journaux, radio, chaînes TV) en agrégeant les documents de thèmes similaires. Elle permet à l'utilisateur d'avoir une information présentée par différents média. Pour permettre cette exploitation, le document doit être préalablement segmenté en fragments thématiquement homogènes : *i.e.* parlant d'un seul sujet. Cette tâche est nommée *segmentation thématique*. Les segments retournés sont identifiés par des labels anonymes (*sujet₁, sujet₂, ...*).

Donner un titre à chaque thème abordé durant l'émission est un atout supplémentaire pour une meilleure diffusion des contenus. En effet, le titre est le meilleur moyen pour décrire les différentes parties de l'émission de façon rapide et précise. D'ailleurs, ce sont les premiers éléments qu'un utilisateur consulte en priorité. Cependant, un titre doit remplir certains critères : il doit être *correct, court, clair et compréhensible*, offrant un aperçu à la fois global et spécifique du contenu. Ces critères rendent la tâche de titrage automatique plus complexe pour la machine. Le contenu même de l'émission peut contenir des informations utiles à la tâche de titrage :

- *Visuel* : un changement de sujet peut être accompagné par un titre vidéo. L'OCR (Optical Character Recognition) permet de détecter et de reconnaître les textes d'une vidéo quand ils sont incrustés.

- *Acoustique* : certains événements sonores, comme les jingles, peuvent aider à la détection de l'énoncé des titres par le présentateur.

- *Linguistique* : les mots les plus pertinents peuvent jouer le rôle d'un titre. Ainsi, une partie du discours contenant le plus grand nombre de mots discriminants peut être aussi considérée comme un titre.

L'extraction des titres à partir du document même n'est pas toujours possible. En effet, les indicateurs visuels et acoustiques sont fortement liés aux choix éditoriaux des chaînes télévisées (le titrage incrusté n'est pas tout le temps disponible, le présentateur principal ne donne pas forcément au début du journal tous les sujets abordés, *etc.*). Le contenu linguistique de l'émission donne des titres moins informatifs (les mots clés peuvent fournir un titre ambigu et demandent un effort supplémentaire de la part de l'utilisateur). De plus, l'extraction d'une partie du discours est un problème compliqué qui dépend du système de reconnaissance automatique de la parole et de la segmentation du discours.

Dans la littérature, malgré l'importance du titrage, peu de travaux traitent des journaux télévisés. Dans (Hsueh & Moore, 2006), les auteurs cherchent à catégoriser des réunions pour cela ils définissent un ensemble fini de catégories. La tâche de titrage est alors vue comme un problème de classification multi-classes. Les auteurs de (Lau *et al.*, 2011) proposent une méthode dans laquelle les titres des articles *Wikipédia* sont des candidats potentiels. Chaque thème est représenté par une liste de *Top10* mots, laquelle est considérée comme une requête. Ainsi, le titre du document le plus proche de la requête est considéré comme un identifiant du thème.

Notre objectif est de développer un système de titrage générique indépendant de toute information structurelle *a priori* sur l'émission tout en respectant la définition d'un bon titre. Nous proposons une nouvelle approche de titrage des segments de journaux télévisés obtenus automatiquement, qui exploite des informations provenant de la presse écrite. Elle consiste à apparier un segment à un article de presse du même jour tout en décrivant le même thème, afin d'attribuer le titre de l'article à ce segment. Cette approche permet de donner à nos segments des titres rédigés par des journalistes professionnels qui d'un point de vue thématique sont proches des segments automatiques.

L'article est structuré comme suit : dans la section 2, on décrit brièvement notre algorithme de segmentation thématique. La section 3 est consacrée à notre approche du titrage. Enfin, les expériences et la conclusion sont présentées respectivement dans les sections 4 et 5.

2 Segmentation thématique

Plusieurs algorithmes de segmentation thématique (ST) ont été proposés dans la littérature. Parmi ces algorithmes, citons *textTiling* (Hearst, 1997), *C99* (Choi, 2000) et *MinCut* (Malioutov & Barzilay, 2006). Un état de l'art plus complet est donné dans (Boucekif *et al.*, 2014b) avec également plus de détails quant à notre approche de ST. Cette approche est inspirée du *TextTiling* mais exploite plus largement les propriétés des données orales traitées. Elle repose sur l'analyse de la distribution des mots entre deux fenêtres textuelles pour déterminer l'existence ou non d'une borne thématique entre deux fenêtres adjacentes. Nous considérons comme segments unitaires les groupes de souffle (GS) qui sont les paroles prononcées par un locuteur entre deux respirations (pauses silencieuses). Ces groupes sont rassemblés en blocs de K GS pour calculer la similarité entre chaque paire de blocs adjacents. À l'aide d'une fenêtre glissante, la similarité est donc calculée tout au long de l'émission entre des blocs adjacents de K GS de part et d'autre des frontières potentielles. Il en résulte ainsi une courbe de cohésion à partir de laquelle sont extraites des frontières.

(Boucekif *et al.*, 2014a) proposent deux façons de pondérer les mots de l'émission dont le principe général est de découper l'émission en N morceaux (ou chunks) de taille différente, où chaque chunk représente la notion d'un document en recherche d'information. Dans la première méthode, le début de chaque intervention du présentateur principal représente le début d'un nouveau chunk. Dans la deuxième méthode, les poids sont calculés de manière itérative. La segmentation obtenue pour une itération donnée fournit un ensemble de documents à partir desquels les poids seront ré-estimés dans la prochaine itération.

Par ailleurs, nous avons introduit dans (Boucekif *et al.*, 2014b) la notion de la cohésion de la parole regroupant la cohésion lexicale et locuteurs dans une seule notion. Une frontière potentielle est valide si la distribution conjointe des mots et des locuteurs diffère suffisamment de part et d'autre de la frontière.

3 Titrage automatique

Notre approche de titrage intervient après la phase de ST de journaux télévisés, elle consiste à associer à chaque segment obtenu le titre de l'article de presse du jour traitant du même sujet. Parmi tous les articles de presse disponibles, le titre associé au segment est celui de l'article le plus proche thématiquement. Le titrage, tel que présenté, peut donc être vu comme une tâche d'ordonnancement d'une liste de titres candidats, incluant les rejets (aucun candidat ne sera retenu).

Étape 1 : Aspirer des articles de presse

La première étape consiste à récolter tous les articles de presse parus sur la page d'accueil de Google Actualités le même jour que le segment à titrer. Les pages étant bruitées par des éléments non-informatifs, nous utilisons l'outil Boilerpipe¹ pour ne garder que le contenu utile de l'article. Toutes les informations concernant les articles sont sauvegardées dans une base de données. Chaque tuple (un article de presse) contient l'identifiant de l'article, le lien de la page web, le titre de l'article, sa date de publication, le contenu de la page et l'identifiant de l'article principal².

Étape 2 : Représentation vectorielle

Le calcul de la similarité entre segment et article de presse nécessite une représentation vectorielle de l'un et l'autre. Or le segment est issu de la transcription de la parole (contenant potentiellement des disfluences et des erreurs de reconnaissance) et l'article est composé de texte écrit dans un style journalistique. La représentation choisie doit être robuste à diverses sources et styles. Avec l'aide du logiciel *Lia-tag*, des prétraitements standards (lemmatisation, filtrage des mots) ont été appliqués sur les articles de presse et les segments. Ensuite, chaque article de presse et chaque segment est remplacé par une liste de mots pertinents selon la mesure *Okapi*. Enfin, une normalisation est réalisée relativement aux mots ayant le score le plus élevé, et un filtrage est appliqué : les mots ayant un score supérieur à +0.25 sont conservés.

Étape 3 : Calcul de similarité

Le but de cette étape est de calculer la similarité entre chaque couple (article, segment). Le choix de la mesure appropriée à la nature de nos documents est très important. Pour cela, nous proposons de comparer les mesures mentionnées dans le tableau 1 telles qu'elles sont définies dans (Curran & Moens, 2002). La mesure *Jaccard* est tout simplement le rapport entre le nombre de termes en commun et le nombre de tous les termes apparaissant dans l'article et le segment. La mesure *Lin* est une version pondérée de la mesure *Jaccard*.

Set_JACCARD	$\frac{ S \cap A }{ S \cup A }$	Extended_JACCARD	$\frac{\sum_{t \in S \cap A} w^S(t) \times w^A(t)}{\ W^S\ _2^2 + \ W^A\ _2^2 - \sum_{t \in S \cap A} w^S(t) \times w^A(t)}$
LIN	$\frac{\sum_{t \in S \cap A} w^S(t) + w^A(t)}{\sum_{t \in S \cup A} w^S(t) + w^A(t)}$	Cosine	$\frac{\sum_{t \in S \cap A} w^S(t) \times w^A(t)}{\ W^S\ _2 \times \ W^A\ _2}$

TABLE 1 – Mesures de similarité utilisées où S est la liste des termes (après filtrage) du segment, A est la liste des termes (après filtrage) de l'article considéré. On note par w_t^S (resp. w_t^A) le poids du terme t dans le segment S (resp. A)

Les valeurs de similarité sont exploitées non seulement pour classer les articles mais aussi pour filtrer les résultats. En effet, un sujet abordé dans un journal télévisé n'est pas forcément traité dans la presse. Par conséquent, le premier document retourné prend une valeur de similarité faible. L'application d'un seuil α permettra de régler ce genre de problème et de prendre uniquement les documents dont on est sûr qu'ils développent la même thématique.

4 Expériences

4.1 Corpus et annotation

Les expériences sont menées sur un corpus constitué de 86 journaux télévisés collectés durant la période du 10 au 16 février 2014 en provenance de 8 chaînes françaises (TF1, France2, france3, M6, Arte, D8, NT1, Euronews). Les émissions ont été transcrites par le biais du système Vocabia (Gauvain *et al.*, 2002) qui obtient 16,1% de taux d'erreurs (WER) sur un corpus équivalent. Dans la même période et avec un rythme d'une fois par heure, nous avons collecté à travers le site web Google Actualités une base de données de 22000 articles. Nous conservons uniquement les articles principaux³ et supprimons les doublons⁴ réduisant le volume à 4600 articles. Pour simplifier la tâche d'annotation nous exploitons seulement les articles principaux, ce qui donne en moyenne 660 articles par jour.

Le grand nombre d'articles de la collection et de segments de JT rend la tâche d'annotation manuelle très longue et fastidieuse. En effet, il faut que l'annotateur évalue la potentielle association thématique entre chaque segment et les articles de la collection. Afin de réduire le nombre d'association à évaluer, nous proposons à l'annotateur pour chaque segment, uniquement l'ensemble des articles de presse du même jour ayant au moins 2 mots en commun avec le segment considéré. Ainsi, l'annotateur a vérifié en moyenne 127 titres par segment (et non les 660 pour chaque segment).

1. <https://code.google.com/p/boilerpipe/>

2. Google Actualités regroupe les articles qui traitent du même sujet, chaque thème est représenté par un article considéré comme le principal.

3. Pour chaque article principal, plusieurs articles portant sur le même sujet sont proposés par Google.

4. Le même article peut être téléchargé plusieurs fois dans la journée.

La tâche d'annotation consiste à indiquer pour chaque article proposé : (i) si le titre de l'article reflète bien le contenu du segment, (ii) s'il résume partiellement le segment (ne couvre pas la totalité du segment, ne suit pas l'actualité, *etc.*), ou (iii) si le titre n'a pas de relation thématique avec le segment. Dans ce travail, uniquement les titres résumant parfaitement le contenu du segment sont considérés comme corrects et les autres comme des titres incorrects. Chaque segment de référence R est associé à un ensemble de titres candidats. Cet ensemble peut être vide si le thème n'a pas été traité dans les articles de la collection, le segment est alors considéré comme un segment *non titrable* (\bar{T}). Dans le cas contraire, il est considéré comme *titrable* (T). Au final, le corpus se compose de 658 segments de type T et 339 segments de type \bar{T} . Le

	Nb	Dur. Moy	T	\bar{T}
Seg. longs	761	131,4	467	294
Seg. courts	236	20,4	191	45
Tous les Seg.	997	105,1	658	339

TABLE 2 – Répartition des segments titrables et non titrables par rapport à la durée des segments (courts et longs)

tableau 2 représente la répartition des segments titrables et non titrables par rapport à la durée des segments. Deux types de segments sont considérés : *long* si la durée est supérieure à 30s, et *court* sinon. Le corpus contient 66,0% de segments titrables et 34,0% non titrables respectivement. Cela s'explique par le fait que les JT traitent non seulement de l'actualité du jour mais aussi de sujets de société qui ne se retrouvent pas nécessairement dans les articles de presse du jour.

4.2 Métriques d'évaluation

4.2.1 Métrique d'évaluation de la segmentation thématique

La qualité de la ST est généralement évaluée en comparant le positionnement des ruptures thématiques de référence avec celles à évaluer. Les mesures les plus utilisées telles F -mesure, p_k (Beeferman *et al.*, 1999) et $windowdiff$ (Pevzner & Hearst, 2002), donnent une valeur numérique reflétant la performance du système sur la totalité de l'émission. Dans (Bouche kif *et al.*, 2014b), une métrique a été proposée pour estimer la performance du système de ST pour chaque segment à détecter. En suivant cette proposition, nous considérons qu'un segment est correct si ses instants de début et fin sont proches de ceux de référence. Pour cela, nous cherchons pour chaque segment de référence R le segment d'hypothèse H correspondant le mieux, c'est à dire le segment ayant une couverture temporelle maximale au segment de référence. La couverture entre deux segments donnés R et H , notée $Couv_{R \leftrightarrow H}$, est définie comme la moyenne harmonique de $Couv_{R \rightarrow H}$ et $Couv_{H \rightarrow R}$ (où $Couv_{R \rightarrow H}$ est le taux de couverture du segment de référence par le segment hypothèse et $Couv_{H \rightarrow R}$ est le taux de couverture du segment hypothèse par le segment de référence). Nous considérons qu'un segment H est correct si $Couv_{R \leftrightarrow H}$ est supérieur à un seuil γ , où

$$Couv_{R \leftrightarrow H} = \frac{2 \times Cov_{R \rightarrow H} \times Cov_{H \rightarrow R}}{Cov_{R \rightarrow H} + Cov_{H \rightarrow R}}. \quad (1)$$

Il ressort de l'exemple de la figure 1 (pour $\gamma = 85\%$) que seul le segment H_1 est correct.

Soit $\#R$ le nombre de segments de référence et $\#H_{Err\gamma}$ le nombre de segments hypothèse dont la couverture harmonique

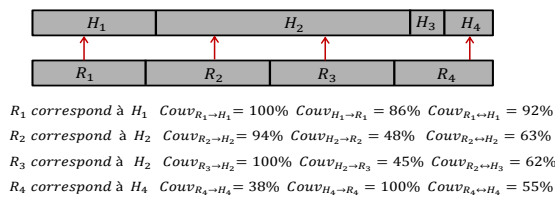


FIGURE 1 – Exemple d'évaluation de la segmentation thématique par *nombre de segments corrects*.

avec son segment de référence ne dépasse pas le seuil γ . Nous définissons SER_γ le taux de segments incorrects tel que

$$SER_\gamma = \frac{\#H_{Err\gamma}}{\#R}. \quad (2)$$

4.2.2 Métrique d'évaluation du titrage

Le système de titrage mis en place peut être évalué selon les erreurs possibles :

- Substitution (*Sub*) : le segment est de type T et le titre affecté au segment n'est pas correct.
- Faux rejet (*FR*) : le segment est de type T et le système ne propose aucun titre.
- Fausse alarme (*FA*) : le segment est de type \bar{T} et le système propose un titre.

Les réponses correctes sont de deux types : (i) le segment est titrable et le titre associé est correct, (ii) le segment est non titrable et aucun titre n'y est associé. TER (Titling Error Rate) est donnée par :

$$TER = \frac{\#Sub + \#FR + \#FA}{\#R} \quad (3)$$

Un titre affecté à un segment est considéré comme bon *si et seulement si* la *ST* est jugée comme correcte ($Couv_{R \rightarrow H} > \gamma$) et le titre affecté figure dans la liste des titres candidats. Pour une évaluation complète, il suffit d'ajouter $H_{Err\gamma}$ comme une source d'erreur.

$$STER_{\gamma} = \frac{\#H_{Err\gamma} + \#Sub_{\gamma} + \#FR_{\gamma} + \#FA_{\gamma}}{\#R} \quad (4)$$

4.3 Expériences et résultats

4.3.1 Évaluation de la segmentation thématique

En terme de nombre de frontières détectées, notre système de ST obtient une F-mesure (de façon similaire à plusieurs travaux dans l'état de l'art, une tolérance de 10s est autorisée entre les frontières d'hypothèses et de références), égale à 71,6% avec un rappel de 73,2% et une précision de 70,0%. Comme il a été mentionné précédemment, nous nous intéressons à l'évaluation en terme de segments corrects. Le tableau 3 résume les résultats obtenus en faisant varier le seuil γ . Nous remarquons une différence de performance entre les segments longs et courts. En effet, contrairement aux segments courts, le système est nettement meilleur sur les segments longs. Ces derniers sont plus importants à extraire que les segments courts (qui ne durent que quelques secondes et correspondent généralement à des brèves).

SEr_{γ} (%)	$\gamma = 80\%$	$\gamma = 85\%$	$\gamma = 90\%$
Seg. longs	22,9	26,3	35,1
Seg. courts	72,0	76,7	81,6
Tous les seg.	34,5	38,2	46,0

TABLE 3 – Performance du système en terme de nombre de segments détectés

4.3.2 Évaluation du titrage sur les segments de référence

Dans un premier temps, nous évaluons le titrage sur les segments de référence (définis manuellement) avec différentes mesures de similarité. La performance du système est présentée sous forme d'une courbe ROC (substitution et faux rejet en fonction de fausse alarme) dans la figure 2. Les résultats ont été obtenus en faisant varier le seuil α appliqué sur les valeurs de similarité. La première observation est que les mesures pondérées donnent de meilleurs résultats que le *Jaccard* standard. En effet, la pondération donne plus de poids aux mots importants et pénalise les moins représentatifs, ceci a une influence directe sur le calcul de la cohésion et donc sur le classement des titres candidats. La mesure *Cosine* donne les meilleurs résultats ($Ter = 13,5\%$) avec une légère différence par rapport à *Extended_Jaccard* ($Ter = 13,9\%$). Par la suite, les performances sont données pour la mesure *Cosine*.

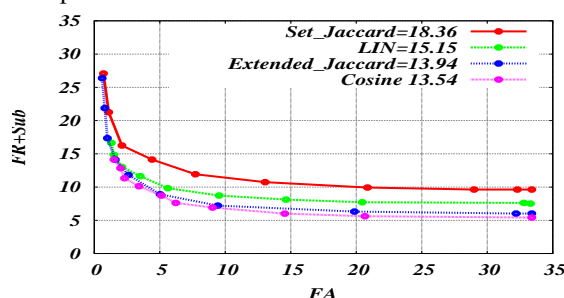


FIGURE 2 – TER pour les différentes variantes de similarité

Le tableau 4 donne la performance du système pour les segments courts et longs. Dans sa globalité, le système donne de très bonnes performances. En effet, entre le titrage et le rejet correct la qualité du titrage est de l'ordre 86,5%. Notre système a la capacité de bien titrer non seulement les segments longs (représentés par un nombre important de termes) mais aussi les segments courts.

Cosine	TER	Sub	FA	FR
Seg. longs	12.6	3.3	2.6	7.6
Seg. courts	16,5	5,1	5,9	5,5
Tous les seg.	13.5	3,7	3,4	6,4

TABLE 4 – Performance du système de titrage sur les segments de référence avec la similarité *Cosine*

Une analyse des erreurs a mis en évidence deux catégories principales d'erreurs : (1) celles qui proviennent de fils d'actualités dont le contenu évolue à court terme, (2) celles qui proviennent d'informations traitant de thématiques proches mais légèrement différentes. Le tableau 5 fournit quelques exemples d'erreurs.

	Titre Correct	Titre erroné	Commentaires
Ex1	Sotchi : la France attend sa première médaille	JO : Martin Fourcade, premier médaillé d'or français.	Dépend de l'heure à laquelle l'information est donnée. Ce genre d'informations peut changer plusieurs fois dans la journée.
Ex2	Le perchiste Renaud Lavillenie accueilli en héros à Roissy.	Saut à la perche : Lavillenie incertain pour les Mondiaux en salle.	
Ex3	Crues : trois départements en vigilance orange.	Tempête : Quelque 10.000 foyers toujours privés d'électricité en Bretagne	Des informations traitent des thématiques proches

TABLE 5 – Exemples d'erreurs de titrage

4.4 Évaluation du titrage sur les segments automatiques

Les résultats du tableau 6 résument le comportement de notre système de titrage sur des segments automatiques. De l'analyse des résultats, il découle que les erreurs de segmentation sont prépondérantes. Plus la qualité de segmentation est bonne plus les erreurs de titrage (FA , Sub , FR) diminuent et donc le système aura la capacité d'affecter de bons titres aux segments. En effet, une faible valeur de couverture signifie que le segment d'hypothèse couvre partiellement le segment de référence (voir le segment H_4 de la figure 1). Par conséquent, la liste des mots décrivant le contenu du segment est incomplète. Dans le cas où le segment couvre partiellement et/ou totalement deux segments de référence (voir le segment H_1 de la figure 1), la liste des mots décrivant le contenu du segment est bruitée.

À noter que les segments courts donnent de mauvaises performances par rapport aux segments longs. En effet, avec une couverture de 85% ($Couv_{R \leftrightarrow H} > 85$), le $STER_{85}$ global est de 46,8%, le système arrive à segmenter et à titrer correctement 63,6% des segments longs et seulement 19.5% des segments courts.

Pour comparer la robustesse du titrage sur des segments générés manuellement et automatiquement, nous évaluons le titrage sur le même ensemble, pour cela nous écartons les segments mal déterminés ($Couv_{R \leftrightarrow H} < 85$). Avec les 615 segments corrects, le taux de titres corrects est respectivement de 87% et 86% pour les segments manuels et automatiques. Ces résultats illustrent bien la robustesse de notre système sur les segments thématiques obtenus automatiquement.

γ	$STER_\gamma$	$H_{err\gamma}$	Sub_γ	FA_γ	FR_γ
80	43,9	34,5	2,4	2,3	4,7
85	46,8	38,2	2,2	2,0	4,4
90	53,3	46,0	1,6	1,8	3,9

TABLE 6 – Taux d'erreur prenant en compte à la fois la segmentation et le titrage.

5 Conclusion

Dans cet article, nous avons décrit notre système de structuration thématique composé de deux tâches complémentaires : la segmentation thématique et le titrage automatique des segments. Après l'étape de ST appliquée à des sorties d'un système de reconnaissance de la parole, les segments obtenus sont titrés. Le titrage consiste à apparier le segment à un article de presse traitant le même sujet. Le titre associé au segment est celui de l'article qui maximise la similarité entre le segment et les articles du jour. Une métrique d'évaluation mesurant conjointement la qualité de la segmentation et du titrage a été proposée. Les résultats obtenus montrent que les erreurs de segmentation restent prépondérantes dans le processus. Le titrage donne de bons résultats et est robuste aux petites imprécisions de la segmentation. Comme perspective à ce travail, il est envisagé d'étudier l'interaction entre ces deux tâches afin d'améliorer la qualité du système de segmentation à partir du titrage, notamment pour les segments courts.

Références

- BEEFERMAN D., BERGER A. L. & LAFFERTY J. D. (1999). Statistical models for text segmentation. *Machine Learning*, **34**(1-3), 177–210.
- BOUCHEKIF A., DAMNATI G. & CHARLET D. (2014a). Intra-content term weighting for topic segmentation. In *39th IEEE International Conference on Acoustics, Speech and Signal Processing, Florence*, p. 7113–7117.
- BOUCHEKIF A., DAMNATI G. & CHARLET D. (2014b). Speech cohesion for topic segmentation of spoken contents. In *INTERSPEECH, Singapore*, p. 1890–1894.
- CHOI F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference, NAACL 2000*, p. 26–33.
- CURRAN J. R. & MOENS M. (2002). Improvements in automatic thesaurus extraction. In *Proceedings of the workshop on Workshop On Unsupervised Lexical Acquisition, Philadelphia*, p. 59–66.
- GAUVAIN J.-L., LAMEL L. & ADDA G. (2002). The LIMSI broadcast news transcription system. *Speech Communication*, **37**(1), 89–108.
- HEARST M. A. (1997). Texttiling : Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, **23**(1), 33–64.
- HSUEH P.-Y. & MOORE J. D. (2006). Automatic topic segmentation and labeling in multiparty dialogue. In *Spoken Language Technology Workshop, Aruba*, p. 98–101.
- LAU J. H., GRIESER K., NEWMAN D. & BALDWIN T. (2011). Automatic labelling of topic models. In *Proceedings of the 49th ACL : Human Language Technologies*, p. 1536–1545.
- MALIOUTOV I. & BARZILAY R. (2006). Minimum cut model for spoken lecture segmentation. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney*.
- PEVZNER L. & HEARST M. (2002). A critique and improvement of an evaluation metric for text segmentation. *Comput. Linguist.*, **28**(1), 19–36.

Un système hybride pour l'analyse de sentiments associés aux aspects

Caroline Brun¹ Diana Nicoleta Popa¹ Claude Roux¹
(1) XRCE, 6 chemin de Maupertuis, 38240 Meylan France
{ Caroline.Brun,Diana.Popa,Claude.Roux }@xrce.xerox.com

Résumé. Cet article présente en détails notre participation à la tâche 4 de SemEval2014 (Analyse de Sentiments associés aux Aspects). Nous présentons la tâche et décrivons précisément notre système qui consiste en une combinaison de composants linguistiques et de modules de classification. Nous exposons ensuite les résultats de son évaluation, ainsi que les résultats des meilleurs systèmes. Nous concluons par la présentation de quelques nouvelles expériences réalisées en vue de l'amélioration de ce système.

Abstract.

An Hybrid System for Aspect-Based Sentiment Analysis

This paper details our participation to the SemEval2014 task 4 (Aspect Based Sentiment Analysis). We present the shared task, and then describe precisely our system, which is a combination of natural processing components and classification modules. We also present its evaluation results and the best system results. We finally expose some new experiments aiming at improving the system.

Mots-clés : Analyse de sentiments associés aux aspects, SemEval2014, système hybride.

Keywords: Aspect Based Sentiment Analysis, SemEval2014, hybrid system.

1 Introduction

L'analyse de sentiments associés aux aspects vise à découvrir les sentiments exprimés par un utilisateur sur les différents aspects d'une entité donnée, (Hu & Liu, 2004), (Liu, 2012). Récemment, la campagne d'évaluation "SemEval" a proposé une tâche d'analyse de sentiments associés aux aspects, (Pontiki *et al.*, 2014), à laquelle nous avons participé. Nous présentons le système que nous avons développé et les résultats obtenus dans le cadre de la campagne (Brun *et al.*, 2014). Pour cet article, en plus de présenter la tâche 4 de la campagne SemEval 2014, nous avons voulu reprendre le travail que nous avons fait pour SemEval afin d'évaluer les points forts et les points faibles de notre système, et de le comparer aux autres systèmes présentés, avec comme objectif l'amélioration des résultats. Pour cela, nous avons réalisé un nouvel ensemble d'expériences pour déterminer les parties les plus pertinentes à modifier.

2 Présentation de la tâche 4 de SemEval 2014

L'analyse de sentiments est une tâche qui connaît actuellement un engouement certain. Cependant, la majorité des approches actuelles tente de détecter la polarité globale d'une phrase ou d'un document dans son ensemble, indépendamment des entités mentionnées (les concepts du domaine visé, par exemple, ordinateurs portables, restaurants, films) et leurs aspects (par exemple la batterie, l'écran, la nourriture, le service, ...). L'objet de la tâche 4 de la campagne d'évaluation SemEval2014 était précisément de s'atteler à l'analyse d'opinions associées aux aspects, c'est-à-dire à détecter les aspects cibles des opinions. Les corpus fournis pour cette tâche étaient constitués de commentaires d'internautes annotés en aspects et polarités. Plus exactement, la tâche était subdivisée en 4 sous-tâches :

1. Extraction des termes dénotant les aspects : par exemple, dans la phrase *I liked the service and the staff, but not the food*, les termes à détecter sont *service*, *staff* et *food*.

2. Extraction des catégories sémantiques dénotant les aspects : étant donné un ensemble prédéfini de catégories par domaine ({ "price", "food", "service", "ambiance", et "anecdote" } dans le domaine des restaurants), associer ces catégories aux phrases ; leur granularité est moins fine que celle des termes de la sous-tâche précédente et elles ne sont pas nécessairement associées à la présence de termes dans la phrase. Par exemple :
 "The restaurant was too expensive" → { price }
 "The restaurant was expensive, but the menu was great" → { price, food }
3. Extraction de la polarité associée aux termes précédemment détectés : la polarité prend ici 4 valeurs : { *positif*, *négatif*, *neutre* et *conflit* }.
 "I loved their **fajitas**" → { fajitas : **positive** }
 "I hated their **fajitas**, but their **salads** were great" → { fajitas : **negative**, salads : **positive** }
 "The **fajitas** are their first plate" → { fajitas : **neutral** }
 "The **fajitas** were great to taste, but not to see" → { fajitas : **conflit** }
4. Extraction de la polarité associée aux catégories précédemment détectées : Par exemple :
 "The restaurant was too expensive" → { price : **negative** }
 "The restaurant was expensive, but the menu was great" → { price : **negative**, food : **positive** }

Deux jeux de données étaient disponibles, l'un concernant des revues d'ordinateurs portables, annotées seulement pour les sous-tâches 1 et 3, l'autre concernant des revues de restaurants, annotées pour les 4 sous-tâches. Nous nous sommes concentrés sur ce dernier domaine afin d'aborder l'ensemble des sous-tâches. Le corpus d'entraînement comprenait 3044 phrases annotées avec 3700 termes et leur polarité et 3715 catégories sémantiques et leur polarité.

3 Notre système d'analyse de sentiments associés aux aspects

3.1 Le système pré-existant

Nous avons d'abord adapté le système d'analyse d'opinions que nous avons développé préalablement (Brun, 2011, 2012). Ce système est fondé sur un analyseur syntaxique robuste XIP, (Ait-Mokhtar *et al.*, 2001), utilisé pour calculer des relations d'opinions, en combinant dépendances syntaxiques et informations lexicales sur la polarité des mots et la sous-catégorisation des prédicats. Ce système génère des dépendances sémantiques appelées SENTIMENT qui sont soit binaires, c'est-à-dire reliant les prédicats de polarité et les cibles des opinions, soit unaires, lorsque la cible de l'opinion n'a pu être détectée. Par exemple, l'analyse de *I was very disappointed by the food and the service*, produit :

SUBJ_N(disappointed, food), SUBJ_N(disappointed, service), OBJ_N(disappointed, I), MANNER_PRE(disappointed, very), SENTIMENT_NEGATIVE(disappointed, service), SENTIMENT_NEGATIVE(disappointed, food)

Ce système préexistant n'extrait pas explicitement les termes dénotant les aspects du domaine, cependant les arguments cibles de la dépendance "SENTIMENT" désignent potentiellement des aspects. De plus, ce système n'extrait que les opinions positives ou négatives, mais ne couvre pas la polarité neutre, ni le conflit. Le système utilisé pour la campagne SemEval est celui de l'anglais, mais un système équivalent est également disponible pour le français.

3.2 Adaptation pour SemEval2014

Ce système a été adapté pour les besoins de SemEval2014 selon deux axes : acquisition lexicale afin de couvrir les termes du domaine, et développement de règles de détection des termes multi-mots et de règles d'extraction d'opinions.

3.2.1 Acquisition lexicale et détection de termes

Le lexique de notre système inclut préalablement du vocabulaire de polarité, tandis que la tâche de détection de termes de SemEval implique des connaissances lexicales du domaine. Nous avons donc encodé du vocabulaire de domaine concernant les revues de restaurant, en utilisant le corpus d'entraînement, les termes y étant explicitement balisés, et en lui associant les traits sémantiques correspondant aux aspects (*food*, *service*, *ambiance*, *price*, *anecdote*). Cette liste a

ensuite été étendue par filtrage de vocabulaire extrait du portail Wikipedia dédié à la nourriture¹. Cela nous a permis d'encoder 761 termes reliés à la nourriture, 31 termes reliés au prix, 105 termes reliés à l'ambiance, et 42 termes reliés au service. Pour détecter les termes complexes, des grammaires locales ont été construites sur la base de ce vocabulaire. Les règles, de type "expressions régulières", détectent les termes à mot multiples, comme *Grilled Chicken special with Edamame Puree* ou encore *staff members* et leur associent une catégorie sémantique.

3.2.2 Adaptation de la grammaire pour la détection de la polarité

La grammaire a été également modifiée pour détecter les opinions liées aux termes et aux aspects sémantiques : si un terme est argument d'une dépendance SENTIMENT, deux nouvelles dépendances sont créées, une associée au terme (OPINION_ON_TERM) et une associée à la catégorie sémantique correspondante (OPINION_ON_CATEGORY) Elles héritent de la polarité de la dépendance SENTIMENT. Si deux dépendances SENTIMENT ciblent un même terme avec des polarités opposées, elles sont alors créées avec le trait de polarité CONFLIT. Si un terme est détecté sans être cible de la dépendance SENTIMENT, elles sont créées avec le trait de polarité "neutre". Enfin, si aucun terme n'est détecté dans la phrase, OPINION_ON_CATEGORY est créée avec la cible "anecdote" (= aspect "fourre-tout"), héritant de la polarité calculée dans la phrase, ou bien du neutre. La dépendance OPINION_ON_TERM relie les termes et leur polarité et constitue la base des sous-tâches 1 et 3.

3.3 Classification

Tandis que la détection des termes et des polarités qui leur sont associées est effectuée par la grammaire, la détection des catégories sémantiques (aspects) et de leur polarité est réalisée par classification automatique via la bibliothèque "LibLinear", (Fan *et al.*, 2008). Nous entraînons tout d'abord un classifieur unique pour associer leurs catégories sémantiques aux phrases, puis pour chaque catégorie, nous entraînons un classifieur de polarité. Dans les deux cas, nous appliquons une validation croisée ("10-fold cross-validation"), qui nous a permis de sélectionner le classifieur de meilleure performance en fonction de la sous-tâche. Nous avons retenu cette méthode pour pallier les erreurs possibles de la grammaire. De cette façon, les informations issues de la grammaire sont utilisées comme des attributs à la fois dans la phase d'entraînement et de prédiction, mais ne sont retenues que les valeurs renvoyées par les classifieurs comme résultats finaux.

3.3.1 Classification des aspects

Ce module associe une ou plusieurs catégories sémantiques aux phrases des revues de restaurant. Pour chaque phrase, le module utilise comme traits le "sac de mots" de la phrase ainsi que des informations fournies par l'analyseur syntaxique. Dans l'étape de prétraitement, les mots outils (déterminants, conjonctions) sont éliminés. Nous utilisons la régression logistique ("L2-regularized") de LibLinear, pour créer le modèle. Les traits considérés sont les lemmes des mots avec leurs fréquences. De plus, l'information fournie par l'analyseur syntaxique est utilisée pour incrémenter la fréquence des termes appartenant aux catégories détectées. Cette information se compose des dépendances associant leurs catégories aux termes et des dépendances d'opinions concernant catégories et termes. Par exemple, pour la phrase suivante : *"Fabulous service, fantastic food, and a chilled out atmosphere and environment"*, les dépendances retenues sont :

```
SERVICE(service), FOOD(food) AMBIENCE(atmosphere), AMBIENCE(environment)
OP_CAT_POS(food), OP_CAT_POS (service), OP_CAT_POS (ambiance),
OP_TERM_POS(food), OP_TERM_POS(service), OP_TERM_POS(atmosphere), OP_TERM_POS(environment).
```

Ceci nous conduit à augmenter les fréquences de la façon suivante : food (+3), service (+3), atmosphere (+2), environment (+2), ambiance (+1). La régression logistique associe alors chaque catégorie à une certaine probabilité. Nous imposons un seuil concernant les valeurs des probabilités des catégories à retenir ; après plusieurs expérimentations le seuil (borne inférieure) que nous avons retenu est de 0,25.

1. http://en.wikipedia.org/wiki/Category:_Listes_of_foods

3.3.2 Classification des polarités associées aux catégories

L'approche pour le calcul des polarités est assez similaire, à quelques détails près. Nous utilisons également les traits correspondant au sac de mots, et la polarité fournie par les dépendances suivantes : `OPINION_ON_CATEGORY` et `SENTIMENT`. Lorsque ces dépendances sont détectées, un trait de la forme `polarity_category` est ajouté pour la classification. Ainsi pour l'exemple précédent : "*Fabulous service, fantastic food, and a chilled out atmosphere and environment*", les dépendances supplémentaires considérées sont :

`SENTIMENT_POSITIVE(atmosphere, chilled out)`, `SENTIMENT_POSITIVE(food, fantastic)`,
`SENTIMENT_POSITIVE(service, fabulous)`.

Après appariement des catégories avec les termes correspondants, les traits supplémentaires sont : *positive_ambience*, *positive_food* et *positive_service*. Puisque la dépendance `OPINION_ON_CATEGORY` est également détectée par l'analyseur, chacun des traits mentionnés ci-dessus aura une fréquence de 2. De plus, la polarité seule est également ajoutée comme trait. L'entraînement des modèles est réalisé avec le solveur SVM "L2-regularized L2-loss" de la bibliothèque Liblinear. Un modèle de classification en polarité est produit pour chaque catégorie. Ainsi, selon les catégories préalablement associées à une phrase donnée, le modèle correspondant est employé pour en prédire la polarité.

3.3.3 Correction de la polarité des termes

Comme précédemment évoqué, les termes et leur polarité sont détectés par la grammaire, la polarité *neutre* étant affectée par celle-ci lorsque ni *positif* ou *negatif* n'est associé. Nous avons utilisé les résultats de la classification des polarités des catégories pour corriger celle des termes, dans le cas des neutres : ainsi, si la classification associe à la catégorie *food* la polarité *positif* dans une phrase pour laquelle seuls des termes de polarité *neutre* sont détectés, ils sont corrigés et prennent la polarité *positif*. Cette correction *a posteriori*, possible du fait que les termes sont associés à leur catégories sémantiques par notre grammaire, s'est montré extrêmement performante (voir table 3).

3.4 Evaluation

Le corpus de test de SemEval14 contient 800 phrases, 1134 occurrences de termes et leur polarité (555 termes distincts), 1025 catégories et leur polarité. L'évaluation s'est déroulée en deux phases : la phase A consistait à annoter les termes et catégories, puis un corpus corrigé pour les termes et les catégories était renvoyé pour la phase B, concernant l'annotation des polarités.

3.4.1 Détection des termes et des catégories (Phase A)

Les mesures considérées pour ces deux sous-tâches étaient les mesures classiques de précision, rappel et F-mesure. Les tableaux 1 et 2 présentent les résultats de notre système (XABSA) pour la détection des termes et des catégories. Les mesures de référence ("baseline") sont décrites en détails dans l'article présentant la tâche (Pontiki *et al.*, 2014). La "baseline" pour la détection des termes consiste simplement à contruire un lexique des termes présents dans le corpus d'entraînement et à les repérer dans le corpus de test. Pour la détection des des catégories, chaque phrase du test est associée au k-phrases les plus proches du corpus d'entraînement (la similarité étant calculée avec le coefficient de Dice). Les catégories les plus fréquentes de ces k phrases sont ensuite associées aux phrases du test.

Méthode	Précision	Rappel	F-Mesure
Baseline	0,627329	0,376866	0,470862
XABSA	0,862453	0,818342	0,839818

TABLE 1: Détection des termes

Méthode	Précision	Rappel	F-Mesure
Baseline	0,637500	0,483412	0,549865
Sac de mots	0,77337	0,799024	0,785988
XABSA	0,832335	0,813658	0,822890

TABLE 2: Détection des catégories

On constate que la combinaison de l'approche "sac-de-mots" avec les sorties de l'analyseur syntaxique permet une amélioration notable des performances pour la détection des catégories. Pour les deux tâches de détection des termes et des

catégories, notre système surpasse largement la référence, se classant parmi les 3 premiers de la compétition pour le corpus des revues de restaurants.

3.4.2 Détection de la polarité (Phase B)

Similairement, les tableaux 3 et 4 décrivent les résultats du système sur la détection de la polarité pour les termes et les catégories. Ici la mesure utilisée est l'exactitude ("accuracy"). Les méthodes de calcul de la référence (baseline) sont également décrites en détails dans (Pontiki *et al.*, 2014). Pour les polarités des termes et des catégories, la polarité la plus fréquente des k-phrases les plus proches du corpus d'entraînement est associée aux phrases du test.

Méthode	Exactitude
Baseline	0,55
XABSA sans correction des termes	0,66
XABSA	0,78

TABLE 3: Polarité des termes

Méthode	Exactitude
Baseline	0,56
Sac de mots	0,68
XABSA	0,78

TABLE 4: Polarité des catégories

A nouveau, notre système s'est bien classé dans la compétition, avec une exactitude globale de 0,78 pour la détection de la polarité des termes et de 0,78 pour la détection de la polarité des catégories. Pour cette dernière sous-tâche, l'utilisation de traits provenant de l'analyse syntaxique combinés au sac de mots se montre là aussi très performante.

3.4.3 Comparaison avec les autres systèmes

32 équipes ont participé à cette campagne d'évaluation. Concernant la détection des termes, le meilleur système, DLIREC (Toh & Wang, 2014) est un tagger CRF, avec une F-mesure de 0,8401, juste devant notre système. Concernant la détection des aspects, le meilleur système, proposé par NRC-Canada (Kiritchenko *et al.*, 2014), obtient la F-mesure de 0,8857 avec 5 classifieurs SVM binaires (1 par aspect), entraînés avec divers n-grams et de l'information lexicale apprise sur le corpus YELP de revues de restaurants. Notre système obtient ici le 3ème F-score. Pour la détection des polarités des termes, DCU (Wagner *et al.*, 2014) et NRC-Canada obtiennent les meilleurs exactitudes, 80,95 et 80,15 respectivement. Ces deux systèmes utilisent à nouveau des classifieurs SVM enrichis avec des lexiques de polarité publiquement disponibles, de l'information syntaxique. Là aussi, notre système obtient le 3ème score. Enfin pour la détection des polarités associées aux aspects, NRC-Canada obtient le meilleur score avec une exactitude de 82,92 en utilisant un SVM enrichi pour capturer l'information relative à chaque aspect. Notre système est ici second.

D'une manière générale, les méthodes les plus performantes combinent des méthodes de classification "état de l'art" avec de l'information linguistique plus ou moins sophistiquée, certains systèmes se restreignant aux ressources fournies (systèmes contraints), d'autre utilisant des corpus additionnels (systèmes non contraints).

3.4.4 Analyse des erreurs

Si notre système a obtenu de bonnes performances, et ce, sur l'ensemble des 4 sous-tâches, les résultats détaillés en montrent les forces et les faiblesses, en particulier pour le calcul de la polarité, c.f. tableaux 5 et 6. On constate tout d'abord

Polarité	Préc.	Rappel	F-Mes.
Conflit	NaN (0/0)	NaN (0/0)	NaN
Négatif	0,78 (143/182)	0,73 (143/196)	0,76
Positif	0,79 (675/845)	0,92 (675/728)	0,86
Neutre	0,58 (62/107)	0,31 (62/196)	0,41

TABLE 5: Résultats par polarité (termes)

Polarité	Préc.	Rappel	F-Mes.
Conflit	0,5(7/14)	0,13(7/52)	0,21
Négatif	0,73(151/208)	0,68(151/222)	0,70
Positif	0,83(599/720)	0,91(599/657)	0,87
Neutre	0,50(42/83)	0,45(42/94)	0,47

TABLE 6: Résultats par polarité (catégories)

que la polarité *conflit* est particulièrement difficile à détecter. En effet, elle est à la fois mal couverte par notre grammaire, car les conflits mettent souvent en jeu des éléments à longue distance, et également assez peu représentée dans le corpus, ce qui pose problème pour l'apprentissage. La polarité *neutre* présente également des résultats relativement faibles : nous pensons que c'est lié au fait que la grammaire l'affecte par défaut, si *positif* ou *négatif* ne sont pas associés à un aspect.

4 Nouvelles expériences

A la suite de SemEval2014, nous avons voulu améliorer les performances de notre système, en particulier pour la phrase B pour lesquelles les résultats étaient un peu plus faibles pour la détection de la polarité des termes et des catégories. Nous avons donc mis en place de nouvelles expériences en intégrant les éléments les plus performants utilisés par les autres équipes lors de la campagne d'évaluation.

Pour les premières expériences, nous avons rajouté le "NRC emotion lexicon", (Mohammad & Turney, 2010) utilisé parmi d'autres par NRC-Canada lors de la campagne, comme lexique supplémentaire de polarité. Nous avons aussi tenté d'intégrer les "ngrams" dans notre système, à l'instar de nombres d'équipes qui ont eu de bons résultats. Nous nous sommes aussi intéressés à la notion de distance entre un mot et un terme de polarité (positif ou négatif) ainsi que la notion de distance entre un terme du domaine (*food*, *service*,...) et un terme de polarité (positif ou négatif).

Les tableaux 7 et 8 synthétisent les résultats de ces expériences pour la polarité des termes et des catégories sémantiques. Les résultats étant très proches, nous donnons, en plus de l'exactitude, le nombre d'occurrences correctes par rapport au nombre total.

Méthode	Exactitude
E0 : XABSA	0,787 (893/1134)
E1 : XABSA+lexique	0,76 (863/1134)
E2 : XABSA+bigrams	0,789 (895/1134)
E3 : XABSA+bigrams+distance	(895/1134)

TABLE 7: Polarité des termes

Méthode	Exactitude
E0 : XABSA	0,787 (807/1025)
E1 : XABSA+lexique	0.78 (788/1025)
E2 : XABSA+bigrams	0.798 (818/1025)
E3 : XABSA+bigrams +distance	0.80 (820/1025)

TABLE 8: Polarité des catégories

L'intégration du nouveau lexique dégrade nos résultats. Ce lexique a été créé via "crowdsourcing", et par exemple, associe le mot *food* à la polarité *positif*. Pour cela, il s'avère être difficilement intégrable directement à notre composant symbolique. L'utilisation des "ngrams" est une piste plus prometteuse, le meilleur résultat correspondant à l'intégration des "bigrams", pour lesquels on observe une nette amélioration des résultats. Concernant la notion de distance, nous rapportons également les meilleurs résultats qui correspondent à la distance entre termes du domaines et mots positifs ou négatifs. Ils améliorent légèrement l'expérience précédente. Ces expériences préliminaires montrent que des améliorations sont possibles, même si l'incrément est relativement réduit.

5 Conclusion

Dans cet article, nous sommes revenus en détails sur notre participation à la campagne d'évaluation SemEval14, pour la tâche 4 concernant l'analyse de sentiments associés aux aspects. Nous avons décrit en détails le système que nous avons conçu : il est fondé sur l'utilisation d'un composant symbolique permettant de détecter les termes du domaine et leur polarité et d'un composant de classification, qui classe les phrases selon leurs catégories sémantiques (aspects) et dans un second temps, leur associe une polarité. Nous montrons également comment les résultats de la classification de la polarité des catégories sémantiques peuvent améliorer a posteriori la polarité des termes. Ce système s'est montré performant, mais nous avons cherché à l'améliorer, en particulier pour certaines polarités, (*neutre* et *conflit*). Pour cela, nous avons réalisé quelques expériences préliminaires qui montrent qu'une amélioration est possible. Ceci constitue une des pistes de recherche future, mais nous souhaitons également travailler sur l'adaptation de ce système au français, pour lequel très peu de données annotées existent pour cette tâche, ainsi que sur l'application d'une méthodologie similaire pour la détection des émotions.

Références

- AIT-MOKHTAR S., CHANOD J.-P. & ROUX C. (2001). A multi-input dependency parser. In T. U. PRESS, Ed., *IWPT*.
- BRUN C. (2011). Detecting opinions using deep syntactic analysis. In *Proceedings of RANLP*, Hissar, Bulgaria.
- BRUN C. (2012). Learning opinionated patterns for contextual opinion detection. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference : Posters, 8-15 December 2012, Mumbai, India*, p. 165–174.

- BRUN C., POPA D. N. & ROUX C. (2014). Xrce : Hybrid classification for aspect-based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 838–842, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.
- FAN R.-E., CHANG K.-W., HSIEH C.-J., WANG X.-R. & LIN C.-J. (2008). Liblinear : A library for large linear classification. *Journal of Machine Learning Research*, **9**, 1871–1874.
- HU M. & LIU B. (2004). Mining and summarizing customer reviews. In *KDD*, p. 168–177.
- KIRITCHENKO S., ZHU X., CHERRY C. & MOHAMMAD S. (2014). Nrc-canada-2014 : Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 437–442, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.
- LIU B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- MOHAMMAD S. M. & TURNEY P. D. (2010). Emotions evoked by common words and phrases : Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, p. 26–34, Stroudsburg, PA, USA : Association for Computational Linguistics.
- PONTIKI M., GALANIS D., PAVLOPOULOS J., PAPAGEORGIOU H., ANDROUTSOPOULOS I. & MANANDHAR S. (2014). Semeval-2014 task 4 : Aspect based sentiment analysis. In *International Workshop on Semantic Evaluation (SemEval)*.
- TOH Z. & WANG W. (2014). Dlirec : Aspect term extraction and term polarity classification system. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 235–240 : Association for Computational Linguistics.
- WAGNER J., ARORA P., CORTES S., BARMAN U., BOGDANOVA D., FOSTER J. & TOUNSI L. (2014). Dcu : Aspect-based polarity classification for semeval task 4. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, p. 223–229, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

La ressource EXPLICADIS, un corpus annoté spécifiquement pour l'étude des relations de discours causales

Caroline Atallah¹

(1) CLLE-ERSS (UMR 5263), Université de Toulouse

caroline.atallah@univ-tlse2.fr

Résumé. Dans le but de proposer une caractérisation des relations de discours liées à la causalité, nous avons été amenée à constituer et annoter notre propre corpus d'étude : la ressource EXPLICADIS (EXplication et ARGumentation en DIScours). Cette ressource a été construite dans la continuité d'une ressource déjà disponible, le corpus ANNODIS. Proposant une annotation plus précise des relations causales sur un ensemble de textes diversifiés en genres textuels, EXPLICADIS est le premier corpus de ce type constitué spécifiquement pour l'étude des relations de discours causales.

Abstract.

A corpus specifically annotated for causal discourse relations studies : the EXPLICADIS resource.

In order to offer a characterization of causal discourse relations, we created and annotated our own corpus : EXPLICADIS (EXplanation and ARGumentation in DIScourse). This corpus was built in the continuity of a readily available corpus, the ANNODIS corpus. Providing a more precise annotation of causal relations in a set of texts that are representative of multiple textual genres, EXPLICADIS is the first corpus of its kind built specifically for causal discourse relations studies.

Mots-clés : annotation de corpus, discours, relations causales.

Keywords: corpus annotation, discourse, causal relations.

1 Introduction

Dans cet article, nous présentons une nouvelle ressource annotée spécifiquement pour l'étude des relations de discours causales. Cette ressource a été constituée dans le cadre du projet qui lui a donné son nom, le projet EXPLICADIS (EXPLICATION et ARGUMENTATION en DISCOURS)¹.

Il nous semble important de distinguer deux types d'objectifs que peut viser la constitution d'un nouveau corpus annoté : celui de fournir des données directement exploitables pour un apprentissage automatique ou celui de tester des hypothèses et de faire émerger des données de nouveaux phénomènes à étudier. Le premier objectif implique que les recherches soient suffisamment avancées sur le sujet et que le modèle théorique utilisé soit stabilisé. L'annotation du corpus vise alors à constituer ce que l'on appelle un *corpus de référence*. Le second objectif, quant à lui, s'inscrit dans une perspective principalement *expérimentale*, il s'agit d'apprendre à partir des données et non de faire apprendre².

Bien que plusieurs projets antérieurs d'annotation discursive, tels que le RST TreeBank (Carlson *et al.*, 2001) ou le Penn Discourse TreeBank (PDTB, Miltsakaki *et al.*, 2004), affichent clairement l'objectif de constituer des corpus de référence, il nous semble, qu'au vu de l'état actuel des recherches sur le discours, l'annotation discursive ne peut, pour l'instant, que s'inscrire dans la seconde perspective.

Le corpus EXPLICADIS a été construit suite à l'exploitation du corpus ANNODIS³ (ANnotation DIScursive de corpus, Péry-Woodley *et al.*, 2009, 2011; Afantenos *et al.*, 2012) pour l'étude spécifique des relations de discours causales. Même si, pour les raisons évoquées plus haut, un corpus tel que ANNODIS ne peut, à notre sens, être qualifié de *corpus de référence*, ce n'est pas pour autant qu'il ne présente pas d'intérêt. Bien au contraire, ce type de ressource ouvre de nombreuses

1. Ce projet a été co-financé par le PRES toulousain et la région Midi-Pyrénées (2010-2013).

2. Bien entendu, les données peuvent être exploitées pour des techniques d'apprentissage automatique, mais il faut alors être conscients que les résultats ne peuvent au mieux qu'être approximatifs.

3. La ressource ANNODIS est disponible sur le site REDAC (Ressources Développées à CLLE-ERSS : <http://redac.univ-tlse2.fr/>), sous licence « Creative Commons ».

possibilités d’exploitations, ou plus précisément d’*explorations*. Nous parlerons pour cela de *corpus exploratoire*, plutôt que de *corpus de référence*.

Dans cet article, nous défendons, à travers la présentation du corpus EXPLICADIS, l’intérêt de constituer mais aussi d’exploiter un corpus de type *exploratoire*. Dans les sections qui suivent, nous présenterons les différentes étapes de notre démarche – exploitation du corpus ANNODIS, puis constitution du corpus EXPLICADIS et, enfin, exploitation de ce dernier – et montrerons comment chacune d’entre elles a contribué à faire avancer les recherches sur les relations de discours causales, recherches qui seront encore amenées à évoluer sur la base de cette ressource prochainement disponible en ligne.

2 Exploitation du corpus ANNODIS pour l’étude des relations causales

Dans le cadre de nos recherches, nous nous intéressons aux relations de discours liées à la causalité et cherchons à caractériser celles-ci, sur la base de données attestées, dans le cadre théorique proposé par la SDRT (*Segmented Discourse Representation Theory*, Asher & Lascarides, 2003). Pour parvenir à un tel objectif, il était nécessaire de nous appuyer sur un corpus annoté en relations de discours, et notamment en relations causales. Le corpus ANNODIS, premier corpus de textes en français enrichis d’annotations discursives, offrait une base pertinente pour l’étude des relations de discours. Nos premières analyses sur les relations causales ont ainsi été menées à partir des annotations réalisées dans le cadre de ce projet. Nous présenterons brièvement cette ressource, puis nous verrons pourquoi l’exploitation de ces données nous a rapidement amenée à envisager la constitution d’un nouveau corpus, un corpus enrichi cette fois spécifiquement pour l’étude des relations causales.

Le projet ANNODIS a été constitué selon deux approches. L’une d’entre elles, dite *approche ascendante*, a abouti à la constitution d’un corpus enrichi avec des relations discursives, corpus sur lequel nos premières analyses se sont appuyées. Il s’agissait de construire une représentation de la structure du discours en liant des unités discursives entre elles par des relations rhétoriques. Pour ce faire, des textes ont d’abord été segmentés en unités de discours élémentaires. Puis, les segments constitués ont été liés entre eux par des relations de discours.

La segmentation ainsi que l’annotation des relations discursives ont été réalisées selon les recommandations fournies par un manuel rédigé spécifiquement pour le projet (Muller *et al.*, 2012). Avant de trouver sa forme définitive, ce manuel a été testé par des annotateurs dits *exploratifs* (notés par la suite “annotateurs A et B”), puis modifié suite à cette première phase d’annotation. Le guide finalisé, de nouveaux textes, au nombre de 42, ont été segmentés, puis triplement annotés par de nouveaux annotateurs. Deux annotations, dites *naïves*, ont été réalisées par des étudiants ne possédant pas de connaissances particulières sur les théories du discours (notés “annotateurs 1, 2 et 3”). Puis, les membres du projet ont eux-mêmes contribué à l’annotation, fournissant une troisième annotation, dite *experte*, de ces 42 textes. En parallèle, les annotateurs A et B ont poursuivi leurs annotations pour fournir au final un ensemble supplémentaire de 44 textes annotés. Les annotateurs experts ont également procédé à une nouvelle annotation de ces 44 textes. La ressource finale comporte ainsi 86 textes segmentés et au moins doublement annotés en relations de discours.

Parmi les dix-sept relations proposées dans le manuel d’annotation, figurent quatre relations causales : *Explication*, *Résultat*, *Explication** et *Résultat**, codées respectivement *explanation*, *result*, *explanation** et *result**. Afin de guider les annotateurs, le manuel propose, pour chaque relation, une définition, des exemples, ainsi que, parfois, une liste de marqueurs potentiels. Nous reprenons ci-dessous les éléments principaux concernant les relations causales.

Explication (*explanation*)

- Définition : La relation d’Explication lie deux segments dont le second (celui qui est attaché) explique le premier (la cible) de façon explicite ou non (indépendamment de l’ordre de présentation). Le premier argument de la relation est le segment expliqué, et le deuxième la cause supposée. Si l’effet est attaché à la cause et non l’inverse, on a la relation de Résultat.
- Exemple : [L’équipe a perdu lamentablement hier.]_1 [Elle avait trop de blessés.]_2 `explanation(1,2)`
- Marqueurs possibles : *car, parce que, à cause de, du fait de, par la faute de, grâce à, si... c’est parce que..., depuis* (si causalité évidente)

Résultat (*result*)

- Définition : La relation Résultat caractérise des liens entre deux segments introduisant deux éventualités (événements ou états) dont la 2ème résulte de la première.
- Exemple : [Nicholas avait bu trop de vin.]_1 [et a donc dû rentrer chez lui en métro.]_2 `result(1,2)`
- Marqueurs possibles : *du coup, donc, par conséquent, en conséquence, par suite, à la suite de quoi*

Explication* (*explanation**) et **Résultat*** (*result**)

- Définition : Dans certains cas, les effets sémantiques du lien rhétorique qui s'établit entre deux segments ne portent pas sur les événements décrits dans les segments, mais sur les actes de parole eux-mêmes.
- Exemples :

[Ferme la porte,]_1 [il fait froid]_2 *explanation**(1, 2)

[Il fait froid,]_1 [ferme la porte.]_2 *result**(1, 2)

Pour nos analyses, nous nous sommes focalisée, parmi les relations annotées dans le cadre du projet ANNODIS, sur ces quatre relations qui correspondent aux relations causales définies par Asher & Lascarides (2003) dans le cadre de la SDRT.

L'analyse de ces dernières nous a permis de faire différents constats. Tout d'abord, la très faible représentation dans le corpus, voire l'absence, des relations d'*Explication** et de *Résultat** dans ANNODIS (table 1) a retenu notre attention.

Nombre de relations annotées	Annot. A (44 textes)	Annot. B (43 textes)	Annot. 1 (28 textes)	Annot. 2 (27 textes)	Annot. 3 (26 textes)	Experts (86 textes)	Total
<i>Explication</i>	39	63	62	38	39	120	361
<i>Résultat</i>	48	97	58	45	28	162	438
<i>Explication*</i>	7	6	8	0	0	0	21
<i>Résultat*</i>	0	0	0	0	0	0	0
Total relations annotées	1390	1426	1060	1110	1114	3353	9453

TABLE 1 – Nombre de relations annotées dans la ressource ANNODIS par chaque annotateur

D'autre part, nous nous sommes rendu compte que les relations causales qui avaient été annotées faisaient l'objet d'un accord inter-annotateurs très faible (voir Atallah, 2014). Dans le but de comprendre ces désaccords, ainsi que la très faible représentation des relations d'*Explication** et de *Résultat**, nous avons fait le choix de nous confronter nous-même à la tâche d'annotation.

Cette expérience, et plus particulièrement l'analyse des quelques occurrences des relations étiquetées *Explication**, nous a rapidement amenée à nous interroger sur la pertinence de la gamme de relations causales envisagées en SDRT et reprises dans ANNODIS. Observons les segments suivants extraits du corpus :

[Arturo a de la chance,]_38 [il arrive en Chine]_39 [au moment de la fête de la nouvelle année.]_40

La relation *explanation**(38, [39-40]) a été annotée et pourtant la relation en jeu ne correspond pas à la relation d'*Explication** telle que définie par la SDRT et reprise dans ANNODIS. En effet, elle ne partage pas grand chose en commun avec celle qui s'établit dans l'exemple cité dans le manuel (*Ferme la porte. Il fait froid.*). Par ailleurs, nous avons jugé que cette relation ne pouvait pas non plus être considérée comme une simple relation d'*Explication* : *Arturo a de la chance* peut être perçu comme un fait subjectif, alors que les arguments des relations d'*Explication* correspondent à des descriptions objectives d'éventualités.

En nous penchant sur les données d'ANNODIS, nous avons relevé d'autres relations pouvant être rapprochées de celle en jeu dans l'exemple que nous avons rapporté. Nous en avons conclu que la gamme de relations envisagée pour l'annotation, et par là-même celle définie en SDRT, ne permettait pas de rendre compte de toutes les relations observables dans les textes. Nous avons alors cherché à réorganiser les relations causales annotées dans le corpus selon la nature du lien en jeu, jusqu'à parvenir à une classification plus pertinente pour rendre compte de la diversité des données. La mise au point d'une nouvelle typologie des relations causales a motivé la constitution du corpus EXPLICADIS annoté sur cette base.

3 Constitution d'une ressource spécifique pour l'étude des relations causales

Si nos premières analyses sur les relations causales ont été menées à partir des annotations réalisées dans le cadre du projet ANNODIS et notamment des situations de désaccords inter-annotateurs, il nous a rapidement semblé nécessaire de procéder à la constitution d'un nouveau corpus enrichi spécifiquement pour l'étude des relations causales. Ce corpus devait être annoté à l'aide d'un jeu d'étiquettes plus important que celui utilisé dans ANNODIS afin d'établir des distinctions plus fines entre les relations causales et de répondre ainsi à certaines difficultés rencontrées lors de la campagne d'annotation précédente. Ce raisonnement suit l'hypothèse posée et vérifiée par Prévot *et al.* (2009) selon laquelle l'introduction d'une gamme de relations plus riche permettrait de rendre l'annotation plus précise et donc moins confuse.

Nous avons ainsi complété la liste des relations retenues lors du projet ANNODIS à l'aide de nouveaux types de relations

causales. Nous avons défini cette nouvelle gamme de relations de discours causales dans (Atallah, 2014). On y retrouve les relations d'*Explication* et de *Résultat*, caractérisées comme des *relations inter-événementielles*, relations dont les effets sémantiques portent sur le contenu. Les relations d'*Explication** et de *Résultat** définies précédemment ont été renommées *relations pragmatiques* dans le but de les distinguer d'un nouveau type de relations causales : les *relations causales épistémiques*. Ces relations introduites plus tôt par Sweetser (1990) ont été caractérisées plus précisément dans le cadre de la SDRT (Atallah, 2014). Enfin, un type de relations causales épistémiques particulier a été distingué des autres sous la dénomination de *relations inférentielles*, suite au rapprochement effectué entre ces relations et celles étudiées par Bras *et al.* (2009).

Nous reprenons ci-dessous cette liste de relations causales, liste qui a servi de base pour l'annotation du corpus EXPLI-CADIS. Chacune de ces relations est associée à une étiquette et illustrée à l'aide d'un exemple extrait du corpus.

Explication (*explanation*)

- Définition : L'éventualité décrite dans le 2nd segment est la cause de l'éventualité décrite dans le 1^{er} segment.
- Exemple : [L'armée est déçue,]_12 [il n'y a aucun viol, aucun pillage, aucun meurtre.]_13

explanation (12, 13)

Résultat (*result*)

- Définition : L'éventualité décrite dans le 1^{er} segment est la cause de l'éventualité décrite dans le 2nd segment.
- Exemple : [Arturo est un petit corbeau]_11 [qui s'ennuie.]_12 [Il décide d'aller visiter le monde.]_13

result (12, 13)

Explication épistémique (*explanation_{ep}*)

- Définition : Le locuteur rapporte ses croyances dans le 1^{er} segment et justifie celles-ci dans le 2nd segment en exposant ses connaissances.
- Exemple : [Ce phénomène semble se confirmer à Mariana,]_37 [où on peut observer deux voies parallèles à la sortie sud de la ville.]_38 *explanation_{ep}* (37, 38)

Résultat épistémique (*result_{ep}*)

- Définition : Les connaissances exposées par le locuteur dans le 1^{er} segment l'entraînent à croire certains faits rapportés dans le 2nd segment.
- Exemple : [Or la psychomécanique répond à ces deux types d'exigences.]_24 [Il serait donc intéressant de regarder si les outils théoriques qu'elle a développés permettent de rendre compte de certaines observations faites par la neuropsychologie.]_25 *result_{ep}* (24, 25)

Explication inférentielle (*explanation_{inf}*)

- Définition : Les deux segments reliés décrivent des connaissances du locuteur. Les faits connus décrits dans le 1^{er} segment découlent logiquement de ceux décrits dans le 2nd segment.
- Exemple : [BITNET était différent d'Internet]_7 [parce que c'était un réseau point-à-point de type « stocké puis transmis ».]_8 *explanation_{inf}* (7, 8)

Résultat inférentiel (*result_{inf}*)

- Définition : Les deux segments reliés décrivent des connaissances du locuteur. Les faits connus décrits dans le 2nd segment découlent logiquement de ceux décrits dans le 1^{er} segment.
- Exemple : [La première exposition avicole de Belfort date de 1922.]_4 [Cela fait donc plus de trois-quarts de siècle que la digne société du même nom encourage, dans la région, les éleveurs amateurs.]_5 *result_{inf}* (4, 5)

Explication pragmatique (*explanation_{prag}*)

- Définition : L'éventualité décrite dans le 2nd segment justifie l'acte de langage accompli lors de l'énonciation du 1^{er} segment.
- Exemple : [Mais que ces derniers se rassurent,]_25 [il y aura encore deux autres tours]_26 [pour se rattraper.]_27 *explanation_{prag}* (25, [26, 27])

Résultat pragmatique (*result_{prag}*)

- Définition : L'éventualité décrite dans le 1^{er} segment justifie l'acte de langage accompli lors de l'énonciation du 2nd segment.
- Exemple : [Suzanne Sequin n'est plus.]_1 [...] [Nos condoléances.]_35 *result_{prag}* (1, 35)

Ce nouveau jeu de relations permet de rendre compte d'une dimension de la causalité non traitée dans ANNODIS. En effet, celui-ci intègre, en plus de la causalité inter-événementielle, des relations qui relèvent de l'argumentation.

En plus des relations que nous venons de présenter, nous avons annoté certains indices linguistiques qui nous semblaient pertinents pour l'étude des relations de discours causales. Ces indices sont de deux types.

Le premier type d'indices correspond à des indices que nous avons associés à l'expression de la causalité. Ces indices (ou faisceaux d'indices) pouvaient correspondre à des connecteurs (*car, parce que, donc, alors...*), mais aussi à des structures

syntactiques particulières (apposition, participe présent, participe passé...) ou à des marques typographiques (deux points, guillemets, parenthèses...).

Le second type d'indices annoté concerne plus spécifiquement les relations causales épistémiques. Nous avons remarqué que ces relations s'accompagnaient souvent de la présence d'éléments exprimant la modalité. Nous avons donc repéré ces indices lorsque ceux-ci étaient présents. Parmi ceux-ci, on trouve, entre autres, des adverbes, comme *probablement*, des verbes modaux, comme *pouvoir*, mais aussi des terminaisons de conditionnel.

4 Présentation de la ressource EXPLICADIS et exploitations

Sur la base des éléments que nous venons de définir, nous avons pu procéder nous-même à l'annotation du corpus EXPLICADIS. Ce corpus a été constitué en trois grandes étapes : une phase exploratoire, suivie de deux phases successives de constitution, puis d'élargissement du corpus.

La première étape a permis de mettre au point la typologie de relations causales présentée plus haut. Pour cela, nous avons ré-annoté l'ensemble des textes annotés lors de la campagne naïve d'annotation d'ANNODIS (42 textes) en nous concentrant sur les relations causales. Dans un premier temps, nous avons cherché à repérer toutes les relations causales présentes en nous appuyant sur les textes segmentés issus d'ANNODIS. Puis, ce n'est que dans un second temps que nous avons pris connaissance des annotations réalisées dans le cadre du projet précédent. Cette démarche avait pour but d'éviter que nos annotations soient trop influencées par celles qui étaient déjà disponibles. Elle nous a permis d'ajouter des relations causales qui n'avaient été repérées par aucun annotateur, mais surtout de nous rendre compte des difficultés posées par l'annotation, étant donnée la gamme restreinte de relations causales considérée dans ANNODIS. Au cours de cette phase exploratoire, nous avons ainsi pu affiner la liste des relations causales nécessaires pour résoudre au mieux les désaccords entre les annotateurs d'ANNODIS.

Une fois les objets à annoter bien définis, nous avons pu procéder à la ré-annotation de l'ensemble des textes annotés lors de la campagne naïve (42 textes précédents) mais aussi exploratoire (44 textes supplémentaires) d'ANNODIS. Ce premier élargissement du corpus nous a permis d'obtenir un corpus plus grand que nous avons nommé "Corpus_86".

Pour l'ensemble des 86 textes, nous avons confronté nos annotations avec les annotations antérieures. Ainsi, à chaque fois qu'au moins un annotateur avait identifié une relation causale entre deux segments, nous avons proposé notre propre annotation, que la relation en jeu soit causale ou non. Nous avons ainsi ré-annoté 533 relations⁴ sur l'ensemble du Corpus_86. Le fait de devoir proposer une relation entre des arguments nous obligeait à réfléchir aux motivations qui nous poussaient à retenir ou non l'annotation d'une relation causale.

Pour qu'un corpus soit le plus représentatif possible, il faut veiller à ce qu'il associe deux caractéristiques (Habert, 2000) : il doit être de taille suffisante et il doit pouvoir rendre compte de la diversité des usages langagiers. Afin de répondre à la première exigence, nous avons, comme indiqué précédemment, élargi notre tout premier corpus de 42 à 86 textes, obtenant ainsi un corpus dont la taille peut être jugée satisfaisante pour mener des analyses quantitatives (27 547 mots).

En ce qui concerne la seconde caractéristique, le Corpus_86 présentait certaines limites. En effet, celui-ci est essentiellement constitué d'extraits de textes issus de brèves de presse (textes à dominante narrative issus de *Est-Républicain* : NEWS) et d'articles encyclopédiques (textes à dominante expositive issus de *Wikipédia* : WIK). Seuls cinq textes à dominante argumentative ont été annotés : deux textes issus d'articles scientifiques de linguistique (LING) et trois de rapports scientifiques concernant la géopolitique (GEOP). Nous avons donc envisagé d'intégrer à notre corpus de nouveaux textes argumentatifs afin d'obtenir une meilleure représentativité des genres textuels au sein de notre corpus. Par ailleurs, cette intégration se voulait pertinente au vu de notre objet d'étude – la *causalité* – et des liens étroits que celui-ci entretient avec l'argumentation.

Notre corpus d'étude, dans sa version finale, comprend, en plus des 86 textes initiaux, 31 extraits de textes supplémentaires. Nous avons sélectionné ces textes parmi ceux qui ont été exploités lors du projet ANNODIS, dans le cadre d'une autre approche. Ceux-ci n'ayant pas été traités par l'approche *ascendante* du projet, nous avons dû procéder à leur segmentation en unités de discours élémentaires avant de les annoter en relations causales.

La table 2 présente l'ensemble de notre corpus d'étude. Les 31 textes supplémentaires y sont représentés sous l'étiquette de "Corpus_31". Nous avons souhaité faire en sorte que les textes issus des sous-corpus LING et GEOP constituent un ensemble de textes argumentatifs comparable, en termes quantitatifs (*cf.* nombre de mots), à l'ensemble des textes narratifs

4. Ces chiffres ne tiennent pas compte par ailleurs des relations causales que nous avons ajoutées et qui n'avaient été repérées par aucun annotateur.

Sous-corpus	Corpus_86		Corpus_31		Total	
	textes	mots	textes	mots	textes	mots
NEWS	39	9 768	3	846	42	10 614
WIK	42	15 983	0	0	42	15 983
LING	2	586	19	6 691	21	7 277 5 229 } 12 506
GEOP	3	1 210	9	4 019	12	
Total	86	27 547	31	11 556	117	39 103

TABLE 2 – Répartition des textes au sein d'EXPLICADIS en fonction des sources dont ils sont issus

issus de NEWS, ainsi qu'à l'ensemble des textes expositifs issus de WIK.

Sur l'ensemble de ces textes, 319 relations causales ont été repérées et annotées à l'aide des huit étiquettes présentées précédemment. La table 3 montre que, même si les relations portant sur le contenu propositionnel (*Explication* et *Résultat*) sont majoritaires dans le corpus, celles-ci ne représentent qu'un peu plus de la moitié des relations causales que nous avons relevées (environ 53 %). La présence des autres relations causales est loin d'être négligeable. Sachant que les relations pragmatiques représentent moins de 1 % de l'ensemble des relations annotées⁵, la nécessité d'intégrer les relations causales épistémiques et inférentielles dans le cadre de la SDRT est confirmée par la réalité des données.

Nombre de relations annotées	Corpus_86	Corpus_31	Total
<i>Explication</i>	77	27	104
<i>Résultat</i>	55	10	65
<i>Explication_épistémique</i>	35	33	68
<i>Résultat_épistémique</i>	9	15	24
<i>Explication_inférentielle</i>	8	4	12
<i>Résultat_inférentiel</i>	26	17	43
<i>Explication_pragmatique</i>	1	1	2
<i>Résultat_pragmatique</i>	1	0	1
Total relations causales annotées	212	107	319

TABLE 3 – Nombre de relations causales annotées dans la ressource EXPLICADIS

Si la constitution de la ressource EXPLICADIS nous a permis d'aboutir à une meilleure caractérisation des relations de discours causales, elle nous a également autorisée à mener, par la suite, des analyses diversifiées. Nous avons pu notamment nous intéresser à la variation relative au genre textuel et mettre en évidence certaines corrélations entre différents paramètres : type de relation causale, choix rhétorique, genre textuel (voir Atallah, 2014).

Cette ressource devrait, par ailleurs, permettre à d'autres utilisateurs de l'exploiter pour leurs propres besoins. Par sa taille, la diversité des textes qui y sont représentés et les annotations proposées, ce corpus devrait constituer une base pertinente et originale pour l'étude des relations causales. En ce qui concerne la diversité des textes, nous notons qu'une telle ressource n'a, à notre connaissance, jamais été conçue. En effet, les projets antérieurs d'annotation discursive ont fait le choix de rester sur la construction de corpus homogènes, et plus spécifiquement de corpus constitués exclusivement de textes journalistiques (textes à dominante narrative)⁶.

Par ailleurs, la gamme de relations retenues pour l'annotation constitue un véritable atout pour mener des études descriptives. D'une part, les annotations proposées tiennent compte de distinctions assez fines et, d'autre part, il s'agit du premier corpus annoté pour le français qui s'appuie sur une vision intégrative de la causalité, considérant non seulement sa dimension événementielle, dimension habituellement traitée, mais également sa dimension argumentative.

Si ce corpus présente un format idéal pour adopter une approche onomasiologique face aux données, c'est-à-dire une approche qui part de la relation elle-même et non de ses marqueurs potentiels, il permettra également, grâce à une projection des indices annotés d'envisager des études selon une perspective sémasiologique (qui part des indices).

Afin de permettre à d'autres utilisateurs de l'exploiter pour leurs propres besoins, la ressource EXPLICADIS sera disponible en ligne, aux côtés du corpus ANNODIS. Tout comme pour ce dernier, les annotations retenues ne peuvent être considérées

5. Cette très faible proportion s'explique par les types de textes retenus dans le corpus. Un corpus rendant compte de la langue parlée, ou impliquant plus généralement des interactions, serait plus approprié pour l'étude de ce dernier type de relations.

6. Par exemple, les textes proposés par le RST TreeBank et par le PDTB sont extraits du *Wall Street Journal*. Quant au French Discourse Tree Bank (FDTB, Danlos *et al.*, 2012), il est prévu que ce corpus soit constitué de textes tirés du journal *Le Monde*.

comme des informations à valeur certaine et stabilisée, elles resteront un reflet de notre propre point de vue sur la causalité. En cela, EXPLICADIS ne se veut pas *corpus de référence* pour l'étude de la causalité, mais bien *corpus exploratoire*. Il pourra ainsi servir de point de départ pour des analyses ultérieures sur la causalité, non comme un objet figé, mais comme un objet que nous invitons à faire évoluer.

Il faut par ailleurs noter que ce corpus n'a fait l'objet que d'une simple annotation. Il serait intéressant de confronter nos annotations à celles d'autres annotateurs. Cela permettrait de tester d'une part la pertinence du nouveau jeu de relations causales que nous avons défini et d'autre part l'hypothèse selon laquelle un jeu de relations plus précis mène à une annotation moins confuse (Prévot *et al.*, 2009). Plus généralement, l'analyse des accords entre annotateurs permettrait de valider notre contribution à l'étude des relations causales, et l'analyse des désaccords de mettre en évidence les améliorations encore nécessaires. Tout comme pour le corpus ANNODIS, les désaccords inter-annotateurs qui pourront être relevés sur le corpus EXPLICADIS ne devront pas être traités comme des erreurs, mais, au contraire, ils devront être considérés et étudiés avec la plus grande attention. Si Habert (2004) recommande au linguiste d'accepter de travailler avec des données imparfaites, nous pensons que le linguiste doit aussi apprendre à tirer de ces imperfections – ici, des désaccords inter-annotateurs – de nouvelles informations qui pourront servir ses recherches.

Références

- AFANTENOS S. D., ASHER N., BENAMARA F., BRAS M., FABRE C., HO-DAC L.-M., LE DRAOULEC A., MULLER P., PÉRY-WOODLEY M.-P., PRÉVOT L., REBEYROLLE J., TANGUY L., VERGEZ-COURET M. & VIEU L. (2012). An empirical resource for discovering cognitive principles of discourse organization : the ANNODIS corpus. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC)*, p. 2727–2734, Istanbul, Turkey.
- ASHER N. & LASCARIDES A. (2003). *Logics of Conversation*. Cambridge University Press.
- ATALLAH C. (2014). *Analyse de relations de discours causales en corpus : étude empirique et caractérisation théorique*. Thèse de Doctorat, Université de Toulouse, Toulouse.
- BRAS M., LE DRAOULEC A. & ASHER N. (2009). A Formal Analysis of the French Temporal Connective *alors*. In BEHRENS & C. FABRICIUS-HANSEN, Eds., *Structuring information in discourse : the explicit/implicit dimension.*, volume 1, p. 149–170. Oslo Studies in Language.
- CARLSON L., MARCU D. & OKUROWSKI M. E. (2001). Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, volume 16, p. 1–10, Stroudsburg, PA, USA : Association for Computational Linguistics.
- DANLOS L., ANTOLINOS-BASSO D., BRAUD C. & ROZE C. (2012). Vers le FDTB : French Discourse Tree Bank. In *TALN 2012 : 19ème conférence sur le Traitement Automatique des Langues Naturelles*, volume 2, p. 471–478, Grenoble, France.
- HABERT B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ? In M. BILGER, Ed., *Linguistique sur corpus. Etudes et réflexions*, volume 31 of *Cahiers de l'université de Perpignan*, p. 11–58. Perpignan : Presses Universitaires de Perpignan.
- HABERT B. (2004). Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs. *Revue française de linguistique appliquée*, **IX**(1), 5–24.
- MILTSAKAKI E., PRASAD R., JOSHI A. & WEBBER B. (2004). The Penn Discourse Treebank. In *In Proceedings of LREC 2004*, Lisbon, Portugal.
- MULLER P., VERGEZ-COURET M., PRÉVOT L., ASHER N., BENAMARA F., BRAS M., LE DRAOULEC A. & VIEU L. (2012). Manuel d'annotation en relations de discours du projet ANNODIS. *Carnets de grammaire*, **21**.
- PRÉVOT L., VIEU L. & ASHER N. (2009). Une formalisation plus précise pour une annotation moins confuse : la relation d'élaboration d'entité. *Journal of French Language Studies*, **19**(2), 207–228.
- PÉRY-WOODLEY M.-P., AFANTENOS S. D., HO-DAC L.-M. & ASHER N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *TAL*, **52**(3), 71–101.
- PÉRY-WOODLEY M.-P., ASHER N., ENJALBERT P., BENAMARA F., BRAS M., FABRE C., FERRARI S., HO-DAC L.-M., LE DRAOULEC A., MATHET Y., MULLER P., PRÉVOT L., REBEYROLLE J., TANGUY L., VERGEZ-COURET M., VIEU L. & WIDLÖCHER A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. In *Actes de TALN 2009*, Senlis, France.
- SWEETSER E. E. (1990). *From Etymology to Pragmatics : Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge : Cambridge University Press.

La séparation des composantes lexicale et flexionnelle des vecteurs de mots

François Lareau Gabriel Bernier-Colborne Patrick Drouin
 OLST, Université de Montréal, C.P. 6128, succ. Centre-Ville, Montréal QC H3C 3J7, Canada
 { francois.lareau | gabriel.bernier-colborne | patrick.drouin }@umontreal.ca

Résumé. En sémantique distributionnelle, le sens des mots est modélisé par des vecteurs qui représentent leur distribution en corpus. Les modèles étant souvent calculés sur des corpus sans pré-traitement linguistique poussé, ils ne permettent pas de rendre bien compte de la compositionnalité morphologique des mots-formes. Nous proposons une méthode pour décomposer les vecteurs de mots en vecteurs lexicaux et flexionnels.

Abstract.

Separating the lexical and grammatical components of semantic vectors

In distributional semantics, the meaning of words is modelled by vectors that represent their distribution in a corpus. Vectorial models being often built from corpora with little linguistic pre-treatment, they do not represent very well the morphological compositionality of words. We propose here a method to decompose semantic vectors into lexical and inflectional vectors.

Mots-clés : Sémantique distributionnelle ; compositionnalité ; flexion.

Keywords: Distributional semantics ; compositionality ; inflection.

1 Introduction

En sémantique distributionnelle, le sens des mots est représenté par des vecteurs qui représentent leur distribution en corpus (Turney & Pantel, 2010). Les modèles étant souvent calculés sur des corpus sans pré-traitement linguistique poussé, ils ne permettent pas de rendre bien compte de la compositionnalité morphologique des mots-formes. Cela mène à des aberrations. Par exemple, dans le modèle de Mikolov *et al.* (2013b), les vecteurs des quasi-synonymes *seems* et *appears* sont plus similaires que ceux de *seems* et *seemed*, qui appartiennent pourtant au même vocable (cf. table 1). Puisque ces vecteurs sont construits en discours, ils contiennent à la fois de l'information sémantique et de l'information d'autre nature tenant plus au fonctionnement des mots dans les phrases et aux propriétés des mots eux-mêmes qu'à leur sens comme tel. Ce bruit ne gêne généralement pas les recherches qui visent une approximation du sens des formes, mais peut gêner considérablement les travaux de nature purement linguistique. Nous cherchons, dans cet article, à éliminer le bruit morphosyntaxique contenu dans les vecteurs dans le but d'isoler le contenu sémantique.

<i>seemed</i>	<i>seems</i>	0,721
<i>appeared</i>	<i>appears</i>	0,657
<i>seemed</i>	<i>appeared</i>	0,723
<i>seems</i>	<i>appears</i>	0,814

TABLE 1 – Aberrations

Notre hypothèse est que, dans le vecteur d'une forme fléchie, on peut en isoler la partie lexicale de sa partie flexionnelle. Par exemple, comme l'illustre la figure 1, les vecteurs des verbes au passé *got*, *began* et *wrote* devraient pouvoir être décomposés en un vecteur lexical (GET, BEGIN ou WRITE) et un vecteur flexionnel représentant le passé (VBD¹) qui est commun à ces formes. De la même façon, les formes *wrote*, *writes* et *write*² partagent un vecteur lexical commun qui représente le vocable WRITE.

Notre objectif est donc d'identifier automatiquement, dans des vecteurs construits *a priori*, des sous-vecteurs qui représentent le contenu flexionnel vs lexical. Dans l'exemple qui précède, nous cherchons donc à isoler le vecteur VBD à partir d'un ensemble d'observations effectuées sur des formes au passé (colonne de gauche dans la figure 1). Une fois ce vecteur

1. Nous utilisons les codes du *Penn Treebank*.

2. Même si *write* n'a pas de marqueur morphologique explicite, nous le considérons comme une forme fléchie puisqu'il porte de l'information grammaticale (infinitif, impératif ou présent de l'indicatif).

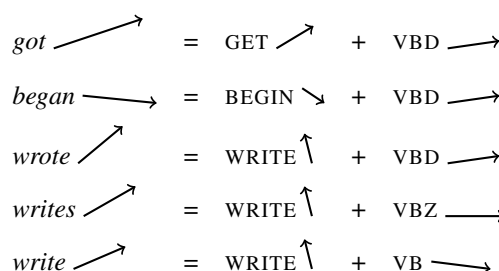


FIGURE 1 – Décomposition de vecteurs de mots en vecteurs lexicaux et flexionnels

isolé, nous pourrions le soustraire de la représentation vectorielle de *wrote* pour isoler le vecteur lexical épuré. En répétant le processus sur des formes à la troisième personne du singulier du présent de l'indicatif ou sur des formes nues, nous croyons qu'il sera possible d'isoler les vecteurs correspondants, qui devraient être relativement proches de ceux isolés à partir des formes au passé. L'intérêt d'une démarche qui rendrait possible l'identification et le retrait d'un vecteur flexionnel n'est pas négligeable puisqu'elle permettrait de maximiser la proximité sémantique entre les formes fléchies des mots et d'identifier les sens lexicaux « purs ».

2 Travaux antérieurs

La méthode que nous décrivons repose sur des représentations vectorielles de mots, qui peuvent être construites de différentes façons ; ici, nous utilisons des représentations apprises au moyen du modèle de langue neuronal *word2vec* (Mikolov *et al.*, 2013a,b). Des modèles de ce type ont été exploités dans de nombreuses applications du TAL, telles que l'étiquetage morphosyntaxique et sémantique, la segmentation, la reconnaissance automatique de la parole et la traduction automatique (Schwenk, 2007; Collobert *et al.*, 2011; Do *et al.*, 2014). Ces modèles apprennent des représentations distribuées de mots qui modélisent leurs propriétés sémantiques et morphosyntaxiques. Ces représentations peuvent notamment être exploitées afin d'estimer la similarité des mots, en utilisant une mesure de similarité de vecteurs telle que le cosinus.

La capacité des modèles de langue neuronaux à modéliser des régularités sémantiques et morphosyntaxiques a été démontrée au moyen de tâches de résolution d'analogies telles que « *homme* est à *roi* ce que *femme* est à *__* ». Mikolov *et al.* (2013c) ont montré qu'il est possible de résoudre de telles analogies au moyen d'un modèle comme *word2vec* en appliquant des opérations simples aux représentations de mots, en l'occurrence en soustrayant le vecteur du mot *homme* de celui de *roi*, puis en additionnant le vecteur de *femme*, le vecteur résultant ayant comme plus proche voisin *reine*. Les auteurs ont évalué leur modèle de langue sur des analogies sémantiques telles que *homme* : *roi* :: *femme* : *reine*, mais ont aussi créé à des fins d'évaluation un ensemble d'analogies morphosyntaxiques telles que *pomme* : *pommes* :: *voiture* : *voitures*. D'autres, comme Lazaridou *et al.* (2013) se sont intéressés à la compositionnalité des mots morphologiquement dérivés. Toutefois, à notre connaissance, personne n'a encore cherché à décomposer les vecteurs de mots pour en isoler les composantes lexicale et flexionnelle.

3 Données et méthodologie

Nous avons choisi de travailler sur les verbes en anglais, et ce, pour deux raisons. D'abord, à cause de la disponibilité d'un très gros modèle pré-entraîné, celui de Mikolov *et al.* (2013b). Ce modèle à 300 dimensions a été entraîné sur 100 milliards de mots du corpus *Google News* à l'aide du logiciel *word2vec* (Mikolov *et al.*, 2013a), et il est disponible gratuitement en ligne³. Pour le manipuler, nous avons utilisé la librairie Python *Gensim* (Řehůřek & Sojka, 2010)⁴. L'autre raison pour laquelle nous avons travaillé sur l'anglais plutôt que le français est qu'il est morphologiquement moins riche, ce qui veut dire que le nombre d'occurrences pour chaque forme est plus élevé dans le corpus, et donc le nombre de contextes dans lesquels elle apparaît, augmentant ainsi la qualité des vecteurs (Bullinaria & Levy, 2007).

Nous avons ciblé trois formes verbales : la forme nue, la troisième personne du singulier du présent de l'indicatif, ainsi que le passé (*become*, *becomes*, *became*). Le gérondif (*-ing*) étant très ambigu, puisqu'il correspond souvent à la fois à

3. <https://code.google.com/p/word2vec>

4. <https://radimrehurek.com/gensim>

un nom et à un verbe, nous ne l'avons pas utilisé.

Un premier jeu de données nous servant à effectuer les tests a été construit à partir d'une liste des mots anglais les plus fréquents (Davies, 2010)⁵. De cette liste, 19 verbes qui n'avaient pas d'homonyme évident d'une autre partie du discours et un qui en avait (*live, lives*) ont été manuellement sélectionnés (cf. table 2).

Une autre liste de verbes a été compilée automatiquement à partir de 10 millions de mots tirés du *British National Corpus* (Burnard, 2007). L'ensemble des verbes du BNC ont été isolés et les variantes morphologiques ont été regroupées autour du lemme. La liste résultante comportait 782 verbes, dont 762 avaient leurs trois formes dans le vocabulaire du modèle de Mikolov *et al.* (2013b) ; ce sont ces 3×762 formes verbales qui constituent le jeu de données étendu.

Nous construisons d'abord à partir du modèle pré-entraîné une matrice où le vecteur de chaque mot-forme du vocabulaire occupe une rangée. Ensuite, pour chaque colonne de la table 2, nous construisons une sous-matrice avec seulement les vecteurs des mots-formes de cette colonne (qui ont tous la même flexion). Nous avons donc une matrice de 3 millions de vecteurs et trois sous-matrices de 20 vecteurs chacune.

Notre hypothèse est que les mots-formes représentés par les vecteurs d'une sous-matrice, puisqu'ils partagent tous la même flexion, doivent avoir certaines propriétés distributionnelles en commun qui les distinguent des autres mots-formes du vocabulaire. Ces propriétés doivent se refléter dans les valeurs des 300 dimensions des vecteurs. Quand on compare les vecteurs d'une sous-matrice flexionnellement homogène à ceux de la matrice complète, certaines de leurs dimensions doivent être relativement homogènes. On peut les identifier en cherchant des dimensions qui varient peu au sein des vecteurs de la sous-matrice comparativement à la matrice générale.

Dans un premier temps, nous avons calculé, dans chaque sous-matrice, un « ratio flexionnel » pour chaque dimension. Ce ratio indique à quel point la dimension en question reflète le contenu flexionnel vs lexical du mot-forme. Plus une dimension est fortement associée à du contenu flexionnel, plus ce ratio est élevé. Il est calculé en comparant la variance de cette dimension dans la sous-matrice à sa variance dans la matrice de tout le modèle⁶. Si une dimension varie peu d'un vecteur à l'autre au sein de la sous-matrice alors qu'elle a une variance élevée dans le modèle en général, c'est un signe qu'elle est fortement liée à la flexion, qui est la propriété commune à tous les vecteurs de la sous-matrice. Le rapport V_m/V_s , où V_m est la variance de la dimension dans la matrice complète et V_s est la variance de cette même dimension dans la sous-matrice, sera plus élevé si la variance dans la sous-matrice est relativement plus faible que dans la matrice générale. Pour ramener ce rapport à des valeurs entre 0 et 1, nous utilisons la formule $\frac{\tanh(V_m/V_s - k) + 1}{2}$ pour calculer le ratio flexionnel de chaque colonne. La tangente hyperbolique (\tanh) donne une courbe en S, et la constante k permet d'ajuster où l'on souhaite que le rapport V_m/V_s croise 0,5, c'est-à-dire à partir de quel rapport V_m/V_s on considère que la dimension est surtout associée à du contenu flexionnel. Nous avons testé plusieurs valeurs de k entre 0,1 et 10, et les résultats étaient systématiquement meilleurs plus on se rapprochait de 0. Nous n'avons pas testé de valeurs négatives parce que les ratios obtenus avec un k quasi-nul étaient déjà très près de 1.

On obtient alors, pour chaque sous-matrice, une liste de n ratios (pour un espace sémantique à n dimensions) entre 0 et 1 qui nous indiquent à quel point chaque dimension est associée à la flexion qui est commune aux mots-formes de la sous-matrice. Ensuite, nous calculons la moyenne par colonne de la sous-matrice afin d'obtenir le vecteur moyen de cet ensemble de mots-formes. Nous multiplions chaque dimension de ce vecteur moyen par les ratios obtenus précédemment, ce qui nous donne alors le vecteur flexionnel qui est commun à tous les vecteurs de la sous-matrice.

Pour les trois formes à l'étude, les ratios obtenus pour chaque dimension étaient très élevés. Pour $k=1$, nous avons des ratios entre 0,61 et 0,99997, avec une médiane de 0,88. Ces ratios diminuent quand on augmente k et augmentent quand on réduit cette constante et nos résultats étaient systématiquement meilleurs plus nous rapprochions k de 0. Comme nous

VB	VBZ	VBD
ask	asks	asked
be	is	was
become	becomes	became
begin	begins	began
bring	brings	brought
continue	continues	continued
follow	follows	followed
get	gets	got
give	gives	gave
have	has	had
hear	hears	heard
live	lives	lived
meet	meets	met
receive	receives	received
seem	seems	seemed
send	sends	sent
speak	speaks	spoke
tell	tells	told
understand	understands	understood
write	writes	wrote

TABLE 2 – Verbes choisis manuellement

5. <http://www.wordfrequency.info>

6. Nous avons testé diverses mesures de la variance : variance, déviation standard, écart médian à la moyenne, et écart médian à la médiane. La variance donnait systématiquement de meilleurs résultats.

multiplions ces ratios par le vecteur moyen de la sous-matrice, il est apparu évident que nos résultats étaient meilleurs quand les vecteurs flexionnels se rapprochaient de la moyenne des vecteurs de la sous-matrice. Nous avons donc poursuivi en utilisant directement la moyenne d'une sous-matrice comme vecteur flexionnel. Cela présente l'avantage de ne pas nécessiter de calcul sur la matrice complète, qui est très lourde. Le vecteur moyen d'un groupe de vecteurs se trouve au « milieu » de ces vecteurs dans l'espace sémantique. En supposant que le contenu lexical des vecteurs les font dévier dans des directions indépendantes, leur milieu devrait correspondre à ce qu'ils ont en commun, c'est-à-dire leur flexion.

Nous avons comparé deux méthodes basées sur la moyenne pour calculer les vecteurs flexionnels. La première (nommée « A » dans les résultats ci-dessous) est la simple moyenne par colonne de la sous-matrice correspondant à une flexion donnée. La seconde méthode (« B » dans les résultats) est plus complexe. D'abord on calcule le vecteur moyen des trois formes fléchies pour chaque mot, ce qui nous donne une approximation du vecteur lexical. Ensuite, pour chaque forme fléchie, nous soustrayons de son vecteur l'approximation du vecteur lexical, ce qui nous donne une approximation de son vecteur flexionnel. Finalement, le vecteur correspondant à une flexion donnée est la moyenne des approximations flexionnelles obtenues à partir de toutes les formes qui portent cette même flexion.

Une fois que nous avons identifié nos trois vecteurs flexionnels (VB, VBZ et VBD) selon une des deux méthodes ci-dessus, nous les soustrayons de chaque vecteur dans la sous-matrice afin d'obtenir les vecteurs lexicaux. Comme nous avons trois formes fléchies pour chaque mot, en soustrayant de chaque vecteur initial les vecteurs flexionnels identifiés, nous obtenons trois vecteurs lexicaux pour chaque mot. On peut en faire la moyenne pour obtenir le vecteur lexical final. Au lieu de cela, nous nous servons de ces trois vecteurs pour évaluer la performance de notre méthode.

4 Évaluation

Nous avons utilisé deux méthodes pour évaluer notre travail. La première consiste à mesurer la différence entre la similarité des vecteurs des formes fléchies d'un même mot et celle entre les vecteurs lexicaux calculés à partir de ces formes. Pour chaque mot, nous avons trois vecteurs initiaux (avant traitement) et trois vecteurs lexicaux (après traitement). Nous calculons la moyenne des similarités cosinus entre chaque paire de vecteurs d'un même mot, avant et après traitement. En principe, si les vecteurs flexionnels que nous avons identifiés sont valides, alors en les soustrayant des vecteurs initiaux on devrait obtenir des vecteurs plus similaires qu'ils ne l'étaient au départ.

Afin de vérifier que ce n'est pas seulement parce qu'on soustrait des valeurs dans les vecteurs que nos vecteurs se rapproche après traitement, nous utilisons une quatrième sous-matrice (de 20 ou 762 verbes), celle-ci remplie de vecteurs choisis au hasard dans la matrice globale. Cette sous-matrice aléatoire est soumise au même traitement que les trois autres, mais comme il n'y a rien de commun aux vecteurs qui la composent, les « vecteurs flexionnels » identifiés ne correspondent à rien. La similarité cosinus moyenne entre ces vecteurs aléatoires et les trois vecteurs associés à chaque mot ne devrait donc pas augmenter significativement avant et après traitement.

La seconde méthode consiste à récupérer dans le modèle les plus proches voisins des vecteurs flexionnels que nous avons identifiés. On s'attend à ce que les voisins du vecteur VBD, par exemple, soient des verbes au passé, et que ceux de VBZ soient des verbes à la troisième personne du singulier.

5 Résultats et discussion

Les tables 3 et 4 ci-dessous donnent les résultats de nos expériences. Les différentes expériences sont identifiées par un code composé d'une lettre qui identifie l'approche utilisée (voir §3), suivie de deux nombres. Le premier indique le jeu de données qui a été utilisé pour identifier les vecteurs flexionnels (20 verbes sélectionnés manuellement ou 762 verbes tirés du corpus BNC). Le second indique le jeu de données dans lequel nous avons soustrait les vecteurs flexionnels pour identifier les vecteurs lexicaux. C'est dans ce jeu de données que nous avons mesuré la similarité des vecteurs avant et après traitement. Dans les deux tables, les trois dernières colonnes donnent les résultats du traitement dans la sous-matrice aléatoire qui sert de témoin dans nos expériences.

La méthode B-20-20 est celle qui donne les meilleurs résultats. On note une augmentation moyenne du score de similarité de 0,14, qui passe de 0,634 à 0,774, soit un gain moyen de 22% (autrement dit, la distance moyenne entre les vecteurs passe de 0,366 à 0,226, soit une réduction de 38%). L'augmentation du score de similarité atteint presque 14% pour B-762-20 et 12% pour A-20-20. Les deux méthodes testées fonctionnent donc bien, mais on obtient systématiquement de

Expérience	Avant	Après	Δ	Avant _{al.}	Après _{al.}	$\Delta_{al.}$
B-20-20	0,634	0,774	0,140	0,046	0,011	-0,035
B-762-20	0,634	0,720	0,086	0,043	0,003	-0,040
A-20-20	0,634	0,711	0,077	0,010	-0,007	-0,017
A-762-20	0,634	0,684	0,050	0,029	-0,015	-0,044
B-762-762	0,579	0,626	0,047	0,051	0,003	-0,048
A-20-762	0,579	0,591	0,013	0,044	-0,004	-0,048
A-762-762	0,579	0,587	0,009	0,052	0,003	-0,049
B-20-762	0,579	0,587	0,008	0,049	0,002	-0,047

TABLE 3 – Résultats de diverses méthodes (en ordre décroissant de performance)

meilleurs résultats quand on applique les vecteurs flexionnels identifiés pour isoler les vecteurs lexicaux dans les 20 verbes sélectionnés manuellement, peu importe la méthode ou le jeu de données utilisés pour identifier les vecteurs flexionnels. Contrairement aux verbes tirés du BNC, ceux de la liste courte ne sont pas ambigus (sauf LIVE) et la soustraction des vecteurs flexionnels verbaux est donc plus pertinente. Dans le cas de formes homographiques, les vecteurs peuvent contenir, pêle-mêle, de l'information flexionnelle verbale, nominale et/ou adjectivale. Une analyse qualitative des résultats obtenus avec les 762 verbes du BNC permet d'ailleurs de vérifier que la tête de la liste triée en ordre décroissant de gain de similarité contient moins de formes ambiguës que la fin de la liste.

En comparaison, l'application des deux méthodes sur la sous-matrice aléatoire donne toujours une diminution du score de similarité. Les augmentations que nous observons dans nos expériences sont donc bien liées à un apprentissage effectué sur des données homogènes.

La table 4 donne les résultats détaillés pour chacun des 20 verbes sélectionnés manuellement. Il est intéressant de noter que les formes de LIVE sont ambiguës et que ce sont elles qui obtiennent l'augmentation la plus négligeable de cette liste.

Mot-forme	Avant	Après	Δ	Avant _{al.}	Après _{al.}	$\Delta_{al.}$
<i>be</i>	0,523	0,715	0,192	0,033	0,005	-0,028
<i>get</i>	0,654	0,845	0,190	0,025	-0,008	-0,033
<i>give</i>	0,692	0,873	0,181	0,031	0,005	-0,026
<i>bring</i>	0,629	0,800	0,171	0,091	0,081	-0,010
<i>have</i>	0,620	0,787	0,167	0,061	0,037	-0,024
<i>receive</i>	0,682	0,841	0,159	0,077	0,053	-0,024
<i>ask</i>	0,684	0,843	0,158	0,001	-0,062	-0,063
<i>begin</i>	0,629	0,785	0,156	0,009	-0,026	-0,034
<i>speak</i>	0,637	0,793	0,155	0,062	0,012	-0,050
<i>send</i>	0,685	0,832	0,148	0,089	0,072	-0,017
<i>hear</i>	0,671	0,816	0,144	0,026	-0,032	-0,057
<i>meet</i>	0,637	0,773	0,136	0,054	0,022	-0,032
<i>tell</i>	0,613	0,745	0,132	0,087	0,023	-0,065
<i>become</i>	0,688	0,819	0,131	0,016	0,002	-0,015
<i>continue</i>	0,720	0,851	0,131	0,014	-0,021	-0,034
<i>follow</i>	0,582	0,709	0,127	0,125	0,108	-0,017
<i>understand</i>	0,616	0,718	0,101	0,141	0,105	-0,035
<i>seem</i>	0,688	0,788	0,099	0,058	0,048	-0,010
<i>write</i>	0,559	0,643	0,085	-0,020	-0,092	-0,072
<i>live</i>	0,471	0,507	0,036	-0,060	-0,106	-0,046
Moyenne	0,634	0,774	0,140	0,046	0,011	-0,035

TABLE 4 – Résultats détaillés de B-20-20 (en ordre décroissant de performance)

On note une corrélation modérée entre la similarité moyenne des formes d'un mot avant traitement et l'augmentation du score de similarité après traitement (les colonnes « Avant » et « Δ » dans les tables 3 et 4). Par exemple, pour B-762-762, cette corrélation est de 0,45. Ce n'est pas surprenant puisque la similarité moyenne des formes d'un mot a tendance à être

plus faible quand il y a une ou plusieurs formes ambiguës. Justement, ces formes ambiguës font obstacle à l'identification d'un vecteur flexionnel clair.

La table 5 donne les 20 plus proches voisins des vecteurs flexionnels VB, VBZ et VBD (en utilisant la méthode B sur 762 verbes). Ces voisins sont relativement distants (similarité cosinus entre 0,36 et 0,48). On voit clairement que le vecteur VB a été moins bien identifié que VBZ et VBD. C'est vraisemblablement dû au fait que la forme nue est elle-même ambiguë : il peut s'agir d'un infinitif, d'un présent de l'indicatif, d'un impératif, etc. On note que la plupart des voisins de VB sont des expressions à mots multiples. Nous croyons que cela s'explique par le fait que ces expressions sont peu fréquentes dans le corpus, et donc que leur vecteur est peu fiable. Autrement dit, notre vecteur VB se retrouve dans une zone mal définie de l'espace sémantique. Les deux autres formes sont moins ambiguës et donnent donc de meilleurs résultats. On peut ainsi supposer que notre méthode fonctionnerait mieux sur une langue morphologiquement plus riche, où les formes ont tendance à être moins ambiguës, à condition que le corpus soit assez gros pour assurer un nombre suffisant d'occurrences pour chaque forme.

VB	VBZ	VBD
<i>OPTION_ONE</i>	<i>sees</i>	<i>moved</i>
<i>tofind</i>	<i>creates</i>	<i>arrived</i>
<i>Bedbugs_Bite_Act</i>	<i>finds</i>	<i>turned</i>
<i>through_emergency_recapitalizion</i>	<i>introduces</i>	<i>returned</i>
<i>outguess_God</i>	<i>initiates</i>	<i>escorted</i>
<i>contacting_Melissa_Medalie</i>	<i>gets</i>	<i>shaken_aware</i>
<i>Yourself_Loan_Modification</i>	<i>uses</i>	<i>pushed</i>
<i>Eat_Spaghetti_Dinner</i>	<i>pushes</i>	<i>chased</i>
<i>Tiger_Wear_Necktie</i>	<i>sustains</i>	<i>hauled</i>
<i>unleash_cyberattacks</i>	<i>interprets</i>	<i>untouchable_Gio_Gonzalez</i>
<i>Diet_Pepsi_Skinny</i>	<i>embraces</i>	<i>exited</i>
<i>cooperate_Ramuglia</i>	<i>manages</i>	<i>snatched</i>
<i>###_###_####_FAX_Afri</i>	<i>amplifies</i>	<i>brought</i>
<i>OK_Vanhaudenhuysse</i>	<i>develops</i>	<i>stood</i>
<i>overcomplicate_things</i>	<i>takes</i>	<i>appeared</i>
<i>Heartaches_Begin</i>	<i>engages</i>	<i>had</i>
<i>Them_Hear</i>	<i>adopts</i>	<i>greeted_enthusiastically</i>
<i>Er_Rip</i>	<i>indulges</i>	<i>chief_Siegfried_Sievert</i>
<i>injure_Pacioretty</i>	<i>nurtures</i>	<i>approached</i>
<i>react_robustly</i>	<i>pulls</i>	<i>picked</i>

TABLE 5 – Vingt plus proches voisins des vecteurs flexionnels (B-762)

Il faut être prudent en comparant les résultats obtenus à partir des 20 verbes manuellement sélectionnés à ceux obtenus à partir du corpus BNC. En effet, il y a ici deux différences importantes : d'une part, le niveau d'ambiguïté, comme nous l'avons noté, mais aussi d'autre part la taille de l'échantillon. Pour vérifier que c'est bien l'ambiguïté qui est en jeu, on pourrait par exemple échantillonner 20 verbe parmi ceux du corpus BNC et répéter l'opération plus fois, puis faire la moyenne des résultats⁷, ce que nous n'avons pas testé.

6 Conclusion

La décomposition d'un vecteur en un vecteur lexical et un vecteur flexionnel est une forme de « séparation aveugle de source » en traitement du signal. Il est probable que l'apprentissage machine soit utile pour cette tâche, mais nous avons proposé des techniques simples basées sur la moyenne des vecteurs. Notre approche donne de bons résultats quand on traite des vecteurs de formes non ambiguës, mais la performance diminue considérablement quand on a des formes homonymiques. Nous croyons que cela reflète la qualité des vecteurs initiaux et est lié aux limites inhérentes à la sémantique distributionnelle.

7. Nous remercions le relecteur anonyme qui a attiré notre attention sur ce point.

On peut se demander si les vecteurs flexionnels obtenus peuvent eux-mêmes être décomposés. Par exemple, il est possible que les vecteurs VBD et VBZ puissent être décomposés en un vecteur correspondant à la partie du discours et un autre correspondant au sens flexionnel.

Ici, nous avons travaillé à partir de vecteurs déjà entraînés sur un corpus non lemmatisé. Nous voulions ainsi éviter le bruit inévitablement introduit par un lemmatisateur. Il serait néanmoins utile de comparer nos résultats avec un modèle entraîné sur un corpus lemmatisé pour voir si les vecteurs ainsi obtenus se rapprochent de nos vecteurs lexicaux. Il serait également intéressant de tester notre approche sur des vecteurs construits à partir de corpus désambiguïsés.

Dans notre approche initiale à base de ratios, nous avons observé, pour les trois formes à l'étude, des ratios toujours très élevés. Cela indique que la distribution d'un mot est surtout conditionnée par sa flexion. Est-ce parce que le sens d'un mot-forme est surtout flexionnel, ou est-ce que les vecteurs contiennent en fait surtout de l'information morphosyntaxique ? Nous croyons que c'est plutôt la deuxième explication qui est la bonne, mais il y a matière à débat.

Références

- BULLINARIA J. & LEVY J. (2007). Extracting semantic representations from word co-occurrence statistics : A computational study. *Behavior Research Methods*, **39**, 510–526.
- L. BURNARD, Ed. (2007). *Reference Guide for the British National Corpus*. Research Technologies Service at Oxford University Computing Services.
- COLLOBERT R., WESTON J., BOTTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, **12**, 2493–2537.
- DAVIES M. (2010). Word frequency data : Corpus of contemporary american english.
- DO Q.-K., ALLAUZEN A. & YVON F. (2014). Modèles de langue neuronaux : une comparaison de plusieurs stratégies d'apprentissage. In *Actes de la 21e conférence sur le traitement automatique des langues naturelles (TALN)*, p. 256–267, Marseille.
- LAZARIDOU A., MARELLI M., ZAMPARELLI R. & BARONI M. (2013). Compositionally derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, p. 1517–1526, Sophia.
- MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.
- MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.
- MIKOLOV T., YIH W.-T. & ZWEIG G. (2013c). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 746–751, Atlanta, Georgia : Association for Computational Linguistics.
- SCHWENK H. (2007). Continuous space language models. *Computer Speech & Language*, **21**(3), 492–518.
- TURNER P. D. & PANTEL P. (2010). From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research*, **37**(1), 141–188.
- ŘEHŮŘEK R. & SOJKA P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, p. 45–50, Valletta, Malta : ELRA.

Traitements pour l'analyse du français préclassique

Sascha Diwersy¹, Achille Falaise², Marie-Hélène Lay³, Gilles Souvay⁴

(1) Université de Cologne, Albertus-Magnus-Platz, D-50923 Köln, Allemagne

(2) ICAR, ENS de Lyon, 15 Parvis René Descartes, 69342 Lyon, France

(3) FoReLL, Université de Poitiers, 5 rue Théodore Lefebvre, 86073 Poitiers, France

(4) ATILF-CNRS, Université Nancy 2, 44 avenue de la Libération, 54063 Nancy, France

sascha.diwery@uni-koeln.de, achille.falaise@ens-lyon.fr, marie-helene.lay@univ-poitiers.fr, gilles.souvay@atilf.fr

Résumé. La période « préclassique » du français s'étend sur tout le XVI^e siècle et la première moitié du XVII^e siècle. Cet état de langue écrite, qui accompagne les débuts de l'imprimerie, est relativement proche du français moderne, mais se caractérise par une grande variabilité graphique. Il s'agit de l'un des moins bien dotés en termes de ressources. Nous présentons ici la construction d'un lexique, d'un corpus d'apprentissage et d'un modèle de langage pour la période préclassique, à partir de ressources du français moderne.

Abstract.

Treatments for Preclassic French parsing

The "Preclassical" French language period extends throughout the sixteenth century and the first half of the seventeenth century. This state of the written French language, which accompanies the beginnings of printing, is relatively close to the modern French, but is characterized by a large graphic variability. It is one of the most underresourced state of the French language. Here we present the construction of a lexicon, a training corpus and a language model for the Preclassic period, built from modern French resources.

Mots-clés : construction de lexique morphologique, annotation et étiquetage de corpus, linguistique diachronique.

Keywords: morphological lexicon construction, corpus annotation and tagging, diachronic linguistics.

1 Introduction

La période « préclassique » de langue française s'étend sur tout le XVI^e siècle et la première moitié du XVII^e siècle. Certaines caractéristiques de cette période, aujourd'hui désuètes, perdurent même jusqu'au XVIII^e siècle. Les écrits de cette époque, qui correspondent au début de l'imprimerie, présentent un début de normalisation graphique, mais la graphie est encore loin d'être stabilisée.

Cet état de la langue est encore peu traité, contrairement à la période médiévale, pour laquelle des ressources ont été créées ces dernières années. Ainsi, la *Base de Français Médiéval*¹ et le *Nouveau Corpus d'Amsterdam* (Gleßgen & Vachon, 2010) offrent des corpus étiquetés pour la période IX^e-XV^e siècles. À l'opposé, les corpus annotés du français moderne sont nombreux, mais la base *Frantext* catégorisée, par exemple, ne remonte pas avant 1850. Entre ces deux périodes, on peut citer principalement le corpus diachronique *Modéliser le changement : les voies du français*², dont quelques textes sont annotés. Toutefois, seuls trois textes annotés (environ 180 000 mots) correspondent à notre période. Il n'existe ainsi guère de ressources et de corpus étiquetés pour le français préclassique et classique.

Il est vrai que cet état de la langue ne présente pas les mêmes difficultés que la langue médiévale. À la différence de cette dernière, le français préclassique reste relativement intelligible pour un locuteur moderne (exemple en Figure 1) ; c'est surtout la variabilité graphique qui pose problème. Il semble ainsi possible d'adapter des ressources et des outils conçus pour le français moderne pour le traitement de cet état de la langue. C'est ce travail d'adaptation qui est présenté

¹ *BFM - Base de Français Médiéval* [En ligne]. Lyon : ENS de Lyon, Laboratoire ICAR, 2012, <http://bfm.ens-lyon.fr>.

² Corpus MCVF annoté xml, sous la direction de France Martineau, avec Paul Hirschbühler, Serge Lusignan, Christiane-Marchello-Nizia, Yves Charles Morin et François Rouget.

ici. Nous nous focaliserons sur la période préclassique (environ 1500–1650), qui est la plus problématique, mais montrerons que cette approche est aussi valable pour la période classique (environ 1651–1800).

2 Le français préclassique

2.1 Caractéristiques

La langue du XVI^e est une langue de transition entre la langue à forte variation graphique du Moyen-Âge, sans norme, avec des marquages régionaux, et une langue normalisée, le français classique, qui n'est pas encore la langue moderne que nous connaissons. Au cours du siècle, les graphies évoluent, sans doute sous l'influence de l'impression : les imprimeurs proposent de nouvelles règles typographiques, par exemple les diacritiques. Les textes de nos corpus sont transmis à travers des éditeurs scientifiques, qui soit respectent le texte (édition diplomatique), soit modernisent partiellement (*u/v*, *il/j*) ou entièrement le texte. Lorsque c'est possible, nous privilégions les éditions les plus proches du texte original, c'est à dire conservant les variantes graphiques, typographie exceptée : le corpus ne garde ainsi pas trace des caractères désuets comme le «s» long.

Pour se rendre compte de l'ampleur de cette variation graphique, on peut se référer aux attestations du lemme « *fruit* » dans le corpus Frantext pour la période 1501-1650 : *frui* (24 attestations), *fruit* (1150), *fruits* (767), *fruitz* (73), *fruis* (2), *fruit* (697), *fruits* (431), *fruitz* (43), *fruyct* (10), *fruycts* (1), *fruyctz* (9), *fruyt* (10). Les traitements vont devoir tenir compte de cette variabilité graphique.

2.2 Constitution d'un corpus et d'un jeu d'étiquettes

Le corpus diachronique du français que nous développons couvre actuellement la période XVI^e–XX^e siècles. Pour la période XVI^e–XVIII^e siècles (périodes préclassique et classique), nous disposons de 189 textes issus de Frantext, des Bibliothèques Virtuelles Humanistes (base Epistémon), des Corpus Électroniques de la Première Modernité et de l'*American and French Research on the Treasury of the French Language*. Comme beaucoup de textes littéraires, même anciens, la plupart d'entre eux sont malheureusement sous droits d'éditeur et ne peuvent pas être redistribués librement. Toutefois, 37 textes (soit environ 2 millions de mots qui constituent le *corpus noyau*, couvrant plusieurs genres littéraires sur toute la période) pourront être diffusés sous licence *Creative Commons* à l'issue du projet.

Le jeu d'étiquettes morpho-syntaxiques est adapté au caractère diachronique du corpus, et vise à simplifier l'annotation de ce dernier. En langue moderne il est parfois délicat d'assigner une forme à une catégorie, même en contexte. Les catégories participe, gérondif et adjectif sont ainsi parfois difficiles à distinguer en langue moderne ; en langue ancienne, c'est même souvent impossible pour un locuteur non spécialiste de la période considérée. Nous avons donc décidé d'adopter un jeu de catégories simple. D'une part, l'annotation s'en tient essentiellement aux parties de discours (verbe, substantif, etc.), et ne « déborde » pas, comme c'est souvent le cas, sur des informations flexionnelles (par exemple mode, temps, etc. pour les verbes). D'autre part, l'annotation introduit des catégories regroupant des cas souvent indécidables ; par exemple, nous avons décidé de créer une étiquette regroupant les adjectifs, participes et gérondifs lorsque leur distinction est difficile, ce qui bien sûr a aussi des conséquences sur la définition des paradigmes verbaux et des auxiliaires.

3 Création de ressources pour le français préclassique

La création de ressources pour les langues anciennes peut s'envisager dans le cadre de la création de ressources pour les langues peu dotées et non standardisées. Notre approche s'inscrit à la suite des travaux de (Sánchez-Marco & al., 2011) pour l'espagnol ancien. Elle consiste à annoter automatiquement un corpus d'apprentissage avec un analyseur pour la langue moderne, puis à corriger manuellement cette annotation. Ce corpus d'apprentissage est ensuite utilisé conjointement à un lexique « archaïsé » à l'aide de règles, pour construire un modèle de langage.

C'est assez dict pour ceste foy.
Quand sçavoir en vous s'assocye,
Monsieur Rien, l'on vous remercy
Du bien qu'avons aprins de vous.
Bazochiens, entendez tous :
Je veulx en triumpant arroy
Eslire et faire ung nouveau roy,
Comme il est coustume de faire ;
Pourtant chacun pense a l'affaire,
Autant les grandz que les petitz,
Et faire les preparatifz ;
Car, ainsi comme liberalle,
Je tendz a monstre generale
Qui, l'esté qui vient, sera faicte.
En honneur du triumphe et feste,
Ne faillez monstrier vos bons cueurs
Qui font de la vertu approche,
Tant que l'on dye par honneurs :
Vive l'excellente Bazoche !

FIGURE 1: Extrait de *Sottie pour le cry de la bazoche*, Anonyme, 1549

Dans notre cas, le corpus d'apprentissage, en graphie ancienne, est « modernisé » à l'aide du lexique archaïsé, *avant* son traitement avec l'analyseur moderne (cf. section 3.2.1), ce qui permet de réduire l'ampleur du travail de correction manuelle.

Ces ressources seront publiées sous licence *Creative Commons* à l'issue du projet.

3.1 Création du lexique

Le lexique associe un lemme (moderne quand il existe) et une partie du discours, avec les formes rencontrées dans le corpus ; on parle de lexique morphologique. La construction de ce lexique s'appuie principalement sur « l'archaïsation » d'un lexique moderne, mais incorpore aussi quelques éléments empruntés à un lexique du français médiéval, surtout utiles en début de période.

3.1.1 Ressources utilisées

Il existe des lexiques morphologiques pour le français moderne. Nous avons choisi le *Lefff* (Sagot, 2010) car il nous semblait mieux adapté à nos besoins initiaux dans sa version adaptée pour *Freeling*³ (en particulier en ce qui concerne les étiquettes, de type EAGLES). Nous avons aussi utilisé *Morphalou* (Romary & al., 2004) qui nous a semblé plus complet dans la nomenclature de lemmes⁴ et pour son appui sur le dictionnaire de référence qu'est le Trésor de la langue Française (TLF). Pour les états médiévaux de la langue, nous nous appuyons sur la nomenclature du Dictionnaire du Moyen Français (1330-1500) (DMF) et sur son lemmatiseur LGeRM (Souvay & Pierrel, 2009). Ce dernier a l'avantage de posséder deux lexiques morphologiques, un pour la période médiévale, et un plus adapté à la langue du XVII^e (appelé « mode » pour « moderne étendu »).

La validation du lexique morphologique va s'appuyer sur le corpus *Frantext*⁵ qui couvre tous les états du français.

3.1.2 Processus de construction

La construction du lexique va se dérouler en 4 grandes étapes qui constituent un cycle. Plusieurs cycles vont être nécessaires pour obtenir un lexique ayant un taux de couverture satisfaisant pour traiter les textes.

La première étape consiste à créer la nomenclature de lemmes et leur flexion moderne. Le lexique de départ *Lefff* est tout d'abord adapté à nos étiquettes morphosyntaxiques. Il est ensuite complété avec les lemmes manquants, pris dans *Morphalou* pour les modernes, et dans LGeRM pour les médiévaux. Une nomenclature complémentaire de lemmes est ajoutée pour couvrir les lemmes absents des deux dictionnaires de référence (TLF et DMF). Ces nouveaux lemmes sont détectés à la fin d'un cycle.

La deuxième étape consiste à archaïser le lexique. Des règles d'archaïsation sont utilisées pour produire la flexion et la variation spécifique à la langue du XVI^e siècle. Elles sont prises dans la base de connaissances du lemmatiseur LGeRM, qui couvre bien les états anciens du français. Afin de ne pas produire un lexique trop volumineux, certaines hypothèses de variations graphiques sont validées par attestation dans le corpus de référence (corpus à annoter et corpus *Frantext*). Les règles sont appliquées les unes à la suite des autres sur tous les mots du lexique. Une seule règle à la fois est appliquée sur un mot. Il faut donc itérer le processus au cas où plusieurs règles pourraient s'appliquer. Le nombre d'itérations est fixé à trois ; en effet, il paraît incertain d'appliquer plus de trois règles sur un mot (le « bruit » devient alors important), alors que le gain de couverture après deux itérations est déjà faible (cf. évaluation).

La troisième étape consiste à compléter le lexique en puisant automatiquement dans les ressources existantes. L'idée est de regarder si les mots absents du lexique mais présents dans les corpus textuels peuvent être analysés. Tous d'abord quelques règles de LGeRM trop générales pour être appliquées sans risque à l'étape deux sont testées (par exemple *an* → *en*, *ain* → *ein*). Ensuite les lexiques LGeRM médiéval et LGeRM moderne étendu sont utilisés, on prend sans vérifier les analyses proposées. Au départ du projet ils étaient les plus riches en termes de variantes graphiques.

³ <http://nlp.lsi.upc.edu/freeling>

⁴ Environ 95 000 lemmes pour *Morphalou* 2, contre par exemple environ 68 000 lemmes dans le *Lefff* 3.2, ou environ 51 000 lemmes dans la version *Freeling* du *Lefff*, si l'on excepte à chaque fois les noms propres, que nous traitons à part.

⁵ 4 515 textes, 270 millions de mots.

La quatrième étape consiste à analyser les résultats pour détecter les règles manquantes et les lemmes absents de la nomenclature. Le lemmatiseur LGeRM est alors configuré pour proposer des hypothèses, qui sont évaluées manuellement. En effet, ce processus ne peut être fait automatiquement car il produirait trop de bruit, en tout cas sur les premiers cycles de la construction du lexique.

3.2 Création du corpus d'apprentissage

Le corpus d'apprentissage comporte environ 62 000 mots, issus de 5 textes, échelonnés entre 1547 et 1776, pour couvrir aussi la période classique. L'annotation de ce corpus s'est effectuée en plusieurs étapes : une initialisation (« *bootstrapping* ») automatique avec des ressources modernes, puis une correction manuelle.

3.2.1 Initialisation du corpus d'apprentissage : projection lexicale et désambiguïsation

Dans un premier temps, le lexique a simplement été « projeté » sur le corpus. Du fait de la grande variété des formes du corpus, cela impliquait une ambiguïté très importante : à chaque forme pouvaient correspondre un grand nombre de parties du discours et de lemmes. Afin de réduire cette ambiguïté, le corpus a ensuite été « modernisé » à l'aide d'une variante du lexique morphologique, c'est à dire qu'à chaque forme ancienne a été associée une forme moderne. Puis il a été annoté automatiquement, sur la base de ces formes modernisées, à l'aide de TreeTagger (Schmid, 1994) et d'un modèle de langage spécifique, développé par Achim Strein sur un corpus de français moderne, en conservant, pour chaque token, toutes les annotations dont la probabilité dépassait 10 % selon TreeTagger. L'intersection entre l'analyse obtenue par projection lexicale et l'analyse obtenue par TreeTagger (en ne conservant que les analyses validées par les deux méthodes) a ainsi permis de réduire fortement l'ambiguïté, ramenant à 20,4 % le nombre de tokens ambigus (10 200 tokens, sur les 62 000 de départ).

3.2.2 Désambiguïsation et validation manuelles

Lors de l'étape suivante, trois annotateurs experts ont vérifié indépendamment l'annotation obtenue pour chacun des tokens désambiguïsés automatiquement, et annoté manuellement les 10 200 tokens qui ne l'avaient pas été. Généralement, il s'agissait alors seulement de sélectionner l'une des analyses parmi lesquelles il avait été impossible de trancher automatiquement.

Mot n°	Forme rencontrée	Variante de	Lemme Vall.	CG Validée	Constellation	Mode Valid.	V	NCM	JQua	NPro	NCF	VAux	NC	Autre	Inconnu
117	maintesfoys														
118	passee						passer(passer)	passé(passe)			passee(passe)				INC
119	vostre														INC
120	temps	temps	temps	NCM		VA/DS		temps(temps)							
121	avecques														INC
122	les	le	le	Autre		VA/DS								le(le)	
123	honorables	honorable	honorable	JQua		VA/DS			honora...						
124	Dames						damer(damer)				dame(dame)				
125	et	et	et	Autre		VA/DS								et(et)	
126	Damoyselles														INC
127				Autre		VA/DS								(.)	
128	leur													leur(leur)lu...	
129	en	en	en	Autre		VA/DS								en(en)	
130	faisans	faisan	faisan	NC		VA/DS								faisa...	
131	beaulx														INC
132	et	et	et	Autre		VA/DS								et(et)	
133	longs							long(long)	long(lo...						
134	narrez	narrer	narrer	V		VA/DS	narrer(narrer)								
135				Autre		VA/DS								(.)	
136	alors	alors	alors	Autre		VA/DS								alors(alors)	
137	que	que	que	Autre		VA/DS								que(que)	
138	estiez														INC
139	hors	hors	hors	Autre		VA/DS								hors(hors)	
140	de	de	de	Autre		VA/DS								de(de)	
141	propos	propos	propos	NCM		VA/DS	propos(propos)			longs narrez, alors que estiez - hors - de propos : dont estes bien					
142				Autre		VA/DS								(.)	
143	dont	dont	dont	Autre		VA/DS								dont(dont)	
144	estiez														INC
145	bien							bien(bien)	bien(bi...					bien(bien)	
146	dignes	digne	digne	JQua		VA/DS			digne(d...						
147	de	de	de	Autre		VA/DS								de(de)	
148	grande														
149	louange						louanger(louanger)		grand(...					gran...	
150														me(loua...	

Figure 2: Capture d'écran d'AnaLog, fenêtre de visualisation du croisement du texte et des ressources sur l'analyse de Pantagruel. À gauche (1), affichage des formes du corpus. Au centre (2), analyse retenue, validée automatiquement (VA/DS) ou manuellement (VM/DS). À droite (3), analyses suggérées pour les cas ambigus.

Pour ce travail d'amendement, de validation et de correction de corpus, nous avons utilisé AnaLog, un outil dédié à l'exploration humaniste des textes (Lay & Pincemin, 2010). Il propose des fonctionnalités de type « *fonctionnalités documentaires*⁶ » disponibles dans les outils d'analyse de données textuelles, en les généralisant et en les adaptant à un environnement de travail du type « linguistique sur corpus ». Ainsi, les index et pourcentages fournis peuvent porter sur toutes les informations disponibles (formes graphiques et variantes, séquences de catégories, ressources dictionnaires utilisées pour l'annotation, etc), et le concordancier permet d'afficher chacune de ces informations.

Par ailleurs, AnaLog a pour ambition de permettre l'exploration des textes en rendant possible leur annotation manuelle systématique : les concordances permettent de mettre au jour des types d'informations récurrentes (ou non, par contraste), et l'étude des données repose sur l'observation de ces récurrences et l'élaboration de catégories permettant d'en rendre compte : les étiquettes. Ces étiquettes sont dynamiques et peuvent être créées ou supprimées librement, par exemple pour créer des étiquettes de travail temporaires.

Ces fonctionnalités, disponibles à partir d'un corpus brut, le sont aussi en partant d'une pré-annotation, comme c'est le cas ici. Dans le cas où les formes rencontrées ne sont pas ambiguës, on peut les valider en une seule fois. À l'inverse, dans le cas où la ressource propose de multiples annotations, elles sont toutes visibles et l'on procède à une désambiguïsation manuelle, soit au fil du texte (Figure 2), soit en validant un choix pour toute une série de formes répondant à une requête *via* le concordancier.

Le corpus d'apprentissage a ainsi pu être corrigé et désambiguïsé manuellement, à l'aide d'AnaLog, par trois experts⁷. Ces derniers se sont avérés en désaccord dans 9 % des cas. Nous avons alors désambiguïsé automatiquement les cas les plus « évidents », notamment lorsque deux des trois annotateurs étaient d'accord, et lorsque la divergence entre annotateurs ne portait que sur les diacritiques du lemme. Les 5,7 % d'ambiguïtés restantes (3 534 tokens) ont ensuite été résolues manuellement par un expert, qui a « tranché », au cas par cas.

4 Analyse du corpus et évaluation

Le lexique et le corpus d'apprentissage ont ensuite été utilisés pour entraîner un modèle de langage spécifique.

4.1 Couverture lexicale

Nous évaluons la qualité du lexique en regardant son taux de couverture en termes de fréquence sur un corpus de référence (Figure 3). En l'occurrence, nous utilisons Frantext, qui permet d'évaluer la couverture sur toute la chronologie du français. En abscisse du graphique on trouve la tranche temporelle (50 ans) et le nombre de textes concernés. Globalement les taux de couverture sont bons à partir de XVe siècle, jusqu'à la période moderne. L'analyse des lacunes du lexique montre une forte proportion de noms propres et de mots étrangers (essentiellement les mots latins). En terme de graphie on remarque une forte proportion d'hapax. Il s'agit souvent de variantes exotiques ou d'erreurs (numérisation, rupture de mots, impression).

4.2 Évaluation de l'annotation

Chaque texte du corpus d'apprentissage a été découpé en trois parties, selon un ratio 8/1/1. La première partie servait pour l'apprentissage proprement dit, la seconde pour le développement, et la dernière pour l'évaluation. Les résultats de cette évaluation sont synthétisés dans le Tableau 1. Dans un premier temps, nous avons évalué l'exactitude de la projection lexicale seule, sans utiliser de modèle de langage. Dans les nombreux cas d'ambiguïté, une analyse était simplement tirée au sort. La précision obtenue, de 60 %, est évidemment très mauvaise. Nous avons ensuite évalué l'approche par modernisation du lexique, et désambiguïsation en utilisant TreeTagger avec un modèle de langage du français moderne (approche décrite dans la partie 3.2.1). La précision est alors meilleure (81,1 %), et finalement assez proche de la précision du modèle entraîné spécifiquement sur notre corpus d'apprentissage (cf. partie 3.2.2, précision 83,3 %). Il s'agit toutefois de résultats intermédiaires, obtenus à partir de ressources non finalisées. À l'issue d'un important travail sur la normalisation des textes, l'optimisation de la chaîne de traitement et l'adaptation du lexique, nous avons pu obtenir une précision de 94,3 %.

⁶ Pour reprendre la distinction faite par Hyperbase entre fonctions statistiques et documentaires.

⁷ Deux experts ont annoté tout le corpus d'apprentissage, et le dernier seulement la moitié.

Projection lexicale	Modernisation	Modèle spécifique intermédiaire	Modèle spécifique finalisé
60 %	81,1 %	83,3 %	94,3 %

Tableau 1: Précision obtenue en fonction de la méthode d'annotation.

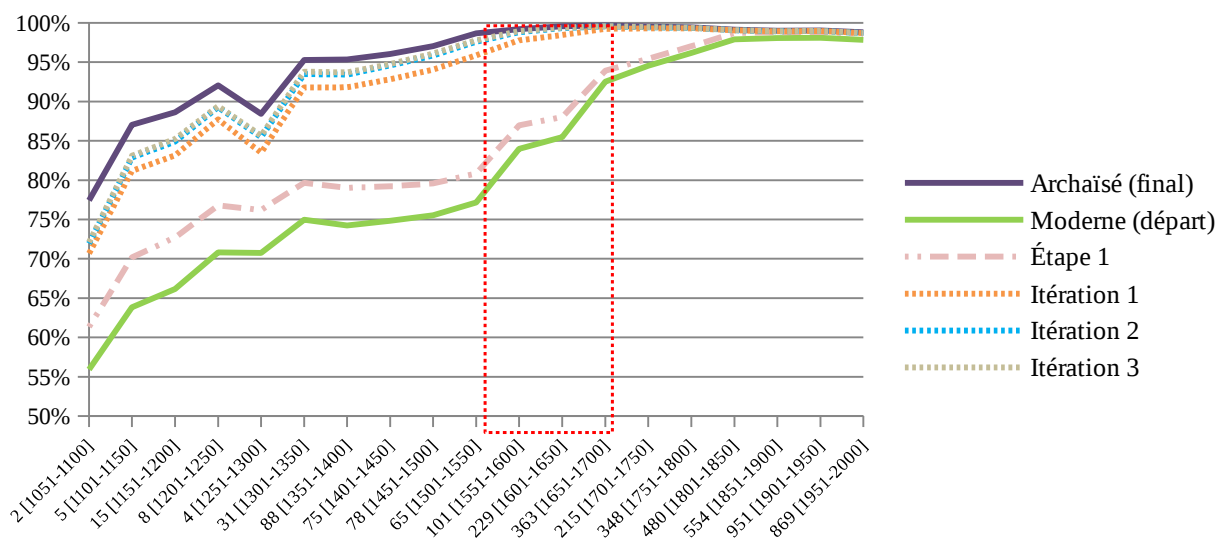


Figure 3: Taux de couverture lexicale du lexique moderne (lexique de départ) et du lexique archaïsé (lexique final), avec étapes intermédiaires, mesurée sur le corpus Frantext (XIème - XXème siècle).

Ces scores sont à mettre en regard de la relative simplicité du jeu d'étiquettes (cf. partie 2.2). De plus, notre évaluation concerne uniquement les étiquettes, et non les lemmes, qui ne sont actuellement pas désambiguïsés. Il reste donc encore une certaine marge de progression, avant de se rapprocher des scores obtenus pour le français moderne, aux environs de 96 % par exemple avec TreeTagger et le jeu d'étiquette GRACE, beaucoup plus complexe que le notre (Allauzen & Bonneau-Maynard, 2008).

Ces résultats sont homogènes sur la période préclassique et classique. Nous avons envisagé la création de modèles de langages distincts en fonction des périodes, mais n'avons pas constaté de gain significatif ; nous préférons donc nous en tenir à la simplicité d'un modèle unique, « panchronique », pour la période préclassique et classique.

5 Conclusion

Ce travail montre qu'il est tout à fait possible d'analyser des textes en français préclassique en adaptant des ressources conçues pour le français moderne. Les outils et les ressources développés dans ce cadre seront librement utilisables à l'issue du projet, ce qui contribuera à combler le manque entre la période médiévale et la période moderne.

Nous envisageons de poursuivre dans cette optique pour le traitement du français moderne, mais surtout du français médiéval. Il est certain que cette approche donnera des résultats de plus en plus dégradés au fur et à mesure que l'on remontera le temps, mais dans quelle mesure ? Enfin, au-delà de la création de ressources, nous envisageons une étude en diachronie longue de la langue française, notamment en adaptant certaines méthodes de *clustering* pour le travail en diachronie. Nous prévoyons aussi une évaluation de cette approche avec d'autres analyseurs : MELt, Morfette, etc.

Remerciements

Ce travail est issu du projet Presto, cofinancé par l'Agence Nationale de la Recherche et la Deutsche Forschungsgemeinschaft.

Références

- ALLAUZEN A., BONNEAU-MAYNARD H. (2008). Training and evaluation of POS taggers on the French MULTITAG corpus. Actes de *International Conference on Language Resources and Evaluation*, Marrakesh, Maroc.
- GLEßGEN M.-D., VACHON C. (2010). *Répertoire bibliographique du Nouveau Corpus d'Amsterdam, établi par Anthonij Dees et Piet Van Reenen (Amsterdam 1987), revu et élargi par M.-D.G. et C.V.*, 3. ed., Stuttgart: Institut für Linguistik/Romanistik.
- LAY M.-H., PINCEMIN B. (2010). Pour une exploration humaniste des textes : AnaLog. Actes des *Journées internationales d'Analyse statistique des Données Textuelles*, Rome, Italie.
- ROMARY L., SALMON-ALT S., FRANCOPOULO G. (2004). Standards going concrete : from LMF to Morphalou. *Workshop on Electronic Dictionaries, Coling 2004*, Genève, Suisse.
- SAGOT B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC 2010)*, Istanbul, Turquie.
- SÁNCHEZ-MARCO CRISTINA, BOLEDA GEMMA, PADRÓ LLUÍS (2011). Extending the tool, or how to annotate historical language varieties. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 1–9, Portland, Oregon, États-Unis.
- SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees, actes de *International Conference on New Methods in Language Processing*, Manchester, Royaume-Uni.
- SOUVAY G., PIERREL J.-M. (2009). LGeRM : lemmatisation de mots en moyen français. *Traitement Automatique des Langues*, 50-2.

Classification de texte enrichie à l'aide de motifs séquentiels

Pierre Holat Nadi Tomeh Thierry Charnois

Université Paris 13, Sorbonne Paris Cité, CNRS, LIPN UMR7030, 93430, France

prenom.nom@lipn.univ-paris13.fr

Résumé. En classification de textes, la plupart des méthodes fondées sur des classifieurs statistiques utilisent des mots, ou des combinaisons de mots contigus, comme descripteurs. Si l'on veut prendre en compte plus d'informations le nombre de descripteurs non contigus augmente exponentiellement. Pour pallier à cette croissance, la fouille de motifs séquentiels permet d'extraire, de façon efficace, un nombre réduit de descripteurs qui sont à la fois fréquents et pertinents grâce à l'utilisation de contraintes. Dans ce papier, nous comparons l'utilisation de motifs fréquents sous contraintes et l'utilisation de motifs δ -libres, comme descripteurs. Nous montrons les avantages et inconvénients de chaque type de motif.

Abstract.

Sequential pattern mining for text classification

Most methods in text classification rely on contiguous sequences of words as features. Indeed, if we want to take non-contiguous (gappy) patterns into account, the number of features increases exponentially with the size of the text. Furthermore, most of these patterns will be mere noise. To overcome both issues, sequential pattern mining can be used to efficiently extract a smaller number of relevant, non-contiguous, features. In this paper, we compare the use of constrained frequent pattern mining and δ -free patterns as features for text classification. We show experimentally the advantages and disadvantages of each type of patterns.

Mots-clés : Fouille de séquences, motifs libres, classification de texte, sélection de descripteurs.

Keywords: Sequence mining, free patterns, text classification, feature selection.

1 Introduction

La classification de séquences est une tâche importante dans beaucoup d'applications où l'information est structurée en séquences (Xing *et al.*, 2010), comme par exemple en biologie pour la classification d'ADN ou de séquences de protéines, et naturellement en traitement automatique des langues où la tâche de classification est un problème classique.

La sélection de descripteurs (Liu & Motoda, 2007), étape importante dans beaucoup d'approches de classification utilisant une représentation des données basée sur les descripteurs, est un problème important. Une approche simple et efficace consiste à considérer chaque mot d'une séquence comme un descripteur. Cependant la nature séquentielle et les dépendances entre les mots d'une phrase sont perdues. Pour capturer ces informations, une stratégie consisterait à générer toutes les sous-séquences possibles de mots au lieu de prendre chaque mot individuellement. Cependant, le nombre des sous-séquences produites qui croît exponentiellement en la taille des séquences rend l'apprentissage difficile et génère trop de bruit et de paramètres à prendre en considération.

Une solution est l'exploitation de techniques de fouille de séquences qui offrent l'avantage de parcourir efficacement l'espace complet des sous-séquences. La fouille de séquences est une des tâches les plus étudiées et les plus complexes en fouille de donnée. Depuis son introduction (Agrawal & Srikant, 1995), beaucoup de chercheurs ont développé des approches pour fouiller les séquences dans de nombreux et différents domaines comme la bio-informatique, le marketing, l'analyses des logs web mais aussi dans la construction de modèles globaux pour la classification (Knobbe *et al.*, 2008).

Dans cet article, nous étudions les performances de différents types de motifs séquentiels utilisés en tant que descripteurs pour la classification de textes, notamment les motifs fréquents sous différentes contraintes (Srikant & Agrawal, 1996) qui permettent de réduire le nombre de motifs extraits, ainsi que les motifs séquentiels δ -libres introduits récemment par (Holat *et al.*, 2014) pour réduire encore le nombre de descripteurs tout en obtenant des résultats compétitifs en terme de F-

mesure pour la classification. Cet article est structuré comme suit. Les techniques d'extraction de différents types de motifs séquentiels sont présentées en 2, et le processus de sélection de descripteurs en section 3. Différentes expérimentations menées en classification sont détaillées en section 4 montrant l'intérêt de ce type d'approche.

2 Extraction de motifs séquentiels

2.1 Définition formelle

Nous introduisons les notions de fouille les plus utiles ici (Agrawal *et al.*, 1993).

Soit $\mathcal{I} = \{i_1, i_2, \dots, i_n\}$ un ensemble fini de littéraux appelés *items*. Une séquence S sur \mathcal{I} est une liste ordonnée $\langle i_1, \dots, i_k \rangle$ non vide, où les i_j sont des items de \mathcal{I} et $j = 1 \dots k$. Un motif séquentiel est simplement une séquence. Une séquence $s_a = \{a_1, a_2, \dots, a_n\}$ est incluse dans une autre séquence $s_b = \{b_1, b_2, \dots, b_n\}$ s'il existe des entiers $1 \leq i_1 \leq i_2 \leq \dots \leq i_n \leq m$ tels que $a_1 = b_{i_1}, a_2 = b_{i_2}, \dots, a_n = b_{i_n}$. Si la séquence s_a est incluse dans s_b , alors s_a est une sous-séquence de s_b et s_b est une super-séquence de s_a , noté $s_a \preceq s_b$.

Une base de séquences SDB est un ensemble de paires (SID, S) où SID est un identifiant de séquence et S est une séquence. Une paire (SID, S) contient une séquence s si S contient au moins une occurrence de s . Le support d'une séquence s dans une base de séquences SDB , noté $Support(s, SDB)$, est défini comme : $Support(s, SDB) = |\{(SID, S) \in SDB | s \preceq S\}|$.

Le problème de la fouille de motifs séquentiels est l'extraction de toutes les séquences s in SDB ayant un support supérieur ou égal à un support minimal σ donné par l'utilisateur. Il existe des propriétés permettant de réduire drastiquement l'espace de recherche. Notamment l'anti-monotonie qui est une propriété centrale pour la construction d'algorithmes efficaces d'extraction de motifs.

Propriété 1 : Anti-monotonie, (Agrawal *et al.*, 1993) :

Soit s' et s deux séquences. Si $s' \preceq s$ alors $Support(s') \geq Support(s)$.

Propriété 2 (conséquence de la propriété 1) :

Soit s' une séquence non fréquente. Quelle que soit s telle que $s' \preceq s$, s est une séquence non fréquente.

Grâce à ces propriétés les premières approches d'extractions de motifs ont adopté une méthode "générer-élaguer". Si un motif n'est pas fréquent, il n'est pas nécessaire de générer les motifs l'incluant puisque ceux-ci ne seront pas fréquents. Cette méthode de parcours en largeur, appelé *Apriori*, est la fondation d'algorithmes d'extractions de motifs comme GSP (Srikant & Agrawal, 1996) et SPADE (Zaki, 2001). Une alternative à Apriori est le paradigme "Frequent Pattern Growth". C'est un parcours en profondeur des motifs qui évite l'étape coûteuse de génération des candidats. On citera notamment PSP (Masseglia *et al.*, 1998), PrefixSpan (Pei *et al.*, 2001) et SPAM (Ayres *et al.*, 2002).

2.2 Contraintes

L'extraction de motifs fréquents pose encore aujourd'hui un problème quant à l'utilité des motifs fréquents extraits. Selon les paramètres utilisés, les résultats peuvent être trop génériques ou être trop nombreux pour pouvoir être traités. Les contraintes introduites par (Srikant & Agrawal, 1996) sont un paradigme puissant pour cibler les motifs pertinents (Pei *et al.*, 2007). Nous allons ici reprendre deux contraintes intéressantes pour notre étude : la contrainte de longueur (qui définit la taille minimale et maximale d'un motif) et le gap (qui définit l'écart minimal et maximal – en nombre d'items de la séquence – entre deux items d'un motif séquentiel).

Malgré l'utilisation des contraintes, le nombre de motifs peut cependant être encore important. Une approche complémentaire consiste à utiliser une représentation condensée (un sous-ensemble) des motifs (Mannila & Toivonen, 1996). Un motif de support σ est dit clos (respectivement libre) si toutes ses sous-séquences (respectivement super-séquences) de support σ ont été élaguées. Un motif δ -libre est un motif libre avec une tolérance plus ou moins δ sur le support permettant de regrouper, donc de réduire, le nombre de motifs.

Les premiers travaux sur les représentations condensées ont été introduits par (Mannila & Toivonen, 1996). Depuis, la plupart des travaux portent sur les motifs ensemblistes non séquentiels, principalement parce qu'il existe des relations fortes entre les motifs ensemblistes et de puissants outils mathématiques comme la théorie des ensembles, la combinatoire et les correspondances de Galois. Ces outils jouent un rôle important dans la construction des représentations condensées

TABLE 1 – Détails du corpus Deft08 avant et après pré-traitement. La longueur fait référence au nombre de mots d'un document. La longueur minimum étant la longueur du plus court document du corpus. La longueur maximum étant la longueur du plus long document du corpus. La longueur moyenne/médiane étant la moyenne et la médiane de la longueur des documents d'un corpus.

Corpus	# documents	# mots	# mots distincts	Long. min	Long. max	Long. moy./méd.
Apprentissage	15.223	6.639.409	185.481	47	14.025	436 / 263
Test	10.596	4.725.358	146.183	17	14.271	446 / 264
App. pré-traité	15.223	3.375.888	161.622	21	6.950	222 / 135
Test pré-traité	10.596	2.306.471	128.377	10	6.779	218 / 132

fondées sur les motifs clos (Pasquier *et al.*, 1999), les motifs essentiels (Casali *et al.*, 2005), les motifs δ -libres (Boulicaut *et al.*, 2003), également appelés clés, générateurs ou libres, dans le cas particulier où $\delta = 0$, et les motifs non-dérivables (Calders & Goethals, 2002).

Les motifs clos ont été étendus aux séquences depuis plusieurs années et des expérimentations en classification ont déjà été réalisées (Kim *et al.*, 2012). Cependant, l'extension aux motifs séquentiels δ -libres (Holat *et al.*, 2014) est relativement récente. Notre objectif est ici d'étudier l'intérêt de ces motifs séquentiels δ -libres par rapport à une approche d'extraction de motifs fréquents sous les différentes contraintes vues précédemment.

3 Motifs et sélection de descripteurs

Un corpus D contient un nombre $|D|$ de documents d . Chaque document d appartient à une classe $c \in C$ et est composé de mots appartenant à un vocabulaire V . Un motif séquentiel extrait m servira de descripteur à un document d selon la fonction caractéristique :

$$f_m(d, c) = \begin{cases} 1 & \text{if } m \preceq d \text{ et } d \in c \\ 0 & \text{sinon} \end{cases}$$

L'ensemble des descripteurs d'un document d formera donc une représentation du document, et sera utilisé par un classifieur statistique pour calculer la probabilité d'un document à appartenir à une classe. Nous avons utilisé le classifieur Wapiti¹ de (Lavergne *et al.*, 2010), une implémentation du modèle Maximum Entropy (MaxEnt). Comme score de classification, nous utilisons la mesure populaire qui combine la précision et le rappel, la F-mesure. Notre approche consiste à utiliser un corpus D dont nous faisons varier le vocabulaire V en ajoutant plusieurs couches d'information pour chaque mot. Nous allons maintenant en voir les détails.

4 Expérimentations

4.1 Données

Les différentes expérimentations sont réalisées sur le jeu de données DEFT'2008² composé d'articles du journal "Le Monde" et d'articles de l'encyclopédie libre "Wikipédia". L'ensemble des classes possibles pour chaque document est $C = \{\text{Art, Économie, Sport, Télévision}\}$.

Le pré-traitement du corpus avant l'étape d'extraction de motifs a consisté à réaliser un étiquetage morpho-syntaxique des mots grâce à l'outil TreeTagger³ de (Schmid, 1994), puis à retirer toute la ponctuation, les mots-vides (mots beaucoup trop communs comme "le", "la", ...) ainsi que de mettre toutes les lettres en minuscule. Les détails du corpus sont en Table 1. Cette étape a réduit la quantité de mots de moitié, notamment en longueur des séquences permettant des étapes d'extractions de motifs beaucoup plus rapides.

1. <http://wapiti.limsi.fr/>

2. <https://deft.limsi.fr/2008/corpus-desc.php>

3. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

TABLE 2 – Meilleurs résultats de classification pour chaque approche. σ est le paramètre de support minimal d'un motif. δ est le paramètre de δ -liberté, la compression de la représentation condensée. La première ligne rappelle le meilleur score de la campagne d'évaluation de DEFT'08 (Charton *et al.*, 2008) qui, pour résumer, utilise une fusion par vote ternaire des résultats de trois classifieurs : un SVM, un Bayésien Naïf et Icsiboost. Notons que les motifs séquentiels obtiennent des scores équivalents voire légèrement supérieurs.

Type vocabulaire	Type descripteur	σ	δ	gap min	gap max	long. min	long. max	F1	# descr.
E. Charton (LIA)	3-gram	-	-	-	-	-	-	87,5	-
Mot_i	Unigramme	-	-	-	-	-	-	86,3	161.622
	Motifs Fréquents	5	-	1	4	1	3	88,8	238.849
	Motifs δ -Libres	0.05%	10%	1	inf.	1	inf.	87,0	6.652
$Mot_i_POS_i$	Unigramme	-	-	-	-	-	-	86,8	186.698
	Motifs Fréquents	5	-	1	5	1	4	88,8	358.042
	Motifs Libres	0.05%	10%	1	inf.	1	inf.	86,7	6.732
$Mot_i_POS_i_POS_{i-1}$	Unigramme	-	-	-	-	-	-	83,8	448.904
	Motifs Fréquents	5	-	1	5	1	3	85,3	119.150
	Motifs Libres	0.05%	50%	1	inf.	1	inf.	84,0	6.067

4.2 Comparaison des différents types de motifs

A partir du corpus pré-traité nous avons généré quatre corpus différents pour ajouter de l'information dans le corpus. En effet, puisque l'extraction de motifs permet de retourner les motifs considérés comme les plus intéressants parmi toutes les combinaisons de "mots" possibles, nous avons donc ajouté plus d'informations dans les données avant de lancer le processus d'extraction. L'utilisation des catégories morpho-syntaxiques des mots est une technique répandue en traitement des langues naturelles. Nous avons donc modifié le vocabulaire du corpus par des expressions plus évoluées utilisant ces deux principes. Une première approche a consisté à ajouter à chaque mot sa catégorie morpho-syntaxique ($Mot_i_POS_i$). Ensuite nous avons généré un corpus dans lequel nous avons ajouté à chaque mot sa catégorie morpho-syntaxique et la catégorie du mot précédent ($Mot_i_POS_i_POS_{i-1}$). Un exemple pour chaque approche est disponible en Figure 1.

Corpus d'entrée : La peinture est une poésie muette !
Corpus Mot_i : peinture est poésie muette
Corpus $Mot_i_POS_i$: peinture_NOM est_VER poésie_NOM muette_ADJ
Corpus de $Mot_i_POS_i_POS_{i-1}$: peinture_NOM est_VER_NOM poésie_NOM_VER muette_ADJ_NOM

FIGURE 1 – Exemple de séquence pour chaque type de corpus

Pour chaque Corpus, nous avons effectué $2 * |C|$ extractions de motifs. Pour rappel, $|C|$ est le nombre de classe d'un corpus. Faire une extraction sur les séquences d'une classe c uniquement nous permet de récupérer un ensemble de motifs plus pertinent pour cette classe. Ces $|C|$ ensembles de motifs vont servir, pour l'apprentissage, à définir les descripteurs des documents de leur classe respective selon la fonction caractéristique vue en Section 3.

Les deux types d'extractions sont l'extraction de motifs fréquents et de motifs δ -libres en faisant varier leurs paramètres respectifs vus en section 2. Une vue d'ensemble des résultats de ces expérimentations est en Figures 2, 3 et 4. Les meilleurs résultats pour chaque type de vocabulaire sont donnés en tableaux 2 et 3.

En Figure 2 et 3, sont montrés les résultats de classification pour les motifs fréquents. Il est évident que les contraintes jouent un rôle important. Cependant leur impact est dépendant des données et il est donc nécessaire de trouver le bon paramétrage sous peine de voir les performances se dégrader, voire s'effondrer. La Figure 2 montre que, pour ce corpus, le meilleur *Gap* maximum est de 4 ou 5, et la figure 3 montre que la meilleure *Longueur* maximum est de 2 ou 3. Une explication probable est que, dans le texte, l'information est plutôt locale, il est rare qu'elle soit étendue sur des centaines de mots. Si les motifs sont trop longs, et surtout si le gap entre chaque mots est important, il est très probable que cela génère une information erronée puisque dénuée de sens. Avec un gap de 5 et une longueur de 4, un motif couvrira au plus une séquence de 19 mots, pour ce corpus c'est le meilleur rapport de couverture sans trop engendrer de bruit.

TABLE 3 – Meilleurs résultats en combinant les différents types de vocabulaire. Le vocabulaire de la "Combinaison par paramètres" est l'union de chaque type de vocabulaire pour un même type de descripteur, avec les mêmes paramètres d'extraction de motifs. Le vocabulaire de la "Combinaison des meilleurs" est l'union de chaque type de vocabulaire pour un même type de descripteur, avec les paramètres donnant le meilleur Fscore pour chaque type de descripteur.

Type vocabulaire	Type descripteur	σ	δ	gap min	gap max	long. min	long. max	F1	# descr.
Combi. par param.	Unigramme	-	-	-	-	-	-	85,5	790.555
	Motifs Fréquents	5	-	1	5	1	4	88,0	882.357
	Motifs Libres	0.05%	10%	1	inf.	1	inf.	87,5	19.015
Combi. meilleurs	Motifs Fréquents	5	-	-	-	-	-	87,3	715.927
	Motifs Libres	0.05%	-	-	-	-	-	87,0	19.003

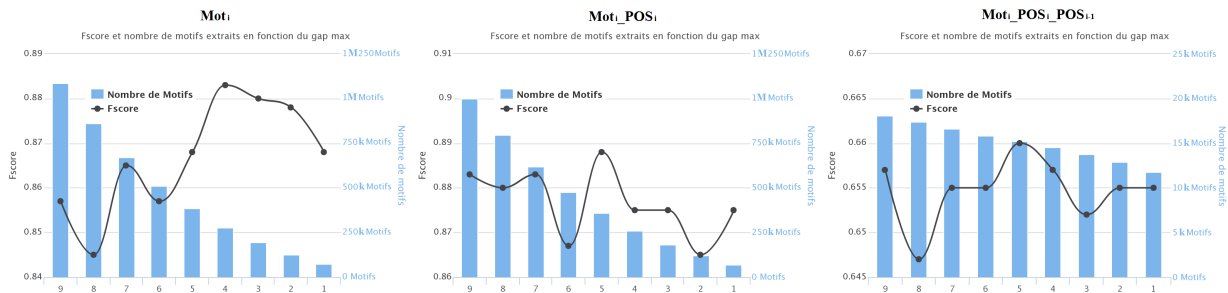


FIGURE 2 – Motifs fréquents : Impact du Gap maximum des motifs sur la taille du vocabulaire et le score de classification. Le support minimal est fixé à 5. La longueur maximum est fixé à 4.

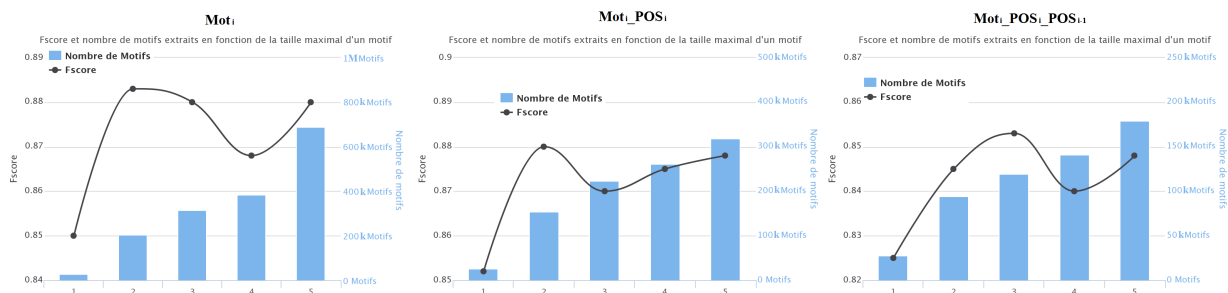


FIGURE 3 – Motifs fréquents : Impact de la longueur maximum des motifs sur la taille du vocabulaire et le score de classification. Le support minimal est fixé à 5. Le Gap maximum est fixé à 4.

Pour les motifs δ -libres, il n'y a pas de contrainte de gap ou de longueur possible (l'outil ne permet pas encore la prise en compte de ces contraintes), mais la contrainte de δ -liberté qui permet d'amplifier la compression de la représentation. En Figure 4, on peut voir que plus le δ est élevé, plus le nombre de motifs extraits sera réduit et plus le score de classification sera bon. Mais comme pour les motifs fréquents, l'effet du paramètre δ sur le score est sujet à variation. Les différentes expérimentations montrent qu'avec un δ trop faible la réduction du nombre de motifs est négative pour la classification. En effet, en augmentant δ on englobe plus de motifs dans une même classe d'équivalence. Cela supprime donc plus de super-motifs dans chaque classe d'équivalence puisque seuls les motifs libres seront retournés. Avec un δ trop faible, on perd donc de l'information potentielle en élaguant ces motifs, et comme il reste toujours trop de super-motifs, ceux qui sont supposés contenir l'information condensée ne sont pas mis en valeur. Mais on remarque que passé un certain niveau de δ , cet élagage devient favorable pour la classification. Une explication probable est que toute l'information spécialisée des motifs de grandes tailles, se retrouve entièrement condensée dans les motifs δ -libres. Il y a donc beaucoup moins de bruit dans les données, chaque motif restant contient l'information de ses super-motifs élagués, améliorant le score de classification avec beaucoup moins de motifs, jusqu'à 35 fois moins de motifs que les fréquents.

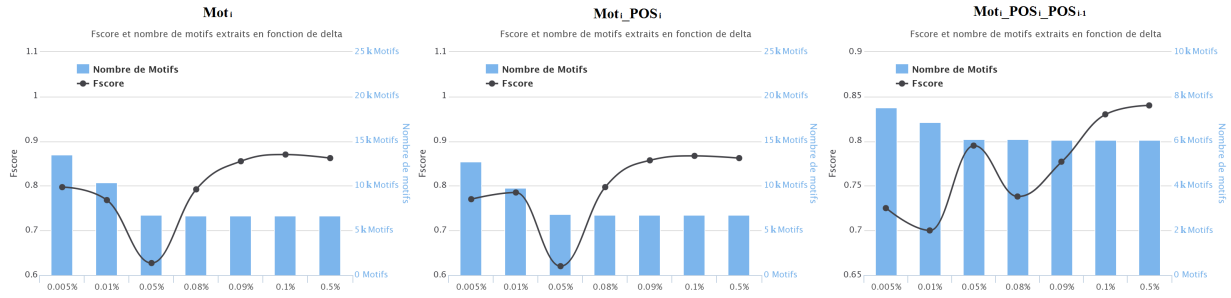


FIGURE 4 – Motifs δ -libres : Impact de la δ -liberté sur la taille du vocabulaire et le score de classification. Le support minimal est fixé à 0.05%

Pour résumer, l'utilisation de motifs séquentiels comme descripteurs a permis d'utiliser un vocabulaire plus évolué de manière efficace. Du point de vue de la F-mesure, ce sont les motifs fréquents sous contrainte de *Gap* et de *Longueur* qui donnent les meilleurs résultats. Notons cependant que la F-mesure n'est améliorée que d'un point par rapport à la F-mesure obtenue avec des δ -libres. En revanche, le nombre de motifs fréquents est 35 fois plus élevé que le nombre de motifs δ -libres. Ces derniers mettent bien en valeur la notion de représentation condensée sans perte d'information.

5 Conclusion

Nous avons étudié l'utilisation de différents types de motifs séquentiels en tant que descripteurs de classifieurs statistiques : les motifs fréquents sous différentes contraintes comme le *gap* et la *longueur*, et les motifs δ -libres qui sont une représentation condensée des motifs fréquents non-contraints. Ces approches permettent de prendre en compte beaucoup plus d'informations issues des données d'apprentissage que l'usage de sac-de-mots ou de n-grammes. Les motifs fréquents ont le meilleur score de classification mais nécessitent un paramétrage très fin, alors que les libres ont un score équivalent avec un paramétrage plus simple et un nombre de descripteurs jusqu'à 35 fois moins nombreux. L'ajout de certaines informations dans les données d'apprentissage, comme les catégories morphosyntaxiques, n'a pas été très concluant. On suppose que cela peut-être dû au corpus qui est d'une taille assez petite, il serait intéressant de poursuivre ce travail sur des corpus beaucoup plus conséquents où l'impact de ces informations supplémentaires serait mis plus en valeur. Une autre perspective est de continuer dans cette lancée et d'ajouter encore plus d'informations, mais sous forme d'itemsets dans les motifs séquentiels. La complexité de cette alternative étant beaucoup plus élevée, une approche possible serait de combiner les contraintes de *Gap* et de *Longueur* avec la δ -liberté.

Remerciements

Nous remercions les relecteurs pour leurs conseils avisés. Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'Avenir portant la référence ANR-10-LABX-0083 et du projet Hybride ANR-11-BS02-002.

Références

- AGRAWAL R., IMIELIŃSKI T. & SWAMI A. (1993). Mining Association Rules Between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93.
- AGRAWAL R. & SRIKANT R. (1995). Mining Sequential Patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, p. 3–14, Washington, DC, USA : IEEE Computer Society.
- AYRES J., FLANNICK J., GEHRKE J. & YIU T. (2002). Sequential Pattern Mining Using a Bitmap Representation. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, p. 429–435 : ACM.

- BOULICAUT J.-F., BYKOWSKI A. & RIGOTTI C. (2003). Free-Sets : A Condensed Representation of Boolean Data for the Approximation of Frequency Queries. *Data Mining and Knowledge Discovery*, **7**(1), 5–22.
- CALDERS T. & GOETHALS B. (2002). Mining All Non-derivable Frequent Itemsets. In T. ELOMAA, H. MANNILA & H. TOIVONEN, Eds., *Principles of Data Mining and Knowledge Discovery*, number 2431 in Lecture Notes in Computer Science, p. 74–86. Springer Berlin Heidelberg.
- CASALI A., CICCETTI R. & LAKHAL L. (2005). Essential Patterns : A Perfect Cover of Frequent Patterns. In A. M. TJOA & J. TRUJILLO, Eds., *Data Warehousing and Knowledge Discovery*, number 3589 in Lecture Notes in Computer Science, p. 428–437. Springer Berlin Heidelberg.
- CHARTON E., CAMELIN N., ACUNA-AGOST R., GOTAB P., LAVALLEY R., KESSLER R. & FERNANDEZ S. (2008). Pré-traitements classiques ou par analyse distributionnelle : application aux méthodes de classification automatique déployées pour deft08. In *Actes DEFT08-TALN'08*.
- HOLAT P., PLANTEVIT M., RAISSI C., TOMEH N., CHARNOIS T. & CREMILLEUX B. (2014). Sequence Classification Based on Delta-Free Sequential Patterns. In *2014 IEEE International Conference on Data Mining (ICDM)*, p. 170–179.
- KIM H. D., PARK D. H., LU Y. & ZHAI C. (2012). Enriching text representation with frequent pattern mining for probabilistic topic modeling. *Proceedings of the American Society for Information Science and Technology*, **49**(1).
- KNOBBE A., CRÉMILLEUX B., FÜRNKRANZ J. & SCHOLZ M. (2008). From local patterns to global models : The lego approach to data mining. In *ECML PKDD 2008 Workshop : From Local Patterns to Global Models*, p. 1–16.
- LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.
- LIU H. & MOTODA H. (2007). *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC.
- MANNILA H. & TOIVONEN H. (1996). Multiple uses of frequent sets and condensed representations (Extended Abstract). In *In Proc. KDD Int. Conf. Knowledge Discovery in Databases*, p. 189–194 : AAAI Press.
- MASSEGLIA F., CATHALA F. & PONCELET P. (1998). The PSP Approach for Mining Sequential Patterns. In *Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery, PKDD '98*.
- PASQUIER N., BASTIDE Y., TAOUIL R. & LAKHAL L. (1999). Discovering Frequent Closed Itemsets for Association Rules. p. 398–416.
- PEI J., HAN J., MORTAZAVI-ASL B., PINTO H., CHEN Q., DAYAL U. & HSU M.-C. (2001). PrefixSpan : Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. p. 215–224.
- PEI J., HAN J. & WANG W. (2007). Constraint-based sequential pattern mining : the pattern-growth methods. *Journal of Intelligent Information Systems*, **28**(2), 133–160.
- SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees.
- SRIKANT R. & AGRAWAL R. (1996). *Mining sequential patterns : Generalizations and performance improvements*.
- XING Z., PEI J. & KEOGH E. (2010). A brief survey on sequence classification. *SIGKDD Explor. Newsl.*, **12**(1), 40–48.
- ZAKI M. J. (2001). SPADE : An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning*, **42**(1-2).

Le traitement des collocations en génération de texte multilingue

Florie Lambrey François Lareau
OLST, Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal QC H3C 3J7, Canada

florie.lambrey@umontreal.ca francois.lareau@umontreal.ca

Résumé. Pour concevoir des générateurs automatiques de texte génériques qui soient facilement réutilisables d'une langue et d'une application à l'autre, il faut modéliser les principaux phénomènes linguistiques qu'on retrouve dans les langues en général. Un des phénomènes fondamentaux qui demeurent problématiques pour le TAL est celui des collocations, comme *grippe carabinée*, *peur bleue* ou *désir ardent*, où un sens (ici, l'intensité) ne s'exprime pas de la même façon selon l'unité lexicale qu'il modifie. Dans la lexicographie explicative et combinatoire, on modélise les collocations au moyen de fonctions lexicales qui correspondent à des patrons récurrents de collocations. Par exemple, les expressions mentionnées ici se décrivent au moyen de la fonction *Magn* : *Magn*(PEUR) = BLEUE, *Magn*(GRIPPE) = CARABINÉE, etc. Il existe des centaines de fonctions lexicales. Dans cet article, nous nous intéressons à l'implémentation d'un sous-ensemble de fonctions qui décrivent les verbes supports et certains types de modificateurs.

Abstract.

The treatment of collocations in multilingual text generation

In order to develop generic natural language generators that could be reused across languages and applications, it is necessary to model the core linguistic phenomena that one finds in language. One central phenomenon that remains problematic in NLP is collocations, such as *heavy rain*, *strong preference* or *intense flavour*, where the same idea (here, intensity), is expressed differently depending on the lexical unit it modifies. In explicative combinatory lexicography, collocations are modelled via lexical functions, which correspond to recurrent patterns of collocations. For instance, the three expressions above are all described with the function *Magn* : *Magn*(RAIN) = HEAVY, *Magn*(PREFERENCE) = STRONG, etc. There are hundreds of lexical functions. In this paper, we focus on the implementation of a subset of them that are used to model support verbs and some modifiers.

Mots-clés : génération automatique de texte multilingue ; collocations ; fonctions lexicales ; théorie sens-texte.

Keywords: multilingual natural language generation ; collocations ; lexical functions ; meaning-text theory.

1 Introduction

La génération automatique de texte (GAT) vise à présenter de façon automatique de l'information que l'on veut communiquer en langue naturelle. À ces fins, l'approche symbolique encode les connaissances linguistiques pertinentes dans des dictionnaires et grammaires, des ressources coûteuses à développer. La plupart des générateurs de texte sont monolingues. Si on veut pouvoir générer des textes dans plusieurs langues en même temps, il faut adapter les modèles linguistiques pour chaque nouvelle langue. Pour créer de tels systèmes de génération automatique de texte multilingue (GATM), il existe deux stratégies. La première consiste à transposer les ressources linguistiques vers la ou les langue(s) visée(s) (Alshawhi & Pulman, 1992; Kim *et al.*, 2003). La seconde consiste à factoriser ce qui peut l'être afin de partager certains éléments de description entre les langues (Avgustinova & Uszkoreit, 2000; Bateman *et al.*, 2005; Lareau & Wanner, 2007; Santaholma, 2008). Notre travail se situe dans la mouvance de cette seconde approche.

Les grammaires et les dictionnaires modélisent divers phénomènes, dont la lexicalisation (Wanner, 1996a; Polguère, 1997, 2000; Reiter & Dale, 2000). Dans le cadre de la GATM, le modèle de lexicalisation doit être conçu de manière à être aussi générique que possible afin de maximiser la quantité de règles communes aux langues à traiter. Pour ce faire, il faut chercher à modéliser les phénomènes linguistiques centraux partagés par plusieurs langues et s'en servir comme base sur

laquelle s’ajoutent les modules spécifiques à chaque langue. Alors que beaucoup de sens se lexicalisent toujours à l’aide d’un même mot dans une langue donnée, d’autres vont se réaliser différemment en fonction du contexte linguistique comme le sens de causation dans *donner le rhume*, *mettre dans le pétrin* ou *jeter un sort*. Ces dépendances lexicales se traduisent par des contraintes combinatoires dans ce qu’on appelle des collocations. Les collocations sont omniprésentes dans les textes naturels mais restent problématiques en traitement automatique des langues (TAL). Malgré leur apparente diversité, les collocations peuvent se généraliser et former des patrons communs à plusieurs langues. L’objectif de cet article est de décrire la méthodologie qui a permis de modéliser certains de ces patrons et de les intégrer dans un système générique de GATM. Ce travail, encore dans sa phase initial se situe dans le cadre du projet GÉCO (GÉNération de COLlocations).

2 Les collocations et les fonctions lexicales en GATM

Plusieurs projets ont porté sur la création et le partage de ressources linguistiques appliquées à la génération de texte multilingue ou à la traduction automatique, notamment Boguslavsky *et al.* (2004), Bateman *et al.* (2005) et Lareau & Wanner (2007). Le processus de lexicalisation est complexe en raison de l’interdépendance entre les choix lexicaux et syntaxiques (Steinlin, 2003). Cela se traduit par des contraintes combinatoires devant être décrites dans les ressources linguistiques. Les collocations sont une parfaite illustration de ces contraintes. Par exemple, *faire un pas*, *dar un paso* (litt. ‘donner un pas’), et *take a step* (litt. ‘prendre un pas’)instancient le même phénomène linguistique (l’emploi d’un verbe support) mais se réalisent différemment d’une langue à l’autre. Ce type d’information est crucial en GATM pour simuler au maximum le langage naturel.

Les collocations, comme *peur bleue*, *brouillard dense* et *remercier chaleureusement*, sont toujours composées d’une base et d’un collocatif, et le choix du collocatif est déterminé par celui de la base. Même si cette préférence lexicale est arbitraire, il est possible d’en tirer une généralisation sous forme de patron. En l’occurrence, dans les exemples donnés ici, on retrouve un même patron d’intensification, que l’on peut représenter sous forme de ratio : $\frac{\text{bleue}}{\text{peur}} = \frac{\text{dense}}{\text{brouillard}} = \frac{\text{chaleureusement}}{\text{remercier}}$. Ce genre de patron se retrouve à travers les langues et correspond à ce qu’on appelle des fonctions lexicales (FL) (Mel’čuk, 1995; Wanner, 1996b; Kahane & Polguère, 2001; Apresjan *et al.*, 2002); dans ce cas-ci, il s’agit de la fonction Magn. Dans notre système, les collocations sont décrites à l’aide de telles fonctions.

Les FL ont déjà été utilisées en GAT (monolingue et multilingue). Heid & Raab (1989), Apresjan *et al.* (2002) et Iordanskaja *et al.* (1996), entre autres, les utilisaient dans le cadre de modèles basés sur la théorie Sens-Texte (TST), alors que Steinlin (2003) les employait dans le cadre de la grammaire d’arbres adjoints (TAG), et Lareau *et al.* (2011, 2012) les implémentaient en grammaire lexicale fonctionnelle (LFG). Enfin, (Fontenelle, 1997) employait les FL en vue de la création d’une base de donnée sémantique-lexicale et faisait référence à d’autres travaux portant sur l’utilisation des FL pour la GAT et en TAL de manière générale (Van der Wouden, 1992; Iordanskaja *et al.*, 1996; Lee & Evens, 1996). Ces implémentations, toutefois, étaient partielles : elles traitaient un sous-ensemble de FL dans un domaine particulier. Un des objectifs du projet GÉCO est d’implémenter de façon exhaustive les FL afin de décrire les patrons de collocations récurrents en langue générale. Dès lors, l’encodage des FL s’oriente autour d’une structuration modulaire de notre système, nous amenant à opérer des regroupements de FL en patrons plus généraux. Le résultat de cette implémentation sera donc un outil répertoriant toutes ces FL et libre d’accès. Le choix d’utiliser les FL pour la GATM s’est imposé car, comme l’expliquent Lareau & Wanner (2007), elles permettent de réduire le nombre de règles spécifiques aux langues. Comme l’a montré Fontenelle (1997), les FL standards permettent d’encoder des distinctions sémantiques fines, comme celle entre les verbes anglais RAISE (=CausPredPlus(PRICE)) et RISE (=IncepPredPlus(PRICE)). Il arrive toutefois que les FL manquent de granularité ; par exemple, *recommander fortement* et *recommander chaleureusement* sont deux valeurs possibles de la même FL Magn(RECOMMANDER), même si leur sens n’est pas tout à fait identique. Par ailleurs, il n’existe pas de consensus solide sur l’inventaire des FL standards (Fontenelle, 1997; Jousse, 2003, 2010).

3 Le système GÉCO

Le système de GATM proposé ici se base sur la grammaire MARQUIS (Lareau & Wanner, 2007; Wanner *et al.*, 2010), implémentée sur la plateforme MATE. MATE est un transducteur de graphes libre conçu par Bohnet & Wanner (2010) et propose un environnement pour le développement de grammaires et dictionnaires. Conçu à la base pour modéliser l’approche multistratale de la TST, il peut manipuler divers types de graphes et construit, à partir d’un graphe source, un graphe correspondant d’un niveau supérieur, sans modifier la structure de départ. Le processus de génération s’articule autour de plusieurs niveaux : conceptuel, sémantique, syntaxique profond et de surface, morphologique profond et de

surface, et phonologique. Des règles de correspondance permettent de passer d'un niveau à l'autre. Étant donné que le travail présent se focalise sur l'étape de lexicalisation en vue de décrire les patrons de collocations récurrents, nous nous sommes focalisés sur l'interface entre la structure sémantique, un graphe orienté acyclique (GOA), et syntaxiques profonde et de surface, des arbres de dépendance non linéarisés. Le niveau sémantique représente le message que l'on souhaite communiquer alors que le niveau syntaxique prend en charge la structure de la phrase. Les règles de correspondance entre niveaux sont développées de manière tripartite incluant une structure source, une structure cible et un ensemble de conditions d'application de la règle. Le modèle est transductif et non transformationnel, c'est-à-dire qu'on ne modifie jamais la structure donnée en entrée, mais on construit une ou plusieurs nouvelle(s) structure(s) d'un niveau de représentation adjacent. La mise en correspondance entre les structures n'est pas déterministe : à une structure donnée en entrée correspondent en général plusieurs structures. En d'autres termes, le paraphrasage fait partie intégrante du système.

4 Implémentation des patrons de collocations

4.1 Sélection de patrons pertinents

La grammaire de MARQUIS (Lareau & Wanner, 2007) établit des règles de l'interface sémantique-syntaxe très générales et capables de traiter six langues européennes avec un faible nombre de règles spécifiques à une langue donnée. Ils ont montré que cette architecture basée sur des dictionnaires riches et des règles aussi génériques que possible accélérerait grandement leur travail. C'est dans cette optique d'optimisation que nous avons généralisé encore plus ces règles dans notre système de GATM à l'aide de patrons de collocations. La création de patrons généraux, des classes de FL partageant des propriétés similaires, s'insère dans une perspective de développement continu de notre plateforme et vise à faciliter l'implémentation de nouvelles FL en plus du maintien des structures existantes.

Les premiers patrons implémentés modélisent la lexicalisation des verbes supports et de certains modificateurs. Ces phénomènes sont très fréquents et ont un impact important sur la structure syntaxique de la phrase. Ils sont décrits au moyen de FL standard simples. Ce sont donc ces FL que nous avons modélisées pour le moment en accord avec les descriptions communément acceptées de ces FL (Mel'čuk *et al.*, 1995; Fontenelle, 1997; Mel'čuk, 2004). C'est pour cela que les regroupements établis jusqu'ici se basent essentiellement sur les propriétés syntaxiques des FL traitées. Par exemple, le patron des verbes supports correspond toujours à un collocatif verbal vide de sens et sa valence. D'autres patrons sont définis en fonction de la relation syntaxique entretenue entre la base et le collocatif ; c'est le cas du patron *modificateur*. Par ailleurs, nos patrons modélisent les informations d'ordre sémantique contenues dans les FL, ce qui se reflète à l'aide d'indices faisant référence aux actants de la base. Par exemple, A_2 retourne une expression adjectivale qui contient la base et qui modifie le deuxième actant de la base. En résumé, l'élaboration des patrons de collocations s'est faite dans l'optique de pouvoir faire des regroupements de comportements syntaxiques similaires. Chaque patron est incarné par une règle de lexicalisation dans MATE. Le projet GÉCO prévoit d'implémenter au fur et à mesure tous les types de relations décrits par les FL standard simples, complexes et non-standard. Les patrons implémentés sont présentés dans la table 1.

Le principal risque de cette approche est d'opérer des regroupements inappropriés de FL menant à un manque de granularité général du système. Par ailleurs, la formulation des règles ainsi que la structuration des dictionnaires auxquels elles font référence est contraint par l'environnement MATE. L'encodage des FL est également dépendant de cette plateforme.

4.2 Implémentation des verbes supports

Les verbes supports ont été bien décrits dans la TST, notamment par Wanner (1996b); Fontenelle (1997); Apresjan *et al.* (2002); Mel'čuk (2004) entre autres. Il existe plusieurs caractéristiques partagées par ce type de verbes, en particulier le fait qu'ils sont sémantiquement vides (dans leur usage en tant que collocatif). Un verbe support n'est donc pas présent dans la structure sémantique de l'énoncé, mais est construit par des règles de la grammaire afin de satisfaire les contraintes syntaxiques et lexicales de la langue. Il en existe trois types majeurs définis en fonction du rôle syntaxique de la base vis-à-vis du verbe support : *Oper* (la base est l'objet direct du collocatif), *Func* (la base est le sujet syntaxique) et *Labor* (la base est l'objet indirect). Ces FL peuvent posséder un indice qui donne la position syntaxique des actants de la base par rapport au verbe support¹. Ainsi, DONNER est l'*Oper*₁ de GIFLE car, d'une part, le sujet syntaxique du verbe support correspond au premier actant de la base et, d'autre part, la base est l'objet direct du verbe support (cf. Mel'čuk 2004).

1. Comme l'indique un relecteur, selon la définition standard des FL les indices font référence aux actants syntaxiques profonds du prédicat. Cependant, baser nos règles sur les actants sémantiques facilite le processus de création de règles.

Patron	Fonction lexicale	Exemple
Modificateur	Magn Ver	<i>peur bleue</i> <i>argument convaincant</i>
Verbe support	Func ₀ Func ₁ Func ₃ Oper ₁ Oper ₂ Oper ₁₂ Labor ₁₂	<i>la neige tombe</i> <i>la faute incombe à X</i> <i>la discussion porte sur Z</i> <i>X essuie un échec</i> <i>Y reçoit un coup</i> <i>X fait un compliment à Y</i> <i>X soumet Y à l'analyse</i>
Nom pléonastique	Figur	<i>rideau de fumée</i>
Préposition	Loc _{ad} Loc _{in}	<i>au front</i> <i>au sein du personnel</i>
Adjectivisation/Adverbialisation	A ₂ Adv ₁	<i>(Y est) couvert de mépris</i> <i>(X fait quelque chose) avec joie</i>

TABLE 1 – Patrons de fonctions lexicales implémentés

Dans la grammaire MARQUIS, à chaque FL décrivant un verbe support correspond une règle de lexicalisation prenant en compte le type de verbe support ainsi que le nombre et le type d'actants réalisés. Il y a en tout six règles pour les verbes supports dans MARQUIS : Func₀, Func_i, Oper₀, Oper_i, Oper_{ij} et Labor_{ij} (chacune de ces règles possède une illustration linguistique dans la table 1). La figure 1 illustre schématiquement la règle Oper_i qui permet de générer des énoncés comme *essuyer un échec*, *donner une baffe* ou *faire un pas*. La partie gauche contient deux nœuds, *X* et *A* ; *X* est un sens prédicatif et *A* est son *i*-ème argument, ce qui est représenté par la flèche partant de *X* vers *A*, symbolisant le fait que *X* domine sémantiquement *A*. Par exemple, pour *Paul essuie un échec*, le sens est 'échouer(Paul)', où *X*= 'échouer', *A*= 'Paul' et *i*=1 (puisque 'Paul' est le premier argument de 'échouer'). La partie droite de Oper_i donne la structure syntaxique profonde correspondante sous forme d'arbre de dépendance. Elle génère trois nœuds dans la structure syntaxique : Oper_i, *C* et "lex(*X*)". Le nœud *C* est la réalisation de *A* (ce que les pointillés indiquent). Cette règle se contente de créer dans la structure syntaxique le nœud correspondant à *A* dans la structure sémantique mais laisse le soin à une autre règle d'opérer sa lexicalisation. Le sens prédicatif (dans notre exemple, 'échouer') peut s'exprimer par le verbe ESSUYER accompagné du nom ÉCHEC ; il a donc deux nœuds correspondants en syntaxe : le verbe Oper_i et le nom qui va lexicaliser *X* ("lex(*X*)"). La lexicalisation de *X*, représentée ici par "lex(*X*)", est récupérée dans un dictionnaire sémantique contenant l'instruction "lex(échouer) = échec" (et potentiellement d'autres lexicalisations pour ce sens, comme ÉCHOUER). La valeur du collocatif est ensuite calculée lors d'une deuxième étape de lexicalisation, alors qu'elle sera récupérée dans un dictionnaire où on trouve l'information "Oper₁(échec) = essuyer" (et potentiellement d'autres collocatifs, comme SUBIR). La structure syntaxique est figée : la base est forcément l'actant syntaxique II (objet direct) et l'argument sémantique de la base est toujours l'actant I (sujet).

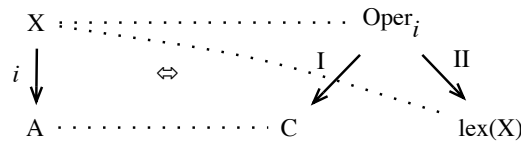
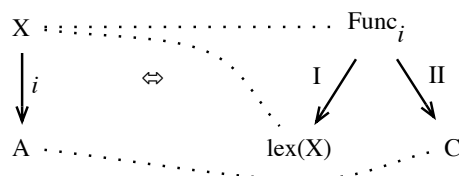


FIGURE 1 – Oper_i dans MARQUIS

Dans notre système, nous avons réduit le nombre de règles pour les verbes supports à trois. Nos règles sont donc encore plus générales que celles de Lareau & Wanner (2007). La figure 4 montre la règle qui traite à la fois Func₁, Func₂, Func₃, etc., et Oper₁, Oper₂, Oper₃, etc. La différence majeure porte sur la description des liens syntaxiques entre le nœud *f* et ses actants syntaxiques profonds. Nos règles exploitent au maximum la structuration des différents dictionnaires. Nous utilisons un dictionnaire LF où nous stockons des entrées comme celle de Oper₁ dans la figure 3. Cette entrée encode toutes les informations requises pour créer une structure syntaxique profonde correcte à l'aide de Oper₁ comme sa partie du discours et son régime ("gp", pour *government pattern*). On apprend donc que la base, *X* dans notre exemple,

FIGURE 2 – $Func_i$ dans MARQUIS

```

Oper1 {
  dpos = V
  gp = { base = II
        1 = I
        I = { dpos = N
              rel = subj } } }

```

FIGURE 3 – $Oper_1$ dans le dictionnaire *LF*

se réalisera nécessairement comme l'actant syntaxique **II** pour chaque $Oper_1$. La relation sémantique **1** correspond à la relation syntaxique **I**. Par ailleurs, cette entrée présente les contraintes syntaxiques imposées sur l'actant **I** de $Oper_1$ (ce doit être un nom introduit par la relation syntaxique de surface **subj**). Dans la figure 4, les instructions "*lf.f.base*" et "*lf.f.i*" sont des chemins permettant de récupérer dans le dictionnaire *LF* les relations syntaxiques relatives (**I** ou **II**) de la base et du premier actant par rapport au verbe support. La FL représentant le verbe support est symbolisée par la variable *f*. L'intérêt de cette règle et qu'elle permet de faire abstraction de la nature exacte de la FL et fonctionne comme un modèle de lexicalisation plus général encore. Cette règle commence par chercher dans le dictionnaire *lexicon*, de façon non-déterministe, l'entrée correspondant à "*lex(X)*" et récupère une FL qui respecte les contraintes de la règle en figure 4. Dans un second temps elle enregistre le nom de la FL dans la variable *f*, à laquelle on peut ensuite faire appel dans la règle. Grâce au dictionnaire *LF*, on peut récupérer la structure syntaxique associée à la FL en question. Cette même règle décrit ainsi les FL $Oper_1$, $Oper_2$, $Oper_3$, etc., et $Func_1$, $Func_2$, $Func_3$, etc.

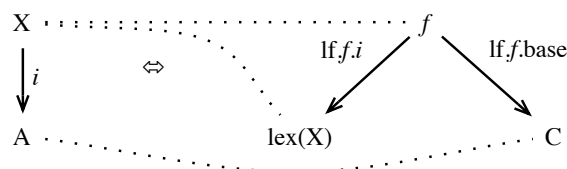


FIGURE 4 – Verbes supports à deux actants syntaxiques dans GÉCO

Nous avons créé des règles comparables pour les patrons illustrés dans la table 1. En plus de ces règles, des ressources lexicographiques sont en cours d'élaboration, notamment un dictionnaire des fonctions lexicales (*LF*), un dictionnaire d'unités sémantiques (*semanticon*) et un dictionnaire contenant les informations relatives à la combinatoire des unités lexicales (*lexicon*) dont la figure 5 illustre une entrée. Lors du processus de lexicalisation, les règles activent l'information

```

échec {
  dpos = N
  gp = { 1 = I
        I = { dpos = N } }
  lf = { name = Oper1
        value = subir }
  lf = { name = Oper1
        value = essuyer } }

```

FIGURE 5 – Entrée pour ÉCHEC dans le dictionnaire *lexicon*

contenue dans chacun de ces dictionnaires. Il est donc nécessaire de développer des dictionnaires détaillés fournissant une

analyse fine des unités lexicales et d’optimiser leur structuration pour accéder facilement à ces données. La spécification exhaustive de la combinatoire des unités lexicales offre également l’avantage d’encoder des relations d’équivalence syntaxique. La figure 5 montre que le lexème ÉCHEC possède deux FL dénotant des verbes supports quasiment équivalents. Les expressions *subir un échec* et *essuyer un échec* peuvent toutes deux être générées à partir d’une structure sémantique unique et ainsi rendre les règles non déterministes. Ainsi, ce système est capable de générer des paraphrases au sein d’une même langue. C’est également ce qui rend notre système multilingue. L’élaboration de tels dictionnaires est évidemment coûteuse en temps et requiert un haut niveau d’expertise en lexicographie. Nous nous servons de ressources existantes, notamment le *Réseau Lexical du Français* (Polguère, 2014), le *DicoEnviro* (L’Homme & Laneville, 2009), le *DicoInfo* (L’Homme, 2009), ainsi que le *DiCE* (Alonso Ramos, 2003).

5 Conclusion

Les fonctions lexicales sont intéressantes pour la GATM car elles permettent de modéliser des phénomènes linguistiques précis et récurrents d’une langue à l’autre, comme la gestion de paraphrases et la lexicalisation complexe. Les patrons modélisés jusqu’ici correspondent à des FL standard simples de base. Plusieurs autres types de FL viendront s’ajouter à l’avenir, notamment avec la modélisation des verbes de réalisation ($Real_i$, $Fact_i$, ...), les FL complexes, et éventuellement les FL non standard. La version finale de notre outil couvrira donc un ensemble conséquent de FL et sera disponible pour toute personne cherchant à travailler avec les FL.

Un des objectifs de notre système est de rendre utilisables pour la GAT des ressources lexicales électroniques existantes qui se servent des FL, comme celles mentionnées ci-dessus. Lors de l’implémentation des dictionnaires à partir de ces ressources, 5% des entrées lexicales existantes seront mises de côté en vue d’effectuer une évaluation de GÉCO. Il sera donc possible de mesurer la qualité de nos patrons à l’aide des mesures de précision et de rappel.

Références

- ALONSO RAMOS M. (2003). Hacia un diccionario de colocaciones del español y su codificación. *Lexicografía computacional y semántica*, **64**, 11–34.
- ALSHAWI H. & PULMAN S. G. (1992). Ellipsis, Comparatives, and Generation. In H. ALSHAWI, Ed., *The Core language engine*, chapter 13, p. 251–275. Cambridge, Massachusetts and London, England : The MIT Press.
- APRESJAN J. D., BOGUSLAVSKY I. M., IOMDIN L. L. & TSINMAN L. L. (2002). Lexical functions in actual NLP applications. In *Computational Linguistics for the New Millennium : Divergence or Synergy ? Festschrift in Honour of Peter Hellwig on the occasion of his 60th Birthday*, p. 55–72. Frankfurt : Peter Lang.
- AVGUSTINOVA T. & USZKOREIT H. (2000). An ontology of systemic relations for a shared grammar of slavic. In *Proceedings of Coling 2000*, Saarbrücken.
- BATEMAN J. A., KRUIJFF-KORBAYOVÁ I. & KRUIJFF G.-J. (2005). Multilingual resource sharing across both related and unrelated languages : An implemented, open-source framework for practical natural language generation. *Research on Language and Computation*, **15**, 1–29.
- BOGUSLAVSKY I., IOMDIN L. & SIZOV V. (2004). Multilinguality in ETAP-3 : Reuse of lexical resources. In *Proceedings of the Workshop on Multilingual Linguistic Resources*.
- BOHNET B. & WANNER L. (2010). Open source graph transducer interpreter and grammar development environment. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC’10)*, p. 211–218, Malta.
- FONTENELLE T. (1997). *Turning a bilingual dictionary into a lexical-semantic database*. Tübingen : Max Niemeyer Verlag.
- HEID U. & RAAB S. (1989). Collocations in multilingual generation. In *Proceedings of the fourth conference on European chapter of the Association for Computational Linguistics (EACL’89)*, p. 130–136, Manchester.
- IORDANSKAJA L., KIM M. & POLGUÈRE A. (1996). Some Procedural Problems in the Implementation of Lexical Functions for Text Generation. In (Wanner, 1996b), p. 279–297.
- JOUSSE A.-L. (2003). Normalisation des fonctions lexicales. Mémoire de DEA, Université Paris 7.
- JOUSSE A.-L. (2010). *Modèle de structuration des relations lexicales basé sur le formalisme des fonctions lexicales*. Thèse de doctorat, Université de Montréal / Université Paris 7.

- KAHANE S. & POLGUÈRE A. (2001). Formal foundation of lexical functions. In *Proceedings of the Workshop on Collocations at ACL 2001*, Toulouse.
- KIM R., DALRYMPLE M., KAPLAN R. M., KING T. H., MASUICHI H. & OHKUMA T. (2003). Multilingual grammar development via grammar porting. In *ESSLLI 2003 Workshop on Ideas and Strategies for Multilingual Grammar Development*.
- LAREAU F., DRAS M., BÖRSCHINGER B. & DALE R. (2011). Collocations in multilingual natural language generation : Lexical functions meet lexical functional grammar. In D. MOLLÁ & D. MARTINEZ, Eds., *Proceedings of the Australasian Language Technology Association Workshop*, p. 95–104, Canberra.
- LAREAU F., DRAS M., BÖRSCHINGER B. & TURPIN M. (2012). Implementing lexical functions in XLE. In M. BUTT & T. H. KING, Eds., *Proceedings of LFG'12*, p. 362–382, Denpasar, Indonesia : CSLI Publications.
- LAREAU F. & WANNER L. (2007). Towards a generic multilingual dependency grammar for text generation. In T. H. KING & E. M. BENDER, Eds., *Proceedings of the GEAF07 Workshop*, p. 203–223, Stanford : CSLI.
- LEE W. & EVENS M. (1996). Generating cohesive text using lexical functions. In (Wanner, 1996b), p. 299–306.
- L'HOMME M.-C. (2009). *DiCoInfo. Le dictionnaire fondamental de l'informatique et de l'Internet*. OLST, Université de Montréal.
- L'HOMME M.-C. & LANEVILLE M.-E. (2009). *DiCoEnviro. Le dictionnaire fondamental de l'environnement*. OLST, Université de Montréal.
- MEL'ČUK I. A. (1995). The future of the lexicon in linguistic description and the explanatory combinatorial dictionary. In I.-H. LEE, Ed., *Linguistics in the morning calm*, volume 3. Seoul : Hanshin.
- MEL'ČUK I. A. (2004). Verbes supports sans peine. *Linguisticae Investigationes*, **27**(2), 203–217.
- MEL'ČUK I. A., CLAS A. & POLGUÈRE A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Universités francophones. Louvain-la-Neuve : Duculot.
- POLGUÈRE A. (1997). Engineering text generation. *La Tribune des industries de la langue et de l'information électronique*, **23-24**, 21–39.
- POLGUÈRE A. (2000). A “natural” lexicalization model for language generation. In *Proceedings of SNLP 2000*, p. 37–50, Chiangmai.
- POLGUÈRE A. (2014). From writing dictionaries to weaving lexical networks. *International Journal of Lexicography*, **27**(4), 396–418.
- REITER E. & DALE R. (2000). *Building Natural Language Generation Systems*. Cambridge : Cambridge University Press.
- SANTAHOLMA M. (2008). Multilingual grammar resources in multilingual application development. In *Coling 2008 : Proceedings of the workshop on Grammar Engineering Across Frameworks*, p. 25–32, Manchester.
- STEINLIN J. (2003). Générer des collocations. Mémoire de DEA, Université Paris 7.
- VAN DER Wouden T. (1992). Prolegomena to a multilingual description of collocations. In H. TOMMOLA & K. VARANTOLA, Eds., *Proceedings of EURALEX 1992*, p. 449–456, Tampere.
- WANNER L. (1996a). Lexical Choice in Text Generation and Machine Translation. *Machine Translation*, **11**(1–3).
- L. WANNER, Ed. (1996b). *Lexical functions in lexicography and natural language processing*, volume 31 of *Studies in language companion series*. Amsterdam/Philadelphia : John Benjamins.
- WANNER L., BOHNET B., BOUAYAD-AGHA N., LAREAU F. & NICKLASS D. (2010). MARQUIS : Generation of user-tailored multilingual air quality bulletins. *Applied Artificial Intelligence*, **24**(10), 914–952.

Médicaments qui soignent, médicaments qui rendent malades : étude des relations causales pour identifier les effets secondaires

François Morlane-Hondère¹ Cyril Grouin¹ Véronique Moriceau^{1,2} Pierre Zweigenbaum¹

(1) LIMSI-CNRS, UPR 3251, rue John von Neumann, 91400 Orsay

(2) Université Paris-Sud, Campus universitaire d'Orsay, 91400 Orsay
{prenom.nom}@limsi.fr

Résumé. Dans cet article, nous nous intéressons à la manière dont sont exprimés les liens qui existent entre un traitement médical et un effet secondaire. Parce que les patients se tournent en priorité vers internet, nous fondons cette étude sur un corpus annoté de messages issus de forums de santé en français. L'objectif de ce travail consiste à mettre en évidence des éléments linguistiques (connecteurs logiques et expressions temporelles) qui pourraient être utiles pour des systèmes automatiques de repérage des effets secondaires. Nous observons que les modalités d'écriture sur les forums ne permettent pas de se fonder sur les expressions temporelles. En revanche, les connecteurs logiques semblent utiles pour identifier les effets secondaires.

Abstract.

Drugs that cure, drugs that make you sick : study of causal links to identify drug side effects

In this paper, we study the textual manifestations of the relation between drugs and side effects in online health forums. Our goal is to find relevant linguistic cues in order to improve the automatic identification of side effects by leveraging the ambiguity between actual side effects and indications (the reason for drug use). We find that the use of discourse markers can be relevant for the identification of indications – a third of indication mentions follow markers like ‘pour’ (‘for’) or ‘dans le but de’ (‘with the aim of’) – while temporal informations are not as discriminating.

Mots-clés : Pharmacovigilance, forums de santé, relations causales.

Keywords: Pharmacovigilance, Health Forums, Causal Links.

1 Introduction

Selon l'Organisation Mondiale de la Santé¹ (OMS), un effet secondaire est une réaction inattendue due à un traitement médical. Bien que des tests cliniques soient réalisés en laboratoire avant la commercialisation des médicaments, il est difficile de prévoir l'ensemble des effets secondaires d'un traitement pendant cette phase de tests, et ce, pour diverses raisons : durée limitée, différences entre patients, modifications des spécifications d'un traitement après les tests cliniques (Megahed, 2014). Il est alors nécessaire de réaliser une veille pharmacologique après l'autorisation de mise sur le marché d'un traitement. Parce qu'il est inattendu, un effet secondaire est généralement négatif (« effet indésirable »), mais peut se révéler positif (le Baclofène, initialement autorisé pour le traitement de troubles musculaires, permet le traitement de l'alcoolisme).

L'identification automatique d'effets indésirables de médicaments est une problématique récente. Alors que seuls 4 à 5% des effets indésirables sont rapportés de façon spontanée² auprès des centres de pharmacovigilance, il est nécessaire d'exploiter d'autres sources d'information, telles que les réseaux sociaux, vers lesquels les patients se tournent désormais pour obtenir des informations. De nombreuses études ont ainsi porté sur l'utilisation des réseaux sociaux pour réaliser une veille épidémiologique (Velardi *et al.*, 2014), suivre les conséquences d'un problème environnemental (Cha & Stow, 2015), ou à des fins de pharmacovigilance (Sampathkumar *et al.*, 2014). Les données issues des réseaux sociaux présentent de nombreux avantages en raison de leur accessibilité et de leur caractère massif. Le fait que ces données sont produites en continu constitue également un intérêt pour le processus de pharmacovigilance, qui nécessite une grande réactivité.

1. <http://www.who.int/fr/>

2. http://www.acadpharm.org/dos_public/GTNotif_Patients_Rap_VF__2015.01.22.pdf

2 État de l'art

Les systèmes développés s'appuient – au moins en partie – sur des lexiques d'effets indésirables (Wang *et al.*, 2009; Sarker *et al.*, 2015). Ces lexiques sont soit projetés directement sur les textes, soit utilisés comme des traits dans un système d'apprentissage automatique. La première approche montre toutefois ses limites face à la variabilité orthographique et stylistique qui caractérise les textes – par nature non contrôlés – issus de forums de discussion en ligne ou de réseaux sociaux. L'entraînement d'un système d'apprentissage automatique à partir d'un corpus annoté manuellement est une autre technique qui consiste à construire un modèle dans lequel les entités à extraire sont caractérisées par un ensemble de traits. Ces derniers portent aussi bien sur l'entité elle-même (présence de majuscules, de chiffres, de certains types de suffixes...) que sur son contexte d'apparition dans le corpus (n-grammes de mots, de parties du discours...).

L'un des problèmes rencontrés lors du repérage automatique d'effets secondaires³ est que ces effets, ou *événements*, se confondent avec un autre type d'entité, les *indications* (Nikfarjam *et al.*, 2015). Alors que les événements sont des symptômes ressentis après la prise d'un médicament (exemple 1), les indications sont les raisons pour lesquelles le médicament a été pris (exemple 2).

- (1) J'ai avalé le médoc a 12h, j'ai mangé et a 12h30 grosse crampes dans le ventre et brusque gonflement suivi d'un urticaire géant...
- (2) En faite c mon endo qui me la prescrit je n'arrivais pas a perdre de poids suite a des pb endocri.

Le fait qu'un symptôme peut souvent aussi bien constituer une indication qu'un événement complique la tâche d'identification des événements et génère du bruit.

Megahed (2014) montre que, du fait de cette ambiguïté, les entités appartenant aux classes "événement" et "indication" sont moins correctement classifiées que d'autres entités non ambiguës comme les noms de traitements. Il ressort également de cette étude que l'importance du phénomène varie en fonction du corpus étudié : le recouvrement entre les entités appartenant à la classe des événements et à celle des indications est plus important dans un corpus composé de messages portant sur la thématique des anti-dépresseurs que sur celle de la migraine. Sarker *et al.* (2015) ont également mis en lumière l'importance de ce phénomène. Ils montrent que le problème de l'ambiguïté entre indication et événement intervient dans 60 % des faux positifs produits par leur système et suggèrent d'analyser la polarité du contexte (les événements auront plus de chances d'être associés à des contextes qui portent une polarité négative que les indications).

L'étude que nous présentons ici se situe en aval d'une annotation préalable : nous partons d'un corpus annoté en entités pertinentes pour le domaine médical et nous cherchons à identifier celles qui relèvent de la cause ou de la conséquence de la prise d'un traitement médicamenteux. Pour ce faire, nous nous proposons d'étudier la pertinence pour ce type de tâche des indices que sont les connecteurs logiques et les marques temporelles. Ces indices ont été utilisés dans de précédentes études. Segura-Bedmar *et al.* (2011) mobilisent ainsi les connecteurs linguistiques pour identifier les interactions entre médicaments tandis que Sun *et al.* (2013) ont souligné l'utilité des éléments temporels pour typer des relations entre concepts médicaux.

3 Corpus

3.1 Constitution

Nous avons limité la thématique abordée dans le corpus au Médiator, pour tester et valider notre méthode d'identification des effets secondaires à un premier traitement. Le Médiator est un hypoglycémiant qui permet de lutter contre les glycémies excessives chez les diabétiques⁴, dont l'usage a été détourné pour permettre la perte de poids chez les personnes non diabétiques. Le choix du Médiator s'appuie sur le fait que les effets secondaires de ce traitement sont connus et documentés (les plus courants concernent des troubles digestifs, de la fatigue et des vertiges), et qu'il est possible d'étudier la manière dont sont exprimés les problèmes rencontrés avant et après la date de retrait du marché en novembre 2009.

3. Nous reprenons la terminologie utilisée par les centres de pharmacovigilance avec lesquels nous interagissons dans le cadre du projet qui soutient ce travail. Pour les centres de pharmacovigilance, un effet secondaire est un événement (quelque chose qui se produit au niveau clinique) qui peut se révéler aussi bien positif que négatif. Dans ce dernier cas, on parlera d'effets indésirables.

4. <http://www.eurekasante.fr/medicaments/vidal-famille/medicament-dmedia01-MEDIATOR.html>

Le corpus se compose de dix fils complets de discussions autour du Médiateur et de son principe actif (benfluorex), publiés à deux périodes différentes (avant et après la date de retrait du marché de ce traitement) et issus de deux forums de santé⁵ en français. Les fils de discussion ont été découpés en messages individuels⁶, pour un total de 157 messages. Une dés-identification⁷ manuelle a été réalisée pour masquer quatre types d'informations identifiantes (nom, prénom, pseudonyme, âge) que nous avons remplacées par une balise typante (e.g. <prenom/>) et nous avons changé les dates⁸ présentes dans les documents lorsqu'elles se rapportent aux patients. Aucun autre traitement n'a été appliqué au corpus : ni tokenisation, ni correction orthographique ou syntaxique.

3.2 Annotation

Pour aborder la problématique de la détection des effets secondaires résultants d'une prise médicamenteuse, nous avons défini un schéma d'annotation composé de 16 catégories sémantiques⁹, inspirées des types sémantiques de l'UMLS (Lindberg *et al.*, 1993). Ces catégories nous permettent de couvrir l'ensemble des informations nécessaires pour détecter la cause d'un effet secondaire (le traitement médical) et l'effet secondaire en lui-même (un symptôme, une maladie, une fonction biologique dégradée, etc.). Le principe d'annotation qui a été retenu consiste à annoter les têtes de syntagme afin (i) de se focaliser sur les entités porteuses de sens et (ii) de limiter l'impact des erreurs de reconnaissances des frontières des annotations. Puisque certaines catégories peuvent être aussi bien la cause que la conséquence d'une prise médicamenteuse, nous avons défini un attribut "rôle" qui permet de spécifier, lorsque le cas s'y prête, si l'annotation renvoie à l'*indication* (cause) ou à l'*événement* (conséquence). La figure 1 donne un exemple d'annotation du corpus.

```
Suite à quelques <SOSY role="indication">malaises</SOSY> avec <SOSY role="indication">perte</SOSY> de <FUNC
role="indication">connaissance</FUNC>, mon <JOB>endocrinologue</JOB> m'a <PROC>prescrit</PROC> du
<CHEM>Médiateur</CHEM>.
Au début c'est vrai j'ai eu le phénomène <PROC role="evenement">perte</PROC> de <FUNC
role="evenement">poids</FUNC> (<WEIGHT role="evenement">6kg</WEIGHT>).
Au fil des années, <DISO role="evenement">migraines</DISO>, <DISO role="evenement">diarrhées</DISO>, <DISO
role="evenement">crampes</DISO>, une <SOSY role="evenement">fatigue</SOSY> de plus en plus grande, une <SOSY
role="evenement">hyper-émotivité</SOSY>, toujours sur les <ANAT role="evenement">nerfs</ANAT>.
```

FIGURE 1 – Exemple d'annotations issues du corpus (SOSY = Sign or Symptom, FUNC = Biological Process or Function, PROC = Medical Procedure, CHEM = Chemical or drugs, DISO = Disorders, ANAT = Anatomy)

Le tableau 1 présente la répartition des annotations par catégorie en fonction de la valeur prise par l'attribut "rôle" (indication/événement) et lorsqu'aucune de ces valeurs n'est pertinente. Sur 16 catégories, seule la moitié est sous-spécifiée avec une valeur d'attribut. Pour ces huit catégories, les annotations en *Anatomy* et *Sign or Symptom* sont majoritairement sous-spécifiées comme "événement" (53,8% pour *Anatomy* et 66,9% pour *Sign or Symptom*). Cette prépondérance correspond également à la manière dont le corpus est annoté, avec des annotations connexes entre ces deux catégories (la portion "mal de tête" sera annotée avec la catégorie *Sign or Symptom* sur "mal" et la catégorie *Anatomy* sur "tête").

5. Nous avons extrait quatre fils de discussion du site atoute.org (période 2004/2006, avant le retrait du marché) et six fils de discussion du site doctissimo.fr (période 2013/2014, après le retrait du marché).

6. Soit 75 messages du site atoute.org et 82 messages du site doctissimo.fr

7. La désidentification consiste à masquer ou modifier toutes informations relevant de catégories prédéfinies (*nom, prénom, adresse, téléphone, date, etc.*) permettant d'identifier l'auteur d'un message ou une personne mentionnée dans un message. S'il peut paraître inutile de désidentifier des documents récupérés sur internet, dans la mesure où il est possible de les retrouver par une simple recherche, la désidentification permet néanmoins de respecter la vie privée des utilisateurs des forums, notamment dans le cas où un utilisateur demanderait à ce que ses messages soient retirés du forum.

8. Nous avons réalisé une antédation aléatoire manuelle des dates en retranchant quelques jours et en conservant le format d'origine.

9. *Anatomy* : parties du corps, y compris fluides et tissus (cerveau, peau, sang) — *Biological Process or Function* : processus ou état qui se produit naturellement, ou résultant d'une activité (respirer) — *Disorders* : maladies (cancer) — *Sign or Symptom* : manifestation observable d'une maladie, condition fondée sur un jugement clinique (fatigue, douleurs, ballonnement) — *Chemical or Drugs* : médicament, principe actif, classe pharmacologique (Médiateur, benfluorex) — *Genes Proteins* : protéines, lipides, acides nucléiques, gènes (insuline, lipase, triglycérides) — *Medical procedure* : activité médicale ou chirurgicale, liée au soin des patients, y compris diagnostiques, procédures et méthodes de traitement (radiothérapie) — *Weight* : poids total ou partiel du patient (82 kg, -5 kgs) — *Job* : activité professionnelle (médecin, gygy). — Des informations posologiques sur le traitement : *Concentration, Dosage, Mode* et des informations temporelles liées au traitement : *Date, Duration, Frequency, Time*.

	Anatomy	Disorders	Duration	Function	Gene	Procedure	Sign or Symptom	Weight
Indication	5	34	0	12	24	30	8	1
Événement	42	17	8	14	1	9	91	20
Aucun	31	52	79	43	13	149	37	25
Total	78	103	87	69	38	188	136	46

TABLE 1 – Répartition des annotations par catégorie selon la valeur de l’attribut “rôle” (indication/événement) et en l’absence de valeur associée à cet attribut (aucun)

4 Études distributionnelles

4.1 Connecteurs logiques

Lorsqu’elles se manifestent dans les textes, les relations discursives peuvent s’accompagner de marqueurs comme des connecteurs logiques, qui explicitent une relation entre deux phrases ou deux segments de phrases (“*parce que*” introduit une explication, “*plutôt que*” une alternative). Nous faisons l’hypothèse que la différence entre indication et événement peut être envisagée comme relevant des relations de discours. Dans ce cas, la présence de connecteurs logiques dans le texte constitue un indice pertinent pour la désambiguïsation des indications et des événements.

Nous avons utilisé Lexconn REF (Roze *et al.*, 2012; Roze, 2013), un lexique contenant 231 connecteurs logiques associés à une ou plusieurs des 20 relations discursives issues de la SDRT (Asher & Lascarides, 2003). Les connecteurs ont été cherchés dans une fenêtre de 10 mots précédant une entité à désambiguïser. Après une première projection des connecteurs, nous avons jugé nécessaire d’apporter trois modifications à Lexconn : (i) les connecteurs que nous avons jugés trop polysémiques (à, en, et, si) n’ont pas été pris en compte, (ii) nous avons regroupé certaines relations comme les relations d’opposition et de contraste, et (iii) nous avons ajouté les connecteurs “*dans le cadre de*” et “*pour cause de*”, associés à la relation *explication*.

Nous rapportons dans le tableau 2 la fréquence des marqueurs extraits pour trois relations (*explication*, *but* et *opposition*) et, entre parenthèses, la proportion qu’elle représente par rapport au nombre total d’indications ou d’événements.

	explication	but	opposition
indication	11 (9,6 %)	38 (33,3 %)	6 (5,3 %)
événement	9 (4,5 %)	8 (4 %)	8 (4 %)

TABLE 2 – Distribution des connecteurs logiques pour les relations *explication*, *but* et *opposition*

Nous n’avons fait apparaître que les relations pour lesquelles au moins dix marqueurs ont été extraits, ce qui ne concerne que trois relations sur les vingt contenues dans Lexconn. Le fait que si peu de relations soient représentées – et la faiblesse relative du nombre de marqueurs extraits en général – peut s’expliquer de plusieurs façons. Une première explication est que Lexconn a été construit en prenant la base Frantext, qui est constituée de textes littéraires. On peut donc prédire un décalage entre la nature des connecteurs utilisés dans Frantext et dans notre corpus. On peut également faire l’hypothèse que le caractère non contrôlé des textes de forums incite les scripteurs à utiliser moins d’indices explicites de la structure textuelle (exemple 3) ou à utiliser des connecteurs atypiques (exemple 4).

(3) Au fil des années, migraines, diarrhées, crampes, une fatigue de plus en plus grande, une hyper-émotivité, toujours sur les nerfs.

(4) j’ai pris médiateur pendant un certain temps = problèmes intestinaux, diarrhée

Le résultat le plus intéressant que nous fournit le tableau 2 est qu’un tiers des indications sont introduites par un marqueur de but (alors que ce n’est le cas que pour 4 % des événements) comme “*pour*” – principalement –, “*afin de*” ou “*dans le but de*”. Cette relation exprime un lien entre la prise d’un médicament et le but de cette prise, à savoir résoudre un problème médical, analysé ici comme une indication (exemple 5).

(5) j’ai moi-même en 1998, pris ce médicament pour maigrir

Les marqueurs d’explication que sont “*car*”, “*dans le cadre de*” et “*pour cause de*” semblent jouer un rôle similaire, mais leur différence d’emploi pour introduire une indication ou un événement est moins flagrante (exemple 6).

(6) ce médicament m'a été prescrit dans le cadre d'un problème d'hyperinsulinisme

La relation d'opposition apparaît potentiellement intéressante en cela qu'elle peut être l'indice d'un phénomène inattendu (comme elle l'est à l'aide du marqueur "or" dans l'exemple 7).

(7) mon endocrinologue me prescrit du LEVOTHYROX 75mg par jour + 3MEDIATOR. or je prends de plus en plus de poids"

La distribution des marqueurs d'opposition n'apparaît toutefois pas potentiellement discriminante.

4.2 Expressions temporelles

Intuitivement, on peut penser que les problèmes de type *indication* se produisent temporellement avant les problèmes de type *événement*. Pour vérifier cette hypothèse, nous nous sommes intéressés aux expressions temporelles associées à ces deux types de problème. Pour cela, nous avons utilisé les annotations manuelles des expressions temporelles du corpus pour les types DATE (*depuis mai 2005*), TIME (*à 12h30, au coucher*), DURATION (*pendant de longs mois*) et FREQUENCY (*par jour, régulièrement*).

Nous avons recensé les expressions temporelles qui se trouvent dans la même phrase qu'une des 8 catégories acceptant un rôle "indication"/"événement". Lorsqu'il existe plusieurs expressions temporelles dans une phrase, nous n'avons considéré que celle la plus proche (en nombre de mots) de l'indication ou l'événement. Nous avons aussi noté la position de l'expression temporelle par rapport à l'indication ou l'événement : par exemple, la date est avant l'indication dans :

Je prend <CHEM>Médiator</CHEM> depuis le mois de <DATE>mars</DATE> pour <PROC role="indication">contrôler</PROC> mon <DISO role="indication">cholestérol</DISO>

ou la durée est après l'événement dans :

J'ai perdu <WEIGHT role="evenement">7 kgs</WEIGHT> en <DURATION>6mois</DURATION>.

Le tableau 3 montre la distribution des expressions temporelles dans le corpus par rapport à leur position vis-à-vis d'une indication ou d'un événement.

		DATE		TIME		DURATION		FREQUENCY		TOTAL
		avant	après	avant	après	avant	après	avant	après	
GENE	INDIC EVENT					1	1		2	4
DISORDER	INDIC EVENT	2	2	2		1		2		5
SIGN OR SYMPTOM	INDIC EVENT	1	1	3	1	8	5	3	1	26
PROCEDURE	INDIC EVENT	4	1	1		2	1			9
FUNCTION	INDIC EVENT			1						1
WEIGHT	INDIC EVENT	2				7	1	1		17
TOTAL	INDIC EVENT	4	2	1	1	5	3	4	2	22
		5	1	6	4	16	12	4	1	51

TABLE 3 – Distribution des expressions temporelles

On remarque que les expressions temporelles, quel que soit leur type, sont très majoritairement associées aux catégories *Sign or symptom* et *Weight* et principalement de rôle *événement*. On note également que les expressions temporelles de type TIME et DURATION sont principalement associées à des rôles *événement* alors que pour les autres types d'expression temporelle, la distribution est assez uniforme. La catégorie *Procedure* est majoritairement de rôle *indication* quelle que soit l'expression temporelle associée. Enfin, la catégorie *Disorders* est de type *indication* si une date est positionnée après alors qu'elle est de type *événement* si la date est avant.

Ces observations sont des indices qui peuvent aider à catégoriser une annotation issue de notre schéma en *indication* ou *événement* même si le petit nombre d'occurrences ne permet pas de tirer de réelle conclusion. De plus, plutôt que la position des expressions temporelles, il faudrait connaître leurs relations de dépendance syntaxique : dans l'exemple suivant, la première date est associée à la catégorie *Procedure* dans la première proposition alors que la seconde date est associée à la catégorie *Sign or symptom* dans la seconde proposition :

```
Ma jeune cousine est sous <CHEM>médiateur</CHEM> pour "<PROC role="indication">s'affiner</PROC>"
depuis <DATE>vendredi 12 mars 2004</DATE> et depuis <DATE>lundi 15</DATE> se plaint de
<SOSY role="evenement">douleurs</SOSY> au <ANATOMY role="evenement">ventre</ANATOMY>
```

On le voit également dans cet exemple, les valeurs normalisées des dates peuvent permettre de typer les catégories : ainsi, la date *vendredi 12 mars 2004* (2004-03-12) précède temporellement la date *lundi 15* (2004-03-15) et confirme qu'une *indication* se produit avant un *événement*. Pour obtenir ces informations, nous avons adapté les règles pour le français de l'outil libre HeidelTime (Moriceau & Tannier, 2014) afin d'extraire et normaliser les expressions temporelles exprimées dans un style propre aux forums en ligne (par exemple, *pdt 3 jrs* pour *pendant 3 jours*).

5 Conclusion

Dans cet article, nous avons présenté les premières études que nous avons menées pour identifier des indices linguistiques permettant de distinguer les entités cliniques qui sont des *indications* (ce pour quoi le traitement est administré) de celles qui sont des *événements* (ce qui est causé par le traitement). De cette étude, il ressort que ces indices permettent plus facilement de repérer les *indications* que les *événements*. Nous prévoyons toutefois de tester les indices que sont les verbes de causation (*provoquer, engendrer, entraîner...*), dont nous faisons l'hypothèse qu'ils sont pertinents pour le repérage des événements.

Nous estimons que ces indices devraient permettre d'identifier automatiquement les effets secondaires dans les messages parus sur les forums de santé, à des fins de pharmacovigilance. A cet effet, nous envisageons de poursuivre ces travaux en utilisant les indices que nous avons mis en évidence pour identifier automatiquement les entités cliniques qui relèvent d'une indication ou d'un événement.

Le corpus que nous avons utilisé est limité à un seul traitement (Médiateur) et se compose d'un nombre réduit de messages (157 messages). Nous avons envisagé la création de ce corpus uniquement dans la perspective d'identifier des indices pour l'extraction d'entités cliniques d'une part, et d'en vérifier leur utilité réelle d'autre part. Si la question de la mise à disposition du corpus et des annotations réalisées est pertinente, la problématique de la redistribution de contenus issus d'internet ne nous permet pas de donner accès à ce corpus. Si le contenu des messages n'est pas redistribuable pour des questions de droit, il est néanmoins possible de donner accès à la liste des URL correspondant aux discussions qui nous ont permis de constituer le corpus. C'est notamment l'approche qui a été suivie dans l'atelier DEFT 2015 (<https://deft.limsi.fr/2015/>) : les organisateurs ont fourni aux participants la liste des tweets sur lesquels ils devaient développer et appliquer leurs méthodes pour répondre à la problématique posée. La méthode présentée dans cet article reste cependant applicable sur n'importe quel corpus issus de forums de santé, ce qui conduit malgré tout à la possibilité de dupliquer la méthode présentée et de confirmer ou de réfuter les conclusions que nous avons présentées.

Remerciements

Ce travail a été réalisé dans le cadre du projet Vigi4MED (ANSM-2013-S-060), financé par l'ANSM (Agence Nationale de Sécurité du Médicament).

Références

- ASHER N. & LASCARIDES A. (2003). *Logics of Conversation*. Cambridge University Press.
- CHA Y. & STOW C. (2015). Mining web-based data to assess public response to environmental events. *Environ Pollut*, **198**, 97–9.

- LINDBERG D. A., HUMPHREYS B. L. & MCRAY A. T. (1993). The Unified Medical Language System. *Methods Inf Med*, **32**(4), 281–91.
- MEGAHED D. (2014). Etude des forums de santé pour la détection d'événements secondaires. Master's thesis, INaLCO.
- MORICEAU V. & TANNIER X. (2014). French resources for extraction and normalization of temporal expressions with heideltime. In *Proc of LREC*, p. 3239–43, Reykjavik, Iceland.
- NIKFARIAM A., SARKER A., O'CONNOR K., GINN R. & GONZALES G. (2015). Pharmacovigilance from social media : mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J Am Med Inform Assoc*.
- ROZE C. (2013). *Vers une algèbre des relations de discours*. PhD thesis, Université Paris-Diderot - Paris VII, Paris, France.
- ROZE C., DANLOS L. & MULLER P. (2012). LEXCONN: a French lexicon of discourse connectives. *Discours*, **10**.
- SAMPATHKUMAR H., CHEN X. & LUO B. (2014). Mining adverse drug reactions from online healthcare forums using hidden markov model. *BMC Med Infor Decis Mak*, **91**(14).
- SARKER A., NIKFARIAM A., O'CONNOR K., GINN R., GONZALEZ G., UPADHAYA T., JAYARAMAN S. & SMITH K. (2015). Utilizing social media data for pharmacovigilance: a review. *J Biomed Inform*. Epub ahead of print.
- SEGURA-BEDMAR I., MARTÍNEZ P. & DE PABLO-SÁNCHEZ C. (2011). A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. *BMC Bioinformatics*, **12**(Suppl 2)(S1).
- SUN W., RUMSHISKY A. & UZUNER O. (2013). Evaluating temporal relations in clinical text : 2012 i2b2 challenge. *J Am Med Inform Assoc*, **20**(5), 806–13.
- VELARDI P., STILO G., TOZZI A. & GESUALDO F. (2014). Twitter mining for fine-grained syndromic surveillance. *Artif Intell Med*, **61**(3), 153–63.
- WANG X., HRIPCSAK G., MARKATOU M. & FRIEDMAN C. (2009). Research paper : Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records : A feasibility study. *JAMIA*, **16**(3), 328–337.

Exploration de modèles distributionnels au moyen de graphes 1-PPV

Gabriel Bernier-Colborne¹

(1) OLST, Université de Montréal, CP 6128, succ. Centre-Ville, Montréal (QC) Canada, H3C 3J7
gabriel.bernier-colborne@umontreal.ca

Résumé. Dans cet article, nous montrons qu'un graphe à 1 plus proche voisin (graphe 1-PPV) offre différents moyens d'explorer les voisinages sémantiques captés par un modèle distributionnel. Nous vérifions si les composantes connexes de ce graphe, qui représentent des ensembles de mots apparaissant dans des contextes similaires, permettent d'identifier des ensembles d'unités lexicales qui évoquent un même cadre sémantique. Nous illustrons également différentes façons d'exploiter le graphe 1-PPV afin d'explorer un modèle ou de comparer différents modèles.

Abstract.

Exploring distributional semantic models using a 1-NN graph.

We show how a 1-NN graph can be used to explore the semantic neighbourhoods modeled by distributional models of semantics. We check whether the connected components of the graph, which represent sets of words that occur in similar contexts, can be used to identify sets of lexical units that evoke the same semantic frame. We then illustrate different ways in which the 1-NN graph can be used to explore a model or compare different models.

Mots-clés : Sémantique distributionnelle, sémantique lexicale, graphe, terminologie, sémantique des cadres.

Keywords: Distributional semantics, lexical semantics, graph, terminology, frame semantics.

1 Introduction

Dans le cadre d'un projet visant à décrire le vocabulaire du domaine de l'environnement, nous cherchons à faciliter l'identification de relations sémantiques paradigmatiques telles que la synonymie ainsi que l'identification d'ensembles d'unités lexicales évoquant un même cadre sémantique, suivant le cadre descriptif proposé par Fillmore (1982). Nous exploitons à cette fin des techniques permettant l'identification semi-automatique de relations sémantiques à partir de corpus spécialisés. L'analyse distributionnelle permet d'estimer la similarité sémantique de deux mots en comparant les contextes dans lesquels ils apparaissent dans un corpus, l'hypothèse sous-jacente étant que les mots qui apparaissent dans des contextes similaires ont tendance à présenter des affinités sémantiques (Harris, 1954). La similarité distributionnelle, qui peut être calculée de différentes façons, est souvent utilisée pour construire des thésaurus distributionnels, ressources associant à chaque entrée une liste de ses plus proches voisins (PPV) selon la mesure de similarité.

Un thésaurus distributionnel peut être considéré comme un graphe de k plus proches voisins (graphe k -PPV), graphe dans lequel chaque mot est relié à ses k PPV. Les graphes k -PPV peuvent servir non seulement à représenter le voisinage d'un mot donné, mais aussi à identifier des ensembles de mots sémantiquement reliés. Différentes techniques peuvent être utilisées à cette fin ; une technique simple consiste à calculer les composantes connexes d'un graphe 1-PPV, celles-ci représentant des ensembles de mots distributionnellement similaires. Dans cet article, nous vérifions si le graphe 1-PPV peut faciliter l'identification d'ensembles d'unités lexicales qui évoquent un même cadre sémantique ; nous montrons également que le graphe nous fournit une perspective intéressante sur les voisinages sémantiques captés par un modèle distributionnel. Nous décrivons le graphe 1-PPV à la Section 2 et les ressources que nous avons utilisées à la Section 3. Nous évaluons le graphe à la Section 4 et présenterons différentes façons d'exploiter le graphe à la Section 5. Enfin, nous présenterons quelques travaux reliés à la Section 6.

2 Le graphe 1-PPV

Tout modèle qui permet d'estimer la similarité de deux mots peut donner lieu à la construction d'un graphe de voisinage qui relie chaque mot à ses PPV. Dans le cadre de ce travail, nous avons utilisé deux modèles différents, à savoir HAL et word2vec (W2V), afin de vérifier si le graphe 1-PPV révèle des différences quant aux voisinages sémantiques qu'ils modélisent. HAL (Lund *et al.*, 1995; Schütze, 1992) est un modèle à fenêtre graphique qui repose sur une matrice mot-mot qui encode la fréquence de cooccurrence des mots. Le modèle de langue neuronal word2vec (Mikolov *et al.*, 2013a,b), qui a été exploité dans plusieurs applications du TAL dans les dernières années, apprend des représentations distribuées de mots qui peuvent servir à estimer la similarité sémantique, entre autres. Dans les deux cas, nous calculons la similarité entre les mots au moyen du cosinus de l'angle de leurs vecteurs ; les PPV d'un mot sont obtenus en calculant la similarité entre ce mot et tous les autres mots dans le vocabulaire.

Le graphe que nous utilisons est un graphe k-PPV symétrique¹, c'est-à-dire un graphe non orienté dans lequel deux mots w_i et w_j sont reliés par une arête si w_i est un des k PPV de w_j ou si w_j est un des k PPV de w_i . Le nombre de composantes connexes² dans ce graphe dépend de la valeur de k , entre autres : en faisant varier la valeur de k , nous avons observé que le graphe k-PPV symétrique devient généralement connexe dès que la valeur de k atteint 3 ou 4. Un graphe connexe peut servir à identifier des ensembles de mots similaires au moyen d'une technique permettant de repérer des sous-graphes de forte densité. Dans cet article, nous utilisons plutôt un graphe qui n'est pas connexe, le graphe 1-PPV symétrique, dans lequel deux mots sont reliés si l'un est le PPV de l'autre. Nous utilisons le graphe 1-PPV parce que le calcul de ses composantes connexes nous fournit un moyen simple d'identifier des ensembles de mots distributionnellement similaires. Nous vérifierons ainsi si ce genre de graphe peut faciliter l'identification d'ensembles d'unités lexicales évoquant un même cadre sémantique (ces ensembles seront décrits à la Section 3) dans le cadre de l'élaboration d'une ressource lexicale spécialisée basée sur la sémantique des cadres.

3 Ressources utilisées

Les modèles ont été construits³ sur le corpus monolingue français PANACEA – domaine de l'environnement (ELRA-W0065), un corpus de pages Web reliées au domaine de l'environnement, contenant 23 514 documents (environ 47 millions de tokens). Ce corpus a été compilé au moyen d'un outil de construction automatique de corpus spécialisés conçu dans le cadre du projet PANACEA et il est distribué librement à des fins de recherche⁴. Le corpus a été lemmatisé au moyen de TreeTagger (Schmid, 1994).

Nous avons également obtenu des données de référence afin d'évaluer le graphe 1-PPV. Ces données ont été extraites du Framed DiCoEnviro⁵ (L'Homme & Robichaud, 2014; L'Homme *et al.*, 2014), une ressource lexicale spécialisée décrivant les termes du domaine de l'environnement au moyen de la sémantique des cadres (Fillmore, 1982). La méthodologie utilisée pour construire le Framed DiCoEnviro est inspirée de celle du projet FrameNet (Ruppenhofer *et al.*, 2010) ; certains des cadres sont basés sur ceux dans FrameNet, d'autres ont été élaborés spécifiquement pour le domaine de l'environnement. Nous avons extrait de cette ressource des ensembles d'unités lexicales qui évoquent le même cadre sémantique ; p. ex. le cadre *Change_of_Temperature* est évoqué par les unités lexicales *réchauffer*, *réchauffement*, *refroidir* et *refroidissement*. Pour chaque cadre, nous avons extrait les unités lexicales qui font partie du vocabulaire pour lequel nous avons construit des modèles, qui est constitué des 10000 formes lemmatisées les plus fréquentes dans le corpus (les mots vides étant exclus au moyen d'un anti-dictionnaire). Parmi les ensembles résultants, nous avons conservé ceux contenant au moins 2 mots, puis nous avons éliminé le recoupement entre les ensembles : lorsque deux ensembles partageaient au moins un mot, nous avons conservé seulement le plus gros ensemble. Nous avons ainsi obtenu 52 ensembles de référence. Le nombre de mots par ensemble varie entre 2 et 10, 42 des 52 ensembles contenant 4 mots ou moins.

1. Il existe différents types de graphes de voisinage, tels que les graphes k-PPV symétriques et mutuels (Maier *et al.*, 2007).

2. Les composantes connexes d'un graphe sont les sous-graphes de taille maximale dans lesquels il existe un chemin entre n'importe quelle paire de sommets.

3. Nous utilisons les bibliothèques python suivantes : gensim (<http://radimrehurek.com/gensim/>) pour l'entraînement des modèles word2vec, scikit-learn (<http://scikit-learn.org/>) pour la SVD et networkx (<http://networkx.github.io/>) pour la construction, l'analyse et la visualisation de graphes.

4. Voir http://catalog.elra.info/product_info.php?products_id=1186&language=fr et <http://panacea-lr.eu/>

5. <http://olst.ling.umontreal.ca/dicoenviro/framed/index.php> (en construction)

4 Évaluation

Les ensembles de référence décrits à la section 3 ont été utilisés afin d'évaluer les graphes 1-PPV symétriques créés à partir des modèles distributionnels. Cette évaluation a été réalisée afin de vérifier si les composantes du graphe 1-PPV permettent d'identifier des ensembles d'unités lexicales évoquant un cadre sémantique ; elle a également servi à choisir un modèle pour les exemples présentés dans cet article. Pour chaque modèle, nous avons évalué différentes paramétrisations⁶. Dans le cas du modèle HAL, nous avons fait varier le type, la forme et la taille de la fenêtre de contexte ainsi que la pondération appliquée aux fréquences de cooccurrence ; l'influence de ces paramètres a été analysée par Bullinaria & Levy (2007), entre autres. Pour chaque paramétrisation, nous avons également appliqué la SVD (300 dimensions, algorithme ARPACK) au modèle, suivant Schütze (1992). Nous avons également évalué différentes paramétrisations du modèle word2vec (W2V) en faisant varier certains de ses principaux paramètres : l'architecture du modèle, l'algorithme d'entraînement, la taille de fenêtre et le nombre de dimensions. Nous avons ainsi évalué 40 paramétrisations de chaque modèle (HAL, SVD et W2V).

Pour chacune des paramétrisations, nous avons calculé différentes caractéristiques du graphe 1-PPV résultant et évalué le graphe au moyen de mesures d'évaluation externes calculées en comparant les composantes connexes du graphe aux ensembles de référence décrits à la Section 3. Dans le cadre de cet article, les mesures que nous utilisons sont deux mesures simples de rappel. Le rappel est simplement le pourcentage des ensembles de référence dont tous les mots se retrouvent dans la même composante. Cette mesure permet d'estimer dans quelle mesure les composantes du graphe permettent d'identifier des ensembles d'unités lexicales évoquant un même cadre sémantique. Pour pénaliser les graphes contenant un nombre faible de composantes (ce qui augmente la probabilité que les mots d'un ensemble de référence se retrouveront dans la même composante), nous avons également calculé une mesure de rappel corrigée en fonction du nombre de composantes, de la façon suivante : $R_{corr} = R \cdot \frac{2|C|}{|V|}$, où R est le rappel, $|C|$ est le nombre de composantes dans le graphe et $|V|$ est le nombre de mots dans le vocabulaire.

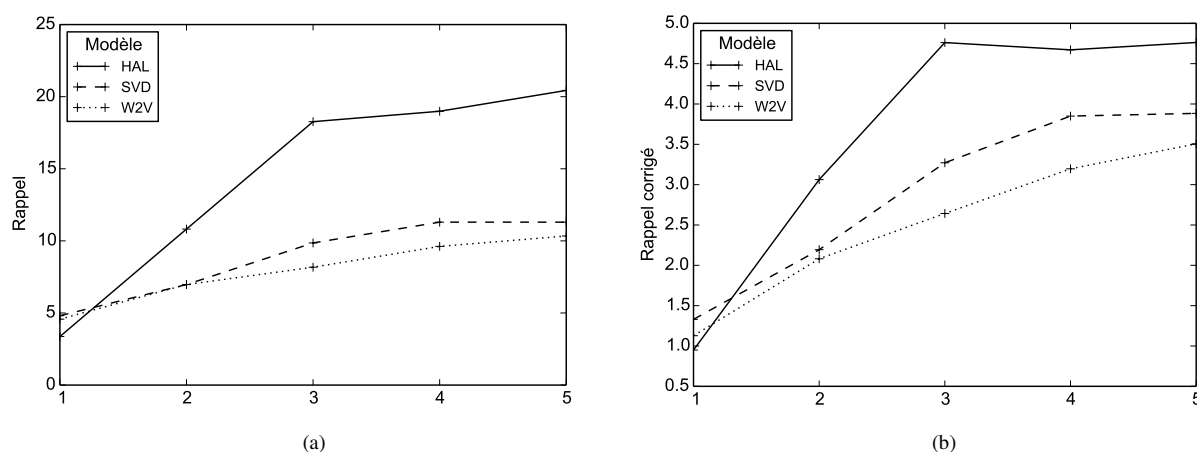


FIGURE 1: Influence de la taille de la fenêtre de contexte sur (a) le rappel et (b) le rappel corrigé. Tous les points sont des moyennes sur l'ensemble des paramétrisations correspondant à une taille de fenêtre et un modèle particuliers.

Le rappel que nous avons observé est de 14.37% en moyenne pour les modèles HAL, 8.85% pour les modèles SVD et 7.93% pour les modèles W2V. Le rappel maximal est de 25% pour les modèles HAL, 17.31% pour les modèles SVD et 15.38% pour les modèles W2V. Ces résultats suggèrent que les composantes du graphe 1-PPV peuvent servir à identifier des ensembles d'unités lexicales qui évoquent un cadre sémantique dans une certaine mesure, bien qu'une augmentation du rappel demeure souhaitable ; à ce titre, il serait intéressant de vérifier comment ce graphe se compare à d'autres types de graphes de voisinage.

Un des paramètres importants des 3 modèles que nous utilisons est la taille de la fenêtre de contexte. La Figure 1 montre l'influence de la taille de fenêtre et du type de modèle sur le rappel et le rappel corrigé. Lorsque le rappel est corrigé en fonction du nombre de composantes, la différence entre les 3 types de modèles est moins importante parce que les

6. Nous ne décrivons pas en détail chacun des paramètres des différents modèles, parce que nous n'examinons pas l'influence des paramètres dans le cadre de cet article, à l'exception de la taille de fenêtre.

modèles HAL produisent un nombre moins élevé de composantes (1318 en moyenne) que les modèles SVD (1611) et W2V (1534). Le fait que HAL offre un rappel plus élevé que les modèles SVD et W2V est lié à plusieurs facteurs, outre le nombre de composantes ; un de ces facteurs est la nature des données de référence, qui peuvent notamment contenir des unités lexicales de différentes parties du discours. Soulignons que le rappel maximal a été atteint avec une fenêtre de 3 mots pour les modèles HAL, 4 mots pour les modèles SVD et 5 mots pour les modèles W2V ; il serait donc intéressant de tester des fenêtres plus larges pour vérifier si le rappel de W2V continue à augmenter.

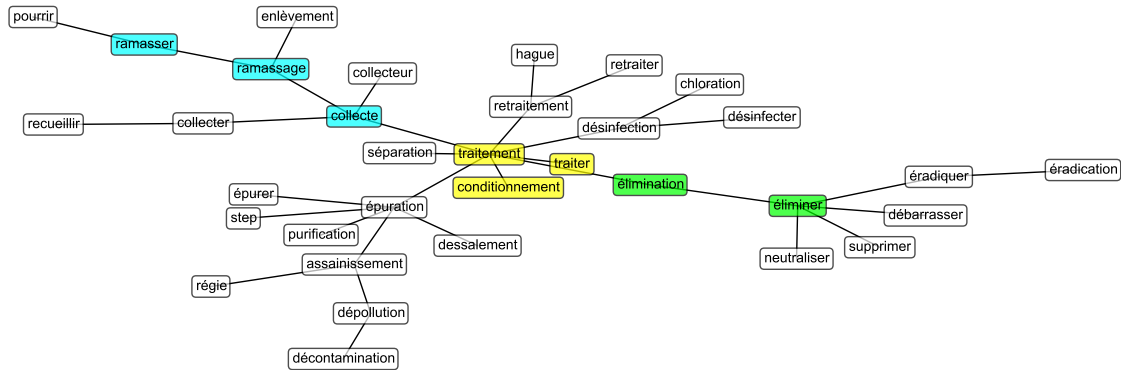


FIGURE 2: Composante du graphe 1-PPV contenant 3 ensembles de référence (mots en bleu, en jaune et en vert).

Le modèle que nous avons retenu pour les exemples présentés dans cet article (sauf indication contraire) est celui qui maximise à la fois le rappel (25%) et le rappel corrigé (6.54) ; il s'agit d'un des modèles HAL calculés au moyen d'une fenêtre de 3 mots. La Figure 2 montre une composante du graphe 1-PPV correspondant. Cette composante contient 3 des ensembles de référence ; ceux-ci sont constitués d'unités lexicales qui évoquent 3 cadres sémantiques liés à la gestion des matières résiduelles (Collecting, Processing_materials et Removing).

La Figure 3 illustre des caractéristiques de ce graphe, à savoir la distribution du nombre de sommets par composante et la distribution du nombre de voisins par sommet. Les graphes 1-PPV contiennent un petit nombre de grosses composantes (la plus grosse contenant en moyenne 179 sommets) et un nombre élevé de petites composantes, dont beaucoup ne contiennent que 2 sommets ; ces composantes d'ordre 2 représentent des PPV *réciroques* ou *mutuels* (paires de mots dont chacun est le PPV de l'autre). Le nombre moyen de composantes d'ordre 2 était de 342 pour les modèles HAL, 360 pour les modèles SVD et 389 pour les modèles W2V, ce qui suggère que W2V produit un nombre plus élevé de PPV mutuels.

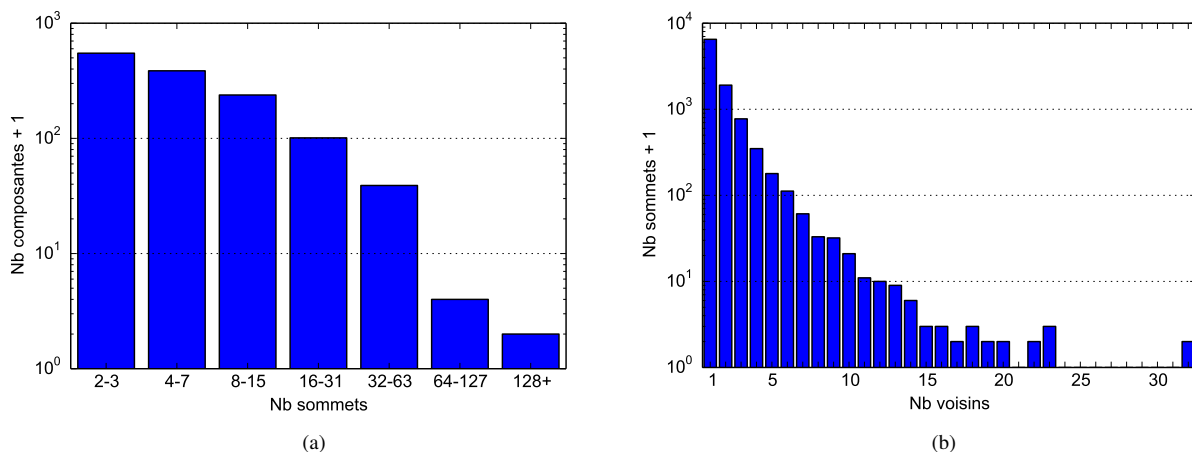


FIGURE 3: (a) Distribution du nombre de sommets par composante. (b) Distribution du nombre de voisins par sommet.

Rang	Sommets	Mots les plus proches du vecteur moyen
1	196	pouvoir, falloir, aller, vouloir, devoir, jamais, faire, déjà, commencer, ...
2	107	of, for, and, the, research, environmental, policy, water, to, ...
3	86	espèce, oiseau, mammifère, animal, poisson, menacer, population, sauvage, rare, ...
4	78	michel, jacques, françois, alain, patrick, paul, dominique, pierre, andré, ...
5	61	substance, toxique, chimique, produit, contenir, polluant, molécule, composé, dangereux, ...
6	61	maïs, céréale, culture, soja, blé, riz, cultiver, colza, pomme, ...
7	61	philippe, jean, bernard, christian, daniel, pascal, jean-pierre, jean-claude, gérard, ...
8	58	déterminer, définir, établir, fixer, évaluer, examiner, décrire, étudier, mesurer, ...
9	55	penser, croire, savoir, peur, apprendre, imaginer, rêver, oublier, douter, ...
10	54	paris, lyon, édition, lille, rennes, toulouse, marseille, bordeaux, montpellier, ...
11	54	autorisation, permis, autoriser, agrément, délivrer, certificat, dérogation, interdiction, interdire, ...
12	53	arbre, plante, feuille, fleur, arbuste, herbe, racine, végétal, graine, ...
13	51	réduire, diminuer, augmenter, limiter, accroître, réduction, baisser, minimiser, croître, ...
14	50	organiser, participer, lancer, réunir, rassembler, annoncer, initier, regrouper, dérouler, ...
15	49	chercheur, scientifique, expert, biologiste, spécialiste, équipe, économiste, auteur, journaliste, ...

TABLE 1: Mots centraux des composantes comprenant le plus grand nombre de sommets.

5 Applications du graphe

Dans la section précédente, nous avons montré que le calcul des composantes connexes du graphe 1-PPV, un moyen simple d’obtenir des ensembles de mots distributionnellement similaires, permet dans certains cas d’identifier des ensembles d’unités lexicales évoquant un même cadre sémantique. Le graphe peut également être utilisé de différentes façons afin d’explorer ou de caractériser les voisinages sémantiques captés par un modèle distributionnel. Par exemple, on peut observer les composantes qui contiennent le plus grand nombre de sommets. Les plus grosses composantes du graphe que nous avons retenu sont présentées dans le Tableau 1, chaque composante étant illustrée au moyen des mots les plus proches du vecteur moyen des mots dans la composante.

Une autre caractéristique intéressante du graphe 1-PPV est le degré (nombre de voisins) des sommets. Par exemple, les sommets ayant le plus de voisins dans le graphe que nous avons retenu sont des mots très fréquents : *pouvoir* (32 voisins), *gens* (23), *aller* (23), *philippe* (22), *oiseau* (20), *of* (19), *insecte* (18), *croire* (18), *falloir* (17) et *chose* (16). En revanche, si on applique la SVD à ce modèle, les sommets qui ont le plus de voisins dans le graphe 1-PPV comprennent des mots beaucoup moins fréquents, dont plusieurs prénoms ; soulignons aussi que le degré maximal du modèle HAL est deux fois plus élevé que celui du modèle SVD correspondant. Dans le cas du modèle W2V qui obtient le meilleur rappel, les sommets ayant le plus de voisins sont des mots à très faible fréquence : *connerie* (26 voisins), *poésie* (16), *sabine* (15), *objectivité* (13), *inadmissible* (12), *cruauté* (12), *economic* (12), *michèle* (12), *with* (11), *cendrer* (11).

Le graphe est également utile à des fins de visualisation. Par exemple, pour visualiser le voisinage d’un mot, on peut vérifier à quelle composante il appartient et produire une figure comme la Figure 2. Si la composante qui contient une requête donnée ne contient pas suffisamment de sommets pour illustrer adéquatement le voisinage de la requête, on peut vérifier quelles autres composantes sont proches de la requête⁷. Par exemple, la composante contenant le mot *absorber* contient 13 mots ; en calculant les 3 autres composantes les plus proches du vecteur de *absorber*, on obtient les 4 composantes illustrées dans la Figure 4.

Le graphe 1-PPV offre de nombreuses possibilités. Par exemple, il serait possible de calculer un deuxième graphe 1-PPV sur les vecteurs moyens des composantes du graphe 1-PPV pour obtenir une représentation plus abstraite des voisinages sémantiques, une possibilité que nous avons commencé à explorer. On pourrait également imaginer différentes façons d’améliorer la cohésion sémantique des ensembles de mots similaires que représentent les composantes du graphe 1-PPV.

7. La distance entre une composante et la requête peut être estimée de différentes façons ; nous utilisons la similarité entre le vecteur de la requête et le vecteur moyen des mots dans une composante donnée, mesurée au moyen du cosinus de l’angle des vecteurs.

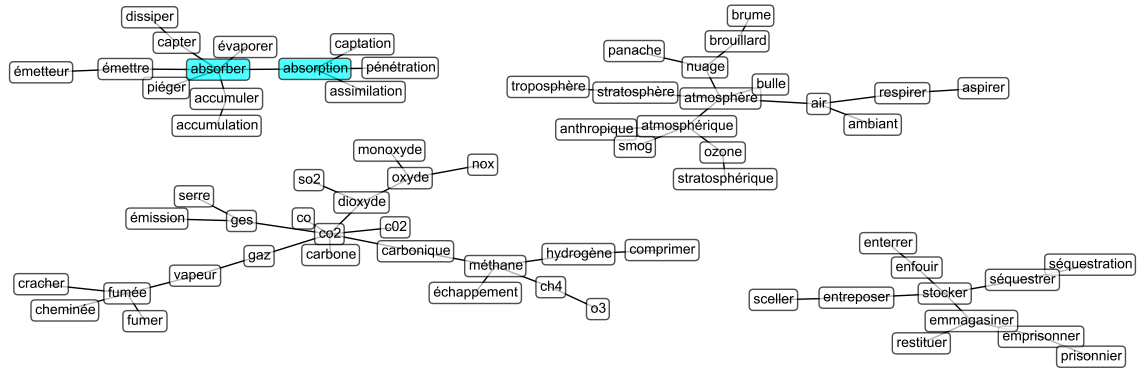


FIGURE 4: Composante contenant la requête *absorber*, accompagnée des 3 autres composantes les plus proches du vecteur de la requête. Les mots *absorber* et *absorption* (en bleu), forment un ensemble de référence (cadre Soaking_up).

6 Travaux reliés

À notre connaissance, il existe peu de travaux qui ont cherché à exploiter les graphes de voisinage afin d'explorer des modèles distributionnels ou d'identifier des ensembles de mots sémantiquement reliés. Gyllensten & Sahlgren (2015) soulignent que la méthode généralement utilisée pour interroger un modèle distributionnel, qui consiste à obtenir une liste ordonnée de voisins pour un mot donné, ne rend pas compte de la structure interne du voisinage du mot. Ils proposent d'utiliser un graphe de voisinage relatif pour décrire le voisinage d'un mot d'une façon qui rend compte de ses différents sens ; ils utilisent notamment cette méthode pour comparer les propriétés de différents modèles distributionnels. Des méthodes basées sur graphe ont été utilisées dans plusieurs travaux visant à découvrir les différents sens ou usages des mots à partir de corpus ; ces méthodes exploitent généralement un graphe de cooccurrence (Dorow & Widdows, 2003; Véronis, 2003; Biemann, 2006; Di Marco & Navigli, 2013), mais des graphes de similarité distributionnelle ont également été utilisés (Feret, 2004). Morardo & Villemonte de La Clergerie (2013) présentent une plateforme de production, de visualisation et de validation de ressources lexicales dont une des composantes principales permet de construire des réseaux lexicaux basés sur la similarité distributionnelle des termes. En ce qui concerne les réseaux lexicaux, Steyvers & Tenenbaum (2005) ont analysé la structure de trois réseaux différents afin de modéliser l'acquisition et l'évolution du lexique, et ont comparé les propriétés de ces réseaux à celles des graphes de voisinage produits au moyen de l'analyse sémantique latente (Landauer & Dumais, 1997). En outre, Claveau *et al.* (2014) considèrent les thésaurus distributionnels comme des graphes k-PPV et exploitent l'information contenue dans ces graphes afin d'améliorer la qualité des thésaurus. Enfin, nous ne connaissons aucun travail portant spécifiquement sur les graphes 1-PPV des modèles distributionnels.

7 Conclusion

Dans cet article, nous avons montré que les composantes connexes d'un graphe 1-PPV symétrique offrent différents moyens d'explorer les voisinages sémantiques captés par un modèle distributionnel et de comparer différents modèles. Nous avons montré que ces composantes, qui représentent des ensembles de mots distributionnellement similaires, permettent dans certains cas d'identifier des ensembles d'unités lexicales qui évoquent un même cadre sémantique. Une évaluation plus approfondie serait nécessaire pour déterminer plus précisément dans quelle mesure les graphes de voisinage distributionnel peuvent faciliter l'identification de ces ensembles. À ce titre, nous comptons mettre à l'épreuve différents types de graphes de voisinage et continuer à développer notre méthodologie d'évaluation afin de mieux évaluer l'efficacité de ces méthodes dans le cadre de l'élaboration de ressources lexicales spécialisées.

Remerciements

Ce projet bénéficie du soutien financier du Conseil de recherches en sciences humaines (CRSH) du Canada.

Références

- BIEMANN C. (2006). Chinese whispers : an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the first workshop on graph based methods for natural language processing*, p. 73–80 : ACL.
- BULLINARIA J. A. & LEVY J. P. (2007). Extracting semantic representations from word co-occurrence statistics : A computational study. *Behavior research methods*, **39**(3), 510–526.
- CLAVEAU V., KIJAK E. & FERRET O. (2014). Explorer le graphe de voisinage pour améliorer les thésaurus distributionnels. In B. BIGI, Ed., *Actes de TALN 2014 (Traitement automatique des langues naturelles)*, p. 220–231, Marseille : ATALA LPL.
- DI MARCO A. & NAVIGLI R. (2013). Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, **39**(3), 709–754.
- DOROW B. & WIDDOWS D. (2003). Discovering corpus-specific word senses. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics – Volume 2*, p. 79–82 : ACL.
- FERRET O. (2004). Découvrir des sens de mots à partir d’un réseau de cooccurrences lexicales. In P. BLACHE, Ed., *Actes de TALN 2004 (Traitement automatique des langues naturelles)*, Fès, Maroc : ATALA LPL.
- FILLMORE C. J. (1982). Frame semantics. In THE LINGUISTIC SOCIETY OF KOREA, Ed., *Linguistics in the Morning Calm : Selected Papers from SICOL-1981*, p. 111–137. Seoul : Hanshin Publishing Co.
- GYLLENSTEN A. C. & SAHLGREN M. (2015). Navigating the semantic horizon using relative neighborhood graphs. *CoRR*, **abs/1501.02670**.
- HARRIS Z. S. (1954). Distributional structure. *Word*, **10**(2–3), 146–162.
- LANDAUER T. K. & DUMAIS S. T. (1997). A solution to Plato’s problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**(2), 211.
- L’HOMME M.-C. & ROBICHAUD B. (2014). Frames and terminology : Representing predicative terms in the field of the environment. In *Proceedings of CogALex*, p. 186–197, Dublin : ACL, DCU.
- L’HOMME M.-C., ROBICHAUD B. & SUBIRATS RÜGGEBERG C. (2014). Discovering frames in specialized domains. In *Proceedings of LREC*, p. 1364–1371, Reykjavik : ELRA.
- LUND K., BURGESS C. & ATCHLEY R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, p. 660–665.
- MAIER M., HEIN M. & VON LUXBURG U. (2007). Cluster identification in nearest-neighbor graphs. In *Algorithmic Learning Theory*, p. 196–210 : Springer.
- MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of ICLR*.
- MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, p. 3111–3119.
- MORARDO M. & VILLEMONT DE LA CLERGERIE É. (2013). Vers un environnement de production et de validation de ressources lexicales sémantiques. In *Actes de SemDis 2013 : Enjeux actuels de la sémantique distributionnelle*, p. 167–180, Les Sables d’Olonne, France.
- RUPPENHOFER J., ELLSWORTH M., PETRUCK M. R. L., JOHNSON C. R. & SCHEFFCZYK J. (2010). FrameNet II : Extended theory and practice. <http://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf>.
- SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.
- SCHÜTZE H. (1992). Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing (Supercomputing’92)*, p. 787–796 : IEEE Computer Society Press.
- STEYVERS M. & TENENBAUM J. B. (2005). The large-scale structure of semantic networks : Statistical analyses and a model of semantic growth. *Cognitive science*, **29**(1), 41–78.
- VÉRONIS J. (2003). Cartographie lexicale pour la recherche d’information. In B. DAILLE, Ed., *Actes de TALN 2003 (Traitement automatique des langues naturelles)*, p. 265–274, Batz-sur-mer : ATALA IRIN.

Apport de l'information temporelle des contextes pour la représentation vectorielle continue des mots

Killian Janod², Mohamed Morchid¹, Richard Dufour¹, Georges Linares¹

¹LIA - University of Avignon (France)

²ORKIS - Aix en Provence (France)

¹firstname.lastname@univ-avignon.fr, ²killian.janod@orkis.com

Résumé. Les représentations vectorielles continues des mots sont en plein essor et ont déjà été appliquées avec succès à de nombreuses tâches en traitement automatique de la langue (TAL). Dans cet article, nous proposons d'intégrer l'information temporelle issue du contexte des mots au sein des architectures fondées sur les sacs-de-mots continus (*continuous bag-of-words* ou *CBOW*) ou sur les Skip-Grams. Ces approches sont manipulées au travers d'un réseau de neurones, l'architecture CBOW cherchant alors à prédire un mot sachant son contexte, alors que l'architecture Skip-Gram prédit un contexte sachant un mot. Cependant, ces modèles, au travers du réseau de neurones, s'appuient sur des représentations en sac-de-mots et ne tiennent pas compte, explicitement, de l'ordre des mots. En conséquence, chaque mot a potentiellement la même influence dans le réseau de neurones. Nous proposons alors une méthode originale qui intègre l'information temporelle des contextes des mots en utilisant leur position relative. Cette méthode s'inspire des modèles contextuels continus. L'information temporelle est traitée comme coefficient de pondération, en entrée du réseau de neurones par le CBOW et dans la couche de sortie par le Skip-Gram. Les premières expériences ont été réalisées en utilisant un corpus de test mesurant la qualité de la relation sémantique-syntactique des mots. Les résultats préliminaires obtenus montrent l'apport du contexte des mots, avec des gains de 7 et 7,7 points respectivement avec l'architecture Skip-Gram et l'architecture CBOW.

Abstract.

Contribution of temporal context information to a continuous vector representation of words

Word embedding representations are gaining a lot of attention from researchers and have been successfully applied to various Natural Language Processing (NLP) tasks. In this paper, we propose to integrate temporal context information of words into the continuous bag-of-words (CBOW) and Skip-gram architectures for computing word-vector representations. Those architectures are shallow neural-networks. The CBOW architecture predicts a word given its context while the Skip-gram architecture predicts a context given a word. However, in those neural-networks, context windows are represented as bag-of-words. According to this representation, every word in the context is treated equally : the word order is not taken into account explicitly. As a result, each word will have the same influence on the network. We then propose an original method that integrates temporal information of word contexts using their relative position. This method is inspired from Continuous Context Models. The temporal information is treated as weights, in input by the CBOW and in the output layer by the Skip-Gram. The quality of the obtained models has been measured using a Semantic-Syntactic Word Relationship test set. Results showed that the incorporation of temporal information allows a substantial quality gain of 5 and 0.9 points respectively in comparison to the classical use of the CBOW and Skip-gram architectures.

Mots-clés : Réseau de neurones, Représentation vectorielle continue, Information contextuelle, Word2vec, Modèle de langue.

Keywords: Neural network, Continuous vectorial representation, Contextual information, Word2vec, language model.

1 Introduction

Les modèles sémantiques représentant des langages projettent leurs termes dans un espace dans lequel les relations sémantiques entre ces termes peuvent être observées ou mesurées. La technique récente des Word2vec (Mikolov *et al.*, 2013a) construit un réseau de neurones permettant de projeter les termes (contenus dans une fenêtre sémantique définie) d’une langue étudiée dans un espace de représentation vectorielle. Cette projection permet aux mots de sens similaires d’être localisés dans une région de l’espace sémantique proche, comme par exemple les termes “Paris” et “Londres”, qui, selon le corpus étudié, peuvent partager l’idée de “Capitale”.

Une telle représentation considère le terme dans un environnement contenant un nombre restreint de mots. De plus, ce voisinage réduit ne permet pas de coder d’éventuelles relations entre les termes. En effet, ce groupe de termes est considéré indépendamment de leurs positions ou en “sac-de-mots”. Malgré les bons résultats observés lors de l’utilisation des méthodes issues des Word2vec lors de la tâche de représentation sémantique, cette représentation vectorielle ne tient pas compte de la disposition des mots dans le contexte. Cette information est pourtant cruciale lorsque l’on définit le “sens” d’un mot en fonction de son contexte. En effet, plus les termes du contexte sont éloignés du terme central à définir, moins leur relation doit avoir d’importance.

Dans ce papier, nous proposons de pallier cette faiblesse en pondérant les termes contenus dans la fenêtre en fonction de leur distance vis-à-vis du terme central à définir. Cette nouvelle représentation sera évaluée lors de tâches similaires à celles définies dans (Mikolov *et al.*, 2013a), pour les deux architectures introduites dans (Mikolov *et al.*, 2013a) : *CBOW* et *Skip-Gram*. Nous montrons ainsi que la position des termes dans un contexte est une information essentielle, permettant de mieux définir le sens d’un terme dans ce contexte.

Le reste de ce papier est organisé comme suit. La section 2 présente un état-de-l’art des différentes représentations de termes. L’approche proposée, fondée sur la technique des Word2vec, est ensuite détaillée dans la section 3. La section 4 présente les expériences ainsi que les résultats observés avant de conclure dans la section 5.

2 Travaux antérieurs

Les “sacs-de-mots” (Salton, 1989) sont aujourd’hui communément utilisés dans de nombreuses tâches du traitement automatique du langage (TAL). Cette représentation a pour particularité de traiter tous les mots de façon identique, ainsi la séquentialité des mots et leurs relations sont ignorées. D’autres approches ont alors été proposées pour réintroduire la séquentialité des mots, comme par exemple l’approche n -grammes. Cette approche vise à prendre en compte, pour un mot donné, les n mots contenus dans son contexte passé. D’autres stratégies ont ensuite émergé pour capturer la proximité sémantique des mots. La plupart d’entre-elles s’appuient sur l’Hypothèse de Distribution (Sahlgren, 2008) qui implique que des mots représentés dans un même contexte ont un sens proche. Certaines de ces méthodes ont débouché sur des représentations vectorielles continues des mots. Ainsi, les méthodes revues par (Baroni & Lenci, 2010) proposent des modèles où les mots sont représentés par leurs relations à l’ensemble des contextes du corpus d’apprentissage. Cette représentation génère cependant souvent des vecteurs de grande dimension et creux. D’autres stratégies, telles que l’allocation latente de Dirichlet (LDA) (Blei *et al.*, 2003), consistent à découvrir les thèmes latents d’un corpus de texte puis à projeter les mots dans ces espaces thématiques. Cette représentation vectorielle associe pour chaque mot sa distribution dans l’ensemble des thèmes. Ces méthodes ont souvent besoin d’être employées avec une décomposition en valeur singulière, pour réduire leur dimensionnalité, éviter les matrices creuses, et conserver un maximum de la significativité.

Récemment, de nouvelles méthodes de représentation de mots fondées sur les réseaux de neurones se sont développées. Chaque mot est alors représenté par un vecteur plein, de taille modéré, qui correspond à une projection du mot dans un espace où les distances modélisent les relations inter-mots. Ces méthodes sont principalement issues des modèles de langue neuronaux (Bengio *et al.*, 2003; Collobert *et al.*, 2011) et sont déjà utilisées dans plusieurs tâches du TAL (Do *et al.*, 2014; Vaswani *et al.*, 2013). L’approche Word2vec (Mikolov *et al.*, 2013a), fondée elle-aussi sur les réseaux de neurones, suscite un intérêt grandissant dans le domaine du TAL. Une méthode similaire, appelée Glove (Pennington *et al.*, 2014), consiste à factoriser une matrice de co-occurrence des mots appris sur une grande quantité de textes. La limite de ces approches Word2vec et Glove est que la position des mots dans une séquence n’est pas prise en compte : l’ensemble de la séquence de mots est considérée comme un “sac-de-mots”, l’ordre des mots étant ignoré.

Nous proposons dans cet article de pallier cette faiblesse en introduisant l’information temporelle des mots (*i.e.* un poids selon leur position) dans la représentation vectorielle continue des mots Word2vec qui l’ignore pour l’instant.

3 Approche proposée

3.1 Méthode d'origine : les sacs-de-mots continus et Skip-Gram de Word2vec

Word2vec est une méthode fondée sur des réseaux de neurones artificiels et définie dans (Mikolov *et al.*, 2013a). Cette méthode propose deux architectures de réseaux de neurones : l'architecture en "sac-de-mots" continu (CBOW) et l'architecture Skip-gram. Ces deux architectures se présentent sous la forme de réseaux de neurones artificiels simples. Ils sont constitués de trois couches : une couche d'entrée, une couche cachée et une couche de sortie. La couche d'entrée contient soit un "sac-de-mots" (CBOW), soit un mot seul (Skip-gram). La couche cachée correspond à la projection des mots d'entrée dans la matrice des poids. Cette matrice est partagée par tous les mots (matrice globale). Enfin, la couche de sortie est composée de neurones "softmax". Pour des raisons de complexité algorithmique due à la couche "softmax", les auteurs dans (Mikolov *et al.*, 2013b) ont introduit deux alternatives appelées "échantillons négatifs" et "softmax hiérarchique". Le couplage de ces fonctions avec la simplicité de ces réseaux leur permettent d'être entraînés sur de très grandes quantités de textes, et ainsi d'obtenir des modélisations de meilleure qualité que les modèles plus complexes à base de récurrence ou de convolution par exemple (Mikolov *et al.*, 2013b, 2010). Un exemple de ces deux architectures est présenté dans la figure 1.

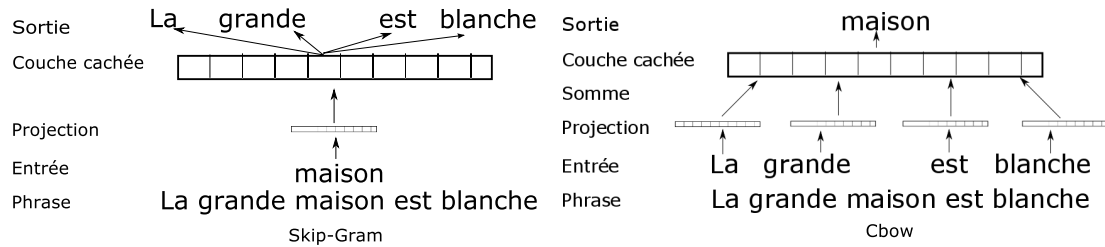


FIGURE 1 – Exemples des architectures CBOW et Skip-gram de Word2vec.

Ces modèles sont capables de capturer des régularités sémantiques et syntaxiques (Mikolov *et al.*, 2013c). En effet, la distance qui sépare la projection de deux mots peut représenter une relation complexe telle que la notion de "singulier-pluriel" ou "masculin-féminin" (Mikolov *et al.*, 2013a) comme le montre la figure 2.

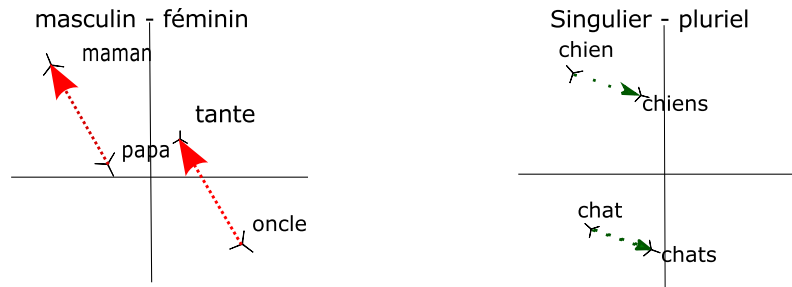


FIGURE 2 – Exemples de relations de mots dans l'espace Word2vec.

3.1.1 Approche par sac-de-mots continus (CBOW)

L'architecture CBOW est un réseau de neurones devant prédire un mot à partir de son contexte. La couche d'entrée représente la présence ou l'absence des mots dans le contexte de manière binaire (*i.e.* 1 pour la présence, 0 pour l'absence). Chaque mot dans le contexte est projeté dans la matrice des poids du modèle. La somme (ou moyenne) de ces représentations passe ensuite par la couche de sortie. Enfin, le modèle compare sa sortie avec le mot seul et corrige sa représentation par rétro-propagation du gradient. Ce modèle cherche à maximiser l'équation :

$$\frac{1}{T} \sum_{t=1}^T \log p(m_t | m_{t-\frac{c}{2}} \dots m_{t+\frac{c}{2}}) \quad (1)$$

où T correspond à l'ensemble des mots dans le corpus et c correspond à la taille de la fenêtre du contexte de chaque mot. Cette architecture présente plusieurs avantages : en effet, en plus d'être efficiente d'un point-de-vue algorithmique (Mikolov *et al.*, 2013a), elle permet à la fois une meilleure modélisation des mots fréquents et une meilleure capture des relations syntaxiques.

3.1.2 Approche par Skip-Gram

L'architecture Skip-Gram tente de prédire, pour un mot donné, le contexte dont il est issu. La couche d'entrée de ce réseau est alors un vecteur ne contenant qu'un seul mot. Le mot est projeté dans la couche cachée puis dans la couche de sortie. Le contexte est ensuite réduit de façon aléatoire à chaque itération. Le vecteur de sortie est ensuite comparé à chacun des mots du contexte réduit et le réseau se corrige par rétro-propagation du gradient. De cette manière, la représentation du mot d'entrée va se rapprocher de chacun des mots présents dans le contexte.

Le réseau de neurones Skip-gram essaie de maximiser une variation de l'équation 1 comme suit :

$$\frac{1}{T} \sum_{t=1}^T \sum_{j=t-c, j \neq t}^{t+c} \log p(m_j | m_t) : \quad (2)$$

Comparativement au CBOW, cette architecture permet une meilleure modélisation des mots peu fréquents et permet de mieux capturer les relations sémantiques (Mikolov *et al.*, 2013a).

3.2 Évolution proposée : Distance inter-contexte

L'architecture CBOW traite les mots du contexte en "sacs-de-mots", négligeant l'information chronologique d'apparition des mots et donc de distance dans le contexte. L'architecture Skip-Gram utilise implicitement cette information en réduisant de façon aléatoire la taille du contexte à chaque itération. Par conséquent, la taille du contexte devient un paramètre d'autant plus important puisqu'une fenêtre trop grande dévalorisera le contexte proche des mots et une fenêtre trop courte ne pourra capturer des relations éloignées. Pour répondre à cette problématique, nous avons ajouté une information temporelle de distance inter-contexte (Bigot *et al.*, 2013a,b), le contexte étant centré sur un mot à prédire. Chacun des éléments du contexte se voit attribuer une pondération selon la distance qui le sépare du mot central comme suit :

$$\frac{\alpha}{b + \beta \log(d)} \quad (3)$$

où d correspond à la distance en nombre de mots entre le mot au centre du contexte et l'élément du contexte à pondérer. α , b , β et d sont utilisés comme coefficients pour faire varier l'importance de l'information temporelle. Ainsi les contextes proches se voient renforcés et les contextes éloignés auront un impact s'ils sont suffisamment fréquents.

4 Expériences

Nous proposons de comparer les performances des modèles classiques (*i.e.* sans tenir compte de l'information temporelle) et des modèles intégrant la position des mots du contexte dans le réseau de neurones (*i.e.* avec information temporelle) pour les architectures CBOW et Skip-gram de la méthode Word2vec. Les modèles sont entraînés à partir d'une très grande quantité de données (partie 4.1) puis évalués au travers de différentes configurations sur un corpus de test permettant de mesurer la qualité de la relation sémantique-syntaxique des mots (voir partie 4.2).

4.1 Protocole expérimental

Quatre corpus ont été utilisés pour l'apprentissage des modèles (Mikolov *et al.*, 2013a) :

- Le corpus One Billion Word Language Modeling Benchmark de 30 914 405 documents (700 260 470 mots).
- Le premier million de caractères de Wikipedia anglais de 124 303 documents (124 301 845 mots).

- Le corpus GigaWord Anglais de 1994 à 2011 de 190 344 429 documents (3 771 326 692 mots).
- Le Brown Corpus de 57 341 documents (1 019 149 mots).

Pour comparer les performances de notre approche, nous avons choisi de faire varier les paramètres d'apprentissages du réseau de neurones, à savoir la taille du contexte utilisé (5 mots, 10 mots, et le document entier), la taille de la couche cachée du réseau de neurones (120 et 300 neurones), et la fonction de distance utilisée pour l'information temporelle ($\frac{1+\log(2)}{1+\log(d)}$ et $\frac{\log(10)}{5*\log(d)}$ appelées respectivement *distance 1* et *distance 2*). Pour chaque condition d'apprentissage, deux modèles sont appris : le modèle classique et le modèle que nous proposons intégrant l'information temporelle du contexte des mots.

Chaque modèle est ensuite évalué avec la tâche de recherche de mots analogues définie dans (Mikolov *et al.*, 2013a). Cette tâche vérifie, pour des ensembles de couples de mots partageant une même relation, que le modèle a bien appris la relation en question. Par exemple, la relation "Capital de" est représentée par la distance entre la paire de mots "Paris" et "France", et par la distance entre la paire de mots "Rome" et "Italie". Pour vérifier qu'une relation similaire lie les deux mots de chaque paire, la question suivante est projetée dans l'espace du modèle par : $Paris - France + Italie = Rome$, qui se traduit par la proximité du vecteur *Rome* avec le vecteur formé par $Paris - France + Italie$. La tâche est constituée de 19 000 paires de couples comme celle-ci modélisant plusieurs relations différentes comme capitale, monnaie, genre, adjectif opposé, singulier-pluriel. . Une fois toutes les relations évaluées, un score est attribué au modèle correspondant au pourcentage de relations correctement modélisées.

4.2 Résultats

Globalement, les tableaux 1, 2 et 3 montrent que les modèles intégrant l'information temporelle sont toujours meilleurs que les modèles de base. Le tableau 2 démontre que plus le contexte utilisé est grand, plus la fonction de distance apporte de l'information jusqu'à l'utilisation des documents entiers comme contexte. Dans notre corpus, moins de 1 % des documents ont plus de 100 mots. Nous utilisons donc un contexte de taille 100 pour prendre en compte le document entier comme contexte. Nos meilleurs gains sont ainsi obtenus avec le document entier (7 points sur le CBOW et 7,7 points sur le Skip-Gram).

Le tableau 2 nous indique que la fonction de distance 2 est plus favorable au modèle CBOW, avec un gain de 4,3 %, mais moins favorable pour le Skip-Gram, avec un gain de 2,1 %. Dans ces mêmes conditions, avec la distance 1, le gain ne dépasse pas 0,9 % pour le CBOW, alors que le gain pour le Skip-Gram atteint 5 %. Enfin dans le tableau 3, nous observons que les modèles possédant une petite couche cachée, ont un gain plus faible que ceux ayant une couche cachée de plus grande taille.

	Skip-gram			CBOW		
Nb de neurones	300					
Taille du contexte	10	15	100	10	15	100
Sans Distance	50,0	50,9	43,7	39	38,9	36,9
Avec Distance 1	55,0	53,7	51,4	39,9	39,6	43,9

TABLE 1 – Performance des modèles selon la taille du contexte des mots (en %).

	Skip-gram	CBOW
Nb de neurones	300	
Taille du contexte	10	
Sans Distance	50,0	39,0
Avec Distance 1	55,0	39,9
Avec Distance 2	52,1	43,3

TABLE 2 – Comparaison des performances selon les distances (en %).

En observant manuellement les exemples de mots des modèles (voir tableau 4), nous remarquons que les modèles avec distance ont tendance à regrouper entre eux des voisins similaires, au contraire des modèles classiques qui ne semblent pas les regrouper. Par exemple, pour "Holidays", l'intégration de la distance permet de regrouper "holiday", "vacation" et "festivities", qui rappellent des mots autour des vacances, "thanksgiving", "easter" et "christmas" à des fêtes religieuses particulières. Ces *catégories* sont moins marquées en l'absence de l'information de distance.

	Skip-gram		CBOW	
Taille du contexte	10			
Taille couche cachée	300	120	300	120
Sans Distance	50,0	43,9	39,0	29,0
Avec Distance 1	55,0	45,1	39,9	30,3

TABLE 3 – Performance sans et avec information temporelle dans un petit espace de projection (10 mots).

Holidays		Meat		Motherboard	
Avec Distance	Sans Distance	Avec Distance	Sans Distance	Avec Distance	Sans Distance
Holiday	vacations	chicken	pork	cpu	cpu
vacation	thanksgiving	beef	not-pasterised	cpus	chipset
festivities	vacation	pork	mutton	microprocessor	geforce4
thanksgiving	christmas	milk	eggs	chips	microprocessor
easter	celebration	eggs	cattle	agp	pentium-m
christmas	easter	seafood	chicken	peripherals	cpus

TABLE 4 – Exemples de similarités obtenues pour des mots particuliers sans et avec information temporelle (distance).

5 Conclusion

Ce papier propose une méthode renforçant l'information temporelle des contextes des mots dans les représentations vectorielles. En effet, ces approches, telles que l'approche Word2vec, considèrent l'ensemble des mots d'une séquence indépendamment de leurs positions dans celle-ci. Cette approche en "sac-de-mots" ne permet donc pas de conserver la structure temporelle de la séquence, chaque mot ayant la même importance dans le réseau de neurones. Nous avons alors proposé de pondérer, dans ce réseau de neurones, les termes de la séquence considérée en fonction de leur distance vis-à-vis du terme central à définir. Les expériences préliminaires, menées sur un corpus de test mesurant la qualité de la relation sémantique-syntactique des mots, montrent l'apport du contexte des mots, avec des gains de 7 et 7,7 points respectivement avec l'architecture Skip-Gram et l'architecture CBOW. Nous prévoyons, en perspective, d'étendre cette étude en évaluant le nombre de mots du contexte de manière plus précise ainsi que d'évaluer l'impact d'autres méthodes de poids pour l'intégration de l'information temporelle.

Références

- BARONI M. & LENCI A. (2010). Distributional memory : A general framework for corpus-based semantics. *Computational Linguistics*, **36**(4), 673–721.
- BENGIO Y., DUCHARME R. & VINCENT P. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, **3**, 1137–1155.
- BIGOT B., LINARÈS G., FREDOUILLE C., DUFOUR R. & LIA C. (2013a). Combining Acoustic Name Spotting and Continuous Context Models to improve Spoken Person Name Recognition in Speech. *Interspeech*, p. 2539–2543.
- BIGOT B., SENAY G., LINARÈS G., FREDOUILLE C. & DUFOUR R. (2013b). Person name recognition in asr outputs using continuous context models. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 8470–8474 : IEEE.
- BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, **3**, 993–1022.
- COLLOBERT R., WESTON J., BOTTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural Language Processing (almost) from Scratch. *The Journal of Machine Learning Research*, **12**, 2493–2537.
- DO Q.-K., ALLAUZEN A. & YVON F. (2014). Modèles de langue neuronaux : une comparaison de plusieurs stratégies d'apprentissage. In *TALN 2014*.
- MIKOLOV T., CORRADO G., CHEN K. & DEAN J. (2013a). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, p. 1–12.

- MIKOLOV T., KARAFIÁT M., BURGET L., CERNOCKÝ J. & KHUDANPUR S. (2010). Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, p. 1045–1048.
- MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, p. 3111–3119.
- MIKOLOV T., YIH W.-T. & ZWEIG G. (2013c). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, p. 746–751.
- PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)*, **12**.
- SAHLGREN M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, **20**(1), 33–54.
- SALTON G. (1989). Automatic text processing : the transformation. *Analysis and Retrieval of Information by Computer*.
- VASWANI A., ZHAO Y., FOSSUM V. & CHIANG D. (2013). Decoding with large-scale neural language models improves translation. In *EMNLP*, p. 1387–1392 : Citeseer.

Etiquetage morpho-syntaxique de tweets avec des CRF

Tian Tian^{1,2} Marco Dinarelli¹ Isabelle Tellier¹ Pedro Cardoso²

(1) Lattice, UMR 8094, 1 rue Maurice Arnoux, 92120 Montrouge

(2) Synthesio, 8-10 rue Villedo, 75001 Paris

ttian@synthesio.com, pedro@synthesio.com, isabelle.tellier@univ-paris3.fr, marco.dinarelli@ens.fr

Résumé. Nous nous intéressons dans cet article à l'apprentissage automatique d'un étiqueteur morpho-syntaxique pour les tweets en anglais. Nous proposons tout d'abord un jeu d'étiquettes réduit avec 17 étiquettes différentes, qui permet d'obtenir de meilleures performances en exactitude par rapport au jeu d'étiquettes traditionnel qui contient 45 étiquettes. Comme nous disposons de peu de tweets étiquetés, nous essayons ensuite de compenser ce handicap en ajoutant dans l'ensemble d'apprentissage des données issues de textes bien formés. Les modèles mixtes obtenus permettent d'améliorer les résultats par rapport aux modèles appris avec un seul corpus, qu'il soit issu de Twitter ou de textes journalistiques.

Abstract.

Part-of-speech Tagging for Tweets with CRFs

We are interested in this paper in training a part-of-speech tagger for tweets in English. We first propose a reduced tagset with 17 different tags, which allows better results in accuracy than traditional tagsets which contain 45 tags. Since we have few annotated tweets, we then try and overcome this difficulty by adding data from other more standard texts into the training set. The obtained models reach better results compared to models trained with only one corpus, whether coming of Twitter or of journalistic texts.

Mots-clés : tweets, CRF, étiquetage morpho-syntaxique.

Keywords: tweets, CRFs, part-of-speech tagging.

1 Introduction

Les réseaux sociaux sont devenus la principale source de textes générés par des utilisateurs sur Internet. Ces textes constituent des données massives potentiellement porteuses de beaucoup d'information, mais aussi difficiles à traiter automatiquement du fait de leur grande variété en genres (textes de blogs, forums, tweets...), domaines et styles. Après la phase préliminaire de tokenisation, l'étiquetage morpho-syntaxique de ces textes apparaît comme une étape fondamentale de traitement, permettant d'éventuelles analyses syntaxiques ultérieures.

Dans cet article, nous nous intéressons à l'étiquetage morpho-syntaxique des tweets, qui présentent souvent le plus grand écart à la norme. Nous évoquons dans un premier temps la spécificité de ces données ainsi que les difficultés majeures qui en découlent pour la tâche d'étiquetage morpho-syntaxique. Puis nous introduisons les corpus (en anglais) utilisés dans nos expériences : l'un d'eux est constitué de tweets, l'autre de textes journalistiques plus respectueux de la norme linguistique standard. Pour les traiter, nous proposons d'abord un jeu d'étiquettes morpho-syntaxiques réduit par rapport à ceux utilisés dans les corpus annotés habituellement disponibles. Nous décrivons ensuite l'approche que nous avons adoptée pour apprendre un étiqueteur morpho-syntaxique de tweets anglais avec des CRF. Nous essayons en particulier d'apprendre des modèles à partir de mélanges de textes issus des deux corpus. Les paramètres que nous faisons varier dans nos expériences sont donc : les propriétés prises en compte dans les textes et les patrons qui définissent les fonctions caractéristiques des CRF, les paramètres de régularisation de l'implémentation et les proportions des différents textes sources dans l'ensemble d'apprentissage. Les résultats de nos expériences montrent que notre jeu d'étiquettes permet d'améliorer les performances de l'étiqueteur et qu'un modèle appris avec un mélange de tweets et de textes de journaux donne de meilleurs résultats que ceux appris sur un seul type de textes.

2 Tâche et état de l’art

Le cadre général de ce travail est celui de l’analyse d’opinion portant sur des noms de produits ou de marques cités dans des tweets. L’étiquetage morpho-syntaxique est un préalable nécessaire car nous souhaitons identifier les produits/marques évoqués (présents sous la forme de noms communs ou de noms propres) ainsi que les mots éventuellement porteurs de sentiments (principalement les adjectifs, les verbes et les adverbess). Le but de notre étiqueteur morpho-syntaxique est donc de différencier les grandes classes de catégories grammaticales, pas de vérifier des propriétés morpho-syntaxiques fines comme les accords en genre et en nombre ni de préparer une analyse syntaxico-sémantique profonde.

Malgré cette simplification, l’étiquetage automatique des tweets reste difficile, surtout à cause de leur caractère mal formé, qu’illustre par exemple la figure 1.

Today wasz Fun cuzz anna Came juss for me <3(: hahaha

FIGURE 1 – Exemple 1 de tweet

Le tweet de la figure 1 est issu du corpus (Ritter *et al.*, 2011). La phrase “correcte” devrait être :

Today was fun because Anna came just for me <3(: hahaha

Dans cet exemple, les difficultés sont multiples :

- présence de fautes d’orthographe : wasz (was), cuzz (because), juss (just)
- inversion majuscule/minuscule : Fun (fun), anna (Anna), Came (came)
- émoticon : <3(:
- interjection : hahaha

Dans un dictionnaire anglais, les mots comme "wasz", "cuzz", "juss", "<3(:" et "hahaha" n’existent probablement pas. Les étiqueteurs à base de règles, fondées sur des listes de mots associés à leurs catégories (comme was : verbe) ne fonctionneront donc pas avec ce genre de tweets, à moins de mises à jour massives des ressources qu’elles exploitent, ou de pré-traitements permettant de “corriger” les textes initiaux. Plutôt que de chercher à réaliser ce genre de pré-traitements, nous choisissons d’étiqueter ce type de textes comme s’ils étaient “bien écrits”. Ceci revient à considérer que "wasz" est une variante possible du mot "was", etc. Pour cela, nous allons compter sur les capacités de généralisation de l’apprentissage automatique.

Eduardo Surita : your a freaking ... <http://tumblr.com/xmciuda0t>

FIGURE 2 – Exemple 2 de tweet

La figure 2 illustre une autre difficulté de l’étiquetage morpho-syntaxique des tweets. En anglais standard, il faudrait écrire "you’re" au lieu de "your". Le mot “your” de ce tweet est ainsi porteur de deux catégories morpho-syntaxiques : pronom et verbe. Pour remédier à ce genre de problèmes, plusieurs solutions sont possibles :

- ajouter un pré-traitement de normalisation afin de substituer "you’re" à "your" et traiter ensuite le tweet comme du texte écrit standard ;
- annoter le texte tel qu’il est, avec une seule étiquette syntaxique pour "your". Dans ce cas, on peut soit créer une étiquette nouvelle spéciale (pronom+verbe) comme dans (Gimpel *et al.*, 2011), soit choisir une étiquette “traditionnelle” (par exemple "pronom") comme dans (Ritter *et al.*, 2011). Cette dernière solution entraînera la possibilité de séquences d’étiquettes en principe interdites pour certaines phrases, comme "pronom déterminant adjectif", signe d’une construction sans verbe.

Ces deux exemples expliquent que les performances des étiqueteurs appris sur des corpus “bien formés” comme le Penn TreeBank (Marcus *et al.*, 1993) (aussi appelé PTB par la suite) ou le French TreeBank (Abeillé *et al.*, 2003) chutent quand ils sont confrontés à des tweets. Le Maximum Entropy POS Tagger¹ appris avec le Penn TreeBank dans (Toutanova & Manning, 2000) a obtenu 96.86% d’exactitude en validation croisée mais seulement 81.3% sur les tweets (Ritter *et al.*, 2011). Les expériences menées sur le français montrent des résultats similaires : l’étiqueteur appris sur le French TreeBank

1. Stanford Pos Tagger : <http://nlp.stanford.edu/software/tagger.shtml>

dans (Constant *et al.*, 2011) atteint 97.3% d'exactitude en validation croisée, alors que celui utilisé dans (Nooralahzadeh *et al.*, 2014) a eu seulement un score de 91.7% sur le corpus French Social Media (Seddah *et al.*, 2012).

Beaucoup de travaux ont été consacrés à l'amélioration du traitement des tweets. (Foster *et al.*, 2011), par exemple, essaient d'apprendre un analyseur syntaxique qui leur est dédié. (Gimpel *et al.*, 2011) propose d'utiliser un tokeniseur spécifique différent et un jeu d'étiquettes particulier (par exemple nom+verbe pour les tokens comme "I'll"). (Ritter *et al.*, 2011) cherchent aussi à construire un étiqueteur morho-syntaxique avec un modèle appris sur un mélange de plusieurs corpus : le Penn TreeBank (Marcus *et al.*, 1993), le NPS IRC Corpus (un corpus de *chatroom* introduit dans (Forsyth, 2007)) et son propre corpus twitter (T-POS). Notre travail se situe dans la même lignée, consistant à mélanger des données issues de corpus de textes bien formés et mal formés pour l'apprentissage. Nous créons ainsi un modèle mixte que nous testons sur les données de Twitter. Notre contribution dans cet article porte sur une proposition de jeu d'étiquettes réduit et sur l'étude des effets de diverses proportions de corpus "standard" et de corpus cible (Twitter) pour apprendre un modèle, associée à une optimisation des paramètres de régularisation du modèle d'apprentissage.

3 Les corpus T-POS, Penn TreeBank et le jeu d'étiquettes universel

(Ritter *et al.*, 2011) ont mis à disposition un corpus Twitter annoté en étiquettes morpho-syntaxiques et en entités nommées. Les sujets de discussions y sont très variés : vie quodidienne, équipes sportives, films, groupes musicaux, etc. Nous avons choisi ce corpus comme référence pour Twitter. La partie annotée est toutefois très limitée : elle ne contient que 787 tweets, soit 15972 tokens. La taille de ce corpus ne permet donc pas d'apprendre un modèle complet.

Ce corpus Twitter contient de nombreux exemples d'une autre spécificité des tweets : la présence de caractères spéciaux désignant des "hashtags" (qui servent à l'indexation des tweets par mots clés), des "retweets" (pour une rediffusion de tweets), des URL et des "usernames" (comptes d'utilisateurs de tweets), aux catégories morpho-syntaxiques souvent ambiguës. Prenons l'exemple des hashtags, repérables à leur symbole "#": ils peuvent se substituer à un mot simple (un adjectif dans l'exemple 3), à un constituant syntaxique complet (exemple 4) ou être simplement un terme d'indexation sans rôle syntaxique précis (exemple 5).

3	My #twitter age is 458 days 0 hours 3 minutes 49 seconds
4	On Thanksgiving after you done eating its #TimeToGetOut unless you wanna help with the dishes
5	New book blogger @GennaSarnak launches weekly feature , Poetry Sunday : http://tinyurl.com/47vbdy5 #Books #Poetry

TABLE 1 – Les hashtags dans les tweets

(Ritter *et al.*, 2011) ont choisi de ne pas traiter les hashtags et autres mots spéciaux utilisés dans Twitter comme des composants linguistiques comme les autres. Ils ont ajouté dans leur corpus de tweets quatre étiquettes spécifiques : USR pour "*at mention*", HT pour "*hashtag*", URL et RT pour "*retweet*" en plus des 45 étiquettes définies dans le Penn TreeBank. Ils n'ont donc pas pris en compte les éventuels rôles syntaxiques des hashtags, illustrés dans les exemples 3 et 4. Ce genre de tweets nécessiterait un étiqueteur morpho-syntaxique particulièrement flexible et robuste.

Notre étiqueteur morpho-syntaxique est construit dans un contexte d'analyse multi-langue. Un seul jeu d'étiquettes pour toutes les langues est dans ce cas nécessaire. Comme, de plus, l'objectif final de notre travail relève de la fouille d'opinion dans des données massives et devrait se passer d'une analyse syntaxique complète de ces données, nous n'avons pas besoin de certaines des distinctions du PennTreebank, comme verbes au passé / verbes au présent, nom commun singulier / pluriel, etc. Cela nous a amené à nous intéresser au jeu d'étiquettes proposé dans (Petrov *et al.*, 2012), qui se veut universel tout en laissant la possibilité de réaliser une analyse syntaxique rudimentaire. Il comporte 12 étiquettes différentes et des correspondances (*mapping*) avec les jeux d'étiquettes utilisés dans 25 TreeBanks de 25 langues différentes sont disponibles. Il a été testé sur le PTB et permet d'obtenir des résultats légèrement meilleurs en exactitude (96.8% contre 96.7% avec les étiquettes originales). Nous l'avons étendu en lui adjoignant les quatre étiquettes dédiées à Twitter utilisées dans (Ritter *et al.*, 2011). Enfin, nous avons rétabli la distinction entre les noms propres et les noms communs (qui sont assimilés dans (Petrov *et al.*, 2012)), afin de préparer le terrain à la tâche ultérieure d'extraction des entités nommées. Le nombre total d'étiquettes que nous considérons est donc finalement de 17 : NUM (nombres), PUNCT, NN (nom commun), NP (nom propre), VB (verbe), ADJ (adjectif), ADV (adverb), DET (déterminant), PRON (pronom), CC (conjonction de coordination), PREPCS (préposition et conjonction de subordination), PRT (particule), X (mot inconnu, interjection, émotion), RT, URL, USR, et HT.

Le tableau 2 montre les correspondances entre ce jeu et les étiquettes du PTB, ainsi qu'avec celles de T-POS. Les diffé-

rences d’étiquettes entre ces deux corpus sont marquées en gras.

Tag universel	Tag Penn TreeBank	Tag T-POS	Remarques
NUM	CD	CD	nombres
PUNCT	" , -LRB- -RRB- . : - ”	" () . , : NONE O LS	
NN	NN NNS	NN NNS	noms communs
NP	NNP NNPS	NNP NNPS	noms propres
VB	MD VB VBD VBG VBN VBP VBZ	MD VB VBD VBG VBN VBP VBZ VPP	verbes
ADJ	AFX JJ JJR JJS	JJ JJR JJS	AFX : <i>Yes</i>
ADV	RB RBR RBS WRB	RB RBR RBS WRB	WRB : <i>where, when</i>
DET	DET PDT PRP\$	PDT DT WDT EX TD	PDT : <i>half</i> , PRP\$: <i>his</i>
PRON	PRP WP	PRP PRP\$ WP WP\$	pronoms
CC	CC	CC	
PREPCS	IN	IN	préposition et CS
PRT	POS TO RP	POS TO RP	<i>particule</i>
X	# \$ FW NIL SYM INTJ	INTJ SYM FW	FW : mot inconnu
RT		RT	<i>Retweet</i>
HT		HT	<i>Hashtag</i>
URL		URL	Adresse d’un site web
USR		USR	Compte d’utilisateur

TABLE 2 – Correspondance entre les étiquettes du PTB et celles de Ritter

4 Les CRF, les fonctions caractéristiques et les patrons

Les CRF (Conditional Random Fields), introduits dans (Lafferty *et al.*, 2001), font désormais partie des méthodes d’apprentissage automatique supervisé standard, ils sont particulièrement efficaces pour l’annotation de séquences. Différents choix sont possibles pour définir les “patrons” qui leur seront utiles pour la tâche d’annotation morpho-syntaxique. (Lavergne *et al.*, 2010) et (Suzuki & Isozaki, 2008) ont ainsi chacun proposé un ensemble de patrons destinés à l’étiquetage morpho-syntaxique de l’anglais, tandis que (Nooralahzadeh *et al.*, 2014) et (Constant *et al.*, 2011) ont construit des modèles pour le français. Nos patrons sont inspirés de ceux de (Constant *et al.*, 2011), définis pour des textes écrits. Etant donnée l’irrégularité des tweets, nous avons utilisé seulement des unigrammes d’étiquettes associés aux propriétés du tableau 3, ainsi que les bigrammes d’étiquettes seules, pour prendre en compte leurs transitions (comme dans un modèle de type *HMM*). Dans ce tableau, le chiffre 0 désigne le token courant, -1 le précédent, 1 le suivant, etc. Toutes nos expériences ont été menées avec le logiciel Wapiti².

Type	Nom de propriété	Fenêtre
valeur de token	valeur de token	[-2, -1, 0, 1, 2]
valeur de token	valeur de token bigramme	[-1, 1], [-1, 0], [0, 1]
type de token en binaire	fstUpper, allUpper, hasDash, hasNumb	0
lowercase	lower	0
prefixe/suffixe	prefixe_n, suffixe_n (n = 1..5)	0
ressource externe en binaire	catétories dans PTB	[-2, -1, 0, 1, 2]

TABLE 3 – Patron des CRF pour les expériences

5 Expériences

Notre but est tout d’abord de mesurer l’effet sur l’apprentissage de passer des jeux d’étiquettes initiaux du T-POS et du PTB (45 étiquettes) à notre jeu d’étiquettes universel spécifique des tweets (17 étiquettes). Nous nous attendons à de meilleures performances en annotation morpho-syntaxique avec moins d’étiquettes. Ensuite, nous essayons d’évaluer les performances d’un modèle mixte appris en mélangeant, dans diverses proportions, des données du PTB et de T-POS.

2. wapiti 1.5.0, <https://wapiti.limsi.fr>

Pour ce faire, nous voulions construire notre propre baseline en apprenant un modèle sur le Penn TreeBank entier et en le testant sur le corpus T-POS. Mais notre serveur ne s’est pas avéré suffisamment puissant pour mener à bout ces expériences. Nous nous contentons donc de reproduire ici le résultat présenté dans (Ritter *et al.*, 2011), obtenu avec un modèle de Maximum d’Entropie : l’exactitude annoncée est de 81.3%.

Nous avons également procédé à une validation croisée en 10 blocs sur le corpus T-POS seul. L’évaluation en exactitude est dans ce cas une moyenne des 10 tests sur 1/10 de T-POS chacun.

Enfin, notre dernière série d’expériences teste des modèles appris avec un mélange de corpus PTB et les 9 blocs de T-POS, avec la même répartition aléatoire du corpus Twitter T-POS que dans la partie précédente.

L’algorithme d’optimisation utilisé pour tous les apprentissages de wapiti est rprop+.

5.1 Validation croisée avec T-POS uniquement

Dans cette section, nous avons testé nos patrons en validation croisée à 10 blocs, en calculant l’exactitude moyenne obtenue sur l’ensemble du jeu d’étiquettes Ritter (49 étiquettes). Pour optimiser la régularisation de la log-vraisemblance du CRF, pondérée par les paramètres L1 et L2, nous avons d’abord fixé L2 à 0.00001 (valeur par défaut dans wapiti) et nous avons testé toutes les valeurs possibles de L1 dans l’ensemble suivant : {0.01, 0.03, 0.1, 0.3, 1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0}. Ensuite, nous avons gardé la valeur de L1 qui donne le meilleur résultat et nous avons fait varier la valeur de L2 dans l’ensemble {0.00001, 0.0001, 0.001, 0.01, 0.1, 0.5, 1} en gardant de nouveau celle qui donne le meilleur résultat.

Jeu d’étiquettes	L1	L2	Moyenne d’exactitude
Ritter	0.1	1	87.21%
Universal	0.01	1	89.25%

TABLE 4 – Validation croisée avec le corpus T-POS

D’après les résultats du tableau 4, nous remarquons que notre jeu d’étiquettes universel permet d’augmenter de 3% l’exactitude des étiquettes morpho-syntaxiques, bien que ce jeu entraîne de nouvelles séquences d’étiquettes auparavant impossibles. Ce résultat nous laisse espérer mieux analyser morpho-syntaxiquement les données de Twitter.

5.2 Modèles mixtes

Pour construire des modèles mixtes, trois différentes proportions de données issues du PTB ont été ajoutées à celles initialement disponibles en apprentissage (en validation croisée) dans T-POS : le même nombre de séquences, 4 fois plus et 9 fois plus. Les données de ces trois parties de PTB sont disjointes. Les ensembles d’apprentissage des corpus mixtes sont construits de la manière suivante : pour chaque itération de la validation croisée, les données provenant du corpus PTB restent les mêmes, seule la partie du corpus T-POS change. Les résultats figurent dans le tableau 5. L’évaluation des modèles est calculée par la moyenne des 10 blocs. Ensuite, l’optimisation des paramètres L1 et L2 se fait de manière identique à la partie précédente.

Jeu d’étiquettes	Proportion	L1	L2	Moyenne d’exactitude
Ritter	1 :1	1.5	0.0001	85.40%
Ritter	4 :1	1.8	0.001	86.72%
Ritter	9 :1	1.9	0.0001	87.18%
Universal	1 :1	1.0	0.01	89.11%
Universal	4 :1	1.0	0.5	89.27%
Universal	9 :1	1.6	0.01	88.95%

TABLE 5 –

Ce résultat montre qu’avec le jeu d’étiquettes Ritter, le fait d’ajouter des textes issus du PTB n’augmente pas vraiment l’exactitude par rapport aux validations croisées (qui était de 87.21%), C’est sans doute dû au fait que les deux corpus ne se ressemblent pas suffisamment : les nouvelles données introduisent de nouveaux tokens et de nouvelles séquences

d'étiquettes qui n'aident pas à mieux reconnaître celles de Twitter. L'effet semble légèrement moindre avec le jeu d'étiquettes universel, qui permet une très légère amélioration. Nous remarquons néanmoins qu'ajouter successivement plus de données du PTB dans l'ensemble d'apprentissage (mélangées avec celles du corpus T-POS) améliore les performances.

6 Conclusion et perspectives du travail

Dans cet article, nous proposons tout d'abord un jeu d'étiquettes réduit par rapport aux 45 étiquettes du PTB, qui sera facilement exploitable pour d'autres langues tout en préservant les spécificités de Twitter. Ce jeu d'étiquettes rend l'étiqueteur morpho-syntaxique plus fiable en regroupant entre elles les catégories similaires/proches. Mais sa limite est qu'il ne distingue pas bien certaines catégories. Une nouvelle version du jeu de Tags universel a d'ors et déjà été proposée dans <http://universaldependencies.github.io/docs/u/pos/index.html>. Cette version sépare les conjonctions de subordination des prépositions, les auxiliaires des verbes, les symboles et les interjections des X et les noms propres des noms communs. Nous avons déjà pris en compte cette dernière distinction. Comme la prochaine étape de notre travail est d'analyser les opinions dans les tweets, nous allons tester l'impact des deux jeux d'étiquettes pour cette analyse d'opinions.

D'autre part, nous avons essayé d'apprendre un étiqueteur morpho-syntaxique à partir de textes écrits et de tweets, pour étiqueter des tweets. Les résultats de ces expériences montrent qu'ajouter des données éloignées de la cible dans la phase d'apprentissage permet d'améliorer l'exactitude. Mais, pour apprendre un étiqueteur plus performant, rien ne vaut l'augmentation de la taille des données d'apprentissage proches de la cible. À défaut, il faudrait envisager l'utilisation de ressources linguistiques externes ou de données non étiquetées.

Une autre piste serait de trouver des patrons plus adaptés aux données de Twitter (nous avons vu que les bigrammes d'étiquettes avec les caractéristiques ne fonctionnent pas). Il peut être aussi intéressant d'ajouter d'autres caractéristiques pertinentes. L'exemple donné au début de l'article dans la figure 1 suggère l'utilisation de transcriptions phonétiques pour normaliser des tokens comme "wasz" ou "cusz", comme proposé dans (Clark, 2003).

Des clusters de grandes quantités de tweets comme dans (Nooralahzadeh *et al.*, 2014), une optimisation des paramètres L1 et L2 à la façon de (Dinarelli & Rosset, 2012) sont aussi des pistes exploitables pour cette tâche. Enfin, un traitement spécifique des hashtags (s'ils jouent un vrai rôle syntaxique) pourrait aussi permettre un étiquetage plus fin et plus régulier des messages.

Références

- ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In A. ABEILLÉ, Ed., *Treebanks*. Dordrecht : Kluwer.
- CLARK A. (2003). Combining distributional and morphological information for part of speech induction. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, p. 59–66, Stroudsburg, PA, USA : Association for Computational Linguistics.
- CONSTANT M., TELLIER I., DUCHIER D., DUPONT Y., SIGOGNE A. & BILLOT S. (2011). Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français. In *TALN*, volume 1, p. 321, Montpellier, France.
- DINARELLI M. & ROSSET S. (2012). Tree-structured named entity recognition on ocr data : Analysis, processing and results.
- FORSYTH E. N. (2007). Improving automated lexical and discourse analysis of online chat dialog.
- FOSTER J., ÇETINOĞLU Ö., WAGNER J., LE ROUX J., HOGAN S., NIVRE J., HOGAN D., VAN GENABITH J. *et al.* (2011). #hardtoparse : Pos tagging and parsing the twitterverse. In *proceedings of the Workshop On Analyzing Microtext (AAAI 2011)*, p. 20–25.
- GIMPEL K., SCHNEIDER N., O'CONNOR B., DAS D., MILLS D., EISENSTEIN J., HEILMAN M., YOGATAMA D., FLANIGAN J. & SMITH N. A. (2011). Part-of-speech tagging for twitter : Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : Short Papers - Volume 2*, HLT '11, p. 42–47, Stroudsburg, PA, USA : Association for Computational Linguistics.

- LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.
- LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale crfs. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, p. 504–513, Stroudsburg, PA, USA : Association for Computational Linguistics.
- MARCUS M. P., SANTORINI B. & MARCINKIEWICZ M. A. (1993). Building a large annotated corpus of english : The penn treebank. *COMPUTATIONAL LINGUISTICS*, **19**(2), 313–330.
- NOORALAHZADEH F., BRUN C. & ROUX C. (2014). Part of speech tagging for french social media data. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference : Technical Papers, August 23-29, 2014, Dublin, Ireland*, p. 1764–1772.
- PETROV S., DAS D. & McDONALD R. (2012). A universal part-of-speech tagset. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, M. U. DOĞAN, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey : European Language Resources Association (ELRA).
- RITTER A., CLARK S., MAUSAM & ETZIONI O. (2011). Named entity recognition in tweets : An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*, p. 1524–1534, Stroudsburg, PA, USA : Association for Computational Linguistics.
- SEDDAH D., C B. S. M., MOUILLERON V. & COMBET V. (2012). The french social media bank : a treebank of noisy user generated content.
- SUZUKI J. & ISOZAKI H. (2008). Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *In ACL*.
- TOUTANOVA K. & MANNING C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora : Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13, EMNLP '00*, p. 63–70, Stroudsburg, PA, USA : Association for Computational Linguistics.

Caractériser les discours académiques et de vulgarisation : quelles propriétés ?

Amalia Todirascu¹, Beatriz Sánchez-Cárdenas²

(1) LiLPa, Université de Strasbourg, 22, Rue René Descartes, BP 80010, 67084 STRASBOURG cedex, France

(2) LexiCon, Universidad de Granada, Calle Buensuceso, 11 18002 Granada, Espagne
todiras@unistra.fr, bsc@ugr.es

Résumé. L'article présente une étude des propriétés linguistiques (lexicales, morpho-syntaxiques, syntaxiques) permettant la classification automatique de documents selon leur genre (articles scientifiques et articles de vulgarisation), dans deux domaines différents (médecine et informatique). Notre analyse, effectuée sur des corpus comparables en genre et en thèmes disponibles en français, permet de valider certaines propriétés identifiées dans la littérature comme caractéristiques des discours académiques ou de vulgarisation scientifique. Les premières expériences de classification évaluent l'influence de ces propriétés pour l'identification automatique du genre pour le cas spécifique des textes scientifiques ou de vulgarisation.

Abstract.

Characterizing scientific genre for academia and general public: which properties should be taken into account?

The article focuses on the study of a set of morpho-syntactic properties for audience-based classification. The linguistic analysis of academic discourse and of popular science discourse reveals that both discourse types are characterized by specific linguistic and textual properties. This research used two French comparable corpora in regards to genre and subject matter. The corpora was composed of scientific articles and popular science texts in the domains of medicine and computer science. The experiments performed as part of our study evaluated the influence of discourse-specific morpho-syntactic properties on genre-based classification, for scientific and popular science texts.

Mots-clés : analyse linguistique, discours scientifique et de vulgarisation, corpus comparables, classification selon le genre.

Keywords: linguistic analysis, academic and popular science discourse, comparable corpora, genre-based classification.

1 Introduction

L'identification automatique du genre du document est une tâche utile pour l'extraction automatique de terminologie ou de néologismes, pour la génération automatique de contenu destiné à un public cible ou pour la simplification automatique. En effet, plusieurs applications peuvent intégrer cette étape de classification par genre. Pour l'extraction de termes ou néologismes, il faut utiliser des textes riches en termes ou productifs en néonymie, tels que les articles scientifiques. Dans le domaine de la didactique des langues, il est nécessaire de proposer aux apprenants des textes adaptés à leur niveau de connaissances, c'est-à-dire des textes simplifiés en terme de lexique, de structures syntaxiques ou discursives. Pour cette raison, nous privilégions la sélection des textes de vulgarisation pour des apprenants ayant des compétences en langues. Etant donné que le genre est caractérisé par un ensemble complexe de paramètres linguistiques et extra-linguistiques, l'identification automatique des genres, en particulier des discours académiques et scientifiques n'est pas une tâche aisée.

Lors de l'identification automatique du genre, la plupart des systèmes proposent une identification basée sur des n-grammes (séquences de mots, de caractères ou de catégories lexicales) (Sebastiani, 2005) (Lee et Myaeng, 2004) (Kessler *et al*, 1997) exploitant le lien entre le thème et le genre traité (Bechet *et al*, 2008). Or, ces approches sont généralement liées au vocabulaire du domaine et, si l'on obtient des performances notables dans un domaine, l'adaptation des systèmes pour d'autres domaines émergents demande une phase de réapprentissage sur un corpus conséquent. Pourtant, quelques travaux se distinguent par l'exploitation des catégories lexicales (Karlgrén et Cutting, 1994) ou des propriétés syntaxiques (Santini, 2007), (Goeuriot *et al*, 2005), (Stamatatos *et al*, 2000) (D'hondt *et al*, 2013) pour classer les documents selon le genre ou l'auteur des textes. Certaines approches proposent des propriétés

indépendantes de la langue (Petrenz et Webber, 2011), (Sun *et al.*, 2012) qui s'avèrent efficaces pour la classification de genres journalistiques. Enfin, plusieurs travaux combinent propriétés stylométriques (longueur de phrases, signes de ponctuation), catégories lexicales et mots-clés pour la classification des genres littéraires (D'hondt *et al.*, 2013), (Lecluze et Lejeune, 2014).

Des nombreux travaux en linguistique proposent des caractérisations des types et des genres textuels par le biais des propriétés morpho-syntaxiques (Biber et Conrad, 2009) (Poudat *et al.*, 2006), (Malrieu et Rastier, 2001). Les résultats de ces études ont été exploités par des systèmes de classification automatique par genre (Charnois *et al.*, 2008, Poudat, 2008, D'hondt *et al.*, 2013). Ces systèmes utilisent des propriétés linguistiques (p.e. la fréquence ou la distribution de noms propres ou de déterminants, la fréquence de certaines catégories de verbes de modalité ou d'opinion, la fréquence des groupes nominaux complexes etc.), mais pas d'annotations linguistiques de plus haut niveau (syntaxe, sémantique discursive).

Nous adoptons une approche de classification automatique exploitant les études des genres textuels disponibles dans la littérature. Dans cet article, nous étudions plusieurs propriétés lexicales (termes du domaine), traits morpho-syntaxiques et structures syntaxiques utiles pour l'identification automatique du genre textuel selon le public visé : communauté d'experts scientifiques ou grand public. Contrairement à d'autres approches qui sélectionnent les propriétés automatiquement, nous partons des travaux en linguistique textuelle (Biber et Conrad, 2009) et des études des genres scientifiques (Hyland, 2009 ; Swales, 2004) et nous comparons les propriétés textuelles des articles scientifiques et de textes de vulgarisation scientifique sur deux corpus comparables¹ en termes de genres et du thème, disponible en français, un corpus médical et un corpus informatique. Les corpus sont comparables en genre : chaque corpus est constitué en parties égales d'articles scientifiques et d'articles de vulgarisation. D'autre part, à l'intérieur du domaine, nous avons sélectionné des textes scientifiques et de vulgarisation traitant des mêmes thèmes. Pour vérifier l'influence des propriétés sélectionnées suite à l'étape de l'analyse de corpus, nous utilisons des techniques de classification automatique et nous comparons avec une approche de classification basée exclusivement sur des termes du domaine ou les mots pleins contenus dans les documents.

2 Discours scientifiques et discours de vulgarisation scientifique

Pour analyser les différences entre discours scientifiques et de vulgarisation, nous adoptons la définition proposée par (Biber et Conrad, 2009), considérant que le genre est défini par un faisceau de propriétés linguistiques et extra-linguistiques (paramètres de production du texte et de réception du texte). Certaines structures et formules sont unanimement reconnues par les utilisateurs comme étant des caractéristiques propres à un genre donné. À ce titre, (Hyland, 2009) propose une classification des genres académiques, du point de vue des pratiques de rédaction et de lecture que la communauté scientifique adopte. Le domaine définit les caractéristiques rhétoriques et stylistiques du discours académique. (Swales, 2004) considère que les genres participent activement à la construction des connaissances d'un champ disciplinaire. Les pratiques disciplinaires se traduisent par la préférence pour certains procédés linguistiques, destinés à faciliter le partage des connaissances. Pour notre étude, nous nous focalisons sur les propriétés linguistiques et extra-linguistiques permettant de distinguer entre les discours académiques et de vulgarisation scientifique.

Des nombreux travaux existent sur la caractérisation des articles scientifiques (Hyland, 2009), (Swales, 2004). Ces recherches ont mis en évidence des procédés spécifiques, reconnus et appliqués par la communauté scientifique, pour exprimer le positionnement de l'auteur (Tutin, 2010) ou pour argumenter des choix méthodologiques (Rinck, 2006). Parmi les procédés linguistiques mentionnés par (Hyland, 2009), plusieurs sont typiques du discours académique : des expressions qui expriment la possibilité d'une hypothèse (verbes de modalité), l'autocitation (pronoms personnels de 1^{ère} personne, articles possessifs), l'expression d'un point de vue (verbes de croyance). (Swales, 2004) propose une analyse détaillée de plusieurs catégories de genres scientifiques (thèse, article, présentation orale) en identifiant les fonctions rhétoriques et linguistiques qui expriment l'argumentation, la citation d'autres travaux ou les connaissances implicites.

Si certains travaux proposent l'identification d'un vocabulaire commun aux textes scientifiques, un meta-langage (Drouin, 2007), d'autres mettent en exergue l'apparition de phénomènes liés à la syntaxe : présence d'adjectifs relationnels (Daille, 1999), préférence pour le passif, utilisation de tournures impersonnelles, utilisation du pronom de 1^{ère} personne. (Kocourek, 1991) souligne comme caractéristique du langage de spécialité les termes du domaine. Aussi, il identifie des phénomènes telles que les propositions participiales ou infinitives, les subordonnées relatives.

¹ Si le terme « corpus comparable » est généralement utilisé pour des corpus multilingue qui partagent les mêmes critères, dans cet article, nous utilisons le terme de corpus comparable pour des corpus qui ont des sous-parties comparables en terme de genre et thème.

D'autre part, le langage de vulgarisation scientifique a également fait l'objet de nombreuses études. (Hyland, 2009) identifie une volonté d'explicitier les notions au niveau de connaissances du public non averti : utilisation des explications, des dispositifs cohésifs tels que la répétition, la synonymie ou les articles démonstratifs pour renforcer la cohésion du texte. Le public auquel l'auteur s'adresse est inclus dans le discours, par l'utilisation du pronom personnel *vous* et par des questions rhétoriques accompagnées de réponses. (Jacobi et Schiele, 1988) proposent plusieurs paramètres linguistiques spécifiques aux textes de vulgarisation telle la reformulation (l'explication d'un terme pivot par d'autres expressions plus simples) pour renforcer les connaissances du public. La reformulation s'exprime par des marqueurs explicites (« *c'est-à-dire* », « *autrement dit* ») ou par des énoncés définitoires (*X est défini comme Y*, *X est nommé Y*, *X est un Y*).

Ces propriétés mises en évidence par des études détaillées des discours académiques ou de vulgarisation scientifiques seront utilisées pour classer automatiquement les deux catégories de discours. Nous évaluons l'influence de ces propriétés décrites dans la littérature dans le cadre d'un système de classification automatique de genres, par une étude de corpus et par quelques expériences de classification.

3 Méthodologie

Nous partons de ces études linguistiques des discours académique et de vulgarisation scientifique pour identifier des propriétés exploitables pour la classification automatique. Pour atteindre cet objectif, nous avons suivi la méthodologie explicitée ci-dessous :

1. Identification des propriétés morpho-syntaxiques, syntaxiques et stylistiques dans la littérature pour l'analyse des discours académiques et de vulgarisation (Hyland, 2009; Swales, 2004; Jacobi et Schiele, 1988) ;
2. Constitution de corpus comparables pour deux domaines différents, composés d'articles scientifiques et de textes de vulgarisation. Les domaines choisis sont le domaine médical et l'informatique, deux domaines dans lesquelles les pratiques d'écriture scientifique sont différentes ;
3. Prétraitement : étiquetage, lemmatisation et analyse syntaxique (Bohnet, 2009) ;
4. Analyse des corpus, à l'aide du concordancier Antconc (Anthony, 2009), appliqué aux corpus annotés, pour l'identification des propriétés utiles pour la classification ;
5. Développement d'un extracteur de propriétés, appliqué sur le corpus annoté ;
6. Expériences de classification avec les propriétés choisies avec la plateforme Weka (Hall *et al*, 2009). Nous avons comparé les résultats obtenus à l'aide des propriétés sélectionnées avec un système utilisant des mots pleins identifiés dans les corpus.

Dans les prochaines sous-sections, nous présentons les étapes de création de corpus et le prétraitement appliqué sur le corpus, l'analyse des propriétés présentées dans la section 2 et les expériences de classification.

3.1 Création de corpus et prétraitement

Notre objectif est d'identifier automatiquement les articles scientifiques et les textes de vulgarisation dans les domaines de la médecine et de l'informatique. Compte tenu du fait que peu de corpus sont disponibles en français (Tutin 2010) nous avons constitué des corpus comparables dans les domaines mentionnés. Les corpus sont comparables par rapport aux genres sélectionnés (textes de vulgarisation, articles scientifiques) et par rapport aux thèmes traités dans les documents. Nous avons cherché les documents à l'aide des mêmes mots-clés (à l'intérieur du domaine) pour sélectionner des documents de genres différents traitant du même thème.

Pour le domaine médical, nous disposons de deux corpus de taille comparable pour le français : 302 textes/genres (environ 1 500 000 mots pour chaque genre). Les articles scientifiques proviennent du corpus Scientext (Tutin, 2010), des sites à destination de spécialistes, gérés par les réseaux de santé et les organismes publics, de quelques revues médicales (la revue « médecine/sciences », Revue française de Rhumatologie). Le corpus du discours de vulgarisation dans le domaine médical été constitué à partir de sites d'information à destination du grand public, créés pour la prévention sur certaines maladies². Le corpus de textes scientifiques du domaine informatique est composé d'articles scientifiques disponibles sur le portail HAL. Le corpus de textes de vulgarisation en informatique est composé des textes disponibles sur des portails ou des magazines en ligne et des tutoriels destinés à l'apprentissage d'un langage de

² Nous remercions Guillaume Bertrand qui a contribué à la constitution du corpus médical dans le cadre de son travail de mémoire du master Linguistique, Informatique, Traduction (Université de Strasbourg, 2011).

programmation pour débutants. Certains textes de vulgarisation ont été retirés du corpus car jugés trop éloignés des textes de vulgarisation (par exemple des brèves annonçant le lancement d'un nouveau logiciel ou produit). Nous disposons au total de 301 textes/genre dans le corpus informatique.

Tout d'abord, nous avons procédé à un nettoyage des corpus recueillis. Pour le corpus informatique il s'agit de supprimer manuellement les images, les tableaux et les formules, ainsi que les parties contenant du code. Les corpus médicaux provenant des sites Web ont été extraits à l'aide d'un outil d'extraction du contenu textuel à partir des pages Web et des fichiers PDF (Todorascu *et al.*, 2012). Les corpus ont été étiquetés, lemmatisés et annotés avec l'analyseur statistique en dépendances de (Bohnet, 2009) disponible pour le français. Nous avons utilisé cet analyseur dans le but d'identifier plusieurs catégories de propriétés syntaxiques des discours scientifiques (Biber et Conrad, 2009; Hyland, 2009). Par ce biais, nous cherchons à identifier des propriétés généralisables pour la classification de textes scientifiques et de vulgarisation entre plusieurs domaines.

3.2 Analyse de corpus

Nous avons réparti les propriétés identifiées dans la littérature dans plusieurs classes :

- propriétés statistiques : la longueur moyenne des phrases, le nombre total d'unités lexicales, la longueur moyenne des mots, la fréquence des mots longs ou courts, les signes de ponctuation (!,?). Il est possible de calculer ces propriétés sans faire appel aux annotations linguistiques ;
- propriétés lexicales : certaines classes de verbes (verbes de cognition, verbes de communication, verbes de modalité) ou d'adjectifs (relationnels). De plus, nous avons utilisé une liste de 100 termes monolexicaux et polylexicaux, extraits à l'aide de Termostat (Drouin, 2007) pour chaque corpus ;
- propriétés morpho-syntaxiques : les catégories lexicales spécifiques (nom, nom propre, verbe, adjectif, adverbe, pronoms personnels de 1^{re} et 2^e personne) ;
- propriétés syntaxiques et sémantiques : les séquences définitoires ou des explications, constructions passives, les tournures impersonnelles, le type de sujet ou d'objet (pronoms, groupes nominaux ou phrase).

Afin d'évaluer le lien entre ces propriétés et leur genre, nous avons étudié les corpus, étiquetés, lemmatisés et analysés en dépendances. Compte tenu des tailles variables des documents composant nos corpus, nous avons calculé la fréquence relative de chaque propriété (FT – le nombre d'occurrences comptées dans le corpus, Nb – le nombre total de mots et de signes de ponctuation du corpus) :

$$Freqrel = \left(\frac{FT}{Nb} \right) * 1000000$$

Pour compter la fréquence des propriétés, nous avons défini des patrons lexico-syntaxiques spécifiques, utilisant le langage d'interrogation de corpus CorpusQueryProcessing (CQP) (Christ, 1994). Pour identifier les sujets et les objets complexes, nous avons exploité les liens de dépendances entre le verbe et le sujet. Nous avons défini des règles heuristiques pour plusieurs phénomènes (tournures impersonnelles, énoncés définitoires). Certaines règles estiment la fréquence relative du phénomène (propositions relatives). Plusieurs patrons identifient des définitions (11 patrons) ou des emplois impersonnels du pronom 'il' (45 patrons) :

[lemma=""être"] [lemma="définir|appeler|nommer"] [word="comme"] [pos="NOM"]

[lemma="il"] [lemma=""être"] [word="nécessaire"] [word="de"] [pos="vinf"]

lemma – impose une contrainte sur le lemme ; word – sur la forme ; pos – sur la catégorie lexicale (vinf – verbe à l'infinitif, NOM – nom).

Nos analyses (tableau 1) montrent que le discours académique est marqué par plusieurs propriétés caractéristiques : les pronoms personnels (*nous*, *je*), les constructions passives, les pronoms impersonnels (*il*, *on*) et la préférence pour des sujets complexes (phrases subordonnées, groupes nominaux modifiés par plusieurs compléments de noms), pour des mots longs (plus de 9 caractères) ou courts. La fréquence relative de ces propriétés est plus importante pour les discours académiques que dans les corpus de vulgarisation scientifique. D'autre part, le discours de vulgarisation est caractérisé par une préférence marquée pour le pronom personnel de la 2^e personne, une forte présence des questions, une préférence pour les marqueurs de reformulation, ainsi que pour les définitions qui éclairent les termes complexes au grand public. Nous avons retenu ces propriétés ayant un comportement similaire dans les deux domaines.

	Je	Nous	passif	définitions	Sujet complexe	Pronoms impersonnel	Reform.	Pronom 2 ^e pers
DS MED	22	2316	3069	814	41536	326	214	0
DV MED	934	274	2546	2212	10451	14	132	315
DS INFO	188	2049	657	1704	15789	187	113	10
DV INFO	0	20	167	11070	4324	20	530	641

TABLE 1 : Quelques propriétés lexicales, morpho-syntaxiques ou syntaxiques et leur fréquence relative (valeurs en partie pour million) (DS – discours académique; DV -discours de vulgarisation)

Certaines propriétés n'ont pas été retenues pour les expériences de classification en raison de leur comportement similaire dans les deux genres (par exemple, la fréquence de noms propres, de noms ou d'adverbes qui ont des fréquences similaires dans les deux genres). D'autres propriétés ont des comportements contradictoires selon le domaine : le pronom personnel *je* est plus fréquent dans le discours de vulgarisation scientifique en médecine, alors qu'en informatique il est spécifique au discours scientifique. Les marqueurs de reformulation semblent plus fréquents dans le discours scientifique médical tandis ce qu'en informatique, ils sont plus fréquents dans les discours de vulgarisation.

Finalement, nous avons comparé le vocabulaire des textes scientifiques et celui de textes de vulgarisation. Nous avons remarqué la présence des termes du domaine dans le corpus de textes scientifiques et le corpus de textes de vulgarisation. Cependant, il existe des différences entre les deux types de discours. D'une part, le langage scientifique se caractérise par la préférence marquée pour des noms abstraits (*analyse, approche, processus, entité*) qui font partie du méta-lexique du domaine informatique et par une préférence marquée pour les noms d'événement liés aux processus d'examen médical et de prise en charge du patient (*examen, étude, traitement*), aux processus des maladies (*infection, apparition, augmentation*). D'autre part, le langage de vulgarisation se manifeste par la fréquence de mots du domaine de la biologie (*cellule, protéine, neurone*) et des parties du corps (*abdomen, foie, ganglion*) pour le corpus médical et par une fréquence des entités propres à l'univers de l'ordinateur (*souris, fenêtre*) ou des programmes (*code, fonction*). Ces classes de noms abstraits ont été rajoutées aux propriétés utilisées pour les expériences de classification.

3.3 Expériences de classification

Après avoir choisi les propriétés à l'issue de l'analyse de corpus, nous avons appliqué ces propriétés pour la classification automatique, afin d'évaluer leur utilité pour l'identification de genres. Les propriétés sélectionnées sont regroupées manuellement et sont utilisées pour représenter les documents :

- TERMES : un ensemble de 100 termes monolexicaux et polylexicaux ont été choisis pour chaque domaine et nous avons appliqué l'union des deux listes pour les expériences de classification entre domaines. Nous utilisons Tfidf calculé pour chaque terme dans le vecteur représentant chaque document.
- STAT : les propriétés statistiques (longueur des mots courts et longs, longueur de la phrase, fréquence des signes d'interrogation ou d'exclamation, fréquence des parenthèses indiquant des explications) ;
- SYN : les propriétés syntaxiques et plus généralement des annotations de haut niveau (la fréquence relative des certains noms abstraits identifiés dans la section précédente, les pronoms *nous* et *vous*, les verbe de cognition ou de modalité, la fréquence des adjectifs relationnels, des constructions passives, des tournures impersonnelles, des sujets et des objets complexes, les définitions et des marqueurs de reformulation).
- ALL : toutes les propriétés statistiques et syntaxiques, à l'exception des termes;
- Unigrams : l'union des mots pleins apparaissant dans le corpus médical ou dans le corpus informatique. Il s'agit du système de base, ces propriétés peuvent être extraites sans recours aux annotations linguistiques de haut niveau.

Pour calculer la fréquence relative de ces propriétés pour chaque document nous avons appliqué la formule et les patrons (présentés en section 3.2) pour identifier certains phénomènes complexes tels que les tournures impersonnelles, les constructions passives, les définitions ou les explications.

Une fois les propriétés extraites, nous avons effectué des tests à l'aide de Weka, une plateforme de classification automatique (Hall *et al.*, 2009). Nous avons classé manuellement les documents comme étant des textes scientifiques ou de vulgarisation. Notre corpus est constitué de 302 textes/genre dans le domaine médical et 301 textes/genre dans le domaine informatique. Les résultats présentés sont obtenus avec l'algorithme SMO, l'implémentation du classifieur SVM disponible sur Weka. Nous avons effectué plusieurs expériences présentées dans le tableau 2 :

Classes	Modèle	Corpus de test	ALL	STAT	SYN	TERMES	Unigrams
2 classes DS, DV	médecine	médecine	96,97%	93,15%	92,51%	87,42 %	99,08%
	informatique	informatique	93,12%	87,30%	83,63%	95,76 %	99,07%
	médecine	informatique	76,16%	74,18%	71,08%	55,43%	65,32%
	informatique	médecine	91,24%	87,89%	83,63%	47,55%	54,25%
4 classes DSMED, DVMED, DSINFO, DVINFO			85,18%	71,36%	77,54%	89,42 %	98,22%

TABLE 2 : L'exactitude (E) obtenue pour les 4 systèmes et le système de base sur un corpus de test du domaine et hors domaine

1) **Expériences de classification à l'intérieur du domaine, et entre les domaines considérant deux classes (discours de vulgarisation et discours scientifiques).** Notre objectif est de vérifier si certaines propriétés choisies peuvent être généralisées entre les domaines pour caractériser les discours scientifiques et de vulgarisation. Pour ces tests, nous avons appliquée la validation croisée ($k=10$). Nous avons construit un modèle pour chaque groupe de propriétés pour le corpus médical et informatique. Pour construire le modèle des termes du domaine sélectionnés manuellement (TERMES), nous avons utilisé l'union des deux ensembles de termes extraits à partir des corpus médical et informatique. Pour la classification à l'intérieur du domaine (corpus de test et d'entraînement du même domaine), on constate que les termes sont plus efficaces pour le corpus informatique ($E=95,76\%$) mais le système ALL obtient 93,12% des instances correctement classées. Pour le corpus médical, le système ALL ($E=96,97\%$) est meilleur que TERMES ($E=87,42\%$). Pour la classification entre domaines, nous constatons que les listes de termes sélectionnés manuellement sont peu efficaces, aussi bien que le système de base Unigram. Pour la classification inter-domaine, le modèle ALL construit sur le corpus informatique est le meilleur (exactitude de 91,24 %), alors que dans l'autre direction, ALL reste toujours le plus performant (76.16%). Les systèmes utilisant exclusivement des annotations de haut niveau (SYN) ont obtenu des résultats plus faibles que les systèmes utilisant des propriétés statistiques (STAT). Le système Unigram reste très efficace pour la classification à l'intérieur du domaine.

2) **Expériences de classification, réalisées en considérant 4 classes (DSMED, DVMED, DSINFO, DVINFO).** Dans ce deuxième cas, nous avons appliqué la méthode de validation croisée (avec $k=10$) sur le corpus d'entraînement. Parmi les quatre configurations de propriétés proposées dans notre approche, il s'avère que les termes restent toujours les meilleurs alors que ALL est en seconde position. Le meilleur score est obtenu par le système Unigram (98,22%).

Les résultats obtenus montrent que la combinaison de propriétés statistiques et syntaxiques ALL semblent généralisables plus facilement entre domaines. Toutefois, la taille du corpus est limitée et les résultats doivent être confirmés sur un corpus de taille plus importante. Le calcul de certaines propriétés s'appuie sur l'existence d'annotations syntaxiques automatiques, dont le résultat n'a pas été corrigé. Certains propriétés (par exemple la fréquence des pronoms relatifs ou les énoncés définitoires) sont évaluées à l'aide d'une catégorie lexicale ou d'une séquence d'étiquettes. Les erreurs provenant de l'analyse automatique peuvent influencer l'extraction de propriétés et donc les résultats de la classification.

4 Conclusion et perspectives

Les expériences de classification à l'intérieur du domaine montrent que l'ensemble de propriétés lexicales, morpho-syntaxiques et syntaxiques sont aussi performantes que les termes. En ce qui concerne la classification entre les domaines, l'ajout des propriétés morpho-syntaxiques se révèle plus effective que la simple sélection manuelle des termes du domaine. Les expériences de classification ont pris en compte des propriétés des genres étudiées dans la littérature par la suite vérifiées à l'aide d'une analyse de corpus détaillée. D'autres comparaisons avec un système de classifications utilisant des n-grams de caractères seront effectuées. Néanmoins, il est évident que ces résultats doivent être validés sur des corpus de taille plus importante, ainsi que l'évaluation des erreurs dues à l'annotation automatique et aux règles d'extraction de propriétés (par exemple des règles qui identifient les définitions ou les tournures impersonnelles), qui s'appuient sur cette annotation. La méthode peut être appliquée à d'autres domaines et elle peut s'avérer utile pour identifier des sous-genres académiques (thèse, article scientifique).

Références

- BIBER D., CONRAD S. (2009). *Register, Genre and Style*, Cambridge University Press.
- BOHNET B. (2009) Efficient Parsing of Syntactic and Semantic Dependency Structures. *Proceedings of Conference on Natural Language Learning (CoNLL)*, Boulder, 67-72.
- CHARNOIS T., DOUCET A., MATHET Y. (2008). Trois approches du GREYC pour la classification de textes. *Actes TALN'08*, Avignon
- CHRIST, O. (1994) A Modular and Flexible Architecture for an Integrated Corpus Query System *Proceedings of COMPLEX'94*, Budapest, Hungary, July 7-10, 1994, PP- 23-32
- DAILLE B. (1999). « Identification des adjectifs relationnels en corpus », Conférence TALN1999, Cargèse
- D'HONDT E., VERBERNE S., KOSTER C., BOVES L. (2013). Text Representations for Patent Classification. *Computational Linguistics*, 39(3), 755–775.
- DROUIN, P. (2007) « Identification automatique du lexique scientifique transdisciplinaire », *Revue française de linguistique appliquée* 2/2007 (Vol. XII) , p. 45-64
- HYLAND, K. (2009). *Academic Discourse*, London: Continuum.
- HALL M., ET AL (2009). « The WEKA Data Mining Software: An Update », *SIGKDD Explorations*, Vol. 11/1.
- GOEURIOT L., MORIN E., ET AL. (2009). « Reconnaissance du type de discours dans des corpus comparables spécialisés », CORIA Conférence en Recherche d'Information et Applications.
- JACOBI D., SCHIELE B. (ÉD.) (1988). *Vulgariser la science*, Seyssel, Éditions Champ Vallon (Milieux).
- KARLGREN J., CUTTING D. (1994). Recognizing text genres with simple metrics using discriminant analysis. *Actes de COLING 94*, 1071-1075.
- KESSLER B., NUNBERG G., SCHÜLTZE H. (1997). Automatic detection of text genre. *Actes de EACL'97*, 32-38.
- KOCOUREK R. (1991). *La langue française de la technique et de la science*, Oscar Brandstetten Verlag.
- LECLUZE, C, LEJEUNE, G. (2014). DEFT 2014, analyse automatique de textes littéraires et scientifiques en langue française (stylométrie et quelques catégories lexicales) , DEFT 2014, Marseille, 1er juillet 2014
- LEE Y.-B., MYAENG S. H. (2004) Automatic identification of text genres and their roles in subject-based categorization, *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*
- MALRIEU D., RASTIER F. (2001). Genres et variations morphosyntaxiques. *Traitement Automatique des langues*, vol. 42, n°2, pp. 548-577.
- PETRENZ P., WEBBER, B. (2011). Stable Classification of Text Genres. *Computational Linguistics* 37:2, 385-393
- POUDAT C., CLEUZIQU G., CLAVIER V. (2006). Catégorisation de textes en domaines et genres : complémentarité des indexations lexicale et morphosyntaxique. *Document numérique*, vol.9, n°1/2006, pp. 61-76.
- RINCK F. (2007). Styles d'auteur et singularité des textes. Approche stylométrique du genre de l'article en linguistique, *Pratiques*, 135/136, 119-136.
- SANTINI M. (2007). Automatic Identification of Genre in Web Pages, Ph.D.Thesis, University of Brighton
- SEBASTIANI F. (2005). « Text categorization » In Alessandro Zanasi (ed.), *Text Mining and its Applications*, WIT Press, Southampton, UK, 2005, pp. 109-129.
- SUN J., YANG Z., LIU S., WANG P. (2012). Applying stylometric analysis techniques to counter anonymity in cyberspace. *Journal of Networks*, 7(2)
- STAMATATOS, E., FAKOTAKIS, N. ET KOKKINAKIS, G. (2000). Automatic Text Categorization in Terms of Genre and Author. In *Computational Linguistics*, Vol.26, No. 4, pages 471–497.
- SWALES, J. (2004) *Research Genres: Explorations and Applications*, Cambridge Applied Linguistics.
- TODIRASCU, A., PADO, S., KISSELEW, K., KRISCH, J., HEID, U. (2012) French and German corpora for audience-based text type classification, *Proceedings of LREC 2012*
- TUTIN A. (2010). Evaluative adjectives in academic writing in the humanities and social sciences. In Lores-Sanz, R., Mur-Duenas, P., Lafuente-Millan, E. *Constructing Interpersonality: Multiple Perspectives on Written Academic Genres*. Cambridge: Cambridge Scholars Publishing..

Extraction et analyse automatique des comparaisons et des pseudo-comparaisons pour la détection des comparaisons figuratives

Suzanne Mpouli^{1,2} Jean-Gabriel Ganascia^{1,2}

(1) Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris

(2) Labex OBVIL, Université Paris-Sorbonne, 1 rue Victor Cousin 75005 Paris

mpouli@acasa.lip6.fr, jean-gabriel.ganascia@lip6.fr

Résumé. Le présent article s'intéresse à la détection et à la désambiguïsation des comparaisons figuratives. Il décrit un algorithme qui utilise un analyseur syntaxique de surface (*chunker*) et des règles manuelles afin d'extraire et d'analyser les (pseudo-)comparaisons présentes dans un texte. Cet algorithme, évalué sur un corpus de textes littéraires, donne de meilleurs résultats qu'un système reposant sur une analyse syntaxique profonde.

Abstract.

Extraction and automatic analysis of comparative and pseudo-comparative structures for simile detection.

This article is focused on automatic simile detection and disambiguation. It describes an algorithm which uses syntactic chunks and handcrafted rules to extract and analyse similes in a given text. This algorithm, which was evaluated on a corpus of literary texts, performs better than a system based on dependency parsing.

Mots-clés : comparaisons figuratives, comparé, comparant, analyse syntaxique de surface, règles manuelles, analyse syntaxique profonde.

Keywords: simile, tenor, vehicle, chunking, handcrafted rules, dependency parsing.

1 Introduction

Avec le nombre sans cesse croissant de textes numérisés disponibles, se pose la question de leur interrogation automatique afin d'en extraire les éléments relevant du style d'un auteur, d'une œuvre ou d'un genre. C'est dans cette optique que s'inscrit le travail que nous présentons dans cette contribution et qui vise à long terme la reconnaissance automatique des comparaisons figuratives dans les textes littéraires. Bien qu'elles aient été très peu étudiées en TAL, les comparaisons figuratives se rattachent à deux phénomènes qui se rencontrent dans la plupart des langues naturelles : la comparaison et le langage figuré. En effet, du point de vue structurel, les comparaisons figuratives sont des constructions comparatives, c'est-à-dire des formes linguistiques utilisées pour exprimer à quel degré et sur quelle base deux entités au minimum peuvent être considérées semblables ou dissemblables. Cependant, du point de vue sémantique, elles se distinguent des autres types de constructions comparatives (appelées par contraste comparaisons littérales) car elles établissent un parallèle entre des termes n'appartenant pas à la même catégorie sémantique en attribuant à l'un de ces termes des attributs propres à la catégorie sémantique de l'autre (Glucksberg & Keysar, 1990). À titre d'illustration, considérons les deux énoncés suivants :

(1) *Céline est aussi revêche que sa sœur.*

(2) *Céline est aussi revêche qu'un cactus.*

Malgré la symétrie parfaite entre ces deux phrases, seule la phrase (2) peut être classée comme une comparaison figurative car elle crée une image en plaçant au même plan un être humain (Céline) et une plante (cactus) à laquelle est conféré un trait de caractère propre aux humains.

Il apparaît donc que la reconnaissance des comparaisons figuratives dans un texte brut comprend trois tâches principales : l'extraction des structures comparatives et pseudo-comparatives contenues dans le texte, l'identification des constituants de ces structures et la désambiguïsation de ces structures. Afin de mieux décrire le problème de la reconnaissance des comparaisons figuratives, nous présentons, dans la section 2, la structure des comparaisons figuratives ainsi que l'état de l'art en matière d'extraction et d'analyse automatiques des comparaisons et des pseudo-comparaisons. Dans la troisième section, nous décrivons un algorithme qui se focalise sur l'extraction des comparaisons et l'identification de leurs constituants. Puis, nous évaluons cet algorithme qui repose sur des règles manuelles et l'analyse syntaxique de surface (*chunking*) sur un corpus de textes littéraires et le comparons à un analyseur syntaxique profond. Dans la dernière section, nous évoquons brièvement la question de la désambiguïsation des comparaisons figuratives. Pour finir, nous concluons notre travail en présentant quelques perspectives futures.

2 État de l'art

2.1 Structures sémantique et syntaxique des comparaisons figuratives

S'inspirant des travaux de Dumarsais, Soublin (1971) définit un processus en deux étapes pour expliquer la formation des comparaisons figuratives à partir de deux phrases ayant la même structure Syntagme nominal + Verbe + Adjectif qualificatif : insertion d'un outil de comparaison entre les deux phrases, puis suppression du verbe et de l'adjectif après le syntagme nominal suivant cet outil de comparaison. On passera donc ainsi de (3a) « *La fille est calme* » et (3b) « *Un lac est calme* » à (4) « *La fille est calme comme un lac est calme* » et finalement à (5) « *La fille est calme comme un lac* ». Typiquement, l'exemple (5) correspond à la forme canonique de la comparaison figurative (Soublin, 1971). En rhétorique, cette forme canonique se compose de cinq éléments :

- le comparé (« fille ») ou terme source qui est décrit totalement ou partiellement par la comparaison;
- le verbe (« est ») qui introduit le motif ou désigne soit une aptitude, soit un comportement sur lequel porte la comparaison ;
- le tertium comparationis ou motif (« calme ») qui représente l'attribut que les entités comparées ont en commun ;
- l'outil de comparaison ou marqueur (« comme ») qui établit le rapport de similitude ou de différence ;
- le comparant (« lac ») ou terme cible qui sert de point de référence pour la comparaison (Hanks, 2012).

Dans la pratique, tous ces éléments, hormis l'outil de comparaison et le comparant, peuvent être omis. Du point de l'ordre des constituants, le comparant, en qualité de complément, suit nécessairement l'outil de comparaison même si la position du reste des constituants peut varier tout en respectant l'ordre syntaxique prévalant dans la langue, dans le cas du français sujet – verbe – objet. Sur le plan grammatical, ce type de constructions se classe parmi les subordonnées comparatives averbales qui sont des versions elliptiques d'une proposition principale et dans lesquelles le comparant occupe la même fonction que le comparé dans la principale (Fuchs et al, 2008). Il existe ainsi une corrélation entre la fonction grammaticale des éléments de la comparaison et leur rôle sémantique : dans la phrase (5), par exemple, « lac » qui est le comparant, remplace dans la subordonnée le comparé « fille », et les deux substantifs sont des sujets.

La comparaison étant avant tout une question de sens, en plus de « comme » qui est incontestablement le marqueur prototypique de la comparaison figurative, la langue française dispose de plusieurs termes susceptibles d'inférer un rapport de similitude ou de dissemblance. Bouverot (1969) oppose ainsi les comparaisons de type I introduites par les comparatifs ou des outils de la forme « déclencheur + que » (ainsi que, de même que...) aux comparaisons de type II reposant sur des adjectifs, des verbes, des suffixes ou des locutions prépositionnelles. Le Tableau I présente l'ensemble des structures possibles pour les comparaisons de type I et de type II en posant la phrase comme contexte de réalisation.

Comparatifs et locutions prépositionnelles	Verbes et locutions verbales	Adjectifs qualificatifs
A/ Marqueur + comparant <i>Il aime briller. Comme les étoiles.</i>	A/ Marqueur + comparant <i>Moi, ressembler à une étoile ?</i>	A/ Marqueur + comparant <i>Il aime briller. Telle une étoile</i>
B/ Comparé + marqueur + comparant <i>Vous êtes revigoré. Souriant. Les yeux comme des étoiles.</i>	B/ Comparé + marqueur + comparant <i>Ses yeux ressemblent à deux étoiles scintillantes.</i>	B/ Comparé + marqueur + comparant <i>Vous êtes revigoré. Souriant. Les yeux pareils à des étoiles.</i>
C/ Verbe + marqueur + comparant <i>Ne jamais filer comme une étoile après le crime.</i>		C/ Verbe + marqueur + comparant <i>Ne jamais filer telle une étoile après le crime.</i>
D/ Comparé + verbe + marqueur + comparant <i>Sa lame luit comme une étoile.</i>		D/ Comparé + verbe + marqueur + comparant <i>Ses yeux sont semblables à des étoiles.</i>
E/ Adjectif motif + marqueur + comparant <i>Tout un peuple. Innombrable comme les étoiles !</i>		
F/ Comparé + adjectif motif + marqueur + comparant <i>Une ville exotique, aussi lointaine que les étoiles dans le ciel.</i>		
G/ Comparé + verbe + adjectif motif + marqueur + comparant <i>Son regard brille ainsi qu'une étoile.</i>		

TABLEAU 1 : Structures des comparaisons figuratives en fonction du marqueur

2.2 Comparaisons figuratives et TAL

En ce qui concerne la recherche sur les comparaisons en traitement automatique des langues, il est possible de délimiter deux phases principales : une première phase linguistique en majorité descriptive et une phase informatique axée sur le développement d'outils pour détecter et analyser des types de comparaisons spécifiques. Bien que relativement récent, le domaine de l'analyse automatique des comparaisons figuratives diffère de celui de la détection des phrases comparatives autant par ses objectifs que par ses méthodes. De manière générale, les approches existantes tentent de tirer parti de la corrélation entre la fonction grammaticale et le rôle sémantique des constituants des comparaisons figuratives. Deux outils ont été testés pour ce faire : GLARF (Meyers *et al.*, 2001 ; Niculae et Yaneva, 2013) qui enrichit la sortie des analyseurs syntaxiques en constituants, notamment en identifiant les sujets, les objets et les noyaux des syntagmes, et un analyseur syntaxique profond, TurboParser (Martins *et al.*, 2010 ; Niculae, 2013). Dans les deux cas, l'identification repose sur les étapes suivantes :

- Établir une liste de marqueurs ;
- Parcourir les nœuds de la structure en arbre de la phrase jusqu'à trouver un substantif 1 étiqueté comme étant un complément d'un des marqueurs ;
- Identifier dans l'arbre un lien rattachant le marqueur à un verbe ;
- Repérer dans l'arbre un lien connectant le verbe identifié en 3 à un substantif 2 étiqueté comme étant son sujet ;

- Trouver dans l'arbre un lien qui subordonne un adjectif qualificatif au verbe identifié en 3 ;
- Si les étapes 2 à 4 ont des résultats positifs, la phrase est extraite, le substantif 1 est considéré comme étant le comparant, le marqueur comme l'outil de la comparaison, le substantif 2 comme le comparé et le verbe est extrait.
- Si l'étape 5 a un résultat positif, l'adjectif qualificatif est considéré comme étant le motif.

De plus, la comparaison des résultats obtenus avec les deux outils montre que l'analyse syntaxique profonde donne de meilleurs résultats : par exemple, sur un set de 53 phrases, on constate un rappel plutôt haut (71 % contre 43 % avec GLARF) pour une précision assez basse, 24 % (Niculae, 2013). Différentes raisons pourraient être avancées pour expliquer cette performance :

- avec les comparatifs, seules deux structures de comparaisons figuratives sont reconnues sur les sept possibles ;
- la polysémie des marqueurs comme « like » qui peut aussi être une forme verbale ;
- l'exploration ne prévoit pas des structures où le comparé est un complément d'objet direct comme dans la phrase « *l'homme ligota ses frères ainsi que des saucissons* » ;
- l'exploration de l'arbre ne considère pas les verbes juxtaposés ou coordonnés à d'autres verbes et les propositions ayant un pronom relatif pour sujet ;
- la fouille de l'arbre ne tient pas compte des comparaisons figuratives ayant plus d'un comparant ou comparé.
- l'extraction des structures comparatives concerne aussi bien les subordonnées comparatives averbales que des subordonnées comparatives contenant des verbes ou d'autres subordonnées introduites par le marqueur.

Ce dernier point a toute son importance puisqu'il existe une différence sémantique non-négligeable entre les subordonnées comparatives verbales mettant en parallèle deux entités et les subordonnées comparatives verbales qui contrastent deux actions ou processus. Pour finir, les méthodes proposées ne recherchent qu'une seule comparaison par phrase et ignorent donc le reste des comparaisons dans le cas de phrases contenant plusieurs comparaisons figuratives.

Au niveau de la désambiguïsation des comparaisons extraites, la méthode proposée s'appuie sur l'apprentissage automatique et la sémantique distributionnelle pour mesurer la similarité sémantique entre le comparé et le comparant qui est combinée avec d'autres attributs tels que le domaine du comparant et la présence d'un article indéfini avant celui-ci (Niculae & Danescu-Niculescu-Mizil, 2014).

3 Extraction et analyse des comparaisons et des pseudo-comparaisons

3.1 Description de l'algorithme

Au regard de notre objectif qui est d'identifier toutes les comparaisons figuratives que renferme un texte littéraire, différents choix méthodologiques ont été faits : identifier les comparaisons de type I mais aussi celles de type II, détecter plus d'une comparaison par phrase le cas échéant, ne pas se limiter aux substantifs comparés, tenir compte de l'ambiguïté des comparaisons figuratives et extraire tous les comparés syntaxiquement possibles. Nous distinguons ainsi quatre groupes de marqueurs de la comparaison :

- les marqueurs traditionnels : *comme, ainsi que, de même que, autant que, plus...que, tel que, moins...que, aussi...que* ;
- les verbes : *ressembler à, sembler, faire l'effet de, faire penser à, faire songer à, donner l'impression de* ;
- les adjectifs qualificatifs : *semblable à, pareil à, tel, similaire à, analogue à, comparable à* ;
- et les locutions prépositionnelles : *à la manière de, à l'image, à l'égal de, à l'instar de, à la façon de*.

L'extraction de comparaisons et de pseudo-comparaisons s'intéresse uniquement aux structures de la forme marqueur + SN ou marqueur,..., SN dans lesquelles le comparant n'est pas un sujet. Afin de mieux circonscrire les phrases concernées, nous avons défini la règle suivante :

Règle 1. Soit un syntagme nominal SN placé immédiatement après un marqueur, le substantif X, noyau de SN, est considéré comme étant un sujet si un verbe conjugué est placé après X et n'est pas séparé de celui-ci par une virgule, un point-virgule, un pronom personnel sujet, un pronom relatif ou une conjonction de subordination.

Une fois le comparant identifié et la phrase extraite, la recherche des constituants de la comparaison se fait vers la gauche uniquement si le marqueur ne se trouve pas en début de phrase, après un signe de ponctuation ou une conjonction de coordination, vers la droite uniquement si le marqueur se trouve en début de phrase, et dans les deux sens s'il est placé directement après un signe de ponctuation ou une conjonction de coordination.

Notre hypothèse de travail principale repose sur le fonctionnement de la syntaxe du français et suppose que si l'on arrive à déterminer la catégorie grammaticale du mot que la structure marqueur + comparant complète syntaxiquement, on peut ainsi inférer la fonction grammaticale du comparé dans la proposition principale et extraire les autres composants de la comparaison. Si ce mot est un verbe, le comparant sera soit le sujet, soit le COD de ce verbe, si c'est un adjectif, le comparant sera forcément le mot que modifie cet adjectif qui peut également être le sujet ou le COD du verbe en fonction de la fonction de l'adjectif et enfin, si ce mot est un substantif, ce substantif est le comparant. De point de vue de la nature des sujets, nous nous sommes limités aux substantifs, aux adjectifs démonstratifs et aux pronoms personnels. Une liste d'indices textuels a été compilée pour vérifier la fonction des constituants recherchés. Par exemple, un adjectif motif ne peut être séparé du marqueur par une conjonction de coordination, un pronom relatif, une préposition ou un syntagme nominal.

3.2 Résultats expérimentaux

L'algorithme présenté dans la section précédente a été testé sur un corpus composé de poèmes en prose écrits par quatre poètes français : Aloysius Bertrand, Stéphane Mallarmé, Charles Baudelaire et Arthur Rimbaud. Ce corpus a été annoté manuellement. Nous nous sommes servis de TreeTagger (Schmid, 1994) pour la tokenisation, l'étiquetage morphosyntaxique et l'analyse syntaxique de surface. Nous avons également écrit des règles qui exploitent la sortie de TreeTagger pour la segmentation en phrases.

Nous avons comparé la performance de notre algorithme à celle d'une version améliorée du système proposé par Niculae (2013) se basant sur des dépendances syntaxiques fournies par le Berkeley Parser (Candito *et al.*, 2010). Les résultats obtenus sont présentés dans le Tableau 2. Pour chaque méthode et chaque classe de constituants, le rappel (vrais positifs/vrais positifs + faux négatifs) et la précision (vrais positifs/vrais positifs + faux positifs) ont été calculés.

	Rp (%)	Pr (%)	VP	FP	FN
Comparé	61,9	46,9	163	184	100
Verbe	55,5	52,8	75	67	60
Adjectif motif	58	69,1	83	37	60
Comparant	90,8	96,7	238	8	24

	Rp (%)	Pr (%)	VP	FP	FN
Comparé	54,3	50,1	143	142	120
Verbe	64,4	47,8	87	95	48
Adjectif motif	44	69,2	63	28	80
Comparant	87	90	228	23	34

TABLEAU 2 : Évaluation du Berkeley Parser (à droite) et de l'algorithme (à gauche). La précision (Pr), le rappel (Rp), les vrais positifs (VP), les faux positifs (FP) et les faux négatifs (FN) sont indiqués.

Contrairement à notre algorithme, le Berkeley Parser peut directement décider si un comparant détecté est utilisé comme sujet ou non. Il commet cependant au niveau de la reconnaissance de comparants plus d'erreurs d'étiquetage morphosyntaxique (36 % des erreurs) que TreeTagger (14%). Les autres erreurs du Berkeley Parser pour cette tâche sont dues à une mauvaise segmentation de phrase, à un comparant faussement identifié comme étant sujet ou à une dépendance erronée. D'autre part, en ce qui concerne l'identification des verbes, le participe passé pose un problème car dans l'annotation manuelle, en fonction de son emploi, il est tantôt considéré comme adjectif, tantôt comme un verbe.

Soulignons également différentes structures qui sont problématiques pour les deux méthodes :

- le participe utilisé comme nom : « ... *coupable à l'égal d'un faux scandalisé* » ;

- les structures « plus de X que de Y » : « *il y a plus de sbires que de citadins* » ;
- l'accumulation de comparaisons dans une même phrase : « *ses cheveux longs comme des saules et peignés comme des broussailles.* »
- l'inversion du sujet : « *cette solide cage de fer derrière laquelle s'agite, hurlant comme un damné, secouant les barreaux comme un orang-outang exaspéré par l'exil, imitant, dans la perfection, tantôt les bonds circulaires du tigre, tantôt les dandinements stupides de l'ours blanc, ce monstre poilu dont la forme imite assez vaguement la vôtre.* »
- les comparés absents : « *Ce soir à Circeto des hautes glaces, grasse comme le poisson, et enluminée comme les dix mois de la nuit rouge, - (son cœur ambre et spunk), - pour ma seule prière muette comme ces régions ...* »
- la présence d'un adjectif non motif avant le marqueur : « *Il est aussi difficile de supposer une mère sans amour maternel qu'une lumière sans chaleur.* »
- un sujet éloigné de son verbe : « *de tous les coins, des fissures des tiroirs et des plis des étoffes s'échappe un parfum singulier, un revenez-y de Sumatra, qui est comme l'âme de l'appartement.* »
- l'accumulation d'adjectifs : « *Les meubles sont vastes, curieux, bizarres, armés de serrures et de secrets comme des âmes raffinées.* »

4 Désambiguïsation des comparaisons figuratives

La phase de la reconnaissance des constituants soulève une question importante pour la désambiguïsation des comparaisons figuratives. En effet, le comparant ou comparé réel dans une phrase n'est pas toujours le noyau du groupe nominal mais souvent son complément comme dans : « *toutes les richesses flambant comme un milliard de tonnerres.* » Cet aspect devrait donc être pris en compte au moment de l'identification des constituants. De plus, sans résolution des pronoms anaphoriques, la désambiguïsation ne peut concerner que les substantifs.

Des structures grammaticales permettent cependant de distinguer les comparaisons littérales des comparaisons figuratives à l'instar de « il y a comme », « c'est comme » et les verbes « élire, nommer, citer, attribuer, considérer, juger, témoigner » employés avec « comme » (Fuchs *et al.*, 2008).

En accord avec les pratiques littéraires de description des comparaisons figuratives, nous avons pris le parti de définir une comparaison figurative par un écart entre les traits sémantiques (concret, abstrait, inanimé, animé) et/ou le domaine du comparant et du comparé. Pour tester cette définition, nous avons utilisé le Dictionnaire électronique des mots de Jean Dubois et François Dubois-Charlier¹ qui renseigne entre autres sur l'animéité et le domaine auquel appartient le substantif. Au vu de la polysémie des substantifs, le problème de leur désambiguïsation se pose. Nous avons donc choisi pour chaque substantif de choisir la première catégorie proposée par le dictionnaire qui est en général la catégorie la plus fréquente. Pour déterminer s'il y a une distance sémantique entre deux substantifs, par défaut, nous attribuons une valeur 1 s'ils ont des natures différentes et une valeur 2 s'ils appartiennent à différents domaines. Cette mesure semble marcher assez bien pour certains termes assez proches comme « haine » et « amour », « charité » et « orgueil » mais pas pour d'autres comme « rue » et « faubourg » ou encore « jour » et « nuit ». On remarque qu'un changement d'animéité est généralement beaucoup plus informatif qu'un changement de domaine mais ne concerne que très peu de couples. Cela laisse supposer qu'une méthode qui s'appuierait sur plus de traits sémantiques ou sur la catégorie naturelle des substantifs donnerait de meilleurs résultats.

5 Conclusion

Notre travail s'inscrivait dans le cadre de la stylistique automatique et avait pour but de présenter ainsi que d'évaluer une méthode pour extraire et analyser automatiquement les (pseudo-)comparaisons. Il nous paraît important dans un premier temps de réduire le bruit, surtout en ce qui concerne la détection des comparés. L'utilisation de ressources linguistiques nous paraît pour ce faire une piste intéressante. Nous songeons aussi à adapter et à tester ce système sur des langues qui possèdent des formes de comparaisons figuratives assez proches de celles du français, comme l'anglais.

¹ Disponible à l'adresse suivante : <http://rali.iro.umontreal.ca/rali/?q=fr/dictionnaire-electronique-des-mots-dem>

Remerciements

Ce travail a bénéficié d'une aide d'Etat gérée par l'Agence Nationale de la Recherche dans le cadre des Investissements d'Avenir portant la référence ANR-11-IDEX-004-02.

Références

- BOUVEROT B. (1969). Comparaison et métaphore. *Le Français moderne* 2, 132-147, 224-238 et 301-316.
- CANDITO M., NIVRE J., DENIS P., ANGUIANO E. H. (2010). Benchmarking of statistical dependency parsers for French. *Proceedings of COLING 2010*, 108-116.
- FUCHS C., FOURNIER, N., LE GOFFIC P. (2008). Structures à subordonnée comparative en français : Problèmes de représentations syntaxiques et sémantiques. *Linguisticae Investigationes* 31:1, 11-61.
- GLUCKSBERG S, KEYSAR B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review* 97:1, 3-18.
- HANKS P. (2012). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review* 97:1, 3-18.
- MARTINS A., SMITH N., XING P., AGUIAR P., FIGUEIREDO M. (2001). Parsing and GLARFing. *Proceedings of RANLP*, 110-114.
- MEYERS A., KOSAKA M., SEKINE S., GRISHMAN R., ZHAO S. (2010). Turbo parsers: Dependency parsing by approximate variational inference. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 34-44.
- NICULAE V. (2013). Comparison pattern matching and creative simile recognition. *Joint Symposium on Semantic Processing, Textual inference and Structure in Corpora*, 110-114.
- NICULAE V., DANESCU-NICULESCU-MIZEL C. (2014). Brighter than gold. *JProceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2008-2018.
- NICULAE V., YANEVA V. (2013). Computational considerations of comparisons and similes. *Proceedings of the ACL Research Student Workshop*, 89-95.
- SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, 44-49.
- SOUBLIN F. (1971). Sur une règle rhétorique d'effacement. *Langue française* 11, 102-109.

Proposition méthodologique pour la détection automatique de Community Manager. Étude multilingue sur un corpus relatif à la Junk Food

Johan Ferguth ¹ Aurélie Jouannet ¹ Asma Zamiti ¹ Yunhe Wu ¹ Jia Li ¹

Antonina Bondarenko ¹ Damien Nouvel ¹ Mathieu Valette ¹

(1) ERTIM, INALCO - 2 rue de Lille, 75007 Paris

jferguth@gmail.com, jouannet.aurelie@gmail.com, zamiti.asma@gmail.com,
yunhe.wu9@gmail.com, maggielleebonnie@gmail.com, tonyabondarenko@gmail.com,
damien.nouvel@inalco.fr, mvalette@inalco.fr

Résumé. Dans cet article, nous présentons une méthodologie pour l'identification de messages suspectés d'être produits par des Community Managers à des fins commerciales déguisées dans des documents du Web 2.0. Le champ d'application est la malbouffe (junkfood) et le corpus est multilingue (anglais, chinois, français). Nous exposons dans un premier temps la stratégie de constitution et d'annotation de nos corpus, en explicitant notamment notre guide d'annotation, puis nous développons la méthode adoptée, basée sur la combinaison d'une analyse textométrique et d'un apprentissage supervisé.

Abstract.

Methodological Proposal for Automatic Detection of Community Manager. Multilingual Study based on a Junk Food corpus

This article describes the methodology for identifying a certain kind of speech in internet forums. The detection of the speech of a Community Manager combines recent issues in the domain of Natural Language Processing, including opinion mining and sentiment analysis, with another more abstract problem. Going beyond detecting the polarity of a message, this project targets the underlying intentions and identity of the author of the message on the forum.

Mots-clés : Community Management, Textométrie, Multilinguisme, Fouille de texte.

Keywords : Community Management, Textometry, Multilingualism, Data Mining.

Introduction

Le métier de gestionnaire de communauté (*community manager* désormais CM) est né de l'essor des sites communautaires et des réseaux sociaux. Ces derniers ont bouleversé la relation hiérarchique, verticale et unilatérale qui était établie jusqu'alors entre les entreprises et les consommateurs. Ainsi, chaque individu est désormais un émetteur légitime. Ce nouveau mode de communication a obligé les entreprises à repenser entièrement leur stratégie pour placer les communautés au centre de leurs dispositifs. Dans ce contexte, le CM a notamment pour mission de développer la visibilité de l'entreprise et de faire fructifier son capital social (e-réputation, etc.). Si, habituellement, dans l'entreprise, l'existence du CM est officielle et son activité non dissimulée, il arrive que certains professionnels valorisent leurs produits ou enseignes en recourant à de faux commentaires positifs. L'objectif est de noyer un nombre important de vrais commentaires négatifs ou de pallier l'absence de commentaires.

Nous relatons dans cet article une expérience de classification automatique de commentaires de forums de discussions ayant trait à la restauration rapide (fast-food). L'objectif applicatif est d'identifier les messages potentiellement écrits par des CM sans que ceux-ci se soient explicitement présentés comme tels. Dans cette optique, nous allons adopter une approche hybride associant une méthode de linguistique de corpus pour la production de descripteurs sémantiques et une phase d'apprentissage supervisé pour la classification des messages. Dans un premier temps nous exposerons notre problématique en mettant le focus sur la difficulté d'identifier le texte d'un énonciateur dont les intentions sont dissimulées.

Nous avons développé quatre corpus de langues différentes : anglais, chinois, français et russe¹. Pour des raisons éditoriales, nous nous focaliserons sur le corpus français pour la description précise de la méthodologie. Pour les autres langues, nous nous contenterons de présenter quantitativement les corpus et de donner les résultats de la classification.

Après un bref état de l’art, nous présenterons en détails notre corpus et la méthodologie de constitution en 2.1. Ensuite, nous détaillerons la méthodologie de sélection des descripteurs et leur caractérisation sémantique en 2.2. Nous indiquerons comment nous avons utilisé l’apprentissage automatique en 3. Enfin, nous concluerons avec un bilan du travail réalisé et présenterons quelques perspectives.

1 État de l’art

La problématique de la tromperie (*deception*) est corrélée à trois applications principales, relativement semblables : l’attribution d’auteurs, la détection de plagiat et la détection de messages à intention dissimulée. Pour les deux premières, les méthodes stylométriques (Juola 2012) ou par mesure statistique de la distance intertextuelle (Luong, éd. 2003) sont souvent utilisées, en plus des méthodes de similarité standard de la fouille de textes (Koppel *et al.* 2009, Afroz *et al.* 2012). D’autres utilisent des méthodes hybrides articulant une approche linguistique, comme par exemple (Rubin et Vashchilko 2012) qui associent une théorie linguistique d’inspiration cognitive (Rhetorical Structure Theory) et des méthodes d’apprentissage supervisé.

En fouille de texte, si le champ des méthodes utilisables est relativement bien balisé et correspond à l’état de l’art (dominé par les méthodes de classification), la constitution des jeux de données pour l’application de méthode supervisée est rendue difficile par la nature intrinsèquement cachée des corpus et fait, en soi, l’objet de recherche. (Gokhman *et al.* 2012) en font un intéressant état de l’art et distinguent deux approches. La première, que nous pourrions qualifier de descendante, consiste à demander à des participants de produire leur corpus dupé, volontairement ou par annotation rétrospective de textes. Cette pratique du panel volontaire est souvent utilisée dans le contexte de recherche en psychologie cognitive. Des approches ascendantes, plus en phase avec les pratiques de la fouille de textes, consistent à élaborer une heuristique pour l’identification de messages suspects. Par exemple, en mesurant sur un site recensant des avis de consommateurs (e.g. Amazon) les commentaires très similaires en fréquences anormalement élevées et/ou en un court laps de temps (Wu *et al.* 2010b) (Jindal et Liu 2008) ou en identifiant des anomalies dans les écarts de satisfaction sur un même produit (Wu *et al.* 2010b). (Gokhman *et al.* 2012) considèrent que ces approches n’offrent pas un véritable *gold standard* mais une approximation acceptable. Nous nous inspirons de cette méthodologie dans notre travail.

2 Corpus et méthodologie

2.1 Stratégies d’élaboration du jeu de données et présentation du corpus

La difficulté inhérente à ce type de tâches réside dans l’obtention d’un corpus annoté de textes écrits par des énonciateurs qui dissimulent leurs intentions. Pour pallier cette difficulté nous avons établi une catégorisation binaire : commentaires *suspectés* d’avoir été rédigés par un CM *vs* commentaires *non suspectés* d’avoir été rédigés par un CM. Nous avons sélectionné manuellement entre 300 et 700 commentaires (de dix mots minimum) suivant les langues de l’étude, principalement sur des forums de discussion². Nous avons effectué une campagne d’annotation comprenant une validation croisée pour chaque langue (2 à 3 annotateurs par langue). Le guide d’annotation a été réalisé avec la participation d’un ancien professionnel du *community management*, qui nous a exposé les stratégies rédactionnelles les plus fréquentes. On pourra en lire le détail dans la liste des règles de sélection ci-après. En complément du corpus web 2.0, nous avons collecté un ensemble de textes émanant des sites web institutionnels des marques considérées dans le contexte applicatif de la *junk-food*, de façon à construire un corpus de référence qui sera intégré dans les tâches de classification. Nous avons fait l’hypothèse que ce corpus nous permettrait de caractériser un discours propre aux CM (nonobstant les variations liées aux genres textuels très différents) et d’établir une échelle de la suspicion.

- Règles de sélection des suspects
 - très (trop) bonne expression écrite,
 - stratégie de mauvaise écriture,

1. Les résultats obtenus sur le corpus russe ayant été jugés peu suffisants, nous ne les présenterons pas ici

2. Exemples pour le français : forum.hardware.fr, Tripadvisor, Doctissimo, www.jeuxvideo.com

- discours uniquement orienté vers le fait de convaincre,
- défense d'une marque contre un scandale,
- valorisation d'un produit et dénigrement d'un produit similaire ou valorise une enseigne plutôt qu'une autre (ex : *les frites de McDo sont pas terribles, je préfère celles de Quick*),
- insiste sur les quantités (ex : *Par exemple, chez Mc Do, je prends le filet-o-fish(un peu moins de 400 cal)*),
- vocabulaire choisi (ex : *Avec gros plan sur ces délicieux beignets de poulets qui vont faire trempette dans des sauces de toutes les couleurs. Miam !*),
- proposition de réductions (ex : *Et avec la carte étudiant, tu peux en avoir 3 de plus pour 1 €si je me souviens bien*),
- conseil (ex : *Essaye d'y aller a plusieurs et prendre un bucket vraiment bon rapport, pour avoir pour son argent*),
- déculpabilisation (ex : *Nous y allons sans aucun complexe ni scrupule !*),
- projection de la vie quotidienne mais générale (ex : *c'est à dire lors des voyages pour partir en vacances, ou lorsque nous faisons plusieurs magasins (ex : meubles, soldes)*),
- reprise claire d'un contre argument (ex : *je suis plutot bon client de Mcdo. . . sans y aller regulierement, j'y ai mes habitudes quand j'ai envie de "gras" avec un gout totalement chimique mais tellement bon. . . pas de fantaisie, j'y commande toujours a peu pres la meme chose (menu maxi best of deluxe/bacon/big mac + croque mcdo ou un autre gros sandwich) pas trop trop cher, ca cale convenablement*),
- très précis (chiffré) (ex : *McDonald's France accueille plus de 1,2 million de clients par jour, soit 440 millions par an ; cela représente 13 clients qui poussent la porte d'un restaurant Mc Donald en moyenne chaque seconde. Pas mal au pays de la gastronomie.*).
- Règles de sélection des non-suspects
 - présence de plusieurs noms de marques ou produits différents,
 - ne pas citer de nom de produit (ex : *j'aime bien aller au fastfood de temps en temps*),
 - parle de soi / sa famille (ex : *en y allant avec ma famille, . . .*),
 - ne valorise pas une enseigne par rapport à une autre (ex : *frites, burger et milkshake : McDo. Quick pour les sauces*),
 - reprise du discours officiel pour le critiquer (comme des listes d'ingrédients et des calories, par ex : *Tout repas supérieur a ce repas "type" est a la fois hypercalorique (on mange facilement 1500kcal au mcdo, donc plus de la moitié des kcal nécessaire journalier !), hyperlipidique (on arrive a 50% lipides avec une part de frites...), et déséquilibré : même une fois par semaine c'est pas tres bon.*).

Le tableau 1 présente le nombre de messages ou de mots collectés pour les différents corpus. Pour le corpus chinois, le module JIEBA³ a été utilisé pour segmenter les textes en mots. Nous constatons des disparités, autant du point de vue du nombre de mots dans les corpus que dans la proportion de messages suspects.

	Français	Anglais	Chinois
Corpus total (mots)	61 205	32 588	8 051
Non-suspects (messages)	249	251	516
Suspects (messages)	61 (20%)	25 (9%)	184 (26%)
Institutionnels (messages/mots)	4 / 30 917	26 / 13 604	3 / 1 850

TABLE 1 – Volume des corpus collectés

2.2 Méthodologie de constitution des descripteurs sémantiques

2.2.1 Sélection textométrique des descripteurs

Nous tentons de mettre en évidence les phénomènes textuels qui différencient les témoignages de nos deux catégories. Notre objectif est de trouver des descripteurs sémantiquement explicables et suffisamment robustes pour servir de caractéristiques aux méthodes d'apprentissage supervisé. Après le test de plusieurs logiciels implémentant les algorithmes de spécificités (Lafon, 1980) utilisés, notre choix s'est porté, pour l'analyse du corpus et l'extraction des critères, sur le logiciel Lexico 3 (Salem *et al.*, 2003), qui s'est avéré le plus robuste dans notre contexte multilingue.

3. <https://github.com/fxsjy/jieba>

Notre méthodologie s'inspire des travaux de Eensoo et Valette (2012, 369-370). Nous avons choisis les descripteurs en effectuant un calcul des spécificités (Lafon, 1980) sur les formes isolées pour chaque sous corpus (suspect/non-suspect). Nous avons ensuite examiné les concodances des candidats descripteurs pour nous assurer de leur pertinence sémantique et de leur monosémie (sauf pour le corpus chinois, les sinogrammes étant très polysémiques). Nous avons ensuite contrôlé visuellement leur répartition uniforme dans le sous corpus considéré en utilisant la fonction de partition (fréquence, spécificités, carte des sections).

En utilisant ce logiciel, nous avons comparé les spécificités de chaque classe à celle des deux autres afin de voir quels termes étaient les plus spécifiques aux suspects, aux non suspects et aux institutionnels. Nous avons ainsi inventorié plusieurs dizaines de critères sémantiques pour chaque classe, puis nous les avons caractérisés en fonction des composantes sémantiques suivant une lecture librement inspirée de (Rastier 2001) : composante *dialectique* (i.e. représentation du temps et du déroulement aspectuel, des rôles et des interactions entre acteurs), composante *dialogique* (positionnement énonciatif des acteurs) et composante *thématique*. Cette dernière inclut une composante *thymique* relative à l'affectivité du locuteur, l'expression de sa subjectivité.

2.2.2 Présentations des descripteurs

La classification de nos descripteurs nous a permis de constater des particularités propres à chacune des catégories. Pour le corpus français, on observe les tendances suivantes. Les messages suspects ont tendance à citer des ingrédients et ont ainsi un propos davantage centré sur des thématiques. Les non suspects, quant à eux, utilisent un contenu thématique plus varié ainsi que des éléments issus de la composante dialectique (emphase, argumentation). Leur positionnement énonciatif est aussi plus prononcé que chez les suspects, par l'usage des pronoms *je* et *tu* (composante dialogique). Enfin, les textes institutionnels, c'est-à-dire issus des sites web des marques, utilisent moins d'éléments caractéristiques, leur discours étant de fait plus formels que le discours provenant des messages de forums. L'utilisation des 1re et 2e personnes du pluriel illustre la relation entreprise-clientèle que l'on retrouve aussi dans leur contenu thématique avec des unités lexicales spécifiques au commerce.

Ci-dessous, on donne le détail de tous les descripteurs pour le français, à titre d'exemple.

- suspects
 - dialogique : (aucun)
 - thématique : *menu, mozzarella, chez, steak, fromage, pizzas, ingrédients, rumeur, nuggets, glaces*
 - thymique : *délicieux, simple*
 - dialectique : (aucun)
- non-suspects
 - dialogique : *je, tu*
 - thématique : *manger, plaisir, malbouffe, santé, semaine, gens, gout, nourriture, alimentation, animaux, plastique, problème*
 - thymique : *mauvais, merde, bon*
 - dialectique : *pense, mais, pas, moins, !, ?, vraiment, très*
- institutionnels
 - dialogique : *vous, vos, nous, votre, notre*
 - thématique : *enfant, formation, préparation, allergène, produits, clients, services*
 - dialectique : (aucun)

3 Classification et évaluation des résultats

Nous utilisons Weka (Hall *et al.*, 2009) pour réaliser l'apprentissage automatique, pour lequel de nombreux algorithmes ont été expérimentés, parmi lesquels *Naïve Bayes* (NB) et *Naïve Bayes Multinomial* (NBM) se sont révélés les plus performants. Selon le corpus, nous disposons de deux ou trois classes. En anglais, la catégorisation doit distinguer entre les messages suspects (S), non-suspects (NS) et institutionnels. Le chinois ne considère que les messages suspects (S) et non-suspects (NS). Pour le français, la stratégie a consisté à fusionner⁴ les classes des messages suspects et institutionnels (S+I) et de les distinguer des messages non-suspects (NS). Il serait de fait intéressant d'aller plus loin en examinant le

4. Cette fusion, basée sur l'hypothèse que les messages les plus suspects se rapprochent du discours institutionnel a permis d'améliorer les performances.

vocabulaire commun aux deux classes, de façon à affiner la stratégie de fusion (on pourrait ainsi considérer uniquement le vocabulaire spécifique aux deux classes). Pour cela, à nouveau, il faudrait obtenir un corpus clairement issu de *CMs* pour pouvoir établir un comparatif de qualité.

Le tableau 2 présente les résultats obtenus pour la classification des textes du corpus d'apprentissage et par validation croisée à dix plis, en les comparant à la stratégie consistant à affecter à tous les messages la classe majoritaire. Nous constatons que les expériences préliminaires, sans descripteurs textométriques et par utilisation du filtre StringToWord-Vector⁵, obtiennent des résultats très proches des classes majoritaires : ceci montre clairement la difficulté de la tâche à laquelle nous sommes confrontés. L'utilisation des descripteurs textométriques permet d'améliorer les résultats en termes d'exactitude pour le français et le chinois.

	Français	Anglais	Chinois
Catégorisation	S + I / NS	S / I / NS	S / NS
Classe majoritaire	80.3%	97.1%	73.7%
Sans descripteurs			
Lemmatisation	non	non	non
Algorithme	NBM	NB	NB
Mots distincts	6000	1000	2854
Exactitude	80,2%	94,9%	74,6%
Avec descripteurs			
Nb descripteurs	51	63	80
Algorithme optimal	NBM	NB	NB
Exactitude	82,8%	91,5%	75,81%

TABLE 2 – Résultats obtenus sur corpus avec ou sans descripteurs sémantiques

Nous remarquons le gain en exactitude pour le français, dont le taux de bonne classification passe de 80,2% à 82,8%. Une analyse plus poussée des résultats sur ce corpus montre que, sur 65 messages suspects (dont 4 institutionnels, par fusion), seuls 3 sont correctement classés sans descripteurs, tandis qu'avec descripteurs, 18 le sont. Cela impacte par effet de bord la bonne classification des messages non-suspects pour lesquels, sur 249 messages, aucun n'était mal classé sans descripteurs, tandis qu'avec les descripteurs 7 deviennent étiquetés de manière erronée comme suspects. Globalement, l'apport est plus remarquable lorsque l'on calcule la f-mesure, qui passe de 72,3% à 79,6% grâce à l'utilisation des descripteurs textométriques.

Pour l'anglais, nos résultats, peu concluants, peuvent s'expliquer de plusieurs manières. D'une part, le grand déséquilibre des commentaires entre la catégorie *suspects* et les deux autres peut rendre la tâche plus difficile (mais nous pensons que ceci est représentatif de ce qui a été trouvé sur internet lors de la constitution du corpus : aussi bien dans les commentaires d'articles que dans les forums, les publications pouvant être catégorisées dans *suspects* selon le guide d'annotation semblent rares et hétérogènes). Il nous a également semblé, au vu des messages plus répandus sur les sites de partage d'avis, que la diffusion de faux commentaires relèverait plutôt d'initiatives personnelles de restaurants que d'une stratégie globale des marques, ce qui peut renforcer leur hétérogénéité. Compte tenu de ces résultats, notre intuition est que les différences culturelles et sociétales (lobbying, culture du fast-food, etc.) dans les pays anglophones, pourrait expliquer que les marques n'aient pas besoin d'intégrer ce type de pratiques dans leurs stratégies globales.

Le cas du corpus chinois présente des particularités : le langage est bien plus hétérogène entre les forums et les sites institutionnels. Ainsi, le registre du corpus institutionnel est beaucoup plus soutenu, ce qui pose des difficultés sur le plan du lexique et impacte fortement les apprentissages automatiques. Il a donc été décidé de ne pas inclure les sites institutionnels, qui n'amélioreraient pas, voire dégraderaient, les résultats. Les expériences portant alors sur deux classes font obtenir un score de 74,61% sur deux classes avec l'algorithme Naive Bayes. En y ajoutant les 60 descripteurs sémantiques repérés à l'aide de Lexico3, nous obtenons 75,81% soit un gain absolu de 1,2 points (malgré la présence de descripteurs sémantiquement ambigus - comme de nombreux mots en chinois).

Nous retenons, essentiellement, deux éléments de ces expérimentations. D'une part, l'utilisation de l'apprentissage automatique, sur une tâche aussi subjective et avec un volume aussi faible de données, est difficile à mettre en œuvre et demande un ajustement méticuleux des paramètres pour éviter le sur-apprentissage. Nous pensons être dans un cas qui se rapproche de la reconnaissance de *signaux faibles*. Nous remarquons que la textométrie aide manifestement à pallier

5. Filtre Weka qui convertit les attributs textuels en vecteurs de mots. Un paramètre permet d'ajuster le nombre de mots distincts sélectionnés (pour le chinois, il correspond au nombre total : 2854).

ces difficultés et permet d’obtenir des résultats qui, s’ils demanderaient à être mieux établis quantitativement, le sont très clairement du point de vue qualitatif : l’utilisation des descripteurs permet d’obtenir plus de messages classés suspects et évite de recourir à une stratégie trop proche de la classe majoritaire (*i.e.* classer tous les messages comme non suspects).

Conclusion

Dans cet article, nous avons tenté d’établir une classification en utilisant des descripteurs sémantiques pour une tâche de reconnaissance des messages dissimulés de gestionnaires de communautés dans un contexte multilingue. Nous avons adopté une méthode hybride associant la linguistique de corpus et la classification supervisée. Les résultats obtenus varient d’une langue à une autre et ne sont pas systématiquement concluants. Les difficultés rencontrées sont cependant inhérentes à la nature même de la problématique initiale, à savoir détecter des émetteurs qui ont pour but la dissimulation. Les résultats sont tout de même encourageants pour une première approche et nous permettent d’envisager des développements ultérieurs.

Références

- AFROZ S., BRENNAN M. & GREENSTADT R. (2012). Detecting hoaxes, frauds, and deception in writing style online. In *Proceedings of the 33rd conference on IEEE Symposium on Security and Privacy*.
- AKOGLU L., CHANDY R. & FALOUTSOS C. (2013). Opinion fraud detection in online reviews by network effects. In *ICWSM*.
- ALMELA Á., VALENCIA-GARCÍA R. & CANTOS P. (2012). Seeing through deception : A computational approach to deceit detection in written communication. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, p. 15–22 : Association for Computational Linguistics.
- BACHENKO J., FITZPATRICK E. & SCHONWETTER M. (2008). Verification and implementation of language-based deception indicators in civil and criminal narratives. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, p. 41–48 : Association for Computational Linguistics.
- DOHSE K. (2013). Fabricating feedback : Blurring the line between brand management and bogus reviews. *Journal of Law, Technology and Policy, Forthcoming*.
- EENSOO E. & VALETTE M. (2012). Sur l’application de méthodes textométriques à la construction de critères de classification en analyse des sentiments. In *TALN 2012*, volume 2, p. 367–374 : GETALP-LIG.
- FENG S., BANERJEE R. & CHOI Y. (2012). Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Short Papers-Volume 2*, p. 171–175 : Association for Computational Linguistics.
- FITZPATRICK E. & BACHENKO J. (2012). Building a data collection for deception research. In *Proceedings of the workshop on computational approaches to deception detection*, p. 31–38 : Association for Computational Linguistics.
- FLEURY S. & ZIMINA M. (2014). Trameur : A framework for annotated text corpora exploration. *COLING 2014*, p.57.
- GOKHMAN S., HANCOCK J., PRABHU P., OTT M. & CARDIE C. (2012). In search of a gold standard in studies of deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, p. 23–30 : Association for Computational Linguistics.
- HALL M., FRANK E., HOLMES G., PFAHRINGER B., REUTEMANN P. & WITTEN I. H. (2009). The weka data mining software : an update. *ACM SIGKDD explorations newsletter*, 11(1), 10–18.
- HAMON T. (2012). Acquisition terminologique pour identifier les mots clés d’articles scientifiques. *Actes du huitième Défi Fouille de Textes*, p.28.
- HAUCH V., MASIP J., BLANDON-GITLIN I. & SPORER S. L. (2012). Linguistic cues to deception assessed by computer programs : a meta-analysis. In *Proceedings of the workshop on computational approaches to deception detection*, p. 1–4 : Association for Computational Linguistics.
- JINDAL N. & LIU B. (2008). Opinion spam and analysis. In *WSDM ’08 Proceedings of the 2008 International Conference on Web Search and Data Mining*, p. 219–230 : ACM.
- JUOLA P. (2012). Detecting stylistic deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, p. 91–96 : Association for Computational Linguistics.

- KOPPEL M., SCHLER J. & ARGAMON S. (2009). Computational methods in authorship attribution. In *Journal of the American Society for Information Science and Technology*, p. 9–26.
- LAFON P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, **1**(1), 127–165.
- LI D. & SANTOS JR E. (2012). Argument formation in the reasoning process : toward a generic model of deception detection. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, p. 63–71 : Association for Computational Linguistics.
- LUONG X. (2003). La distance intertextuelle. *Revue Corpus*.
- MO Q. & YANG K. (2014). Overview of web spammer detection. *Journal of Software* 25(7).
- RUBIN V. L. & VASHCHILKO T. (2012). Identification of truth and deception in text : Application of vector space model to rhetorical structure theory. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, p. 97–106 : Association for Computational Linguistics.
- SALEM A., LAMALLE C., MARTINEZ W., FLEURY S., FRACCHIOLO B., KUNCOVA A. & MAISONDIEU A. (2003). Lexico3—outils de statistique textuelle. manuel d'utilisation. *Syled-CLA2T, Université de la Sorbonne nouvelle—Paris*, **3**.
- SPORER S. L. (2012). Making the subjective objective ? : computer-assisted quantification of qualitative content cues to deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, p. 78–85 : Association for Computational Linguistics.
- VARTAPETIANCE A. & GILLAM L. (2012). I don't know where he is not : does deception research yet offer a basis for deception detectives ? In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, p. 5–14 : Association for Computational Linguistics.
- WU, GREENE, SMYTH & CUNNINGHAM (2010a). Distortion as a validation criterion in the identification of suspicious reviews. In *Proceedings of the First Workshop on Social Media Analytics*, p. 10–13 : ACM.
- WU, GREENE, SMYTH & CUNNINGHAM (2010b). *Distortion as a validation criterion in the identification of suspicious reviews*. Rapport interne UCD-CSI-2010-04.

MEDITE : logiciel d'alignement de textes pour l'étude de la génétique textuelle

Zied Sellami¹ Jean-Gabriel Ganascia¹ Mohamed-Amine Boukhaled¹
(1) Laboratoire d'Informatique de Paris 6 (LIP6) 4 Place Jussieu, 75005 Paris
{zied.sellami, jean-gabriel.ganascia, mohamed.boukhaled}@lip6.fr

Résumé. MEDITE est un logiciel d'alignement de textes permettant l'identification de transformations entre une version et une autre d'un même texte. Dans ce papier nous présentons les aspects théoriques et techniques de MEDITE.

Abstract.

MEDITE: text alignment software for the study of textual genetics

MEDITE is an alignment software able to identifying transformations between two versions of a same text. In this paper we show the theoretical and technical aspects of this tool.

Mots-clés : Alignement de textes, Génétique textuelle, Détection d'homologies dans les séquences textuelles

Keywords: Text alignment, Textual genetics, Homology detection in text sequences

1 MEDITE : aspects théoriques

MEDITE est un logiciel d'alignement de textes issu d'une collaboration entre l'ITEM (Institut des Textes et Manuscrits Modernes) et l'équipe ACASA du LIP6.

Initialement MEDITE était prévu pour aligner des transcriptions linéarisées d'avant-textes afin de mettre en évidence les différences et les invariances. Il s'est révélé utile dans de nombreuses autres applications, comme pour établir l'appareil critique d'éditions savantes en comparant les différentes versions publiées d'une œuvre, pour l'étude des variations de textes collectifs, ou encore pour la comparaison de bi-textes afin d'améliorer les outils de traduction statistique (Ganascia, Bourdaillet, 2006). MEDITE est construit sur un algorithme d'alignement de textes par fragments qui recourt à une détection des homologies par la méthode des arbres de suffixes. Il met en évidence les suppressions, les insertions, les remplacements et les déplacements. La première étape de l'algorithme identifie les blocs homologues maximaux. Il s'agit ensuite de distinguer, parmi ces blocs, des pivots et des blocs dits déplacés. Le processus est itéré de façon récursive afin d'éviter les phénomènes de masquage. Enfin, les insertions, les suppressions et les remplacements se déduisent de l'alignement des blocs non répétés (Fenoglio, Ganascia, 2008). L'algorithme de MEDITE étant fondé sur des principes d'algorithmique des séquences, il est indépendant de la langue et peut donc traiter n'importe quel texte, sans ressources spécifiques. En outre, il peut repérer des réutilisations de parties de mots, ce qui s'avère très utile, en particulier pour les langues flexionnelles (Bourdaillet, Ganascia, 2007).

2 MEDITE : aspects pratiques

Une version en ligne de MEDITE est disponible dans <http://obvil.paris-sorbonne.fr/developpements/medite>. La page d'accueil de MEDITE permet à l'utilisateur d'introduire deux textes à comparer et de paramétrer l'outil (voir Figure 1). Les paramètres de MEDITE sont aux nombres de 8 :

- **Sensible à la casse** : l'alignement est sensible ou pas aux caractères majuscules/minuscules ;
- **Sensible aux signes diacritiques** : l'alignement est sensible ou pas aux caractères accentués ;
- **Sensible aux séparateurs** : l'alignement est sensible ou pas à la ponctuation ;
- **Algorithme mots (coché) ou caractères (non coché)** : l'algorithme de comparaison effectue un découpage par mots ou par caractères des segments à comparer ;
- **Colorer uniquement les blocs en commun** : il s'agit d'une option d'affichage des résultats. En cochant cette option, uniquement les blocs communs entre les deux textes seront colorés ;

- **Longueur minimale des chaînes pivots** : paramètre lié aux arbres de suffixes et fixé à une valeur de 5 ;
- **Ratio minimal des chaînes remplacés** : Il s'agit d'indiquer en pourcentage le taux de textes différents entre deux blocs comparés pour considérer cela en tant qu'un remplacement ;
- **Seuil de longueur pour validation lissage** : paramètre lié aux arbres de suffixes et fixé à une valeur de 50 %.

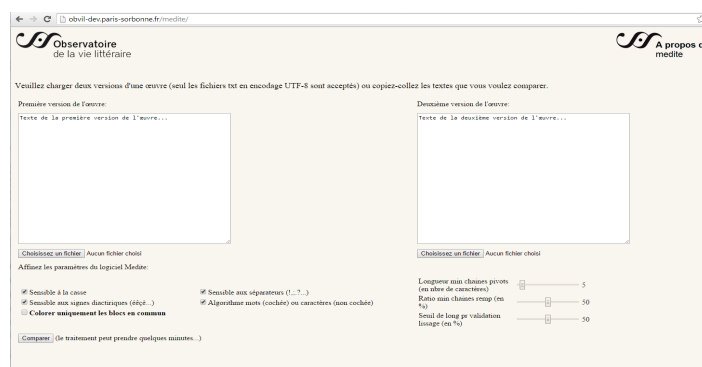


FIGURE 1: Capture d'écran de la page d'accueil de l'outil MEDITE

Les résultats sont affichés après que l'utilisateur clique sur le bouton comparer. Les résultats des alignements sont affichés dans une page de résultats (voir Figure 2). Les deux textes sont présentés côte à côte sur une interface qui met en évidence, au moyen de différentes couleurs, les blocs insérés, supprimés, remplacés et déplacés. Les blocs en commun entre les deux textes sont reliés entre eux par un simple clic de souris. Les résultats peuvent être sauvegardés en cliquant sur l'icône disquette de la page des résultats.

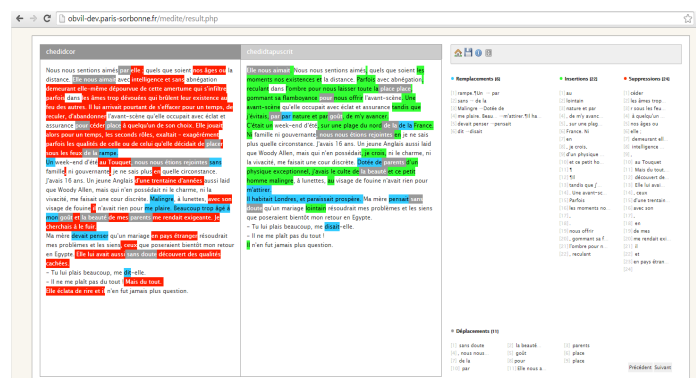


FIGURE 2: Capture d'écran de la page des résultats de MEDITE

Remerciements

Ce travail a bénéficié d'une aide d'État gérée par l'Agence Nationale de la Recherche dans le cadre des Investissements d'Avenir portant la référence ANR-11-IDEX-0004-02

Références

FENOGLIO I., GANASCIA J.-G. (2008). "Le logiciel MEDITE: approche comparative de documents de genèse", in *L'édition du manuscrit - De l'archive de création au scriptorium électronique*, Aurèle Crasson, Academia A|B Bruylant, col. *Au coeur des textes*, n°10, 209-228.

BOURDAILLET J., GANASCIA J.-G. (2007) Alignment of Noisy Unstructured Text Data, *IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data*, Hyderabad, India.

GANASCIA, J.-G., BOURDAILLET, J. (2006). Alignements unilingues avec MEDITE. *Actes des Huitièmes Journées Internationales d'Analyse Statistique des Données Textuelles*.

Phœbus : un Logiciel d'Extraction de Réutilisations dans des Textes Littéraires

Mohamed-Amine Boukhaled, Zied Sellami, Jean-Gabriel Ganascia

LIP6 (Laboratoire d'Informatique de Paris 6), Université Pierre et Marie Curie and CNRS (UMR7606),
ACASA Team, 4, place Jussieu,
75252-PARIS Cedex 05 (France),
{mohamed.boukhaled, zied.sellami, jean-gabriel.ganascia}@lip6.fr

Résumé. Phœbus est un logiciel d'extraction de réutilisations dans des textes littéraires. Il a été développé comme un outil d'analyse littéraire assistée par ordinateur. Dans ce contexte, ce logiciel détecte automatiquement et explore des réseaux de réutilisation textuelle dans la littérature classique.

Abstract.

Phœbus: a Reuse Extraction Software for Literary Text

Phœbus is a reuse extraction software for literary text. It was developed as a computer-assisted literary analysis tool. In this context, the software automatically detects and explores textual reuse networks in classical literature.

Mots-clés : Extraction de réutilisations, empreintes digitales textuelles, analyse littéraire assistée par ordinateur

Keywords: Reuse Extraction, textual fingerprinting, computer-assisted literary analysis.

1 Introduction et Motivation

Les études littéraires et le néo-structuralisme des années soixante et soixante-dix mirent en évidence l'importance des influences mutuelles dans la production littéraire et intellectuelle. Celles-ci se traduisent par des réemplois plus ou moins littéraires et par l'usage d'un vocabulaire similaire. Des notions comme l'intertextualité (Kristeva, 1974) et l'hyper-textualité (Genette, 1982), ont été introduites dès les années soixante-dix pour approcher et formaliser ces phénomènes à partir de l'étude des paraphrases, les réécritures et les citations. L'extraction des réutilisations dans les textes littéraires permet aux chercheurs en littérature de tracer les réemplois d'écrit antérieurs, à regarder l'origine des citations partagées et la façon dont elles sont introduites, d'énumérer les utilisations de proverbes dans les romans et d'étudier d'une manière plus contrôlée la notion d'influence littéraire en termes d'idées et de styles.

L'augmentation de la puissance de calcul des machines et la numérisation des corpus de très grande taille, y compris les corpus de textes littéraires, ont participé au développement des possibilités d'extraction automatique de la réutilisation dans de tel corpus. Ainsi plusieurs méthodes basées sur différentes approches ont été proposées pour effectuer de telles tâches. L'une des approches les plus réussies est l'approche basée sur les empreintes digitales textuelles (Broder, 1997). Cette approche consiste tout d'abord en l'indexation des textes avec des séquences récurrentes de mots. Puis comparer les séquences appartenant à deux textes différents pour extraire les segments textuels communs.

Dans ce contexte, Phœbus a été conçu comme un outil d'analyse littéraire assistée par ordinateur permettant l'extraction de réutilisations dans des textes littéraires. Ce logiciel détecte automatiquement et explore des réseaux de réutilisation textuelle dans la littérature classique.

2 L'approche utilisée

L'approche d'extraction des réutilisations dans Phœbus est basée sur une mise en place et une adaptation de la méthode des empreintes digitales textuelles (Ganascia et al., 2014). Plus précisément, elle se compose de quatre étapes principales :

1. Préparation du texte en utilisant des techniques de traitement automatique du langage naturel ;

2. Extraction de séquences récurrentes élémentaires de mots ;
3. Regroupement des séquences récurrentes élémentaires qui se chevauchent en séquences récurrentes de plus grande taille ;
4. Filtrage des séquences résultantes et élimination des redondances.

3 Fonctionnement de Phœbus

Phœbus a été développé comme une application web avec une architecture client/serveur. L'utilisateur peut y accéder en utilisant un navigateur web via l'adresse suivante : <http://obvil-dev.paris-sorbonne.fr/phoebus>. L'interface utilisateur de Phœbus est assez intuitive, elle se compose de deux champs textuels et de trois contrôleurs de paramètres : Les deux champs textuels servent à copier ou à charger les textes à comparer (textes dont on voudrait extraire des parties similaires présentant une réutilisation). Les contrôleurs permettent de choisir les trois paramètres du programme qui définissent la granularité et la finesse d'analyses à travers les trois propriétés suivantes :

- Le nombre de mots réutilisés à considérer dans les empreintes digitales textuelles, ce qui correspond au nombre de mots en commun entre une réutilisation et son texte original.
- Le nombre de trous autorisés dans ces empreintes pour leurs donner une plus grande capacité de généralisation à des réutilisations non littérales (réutilisations avec changement de quelques mots)
- Le fait de respecter ou pas l'ordre des mots réutilisés entre la réutilisation et le texte original.

Une fois les deux textes choisis et les paramètres définis, l'utilisateur lance l'extraction et les résultats seront affichés et synthétisés (voir Figure.1) d'une façon très ergonomique en incluant entre autre :

- Un alignement automatique des réutilisations entre texte source et texte cible
- La possibilité de navigation dans les différentes réutilisations.
- Une valeur d'importance donnée à chaque réutilisation par critère de taille (Plus la couleur verte est foncée, plus la réutilisation est importante).
- Le couplage avec le logiciel MEDITE qui compare en finesse les transformations textuelles élémentaires (suppressions, insertions, remplacements et déplacements) qui font passer d'un bloc à son semblable.

Figure 1. Exemple d'une page de résultats produite par Phœbus

Remerciement

Ce travail a bénéficié d'une aide d'État gérée par l'Agence Nationale de la Recherche dans le cadre des Investissements d'Avenir portant la référence ANR-11-IDEX-0004-02

Références

Broder, A. Z. (1997). On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings* (pp. 21–29)

Ganascia, J.-G., Glaudes, P., & Del Lungo, A. (2014). Automatic detection of reuses and citations in literary texts. *Literary and Linguistic Computing*, 29(3), 412–421.

Genette, G. (1982), *Palimpsestes : La Littérature au second degré*, Seuil, coll. « Essais », Paris.

Kristeva, J. (1974). *La Révolution du langage poétique*, col. “Tel Quel”, Editions du Seuil.

YADTK : Une plateforme open-source à base de règles Pour développer des systèmes de dialogue oral

Jérôme Lehuen – Carole Lailler – Julien Stenzhorn
Équipe LST, LIUM, Université du Mans
Avenue Laennec, 72085 Le Mans Cedex 9
{Jerome.Lehuen, Carole.Lailler}@lium.univ-lemans.fr

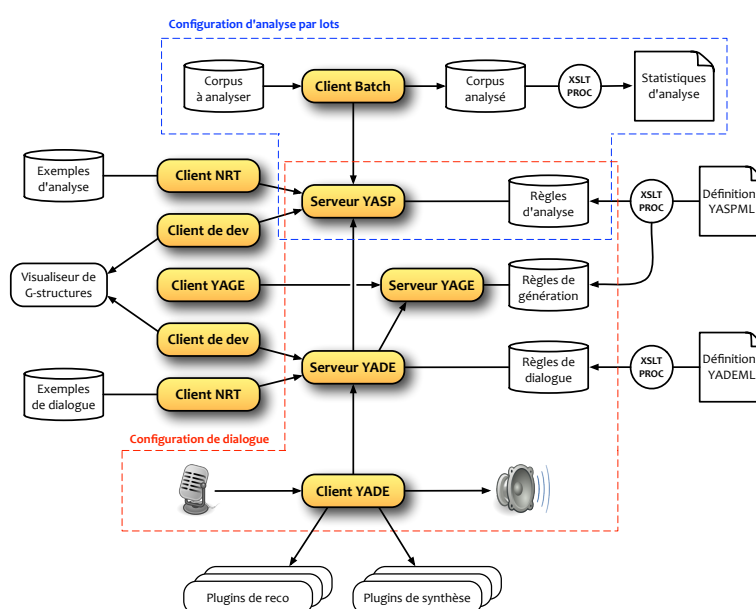
Résumé. YADTK est une plateforme open-source pour développer des systèmes de dialogue oral. De part son caractère déclaratif et unifié, le modèle de représentation des connaissances permet un développement rapide et facilité.

Abstract. YADTK is an open-source, rule-based framework to build spoken dialogue systems. The declarative and unified nature of the model of knowledge representation allows a rapid and easier development process.

Mots-clés : Dialogue Oral Personne-Système, Analyse Sémantique, Logiciel Open-Source.

Keywords: Spoken Dialogue System, Semantic Parsing Tool, Open-Source Software.

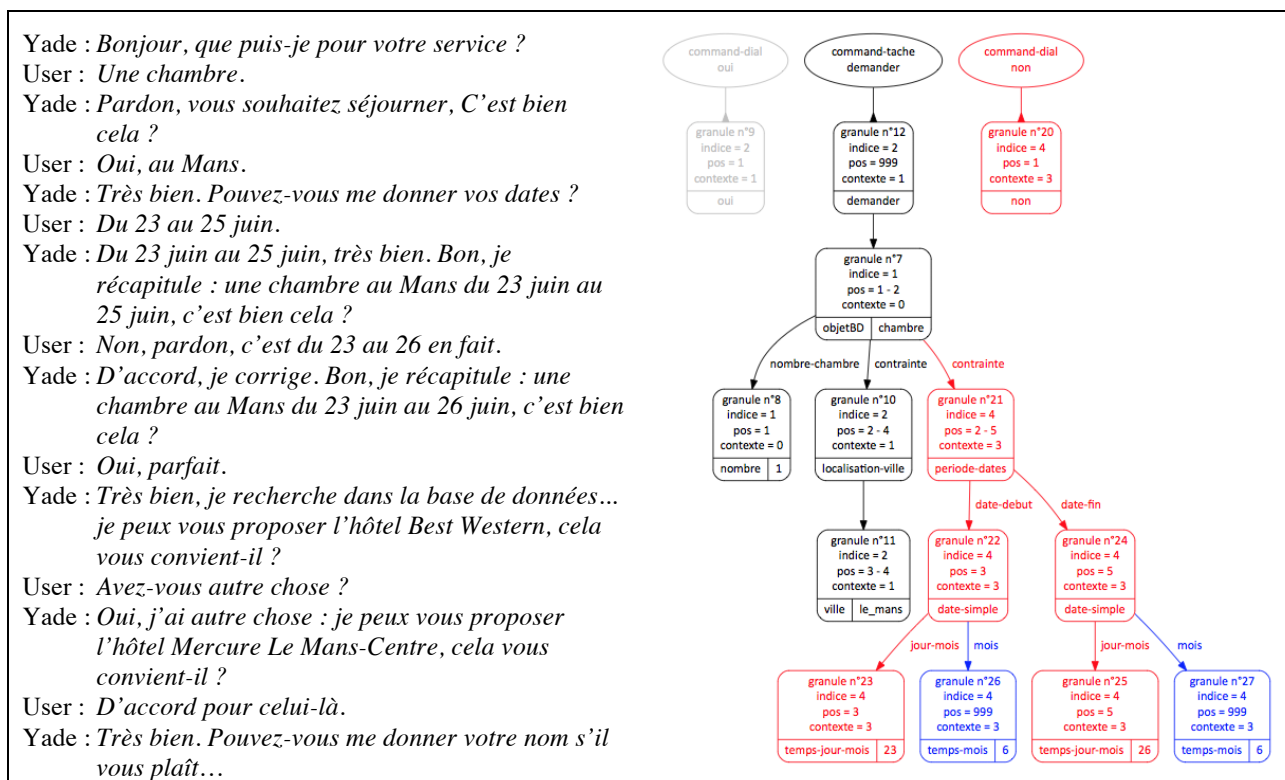
L'objectif de la démonstration est de présenter YADTK, une plateforme pour développer des systèmes de dialogue oraux. Ce logiciel offre la possibilité d'obtenir une représentation sémantique immédiate, clés en mains. Dans la mesure où il ne nécessite en amont ni prétraitements, ni analyses morphosyntaxiques, ni processus d'annotation coûteux, cet environnement intégré permet une fluidité dans le développement. Entièrement symbolique, il se fonde sur un système à base de règles, ce qui autorise un développement incrémental à cycles courts : chaque modification, même mineure, peut donner lieu à des tests. YADTK comporte un module de compréhension (YASP), un module de génération (YAGE), et un moteur de dialogue (YADE) avec entrées et sorties vocales. Il comporte également des modules permettant des tests unitaires, des tests de non-régression, ainsi que des analyses par lots d'énoncés. La figure suivante montre les différents modules et plugins, ainsi que leurs interdépendances, au sein d'une architecture client-serveur :



Ces composants reposent sur un modèle commun de représentation des connaissances appelé **Modèle des Granules** qui permet non seulement de représenter le sens des énoncés sous la forme de **structures conceptuelles hiérarchiques** (G-structures), mais aussi d'opérer des inférences et de déclencher des actions dans le cadre d'un dialogue finalisé. Ce modèle est associé à un formalisme XML de représentation des connaissances appelé YASPML qui décrit une grammaire sémantique réversible. Celle-ci s'inspire des grammaires génératives (Ruwet, 1967), des grammaires de cas

(Fillmore, 1982), et des grammaires de dépendances (Tesnière, 1988). Son caractère réversible lui permet d'être utilisée à la fois en compréhension et en génération.

Le cadre suivant contient un dialogue qui illustre les capacités dialogiques de YADE, ainsi que la structure de Granules construite de façon incrémentale, en parallèle avec la progression du dialogue :



Lors de cette démonstration, nous nous focaliserons sur deux aspects fondateurs de YADTK :

- **L'interactivité** dans le processus de construction des connaissances, induite par des cycles de développement courts et par les outils proposés : **visualisation graphique immédiate** des représentations sémantiques, modifications et amendements facilités, **contrôle de la cohérence globale** grâce aux tests de non-régression.
- **La déclarativité** du langage d'exploitation des représentations sémantiques. Elle prend notamment effet dans le cadre du développement d'un système de dialogue finalisé. En effet, ce langage à base de règles propose des primitives de manipulation des G-structures qui permettent de décrire facilement **inférences** et **prises de décision**.

Pour notre démonstration, nous nous fonderons sur le corpus MEDIA (Bonneau-Maynard et al., 2006). Ce corpus porte sur la réservation de séjours dans des établissements hôteliers partout en France. Nous nous permettrons cependant d'éprouver la portabilité de l'environnement en ayant recours au corpus RITEL (Rosset et al., 2006). Nous montrerons ainsi qu'il est tout-à-fait possible d'exploiter tout ou partie d'une base de connaissances déjà constituée.

Références

- BONNEAU-MAYNARD H. ET AL. (2006). Results of the french evalda-media evaluation campaign for literal understanding. In: *Proceedings of LREC'2006*, Genova, Italy, pp. 2054–2059.
- FILLMORE C. (1982). *Frame semantics*. In: *Linguistics in the Morning Calm*, Linguistic Society of Korea (ed.), Hanshin Publishing, Seoul, 111–37.
- ROSSET S. ET AL. (2006). Interaction et recherche d'information : le projet RITEL. In: *Traitement Automatique des Langues*, 46 (3), 155–179.
- RUWET N. (1967). *Introduction à la grammaire générative*, Plon, Paris.
- TESNIERE L. (1988). *Éléments de syntaxe structurale*, Klincksieck, Paris.

TermLis : un contexte d'information logique pour des ressources terminologiques.

Annie Foret
IRISA, Université Rennes 1, France
foret@irisa.fr

Résumé. Nous présentons TermLis un contexte d'information logique construit à partir de ressources terminologiques disponibles en xml (FranceTerme), pour une utilisation flexible avec un logiciel de contexte logique (CAMELIS). Une vue en contexte logique permet d'explorer des informations de manière flexible, sans rédaction de requête a priori, et d'obtenir aussi des indications sur la qualité des données. Un tel contexte peut être enrichi par d'autres informations (de natures diverses), mais aussi en le reliant à d'autres applications (par des actions associées selon des arguments fournis par le contexte). Nous montrons comment utiliser TermLis et nous illustrons, à travers cette réalisation concrète sur des données de FranceTerme, les avantages d'une telle approche pour des données terminologiques.

Abstract.

TermLis : a logical information context for terminological resources.

We present TermLis a logical information context constructed from terminological resources available in XML (FranceTerme), for a flexible use with a logical context system (CAMELIS). A logical view of a context allows to explore information in a flexible way, without writing explicit queries, it may also provide insights on the quality of the data. Such a context can be enriched by other information (of diverse natures), it can also be linked with other applications (according to arguments supplied by the context). We show how to use TermLis and we illustrate, through this concrete realization from FranceTerme data, the advantages of such an approach with terminological data.

Mots-clés : Applications multilingues, Classification, Extraction d'information, Fouilles de données textuelles, Recherche d'information, Ressources du langage, Données Ouvertes, Qualité des données, Données légales.

Keywords : Multilingual applications, Classification, Information extraction, Textual data mining, Information retrieval, Linguistic resources, Open Data, Information Quality, Legal Information.

Introduction. Cette réalisation se situe dans le cadre de travaux visant à rendre exploitables des données linguistiques par l'approche des systèmes d'information logiques : de telles données ne sont pas toujours simples d'utilisation sans aide, par ailleurs les systèmes d'information logique sont conçus pour permettre une navigation flexible dans des données organisées comme un contexte logique. D'autres travaux se situent dans un cadre similaire (Cellier *et al.*, 2011; Quiniou *et al.*, 2012; Foret & Ferré, 2010; Falk *et al.*, 2014) mais les données y sont de natures différentes, et les buts visés aussi.

Cette démonstration illustre comment une vue en contexte logique permet d'explorer des données terminologiques de manière flexible, avec des facettes sémantiques, des possibilités d'inférences logiques et sans rédaction de requête a priori. Nous traitons le cas des données FranceTerme avec l'outil de gestion de contexte Camelis (le seul outil de gestion de contexte logique disponible à notre connaissance).

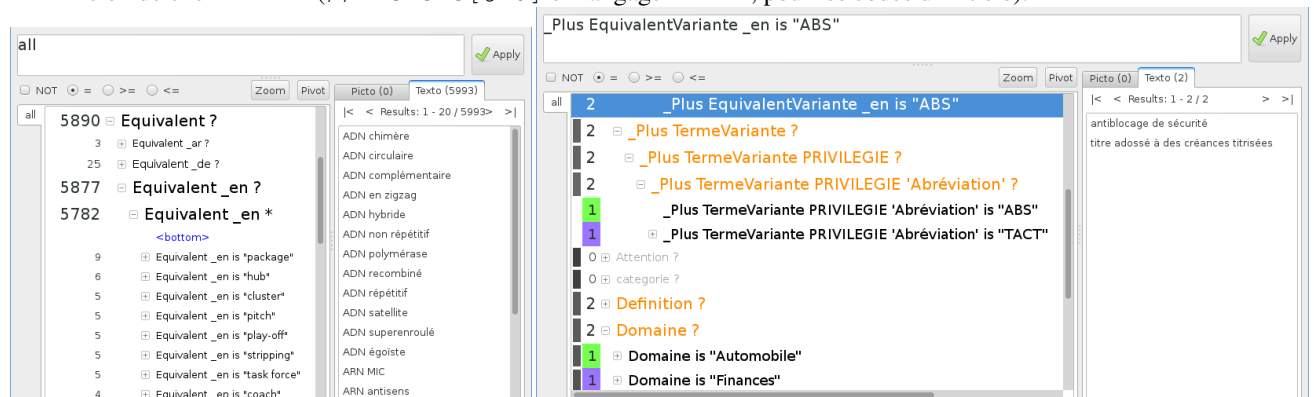
Ressource linguistique. La ressource utilisée présente une richesse de structure : aspects multilingues, des définitions, des relations synonymes etc., une variabilité en fonction d'un domaine/sous-domaine ou en fonction de critères linguistiques (plusieurs variantes d'anglais par exemple), l'absence possible ou la répétition possible de certains types d'informations ; son caractère règlementaire (avec des informations source et date de publication) nous intéresse aussi. Des extensions ultérieures peuvent être envisagées : pour de nouvelles données que l'on organisera selon un schéma analogue.

Contexte logique. Un contexte logique est défini par un ensemble fini d'objets \mathcal{O} , chaque objet o_i ayant un label l_i (a priori plusieurs objets peuvent avoir le même label), et pour chaque objet o_i , un ensemble fini de descriptions logiques $d(o_i)$ chaque expression étant une formule bien formée pour un langage logique donné L . Un *système de gestion de contexte logique* permet de charger et d'exploiter un tel contexte, en permettant l'interrogation d'un contexte par des requêtes logiques (explicites ou interactives), la réponse est alors un sous-contexte d'objets satisfaisant cette requête. Nous avons utilisé Camelis (version 1, accessible à <http://www.irisa.fr/LIS/ferre/camelis/>) ce logiciel est basé sur l'analyse de concept logique (LCA) définie en (Ferré & Ridoux, 2004), une extension de l'analyse de concept formel

(FCA, voir (Ganter & Wille, 1999)) : un *concept logique*, noté c , est un couple formé d'une extension $ext(c)$ (un ensemble d'objets) et d'une intension $int(c)$ (une formule) tel que les éléments de $ext(c)$ sont exactement ceux qui vérifient $int(c)$; ces concepts forment un treillis auquel correspond l'*arbre de navigation logique* et incrémental dans la fenêtre gauche du logiciel. Le logiciel Camelis est aussi prévu pour gérer des ensembles d'objets de types différents.

Contexte TermLis. Le contexte logique a été obtenu en réalisant un transducteur, appliqué au fichier source XML pour FranceTerme. Cette réalisation, dont nous illustrons l'usage avec Camelis (cf figures), s'est appuyée sur :

- un modèle du document XML (une DTD générée automatiquement) ;
- un modèle de contexte logique visé ; avec notamment : un choix de types d'objets ; une sélection de propriétés ;
- un lien entre les deux modèles, comprenant en particulier la spécification d'une clé dans le source (par une expression de chemin XML (`//Article[@id]` en langage XPATH, pour les codes d'Article).



L'arbre de navigation à gauche permet des scénarios variés selon les facettes ouvertes et sélectionnées successivement (all correspond au contexte initial) : recherches simples avec divers types de données (chaines, dates etc.) ; combinaisons logiques ; mise en évidence de variations/faux-amis, d'un élément (comme <Attention>), voire des défauts (redondances).

La *cohérence* est assurée entre les 3 fenêtres (requête/formule logique en haut ; objets à droite ; index de propriétés/arbre de navigation à gauche). Un autre avantage de l'exploration avec Camelis est *sa navigation sûre* évitant les résultats vides.

Des capacités complémentaires pour le contexte sont introduites par : des axiomes logiques pour des recherches avec inférences ; des règles ajoutant des propriétés ; des actions liant à d'autres ressources (CNRTL) ou traitements (analyseur,...). Nous illustrons aussi ces points, un découpage modulaire permettant d'intégrer tout ou une sélection de ces compléments.

Références

- CELLIER P., FERRÉ S., DUCASSÉ M. & CHARNOIS T. (2011). Partial orders and logical concept analysis to explore patterns extracted by data mining. In S. ANDREWS, S. POLOVINA, R. HILL & B. AKHGAR, Eds., *Conceptual Structures for Discovering Knowledge - 19th International Conference on Conceptual Structures, ICCS 2011, Derby, UK, July 25-29, 2011. Proceedings*, volume 6828 of *Lecture Notes in Computer Science*, p. 77–90 : Springer.
- FALK I., BERNHARD D. & GÉRARD C. (2014). From non word to new word : Automatically identifying neologisms in french newspapers. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, p. 4337–4344 : European Language Resources Association (ELRA).
- FERRÉ S. & RIDOUX O. (2004). Introduction to logical information systems. *Inf. Process. Manage.*, **40**(3), 383–419.
- FORET A. & FERRÉ S. (2010). On categorial grammars as logical information systems. In L. KWUIDA & B. SERTKAYA, Eds., *Formal Concept Analysis, 8th International Conference, ICFCA 2010, Agadir, Morocco, March 15-18, 2010. Proceedings*, volume 5986 of *Lecture Notes in Computer Science*, p. 225–240 : Springer.
- GANTER B. & WILLE R. (1999). *Formal concept analysis - mathematical foundations*. Springer.
- QUINIOU S., CELLIER P., CHARNOIS T. & LEGALLOIS D. (2012). What about sequential data mining techniques to identify linguistic patterns for stylistics ? In A. F. GELBUKH, Ed., *Computational Linguistics and Intelligent Text Processing - 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part I*, volume 7181 of *Lecture Notes in Computer Science*, p. 166–177 : Springer.

Etude de l'image de marque d'entités dans le cadre d'une plateforme de veille sur le Web social.

Leila Khouas¹ Caroline Brun² Anne Peradotto³ Jean-Valère Cossu⁴ Julien Boyadjian⁵
Julien Velcin⁶

(1) AMI Software R&D, 1475 av. A. Einstein 34000 Montpellier, France

(2) Xerox Research Centre Europe, 6 chemin de Maupertuis, 38240 Meylan, France

(3) EDF R&D, ICAME, 1 av. du Général de Gaulle, 92140 Clamart, France

(4) LIA CERI, 339 chemin des Meinajariès, 84911 Avignon, France

(5) Centre d'Études Politiques de l'Europe Latine, 39 rue de l'Université, 34060 Montpellier, France

(6) ERIC, Université de Lyon 2, 5 av. P. Mendès-France, 69676 Bron, France

lkh@amisw.com, caroline.brun@xrce.xerox.com, anne.peradotto@edf.fr, jean-valere.cossu@univ-avignon.fr, julien.boyadjian@hotmail.fr, julien.velcin@univ-lyon2.fr

Résumé. Ce travail concerne l'intégration à une plateforme de veille sur internet d'outils permettant l'analyse des opinions émises par les internautes à propos d'une entité, ainsi que la manière dont elles évoluent dans le temps. Les entités considérées peuvent être des personnes, des entreprises, des marques, etc. Les outils implémentés sont le produit d'une collaboration impliquant plusieurs partenaires industriels et académiques dans le cadre du projet ANR ImagiWeb.

Abstract.

Study the brand image of entities as part of a social media-monitoring platform.

The work presented here is about a Web monitoring software providing powerful tools for the analysis of opinions expressed in social media about an entity, such as a celebrity, a company or a brand. The implemented tools result from the research ANR project ImagiWeb involving several industrial and academic partners.

Mots-clés : Plateforme de veille sur internet, médias sociaux, analyse d'opinion, fouille de données.

Keywords: Web monitoring software, social media, opinion analysis, data mining.

1 Introduction et contexte

La multiplication de ressources Web riches en opinions comme les blogs personnels, les médias sociaux ou les commentaires sur l'actualité fournit un important gisement d'information subjective. Celle-ci peut être exploitée pour accéder aux opinions exprimées spontanément par les internautes sur différentes entités comme des personnalités, des entreprises ou des marques de produits. L'analyse de ces données permet d'identifier l'image de l'entité telle qu'elle est perçue ainsi que son positionnement par rapport à d'autres entités. Cette connaissance est précieuse pour mener les actions adaptées en vue de maintenir une bonne image, corriger certains aspects et prévenir d'éventuelles crises d'image. Le travail présenté ici concerne l'intégration, au sein de la plateforme AMIEI¹, d'outils d'analyse permettant l'exploration des opinions émises par les internautes à propos d'une entité donnée, ainsi que l'étude des mécanismes de leur évolution dans le temps. AMIEI est une solution logicielle développée par AMI Software², permettant la mise en place d'un cycle de veille complet dans des contextes divers tels que l'intelligence économique et l'e-réputation. Les outils d'analyse d'opinion intégrés résultent d'une collaboration avec plusieurs partenaires industriels et académiques dans le cadre du projet ANR ImagiWeb³. Ce projet a permis l'élaboration d'une approche générale incluant la construction d'un modèle de connaissance pour la représentation de l'image, le développement d'algorithmes d'apprentissage automatique pour l'extraction des différents aspects de l'opinion et enfin l'application d'algorithmes de clustering permettant de regrouper des internautes ayant des opinions similaires pour en déduire des profils d'opinion.

¹ AMI Enterprise Intelligence : <http://www.amisw.com/fr/wp-content/uploads/2015/01/SPD-AMI-EI7.01.pdf>

² AMI Software : <http://www.amisw.com>

³ Projet ANR (Agence nationale de la recherche) : <http://mediamining.univ-lyon2.fr/velcin/imagiweb/>

2 Analyse d'image de marque avec AMIEI

Une plateforme AMIEI consiste en une suite de modules permettant la mise en œuvre des quatre principales phases d'un processus de veille, à savoir : l'acquisition de l'information par une collecte automatique des données ; la capitalisation et le traitement des données collectées et leur organisation dans une base de données dédiée (*mémoire d'entreprise*); l'analyse des données pour en extraire l'information utile; enfin, le partage et la diffusion de l'information auprès des collaborateurs. L'utilisation de la plateforme se fait via une application Web. Les outils d'analyse d'image issus du projet ImagiWeb sont intégrés sous forme d'un composant d'analyse (*application ImagiWeb*) applicable à des données collectées (*corpus*) au sein de la plateforme (voir FIGURE 1).

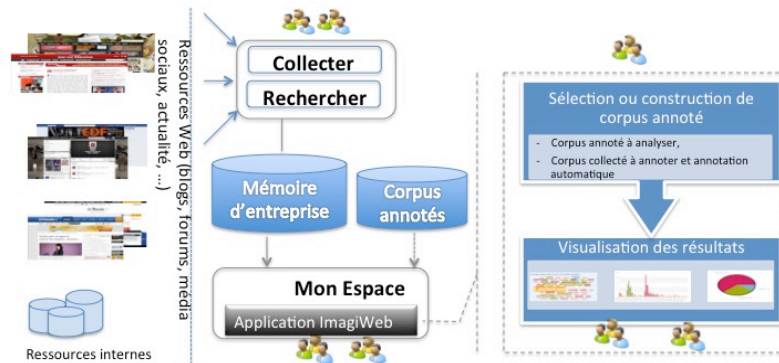


FIGURE 1 : Architecture générale de l'analyse ImagiWeb et son intégration dans la plateforme AMIEI

L'application ImagiWeb permet l'annotation automatique d'un corpus de données. Ceci consiste à identifier des opinions et à en extraire les différentes caractéristiques conformément au modèle de connaissance représentant l'image de l'entité. Pour les cas deux d'études traités, chaque opinion est décrite principalement par une polarité (une échelle de 5 valeurs entre *très négatif* et *très positif*) et une cible (aspect principal sur lequel porte l'opinion, tels que la *personne* ou le *positionnement* pour le cas d'hommes politiques). Une fois le corpus annoté, l'application permet à l'utilisateur final, à l'aide d'un ensemble d'outils d'affichage et de visualisation adaptés, d'explorer les annotations, générer des statistiques selon les différents critères disponibles et visualiser des profils d'opinion produits par les algorithmes de clustering ainsi que leur dynamique dans le temps (voir FIGURE 2).

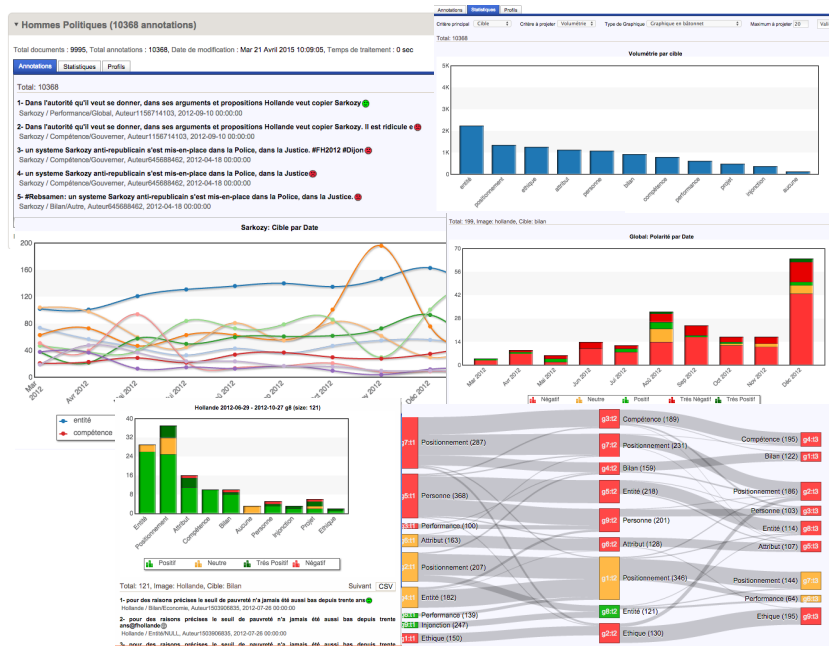


FIGURE 2 : Exemples d'outils de navigation et de visualisation proposés pour l'analyse de l'image dans l'application ImagiWeb. Les données portent sur l'image d'hommes politiques sur Twitter.

Building a Bilingual Vietnamese-French Named Entity Annotated Corpus through Cross-Linguistic Projection

Ngoc Tan Le¹, Fatiha Sadat¹

(1) Département Informatique, Université du Québec à Montréal,
201 avenue Président Kennedy, H2X 3Y7 Montréal, Québec, Canada
le.ngoc_tan@courrier.uqam.ca, sadat.fatiha@uqam.ca

Résumé. La création de ressources linguistiques de bonne qualité annotées en entités nommées est très coûteuse en temps et en main d'œuvre. La plupart des corpus standards sont disponibles pour l'anglais mais pas pour les langues peu dotées, comme le vietnamien. Pour les langues asiatiques, cette tâche reste très difficile. Le présent article concerne la création automatique de corpus annotés en entités nommées pour le vietnamien-français, une paire de langues peu dotée. L'application d'une méthode basée sur la projection cross-lingue en utilisant des corpus parallèles. Les évaluations ont montré une bonne performance (F-score de 94.90%) lors de la reconnaissance des paires d'entités nommées dans les corpus parallèles et ainsi la construction d'un corpus bilingue annoté en entités nommées.

Abstract. The creation of high-quality named entity annotated resources is time-consuming and an expensive process. Most of the gold standard corpora are available for English but not for less-resourced languages such as Vietnamese. In Asian languages, this task is remained problematic. This paper focuses on an automatic construction of named entity annotated corpora for Vietnamese-French, a less-resourced pair of languages. We incrementally apply different cross-projection methods using parallel corpora, such as perfect string matching and edit distance similarity. Evaluations on Vietnamese –French pair of languages show a good accuracy (F-score of 94.90%) when identifying named entities pairs and building a named entity annotated parallel corpus.

Mots-clés : Entité nommée, corpus parallèle, projection cross-lingue.

Keywords: Named entity, parallel corpus, cross-projection.

This demonstration concerns a Named Entity Recognizer (NER) tool for Vietnamese-French, a less-resourced pair of languages using bilingual parallel corpora. Our approach is based on the assumption that a word A (or a phrase A') in the source language L_1 is often translated to a word B (or a phrase B') in the target language L_2 . Thus, we can expect that their translations also co-occur more often in the target language (Fung, 2000). Based on this assumption, different steps were proposed for an automatic extraction of bilingual NER pairs in a parallel corpus and the construction of a NER tool for Vietnamese-French, a less-resourced pair of languages.

Our proposed strategy of best matching criterion for the extraction of bilingual NER pairs from parallel corpora relies on the following steps. First, for each word w_i in the source and target corpora, we build context vectors by considering a window size in the two corpora, the content words, the label and the co-occurrence of all words. The context vectors of the target words are translated using a bilingual dictionary and some gazetteers. Finally, a similarity for each source term S and for each target term T is computed on the basis of the Levenshtein distance with S_1 and T_1 are a word or a phrase respectively in source term S and in target term T:

$$similarity(S_1, T_1) = 1 - \frac{edit_distance(S_1, T_1)}{maxlength(|S_1|, |T_1|)} \quad (1)$$

As a preprocessing step, we tag and lemmatize the text in both languages. The bilingual French-Vietnamese corpora contain 20,000 pairs of sentences. This step allows us to focus on content words only (nouns, verbs, adjectives and adverbs) and thus reduces the noise in our model. Content words are the primary focus for thesaurus enrichment. Word segmentation for Vietnamese is completed using the VCL_WS tool of the VCL group (Vu et al., 2011). Moreover, the Vietnamese corpus is annotated using the VCL_POS tagger, which relies on the maximum entropy approach (Nguyen et al., 2011). The TagEN tool (Tagueur Entités Nommées – Named Entity Tagger) for French is a tool for recognizing

named entities developed by Jean-François Berroyer and Thierry Poibeau at Laboratoire d'Informatique de Paris-Nord (LIPN) (Poibeau, 2003).

In this demonstration, our interest concerns the automatic extraction of NE pairs using a cross-linguistic method and thus the construction of bilingual NE lexicons of proper names for location, organization and person. We evaluate the quality of the bilingual NE pairs for French-Vietnamese bilingual sentence pairs with the help of linguistic experts.

Our evaluation uses a test file of 1,060 pairs of French-Vietnamese sentences pairs and shows a good accuracy (F-score of 94.90%). The bilingual extracted NE pairs can be used to enrich a bilingual dictionary and gazetteers. Table 1, Figures 1 and Figure 2 show the results using the precision, recall and F-score values of the developed NER tool.

The developed NER tool will be used in a complete information extraction system, which can be used in many NLP and machine learning applications such as text classification, machine translation, information retrieval, summarization, etc.

	#Correct-Found	#Noise	#To find	Precision	Recall	F-score
Location	237	4	246	98.34%	96.34%	97.33%
Organization	17	11	55	60.71%	30.91%	40.96%
Person	86	4	96	95.56%	89.58%	92.47%
Average				96.95%	92.96%	94.90%

TABLE 1 : Results of the experimentation with a test set of 1,060 bilingual French-Vietnamese sentences pairs

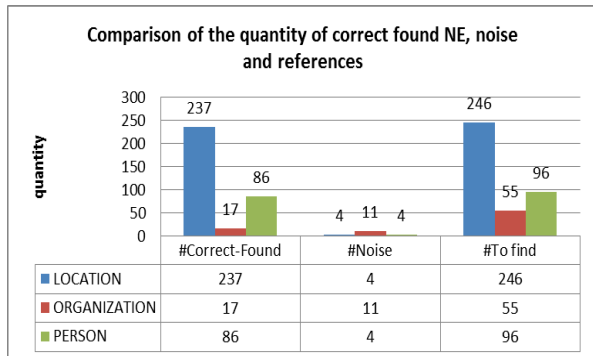


FIGURE 1: Comparison of the quantity of correct found NE, noise and references

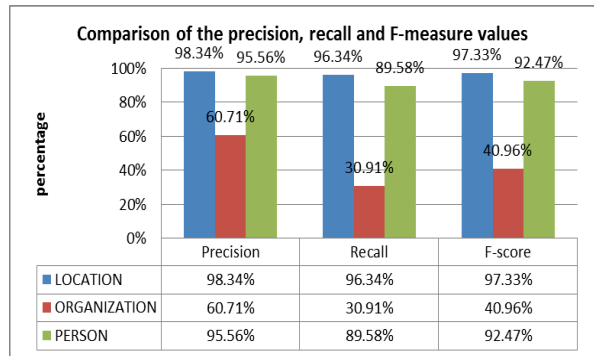


FIGURE 2: Comparison of the precision, recall and F-measure

References

- CHINCHOR, N. (1998). MUC-7 named entity task definition. *In Proceedings of the 7th Message Understanding Conference*.
- FUNG, P. (2000). A statistical view of bilingual lexicon extraction: From parallel corpora to non-parallel corpora. *In Jean Véronis, editor, parallel text processing*.
- VU DINH HONG (2011). Phân đoạn từ tiếng việt ngữ dụng. *Master Thesis of University of Sciences, National University of Ho Chi Minh city*.
- NGUYEN KHUONG AN, DINH DIEN (2011). Tích hợp thông tin từ loại vào hệ dịch máy thống kê. *National Conference, Cần Thơ*.
- THIERRY POIBEAU (2003). The multilingua named entity recognition framework. *Association for Computational Linguistics*, 155-158.

Recherche de motifs de graphe en ligne

Bruno Guillaume
LORIA, Inria Nancy Grand-Est *
bruno.guillaume@loria.fr

Résumé. Nous présentons un outil en ligne de recherche de graphes dans des corpus annotés en syntaxe.

Abstract.

Online Graph Matching

We present an online tool for graph pattern matching in syntactically annotated corpora.

Mots-clés : Syntaxe de dépendances, Corpus, Graphes.

Keywords: Dependency Syntax, Corpus, Graph matching.

Contexte

Les annotations linguistiques, par exemple en syntaxe sont souvent représentées par des arbres, soit en constituants, soit en dépendances. Le fait de se retenir aux arbres a des avantages pratiques notamment pour calculer ces structures. Cependant, du point de vue linguistique, les arbres ne sont souvent pas suffisants lorsque l'on veut enrichir les structures. Le corpus DEEP-SEQUOIA (Candito *et al.*, 2014), par exemple, propose une annotation en dépendances profondes de phrases en français. Dans ce corpus, aucune hypothèse n'est faite sur les structures employées et il y a donc de très nombreux cas d'annotations qui ne se représentent pas comme des arbres : par exemple certaines unités lexicales ont plusieurs gouverneurs (jusqu'à 7 dans la version 1.1 du corpus) et il existe de nombreux cycles.

C'est pour ces raisons que nous avons proposé d'utiliser la réécriture de graphes comme cadre formel pour décrire des processus de transformations de structures syntaxiques. Le logiciel GREW (Guillaume *et al.*, 2012) implémente ce modèle de calcul et permet de faire ce type de transformation. Pour déclencher l'application d'une règle, GREW utilise une recherche de motifs de graphes (pattern matching). C'est cette fonctionnalité de GREW qui est exploitée dans la version en ligne GREW-WEB¹. Dans cet outil, on écrit un motif de graphe (généralement un petit graphe) et on peut visualiser les occurrences correspondantes dans un corpus donné. GREW-WEB est disponible avec quelques corpus libres de droits : SEQUOIA (Candito & Seddah, 2012), DEEP-SEQUOIA (Candito & Seddah, 2012) en français, UNIVERSAL DEPENDENCY TREEBANK (McDonald *et al.*, 2013) en français et en coréen et TIGER (Brants *et al.*, 2004) en allemand.

Exemples de recherche

Recherche d'une sous-catégorisation On recherche, dans SEQUOIA, un verbe avec à la fois un argument `a_obj` et un argument `de_obj`. Le résultat obtenu (6 occurrences) est représenté dans la Figure 1.

```
1 match { V [cat=V]; A []; DE []; % les 3 nœuds recherchés
2       V -[a_obj]-> A; V -[de_obj]-> DE; } % les relations entre les nœuds
```

Utilisation des contraintes négatives On peut filtrer les résultats obtenus en ajoutant des contraintes négatives. Ici, on recherche, toujours dans SEQUOIA, les occurrences de *prendre* avec un objet nominal sans déterminant (11 occurrences).

```
1 match { V[lemma="prendre"]; OBJ[cat=N]; V -[obj]-> OBJ } % "prendre" + OBJ nominal
2 without { D[]; N -[det]-> D } % sans det pour l'OBJ
```

*. Ce travail a bénéficié du soutien du projet Ortolang (ortolang.fr). L'auteur remercie Antoine Chemardin pour son aide dans le développement de l'interface Web.

1. <http://grew.loria.fr/demo>

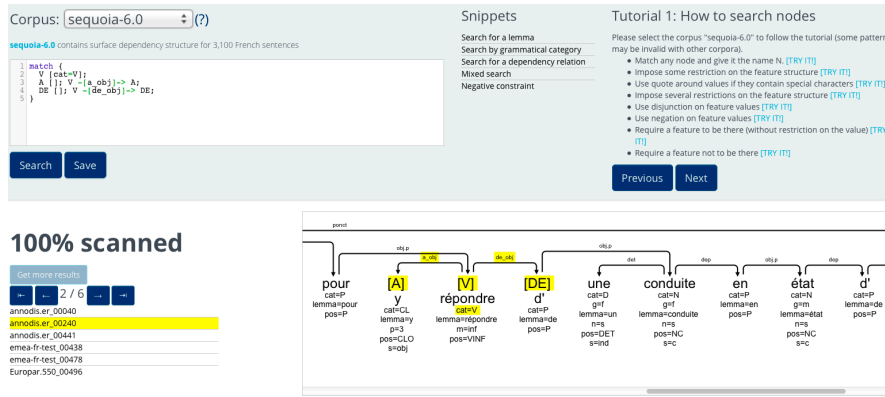


FIGURE 1 – Capture d’écran de l’interface

Recherche d’erreurs dans un corpus GREW-WEB permet de rechercher systématiquement des motifs qui sont susceptibles d’être des erreurs. Par exemple, dans SEQUOIA, on peut vérifier l’accord sujet-verbe. On trouve 23 occurrences du motif suivant, ce qui permet de repérer une dizaine d’erreurs d’annotation.

```

1 match { S [n=*]; V [cat=V, n=*]; V -[su] -> S } % le motif sujet-verbe
2 without { S.n = V.n } % les traits "n" différents
3 without { V[m=part, t=past]; A[lemma=avoir]; V -[aux.tps] -> A } % on élimine l'aux avoir
4 without { S[n=s]; V[n=p]; S -[coord] -> * } % pas de coord. comme sujet
5 without { S[cat=N, lemma="minorité"|"dizaine"|. . . ] } % exceptions lexicales

```

Recherche de graphes Dans DEEP-SEQUOIA, les structures sont des graphes et on peut donc rechercher des motifs qui sont eux-aussi des graphes. Ci-dessous, on recherche les cycles de longueur 8, on en trouve 2 occurrences dans le corpus.

```

1 match { N1[]; N2[]; N3[]; N4[]; N5[]; N6[]; N7[]; N8[];
2 N1 -> N2; N2 -> N3; N3 -> N4; N4 -> N5; N5 -> N6; N6 -> N7; N7 -> N8; N8 -> N1 }

```

Conclusion

L’outil en ligne GREW-WEB permet de trouver rapidement des exemples en corpus de constructions particulières ou de rechercher de façon systématique des erreurs d’annotation. En fait, rien ne restreint l’usage de GREW-WEB à des structures en dépendances, il peut être utilisé sur tout type de graphes comme des analyses en constituants, des graphes de représentation sémantique par exemple.

Références

- BRANTS S., STEFANIE D., EISENBERG P., HANSEN S., KÖNIG E., LEZIUS W., ROHRER C., SMITH G. & USZKO-REIT H. (2004). TIGER : Linguistic Interpretation of a German Corpus. *J. of Language and Computation*, **2**, 597–620.
- CANDITO M., PERRIER G., GUILLAUME B., RIBEYRE C., FORT K., SEDDAH D. & VILLEMONT DE LA CLERGE-RIE É. (2014). Deep Syntax Annotation of the Sequoia French Treebank. In *LREC*, Reykjavik, Iceland.
- CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l’adaptation d’analyseur par pont lexical. In *Proc. of TALN*, Grenoble, France.
- GUILLAUME B., BONFANTE G., MASSON P., MOREY M. & PERRIER G. (2012). Grew : un outil de réécriture de graphes pour le TAL. In *12ième conférence TALN*, Grenoble, France : ATALA.
- MCDONALD R., NIVRE J., QUIRMBACH-BRUNDAGE Y., GOLDBERG Y., DAS D., GANCHEV K., HALL K., PETROV S., ZHANG H., TACKSTROM O., BEDINI C., CASTELLO N. B. & LEE J. (2013). Universal dependency annotation for multilingual parsing. In *Proc. of ACL 2013*.

Un patient virtuel dialogant

Leonardo Campillos Dhouha Bouamor Éric Bilinski
Anne-Laure Ligozat Pierre Zweigenbaum Sophie Rosset
LIMSI - CNRS, Orsay
prenom.nom@limsi.fr

Résumé. Le démonstrateur que nous décrivons ici est un prototype de système de dialogue dont l'objectif est de simuler un patient. Nous décrivons son fonctionnement général en insistant sur les aspects concernant la langue et surtout le rapport entre langue médicale de spécialité et langue générale.

Abstract.

An Interactive Virtual Patient

This paper describes the work-in-progress prototype of a dialog system that simulates a virtual patient consultation. We describe the general architecture and specifically the mapping between technical and lay terms in the medical domain.

Mots-clés : Patient virtuel, système de dialogue, langage spécialisé, langage grand public.

Keywords: Virtual patient, dialog system, specialised language, lay language.

1 Introduction

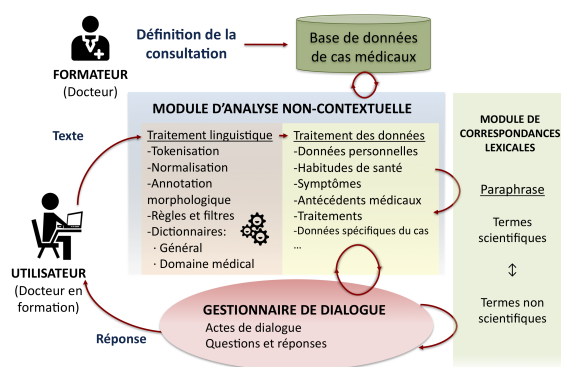
L'objectif du projet PatientGenesys est de développer un outil de création de cas cliniques numériques de simulation médicale pour la formation des personnels de santé. Dans ce cadre, nous développons un agent conversationnel dont l'objectif est de simuler un patient qui, pendant une consultation médicale, répond à un professionnel de santé.

Dans le domaine de la santé, les patients virtuels sont utilisés notamment pour l'aide au développement de certaines compétences comme la communication (Deladisma *et al.*, 2007), l'interaction avec des patients atteints de pathologies mentales (Hubal *et al.*, 2003; Kenny & Parsons, 2011) ou encore la formation d'étudiants en pharmacologie (Park & Summons, 2013). La plupart de ces systèmes sont développés pour l'anglais. Celui du projet PatientGenesys concerne, dans un premier temps, le français. Néanmoins, la plupart des défis que présente un tel projet sont indépendants de la langue. La première difficulté est liée à l'absence de corpus spécifique qui nous interdit, dans un premier temps, d'utiliser des approches fondées sur de l'apprentissage statistique. Une autre difficulté tient à la variabilité des termes et en particulier à la différence entre termes techniques et termes grand public. Un autre aspect, réservé à des travaux futurs, est d'aller vers la programmation automatique de systèmes spécialisés dans un domaine à partir des spécifications du cas fournies par le formateur, le contenu de ces spécifications n'étant pas connu à l'avance.

2 Architecture du système de dialogue

La figure 1 illustre le système dans son ensemble. L'initiative de l'interaction revient à l'utilisateur (le médecin en formation). L'entrée est textuelle, la sortie est audio et textuelle. Quatre modules composent ce système :

- La base des cas patients, décrits par le formateur, comprend une formalisation du dossier médical. Cette formalisation s'appuie sur des schémas cognitifs comme (Evans *et al.*, 1986; Patel *et al.*, 1989).
- L'analyse non-contextuelle implique une analyse linguistique générale et un traitement des connaissances spécifiques au domaine. L'analyse générale prend en charge la segmentation en mots, l'analyse en parties du discours, la correction orthographique et l'identification des variantes morphologiques de certaines catégories de mots. Le traitement dépendant du domaine prend en charge la détection des entités générales et médicales, la classification



Salutations

> Bonjour.

— Bonjour, docteur.

Questions sur le but de la consultation

> Pouvez-vous me dire pourquoi vous consultez ?

— Alors euh, je viens pour douleur thoracique, toux et fièvre.

Introduction de la consultation

> D'accord. Avant tout je vais vous poser quelques questions.

— OK.

Questions sur les données générales du patient

> Quels métiers avez-vous effectués ?

— Alors, je suis retraité.

> Est-ce que vous fumez ?

— Et bien, oui, euh... je fume

> Combien ?

— Je fume 20 cigarette par jour.

Questions sur les antécédents

> D'accord, quels problèmes de santé avez-vous déjà eu ?

— J'ai une hypertension artérielle.

> Vous êtes en traitement pour l'hypertension ?

— Euh... je prends Co-renaltec oral à dose de 20 mg, 1 par jour.

FIGURE 1 – Présentation générale du système, et exemple de dialogue entre un médecin et le système actuel (en italiques). Les annotations en bleu ajoutées à la main montrent le schéma suivi.

des questions, la détection du focus et la détection des actes de dialogue.

- La détection des correspondances lexicales intervient dans deux situations : (i) à partir d'une représentation hiérarchique des concepts et des termes associés, elle permet de trouver une correspondance entre une question portant sur une catégorie d'affections et une affection précise du patient ; (ii) à partir d'un lexique liant termes spécialisés et termes grand public, elle permet de privilégier les termes grand public dans la réponse du patient.
- Le gestionnaire de dialogue s'appuie sur des schémas dépendant pour partie de la spécialité médicale concernée par le cas patient. Il prend en charge tout ce qui concerne la gestion du flux dialogique.

3 Conclusion

Nous avons présenté l'architecture générale d'un système de dialogue dont l'objectif est de simuler un patient dans le cas de différentes spécialités médicales. Si le système est dans son ensemble classique, il présente certaines spécificités dont la plus importante concerne les correspondances entre termes techniques (présents dans le dossier médical) et termes grand public (que doit utiliser le patient), le médecin étant susceptible d'utiliser l'un ou l'autre.

Remerciements : Ce travail a été réalisé dans le cadre du projet FUI PatientGenesys, contrat d'aide N° F1310002 P.

Références

- DELADISMA A. M., COHEN M., STEVENS A., WAGNER P., LOK B., BERNARD T., OXENDINE C., SCHUMACHER L., JOHNSEN K., DICKERSON R. *et al.* (2007). Do medical students respond empathetically to a virtual patient ? *The American Journal of Surgery*, **193**(6), 756–760.
- EVANS D. A., BLOCK M. R., STEINBERG E. R. & PENROSE A. M. (1986). Frames and heuristics in doctor-patient discourse. *Social science & medicine*, **22**(10), 1027–1034.
- HUBAL R. C., FRANK G. A. & GUINN C. I. (2003). Lessons learned in modeling schizophrenic and depressed responsive virtual humans for training. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, p. 85–92, New York, NY, USA : ACM.
- KENNY P. & PARSONS T. (2011). Embodied conversational virtual patients. *Conversational Agents and Natural Language Interaction : Techniques and Effective Practices*. Information Science Reference, p. 254–281.
- PARK M. & SUMMONS P. (2013). A computer-generated digital patient for oral interview training in pharmacy. *Advanced Science and Technology Letters*, p. 28 :126–131.
- PATEL V. L., EVANS D. A. & KAUFMAN D. R. (1989). Cognitive framework for doctor-patient interaction. *Cognitive science in medicine : Biomedical modeling*, p. 253–308.

Intégration du corpus des actes de TALN à la plateforme ScienQuest

Achille Falaise¹

(1) ICAR, ENS de Lyon, 15 parvis René Descartes, 69342 LYON cedex 7
achille.falaise@ens-lyon.fr

Résumé. Cette démonstration présente l'intégration du corpus arboré des Actes de TALN à la plateforme ScienQuest. Cette plateforme fut initialement créée pour l'étude du corpus de textes scientifiques Scientext. Cette intégration tient compte des méta-données propres au corpus TALN, et a été effectuée en s'efforçant de rapprocher les jeux d'étiquettes de ces deux corpus, et en convertissant pour le corpus TALN les requêtes prédéfinies conçues pour le corpus Scientext, de manière à permettre d'effectuer facilement des recherches similaires sur les deux corpus.

Abstract.

Integration of the TALN proceedings treebank to the ScienQuest platform

This demonstration shows the integration of the TALN proceedings Treebank to the ScienQuest platform. This platform was initially created for the study of the Scientext scientific texts corpus. This integration takes into account the metadata to the TALN corpus, and was done in an effort to reconcile these two corpora's sets of labels, and to convert for the TALN corpus the predefined queries designed for the Scientext corpus, in order to easily perform similar queries on the two corpora.

Mots-clés : corpus, corpus arborés, environnement d'étude de corpus.

Keywords: corpora, treebanks, corpus study environment.

1 Introduction

Les corpus de textes disciplinaires permettent d'étudier, en diachronie, l'historique d'une discipline et les évolutions de sa phraséologique, et d'un point de vue synchronique, de comparer les différents types de communications au sein de la discipline, mais aussi, pour peu que l'on dispose des corpus correspondants, par rapport à d'autres disciplines. Le corpus des actes de TALN offre ainsi l'opportunité à la communauté d'effectuer un peu d'introspection. Cette démonstration présente l'intégration de ce corpus dans la plateforme ScienQuest¹, qui intègre déjà un corpus pluridisciplinaire de textes scientifiques², et permet d'effectuer simplement des recherches et comparaisons sur des corpus arborés.

2 Le corpus TALN

Le corpus « TALN Archives »³ a été collecté par Florian Bourdin (Bourdin, 2013) à partir des différents sites Web des conférences TALN et RÉCITAL (1997-2014). Il s'agit d'un corpus de textes au format *pdf*, accompagnés de méta-données (notice *bibtex* et résumé).

Un sous-ensemble de 586 articles a été sélectionné et traité par Ludovic Tanguy (Tanguy, 2013), afin d'en extraire le texte intégral, et de l'analyser avec TALISMANE (Urieli & Tanguy, 2013). Le corpus arboré ainsi obtenu contient 2,3 millions de tokens, annotés en parties du discours, en lemmes et en dépendances syntaxiques.

¹ <http://corpora.aiakide.net/link/taln>

² <http://corpora.aiakide.net/link/sc texts-fr>

³ <https://github.com/boudinfl/taln-archives>

3 Intégration à la plateforme ScienQuest

La plateforme ScienQuest (Falaïse *et al.*, 2011) fut initialement créée pour l'étude linguistique du positionnement et du raisonnement dans le corpus de textes scientifiques Scientext (Tutin *et al.*, 2009), analysé avec Syntex. Cette plateforme, qui se veut simple à utiliser pour des non-TAListes, permet de rechercher en ligne des concordances dans un corpus, en fonction de critères linguistiques.

Méta-données. L'interface de ScienQuest permet de créer facilement des sous-corpus en fonction des méta-données du corpus. En outre, lors d'une recherche de concordances, des statistiques concernant la répartition des occurrences en fonction de ces méta-données sont calculées. Certaines des méta-données présentes dans le corpus TALN ont ainsi été intégrées dans la plateforme : conférence (TALN ou RECITAL), année et type d'article (court, long, etc.). Cela permet ainsi par exemple de sélectionner un sous-corpus en fonction de la conférence, ou de distinguer la fréquence relative d'un token ou d'un motif en fonction de l'année.

Étiquettes. Le jeu d'étiquettes utilisé par TALISMANE est assez riche (28 étiquettes morphosyntaxiques et 23 relations syntaxiques). Le mode de recherche privilégié dans ScienQuest (« mode libre ») utilise un assistant, qui présente ces étiquettes de manière conviviale. En général, le jeu d'étiquettes qui y est présenté est plus simple que le jeu d'étiquettes du corpus, et regroupe souvent plusieurs étiquettes sous un même nom de manière transparente pour l'utilisateur, qui ne voit que la « méta-étiquette » ; le jeu d'étiquettes complet n'est disponible dans sa totalité que dans le « mode avancé ». Pour le corpus TALN, afin de permettre une certaine compatibilité avec le corpus de textes scientifiques Scientext, nous nous sommes alignés, dans la mesure du possible, sur les 9 étiquettes morphosyntaxiques et les 13 relations syntaxiques présentées dans l'assistant « mode libre » pour ce corpus.

Grammaires. Enfin, le corpus Scientext est accompagné de « grammaires », c'est à dire de requêtes complexes pré-enregistrées, visant l'étude de la phraséologie du raisonnement et du positionnement dans les textes scientifiques. Ces grammaires, développées pour un corpus analysé avec Syntex, ont été « traduites » pour correspondre à l'analyse TALISMANE du corpus TALN.

4 Conclusion

L'intégration du corpus TALN dans ScienQuest permet ainsi d'effectuer facilement des recherches sur ce corpus. Cela peut aller de requêtes très simples, comme l'observation de l'évolution du nombre d'occurrences de « corpus » ou « grammaire » au fil des années, à des requêtes plus complexes, par exemple en utilisant les grammaires pré-enregistrées. On peut ainsi remarquer que la phraséologie semble assez homogène entre TALN et RECITAL, mais varie nettement en fonction du type d'article.

Références

- BOUDIN F. (2013). TALN Archives : une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue. *Actes de TALN 2013*, Les Sables d'Olonne, pages 507-514.
- FALAISE A., TUTIN A., KRAIF O. (2011). Exploitation d'un corpus arboré pour non spécialistes par des requêtes guidées et des requêtes sémantiques. *Actes de TALN 2011*, Montpellier, pages 187-215.
- TANGUY L. (2013). Corpus TALN, en ligne : <http://redac.univ-tlse2.fr/corpus/taln.html> (consulté le 8 mai 2015).
- TUTIN A., GOSSMANN F., FALAISE A., KRAIF O. (2009). Autour du projet Scientext : étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques. *Journées Linguistique de Corpus*, Lorient.
- URIELI A. ET TANGUY L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. *Actes de TALN 2013*, Les Sables d'Olonne, pages 188-201.

Une aide à la communication par pictogrammes avec prédiction sémantique

Aurélie Merlo^{1,2}

(1) Savoirs, Textes, Langage (STL), rue du Barreau, 59653 Villeneuve d'Ascq Cedex

(2) Ergonotics SAS, 165 avenue de Bretagne, 59000 Lille, France

aurelie@ergonotics.com

Résumé. Cette démonstration présente une application mobile (pour tablette et smartphone) pour des personnes souffrant de troubles du langage et/ou de la parole permettant de générer des phrases à partir de la combinaison de pictogrammes puis de verbaliser le texte généré en Text-To-Speech (TTS). La principale critique adressée par les patients utilisant les solutions existantes est le temps de composition trop long d'une phrase. Cette limite ne permet pas ou très difficilement d'utiliser les solutions actuelles en condition dialogique. Pour pallier cela, nous avons développé un moteur de génération de texte avec prédiction sémantique ne proposant à l'utilisateur que les pictogrammes pertinents au regard de la saisie en cours (e.g. après le pictogramme [manger], l'application propose les pictogrammes [pomme] ou encore [viande] correspondant à des concepts comestibles). Nous avons ainsi multiplié de 5 à 10 la vitesse de composition d'une phrase par rapport aux solutions existantes.

Abstract.

An Augmentative and Alternative Communication with semantic prediction.

This demo shows a mobile app (for smartphone and tablet) for people with language and/or speech disorders for generating sentences from the combination of icons and verbalize text generated by Text-To-Speech (TTS). The main criticism expressed by patients using existing solutions is the time too long required to compose a sentence. Hence, existing solutions are hard to use in dialogic conditions. To alleviate this problem, we have developed a text generation engine with semantic prediction. This engine only proposes to use the relevant pictograms in view of the current entry (e.g. after the pictogram [eat], the application offers pictograms [apple] or [meat] corresponding to edible concepts). We have multiplied from 5 to 10 the speed of composing a sentence compared to existing solutions.

Mots-clés : génération de texte, prédiction sémantique, frame semantics, handicap, aide à la communication

Keywords: natural language generation, semantic prediction, frame semantics, disability, AAC

1. Introduction

Selon l'American Speech-Language Hearing Association, une personne sur 100 dans le monde est concernée par des troubles du langage et/ou de la parole. Ces troubles peuvent faire suite à des causes diverses telles que des maladies neuro-dégénératives (e.g. Sclérose Latérale Amyotrophique), des accidents (e.g. accidents cardio-vasculaires) ou encore des opérations chirurgicales lourdes de la sphère ORL (e.g. laryngectomie, trachéotomie, glossectomie).

La communication parlée étant l'interface majeure des interactions humaines, les troubles du langage et/ou de la parole entraînent un isolement social et une perte de l'autonomie des personnes touchées. Des solutions techniques existent pour pallier ces troubles ; ce sont des aides à la communication ou Augmentative and Alternative Communication (AAC) selon la terminologie anglo-saxonne. Ces aides permettent aux patients, à l'aide de saisies textuelles ou pictogrammiques, de composer des phrases et de les verbaliser en sortie à l'aide d'un système de Text-To-Speech (TTS). Néanmoins, les solutions existantes de suppléance ne sont actuellement pas considérées comme satisfaisantes par les patients eux-mêmes. La principale critique émise est le temps de composition d'une phrase beaucoup trop long pour une utilisation dans un contexte dialogique naturel. En effet, les solutions existantes ne permettent de générer que 5 à 10 mots par minute contrairement à la vitesse de la parole non-altérée estimée à environ 200 mots par minute. Seuls alors la famille, les amis et le corps médical sont assez patients pour interagir avec le patient.

Cette démonstration présente une aide à la communication par pictogrammes pour support mobile (smartphone et

tablette). Nous avons augmenté sensiblement la vitesse de composition d'une phrase à l'aide de pictogrammes (de 5 à 10 fois par rapport aux solutions existantes) grâce à la mise en place de solutions innovantes telles qu'un moteur de prédiction sémantique. L'objectif principal de ce moteur est de ne prédire uniquement que les pictogrammes pertinents au regard de la saisie de l'utilisateur pour générer en sortie des phrases pertinentes du point de vue du sens et de la syntaxe (ex : après le pictogramme [manger]¹, le moteur prédit uniquement des pictogrammes comme [pomme], [viande] ou encore [chocolat] représentant des concepts comestibles).

Nous proposons ici de décrire la prédiction sémantique en place dans l'application. Nous avons développé une ontologie d'affichage des pictogrammes et un module de gestion de la surface pour rendre les phrases grammaticalement et orthographiquement correctes selon la langue. Ces éléments seront présentés en démonstration.

2. Présentation générale de l'application

L'application est une aide à la communication par pictogrammes à destination de personnes ayant des troubles du langage et/ou de la parole avec des troubles cognitifs et gestuels légers. Contrairement aux solutions existantes dont les pictogrammes représentent des mots-formes, elle repose sur le principe qu'un pictogramme représente un concept, soit une information sur le monde qui peut être définie par l'association avec d'autres informations. Ainsi, le concept AVOCAT² représenté par un fruit se définit par ses propriétés comme /comestible/, /avec un noyau/, /peut être vert/ ou encore /avec une peau/. Ces propriétés distinguent ce concept du concept AVOCAT représenté par un humain et possédant des propriétés différentes. Le concept AVOCAT fruit peut être mis en relation avec d'autres concepts selon ses propriétés comme les concepts VERT, MANGER ou encore ÉPLUCHER. Lorsque l'utilisateur sélectionne un pictogramme, il sélectionne un concept et génère une forme fléchie. Par exemple, la sélection du concept MANGER, génère la forme fléchie *je mange*³. Il peut à tout moment modifier cette forme fléchie ou cette forme peut être modifiée automatiquement selon des règles de surface (e.g. accords, conjugaison).

Les pictogrammes sont présentés sur une interface façon mind map (cf. Figure 1 ci-dessous). L'interface est composée de 6 sections : les actions, les circonstanciels (temps, lieux, instruments...), les objets concrets, les objets abstraits, les êtres vivants et les favoris. Les sections des adjectifs et des adverbes de manière apparaissent uniquement à la sélection d'un nom ou d'un verbe. Chaque section contient des catégories qui peuvent contenir elles-mêmes des catégories ou des concepts. Le concept sélectionné par l'utilisateur (e.g. MANGER sur l'interface ci-dessous) est placé au centre autour duquel s'organise la prédiction sémantique des concepts pertinents (e.g. des concepts comestibles, des lieux, des temps).



FIGURE 1 : Interface de l'application

3. Prédiction sémantique et Frame Semantics

La prédiction sémantique des concepts permet de ne proposer que les concepts pertinents au regard de la saisie en cours de l'utilisateur. Pour cela, nous avons créé une grammaire par frames dans la lignée de la Frame Semantics (Fillmore

¹ Par convention, nous mentionnons les pictogrammes entre crochets.

² Par convention, nous mentionnons les concepts en petites majuscules.

³ Le pronom personnel de la première personne du singulier est généré par défaut pour les verbes.

1968, 1982). La Frame Semantics est une théorie initiée par Fillmore à la fin des années 1960. C’est une sémantique empirique mettant en avant la continuité entre langue et expérience du monde. Une frame correspond à une situation prototypique évoquée par des lexèmes (e.g. *manger, dévorer, grignoter*). Chaque frame peut être décrite par une définition et un typage sémantique des participants de la situation ou *frame elements* (cf. Table 1 ci-dessous).

Frame	Type sémantique du sujet	Verbes du frame	Types sémantiques des objets	Exemple de phrases générées
manger	humain	manger dévorer grignoter	aliment instrument pour manger accompagnant lieu temps (durée, date et fréquence)	<i>Je mange une pomme</i> <i>Je mange avec une fourchette</i> <i>Je mange avec Paul</i> <i>Je mange dans la cuisine</i> <i>Je mange à 14h</i> <i>Je mange</i> <i>Je mange une pomme dans la cuisine</i> <i>Je mange avec Paul dans la cuisine à 14h</i>

TABLE 1 : Exemple du frame « manger »

Notre lexique contient 684 verbes correspondant à environ 250 frames. Les 5800 concepts de notre lexique sont annotés en types sémantiques appelés par les frames (e.g. les concepts POMME, AVOCAT, VIANDE, CHOCOLAT correspondent au type « aliment »). Un concept peut avoir plusieurs types sémantiques selon que la frame qui l’appelle est spécifique ou générique. Par exemple, la frame « manger » appelle le type sémantique « instrument pour manger » du concept FOURCHETTE contrairement à la frame « acheter » plus générique qui appelle le type sémantique « artefact » de FOURCHETTE.

4. Niveaux de prédiction sémantique

La prédiction sémantique à base de frames peut se faire à plusieurs niveaux dans la saisie. Nous venons de voir qu’elle peut se faire sur la relation entre le sujet et le verbe, entre le verbe et ses objets mais également sur les propriétés du concept comme la relation entre le nom et ses adjectifs (e.g. le frame « aliment » a pour type sémantique « couleur ») et entre le verbe et ses adverbes (e.g. le frame « manger » a pour frame element « vitesse »). Un système de restrictions sémantiques vient affiner la prédiction des concepts associés aux propriétés (e.g. le type sémantique « couleur » pour AVOCAT est restreint au concept VERT et non à toutes les couleurs). Nous sommes également capable d’identifier la gradabilité de certains concepts et de proposer ainsi à l’utilisateur des adverbes d’intensité quand cela est pertinent (e.g. prédictions des concepts : POMME ROUGE TRÈS, *HOMME MORT TRÈS, ROULER VITE TRÈS). Enfin, nous prenons en considération le caractère massif ou comptable des concepts pour la prédiction des numéraux (e.g. prédictions des concepts : POMME TROIS, *PEUR TROIS).

Références

FILLMORE C. (1968). “The case for case”, In: E. Bach and R.T. Harms (eds) (1968) *Universals in Linguistic Theory*. London: Holt, Rinehart and Winston, 1–88.

FILLMORE C.(1982). “Frame Semantics”, In H. P. CO, Ed., *Linguistics in the morning calm*, 111–137.

Un système expert fondé sur une analyse sémantique pour l'identification de menaces d'ordre biologique

Cédric Lopez¹, Aleksandra Ponomareva¹, Cécile Robin², André Bittar², Paolo Curtoni², Xavier Larrucea³,
Frédérique Segond¹, Marie-Hélène Metzger⁴

(1) Viseo Technologies, 4 avenue Doyen Louis Weil, Grenoble (France)

(2) Holmes Semantic Solutions, 12-14, rue Claude Genin, Grenoble (France)

(3) Tecnia, Parque tecnologico de Bizkaia, Edif. 202, Zamudio (Espagne)

(4) Université Lyon 1, CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne (France)

cedric.lopez@viseo.com, aleksandra.ponomareva@viseo.com, robin@ho2s.com, bittar@ho2s.com,
curtoni@ho2s.com, xavier.larrucea@tecnalia.com, frederique.segond@viseo.com,
marie-helene.metzger@chu-lyon.fr

Résumé. Le projet européen TIER (Integrated strategy for CBRN – Chemical, Biological, Radiological and Nuclear – Threat Identification and Emergency Response) vise à intégrer une stratégie complète et intégrée pour la réponse d'urgence dans un contexte de dangers biologiques, chimiques, radiologiques, nucléaires, ou liés aux explosifs, basée sur l'identification des menaces et d'évaluation des risques. Dans cet article, nous nous focalisons sur les risques biologiques. Nous présentons notre système expert fondé sur une analyse sémantique, permettant l'extraction de données structurées à partir de données non structurées dans le but de raisonner.

Abstract.

An Expert System Based on a Semantic Analysis for Identifying Biological Threats

The European project TIER (Integrated strategy for CBRN – Chemical, Biological, Radiological and Nuclear - Threat Identification and Emergency Response) aims at developing a comprehensive and integrated strategy for emergency response in case of chemical, biological, radiological and nuclear danger, as well as explosives use, based on threat identification and risk assessment. In this article, we focus on the biological risks. We introduce our business rules management system based on a semantic analysis, that enables the extraction of structured data from unstructured data with the aim to make reasoning.

Mots-clés : TIER, SGRM, Système de Gestion de Règles Métier, analyse sémantique.

Keywords: TIER, BRMS, Business Rules Management System, semantic analysis.

1 Introduction

L'un des objectifs du projet européen TIER (*Integrated strategy for CBRN Threat Identification and Emergency Response*) est d'identifier les menaces et les risques chimiques, biologiques, radiologiques et nucléaires. Nous nous focalisons dans un premier temps sur les risques biologiques à des fins de détection de menaces liées au bioterrorisme. Le défi consiste à structurer les informations recueillies depuis différentes sources de données non structurées du Web concernant l'apparition ou l'évolution des pathologies infectieuses relevant des catégories A à C définies par les *Centers for Disease Control and Prevention* (CDC) et de leurs informations relatives (symptômes, nombre de cas détectés, cas mortels, etc.).

2 Approche

L'originalité de notre approche consiste à utiliser un BRMS (*Business Rules Management System*) pour la gestion de règles linguistiques dites « de transition ». Celles-ci s'appuient sur une analyse syntaxique et sémantique pour générer des données structurées en vue de leur intégration dans la base de connaissance. Un BRMS est un outil composé d'un moteur de règles et de l'environnement nécessaire permettant de les manipuler. On peut ainsi définir une base de connaissances et raisonner sur des faits à partir d'un ensemble de règles métier, ici nos règles de transition. L'outil *open source* Drools (Browne, 2009) a été retenu dans le cadre de ce projet. Notre approche est constituée de 5 étapes :

1. **Constitution d'un jeu de données textuelles** (à partir du Web : Direction générale de la Santé, Institut de veille sanitaire, ...) et **définition du modèle de données**. Les médecins impliqués dans TIER ont identifié les libellés des germes correspondant aux catégories A à C du CDC et ont défini les entités d'intérêt pour caractériser l'événement : date de début et date de fin de l'événement, germe impliqué dans l'événement, nombre de personnes infectées, nombre de personnes décédées, lieu (pays ou région) de l'événement, *etc.* Dans la suite, nous extrayons automatiquement des données correspondant à ce modèle.
2. **Reconnaissance des entités d'intérêts**. Celle-ci est fondée sur une approche hybride (symbolique et statistique). Concernant la partie symbolique, nous avons conçu une grammaire composée de lexiques et de règles basées sur l'analyse morphosyntaxique. Par exemple, la présence du lemme « cas » implique souvent la mention d'une quantité (nombre de cas mortels d'un événement par exemple). La partie statistique se concentre sur la localisation des événements biologiques par le biais d'un classifieur de type CRF (*Conditional Random Field statistical modelling method*) (Lefferty et al., 2001).
3. **Analyse sémantique**. La détection des relations sémantiques entre les prédicats et leurs arguments s'opère sur un graphe de dépendances syntaxiques (fourni par HOLMES¹) et se fait par l'application d'un ensemble de grammaires de transformation de graphe. Les grammaires exploitent les informations linguistiques présentes sur les nœuds (tokens) du graphe pour convertir les dépendances syntaxiques en relations sémantiques telles que AGENT, CAUSE, LOCALIZATION, MANNER, MODALITY, NEGATION, *etc.* et des relations temporelles, telles que AFTER, BEFORE, DURING, *etc.*
4. **Développement des règles**. Fondées sur les résultats de la reconnaissance d'entités d'intérêts et sur l'analyse sémantique, les règles permettent de générer des données structurées candidates pour peupler la base de connaissance. Nos règles « de transition », développées avec Java/MVEL dans Drools, ont pour objectif de transformer le résultat de l'analyse linguistique en des objets correspondants au modèle précédemment défini, permettant ainsi de peupler notre base de connaissance. Nos règles s'appuient à la fois sur les entités d'intérêt détectées, et sur des éléments linguistiques d'ordre syntaxique et sémantique. Par exemple, pour la phrase « En conséquence, on considère désormais que 18 cas de fièvre hémorragique à virus Ebola et 6 décès ont été notifiés », on pourra appliquer la règle suivante : « si l'annotation sémantique CASE_NB (nombre de cas) a une relation de préposition avec le nom du germe, alors le nom du germe et le nombre de cas sont intégrés dans le même fait candidat.
5. **Sélection des candidats pertinents et peuplement de la base de connaissance**. À chaque règle est associé un score de saillance : plus la règle est précise (i.e. contraignante), plus le score est élevé. Une forte précision des faits extraits est assurée en imposant la présence de certaines relations sémantiques. Au contraire, des contraintes faibles permettent de gagner en rappel au détriment de la précision. La sélection des faits est réalisée via le calcul d'un score qui permet de classer les faits par ordre de pertinence. Ce score dépend de la saillance des règles appliquées. Le seuil S a pour objectif de distinguer les faits pertinents des faits non pertinents.

L'évaluation a été effectuée sur 166 faits extraits automatiquement dans 50 textes jusque-là non exploités. Nous avons annoté manuellement chaque fait selon deux classes : « pertinent » et « non pertinent ». Un fait est considéré « pertinent » lorsque les informations sont cohérentes avec l'information véhiculée dans le texte (dans les autres cas, le fait est jugé « non pertinent »). Nous avons utilisé les mesures de micro-moyenne de précision, de rappel et de F-score en variant le seuil S du score de saillance. Les résultats indiquent que le meilleur F-score (0,73) est atteint pour $S=0,65$, avec une précision (0,74) et un rappel (0,73) équivalents.

Remerciements

Avec le soutien financier du programme Prévenir et combattre la criminalité (ISEC) Commission européenne – DG Affaires Intérieures.

Références

BROWNE P. (2009) *JBoss Drools Business Rules*, Packt Publishing, pp. 304, ISBN 1847196063, 2009.

LAFFERTY, J., MCCALLUM, A., PEREIRA, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001* (ICML 2001), Williamstown, MA, USA.

¹ <http://www.ho2s.com/fr/>

DisMo : un annotateur multi-niveaux pour les corpus oraux

George Christodoulides¹, Giulia Barreca², Mathieu Avanzi³

(1) Centre Valibel, Institut Langue & Communication, Université de Louvain
Place Blaise Pascal 1, B-1348 Louvain-la-Neuve, Belgique

(2) Laboratoire MoDyCo, CNRS, Université Paris Ouest Nanterre La Défense
200, Avenue de la République, FR-92001 Nanterre, France

Université Catholique de Milan
1, Largo A. Gemelli, 20123 Milan, Italie

(3) DTAL, Faculty of Modern & Medieval Languages, University of Cambridge
Sidgwick Avenue, CB3 9DA Cambridge, Royaume-Uni

george@mycontent.gr; giulia.barreca@gmail.com; mathieu.avanzi@gmail.com

Résumé. Dans cette démonstration, nous présentons l'annotateur multi-niveaux *DisMo*, un outil conçu pour faire face aux spécificités des corpus oraux. Il fournit une annotation morphosyntaxique, une lemmatisation, une détection des unités poly-lexicales, une détection des phénomènes de disfluece et des marqueurs de discours.

Abstract.

DisMo: a multi-level annotator for spoken language

In this demonstration we present the multi-level automatic annotator *DisMo* which is specifically designed for the challenges posed by spoken language corpora. Its output comprises of part-of-speech tagging, lemmatization, multi-word unit detection, detection of disfluency phenomena and discourse markers.

Mots-clés : annotation morphosyntaxique, corpus oraux, disfluences, unités poly-lexicales

Keywords: part-of-speech tagging, spoken corpora, disfluencies, multi-word expressions

1 Introduction

L'annotation des corpus oraux présente des défis particuliers, liés aux caractéristiques de la langue parlée et sa transcription, notamment : l'absence de ponctuation, les unités de segmentation multiples, les disfluences et la syntaxe souvent non-canonique. Si la méthodologie d'analyse et les outils d'annotation automatique doivent être adaptés, il est toutefois souhaitable de pouvoir comparer un corpus oral avec un corpus écrit, sur base d'un « dénominateur commun », et d'enrichir l'annotation avec des couches supplémentaires pour décrire les phénomènes propres à l'oral. Des études antérieures sur l'annotation morphosyntaxique des corpus oraux ont eu recours à des étiqueteurs conçus pour l'écrit, et adaptés à la suite d'un prétraitement des transcriptions des données orales (Blanc et al. 2008), ou d'un ajustement du corpus (Valli et Véronis, 1999). Certains auteurs ont également opté pour des solutions qui comportent soit un apprentissage automatique à partir de corpus oraux corrigés manuellement (Eshkol et al. 2010), soit l'utilisation de fichiers de paramétrage spécifiques pour les données orales (Benzitoun et al. 2012).

Cet article de démonstration présente *DisMo* (Christodoulides et al. 2014), un outil d'annotation automatique spécifiquement conçu pour les corpus oraux, et qui fournit une analyse multi-niveaux : étiquetage morphosyntaxique, lemmatisation, détection des unités poly-lexicales, détection et annotation des phénomènes de disfluece et des marqueurs de discours. Actuellement, trois grands corpus de référence du français parlé ont été annotés à l'aide de *DisMo*: le corpus Phonologie du Français Contemporain (PFC) (Durand et al. 2009) (1,4 million tokens), la collection des corpus du centre VALIBEL (Simon et al. 2014) (environ 6 million tokens), et le Corpus Oral de français de Suisse Romande (OFRON) (0,5 million tokens) (Avanzi et al. 2012). Les modèles statistiques de *DisMo* ont été entraînés d'abord sur le corpus CPROM-PFC (Avanzi 2014), lui-même contenant des échantillons du corpus PFC. Deux annotateurs experts ont corrigé l'annotation de 57 mille tokens. Ces premiers modèles ont été utilisés pour annoter 127 mille tokens du corpus PFC, et un annotateur expert a alors corrigé manuellement cet échantillon. L'ensemble de ces données a enfin permis d'entraîner des modèles statistiques. Après ces campagnes d'annotation et de correction, la précision du système est de l'ordre de 97% pour le jeu d'étiquettes complet, et 98% pour un jeu d'étiquettes simplifié.

2 Architecture du système

L'annotateur accepte plusieurs types d'entrées : l'analyse complète se base sur une transcription orthographique alignée au signal de la parole, au moins au niveau de l'énoncé. Dans ce cas, le système prend en compte dans son analyse des paramètres prosodiques calculés automatiquement (pauses silencieuses, débit de parole et mouvements mélodiques), notamment pour l'identification des disfluences et des unités de segmentation. Un alignement d'unités à un niveau plus fin (p.ex. au niveau des mots ou même des syllabes) peut permettre d'améliorer la performance. Il est aussi possible d'annoter une transcription non alignée, ou même un texte écrit. Le système est capable d'annoter des interactions, impliquant plusieurs locuteurs : les tours de parole sont considérés comme des indices de segmentation, et les annotations sont stockées séparément. Nous avons d'ailleurs développé des scripts qui facilitent le traitement des transcriptions selon certaines conventions (indices de locuteurs, conventions de transcription etc.). Les formats que le système accepte en entrée sont des fichiers *Praat TextGrid*, *TranscriberAG*, *ELAN*, *Exmaralda Partitur*, ou des fichiers texte. *DisMo* peut ajouter des tiers avec les résultats d'annotation et stocker des fichiers dans les formats mentionnés, produire en sortie des fichiers XML et OpenDocument, ou effectuer des modifications dans une base de données relationnelle (SQL, selon le schéma du logiciel *Praaline* ; Christodoulides 2014). *DisMo* est structuré autour de six modules, chaque module ajoutant ou modifiant des annotations sur les différents niveaux. Les opérations suivantes sont appliquées en cascade :

- Prétraitement et découpage en unités lexicales (tokenisation) ;
- Application de ressources linguistiques: les unités non-ambiguës sont annotées, la liste des étiquettes possibles est établie pour les autres. Certaines disfluences et unités poly-lexicales sont reconnues à ce stade, ainsi que les marqueurs de discours et les unités poly-lexicales potentiels ;
- Annotation morphosyntaxique (en partie du discours) préliminaire, à l'aide d'un modèle statistique CRF ;
- Détection des disfluences et de la segmentation, à l'aide de règles et d'un modèle CRF ;
- Annotation morphosyntaxique finale, combinée avec la détection des unités poly-lexicales, à l'aide d'un modèle statistique CRF.
- Post-traitement des annotations, à l'aide des règles de cohérence.

DisMo est écrit en C++ et utilise plusieurs bibliothèque de source ouverte, notamment *OpenFST*, *Helsinki Finite-State Transducer Technology (HFST)*¹ et *CRF++ toolkit*². Les ressources lexicales sont basées sur les dictionnaires DELA (Courtois et al., 1997), GLÀFF (Sajous et al., 2013) et des dictionnaires des unités nommés créés par les auteurs.

3 Niveaux d'annotation

L'articulation de l'annotation sur plusieurs niveaux est présentée dans la Figure 1. L'étiquetage morphosyntaxique attribue une des 64 catégories d'étiquettes (12 catégories principales), ainsi que des informations supplémentaires (genre, nombre, lemme) à chaque unité lexicale minimale et aussi à chaque unité poly-lexicale identifiée. Le jeu d'étiquettes complet, ainsi qu'une comparaison avec d'autres annotateurs est disponible sur le site web du logiciel (www.corpusannotation.org/dismo/tagset). Le schéma d'annotation pour les phénomènes de disfluence (hésitations, amorces, allongements, répétitions, insertions, substitutions, interruptions, etc.), ainsi que les algorithmes utilisés pour leur détection, sont détaillés dans (Christodoulides & Avanzi 2015).

FIGURE 1 : Les différents niveaux d'annotation sous forme de TextGrid.

–	un	mois	dans	une	agence	Saint	Cloud	donc	là	on	travaillait	euh	–	à la	main	–	tok-min (518)
–	NO	P	NOM	NO	NOM	AD	AD	VER	ITJ	–	D	NOM	–	D	NOM	–	pos-min (518)
SIL:1	M:co	R	com	Mp	o	V	D	impf	FIL	SIL:1			SIL:1				disfluency (518)
–	un	mois	dans	une	agence	Saint	Cloud	donc	là	on	travaillait	euh	–	à la	main	–	tok-mwu (497)
–	NO	P	NOM	NO	NOM	AD	AD	VER	ITJ	–	D	NOM	–	D	NOM	–	pos-mwu (497)
SIL:1	M:co	R	com	pro		V	D	impf		SIL:1			SIL:1				discourse (102/497)
–	un	mois	dans	une	agence	à Saint	Cloud	donc	là	on	travaillait	euh	–	à la	main	–	ortho (119)

¹ <http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/>

² <http://crfpp.googlecode.com/svn/trunk/doc/index.html>

4 Conclusion

Dans cet article, nous avons présenté une synthèse des travaux antérieurs portant sur l'annotation des corpus oraux, qui a permis d'aboutir à un nouveau système, conçu spécifiquement pour traiter les phénomènes propres aux corpus oraux. Ce système nous a permis d'annoter un grand ensemble de données de corpus oraux. Grâce à cette campagne d'annotation, plusieurs études ont pu être conduites (notamment dans sur la question de la variation régionale, l'étude de phénomènes phonotactiques, p. ex. la liaison, la chute des liquides post-obstruantes), les phénomènes syntaxiques propres à l'oral, la fluence et la disfluence, etc. Des interfaces web pour faciliter l'accès à ces données, ainsi qu'un service web pour l'utilisation de l'annotateur, sont en cours de conception et seront bientôt mises à la disposition de la communauté. *DisMo* est disponible (licence GPL3) sur le site web www.corpusannotation.org, en version autonome (plateformes Windows, Mac et Linux) ou comme plug-in pour le logiciel de gestion et d'annotation de corpus *Praaline*.

Références

- AVANZI M. (2014), A Corpus-Based Approach to French Regional Prosodic Variation, *Nouveaux cahiers de linguistique française*, 31, 309-323.
- AVANZI M., BEGUELIN M.-J., DIEMOZ F. (2012). *Présentation du corpus OFROM – corpus oral de français de Suisse romande*, Université de Neuchâtel, <http://www.unine.ch/ofrom>
- BENZITOUN C., FORT K., SAGOT B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. Actes de *JEP – TALN – RECITAL 2012*, Vol. 2: TALN, 99-112.
- BLANC O., CONSTANT M., DISTER A., WATRIN P. (2008). Corpus oraux et chunking. Actes de Journées d'étude sur la parole (JEP), Avignon, France.
- CHRISTODOULIDES G., AVANZI M. (2015). Automatic Detection and Annotation of Disfluencies in Spoken French Corpora, *Proceedings of Interspeech 2015*, Dresde, Allemagne, 6-10 septembre 2015, 5 pp.
- CHRISTODOULIDES G., AVANZI M., GOLDMAN, J-PH. (2014). DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. An Evaluation on a Corpus of French Spontaneous and Read Speech, *Proceedings of LREC 2014*, Reykjavik, Islande, 3902-3907.
- CHRISTODOULIDES, G. (2014). Praaline: Integrating tools for speech corpus research. Actes de *IX Language Resources and Evaluation Conference (LREC 2014)*, 26-31 mai 2014, Reykjavik, Islande, 31-34.
- COURTOIS B., GARRIGUES M., GROSS G., GROSS M., JUNG R., MATHIEU-COLAS M., MONCEAUX A., PONCET-MONTANGE A., SILBERZTEIN M., VIVÈS R. (1997). Dictionnaires électronique DELAC : les mots composés binaires. Rapport technique 56, LADL, Université Paris 7.
- DURAND J., LAKS B., LYCHE C., (EDS) (2009). *Phonologie, variation et accents du français*, Paris, Hermès.
- ESHKOL I., TELLIER I., TAALAB S., BILLOT S. (2010). Étiqueter un corpus oral par apprentissage automatique à l'aide de connaissances linguistiques. Actes de *10es Journées Internationales d'analyse statistique des données textuelles*.
- SAJOUS F., HATHOUT N., CALDERONE B. (2013). GLÀFF, un Gros Lexique À tout Faire du Français. Actes de *TALN*.
- SCHMID H. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
- SIMON A.C., FRANCARD M., HAMBYE P. (2014). "The VALIBEL Speech Database", In: Durand J., Gut U., Kristoffersen G., *The Oxford Handbook of Corpus Phonology*, DOI: 10.1093/oxfordhb/9780199571932.013.017.
- VALLI A., VERONIS J. (1999). Étiquetage automatique de corpus oraux : problèmes et perspectives. *Revue française de linguistique appliquée*, 4(2), 113-133.