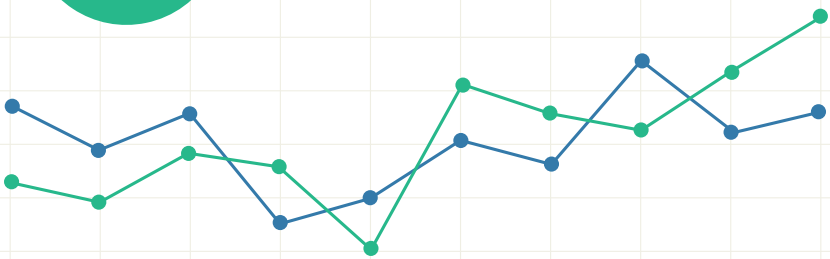
TALN

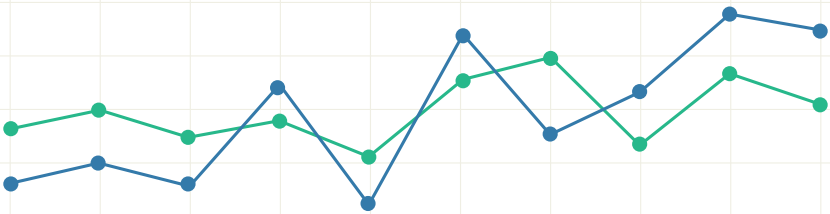
& RÉCITAL

22e Conférence sur le Traitement Automatique des Langues Naturelles



Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues

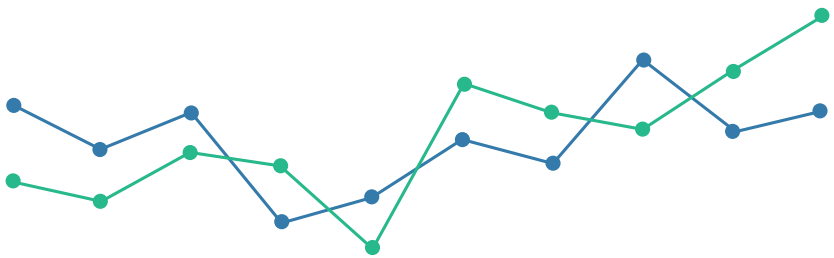
17e



22-25 JUIN 2015 - CAEN

Université de Caen Basse-Normandie

SOMMAIRE



AVANT PROPOS

Cette vingt-deuxième édition de TALN et cette dix-septième édition de RÉCITAL ont lieu pour la première fois en terre normande, sur le campus universitaire de Caen. Leur organisation a été assurée par l'équipe HULTECH du GREYC (Groupe de recherche en informatique, image, automatique et instrumentation de Caen) et le CRISCO (Centre de Re- cherches Inter-langues sur la Signification en COntexte). Après Marseille l'an passé, nous traversons donc la France pour arriver dans la cité de Guillaume Le Conquérant. Les participants auront l'occasion de découvrir quelques-uns des aspects de notre région au fil de ces quatre journées.

Cette édition 2015 de TALN et RÉCITAL permettra à de nombreux chercheurs et doctorants de présenter leurs travaux sous la forme de communications orales, de posters ou de démonstrations. Sur 48 articles longs soumis, 26 seront présentés lors des sessions orales. En ce qui concerne les articles courts, 46 articles sur 63 ont été acceptés. Le comité de programme a choisi de favoriser les échanges au sein de la communauté de recherche en acceptant des articles aux thématiques très diversifiées. Nous remercions ici les membres du comité de programme et de lecture pour le travail réalisé, parfois avec des délais serrés.

La conférence RÉCITAL est un lieu d'échange privilégié entre la communauté et ses jeunes chercheurs. Cette année encore, elle fait naturellement la part belle aux travaux prospectifs et aux états de l'art. Le processus d'acceptation est sélectif (sur 12 soumissions, 7 articles ont été acceptés, dont 3 sous forme d'articles longs et 4 sous forme de posters), mais les relectures ont aussi largement vocation à conseiller et à encourager les jeunes chercheurs, que leur soumission soit acceptée ou refusée. Nous remercions les membres du comité de programme de RÉCITAL pour ce travail particulièrement important pour la relève du TAL.

Nous recevrons deux conférenciers invités, Roberto Navigli, Professeur Associé à Sapienza – Università di Roma, et Marie-Claude L'Homme, Professeure à l'Université de Montréal et directrice de l'Observatoire de linguistique Sens- Texte. Nous les remercions tous deux chaleureusement d'avoir accepté de partager avec nous leurs réflexions et leur expérience.

La conférence accueille cette année cinq ateliers : DEFT – DÉfi Fouille de Textes (11e édition) ; ETeRNAL – Éthique et TRaitemeNt Automatique des Langues ; ITI – Interface Tal-Ihm ; TALaRE -Traitement Automatique des Langues Régionales de France et d'Europe et enfin TASLA – Traitement Automatique des langues SLAves. Plusieurs de ces ateliers sont de nouvelles créations, preuve s'il en faut du dynamisme de notre communauté. Nous remercions les organisateurs de ces ateliers pour leur travail d'animation et d'organisation. Enfin, douze démonstrations seront présentées, qui mettent en valeur le dynamisme de la communauté de TALN dans le domaine de la production logicielle.

Afin de faire le lien entre la Recherche et l'Industrie au moyen de rencontres informelles, de présentations de projets et solutions logicielles, TALN accueille de nouveau cette année le Salon de l'Innovation. L'objectif de ce salon est d'offrir la possibilité de présenter les principaux projets de recherche publics et privés conduits en France et dans le monde francophone. Cela concerne les grands produits ou projets développés par les sociétés privées ou en collaboration avec elles, mais également les grands projets publics tels ceux découlant des contrats ANR. Ce salon est donc un lieu d'échange entre recherche publique et privée.

Une telle conférence, réunissant près de 200 personnes, ne peut se faire sans le soutien des institutions (Université de Caen Normandie, ENSICAEN, Région Basse-Normandie, CNRS, Agence Nationale de la Recherche, Ministère de la Culture et de la Communication) et de partenaires privés (Noopsis, Syllabs, SucceedTogether). Nous les en remercions.

L'équipe d'organisation tient également à remercier chaleureusement le comité permanent de l'ATALA qui a supervisé la sélection scientifique, assurant la continuité scientifique de TALN, événement clé de l'animation de la communauté francophone du traitement automatique des langues.

Co-Présidents de TALN2015
Jean-Marc Lecarpentier et Nadine Lucas

Présidente et Vice-Président de RÉCITAL2015
Charlotte Lecluze et José G. Moreno

Présidence par intérim de TALN 2015 - les membres cooptés du CPERM
Philippe Blache, Emmanuel Morin, Pascale Sébillot et Pierre Zweigenbaum

PARTENAIRES

GREYC HULTECH

Centre de Recherches Inter-langues sur la Signification en COntexte
EA 4255

ATALA

ENSICAEN
ÉCOLE NATIONALE SUPÉRIEURE D'INGÉNIEURS DE CAEN & CENTRE DE RECHERCHE

UNICAEN université de Caen Basse-Normandie

cnrs dépasser les frontières

syllabs

NOOPSIS
sense mining solutions

AGENCE NATIONALE DE LA RECHERCHE
ANR
projet ART-ADN : ANR-12-SOIN-0003-02

RÉGION BASSE NORMANDIE

SucceedTogether
Générateur d'intélligence collaborative

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Ministère Culture Communication

Conception graphique : LOUISE MARNAI .free.fr

INVITÉS

ROBERTO NAVIGLI

Multilinguality at Your Fingertips: BabelNet, Babelfy and Beyond!

Associate professor in the Department of Computer Science at the Sapienza University of Rome and member of the Linguistic Computing Laboratory

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.

MARIE-CLAUDE L'HOMME

Pourquoi construire des ressources terminologiques et pourquoi le faire différemment ?

Professeure à l'Université de Montréal, directrice de l'Observatoire de linguistique Sens-Texte

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 portait 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.

COMITÉ D'ORGANISATION TALN

PRÉSIDENCE :

Jean-Marc Lecarpentier
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Nadine Lucas
(Université de Caen Basse-Normandie, GREYC/HULTECH)

VICE-PRÉSIDENCE :

Gaël Dias
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Pierre Larrivée
(Université de Caen Basse-Normandie, CRISCO)

Pierre Beust
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Lilia Boughchiche
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Virginie Desnos-Carreau
(Université de Caen Basse-Normandie)

Gaël Dias
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Stéphane Ferrari
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Pierre Larrivée
(Université de Caen Basse-Normandie, CRISCO)

Charlotte Lecluze
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Jean-Marc Lecarpentier
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Nadine Lucas
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Paul Martin
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Yann Mathet
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Serge Mauger
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Fabrice Maurel
(Université de Caen Basse-Normandie, GREYC/HULTECH)

José G. Moreno
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Marc Spaniol
(Université de Caen Basse-Normandie, GREYC/HULTECH)

COMITÉ DE PROGRAMME TALN

Vincent Claveau
(CNRS, IRISA)

Gaël Dias
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Claire Gardent
(CNRS, LORIA)

Nabil Hathout
(CNRS, CLLE/ERSS)

Philippe Langlais
(Université de Montréal, RALI)

Pierre Larrivée
(Université de Caen Basse-Normandie, CRISCO)

Emmanuel Morin
(Université de Nantes, LINA)

Adeline Nazarenko
(Université Paris-Nord, LIPN)

Marc Spaniol
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Pierre Zweigenbaum
(CNRS, LIMSI)

Pascale Sébillot
(INSA de Rennes, IRISA)

COMITÉ SCIENTIFIQUE TALN

Stergos Afantenos
(Université Paul Sabatier, IRIT)

Salah Aït-Mokhtar
(Xerox Research Centre Europe)

Maxime Amblard
(Université de Lorraine, LORIA)

Jean-Yves Antoine
(Université François Rabelais Tours/Blois, LI)

Delphine Battistelli
(Université Paris Ouest – Nanterre, MoDyCo)

Denis Bechet
(Université de Nantes, LINA)

Delphine Bernhard
(Université de Strasbourg, LiLPa)

Romaric Besançon
(CEA LIST)

Yves Bestgen
(Université catholique de Louvain)

Brigitte Bigi
(CNRS, LPL)

Philippe Blache
(CNRS, LPL)

Hervé Blanchon
(Université Grenoble 2, LIG)

Florian Boudin
(Université de Nantes, LINA)

Julien Bourdaillet
(Xerox Corporation USA)

Francis Brunet-Manquat
(Université Grenoble 2, LIG)

Nathalie Camelin
(Université du Maine, LIUM)

Benoit Crabbé
(Université Paris 7, INRIA)

Beatrice Daille
(Université de Nantes, LINA)

Géraldine Damnati
(Orange Labs)

Laurence Danlos
(Université Paris 7, INRIA)

Marc Dymetman
(Xerox Research Centre Europe)

Iris Eshkol
(Université d'Orléans, LLL)

Cécile Fabre
(Université Toulouse 2, CLLE/ERSS)

Benoit Favre
(Université d'Aix-Marseille, LIF)

Olivier Ferret
(CEA LIST)

Dominic Forest
(Université de Montréal)

Karën Fort
(Université Paris-Sorbonne, STIH)

Nathalie Friburger
(Université François Rabelais Tours/BLOIS, LI)

Michel Gagnon
(École Polytechnique de Montréal)

Éric Gaussier
(Université Joseph Fourier, LIG)

Kim Gerdes
(Université Sorbonne Nouvelle)

Jérôme Goulian
(Université de Grenoble 2, LIG)

Natalia Grabar
(CNRS, STL)

Lamia Hadrich Belguith
(Université de Sfax, MIRACL)

John Hale
(Cornell University)

Thierry Hamon
(Université Paris Nord, LIMSI)

Nicolas Hernandez
(Université de Nantes, LINA)

Stéphane Huet
(Université d'Avignon et des Pays de Vaucluse, LIA)

Christine Jacquin
(Université de Nantes, LINA)

Sylvain Kahane
(Université Paris Ouest – Nanterre, MoDyCo)

Olivier Kraif
(Université Stendhal Grenoble 3, LIDILEM)

Mathieu Lafourcade
(Université Mon 2, LIRMM)

Éric Laporte
(Université Paris-Est Marne-la-Vallée, LIGM)

Joseph Le Roux
(Université Paris 13, LIPN)

Anne-Laure Ligozat
(ENSIIE, LIMSI)

Denis Maurel
(Université Francois Rabelais Tours, LI)

Aurélien Max
(Université Paris-Sud, LIMSI)

Richard Moot
(CNRS, LaBRI)

Erwan Moreau
(Trinity College Dublin)

Véronique Moriceau
(Université Paris-Sud, LIMSI)

Jean-Yves Morin
(Université de Montréal)

Philippe Muller
(Université Paul Sabatier Toulouse, IRIT)

Luka Nerima
(Université de Genève)

Jian-Yun Nie
(Université de Montréal)

Aurélie Névéol
(CNRS, LIMSI)

Yannick Parmentier
(Universite d'Orléans, LIFO)

Thierry Poibeau
(CNRS, LaTTiCe)

Alain Polguère
(Université de Lorraine, ATILF)

Andrei Popescu-Belis
(IDIAP Research Institute)

Jean-Philippe Prost
(Université Montpellier 2, LIRMM)

Solen Quiniou
(Université de Nantes, LINA)

Christian Raymond
(INSA de Rennes, IRISA)

Christian Retoré
(Université de Montpellier, LIRMM)

Mathieu Roche
(Cirad, TETIS)

Didier Schwab
(Université Grenoble 2, LIG)

Djamé Seddah
(Université Paris-Sorbonne, INRIA)

Michel Simard
(National Research Council Canada)

Kamel Smaïli
(Université de Lorraine, LORIA)

Xavier Tannier
(Université Paris-Sud, LIMSI)

Isabelle Tellier
(Université Sorbonne Nouvelle, LaTTiCe)

Juan-Manuel Torres-Moreno
(Université d'Avignon et des Pays de Vaucluse, LIA)

Guillaume Wisniewski
(Université, Paris-Sud, LIMSI)

François Yvon
(Université Paris-Sud, LIMSI)

Michael Zock
(CNRS, LIF)

Mounir Zrigui
(Université Monastir, UTIC)

RELECTEURS ADDITIONNELS TALN

Mohamed Achraf Ben Mohamed
(University of Monastir, UTIC)

Thierry Charnois
(Université Paris 13, LIPN)

Marco Dinarelli
(CNRS, LaTTiCe)

Yannick Estève
(Université du Maine, LIUM)

Joseph Leroux
(Université Paris 13, LIPN)

Elizaveta Loginova
(Université de Nantes, LINA)

Seifeddine Mechti
(University of Sfax, MIRACL)

Laroussi Merhbène
(University of Monastir, LATICE)

Vassilina Nikoulina
(Xerox Research Centre Europe)

Corentin Ribeyre
(Université Paris 7, ALPAGE)

Sophie Rosset
(CNRS, LIMSI)

Benoît Sagot
(INRIA, ALPAGE)

Guillaume Wisniewski
(Université Paris-Sud, LIMSI)

COMITÉ D'ORGANISATION RÉCITAL

PRÉSIDENTE :

Charlotte Lecluze
(Université de Caen Basse-Normandie, GREYC/HULTECH)

VICE-PRÉSIDENT :

José G. Moreno
(Université de Caen Basse-Normandie, GREYC/HULTECH)

COMITÉ DE PROGRAMME RÉCITAL

Jean-Yves Antoine
(Université François Rabelais Tours/Blois, LI)

Adrien Barbaresi
(ENS Lyon)

Loïc Barrault
(Université du Maine)

Patrice Bellot
(Polytech Marseille – Aix-Marseille Université, LSIS)

Pierre Beust
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Frederik Bilhaut
(NOOPSIS)

Florian Boudin
(Université de Nantes, LINA)

Romain Brixtel
(University of Lausanne, OB)

Marie Candito
(Université Paris Diderot, Alpage)

Thierry Charnois
(Université Paris 13, LIPN)

Benoit Crabbé
(Université Paris 7, INRIA)

Hervé Déjean
(Xerox Center)

Eliane Delente
(Université de Caen Basse-Normandie, CRISCO)

Antoine Doucet
(University of La Rochelle, L3i)

Cécile Fabre
(Université Toulouse 2, CLLE/ERSS)

Benoit Favre
(Université d'Aix-Marseille, LIF)

Dominic Forest
(Université de Montréal)

Thomas Francois
(UCLouvain), CENTAL)

Nuria Gala
(Aix-Marseille Université, LIF)

Thierry Hamon
(Université Paris Nord, LIMSI)

Nicolas Hernandez
(Université de Nantes, LINA)

Olivier Kraif
(Université Stendhal Grenoble 3, LIDILEM)

Jean-Charles Lamirel
(Université Henri Poincaré Nancy 1, LORIA)

Philippe Langlais
(Université de Montréal, RALI)

Anais Lefeuvre
(Université François Rabelais, Tours)

Gaël Lejeune
(Université de nantes, LINA)

Cédric Lopez
(VISEO)

Yann Mathet
(Université de Caen Basse-Normandie, GREYC/HULTECH)

Yayoi Nakamura-Delloye
(INALCO)

Damien Nouvel
(INALCO)

Jean-Philippe Prost
(Université Montpellier 2, LIRMM)

Richard Renault
(Université de Caen Basse-Normandie, CRISCO)

Lois Rigouste
(MyScript)

Fatiha Sadat
(Université du Québec à Montréal)

Gregory Smits
(Université de Rennes 1, IRISA)

Juan-Manuel Torres-Moreno
(Université d'Avignon et des Pays de Vaucluse, LIA)

Anne Vilnat
(Université Paris-Sud, LIMSI)

Lei Zhang
(Hong Kong Polytechnic University)

Maria Zimina
(Université Paris Diderot, CLILLAC-ARP)

Michael Zock
(Aix-Marseille Université, LIF)

Mounir Zrigui
(Faculté des Sciences de Monastir et de l'Informatique à Carthage)

Pierre Zweigenbaum
(CNRS, LIMSI)

LUNDI 22 ATELIERS

Les ateliers portent sur une thématique particulière de TAL afin de rassembler quelques exposés plus ciblés que lors de la conférence plénière. Chaque atelier a son propre président et son propre comité de programme.
Le responsable de l'atelier est chargé de l'appel à candidatures et de la coordination de son comité de programme. Les ateliers ont lieu en parallèle durant une journée ou une demi-journée (2 à 4 sessions de 1h30) à l'Université de Caen Basse-Normandie le lundi 22 juin 2015.

9h00 - 17h30

ATELIER DEFT

DÉfi Fouille de Textes

Créé en 2005 à l'image des campagnes TREC et MUC, le DÉfi Fouille de Textes est une campagne d'évaluation francophone qui propose chaque année de confronter les méthodes de plusieurs équipes de recherche sur une thématique régulièrement renouvelée. Cette onzième édition portera sur l'analyse de l'opinion, des sentiments et des émotions dans des tweets r rédigés en français.

● *Amphithéâtre S3-043*

ATELIER ETERNAL

Éthique & TRaitemeNt Automatique des Langues

Les questions sous-jacentes que nous souhaiterions voir aborder concernent aussi bien les apports du TAL à l'ethique que nos responsabilites en tant que producteurs d'outils. Nous ne pouvons en effet pas faire semblant de ne pas savoir que ceux-ci rendent possibles des abus, des actes criminels, des violations des droits individuels. Aujourd'hui, de quoi les outils de TAL sont-ils capables ? Jusqu'où s'etend notre responsabilité morale ? Devons-nous etre des lanceurs d'alertes ? Quels mesures peut-on prendre pour limiter les effets potentiellement negatifs de nos recherche ?

● *Amphithéâtre S3-044*

ATELIER ITI

Interface Tal-Ihm

La coopération entre les disciplines « Traitement Automatique des Langues » (TAL) et « Interaction Homme Machine » (IHM) a toujours semblee évidente et sur un mode que nous pourrions qualifier de fusionnel. Plus récemment on observe également une forte intrication de ces deux domaines de recherche pour la conception d'aides techniques pour pallier les situations de handicaps sensoriels ou situationnels. Certaines thématiques sont remarquables par les relations qu'elles impliquent entre traitement automatique d'une langue et interaction avec une machine. Les recherches proposées devront approcher au moins deux des trois thèmes suivants : TAL, IHM et handicap ; TAL, IHM et texte ; TAL, IHM et interprétation.

● *Amphithéâtre S3-045*

9h00 - 12h30

ATELIER TALaRE

Traitement Automatique des Langues Régionales de France et d'Europe

Les recherches en traitement automatique des langues peu et moyennement dotées connaissent actuellement un regain d'intérêt à travers la constitution de corpus et de lexiques dans une perspective globale de préservation du patrimoine culturel. Les langues régionales sont généralement à ranger dans cette catégorie, car les ressources électroniques pour ces langues sont rares, peu visibles et sous exploitées, parfois inexistantes. Nous appelons à la soumission de travaux de recherche autour de la constitution de ressources et d'outils pour les langues régionales ou minoritaires de France et d'Europe (y compris les langues d'Outre Mer).

● *Salle S3-161*

14h00 - 17h30

ATELIER TASLA

Traitement Automatique des langues SLAves

Les langues slaves suscitent l'intérêt de plusieurs chercheurs en linguistique informatique dans différents contextes scientifiques, et notamment dans le cadre francophone. Réputées pour leur complexité, elles font l'objet de différentes études qui ont pour but leur modélisation et leur exploration à tous les niveaux linguistiques. Dans le paysage de la recherche francophone, les études sont menées sur des langues très hétérogènes dans ce groupe, allant des langues slaves du Sud (bulgare, macédonien, BCMS) aux langues slaves de l'Est (russe, ukrainien, biélorusse) et de l'Ouest (polonais, tchèque, slovaque, slovène). L'atelier que nous proposons a pour objectif de rapprocher des chercheurs en TAL des langues slaves, pour mettre en commun les méthodologies et les expériences en vue de (i) rassembler et systématiser les recherches existantes et de (ii) réfléchir à la construction d'un projet fédérateur autour de la problématique des convergences/divergences entre les langues slaves et des méthodes de leur traitement dans le cadre francophone.

● *Salle S3-161*

MARDI 23 SESSIONS ORALES

9h00 - 10h00

CONFÉRENCE INVITÉ ROBERTO NAVIGLI

● *Amphithéâtre S3-057*

Président : Marc Spaniol

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.

10h30 - 12h00

TRADUCTION

● *Amphithéâtre S3-057*

Président : David Langlois

1° Utilisation de mesures de confiance pour améliorer le décodage en traduction de parole.

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 traduc- tion 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).

Par Laurent Besacier, Benjamin Lecouteux et Luong Ngoc Quang.

2° Multi-alignement vs bi-alignement : à plusieurs, c'est mieux !

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.

Par Olivier Kraif.

3° Apprentissage discriminant des modèles continus de traduction

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 les tâches de traduction des séminaires TED Talks de l'anglais vers le français montrent la pertinence d'un cadre discriminant d'apprentissage mais dont les performances sont liées au choix d'une stratégie d'initialisation adéquate. Nous

montrons qu'avec une initialisation judicieuse des gains significatifs en terme de score peuvent être obtenus.
Par Quoc-Khanh Do, Alexandre Allauzen et François Yvon.

DÉSAMBIGUÏSATION

● *Amphithéâtre S3-049*
Présidente : Charlotte LECLUZE

1° Désambiguïsation d'entités pour l'induction non supervisée de schémas événementiels

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.
Par Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret et Romaric Besançon.

2° Création rapide et efficace d'un système de désambiguïsation lexicale pour une langue peu dotée

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 généricité en traduisant le \semcor, un corpus en anglais annoté grâce au \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.
Par Mohammad Nasiruddin, Andon Tchechmedjiev, Hervé Blanchon et Didier Schwab.

RÉCI TAL

3° Désambiguïsation lexicale à base de connaissances par sélection distributionnelle et traits sémantiques (Article RÉCITAL)

La désambiguïsation lexicale permet d'améliorer de nombreuses applications en traitement automatique des langues (TAL) comme la recherche d'information, l'extraction d'information, la traduction automatique, ou la simplification lexicale de textes. Schématiquement, il s'agit de choisir quel est le sens le plus approprié pour chaque mot d'un texte. Une des approches classiques consiste à estimer la similarité sémantique qui existe entre les sens de deux mots puis de l'étendre à l'ensemble des mots du texte. La méthode la plus directe donne un score de similarité à toutes les paires de sens de mots puis choisit la chaîne de sens qui retourne le meilleur score (on imagine la complexité exponentielle liée à cette approche exhaustive). Dans cet article, nous proposons d'utiliser une méta-heuristique d'optimisation combinatoire qui consiste à choisir une fenêtre contenant les voisins les plus proches par sélection distributionnelle autour du mot à désambiguïser. Le test et l'évaluation de notre méthode portent sur un corpus écrit en langue française en se servant du réseau sémantique BabelNet. Le taux d'exactitude obtenu est de 78% sur l'ensemble des noms et des verbes choisis pour l'évaluation.
Par Mokhtar Boumedyen Billami.

13h30 - 15h30

SYNTAXE & PARAPHRASE

● *Amphithéâtre S3-057*
Président : Jean-Yves Antoine

1° Grammaires phrastiques et discursives fondées sur TAG : une approche de D-STAG avec les ACG

Nous présentons une méthode pour articuler grammaire de phrase et grammaire de discours. Cette méthode permet à la fois l'intégration des deux grammaires sans recourir à une étape de traitement intermédiaire et de 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 utilisant les Grammaires d'Arbres Adjoint (TAG), Discourse Synchronous TAG (D-STAG). Nous utilisons pour ce faire un encodage des TAG dans les Grammaires Catégorielles Abstraites (ACG). Cela 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 afin d'articuler naturellement grammaire phrastique et grammaire discursive. Tous les exemples peuvent être exécutés avec le logiciel approprié.
Laurence Danlos, Aleksandre Maskharashvili et Sylvain Pogodalla.

2° Analyse syntaxique de l'ancien français : quelles propriétés de la langue influent le plus sur la qualité de l'apprentissage ?

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. 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 par nature hétérogènes et aussi quantitativement limitées. Nous explorons donc diverses stratégies, fondées sur différents critères (variabilité du lexique, forme Vers/Prose des textes, époque de rédaction), pour constiter des corpus d'entrainement menant aux meilleurs résultats possibles.

Par Gaël Guibon, Isabelle Tellier, Matthieu Constant et Kim Gerdes.

3° Noyaux de réécriture de phrases munis de types lexico-sémantiques

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.

Par Martin Gleize et Brigitte Grau.

4° Extraction automatique de paraphrases grand public pour les termes médicaux

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 (eg, 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 (eg, recherche d'information grand public, lisibilité et simplification de textes, systèmes de question-réponses).

Par Natalia Grabar et Thierry Hamon.

EXTRACTION D'INFORMATION

● ***Amphithéâtre S3-049***

Président : Aurélien Bossard

1° Apprentissage par imitation pour l'étiquetage de séquences : vers une formalisation des méthodes d'étiquetage easy-first

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 et 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.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.

Par Elena Knyazeva, Guillaume Wisniewski et François Yvon.

2° 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

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.

Par Nicolas Pécheux, Alexandre Allauzen, Thomas Lavergne, Guillaume Wisniewski et François Yvon.

3° Stratégies de sélection des exemples pour l'apprentissage actif avec des CRF

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 à 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 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 proportion 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.

Par Vincent Claveau et Ewa Kijak.

4° Identification de facteurs de risque pour des patients diabétiques à partir de comptes-rendus cliniques par des approches hybrides

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.

Par Cyril Grouin, Véronique Moriceau, Sophie Rosset et Pierre Zweigenbaum.

16h00 - 17h30

CLASSIFICATION & ALIGNEMENT

● *Amphithéâtre S3-057*

Président : Florian Boudin

1° Typologie des langues automatique à partir de treebanks

La typologie des langues consiste à identifier certaines propriétés syntaxiques et de les comparer au travers de plusieurs langues. Nous proposons dans cet article d'extraire automatiquement ces propriétés à partir de treebanks et de les analyser en vue de dresser une typologie. Nous décrivons cette méthode ainsi que les outils développés pour la mettre en œuvre. Nous appliquons la méthode 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 langue.

Par Philippe Blache, Grégroie de Montcheuil et Stéphane Rauzy.

2° Attribution d'Auteur : approche multilingue fondée sur les répétitions maximales

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-ème ordre (c-à-d 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.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.

Par Romain Brixtel, Charlotte Lecluze et Gaël Lejeune.

3° Mesurer la similarité entre phrases grâce à Wikipédia en utilisant une indexation aléatoire.

Ce papier 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.

Par Hai Hieu Vu, Jeanne Villaneau, Farida Saïd et Pierre-François Marteau.

COMPRÉHENSION & PARAPHRASE

● ***Amphithéâtre S3-049***

Président : Olivier Ferret

1° Compréhension automatique de la parole sans données de référence

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 de l'utilisateur.

Par Emmanuel Ferreira, Bassam Jabaian et Fabrice Lefèvre.

RÉCI TAL

2° fr2sql : Interrogation de bases de données en français (Article RECITAL)

Les bases de données sont de plus en plus courantes et prennent de plus en plus d'ampleur au sein des applications et sites Web actuels. Elles sont souvent amenées à être utilisées par des personnes n'ayant pas une grande compétence en la matière et ne connaissant pas rigoureusement leur structure. C'est pour cette raison que des traducteurs du langage naturel aux requêtes SQL sont développés. Malheureusement, la plupart de ces traducteurs se cantonnent à une seule base du fait de la spécificité de l'architecture de celle-ci. Dans cet article, nous proposons une méthode visant à pouvoir interroger n'importe quelle base de données à partir du français. Nous évaluons notre application sur deux tables à la structure différente et nous montrons également qu'elle supporte plus d'opérations que la plupart des autres traducteurs.

Par Jérémy Ferrero.

3° Analyse d'expressions temporelles dans les dossiers électroniques patients

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 s'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 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.

Par Mike Donald Tapi Nzali, Aurélie Névéol et Xavier Tannier.

MERCREDI 24 SESSIONS ORALES

9h00 - 10h30

OPINIONS & SENTIMENTS

● *Amphithéâtre S3-057*

Président : Patrick Paroubek

1° Méthode faiblement supervisée pour l'extraction d'opinion ciblée dans un domaine spécifique

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. 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 vidéos.

Par Romaric Besançon.

2° 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

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 études en fouille de textes subjectifs (analyse de sentiments et fouille d'opinions). Ces études 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.

Par Mathieu Valette et Egle Eensoo.

RÉCI TAL

3° Vers un modèle de détection des affects, appréciations et jugements dans le cadre d'interactions humain-agent (Article RECITAL)

Cet article aborde la question de la détection des expressions d'attitude — affect, d'appréciation et de jugement (Martin and White, 2005)– dans le contenu verbal de l'utilisateur au cours d'interactions en face-à-face avec un agent conversationnel animé. Il propose un positionnement en terme de modèles et de méthodes pour le développement d'un système de détection adapté aux buts communicationnels de l'agent et à une parole conversationnelle. Après une description du modèle théorique de référence choisi, l'article propose un modèle d'annotation des attitudes dédié l'exploration de ce phénomène dans un corpus d'interaction humain-agent. Il présente ensuite une première version de notre système. Cette première version se concentre sur la détection des expressions d'attitudes pouvant référer à ce qu'aime ou n'aime pas l'utilisateur. Le système est conçu selon une approche symbolique fondée sur un ensemble de règles sémantiques et de représentations logico-sémantiques des énoncés.

Par Caroline Langlet.

SÉMANTIQUE

● *Amphithéâtre S3-049*

Présidente : Delphine Bernard

1° Estimation de l'homogénéité sémantique pour les Questionnaires à Choix Multiples

L'homogénéité sémantique stipule que des termes sont sémantiquement proches mais non similaires. Cette notion est au coeur 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.

Par Van-Minh Pho, Anne-Laure Ligozat et Brigitte Grau.

2° Extraction automatique de relations sémantiques dans les définitions : approche hybride, construction d'un corpus de relations sémantiques pour le français

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.

Par Emmanuel Cartier.

3° Déclasser les voisins non sémantiques pour améliorer les thésaurus distributionnels

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 cooccurrents syntaxiques et nous montrons l'intérêt de la combiner avec les méthodes décrites dans (Ferret, 2013) selon une stratégie de type vote.

Par Olivier Ferret.

14h00 - 15h30

PLÉNIÈRE

● ***Amphithéâtre S3-057***

Président : Pierre Zweigenbaum

1° Comparaison d'architectures neuronales pour l'analyse syntaxique en constituants

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.L'article traite de l'analyse syntaxique lexicalisée pour les grammaires de constituants.

Par Maximin Coavoux et Benoît Crabbé.

2° ...des conférences enfin disons des causeries... Détection automatique de segments en relation de paraphrase dans les reformulations de corpus oraux

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 d'autres points de vue.

Par Natalia Grabar et Iris Eshkol.

3° Utiliser les interjections pour détecter les émotions

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.

Par Amel Fraisse et Patrick Paroubek.

JEUDI 25 • SESSIONS POSTERS & DÉMOS

9h00 - 10h00

CONFÉRENCE INVITÉE MARIE-CLAUDE L'HOMME

● *Amphithéâtre S3-057*

Président : Pierre Beust

Pourquoi construire des ressources terminologiques et pourquoi le faire différemment ?

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 portait 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.

> SESSION 1

10h00 - 10h30

INTRODUCTION POSTERS/DÉMONSTRATIONS

● *Amphithéâtre S3-057*

10h30 - 12h00

POSTERS

● *1er étage*

A Simple Discriminative Training Method for Machine Translation with Large-Scale Features

Margin infused relaxed algorithms (MIRAs) dominate 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.

Tian Xia, Shaodan Zhai, Zhongliang Li et Shaojun Wang.

Natural Language Reasoning using Coq : Interaction and Automation

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.

Stergios Chatzikyriakidis.

Vous aimez ?...ou pas ? LikeIt, un jeu pour construire une ressource lexicale de polarité

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.

Mathieu Lafourcade, Nathalie Le Brun et Alain Joubert.

Étude des verbes introducteurs de noms de médicaments dans les forums de santé

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.

François Morlane-Hondère, Cyril Grouin et Pierre Zweigenbaum.

Initialisation de Réseaux de Neurones à l'aide d'un Espace Thématique

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 amélioration 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.

Mohamed Morchid, Richard Dufour et Georges Linarès.

FDTB1: Repérage des connecteurs de discours en corpus

Cet article présente le repérage 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.

Jacques Steinlin, Margot Colinet et Laurence Danlos.

ROBO : Une mesure d'édition pour la comparaison de phrases - Application au résumé automatique

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.

Aurélien Bossard et Christophe Rodrigues.

Classification d'entités nommées de type « film »

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.

Olivier Collin et Aleksandra Guerraz.

A critical survey on measuring success in rank-based keyword assignment to documents

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.

Natalie Schluter.

Effects of Graph Generation for Unsupervised Non-Contextual Single Document Keyword Extraction

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.

Natalie Schluter.

Adaptation par enrichissement terminologique en traduction automatique statistique fondée sur la génération et le filtrage de bi-segments virtuels

Nous proposons des travaux préliminaires sur une approche permettant d'ajouter des termes bilingues à un système de Traduction Automatique Statistique (TAS) à base de segments. Ces termes sont, non seulement, inclus individuellement, mais aussi avec des contextes induits autour de ces mots. 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.

Christophe Servan et Marc Dymetman.

Une mesure d'intérêt à base de surreprésentation pour l'extraction des motifs syntaxiques stylistiques

Dans cette contribution, nous présentons une étude sur la stylistique computationnelle des textes de la littérature classiques 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.

Mohamed Amine Boukhaled, Francesca Frontini et Jean-Gabriel Ganascia.

Une approche évolutionnaire pour le résumé automatique

Dans cet article, nous proposons une méthode de résumé automatique fondés sur l'utilisation d'un algorithme génétique pour parcourir l'espace des résumés candidats couplé à un calcul de divergence 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.

Aurélien Bossard et Christophe Rodrigues.

Identification des unités de mesure dans les textes scientifiques

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.

Soumia Lilia Berrahou, Patrice Buche, Juliette Dibie-Barthélemy et Mathieu Roche.

Évaluation intrinsèque et extrinsèque du nettoyage de pages Web

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.

Gaël Lejeune, Romain Brixtel et Charlotte Lecluze.

CANÉPHORE : un corpus français pour la fouille d'opinion ciblée

La fouille d'opinion ciblée (aspect-based sentiment analysis) connaît ces dernières années 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. L'annotation fournie permet l'évaluation de systèmes de fouille d'opinion à plusieurs niveaux de granularité. Nous accompagnons ce corpus de résultats de référence pour l'extraction de marqueurs d'opinion non supervisée. Dans un deuxième temps nous présentons une méthode améliorant les résultats de cette extraction, en suivant une approche semi-supervisée.

Joseph Lark, Emmanuel Morin et Sebastian Peña Saldarriaga.

Extraction des Contextes Riches en Connaissances en corpus spécialisés

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 \oe uvre la notion de Contextes Riches en Connaissance 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.

Firas Hmida, Emmanuel Morin et Béatrice Daille.

Traitement automatique des formes métriques des textes versifiés

L'objectif de cet article est de présenter tout d'abord dans ses grandes lignes le projet qui a pour objet le traitement automatique des formes métriques de la poésie et du théâtre français du début du XVIIe au début du XXe 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. Enfin, nous présenterons la procédure d'appariement des rimes et de détermination des schémas de strophes dans les suites périodiques et les formes fixes.

Eliane Delente et Richard Renault.

Apprentissage automatique d'un modèle de résolution de la coréférence à partir de données orales transcrites du français : le système CROC

Cet article présente CROC (Coreference Resolution for Oral Corpus), le 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.

Adèle Désoyer, Frédéric Landragin et Isabelle Tellier.

Vers un diagnostic d'ambiguïté des termes candidats d'un texte

Les recherches autour de la désambiguïsation 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 diagostic 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.

Gaël Lejeune et Béatrice Daille.

Augmentation d'index par propagation sur un réseau lexical - Application aux comptes rendus de radiologie

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.

Mathieu Lafourcade et Lionel Ramadier.

Détection automatique de l'ironie dans les tweets en français

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 de polarité explicites/implicites. Les résultats obtenus sont encourageants.

Jihen Karoui, Farah Benamara Zitoune, Véronique Moriceau, Nathalie Aussenac-Gilles et Lamia Hadrich Belguith.

Création d'un nouveau treebank à partir de quatrièmes de couverture

Nous présentons ici 4-couv, un nouveau corpus arboré d'environ 3500 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.

Philippe Blache, Grégoire Moncheuil, Stéphane Rauzy et Marie-Laure Guénot.

POSTERS RÉCITAL

● *1er étage*

Résumé Automatique Multi-Document Dynamique : État de l'art

Les travaux menés dans le cadre du résumé automatique de texte ont montré des résultats à la fois très encourageants mais qui sont toujours à améliorer. La problématique du résumé automatique ne cesse d'évoluer avec les nouveaux champs d'application qui s'imposent, ce qui augmente les contraintes liées à cette tâche. Nous nous intéressons au résumé extractif multi-document dynamique. Pour cela, nous examinons les différentes approches existantes en mettant l'accent sur les travaux les plus récents. Nous montrons ensuite que la performance des systèmes de résumé multi-document et dynamique est encore modeste. Trois contraintes supplémentaires sont ajoutées : la redondance inter-document, la redondance à travers le temps et la grande taille des données à traiter. Nous essayons de déceler les insuffisances des systèmes existants afin de bien définir notre problématique et guider ainsi nos prochains travaux.

Maâli Mnasri

Alignement multimodal de ressources éducatives et scientifiques

Cet article présente certaines questions de recherche liées au projet (anonymisé). L'ambition de ce projet est de valoriser les ressources éducatives et académiques en exploitant au mieux les différents médias disponibles (vidéos de cours ou de présentations d'articles, manuels éducatifs, articles scientifiques, présentations, etc). Dans un premier temps, nous décrirons le problème d'utilisation jointe de ressources multimédias éducatives ou scientifiques pour ensuite introduire l'état de l'art dans les domaines concernés. Cela nous permettra de présenter quelques questions de recherche sur lesquelles porteront des études ultérieures. Enfin nous finirons en introduisant trois prototypes développés pour analyser ces questions.

Hugo Mougard

DÉMONSTRATIONS

● *Salle S3-162*

MEDITE : logiciel d'alignement de textes pour l'étude de la génétique textuelle

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.

Zied Sellami, Jean-Gabriel Ganascia et Mohamed Amine Boukhaled.

Phœbus : un Logiciel d'Extraction de Réutilisations dans des Textes Littéraires

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.

Mohamed Amine Boukhaled, Zied Sellami et Jean-Gabriel Ganascia.

YADTK : Une plateforme open-source à base de règles pour développer des systèmes de dialogue oral

YADTK est une plateforme de développement open-source pour construire et maintenir des systèmes de dialogue oral. En outre, elle permet de procéder à des tests unitaires, à des tests de non-régression, ainsi qu'à des analyses par lots d'énoncés. De part son caractère déclaratif et unifié, le modèle de représentation des connaissances permet un développement rapide (cycles de dev courts) et facilité (représentations sémantiques graphiques).

Jérôme Lehuen et Carole Lailler.

TermLis : un contexte d'information logique pour des ressources terminologiques.

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.

Annie Foret.

Etude de l'image de marque d'entités dans le cadre d'une plateforme de veille sur le Web social.

Le travail présenté ici concerne l'intégration à une plateforme de veille sur internet d'un ensemble 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.

Leila Khouas, Caroline Brun, Anne Peradotto, Jean-Valère Cossu, Julien Boyadjian et Julien Velcin.

Building a Bilingual Vietnamese-French Named Entity Annotated Corpus through Cross-Linguistic Projection

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.

Ngoc Tan Le et Fatiha Sadat.

> SESSION 2

13h30 - 14h00

INTRODUCTION POSTERS/DÉMONSTRATIONS

● ***Amphithéâtre S3-057***

14h00 - 15h30

POSTERS

● ***1er étage***

Dictionnaires morphologiques du français contemporain : présentation de Morfetik, éléments d'un modèle pour le TAL

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 quelques éléments d'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.

Michel Mathieu-Colas, Emmanuel Cartier et Aude Grezka.

Une métagrammaire de la morphologie verbale de l'arabe

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.

Simon Petitjean, Younes Samih et Timm Lichte.

Entre écrit et oral ? Analyse comparée de conversations de type tchat et de conversations téléphoniques dans un centre de contact client

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.

Géraldine Damnati, Aleksandra Guerraz et Delphine Charlet.

Construction et maintenance d'une ressource lexicale basées sur l'usage

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.

Laurie Planes.

Utilisation d'annotations sémantiques pour la validation automatique d'hypothèses dans des conversations téléphoniques

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.

Carole Lailler, Yannick Estève, Renato De Mori, Mohamed Bouallègue et Mohamed Morchid.

Etiquetage morpho-syntaxique en domaine de spécialité : le domaine médical

L'étiquetage morpho-syntaxique est une tache 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 entrainé 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.

Christelle Rabary, Thomas Lavergne et Aurélie Névéol.

Vers une typologie de liens entre contenus journalistiques

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.

Remi Bois, Guillaume Gravier, Emmanuel Morin et Pascale Sébillot.

CDGFr, un corpus en dépendances non-projectives pour le français

Dans le cadre de l'analyse en dépendances du français, le phénomène de la non-projectivité est peu pris en compte, en majeur 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.

Denis Béchet et Ophélie Lacroix.

Utilisation des réseaux de neurones récurrents pour la projection interlingue d'étiquettes morpho-syntaxiques à partir d'un corpus parallèle

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. 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.

Othman Zennaki, Nasredine Semmar et Laurent Besacier.

Segmentation et titrage automatique de journaux télévisés

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.

Abdessalam Bouchekif, Géraldine Damnati, Nathalie Camelin, Yannick Estève et Delphine Charlet.

Un système hybride pour l'analyse de sentiments associés aux aspects

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.

Caroline Brun, Diana Nicoleta Popa et Claude Roux.

La ressource EXPLICADIS, un corpus annoté spécifiquement pour l'étude des relations de discours causales

Dans le but de proposer une caractérisation des relations de discours liées à la causalité, nous avons été amenés 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.

Caroline Atallah.

La séparation des composantes lexicale et flexionnelle des vecteurs de mots

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 compositionalité morphologique des mots-formes. Nous proposons une méthode pour décomposer les vecteurs de mots en vecteurs lexicaux et flexionnels.

François Lareau, Gabriel Bernier-Colborne et Patrick Drouin.

Traitements pour l'analyse du français préclassique

La période « préclassique » du français s'étend sur tout le XVIe siècle et la première moitié du XVIIe 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.

Sascha Diwersy, Achille Falaise, Marie-Hélène Lay et Gilles Souvay.

Classification de texte enrichie à l'aide de motifs séquentiels

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 delta-libres, comme descripteurs. Nous montrons les avantages et inconvénients de chaque type de motif.

Pierre Holat, Nadi Tomeh et Thierry Charnois.

Le traitement des collocations en génération de texte multilingue

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.

Florie Lambrey et François Lareau.

Médicaments qui soignent, médicaments qui rendent malades : étude des relations causales pour identifier les effets secondaires

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 mettons en évidence 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.

François Morlane-Hondère, Cyril Grouin, Véronique Moriceau et Pierre Zweigenbaum.

Exploration de modèles distributionnels au moyen de graphes 1-PPV

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.

Gabriel Bernier-Colborne.

Apport de l'information temporelle des contextes pour la représentation vectorielle continue des mots

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.

Killian Janod, Mohamed Morchid, Richard Dufour et Georges Linares.

Etiquetage morpho-syntaxique de tweets avec des CRF

Nous nous intéressons dans cet article à l'apprentissage automatique d'un étiqueteur mopho-syntaxique pour les tweets en anglais. Nous proposons tout d'abord un jeu d'étiquettes réduit, qui permet d'obtenir de meilleures performances par rapport au jeu d'étiquettes traditionnel. Comme nous disposons de peu de tweets étiquetées, nous essayons ensuite de compenser ce handicap en ajoutant des données issues de textes bien formés dans l'ensemble d'apprentissage. Les modèles mixtes obtenus permettent d'améliorer légèrement les résultats, au prix d'un temps d'apprentissage plus long.

Tian Tian, Dinarelli Marco, Tellier Isabelle et Cardoso Pedro.

Caractériser les discours académiques et de vulgarisation : quelles propriétés ?

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 différentes (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.

Amalia Todirascu et Beatriz Sanchez Cardenas.

Extraction et analyse automatique des comparaisons et des pseudo-comparaisons pour la détection des comparaisons figuratives

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.

Suzanne Mpouli et Jean-Gabriel Ganascia.

Proposition méthodologique pour la détection automatique de Community Manager. Étude multilingue sur un corpus relatif à la Junk Food

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.

Johan Ferguth, Aurélie Jouannet, Asma Zamiti, Damien Nouvel, Mathieu Valette et Yunhe Wu.

POSTERS RÉCITAL

● *Amphithéâtre S3-057*

État de l'art : l'analyse du dialogue appliquée aux conversations écrites en ligne porteuses de demandes d'assistance

Le développement du Web 2.0 et le processus de création et de consommation massive de contenus générés par les utilisateurs qu'elle a enclenché a permis le développement de nouveaux types d'interactions chez les internautes. En particulier, nous nous intéressons au développement du support en ligne et des plate-formes d'entraide. En effet, les archives de conversations en ligne porteuses de demandes d'assistance représentent une ressource inestimable, mais peu exploitée. L'exploitation de cette ressource permettrait non seulement d'améliorer les systèmes liés à la résolution collaborative des problèmes, mais également de perfectionner les canaux de support client proposés par les entreprises opérant sur le web. Pour ce faire, il est cependant nécessaire de définir un cadre formel pour l'analyse discursive de ce type de conversations. Cet article a pour objectif de présenter l'état de la recherche en analyse des conversations écrites en ligne, sous différents médiums, et de montrer dans quelle mesure les différentes méthodes exposées dans la littérature peuvent être appliquées à des conversations fonctionnelles inscrites dans le cadre de la résolution collaborative des problèmes utilisateurs.

Soufian Salim

DÉMONSTRATIONS

● *Salle S3-162*

Recherche de motifs de graphe en ligne

Nous présentons un outil en ligne de recherche de graphes dans des corpus annotés en syntaxe.

Bruno Guillaume.

Un patient virtuel dialogant

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.

Leonardo Campillos, Dhouha Bouamor, Éric Bilinski, Anne-Laure Ligozat, Pierre Zweigenbaum and Sophie Rosset.

Intégration du corpus des actes de TALN à la plateforme ScienQuest

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.

Achille Falaise.

Une aide à la communication par pictogrammes avec prédiction sémantique

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.

Aurélie Merlo.

Un système expert fondé sur une analyse sémantique pour l'identification de menaces d'ordre biologique

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.

Cédric Lopez, Aleksandra Ponomareva, Cécile Robin, André Bittar, Xabier Larrucea, Frédérique Segond et Marie-Hélène Metzger

DisMo : Un annotateur multi-niveaux pour les corpus oraux

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 disfluence et des marqueurs de discours.

Giulia Barreca.

> ENTREPRISES

EPTICA

- *95b, rue de Bellevue*
  > 92100 Boulogne-Billancourt

Domaines : Support client, Relation client, Voix du client, TALN

Le projet ODISAE a pour objectif de réaliser un analyseur sémantique de conversations en ligne entre agents et clients, afin d'enrichir les systèmes de gestion de la relation client de fonctionnalités jusqu'à présent non disponibles sur le marché. Le projet ODISAE offre la possibilité, aux clients d'Eptica, de contribuer à une nouvelle génération d'outils de Gestion de la Relation Client. Le projet réunit autour d'Eptica plusieurs partenaires industriels et universitaires : Jamespot pour les réseaux sociaux d'entreprise, Kwaga pour la gestion des contacts emails ou web pour récupérer les données et créer des fiches de contact correspondantes, Cantoche pour les avatars conversationnels, TokyWoky pour la constitution de communautés, le LINA de l'Université de Nantes pour sa contribution en analyse de contenu, et l'APROGED pour son expertise dans l'évaluation industrielle. Les partenaires valideurs sont le Centre Départemental du Tourisme de l'Aube et l'INSEE. Créé en 2001, Eptica conçoit, édite et commercialise une gamme de solutions logicielles permettant aux sociétés de créer, développer et gérer en temps réel la relation avec leurs clients, fournisseurs et partenaires par internet. La plateforme est proposée en mode SaaS (pour 60% des clients) ou hébergée par lec lient. Eptica est le leader européen des solutions de relation client multi canal et multilingue : web self service, email management, gestion des SMS/ Fax/Courrier et des appels, chat, réseaux sociaux et gestion de la base de connaissance pour le service clients. En créant des synergies entre le site web et le service clients, Eptica permet de répondre rapidement aux questions et de maximiser ainsi chaque opportunité de vente. En 2012, Eptica a acquis la société Lingway, leader français de l'analyse sémantique multilingue. Ainsi, Eptica a pu enrichir sa gamme de produits de puissants composants linguistiques (moteur de recherche, analyse du sentiment, etc.) et s'adjoindre d'une équipe d'experts en Traitement Automatique des Langues (TALN). Le moteur ELS (Eptica Linguistic Services), issu de cette fusion, est l'un des piliers des versions actuelles de Eptica Server.

Contact : Hugues de Mazancourt
hugues.de-mazancourt@eptica.com
06.72.78.70.33
www.odisae.com

NOOPSIS

- *9, Longue Vue des Astronomes*
  > 14111 Louvigny

Noopsis propose des solutions logicielles à haute valeur ajoutée sur un coeur de métier très spécifique : l'analyse sémantique. Il s'agit principalement d'automatiser l'extraction et la recherche d'informations à partir de contenus textuels en langue naturelle (ce que l'on appelle « text-mining », « fouille de texte »). Nous nous adressons en priorité aux entreprises et organismes à qui nous proposons des applications d'intelligence économique, de veille, ou de collecte automatique d'informations de toutes natures. La société a été fondée en 2008 par un groupe de chercheurs et d'ingénieurs spécialisés en ingéniere des langues, en analyse de donneés, et en intelligence économique. Les associés fondateurs sont issus du GREYC (CNRS-Université de Caen), laboratoire largement reconnu dans la communauté scientifique internationale pour ses activités en traitement automatique des langues. Lauréate du concours national de création d'entreprises innovantes en 2007, la société poursuit son développement technologique avec le soutien de l'Agence Nationale de la Recherche (ANR), d'Oséo Innovation, de Synergia et de la Région Basse-Normandie.

Contact : Frédérik Bilhaut - contact@noopsis.fr
02 90 92 05 70 - http://noopsis.fr

SEMIOTIME

- *58, avenue Pierre Berthelot*
  > 14000 Caen

Domaines : Text Analytics & Veille stratégique internationale

SEMIOTIME est une startup du CNRS spécialisée dans la recherche et l'analyse de l'information. Son savoir-faire technologique réside dans une triple expertise : la collecte et l'analyse des données textuelles multilingues, la gestion des grands volumes de données, et la production de synthèses interprétables. Semiotime conçoit et développe des solutions logicielles de recherche, de surveillance et d'analyse de l'information pour les professionnels souhaitant sécuriser ou dynamiser leurs activités. La mission de Semiotime est de délivrer des informations fiables, interprétables, dans des secteurs d'activité ciblés, quelle que soit la langue des sources d'information.Le premier produit commercialisé par Semiotime est un système de veille média. La solution est reconnue pour la précision et la fiabilité des résultats qu'elle produit. Les informations sont captées en toute langue puis présentées au lecteur dans la langue de son choix et sous forme de synthèses graphiques interactives.

Contact : Emmanuel Giguet
semiotime@semiotime.fr - www.semiotime.fr

SUCCEED TOGETHER

- *60, bis rue de Rochechouart*
  > 75009 Paris

Domaines : Traitement sémantique de messages courts par analyse ultra-rapide

En partenariat avec plusieurs unités de recherche universitaire, Succeed Together exploite les possibilités nouvelles offertes par le traitement automatique du langage dans le cadre du développement de la performance collective. Il s'agit de :
> Recueillir les avis et idées de tous, facteur clé de leur engagement au travail aujourd'hui
> Les rendre exploitable immédiatement
Cette accélération des échanges entre la direction et la base de l'entreprise permet ainsi de combler le fossé qui se creuse naturellement entre la stratégie et sa mise en œuvre. C'est principalement lors des temps collectifs, qu'ils soient en face à face ou à distance, que les points de vue peuvent s'aligner, pour ainsi améliorer la performance collective.

Contact : Andrianiaina Herizo - herizo@succeed-together.eu
06 82 13 25 37 - www.succeed-together.eu

SYLLABS

- *26, rue Notre Dame de Nazareth*
  > 75003 Paris

Domaines : Classification / catégorisation, Extraction d'information et fouille de texte, Génération automatique de textes, Résumé, Recherche d'information, Web mining

Syllabs est une startup technologique spécialisée en sémantique appliquée au web. Nous proposons des solutions « Semantics as a Service », en particulier pour le e-commerce, le e-tourisme, les annuaires et les médias : faire venir les visiteurs, les faire revenir, les garder et les comprendre.

Nous avons 3 types de technologies :
> web mining : collecte d'information structurée ou non ;
> text mining : extraction d'information, catégorisation, clustering, tagging, etc. (solution robuste adaptée aux gros volumes - Big Data) ;
> production de contenus : génération de textes de qualité, très variables et avec un style adapté au client.

Contact :
hello@syllabs.com
01 55 28 67 34 - www.syllabs.fr

SYNAPSE DÉVELOPPEMENT

- *5, rue du Moulin Bayard*
  > 31000 Toulouse

Domaines : Traitement du langage, Analyse syntaxique, Analyse sémantique, Correction orthographique et grammaticale

Synapse Développement crée des applications de valorisation de contenus et d'analyse de données textuelles. Des entreprises comme Amazon et Microsoft embarquent nos technologies dans leurs produits, celles-ci étant donc utilisées aujourd'hui par des millions de personnes.

Synapse a été créée en 1994 et sa première application a été le logiciel Cordial, correcteur d'orthographe et grammaire de référence pour le français. Dès 1994, l'accent a été mis sur l'innovation et sur la qualité des produits. Le correcteur Cordial est aujourd'hui un outil qui surclasse par sa précision les outils de correction intégrés dans les suites bureautiques ou gestionnaires de courriels du marché.

Passionnés par la langue et souhaitant accompagner nos clients dans l'aide à la rédaction, nous avons édité de nombreuses ressources numériques tels que des dictionnaires de définitions et de synonymes. Ces ressources sont aujourd'hui les plus complètes pour le français et sont largement diffusées sous les formes les plus diverses : sites Internet, applications mobiles ou encore marque-pages numériques !

Par ailleurs, nous offrons des solutions aux entreprises qui souhaitent valoriser leurs contenus Internet ou analyser d'importants volumes de textes, issus des réseaux sociaux par exemple. Depuis 20 ans notre équipe R&D développe des algorithmes qui permettent à une machine de comprendre un texte. Nous sommes aujourd'hui capables de comprendre, résumer, classer, analyser la tonalité d'un contenu, l'indexer automatiquement et relier les documents entre eux, afin de présenter ces analyses de façon originale dans des scénarios de publication apportant une réelle plus-value au lecteur.

La société emploie aujourd'hui une dizaine de personnes.

Contact : Baptiste Chardon
baptiste.chardon@synapse-fr.com
06 75 81 48 78
www.synapse-developpement.fr

> PROJETS

ART-ADN

- *GREYC*
  > Université de Caen Basse-Normandie
  > Campus Côte de Nacre, Bd du Maréchal Juin
  > CS 14032 - 14032 CAEN cedex 5

Accès par Retour Tactilo-oral Aux Documents Numériques

Domaines : Informatique de l'image et du langage - Psychologie - Electronique - Accessibilité

Un non voyant ne pourra « pointer » directement une partie de l'écran comme un voyant le ferait avec les yeux ; et cela dès le premier regard sur la page, avant même l'accès à son contenu articulable (rendu généralement aux aveugles par une lecture ou une description produite par la combinaison d'une synthèse de parole et d'un retour braille ; et dans le cas des dispositifs nomades, la limitation est plus dure puisque l'ajout d'un terminal braille s'avère peu pratique). De plus il a été montré que la posture cognitive que permet de prendre ce « first glance » joue un rôle prépondérant dans l'efficacité de la compréhension et la mémorisation des informations. Nous nous proposons à travers ce projets d'évaluer une stratégie différente pour donner un accès tactile palliant cette difficulté lors de la navigation web des non-voyants : il s'agira de donner au non voyant une représentation tactile de la structure visuelle des interfaces présentés sur une tablette numérique. Cela grâce à des actionneurs placés sur la main non active que l'on commandera via le bluetooth : les informations en terme de pixels et de structure logique du document seront fournies par la tablette au survol des doigts puis traduites par les actionneurs sous la forme de vibrations localisées, à fréquences et amplitudes variables. Autrement dit, notre ambition est de remplacer la capacité d'exploration visuelle d'un individu, qui s'appuie sur la vibration lumineuse de l'écran, par une capacité d'exploration manuelle, qui s'appuie sur la vibration tactile des actionneurs.

Contact Fabrice Maurel
fabrice.maurel@unicaen.fr
https://art-adn.greyc.fr/

CABERNET

Compréhension Automatique de Textes Biomédicaux pour la Recherche Translationnelle

- *LIMSI-CNRS UPR*
  > 3251, rue John von Neuman
  > 91400 Orsay

Domaines : Extraction d'information, Traitement Automatique de la Langue Biomédicale, Dossier électronique Patient, Santé Publique

Dans le domaine biomédical, les informations cliniques et institutionnelles sont contenues dans le texte de publications scientifiques ou de dossiers patients et ne sont pas directement accessibles à des fins de traitement automatique. Pour pallier cela, des méthodes de Traitement Automatique de Langue Naturelle (TALN) ont été développées avec succès afin d'extraire des informations pertinentes des textes libres et de les convertir en représentations formelles exploitables par l'homme et par la machine. Ce projet propose une analyse qui va au dela de la simple extraction de concepts isolés en permettant d'inclure le contexte d'occurrence ainsi que les relations entre concepts. Par ailleurs, nombre de travaux antérieurs sont limités à l'analyse de textes du domaine biomédical rédigés en anglais. Ce projet participera au nécessaire développement de méthodes permettant d'analyser les dossiers électroniques patient en français afin d'en extraire des représentations formelles compatibles avec celles disponibles pour l'anglais.

Ce projet de recherche se donne pour objectif de :
1 Mettre à disposition de la communauté scientifique des ressources dans un domaine de spécialité (le domaine biomédical) en français
2 Étudier l'adaptation en domaine de spécialité d'outils développés pour la langue générale
3 Appliquer ces outils à l'analyse automatique de dossiers électroniques patient et à la detection de liens entre données cliniques et littérature

Ce projet innovant permettra une analyse fine du contenu des textes du domaine biomedical, et en particuliers les textes cliniques. Il repose sur des principes issus de la linguistique et sera guide par des applications en medicine personnalisée. Une approche globale des problématiques de TAL sous l'angle de l'adaptation permettra d'assurer la portabilité des méthodes utilisées à d'autres applications dans le domaine biomédical.

Contact : Aurélie Névéol
neveol@limsi.fr
01 69 85 80 10

CRISTAL

Contextes RIches en connaissanceS pour la TrAduction terminoLogique

- *LINA*
  > 2, rue de la Houssinière
  > BP 92208 - 44322 Nantes Cedex 03

Domaines : Multilinguisme, Corpus comparable, Domaines spécialisés, TAO

Il est primordial au sein de chaque entreprise de pouvoir échanger avec ses partenaires, clients, employés dans leur langue maternelle mais aussi de donner à ses employés la possibilité de communiquer aisément dans une langue étrangère, en particulier dans leur domaine d'expertise.

L'accès aux expressions, termes techniques et à leurs traductions, qui est indispensable à tout processus de communication, nécessite une « contextualisation » de ces derniers pour en assurer la compréhension. En effet, avoir accès à un terme ou à sa traduction ne suffit pas, encore faut-il être capable (1) de l'employer correctement et (2) d'en appréhender le sens exact. Cette contextualisation a donc lieu à deux niveaux : 1) dans les textes : l'utilisateur doit avoir accès à des informations concernant l'usage des termes. Pour cela, il faut pouvoir extraire automatiquement des contextes riches en connaissances linguistiques (CRCL) ; 2) dans le domaine conceptuel : l'utilisateur doit avoir accès aux relations sémantiques ou conceptuelles entre termes afin de mieux en saisir le sens. Pour cela, il faut pouvoir extraire automatiquement des contextes riches en connaissances conceptuelles (CRCC).

La technologie d'extraction de contextes que nous abordons dans le cadre du projet CRISTAL vise la conception de nouveaux dictionnaires qui présenteront, pour chaque terme et ses traductions éventuelles, une fiche terminologique listant ses contextes et explicitant les connaissances qu'ils contiennent.

Les perspectives attendues sont, d'un point de vue scientifique, la mise en œuvre d'une nouvelle génération d'outils d'aide à la traduction et à la gestion terminologique, et d'un point de vue commercial, l'amélioration de la communication multilingue des entreprises en termes de qualité, rapidité et confort.

Contact : Emmanuel Morin
emmanuel.morin@univ-nantes.fr
02.51.12.58.39

LIMAH

Labex CominLabs

- *IRISA*
  > Campus de Beaulieu
  > 35042 Rennes Cedex

Domaines : traitement automatique des langues, psychologie, ergonomie, sociologie, droit

Si le volume et la diversité des contenus multimédias disponibles augmentent rapidement, aujourd'hui encore les données multimédias sont pour la plupart non reliées, c'est-à-dire sans liens explicites entre fragments en relation. Le projet Linking Media in Acceptable Hypergraphs (LIMAH) vise à explorer les structures d'hypergraphes pour les collections multimédias, créant des liens entre des fragments de documents multimédias, où la notion de lien reflète une proximité fondée sur le contenu — proximité thématique, opinion exprimée, réponse à une question, etc. Exploitant et développant des techniques de comparaison de paires de contenus multimédias, LIMAH aborde deux questions-clés de la structuration par graphes de collections multimédias : comment construire automatiquement, à partir d'une collection de documents, un hypergraphe (c'est-à-dire un graphe combinant des arcs de différentes natures) qui fournisse des liens exploitables pour des cas d'usage particuliers ? Comment des collections avec des liens explicites modifient les usages des données multimédias, que ce soit d'un point de vue technologique ou d'un point de vue utilisateur ? Le projet LIMAH étudie la création d'hypergraphes et leur acceptabilité dans deux cas d'usage distincts et complémentaires, à savoir la navigation dans des données d'actualité et l'apprentissage fondé sur des cours en ligne.

Contact : Pascale Sébillot
pascale.sebillot@irisa.fr
06 13 21 60 61

SENSEI

● *University of Trento*

Giuseppe Riccardi, Department of Information Engineering and Computer Science

La notion de conversation entre des usagers ou des clients et des entitées commerciales ou administratives a récemment connue une très forte mutation tout en restant le paradigme principal de contact. Chaque jour des millons de conversations téléphoniques sont gérées par des centres d'appels, alors que dans les plateformes collaboratives sur le WEB et les réseaux sociaux des millions d'échanges ont lieu entre utilisateurs de services. Le traitement de vastes collections de ces deux formes de conversation, orales dans le cadre des centres d'appels et écrites pour le WEB, posent de nouveaux défi au Traitement Automatique de la Langue : pouvons-nous tirer partie de la masse de données disponibles pour extraire de nouvelles connaissances qui pourront être utiles aux responsables des services faisant l'objet de ces conversations ? et plus généralement à toute personne interessé par ces contenus interactifs ?

Dans ce cadre, le projet SENSEI poursuit deux objectifs : premièrement développer des modèles et des méthodes permettant l'analyse de ces corpus de conversation afin d'aider les utilisateurs à extraire de la connaissance à partir de ces diverses sources ; deuxièmement développer et évaluer des prototypes de systèmes d'analyse, de production de rapport et de résumé automatique dans des cadres réalistes, avec des utilisateurs finaux permettant de prouver l'intéret des méthodes développées sur des tâches existantes.

Contexte Applicatif :
Deux contextes applicatifs ont été sélectionnés en priorité dans le cadre de SENSEI : les centres d'appels téléphoniques et les média sociaux centrés autour de la presse en ligne (commentaires sur des articles publiés, débats sur des réseaux sociaux à propos de sujets d'actualités).

Dans le cadre des centres d'appels, les utilisateurs finaux sont les professionnels en charge du contrôle-qualité des conversations ; pour les média sociaux il s'agit des journalistes, éditeurs de presse en ligne, analyste du WEB ou « blogger » désireux d'explorer les débats autour d'un article ou d'un sujet sociétal.

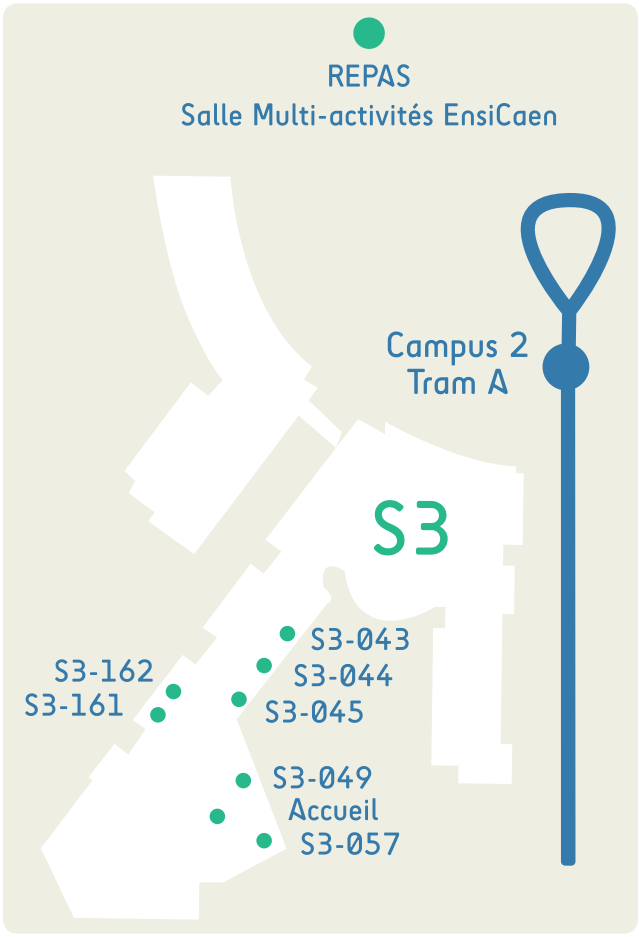
Objectifs et résultats attendus :
Les objectifs scientifiques et technologiques du projet SENSEI sont de développer de nouvelles méthodes d'accès au sens de toutes ces collections de conversations à travers les aspects suivants :
> Analyse des conversations, à la fois sur leur contenu sémantique, mais aussi sur les dimensions dialogiques et comportementales des participants
> Développement de méthodes permettant d'adapter les modèles d'analyse rapidement à la diversité des contenus et des medias véhiculant ces nouveaux types de conversation.
> Génération de rapport et de résumé permettant de présenter à un utilisateur sous une forme synthétique une collection de conversations entre deux ou plusieurs participants.
> Evaluation "écologique" des technologies développées en concertation avec les utilisateurs finaux identifiés dans les différents cadres d'études définis dans SENSEI.

Impact :
Le projet SENSEI a pour but d'avancer l'état de l'art dans le domaine de la compréhension automatique de conversations, orales ou écrites, entre humains. Dépassant le paradigme du simple "sac de mots" ou "sac de concepts" utilisé habituellement par les méthodes actuelles de fouille de données dans ce type de corpus, le projet SENSEI a pour but de montrer qu'une prise en compte plus profonde de l'interactivité inhérente à ce type de données peut permettre de dépasser les performances actuelles des systèmes "état de l'art".

Etant donné les différents cadre d'étude et utilisateurs finaux identifiés, SENSEI aura un impact dans plusieurs secteurs d'activités, tel que les centres d'appels, les organes de presse et plus généralement tout acteur s'intéressant à la fouille de données provenant de réseaux sociaux.

Contact en France : Benoît Favre & Frédéric Béchet
Université d'Aix Marseille / LIF-CNRS
benoit.favre@lif.univ-mrs.fr
frederic.bechet@lif.univ-mrs.fr

INFOS PRATIQUES



PROGRAMME ACTES TALN ACTES RÉCITAL

TALN2015.GREYC.FR • FACEBOOK.COM/TALN.GREYCCAEN • #TALN2015

TALN
CAEN 2015
& RÉCITAL