Résumés de thèses

Rubrique préparée par Fiammetta Namer

Université Nancy2 de Nancy, UMR « ATILF »
Fiammetta.Namer@univ-nancy2.fr

Florian BOUDIN (boudjinn@gmail.com)

Titre : Exploration d'approches statistiques pour le résumé automatique de texte

Mots-clés : traitement automatique du langage naturel, résumé automatique, méthodes statistiques, chimie organique, *maximal marginal relevance*, *document understanding conference*, *text analysis conference*.

Title : *Exploring Statistical Approaches for Automatic Text Summarization .*

Keywords : *natural language processing, text summarization, statistical methods, organic chemistry, Maximal Marginal Relevance, document understanding conference, text analysis conference.*

Thèse de doctorat en Informatique, Université d'Avignon et des Pays de Vaucluse, École doctorale 380 « Sciences et Agronomie », Laboratoire Informatique d'Avignon, Avignon, sous la direction de M. Juan-Manuel Torres Moreno, (MdC HDR) et M. Marc El-Bèze, (professeur). Soutenue le 05/12/2008.

Jury : M Jean-Manuel Torres(MC-HDR, Université d'Avignon, codirecteur), M. Marc El Bèze (Pr, Université d'Avignon, codirecteur), M. Thierry Poibeau (DR-HDR, LIPN, président), M. Guy Lapalme (Pr, RALI-Montréal, rapporteur), M. Horacio Saggion (Research Fellow, NLPG, Sheffield, rapporteur), M. Patrick Gallinari (Pr, LIP6, examinateur).

Résumé *: Un résumé est un texte reformulé dans un espace plus réduit. Il doit exprimer avec un minimum de mots le contenu essentiel d'un document. Son but est d'aider le lecteur à repérer les informations qui peuvent l'intéresser sans pour autant devoir lire le document en entier. Mais pourquoi avons-nous tant besoin de résumés ? Simplement parce que nous ne disposons pas d'assez de temps et d'énergie pour tout lire. La masse d'information textuelle sous forme électronique ne cesse d'augmenter, que ce soit sur Internet ou dans les réseaux des entreprises.*

Ce volume croissant de textes disponibles rend difficile l'accès à l'information désirée sans l'aide d'outils spécifiques. Produire un résumé est une tâche très complexe car elle nécessite des connaissances linguistiques ainsi que des connaissances du monde qui restent très difficiles à incorporer dans un système automatique. Dans cette thèse de doctorat, nous explorons la problématique du résumé automatique par le biais de trois méthodes statistiques permettant chacune la production de résumés répondant à une tâche différente.

Nous proposons une première approche pour la production de résumés dans le domaine spécialisé de la chimie organique. Un prototype nommé Yachs a été développé pour démontrer la viabilité de notre approche. Ce système est composé de deux modules, le premier applique un prétraitement linguistique particulier afin de tenir compte de la spécificité des documents de chimie organique tandis que le second sélectionne et assemble les phrases à partir de critères statistiques dont certains sont spécifiques au domaine. Nous proposons ensuite une approche répondant à la problématique du résumé automatique multidocument orienté par une thématique. Nous détaillons les adaptations apportées au système de résumés génériques Cortex ainsi que les résultats observés sur les données des campagnes d'évaluation DUC. Les résultats obtenus par la soumission du LIA, lors des participations aux campagnes d'évaluations DUC 2006 et DUC 2007, sont discutés. Nous proposons finalement deux méthodes pour la génération de résumés mis à jour. La première approche dite de maximisation-minimisation a été évaluée par une participation à la tâche pilote de DUC 2007. La seconde méthode est inspirée de Maximal Marginal Relevance *(MMR), elle a été évaluée par plusieurs soumissions lors de la campagne TAC 2008.*

URL où la thèse pourra être téléchargée :
http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/boudin/downloads/thesis_florian_boudin.pdf

---

Maud EHRMANN (Maud.Ehrmann@xrce.xerox.xom)

Titre : Les Entités Nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation

Mots-clés : définition des entités nommées, sens et référence en TAL, désambiguïsation, annotation fine, résolution de métonymie.

Title **:** *Named Entities, from linguistics to Natural Language Processing : theoretical status and disambiguation methods.*

Keywords **:** *named entities definition, sense and reference in computational linguistics, disambiguation, fine-grained annotation, metonymy resolution.*

Thèse de doctorat en Linguistique théorique, descriptive et automatique, Université de Paris 7 Denis Diderot, LaTTiCe, UFR Linguistique, sous la direction de Bernard Victorri, DR CNRS. Soutenue le 02/06/2008. Thèse réalisée dans le cadre d'une convention CIFRE, en partenariat avec le laboratoire LaTTiCe (Paris 7) et le Centre de Recherche Xerox à Grenoble (XRCE).

Jury : Mme Laurence Danlos (Pr, Université de Paris 7, présidente), M. Bernard Victorri (DR, LaTTiCe, directeur), Mme Adeline Nazarenko, (Pr, Université de Paris-Nord, rapporteur), M. Pierre Zweigenbaum (DR, LIMSI-CNRS, rapporteur), Mme Caroline BRUN (Xerox Research Centre Europe, examinatrice-tutrice), M. Marcel Cori (Pr, Université Paris 10, examinateur).

Résumé : *Le traitement des entités nommées fait aujourd'hui figure d'incontournable en traitement automatique des langues. Apparue au milieu des années 1990 à la faveur des dernières conférences MUC (*Message Understanding Conferences*), la tâche de reconnaissance et de catégorisation des noms de personnes, de lieux, d'organisations, etc. apparaît en effet comme fondamentale pour diverses applications participant de l'analyse de contenu et nombreux sont les travaux se consacrant à sa mise en œuvre, obtenant des résultats plus qu'honorables. Fort de ce succès, le traitement des entités nommées s'oriente désormais vers de nouvelles perspectives avec, entre autres, la désambiguïsation et une annotation enrichie de ces unités. Ces nouveaux défis rendent cependant d'autant plus cruciale la question du statut théorique des entités nommées, lequel n'a guère été discuté jusqu'à aujourd'hui.*

Deux axes de recherche ont par conséquent été investis durant ce travail de thèse: nous avons, d'une part, tenté de proposer une définition des entités nommées et, d'autre part, expérimenté des méthodes de désambiguïsation. À la suite d'un état des lieux de la tâche de reconnaissance de ces unités et d'un exposé des difficultés pouvant se présenter à l'occasion d'une telle entreprise, il fut avant tout nécessaire d'examiner, d'un point de vue méthodologique, comment aborder la question de la définition les entités nommées. La démarche adoptée invita à se tourner du côté de la linguistique, avec les noms propres et les descriptions définies, puis du côté du traitement automatique, ce parcours visant au final à proposer une définition tenant compte tant des aspects du langage que des capacités et exigences des systèmes informatiques. La suite du mémoire rend compte d'un travail davantage expérimental, avec l'exposé d'une méthode d'annotation fine tout d'abord, de résolution de métonymie enfin. Ces travaux, combinant approche symbolique et approche distributionnelle, rendent compte de la possibilité d'une double annotation (catégories générales et catégories fines) et d'une désambiguïsation des entités nommées.

Abstract : *Introduced as part of the last Message Understanding Conferences dedicated to information extraction, Named Entity extraction is a well-studied task in Natural Language Processing. The recognition and the categorization of person*

names, location names, organisation names, etc. is regarded as a fundamental process for a wide variety of natural language processing applications dealing with content analysis and many research works are devoted to it, achieving very good results. Following this success, named entity treatment is moving towards new research propects with, among others, disambiguation and fined-grained annotation. However, this new challenges make even more crucial the question of named entity definition, which was not much discussed until now.

Two main lines were explored during this PhD project : first we tried to propose a definition of named entities and then we experimented disambiguation methods. After a presentation and a state of the art of the named entity recognition task, we had to examine, from a methodological point of view, how to tackle the question of the definition of named entities. Our approach led us to study, firstly, the linguistic side, with proper names and definite descriptions and, secondly, the computing side, this development aiming at, finally, proposing a named entity definition that takes into account language aspects but also informatic systems capacities and requirements. The continuation of the dissertation is about more experimental works, with a presentation of experiments about fined-grained named entity annotation and metonymy resolution methods.

URL où la thèse pourra être téléchargée :
http://www.xrce.xerox.com/people/ehrmann/home.html

François LAREAU (francois.lareau@umontreal.ca**)**

Titre : Vers une grammaire d'unification Sens-Texte du français : le temps verbal dans l'interface sémantique-syntaxe

Mots-clés : temps grammatical, grammaire d'unification Sens-Texte, interface sémantique-syntaxe, flexion, verbes, français (langue)

Title : *Towards a Meaning-Text unification grammar of French : verbal tense in the semantics-syntax interface..*

Keywords : *tense, Meaning-Text unification grammar, semantics-syntax interface, inflection, verbs, french (language).*

Thèse de doctorat en linguistique, en cotutelle entre l'Université de Paris 7, LLF, et l'Université de Montréal, OLST, sous la direction de M. Sylvain Kahane, (Professeur, U. de Paris 10), et M. Igor Mel'čuk (Professeur, U de Montréal, Canada). Soutenue à Montréal le 25/11/2008.

Jury : M. Sylvain Kahane (Pr, Université Paris 10, codirecteur), M. Igor Mel'čuk (Pr, Université de Montréal, co-directeur), M. Alain Polguère (Pr., Université de

Montréal, président et rapporteur), M. Owen Ranbow (Pr, research scientist, Columbia University, rapporteur), Mme Laurence Danlos (Pr, Université de Paris 7, examinatrice), M. Antoine Soare (Pr, Université de Montréal, examinateur).

Résumé *: Cette thèse vise deux objectifs principaux. D'une part, nous cherchons à décrire la flexion verbale en français, et plus particulièrement celle en temps dans l'interface sémantique-syntaxe. D'autre part, nous voulons, à travers ce phénomène linguistique central, développer le formalisme de la grammaire d'unification Sens-Texte (GUST) et en vérifier l'utilité pour la description formelle des langues.*

Pour atteindre notre premier objectif, nous proposons une méthodologie, inspirée de la lexicographie, pour l'étude des signes grammaticaux. Cette méthodologie met l'accent sur les signes dans leur ensemble, et non sur une seule de leurs composantes (signifié, signifiant ou combinatoire).

Au niveau descriptif, notre thèse principale est que le français possède non pas une, mais bien deux catégories flexionnelles de temps complémentaires :

- la catégorie de décalage, qui regroupe deux grammèmes situant un point de repère par rapport au moment d'élocution :

- le grammème ***non décalé*** *signifie que le point de repère est « maintenant » ou un fait dans le futur. Son signifiant est nul,*

- le grammème ***décalé*** *indique que le repère est un fait dans le passé. Il s'exprime par le suffixe –ai– ;*

- la catégorie de temps comme tel, qui regroupe trois grammèmes situant les faits par rapport à ce point de repère :

- le grammème ***simultané*** *signifie que le fait dénoté par le verbe est simultané à son point de repère. Son signifiant est nul,*

- le grammème profond ***antérieur*** *indique que le fait est antérieur à son point de repère. Son signifiant est l'auxiliaire avoir (ou être). Dans un registre littéraire, il s'exprime conjointement avec le grammème* ***non décalé*** *par le suffixe du passé simple.*

Le grammème ***postérieur*** *pose le fait comme postérieur à son point de repère. Il s'exprime par le suffixe –r–.*

Au niveau formel, nos principaux apports au formalisme de GUST sont la représentation des structures en termes d'objets et de fonctions, ainsi que la modélisation des décompositions sémantiques. La formalisation de notre modèle de la flexion verbale s'avère un moyen intéressant de vérifier l'utilité de GUST puisqu'elle met en jeu des signes variés. Elle nous donne également l'occasion d'observer le mécanisme d'articulation des modules de la grammaire.

URL où la thèse pourra être téléchargée :
https://www.webdepot.umontreal.ca/Usagers/lareauf/these/lareau-these.pdf

---

Jean-Philippe PROST (JPProst@gmail.com**)**

Titre : Modélisation de la gradience syntaxique par analyse relâchée à base de contraintes

Mots-clés : gradience, acceptabilité, grammaticalité, optimalité, configuration, syntaxe modèle-théorique (*Model-Theoretic Syntax*), Grammaires de Propriétés, analyse syntaxique tabulaire par contraintes, robustesse, satisfaction relâchée de contraintes.

Title : *Modelling Syntactic Gradience with Loose Constraint-based Parsing.*

Keywords : *Gradience, acceptability, grammaticality, optimality, configuration, Model-Theoretic Syntax, Property Grammars, characterisation, constraint-based chart parsing, robustness, loose constraint satisfaction.*

Thèse de doctorat en Informatique, en cotutelle entre l'Université de Provence, UFR MIM, Laboratoire Parole et Langage, et Macquarie University, Australie, Department of Computing, Division of Information and Communication Sciences, Centre for language Technology, sous la direction de M. Philippe Blache, (DR, U. de Provence), M. Diego Mollà Aliod, (Senior Lecturer, Macquarie U., Australie) et M. Mark Dras (Senior Lecturer, Macquarie U., Australie). Soutenue le 10/12/2008.

Jury : M. Philippe Blache (DR, LPL, codirecteur), M. Diego Mollà Aliod (senior lecturer, Macquarie University, Australie, codirecteur), M. Alexis Nasr (Pr, Université de la Méditerranée, président), M. Denys Duchier (Pr, Université d'Orléans, rapporteur), M. Gerald Penn (Associate Professor, University of Toronto, rapporteur), M. Eric de la Clergerie (CR, INRIA, examinateur).

Résumé *: La grammaticalité d'une phrase est habituellement conçue comme une notion binaire : une phrase est soit grammaticale, soit agrammaticale. Cependant, bon nombre de travaux se penchent de plus en plus sur l'étude de degrés d'acceptabilité intermédiaires, auxquels le terme de gradience fait parfois référence. À ce jour, la majorité de ces travaux s'est concentrée sur l'étude de l'évaluation humaine de la gradience syntaxique. Cette étude explore la possibilité de construire un modèle robuste qui s'accorde avec ces jugements humains.*

Nous suggérons d'élargir au langage mal formé les concepts de Gradience Intersective et de Gradience Subsective, proposés par Aarts pour la modélisation de jugements graduels. Selon ce nouveau modèle, le problème que soulève la gradience concerne la classification d'un énoncé dans une catégorie particulière, selon des critères basés sur les caractéristiques syntaxiques de l'énoncé. Nous nous attachons à étendre la notion de Gradience Intersective (GI) afin qu'elle concerne le choix de la meilleure solution parmi un ensemble de candidats, et celle de Gradience Subsective (GS) pour qu'elle concerne le calcul du degré de typicité de cette structure au sein de sa catégorie. La GI est alors modélisée à l'aide d'un

*critère d'optimalité, tandis que la GS est modélisée par le calcul d'un degré d'acceptabilité grammaticale. Quant aux caractéristiques syntaxiques requises pour permettre de classer un énoncé, notre étude de différents cadres de représentation pour la syntaxe du langage naturel montre qu'elles peuvent aisément être représentées dans un cadre de syntaxe modèle-théorique (*Model-Theoretic Syntax*). Nous optons pour l'utilisation des Grammaires de Propriétés (GP), qui offrent, précisément, la possibilité de modéliser la caractérisation d'un énoncé. Nous présentons ici une solution entièrement automatisée pour la modélisation de la gradience syntaxique, qui procède de la caractérisation d'une phrase bien ou mal formée, de la génération d'un arbre syntaxique optimal, et du calcul d'un degré d'acceptabilité grammaticale pour l'énoncé.*

À travers le développement de ce nouveau modèle, la contribution de ce travail comporte trois volets.

Premièrement, nous spécifions un système logique pour les GP qui permet la révision de sa formalisation sous l'angle de la théorie des modèles. Il s'attache notamment à formaliser les mécanismes de satisfaction et de relâche de contraintes mis en œuvre dans les GP, ainsi que la façon dont ils permettent la projection d'une catégorie lors du processus d'analyse. Ce nouveau système introduit la notion de satisfaction relâchée, et une formulation en logique du premier ordre permettant de raisonner au sujet d'un énoncé.

*Deuxièmement, nous présentons notre implantation du processus d'analyse syntaxique relâchée à base de contraintes (*Loose Satisfaction Chart Parsing*, ou LSCP), dont nous prouvons qu'elle génère toujours une analyse syntaxique complète et optimale. Cette approche est basée sur une technique de programmation dynamique (*dynamic programming*), ainsi que sur les mécanismes décrits ci-dessus. Bien que d'une complexité élevée, cette solution algorithmique présente des performances suffisantes pour nous permettre d'expérimenter notre modèle de gradience.*

Et troisièmement, après avoir postulé que la prédiction de jugements humains d'acceptabilité peut se fonder sur des facteurs dérivés de la LSCP, nous présentons un modèle numérique pour l'estimation du degré d'acceptabilité grammaticale d'un énoncé. Nous mesurons une bonne corrélation de ces scores avec des jugements humains d'acceptabilité grammaticale. Qui plus est, il s'avère que notre modèle obtient de meilleures performances que celles d'un modèle préexistant que nous utilisons comme référence, et qui, quant à lui, a été expérimenté à l'aide d'analyses syntaxiques générées manuellement.

URL où la thèse pourra être téléchargée :
http://tel.archives-ouvertes.fr/tel-00352828/fr/