Résumés de thèses

Rubrique préparée par Fiammetta Namer

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

Rémi BOVE (Remi.Bove@univ-provence.fr)

Titre : Analyse syntaxique automatique de l'oral : étude des disfluences

Mots-clés : traitement automatique des langues, analyse syntaxique automatique, disfluences, oral, français parlé, étiquetage automatique, grammaire de *chunking*, linguistique de corpus.

Title : *Speech parsing : a study of disfluencies .*

Keywords : *natural language processing, robust parsing, disfluencies, speech, spoken french, tagging, chunking grammar, corpus linguistics.*

Thèse de doctorat en Sciences du Langage, mention Traitement Automatique des Langues, Université de Provence, Aix-Marseille I, Département de Lettres, Arts, Communication et Sciences du Langage, Laboratoire d'Informatique Fondamentale (LIF), UMR 6166 CNRS, Aix-en-Provence, sous la direction de Jean Veronis (professeur). Soutenue le 25/11/2008.

Jury : M. Jean Véronis (Pr, Université d'Aix en Provence, directeur), Mme Martine Adda-Decker (CR., LIMSI-CNRS, rapporteur), M. Jean-Yves Antoine (Pr, Université François Rabelais de Tours, rapporteur), M. Jacques Vergne (Pr, Université de Caen, examinateur), M. Henri-José Delofeu (Pr., Université de Provence, examinateur).

Résumé *: Ce travail de recherche porte sur l'étude des phénomènes de disfluences (répétitions, autocorrections, amorces, etc.) en français parlé spontané transcrit.*
À première vue, le terme « disfluence » semble évoquer une anormalité, pourtant ces phénomènes font partie des modes de production tout à fait normaux de l'oral. Le but de cette thèse est donc d'étudier de façon détaillée l'impact de ces phénomènes récurrents en français parlé sur l'analyse syntaxique automatique de l'oral, et de proposer un modèle théorique permettant de les intégrer dans cette analyse. Notre axe de recherche se fonde sur l'hypothèse selon laquelle une analyse

détaillée des énoncés oraux (principalement en termes morphosyntaxiques) peut permettre un traitement efficace pour ce type de données et s'avère incontournable dans une optique de développement d'applications génériques dans le domaine des technologies de la parole.

Dans le cadre de ce travail, nous présentons tout d'abord certaines particularités du français parlé, puis nous proposons une typologie détaillée des disfluences. Divers auteurs ont abordé l'étude de disfluences particulières en français, mais une vision d'ensemble, montrant notamment l'interaction des différents phénomènes, fait pour l'instant défaut. Cette typologie est basée sur un corpus qui sert à étudier de façon précise la distribution et le comportement syntaxique des différents types de disfluences. Nous positionnons ensuite notre approche et notre vision structurelle des disfluences, en présentant la banque de données de représentations arborescentes que nous avons créée à partir de notre modèle théorique.

La dernière partie est consacrée au développement d'un analyseur syntaxique partiel pour l'oral, essentiellement focalisé sur les disfluences. Notre analyse comporte trois phases principales. Tout d'abord une analyse morphosyntaxique est réalisée sur le corpus à l'aide d'un étiqueteur existant (Treetagger) que nous avons adapté au cas des transcriptions orales. Ensuite, une détection « brute » des disfluences est implémentée en préalable à la dernière étape qui consiste à regrouper les énoncés en syntagmes minimaux non récursifs (ou « chunks *»). Le corpus final est ainsi segmenté en* chunks *non disfluents d'une part, à côté des* chunks *disfluents d'autre part. Ces différentes étapes se justifient par la prise en compte explicite des régularités observées dans notre corpus.*
Les résultats de l'analyse automatique sont finalement évalués de façon quantitative sur le corpus permettant ainsi de valider le modèle théorique de façon empirique.

URL où la thèse pourra être téléchargée :
http://bove.remi.free.fr/These_RB.pdf

Bruno CARTONI (cartonib@gmail.com)

Titre : De l'incomplétude lexicale en traduction automatique : vers une approche morphosémantique multilingue

Mots-clés : mots inconnus, traduction automatique, morphologie constructionnelle, néologie, méthode contrastive.

Title : *lexical incompleteness in machine translation: a multilingual morphosemantic approach.*

Keywords : *Unknown words, machine translation, constructional morphology, neology, contrastive method.*

Thèse de doctorat en Traitement informatique multilingue, Université de Genève, École de traduction et d'interprétation, Département de Traitement informatique multilingue, (Genève, Suisse), sous la direction de Margaret King (professeur). Soutenue le 27/06/2008.

Jury : Mme Pierrette Bouillon (Pr, Université de Genève, présidente), Mme Margaret King (Pr, Université de Genève, directrice), Mme Fiammetta Namer (Pr, Université de Nancy2, examinatrice), M. Anthony Hartley (Pr, Université de Leeds, examinateur).

Résumé *: L'ambition de ce travail est d'évaluer la faisabilité d'une implémentation informatique de l'inférence entre les langues. Nous avons concentré notre attention sur les phénomènes d'inférence dans la construction des mots, que nous avons regroupés sous le terme de liens morphosémantiques multilingues, et sur la faisabilité de leur implémentation en traduction automatique. Ce travail se veut à la fois théorique et pratique. D'un point de vue théorique, il questionne les fondements de cette inférence et propose une première ébauche de modélisation. D'un point de vue pratique, il montre comment cette inférence pourrait être exploitée pour résoudre en partie un problème important : l'incomplétude lexicale en traduction automatique*

Toutes les applications de traitement de la langue basées sur les lexiques dépendent de la richesse de cette ressource. Un mot absent du lexique ne peut en effet pas être traité par le système, ce qui a des conséquences plus ou moins dommageables sur la qualité de la sortie. Suivant les applications, de nombreuses solutions ont été envisagées pour pallier cette incomplétude lexicale et deviner l'inconnu. Dans un système de traduction automatique, où l'on passe d'une langue à l'autre, deviner l'inconnu est une tâche très complexe, qui recouvre une étape d'analyse du mot inconnu et une étape de génération de la traduction de ce mot.

Les mots inconnus des systèmes de traduction automatique sont de différentes sortes (noms propres, mots issus de la créativité lexicale, mots erronés), mais ce sont les mots issus de la créativité lexicale qui nous intéressent dans ce travail. Ces mots constituent un ensemble dynamique : certains vont un jour entrer dans le lexique, d'autres n'existeront que dans le temps de leur production. L'exploitation des liens morphosémantiques multilingues en traduction automatique a donc pour but, in fine, de proposer une traduction pour les mots construits néologiques, sans devoir forcément les enregistrer dans le lexique. D'un point de vue pratique, nous nous sommes volontairement concentré sur un procédé de construction (la préfixation) et sur deux langues (l'italien et le français). Il n'en reste pas moins, que les méthodes et les solutions proposées sont applicables à d'autres procédés de formation néologique et à d'autres paires de langues.

Dans un premier temps, cette recherche présente différentes études sur l'incomplétude lexicale dans différents systèmes de traduction automatique et dans

d'autres lexiques d'applications informatiques de traitement de la langue. Ces études ont montré que ce phénomène était constant et que la solution à l'incomplétude lexicale ne pouvait résider dans une simple alimentation du lexique. Par ailleurs, l'analyse qualitative de ce phénomène a souligné la présence d'un nombre important de néologismes formés selon des procédés réguliers. Ces néologismes construits sont en outre influencés par le contact entre les langues, ce qui permet d'envisager un certain parallélisme entre les constructions néologiques et donc d'imaginer une traduction automatique des néologismes.

Dans un deuxième temps, nous définissons plus précisément la notion de lien morphosémantique multilingue, qui permet de rendre compte des similitudes de construction entre deux langues. Ce lien est défini selon une double reproductibilité, à la fois au sein d'une même langue et entre les langues. Pour être exploités dans la traduction automatique des néologismes construits, ces liens sont formalisés par l'intermédiaire de règles de construction des lexèmes (RCL) bilingues, en adoptant l'approche lexématique de la morphologie, qui dispose d'outils descriptifs idéaux pour le traitement de la néologie. L'élaboration de ces RCL passe nécessairement par une étude approfondie des systèmes morphologiques des deux langues et une étude contrastive des procédés de construction. Cette démarche contrastive se fonde sur l'utilisation d'un tertium comparationis, *qui joue le rôle d'un point de comparaison sur lequel nous pouvons projeter les éléments des deux langues. Cette projection nous a fourni le matériel traductionnel permettant d'implémenter les règles de construction des lexèmes bilingues. Elle a également permis, dans les étapes d'affinage, de rendre compte des divergences structurelles présentes dans les règles de préfixation des différentes langues.*

La troisième partie de ce travail porte sur l'implémentation informatique de ces RCL bilingues dans le contexte de la traduction automatique des mots construits. Pour ce faire, nous avons mis au point un prototype de traducteur automatique, permettant de traduire des néologismes préfixés. Ce prototype nous a permis d'expérimenter pas à pas les étapes de la traduction automatique, en évaluant chaque principe et chaque contrainte implémentés. Nous montrons que le défi principal réside dans la partie « analyse » des mots inconnus, étape sur laquelle nous avons concentré nos efforts pour implémenter des mécanismes de contrainte permettant d'assurer une correction optimale de cette analyse. La partie génération, pour sa part, requiert avant tout un lexique bilingue approprié pour la traduction automatique des néologismes construits. Mais la génération morphologique est également confrontée à un certain nombre de problématiques inhérentes à la préfixation, à savoir l'alternance entre préfixes (multidimensionnel ou pluridimensionnel) et l'alternance entre bases (anticancer ou anticancéreux).

Enfin, dans la quatrième partie, nous avons évalué notre démarche, d'une part, sous l'angle de la qualité de la traduction des néologismes construits et de l'influence de leur résolution sur la qualité de la phrase et, d'autre part, d'un point de vue plus global, en posant des questions de faisabilité et de portabilité de notre approche. Nous avons pu ainsi souligner que les fondements théoriques forts, les contraintes adéquates et des ressources appropriées étaient les conditions essentielles à l'exploitation des liens morphosémantiques multilingues en traduction automatique.

URL où la thèse pourra être téléchargée :
http://www.issco.unige.ch/staff/bruno/recherche.html

Mehdi EMBAREK (**embarekm@gmail.com)**

Titre : Un système de question-réponse dans le domaine médical : Le système Esculape

Mots-clés : systèmes de question-réponse, domaine médical, ontologie, patrons linguistiques.

Title : *A question answering system in the medical domain : The Esculape system.*

Keywords : *question-answering systems, medical domain, ontology, linguistic patterns.*

Thèse de doctorat en Informatique, LIC2M, INSTN, LIST-CEA, Saclay, sous la direction de Christian Fluhr (professeur). Soutenue le 04/07/2008.

Jury : M. Pierre Zweigenbaum (DR, LIMSI-CNRS, président et rapporteur), M. Christian Fluhr (Pr, LIST-CEA, directeur), Mme Brigitte Grau (Pr, Université d'Evry, rapporteur), M. Patrice Bellot (MC, Université d'Avignon, examinateur), M. Olivier Ferret (IC, LIST-CEA, examinateur).

Résumé *: Le domaine médical dispose aujourd'hui d'un très grand volume de documents électroniques permettant ainsi la recherche d'une information médicale quelconque. Cependant, l'exploitation de cette grande quantité de données rend la recherche d'une information précise complexe et coûteuse en termes de temps. Cette difficulté a motivé le développement de nouveaux outils de recherche adaptés, comme les systèmes de question-réponse. En effet, ce type de système permet à un utilisateur de poser une question en langage naturel et de retourner une réponse précise à sa requête au lieu d'un ensemble de documents jugés pertinents, comme c'est le cas des moteurs de recherche. Les questions soumises à un système de question-réponse portent généralement sur un type d'objet ou sur une relation entre objets. Dans le cas d'une question telle que « Qui a découvert l'Amérique ? » par exemple, l'objet de la question est une personne. Dans des domaines plus spécifiques, tels que le domaine médical, les types rencontrés sont eux-mêmes plus spécifiques. La question « Comment rechercher l'hématurie ? » appelle ainsi une réponse de type examen médical.*

L'objectif de ce travail est de mettre en place un système de question-réponse pour des médecins généralistes portant sur les bonnes pratiques médicales. Ce système permettra au médecin de consulter une base de connaissances lorsqu'il se trouve en

consultation avec un patient. Ainsi, dans ce travail, nous présentons une stratégie de recherche adaptée au domaine médical. Plus précisément, nous exposerons une méthode pour l'analyse des questions médicales et l'approche adoptée pour trouver une réponse à une question posée. Cette approche consiste à rechercher en premier lieu une réponse dans une ontologie médicale construite à partir de ressources sémantiques disponibles pour la spécialité. Si la réponse n'est pas trouvée, le système applique des patrons linguistiques appris automatiquement pour repérer la réponse recherchée dans une collection de documents candidats. L'intérêt de notre approche a été illustré au travers du système de question-réponse « Esculape » qui a fait l'objet d'une évaluation montrant que la prise en compte explicite de connaissances médicales permet d'améliorer les résultats des différents modules du processus de traitement.

URL où la thèse pourra être téléchargée :
http://www.univ-mlv.fr/fr/index.php?rub=recherche&srub=actupub&actu=thesehdr&numactu=475

Tonio WANDMACHER (tonio.wandmacher@univ-tours.fr)

Titre : Prédiction de mots adaptative pour l'aide à la communication pour personnes handicapées

Mots-clés : prédiction de mots, modélisation du langage, modélisation de l'utilisateur, communication assistée, Analyse Sémantique Latente.

Title : *Adaptive word prediction and its application in an assistive communication system.*

Keywords : *word prediction, language modeling, user modeling, augmentative and alternative communication, Latent Semantic Analysis.*

Thèse de doctorat en Informatique ; co-tutelle Université de Tübingen/Université François Rabelais de Tours, BdTln, Département d'Informatique, Tours, sous la direction de Jean-Yves Antoine (professeur) et Uwe Mönnig (professeur). Soutenue le 22/10/2008.

Jury : M. Denis Maurel (Pr, Université de Tours, président), M. Jean-Yves Antoine (Pr, Université de Tours, codirecteur), M Uwe Mönnig (Pr, Université de Tübingen, codirecteur), M. Frédéric Béchet (MC-HDR, Université d'Avignon, rapporteur), Mme Béatrice Daille (Pr, Université de Nantes, rapporteur), M. Detmar Meurers (Pr, Université de Tübingen, rapporteur), M. Stefan Langer (Privatdozent (HDR), Université de Tübingen, rapporteur).

***Abstract**: This work investigates the capacities of adaptive methods for word prediction, which represents a central strategy in the context of alternative and augmentative communication (AAC) for speech- and motion-impaired persons (due to e.g. cerebral palsy, brain traumatisms, locked-in syndrome).*

First an introduction to the respective research fields of AAC and word prediction is given. Here we address in particular stochastic (n-gram) language models, which have been shown to be very appropriate for the task of prediction. We then explore two major classes of adaptive methods: In a first part we consider strategies enabling to adapt to the lexical and syntactic preferences of the user of an AAC system. Here we investigate the recency promotion or cache model, an auto-adaptive user lexicon automatically adding unknown words and the dynamic user model (DUM), which integrates every input of the user. In order to assess the adaptive capacities of the different approaches we apply cross-register evaluation, i.e. we use test corpora from several registers (newspaper, literature, transcribed speech and e-mail); moreover we test all models on three languages: French, German and English.
The results for the cache model and the user lexicon are quite moderate; however we find important gains for the dynamic user model (up to +9.4% in ksr). In addition we could show that this model adapts rather quickly: After integrating 2000 words only the model performs significantly better than the baseline.

The second class of adaptation methods aims to exploit to the semantic or topical context. After presenting a number of approaches (such as trigger or topic models) we focus in particular on Latent Semantic Analysis (LSA), a vectorial model establishing semantic similarity from distributional properties of lexical units in large corpora. A difficult aspect however arises from the integration of LSA-based information to the general prediction model; for this reason we discuss and evaluate here several integration methods, and we propose a new confidence scheme, controlling the influence of the LSA component with respect to the word to predict. Moreover we thoroughly test several parameters which are decisive for an LSA-based semantic space, such as dimensionality or the size of the co-occurrence window.
The semantic adaptation by LSA was assessed in the same way as before (cross-register evaluation). Here, the advantages are moderate (up to +1.7% in ksr) but rather stable over the different registers and languages. We could also show that the gains achieved by the dynamic user model and the LSA component are nearly additive, when both are active.

In the last part of this work we present Sibylle, an assistive communication system that implements the most successful of the previously investigated adaptation methods (DUM and LSA). After a description of the user interface as well as the communication enhancing components we report results from the application of this system in a rehabilitation center, where it has been in use for several years. One of

the most important results from this application was to see how crucial the configurability of an AAC system is: The clinical patterns of AAC users vary strongly, and likewise their individual preferences and needs. For this reason we aimed to make every interactive component as configurable as possible. This makes Sibylle a highly adaptive and adaptable AAC system, integrating an important variety of communication-enhancing functions.

URL où la thèse pourra être téléchargée :
Contacter l'auteur