Résumés de thèses

Rubrique préparée par Fiammetta Namer

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

Marianna APIDIANAKI (marianna@linguist.jussieu.fr)

Titre : Acquisition automatique de sens pour la désambiguïsation et la sélection lexicale en traduction

Mots-clés : désambiguïsation lexicale, induction de sens, apprentissage non supervisé, clustering, prédiction de traduction.

Title : *Automatic sense acquisition for word sense disambiguation and lexical selection in translation.*

Keywords : *Word Sense Disambiguation, sense induction, unsupervised learning, clustering, translation prediction.*

Thèse de doctorat en Linguistique théorique, descriptive et automatique (Sciences du langage), Université de Paris 7 Denis Diderot, LaTTiCe, UFR Linguistique, sous la direction de Catherine FUCHS (DR CNRS). Soutenue le 05/09/2008.

Jury : M. Eric Gaussier (Pr, Université de Grenoble, président), Mme Catherine Fuchs (DR, CNRS-LaTTiCe, directrice), M. Philippe Langlais, (Pr, Université de Montréal, rapporteur), M. Daniel Kayser (Pr, Université de Paris 13, rapporteur), M. Helge Dyvik (Pr, Université de Bergen, Norvège, examinateur), Mme Elsa Skavounou (*Technical Account Manager*, SYSTRAN, examinatrice).

Résumé *: Le travail présenté dans cette thèse explore la question de l'acquisition automatique de sens pour la désambiguïsation lexicale dans un cadre de traduction. Partant de l'hypothèse du besoin de conformité des inventaires sémantiques utilisés pour la désambiguïsation dans le cadre d'applications précises, la problématique du repérage des sens se situe dans un cadre bilingue et le traitement s'oriente vers la traduction.*
Une méthode d'acquisition de sens est proposée permettant d'établir des correspondances sémantiques entre les mots de deux langues en relation de traduction. L'induction de sens est effectuée par une combinaison d'informations

distributionnelles et traductionnelles extraites d'un corpus bilingue parallèle. Une grande partie de l'analyse au niveau théorique est ainsi consacrée à l'étude de l'applicabilité des hypothèses distributionnelles du sens et de la similarité sémantique dans un cadre bilingue. Ces hypothèses étant sous-jacentes à maintes méthodes d'analyse sémantique s'appuyant sur des informations monolingues, nous montrons qu'elles peuvent être également exploitées dans un cadre impliquant des langues différentes, permettant aux méthodes qui s'en servent de fournir des résultats pertinents. Les hypothèses distributionnelles en question constituent ainsi les principes sur lesquels s'appuient les méthodes implémentées.

Dans la méthode d'acquisition de sens proposée, elles sont couplées avec l'hypothèse d'une correspondance sémantique entre mots en relation de traduction. L'analyse effectuée par cette méthode met au jour des relations de similarité sémantique entre les équivalents des mots ambigus de la langue source. Ces relations servent au clustering sémantique des équivalents ; les clusters générés de cette manière sont projetés sur les mots ambigus source, projection qui rend possible l'induction de distinctions sémantiques au sein de ces mots.

La méthode proposée étant à la fois non supervisée et entièrement fondée sur des données, elle est, par conséquent, indépendante de la langue et permet l'élaboration d'inventaires sémantiques relatifs aux domaines représentés dans les corpus traités. Les sens mis en évidence, dans la mesure où ils sont représentés par des correspondances sémantiques interlangues, s'avèrent pertinents pour la désambiguïsation dans un cadre de traduction. Contrairement aux méthodes établissant des correspondances biunivoques entre sens et équivalents de traduction, les correspondances sont ici élaborées sur la base des relations paradigmatiques repérées entre les équivalents, ce qui constitue une des originalités de cette représentation sémantique. Cette particularité de la méthode lui permet, à la fois, de capter de véritables correspondances de sens et de prendre en compte une caractéristique essentielle à tout processus de traduction, présente dans les textes traduits : celle qui consiste à recourir, de la part du traducteur, à des mots sémantiquement similaires pour traduire le sens d'un mot ambigu source, dans un contexte donné.

L'inventaire ainsi généré se différencie de manière importante des inventaires « classiques », d'une part, par sa structure et, d'autre part, par son contenu. Au sein de cet inventaire, des liens entre sens lexicaux sont modélisés, liens pouvant être exploités pour la modification de la granularité des descriptions sémantiques fournies. De cette manière, en utilisant uniquement des informations internes, la méthode parvient à révéler des sens lexicaux plus ou moins grossiers. La modélisation sémantique effectuée est donc dynamique, dans la mesure où elle permet d'accéder à des informations de quantité et de qualité variables, en fonction des besoins qui se présentent dans un cadre donné. Par ailleurs, étant donné leur caractère interlangue, les descriptions incluses au sein de l'inventaire de sens peuvent être exploitées pour la désambiguïsation et la sélection lexicale dans un cadre de traduction. Ces deux méthodes, de désambiguïsation et de sélection lexicale, ont été mises en place dans le cadre de ce travail et exploitent le contenu de l'inventaire que nous avons élaboré.

Les avantages liés à la nature de la modélisation de sens que nous proposons sont

également manifestes au niveau de l'évaluation des méthodes de désambiguïsation et de sélection lexicale. En effet, dans la mesure où elle met en évidence les relations sémantiques entre les équivalents des mots ambigus et où elle opère une distinction entre sens proches et distants, la modélisation proposée rend possible l'élaboration d'une métrique d'évaluation pondérée. Cette métrique se différencie des métriques classiques d'évaluation du résultat de la traduction automatique, ainsi que de celles utilisées pour évaluer le résultat des méthodes de désambiguïsation multilingue. Basée sur le principe de précision enrichie, notre métrique tient compte de la pertinence sémantique des résultats du processus de sélection lexicale.
Ainsi, les apports qualitatifs – qui consistent en une modélisation dynamique de la sémantique lexicale, permettant la prise en compte de relations intersens – se conjuguent à une augmentation de la qualité de l'évaluation des tâches de désambiguïsation et de sélection lexicale, qui permet, quant à elle, la prise en compte de propositions sémantiquement similaires. Les conclusions sur l'évaluation quantitative des méthodes de désambiguïsation et de sélection lexicale montrent en effet que la modélisation sémantique fournie par notre méthode d'acquisition de sens est bénéfique à ces tâches dans un cadre bilingue.

URL où la thèse pourra être téléchargée :

http://tel.archives-ouvertes.fr/tel-00322285/fr

Lucie BARQUE (lbarque@linguist.jussieu.fr)

Titre : Description et formalisation de la polysémie régulière en français

Mots-clés : sémantique lexicale, polysémie régulière, définitions lexicographiques, patrons de polysémie, sous-spécification sémantique.

Title : *Description and formalization of regular polysemy in French.*

Keywords : *lexical semantics, regular polysemy, lexicographic definitions, polysemy patterns, semantic underspecification.*

Thèse de doctorat en Sciences du Langage, Université de Paris 7 Denis Diderot, LaTTiCe, UFR Linguistique, sous la direction de Sylvain Kahane, Pr, et Alain Polguère, Pr. Université de Montréal. Soutenue le 14/03/2008.

Jury : Mme Laurence Danlos (Pr, Université de Paris 7, présidente), M. Sylvain Kahane (Pr, Université Paris 10, codirecteur), M. Alain Polguère (Pr, Université de Montréal, codirecteur), Mme Pierrette Bouillon, (Pr, Université Genève,

rapporteur), M. Achim Stein (Pr, Université de Stuttgart, rapporteur), M. Georges Kleiber (Pr, Université de Strasbourg, examinateur).

Résumé *: La thèse propose une réflexion théorique sur la nature des liens de polysémie et sur leur rôle dans la structuration du lexique. Son apport principal est de préciser le concept de polysémie régulière défini une première fois dans (Apresjan, 1974) en s'appuyant sur des descriptions lexicales riches et explicitement structurées produites dans le cadre de la lexicologie explicative et combinatoire. La méthode adoptée consiste à décrire en parallèle des sens lexicaux (sous la forme de définitions lexicographiques structurées) et les liens sémantiques entre sens associés à un même signifiant (sous la forme de paires de définitions sous-spécifiées appelées patrons de polysémie). Par exemple, les sens TAUPE(animal)~TAUPE(espion), GORILLE(singe)~GORILLE(garde du corps), etc. ont été décrits en même temps que le lien de polysémie animal → individu ayant une fonction qui lie ces paires. Cette méthode a donné lieu à la production d'un fragment de lexique sémantique formalisé décrivant la polysémie active dans les champs sémantiques des animaux et des sentiments. Il pourra servir de base au développement rapide et systématique d'un lexique du français plus étendu qui mettra en avant les régularités en matière de polysémie.*

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

Daniel DECHELOTTE (dechelot@limsi.fr)

Titre : Traduction automatique de la parole par méthodes statistiques

Mots-clés : traduction par méthodes statistiques, traduction de la parole, reconnaissance de la parole.

Title : *Automatic speech translation by statistical methods.*

Keywords : *statistical machine translation, speech translation, automatic speech recognition.*

Thèse de doctorat en Informatique, Université de Paris 11, LIMSI-CNRS (UPR 3251), UFR de Sciences, sous la direction de Holger Schwenk, Pr et de Jean-Luc Gauvain, DR. Soutenue le 17/12/2007.

Jury : M. Joseph Mariani (Pr, Université de Paris 11, président), M. Holger Schwenk (Pr, Université de Paris 11, codirecteur), M. Jean-Luc Gauvain (DR, LIMSI-CNRS, codirecteur), M. Laurent Besacier, (MC-HDR, Université Joseph-Fourier-ISTG, rapporteur), M. Roland Kuhn (CR, NRC Institute for Information

Technology (Canada), rapporteur), M. Philipp Koehn (professeur assistant, Université d'Edimburgh, examinateur).

***Résumé** : La traduction de la parole est un thème de recherche récent, car il combine deux problèmes scientifiques complexes : la reconnaissance de la parole et la traduction automatique. On imagine pourtant sans mal les applications potentielles : systèmes de réservation multilingues, aide au tourisme, indexation cross-lingue de contenus multimédias, assistant pour l'échange d'informations et la négociation, etc. Cette thèse a porté sur la traduction automatique et plus particulièrement sur la traduction de la parole reconnue automatiquement. La tâche retenue est la traduction des discours des députés européens aux sessions plénières du parlement européen, entre l'anglais et l'espagnol.*

Nos recherches ont débuté par la conception d'un décodeur pour le modèle « IBM-4 », un modèle statistique performant à base de mots. Ce décodeur a été entièrement développé au cours de cette thèse. Au milieu de l'année 2006, l'avènement de Moses, un décodeur par groupes de mots libre et à l'état de l'art, nous a donné l'opportunité de poursuivre nos recherches avec un autre modèle de traduction. Nous avons envisagé une collaboration entre les deux décodeurs, mais elle n'a malheureusement pas produit l'amélioration espérée.

Dans nos expériences avec les deux décodeurs, le modèle de langage quadrigramme neuronal, développé originellement pour la reconnaissance de la parole, s'est avéré très performant dans les deux sens de traduction, amenant des améliorations sensibles pour toutes les mesures automatiques. Les systèmes de traduction mis en œuvre dans cette thèse ont été très compétitifs à la dernière évaluation TC-Star, en février 2007.

De plus, nous avons tenté d'améliorer les performances des systèmes de traduction par groupes de mots comme Moses. Au cœur de ces systèmes se trouve la table de traduction, sorte de « dictionnaire bilingue ». Les scores qu'elle contient sont le résultat de choix heuristiques. Nous avons proposé un algorithme inspiré de celui du Perceptron pour modifier de façon discriminante ces scores en observant les erreurs de traduction sur un ensemble de développement. Un gain substantiel a été observé dans un sens de traduction mais n'a pas été confirmé sur une autre tâche. Nous pensons que ces résultats contrastés pourraient être dus à un défaut de cohérence, ou de lissage, entre les scores de la table de traduction.

L'amélioration de l'interaction entre la reconnaissance de la parole et la traduction présente plusieurs aspects. À notre connaissance, nous avons été les premiers à mesurer l'impact du taux de mots erronés de la reconnaissance sur les performances de la traduction, et d'évaluer séparément les impacts respectifs du modèle de langage source et du modèle acoustique.

Un autre aspect est la prise en compte de l'ambiguïté de la sortie de la reconnaissance automatique, c'est-à-dire les mots entre lesquels le système de

reconnaissance « hésite ». D'après notre propre expérience et les articles parus à ce sujet, toutes les méthodes imposent leurs compromis respectifs, que l'on traduise le treillis de mots produit par la reconnaissance, un réseau de confusion ou une liste de n meilleures hypothèses.

Nous nous sommes ensuite placés dans le cadre de la traduction d'un flux de mots produit par un système de reconnaissance « inconnu ». Plusieurs traitements spécifiques à la parole sont utiles, pour gérer les mots répétés et autres disfluences. Nous avons constaté l'importance de transformer les données à traduire pour les faire ressembler aux données d'entraînement du système. Mais de façon surprenante, en matière de divergence entre données d'entraînement et données de test, nos expériences ont montré que la ponctuation était au moins aussi importante que les mots. Nous avons proposé un algorithme d'insertion de ponctuations dont le seul critère est la quantité de ponctuations insérées. Cet algorithme a permis d'améliorer très nettement le score Bleu, une mesure automatique populaire de la qualité de traduction.
Enfin, nous avons modifié le système de reconnaissance de manière à lui faire insérer ou supprimer plus de mots. Bien que le score Bleu puisse légèrement bénéficier d'un taux d'insertion plus élevé, le taux de mots erronés WER semble le bon critère à minimiser par la reconnaissance pour obtenir les meilleures performances de traduction de la parole.

URL où la thèse pourra être téléchargée :

http://www.limsi.fr/Individu/dechelot/archives/Dechelotte_These07.pdf
http://yo.dan.free.fr/archives/Dechelotte_These07.pdf

Juan-Manuel TORRES-MORENO (juan-manuel.torres@univ-avignon.fr)

Titre : Du textuel au numérique : analyse et classification automatiques

Mots-clés : apprentissage automatique, traitement automatique de la langue naturelle, classification textuelle, résumé automatique de textes.

Title : *From texts to numbers: automatic analysis and classification.*

Keywords : *Machine Learning, Natural Processing Language, text classification, automatic text summarization.*

Mémoire d'HDR en Informatique, Université d'Avignon et des Pays de Vaucluse, Laboratoire Informatique d'Avignon (UPRES 4128), UFR de Sciences. Soutenue le 12/12/2007.

Jury : Mme Violaine Prince (Pr, LIRMM Montpellier, présidente), M. Frédéric Alexandre (DR, INRIA-Loria Nancy, rapporteur), M. Joan Cabestany, (Pr, UPC Barcelona, rapporteur), M. Jean-Paul Haton (Pr, Université Henri Poincaré Nancy, rapporteur), M. Guy Lapalme (Pr, RALI Montréal, rapporteur), M. Eitan Altman (DR, INRIA-Sophia Antipolis, examinateur), M. Marc Elbèze (Pr, Université d'Avignon, examinateur), Mme Mirta Gordon (Pr, LANCI, Montréal, examinatrice), M. Jean-Guy Meunier (Pr, LANCI, Montréal, examinateur).

***Résumé** : La présentation en vue de l'obtention de l'habilitation à diriger des recherches (HDR) est une synthèse de nos travaux de recherche menés depuis la fin de notre thèse de doctorat (réseaux de neurones incrémentaux, soutenue fin 1997). Elle couvre l'année de stage postdoctoral au LANCI au Canada, les trois années comme professeur à l'UQAC et à l'École polytechnique (Montréal, Canada) ainsi que les trois années consécutives à notre recrutement au sein de l'Université d'Avignon, puis comme responsable de la thématique TALNE du LIA. Nos travaux sont à l'intersection de trois domaines: l'apprentissage automatique, le traitement automatique de la langue naturelle écrite (TAL) et les méthodes probabilistes. Les modèles utilisés sont pour la plupart issus d'apprentissages automatiques. Ils essaient de capturer les connaissances cachées dans les corpus documentaires. Cela n'est pas incompatible avec notre formation doctorale. On a ainsi posé des problèmes de classification d'opinions comme un problème de classification où tout le cadre formel de l'apprentissage et généralisation par des réseaux de neurones ou SVM peut être appliqué avec succès. L'hypothèse de base de nos travaux en TAL est qu'il n'y a rien de plus concret que les textes. Ainsi, nos travaux se sont concentrés sur la classification et la catégorisation de textes, le résumé automatique de documents, la compression automatique de phrases et la génération automatique de textes, toujours appliqués sur des grandes masses de textes. Ces méthodes ont été validées en utilisant des approches pragmatiques : les campagnes d'évaluation nationales et internationales. Nous avons participé à plusieurs campagnes d'évaluation (DEFT en détection d'opinions et identification d'auteur, DUC en résumé orienté par une thématique) et, dans ces campagnes, les performances des méthodes numériques surpassent ou égalent celles des méthodes symboliques sans utiliser de lourdes ressources linguistiques. Pendant ces années de recherche, plusieurs fois nous nous sommes posé la question de savoir si la linguistique pouvait encore jouer un rôle dans le traitement de la langue naturelle.*

Peut-on aller vers le tout numérique ?

Au-delà des promesses théoriques d'indépendance, l'approche numérique est fortement dépendante des corpora annotés (souvent à la main). Les corpora sont parfois insuffisants face aux tâches complexes et alors les unités, telles que les n-grammes, deviennent des événements très rares. On peut, certes, pallier leur manque par des algorithmes de lissage (Good-Turing, Backoff, Katz), mais ces derniers induisent parfois des biais non évidents. Enfin, le modèle de sac de mots est une simplification exagérée qui néglige la structure de la phrase, ce qui implique une perte importante d'information. Cette approche a donc ses limites. Les

approches linguistiques et les méthodes numériques peuvent-elles jouer un partenariat dans les tâches du TAL? La réponse positive à cette question ouvre une voie intéressante aux recherches que nous comptons entreprendre : la conception de systèmes TAL hybrides, notamment pour la génération automatique de textes et pour la compression de phrases. On peut difficilement envisager de dépasser le plafond auquel les méthodes numériques se heurtent sans faire appel à la finesse des approches linguistiques, mais sans négliger pour autant de les valider et de les tester sur des corpora. Les méthodes d'apprentissage son capables de traiter des grandes masses de documents à un certain niveau de granularité, mais l'analyse linguistique est plus fine et cela a ses avantages. Nous avons montré à travers nos travaux, en particulier ceux consacrés au résumé automatique et au raffinement de requêtes, qu'un système hybride combinant des approches numériques à la base et une analyse linguistique au sommet, donne de meilleures performances que les systèmes pris de façon isolée.

Enfin, cette HDR se conclut par la présentation d'une approche lointaine au TAL et aux méthodes numériques : celle de la physique statistique. À partir de notions intuitives de phénomènes de magnétisation, nous avons utilisé le formalisme des systèmes de spin et les réseaux de neurones pour introduire le concept d'énergie textuelle d'un document. Cette nouvelle mesure de similarité a été appliquée avec succès aux tâches de résumé générique et guidé par une thématique, ainsi qu'à celui de la segmentation thématique de documents. D'autres travaux en cours devraient confirmer d'autres applications de cette nouvelle notion de similarité.

URL où la thèse pourra être téléchargée :
http://lia.univ-avignon.fr/fileadmin/documents/Users/Intranet/chercheurs/torres/recherche/torres_hdr_2008.pdf

http://daniel.iut.univ-metz.fr/~cortex/torres_hdr_2008.pdf