Résumés de thèses

Rubrique préparée par Fiammetta Namer

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

Houda BOUAMOR : (houda.bouamor@limsi.fr)

Titre : Étude de la paraphrase sous-phrastique en traitement automatique des langues

Mots-clés : corpus monolingues, acquisition de paraphrase, classification automatique de paraphrases, typologie de paraphrases.

Title: *A study of phrasal paraphrases in Natural Language Processing*

Keywords: *monolingual corpora, paraphrase acquisition, paraphrase classification, paraphrase typology.*

Thèse de doctorat en Informatique, LIMSI-CNRS, École doctorale d'informatique de Paris-Sud, Université Paris-Sud, Orsay, sous la direction d'Aurélien Max (MCF, Université Paris-Sud) et d'Anne Vilnat (Pr, Université Paris-Sud). Thèse soutenue le 11/06/2012.

Jury : M. Aurélien Max, (MC, Université Paris-Sud, codirecteur), Mme Anne Vilnat (Pr, Université Paris-Sud, codirectrice), M. Yves Lepage (Pr, Université Waseda, Japon, rapporteur), M. Emmanuel Morin (Pr, Université de Nantes, rapporteur), M. Philippe Langlais (Pr, Université de Montréal, examinateur), Mme Adelne Nazarenko (Pr, Université Paris-Nord, examinatrice), M. François Yvon (Pr, Université Paris-Sud, examinateur).

Résumé : *La variabilité en langue est une source majeure de difficultés dans la plupart des applications du traitement automatique des langues. Elle se manifeste dans le fait qu'une même idée ou un même événement peut être exprimé avec des mots ou des groupes de mots différents ayant la même signification dans leur contexte respectif. Capturer automatiquement des équivalences sémantiques entre des unités de texte est une tâche complexe mais s'avère indispensable dans de nombreux contextes. L'acquisition* a priori *de listes d'équivalences met à disposition des ressources utiles pour, par exemple, améliorer le repérage d'une réponse à une*

question, autoriser des formulations différentes en évaluation de la traduction automatique, ou encore aider des auteurs à trouver des formulations plus adaptées.

Dans cette thèse, nous proposons une étude détaillée de la tâche d'acquisition de paraphrases sous-phrastiques à partir de paires d'énoncés sémantiquement liés. Nous démontrons empiriquement que les corpus parallèles monolingues, bien qu'extrêmement rares, constituent le type de ressources le plus adapté pour ce genre d'étude. Nos expériences mettent en jeu cinq techniques d'acquisition, représentatives de différentes approches et connaissances, en anglais et en français. Afin d'améliorer la performance en acquisition, nous réalisons la combinaison des paraphrases produites par ces techniques par une validation reposant sur un classifieur automatique à maximum d'entropie biclasse. Un résultat important de notre étude est l'identification de paraphrases qui défient actuellement les techniques étudiées, lesquelles sont classées et quantifiées en anglais et français. Nous examinons également dans cette thèse l'impact de la langue, du type du corpus et la comparabilité des paires des énoncés utilisés sur la tâche d'acquisition de paraphrases sous-phrastiques. Nous présentons le résultat d'une analyse de la performance des différentes méthodes testées en fonction des difficultés d'alignement des paires de paraphrases d'énoncés. Nous donnons, ensuite, un compte-rendu descriptif et quantitatif des caractéristiques des paraphrases trouvées dans les différents types de corpus étudiés ainsi que celles qui défient les approches actuelles d'identification automatique.

URL où la thèse pourra être téléchargée : http://tel.archives-ouvertes.fr/tel-00717702

Marie CALBERG-CHALLOT : (marie.calberg-challot@orange.fr)

Titre : Dynamique de la langue et de la terminologie dans le domaine de l'ingénierie nucléaire

Mots-clés : ingénierie nucléaire, dictionnaire, vocabulaire, langue de spécialité, terminologie.

Title: *Dynamics of language and terminology in the field of nuclear energy: a case study.*

Keywords: *monolingual corpora, paraphrase acquisition, paraphrase classification, paraphrase typology.*

Thèse de doctorat en Sciences du Langage, UMR HTL, UFRL, École doctorale sciences du langage, Université Paris-Diderot, sous la direction de Danielle Candel (CR, HTL, CNRS) et de John Humbley (Pr, Université Paris-Diderot). Thèse soutenue le 24/01/2012.

Jury : Mme Danielle Candel, (CR, HTL, codirectrice), M. John Humbley (Pr, Université Paris-Diderot, codirecteur), M. Christophe Roche (Pr, Université de Savoie, rapporteur et président), Mme Maria Teresa Cabré (Pr, Universitat Pompeu Fabra, Barcelone, rapporteur), M. Xavier Dumont (ingénieur, GDF Suez, examinateur).

Résumé : *Le but de ce travail est d'aider à une meilleure diffusion des résultats de la recherche scientifique et technologique dans le domaine de l'énergie nucléaire, à un meilleur échange, et à obtenir une information scientifique et technique de qualité.*

Un rappel des grandes étapes de l'histoire de l'énergie nucléaire permettra d'en dégager quelques faits marquants. Ces étapes sont aussi l'occasion de voir émerger de premières créations et innovations terminologiques dans le choix de termes, dans l'art de les définir ou dans la manière de les employer en contexte.

Tout au long de l'étude, une attention toute particulière sera accordée, par ailleurs, aux premiers témoignages ou réactions de linguistes, tels qu'ils peuvent être recueillis dans les dictionnaires de langue. En effet, l'enregistrement ou l'absence de termes dans le dictionnaire peuvent eux-mêmes être significatifs du jugement porté sur le vocable, tout comme peut l'être son traitement lexicographique. C'est aussi le problème de la diffusion et de la vulgarisation des connaissances qui est ainsi abordé. Après avoir exposé quelques modèles de l'histoire, on décrira surtout l'état de la terminologie contemporaine, les modes de création terminologique en ingénierie nucléaire et les besoins actuels dans ce domaine.

Le corpus d'étude, essentiel pour la recherche présentée, sera constitué de trois ensembles : des textes disponibles dans le domaine du nucléaire, notamment dans les sociétés et organismes de recherche, des dictionnaires collectés dans ces mêmes sources, et enfin des témoignages d'experts. La constitution d'un réseau d'experts du domaine est déjà en cours, ce qui représente un atout certain pour une telle recherche.

Chacun de ces trois ensembles s'appuie sur des ressources originales et complémentaires à la fois, qui seront exploitées pour l'étude de la création terminologique aussi bien que pour la méthodologie de formalisation des termes du nucléaire. Des recherches et des solutions de formalisation préexistantes seront réutilisées, et nous nous proposons de contribuer à l'amélioration d'outils disponibles en ingénierie des langues, en améliorant leur réutilisation et leur exploitation. On comprendra alors que c'est la création d'un nouveau dictionnaire de l'ingénierie nucléaire qui sera l'aboutissement souhaité de cette thèse.

URL où la thèse pourra être téléchargée : s'adresser à l'auteur.

Antoine DOUCET : (antoine.doucet@unicaen.fr)

Titre : Extraction, exploitation et évaluation de connaissances à partir de documents

Mots-clés : fouille de données séquentielles, unités multimots, recherche d'information, évaluation des systèmes d'information, méthodes multilingues, passage à l'échelle.

Title: *Extraction, Exploitation and Evaluation of Document-based Knowledge*.

Keywords: s*equential data mining, multiword units, information retrieval, evaluation of information systems, any language techniques, scalability.*

HDR en Informatique, CREYC-CNRS, département informatique, UFR de Sciences, Université de Caen Basse-Normandie, Caen, sous la direction de Gaël Dias (Pr, Université de Caen). HDR soutenue le 30/04/2012.

Jury : M. Gaël Dias, (Pr, Université de Caen, directeur), Mme Isabelle Tellier (Pr, Université Paris 3, présidente), M. Massih-Reza Amini (MC-HDR, Université Pierre et Marie Curie, Paris 6, rapporteur), M. Pavel Bradzil (Pr, Université de Porto, rapporteur), M. Manuel Vilares Ferro (Pr, Université de Vigo, Espagne, rapporteur), M. Bruno Crémilleux (Pr, Université de Caen Basse-Normandie, examinateur), Mme Mounia Lalmas (chercheur, Yahoo! Research Barcelone, examinatrice).

Résumé **:** *Les travaux présentés dans ce mémoire gravitent autour du document numérique : extraction de connaissances, utilisation de connaissances et évaluation des connaissances extraites, d'un point de vue théorique aussi bien qu'expérimental.*

Le noyau commun à mes travaux de recherche est la généricité des méthodes avec une attention particulière apportée à la question du passage à l'échelle. Ceci implique que les algorithmes, principalement appliqués au texte dans ce mémoire, fonctionnent en réalité pour tout type de données séquentielles.

Sur le matériau textuel, la généricité et la robustesse algorithmique des méthodes permettent d'obtenir des approches endogènes, fonctionnant pour toute langue, pour tout genre et pour tout type de documents (et de collection de documents). Le matériau expérimental couvre ainsi des langues utilisant différents alphabets, et des langues appartenant à différentes familles linguistiques. Les traitements peuvent d'ailleurs être appliqués de la même manière au grain phrase, mot, ou même caractère.

Les collections traitées vont des dépêches d'agences de presse aux ouvrages numérisés, en passant par les articles scientifiques.

Ce mémoire présente mes travaux en fonction des différentes étapes du pipeline de traitement des documents, de leur appréhension à l'évaluation applicative. Le document est ainsi organisé en trois parties décrivant des contributions en :

- *extraction de connaissances (fouille de données séquentielle et veille multilingue) ;*
- *exploitation des connaissances acquises, par des applications en recherche*

d'information, classification et détection de synonymes via *un algorithme efficace d'alignement de paraphrases ;*
- *méthodologie d'évaluation des systèmes d'information dans un contexte de données massives, notamment évaluation des performances des systèmes de recherche d'information sur des bibliothèques numérisées.*

URL où l'HDR pourra être téléchargée : https://doucet.users.greyc.fr/

Selja SEPPÄLÄ : (selja.seppala.unige@gmail.com)

Titre : Contraintes sur la sélection des informations dans les définitions terminographiques : vers des modèles relationnels génériques pertinents

Mots-clés : définitions terminographiques, sélection d'informations, traits conceptuels, pertinence des traits, concepts, terminologie, terminographie, dictionnaires.

Title: *Constraints on information selection in terminographic definitions: towards relevant generic relational models.*

Keywords: *terminographic definitions, information selection, conceptual features, feature relevance, concepts, terminology, terminography, dictionaries.*

Thèse de doctorat en Traitement Informatique Multilingue (terminologie et traitement automatique des langues), département de traitement informatique multilingue, faculté de traduction et d'interprétation, Université de Genève, sous la direction de Bruno de Bessé (Pr honoraire, Université de Genève). Thèse soutenue le 29/02/2012.

Jury : M. Bruno de Bessé, (Pr hon., Université de Genève, directeur), Mme Pierrette Bouillon (Pr, Université de Genève, présidente), Mme Marie-Claude L'Homme (Pr, Université de Montréal, rapporteur externe), Mme Anne Reboul (DR, L2C2-CNRS, rapporteur externe).

Résumé : *La rédaction de définitions est une activité essentielle dans le développement de ressources terminologiques de qualité. Cette activité, qui est pour l'heure surtout réalisée à la main, gagnerait cependant en efficacité et en cohérence si l'on parvenait à l'automatiser. Notre objectif général est donc de concevoir et de mettre en œuvre des outils d'aide à la rédaction de définitions génériques qui puissent être utilisés dans tout type de contexte terminographique, indépendamment du domaine traité ou de la langue de rédaction.*

Pour ce faire, nous proposons une réflexion de fond sur la définition et l'activité définitoire. Automatiser la rédaction de définitions exige, en effet, de préciser ce

qu'est une définition terminographique, ce que l'on définit, comment on définit et, surtout, quelles sont les questions que soulève cette activité. L'examen de ces questions nous amène, ainsi, à spécifier la problématique centrale de notre thèse, à savoir la sélection des informations définitoires.

La plupart du temps, le terminologue compose les définitions à partir de textes spécialisés rédigés par des experts. Or, toutes les informations contenues dans ces textes ne sont pas définitoires et, parmi celles qui le sont, toutes ne sont pas pertinentes pour définir. Les principales questions que soulève donc cette activité et auxquelles il y a lieu de répondre si l'on veut l'automatiser sont :

- *Qu'est-ce qui détermine ou influence la sélection des informations ?*
- *Quels types d'informations sont pertinents pour définir ?*

Sur la base d'un inventaire des différents facteurs connus pour contraindre la sélection des informations définitoires, nous identifions celui qui devrait, a priori*, être le plus indépendant des domaines et des langues. Nous faisons, ainsi, l'hypothèse que la sélection des informations dépend, en partie, du type d'entité défini. Si cette hypothèse est vérifiée, alors il est possible de proposer des modèles définitoires relationnels fondés sur les propriétés des différents types d'entités existants.*

Pour tester cette hypothèse, nous proposons d'adopter les catégories de haut niveau d'une ontologie formelle réaliste, la Basic Formal Ontology (BFO), qui reposent sur des distinctions philosophiques et qui sont conformes aux connaissances scientifiques du monde. Nous élaborons des modèles relationnels à partir des spécifications de cette ontologie et utilisons ces modèles pour annoter un corpus multidomaine de définitions existantes. Des analyses de corpus quantitatives permettent de voir dans quelle mesure les propriétés de chaque type d'entité ont été jugées pertinentes pour définir en terminologie. Les résultats très encourageants d'une étude empirique pilote visant à tester l'applicabilité de notre proposition tendent à confirmer notre hypothèse. Ils ouvrent, ainsi, la voie à de futurs travaux de consolidation de notre proposition et de mise en œuvre des modèles dans des outils d'aide à la rédaction de définitions. Par les réflexions théoriques de ce travail, nous contribuons également aux fondements d'une théorie intégrée des définitions en terminologie.

URL où la thèse pourra être téléchargée : http://archive-ouverte.unige.ch

---

Nadi TOMEH : (nadi.tomeh@gmail.com)

Titre : Modèles discriminants d'alignement pour la traduction automatique statistique.

Mots-clés : traduction statistique, modèles d'alignement mot à mot, modèles discriminants.

Title: *Discriminative alignment models for statistical machine translation.*

Keywords: *Statistical machine translation, Word alignement, Discriminative models.*

Thèse de doctorat en Informatique, UFR des Sciences, LIMSI-CNRS, Université Paris-Sud, Orsay, sous la direction d'Alexandre Allauzen (MC, Université Paris-Sud) et de François Yvon (Pr, Université de Paris-Sud). Thèse soutenue le 27/06/2012.

Jury : M. Alexandre Allauzen (MC, Université Paris-Sud, codirecteur), M. François Yvon (Pr, Université de Paris-Sud, codirecteur), Mme Anne Vilnat (Pr, Université de Paris-Sud, présidente), M. Eric Gaussier (Pr, Université Joseph-Fourier, rapporteur), M. Philippe Langlais (Pr, Université de Montréal, rapporteur), M. Hermann Ney (Pr, University RWTH Aachen, examinateur).

Résumé : *La tâche d'alignement d'un texte dans une langue source avec sa traduction en langue cible est souvent nommée alignement bitexte. Elle a pour but de faire émerger les relations de traduction qui peuvent s'exprimer à différents niveaux de granularité entre les deux faces du bitexte. De nombreuses applications de traitement automatique des langues naturelles s'appuient sur cette étape afin d'accéder à des connaissances linguistiques de plus haut niveau. Parmi ces applications, nous pouvons citer bien sûr la traduction automatique, mais également l'extraction de lexiques et de terminologies bilingues, la désambiguïsation sémantique ou l'apprentissage des langues assisté par ordinateur.*

La complexité de la tâche d'alignement de bitextes s'explique par les différences linguistiques entre les langues. Ces différences peuvent être d'ordre sémantique, syntaxique, ou morphologique. Dans le cadre des approches probabilistes, l'alignement de bitextes est modélisé par un ensemble de variables aléatoires cachées. Afin de réduire la complexité du problème, le processus aléatoire sous-jacent fait l'hypothèse simplificatrice qu'un mot en langue source est lié à au plus un mot dans la langue cible, ce qui induit une relation de traduction asymétrique. Néanmoins, cette hypothèse est simpliste, puisque les alignements peuvent de manière générale impliquer des groupes de mots dans chacune des langues. Afin de rétablir cette symétrie, chaque langue est considérée tour à tour comme la langue source et les deux alignements asymétriques résultants sont combinés à l'aide d'une heuristique. Cette étape de symétrisation revêt une importance particulière dans l'approche standard en traduction automatique, puisqu'elle précède l'extraction des unités de traduction, c'est-à-dire les paires de segments.

L'objectif de cette thèse est de proposer de nouvelles approches, d'une part pour l'alignement de bitextes, et d'autre part pour l'extraction des unités de traduction.

*La spécificité de notre approche consiste à remplacer les heuristiques utilisées par des modèles d'apprentissage discriminant. Nous présentons un modèle « Maximum d'entropie » (ou MaxEnt) pour l'alignement de mots pour lesquels chaque lien d'alignement est prédit de manière indépendante. L'interaction entre les liens d'alignement est alors prise en compte par l'empilement (*stacking*) d'un second modèle prenant en compte la structure à prédire sans pour autant augmenter la complexité globale. Cette formulation peut être vue comme une manière d'apprendre la combinaison de différentes méthodes d'alignement : le modèle considère ainsi l'union des alignements d'entrées pour en sélectionner les liens jugés fiables. Le modèle MaxEnt proposé permet d'améliorer les performances d'un système de traduction automatique de l'arabe vers l'anglais en considérant le jeu de données de la tâche NIST'09. Ces améliorations sont mesurées en termes de taux d'erreurs sur les alignements ainsi qu'en termes de qualité de traduction* via *la métrique automatique BLEU.*

Nous proposons également un modèle permettant à la fois de sélectionner et d'évaluer les unités de traduction extraites d'un bitexte aligné. Ces deux étapes sont reformulées dans le cadre de l'apprentissage supervisé, afin de modéliser la décision de garder ou pas une paire de segments comme une unité fiable de traduction. Ce cadre permet l'utilisation de caractéristiques riches et nombreuses favorisant ainsi une décision robuste. Nous proposons une méthode simple et efficace pour annoter les paires de segments utiles pour la traduction. Le problème d'apprentissage automatique qui se pose alors est particulier, puisque nous ne disposons que d'exemples positifs. Nous proposons d'utiliser l'approche SVM à une classe afin de modéliser la sélection des unités de traduction. Grâce à cette approche, nous obtenons des améliorations significatives en termes de scores BLEU pour un système entraîné avec un petit ensemble de données.

URL où la thèse pourra être téléchargée : http://perso.limsi.fr/Individu/nadi/wp-content/uploads/Nadi-Tomeh_thesis.pdf