Résumés de thèses

Rubrique préparée par Fiammetta Namer

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

Florence AMARDEILH (florence.amardeilh@mondeca.com)

Titre : Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle.

Mots-clés : Web sémantique, informatique linguistique, extraction d'information, arbres conceptuels, acquisition de connaissances, représentation des connaissances, ontologies, annotation sémantique, base de connaissance.

Title : *Semantic Web and Computational Linguistic : methodological proposals and conception of a software platform.*

Keywords : *Semantic Web, computational linguistic, information extraction, conceptual trees, knowledge acquisition, knowledge representation, ontologies, semantic annotation, knowledge base.*

Thèse de doctorat en Informatique et Sciences du Langage, Université de Paris-X Nanterre, MoDyCo (Modèles, Dynamiques, Corpus) – UMR CNRS 7114, sous la direction de M. Jean-Luc MINEL, Ingénieur de Recherche LaLLIC-CNRS (HDR) et M. Philippe Laublet, MC, Université Paris IV – Sorbonne. Soutenue le 10/05/2007.

Jury : M. Benoît Habert (Pr, Université de Paris X Nanterre, président), M. Jean-Luc Minel (IR HDR, MoDyCo-CNRS, codirecteur), M. Philippe Laublet, (MC, Université Paris IV, codirecteur), Mme Nathalie Aussenac (CR HDR, IRIT-CNRS, rapporteur), M Gilles Kassel (Pr, Université de Picardie, rapporteur), Mme Teresa Pazienza (Pr, Università Roma Tor Vergata, Italie, examinatrice), M Jean Delahousse (PDG, Mondeca, Paris, invité).

Résumé *: Cette thèse aborde les problématiques liées à l'annotation sémantique et au peuplement d'ontologies dans le cadre défini par le Web Sémantique (WS). La vision du Web Sémantique initiée en 1998 par Sir Tim Berners-Lee a pour objectif de permettre une meilleure exploitation des informations disponibles sur le Web par les agents logiciels. Pour cela, les ressources, textuelles ou multimédias, doivent être sémantiquement étiquetées par des annotations structurées. Dans ce processus*

d'annotation sémantique, les ontologies jouent un rôle primordial puisqu'elles modélisent les concepts, attributs et relations utilisées pour annoter le contenu des ressources. Mais, il est tout aussi important que la base de connaissance, associée à cette ontologie, contienne les instances à utiliser pour l'annotation sémantique. C'est pourquoi la tâche de peuplement d'ontologie a pour but d'enrichir (semi-) automatiquement la base de connaissance avec les nouvelles instances de concepts, d'attributs et de relations.

La réalisation de ces deux tâches consiste à combiner les outils d'extraction d'information (EI) avec les outils de représentation des connaissances du WS. En effet, le principal mode de transfert de la connaissance se fait par l'utilisation du langage naturel dans les ressources documentaires. Malgré tout, il existe actuellement un fossé entre les formats de représentation des analyses linguistiques et ceux de représentation des connaissances. Cette thèse propose de combler ce fossé en concevant un médiateur capable de transformer les étiquettes linguistiques générées par les outils d'EI en des représentations plus formelles, annotations sémantiques des textes ou instances d'une ontologie donnée et relations entre celles-ci. L'enjeu consiste aussi bien à proposer une réflexion méthodologique sur l'interopérabilité des différentes technologies, qu'une conception de solutions opérationnelles dans le monde des entreprises, et à plus large échelle du Web.

Dans le cadre de cette thèse, nous avons donc conçu une démarche, nommée OntoPop pour « Ontology Population », qui met en place une passerelle reposant sur un ensemble de règles, dites « d'acquisition de connaissances » et sur un langage d'implémentation de ces règles, OPAL (Ontology Population and Annotation Language. Nous montrons comment cette passerelle peut être utilisée dans un cycle complet d'extraction d'information, d'enrichissement des ressources terminologiques et ontologiques, d'annotations sémantiques et de mise à jour des lexiques utilisés par l'outil d'EI. L'accent est porté sur la résolution des problèmes soulevés par un tel cycle de vie, notamment à propos de la consolidation des nouvelles annotations et instances vis-à-vis du modèle de l'ontologie. Enfin, nous soumettons des propositions pour l'opérationnalisation de la démarche OntoPop à travers une méthodologie et une plateforme logicielle basée sur l'outil de représentation des connaissances ITM de la société Mondeca. La méthodologie a pour objectif de fournir un mode d'emploi simple et efficace pour la réalisation d'une application concrète d'annotations sémantiques ou de peuplement d'ontologie au sein d'une entreprise. La plateforme logicielle offre des exemples de composants logiciels modulaires, autorisant un maximum de flexibilité vis-à-vis des besoins et objectifs de chaque nouvelle application d'annotation sémantique ou de peuplement d'ontologie.

***Abstract**: This thesis deals with the issues related to semantic annotation and ontology population within the framework defined by the Semantic Web (SW). The vision of the Semantic Web initiated in 1998 by Sir Tim Berners-Lee aims to structure information available on the Web. To achieve that goal, the resources, textual or multimedias, must be semantically tagged by metadata so that the*

software agents can exploit them. The explicit representation of the contents of the Web documentary resources is made possible thanks to the ontologies. In the process of semantic annotation, ontologies play a major part since they model the concepts, their attributes and the relations used to annotate the contents of the documents. Ontology constrains the applications on the authorized vocabularies and instances authorized as metadata. But if it is essential for a Semantic Web application to rely on an ontology for the realization of this semantic annotation task, it is also important that the knowledge base, associated with this ontology, contains the instances to be used for semantic annotation. This is why the purpose of the ontology population task aims to enrich (semi-)automatically the knowledge base with new instances of concepts, attributes and relations as defined by the ontology model.

The idea suggested in this thesis is to combine the information extraction (IE) tools with the knowledge representation tools of the WS for the achievement of these two tasks. Indeed, the resolution of the challenges posed by these tasks depends of these IE tools considering that the principal mode of knowledge transfer is done through the use of natural language in the documentary resources. Despite all integration efforts, there is currently a gap between the representation formats of the linguistic tools and those of the knowledge representation tools in the field of the Semantic Web. This thesis proposes to fill this gap by designing a mediator able to transform the tags generated by the IE tools into a more formal representation, being the semantic annotations or the ontology instances. In other words, we try to answer the following issue: how can we map a certain textual representation into a semantic knowledge representation? The stake consists in proposing a methodological reflexion on the interoperability of various technologies as well as a design of operational solutions in the world of the companies, and on broader scale of the Web.

Within this thesis, we thus conceived a framework named OntoPop for "Ontology Population". This framework proposes a bridge in the form of rules, known as "Knowledge Acquisition Rules". They will allow the transformation of the conceptual trees from a linguistic extraction into a formal semantic knowledge representation in the form of annotations or instances. This transformation strongly relies on the concept of context within the conceptual trees. The OPAL language (Ontology Population and Annotation Language) defines a grammar for the implementation of these rules. Besides, we illustrate the way they can be used in a complete lifecycle including information extraction, terminological and ontological resources enrichment, semantic annotation and lexicons update. We stress the resolution of the problems implied by such a lifecycle, such as the transformation itself, the annotation and instances consolidation with respect to the ontology model and the maintenance of the lexicons. Lastly, we submit proposals for the implementation of the OntoPop through a methodology in five stages and a software platform based on the knowledge repository ITM designed by Mondeca. The methodology aims to provide simple and effective instructions for the realization of a concrete semantic annotation or ontology population application within a

company. The software platform offers examples of modular software components, allowing a maximum of flexibility with respect to the needs and objectives for each new application.

URL où la thèse pourra être téléchargée :

http://tel.archives-ouvertes.fr/tel-00146213

Maxime AMBLARD (maxime.amblard@gmail.com)

Titre : Calculs de représentations sémantiques et syntaxe générative : les grammaires minimalistes catégorielles

Mots-clés : grammaires génératives, interface syntaxe/sémantique, isomomrphisme de Curry-Howard, lambda-calcul, lambda mu-calcul, logique linéaire, langages formels, grammaires catégorielles, grammaires minimalistes, types de Montague.

Title : *Semantic representations and generative grammars : the Categorial Minimalist Grammars.*

Keywords : *generative grammars, syntax/semantic interface, Curry-Howard isomorphism, lambda-calculus, lambda mu-calculus, linear logic, formal languages, categorial grammars, minimalist grammars, Montague semantic.*

Thèse de doctorat en Informatique, Université de Bordeaux 1, LaBRI (Laboratoire Bordelais de Recherche en Informatique) – UMR CNRS 5800, sous la direction de M. Christian Retore, Pr, Université Bordeaux 1 et M. Alain Lecomte, Pr, Université Paris 8. Soutenue le 21/09/2007.

Jury : M. Géraud Sénizergues (Pr, Université Bordeaux 1-LaBRI, président), M. Christian Rétoré (Pr, Université Bordeaux 1-LaBRI, co-directeur), M. Alain Lecomte, (Pr, Université Paris 8, co-directeur), Mme Isabelle Tellier (MC HDR, Université Lille3, rapporteur), M. Uwe Mönnich (Pr, Université de Tübingen, Allemagne, rapporteur), M. Nicholas Asher (DR., CNRS-IRIT, examinateur), M. Gregory Kobele (MC, Université de Berlin, Allemagne, invité).

Résumé *: Les travaux de cette thèse se situent dans le cadre de la linguistique computationnelle. La problématique est de définir une interface syntaxe/sémantique basée sur les théories de la grammaire générative.*

Une première partie concernant le problème de l'analyse syntaxique présente tout

d'abord la syntaxe générative, puis un formalisme la réalisant : les grammaires minimalistes de Stabler.
À partir de ces grammaires, nous réalisons une étude sur les propriétés de l'opération de fusion pour laquelle nous définissons des notions d'équivalence, ainsi qu'une modélisation abstraite des lexiques.

Une seconde partie revient sur le problème de l'interface. Pour cela, nous proposons un formalisme de type logique, basé sur la logique mixte (possédant des connecteurs commutatifs et non commutatifs), qui équivaut, sous certaines conditions, aux grammaires de Stabler.
Dans ce but, nous introduisons une normalisation des preuves de cette logique, normalisation permettant de vérifier la propriété de la sous-formule. Ces propriétés sont également étendues au calcul de Lambek avec produit.

À partir de l'isomorphisme de Curry-Howard, nous synchronisons un calcul sémantique avec les preuves réalisant l'analyse syntaxique. Les termes de notre calcul font appel aux propriétés du lambda mu-calcul, ainsi qu'à celles de la DRT (Discourse Representative Theory).

Une dernière partie applique ces formalismes à des cas concrets. Nous établissons des fragments d'une grammaire du français autour du problème des clitiques.

URL où la thèse pourra être téléchargée :

http://tel.archives-ouvertes.fr/tel-00185844/fr/
http://maxime.amblard.googlepages.com/these

Didier BOURIGAULT (didier.bourigault@univ-tlse2.fr)

Titre : Un analyseur syntaxique opérationnel : SYNTEX

Mots-clés : analyse syntaxique automatique, philosophie de la technique.

Mémoire d'HDR en Sciences du Langage, Université de Toulouse-Le Mirail, CLLE-ERSS – UMR 5263, sous la direction de M. Benoît Habert, Pr, Université Paris 10- Nanterre. Soutenue le 09/06/2007.

Jury : M. Benoît Habert (Pr, Université de Paris X Nanterre, rapporteur), M. Sylvain Kahane (Pr, Université de Paris X Nanterre, rapporteur), Mme Marie-Paule Péry-Woodley (Pr, Université de Toulouse Le Mirail, rapporteur), M. Jean-

Pierre Chanod (Manager, Xerox Research Centre Europe, examinateur), M Jean Véronis (Pr, Université d'Aix-en-Provence, examinateur), M. Bernard Victorri (DR, École Normale Supérieure, examinateur).

Résumé *: Dans ce mémoire d'habilitation à diriger les recherches, nous présentons les recherches que nous avons menées ces dix dernières années autour de la réalisation, l'évaluation et l'utilisation du logiciel Syntex, un analyseur syntaxique automatique du français. Dans la première partie du mémoire, nous retraçons le chemin qui nous a conduit de Lexter, un analyseur syntaxique robuste dédié au repérage des syntagmes nominaux terminologiques dans les corpus spécialisés, à Syntex, un analyseur à plus large couverture. La deuxième partie du mémoire est consacrée à un panorama historique du domaine du traitement automatique des langues, dans lequel nous montrons que les recherches dans ce domaine ont toujours été partagées entre les travaux théoriques et les applications à visée industrielle. Ce panorama est suivi d'une revue de travaux en analyse syntaxique robuste, qui identifie une lignée dans laquelle s'inscrit notre recherche. Dans la troisième partie, nous présentons les concepts clés qui ont guidé la conception de l'analyseur Syntex. L'analyse syntaxique automatique est présentée comme un problème de reconnaissance de formes, représentées par des structures de dépendance syntaxique. Syntex est un analyseur procédural à cascades. Sur le plan épistémologique, il peut être caractérisé comme un objet technique, au sens de la philosophie des techniques de G. Simondon, en tant que ses progrès se développent selon les deux dimensions de l'adaptation et de l'auto-corrélation.*

URL où la thèse pourra être téléchargée :

http://w3.univ-tlse2.fr/erss/textes/pagespersos/bourigault/hdr.html

---

Yayoi NAKAMURA-DELLOYE (yayoi@free.fr)

Titre : Alignement automatique de textes parallèles français-japonais

Mots-clés : alignement, corpus parallèles, analyse morphologique japonaise partielle, mémoire de traduction, analyse syntaxique partielle, proposition syntaxique, subordination, Prolog, CFG, DCG, appariement de graphes, classification ascendante hiérarchique, linguistique contrastive.

Title : *French-Japanese parallel texts automatic alignment*

Keywords : *Alignment, parallel corpora, partial japanese morphological analysis, translation memory, partial syntactic analysis, syntactic clause, subordination, Prolog, CFG, DCG, graph matching, agglomerative hierarchical clustering, contrastive linguistics.*

Thèse de doctorat en Sciences du Langage, Université de Paris 7, Département de linguistique, UMR 8094 Lattice, sous la direction de Mme Catherine Fuchs, Dr, et de Mme Catherine Garnier, Pr, INALCO. Soutenue le 17/12/2007.

Jury : Mme Catherine Fuchs, (Dr, CNRS-Lattice, directrice), Mme Catherine Garnier (Pr, INALCO, codirectrice), M. Philippe LANGLAIS, (Pr, Université de Montréal, rapporteur), M. Yves Lepage (Pr, Université de Caen, rapporteur), M. Pierre Le Goffic (Pr, Université Paris III, examinateur), M. Pierre Zweigenbaum (Dr, CNRS-LIMSI, examinateur).

***Résumé** : L'alignement automatique consiste à trouver une correspondance entre des unités de « textes parallèles » — ensemble de textes de langues différentes, constitué d'un texte original et de ses traductions. L'alignement peut être réalisé à différents niveaux, mais nous nous intéressons plus particulièrement à la réalisation d'un système qui procède à l'alignement au niveau des propositions, unités profitables dans beaucoup d'applications.*

La présente thèse est constituée de deux types de travaux : les travaux introducteurs permettant d'instaurer les bases nécessaires pour notre objectif principal, et ceux constituant le noyau central. Le noyau de la thèse s'articule autour de la notion de proposition syntaxique. Il est composé de deux types de travaux, études linguistiques et réalisations informatiques.

Travaux introducteurs
Les travaux introducteurs comprennent l'étude des généralités sur l'alignement ainsi que des travaux consacrés à l'alignement des phrases, opération élémentaire de tout type d'alignement. Ces travaux ont conduit à la réalisation d'un système d'alignement des phrases adapté au traitement des textes français et japonais. Notre système est caractérisé par l'absence de recours à des moyens extérieurs tels que des analyseurs morphologiques ou des dictionnaires bilingues.

Études linguistiques
La première moitié du noyau de la thèse est consacrée aux études linguistiques qui se divisent elles-mêmes en deux sujets qui constituent chacun une partie indépendante : la proposition en français et la proposition en japonais.
Le but de nos études sur la proposition française est de définir une grammaire pour la détection des propositions, non pas très précise avec calcul d'informations diverses, mais une grammaire simple, efficace et opérationnelle avec des informations disponibles et restreintes. Face aux problèmes posés par les travaux existants qui sont peu adaptés à notre opération de détection des propositions, nous avons cherché à définir une typologie des propositions, basée sur des critères uniquement formels. Nous avons ensuite réalisé un classement des subordonnées selon les critères combinés de catégorie/position. Chaque classe a été décrite de manière systématique et précise, à l'aide des travaux de P. Le Goffic.

Dans les études sur le japonais, nous cherchons à cerner les deux unités importantes pour nos travaux : la phrase et la proposition. Nous définissons d'abord la phrase japonaise sur la base de la thèse que nous tenterons de défendre : la constitution par l'opposition thème-rhème de la structure fondamentale de la phrase japonaise. Nous tentons ensuite d'élucider la notion de proposition en résolvant deux problèmes : distinction des formes du mot variable constituant une proposition de celles formant les syntagmes non propositionnels ; définition des connecteurs assurant la jonction de deux propositions.

Réalisations informatiques

Les réalisations informatiques comportent trois tâches, composant ensemble au final l'opération d'alignement des propositions, incarnées par trois systèmes informatiques distincts : deux détecteurs de propositions (un pour le français et un pour le japonais), ainsi qu'un système d'alignement des propositions.

Toutes nos réalisations informatiques partent d'une étude des techniques existantes. Elles sont évaluées et leurs échecs sont abondamment discutés en vue d'améliorations futures.

Les deux systèmes de détection des propositions du français et du japonais ont été réalisés de manière à identifier les propositions à partir du résultat des systèmes d'analyse syntaxique existants : chunker de Paris 7 pour le français et l'analyseur de relations dépendancielles CaBoCha pour le japonais. Pour le français, nos études linguistiques ont permis de définir une grammaire pour la détection des propositions de type CFG. Cette grammaire a ensuite été réécrite selon le formalisme DCG et incluse dans un module principal développé en Prolog. Notre détecteur de propositions du japonais réalise la reconnaissance par regroupement en cascade des chunks. L'arrêt du regroupement est contrôlé par les critères définis sur la base de notre définition de la proposition issue de nos études linguistiques. Notre système se différencie par rapport au système antérieur par sa capacité de reconnaissance des propositions même imbriquées.

Le système d'alignement des propositions, fruit final de notre thèse, recourt pour sa phase de prétraitement aux deux précédents détecteurs de propositions ainsi qu'à l'aligneur de phrases, développé dans le cadre des travaux introducteurs. Le caractère non parallèle des propositions en relation de traduction dans les textes parallèles français-japonais est un problème crucial pour l'opération d'alignement. De cette observation, nous avons réalisé trois méthodes d'alignement des propositions, caractérisées par leur capacité d'alignement des traductions croisées, ce qui était impossible pour beaucoup de méthodes classiques d'alignement. Les deux premières se basent sur l'approche spectrale d'appariement des graphes inexacts, qui consiste à projeter les graphes sur un sous-espace propre, l'une de ces deux méthodes prenant en compte uniquement les propriétés topologiques des graphes et l'autre traitant les graphes valués par les types de propositions. La troisième méthode, exploitant plus de données disponibles telles que les informations lexicales, est une technique inspirée de la classification ascendante hiérarchique (CAH).

URL où la thèse pourra être téléchargée :

http://yayoi.free.fr/TAL/intro.html

Mehdi YOUSFI-MONOD (mehdi.yousfi@laposte.net)

Titre : Compression automatique ou semi-automatique de textes par élagage des constituants effaçables : une approche interactive et indépendante des corpus

Mots-clés : TALN, résumé automatique, compression de phrases, théorie du gouvernement et du liage, arbre syntaxique, grammaire de constituants, outil interactif.

Title : *Automatic or semi-automatic text compression through removable constituent pruning : an interactive and corpus-free approach*

Keywords : *NLP, automatic summarization, sentence compression, government and binding theory, syntactic tree, constituent grammar, interactive tool.*

Thèse de doctorat en Informatique, Université de Montpellier 2, Département d'informatique, UMR 5506 LIRMM, sous la direction de Mme Violaine Prince, Pr. Soutenue le 16/11/2007.

Jury : Mme Violaine Prince, (Pr, Université Montpellier 2, directrice), M. Jacques Vergne (Pr, Université de Caen, rapporteur), M. Jean-Luc Minel (IR HDR, MoDyCo-CNRS, rapporteur), Mme Augusta Mela (MC, Université de Montpellier 3, examinatrice), M. Juan-Manuel Torres-Moreno (MC, Université Avignon, examinateur), M. Jacques Chauché (Pr, Université Montpellier 2, examinateur).

Résumé *: Le travail s'inscrit dans le domaine du traitement automatique du langage naturel et traite plus spécifiquement d'une application de ce dernier au résumé automatique de textes.*
L'originalité de la thèse consiste à s'attaquer à une variété fort peu explorée, la compression de textes, par une technique non supervisée.
Ce travail propose un système incrémental et interactif d'élagage de l'arbre syntagmatique des phrases, tout en préservant la cohérence syntaxique et la conservation du contenu informationnel important.
Sur le plan théorique, le travail s'appuie sur la théorie du gouvernement de Noam Chomsky et plus particulièrement sur la représentation formelle de la théorie X-barre pour aboutir à un fondement théorique important pour un modèle computationnel compatible avec la compression syntaxique de phrases.

Le travail a donné lieu a un logiciel opérationnel, nommé COLIN, qui propose deux modalités : une compression automatique, et une aide au résumé sous forme semi-automatique, dirigée par l'interaction avec l'utilisateur.
Le logiciel a été évalué grâce à un protocole complexe par vingt-cinq utilisateurs bénévoles.
Les résultats de l'expérience montrent que 1) la notion de résumé de référence qui sert aux évaluations classiques est discutable ; 2) les compressions semi-automatiques ont été fortement appréciées ; 3) les compressions totalement automatiques ont également obtenu de bons scores de satisfaction.
À un taux de compression supérieur à 40 % tous genres confondus, COLIN fournit un support appréciable en tant qu'aide à la compression de textes, ne dépend d'aucun corpus d'apprentissage, et présente une interface conviviale.

URL où la thèse pourra être téléchargée :

http://tel.archives-ouvertes.fr/tel-00185367/fr/