Résumés de thèses

Rubrique préparée par Fiammetta Namer

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

Adrien LARDILLEUX : (Adrien.Lardilleux@info.unicaen.fr**)**

Titre : Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle.

Mots-clés : traitement automatique des langues, hapax, multilinguisme, traduction automatique, alignement, événements rares.

Title : *The contribution of low frequencies to multilingual sub-sentential alignment : a differential approach.*

Keywords : *natural language processing – hapax legomenon – multilingualism – machine translation – alignment – rare events*

Thèse de doctorat en Informatique, Université de Caen, département d'informatique, UFR de SciencesGREYC (UMR 6072), Caensous la direction de Yves Lepage (Pr, Université de Caen Basse-Normandie, Université de Waseda, Japon). Thèse soutenue le 14/09/2010.

Jury : M. Yves Lepage (Pr, Université de Caen Basse-Normandie, Université de Waseda, Japon, directeur), M. Jacques Vergne (Pr, Université de Caen, président et examinateur), M. Christian Boitet (Pr, Université Joseph Fourier-Grenoble I, rapporteur), M. Philippe Langlais (Pr., Université de Montréal, rapporteur), M. François Yvon (Pr, Université Paris-Sud XI, rapporteur), Mme Béatrice Daille (Pr, Université de Nantes, examinatrice), M. Andy Way (Pr associé, Dublin City University, examinateur).

Résumé *: L'objectif de cette thèse est de montrer que, contrairement aux idées reçues, les mots de basses fréquences peuvent être mis à profit de façon efficace en traitement automatique des langues. Nous les mettons à contribution en alignement sous-phrastique, tâche qui constitue la première étape de la plupart des systèmes de traduction automatique fondée sur les données (traduction probabiliste ou par l'exemple). Nous montrons que les mots rares peuvent servir de fondement même dans la conception d'une méthode d'alignement sous-phrastique multilingue, à*

l'aide de techniques différentielles proches de celles utilisées en traduction automatique par l'exemple. Cette méthode est réellement multilingue, en ce sens qu'elle permet le traitement simultané d'un nombre quelconque de langues. Elle est de surcroît très simple, et permet un passage naturel à l'échelle. Nous comparons notre implémentation, Anymalign, à deux ténors statistiques du domaine sur des tâches bilingues. Bien qu'à l'heure actuelle ses résultats soient en moyenne légèrement en retrait par rapport à l'état de l'art en traduction automatique probabiliste par segments, nous montrons que la qualité propre des lexiques produits par notre méthode est en fait supérieure à celle de l'état de l'art.

URL où la thèse pourra être téléchargée :
http : //tel.archives-ouvertes.fr/tel-00520787/fr/

Huei-Chi LIN : (linhueichi@gmail.com**)**

Titre : Un module NooJ pour le traitement automatique du chinois : formalisation du vocabulaire et des têtes des groupes nominaux.

Mots-clés : traitement automatique des langues naturelles, formalisation du chinois, dictionnaire électronique du chinois, description syntaxique des groupes nominaux chinois.

Title : *A NooJ Module for the Automatic Processing of Chinese : Formalising the Chinese Vocabulary and Noun Phrases.*

Keywords : *Natural Language Processing. Automatic Processing of Chinese. Electronic Dictionaries for Chinese. Syntactic Description of Chinese Noun Phrases*

Thèse de doctorat en Sciences du Langage, Université de Franche-Comté, UFR SLHS Laboratoire de sémiotique, linguistique et informatique (LASELDI), Besançon sous la direction de Max Silberztein (Pr, Université de Franche-Comté). Thèse soutenue le 15/06/2010.

Jury : M. Max Silberztein (Pr, Université de Franche-Comté, directeur), Mme Andrée Chauvin-Vileno (Pr, Université de Franche-Comté, présidente), M. Joël Bellassen (Dr, INaLCO, rapporteur), Mme Anaïd Donabédian-Demopoulos (Pr, INaLCO, rapporteur), Mme Zhitang Yang-Drocourt (MC HDR, INaLCO, examinatrice).

Résumé *: Cette étude présente le développement du module d'analyse automatique du chinois qui permet de reconnaître dans les textes les unités lexicales en chinois moderne puis les groupes nominaux noyaux.*

Pour atteindre ces deux objectifs principaux, nous devions résoudre les problèmes suivants :

1) identifier les unités lexicales en chinois moderne ;

2) déterminer leurs catégories ;

3) décrire la structure de syntaxe locale et des groupes nominaux noyaux.

C'est ainsi que nous avons été amenée à constituer d'abord un corpus regroupant des textes littéraires et journalistiques publiés au XX^e siècle. Ces textes sont écrits en chinois moderne avec des caractères traditionnels. Grâce à ces données textuelles, nous avons pu recueillir des informations linguistiques telles qu'unités lexicales, structures syntagmatiques ou règles grammaticales. Ensuite, nous avons construit des dictionnaires électroniques dans lesquels chaque unité lexicale est représentée par une entrée, à laquelle sont associées des informations linguistiques telles que catégories lexicales, classes de distribution sémantique ou descriptions formelles de certaines formes lexicales. À ce stade, nous avons cherché à identifier les unités lexicales du lexique chinois et leurs catégories en les recensant. Grâce à cette liste, l'analyseur lexical peut traiter des unités lexicales de différents types, en bloc, sans les découper en composants. Ainsi, on traite les unités lexicales suivantes comme des unités atomiques :

理髮 lǐfà / fǎ <arranger-cheveux> 'faire la coiffure'

放假 fàngjià <distribuer-vacance> 'être en vacances'

刀子口 dāozikǒu <couteau-bouche> 'parole cruelle'

研究員 yánjiū / jiù yuán <effectuer des recherches-K> 'chercheur'

翻譯系統 fānyì xìtǒng <traduire-système> 'système de traduction'

浪漫主義 làngmàn zhǔyì <romantique- -isme> 'romantisme'

Puis, nous avons décrit de manière formelle un certain nombre de syntagmes locaux, ainsi que cinq types de groupes nominaux noyaux. Enfin, nous avons utilisé le module chinois ainsi développé pour étudier l'évolution thématique dans les textes littéraires.

URL où la thèse pourra être téléchargée :
http : //scd.univ-fcomte.fr

Stéphanie WEISER : (stephanie.weiser@gmail.com)

Titre : Repérage et typage d'expressions temporelles pour l'annotation sémantique automatique de pages Web – Application au e-tourisme.

Mots-clés : extraction automatique d'information, ontologie, schéma d'annotation, expressions temporelles, e-tourisme, transducteurs.

Title : *Extraction and mark-up of temporal expressions for automatic semantic annotation of Web pages – Application to E-tourism*

Keywords : *Automatic information extraction, ontology, annotation scheme, temporal expressions, e tourism, transducers*

Thèse de doctorat en Sciences du Langage, option traitement automatique des langues, Université de Paris-Ouest Nanterre la Défense, école doctorale connaissance, langage, modélisation, Laboratoire MoDyCo – UMR7114, Nanterresous la codirection de Jean-Luc Minel (Pr, Université de Paris Ouest) et Philippe Laublet (MC, Université de Paris Ouest). Thèse soutenue le 30/06/2010.

Jury : M. Jean-Luc Minel (Pr, Université de Paris Ouest, directeur), M. Philippe Laublet (MC, Université de Paris Ouest, codirecteur), M. Eric Laporte (Pr, Université Marne-la-Vallée, président et rapporteur), M. Cédric Fairon (Pr, Université Catholique de Louvain, rapporteur), Mme Delphine Battistelli (MC HDR, Université Paris IV Sorbonne, examinatrice), Mme Florence Armadeilh (MC, Université Paris Ouest, examinatrice).

Résumé *: Cette thèse présente Adetoa, système dédié au repérage et à l'annotation sémantique automatique d'expressions temporelles dans des pages Web pour une application du e-tourisme. La réalisation de ce système de TAL s'appuie sur une étude linguistique détaillée menée à partir d'une réflexion générale sur l'expression de la temporalité dans ce type de textes. Cette étude, réalisée sur des cas réels, a permis de mettre en évidence la complexité des formes linguistiques ayant une double spécificité : elles se trouvent sur des pages Web et sont propres au domaine du tourisme. Présentée dans les premiers chapitres, elle est à la base d'Adetoa qui s'intègre dans la plate-forme du projet Eiffel pour laquelle les modèles ontologiques et les langages du Web sémantique pour la représentation des connaissances et la recherche sémantique d'information sont utilisés.*

Sur un plan linguistique, les contenus ont des particularités propres : les informations temporelles apparaissent rarement dans un texte rédigé, la syntaxe du français n'est pas toujours respectée et il y a peu de prédications. Leur placement dans les pages Web, organisées de manière fort variée, ne présente aucune régularité, rendant difficile voire parfois impossible l'automatisation de leur analyse, comme l'a montré l'analyse sémiotique des pages (par opposition par exemple avec les guides touristiques papier).

Le développement d'Adetoa s'est nourri de ces études théoriques. L'analyse linguistique a permis de construire un ensemble important de transducteurs (avec Unitex) pour les tâches de repérage et d'annotation des expressions temporelles, ce

qui constitue une ressource pouvant être généralisée. De plus, d'autres informations du domaine touristique sont repérées : les objets du tourisme et les adresses. Des transducteurs de liage permettent de grouper toutes les informations concernant une même offre touristique.

Un schéma d'annotation a été mis au point. Il est lié à une ontologie du tourisme, mais n'en est pas un calque direct car sa finalité est de rester au plus près des expressions linguistiques de manière à les caractériser finement. Pour l'intégration d'Adetoa au sein de la chaîne de traitement d'Eiffel, des règles de transformation permettant de faire le pont entre les expressions annotées et les données à stocker dans la base de connaissances ont été élaborées. Dans le cadre de cette thèse, parallèlement à ces développements, l'ontologie du projet a été adaptée de manière à ce que les données annotées puissent prendre place dans la base de connaissance qui lui correspond.

L'évaluation d'Adetoa, présentée dans le dernier chapitre, a montré des résultats satisfaisants aussi bien d'un point de vue théorique que pour cette application industrielle.

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