Résumés de thèses

Rubrique préparée par Fiammetta Namer

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

Aurélien BOSSARD : (aurelien.bossard@gmail.com)

Titre : Contribution au résumé automatique multi-documents.

Mots-clés : résumé automatique, extraction d'information, traitement automatique de la langue.

Title: *Contribution to Automatic Multi-Document Summarization.*

Keywords: *automatic summarization, information extraction, natural language processing.*

Thèse de doctorat en Informatique, Institut Galilée, Laboratoire d'Informatique de Paris-Nord & CNRS, UMR 7030, Université Paris-Nord, Villetaneusesous la direction de Thierry Poibeau (CR HDR, CNRS-LaTTiCe). Thèse soutenue le 12/07/2010.

Jury : M. Thierry Poibeau (CR HDR, CNRS-LaTTiCe, directeur), Mme Anne Vilnat (Pr, LIMSI, présidente), M. Juan-Manuel Torres Moreno (MC HDR, LIA, Université d'Avignon et des Pays de Vaucluse, rapporteur), M. Guy Lapalme (Pr, Université de Montréal, rapporteur), Mme Céline Rouveirol (Pr, LIPN-Université Paris 13, examinatrice), M. Daniel Kayser (Pr, LIPN-Université Paris 13, examinateur).

Résumé *: Résumer un texte consiste à réduire ce texte en un nombre limité de mots. Le texte ainsi réduit doit rester fidèle aux informations et idées du texte original. Que ce soit pour des professionnels qui doivent prendre connaissance du contenu de documents en un temps limité ou pour un particulier désireux de se renseigner sur un sujet donné sans disposer du temps nécessaire pour lire l'intégralité des textes qui en traitent, le résumé est une aide contextuelle importante. Avec l'augmentation de la masse documentaire disponible électroniquement, résumer des textes automatiquement est devenu un axe de recherche important dans le domaine du traitement automatique de la langue. La production automatique de résumés pose le problème de la détection et de la modélisation des informations contenues*

dans les textes. Elle suppose également la hiérarchisation de ces informations afin d'intégrer les plus importantes au sein du résumé. Cette thèse de doctorat propose une méthode statistique pour le résumé automatique par extraction ainsi que l'intégration d'analyses linguistiques au processus de sélection de phrases.

La méthode que nous proposons est fondée sur une classification des phrases à résumer en classes sémantiques en utilisant des calculs de similarité entre les phrases. Cette étape nous permet d'identifier les phrases qui risquent de présenter des éléments d'information similaires et ainsi de supprimer toute redondance du résumé généré. Une seconde étape vise à sélectionner une phrase par classe, en tenant compte de la similarité des phrases à une éventuelle requête utilisateur, de la longueur des phrases ainsi que de la centralité dans leur classe. Les résumés ainsi générés doivent maximiser la centralité et la diversité des informations.

Cette méthode a été évaluée sur deux tâches de la campagne d'évaluation TAC 2008 : le résumé de dépêches et le résumé d'opinions issues de blogs. Les résultats mitigés sur la première tâche et encourageants sur la seconde nous ont poussés à prendre en compte des critères de sélection de phrases spécifiques aux types de documents traités. Nous avons alors proposé d'établir une catégorisation des dépêches de presse ainsi que l'annotation automatique de leur structure afin d'améliorer la qualité des résumés générés par notre système. Nous avons également étudié l'apport de l'annotation en entités nommées et de la résolution d'anaphores pour le résumé automatique. Le système et ces trois derniers modules ont été évalués sur la tâche de résumé et mise à jour de résumés de dépêches de la campagne TAC 2009, se classant dans le premier quart des participants. Si l'intégration de la structure a fait évoluer positivement la qualité des résumés, l'intégration de l'annotation en entités nommées et de la résolution d'anaphores s'est révélée décevante, sans doute à cause de la qualité insuffisante des annotations.

Notre méthode de résumé a également fait l'objet d'une intégration à un système applicatif plus large visant à aider un possesseur de corpus à visualiser les axes essentiels et à en retirer automatiquement les informations importantes.

URL où la thèse pourra être téléchargée : http://www-lipn.univ-paris13.fr/~abossard/indexpublicationsfr.php

---

Cécile FABRE : (cecile.fabre@univ-tlse2.fr)

Titre : Affinités syntaxiques et sémantiques entre mots : apports mutuels de la linguistique et du TAL.

Mots-clés : linguistique de corpus, analyse syntaxique automatique, complémentation verbale, acquisition de relations sémantiques, lexique et discours.

Title: *Syntactic and semantic affinities between words: mutual contribution of linguistics and NLP.*

Keywords: *corpus linguistics, parsing, verb complementation, acquisition of semantic relations, lexicon and discourse.*

HDR en Sciences du Langage, Université de Toulouse 2 Le Mirail, département de Sciences du Langage, CLLE-ERSS UMR 5263, Toulouse, sous la direction de Marie-Paule Péry-Woodley (Pr, Université Toulouse2). HDR soutenu le 29/11/2010.

Jury : Mme Marie-Paule Péry-Woodley (Pr, Université de Toulouse 2, directrice), M. Benoît Habert (Pr, ENS de Lyon, président), Mme Adeline Nazarenko (Pr, Université Paris 13, rapporteur), M. Alain Polguère (Pr, Université Nancy2, rapporteur), M. Nabil Hathout (CR1, CLLE-CNRS, examinateur), Mme Paola Merlo (maître d'enseignement et de recherche, Université de Genève, examinatrice), Mme Agnès Tutin (MC HDR, Université de Grenoble 3, examinatrice).

Résumé *: Je présente un bilan des travaux que j'ai menés depuis mon recrutement à l'Université de Toulouse 2 - Le Mirail (1997) dans le domaine de la linguistique et du traitement automatique des langues (TAL). J'ai exploré le lien entre ces deux disciplines de deux façons que j'estime complémentaires : tout d'abord, je considère le champ applicatif du TAL comme un terrain d'investigation important pour la linguistique. Le TAL, et, de façon générale, les applications relevant du domaine de l'ingénierie des langues, sollicitent un renouvellement des objets d'étude de la linguistique et élargissent le champ de ses questionnements. En retour, la linguistique gagne à s'appuyer sur des procédures de découverte issues du TAL, fondées sur le traitement de corpus numérisés et annotés et sur le recours à des techniques de quantification adaptées aux besoins de la description linguistique. Au sein de ce cadre général, les travaux que j'ai menés ont porté principalement sur deux thématiques de recherche que j'ai résumées sous les termes d'affinités sémantiques et syntaxiques. Le premier concerne la question du repérage des rapports de proximité sémantique entre différents types d'unités (mots, termes, structures prédicatives). Identifier sous la diversité des formulations des éléments de contenu similaire est un objectif crucial pour de nombreuses applications qui visent l'accès à l'information dans les textes. Dans cette perspective, j'ai cherché à considérer sur le plan linguistique cette question de la proximité sémantique, en faisant en particulier appel à des techniques d'analyse distributionnelle automatique qui visent à calculer les rapprochements sémantiques entre mots sur la base de la similarité de leur comportement syntaxique dans les corpus. Cette approche inductive des relations de sens déborde largement les limites des relations classiquement décrites en linguistique et sollicite des procédures nouvelles de description et de validation. Le second volet concerne la question des*

affinités syntaxiques entre mots : impliquée dans le projet de développement et d'exploitation d'un analyseur syntaxique automatique, Syntex, je me suis intéressée à une question qui est au cœur des problèmes d'ambiguïté syntaxique, à savoir le rattachement des groupes prépositionnels. J'ai travaillé en particulier à la mise au point d'une méthode permettant de distinguer des types différents de rattachement prépositionnel, de nature argumentale ou adjonctive. Dans ce cas également, mon travail est guidé par un objectif qui relève du TAL (améliorer les performances d'un analyseur), et ce projet m'a amenée en retour à retravailler une question linguistique centrale en syntaxe, la distinction entre arguments et circonstants, et à développer des méthodes d'analyse de corpus qui permettent de substituer à une conception binaire de ces notions une appréciation plus graduelle de l'autonomie du groupe prépositionnel par rapport au verbe. Je propose donc de montrer comment les outils de TAL appliqués aux corpus offrent à la linguistique des instruments d'observation et d'expérimentation qui permettent d'aborder les faits langagiers par le biais de l'observation des usages et sous l'angle de la quantification. Ma conviction est que la linguistique ainsi outillée peut jouer un rôle plus important sur les nombreux terrains applicatifs qui nécessitent l'analyse de données langagières.

URL où la thèse pourra être téléchargée :
http://hal.archives-ouvertes.fr/tel-00552097/

Nicolas OBIN (nobin@ircam.fr**)**

Titre : MeLos : analyse et modélisation de la prosodie et du style de parole.

Mots-clés : prosodie, style de parole, synthèse de la parole, modèle de Markov caché (HMM) à observation discrète/continue, stylisation, modèle de trajectoire, linguistique.

Title: *MeLos: Analysis and Modelling of Speech Prosody and Speaking Style.*

Keywords: *speech prosody, speaking style, speech synthesis, discrete/continuous HMMs, stylization, trajectory modelling, linguistics.*

Thèse de doctorat en Sciences du Langage (traitement du signal), IRCAM, équipe analyse et synthèse des sons, Université Paris 6, Parissous la codirection de Xavier Rodet (DR, IRCAM) et Anne Lacheret (Pr, Modyco, Université Paris-Ouest Nanterre). Thèse soutenue le 23/06/2011.

Jury : M. Xavier Rodet (DR, IRCAM, codirecteur) Anne Lacheret (Pr, Modyco, Université Paris-Ouest Nanterre, codirectrice), M. Nick Campbell (Pr, CLCS &

University of Dublin, rapporteur), M. Simon King (Pr, CSTR & University of Edimburgh, rapporteur), M. Jean-François Bonastre (Pr, LIA & Université d'Avignon, examinateur), M. Éric de la Clergerie (Dr, INRIA-ALPAGE, examinateur), M. David Wessel (Pr, CNMAT & University of California Berkeley, examinateur), M. Jean-Luc Zarader (Pr, ISIR & Université de Paris 6, examinateur).

Résumé : *Cette thèse a pour objet la modélisation de la prosodie dans le cadre de la synthèse de la parole. Nous présentons MeLos : un système complet d'analyse et de modélisation de la prosodie, « la musique de la parole ». C'est un système unifié fondé sur des modèles de Markov cachés (HMMs) à observation discrète/continue pour modéliser les caractéristiques symbolique et acoustique de la prosodie. Il comporte :*
1) une chaîne de traitement linguistique de surface et profonde, introduite pour enrichir la description des caractéristiques du texte ;
2) un modèle segmental associé à la fusion de Dempster-Shafer, utilisé pour combiner les contraintes linguistique et métrique dans la production des pauses ;
3) un modèle de trajectoire fondé sur la stylisation des contours prosodiques, qui permet de modéliser simultanément les variations à court et long terme de la F0.
Le système proposé est utilisé pour modéliser les stratégies et le style d'un locuteur. Il est étendu à la modélisation du style de parole par des méthodes de modélisation en contexte partagé et de normalisation du locuteur.

Abstract*: This thesis addresses the issue of modelling speech prosody for speech synthesis and presents MeLos: a complete system for the analysis and modelling of speech prosody, "the music of speech".*

Research into the analysis and modelling of speech prosody has increased dramatically in recent decades, and speech prosody has emerged as a crucial concern for speech synthesis. The issue of speech prosody modelling is to model speech prosody variations depending on the context - linguistic (e.g. linguistic structure), para-linguistic (e.g., emotion), or extra-linguistic (e.g., socio-geographical origins, situation of a communication). Modelling the variability of speech prosody is required to provide natural, expressive, and varied speech in many applications of high-quality speech synthesis such as multi-media (avatars, video games, story telling, dialogue systems) and artistic (cinema, theatre, music, digital arts) applications. The objective of this thesis is to model the strategy, alternatives, and speaking style of a speaker for natural, expressive, and varied speech synthesis. The present study presents original contributions with special attention paid to the combination of theoretical linguistic and statistical modelling to provide a complete speech prosody system. A unified discrete/continuous context-dependent HMM is presented to model the symbolic and the acoustic characteristics of speech prosody:

1) A rich description of the text characteristics based on a linguistic processing chain that includes surface and deep syntactic parsing is proposed to refine the

modelling of the speech prosody in context.

2) Segmental HMMs and Dempster-Shafer fusion are used to balance linguistic and metric constrains in the production of a pause.

3) A trajectory model is proposed based on the stylization and the simultaneous modelling of short and long-term F0 variations over various temporal domains.

The proposed system is used to model the strategies, alternatives and speaking style of a speaker, and is extended to model the speaking style of any arbitrary number of speakers using shared-context-dependent modelling and speaker normalization techniques.

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

Fabien POULARD : (fabien.poulard@univ-nantes.fr)

Titre : Détection de dérivation de texte.

Mots-clés : détection de dérivation, révisions, plagiat, approche par signature, mesures de similarité, recherche d'information.

Title: Detecting textual derivatives.

Keywords: detection of derivation, revisions, plagiarism, signature approach, similarity metrics, information retrieval

Thèse de doctorat en Informatique, Université de Nantes – UFR des Sciences et Techniques, LINA – UMR CNRS 6241, Nantes, sous la direction de Béatrice Daille (Pr, Université de Nantes). Thèse soutenue le 24/03/2011.

Jury : Mme Béatrice Daille (Pr, Université de Nantes, directrice), Mme Josiane Mothe (Pr, IUFM de Toulouse, directrice), M. François Yvon (Pr, LIMSI & Université Paris 11, rapporteur), M. Patrice Bellot (MC, Université d'Avignon, rapporteur), M. Claude de Loupy (directeur, Syllabs, examinateur), M. Nicolas Hernandez (MC, Université of Nantes, examinateur).

Résumé : L'Internet permet la production et la diffusion de contenu sans effort et à grande vitesse. Cela pose la question du contrôle de leur origine. Ce travail s'intéresse à la détection des liens de dérivation entre des textes. Un lien de dérivation unit un texte dérivé et les textes préexistants à partir desquels il a été écrit. Nous nous sommes concentrés sur la tâche d'identification des textes dérivés étant donné un texte

source, et ce pour différentes formes de dérivations. Notre première contribution consiste en la définition d'un cadre théorique posant les concepts de la dérivation ainsi qu'un modèle multidimensionnel cadrant les différentes formes de dérivations. Nous avons ensuite mis en place un cadre expérimental constitué d'une infrastructure logicielle libre, de corpus d'évaluation et d'un protocole expérimental inspiré de la RI. Les corpus Piithie et Wikinews que nous avons développés sont, à notre connaissance, les seuls corpus en français pour la détection de dérivation. Finalement, nous avons exploré différentes méthodes de détections fondées sur l'approche par signature. Nous avons notamment introduit les notions de singularité et d'invariance afin de guider le choix des descripteurs utilisés pour la modélisation des textes en vue de leur comparaison. Nos résultats montrent que le choix motivé des descripteurs, linguistiques notamment, permet de réduire la taille de la modélisation des textes et, par conséquent, des coûts de la méthode, tout en offrant des performances comparables à l'approche de l'état de l'art beaucoup plus volumineuse.
.

URL où la thèse pourra être téléchargée :
http://www.fabienpoulard.info/download/Recherche/These/these.pdf

---

Elsa TOLONE : (elsa.tolone@univ-paris-est.fr**)**

Titre : Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français.

Mots-clés : traitement automatique des langues, ressources linguistiques, lexiques syntaxiques, Lexique-Grammaire, analyse syntaxique, évaluation.

Title: *Parsing with French Lexicon-Grammar tables.*

Keywords: *Natural Language Processing, language resources, syntactic lexica, Lexicon-Grammar, parsing, evaluation .*

Thèse de doctorat en Informatique Linguistique, Université de Paris Est, département d'Informatique, UMR 8049 LIGM, Marne-la-Vallée, sous la direction d'Éric Laporte (Pr, Université de Paris Est). Thèse soutenue le 31/03/2011.

Jury : M. Éric Laporte (Pr, Université de Paris Est, directeur), M. Denys Duchier (Pr, Université d'Orléans, président), Mme Laurence Danlos (Pr, Université de Paris 7, rapporteur), Mme Laura Kallmeyer (Pr, Université de Düsseldorf, rapporteur), M. Éric de la Clergerie (CR, INRIA Paris-Rocquencourt, examinateur), M. Mathieu Constant (MC, Université Paris-Est, examinateur).

Résumé *: Les tables du Lexique-Grammaire, dont le développement a été initié par M. Gross (1975), constituent un lexique syntaxique très riche pour le français. Elles couvrent diverses catégories lexicales telles que les verbes, les noms, les adjectifs et les adverbes. Cette base de données linguistiques n'est cependant pas directement exploitable informatiquement car elle est incomplète et manque de cohérence. Chaque table regroupe un certain nombre d'entrées jugées similaires car elles acceptent des propriétés communes. Ces propriétés ont pour particularité de ne pas être codées dans les tables même, mais uniquement décrites dans la littérature. Pour rendre ces tables exploitables, il faut expliciter les propriétés intervenant dans chacune d'entre elles. De plus, un grand nombre de ces propriétés doivent être renommées dans un souci de cohérence.*

Notre objectif est d'adapter les tables pour les rendre utilisables dans diverses applications de traitement automatique des langues (TAL), notamment l'analyse syntaxique. Nous expliquons les problèmes rencontrés et les méthodes adoptées pour permettre leur intégration dans un analyseur syntaxique. Nous proposons LGExtract, un outil générique pour générer un lexique syntaxique pour le TAL à partir des tables du Lexique-Grammaire. Il est relié à une table globale dans laquelle nous avons ajouté les propriétés manquantes et à un unique script d'extraction incluant toutes les opérations liées à chaque propriété devant être effectuées pour toutes les tables. Nous présentons également LGLex, le nouveau lexique syntaxique généré des verbes, des noms prédicatifs, des expressions figées et des adverbes. Ensuite, nous montrons comment nous avons converti les verbes et les noms prédicatifs de ce lexique au format Alexina, qui est celui du lexique Lefff (Lexique des formes fléchies du français) (Sagot, 2010), un lexique morphologique et syntaxique à large couverture et librement disponible pour le français. Ceci permet son intégration dans l'analyseur syntaxique FRMG (French MetaGrammar) (Thomasset et de la Clergerie, 2005), un analyseur profond à large couverture pour le français, fondé sur les grammaires d'arbres adjoints (TAG), reposant habituellement sur le Lefff. Cette étape de conversion consiste à extraire l'information syntaxique codée dans les tables du Lexique-Grammaire. Nous présentons les fondements linguistiques de ce processus de conversion et le lexique obtenu. Nous évaluons l'analyseur syntaxique FRMG sur le corpus de référence de la campagne d'évaluation d'analyseurs du français Passage (Produire des annotations syntaxiques à grande échelle) (Hamon et al.*, 2008), en comparant sa version fondée sur le Lefff avec notre version reposant sur les tables du Lexique-Grammaire converties.*

URL où la thèse pourra être téléchargée :
http://www-igm.univ-mlv.fr/~tolone/phd.pdf