Résumés de thèses

Rubrique préparée par Fiammetta Namer

Université de Lorraine, UMR « ATILF »
Fiammetta.Namer@univ-lorraine.fr

Romain DEVEAUD : romain.deveaud@gmail.com

Titre : Vers une représentation du contexte thématique en Recherche d'Information

Mots-clés : recherche d'information, contextualisation, concepts implicites, modélisation thématique probabiliste, sources d'information, retour de pertinence simulé, modèles de pertinence, TREC.

***Title**: Generative models of topical context for Information Retrieval*.

***Keywords:** Information retrieval, contextualization, latent concepts, probabilistic topic modeling, information sources, pseudo-relevance feedback, relevance models, TREC.*

Thèse de doctorat en Informatique, Laboratoire d'Informatique d'Avignon (LIA), Centre d'enseignement et de recherche en informatique (CERI), Université d'Avignon et des Pays de Vaucluse, Avignon, sous la direction de Pierre Bellot (Pr, Université d'Aix-Marseille) et Éric San Juan (MC, Université d'Avignon). Thèse soutenue le 29/11/2013.

Jury : M. Éric SanJuan (MC, Université d'Avignon, codirecteur), M. Patrice Bellot (Pr, Université d'Aix-Marseille, codirecteur), Mme Josiane Mothe (Pr, ESPE Toulouse, présidente), M. Jian-Yun Nie (Pr, Université de Montréal, rapporteur), M. Philippe Mulhem (CR-HDR, LIG, Grenoble, rapporteur), M. Jaap Kamps (Associate Professor, University of Amsterdam, examinateur), M. Benjamin Piwowarski (CR, LIP6, examinateur).

***Résumé* :** *Quand des humains cherchent des informations au sein de bases de connaissances ou de collections de documents, ils utilisent un système de recherche d'information (SRI) faisant office d'interface. Les utilisateurs doivent alors transmettre au SRI une représentation de leur besoin d'information afin que celui-ci puisse trouver des documents contenant des informations pertinentes. De nos jours, la représentation du besoin d'information est constituée d'un petit ensemble de*

mots-clés plus souvent connu sous la dénomination de « requête ». Or, quelques mots peuvent ne pas être suffisants pour représenter précisément et efficacement l'état cognitif complet d'un humain par rapport à son besoin d'information initial. Sans une certaine forme de contexte thématique complémentaire, le SRI peut ne pas renvoyer certains documents pertinents exprimant des concepts n'étant pas explicitement évoqués dans la requête.
Dans cette thèse, nous explorons et proposons différentes méthodes statistiques, automatiques et non supervisées pour la représentation du contexte thématique de la requête. Plus spécifiquement, nous cherchons à identifier les différents concepts implicites d'une requête formulée par un utilisateur sans qu'aucune action de sa part ne soit nécessaire. Nous expérimentons pour cela l'utilisation et la combinaison de différentes sources générales d'information représentant les grands types d'information auxquels nous sommes confrontés quotidiennement sur Internet. Nous tirons également parti d'algorithmes de modélisation thématique probabiliste (tels que l'allocation de Dirichlet latente) dans le cadre d'un retour de pertinence simulé. Nous proposons, par ailleurs, une méthode permettant d'estimer conjointement le nombre de concepts implicites d'une requête ainsi que l'ensemble de documents pseudo-pertinents le plus approprié afin de modéliser ces concepts. Nous évaluons nos approches en utilisant quatre collections de tests TREC de grande taille. En annexes, nous proposons également une approche de contextualisation de messages courts exploitant des méthodes de recherche d'information et de résumé automatique.

URL où la thèse pourra être téléchargée :
romaindeveaud.info/publis/thesis_romain_deveaud.pdf

Nicolas FOUCAULT : foucault@limsi.fr

Titre : Questions-Réponses en domaine ouvert : sélection pertinence de documents en fonction du contexte de la question

Mots-clés : traitement automatique des langues, questions-réponses, recherche d'information, Ritel, Quaero, sélection de documents, modèlisation des langues, classification de pages Web, segmentation de pages Web, apprentissage automatique.

***Title**: Open domain question-answering : relevant document selection geared to the question*

***Keywords:** Natural Language Processing, Question & Answering, Information Retrieval, Quaero, Ritel, document selection, language modeling, web page classification, web page segmentation, machine learning.*

Thèse de doctorat en Informatique, spécialité Traitement automatique des langues naturelles, LIMSI-CNRS, Département d'Informatique, Université Paris-Sud, Orsay, sous la direction de Sophie Rosset (DR, LIMSI) et Gilles Ada (IR-HC, LIMSI). Thèse soutenue le 16/12/2013.

Jury : Mme Sophie Rosset (DR, LIMSI, codirectrice), M. Gilles Ada (IR-HC, LIMSI, codirecteur), Mme Brigitte Grau (Pr, ENSIIE, présidente), Mme Pascale Sébillot (Pr, INSA, rapporteur), M. Patrice Bellot (Pr, Université Aix-Marseille, rapporteur), Mr Thierry Baccino (Pr, Université Paris 8, examinateur).

Résumé **:** *Les problématiques abordées dans cette thèse sont de définir une adaptation unifiée entre la sélection des documents et les stratégies de recherche de la réponse à partir du type des documents et de celui des questions, d'intégrer la solution au système de Questions-Réponses (QR) RITEL du LIMSI et d'évaluer son apport. Nous développons et étudions une méthode fondée sur une approche de recherche d'information pour la sélection de documents en QR. Celle-ci s'appuie sur un modèle de langue et un modèle de classification binaire de textes en catégories, pertinent ou non pertinent d'un point de vue QR. Cette méthode permet de filtrer les documents sélectionnés pour l'extraction de réponses par un système QR. Nous présentons la méthode et ses modèles, et la testons dans le cadre QR à l'aide de RITEL. L'évaluation est faite en français en contexte Web sur un corpus de 500 000 pages Web et de questions factuelles fournies par le programme Quaero. Celle-ci est menée soit sur des documents complets, soit sur des segments de documents. L'hypothèse suivie est que le contenu informationnel des segments est plus cohérent et facilite l'extraction de réponses. Dans le premier cas, les gains obtenus sont faibles comparés aux résultats de référence (sans filtrage). Dans le second cas, les gains sont plus élevés et confortent l'hypothèse, sans pour autant être significatifs. Une étude approfondie des liens existant entre les performances de RITEL et les paramètres de filtrage complète ces évaluations.*

Le système de segmentation créé pour travailler sur des segments est détaillé et évalué. Son évaluation nous sert à mesurer l'impact de la variabilité naturelle des pages Web (en taille et en contenu) sur la tâche QR, en lien avec l'hypothèse précédente.

En général, les résultats expérimentaux obtenus suggèrent que notre méthode aide un système QR dans sa tâche. Cependant, de nouvelles évaluations sont à mener pour rendre ces résultats significatifs, et notamment en utilisant des corpus de questions plus importants.

URL où la thèse pourra être téléchargée : tel.archives-ouvertes.fr/tel-00944622

Rima HARASTANI : rima.harastani@gmail.com

Titre : Alignement lexical en corpus comparables : le cas des composés savants et des adjectifs relationnels

Mots-clés : corpus comparables, langue de spécialité, alignement multilingue, composés savants, adjectifs relationnels.

Title*: Lexical Alignment from comparable corpora : neoclassical compounds and relational adjectives*.

Keywords: *comparable corpora, specialized language, multilingual alignment, neoclassical compounds, relational adjectives.*

Thèse de doctorat en Informatique, LINA, UFR Sciences et Techniques, Université de Nantes, Nantes, sous la direction de Béatrice Daille (Pr, Université de Nantes) et d'Emmanuel Morin (Pr, Université de Nantes). Thèse soutenue le 10/02/2014.

Jury : Mme Béatrice Daille (Pr, Université de Nantes, codirectrice), M. Emmanuel Morin (Pr, Université de Nantes, codirecteur), M. Holger Schwenk (Pr, Université du Maine, président), M. Hervé Blanchon (MC, Université Pierre-Mendès-France, rapporteur), M. Ulrich Heid (Pr, Université de Stuttgart, rapporteur), M. Vincent Claveau (CR, CNRS-IRISA, examinateur).

Résumé : *Notre travail concerne l'extraction automatique d'une liste de termes alignés avec leurs traductions (c'est-à-dire un lexique bilingue spécialisé) à partir d'un corpus comparable dans un domaine de spécialité. Un corpus comparable comprend des textes écrits dans deux langues différentes sans aucune relation de traduction entre eux mais dont les textes appartiennent à un même domaine. Les contributions de cette thèse portent sur l'amélioration de la qualité d'un lexique bilingue spécialisé extrait à partir d'un corpus comparable. Nous proposons des méthodes consacrées à la traduction de deux types de termes, qui ont des caractéristiques en commun entre plusieurs langues ou qui posent, de par leur nature, des problèmes pour la traduction : les composés savants (termes contenant au moins une racine empruntée au grec ou au latin) et les termes composés d'un nom et d'un adjectif relationnel. Nous développons également une méthode, qui exploite des contextes riches en termes spécifiques au domaine du corpus, pour réordonner dans un lexique bilingue spécialisé des traductions candidates fournies pour un terme. Les expériences sont réalisées en utilisant deux corpus comparables spécialisés, le premier dans le domaine du cancer du sein et le deuxième dans le domaine des énergies renouvelables, en français, anglais, allemand et espagnol.*

URL où la thèse pourra être téléchargée : tel.archives-ouvertes.fr/tel-00949025

Gaël LEJEUNE : gael.lejeune@unicaen.fr

Titre : Veille épidémiologique multilingue : une approche parcimonieuse au grain caractère fondée sur le genre textuel

Mots-clés : traitement du langage naturel, multilinguisme, recherche d'information, extraction d'information.

Title*: Multilingual epidemic surveillance : a parsimonious caracter-based approach*.

Keywords: *Natural Language Processing; Information Extraction; Multilingualism; Information Retrieval.*

Thèse de doctorat en Informatique, GREYC UMR 6072, UFR Sciences, Université de Caen, Caen, sous la direction de Nadine Lucas (CR-HDR, GREYC) et d'Antoine Doucet (MC, Université de Caen). Thèse soutenue le 16/10/2013.

Jury : Mme Nadine Lucas (CR-HDR, GREYC, codirectrice), M. Antoine Doucet (MC, Université de Caen, codirecteur), Mme Florence Sèdes (Pr, Université de Toulouse 3, présidente), M. Luigi Lancieri (Pr, Université de Lille 1, rapporteur), M. Gabriel Pereira Lopes (Pr, Université Nouvelle de Lisbonne, rapporteur), M. Gaël Dias (Pr, Université de Caen, examinateur), Mme Natalia Grabar (CR, CNRS-STL, examinatrice)., M. Ludovic Tanguy (CR-HDR, CLLE-ERSS-CNRS, examinateur).

Résumé : *Ce travail de thèse porte sur la définition et l'application de méthodes de traitement automatique des langues adaptées au traitement de corpus multilingues. L'application principale visée est la collecte en temps réel d'informations sur un domaine spécifique : la détection de la propagation de maladies infectieuses à partir d'articles de presse. Dans ce domaine, il est particulièrement dommageable de devoir attendre qu'un article en anglais (ou dans une autre langue de grande diffusion) signale une épidémie avant de pouvoir réagir. L'émission d'une alerte dès le premier article publié en langue vernaculaire serait donc une avancée considérable. Seulement, les systèmes spécialisés dans la veille épidémiologique ne sont pas, à l'heure actuelle, véritablement multilingues : ils sont en définitive constitués par parallélisation de systèmes monolingues. La majorité des étapes de traitement étant entièrement dépendantes de la langue traitée, la factorisation des procédures est très limitée. En effet, les modules d'analyse locale (lemmatiseurs, étiqueteurs...) sont spécifiques à une langue. Or, si de tels analyseurs sont disponibles pour certaines langues de grande diffusion, pour la plupart des langues, ces analyseurs restent à créer. Traiter une langue supplémentaire devient alors chaque fois un peu plus coûteux, en temps ou en ressources financières, à mesure que l'on cherche à traiter des langues « moins dotées en ressources ». Les systèmes*

ainsi créés sont en fait « multi-monolingues », la progression se fait pas à pas, langue après langue.

A contrario*, nous mettons en avant dans la thèse un système multilingue « par essence » où le coût marginal de traitement d'une nouvelle langue serait quasi nul. C'est par une approche textuelle, fondée sur des invariants du genre textuel, que nous avons décidé d'attaquer le problème de la variété des langues. Parmi ces invariants figurent la structuration interne des documents et les stratégies d'écriture utilisées par les journalistes : le « style collectif ». L'approche utilisée se fonde sur les stratégies de communication utilisées dans ces textes.*

*Le système ainsi développé comporte un noyau central indépendant de la langue traitée. Ce système, baptisé DAnIEL (*Data Analysis for Information Extraction in any Language *[analyse de données pour l'extraction d'information multilingue]), utilise des règles très robustes sur le plan multilingue, et s'appuie sur une description minimale de la langue. Le coût marginal de traitement d'une nouvelle langue se limite à la création d'une liste de quelques dizaines de noms usuels de maladies, sans contrainte d'organisation sous forme d'ontologie. Les noms scientifiques des maladies sont utilisés comme supplétifs à l'échelle du texte. Les résultats obtenus par DAnIEL sont très prometteurs, le système est efficace dans trois sous-tâches de la veille épidémiologique :*

– le filtrage des documents, aboutissant à une classification « pertinents vs *non pertinents » ;*
– le typage, à des fins d'alerte, des événements épidémiologiques qu'ils décrivent ;
– le regroupement de ces événements, afin de limiter la redondance pour l'utilisateur final.

DAnIEL complète la couverture des systèmes actuels en permettant de traiter à moindre coût des langues à morphologie riche (grec, finnois, polonais) ou munies d'un système d'écriture différent (chinois, arabe, hindi). DAnIEL a été confronté à un corpus annoté par des humains contenant un peu plus de 2 000 documents en cinq langues (anglais, chinois, grec, polonais, russe). Ce corpus a permis de montrer que DAnIEL offrait un très bon rappel avec très peu de ressources lexicales impliquées tout en offrant une bonne précision.

URL où la thèse pourra être téléchargée : https://lejeuneg.users.greyc.fr/?these

Laurence LONGO : laurence.longo@gmail.com

Titre : Vers des moteurs de recherche « intelligents » : un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaînes de référence

Mots-clés : détection automatique de thèmes, chaînes de référence, TAL, sémantique lexicale, coréférence, genres textuels, segmentation thématique, marqueurs linguistiques, cohésion, linguistique de corpus, *RefGen*.

***Title**: Toward "intelligent" search engines: an automatic topic detection tool. A reference chains identification based method*.

***Keywords:** Topic detection, reference chains, NLP, lexical semantics, coreference, textual genre, topic segmentation, linguistic markers, cohesion, corpus linguistics, RefGen.*

Thèse de doctorat en Sciences du Langage, mention Linguistique et Informatique, EA 1339 LiLPa (Linguistique, Langues, Parole), UFR LSHA (Langues et sciences humaines appliquées), département d'informatique, Université de Strasbourg, Strasbourg, sous la direction de Catherine Schnedecker (Pr, Université de Strasbourg). Thèse soutenue le 12/12/2013.

Jury : Mme Catherine Schnedecker (Pr, université de Strasbourg, directrice), Mme Agnès Tutin (Pr, Université Stendhal Grenoble, présidente), M. Yves Bestgen (Pr, Université catholique de Louvain, rapporteur), M. Denis Maurel (Pr, Université François Rabelais, rapporteur), Mme Amalia Todirascu (MCF, Université de Strasbourg, examinatrice), M. Frédéric Landragin (CR, CNRS-UMR Lattice, examinateur), M. Christian Dhinaut (responsable marketing, société Divalto, Strasbourg, examinateur).

***Résumé :** L'objectif de cette thèse est de proposer une nouvelle méthode d'indexation des documents, basée sur leur structure thématique, prenant en compte des paramètres comme la cohésion et la cohérence du texte, ainsi que le genre de discours utilisé.*

Si, jusqu'à présent, la notion de thème ne fait toujours pas l'unanimité dans les études linguistiques, le point de vue adopté du côté du TAL demeure plutôt consensuel, bien qu'assez réducteur, car le thème est considéré comme ce sur quoi porte la phrase ou le texte. Cette « simplification » de la notion de thème est liée à la relative complexité de la tâche de détection automatique de thèmes. Pourtant, les thèmes des documents sont liés les uns aux autres par des connecteurs, par des procédés de répétitions, et par des marqueurs référentiels explicites. Ce sont ces marques de cohésion, qui participent à la détection des thèmes des documents, que nous nous proposons d'identifier automatiquement.

*La visée applicative de la thèse consiste alors à l'amélioration substantielle d'un moteur de recherche global par l'ajout de l'outil de détection automatique de thèmes ATDS-Fr (*Automatic Topic Detection System for French*). Cet outil adopte une approche hybride statistique-linguistique qui 1) découpe les documents en segments thématiquement homogènes et 2) identifie, par le biais de marqueurs linguistiques, les thèmes des documents afin de les décrire.*
Dans la thèse, nous avons focalisé notre développement sur RefGen, le module d'identification automatique de suites d'expressions référentielles ou chaînes de référence (CR) inclues dans ATDS-Fr. RefGen utilise des méthodes classiques pour identifier les expressions référentielles (noms propres, pronoms, groupes nominaux, etc.), mais aussi d'autres paramètres tels que le genre textuel du document qui influe sur le matériau linguistique des CR. Ainsi, avons-nous posé comme hypothèse que les CR posséderaient des propriétés déterminées par leur genre d'occurrence. Le genre conditionnerait la manière dont elles sont construites. Pour vérifier notre hypothèse, nous avons mené une étude comparative des CR sur un corpus composé de divers genres textuels (éditoriaux, romans, lois européennes, faits divers, rapports publics). Nous avons utilisé les paramètres identifiés dans cette étude dans le calcul de la référence de RefGen.
Le module RefGen a été évalué, d'une part en comparant les CR proposées automatiquement à une annotation manuelle (sur un corpus de référence de divers genres textuels) et, d'autre part, en le soumettant aux diverses métriques d'évaluation actuelles de la coréférence (MUC, B^3, CEAF, BLANC). Lors de l'évaluation automatique de RefGen, un corpus multigenre de 50 000 mots, annoté en relations de référence a été constitué. Les performances obtenues par notre système oscillent entre 63 et 73 % suivant les métriques.

URL où la thèse pourra être téléchargée : tel.archives-ouvertes.fr/tel-00939243

François Morlane-Hondère : francois.morlane@univ-tlse2.fr

Titre : Une approche linguistique de l'évaluation des ressources extraites par analyse distributionnelle automatique

Mots-clés : sémantique distributionnelle, lexicologie, linguistique de corpus.

Title*: Evaluating distributional thesauri: a linguistic approach*.

Keywords: *distributional semantics, lexicology, corpus linguistics*.

Thèse de doctorat en Sciences du Langage, CLLE-ERSS (UMR 5263), Département de Sciences du Langage, UFR Langues Lettres Littératures et Civilisations Étrangères (LLCE), Université Toulouse II, Toulouse, sous la direction de Cécile Fabre (Pr, Université de Toulouse II). Thèse soutenue le 10/07/2013.

Jury : Mme Cécile Fabre (Pr, université de Toulouse II, directrice), Mme Béatrice Daille (Pr, Université de Nantes, rapporteur), M. Alain Polguère (Pr, Université de Lorraine, rapporteur), M. Nabil Hathout, (DR, CLLE-ERSS CNRS, examinateur), M. Pierre-André Buvet (MC, Université Paris 13, examinateur).

Résumé : *Dans cette thèse, nous abordons du point de vue linguistique la question de l'évaluation des bases lexicales extraites par analyse distributionnelle automatique (ADA). Les méthodes d'évaluation de ces ressources qui sont actuellement mises en œuvre (comparaison à des lexiques de référence, évaluation par la tâche, test du TOEFL...) relèvent en effet d'une approche quantitative des données qui ne laisse que peu de place à l'interprétation des rapprochements générés. De ce fait, les conditions qui font que certains couples de mots sont extraits alors que d'autres ne le sont pas restent mal connues. Notre travail vise une meilleure compréhension des fonctionnements en corpus qui régissent les rapprochements distributionnels. Pour cela, nous avons dans un premier temps adopté une approche quantitative qui a consisté à comparer plusieurs ressources distributionnelles, calculées sur des corpus différents, à des lexiques de référence (le Dictionnaire électronique des synonymes du CRISCO et le réseau lexical JeuxDeMots). Cette étape nous a permis, premièrement, d'avoir une estimation globale du contenu de nos ressources, et, deuxièmement, de sélectionner des échantillons de couples de mots à étudier d'un point de vue qualitatif.*

Cette deuxième étape constitue le cœur de la thèse. Nous avons choisi de nous focaliser sur les relations lexico-sémantiques que sont la synonymie, l'antonymie, l'hyperonymie et la méronymie, que nous abordons en mettant en place quatre protocoles différents. En nous appuyant sur les relations contenues dans les lexiques de référence, nous avons comparé les propriétés distributionnelles des couples de synonymes/antonymes/hyperonymes/méronymes qui ont été extraits par l'ADA avec celles des couples qui ne l'ont pas été. Nous mettons ainsi au jour plusieurs phénomènes qui favorisent ou bloquent la substituabilité des couples de mots (donc leur extraction par l'ADA). Ces phénomènes sont considérés au regard de paramètres comme la nature du corpus qui a permis de générer les bases distributionnelles étudiées (corpus encyclopédique, journalistique ou littéraire) ou les limites des lexiques de référence.

Ainsi, en même temps qu'il questionne les méthodes d'évaluation des bases distributionnelles actuellement employées, ce travail de thèse illustre l'intérêt qu'il y a à considérer ces ressources comme des objets d'études linguistiques à part entière. Les bases distributionnelles sont en effet le résultat d'une mise en œuvre à grande échelle du principe de substituabilité, ce qui en fait un matériau de choix pour la description des relations lexico-sémantiques.

URL où la thèse pourra être téléchargée : w3.erss.univ-tlse2.fr/membre/morlane

Laurie SERRANO : laurie.serrano9@gmail.com

Titre : Vers une capitalisation des connaissances orientée utilisateur : extraction et structuration automatiques de l'information issue de sources ouvertes

Mots-clés : gestion des connaissances, exploration de données, représentation des connaissances, renseignement d'origine sources ouvertes, ontologies (informatique), Web sémantique.

Title*: User-driven knowledge gathering: automatic extraction and structuring of open sources information*.

Keywords: *Knowledge management, data mining, knowledge representation (information theory), open source intelligence, ontologies (information retrieval), Semantic Web.*

Thèse de doctorat en Informatique, GREYC (UMR 6072), UFR d'Informatique, Université de Caen, Caen, sous la direction de Maroua Bouzid (Pr, Université de Caen), Stephan Brunessaux (Senior Expert, Cassidian-EADS), et Thierry Charnois (Pr, Université Paris 13). Thèse soutenue le 24/01/2014.

Jury : Mme Maroua Bouzid (Pr, Université de Caen, codirectrice), M. Stephan Brunessaux (Senior Expert, Cassidian-EADS, codirecteur), M. Thierry Charnois (Pr, Université Paris 13, codirecteur), M. Gaël Dias (Pr, Université de Caen, président), Mme Laurence Cholvy (DR, ONERA, rapporteur), M. Thierry Poibeau (DR, LaTTiCe-CNRS, rapporteur), Mme Fatiha Saïs (MC, Université Paris 11, examinatrice).

Résumé : *Face à l'augmentation vertigineuse des informations disponibles librement (notamment sur le Web), repérer efficacement celles qui présentent un intérêt s'avère une tâche longue et complexe. Les analystes du renseignement d'origine sources ouvertes sont particulièrement concernés par ce phénomène. En effet, ceux-ci recueillent manuellement une grande partie des informations d'intérêt afin de créer des fiches de connaissance résumant le savoir acquis à propos d'une entité. Dans ce contexte, cette thèse a pour objectif de faciliter et réduire le travail des acteurs du renseignement et de la veille. Nos recherches s'articulent autour de trois axes : la modélisation de l'information, l'extraction d'information et la capitalisation des connaissances. Nous avons réalisé un état de l'art de ces différentes problématiques afin d'élaborer un système global de capitalisation des connaissances. Notre première contribution est une ontologie dédiée à la représentation des connaissances spécifiques au renseignement et pour laquelle nous avons défini et modélisé la notion d'événement dans ce domaine. Par ailleurs, nous avons élaboré et évalué un système d'extraction d'événements fondé sur deux approches actuelles en extraction d'information : une première méthode symbolique*

et une seconde basée sur la découverte de motifs séquentiels fréquents. Enfin, nous avons proposé un processus d'agrégation sémantique des événements afin d'améliorer la qualité des fiches d'événements obtenues et d'assurer le passage du texte à la connaissance. Celui-ci est fondé sur une similarité multidimensionnelle entre événements, exprimée par une échelle qualitative définie selon les besoins des utilisateurs.

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

---

Assaf URIELI : assaf.urieli@gmail.com

Titre : Analyse syntaxique robuste du français : concilier méthodes statistiques et connaissances linguistiques dans l'outil Talismane

Mots-clés : analyse syntaxique du français, apprentissage automatique supervisé, parsing par transitions.

***Title**: Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*.

***Keywords:** French parsing, supervised machine learning, transition-based parsing.*

Thèse de doctorat en Sciences du Langage, CLLE-ERSS UMR 5263, département de Sciences du Langage, UFR LLCE, Université de Toulouse II, Toulouse, sous la direction de Ludovic Tanguy (MC HDR, Université de Toulouse II). Thèse soutenue le 17/12/2013.

Jury : M. Ludovic Tanguy (MC HDR, Université de Toulouse II, directeur), M. Nabil Hathout (DR, CLLE-ERSS, Toulouse, président), M. Alexis Nasr (Pr, Université Aix Marseille, rapporteur), M. Eric Wehrli (Pr, LATL Genève, rapporteur), Mme Marie Candito (MC, Université Paris 7, examinatrice).

***Résumé :** Dans cette thèse, nous explorons l'analyse syntaxique statistique robuste du français. Notre principal souci est de trouver des méthodes qui permettent au linguiste d'injecter des connaissances et/ou des ressources linguistiques dans un moteur statistique afin d'améliorer les résultats à la fois globalement et pour certains phénomènes spécifiques. D'abord, nous décrivons le schéma d'annotation en dépendances du français, et les algorithmes capables de produire cette annotation, en particulier le parsing par transitions. Après avoir exploré les algorithmes d'apprentissage automatique supervisé pour les problèmes de classification en TAL, nous présentons l'analyseur syntaxique Talismane, développé dans le cadre de cette thèse, qui comprend quatre modules statistiques – le découpage en phrases, la segmentation en mots, l'étiquetage morpho-syntaxique et le parsing – ainsi que les diverses ressources linguistiques utilisées par les modèles*

*de base. Nos premières expériences ont permis d'identifier la meilleure configuration d'apprentissage parmi les nombreuses configurations possibles. Ensuite, nous explorons les améliorations apportées par le principe de recherche par faisceau (*beam search*) et la propagation du faisceau d'un module à un autre. Finalement, nous présentons une série d'expériences dont le but est de corriger des erreurs linguistiques spécifiques au moyen de descripteurs ciblés pour l'apprentissage. Une de nos innovations est l'introduction des règles qui imposent ou interdisent certaines décisions locales, permettant ainsi de contourner le modèle statistique, nous explorons l'utilisation de règles pour les erreurs que les descripteurs n'ont pas pu corriger. Finalement, nous explorons également l'utilisation de ressources linguistiques à large couverture, au travers d'une expérience d'apprentissage semi-supervisé avec une ressource lexicale produite par une analyse sémantique distributionnelle.*

URL où la thèse pourra être téléchargée :
w3.erss.univ-tlse2.fr/textes/pagespersos/urieli/URIELI-thesis-2013.pdf