Résumés de thèses

Rubrique préparée par Sylvain Pogodalla

INRIA, Villers-lès-Nancy, F-54600, France
Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54500, France
CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54500, France
sylvain.pogodalla@inria.fr

Haithem AFLI : Haithem.Afli@dcu.ie

Titre : La traduction automatique statistique dans un contexte multimodal

Mots-clés : Corpus comparable multimodal, traduction automatique, extraction des données parallèles.

Title: *Statistical Machine Translation in a Multimodal Context*

Keywords: *Multimodal comparable corpora, machine translation, parallel data extraction.*

Thèse de doctorat en Informatique, Laboratoire d'informatique de l'université du Maine (LIUM), UFR Sciences et Technique, Université du Maine, Le Mans, sous la direction de Holger Schwenk (Pr, Université du Maine, Le Mans) et Loïc Barrault (MC, Université du Maine, Le Mans). Thèse soutenue le 07/07/2014.

Jury : M. Holger Schwenk (Pr, Université du Maine, Le Mans, codirecteur), M. Loïc Barrault (MC, Université du Maine, Le Mans, codirecteur), M. Emmanuel Morin (Pr, Université de Nantes, président), M. Kamel Smaïli (Pr, Université de Lorraine, Nancy, rapporteur), M. Philippe Langlais (Pr, Université de Montréal, Canada, rapporteur), M. Alexandre Allauzen (MC HDR, Université de Paris Sud – Paris 11, examinateur).

Résumé : *Les performances des systèmes de traduction automatique statistique dépendent de la disponibilité de textes parallèles bilingues, appelés aussi bitextes. Cependant, les textes parallèles librement disponibles sont aussi des ressources rares : la taille est souvent limitée, la couverture linguistique insuffisante ou le domaine des textes n'est pas approprié. Il y a relativement peu de paires de langues pour lesquelles*

des corpus parallèles de tailles raisonnables sont disponibles pour certains domaines. L'une des façons pour pallier au manque de données parallèles est d'exploiter les corpus comparables qui sont plus abondants.

Les travaux précédents dans ce domaine n'ont été appliqués que pour la modalité texte. La question que nous nous sommes posée durant cette thèse est de savoir si un corpus comparable multimodal permet d'apporter des solutions au manque de données parallèles dans le domaine de la traduction automatique. Dans cette thèse, nous avons étudié comment utiliser des ressources provenant de différentes modalités (texte ou parole) pour le développement d'un système de traduction automatique statistique. Une première partie des contributions consiste à proposer une technique pour l'extraction des données parallèles à partir d'un corpus comparable multimodal (audio et texte). Les enregistrements sont transcrits avec un système de reconnaissance automatique de la parole et traduits avec un système de traduction automatique. Ces traductions sont ensuite utilisées comme requêtes d'un système de recherche d'information pour sélectionner des phrases parallèles sans erreur et générer un bitexte.

Dans la deuxième partie des contributions, nous visons l'amélioration de notre méthode en exploitant les entités sous-phrastiques créant ainsi une extension à notre système en vue de générer des segments parallèles. Nous améliorons aussi le module de filtrage. Enfin, nous présentons plusieurs manières d'aborder l'adaptation des systèmes de traduction avec les données extraites.

Nos expériences ont été menées sur les données des sites web TED et Euronews qui montrent la faisabilité de nos approches.

URL où le mémoire pourra être téléchargé : http://www.afcp-parole.org/doc/theses/these_HA14.pdf

---

Dhouha BOUAMOR : dhouha.bouamor@limsi.fr

Titre : Constitution de ressources linguistiques multilingues à partir de corpus de textes parallèles et comparables

Mots-clés : Alignement multilingue, corpus parallèles, corpus comparables.

Title: *Multilingual Linguistic Resources Constitution from Parallel and Comparable Corpora*

Keywords: *Multilingual alignment, parallel corpora, comparable corpora.*

Thèse de doctorat en Informatique, LIMSI-CNRS, Informatique, Université Paris Sud – Paris 11, Orsay, sous la direction de Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay) et Nasredine Semmar (IR, CEA-LIST). Thèse soutenue le 21/02/2014.

Jury : M. Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay, codirecteur), M. Nasredine Semmar (IR, CEA-LIST, codirecteur), M. Reinhard Rapp (Pr, Johannes

Gutenberg-Universität Mainz, Allemagne, rapporteur), M. Éric Gaussier (Pr, Université Joseph Fourier, Grenoble, rapporteur), M. Philippe Langlais (Pr, Université de Montréal, Canada, examinateur), M. François Yvon (Pr, Université Paris Sud, président).

Résumé : *Les lexiques bilingues sont des ressources particulièrement utiles pour la Traduction Automatique et la Recherche d'Information Translingue. Leur construction manuelle nécessite une expertise forte dans les deux langues concernées et est un processus coûteux. Plusieurs méthodes automatiques ont été proposées comme une alternative, mais elles ne sont disponibles que dans un nombre limité de langues et leurs performances sont encore loin derrière la qualité des traductions manuelles. Notre travail porte sur l'extraction de ces lexiques bilingues à partir de corpus de textes parallèles et comparables, c'est à dire la reconnaissance et l'alignement d'un vocabulaire commun multilingue présent dans ces corpus.*

En nous basant sur des corpus parallèles, nous présentons une approche qui porte sur le traitement d'expressions polylexicales, allant de leur acquisition automatique à leur intégration dans un système de traduction automatique statistique. Notre intérêt se porte par ce type d'unités car, en plus du fait qu'elles soient fréquemment utilisées dans le langage oral et écrit de tous les jours ainsi que dans les communications spécialisées techniques et scientifiques, leur identification est fondamentale pour les applications faisant intervenir les aspects sémantiques de la langue et surtout la traduction automatique.

Pour les corpus comparables, nous proposons deux approches innovantes dont le but est d'extraire des lexiques bilingues spécialisés dans les domaines de la finance des entreprises, du cancer du sein, de l'énergie éolienne et de la technologie mobile. La première approche étend l'approche distributionnelle par un processus de désambiguïsation lexicale. Le but de cette approche est de ne garder que les éléments du contexte les plus susceptibles de donner la meilleure représentation du mot à traduire. Notre deuxième approche repose sur Wikipédia et l'analyse explicite sémantique. L'originalité de cette approche réside dans le fait que, au lieu de considérer l'espace des mots d'un corpus pour la représentation des mots que l'on souhaite traduire, ces derniers sont représentés dans l'espace des titres des articles de Wikipédia. Les approches nouvellement introduites se comparent favorablement aux méthodes existantes dans la plupart des configurations testées.

URL où le mémoire pourra être téléchargé : http://www.theses.fr/2014PA112032

Camille DUTREY : camille@dutrey.fr

Titre : Analyse et détection automatique de disfluences dans la parole spontanée conversationnelle

Mots-clés : Traitement automatique des langues, traitement automatique de la parole, oral spontané, parole conversationnelle, disfluences, analyse robuste, centres d'appels.

Title: *Disfluency Analysis and Automatic Detection in Conversational Spontaneous Speech*

Keywords: *Natural language processing, speech processing, spontaneous speech, conversational speech, disfluency, robust analysis, call centre.*

Thèse de doctorat en Informatique, LIMSI-CNRS (UPR 3251), Université Paris Sud – Paris 11, Orsay, sous la direction de Sophie Rosset (DR, CNRS, LIMSI, Orsay). Thèse soutenue le 16/12/2004.

Jury : Mme Sophie Rosset (DR, CNRS, LIMSI, Orsay, directrice), Mme Anne Vilnat (Pr, Université Paris Sud – Paris 11, présidente), M. Frédéric Béchet (Pr, Aix Marseille Université, rapporteur), M. Frédéric Landragin (CR, CNRS, Lattice, Montrouge, rapporteur), Mme Katarina Bartkova (MC, Université de Lorraine, Nancy, examinatrice), Mme Martine Adda-Decker (DR, CNRS, LPP, Paris, examinatrice).

Résumé : *Extraire de l'information de données langagières est un sujet de plus en plus d'actualité compte tenu de la quantité toujours croissante d'information qui doit être régulièrement traitée et analysée. Nous assistons également, depuis les années 90, à l'essor des recherches sur des données de parole, qui posent des problèmes supplémentaires par rapport à l'écrit, notamment du fait de la présence de phénomènes propres à l'oral (hésitations, reprises, corrections) mais aussi parce qu'elles sont traitées par un système de reconnaissance automatique de la parole qui génère potentiellement des erreurs. Ainsi, extraire de l'information de données audio implique de tenir compte du « bruit » intrinsèque à l'oral ou généré par le système de reconnaissance de la parole. Il ne peut s'agir d'une simple application de méthodes qui ont fait leurs preuves sur de l'écrit. L'utilisation de techniques adaptées au traitement des données issues de l'oral et prenant en compte leurs spécificités liées au signal de parole et à la transcription — manuelle comme automatique — de ce dernier représente un thème de recherche en plein développement, qui soulève de nouveaux défis scientifiques. Ces défis sont liés à la gestion de la variabilité dans la parole et des modes d'expressions spontanés. Par ailleurs, l'analyse robuste de conversations téléphoniques a également fait l'objet de travaux, dans la continuité desquels s'inscrivent ces travaux de thèse.*

Cette thèse porte plus spécifiquement sur l'analyse des disfluences et de leur réalisation dans des données conversationnelles issues des centres d'appels EDF, à partir du signal de parole et des transcriptions de ce dernier. Ce travail convoque différents domaines, de l'analyse robuste de données issues de la parole à l'analyse et la gestion des aspects liés à l'expression orale. L'objectif de la thèse est de proposer des mé-

thodes adaptées à ces données, qui permettent d'améliorer les analyses de fouille de texte réalisées sur les transcriptions. Pour répondre à ces problématiques, nous avons analysé finement le comportement de phénomènes caractéristiques de l'oral spontané (disfluences) dans des données orales issues de centres d'appels EDF, et nous avons mis au point une méthode automatique pour leur détection, en utilisant des indices linguistiques, acoustico-prosodiques, discursifs et para-linguistiques.

Les apports de cette thèse s'articulent donc selon trois axes de recherche. Premièrement, nous proposons une caractérisation des conversations en centres d'appels du point de vue de l'oral spontané et des phénomènes qui le caractérisent. Deuxièmement, nous avons mis au point (i) une chaîne d'enrichissement des données orales effective sur plusieurs plans d'analyse (linguistique, prosodique, discursif, para-linguistique) ; (ii) un système de détection automatique des disfluences d'édition adapté aux données orales conversationnelles, utilisant le signal et les transcriptions (manuelles ou automatiques). Troisièmement, d'un point de vue « ressource », nous avons produit un corpus de transcriptions automatiques de conversations issues de centres d'appels annoté en disfluences d'édition (méthode semi-automatique).

URL où le mémoire pourra être téléchargé : https://tel.archives-ouvertes.fr/tel-01164385

Nuria GALA PAVIA : nuria.gala@lif.univ-mrs.fr

Titre : Approches multidisciplinaires pour l'étude du lexique et la création de ressources lexicales nouvelles

Mots-clés : Lexicologie, lexicographie, ressources lexicales, traitement automatique des langues, psycholinguistique.

Title: *Multidisciplinary Approaches to the Study of the Lexicon and to Building New Lexical Resources*

Keywords: *Lexicology, lexicography, lexical resources, natural language processing, psycholinguistics.*

Habilitation à diriger des recherches en Sciences du Langage, LIF-CNRS, Arts, Lettres, Langues et Sciences Humaines, Aix Marseille Université, Marseille, sous la direction de Philippe Blache (DR, CNRS, LPL, Aix-en-Provence). Habilitation soutenue le 05/06/2015.

Jury : M. Philippe Blache (DR, CNRS, LPL, Aix-en-Provence, directeur), M. Nabil Hathout (DR, CNRS, CLLE-ERSS, Toulouse, rapporteur), M. Alain Polguère (Pr, Université de Lorraine, Nancy, rapporteur), M. Horaccio Saggion (Pr, Universitat Pompeu Fabra (UPF), Barcelone, Espagne, rapporteur), M. Patrice Bellot (Pr,

Aix Marseille Université, examinateur), Mme Agnès Tutin (Pr, Université Stendhal Grenoble 3, examinatrice).

Résumé : *L'intérêt pour les ressources lexicales n'a cessé d'évoluer en fonction des besoins et des technologies. Le lexique se trouve, ainsi, au cœur de nombreuses recherches dans des domaines variés : construction de dictionnaires, apprentissage du vocabulaire, aide à la lecture, etc. Il est également le socle des outils de traitement automatique des langues et des technologies du langage au sens large.*

La construction et l'enrichissement de ressources reste une tâche coûteuse et requiert des compétences dans différentes disciplines. À ce jour, les traitements automatiques permettent d'améliorer la couverture des lexiques et le caractère explicite, détaillé et approfondi des informations qu'ils contiennent. Les méthodes de construction sont ainsi diversifiées (semi-automatiques, collaboratives) et les lexiques qui en résultent sont de plus en plus dynamiques, dans une perspective de partage de données à grande échelle.

Enfin, les ressources lexicales représentent un enjeu sociétal important, parce qu'elles sont nécessaires pour développer des applications d'assistance à l'apprentissage des langues comme dans l'aide à la remédiation de pathologies de la lecture et de la parole, etc.

Dans ce mémoire pour l'obtention de l'Habilitation à Diriger des Recherches, c'est par le biais de la multidisciplinarité que nous abordons le lexique et la construction de ressources. Outre la description de quelques ressources que nous avons eu l'occasion de créer et/ou enrichir, nous apportons une mise en perspective historique et méthodologique de quelques approches et applications où le lexique reste au centre des problématiques.

URL où le mémoire pourra être téléchargé : `http://pageperso.lif.univ-mrs.fr/~nuria.gala/publis/HdR_NGala_juin2015.pdf`

Valérie HANOKA-MAITENAZ : valerie.hanoka@gmail.com

Titre : Extraction et complétion de terminologies multilingues

Mots-clés : Extraction terminologique multilingue, extension de terminologie multilingue, graphe de traduction, traitement automatique des langues.

Title: *Multilingual Terminologies Extraction and Extension*

Keywords: *Multilingual terminology extraction, multilingual terminology extension, translation graph, natural language processing.*

Thèse de doctorat en Sciences du Langage, Alpage, Université Paris Diderot – Paris 7, Paris, sous la direction de Laurence Danlos (Pr, Université Paris Diderot –

Paris 7) et Benoît Sagot (CR, INRIA Paris – Rocquencourt). Thèse soutenue le 07/07/2015.

Jury : Mme Laurence Danlos (Pr, Université Paris Diderot – Paris 7, codirectrice), M. Benoît Sagot (CR, INRIA Paris – Rocquencourt, codirecteur), M. Mathieu Lafourcade (MC HDR, Université de Montpellier 2, président), Mme Marie-Claude L'Homme (Pr, Université de Montréal, Canada, rapporteur), M. Bruno Gaume (CR, CNRS, CLLE-ERSS, Toulouse, examinateur).

Résumé : *Les processus d'extraction terminologique automatique ont été jusqu'ici majoritairement conçus pour être appliqués à des corpus monolingues et dans des registres de langue uniformes. Cette thèse, réalisée dans le cadre d'une convention CIFRE, prolonge cet objectif pour une application à des données textuelles bruitées et issues de langues de plus en plus variées, pour l'extraction de « termes de terrain ».*

Ce travail s'inscrit dans le cadre de l'analyse de verbatim issus d'enquêtes internes au sein de multinationales traitées par l'entreprise Verbatim Analysis - VERA ; il consiste à élaborer une séquence de traitements pour l'extraction automatique de terminologies qui soit faiblement dépendante de la langue, du registre de langue ou du domaine.

Suivant une réflexion fondée sur différents aspects de typologie linguistique appliquée à sept langues, nous proposons des prétraitements textuels préliminaires à l'entraînement de modèles. Ces derniers sont soit indispensables (segmentation en tokens), soit optionnels (amputation d'une partie de l'information morphologique). Sur l'ensemble des données ainsi produites, nous calculons des traits numériques (statistiques ou fréquentiels) pour l'entraînement des modèles statistiques de type CRF. Nous sélectionnons un ensemble de meilleurs modèles grâce à une évaluation automatisée, au moyen d'une métrique adaptée, des termes extraits par les modèles produits pour l'ensemble des cadres expérimentaux envisagés pour chaque langue. Nous réalisons alors une seconde série d'évaluations pour étudier l'exploitabilité de ces modèles pour d'autres langues que celles sur lesquelles ils ont été entraînés. Il ressort de ces expériences que cette méthode aboutit à une extraction de termes de terrain de qualité satisfaisante. Les meilleurs scores obtenus (pour une évaluation monolingue des modèles) se situent, pour la majorité des langues, au-dessus de l'isoligne de f-score 0,9. Ces scores peuvent même être améliorés pour certaines langues grâce à l'application trans-lingue des meilleurs modèles d'autres langues ; il en ressort que notre approche constitue potentiellement un bon levier à des extractions terminologiques pour des langues ne disposant pas de leurs propres modèles.

La seconde partie de notre travail présente nos travaux relatifs à la complétion automatique de terminologies structurées multilingues. Nous avons proposé et évalué deux algorithmes de complétion qui prennent en entrée un graphe de traduction multilingue (que nous construisons à partir de ressources libres) et une terminologie multilingue structurée. Ils proposent alors de nouveaux candidats termes pour cette dernière. Notre approche permet de compléter la terminologie structurée dans une langue qu'elle couvre déjà, mais également d'étendre sa couverture à de nouvelles langue.

L'un de ces algorithmes est également appliqué au wordnet du français WOLF, ce qui en permet une amélioration importante de la couverture.

URL où le mémoire pourra être téléchargé : https://www.academia.edu/15147854/Extraction_et_Comple_tion_de_Terminologies_Multilingues

---

Ophélie LACROIX : ophelie.lacroix@limsi.fr

Titre : De l'étiquetage syntaxique pour les grammaires catégorielles de dépendances à l'analyse par transition dans le domaine de l'analyse en dépendances non-projectives

Mots-clés : Analyse syntaxique en dépendances, grammaires catégorielles de dépendances, étiquetage syntaxique, analyse par transition.

Title: *From Syntactic Tagging for Categorial Dependency Grammars to Transition-Based Parsing in the Domain of Non-Projective Dependency Parsing*

Keywords: *Dependency parsing, categorial dependency grammar, syntactic tagging, transition-based parsing.*

Thèse de doctorat en Informatique, LINA (UMR 6241), UFR Sciences et Techniques, Université de Nantes, sous la direction de Colin De La Higuera (Pr, Université de Nantes) et Denis Béchet (MC, Université de Nantes). Thèse soutenue le 08/12/2014.

Jury : M. Colin De La Higuera (Pr, Université de Nantes, codirecteur), M. Denis Béchet (MC, Université de Nantes, codirecteur), M. Christian Retoré (Pr, Université de Montpellier 2, président), M. Matthieu Constant (MC HDR, Université Paris-Est Marne-la-Vallée, rapporteur), M. Alexis Nasr (Pr, Université Aix Marseille, rapporteur), Mme Marie Candito (MC, Université Paris Diderot – Paris 7, examinatrice).

Résumé : *Cette thèse prend place dans le domaine de l'analyse syntaxique en dépendances. D'une part nous étudions l'impact d'une méthode statistique d'étiquetage syntaxique sur un analyseur basé sur les grammaires catégorielles de dépendances. Nous proposons en ce sens un processus complet de pré-annotation comprenant la segmentation des phrases en mots (incluant les mots composés), l'étiquetage grammatical et syntaxique de ces mots et l'analyse en dépendances de la phrase dans le but d'alléger le travail des annotateurs dans le cadre de la construction de corpus en dépendances non-projectives pour le français. D'autre part, nous étudions également les méthodes intégralement dirigées par les données dans le domaine de l'analyse en dépendances à travers l'adaptation d'un analyseur par transition à la représentation en dépendances des grammaires catégorielles de dépendances. Puis nous proposons une méthode séparant les étapes de prédiction des dépendances projectives et non-projectives dans le but d'améliorer la prédiction des dépendances non-projectives.*

Nous montrons que cette méthode est adaptable à n'importe quel corpus en dépendances standard.

URL où le mémoire pourra être téléchargé : https://hal.archives-ouvertes.fr/tel-01112072

Marilyne LATOUR : marilyne.latour@reportlinker.com

Titre : Du besoin d'informations à la formulation des requêtes : étude des usages de différents types d'utilisateurs visant l'amélioration d'un système de recherche d'informations

Mots-clés : Termes recherche, comportement utilisateur, langage naturel, expression et interprétation des besoins, formulation question, recherche d'informations, besoin informationnel.

Title: *The Process from Informational Need to Query Formulation: The Study of User Expertise Differentiation to Improve an Information Retrieval System Results*

Keywords: *Query expression, information retrieval, information need, query formulation, natural language, search team, user behavior.*

Thèse de doctorat en Sciences de l'Information et de la Communication, GRESEC (EA 608), Université Stendhal, Grenoble, sous la direction de Laurence Balicco (Pr, Université Stendhal Grenoble 3). Thèse soutenue le 24/06/2014.

Jury : Mme Laurence Balicco (Pr, Université Stendhal Grenoble 3, directrice), Mme Sylvie Lainé-Cruzel (Pr, Université Jean Moulin Lyon 3, rapporteur, présidente), Mme Josiane Mothe (Pr, Université Paul Sabatier Toulouse 3, rapporteur), Mme Céline Paganelli (MC, Université Paul Valéry Montpellier 3, examinatrice).

Résumé : *A l'instar des comportements d'utilisateurs aussi variés qu'imprévisibles en matière de recherches d'Informations (RI), l'objectif de notre travail a été d'évaluer la façon dont un même utilisateur verbalise un besoin informationnel à travers un énoncé de type « expression libre » (appelé langage naturel — LN) et un énoncé de type mots-clés (appelé langage de requêtes). Nous nous situons dans un contexte applicatif, à savoir des demandes de remboursement des utilisateurs d'un moteur de recherche dédié à des études économiques en français. Nous avons recueilli via ce moteur, les deux types d'énoncés sur 5 années consécutives totalisant un corpus de 1398 demandes en langage naturel et de 3427 requêtes.*

Une première approximation serait de penser que l'expression en LN est une verbalisation plus riche que le besoin informationnel à proprement parler. Nous avons développé des méthodes entièrement automatiques permettant de déterminer quelles informations sont réellement ajoutées, modifiées, supprimées entre la demande en LN et la requête. Pour cela, nous avons dégagé les régularités des structures énonciatives afin de mettre au point des procédures qui s'appuient sur les principaux motifs

séquentiels fréquents des demandes en LN. Nous avons dégagé les [REFERENTS] des demandes en LN i.e. les mots-clés ou syntagmes principaux de la demande (ex : « revêtement de sol », « télésurveillance », « lavage auto »). Notre hypothèse H1 qui consistait à valider si le choix du vocabulaire fait pour la saisie de la requête correspond aux termes employés lors de la formulation de la demande en LN a été validée. L'hypothèse H2 qui consistait à tester si l'ordre des termes de la question en LN est identique à la structure de la requête a elle aussi été validée puisque très peu d'utilisateurs modifient l'ordre d'apparition des termes. Notre hypothèse H3 qui consistait à tester si la demande en LN ainsi que la requête conservaient des éléments inhérents à la tâche à effectuer par l'utilisateur a été partiellement validée. Les demandes en LN comportent bien des indices linguistiques et structurels permettant de repérer les buts explicites des tâches de RI des utilisateurs ; mais ces indices ne se retrouvent que très peu dans les requêtes elles-mêmes. D'après nos conclusions, toutes les requêtes ne sont pas à traiter de façon identiques par les systèmes de RI : certaines requêtes sont suffisamment explicites alors qu'il serait intéressant de proposer une aide à la reformulation des requêtes pour celles dont la tâche sous jacente de RI n'a pas été identifiée.

URL où le mémoire pourra être téléchargé : https://tel.archives-ouvertes.fr/tel-01100752/document

---

Aurélien LAUF : aurelienlauf@gmail.com

Titre : Propagation du buzz sur Internet — Identification, analyse, modélisation et représentation dans un contexte de veille

Mots-clés : Buzz, veille, intelligence d'entreprise, internet, graphe, autorité, clustering, cooccurrence, linguistique de corpus.

Title: *Buzz Lifecyle on the Web—Identification, Analysis, Modeling and Representation in the Context of Strategic and Competitive Intelligence*

Keywords: *Buzz, competitive intelligence, internet, graph, authority, clustering, cooccurrence, corpus linguistics.*

Thèse de doctorat en Traitement automatique de langues, Langues, littératures et sociétés du monde, Institut National des Langues et Civilisations Orientales (INALCO), Paris, sous la direction de Mathieu Valette (Pr, INALCO, Paris). Thèse soutenue le 14/10/2014.

Jury : M. Mathieu Valette (Pr, INALCO, Paris, directeur), M. Pascal Marchand (Pr, Université de Toulouse 3, rapporteur), M. Mathieu Roche (DR, CIRAD, rapporteur), Mme Frédérique Segond (Pr associé, INALCO, Paris, présidente), M. Julien Velcin

(MC, Université de Lyon 2, examinateur), Mme Leila Khouas (Dr, AMI Software, examinatrice).

Résumé : *Pour une entreprise, il est nécessaire de faire de la veille afin de devancer ses concurrents, innover, comprendre les attentes de ses clients, et réagir le plus rapidement possible.*

Le Web ne cesse de grossir, atteignant des proportions incommensurables. Par ailleurs, on assiste à une perte de confiance vis-à-vis des médias classiques et à un changement majeur dans les habitudes d'accès à l'information. Les nouveaux intermédiaires de l'information que sont les blogs, les forums et les réseaux sociaux sont des mines d'informations qui doivent être exploitées. Ces nouvelles sources numériques posent néanmoins la question de l'évaluation de la pertinence ou de la crédibilité d'une information et impliquent de nouveaux modes d'extraction et d'analyse en continu des données.

S'inscrivant dans un contexte de veille informationnelle et d'intelligence d'entreprise sur Internet, l'objectif de cette thèse est d'élaborer des outils et des méthodes permettant d'identifier, analyser, modéliser et représenter le cheminement des buzz sur Internet. Tout buzz a un ou plusieurs points d'origine : les sources primaires. L'information est ensuite relayée par des sources secondaires qui vont accélérer ou non la propagation en fonction de leur degré d'influence. Tout au long du cycle de vie du buzz, le contenu sémantique est amené à évoluer. Il s'agira donc de détecter des buzz en ligne, remonter jusqu'aux sources primaires et aux sources secondaires ayant joué un rôle majeur dans sa propagation, d'en dégager des sous-thématiques ainsi que des communautés de discours, et d'analyser les différences sémantiques pouvant apparaître dans le temps.

La compréhension d'un buzz sur Internet passe ainsi par l'analyse de ce qui se dit (grands thèmes abordés, sous-thématiques, etc.) et la qualification des émetteurs : qui est à l'origine du buzz, qui parle, quelles sont les personnes les plus influentes, quelles interactions entretiennent-elles au sein des réseaux ? Cette thèse est axée autour de deux types d'analyses complémentaires : une analyse topologique des émetteurs (théorie des graphes et des réseaux) et une analyse du contenu textuel (linguistique de corpus).

De fait, de nombreuses approches complémentaires seront considérées dans cette étude. L'analyse conjointe de la topologie du Web et du contenu textuel est intéressante pour tenter de comprendre comment le sens se propage. Les personnes « parlant de la même chose » ont-elles tendance à se connaître ? L'apparition sur le Web d'un buzz donné s'explique-t-elle par une diffusion virale au sein des réseaux ou s'agit-il d'une appropriation commune « accidentelle », à un moment donné et à divers endroits du Web ?

URL où le mémoire pourra être téléchargé : https://tel.archives-ouvertes.fr/tel-01126913

Maxime LEFRANÇOIS : maxime.lefrancois.86@gmail.com

Titre : Représentation des connaissances sémantiques lexicales de la Théorie Sens-Texte : Conceptualisation, représentation, et opérationnalisation des définitions lexicographiques

Mots-clés : Représentation de connaissances, connaissances linguistiques, théorie sens-texte, prédicats linguistiques, représentations linguistiques, définitions lexicographiques, sémantique décompositionnelle, web des données liées.

Title: *Meaning-Text Theory Lexical Semantic Knowledge Representation: Conceptualization, Representation, and Operationalization of Lexicographic Definitions*

Keywords: *Knowledge representation, linguistic knowledge, meaning-text theory, linguistic predicates, linguistic representations, lexicographic definitions, decompositional semantics, web of linked data.*

Thèse de doctorat en Informatique, i3S (UMR 7271), Université de Nice Sophia Antipolis, Sophia Antipolis, sous la direction de Fabien Gandon (DR, Inria) et Christian Boitet (Pr, Université de Grenoble 1). Thèse soutenue le 24/06/2014.

Jury : M. Fabien Gandon (DR, Inria, codirecteur), M. Christian Boitet (Pr, Université de Grenoble 1, codirecteur), Mme Nathalie Aussenac-Gilles (DR, CNRS, IRIT, Toulouse, rapporteur), M. Igor Boguslavsky (Pr, Universidad Politécnica de Madrid, Espagne & Russian Academy of Sciences, Moscou, Russie, rapporteur), Mme Marie-Laure Mugnier (Pr, Université de Montpellier 2, rapporteur), M. Andrea Tettamanzi (Pr, Université de Nice Sophia Antipolis, président).

Résumé : *Les linguistes, comme toute communauté d'intérêt, produisent des connaissances. Des besoins récurrents émergent de ces connaissances produites, auxquels le domaine de l'ingénierie des connaissances cherche à répondre. Dans cette thèse, nous proposons une démarche d'ingénierie des connaissances appliquée aux connaissances lexicales sémantiques du Dictionnaire Explicatif et Combinatoire de la Théorie linguistique Sens-Texte (TST). Nous nous intéressons en particulier à trois briques élémentaires de connaissances dans le DEC : les prédicats linguistiques, définis dans la théorie des Actants Sémantiques de Mel'čuk, les représentations linguistiques, et les définitions lexicographiques, symbolisées par une représentation sémantique dans la TST. Nous adoptons une méthodologie en trois étapes : extension de la conceptualisation, élaboration d'un formalisme de représentation des connaissances, et opérationnalisation de ce formalisme.*

Nous étudions dans un premier temps la conceptualisation de la TST, et montrons en quoi elle devrait être étendue pour faciliter une formalisation ultérieure. Nous justifions en particulier la nécessité de définir un nouveau niveau de représentation sémantique profond, basé sur des graphes. Nous y définissons la notion de type d'unité sémantique profonde et sa structure actancielle : un ensemble de positions actancielles signées, qui peuvent être obligatoires, optionnelles, ou interdites, et étiquetées

par des rôles sémantiques lexicalisés. Nous montrons que l'organisation hiérarchique des types d'unité sémantique profonde peut correspondre à une hiérarchie de sens au sein de laquelle les structures actancielles sont héritées et spécialisées. Nous reconceptualisons les définitions lexicographiques au niveau sémantique profond, et au niveau du dictionnaire. Finalement, nous présentons un prototype d'éditeur de définitions basé sur la manipulation directe de graphes, qui permettera une intégration future de nos travaux dans des projets de lexicographie explicative et combinatoire.

Ensuite, nous proposons un formalisme de représentation des connaissances adapté à cette conceptualisation. Nous démontrons que les logiques de description et le formalisme des Graphes Conceptuels ne sont pas adaptés pour représenter les connaissances de la TST. Nous construisons alors un nouveau formalisme de représentation des connaissances adapté, dit des Graphes d'Unités.

Enfin nous étudions l'opérationnalisation du formalisme des Graphes d'Unités. Nous lui associons une sémantique formelle basée sur la théorie des modèles et l'algèbre relationnelle, et montrons que les conditions de décidabilité du raisonnement logique correspondent aux intuitions des lexicographes. Nous proposons également une implémentation du formalisme avec les standards du web sémantique, ce qui permet de profiter des architectures existantes pour le partage, l'interopérationnalisation, et l'interrogation des connaissances sur le web des données lexicales liées.

Nos travaux de recherche ouvrent de nombreuses perspectives, tant sur l'enrichissement du formalisme des Graphes d'Unités pour représenter plus de connaissances de la TST, que sur des applications en ingénierie des connaissances, en lexicographie, et en TALN.

URL où le mémoire pourra être téléchargé : http://www.maxime-lefrancois.info/docs/Lefrancois-these-2014.pdf

---

Sandrine OLLINGER : sandrine.ollinger@atilf.fr

Titre : Le raisonnement analogique en lexicographie, son informatisation et son application au Réseau Lexical du Français

Mots-clés : Lexicographie, analogie, raisonnement analogique, système lexical, graphe petit monde.

Title: *The Analogical Reasoning in Lexicography and its Computerisation: Application to the French Lexical Network*

Keywords: *Lexicography, analogy, analogical reasoning, lexical system, small word network.*

Thèse de doctorat en Sciences du Langage, ATILF, Université de Lorraine, Nancy, sous la direction de Alain Polguère (Pr, Université de Lorraine, Nancy). Thèse soutenue le 15/12/2014.

Jury : M. Alain Polguère (Pr, Université de Lorraine, Nancy, directeur), M. Sylvain Kahane (Pr, Université Paris Ouest Nanterre La Défense – Paris 10, rapporteur), M. Mathieu Lafourcade (MC HDR, Université de Montpellier 2, rapporteur), Mme Marie Candito (MC, Université Paris Diderot – Paris 7, examinatrice), M. Bruno Gaume (CR, CNRS, CLLE-ERSS, Toulouse, examinateur), M. Jean-Marie Pierrel (Pr, Université de Lorraine, Nancy, président).

Résumé : *La lexicographie contemporaine s'est attachée à définir de nouveaux modèles de description et met aujourd'hui à disposition de la communauté des ressources formelles et cohérentes offrant de multiples possibilités d'exploitations automatiques. Cette thèse concentre son attention sur le modèle des systèmes lexicaux proposé par la Lexicographie Explicative et Combinatoire. Plus précisément, elle s'intéresse au Réseau Lexical du Français, en cours de développement. En tant que système lexical, cette ressource est un graphe monolingue. Elle est constituée d'un ensemble de sommets, les unités lexicales du français, entre lesquels sont encodées de nombreuses relations, en grande majorité syntaxico-sémantiques. La présente thèse pose les bases d'une exploration de cette ressource par raisonnement analogique. Elle débute par une revue sélective de la formalisation et de l'informatisation de l'analogie en traitement automatique des langues, dans le cas précis de l'étude du lexique. Elle définit ainsi le principe de l'exploration réalisée comme un regroupement de structures unifiables. Les sommets du graphe lexical s'apparentent alors à des objets disposant d'un certain nombre d'Attributs, disponibles dans leur description lexicographique. Ils entretiennent des Relations, représentées par les arcs. Une réflexion est menée sur la nature des différents éléments composant le réseau et sur les rapports qu'ils entretiennent. Elle est réalisée en prenant en compte l'évolution de la ressource sur une période de trente mois et est accompagnée d'une analyse topologique, relevant des propriétés proches de celles des graphes petit monde. Deux séries d'expériences exploratoires sont ensuite réalisées. La première permet de conforter l'idée selon laquelle la formalisation en œuvre dans la ressource permet de détecter automatiquement des analogies. Elle met en avant la possibilité de réaliser différents types d'exploration par raisonnement analogique, en fonction des points d'entrée et des éléments d'informations comparés. Elle montre également l'apport de telles explorations en terme de vérification de la cohérence du réseau et d'émergence de règles lexicales. La seconde série d'expériences se concentre autour de la notion de configurations de dérivations. Elle montre comment le regroupement de sous-graphes analogues met en avant l'existence de connexions lexicales récurrentes. L'état d'avancement de la ressource exploitée ne permet pas d'obtenir des règles et des modèles aboutis. Les résultats obtenus sont toutefois encourageants. Les observations réalisées amènent à considérer l'analogie comme un guide permettant de s'assurer de la bonne qualité de la représentation du lexique proposée par une ressource. Elle permet également d'acquérir automati-*

quement des connaissances sur son organisation. De telles connaissances permettent d'identifier des phénomènes linguistiques et d'instrumenter l'activité lexicographique.

URL où le mémoire pourra être téléchargé : https://hal.archives-ouvertes.fr/tel-01107631

Quentin PRADET : quentin.pradet@gmail.com

Titre : Annotation en rôles sémantiques du français en domaine spécifique

Mots-clés : VerbNet, analyse sémantique, traduction automatique, ressource lexicale.

Title: *Domain-specific French Semantic Role Labeling*

Keywords: *VerbNet, semantic analysis, machine translation, lexical resource.*

Thèse de doctorat en Informatique, Alpage, Université Paris Diderot – Paris 7, Paris, sous la direction de Laurence Danlos (Pr, Université Paris Diderot – Paris 7) et Gaël de Chalendar (IR, CEA). Thèse soutenue le 06/02/2015.

Jury : Mme Laurence Danlos (Pr, Université Paris Diderot – Paris 7, codirectrice), M. Gaël de Chalendar (IR, CEA, codirecteur), M. Guy Lapalme (Pr, Université de Montréal, Canada, rapporteur), M. Patrick Saint-Dizier (DR, CNRS, IRIT, Toulouse, rapporteur), Mme Brigitte Grau (Pr, ENSIIE, LIMSI, Orsay, présidente).

Résumé : *Cette thèse de Traitement Automatique des Langues a pour objectif l'annotation automatique en rôles sémantiques du français en domaine spécifique. Cette tâche consiste à la fois à désambiguïser le sens des verbes d'un texte et à annoter leurs syntagmes avec des rôles sémantiques tels qu'Agent, Patient, ou Destination. Elle aide de nombreuses applications dans les domaines où des corpus annotés existent : on peut alors entraîner des algorithmes supervisés performants. Nous cherchons au contraire à annoter des domaines ne disposant pas de tels corpus annotés. Nous considérons ici trois domaines : le réchauffement climatique, Informatique/Internet, et le football, leurs corpus annotés ne nous servant que pour l'évaluation. Nous montrons que nos traductions vers le français de lexiques sémantiques pour l'anglais donnent la possibilité d'annoter en rôles sémantiques des textes aussi bien en domaine général qu'en domaine spécifique sans avoir à entraîner un modèle statistique.*

Nos travaux portent sur deux grands axes : les ressources puis les méthodes servant à l'annotation en rôles sémantiques.

Concernant les ressources, nous commençons par traduire la base de données lexicales WordNet vers le français à l'aide d'un modèle de langue syntaxique issu du web. Cette ressource, WoNeF, est disponible en trois versions : une à haute précision (93,3 %), une à haut F-score (70,9 %), et l'autre à haute couverture, plus large mais plus bruitée. Nous traduisons ensuite le lexique VerbNet dans lequel les verbes sont regroupés suivant leur traits syntaxiques, morphologiques et sémantiques. La traduc-

tion, nommée Verb∋Net, a été obtenue à la fois en réutilisant au maximum les lexiques verbaux du français (le Lexique-Grammaire et Les Verbes Français) mais aussi avec un travail manuel important pour contrôler au mieux son contenu.

Concernant les méthodes, nous commençons par évaluer notre méthode basée sur VerbNet sur le corpus annoté FrameNet en suivant les travaux de Swier and Stevenson [2005]. Nous montrons que des améliorations conséquentes peuvent être obtenues à la fois d'un point de vue syntaxique avec la prise en compte de la voix passive et d'un point de vue sémantique en filtrant les syntagmes ne correspondant pas aux restrictions de sélection indiquées dans VerbNet et en réutilisant les résultats des premières annotations automatiques non ambiguës.

Enfin, une fois ces briques en place, nous évaluons la faisabilité de l'annotation en rôles sémantiques du français dans nos trois domaines spécifiques. Nous évaluons en effet quels sont les avantages et inconvénients de se baser sur VerbNet et Verb∋Net pour annoter ces domaines en anglais et en français.

URL où le mémoire pourra être téléchargé : https://hal.archives-ouvertes.fr/tel-01182711

Driss SADOUN : driss.sadoun@yahoo.fr

Titre : Des spécifications en langage naturel aux spécifications formelles via une ontologie comme modèle pivot

Mots-clés : Spécifications en langage naturel, représentation des connaissances, ontologie, vérification formelle.

Title: *From Natural Language Specifications to Formal Specifications via an Ontology as a Pivot Model*

Keywords: *Natural language requirements, knowledge representation, ontology, formal verification.*

Thèse de doctorat en Informatique, LIMSI, Université Paris Sud – Paris 11, Orsay, sous la direction de Brigitte Grau (Pr, ENSIIE), Catherine Dubois (Pr, ENSIIE), et Yacine Ghamri-Doudane (Pr, Université de La Rochelle). Thèse soutenue le 17/06/2014.

Jury : Mme Brigitte Grau (Pr, ENSIIE, codirectrice), Mme Catherine Dubois (Pr, ENSIIE, codirectrice), M. Yacine Ghamri-Doudane (Pr, Université de La Rochelle, codirecteur), Mme Chantal Reynaud (Pr, Université Paris Sud – Paris 11, présidente), M. François Lévy (Pr, Université Paris Nord – Paris 13, rapporteur), M. Yamine Ait-

Ameur (Pr, ENSEEIHT, Toulouse, rapporteur), M. Yannick Toussaint (CR HDR, INRIA Nancy – Grand Est, examinateur).

Résumé : *Le développement d'un système a pour objectif de répondre à des exigences. Aussi, le succès de sa réalisation repose en grande partie sur la phase de spécification des exigences qui a pour vocation de décrire de manière précise et non ambiguë toutes les caractéristiques du système à développer.*

Les spécifications d'exigences sont le résultat d'une analyse des besoins faisant intervenir différentes parties. Elles sont généralement rédigées en langage naturel (LN) pour une plus large compréhension, ce qui peut mener à diverses interprétations, car les textes en LN peuvent contenir des ambiguïtés sémantiques ou des informations implicites. Il n'est donc pas aisé de spécifier un ensemble complet et cohérent d'exigences. D'où la nécessité d'une vérification formelle des spécifications résultats.

Les spécifications LN ne sont pas considérées comme formelles et ne permettent pas l'application directe de méthodes de vérification formelles. Ce constat mène à la nécessité de transformer les spécifications LN en spécifications formelles. C'est dans ce contexte que s'inscrit cette thèse.

La difficulté principale d'une telle transformation réside dans l'ampleur du fossé entre spécifications LN et spécifications formelles. L'objectif de mon travail de thèse est de proposer une approche permettant de vérifier automatiquement des spécifications d'exigences utilisateur, écrites en langage naturel et décrivant le comportement d'un système. Pour cela, nous avons exploré les possibilités offertes par un modèle de représentation fondé sur un formalisme logique.

Nos contributions portent essentiellement sur trois propositions : 1) une ontologie en OWL-DL fondée sur les logiques de description, comme modèle de représentation pivot permettant de faire le lien entre spécifications en langage naturel et spécifications formelles ; 2) une approche d'instanciation du modèle de représentation pivot, fondée sur une analyse dirigée par la sémantique de l'ontologie, permettant de passer automatiquement des spécifications en langage naturel à leur représentation conceptuelle ; et 3) une approche exploitant le formalisme logique de l'ontologie, pour permettre un passage automatique du modèle de représentation pivot vers un langage de spécifications formelles nommé Maude.

URL où le mémoire pourra être téléchargé : http://www.theses.fr/2014PA112116

Agata SAVARY : agata.savary@univ-tours.fr

Titre : Representation et traitement automatique de la composition, de la variation et de l'approximation dans des ressources et outils linguistiques

Mots-clés : Traitement automatique des langues, unités polylexicales, entités nommées, ressources linguistiques, composition, variation, annotation de corpus, multilin-

guisme, polonais, morphologie computationnelle, outils à état finis, correction XML, distance d'édition.

Title: *Representation and Processing of Composition, Variation and Approximation in Language Resources and Tools*

Keywords: *Natural language processing, multi-word expressions, named entities, language resources, composition, variation, corpus annotation, multilingualism, Polish, computational morphology, finite-state tools, XML, tree-to-language correction, edit distance.*

Habilitation à diriger des recherches en Informatique, Laboratoire d'Informatique, Université François Rabelais, Tours. Habilitation soutenue le 27/03/2014.

Jury : Mme Anne Abeillé (Pr, Université Paris Diderot – Paris 7, rapporteur), M. Jean-Yves Antoine (Pr, Université François Rabelais Tours, examinateur), Mme Béatrice Daille (Pr, Université de Nantes, rapporteur), M. Jan Hajič (Pr, Charles University in Prague, République Tchèque, rapporteur), M. Denis Maurel (Pr, Université François Rabelais, Tours, examinateur), Mme Agnieszka Mykowiecka (CR HDR, Polish Academy of Sciences, Varsovie, Pologne, examinatrice), M. Joachim Niehren (DR, INRIA Lille – Nord Europe, examinateur).

Résumé : *Ce mémoire présente un panorama de plusieurs sujets liés au traitement automatique des langues (TAL), à la linguistique et à l'informatique. Les phénomènes de composition et de variabilité des unités linguistiques sont parmi les défis principaux du TAL auxquels je m'intéresse. Ils sont étroitement liés au postulat de la compositionnalité des langues naturelles, qui permet notamment d'éviter l'explosion combinatoire des cas lexicalisés, ce que je démontre sur un exemple de calcul automatique de valence émotionnelle d'un énoncé. Certaines unités linguistiques, et plus particulièrement les expressions polylexicales (EP), remettent en cause ce principe de compositionnalité. Leur variabilité partiellement régulière et partiellement idiosyncratique, les place aux frontières des différents niveaux de modélisation linguistique, tels que le lexique et la syntaxe. Un chapitre de ma dissertation présente un large état d'art de la description et du traitement automatique des unités polylexicales. Mes propres contributions dans ce domaine consistent notamment en un formalisme et un outil pour la description morphosyntaxique des EP, nommé Multiflex et applicable à des langues de caractéristiques morphologiques variables. Il emploie une approche à base de graphes, enrichis d'un mécanisme d'unification, où les paradigmes de flexion des EP sont basés sur des descriptions morphologiques des composants simples et sur des règles de transformations morphosyntaxiques. Les variations orthographiques, flexionnelles et syntaxiques sont exprimées dans un cadre commun.*

Mon autre domaine de spécialité concerne les entités nommées (EN), dont la plupart peuvent être considérées comme des instances particulières d'EP. Leur charge sémantique et dénotationnelle importante les place au centre de l'intérêt de la communauté TAL, dont témoigne l'état d'art du domaine. Je présente les hypothèses, les processus et les résultats de deux tâches d'annotation — celle des entités nommées et celle de la

coréférence — d'un large projet d'annotation dédié à la création du Corpus National du Polonais. Je décris également mes contributions dans le développement d'outils de reconnaissance d'entités nommées, probabilistes comme à base de règles, ainsi que l'enrichissement automatisé de Prolexbase — une base de données multilingue de noms propres — à partir de ressources libres.

Le problème relativement nouveau auquel je m'intéresse, lié à la description des expressions polylexicales, des entités nommées et des mentions est celui des structures imbriquées. Il permet de voir le traitement automatique des unités linguistiques complexes sous un nouveau jour, car elles doivent alors être représentées par des structures arborescentes (ou, plus généralement, par des graphes acycliques) plutôt que par des séquences non structurées de mots. Ainsi, des dépendances entre éléments distants dans un énoncé, des discontinuités, des chevauchements et d'autres phénomènes linguistiques fréquents deviennent plus faciles à modéliser.

En parallèle, je mène une activité axée sur les méthodes à états finis pour le traitement automatique des langues et des documents XML. Ma contribution principale dans ce domaine, réalisée en collaboration avec deux collègues, consiste en la première méthode relativement complète de correction d'un arbre par rapport à un langage d'arbres, et plus précisément de correction des documents XML par rapport à une DTD.

Les thèmes abordés par ma dissertation ouvrent plusieurs perspectives scientifiques, où l'accent est mis sur les liens entre différents domaines et communautés. Ces perspectives incluent : (i) l'intégration de données linguistiques fines dans le Web des données libres (linked open data), (ii) l'analyse syntaxique profonde des expressions polylexicales, (iii) l'identification des EP dans des corpus arborés comme instance du problème de correction d'un arbre par rapport à un langage d'arbres, (iv) une taxonomie et un benchmark pour les méthodes de correction d'un arbre par rapport à un langage d'arbres.

URL où le mémoire pourra être téléchargé : http://www.info.univ-tours.fr/~savary/English/papersASgb.html#HDR

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: *Towards a User-Oriented Knowledge Capitalization: Automatic Extraction and Structuring of Information from Open Sources*

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

Thèse de doctorat en Informatique et applications, SIMEM, GREYC (UMR 6072), Université de Caen Basse-Normandie, Caen, sous la direction de Maroua Bouzid (Pr, Université de Caen), Stephan Brunessaux (Senior expert, Cassidian, EADS), et Thierry Charnois (Pr, Université Paris Nord – 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 Nord – Paris 13, codirecteur), M. Gaël Dias (Pr, Université de Caen, président), Mme Laurence Cholvy (DR, DTIM, ONERA, rapporteur), M. Thierry Poibeau (DR, CNRS, Lattice, Paris, rapporteur), Mme Fatiha Saïs (MC, Université Paris Sud – Paris 11, Orsay, 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 le mémoire pourra être téléchargé : https://hal.archives-ouvertes.fr/tel-01082975

Xavier TANNIER : xtannier@limsi.fr

Titre : Traitement des événements et ciblage d'information

Mots-clés : Traitement automatique des langues, fouille de texte, extraction d'information, extraction d'événements.

Title: *Event Extraction and Focused Retrieval*

Keywords: *Natural language processing, text mining, information extraction, event extraction.*

Habilitation à diriger des recherches en Informatique, LIMSI-CNRS (UPR), UFR de Sciences, Université Paris Sud – Paris 11, Orsay, sous la direction de Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay). Habilitation soutenue le 18/06/2015.

Jury : M. Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay, directeur), Mme Béatrice Daille (Pr, Université de Nantes, rapporteur), Mme Marie-Francine Moens (Pr, Université Catholique de Louvain, Belgique, rapporteur), Mme Pascale Sébillot (Pr, INSA de Rennes, rapporteur), Mme Catherine Berrut (Pr, Université Joseph Fourier de Grenoble, examinatrice), M. Patrice Bellot (Pr, Aix Marseille Université, examinateur), Mme Sophie Rosset (DR, CNRS, LIMSI, Orsay, examinatrice).

Résumé : *Dans ce mémoire, nous organisons nos travaux principaux autour de quatre axes de traitement des informations textuelles : le ciblage, l'agrégation, la hiérarchisation et la contextualisation d'information. La majeure partie du document est dédiée à l'analyse des événements.*

Nous introduisons d'abord la notion d'événement à travers les diverses spécialités du traitement automatique des langues qui s'en sont préoccupées. Nous proposons ainsi un survol des différents modes de représentation des événements, tout en instaurant un fil rouge pour l'ensemble de la première partie. Nous distinguons ensuite deux grandes classes de travaux autour des événements, deux grandes visions que nous avons nommées, pour la première, « l'événement dans le texte », et pour la seconde, « l'événement dans le monde. »

Dans la première, nous considérons l'événement comme la désignation linguistique de quelque chose qui se passe, et nous tentons d'une part d'identifier ces désignations dans les textes, et d'autre part d'induire les relations temporelles existant entre ces événements, que ce soit dans des textes journalistiques ou médicaux. Nous réfléchissons enfin à une métrique d'évaluation adaptée à ce type d'informations.

Pour ce qui est de « l'événement dans le monde », nous envisageons plus l'événement tel qu'il est perçu par le citoyen, et nous proposons plusieurs approches originales pour aider celui-ci à mieux appréhender la quantité écrasante d'événements dont il prend connaissance chaque jour : les chronologies thématiques, les fils temporels, et une approche automatisée du journalisme de données.

La deuxième partie revient sur des travaux en lien avec le ciblage d'information. Nous décrivons tout d'abord nos travaux sur les systèmes de questions-réponses, dans lesquels nous avons eu recours à l'analyse syntaxique pour aider à justifier les réponses trouvées à une question en langage naturel. Enfin, nous abordons le sujet de la collecte thématique de documents sur le Web, dans le but de créer automatiquement des corpus et des lexiques spécialisés.

Enfin, nous concluons et revenons sur les perspectives associées aux travaux présentés sur les événements, avec pour but d'abolir partiellement la frontière qui séparent les différents axes présentés.

URL où le mémoire pourra être téléchargé : https://perso.limsi.fr/Individu/xtannier/Publications/files/Tannier_HDR.pdf

---

Wajdi ZAGHOUANI : wajdiz@cmu.edu

Titre : Le développement de corpus annotés pour la langue arabe

Mots-clés : Annotation de corpus, guides d'annotation, treebank, propbank, langue arabe.

Title: *Building Annotated Corpus for the Arabic Language*

Keywords: *Corpus annotation, annotation guidelines, treebank, propbank, Arabic language.*

Thèse de doctorat en Sciences du Langage, Connaissance Langage Modélisation, MoDyCo (UMR 7114), Université Paris Ouest Nanterre La Défense – Paris 10, Paris, sous la direction de Sylvain Kahane (Pr, Université Paris Ouest Nanterre La Défense – Paris 10). Thèse soutenue le 06/01/2015.

Jury : M. Sylvain Kahane (Pr, Université Paris Ouest Nanterre La Défense – Paris 10, directeur), M. Khalid Choukri (Directeur, E.L.D.A., examinateur), M. Jean-Luc Minel (Pr, Université Paris Ouest Nanterre La Défense – Paris 10, examinateur), M. Jean-Luc Muller (Directeur, CREFOP, examinateur), M. Alexis Nasr (Pr, Aix Marseille Université, rapporteur), M. Thierry Poibeau (DR, CNRS, Lattice, Montrouge, rapporteur), M. Benoît Sagot (CR, INRIA Paris – Rocquencourt, examinateur).

Résumé : *L'annotation linguistique de corpus joue un rôle important dans le développement d'applications en traitement automatique des langues naturelles telles que la recherche d'informations, l'extraction d'informations, la traduction automatique, les systèmes de questions/réponses et le résumé automatique. Cette thèse vise à présenter et à mettre en perspective nos travaux scientifiques sur l'annotation de corpus et sur la création de ressources lexicales dans la langue arabe. Nous discutons des méthodes, des difficultés linguistiques ainsi que de l'importance de ces travaux pour le traitement automatique des langues en illustrant quelques exemples où des ressources ont été intégrées dans des applications.*

Voici quelques-unes des questions auxquelles nous avons été confrontés au cours des 10 ans où nous avons travaillé au développement de corpus annotés pour l'arabe et pour lesquelles nous avons essayé d'apporter des éléments de réponses.

Tout d'abord, est-il possible de développer des corpus annotés pour la langue arabe en se servant majoritairement des méthodes et des approches d'annotation existantes ? Quelle est l'ampleur des adaptations à faire ? Quelles sont les difficultés que l'on rencontre lorsqu'on cherche à adapter des méthodes développées pour d'autres langues ? Comment peut-on optimiser les procédures et l'effort d'annotation dans les projets d'annotation d'envergure ? Et comment se fait le lien entre les différentes couches d'annotation dans les corpus ? Est-ce que les annotations peuvent se compléter les unes les autres ? Ensuite, quelles sont les particularités linguistiques de la langue arabe dont on doit tenir compte lors d'un projet d'annotation ? Enfin, quelles sont les exploitations possibles des corpus annotés pour la langue arabe ?

URL où le mémoire pourra être téléchargé : https://bdr.u-paris10.fr/theses/internet/2015PA100002.pdf

---