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

Anaïs LEFEUVRE-HALFTERMEYER : anais.lefeuvre@live.fr

Titre : Sémantique des temps du français : une formalisation compositionnelle

Mots-clés : Sémantique formelle, temporalité des éventualités, lexique compositionnel, récits de voyage, lambda-calcul simplement typé.

Title: *French Tenses Semantics: a Compositional Formalization*

Keywords: *Formal semantics, eventualities temporality, compositional lexicon, travel novels, simply typed lambda-calculus.*

Thèse de doctorat en Informatique, LaBRI/INRIA, unité de formation d'informatique, Université de Bordeaux, sous la direction de Christian Retoré (Pr, Université de Bordeaux) et Mauro Gaio (Pr, Université de Pau et des Pays de l'Adour). Thèse soutenue le 23/06/2014.

Jury : M. Christian Retoré (Pr, Université de Bordeaux, codirecteur), M. Mauro Gaio (Pr, Université de Pau et des Pays de l'Adour, codirecteur), M. Henk Verkuyl (Pr émérite, Université d'Utrecht, Pays-Bas, rapporteur), M. Patrice Enjalbert (Pr émérite, Université Paris Ouest Nanterre La Défense – Paris 10, rapporteur), Mme Delphine Battistelli (Pr, Université Paris Ouest Nanterre La Défense – Paris 10, examinatrice), M. Richard Moot (CR, CNRS, LaBRI, Bordeaux, examinateur), M. Jean-Yves Antoine (Pr, Université François Rabelais, Tours, examinateur), M. Nicolas Hannusse (DR, CNRS, LaBRI, Bordeaux, président).

Résumé : *Cette thèse s'inscrit dans le cadre du projet Région Aquitaine–INRIA : ITIPY dont le but est à terme l'extraction automatique d'itinéraires à partir de récits de voyage du XIXème et du début du XXème siècle.*

Notre premier travail fut de caractériser le corpus comme échantillon du français, par une étude contrastive d'une part de données quantitatives, et d'autre part de la structure des récits de voyage. Cette étude montrant que les segments textuels portant la narration de l'itinéraire sont difficiles à isoler du reste du récit, nous avons adopté une approche centrée sur le verbe comme support privilégié à l'expression du déplacement ou de la localisation.

Nous nous sommes consacrée à l'étude de la composante temporelle de la sémantique de ces verbes, et plus particulièrement à l'analyse automatique des temps verbaux du français. Disposant d'un analyseur syntaxique et sémantique à large échelle du français, Grail, basé sur les grammaires catégorielles et la sémantique compositionnelle en λ-DRT, notre tâche a été de prendre en compte les temps des verbes pour reconstituer la temporalité des événements et des états, notions regroupées sous le terme d'éventualité.

Cette thèse se concentre sur la construction d'un lexique sémantique traitant des temps verbaux du français. Nous proposons une extension et une adaptation d'un système d'opérateurs compositionnels conçu pour les temps du verbe néerlandais et anglais, aux temps et à l'aspect du verbe français du XIXème siècle à nos jours. Pour cela, nous nous appuyons sur une étude sémantique des temps et aspect du français d'un point de vue diachronique. Nous proposons une modélisation en terme d'intervalles de cette nouvelle version du système et nous proposons les entrées du lexique sémantique pour quelques adverbiaux de temps, eux aussi adaptés dans le cadre de ce système.

Cette formalisation est de facto opérationnelle, car elle est définie en terme d'opérateurs du λ-calcul dont la composition et la réduction, déjà programmées, calculent automatiquement les représentations sémantiques souhaitées : des formules multisortes de la logique d'ordre supérieur.

Le passage de l'énoncé comportant une éventualité seule au discours, dont le maillage référentiel est complexe, est discuté, et nous concluons par les perspectives qu'ouvrent nos travaux pour l'analyse du discours.

URL où le mémoire pourra être téléchargé :
https://tel.archives-ouvertes.fr/tel-01136420

Elisa OMODEI : elisa.omodei@ens.fr

Titre : Modélisation des dynamiques socio-sémantiques dans les communautés scientifiques

Mots-clés : Linguistique computationelle, modélisation statistique, réseaux sémantiques, extraction lexicale, dynamiques socio-sémantiques, réseaux de collaboration.

Title: *Modeling the Socio-semantic Dynamics of Scientific Communities*

Keywords: *Computational linguistics, statistical modeling, semantic networks, automatic term extraction, socio-semantic dynamics, co-authorship networks.*

Thèse de doctorat en Mathématiques appliquées aux sciences sociales, école doctoral transdisciplinaire lettres/sciences, LaTTiCe (UMR 8094), Institut des Systèmes Complexes de Paris Île-de-France (ISC-PIF), École Normale Supérieure, Paris, sous la direction de Thierry Poibeau (DR, CNRS, LaTTiCe) et Jean-Philippe Cointet (IR, INRA). Thèse soutenue le 19/12/2014.

Jury : M. Thierry Poibeau (DR, CNRS, LaTTiCe, codirecteur), M. Jean-Philippe Cointet (IR, INRA, codirecteur), M. Jean-Pierre Nadal (DR, CNRS, Laboratoire de Physique Statistique — UMR 8550 — et Centre d'Analyse et de Mathématique Sociales — UMR 8557, président), Mme Clémence Magnien (DR, CNRS, LIP6, rapporteur), M. Roger Guimera (Adjunct Professor, Rovira i Virgili University, Tarragona, Espagne, rapporteur), M. Emmanuel Lazega (Pr, Sciences Po, Paris, examinateur).

Résumé : *Comment les structures sociales et sémantiques d'une communauté scientifique guident-elles les dynamiques de collaboration à venir ? Dans cette thèse, nous combinons des techniques de traitement automatique des langues et des méthodes provenant de l'analyse de réseaux complexes pour analyser une base de données de publications scientifiques dans le domaine de la linguistique computationnelle : l'*ACL Anthology. *Notre objectif est de comprendre le rôle des collaborations entre les chercheurs dans la construction du paysage sémantique du domaine, et, symétriquement, de saisir combien ce même paysage influence les trajectoires individuelles des chercheurs et leurs interactions. Nous employons des outils d'analyse du contenu textuel pour extraire des textes des publications les termes correspondant à des concepts scientifiques. Ces termes sont ensuite connectés aux chercheurs pour former un réseau socio-sémantique, dont nous modélisons la dynamique à différentes échelles. Nous construisons d'abord un modèle statistique, à base de régressions logistiques multivariées, qui permet de quantifier le rôle respectif des propriétés sociales et sémantiques de la communauté sur la dynamique microscopique du réseau socio-sémantique. Nous reconstruisons par la suite l'évolution du champ de la linguistique computationelle en créant différentes cartographies du réseau sémantique, représentant les connaissances produites dans le domaine, mais aussi le flux d'auteurs entre les différents champs de recherche du domaine. En résumé, nos travaux ont montré que la combinaison des méthodes issues du traitement automatique des langues et de l'analyse des réseaux complexes permet d'étudier d'une manière nouvelle l'évolution des domaines scientifiques.*

URL où le mémoire pourra être téléchargé :
https://tel.archives-ouvertes.fr/tel-01097702

Amandine PÉRINET : amandine.perinet@yahoo.fr

Titre : Analyse distributionnelle appliquée aux textes de spécialité : réduction de la dispersion des données par abstraction des contextes

Mots-clés : Traitement automatique des langues, textes de spécialité, terminologie, analyse distributionnelle, modèle vectoriel, groupements sémantiques, termes complexes, relations sémantiques, abstraction de contextes.

Title: *Distributional Analysis Applied to Specialized Corpora: Reduction of Data Sparsity through Context Abstraction*

Keywords: *Natural language processing, specialised corpora, terminology, distributional analysis, vector space model, semantic cluster, complex terms, semantic relations, context abstraction.*

Thèse de doctorat en Informatique, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé (LIMICS-INSERM), Université Paris Nord – Paris 13, sous la direction de Sylvie Després (Pr, Université Paris Nord – Paris 13, LIMICS-INSERM) et Thierry Hamon (MC, Université Paris Nord – Paris 13, LIMSI). Thèse soutenue le 17/03/2015.

Jury : Mme Sylvie Després (Pr, Université Paris Nord – Paris 13, LIMICS-INSERM, codirectrice), M. Thierry Hamon (MC, Université Paris Nord – Paris 13, LIMSI, codirecteur), Mme Cécile Fabre (Pr, Université de Toulouse II, rapporteur), M. Emmanuel Morin (Pr, Université de Nantes, rapporteur), M. Thierry Charnois (Pr, Université Paris Nord – Paris 13, président), M. Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay, examinateur), M. Olivier Ferret (CR, CEA LIST, examinateur).

Résumé : *Dans les domaines de spécialité, les applications telles que la recherche d'information ou la traduction automatique s'appuient sur des ressources terminologiques pour prendre en compte les termes, les relations sémantiques ou les regroupements de termes. Pour faire face au coût de la constitution de ces ressources, des méthodes automatiques ont été proposées. Parmi celles-ci, l'analyse distributionnelle s'appuie sur la redondance d'informations se trouvant dans le contexte des termes pour établir une relation. Alors que cette hypothèse est habituellement mise en œuvre grâce à des modèles vectoriels, ceux-ci souffrent du nombre de dimensions considérable et de la dispersion des données dans la matrice des vecteurs de contexte.*

En corpus de spécialité, ces informations contextuelles redondantes sont d'autant plus dispersées et plus rares que les corpus ont des tailles beaucoup plus petites.

De même, les termes complexes sont généralement ignorés étant donné leur faible nombre d'occurrences. Dans cette thèse, nous nous intéressons au problème de la limitation de la dispersion des données sur des corpus de spécialité et nous proposons une méthode permettant de densifier la matrice des contextes en réalisant une abstraction des contextes distributionnels. Des relations sémantiques acquises en corpus sont utilisées pour généraliser et normaliser ces contextes. Nous avons évalué la robus-

tesse de notre méthode sur quatre corpus de tailles, de langues et de domaines différents. L'analyse des résultats montre que, tout en permettant de prendre en compte les termes complexes dans l'analyse distributionnelle, l'abstraction des contextes distributionnels permet d'obtenir des groupements sémantiques de meilleure qualité mais aussi plus cohérents et homogènes.

URL où le mémoire pourra être téléchargé :
https://tel.archives-ouvertes.fr/tel-01202371

Simon PETITJEAN : simon.petitjean@hhu.de

Titre : Génération modulaire de grammaires formelles

Mots-clés : Langages dédiés, modularité.

Title: *Modular Generation of Formal Grammars*

Keywords: *Domain specific languages, modularity.*

Thèse de doctorat en Informatique, Laboratoire d'Informatique Fondamentale d'Orléans (LIFO), UFR Sciences, Université d'Orléans, sous la direction de Denys Duchier (Pr, Université d'Orléans, LIFO). Thèse soutenue le 11/12/2014.

Jury : M. Denys Duchier (Pr, Université d'Orléans, LIFO, directeur), Mme Claire Gardent (DR, CNRS, LORIA, Nancy, présidente), Mme Laura Kallmeyer (Pr, Université de Düsseldorf, Allemagne, rapporteur), M. Kim Mens (Pr, Université catholique de Louvain, Belgique, rapporteur), M. Olivier Bonami (MC, Université Paris-Sorbonne, examinateur), M. Éric Villemonte de la Clergerie (CR, INRIA Paris – Rocquencourt, examinateur), Mme Ann Copestake (Pr, Université de Cambridge, Royaume-Uni, examinatrice), M. Yannick Parmentier (MC, Université d'Orléans, LIFO, examinateur).

Résumé : *Les travaux présentés dans cette thèse visent à faciliter le développement de ressources pour le traitement automatique des langues. Les ressources de ce type prennent des formes très diverses, en raison de l'existence de différents niveaux d'étude de la langue (syntaxe, morphologie, sémantique, ...) et de différents formalismes proposés pour la description des langues à chacun de ces niveaux. Les formalismes faisant intervenir différents types de structures, un unique langage de description n'est pas suffisant : il est nécessaire pour chaque formalisme de créer un langage dédié (ou DSL), et d'implémenter un nouvel outil utilisant ce langage, ce qui est une tâche longue et complexe.*

Pour cette raison, nous proposons dans cette thèse une méthode pour assembler modulairement et adapter des cadres de développement spécifiques à des tâches de génération de ressources langagières. Les cadres de développement créés sont construits autour des concepts fondamentaux de l'approche XMG (eXtensible MetaGrammar), à savoir disposer d'un langage de description permettant la définition modulaire d'abs-

tractions sur des structures linguistiques, ainsi que leur combinaison non-déterministe (c'est-à-dire au moyen des opérateurs logiques de conjonction et disjonction). La méthode se base sur l'assemblage d'un langage de description à partir de briques réutilisables, et d'après un fichier unique de spécification. L'intégralité de la chaîne de traitement pour le DSL ainsi défini est assemblée automatiquement d'après cette même spécification.

Nous avons dans un premier temps validé cette approche en recréant l'outil XMG à partir de briques élémentaires. Des collaborations avec des linguistes nous ont également amené à assembler des compilateurs permettant la description de la morphologie de l'Ikota (langue bantoue) et de la sémantique (au moyen de la théorie des frames*).*

URL où le mémoire pourra être téléchargé :
https://tel.archives-ouvertes.fr/tel-01202647

Marc SPANIOL : marc.spaniol@unicaen.fr

Titre : Un cadre pour l'analyse temporelle d'Internet

Mots-clés : Analyse temporelle d'Internet, entités nommées, l'analyse de l'Internet sur le niveau entité.

Title: *A Framework for Temporal Web Analytics*

Keywords: *Temporal Web analytics, named entities, entity-level Web analytics.*

Habilitation à diriger des recherches en Informatique, GREYC (UMR 6072), UFR Sciences, Université de Caen Basse-Normandie, sous la direction de Gaël Dias (Pr, Université de Caen Basse-Normandie, GREYC). Habilitation soutenue le 09/12/2014.

Jury : M. Gaël Dias (Pr, Université de Caen Basse-Normandie, GREYC, directeur), M. Patrice Bellot (Pr, Aix Marseille Université, LSIS, rapporteur, président), M. Éric Gaussier (Pr, Université Joseph Fourier, LIG/AMA, Grenoble, rapporteur), M. Mathieu Roche (Chercheur HDR, Cirad, TETIS, Montpellier, rapporteur), M. Mohand Boughanem (Pr, Université Paul Sabatier, IRIT, Toulouse, examinateur), M. Aldo Gangemi (Pr, Université Paris Nord – Paris 13, LIPN, examinateur).

Résumé : *Web-preservation organizations like the Internet Archive not only capture the history of born-digital content but also reflect the* zeitgeist *of different time periods over more than a decade. This longitudinal data is a potential gold mine for researchers like sociologists, politologists, media and market analysts, or experts on intellectual property.*

Longitudinal data analytics—the Web of the Past—poses research challenges, but has not received due attention. The sheer size and content of Web archives render them relevant to analysts within a range of domains. The Internet Archive holds more than 350 billion versions of Web pages, captured since 1996. This coverage can no longer

be maintained, as Web content is growing at enormous rates. A high-coverage archive would have to be an order of magnitude larger.

A Web archive of timestamped versions of Web sites over a long-term time horizon opens up great opportunities for analysts. However, difficulties arise from name ambiguities, requiring a disambiguation mapping of mentions (noun phrases in the text) onto entities. For example, "Bill Clinton" might be the former US president William Jefferson Clinton, or any other William Clinton contained in Wikipedia. Ambiguity further increases if the text only contains "Clinton" or a phrase like "the US president". The temporal dimension introduces additional complexity, for example when names of entities have changed over time (e.g. people getting married or divorced, or organizations that undergo restructuring in their identities). By mapping names and phrases onto canonicalized entities, we raise the entire analytics to a semantic rather than keyword-level in order to make sense of the raw and often noisy Web contents.

URL où le mémoire pourra être téléchargé :
https://spaniol.users.greyc.fr/HDR(Spaniol).pdf