Résumés de thèses

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Elizaveta CHERNYSHOVA : elizaveta.chernyshova@gmail.com

Titre : Expliciter et inférer dans la conversation. Modélisation de la séquence d'explicitation dans l'interaction

Mots-clés : Implicite, explicite, inférence, interaction, analyse conversationnelle, modélisation.

Title: *Making Explicit and Inferring in Conversations. A Model of Explicitation Sequence in Interaction*

Keywords: *Implicit, explicit, inference, interaction, conversation analysis, modeling.*

Thèse de doctorat en sciences du langage, UFR Sciences du Langage, ICAR, UMR 5191, Université Lumière Lyon 2. Thèse soutenue le 17/12/2018.

Jury : Mme Véronique Traverso (DR, CNRS, codirectrice), M. Sylvain Kahane (Pr, Université Paris Ouest Nanterre la Défense, codirecteur), Mme Claire Beyssade (Pr, Université Paris 8, rapporteur), Mme Maj-Britt Mosegaard Hansen (Pr, University of Manchester, Royaume-Uni, rapporteur), Mme Nathalie Rossi-Gensane (Pr, Université Lumière Lyon 2, présidente), M. Arnulf Deppermann (Pr, Universität Mannheim, Allemagne, examinateur).

Résumé : *Cette thèse porte sur la co-construction de la signification en interaction et les manifestations des processus interprétatifs des participants. En s'intéressant au processus d'explicitation, c'est-à-dire le processus par lequel un contenu informationnel devient explicite dans la conversation, elle propose une étude pluridimensionnelle de la séquence conversationnelle en jeu dans ce processus. La co-construction de la signification est ici abordée comme relevant d'une transformation informationnelle et de l'inférence.*

Nos analyses ont porté sur un corpus de français parlé en interaction, en contexte de repas et apéritifs entre amis. À partir d'une collection de séquences d'explicitation, définies comme des configurations dans lesquelles une inférence est soumise à validation, ce travail propose une analyse multidimensionnelle, portant un double regard sur les données : celui de l'analyse conversationnelle et celui d'une modélisation de la pratique d'explicitation. Ainsi, nous proposons de parcourir cette pratique selon trois axes d'analyse : (a) une analyse séquentielle, s'intéressant au déploiement de la séquence d'explicitation et des éléments la composant ; (b) une analyse reposant sur une modélisation de la gestion informationnelle dans cette séquence ; et (c) une analyse des formats linguistiques employés pour l'exhibition du processus inférentiel. Un des enjeux de ce travail est l'élaboration d'un modèle conversationnaliste pour la gestion informationnelle et son application à l'analyse des données de langue parlée en interaction.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-02070720

Arnaud FERRÉ : arnaud.ferre.pro@gmail.com

Titre : Représentations vectorielles et apprentissage automatique pour l'alignement d'entités textuelles et de concepts d'ontologie : application à la biologie

Mots-clés : Extraction d'information, normalisation, plongement lexical, intelligence artificielle, traitement automatique des langues.

Title: *Vector Representations and Machine Learning for Alignment of Text Entities with Ontology Concepts: Application to Biology*

Keywords: *Information extraction, normalization, word embedding, artificial intelligence, natural language processing.*

Thèse de doctorat en informatique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), UPR 3251, département STIC, Université Paris-Sud, Orsay, sous la direction de Claire Nédellec (DR, INRA), Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay). Thèse soutenue le 24/05/2019.

Jury : Mme Claire Nédellec (DR, INRA, codirectrice), M. Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay, codirecteur), M. Alexandre Allauzen (Pr, Université Paris-Sud, LIMSI, Orsay, président), Mme Nathalie Aussenac (DR, CNRS, IRIT, Toulouse, rapporteur), M. Emmanuel Morin (Pr, Université de Nantes, LS2N, rapporteur), M. Vincent Claveau (CR, CNRS, IRISA, examinateur).

Résumé : *L'augmentation considérable de la quantité des données textuelles rend aujourd'hui difficile leur analyse sans l'assistance d'outils. Or, un texte rédigé en langue naturelle est une donnée non structurée, c'est-à-dire qu'elle n'est interprétable que par un programme informatique spécialisé, sans lequel les informations des textes*

restent largement sous-exploitées. Parmi les outils d'extraction automatique d'information, nous nous intéressons aux méthodes d'interprétation automatique de textes pour la tâche de normalisation d'entité, qui consiste en la mise en correspondance automatique des mentions d'entités de textes avec des concepts d'un référentiel.

Pour réaliser cette tâche, nous proposons une nouvelle approche par alignement de deux types de représentations vectorielles d'entités capturant une partie de leur sens : les plongements lexicaux pour les mentions textuelles et des « plongements ontologiques » pour les concepts, conçus spécifiquement pour ce travail. L'alignement entre les deux se fait par apprentissage supervisé. Les méthodes développées ont été évaluées avec un jeu de données de référence du domaine biologique et elles représentent aujourd'hui l'état de l'art pour ce jeu de données. Ces méthodes sont intégrées dans une suite logicielle de traitement automatique des langues et les codes sont partagés librement.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-02166253

Natalia GRABAR : natalia.grabar@univ-lille.fr

Titre : Adaptation de documents techniques pour les locuteurs non spécialisés

Mots-clés : Simplification, domaine médical, acquisition de ressources, apprentissage automatique.

Title: *Adaptation of Technical Documents for Non-Specialized Speakers*

Keywords: *Simplification, medical domain, acquisition of resources, machine learning.*

Habilitation à diriger des recherches en informatique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), UPR 3251, Université Paris-Sud, Orsay, sous la direction de François Yvon (DR, CNRS). Habilitation soutenue le 17/05/2019.

Jury : M. François Yvon (DR, CNRS, directeur), Mme Pascale Sébillot (Pr, INSA de Rennes, rapporteur), M. Cédrick Fairon (Pr, Université Catholique de Louvain, Belgique, rapporteur), Mme Pierrette Bouillon (Pr, Université de Genève, Suisse, rapporteur), Mme Chantal Reynaud (Pr, Université Paris-Sud, présidente), M. Stefan Schulz (Pr, Graz General Hospital and University Clinics, Autriche, examinateur), M. Nabil Hathout (DR, CNRS, examinateur).

Résumé : *Comme tout domaine de spécialité, le domaine médical manipule des notions très spécifiques (blépharospasme, alexitymie, appendicectomie), qui sont difficiles à comprendre par les non-spécialistes. Nous proposons un ensemble de travaux dont l'objectif général consiste à adapter les documents techniques de santé et à en assurer une meilleure compréhension par les non-spécialistes. Pour atteindre cet ob-*

jectif, nous proposons une série d'expériences qui font partie d'un processus complexe et ambitieux : (1) la catégorisation des documents selon la difficulté qu'ils présentent ; (2) la détection de passages difficiles au sein des documents ; (3) l'acquisition de ressources pour la simplification lexicale et sémantique des documents ; (4) l'alignement de phrases parallèles à partir de corpus comparables pour engendrer des règles de transformation syntaxique. De plus, une partie non expérimentale du travail est dédiée à l'analyse des travaux de l'état de l'art autour de l'évaluation de la simplification de documents. De manière générale, la recherche que nous présentons ici est une recherche appliquée, motivée par des besoins réels. Chaque étape est effectuée avec une méthode clairement décrite et testée, dont les résultats sont évalués, positionnés par rapport à l'état de l'art et discutés. En fonction des étapes et des tâches, différentes méthodes sont exploitées (à base de règles, par apprentissage supervisé, avec ou sans connaissances linguistiques...). À différentes étapes de ce travail, il a également été nécessaire de construire de nouvelles ressources (lexique, corpus...) dont la genèse est également retracée. En dehors de la simplification lexicale et de la compréhension de textes de spécialité, les résultats et ressources obtenus peuvent être utiles pour d'autres applications et tâches du traitement automatique des langues (TAL) : recherche et extraction d'information, systèmes de questions-réponses, implication textuelle...

URL où le mémoire peut être téléchargé :
http://natalia.grabar.free.fr/publications/grabar-HDR2019.pdf

---

Aurélie NÉVÉOL : aurelie.neveol@limsi.fr

Titre : Traitement automatique de la langue biomédicale

Mots-clés : Extraction d'information, représentation des connaissances, recherche translationnelle.

Title: *Biomedical Natural Language Processing*

Keywords: *Information extraction, knowledge representation, translational research.*

Habilitation à diriger des recherches en informatique, UFR d'informatique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), UPR 3251, CNRS, Université Paris-Sud, Orsay, sous la direction de Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay). Habilitation soutenue le 26/11/2018.

Jury : M. Pierre Zweigenbaum (DR, CNRS, LIMSI, Orsay, directeur), M. Marc Cuggia (Pr et praticien hospitalier, CHU Pontchaillou, Université de Rennes 1, examinateur), M. Cédrick Fairon (Pr, Université Catholique de Louvain, Belgique, rapporteur), Mme Christine Froidevaux (Pr, Université Paris-Sud, présidente), M. Emmanuel Morin (Pr, Université de Nantes, LS2N, rapporteur), Mme Lynda Tamine Lechani (Pr, Université Paul Sabatier, Toulouse, rapporteur).

Résumé : *Dans le domaine biomédical, les informations cliniques et institutionnelles sont contenues dans le texte de publications scientifiques ou de dossiers patients et ne sont pas directement accessibles à des fins de traitement automatique. Pour pallier cela, le traitement automatique de la langue naturelle peut offrir des méthodes d'extraction d'information afin de convertir des textes libres en représentations exploitables pour la recherche médicale et la santé publique.*

Cependant, ces méthodes doivent être robustes face au volume, à la technicité et à la diversité des textes à traiter.

Le traitement automatique de la langue biomédicale (ou TAL biomédical) est un champ de recherche pluridisciplinaire qui mobilise l'informatique, la linguistique ainsi que la médecine. Il s'inscrit dans le champ du traitement automatique de la langue, tout en allant au-delà du service rendu à la médecine.

Trois thématiques ont particulièrement fait l'objet de mon travail ces dernières années : (1) la modélisation des informations ; (2) l'analyse de textes en langue de spécialité ; et (3) les applications biomédicales concrètes.

Tout ce travail repose sur l'analyse de corpus variés du domaine. Ainsi, le développement de ressources en soutien du TAL biomédical, en particulier pour les langues autres que l'anglais comparativement peu dotées, est un défi scientifique majeur. Mes contributions sur ce point s'appuient sur une analyse des schémas de représentation des connaissances dans le domaine, qui a permis le développement de corpus annotés destinés à être partagés par la communauté à des fins de développement méthodologique et d'évaluation. Une autre série de contributions a porté sur la proposition de méthodes d'analyse de textes médicaux, par exemple avec l'extraction d'entités et de relations. Ce travail a permis de montrer l'importance de la question de l'adaptation en domaine et de l'adaptation multilingue. Enfin, mes contributions à des études ciblées sur des applications biomédicales permettent de souligner l'impact attendu du traitement automatique de la langue en épidémiologie, en santé publique et sur les pratiques de recherche au-delà de ces disciplines. L'un des défis du TAL biomédical est de réaliser pleinement ce potentiel en mettant des outils à disposition de la communauté médicale afin de devenir un levier incontournable de la recherche translationnelle.

Ces travaux ont été réalisés en collaboration avec de nombreux collègues notamment au LIMSI CNRS et à la U.S. National Library of Medicine dans le cadre de plusieurs thèses, post-docs, stages de masters, et pour certains dans le cadre de projets de recherche ANR, H2020 et Digicosme, impliquant différents organismes, tels que l'Inserm, le CEA et des partenaires hospitaliers.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-02167096