Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

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

William BABONNAUD : willibab@hotmail.fr

Titre : Sémantique lexicale, compositionnalité et coercions. Fondements théoriques des types sémantiques

Mots-clés : sémantique formelle, sémantique lexicale, théorie des types, théorie des catégories, théorie des topos, coercion, compositionnalité, inférence de types.

Title: *Lexical Semantics, Compositionality and Coercions. Theoretical Foundations of Semantic Types*

Keywords: *formal semantics, lexical semantics, type theory, category theory, topos theory, coercion, compositionality, type inference.*

Thèse de doctorat en informatique, LORIA, UMR 7503, école doctorale IAEM, Université de Lorraine, sous la direction de M. Philippe de Groote (DR, Inria Nancy – Grand Est, LORIA). Thèse soutenue le 22/11/2022.

Jury : M. Philippe de Groote (DR, Inria Nancy – Grand Est, LORIA, directeur), Mme Laurence Danlos (Pr émérite, Université Paris Cité, présidente), M. Paul-André Melliès (DR, CNRS, IRIF, rapporteur), M. Christian Retoré (Pr, Université de Montpellier, rapporteur), Mme Alda Mari (DR, CNRS, examinatrice), M. Mathieu Constant (Pr, Université de Lorraine, ATILF, examinateur), Mme Laura Kallmeyer (Pr, Heinrich-Heine-Universität Düsseldorf, Allemagne, examinatrice).

Résumé : *Cette thèse s'intéresse à l'utilisation des théories de types et des types eux-mêmes dans les formalismes sémantiques compositionnels en traitement automatique des langues. Les types sémantiques jouent un rôle essentiel dans la détection et la représentation de certains phénomènes sémantiques, incluant les usages créatifs de la langue, la polysémie et la coprédication, et nécessitent pour cela une certaine précision linguistique ainsi que des mécanismes théoriques capables de manipuler des*

structures complexes et des coercions de types. Afin de répondre à ces exigences, l'objectif de ce travail de thèse est de proposer une base minimaliste à la construction des théories de types sémantiques, qui soit capable dans une certaine mesure d'unifier les différentes approches actuelles à la sémantique lexicale et formelle. Une première partie est dédiée à l'examen des contraintes linguistiques qui pèsent sur la notion même de type sémantique, et aboutit à l'élaboration de principes généraux destinés à encadrer l'élaboration de théories de types sémantiques. Dans une seconde partie, ces principes sont confrontés aux fondements mathématiques de telles théories, conduisant à la construction d'une théorie de types dans un style montagovien, augmenté d'une relation de sous-typage et de coercions, à partir d'un modèle catégorique de topos. Enfin, une dernière partie traite du choix des types sémantiques de base, et tente d'évaluer expérimentalement si l'acquisition de tels types à partir de données empiriques est envisageable.

URL où le mémoire peut être téléchargé :
https://hal.science/tel-03935669

Cyril GROUIN : cyril.grouin@lisn.upsaclay.fr

Titre : Le traitement automatique des langues face à l'évolution des usages de la langue

Mots-clés : domaine de spécialité, hétérogénéité, modèles pré-entraînés, modélisation, réseaux sociaux, sémantique.

Title: *Natural Language Processing Facing the Language Uses Evolution*

Keywords: *heterogeneity, modelling, pre-trained models, semantics, social media, specialty domain.*

Habilitation à diriger des recherches en informatique, Laboratoire Interdisciplinaire des Sciences du Numérique (LISN), UMR 9015, Université Paris-Saclay, sous la direction de Mme Anne Vilnat (Pr, Université Paris-Saclay). Habilitation soutenue le 23/03/2023.

Jury : Mme Anne Vilnat (Pr, Université Paris-Saclay, examinatrice et marraine scientifique), Mme Béatrice Daille (Pr, Université de Nantes, rapporteuse), M. Patrick Ruch (Pr, Haute École Spécialisée de Suisse Occidentale, HES-SO, Genève, Suisse, rapporteur), M. Mathieu Valette (Pr, Institut National des Langues et Civilisations Orientales, INaLCO, Paris, rapporteur), Mme Pascale Sébillot (Pr, Institut National des Sciences Appliquées, INSA, Rennes, examinatrice).

Résumé : *Dans ce manuscrit, je présente les recherches que j'ai menées sur les productions langagières des locuteurs d'une langue sur les réseaux sociaux. Mon manuscrit s'articule autour de deux angles d'analyse : l'impact des utilisateurs sur la langue, cette dernière étant alors envisagée comme objet d'étude, et l'impact des utilisateurs, au travers de leurs productions langagières, sur les outils et ressources*

utilisés pour le traitement automatique des langues. Face aux différences culturelles d'une part et aux évolutions techniques et sociétales d'autre part, telles que l'apparition du français inclusif, le traitement automatique des langues est lui-même en constante évolution pour faire face à cette variabilité linguistique, représentative de la diversité individuelle. Nous avons constaté l'opportunité d'étudier les inférences pour de la fouille d'opinion en chinois en complément des mots porteurs d'opinion/sentiment/émotion. Les réseaux sociaux constituent une source de témoignages pertinente en pharmacovigilance pour la détection des effets secondaires ou du mésusage médicamenteux, ou encore en contexte pandémique. Alors que l'informatique permet désormais d'encoder davantage d'informations, notamment d'ordre statistique, et bien que des stéréotypes de genre aient été identifiés dans les modèles transformers *actuels, les travaux combinant des informations morphosyntaxiques aux représentations vectorielles confirment la complémentarité des informations linguistiques dans plusieurs tâches classiques du TAL. Les prochains verrous scientifiques à lever viendront de l'imbrication désormais plus marquée de la multimodalité dans les productions langagières.*

URL où le mémoire peut être téléchargé :
https://hal.science/tel-04217062

Chuyuan Li : lisa27chuyuan@gmail.com

Titre : Détecter et prédire la structure du discours dans les dialogues face à la rareté des données

Mots-clés : analyse du discours, apprentissage automatique, dialogue, rareté des données, apprentissage auto supervisé.

Title: *Facing Data Scarcity in Dialogues for Discourse Structure Discovery and Prediction*

Keywords: *discourse analysis, machine learning, dialogue, data scarcity, self-supervised learning.*

Thèse de doctorat en informatique, LORIA, UMR 7503, Université de Lorraine, sous la direction de M. Maxime Amblard (Pr, Université de Lorraine, LORIA) et Mme Chloé Braud (CR, CNRS, IRIT). Thèse soutenue le 24/08/2023.

Jury : M. Maxime Amblard (Pr, Université de Lorraine, LORIA, codirecteur), Mme Chloé Braud (CR, CNRS, IRIT, codirectrice), M. Mathieu Constant (Pr, Université de Lorraine, ATILF, président), M. Benoît Crabbé (Pr, Université Paris Cité, rapporteur), Mme Junyi Li (*associate professor*, University of Texas, Austin, États-Unis, rapporteuse), Mme Chloé Clavel (Pr, Télécom-Paris, examinatrice), M. Giuseppe Carenini (Pr, University of British Columbia, Vancouver, Canada, examinateur).

Résumé : *A document is more than a random combination of sentences. It is, instead, a cohesive entity where sentences interact with each other to create a coherent struc-*

ture and convey specific communicative goals. The field of discourse examines the sentence organization within a document, aiming to reveal its underlying structural information. Discourse analysis plays a crucial role in Natural Language Processing (NLP) and has demonstrated its usefulness in various downstream applications like summarization and question answering. Existing research efforts have focused on automatically extracting discourse structures through tasks such as discourse relation identification and discourse parsing. However, these data-driven methods have predominantly been applied to monologue scenarios, leading to limited availability and generalizability of discourse parsers for dialogues. In this thesis, we address this challenging problem: discourse analysis in dialogues, which presents unique difficulties due to the scarcity of suitable annotated data.

We approach discourse analysis along two research lines: discourse feature discovery, and discourse structure prediction. In the first research line, we conduct experiments to investigate linguistic markers, both lexical and non-lexical, in text classification tasks. We are particularly interested in the context of mental disorder identification since it reflects a realistic scenario. To address the issue of data sparsity, we propose techniques for enhancing data representation and feature engineering. Our results demonstrate that non-lexical and discourse-level (even though shallow) features are reliable indicators in developing more general and robust classifiers. In the second research line, our objective is to directly predict the discourse structure of a given document. We adopt the Segmented Discourse Representation Theory (SDRT) framework, which represents a document as a graph. The task of extracting this graph-like structure using machine learning techniques is commonly known as discourse parsing. Taking inspiration from recent studies that investigate the inner workings of transformer-based models ("BERTology"), we leverage discourse information encoded in Pre-trained Language Models (PLMs) such as Bidirectional Encoder Representations from Transformers (BERT) and propose innovative extraction methods that require minimal supervision. Our discourse parsing approach involves two steps: first, we predict the discourse structure, and then we identify the relations within the structure. This two-stage process allows for a comprehensive analysis of the parser's performance at each stage. Using self-supervised learning strategies, our parser achieves encouraging results for the full parsing. We conduct extensive analyses to

evaluate the parser's performance across different discourse structures and propose directions for future improvements.

URL où le mémoire peut être téléchargé :
https://theses.fr/s288507

Carlos RAMISCH : carlos.ramisch@lis-lab.fr

Titre : Expressions polylexicales en traitement automatique des langues : sauter dans l'inconnu et faire la mise au point

Mots-clés : expressions polylexicales, traitement automatique des langues, compositionnalité, sémantique, PARSEME, corpus annotés, identification d'expressions.

Title: *Multiword Expressions in Computational Linguistics: Down the Rabbit Hole and Through the Looking Glass*

Keywords: *multiword expressions, computational linguistics, compositionality, semantics, PARSEME, annotated corpora, MWE identification.*

Habilitation à diriger des recherches en informatique, Laboratoire d'Informatique et Systèmes, LIS, UMR 7020, UFR Sciences, campus de Luminy, Aix Marseille Université, sous la direction de M. Alexis Nasr (Pr, Aix Marseille Université). Habilitation soutenue le 05/09/2023.

Jury : M. Alexis Nasr (Pr, Aix Marseille Université, directeur), M. Alain Polguère (Pr, Université de Lorraine, rapporteur et président), M. Leo Wanner (Pr, Universitat Pompeu Fabra, Barcelone, Espagne, rapporteur), M. Francis Bond (Pr, Palacký University Olomouc, République tchèque, rapporteur), Mme Agnès Tutin (Pr, Université de Grenoble Alpes, examinatrice).

Résumé : *Un des phénomènes les plus fascinants des langues humaines est la création et l'utilisation d'expressions idiomatiques qui défient toutes les règles de composition logique. Par exemple, en portugais brésilien, on peut exprimer un désaccord avec "nem aqui nem na China" (lit. "et-pas ici et-pas en-la Chine" : "absolument pas") ou "nem que a vaca tussa" (lit. "et-pas si la vache tousse" : "absolument jamais"). Les expressions idiomatiques de ce type sont des expressions polylexicales (EP) prototypiques, c'est-à-dire des interprétations idiosyncrasiques associées à des combinaisons de mots particulières.*

Beaucoup d'encre a coulé sur le traitement informatique des EP dans le TAL depuis le célèbre article de Sag et al. (2002)[1]*. Le présent manuscrit donne un aperçu de la recherche sur ce sujet, en mettant l'accent sur mes propres intérêts scientifiques. Je commence par un chapitre descriptif couvrant à la fois le phénomène linguistique*

1. Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, Dan Flickinger. (2002). *Multiword Expressions: A Pain in the Neck for NLP.* DOI : 10.1007/3-540-45715-1_1.

et son traitement informatique, motivant et illustrant les notions abstraites par des exemples pédagogiques.

Les deux chapitres suivants couvrent les tâches d'identification et de découverte automatique d'EP. Pour ces deux chapitres, je commence par passer en revue les ressources (jeux de données et corpus), notamment celles auxquelles j'ai contribué. Ensuite, je présente les modèles utilisés pour (a) prédire la compositionnalité des EP nominales en anglais, français et portugais, et (b) identifier les EP verbales en contexte, dans le cadre du projet PARSEME. Les deux chapitres détaillent les défis posés par l'évaluation de ces tâches et contiennent des résultats d'évaluation empiriques.

Enfin, je résume mes principales contributions et explore les pistes de recherche futures qui me semblent prometteuses. Celles-ci incluent la poursuite du travail sur les EP, l'induction de lexiques sémantiques, et le TAL orienté diversité. Plus qu'une synthèse, ce manuscrit contient des études originales de travaux connexes, contextualise, étend et articule mes contributions au domaine.

URL où le mémoire peut être téléchargé :
https://theses.hal.science/tel-04216223

Aline ÉTIENNE : acm.etienne@gmail.com

Titre : Analyse automatique des émotions dans les textes : contributions théoriques et applicatives dans le cadre de l'étude de la complexité des textes pour enfants

Mots-clés : émotions, schéma d'annotation, apprentissage automatique, complexité linguistique, textes pour enfants.

Title: *Automatic Emotion Analysis of Texts: Theoretical and Applicative Contributions to the Study of Complexity of Texts for Children*

Keywords: *emotions, annotation scheme, machine learning, linguistic complexity, texts for children.*

Thèse de doctorat en sciences du langage : traitement automatique des langues, MoDyCo, UMR 7114, département de sciences du langage, UFR PHILLIA, Université Paris Nanterre, sous la direction de Mme Delphine Battistelli (Pr, Université Paris Nanterre) et M. Gwénolé Lecorvé (chercheur, HDR, Orange). Thèse soutenue le 21/06/2023.

Jury : Mme Delphine Battistelli (Pr, Université Paris Nanterre, codirectrice), M. Gwénolé Lecorvé (chercheur, HDR, Orange, codirecteur), Mme Sophie Rosset (DR, Université Paris Saclay, présidente), Mme Núria Gala (Pr, Université d'Aix-Marseille, rapporteuse), M. Dominique Legallois (Pr, Université Paris 3 Sorbonne Nouvelle, rapporteur), Mme Nathalie Blanc (Pr, Université Montpellier 3 Paul-Valéry, examinatrice), Mme Anne Lacheret-Dujour (Pr, Université Paris Nanterre, examinatrice), Mme Iva Novakova (Pr, Université de Grenoble Alpes, examinatrice).

Résumé : *Notamment de par la diversité des moyens linguistiques employés pour les dénoter, les émotions exprimées dans un texte constituent un objet difficile à circonscrire. Leur étude, abordée ici dans le contexte de l'analyse de la complexité linguistique de textes jeunesse, pose alors de nombreux défis en linguistique comme en traitement automatique des langues (TAL). Cette thèse vise à déterminer comment explorer la dimension émotionnelle d'un texte de sorte à opérer une analyse qui rende compte de la diversité des marqueurs linguistiques des émotions (c'est-à-dire ne se limitant pas au lexique émotionnel), qui soit automatisable, et qui puisse contribuer à mettre au jour des éléments de complexité des textes. L'objectif est donc de proposer des outils théoriques opératoires pour l'analyse linguistique des émotions, mobilisables pour évaluer le caractère plus ou moins accessible — c'est-à-dire compréhensible — d'un texte pour un enfant. La méthodologie mise en œuvre pour cela repose sur la définition d'un schéma d'annotation des émotions, intégrant des notions pertinentes à la fois sur le plan linguistique, TAL et psycholinguistique pour caractériser la notion d'émotion. Son application manuelle sur un corpus de plus de 1500 textes a permis l'élaboration d'un corpus annoté en émotions, à partir duquel un outil d'analyse automatique des émotions dans les textes a été développé grâce aux techniques d'apprentissage automatique profond (modèle transformeur CamemBERT). Ce corpus annoté a aussi donné lieu à de nombreuses observations linguistiques aidant à mieux cerner le fonctionnement de l'expression des émotions.*

URL où le mémoire peut être téléchargé :
https://theses.hal.science/tel-04210908

Laure SOULIER : laure.soulier@isir.upmc.fr

Titre : Modèles de langue neuronaux pour la génération fidèle de texte à partir de données structurées et la recherche d'information conversationnelle proactive

Mots-clés : modèles de langue neuronaux, génération data-to-text, information structurée, recherche d'information conversationnelle, compréhension de la requête, clarification de la requête, apprentissage continu.

Title: *Neural Language Models for Faithful Data-to-Text Generation and Proactive Conversational Search*

Keywords: *neural language models, data-to-text generation, structured information, conversational search, query understanding, query clarification, continual learning.*

Habilitation à diriger des recherches en informatique, ISIR, UMR 7222, UFR Ingénierie, Sorbonne Université. Habilitation soutenue le 20/03/2023.

Jury : M. Laurent Besacier (*principal scientist*, Naver Labs Europe, France, examinateur), M. Éric Gaussier (Pr, Université Grenoble Alpes, rapporteur), M. Fabio Crestani (Pr, Università della Svizzera italiana, Lugano, Suisse, examinateur), M. Evangelos Kanoulas (Pr, University of Amsterdam, Pays-Bas, rapporteur), Mme Marie-Francine

Moens (Pr, KU Leuven, Louvain, Belgique, rapporteuse), Mme Catherine Pelachaud (DR, CNRS, présidente).

Résumé : *Les grands modèles de langue sont désormais prédominants dans la plupart des travaux de recherche en traitement du langage naturel, en recherche d'information ou encore en vision par ordinateur. Ces modèles ont démontré de grandes capacités à capturer la sémantique des éléments et à générer des textes ou des images plausibles. Cependant, leur entraînement guidé par des probabilités et la détection de co-occurrences nuit parfois à la pertinence de leurs résultats. L'ambition de ce manuscrit est de discuter et de contribuer à trois enjeux majeurs sous-jacents aux modèles de langue neuronaux dans le cadre d'une tâche de génération de descriptions à partir de données structurées et de recherche d'information conversationnelle. Le premier défi se concentre sur la fidélité et la pertinence de la génération de texte, discutant la modélisation des différentes parties des architectures des modèles de langue (i.e., l'encodeur et le décodeur). La deuxième question de recherche porte sur la contextualisation des modèles de langue, et notamment sur la contextualisation des besoins en information pour la recherche conversationnelle. Enfin, nous étudions la capacité des modèles de langue à s'adapter continuellement aux nouvelles connaissances lorsqu'ils sont utilisés pour effectuer des tâches d'ordonnancement de documents. Nous concluons par une discussion sur les perspectives prometteuses de ces questions de recherche, et ouvrons également de nouvelles directions pour l'apprentissage automatique et la robotique.*

URL où le mémoire peut être téléchargé :
https://hal.science/tel-04040213

Georgios ZERVAKIS : gonconist@gmail.com

Titre : Enrichir des modèles de langue de grande taille avec des lexiques sémantiques et des analogies

Mots-clés : lexiques sémantiques, analogies, BERT, modèles de langue de grande taille.

Title: *Enriching Large Language Models with Semantic Lexicons and Analogies*

Keywords: *semantic lexicons, analogies, BERT, large language models.*

Thèse de doctorat en informatique, LORIA, UMR 7503, école doctorale IAEM, Université de Lorraine, sous la direction de M. Miguel Couceiro (Pr, Université de Lorraine), M. Emmanuel Vincent (DR, Inria Nancy – Grand Est) et M. Marc Schoenauer (DR, Inria Saclay – Île-de-France). Thèse soutenue le 08/03/2023.

Jury : M. Miguel Couceiro (Pr, Université de Lorraine, codirecteur), M. Emmanuel Vincent (DR, Inria Nancy – Grand Est, codirecteur), M. Marc Schoenauer (DR, Inria Saclay – Île-de-France, codirecteur), M. Salvatore Ruggieri (Pr, Università di Pisa,

Pise, Italie, rapporteur), M. Christian Müller (DR, DFKI, Allemagne, rapporteur), Mme Élisa Fromont (Pr, Université Rennes 1, examinatrice).

Résumé : *Les progrès récents de l'apprentissage profond et des réseaux de neurones ont permis d'aborder des tâches complexes de traitement du langage naturel, qui sont appliquées à une pléthore de problèmes réels allant des assistants intelligents dans les appareils mobiles à la prédiction du cancer. Néanmoins, les systèmes modernes basés sur ces approches présentent plusieurs limitations qui peuvent compromettre leurs performances et leur fiabilité, les rendre injustes envers les minorités ou exposer des données personnelles. Nous sommes convaincus que l'intégration de connaissances et de raisonnement symboliques dans le cadre de l'apprentissage profond est une étape nécessaire vers la résolution de ces limitations. Par exemple, les ressources lexicales peuvent enrichir les réseaux de neurones profonds avec des connaissances sémantiques ou syntaxiques, et les règles logiques peuvent fournir des mécanismes d'apprentissage et de raisonnement. Par conséquent, l'objectif de cette thèse est de développer et d'évaluer des moyens d'intégrer différents types de connaissances et de raisonnement symboliques dans un modèle de langage largement utilisé, le* Bidirectional Encoder Representations from Transformers *(BERT). Dans un premier temps, nous considérons le* retrofitting*, une technique simple et populaire pour raffiner les plongements lexicaux de mots grâce à des relations provenant d'un lexique sémantique. Nous présentons deux méthodes inspirées par cette technique pour incorporer ces connaissances dans des plongements contextuels de BERT. Nous évaluons ces méthodes sur trois jeux de données biomédicales pour l'extraction de relations et un jeu de données de critiques de films pour l'analyse des sentiments, et montrons qu'elles n'ont pas d'impact substantiel sur les performances pour ces tâches. En outre, nous effectuons une analyse qualitative afin de mieux comprendre ce résultat négatif. Dans un second temps, nous intégrons le raisonnement analogique à BERT afin d'améliorer ses performances sur la tâche de vérification du sens d'un mot, et de le rendre plus robuste. Pour cela, nous reformulons la vérification du sens d'un mot comme une tâche de détection d'analogie. Nous présentons un modèle hybride qui combine BERT pour encoder les données d'entrée en quadruplets et un classifieur neuronal convolutif pour décider s'ils constituent des analogies valides. Nous testons notre système sur un jeu de données de référence et montrons qu'il peut surpasser les approches existantes. Notre étude empirique montre l'importance de l'encodage d'entrée pour BERT, et comment cette dépendance est atténuée en intégrant les propriétés axiomatiques des analogies lors de l'apprentissage, tout en préservant les performances et en améliorant la robustesse.*

URL où le mémoire peut être téléchargé :
https://www.theses.fr/2023LORR0039
