Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

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

Maria BORITCHEV : maria_boritchev@yahoo.fr

Titre : Modélisation dynamique des dialogues

Mots-clés : sémantique formelle, linguistique computationnelle, dialogue, questions.

Title: *Dialogue Modeling in a Dynamic Framework*

Keywords: *formal semantics, computational linguistics, dialogue, questions.*

Thèse de doctorat en informatique, LORIA, UMR 7503, Université de Lorraine, Nancy, sous la direction de Maxime Amblard (MC HDR, Université de Lorraine, LORIA) et Philippe de Groote (DR, INRIA, LORIA). Thèse soutenue le 22/11/2021.

Jury : M. Maxime Amblard (MC HDR, Université de Lorraine, LORIA, codirecteur), M. Philippe de Groote (DR, INRIA, LORIA, codirecteur), M. Miguel Couceiro (Pr, Université de Lorraine, LORIA, président), Mme Farah Benamara Zitoune (MC HDR, Université Paul Sabatier, IRIT, rapporteuse), M. Jonathan Ginzburg (Pr, Université Paris Diderot – Paris 7, rapporteur), Mme Ellen Breitholtz (MC, University of Gothenburg, Suède, examinatrice), Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre, examinatrice).

Résumé : *L'étude formelle du discours soulève de nombreuses interrogations liées à la nature et à la définition des phrases et de la manière dont une suite de phrases s'articule pour former un discours cohérent. Le langage est intrinsèquement dynamique : dans sa sémantique en contexte (par exemple, l'utilisation de références) et dans l'interaction (par exemple, les liens entre les actes de dialogue). Le passage du discours au dialogue donne lieu à des questions plus spécifiques en particulier liées à la relation entre les questions et les réponses. Afin d'aborder ces thématiques, nous nous concentrons sur la sémantique des questions.*

Il existe de nombreux formalismes et cadres de travail pour la sémantique formelle des phrases déclaratives et du discours. Le dialogue, pour sa part, est largement étudié d'un point de vue linguistique et traitement automatique des langues. L'objectif de notre travail est d'utiliser les théories classiques de sémantique formelle dans un cadre orienté vers le dialogue réel. Cette thèse présente une formalisation sémantique du dialogue dans une théorie dynamique des types simples.

Nous produisons des modèles du dialogue, et en particulier de l'articulation des questions et des réponses, en mêlant la Neo-Davidsonian Event Semantics *à la* Inquisitive Semantics *de manière compositionnelle et dynamique à travers l'usage de la* Continuation Style Dynamic Semantics*. Notre modèle est ancré dans une implémentation d'interface syntaxe-sémantique appelée* Abstract Categorial Grammars*.*

Une autre façon d'aborder la sémantique du dialogue est de s'intéresser aux données réelles, ce qui permet de mettre en perspective nos idées formelles et de les confronter aux observations. Pour ce faire, nous avons constitué un corpus, appelé Dialogues in Games *(DinG), composé de transcriptions d'enregistrements de personnes jouant au jeu de société Catane (en français). Nos études, centrées sur les questions et les réponses dans des données orales multilingues (anglais, français, néerlandais, espagnol mexicain, italien du nord et mandarin), a donné lieu à plusieurs schémas d'annotation, dont une partie a été appliquée à DinG.*

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03541628

Marie CANDITO : marie.candito@univ-paris-diderot.fr

Titre : Annoter et prédire des représentations linguistiques de phrases

Mots-clés : annotations linguistiques, analyse syntaxique automatique, analyse sémantique automatique.

Title: *Annotate and Predict Linguistic Representations of Sentences*

Keywords: *linguistic annotation, syntactic parsing, semantic parsing.*

Habilitation à diriger des recherches en informatique, Laboratoire de Linguistique Formelle, UMR 7110, Département de linguistique, Université de Paris. Habilitation soutenue le 19/01/2022.

Jury : M. Sylvain Schmitz (Pr, Université de Paris, président), M. Frédéric Béchet (Pr, Aix-Marseille Université, rapporteur), Mme Claire Gardent (DR, CNRS, LORIA, rapporteuse), Mme Paola Merlo (Pr, Université de Genève, Suisse, rapporteuse), M. Benoît Crabbé (Pr, Université de Paris, examinateur), M. Pierre Zweigenbaum (DR, CNRS, LISN, examinateur).

Résumé : *Le travail présenté, réalisé pour la majeure part en collaboration, concerne principalement l'explicitation de représentations linguistiques de phrases,*

qu'il s'agisse de la méthodologie de constitution manuelle de telles ressources, ou de la définition de modèles permettant de prédire de telles représentations, par apprentissage supervisé ou semi-supervisé.

Le mémoire présente, à divers degrés de détail :

– des contributions en termes de ressources annotées, pour le français, qu'il s'agisse d'expressions polylexicales, d'arbres de dépendances, de graphes de dépendances profondes, de cadres et rôles sémantiques FrameNet. Ces ressources sont définies avec exigence quant à la finesse des analyses linguistiques, et quant à leur utilisabilité comme données d'apprentissage supervisé ;

– des contributions en analyse syntaxique en dépendances, d'une part sur la problématique de la robustesse des analyseurs supervisés face aux mots inconnus et aux changements de domaine, d'autre part sur l'exploitation d'un contexte plus large et de modèles spécialisés pour la correction automatique d'arcs, pour les phénomènes le plus fréquemment source d'erreurs (rattachement prépositionnel et coordination) ;

– la proposition d'un modèle pour l'analyse automatique en graphes de dépendances reposant sur un apprentissage multitâche, où la tâche principale est réalisée par un parseur biaffine neuronal, et où des tâches auxiliaires sont définies pour ajouter de l'interdépendance dans la prédiction des arcs.

Ce mémoire couvre une période longue, marquée par l'arrivée de méthodes neuronales en TAL. L'apprentissage par transfert permet de fournir des représentations vectorielles de mots, hors ou en contexte, en utilisant des objectifs génériques, en particulier la prédiction d'un mot sachant son contexte. Il est fascinant de constater qu'un objectif aussi simple et brut permet de construire des modèles apportant des gains très importants dans à peu près toutes les tâches de TAL. Le transfert se fait en utilisant des corpus à l'état brut, ne nécessitant pas de modélisation linguistique (outre la définition des unités considérées). C'est ainsi l'objectif même d'analyse automatique de phrases qui est remis en cause. Certaines tâches, comme la traduction automatique, le résumé automatique, l'analyse de sentiments sont actuellement mieux gérées par des modèles « de bout-en-bout », ne nécessitant pas d'expliciter des représentations linguistiques traditionnelles. On assiste même à une ingénierie inversée, où ce sont les modèles de langue préentraînés sur corpus bruts qui sont sondés, pour voir si et où s'y cachent les concepts linguistiques traditionnels.

Cela dit, même s'il est difficile de prédire l'avenir du concept même d'analyse automatique de phrases, les besoins d'interprétabilité des modèles et de quantification des phénomènes linguistiques font que le concept reste d'actualité. On peut même espérer que les sondes linguistiques des modèles neuronaux permettent d'éclairer d'un jour nouveau certains concepts linguistiques.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03544267

Julie HUMBERT-DROZ : julie.humbertdroz@gmail.com

Titre : Définir la déterminologisation : approche outillée en corpus comparable dans le domaine de la physique des particules

Mots-clés : déterminologisation, linguistique outillée, linguistique de corpus, terminologie textuelle, langues de spécialité.

Title: *Defining Determinologisation: Tool-Based Approach in a Comparable Corpus in the Field of Particle Physics*

Keywords: *determinologisation, tool-based approach, corpus linguistics, textual terminology, languages for special purposes.*

Thèse de doctorat en sciences du langage & traitement informatique multilingue, Département de traitement informatique multilingue, Faculté de traduction et d'interprétation, Université de Genève, Suisse. Thèse soutenue le 07/09/2021.

Jury : Mme Aurélie Picton (Pr, Université de Genève, Suisse, codirectrice), Mme Anne Condamines (DR, CNRS, codirectrice), Mme Agnès Tutin (Pr, Université Grenoble Alpes, rapporteuse), M. Patrick Drouin (Pr, Université de Montréal, Canada, rapporteur), Mme Amélie Josselin-Leray (MC, Université Toulouse – Jean Jaurès, examinatrice), Mme Mathilde Fontanet (Pr, Université de Genève, Suisse, présidente).

Résumé : *Cette thèse porte sur la question de la déterminologisation dans le domaine de la physique des particules. La déterminologisation peut s'envisager à la fois comme un processus de passage de termes d'une langue de spécialité à la langue générale et comme le résultat de ce processus, c'est-à-dire le fonctionnement de termes dans la langue générale. Face au manque de travaux traitant spécifiquement le processus, notre travail vise à combler cette lacune en abordant cette question de manière systématique et approfondie. Nous nous intéressons particulièrement aux phénomènes sémantiques qui se produisent au cours du processus. Par la description de ces phénomènes, nous cherchons à mieux cerner la notion de déterminologisation et ses manifestations dans les textes.*

Dans ce but, nous avons développé une méthode de linguistique outillée pour l'analyse des termes dans des textes représentant différentes étapes du processus de déterminologisation. Cette méthode repose sur un corpus comparable et sur deux indices. Les données sont organisées en cinq sous-corpus, constitués de textes qui relèvent de différents genres textuels et degrés de spécialisation. Ces textes sont sélectionnés dans le but d'approcher le continuum entre langue de spécialité et langue générale. En outre, afin de caractériser les différents fonctionnements des termes dans le corpus, deux indices basés sur les contextes distributionnels des termes sont définis.

L'exploration de ces indices dans le corpus révèle la diversité des fonctionnements des termes dans la presse, en comparaison avec les textes spécialisés. Les phénomènes repérés permettent d'alimenter la réflexion sur la nature des changements sémantiques

résultant de la déterminologisation et montrent que ces changements reflètent des différences de points de vue ou de conceptualisation, inhérents à la multidimensionnalité des concepts. Les données mettent également en lumière les mécanismes à l'œuvre dans la création des emplois métaphoriques des termes qui se fondent non seulement sur les usages des termes dans la presse, lorsque de nouvelles connotations sont véhiculées, mais aussi sur la coexistence dans la langue générale des composants du terme ou d'unités de la même famille morphologique.

Nos observations contribuent ainsi à définir la déterminologisation comme un processus complexe et non linéaire de passage de termes dans la langue générale, qui fait intervenir de nombreux intermédiaires et qui est influencé par au moins cinq facteurs, aussi bien linguistiques qu'extralinguistiques. Plus largement, nous montrons l'ampleur des questions liées au processus d'intégration et au fonctionnement des termes dans la langue générale. En ce sens, notre thèse participe à reconnaître la transversalité de la déterminologisation dans les questions de circulation des termes, de diffusion des connaissances et de néologie.

URL où le mémoire peut être téléchargé :
https://archive-ouverte.unige.ch/unige:157351

Martin LENTSCHAT : martin.lentschat@gmail.com

Titre : Instanciation de relations n-aires dans des articles scientifiques guidée par une Ressource Termino-Ontologique de domaine

Mots-clés : relations n-aires, extraction d'information, extraction de relations, ingénierie des connaissances, ressource termino-ontologique, représentation de données, mesure de pertinence.

Title: *n-Ary Relations Instantiation from Scientific Articles Driven by a Domain Ontological and Terminological Resource*

Keywords: *n-ary relations, information extraction, relation extraction, knowledge engineering, ontological and terminological resource, data representation, relevance measure.*

Thèse de doctorat en informatique, UMR TETIS, Université de Montpellier, sous la direction de Patrice Buche (IR, INRAE, IATE, UMR 1208), Juliette Dibie (Pr, INRAE, MIA-Paris, UMR 518) et Mathieu Roche (DR, CIRAD, TETIS). Thèse soutenue le 14/12/2021.

Jury : M. Patrice Buche (IR, INRAE, IATE, UMR 1208, codirecteur), M. Patrice Bellot (Pr, Université Aix-Marseille, LIS, UMR 7020, rapporteur), Mme Nathalie Pernelle (Pr, Université Sorbonne Paris Nord, LIPN, UMR 7030, rapporteuse), Mme Nathalie Aussenac-Gilles (DR, CNRS, IRIT, UMR 5505, présidente), M. Konstantin Todorov (MC HDR, Université de Montpellier, LIRMM, UMR 5506, examinateur),

Mme Juliette Dibie (Pr, INRAE, MIA-Paris, UMR 518, codirectrice), M. Mathieu Roche (DR, CIRAD, TETIS, codirecteur).

Résumé : *Cette thèse s'inscrit dans le domaine de recherche des* smart data *et consiste à proposer de nouvelles méthodes de représentation et d'extraction de données expérimentales à partir d'articles scientifiques, évaluées sur un corpus dans le domaine des emballages alimentaires.*

L'objectif de cette thèse est de peupler une base de connaissances d'instances de relations n-aires extraites de documents scientifiques textuels. Les connaissances expérimentales visées sont représentées sous forme de relations n-aires composées d'arguments symboliques (un syntagme) et quantitatifs (une valeur numérique et une unité de mesure).

L'approche proposée s'appuie sur une ressource termino-ontologique (RTO) et se décompose en deux phases :

1) extraction des instances d'arguments,

2) mise en relation de ces instances dans des relations n-aires.

*La phase 1 propose une représentation des instances d'arguments extraites : SciPuRe (*Scientific Publication Representation*). Celle-ci intègre des descripteurs ontologiques, lexicaux et structurels. La phase 2 s'appuie sur les informations présentes dans les tableaux des documents, extraits automatiquement, pour guider l'extraction des relations n-aires à partir de relations partielles devant être complétées par les instances d'arguments de la phase 1. Trois approches sont proposées et évaluées afin d'identifier les instances d'arguments qui doivent compléter les relations : l'utilisation de la structure des documents, l'analyse des cooccurrences entre instances d'arguments, et l'utilisation de mesure de similarité donnée par des modèles de* word embeddings*.*

Nos résultats montrent l'importance du filtrage des instances à l'issue de la phase 1. Les deux critères mesurant la pertinence d'une instance d'argument symbolique étant sa spécificité et sa fréquence. La pertinence des arguments quantitatifs est déterminée par la discrimination de l'instance d'argument selon les sections des articles. Les résultats montrent un effet positif lors du filtrage de 20% des instances les moins pertinentes. En phase 2, nous avons expérimenté une approche d'assistance aux experts en sélectionnant plusieurs candidats pour chaque instance d'argument manquante dans une relation partielle. Nos résultats montrent que la méthode à adopter varie selon l'approche souhaitée. Lors de la sélection d'un seul candidat, l'approche fondée sur les analyses des cooccurrences donne les meilleurs résultats. Avec une sélection de trois ou cinq candidats, l'analyse des similarités sémantiques par des modèles BERT fournit de bons résultats. Enfin, lors de la sélection de dix candidats, l'approche fon-

dée sur la structure des documents est la plus efficace pour compléter les relations n-aires.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03587319

Didier SCHWAB : didier.schwab@univ-grenoble-alpes.fr

Titre : Contributions au Traitement Automatique des Langues et à un domaine d'application, la Communication Alternative et Augmentée

Mots-clés : traitement automatique des langues et de la parole, communication alternative et augmentée, clarification de sens, représentation du sens, acquisition du sens, exploitation du sens.

Title: *Contributions to Speech and Natural Language Processing and to an Application Domain, Alternative and Augmentative Communication*

Keywords: *speech and natural language processing, alternative and augmentative communication, meaning clarification, meaning representation, meaning acquisition, meaning exploitation.*

Habilitation à diriger des recherches en informatique, Laboratoire d'informatique de Grenoble, Université Grenoble Alpes. Habilitation soutenue le 08/12/2021.

Jury : M. Emmanuel Morin (Pr, Université de Nantes, rapporteur), M. Pierre Zweigenbaum (DR, CNRS, LISN, rapporteur), M. Andrei Popescu-Belis (Pr, Haute École spécialisée de Suisse occidentale, Vaud, Suisse, rapporteur), M. Mathieu Lafourcade (MC HDR, Université de Montpellier, examinateur), M. Laurent Besacier (scientifique principal, Naver Labs Europe, Grenoble, examinateur), Mme Pierrette Bouillon (Pr et doyenne, Université de Genève, Suisse, examinatrice), Mme Sophie Dupuy-Chessa (Pr, Université Grenoble Alpes, présidente).

Résumé : *Que feriez-vous si vous étiez privé de la parole, peut-être même incapable de faire le moindre geste ? Cette situation peut arriver à tout le monde de manière transitoire (voyage dans un pays, problème de santé) ou de manière permanente (handicap). Elle peut arriver simplement après une opération chirurgicale relativement lourde ou un accident. Dans une moindre mesure, elle peut également arriver si vous êtes dans un pays dont vous ne parlez pas la langue. On vous montrera des images, des pictogrammes pour que vous puissiez faire passer votre message, avoir besoin de quelque chose, chercher son chemin, indiquer un problème de santé. Une telle communication est appelée* communication alternative et augmentée *(CAA).*

Les recherches présentées ici se situent dans le domaine du traitement automatique des langues et de la parole (TALP). Elles concernent la représentation, l'acquisition et l'exploitation du sens pour et par la clarification des données langagières. L'exploration de la communication alternative et augmentée a rapidement conduit à se

demander comment les recherches menées au GETALP pourraient être bénéfiques à la CAA et, suivant un schéma de pensée classique, comment, en retour, le traitement automatique des langues et de la parole pourrait bénéficier de la CAA.

Ce document explique comment ces travaux, tout en restant axés sur le sens et sa clarification, sont passés petit à petit de l'écrit aux pictogrammes en intégrant la parole et le regard ; comment ils ont porté sur la désambiguïsation lexicale, les représentations vectorielles pour le TALP, la traduction automatique du texte et de la parole jusqu'à des travaux autour du dialogue. Enfin, il présente les hypothèses et pistes de recherches qui pourront être suivies dans les années qui viennent.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03535726