Résumés de thèses

Rubrique préparée par Sylvain Pogodalla

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

Maximin COAVOUX : mcoavoux@inf.ed.ac.uk

Titre : Analyse syntaxique automatique en constituants discontinus des langues à morphologie riche

Mots-clés : Traitement automatique des langues, analyse syntaxique automatique, arbres en constituants discontinus, systèmes de transitions, apprentissage profond, apprentissage multi-tâches.

Title: *Discontinuous Constituency Parsing of Morphologically Rich Languages*

Keywords: *Natural Language Processing, syntactic parsing, discontinuous constituency trees, transition systems, deep learning, multitask learning.*

Thèse de doctorat en Sciences du langage, UFR de linguistique, Laboratoire de Linguistique Formelle (LLF), UMR 7110, Université Paris Diderot, sous la direction de Benoît Crabbé (MC HDR, Université Paris Diderot). Thèse soutenue le 11/12/2017.

Jury : M. Benoît Crabbé (MC HDR, Université Paris Diderot, directeur), Mme Claire Gardent (DR, CNRS, LORIA, Nancy, rapporteur), M. Alexis Nasr (Pr, Aix-Marseille Université, rapporteur et président), M. Alexandre Allauzen (MC HDR, Université Paris-Sud, examinateur), M. Carlos Gómez Rodríguez (Profesor Contratado Doctor, Universidade da Coruña, La Corogne, Espagne, examinateur).

Résumé : *L'analyse syntaxique consiste à prédire la représentation syntaxique de phrases en langue naturelle sous la forme d'arbres syntaxiques. Cette tâche pose des problèmes particuliers pour les langues non configurationnelles ou qui ont une morphologie flexionnelle plus riche que celle de l'anglais. En particulier, ces langues manifestent une dispersion lexicale problématique, des variations d'ordre des mots plus fréquentes et nécessitent de prendre en compte la structure interne des mots-formes pour permettre une analyse syntaxique de qualité satisfaisante.*

*Dans cette thèse, nous nous plaçons dans le cadre de l'analyse syntaxique robuste en constituants par transitions. Dans un premier temps, nous étudions comment intégrer l'analyse morphologique à l'analyse syntaxique, à l'aide d'une architecture de réseaux de neurones basée sur l'apprentissage multitâches. Dans un second temps, nous proposons un système de transitions qui permet de prédire des structures générées par des grammaires légèrement sensibles au contexte telles que les LCFRS (*Linear Context-Free Rewriting System*). Enfin, nous étudions la question de la lexicalisation de l'analyse syntaxique. Les analyseurs syntaxiques en constituants lexicalisés font l'hypothèse que les constituants s'organisent autour d'une tête lexicale et que la modélisation des relations bilexicales est cruciale pour désambiguïser. Nous proposons un système de transition non lexicalisé pour l'analyse en constituants discontinus et un modèle de scorage basé sur les frontières de constituants et montrons que ce système, plus simple que des systèmes lexicalisés, obtient de meilleurs résultats que ces derniers.*

Yann MATHET : yann.mathet@unicaen.fr

Titre : Une contribution à la linguistique computationnelle et au TAL : de la sémantique de l'espace et du temps à l'annotation et aux mesures d'accord inter-annotateurs.

Mots-clés : Annotation, mesures d'accord inter-annotateurs, sémantique, temps.

Title: *A Contribution to Computational Linguistics and Natural Language Processing: from the Semantics of Space and Time to Annotations and Agreement Measures*

Keywords: *Annotation, agreement-measures, semantics, time.*

Habilitation à diriger des recherches en Informatique, GREYC, UMR 6072, Université de Caen Normandie, sous la direction de Marc Spaniol (Pr, Université de Caen). Habilitation soutenue le 05/12/2017.

Jury : M. Marc Spaniol (Pr, Université de Caen, directeur), M. Jean-Yves Antoine (Pr, Université François Rabelais, Tours, rapporteur), Mme Delphine Battistelli (Pr, Université Paris X, rapporteur), M. Massimo Poesio (Pr, University of Essex, Royaume-Uni, rapporteur), M. Frédéric Landragin (DR, CNRS, Lattice, Montrouge, examinateur), M. Pierre Zweigenbaum (DR, CNRS, Limsi, Orsay, examinateur).

Résumé : *This study addresses two different questions in the fields of Computational Linguistics (CL) and Natural Language Processing (NLP): the question of how to model natural language semantics, especially in space and time paradigms, and the question of how to annotate corpora. These seemingly different questions are tied by the fact that when studying how to model some linguistic phenomena, for instance in semantics, it is necessary to get annotated data related to these phenomena, first to get inspiring examples of what is really studied, and second to assess our models by confronting their productions with reference annotations. Precisely, because of these*

ties, my research domain has progressively widened from pure semantics questions to questions about annotation.

In my PhD thesis, I addressed spatio-temporal semantics as it appears in natural language. Most available models rely on so-called topological relations, where the very questions is "in what place is X located?" These models fail to render the semantics of many expressions which cannot be described in terms of being located into a place, nor in terms of going into (or getting out of) a place. For instance, the sentence "(the road / the car) circumvents the city" involves a complex relationship between the shape of the road or of the trajectory of the car and the city (in addition to a topological relation of exteriority). I introduced the limits of these models and proposed solutions. Subsequently, my work has been focusing more and more on the semantics of time, in collaboration with other computer scientists, and also a linguist. In particular, we have addressed the question of how repetition (iterative events) is conveyed in natural language, in such examples as: "every Thursday, they played cards. The game lasted about 2 hours", and how to model it. One of the main results of this study is that natural language is able to handle an iterative event as if it were a sole generic event. This is clearly visible in the second sentence by the use of the singular "the game" which surprisingly refers to a plurality of games. We have designed a model which accounts for that, and for a wide range of related phenomena.

At the same time, several collaborations in CL and NLP research projects led us to focus more and more on annotation process. In particular, the ANNODIS project consisted in creating and providing a discourse relations corpus, and made appear the need for new methods and tools to annotate texts. Together with a colleague, Antoine Widlöcher, we designed and developed a versatile annotation platform, namely Glozz, which not only fulfills the ANNODIS requirements, but also fits a wide range of projects worldwide. Producing annotations brings another question: how to make sure that annotations are valid? Consequently, we have studied the existing methods to assess annotations, and we found that most of them do not fit CL nor NLP purposes. In particular, CL and NLP mainly refer to linguistic streams (texts, videos), whereas most used assessment methods concern sets of independent items. As a consequence, in many cases, scholars do not use relevant measures to assess their annotations, which leads to strong biases in the results. Here again, we have proposed solutions with a new set of agreement measures, namely the Gamma family. Besides, this work goes along with a more general reflection on the principles of assessment methods, which is an additional contribution.

URL où le mémoire peut être téléchargé :
https://mathet.users.greyc.fr/