Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

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

Vincent CLAVEAU : vincent.claveau@irisa.fr

Titre : Du traitement des langues en recherche d'information et vice versa

Mots-clés : traitement automatique des langues, recherche d'information, intelligence artificielle.

Title: *About Natural Language Processing for Information Retrieval and vice versa*

Keywords: *natural language processing, information retrieval, artificial intelligence.*

Habilitation à diriger des recherches en informatique, IRISA, Université de Rennes 1. Habilitation soutenue le 10/01/2020.

Jury : Mme Adeline Nazarenko (Pr, Université Paris-Nord, présidente), Mme Catherine Berrut (Pr, Université de Grenoble Alpes, rapporteuse), M. Philippe Langlais (Pr, Université de Montréal, Canada, rapporteur), M. Jacques Savoy (Pr, Université de Neuchâtel, Suisse, rapporteur), M. Patrice Bellot (Pr, Université Aix-Marseille, examinateur), M. Olivier Dameron (MC, Université Rennes 1, examinateur).

Résumé : *La recherche d'information (RI) et le traitement automatique des langues (TAL) sont deux domaines de recherche de l'informatique partageant en commun leur matériau premier : la langue. Pourtant, à quelques exceptions notables près, ces deux domaines ont longtemps évolué indépendamment, avec peu d'interactions.*

C'est ce rapprochement entre TAL et RI, pour peu que l'on veuille les distinguer, qui est le fil conducteur principal de ce manuscrit. Au travers de la présentation d'une partie de nos travaux, nous montrons les allers-retours, les convergences, les synergies, qu'il peut y avoir entre ces deux domaines.

Ce document n'offre donc pas une vue exhaustive de nos travaux, ni en largeur (tous n'y sont pas présentés), ni en profondeur (tous les détails techniques n'y sont pas

reportés). Ce document n'est pas non plus une revue de l'état de l'art, mais pour situer les travaux présentés dans un contexte plus large, nous proposons deux brefs panoramas des interactions entre TAL et RI.

Au travers d'une sélection de nos travaux passés, nous montrons ainsi tous les bénéfices à croiser les connaissances acquises dans chacun de ces domaines. Précisément, nous avons articulé ce mémoire en deux parties, l'une dédiée aux apports du TAL pour la RI, et l'autre aux apports de la RI pour le TAL. Nous revisitons ainsi plusieurs de nos contributions sur, d'une part, la morphologie, la translittération, la segmentation thématique, l'analyse fine de termes médicaux dans un contexte de RI, et d'autre part, sur l'utilisation des moteurs de recherche comme classifieurs, les tâches de RI comme techniques d'évaluation de techniques de TAL, la sémantique distributionnelle et les plongements de mots par et pour la RI. Nous discutons également de la pertinence de cette dichotomie entre ces deux domaines à l'heure de l'intelligence artificielle, et de la convergence de leur corpus technique (notamment les approches neuronales). Nous présentons enfin quelques enjeux de recherche à la croisée de ces domaines.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03027676

Alice MILLOUR : alice.millour@abtela.eu

Titre : Myriadisation de ressources linguistiques pour le traitement automatique de langues non standardisées

Mots-clés : myriadisation, traitement automatique des langues, langues peu dotées, langues non standardisées, corpus annoté, morphosyntaxe, annotation manuelle.

Title: *Crowdsourcing Linguistic Resources for Non-standardised Languages Processing*

Keywords: *crowdsourcing, natural language processing, less-resourced languages, non-standardized languages, annotated corpora, part-of-speech, manual annotation.*

Thèse de doctorat en informatique, Sens Texte Informatique Histoire, Sociologie et Informatique pour les Sciences Humaines, Sorbonne Université, Paris, sous la direction de Karën Fort (MC, Sorbonne Université) et Claude Montacié (Pr, Sorbonne Université). Thèse soutenue le 14/12/2020.

Jury : Mme Karën Fort (MC, Sorbonne Université, codirectrice), M. Claude Montacié (Pr, Sorbonne Université, codirecteur), M. Laurent Besacier (Pr, Université Grenoble Alpes, Laboratoire d'Informatique de Grenoble, rapporteur), M. Benoît Sagot (DR, Inria, rapporteur), Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre, présidente), Mme Delyth Prys (Pr, Bangor University, Royaume-Uni, examinatrice).

Résumé : *Les sciences participatives, et en particulier la myriadisation (*crowdsourcing*) bénévole, représentent un moyen peu exploité de créer des*

ressources langagières pour certaines langues encore peu dotées, et ce malgré la présence de locuteurs sur le Web. Or, le développement de technologies du langage est très fortement dépendant de l'existence de ressources pérennes, qu'elles soient brutes ou annotées.

Nous présentons dans ce travail de thèse nos expériences de production participative de ressources et de développement — grâce à ces ressources — d'outils d'annotation automatique en parties du discours. Nous avons appliqué notre méthodologie à trois langues non standardisées, en l'occurrence l'alsacien, le créole guadeloupéen et le créole mauricien. Pour des raisons historiques différentes, de multiples pratiques (ortho)graphiques co-existent en effet pour ces trois langues. Les contextes linguistiques choisis nous ont confrontée à l'adaptabilité des méthodes habituellement employées pour développer des outils en TAL. En particulier, les difficultés posées par l'existence de cette variation nous ont menée à proposer trois tâches de myriadisation permettant respectivement la collecte de corpus bruts, l'annotation en parties du discours de ces corpus, et la production de variantes graphiques.

L'analyse intrinsèque et extrinsèque de ces ressources recueillies auprès de locuteurs montre l'intérêt d'utiliser la myriadisation dans un cadre linguistique non standardisé : les locuteurs ne sont pas considérés dans notre travail comme un ensemble uniforme de contributeurs dont les efforts cumulés permettent d'achever une tâche particulière, mais comme un ensemble de détenteurs de connaissances complémentaires. En outre, la variation graphique observée tend à dégrader les performances des outils reconnus comme performants dans des contextes standardisés. Ainsi, parallèlement à la définition d'une méthodologie de collecte de ressources variées, nous menons une évaluation de l'impact de la variation sur les performances des outils entraînés, puis nous proposons une démarche qui vise à intégrer ces ressources variées au développement d'outils plus robustes.

La qualité des ressources produites au cours de ce travail et les gains observés quant aux performances des outils entraînés nous permettent de conclure au bien-fondé de l'utilisation de la myriadisation pour le développement de ressources langagières dans ces contextes linguistiques particuliers. Les plateformes développées, les ressources langagières, ainsi que les modèles d'outils d'annotation entraînés sont librement disponibles.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-03083213

Bénédicte PIERREJEAN : benedicte.pierrejean@gmail.com

Titre : Évaluation qualitative des *word embeddings* : étude de l'instabilité dans les modèles neuronaux

Mots-clés : *word embeddings*, sémantique distributionnelle, évaluation qualitative.

Title: *Qualitative Evaluation of Word Embeddings: Investigating the Instability in Neural-Based Models*

Keywords: *word embeddings, distributional semantics, qualitative evaluation.*

Thèse de doctorat en sciences du langage, CLLE-ERSS, UMR 5263, Université Toulouse 2 - Jean Jaurès, sous la direction de Ludovic Tanguy (MC, Université Toulouse 2 - Jean Jaurès). Thèse soutenue le 08/01/2020.

Jury : M. Ludovic Tanguy (MC, Université Toulouse 2 - Jean Jaurès, directeur), M. Olivier Ferret (IR, CEA LIST, rapporteur), M. Alessandro Lenci (Pr, University of Pisa, Pise, Italie, rapporteur), Mme Cécile Fabre (Pr, Université Toulouse 2 - Jean Jaurès, examinatrice), Mme Aurélie Herbelot (*assistant professor*, University of Trento, Trente, Italie, examinatrice).

Résumé : *Distributional semantics has been revolutionized by neural-based word embeddings methods such as word2vec that made semantics models more accessible by providing fast, efficient and easy-to-use training methods. These dense representations of lexical units based on the unsupervised analysis of large corpora are more and more used in various types of applications. They are integrated as the input layer in deep learning models, or they are used to draw qualitative conclusions in corpus linguistics. However, despite their popularity, there still exists no satisfying evaluation method for word embeddings that provides a global yet precise vision of the differences between models. In this PhD thesis, we propose a methodology to qualitatively evaluate word embeddings and provide a comprehensive study of models trained using word2vec. In the first part of this thesis, we give an overview of distributional semantics evolution and review the different methods that are currently used to evaluate word embeddings. We then identify the limits of the existing methods and propose to evaluate word embeddings using a different approach based on the variation of nearest neighbors. We experiment with the proposed method by evaluating models trained with different parameters or on different corpora. Because of the non-deterministic nature of neural-based methods, we acknowledge the limits of this approach and consider the problem of nearest neighbor's instability in word embeddings models. Rather than avoiding this problem we embrace it and use it as a means to better understand word embeddings. We show that the instability problem does not impact all words in the same way and that several linguistic features are correlated. This is a step towards a better understanding of vector-based semantic models.*

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-02628954