L’étiquetage sémantique automatique de textes littéraires au regard de l'apprenant d'une langue seconde

Encadrant de thèse :
Mathieu Constant (Université de Lorraine, CNRS, ATILF)
Mathieu.Constant@univ-lorraine.fr

École Doctorale :
École Doctorale Stanislas

Laboratoire d’accueil :
Analyse et Traitement Informatique de la Langue Française (ATILF), UMR 7118

Domaine :
Traitement automatique des langues

Contexte :

Le numérique a révolutionné les usages des citoyens pour accéder aux connaissances et à la culture. Parallèlement, le développement de l’intelligence artificielle a amplifié le phénomène en permettant la conception d’outils de traitement de données de plus en plus pointus et accessibles au grand public. Par exemple, l’intégration de techniques d’apprentissage automatique issues de l’intelligence artificielle pour développer des outils du traitement automatique des langues (TAL) a permis au grand public d’accéder encore plus facilement aux ressources langagières dont la quantité a explosé ces dernières années : ex. moteurs de recherche, outils de traduction automatique, systèmes de résumé automatique, …

L’intelligence artificielle est à la fois source d’espoirs et de craintes pour le grand public. D’un côté, les outils font des avancées significatives sur certaines tâches contraignantes que l’on pensait réservées aux humains. D’un autre côté, avec le développement de modèles de complexité croissante, il est de plus en plus difficile pour les chercheurs (et a fortiori pour le citoyen) de comprendre en profondeur le comportement et les limites des modèles appris automatiquement. Le domaine du TAL n’échappe pas à ce constat. Les performances sur les tâches classiques de ce domaine ont été spectaculairement améliorées. Mais, leur évaluation se borne très souvent à la comparaison quantitative entre les résultats obtenus automatiquement en utilisant le modèle appris et les résultats attendus par un expert de la tâche sur des jeux de données de référence. Une étude qualitative plus fine des modèles est la plupart du temps oubliée, mauvaise pratique qui ouvre la porte à de nombreux fantasmes sur l’intelligence artificielle.
Devant ce problème, un nombre croissant de chercheurs prend lentement conscience de la nécessité de faire évoluer les pratiques d’évaluation pour une compréhension plus profonde des modèles appris, avec, notamment, pour objectif de démystifier l’intelligence artificielle aux yeux du grand public.

Sujet :

Le projet doctoral proposé s'inscrit dans le domaine du traitement automatique des langues (TAL). Il consiste, en premier lieu, à étudier une approche expérimentale originale ayant pour objectif à la fois de développer de nouvelles méthodes performantes d'apprentissage automatique pour une tâche donnée du TAL, et de mieux appréhender qualitativement les modèles appris à partir de ces méthodes.

Notre cas d'étude est l'étiquetage sémantique de textes littéraires. Etant donné une unité lexicale en contexte, le but est de prédire automatiquement son sens parmi un ensemble de sens possibles trouvés dans une ressource lexicale. C'est une tâche historique du TAL, aussi connue sous le nom de *levée d'ambiguïté sémantique*, qui est difficile à cause de la forte ambiguïté sémantique des unités lexicales (quand elles sont prises hors contexte). De nombreux travaux de recherche ont déjà été réalisés sur ce sujet avec différentes approches (par ex. Lesk 1986, Schutze 1998, Mihalcea et al. 2004, Véronis 2004, Audibert 2004, Marquez et al. 2006, Agirre et al. 2014, Parsini et Navigli 2017). L'originalité de la thèse résidera dans la combinaison de trois points :

(i) l'exploitation d'une ressource lexicale de nouvelle génération : le Réseau Lexical du Français (Polguère 2014), construit à l'ATILF. Cette ressource encore en construction est incomplète. Une des difficultés consistera notamment à repérer quand un sens n'est pas présent dans la ressource.

(ii) l'exploitation des nouvelles techniques d'apprentissage automatique s'appuyant sur des algorithmes à base de graphes (ex. Agirre et al. 2014) via des méthodes semi-supervisées (Yuan et al. 2016, Parsini et Navigli 2017).

(iii) l'adaptation des modèles aux textes littéraires, dont la finesse lexicale peut poser de sérieux problèmes aux systèmes d'étiquetage automatique. Par exemple, une occurence d'unité lexicale dans un contexte donné pourra recouvrir plusieurs sens dans la ressource. Des méthodes génériques d'adaptation existent (ex. Sens et Tou 2007). Il s'agira de spécialiser ces méthodes au domaine étudié. En particulier, elles seront appliquées à la base de données textuelles Frantext (http://www.frantext.fr) hébergée et maintenue à l'ATILF.

L'autre originalité de cette thèse résidera dans le protocole d'évaluation. En particulier, il s'agira d'analyser comparativement les modèles appris automatiquement et les résultats obtenus sur un jeu de données, à la lumière du comportement et des productions d'apprenants humains de langue seconde (L2) en effectuant la même tâche sur les mêmes données dans un environnement d'apprentissage dédié. L'idée est d'exploiter des méthodes statistiques de calcul de corrélation afin d'établir des liens (ou pas) entre les méthodes d'apprentissage automatique et humain. Il sera fondamental de définir un protocole expérimental rigoureux pour introduire le moins de biais possible. En particulier, il sera nécessaire d'étudier scrupuleusement l'environnement de travail de l'apprenant, son niveau de langue et autres caractéristiques sociologiques. Les conclusions tirées permettront d'avoir une meilleure compréhension des modèles produits. Si elle est observée, la complémentarité entre l'apprentissage automatique et humain, pourra être directement exploitée pour améliorer les modèles appris, en prenant en compte les productions des apprenants L2 dans les méthodes d'apprentissage automatique.

L'intérêt du TAL pour l'apprentissage des langues assisté par ordinateur a été montré par de nombreux travaux de recherche (ex. Segond et Parmentier 2004, Nerbonne 2012, Kaya et Erygit 2015). L'inverse reste, à notre connaissance, une voie prometteuse de recherche à explorer dans la lignée de certains travaux dans (Poibeau et Vilavicencio 2017), de la même manière que le *crowdsourcing* pour le TAL (Fort 2016).

Environnement de travail :

La personne recrutée intégrera l'équipe *Ressources* de l'ATILF. Elle disposera d'un bureau et d'un ordinateur de travail. La pluridisciplinarité du sujet l'amènera à collaborer avec des chercheurs d'autres équipes du laboratoire (ex. équipes *Lexique* et *Didactique des langues*), et des chercheurs d'autres laboratoires (ex. IECL).

Profil attendu du candidat ou de la candidate :

Le/la candidat(e) devra posséder au moins un Master en Traitement Automatique des Langues (ou équivalent). Il ou elle devra démontrer de fortes compétences en informatique, et d'un intérêt certain pour la linguistique et la didactique des langues.

Bibliographie :

E. Agirre, O. López de Lacalle, A. Soroa (2004). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1), pp. 57–84

L. Audibert (2004). Word sense disambiguation criteria: a systematic study. *Proceedings of the 20th International Conference in Computational Linguistics* (COLING 2004).

K. Fort (2016). *Collaborative Annotation for Reliable Natural Language Processing : Technical and Sociological Aspects*, Wiley-ISTE, 196 p.

H. Kaya, G. Erygit (2015). Using Finite State Transducers for Helping Foreign Language Learning. *Proceedings of ACL-IJCNLP 2015*.

M. Lesk (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of the 5th SIGDOC, pp.* 24-26.

L. Marquez, G. Escudero, D. Martinez, G. Rigau (2006). Supervised corpus-based methods for WSD. In *Word Sense Disambiguation: Algorithms and Applications*, E. Agirre and P. Edmonds, Eds. Springer, New York, NY, 167–216.

R. Mihalcea (2004). Co-training and self-training for word sense disambiguation. *Proceedings of the 8th Conference on Computational Natural Language Learning, pp.* 33-40.

A. Mohamed (2017). Exposure frequency in L2 reading : an eye-movement perspective in incidental vocabulary. *Studies in Second Language Acquisition*, pp. 1 – 25.

J. Nerbonne (2012). Natural Language Processing in Computer-Assisted Language Learning. *The Oxford Handbook of Computational Linguistics*. Edited by Ruslan Mitkov. Oxford handbooks.

T. Pasini, R. Navigli (2017). Train-O-Matic: Large-Scale Supervised Word Sense Disambiguation in Multiple Languages without Manual Training Data. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017).*

T. Poibeau, A. Villavicencio (Eds.). (2017). *Language, Cognition, and Computational Models.* Studies in Natural Language Processing. Cambridge: Cambridge University Press.

A. Polguère (2014). From Writing Dictionaries to Weaving Lexical Networks, *International Journal of Lexicography,* 27(4), pp. 396-418.

H. Schutze (1998). Automatic word sense discrimination.*Computational Linguistics, 24(*1), pp. 97-124.

J. Veronis (2004). Hyperlex: Lexical cartography for information retrieval. *Comput. Speech Lang. 18 (*3), pp. 223-252.

Y. Seng Chan, H. Tou Ng (2007). Domain Adaptation with Active Learning for Word Sense Disambiguation. *Proceedings of ACL 2007*.

F. Segond, T. Parmentier (2004). NLP serving the cause of language learning, *Proceedings of COLING 2004 workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning*, pp. 11-17

D. Yuan, J. Richardson, R. Doherty, C. Evans, E. Altendorf (2016). Semi-supervised Word Sense Disambiguation with Neural Models. *Proceedings of COLING 2016*.