Proposition de sujet de thèse prioritaire pour l'ATILF
mai 2021

Titre : Lier des ressources lexicales du français en vue d'une interopérabilité entre niveaux linguistiques

Encadrement :

- Mathieu Constant (Professeur des universités, ATILF)
- Karën Fort (Maîtresse de conférences, LORIA)
- Bruno Guillaume (Chargé de recherche, LORIA)

Sujet :

Les ressources lexicales sont au cœur de la linguistique et du traitement automatique des langues. Cependant, les ressources existantes pour le français restent très hétérogènes dans leur conception, comme dans leur couverture. Ces ressources présentent des vues spécifiques mais complémentaires de la réalité linguistique. Par exemple, la vue dictionnairique du *Trésor de la Langue Française* vs. la vue en réseau du RL-Fr (Polguère 2014). Les ressources couvrent parfois des niveaux linguistiques différents : ex. des lexiques morphologiques (Morphalou, ATILF 2019), des lexiques syntaxiques (Dicovalence, Eynde et Mertens 2003 ; tables du lexique-grammaire, Gross 1975), des lexiques sémantiques (VerbeNet, Danlos et al. 2016). Il existe par ailleurs des modes différents de construction : ex. construction participative vs. construction par expertise linguistique.
Lier ou combiner de manière cohérente ces ressources peut permettre de donner un point de vue linguistique plus large et de mettre en valeur certains phénomènes sous différentes perspectives. Par ailleurs, les liens créés peuvent être utiles pour des tâches de traitement automatique des langues telles que l'annotation automatique de corpus (ex. étiquetage sémantique).

Dans cette thèse, nous souhaitons évaluer et comparer un certain nombre de méthodes de liage entre ressources lexicales très diverses. Généralement, lier automatiquement deux ressources consiste à lier leurs entrées respectives. La tâche revient à calculer, pour chaque entrée, un profil linguistique, puis à évaluer la proximité entre chaque paire d'entrées en comparant leurs profils. Si les profils des entrées sont considérés comme suffisamment proches, alors les entrées sont liées. Techniquement, on passe le plus souvent par une représentation par graphes[1], déjà très étudiée d'un point de vue algorithmique et en termes de modélisation informatique.
Les différentes méthodes de liage diffèrent le plus souvent dans le calcul et la représentation du profil des entrées, et dans les mesures de comparaison entre ces profils. On compte trois grands types d'approches :

(1) les approches fondées sur des heuristiques à partir du voisinage lexical au sein des ressources -- ex. BabelNet (Navigli et Ponzatto 2012) -- , et/ou à partir de la structure de la ressource -- ex. mesure de la confluence (Gaume et al. 2016) --. Elles permettent d'avoir un certain contrôle sur la procédure de liage et ont montré des

[1] Un graphe est formé d'un ensemble d'entités qui peuvent être connectées entre elles par des liens. Ces liens peuvent être typés. Ils peuvent être unidirectionnels ou bidirectionnels.

résultats intéressants. Il existe cependant des questions sur le fait qu'elles arrivent à passer à l'échelle avec des ressources vraiment différentes.

(2) les approches numériques fondées sur des représentations vectorielles (plongements[2]) apprises automatiquement en utilisant des méthodes auto-supervisées à partir de grandes quantités de textes et/ou de la structure des ressources lexicales : ex. plongements de mots (Mikolov et al. 2013, Devlin et al. 2019), plongements de sens (Gurevych et al. 2016), plongements de graphes (Cai et al. 2018). La proximité entre les profils se calcule à partir de distances géométriques entre vecteurs. Ces méthodes présentent plus de robustesse, mais moins de contrôle, car elles sont plus opaques.

(3) les approches hybrides qui intègrent plongements et heuristiques, afin de combiner robustesse et contrôle.

Les différentes méthodes seront évaluées sur un ensemble varié de ressources du français: TLFi (http://atilf.atilf.fr, dictionnaire, construction experte), wiktionnaire (https://fr.wiktionary.org, dictionnaire, construction participative), RL-Fr (Polguère 2014, réseau lexico-sémantique, construction experte), JeuxDeMots (Lafourcade 2020, réseau lexico-sémantique, construction participative), Dicovalence (Eynde et Mertens 2003, lexique syntaxique, construction experte), Verbenet (Danlos et al. 2016, lexique sémantique, construction experte). Une approche incrémentale sera proposée, en commençant par deux ressources et en étendant au fur-et-à-mesure à d'autres ressources.

L'évaluation des méthodes de liage sera effectuée de manière intrinsèque et extrinsèque. Concernant l'évaluation intrinsèque, étant donné la taille des ressources, il n'est pas réaliste d'effectuer une évaluation sur l'ensemble de celles-ci. Nous proposons les modes d'évaluation partielle suivants :

- *via* la comparaison avec des jeux réduits de données parfaitement annotées,
- en faisant un focus sur des phénomènes particuliers, ou certaines entrées spécifiques,
- *via* une analyse des données produites, à l'aide d'outils d'exploration de graphes du type Grew (Bonfante et al. 2018),
- *via* une plateforme de sciences participatives comme *Language Arc* du Linguistic Data Consortium (https://languagearc.com/).

Concernant l'évaluation extrinsèque, nous utiliserons les liens créés automatiquement pour enrichir les ressources utilisées pour des tâches de traitement automatique des langues, comme la levée d'ambiguïté lexicale. La comparaison des résultats produits avec et sans les liens créés permettra d'évaluer en partie la qualité des méthodes de liage.

Une des perspectives de ce travail sera de contribuer au réseau des *Linguistic Linked Open Data* (McCrae 2012) qui relie des ressources langagières multilingues à grande échelle.

[2] Les plongements permettent de représenter des objets linguistiques (ex. mots, sens) sous la forme de séquences de plusieurs centaines de nombres réels (au sens mathématique).

Environnement de recherche :
La thèse sera co-encadrée par des chercheurs de l'ATILF et du LORIA et s'inscrira dans le projet OLKi, financé par Lorraine Université Excellence. Elle bénéficiera donc d'un environnement pluridisciplinaire, qui est un point essentiel pour ce sujet de thèse. Le ou la doctorante pourra profiter de l'environnement de recherche international offert par l'action COST NexusLinguarum (https://www.cost.eu/actions/CA18209/), notamment pour financer des missions scientifiques avec différents partenaires européens.
Les co-encadrants de thèse ont déjà collaboré dans des projets de création de ressources langagières, notamment RigorMortis (Fort et al. 2018, 2020) et PARSEME-FR (Candito et al. 2020).

Références

Analyse et traitement informatique de la langue française - UMR 7118 (ATILF) (2019). *Morphalou* [Lexique]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v3.1, https://hdl.handle.net/11403/morphalou/v3.1.

H. Cai, V. W. Zheng, K. C. Chang (2018). A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. In: IEEE Transactions on Knowledge and Data Engineering 30 (9), pp. 1616–1637.

M. Candito, M. Constant, C. Ramisch, A. Savary, B. Guillaume, Y. Parmentier, S. Cordeiro (2020). A French corpus annotated for multiword expressions and named entities. *Journal of Language Modelling* 8 (2), Institute of Computer Science, Polish Academy of Sciences, Poland, pp.415-479.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: Association for Computational Linguistics.

K. van den Eynde, Mertens, Piet (2003). La valence: l'approche pronominale et son application au lexique verbal. Journal of French Language Studies 13, 63-104.

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

K. Fort, B. Guillaume, Y.-A. Pilatte, M. Constant, N. Lefèbvre (2020). Rigor Mortis: Annotating MWEs with a Gamified Platform. LREC 2020 - Language Resources and Evaluation Conference.

K. Fort, B. Guillaume, M. Constant, N. Lefèbvre, Y.-A. Pilatte (2018). "Fingers in the Nose": Evaluating Speakers' Identification of Multi-Word Expressions Using a Slightly GamifiedCrowdsourcing Platform. Proceedings of the Joint Workshop on Linguistic Annotation, MultiwordExpressions and Constructions (LAW-MWE-CxG-2018), Aug 2018, Santa Fe, United States. pp.207- 213.

B. Gaume, K. Duvignau, E. Navarro, Y. Desalle, H. Cheung, S.K. Hsieh, P. Magistry, L. Prévot (2016). Skillex: a graph-based lexical score for measuring the semantic efficiency of used verbs by human subjects describing actions. T.A.L. 55 (3), Numéro spécial sur Traitement Automatique des Langues et Sciences Cognitives, pp. 97–121.

M. Gross (1975). *Méthodes en syntaxe. Le régime des constructions complétives*. Paris : Hermann.

B. Guillaume, M.-C. de Marneffe, G. Perrier (2019). “Conversion et améliorations de corpus du français annotés en Universal Dependencies”. In: Traitement Automatique des Langues 60.2, pp. 71–95.

I. Gurevych, J. Eckle-Kohler, M. Matuschek (2016). Linked Lexical Knowledge Bases: Foundations and Applications. In: Synthesis Lectures on Human Language Technologies 9, pp. 1–146.

M. Lafourcade, N. Le Brun (2020). JeuxDeMots : Un réseau lexico-sémantique pour le français, issu de jeux et d’inférences. Lexique 27.

J. McCrae, G. Aguado-de-Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel- Ponsoda, D. Spohr, T. Wunner (2012). Interchanging lexical resources on the Semantic Web. In: Language Resources and Evaluation 46.4, pp. 701–719.

T. Mikolov, I. Sutskever, K. Chen, G. S Corrado, J. Dean (2013). Distributed Representations of Words and Phrases and their Compositionality. In: Advances in Neural Information Processing Systems 26, pp. 3111–3119.

R. Navigli, S. P. Ponzetto (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. In: Artif. Intell. 193, pp. 217–250.

A. Polguère (2014). “From Writing Dictionaries to Weaving Lexical Networks”. In: International Journal of Lexicography 27 (4), pp. 396–418.

A. Tchechmedjiev, T. Mandon, M. Lafourcade, A. Laurent, K. Todorov (2017). Ontolex JeuxDeMots and Its Alignment to the Linguistic Linked Open Data Cloud. In: The Semantic Web – ISWC 2017. Cham: Springer International Publishing, pp. 678–693.