Vers l'élaboration d'une ontologie interlingue pour le Lexique Scientifique Transdiciplinaire

Olivier Kraif, Brooke Stephenson

Vers l'élaboration d'une ontologie interlingue pour le Lexique Scientifique Transdiciplinaire

Olivier Kraif & Brooke Stephenson – 19 novembre 2018

LIDILEM – Université Grenoble Alpes

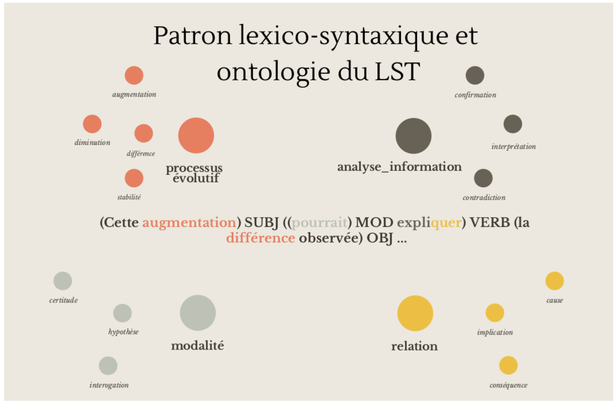
NeuroCoG – Univ. Grenoble Alpes

Contexte scientifique et objectifs du projet

Le discours scientifique intègre un lexique relevant de catégories sémantiques et épistémologiques spécifiques, le lexique scientifique transdisciplinaire (Pecman 2004, Paquot 2010, Hatier et al. 2016). Ce lexique intègre des unités lexicales comme *hypothèse*, *montrer*, *quantitatif* mais aussi des expressions polylexicales et des routines plus larges comme *obtenir des résultats encourageants*, *comme on l'a vu précédemment*, *les résultats montrent que* … La constitution d'un tel lexique est particulièrement utile pour plusieurs types d'applications : indexation, fouille de données, didactique du FOS, aide à la rédaction et à la lecture de textes scientifiques, aides à la traduction, etc.

De la lexico-syntaxe au niveau sémantique

Dans le cadre du projet ANR Termith (2012-2016), le LIDILEM a élaboré un lexique sémantique de ce type discours intégrant des étiquettes sémantiques et une organisation ontologique, à partir d'informations obtenues à partir de techniques distributionnelles appliquées à des corpus du français (Hatier et al. 2016). Dans ce type d'approche, on identifie les classes sémantiques en lien avec les patterns lexico-syntaxiques observés dans les corpus, comme dans l'exemple ci-dessous.



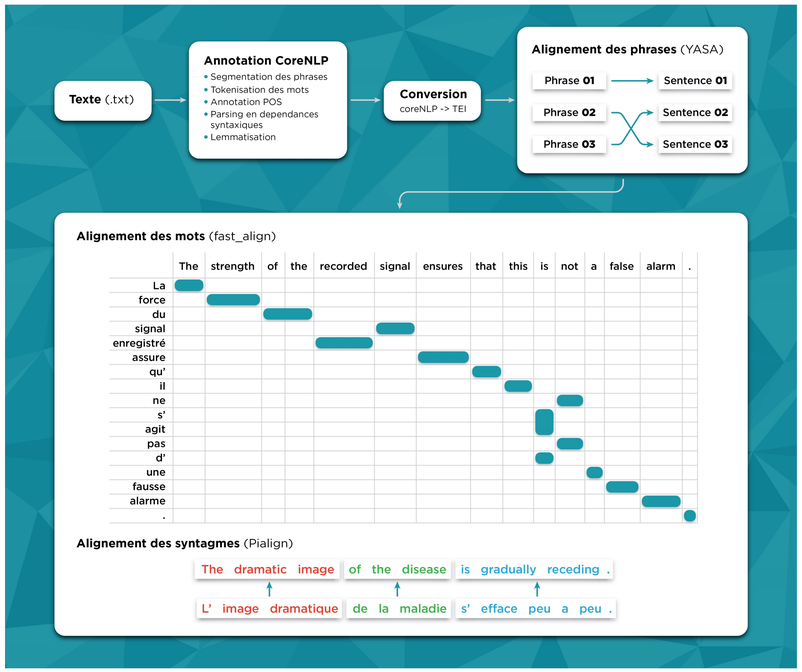
Dans le cadre du présent projet, nous souhaitons étendre ce lexique à une version anglaise, en exploitant des techniques d'alignement de corpus (Schulz et al. 2016, Och et al. 1999) et des méthodes d'analyse distributionnelle sémantique, permettant de caractériser le sens des mots à partir de leurs contextes phrastiques (Mikolov et al. 2013, Itszyler et al., 2016).

Etapes

- Elargissement d'un corpus bilingue parallèle anglais-français de textes scientifiques.
- Annotation structurelle (balisage TEI) et analyse linguistique automatique du corpus (parsing en dépendances syntaxiques).
- Alignement au plan phrastique et lexical.
- Projection du lexique scientifique transdisciplinaire (LST) sur la partie française du corpus.
- Extraction des équivalents en anglais par filtrage des expressions alignées avec des expressions du LST en français. On pourra à ce stade faire intervenir une ressource complémentaire, tel que l'interlexique établi par Gilles (2017), afin d'évaluer la qualité du lexique ainsi obtenu.
- Catégorisation sémantique. On cherchera notamment à confronter les catégories sémantiques élaborées par Hatier et al. (2016) aux mesures données par l'analyse distributionnelle sur des corpus comparables.

A terme, cette étude préparatoire permettra de jeter les bases d'une ontologie interlinguistique indépendante du français ou de l'anglais.

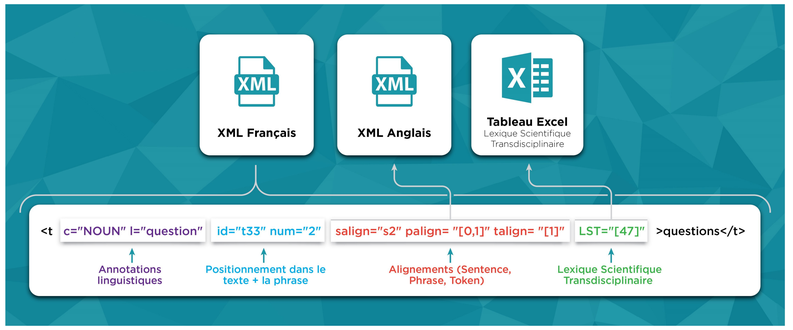
Chaine de traitement effectuée



Résultats

Les quatre premières étapes ont été accomplies :

- Un corpus d'article a été constitué depuis le portail de revue OpenEdition.org et d'autres sources → Environ 900 articles téléchargés, pour un volume d'environ 4M de mots par langue.
- L'annotation syntaxique a été faite avec CoreNLP (Manning et al. 2014), et l'alignement a été réalisé avec Yasa (Lamraoui & Langlais, 2013), FastAlign (Dyer et al. 2013) et pialign (Neubig et al. 2011).
- Les entrées du LST ont été projetées sur le corpus comme attribut additionnel. On obtient l'annotation XML schématisée ci-dessous.



Perspectives

A partir de cette ressource, nous prévoyons d'extraire les équivalents anglais du LST. Cette tâche implique une étape préalable de désambiguïsation des marques du LST. Les équivalents obtenus pourront être complétés par des données issues du lexique multilingue élboré par Gilles (2017).

La catégorisation sémantique proposée par Hatier et al. (2016) pourra alors être confontées aux données issues de la sémantique distributionnelle, en utilisant par exemple les méthodes de plongement lexical (Mikolov et al. 2013). Nous nous appuyerons enfin sur le caractère bilingue de cette catégorisation pour élaborer une ontologie visant le niveau conceptuel indépendamment de l'anglais ou du français.

Références

Dyer C., Chahuneau V., Smith N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Proc. of NAACL-HLT*, pages 644–648.

Gilles, F. (2017) Valorisation des analogies lexicales entre l'anglais et les langues romanes : étude prospective pour un dispositif plurilingue d'apprentissage du FLE dans le domaine de la santé, Thèse de doctorat, sous la dir. de C. Degache et O. Kraif, Université Grenoble Alpes.

Hatier, S., Augustyn, M., Tran, T. T. H., Yan, R., Tutin, A., Jacques, M.-P. (2016). "French Cross-disciplinary Scientific Lexicon: Extraction and Linguistic Analysis", Euralex 2016, Tbilissi, Géorgie, 6-10 September 2016.

Itszyler, E., Ribeiro, S., Sigman, M., Fernández Slezak, D. (2016) "Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database". *arXiv:1610.01520*

Lamraoui, F., and P. Langlais (2013) Yet Another Fast, Robust and Open Source Sentence Aligner. Time to Reconsider Sentence Alignment?, *XIV Machine Translation Summit*, Nice, France.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean J. (2013) Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

Neubig G., Watanabe T., Sumita E., Mori S., Kawahara T. (2011). An unsupervised model for joint phrase alignment and extraction. In *Proc. of ACL-HLT*, pages 632–641.

Och, F.J. and Tillmann, C. and Ney, H. and others (1999) Improved alignment models for statistical machine translation, *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*

Paquot, M. (2010). Academic vocabulary in learner writing: From extraction to analysis. London: Continuum.

Pecman M. (2004). Phraséologie contrastive anglais-français : analyse et traitement en vue de l'aide à la rédaction scientifique, Thèse en Sciences du Langage, Université Sophia Antipolis, UFR Lettres, Arts et Sciences Humaines

Schulz, P., Wilker A. and Sima'an, K. (2016): Word Alignment without NULL Words, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)