Chercheur/e postdoctoral/e en Sémantique Diachronique Computationnelle

Labex EFL (Empirical Foundations of Linguistics, Paris, https://www.labex-efl.fr/)

Axe 5 : Analyse Sémantique Computationnelle

Sujet: modèles computationnels interprétables pour la détection et le suivi automatiques des évolutions sémantiques : combinaison des approches *Contextual Embeddings* et *Pattern Mining*

Durée du contrat : 18 mois

Localisation : Paris

Établissement et laboratoire de rattachement : Université Sorbonne Paris Nord, LIPN UMR7030 CNRS

Date limite pour candidater : 30 avril 2022

Période des auditions : 15-30 avril 2022

Date de prise de fonctions : à partir du 1er mai 2022

Contexte, problématiques et axes de la recherche

Les langues évoluent continuellement, poussées par la double nécessité de s'adapter aux développements socioculturels et technologiques et de rendre la communication plus efficace et expressive. En particulier, des mots nouveaux sont forgés ou empruntés à d'autres langues, certains mots deviennent obsolètes, d'autres enfin acquièrent de nouvelles significations ou perdent des significations existantes.

En TAL, l'étude du dynamisme des langues, notamment du point de vue lexical, est devenu depuis quelques années un sujet de recherche important qui complète les approches synchroniques. Le champ de recherche se structure, avec des états de l'art récents (Monteirol et al., 2021; Tahmasebi et al., 2021) et plusieurs manifestations scientifiques (International Workshop on Computational Approaches to Historical Language Change 2019 et 2021, ACL 2019 et 2020). Deux premières tâches d'évaluation des systèmes de détection ont été proposées (Unsupervised Lexical Semantic Change Detection Task, SemEval2020) et des jeux de références ont été mis en place pour quatre langues (anglais, latin, suédois et allemand).

Les systèmes de détection des changements lexicaux ont suivi les avancées des méthodes de TAL : après les premiers systèmes essentiellement basés sur les évolutions de fréquence (par exemple Gulordova & Baroni, 2011), les systèmes ont utilisé les *word embeddings* (Kim et al., 2014, Schletchweg et al., 2019) puis les *contextual embeddings* (Hu et al., 2019; Martinc et al., 2019; Giulianelli et al., 2020). Ces derniers systèmes procèdent généralement

en regroupant les représentations vectorielles contextuelles des différents usages en clusters de sens, puis détectent les évolutions selon différentes métriques (Monteirol et al. 2021). Les systèmes actuels connaissent encore de nombreuses limitations. Principalement, l'opacité des modèles neuronaux ne permet pas de caractériser ces évolutions, en particulier il est difficile, voire impossible de lier les changements sémantiques à des caractéristiques linguistiques (morphologique, syntaxique, lexico-syntaxique), ou de catégoriser les types de changements (extension, restriction, métaphore, métonymie, etc.). Dans ce but, une piste serait de combiner les approches neuronales avec des approches *Pattern Mining* ou de fouille de motifs (Béchet et al. 2015) ou encore des méthodes issues de la linguistique de corpus (par exemple Gries, 2012) qui permettent d'extraire les constructions lexico-syntaxiques les plus saillantes d'un corpus d'occurrences et d'identifier leurs évolutions. Il serait également intéressant d'exploiter les informations contextuelles des occurrences des nouveaux emplois (date, type de source, de domaine, origine géographique, etc.) pour caractériser et suivre l'évolution des emplois.

L'objectif est donc de mettre en place une approche permettant de caractériser automatiquement les évolutions sémantiques. Une première étape consistera à expérimenter les travaux de l'état de l'art pour la détection des évolutions. Il s'agira ensuite à partir des *embeddings* contextuels et des corpus diachroniques de concevoir une approche pour mettre en évidence les caractéristiques linguistiques de chacun des clusters de sens et leur évolution. Les corpus étudiés seront principalement en anglais et en français. Le postdoctorant ou la postdoctorante travaillera en collaboration avec des informaticiens et des linguistes qui construisent actuellement un corpus de référence d'évolutions sémantiques pour le français (méthodologie *Durel* : Schlechtweg et al., 2018).

D'autres problématiques pourront, dans un second temps, également être abordées par la personne recrutée et notamment : les systèmes actuels ne tiennent pas compte de l'évolution graduelle, se limitant généralement à comparer deux états de langue synchroniques ; pour obtenir la représentation vectorielle d'une lexie dans un contexte, il est possible d'utiliser l'une des couches cachées ou une combinaison de celles-ci. Il n'existe pas aujourd'hui de consensus sur la couche à prendre en compte pour obtenir la représentation sémantique la plus adéquate.

La personne recrutée rejoindra, dans l'axe 5 du Labex "Sémantique computationnelle", l'équipe de chercheurs et d'enseignants-chercheurs du Labex qui travaillent sur l'opération "Variation et changement sémantique" qui vise à :

- développer de nouveaux modèles et méthodes pour la détection automatique des changements sémantiques lexicaux, la typologie des changements des points de vue intra-linguistiques, diachroniques et diastratiques ;
- développer un jeu de référence d'évolutions sémantiques pour le français contemporain, en s'appuyant sur les corpus diachroniques disponibles.

Profil recherché

– doctorat en informatique spécialisé en Traitement Automatique des Langues et Apprentissage Automatique

- maîtrise des méthodes d'apprentissage profond et des modèles de langue

– langue de travail : français et/ou anglais

Composition du dossier

- une lettre de motivation
- un descriptif du projet de recherche en lien avec la problématique à résoudre
- un CV avec liste de publications et 3 publications représentatives (pdf ou lien),
- lettres de recommandations ou noms de deux référents.

Le dossier sera envoyé à emmanuel.cartier@lipn.univ-paris13.fr et thierry.charnois@lipn.univ-paris13.fr avant le 15 janvier 2022. Les auditions des candidat(e)s pré-sélectionné(e)s auront lieu fin janvier 2022.

********** English version

Postdoctoral researcher in Computational Diachronic Semantics

Labex EFL (Empirical Foundations of Linguistics, Paris, https://www.labex-efl.fr/)

Strand 5: Computational Semantic Analysis

Research area : interpretable computational models for automatic detection and monitoring of semantic evolutions: combination of Contextual Embeddings and Pattern Mining approaches

Contract duration: 18 months

Location: Paris

Research Laboratory: Sorbonne Paris Nord University, LIPN UMR7030 CNRS

Application deadline: April 30, 2022

Audition period: April 15-30, 2022

Job Starting date: from May 1, 2022

Context, Issues and research axes

Languages are constantly evolving, driven by the need to adapt to socio-cultural and technological developments and to make communication more efficient and expressive. In particular, new words are forged or borrowed from other languages, some words become obsolete, others acquire new meanings or lose existing meanings.

In NLP, the study of language dynamics, especially from the lexical point of view, has gained audience in recent years, complementing synchronic approaches. The field of research is structuring itself, with recent state of the art (Monteirol et al., 2021; Tahmasebi et al., 2021) and several scientific events (International Workshop on Computational Approaches to Historical Language Change 2019 and 2021, ACL 2019 and 2020). Two initial evaluation tasks have been proposed (Unsupervised Lexical Semantic Change Detection Task, SemEval2020) and reference sets have been set up for four languages (English, Latin, Swedish and German).

Lexical change detection systems have followed advances in NLP methods: after the first systems essentially based on frequency changes (for example Gulordova & Baroni, 2011), systems used word embeddings (Kim et al., 2014, Schletchweg et al., 2019) and more recently contextual embeddings (Hu et al., 2019; Martinc et al., 2019; Giulianelli et al., 2020). These latter systems generally proceed by grouping the contextual vector representations of the different uses into clusters of meaning, then detect changes according to different metrics (Monteirol et al. 2021). Current systems still face many limitations. Mainly, the opacity of neural models does not make it possible to characterize these evolutions, in particular it is difficult, if not impossible, to link the semantic changes to linguistic morphological, syntactic or lexico-syntactic features, or to categorize the types of changes (extension, restriction, metaphor, metonymy, etc.). To this end, one avenue would be to combine neural approaches with *Pattern Mining* (Béchet et al. 2015) or collocation extraction approaches from corpus linguistics (for example Gries, 2012) which make it possible to extract the most salient lexico-syntactic patterns of a given meaning from a corpus of occurrences and thus identify the evolution. It would also be interesting to use the contextual information of the occurrences (date, type of source, domain, diatopic and diastratic features, etc.) to characterize and follow the evolution of usages.

The job main objective is therefore to set up a system combining these approaches to allow an automatic characterization of semantic evolutions. The first step will consist in experimenting with state-of-the-art models for detecting changes. The second step will then try to combine contextual embeddings and pattern mining approaches / collocation extraction to highlight the linguistic characteristics of each of the meaning clusters and their evolution. The studied corpora will be mainly in English and French. The postdoctoral fellow will work in collaboration with computer scientists and linguists from the Labex who are currently building a reference corpus of semantic evolutions for French (following the Durel methodology: Schlechtweg et al., 2018).

Other issues may also be addressed by the recruited person, and in particular: current systems do not take into account the graduality of evolutions, generally being limited to comparing two synchronic language states; to get the vector representation of a lexis in a

context, it is possible to use one of the hidden layers or a combination of them. There is currently no consensus on the most adequate layer to take into account to obtain the most adequate semantic representation.

The recruited person will join the strand 5 (“Computational Semantics”) of the Labex, specifically the research team working on the “Semantic Variation and Change” operation which aims to:

- develop new models and methods for the automatic detection of lexical semantic changes, the typology of changes from intra- and extra-linguistic points of view;
- develop a reference dataset of semantic evolutions in contemporary French, based on available diachronic corpora.

Candidate profile

- PhD in computer science specialised in Computational Linguistics and Machine Learning

- deep learning methods and language models attested training and experience

- working language: French and / or English

Application

Please send :

- a cover letter
- a description of the research project related to the research questions
- a CV with a list of publications and 3 representative publications (pdf or link),
- letters of recommendation or names of two referees.

to emmanuel.cartier@lipn.univ-paris13.fr and thierry.charnois@lipn.univ-paris13.fr before January 15, 2022. The auditions of the pre-selected candidates will take place at the end of January 2022.

Références

Béchet N., Cellier P., Charnois T. and Crémilleux B. (2015). “Sequence mining under multiple constraints”. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing (SAC 2015),* ACM Press, Salamanca, Spain, pages. 908--914.

Giulianelli, M., Tredici, M.D., & Fernández, R. (2020). “Analysing Lexical Semantic Change with Contextualised Word Representations”. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973 July 5 - 10, 2020. https://www.aclweb.org/anthology/2020.acl-main.365.pdf

Gries Stefan Th. (2012). "Behavioral Profiles: a fine-grained and quantitative approach in corpus-based lexical semantics". In Gonia Jarema, Gary Libben, Chris Westbury (eds.),

Methodological and analytic frontiers in lexical research, 57-80. Amsterdam Philadelphia: John Benjamins.

Montariol, S. (2021). *Models of diachronic semantic change using word embeddings. (Modèles diachroniques à base de plongements de mot pour l'analyse du changement sémantique)*. PhD Thesis, Paris-Saclay. 223 pages https://tel.archives-ouvertes.fr/tel-03199801/document

Montariol S., Doucet A. and Allauzen A. (2021). "Etat de l'art du changement sémantique à partir de plongements contextualisés". In *Coria 2021*, http://coria.asso-aria.org/2021/articles/court_27/main.pdf

Montariol, S., Martinc, M., & Pivovarova, L. (2021). "Scalable and Interpretable Semantic Change Detection". *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652 June 6–11, 2021. . https://www.aclweb.org/anthology/2021.naacl-main.369.pdf

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). "SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection". *Proceedings of the 14th International Workshop on Semantic Evaluation*, pages 1–23 Barcelona, Spain (Online), December 12, 2020. https://www.aclweb.org/anthology/2020.semeval-1.1.pdf

Schlechtweg, D., & Walde, S.S. (2020). "Simulating Lexical Semantic Change from Sense-Annotated Data". In Ravignani, A. and Barbieri, C. and Martins, M. and Flaherty, M. and Jadoul, Y. and Lattenkamp, E. and Little, H. and Mudd, K. and Verhoef, T. (Eds.): *The Evolution of Language: Proceedings of the 13th International Conference (EvoLang13)*. http://brussels.evolang.org/proceedings/paper.html?nr=9

Tahmasebi, N., Borin, L., & Jatowt, A. (2018). "Survey of Computational Approaches to Lexical Semantic Change". *Computational Linguistics*, vol. 1, n°1, https://arxiv.org/pdf/1811.06278.pdf

Tahmasebi N., Borin L., Jatowt A., Xu Y. and Hengchen S. (éds, 2021). *Computational approaches to semantic change*, Language Science Press, 396p. https://langsci-press.org/catalog/book/303

Schlechtweg D., Schulte im Walde S. and Eckmann S. (2018). Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics. https://www.aclweb.org/anthology/N18-2027.pdf