A French Reference Dataset of Semantic Innovations for Computational Linguistics

Emmanuel Cartier, University Sorbonne Paris Nord, LIPN UMR7030 CNRS,
emmanuel.cartier@lipn.univ-paris13.fr

Keywords : Lexical sematic change, French reference dataset, computational linguistics

This presentation will focus on the methodology used to build a reference dataset of nouns and verbs that have undergone a semantic evolution between 1800 and today in modern and contemporary French. This work is the first step of a project aiming at developing an interpretable automatic detection system of semantic innovations in corpora combining the most promising current computational approach, *Contextual Embeddings* - which manages to encode fine-grained linguistic characteristics of lexemes in context but remains a black box that cannot be directly interpreted -, with a more transparent method, inspired by the *behavorial profile* (Gries, 2010) and the *dynamic behavioral profile* concepts (Jensergers and Gries, 2017), which aim at spotting the lexico-syntactic prototypical uses of lexemes synchronically and diachronically.

In NLP, reference datasets allow the evaluation of automatic systems acccording to specific tasks. In lexical semantics, within the framework of the yearly SemEval competition, several tasks and gold standard datasets have been setup around Word Sense Disambiguation (see Raganato et al. 2017 for a review) with more recently a task devoted to Lexical Semantic Change (Schlechtweg et al., 2020). To build the datasets, different methods have been used, with different levels of granularity and various goals. The most intuitive method consists in establishing, for a given lexeme, the list of its meanings and then in annotating the meaning in different contexts. This method, notably used in projects such as SemCor3.0 (Miller et al., 1993) or BabelNet (Navigli and Ponzetto, 2012), has a few drawbacks : a low inter-annotator agreement, the inherent subjectivity of the initial meaning categorisation, the lack of consideration of the gradual nature of the senses in contexts, and a time-consuming annotation work. To alleviate these difficulties, another method proposes not to make any hypothesis on the meanings, but to present to the annotators pairs of contexts including the lexeme and to ask them to evaluate the similarity of meaning between the contexts, allowing on the one hand a better inter-annotator agreement, and then to generate from the evaluation results clusters of meanings. This method has been used for semantic disambiguation (Erk et al, 2013; Pilehvar et Camacho-Collados, 2019) and for semantic innovations detection tasks (Schlechtweg et al., 2018; Schlechtweg et al., 2020), by adding a temporal feature to each context. These manual methods remain expensive to implement, and do not allow the identification of prototypical meanings. Automatic methods have also been proposed to automatically generate reference sets from resources including both meanings and illustrative contexts, such as WordNet and Wiktionary, for example the X-WIC (Raganato et al., 2021) reference datasets. A limitation of this approach is that the degree of similarity between senses cannot be automatically inferred, leading to binary judgments that do not account for the semantic links and gradual semantic similarity between occurrences.

Our reference dataset combines the similarity-based and the meaning-based approaches by providing : 1/ a list of words with their meaning and prototypical contexts of use; 2/ a list of 200 contexts for each word, their date of occurrence and the similarity of usage with one of the defined meanings.

The identification of lexemes having undergone an evolution is not an easy task: the semantic evolution is generally progressive, the existing dictionaries do not immediately allow the identification of clearly distinct meanings and they adopt various definition granularity strategies. In this work, we used the French Wiktionary, known to import and simplify the TLFi data, the most detailed dictionary of contemporary French. Assuming that a polysemous lexeme (i.e. having at least two meanings) has a great chance to have undergone a semantic evolution, we extracted all the nouns and verbs from this database, to obtain a first list including both the different meanings and the associated illustrative examples. We also filtered the lexemes marked, for at least one of their meanings, as 'dated' or 'obsolete'. A manual post-processing allowed us to remove irrelevant lexemes (proper nouns, meanings too close, obsolete words, etc.). For the nearly 30,000 lexemes retained, the contexts were automatically extracted from a diachronic press corpus consisting of the newspress part of Gallica (period 1800-1940) and the JSI corpus (2014-2021). Then, starting from the senses identified by the Wiktionary, and for a sample of 30 nouns and 30 verbs, a team of linguists sought to identify exemplars of each sense in each corpus. This method made it possible to validate or not the meanings identified by the Wiktionary, to obtain a certain number of prototypical contexts for each meaning and to infer words for which meanings disappeared or appeared through time.

At the end of this manual annotation, for exploratory and validation purposes, two automatic methods were carried out: on the one hand, contextual embeddings were generated for each context of each retained lexeme and a clustering allowed to represent the meaning clusters, as well as the anchor points represented by the contextual representations of the prototypical contexts; on the other hand, a Behavioral Profile - inspired approach, focusing on typical lexico-syntactic patterns for nouns and verbs, allowed to obtain for each word the most representative pattern(s) and to plot their evolution over the period.

We will report the results obtained so far, from a sample of 30 nouns and 30 verbs, by pointing out various recurrent phenomena, in particular the semantic relations between meanings (extension/restriction of meaning, metonymy and metaphor), the existence of many ambiguous contexts, the continuum denoted by clustering through Contextual Embeddings and the joint contribution of the Embeddings and behavioral profiles approaches.

References

Erk, K. McCarthy, D. and Gaylord N. (2013). Measuring Word Meaning in Context. Computational Linguistics, 39(3):511–554.

Gries, S.T. (2010). Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. The Mental Lexicon, 5, 323-346.

Jansegers, M., & Gries, S.T. (2017). Towards a dynamic behavioral profile: A diachronic study of polysemous sentir in Spanish. Corpus Linguistics and Linguistic Theory, 16, 145 - 187.

Miller, G.A., Leacock, C., Tengi, R. and Bunker R.T. (1993). A Semantic Concordance. In Human Language Technology: Proceedings of aWorkshop Held at Plainsboro, New Jersey, March 21-24, 1993.

Navigli, R. and Ponzetto S.P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193:217–250.

Pasini, T., & Camacho-Collados, J. (2020). A Short Survey on Sense-Annotated Corpora. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5759–5765 Marseille, 11–16 May 2020*. https://aclanthology.org/2020.lrec-1.706.pdf

Pilehvar M.T. and Camacho Collados J. (2019). WiC: the Word-in-Context dataset for evaluating context-sensitive meaning representations. In Proceedings of the 2019 Conference of the NaaCL: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1267–1273, Minneapolis, Minnesota.

Raganato, A., Pasini, T., Camacho-Collados, J., & Pilehvar, M.T. (2020). XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization. proceedings of EMNLP 2020. https://aclanthology.org/2020.emnlp-main.584.pdf

Sajous, Franck S. & Nabil Hathout (2015). GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In Proceedings of the eLex 2015 conference, 405–426. Herstmonceux, England.

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, pages 169–174, New Orleans, Louisiana.

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.