Semantic Shifts

Fréjus - Destremau
12-23 Sept. 2022

Semantic Shifts: From lexicon to grammar. Diachronic and typological perspectives

Workshops, Methods & Tools 1

NLP for Semantic Change

Emmanuel Cartier (LIPN, Univ. Sorbonne Paris Nord)

Lexical Change and Variation are both essential characteristics of linguistic systems. In NLP, the topics is a recent growing area of research, with reference datasets, experimental protocols and algorithms (see Kutusov et al., 2018; Tahmasebi et al., 2019 for a review).

During the training course, we will present and illustrate three main NLP approaches to track semantic change and variation, as well as to follow the life-cycle of emerging new words and/or meaning. They rely respectively on cognitive, linguistical and sociolinguistical properties of the evolution of form-meaning pairs.

The first and most evident approach consists in tracking the frequency evolution of words through time and through varieties of language. Frequency has long been recognized as a huge signal of exposure and entrenchment of lexical usage (Ellis et al., 2012, for example). We will detail several simple and nevertheless powerful techniques enabling to detect emergence, obsolescence, trends of evolution, and temporal clusters (Koplening, 2018; Kulkarni et al., 2015; Hilpert and Gries, 2016). The model of successful innovation (Rogers, 2003 [1962]) will be critically presented, showing the various paths of evolution (Nevalainen, 2015, 2018; Feltgen et al., 2017).

A more linguistically-grounded approach consists in studying the combinatorial profile (Gries, 2010) of lexemes and its evolution through time. Based on notions like collocation and collostruction (Stefanovistsch & Gries, 2003) , from quantitatively significant corpus, several measures have been proposed to approximate the behavior of lexemes and detect lexical, syntactic or lexico-syntactic change, signaling new meanings.

A second and complementary approach, grounded on the distributional hypothesis that words sharing the most contexts are most semantically similar (Harris, 1954; see Turney & Pantel, 2010 and Baroni & Lenci, 2010 for a computational presentation), enables to follow semantic change through the evolution of the cluster of similar words (i.e. notably synonyms, antonyms, hypernyms, hypernyms, co-hyponyms and meronyms). From the word2vec (Mikolov et al., 2013a and 2013b) initial popular model to the BERT family transformers (Vaswani et al., 2018; Devlin et al. 2018), which have become an essential initial step in most NLP systems, we will show through concrete examples how these models can help detect and follow the evolving linguistic properties of lexemes.

A more sociolinguistical-based approach, notably grounded on the social network structure (Milroy and Milroy, 1985) and the community practice concept (Eckert, 2012), enables to follow the life-cycle

of lexical usage through the linguistic communities. Some preliminary experiments have been setup, mainly from online social networks (Eisenstein et al, 2014; Grieve et al., 2018), showing the paths of diffusion by using sociological properties of people and the structure of communication ties (see Nguyen et al., 2016; Clem, 2016 for a review).

Practical examples and codes demonstrating the current state-of-the-art in semantic change automatic tracking will illustrate the methods, their strengths and limits, and avenues for future research and collaboration.

Provisional Planning of sessions

Session 1

- frequency evolution : 20 mns
- combinatorial profile evolution (Collocations, Collostructions and their combination) : 30 minutes
- social network analysis for semantic change tracking : 30 minutes

Session 2

- distributional profile evolution (Word and Neural Contextual Embeddings) : 30 minutes
- Hands-on session on Word Embeddings : 30-45 minutes.

Recommended readings

Hilpert Martin & Gries Stefan T., 2016, « Quantitative approaches to diachronic corpus linguistics ». The Cambridge handbook of English historical linguistics, 36-53.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, & Erik Velldal. 2018. « Diachronic word embeddings and semantic shifts: a survey ».

Nguyen, Dong, Doğruöz, A. Seza, Rosé, Carolyn P. et al., 2016, « Computational sociolinguistics: A survey ». Computational linguistics, 42(3), 537-593.

Tahmasebi Nina, Borin Lars & Jatowt Adam, 2018, « Survey of Computational Approaches to Lexical Semantic Change ». arXiv preprint arXiv:1811.06278.

References

Baroni Marco & Lenci Alessandro, 2010, « Distributional memory: A general framework for corpus-based semantics ». Computational Linguistics, 36(4):673-721.

Clem, Emily, 2016, « Social network structure, accommodation, and language change ». UC Berkeley PhonLab Annual Report, 12(1).

Devlin Jacob, Chang Ming-Wei, Lee Kenton & Toutanova Katarina, 2018, « Bert: Pre-training of deep bidirectional transformers for language understanding ». In Proceedings of the 2019 Conference of NAACL-HLT, 2018.

Eckert Penelope, 2012, « Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation ». Annual review of Anthropology, 41:87-100.

Eisenstein, Jacob, O'Connor, Brendan, Smith, Noah A., et al., 2014, « Diffusion of lexical change in social media ». PloS one, 9(11).

Ellis Nick C., 2012, « What can we count in language, and what counts in language acquisition, cognition, and use?" In S. Th. Gries ; D. S. Divjak (Eds.) Frequency effects in language learning and processing (Vol. 1). (pp. 7-34). Berlin: De Gruyter Mouton.

Feltgen Quentin, Fagard Benjamin & Nadal Jean-Pierre, 2017, « Frequency patterns of semantic change: corpus-based evidence of a near-critical dynamics in language change ». Royal Society Open Science, The Royal Society, 2017, 170830, 4 (11).

Goldberg, Yoav, & Omer Levy, 2014, « word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method », arXiv preprint arXiv:1402.3722.

Gries Stefan Th., 2012, « Behavioral Profiles: a fine-grained and quantitative approach in corpus-based lexical semantics ». In Gonia Jarema, Gary Libben, and Chris Westbury (eds.), Methodological and analytic frontiers in lexical research, 57-80. Amsterdam and Philadelphia: John Benjamins.

Grieve, Jack, Andrea Nini, and Diansheng Guo, 2018, « Mapping lexical innovation on American social media ». Journal of English Linguistics 46.4 (2018): 293-319.

Hamilton, William L., Jure Leskovec, & Dan Jurafsky, 2016, « Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change », Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1489–1501,

Berlin, Germany, August 7-12, 2016

Hilpert Martin & Gries Stefan T., 2016, « Quantitative approaches to diachronic corpus linguistics ». The Cambridge handbook of English historical linguistics, 36-53.

Jawahar Ganesh & Seddah Djamé, 2019, « Contextualized Diachronic Word Representations ». 1st International Workshop on Computational Approaches to Historical Language Change 2019 (colocated with ACL2019), Aug 2019, Florence, Italy.

Koplenig Alexander, 2018, « Using the parameters of the Zipf-Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes - a large-scale corpus analysis ». Corpus Linguistics and Linguistic Theory, 14(1), 1-34.

Kulkarni Vivek, Al-Rfou Rami, Perozzi Bryan & Skiena Steven, 2015, « Statistically significant detection of linguistic change ». In Proceedings of the 24th International Conférence on World Wide Web, WWW '15, pages 625-635.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, & Erik Velldal. 2018. « Diachronic word embeddings and semantic shifts: a survey ».

Lim, Kyungtae, Niko Partanen, & Thierry Poibeau. 2018. « Multilingual Dependency Parsing for Low-Resource Languages: Case Studies on North Saami and Komi-Zyrian ».

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, & Jeffrey Dean, 2013a, « Distributed Representations of Words and Phrases and their Compositionality », Advances in neural information processing systems, pp. 3111-3119.

Mikolov, Tomas, Kai Chen, Greg Corrado, & Jeffrey Dean, 2013, « Efficient Estimation of Word Representations in Vector Space », arXiv preprint arXiv:1301.3781.

Milroy James & Milroy Lesley, 1985, « Linguistic change, social network and speaker innovation ». Journal of Linguistics, 21(2):339-384.

Nevalainen Terttu, 2015, « Descriptive adequacy of the S-curve model in diachronic studies of language change ». Studies in Variation, Contacts and Change in English 16.

Nevalainen Terttu, Laitinen Mikko, Nevala Minna et al., 2018, « Changes in different stages: From nearing completion to completed ». In T. Nevalainen, M. Palander-Collin, and T. Saily (Eds.), Patterns of Change in 18th-century English: A Sociolinguistic Approach (pp. 251-254).

Nguyen, Dong, Doğruöz, A. Seza, Rosé, Carolyn P. et al., 2016, « Computational sociolinguistics: A survey ». Computational linguistics, 42(3), 537-593.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, & Luke Zettlemoyer. 2018. « Deep contextualized word representations », Proceedings of NAACL-HLT 2018, pages 2227–2237, New Orleans, Louisiana, June 1 - 6, 2018.

Rogers Everett M., 2003 [1962], Diffusion of innovations (5th ed.). New York, NY: Free Press.

Ruder, Sebastian, Ivan Vulić, & Anders Søgaard. 2017. « A Survey Of Cross-lingual Word Embedding Models », Journal of Artificial Intelligence Research, 65, 569-631.

Stefanowitsch Anatol & Gries Stefan T., 2003, « Collostructions: Investigating the interaction of words and constructions ». International journal of corpus linguistics, 8(2):209-243.

Tahmasebi Nina, Borin Lars & Jatowt Adam, 2018, « Survey of Computational Approaches to Lexical Semantic Change ». arXiv preprint arXiv:1811.06278.

Turney Peter D. & Pantel Patrick, 2010, « From frequency to meaning: Vector space models of semantics ». Journal of artificial intelligence research, 37, 141-188.

Vaswani Ashish, Shazeer Noam, Parmar Niki et al., 2017, « Attention Is All You Need ». In Advances in Neural Information Processing Systems.

Yao Zijun, Sun Yifan, Ding Weicong et al., 2018, « Dynamic word embeddings for evolving semantic discovery ». WSDM ’18, pages 673-681, ACM.