Andres Karjus, Richard A. Blythe, Simon Kirby and Kenny Smith
University of Edinburgh

RUSE 2019

Changing communicative need predicts lexical competition and contributes to language change

Large diachronic text corpora, as samples of utterances produced by populations over time, enable a usage-based approach to the study of evolutionary dynamics in languages, while advances in machine learning allow for automatic inference of semantic proximity and meaning change (cf. Hamilton et al. 2016, Xu and Kemp 2015, Schlechtweg et al. 2017, Turney et al 2019). We present work on identifying and quantifying interactions between words and test the hypothesis that increased communicative need in a semantic subspace supports co-existence of similar words, while low communicative need leads to a survival of the fittest situation.

'Competition' is used here to refer to any processes where the usage dynamics of some word or words affect the usage of other words. An example of direct competition would be a word that goes out of usage due to being replaced by a new borrowing, or by another native word that has undergone semantic change and is being used in an overlapping sense. An example of less direct competition would be a word that goes out of usage because its entire discourse topic is going out of usage, in turn due to the rise of new topics.

Our approach to detect competitive dynamics is based on the following components. Semantic proximity is estimated using diachronic word embeddings (cf. Yao et al. 2018). Distributionally similar words are more likely to be in direct competition to be used in an utterance than unrelated ones. Frequency change is the obvious indicator of change in usage. The strength of competition is quantified using a metric based on the idea that as the usage of a word increases, some other word(s) must decrease for the probability mass to be equalized (as occurrence probabilities in a corpus segment sum up to 1). If the top synonym(s) of a target have decreased as much as the target increased, then this indicates likely competition between them.

Often entire clusters of semantically similar words increase (or decrease) together, exhibiting no sign of competition between them. We have previously quantified this effect in the topical-cultural advection model ([anonymized]), which measures the extent that a change in the frequency of a word can be attributed to the mean frequency change in its topic. We use this measure as a proxy to changing communicative need (cf. Regier et al. 2016, Gibson et al. 2017) in a semantic subspace.

We test our approach on subsets of data drawn from large diachronic corpora of multiple languages: English, German, Estonian, and Scottish Twitter English. In addition to the advection variable, we also control for a number of lexico-statistical measures, including commonality of the form, formal similarity to nearest neighbours, semantic subspace density, semantic change, frequency, momentum, and dissemination in the corpus segment. We also include a simple measure of polysemy, to control for the distorted signal of polysemous words that are typically not well captured by simple word embeddings. Semantic change is also how competition could manifest: the 'loser' could acquire a new meaning and continue to be used.

These variables are subsequently modelled in a standard regression framework. The technical complexity underlying the data collection process requires careful control for confounds and making sure the observed effects are not artefacts of the machine learning models (Dubossarsky et al. 2016). To that end, the models are compared with randomized baselines, based on randomized embeddings. Increasing communicative need in a semantic subspace, operationalized by the advection model, does predict decreased competition, while high semantic similarity in subspaces with neutral or negative advection often leads to competition. In addition to presenting completed work, we will discuss ongoing work on testing these findings experimentally.

References

Dubossarsky, H., Weinshall, D., & Grossman, E. (2017). Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1147–1156.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., Gibson, M., Piantadosi, S., Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*.

Hamilton, W.L., Leskovec, J., Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016. Volume 1: Long Papers*.

[anonymized]

Regier, T., Carstensen, A., Kemp, C. (2016). Languages Support Efficient Communication about the Environment: Words for Snow Revisited. *PLOS ONE 11, 1–17*.

Schlechtweg, D., Eckmann, S., Santus, E., im Walde, S. S., & Hole, D. (2017). German in Flux: Detecting Metaphoric Change via Word Entropy. *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 354–367.

Stewart, I., Eisenstein, J. (2018). Making "fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium*, pp. 4360–4370.

Turney, P.D., Mohammad S.M. (2019) The natural selection of words: Finding the features of fitness. *PLOS ONE 14(1): e0211512*.

Xu, Y., Kemp, C., 2015. A Computational Evaluation of Two Laws of Semantic Change. *CogSci 37*.

Yao, Z., Sun, Y., Ding, W., Rao, N., Xiong, H. (2018). Dynamic Word Embeddings for Evolving Semantic Discovery. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18*, pp. 673–681.