Cloudspotting: Visual analytics for distributional semantics

Mariana Montes
Supervised by Dirk Geeraerts, Dirk Speelman & Benedikt Szmrecsanyi

The present study is part of the Nephological Semantics research project at QLVL, which aims to develop tools for large-scale corpus-based semantic analyses. A core aspect of the project involves representing semantic structure with vector space models (VSMs), a computational tool that currently requires a deeper understanding of its inner workings and how its results relate to cognitive theories of meaning.

Count-based VSMs represent words[1] as vectors of co-occurrence frequencies in a multidimensional space (Turney and Pantel 2010; Lenci 2018). Basically, a word is represented by its association strength to other words. They can be generated at both type- and token-level (Heylen, Speelman, and Geeraerts 2012; Heylen et al. 2015; De Pascale 2019). At type level, two words are represented as more similar if they are attracted to the same contextual features (e.g. other words) and repelled by the same contextual features. This should allow us to identify semantic fields and other relationships between words, but collapses the full range of contexts of each word into one representation. At the token level, instead, we look at individual *occurrences*, and define them as more similar if the words in their contexts are attracted to and repelled by the same contextual features. This way we should be able to map the internal variation of the behavior of individual words, i.e. their semasiological structure.

Within the larger Nephological Semantics project, this work package is dedicated to the understanding of token-level vector space models as a tool for the study of polysemy. Concretely, we explore a number of parameter settings for the models, i.e. ways of defining the context used to represent each tokens, and their impact on the resulting representation, by means of visual analytics. We used manual annotation of sense tags as a heuristic, but without considering them a golden standard. Instead, we aim to map parameter settings to various semantic phenomena coded in the annotations, such as meaning granularity (e.g. distinguishing homonyms and senses within the homonyms). The vector space models, which take the form of large matrices, can be reduced to two dimensions via different methods, such as t-SNE (van der Maaten and Hinton 2008; Krijthe 2015). These coordinates can then be mapped onto a scatterplot, resulting in a variety of shapes, which we call *clouds.*

The workflow was applied to a set of 32 Dutch nouns, verbs and adjectives exhibiting a range of semantic phenomena. For each of them, 240-320 concordance lines were extracted, annotated and modeled. The combination of parameter settings, some of which included syntactic information, resulted in 200-212 different models. The models were clustered with Partition Around Medoids (Maechler et al. 2021) so that a manageable, representative set could be explored in more depth, in particular visualizing their t-SNE representations.

Preliminary results suggest that the shape of the clouds depends on the distinctiveness of the collocational patterns, which may or may not match the sense annotations. Noisier models can smooth over these sharp distinctions, while more refined models emphasize them. More importantly, there is no set of parameters that works across the board.

References

[1] The term *word* is used very loosely here to encompass different possible definitions.

De Pascale, S. 2019. "Token-Based Vector Space Models as Semantic Control in Lexical Lectometry." PhD thesis. https://lirias.kuleuven.be/retrieve/549451.

Heylen, Kris, Dirk Speelman, and Dirk Geeraerts. 2012. "Looking at Word Meaning. An Interactive Visualization of Semantic Vector Spaces for Dutch Synsets." In *Proceedings of the Eacl 2012 Joint Workshop of LINGVIS & UNCLH*, 16–24. Avignon.

Heylen, Kris, Thomas Wielfaert, Dirk Speelman, and Dirk Geeraerts. 2015. "Monitoring Polysemy: Word Space Models as a Tool for Large-Scale Lexical Semantic Analysis." *Lingua* 157 (April): 153–72. https://doi.org/10.1016/j.lingua.2014.12.001.

Krijthe, Jesse H. 2015. *Rtsne: T-Distributed Stochastic Neighbor Embedding Using Barnes-Hut Implementation*. https://github.com/jkrijthe/Rtsne.

Lenci, Alessandro. 2018. "Distributional Models of Word Meaning." *Annual Review of Linguistics* 4 (1): 151–71. https://doi.org/10.1146/annurev-linguistics-030514-125254.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2021. *Cluster: Cluster Analysis Basics and Extensions*. https://CRAN.R-project.org/package=cluster.

Turney, Peter D, and Patrick Pantel. 2010. "From Frequency to Meaning: Vector Space Models of Semantics." *Journal of Artificial Intelligence Research* 37: 141–88.

van der Maaten, L. J. P., and G. E. Hinton. 2008. "Visualizing High-Dimensional Data Using t-SNE." *Journal of Machine Learning Research* 9: 2579–2605.