Identifying shifts in metaphorical use of compound nouns over time

Chris Jenkins, Filip Miletić, Sabine Schulte im Walde
Universität Stuttgart
{christopher.jenkins, filip.miletic, schulte}@ims.uni-stuttgart.de

Nominal compounding is a productive process in many languages, providing a useful domain for computational semantic models to identify changes in literal or metaphorical use vis-à-vis the distributional similarity between compounds and their constituents. We hypothesize that compounds' senses should be distinguishable by their varying relationships to their constituents' distributions. The core of our approach is to run clustering algorithms on vector representations of compounds and their constituents (derived from BERT (Devlin et al., 2019)). The representations depend on the context that the item appears in, drawn from diachronic corpora in English and German (Davies, 2012; Berlin-Brandenburgische Aka. der Wissenschaften., 2022). Similarity between cluster distributions (for a given time period) can be correlated with present-day compositionality judgments (Cordeiro et al., 2019; Schulte im Walde et al., 2016), forming a basis to predict whether the compound was used more or less metaphorically over time. Alternatively, the divergence between cluster distributions of compounds and constituents can be used to predict clusters of sentences containing metaphorical uses of target compounds. We draw on recent work using contextual embeddings to cluster words according to their sense (e.g. Montariol et al. (2021)), and experiment with additional cluster features: a simpler vector representation (Basile et al., 2015), and frequency and productivity statistics (across different eras in the corpora). Combining representations is a means to facilitate separating groups of items during clustering (each connected to a sentence from the diachronic corpora). Overall, our focus on compounds provides a foundation to develop semantic models that can distinguish between literal and metaphorical language.

References: • Basile, P., A. Caputo, & G. Semeraro. Temporal random indexing: A system for analysing word meaning over time. *Italian J. of Comp. Ling.*, 1:55–68, 12 2015. • Cordeiro, S., A. Villavicencio, M. Idiart, & C. Ramisch. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57, 2019. • Davies, M. Expanding horizons in hist. ling. with 400M word corpus of historical Am. English. *Corpora*, 7:121–157, 2012. • Devlin, J., M. Chang, K. Lee, & K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of ACL*, pp. 4171–4186, Minneapolis, 2019. Association for Computational Linguistics • Montariol, S., M. Martinc, & L. Pivovarova. Scalable & interpretable semantic change detection. *Proceedings of NAACL.*, pp 4642–4652, 2021. Association for Computational Linguistics • Schulte im Walde, S., A. Hätty, S. Bott, & N. Khvtisavrishvili. GhoSt-NN: A representative gold std. of German noun-noun compounds. *Proceedings of LREC.*, pp. 2285–2292, Portorož, 2016. Euro. Lang. Resources Association. (ELRA). • Hrsg. von der Berlin-Brandenburgische Aka. der Wissenschaften. Deutsches Textarchiv., 2022.