Evaluating semantic composition of German compounds

Corina Dima
Universität Tübingen
cdima@sfs.uni-tuebingen.de

Jianqiang Ma
Universität Tübingen
jma@sfs.uni-tuebingen.de

Erhard Hinrichs
Universität Tübingen
eh@sfs.uni-tuebingen.de

Evaluation of composition models. Creating meaningful, reusable representations for multi-word expressions remains an open problem for distributional semantics. This work quantitatively evaluates composition functions that can construct a *composite* representation for German compounds e.g. *Apfelbaum* 'apple tree' from the representations of their constituents, e.g. *Apfel* and *Baum* (see Dima, 2015). The composite representation of a compound should ideally be indistinguishable from its *observed* representation. The observed representations of both the compounds and their constituent words are built using the GloVe method and a 10 billion token raw-text corpus. We use the rank metric to evaluate 12 composition functions on a frequency-filtered subset of the compounds available in GermaNet 9.0. The representation of the head (model 1) was used as a strong baseline, which was slightly outperformed by the weighted vector addition (model 6). The *Wmask* model (model 12) produced the best results.

Transparency and composition. Given that the meaning of non-transparent compounds cannot be inferred from that of their constituents, how does transparency affect the performance of composition functions? To answer this question, we compared the human judgments for compound transparency (Schulte im Walde et al., 2013) with the composition results of model 12 for a subset of the compounds. We found that for less transparent compounds, the composition yielded lower quality representations than for more transparent ones. While this is to be expected, the composition functions also struggled with transparent compounds whose constituents have either a metaphoric meaning (e.g. Schnee*mann* 'snowman') or multiple senses (e.g. Kaffee*mühle* 'coffee grinder'). As future work, we will address these issues by (1) identifying opaque compounds and building their representations directly and (2) using sense-aware word representations.

References: • Dima, C. (2015): Reverse-engineering Language: A Study on the Semantic Compositionality of German Compounds. In: *Proceedings of EMNLP*, 17–21. • Schulte im Walde, S., Müller, S., & Roller, S. (2013) Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In: *Proceedings of *SEM*, 255–265.