Features of Compositionality in English and German Noun-Noun-Compounds

Anna Hätty, Sabine Schulte im Walde, Stefan Bott, Nana Khvtisavrishvili
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Noun-noun compounds are complex words with two simplex nouns as constituents. In English and German, the first constituent represents the modifier of the compound, and the second constituent represents the head. A compound may have different degrees of compositionality regardings its constituents. For example, the German compound *Lederhose* ('leather trousers') is highly compositional, because the meaning of the compound can be obtained by combining the meanings of the constituent words. The compound *Sündenbock* ('sin buck', meaning scapegoat), in contrast, is rather non-compositional, because there is no obvious synchronic relation between the meaning of the compound and the meanings of the constituents. Understanding the principles of compositionality is important to understand the processing of compounds in human mind, for their semantic interpretation, and for their translation to other languages.

– A common NLP approach to compute the degree of compositionality are vector space models (Reddy et al., 2011; Schulte im Walde et al., 2013; among others): Context words for the compounds and constituents are extracted from large corpora, and the similarity between the compound and the constituent context vectors predicts the degree of compositionality of the compound. Our work considers German and English noun-noun compounds, and as the basis for our models, the German and English COW corpora[1] are used.

– As compound datasets, we rely on three existing resources: the 90 English noun-noun compounds introduced by Reddy et al. (2011), a subset of the 1,443 English compounds introduced by O'Séaghdha (2007), and the 244 German noun-noun compounds introduced by Schulte im Walde et al. (2013). In addition, we created a new dataset of German compounds comprising 1,208 noun-noun compounds. All compounds datasets have been extended to include human compositionality ratings, and semantic relations between modifiers and heads.

– Our specific interest is to determine how different properties of the compounds influence their degree of compositionality. One factor is the productivity of a constituent (i.e., the number of words a constituent can be combined with in a compound, either in the modifier or in the head position). For example, it is investigated how the compositionality of a compound like *Brunnenwasser* 'well water' with a productive modifier (*Brunnen* 'well' is the modifier in 59 different noun-noun compound types in the COW corpus: *Brunnenhaus* 'well house', *Brunnenschacht* 'well shaft', …) differs from a compound like *Ansichtskarte* 'view card' with a non-productive modifer. Other factors we explore are corpus frequency (e.g. *Filmpreis* 'film award' containing frequent constituents vs. *Torfmull* 'peat dust' containing low-frequent ones); ambiguity (e.g. *Porzellanrohling* 'porcelain blank', where the head can stand for brute or blank) and semantic relations (e.g. *Rechtsgeschichte* 'legal history' with the relation ABOUT). We use the vector space models to look into these factors regarding the various factors and both English and German data.

References

Diarmuid Ó Séaghdha. 2007. Designing and evaluating a semantic annotation scheme for compound nouns. In *Proceedings of Corpus Linguistics*.

Silva Reddy, Diana McCarthy & Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of IJCNLP,* 210-218.

Sabine Schulte im Walde, Stefan Müller & Stephen Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Proceedings of *SEM,* 255-265.

1 http://corporafromtheweb.org/