Potential and limits of distributional approaches for semantic relatedness

Sabine Schulte in Walde
University of Stuttgart, Germany

Distributional models assume that the contexts of a linguistic unit (such as a word, a multi-word expression, a phrase, a sentence, etc.) provide information about the meaning of the linguistic unit (Firth, 1957; Harris, 1968). They have been widely applied in data-intensive lexical semantics (among other areas), and proven successful in diverse research issues, such as the representation and disambiguation of word senses (Schütze, 1998; McCarthy et al., 2004; Springorum et al., 2013), selectional preference modelling (Herdagdelen and Baroni, 2009; Erk et al., 2010; Schulte im Walde, 2010), the compositionality of compounds and phrases (McCarthy et al., 2003; Reddy et al., 2011; Boleda et al., 2013), or as a general framework across semantic tasks ('distributional memory', cf. Baroni and Lenci, 2010; Pado and Utt, 2012), to name just a few examples.

While it is clear that distributional knowledge does not cover all the cognitive knowledge humans possess with respect to word meaning (Marconi, 1997; Lenci, 2008), distributional models are very attractive, as the underlying parameters are accessible from even low-level annotated corpus data. We are thus interested in maximising the benefit of distributional information for lexical semantics, by exploring the meaning and the potential of comparatively simple distributional models.

In this respect, this talk will present four case studies on semantic relatedness tasks that demonstrate the potential and the limits of distributional models.

1. Motivation: Assuming that associations reflect semantic knowledge that can be captured by distributional information, I will present a study that explores the availability of various German association norms in window co-occurrence of standard web and newspaper corpora (Schulte im Walde and Müller, 2013).

2. Compositionality: I will compare two studies on predicting the compositionality for a set of German noun-noun compounds, i.e., the degree of semantic relatedness between a compound and its constituents. One model relies on simple corpus co-occurrence features to instantiate a distributional model of the compound nouns and their nominal constituents (Schulte im Walde et al., 2013); the other model integrates the lexical information into a multimodal LDA model, accomplished by cognitive and visual modalities (Roller and Schulte im Walde, 2013).

3. Paradigmatic relations: I will present two case studies relying on word co-occurrences to distinguish between the paradigmatic relations synonymy, antonymy and hypernymy with regard to German nouns, verbs and adjectives. The first study combines a word space model with a simple co-disambiguation approach, and uses decision trees to distinguish between the relations (Scheible et al., 2013); the second study is a pattern-based approach, and uses nearest-centroid classification (Schulte im Walde and Kper, 2013).

4. Application to Statistical Machine Translation (SMT): I will describe the integration and evaluation of source-side and target-side subcategorisation information into a hierarchical English-to-German SMT system (Weller et al., 2013).