Distinguishing the Similar: An analysis of the semantic distinctions captured by distributional models of verb meaning.

Kris Heylen
kris.heylen@arts.kuleuven.be
University of Leuven

As in lexical semantics in general, distributional methods have also proven a successful technique for the automatic modeling of verb meaning. However, much more than with other lexical categories, the research into verb semantics has been based on the idea that a verb’s meaning is strongly linked to its syntactic behavior and more specifically, to its selectional preferences. This has led distributional methods of verb meaning to make use of two distinct types of syntactic contexts to automatically retrieve semantically similar verbs.

The first approach is in principle purely syntactical and looks at a verb’s distribution over subcategorization frames, i.e. the possible combinations of syntactic verb arguments like subject, direct object, indirect object etc. This purely syntactic information can be extended with some high-level semantic information like the animacy of the verb arguments (see Schulte im Walde, 2006 for an overview). Whereas this first approach is specifically geared towards verbs and is inspired by the long linguistic research tradition on ‘verb valency’, the second approach is more generally applicable to all lexical categories and was mainly developed within computational linguistics. These so-called word space models use other words as context features with a specific implementation only using those context words that co-occur in a given dependency relation to the target word (see Padó and Lapata, 2007 for an overview). Whereas this first approach is specifically geared towards verbs and is inspired by the long linguistic research tradition on ‘verb valency’, the second approach is more generally applicable to all lexical categories and was mainly developed within computational linguistics. These so-called word space models use other words as context features with a specific implementation only using those context words that co-occur in a given dependency relation to the target word (see Padó and Lapata, 2007 for an overview). In this second approach, one specific context feature corresponds to one lexeme plus its syntactic relation to the target verb, whereas in the first approach, one context feature is a possible combination of syntactic arguments that a verb can govern. Whereas the first approach is mostly used to automatically induce Levin-style verb classes, the second approach is typically applied to retrieve semantic equivalents for specific verbs (but see Li and Brew, 2008 for a comparison of the two methods on the task of inducing Levin-style classes).

In this presentation we will look more closely at the kind of semantic information that is captured by these two distinct types of distributional methods for verb meaning. For a sample of 1000 highly frequent Dutch verbs we constructed the two basic models described above from an automatically parsed corpus of Dutch newspapers. In a first step, we used all of the verb-specific dependency relations covered by the parser and in a second step we reduced the number of different dependency relations to only include core arguments (excluding so-called complements). For the comparison of the models, we first looked at the overall correlation between the verb similarities calculated based on the different models and see that they, at least partially, capture comparable semantic distances. In second analysis, we zoomed in on a number of specific verbs and compared the semantic aspects captured by the different models. The models based on subcategorization frames showed a tendency to reflect a verb’s aspectual properties whereas the models based on co-occurring lexemes demonstrate more strictly semantic and topical information.

References

