Experiments on the Automatic Induction of German Semantic Verb Classes

PhD-Thesis: Abstract

Sabine Schulte im Walde, IMS, Universität Stuttgart

This thesis investigates the potential and the limits of an automatic acquisition of semantic classes for German verbs. Semantic verb classes are an artificial construct of natural language which generalises over verbs according to their semantic properties; the class labels refer to the common semantic properties of the verbs in a class at a general conceptual level, and the idiosyncratic lexical semantic properties of the verbs are either added to the class description or left underspecified. Examples for conceptual structures are *Position* verbs such as *liegen* 'to lie', *sitzen* 'to sit', *stehen* 'to stand'. On the one hand, verb classes reduce redundancy in verb descriptions, since they encode the common properties of verbs. On the other hand, verb classes can predict and refine properties of a verb that received insufficient empirical evidence, with reference to verbs in the same class; under this aspect, a verb classification is especially useful for the pervasive problem of data sparseness in NLP, where little or no knowledge is provided for rare events. To my knowledge, no German verb classification is available for NLP applications. Such a classification would therefore provide a principled basis for filling a gap in available lexical knowledge.

The construction of semantic classes typically benefits from a long-standing linguistic hypothesis which asserts a tight connection between the lexical meaning of a verb and its behaviour, cf. Levin (1993). We can utilise this meaning-behaviour relationship in that we induce a verb classification on basis of verb features describing verb behaviour (which are easier to obtain automatically than semantic features) and expect the resulting behaviour-classification to agree with a semantic classification to a certain extent. A common approach to define verb behaviour is captured by the diathesis alternation of verbs. I have developed, implemented and trained a statistical grammar model for German which provides empirical lexical information, specialising on but not restricted to the subcategorisation behaviour of verbs. The grammar model serves as source for a German verb description at the syntax-semantic interface: The verbs are distributionally described on three levels, each of them refining the previous level by additional information. The first level encodes a purely syntactic definition of verb subcategorisation, the second level encodes a syntactico-semantic definition of subcategorisation with prepositional preferences, and the third level encodes a syntactico-semantic definition of subcategorisation with prepositional and selectional preferences. The most elaborated description comes close to a definition of verb alternation behaviour.

The automatic induction of the German verb classes is performed by the k-Means algorithm, a standard unsupervised clustering technique as proposed by Forgy (1965). The algorithm uses the syntactico-semantic descriptions of the verbs as empirical verb properties and learns to induce a semantic classification from this input data. The clustering outcome cannot be a perfect semantic verb classification, since (i) the meaning-behaviour relationship on which we rely for the clustering is not perfect, and (ii) the clustering method is not perfect for the ambiguous verb data. But the goal of this thesis is not necessarily to obtain the optimal clustering result, but to understand the potential and the restrictions of the natural language clustering approach. Only in this way we can develop a methodology which can be applied to large-scale data. Key issues of the clustering methodology refer to linguistic aspects on the one hand, and to technical aspects on the other hand.