Extracting Inflectional Paradigms from Raw Text: German as a Test Case

Amit Kirschenbaum
Universität Leipzig
amit@informatik.uni-leipzig.de

Unsupervised learning of morphology aims to capture the structure of words based on statistical methods and machine-learning applied to an unannotated corpus as input. From a theoretical perspective, unsupervised methods can provide empirical evidence in support of a morphological theory. In addition, they can have practical applications, e.g., in information retrieval and language documentation (Hammarström and Borin, 2011).

The current study focuses on extraction of inflectional paradigms from raw text in an unsupervised manner. It is inspired by Bybee's Network Model (1995), which describes the lexicon as a network of elements, where words are linked with each other through sets of phonological and semantic features. According to this model, if parallel phonological and semantic connections represent a pattern found in multiple sets of items, then these connections constitute morphological relations. Inflectional paradigms are then viewed as clusters of highly connected words.

The proposed method uses word embeddings to model semantic similarity between words, following the distributional hypothesis (Harris, 1954). Form similarity is computed by applying an approximate string matching to pairs of word forms. The two measures are combined to a single similarity score, which is then used to create a word similarity graph. Our hypothesis is that cohesive sub-graphs which are sparsely connected to each other, also termed *communities*, correspond to groups of morphologically similar words, which are candidates for inflectional paradigms. We apply the Clique Percolation Method (Palla et al., 2005) that uses cliques, fully connected sub-graphs, to identify communities in the constructed graph. The method detects morphological patterns by exploring relations within and across communities, using multiple-sequence alignment, a technique used in computational biology to detect similarities among bio-sequences (e.g., DNA, proteins). It then constructs inflectional classes where inter-communities relations are found.

The method is tested on German, a morphologically challenging language, due to the variety of the morphological processes it involves (concatenative affixation, circumfixation, stem alternation). The results show that the method outperforms equivalent unsupervised methods addressing the same task.

References: • Bybee J. (1995): Regular morphology and the lexicon. *Language and Cognitive Processes*, 10(5), 425–455 • Hammarström, H. and Borin, L. (2011): Unsupervised learning of morphology. *Computational Linguistics*, 37(2), 309–350. • Harris, Z. (1954): Distributional Structure. *Word* 10(2/3), 146–162 • Palla, G., Derényi, I., Farkas, I., Vicsek, T. (2005): Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814–818.