Unsupervised and Knowledge-free Learning of Compound Splits and Periphrases

Florian Holz, Chris Biemann

NLP Group, Department of Computer Science, University of Leipzig
{holz|biem}@informatik.uni-leipzig.de

Abstract. We present an approach for knowledge-free and unsupervised recognition of compound nouns for languages that use one-word-compounds such as Germanic and Scandinavian languages. Our approach works by creating a candidate list of compound splits based on the word list of a large corpus. Then, we filter this list using the following criteria:
(a) frequencies of compounds and parts,
(b) length of parts.
In a second step, we search the corpus for periphrases, that is a reformulation of the (single-word) compound using the parts and very high frequency words (which are usually prepositions or determiners). This step excludes spurious candidate splits at cost of recall. To increase recall again, we train a trie-based classifier that also allows splitting multi-part-compounds iteratively.
We evaluate our method for both steps and with various parameter settings for German against a manually created gold standard, showing promising results above 80% precision for the splits and about half of the compounds periphrased correctly. Our method is language independent to a large extent, since we use neither knowledge about the language nor other language-dependent preprocessing tools.
For compounding languages, this method can drastically alleviate the lexicon acquisition bottleneck, since even rare or yet unseen compounds can now be periphrased: the analysis then only needs to have the parts described in the lexicon, not the compound itself.

1 Introduction

A number of languages extensively use compounding as an instrument of combining several word stems into one (long) tokens, e.g. Germanic languages, Korean, Greek and Finnish. Compared to languages such as English, where (noun) compounds are expressed using several tokens, this leads to a tremendous increase in vocabulary size. In applications, this results in sparse data, challenging a number of NLP applications. For IR experiments with German, Braschler et al. report that decompounding results in higher text retrieval improvements than stemming [1].

As an example, consider the German compound "Prüfungsvorbereitungsstress" (stress occurring when preparing for an the exam) - without an analysis,