German decompounding in a difficult corpus

Enrique Alfonseca, Slaven Bilac and Stefan Pharies

Google, Inc.
{ealfonseca,slaven,stefanp@google.com}

Abstract. Splitting compound words has proved to be useful in areas such as Machine Translation, Speech Recognition or Information Retrieval (IR). In the case of IR systems, they usually have to cope with noisy data, as user queries are usually written quickly and submitted without review. This work attempts at improving the current approaches for German decompounding when applied to query keywords. The results show an increase of more than 10% in accuracy compared to other state-of-the-art methods.

1 Introduction

The so-called compounding languages, such as German, Dutch, Danish, Norwegian or Swedish, allow the generation of complex words by merging together simpler ones. So, for instance, the German word *Blumensträuße* (flower bouquet) is made up of *Blumen* (flower) and *sträuße* (bouquet). This allows speakers of these languages to easily create new words to refer to complex concepts by combining existing ones, whereas in non-compounding languages these complex concepts would normally be expressed using multiword syntactic constituents.

For many language processing tools that rely on lexicons or language models it is very useful to be able to decompose compounds to increase their coverage. In the case of German, the amount of compounds in medium-size corpora (tens of millions of words) is large enough that they deserve special handling: 5-7% of the tokens and 43-47% of the word forms in German newswire articles are compounds [1, 2]. When decompounding tools are not available, language processing systems for compounding languages must use comparatively much larger lexicons [3]. German decompounders have been used successfully in Information Retrieval [4, 5], Machine Translation [6–8], word prediction systems [1] and Speech Recognition [3, 9].

When decompounding German words from well-formed documents that have undergone editorial review, one can assume that most of the words or compound parts can be found in dictionaries and thesauri and the number of misspellings is low, which greatly simplifies the problem of decompounding. For example, Marek [10] observed that only 2.8% of the compounds in texts from the German computer magazine called *c't* contain at least one unknown part.

On the other hand, when working with web data, which has not necessarily been reviewed for correctness, many of the words are more difficult to analyze. This includes words with spelling mistakes, and texts that, being mostly written