A Probabilistic Model for Guessing Base Forms of New Words by Analogy

Krister Lindén

Department of General Linguistics, P.O. Box 9, FIN-00014 University of Helsinki
Krister.Linden@Helsinki.fi

Abstract. Language software applications encounter new words, e.g., acronyms, technical terminology, loan words, names or compounds of such words. Looking at English, one might assume that they appear in base form, i.e., the lexical look-up form. However, in more highly inflecting languages like Finnish or Swahili only 40-50 % of new words appear in base form. In order to index documents or discover translations for these languages, it would be useful to reduce new words to their base forms as well. We often have access to analyzes for more frequent words which shape our intuition for how new words will inflect. We formalize this into a probabilistic model for lemmatization of new words using analogy, i.e., guessing base forms, and test the model on English, Finnish, Swedish and Swahili demonstrating that we get a recall of 89-99 % with an average precision of 76-94 % depending on language and the amount of training material.

1 Introduction

New words and new usages of old words are constantly finding their way into daily language use. This is particularly prominent in quickly developing domains such as biomedicine and technology. Humans deal with new words based on previous experience: we treat them by analogy to known words. The new words are typically acronyms, technical terminology, loan words, names or compounds containing such words. They are likely to be unknown by most hand-made morphological analyzers. In some applications, hand-made guessers are used for covering this low-frequency vocabulary.

Unsupervised acquisition of morphologies from scratch has been studied as a general problem of morphology induction in order to automate the morphology building procedure. For overviews, see [8] and [2]. The problem is alleviated by the fact that there often are dictionaries available with common base forms or word roots for the most frequent words. If the inflectional patterns can be learned approximately from a corpus, the most common base forms can be checked against a dictionary in order to boost the performance of the methods. However, when we approach the other end of the spectrum, we have very rare words for which there are no ready base forms available in dictionaries and for heavily inflecting languages only 40-50 % of the words appear in base form in a