

Dietrich Klakow

Practical applications for language models

The talk will focus on practical applications for language models rather than fundamental algorithms.

The first step in building a language model is defining a vocabulary. English is a morphologically simple language. Language modelling techniques suitable for English often do not work well for other languages like Turkish, Tamil or Finish. In a first step a "morphological" analysis of the language is necessary. Simple automatic techniques for this are discussed.

The second part will first introduce a method that helps to find suitable parts (e.g. documents) of a large heterogeneous training corpus matching a specific narrow and possibly mismatched target domain. Automatically selecting articles from a language model training corpus allows for smaller models having better perplexities. Based on this and two other language model adaptation techniques so called Fast Marginal Adaptation and Class Based Adaptation are introduced and it is shown how ten utterances of a small target corpus (e.g. Switchboard) and a large mismatched background corpus (e.g. WSJ) can be used to build a competitive Switchboard LM.