Abstract

Standard machine learning approaches in NLP require large amounts of data for training. These do not exist for the majority of languages. Creating annotated data, which is crucial for supervised approaches, can be both expensive and time-consuming. The result is that languages for which such data is missing may not receive the attention of the language technology community. This thesis addresses the question whether it is possible to build an accurate NLP tool for a low-resource language using small amounts of data and some linguistic information. It also investigates whether such a tool can be further adapted to perform on a genetically related language by harnessing the power of crosslingual similarities between both languages.

We use an off-the-shelf NLP toolkit (OpenNLP) to train a number of models for the morphological analysis and part-of-speech (POS) tagging of Zulu - a South African Bantu language of the Nguni group with 10.3 million speakers. The training is done with a linguistically informed semi-supervised approach, where each POS model is incrementally augmented with different features derived by our linguistic knowledge of Zulu. The models are then tested and compared. Some implications are then discussed, having to do with the possibility to linguistically adapt and apply the best-performing model to another, closely related Bantu language.