Abstract CLIN 25:

keywords: language modeling, n-grams, word order, syntactic equivalence

title: Extending n-gram language models based on equivalent syntactic patterns

abstract:

In automatic speech recognition, the language model helps to disambiguate between words with a similar pronunciation. A standard language model is typically based on n-grams (sequences of n consecutive words) and their probabilities of occurrence. These n-gram models however suffer from data sparsity and cannot model long-span dependencies. The purpose of this research is to alleviate the former problem by automatically generating more n-grams than the ones based on surface structure.
Like many other languages, Dutch often can have the same syntactic pattern in different word orders (e.g. subject – verb inversion, switching between verb – object in head clause and object – verb in subordinate clause, conjunctions). In this work, we investigate whether we can generate new, meaningful n-grams based on these word order switches in an attempt to increase n-gram coverage. We do this in the following way: first our training data is parsed by Alpino, a dependency parser for Dutch. Based on these parses, we then extract those patterns for which the word order can be reversed and add the corresponding reversed n-grams to the language model. Some probability needs to be assigned to these extra n-grams and to that end several existing ways of redistributing probability mass are investigated. Finally, we compare the performance of the extended language model with the original n-gram model by evaluating their predictive power on a test set of Southern Dutch newspaper material.