Translation-based Word Clustering for Language Models

Joris Pelemans
KU Leuven
joris.pelemans@esat.kuleuven.be

Hugo van Hamme
KU Leuven
hugo.vanhamme@esat.kuleuven.be

Patrick Wambacq
KU Leuven
patrick.wambacq@esat.kuleuven.be

One of the major challenges in the field of language modelling (and others) is data sparsity. Even with the increasing amount of data, there is simply not enough data to reliably estimate probabilities for short word sequences, let alone full sentences. Hence, research in this field has focused largely on finding relations between words or word sequences, inferring probabilities for unseen events from seen events. In this work we focus on a new approach to cluster words by examining their translations in multiple languages. That is, if two words share the same translation in many languages, they are likely to be (near) synonyms. By adding some context to the hypothesized synonyms and by filtering out those that do not belong to the same part of speech, we are able to find meaningful word clusters. The clusters are incorporated into an n-gram language model by means of class expansion i.e. the contexts of similar words are shared to achieve more reliable statistics for infrequent words. We compare the new model to a baseline word n-gram language model with interpolated Kneser-Ney smoothing.