THE EFFECT OF WORD SIMILARITY ON N-GRAM LANGUAGE MODELS IN NORTHERN AND SOUTHERN DUTCH

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

Bruno De Laet
bruno.delaet@student.kuleuven.be
KU Leuven

Kris Demuynck
kris.demuynck@elis.ugent.be
Ghent University

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

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

In this paper we examine several combinations of classical n-gram language models with more advanced and well known techniques based on word similarity such as cache models, Latent Semantic Analysis, probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation. We compare the efficiency of these combined models to a model that combines n-grams with the recently proposed, state-of-the-art neural network-based hierarchical softmax skip-gram. We discuss the strengths and weaknesses of each of these models, based on their predictive power of the Dutch language.

In addition, we investigate whether and in what way the effect of Southern Dutch training material on these combined models differs when evaluated on Northern and Southern Dutch material. Experiments on Dutch news paper and magazine material show that topics extend well over these languages: the addition of topic models trained on Southern Dutch achieves a substantial improvement compared to an n-gram baseline, when evaluated on Northern Dutch data. On the other hand, n-gram language models trained on Southern Dutch perform worse on Northern Dutch data than they do on Southern Dutch data. This leads us to conclude that Southern and Northern Dutch differ mostly in local, more syntactic phenomena than in global phenomena like word usage and topic.