

This thesis is devoted to an empirical study of lexical association measures and their application to collocation extraction. We focus on two-word (bigram) collocations only. We compiled a comprehensive inventory of 82 lexical association measures and present their empirical evaluation on four reference data sets: dependency bigrams from the manually annotated Prague Dependency Treebank, surface bigrams from the same source, instances of surface bigrams from the Czech National Corpus provided with automatically assigned lemmas and part-of-speech tags, and distance verb-noun bigrams from the automatically part-of-speech tagged Swedish Parole corpus. Collocation candidates in the reference data sets were manually annotated and labeled as collocations and non-collocations. The evaluation scheme is based on measuring the quality of ranking collocation candidates according to their chance to form collocations. The methods are compared by precision-recall curves and mean average precision scores adopted from the field of information retrieval. Tests of statistical significance were also performed. Further, we study the possibility of combining lexical association measures and present empirical results of several combination methods that significantly improved the performance in this task. We also propose a model reduction algorithm significantly reducing the number of combined measures without a statistically significant difference in performance.