Lexical Association Measures: Collocation Extraction

Pavel Pecina

Abstract of Doctoral Thesis

This thesis is devoted to an empirical study of lexical association measures and their application for 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 the previous 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 identified 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 feature selection algorithm significantly reducing the number of combined measures.

Abstrakt doktorské disertační práce

Tato práce je věnovaná empirické studii lexikálních asociačních měr a jejich aplikaci v úloze automatické extrakce kolokací. Pozornost je soustředěna na dvouslovné (bigramové) kolokace. V rámci práce byl sestaven vyčerpávající seznam 82 lexikálních asociačních měr a provedena jejich evaluace na celkem čtyřech referenčních datových množinách: závislostních bigramech z ručně anotovaného Pražského závislostního korpusu, povrchové bigramy ze stejného korpusu, instance prvků předchozí množiny z Českého národního korpusu opatřeného automatickou lemmatizací a morfologickým značkováním a vzdálenostními verbnominálními bigramy z automaticky značkovaného švédského korpusu Parole. Kolokační kandidáti v referenčních množinách byli manuálně anotováni jako kolokace nebo nekolokace. Použité evaluační schéma je založeno na měření kvality seřazení kolokačních kandidátů dle jejich pravděpodobnosti tvořit kolokaci. Metody jsou porovnány pomocí precision-recall křivek a hodnot mean average precision, které jsou převzaty z oboru vyhledávání informací. Provedeny byly i testy signifikance výsledků. Dále je zkoumána možnost kombinování lexikálních asociačních měr a presentovány výsledky několika kombinačních metod, jejichž použití vedlo k výraznému zlepšení úspěšnosti řešení této úlohy. Dále je v práci navržen algoritmus významně redukující složitost použitých kombinačních modelů.