Abstract

Title: Processing of Turkic Languages
Author: Sibel Ciddi
Department: Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics, Charles University in Prague
Supervisor: RNDr. Daniel Zeman, Ph.D.
Abstract: This thesis presents several methods for the morphological processing of Turkic languages, such as Turkish, which pose a specific set of challenges for natural language processing. In order to alleviate the problems with lack of large language resources, it makes the data sets used for morphological processing and expansion of lexicons publicly available for further use by researchers. Data sparsity, caused by highly productive and agglutinative morphology in Turkish, imposes difficulties in processing of Turkish text, especially for methods using purely statistical natural language processing. Therefore, we evaluated a publicly available rule-based morphological analyzer, TRmorph, based on finite state methods and technologies. In order to enhance the efficiency of this analyzer, we worked on expansion of lexicons, by employing heuristics-based methods for the extraction of named entities and multi-word expressions. Furthermore, as a preprocessing step, we introduced a dictionary-based recognition method for tokenization of multi-word expressions. This method complements tokens in the larger multi-word expression lexicons, and prepares them for further morphological processing. Experiment results point out that the new addition of lexical tokens provide promising coverage increase for the text processed by TRmorph. The dictionary-based recognition method enables tokenization of multi-word expressions; and tokenized multi-word expressions help reducing morphological ambiguity. The proposed method enhances the efficiency of a purely rule-based morphological analyzer with finite-state transducers.
Keywords: morphological analysis, finite-state transducer, finite-state automata, recognition and tokenization of named entities, and multi-word expressions, morphological & lexical ambiguity.