

## Abstract

Most of the world's natural languages have complex morphology. But the expense of building a morphological analyzer by hand has prevented the development of morphological analysis systems for most of the world's languages. Unsupervised induction techniques, that learn the morphology of a language from unannotated text data, can facilitate the development of computational morphology systems for new languages. Such unsupervised morphological analysis systems have been shown to help natural language processing tasks including speech recognition (Creutz, 2006) and information retrieval (Kurimo et al., 2008b). This thesis describes ParaMor, an unsupervised induction algorithm for learning morphological paradigms from large collections of words in any natural language. Paradigms are sets of mutually substitutable morphological operations that organize the inflectional morphology of natural languages. ParaMor focuses on the most common morphological process, suffixation.

ParaMor learns paradigms in a three step algorithm. First, a recall-centric search scours a space of candidate partial paradigms for those which possibly model suffixes of true paradigms. Second, ParaMor merges selected candidates which appear to model portions of the same paradigm. And third, ParaMor discards those clusters which most likely do not model true paradigms. Based on the acquired paradigms, ParaMor builds a morphological segmentation algorithm. ParaMor, by design, is particularly effective for inflectional morphology. Other systems, such as Morfessor (Creutz, 2006), better identify derivational morphology. This thesis leverages the complementary strengths of ParaMor and Morfessor by combining the analyses from the two systems into a single compound analysis.

ParaMor and its combination with Morfessor were evaluated by participating in Morpho Challenge 2007, a peer operated competition for morphology analysis systems (Kurimo et al., 2008a). Morpho Challenge 2007 evaluated each system's morphological analyses in four languages, English, German, Finnish, and Turkish. When ParaMor's morphological analyses are merged with those of Morfessor, the resulting morpheme recall in all four languages is higher than that of any system which competed in the Challenge; in Turkish, ParaMor's recall, at 52.1%, is twice that of the next highest system.