

2007 Student Research Symposium

Language Technologies Institute



ParaMor: Finding Paradigms across Morphology

Christian Monson

Abstract:

Performance at natural language processing tasks from speech recognition (Creutz, 2006) to machine translation (Goldwater and McClosky, 2005) can improve with careful morphological analysis. But building a morphological analyzer for a language typically requires expert language knowledge that may be in short supply. Unsupervised morphology induction overcomes this knowledge bottleneck by learning to analyze the morphology of a language from nothing more than raw text in that language. My unsupervised morphology induction algorithm, ParaMor (Monson et al., 2007a), fared well in Morpho Challenge 2007 a peer operated competition on unsupervised morphology induction (Kurimo et al., 2007; Monson et al., 2007b),.

ParaMor consists of two phases. In the first phase, ParaMor identifies sets of candidate affixes from the input text that closely mimic the inflectional paradigms of that text's language. To propose these paradigms, ParaMor bundles sets of candidate stems with sets of candidate affixes. ParaMor searches among the many paradigm candidates, identifying and isolating those which propose the most likely morpheme boundaries and which explain the largest number of observed surface forms. ParaMor's second phase straightforwardly segments the corpus word forms at morpheme boundaries that are suggested by the reconstructed paradigms.

Morpho Challenge 2007 evaluated systems on their precision, recall, and balanced F_1 at identifying inflectional and derivational morphology. In the English track, ParaMor outperformed at F_1 an already sophisticated baseline induction algorithm, Morfessor (Creutz, 2006). In German, ParaMor suffered from low morpheme recall. But combining ParaMor's analyses with analyses from Morfessor resulted in a set of analyses that outperformed either algorithm alone, and that placed first in F_1 among all algorithms submitted to the German track of Morpho Challenge 2007. Deeper error analysis indicates that ParaMor and Morfessor are complementary algorithms. Morfessor focuses on precision and does not discriminate between inflectional and derivational morphology. Conversely, ParaMor's design closely models the inflectional paradigm structure of language, and attains high recall at isolating inflectional affixes.

Creutz, Mathias. *Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition*. Ph.D. Thesis in Computer and Information Science, Report D13. Helsinki: University of Technology, Espoo, Finland, 2006.

- Kurimo, Mikko, Mathias Creutz, and Matti Varjokallio. *Unsupervised Morpheme Analysis - Morpho Challenge 2007*. <<http://www.cis.hut.fi/morphochallenge2007/>>, March 26, 2007.
- Goldwater, Sharon, and David McClosky. *Improving Statistical MT through Morphological Analysis*. Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. Vancouver, B.C., Canada, 2005.
- Monson, Christian, Jaime Carbonell, Alon Lavie, and Lori Levin. *ParaMor: Minimally Supervised Induction of Paradigm Structure and Morphological Analysis*. In Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology (SIGMORPHON). Prague, Czech Republic, 2007a.
- Monson, Christian, Jaime Carbonell, Alon Lavie, and Lori Levin. *ParaMor: Finding Paradigms across Morphology*. In Proceedings of the Morpho Challenge 2007 Workshop in the CLEF 2007 Working Notes. Budapest, Hungary, 2007b. In Press.