

UNSUPERVISED INDUCTION OF NATURAL LANGUAGE MORPHOLOGICAL STRUCTURE

(Synopsis)

Ph.D. Thesis Proposal

April 8, 2005

Christian Monson

Language Technologies Institute
School of Computer Science
Carnegie Mellon University

Thesis Committee

Jaime Carbonell (Co-Chair)
Alon Lavie (Co-Chair)
Lori Levin
Ron Kaplan (PARC)

Abstract

Most natural languages exhibit inflectional morphology, that is, the surface forms of words change to express syntactic features—I *run* vs. She *runs*. The ubiquity of morphology in natural language calls for language processing systems to possess knowledge of morphological structure. Presently, most computational applications hand-encode the requisite structural morphological facts, but manually describing the morphology of a language demands linguistic and computational expertise that is often in short supply. This knowledge acquisition bottleneck can be ameliorated through the automatic induction of morphological structure from readily available machine readable natural language data.

This thesis proposal describes a methodology for acquiring morphology structure from unannotated text corpora. The unsupervised algorithms exploit constraints inherent in natural language morphology to build and search a space of candidate morphological analyses. The preliminary system, trained on an initial corpus of Spanish newswire text, already successfully identifies paradigmatic morphological structure—surpassing an F_1 measure of 0.5 against ideal Spanish inflectional sub-classes. This thesis proposal both describes the current search space creation and inference algorithms, and outlines extensions designed to speed up creation of the search space, improve the quality of the induced morphological structures, segment words into morphemes, and generalize the algorithms to additional languages.

Thesis Claims

1. A space of candidate morphological analyses can be laid out in an unsupervised fashion that includes, for any natural language, the major underlying paradigmatic sub-classes of that language. Search algorithms can scour this space for those candidates which usefully capture morphological structure.
2. The space of morphology candidates displays specific properties that search algorithms can exploit to identify morphological relationships. This thesis will leverage local paradigmatic and syntagmatic properties of the search space into novel and effective search strategies.
3. The search algorithms developed for this thesis will discover the major paradigmatic sub-classes of individual natural languages. The discovered sub-classes will then aid word-to-morpheme segmentation in those languages.
4. The developed search algorithms will reliably avoid hypothesizing as morphological sub-classes unrelated suffix sets. Similarly, the morphological segmentation algorithms developed for this dissertation will have recall and precision performance at least on par with other state-of-the-art systems.

For the full version of this thesis proposal please contact Christian Monson at:

cmonson@cs.cmu.edu