Using Hand-Written Rewrite Rules to Induce Underlying Morphology

Michael A. Tepper
University of Washington
mtepper@u.washington.edu

Abstract

Allomorphic variation, or form-variation among morphemes with the same referential meaning, is often mentioned as a stumbling block to unsupervised morphological induction. To address this problem head-on, we present a hybrid approach that uses a small amount of linguistic knowledge in the form of orthographic rewrite rules to help refine an existing segmentation. Our goal is to learn when surface morphs (units of the segmentation) should really be counted together as the same underlying morpheme. In order to do this, we customize the Morfessor algorithm and model developed by Mathias Creutz and Krista Lagus, adding segmentation analyses generated by orthographic rewrite rules along with a statistical framework to predict when analyses should be used as underlying morphemes. An initial segmentation produced by Morfessor Categories-MAP 0.9.2 is used as input. To suggest underlying morphemes, a set of language-specific orthographic rules is currently needed. Though we are not officially a part of the Challenge competition, for English and Turkish we report 62.22% and 54.83% contest F-measures, which amount to 2% and 48% improvements respectively over top unsupervised entrants for those languages.

Keywords

Morphological induction, Allomorphic variation, Knowledge-lite, Word segmentation