Towards unsupervised induction of morphophonemic rules

Erwin Chan
University of Pennsylvania
echan3@seas.upenn.edu

Abstract

The task we are investigating is unsupervised learning of natural language morphology for inflectional languages. The target morphological grammar consists of a lexicon of morphological base forms and transforms. A base form represents all inflections of a lexeme, and all base forms of the same POS category share the same fine-grained morphosyntactic type. Transforms are morphophonemic rewrite rules that convert base forms to derived forms, and whose context of application is limited to a specific set of base forms.

We have developed a greedy algorithm to induce such a grammar. At each iteration, suffixal transforms to convert between base and derived forms of lexemes are hypothesized. The algorithm chooses the transform that maximizes vocabulary coverage, while minimizing the number of conflicts resulting from proposing as base forms words previously found to be derived forms. After base forms and transforms have been learned, a distributional clustering step assigns the base forms to POS classes. In future work, the transforms will be converted to generalized rewrite rules by inducing phonological characteristics common to the base forms.

We have tested this algorithm on a version of the Penn Treebank annotated for inflectional morphology. The algorithm achieves 71.7% recall and 92.9% precision on inflectional relations, where both a base and derived form occur in the corpus. We are currently testing the algorithm on other languages, and will present results on the Morphochallenge gold standards.