Unsupervised Learning for Natural Language Processing

Dan Klein
University of California, Berkeley
klein@cs.berkeley.edu

Abstract

Given the abundance of text data, unsupervised approaches are very appealing for natural language processing. We present three latent variable systems which achieve state-of-the-art results in domains previously dominated by fully supervised systems. For syntactic parsing, we describe a grammar induction technique which begins with coarse syntactic structures and iteratively refines them in an unsupervised fashion. The resulting coarse-to-fine grammars admit efficient coarse-to-fine inference schemes and have produced the best parsing results in a variety of languages. For coreference resolution, we describe a discourse model in which entities are shared across documents using a hierarchical Dirichlet process. In each document, entities are repeatedly rendered into mention strings by a sequential model of attentional state and anaphoric constraint. Despite being fully unsupervised, this approach is competitive with the best supervised approaches. Finally, for machine translation, we present a model which learns translation lexicons from non-parallel corpora. Alignments between word types are modeled by a prior over matchings. Given any fixed alignment, a joint density over word vectors derives from probabilistic canonical correlation analysis. This approach is capable of discovering high-precision translations, even when the underlying corpora and languages are divergent.