Labeled Grammar Induction with Minimal Supervision
Yonatan Bisk, Christos Christodoulopoulos, and Julia Hockenmaier
University of Illinois at Urbana-Champaign

1. Induce and Label Clusters: Noun, Verb, Other

We use the Bayesian Mixture of Multinomials model (BMMM) of Christodoulopoulos et al. 2011 to induce word clusters. BMMM performs a type-based clustering based on token-level features and automatically inferred morphology [Morfessor (Creutz & Lagus 2006)]. Based on the Universal POS tags of the three most common words, clusters are labeled as N(noun), V(erb) or O(ther).

2. Induce a Grammar and Learn Labeled Dependencies

We train a parsing model (Bisk & Hockenmaier 2013:2015) on the induced parse forests. The parser returns CCG derivations and hence labeled dependencies.

3. Parsing Evaluation

Bisk & Hockenmaier 2015 produce labeled dependencies with an unsupervised CCG system based on gold POS tags. We show that performance degrades only slightly (less than 1/3 on average) with induced word clusters.

Analysis & Future Work

Every language poses its own challenges. In panel 2 we see that identifying verbs proves difficult in Chinese. Additionally, in panel 4 we find the largest gaps in languages with rich morphology. Better clustering or feedback from the syntax may help address these issues.

References:
Bisk & Hockenmaier 2013 An HDP Model for Inducing Combinatory Categorial Grammars. Trans. of the ACL 2013