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

Most approaches to grammar induction are based on the assumption that gold POS tags are available to the induction system. POS tags are arbitrary, relatively clean, which we replace with induced clusters.

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).

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

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