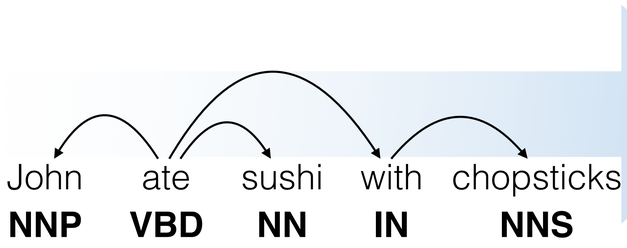
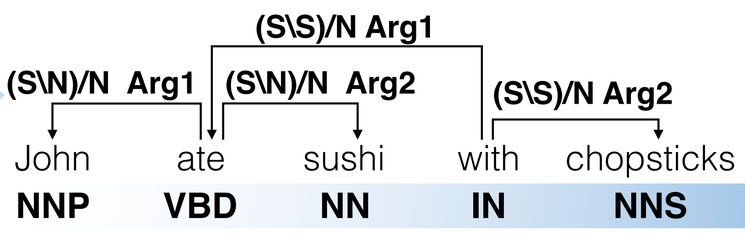
Labeled Grammar Induction with Minimal Supervision

Yonatan Bisk, Christos Christodoulopoulos, and Julia Hockenmaier
University of Illinois at Urbana-Champaign

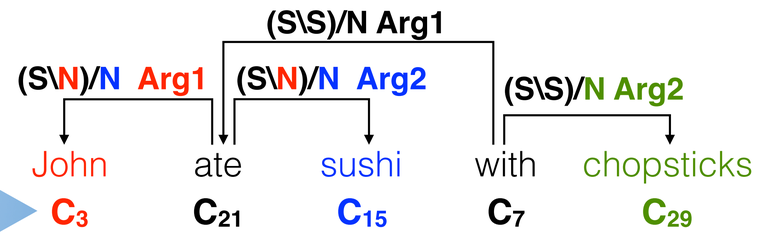
Dependency Grammar Induction:
Unlabeled Dependencies from POS tags



CCG Induction (Bisk & Hockenmaier 2013; 2015):
Labeled Dependencies from POS tags

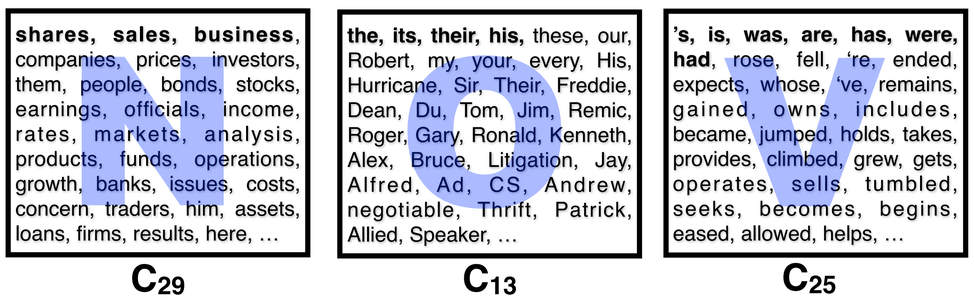


This paper:
Labeled dependencies without POS tags



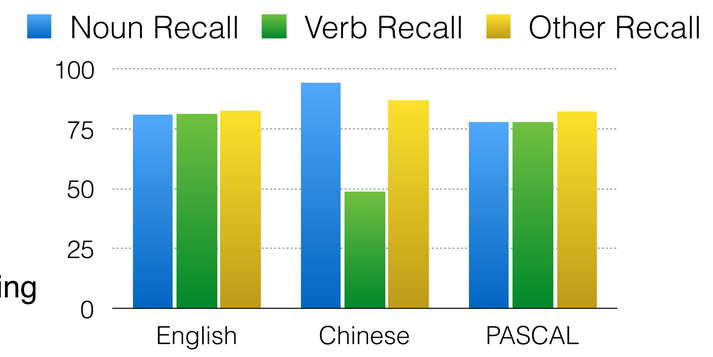
Most approaches to **grammar induction** are based on the assumption that **gold POS tags are available** to the induction system.
POS tags are arbitrary, relatively clean, clusters, which we **replace with induced clusters**.

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



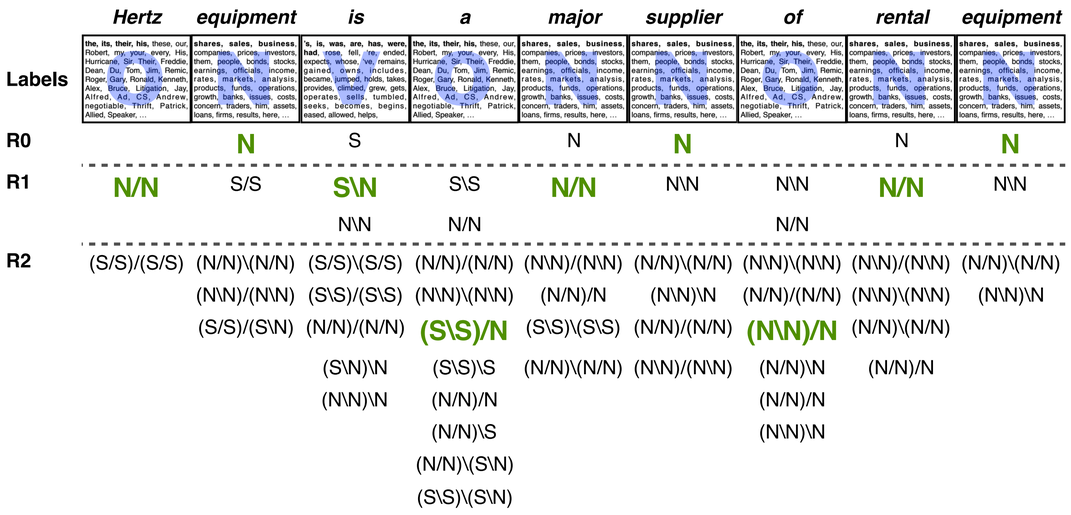
Data:
- **CCGbanks:** English and Chinese
- **Dependency Corpora**: 10 PASCAL Challenge Languages

Metric:
Recall from majority vote cluster labeling from 3 annotated words per cluster.

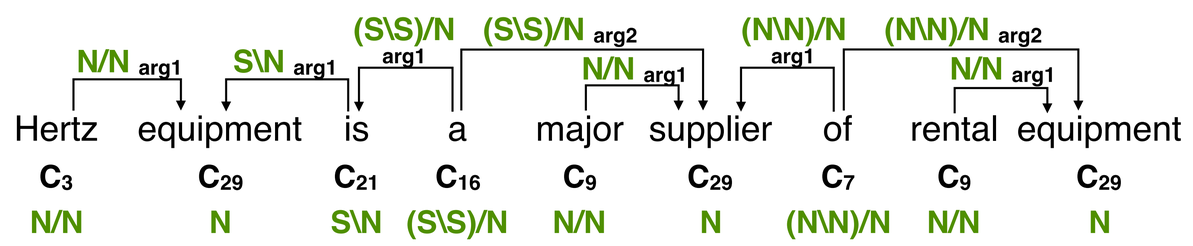


We use the Bayesian Mixture of Multinomials model (BMMM) of Christodoulopoulos et al. 2011 to induce word clusters. BMMM performs a **type-based clustering** based on **token-level features** and automatically inferred **morphology** [Morfessor (Creutz & Lagus 2006)]. Based on the Universal POS tags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

2. Induce a Grammar and Learn Labeled Dependencies



CCG Induction: Nouns can have the CCG **category N**, verbs can have the CCG **category S**, and may take adjacent nouns as arguments (S\N, S/N, (S\N)/N, etc.). All words can modify (XIX) adjacent N and S.

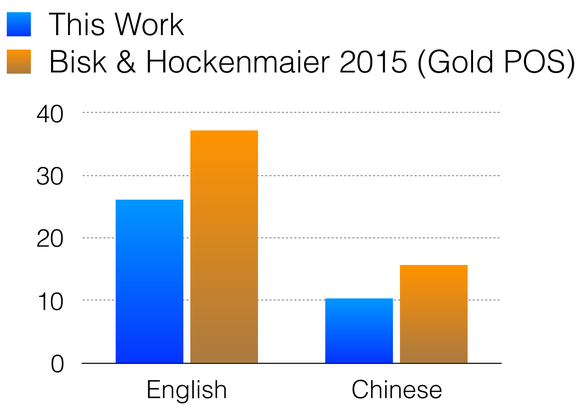


We train a parsing model (Bisk & Hockenmaier 2013;2015) on the induced parse forests. The parser returns CCG derivations and hence **labeled dependencies**.

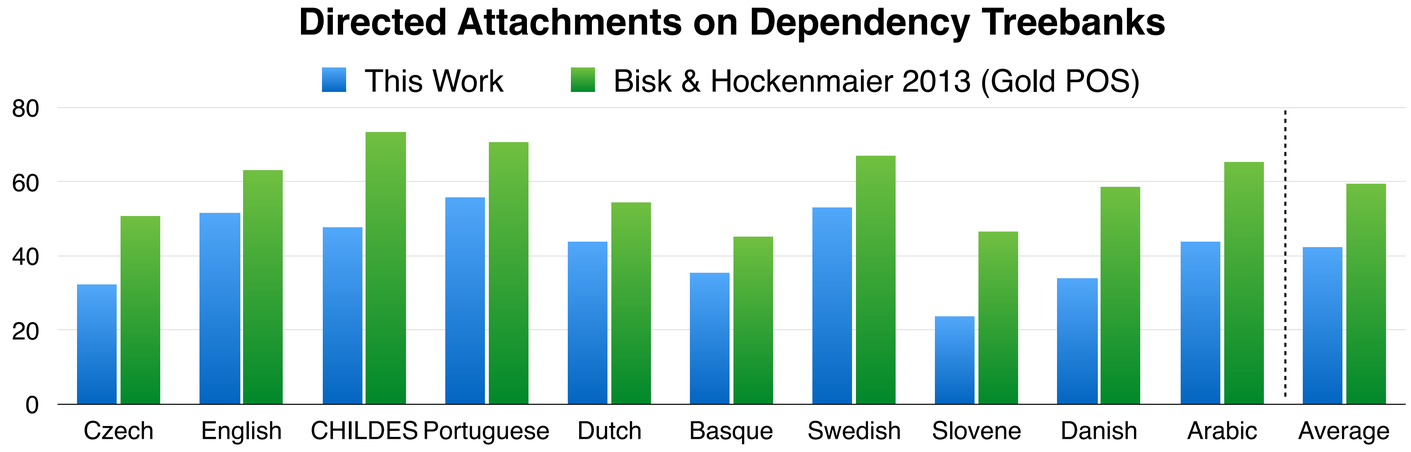
3. Parsing Evaluation

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**.

Labeled F1 on CCGbank



Directed Attachments on Dependency Treebanks



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:

Creutz & Lagus. Morfessor in the Morpho challenge. *Proc of PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes.* 2006
Christodoulopoulos et al. A Bayesian mixture model of PoS induction using multiple features. *Proc of EMNLP* 2011
Bisk & Hockenmaier 2012 Simple Robust Grammar Induction with Combinatory Categorial Grammars. *Proc of AAAI 2012*
Bisk & Hockenmaier 2013 An HDP Model for Inducting Combinatory Categorial Grammars. *Trans. of the ACL. 2013*
Bisk & Hockenmaier 2015 Probing the Linguistic Strengths and Limitations of Unsupervised Grammar Induction. *Proc of ACL* 2015