Unsupervised Constituency Grammar Induction: Learning Bracketing and Phrasal Categories

Peter Lubell-Doughtie

University of Amsterdam

lubell@science.uva.nl

May 13th, 2011

Motivation

Why unsupervised?

Comprehend intelligence and cognition via understanding language:

> *"Really knowing semantics is a prerequisite for anything to be called intelligence" – Partee*

But why unsupervised labeling?

- The vast majority of text is neither bracketed nor labeled.
- Knowing the labels of constituents and words is a step towards knowing their semantics.
- Many applications of knowing the relationships between constituents (e.g. Information Retrieval, Machine Translation)

Reichart and Rappoport: CCL+BMM

Bracketing

Use Seginer's CCL algorithm to generate a bracketing from raw sentences

Initial Labeling

Use BMM to label each constituent

Reduce number of labels by clustering

- features are label parent to/from child and sibling relationships
- features are POS tag left-most frequency
- use cosine similarity as the distance metric between feature vectors
- assign all other labels to the top D most frequent

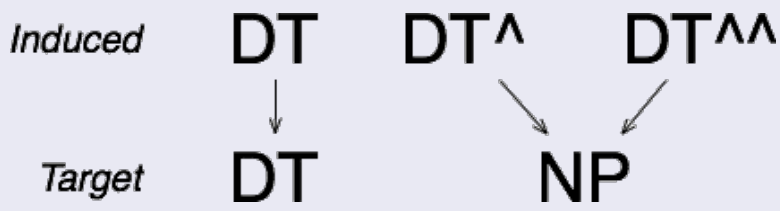
Reichart and Rappoport: Evaluation

Definition

Given an induced and target label pair, (X_i, Y_j), let C_{X_i,Y_j} be the number of times (X_i, Y_j) label a constituent having the same span in the same sentence and 0 if they share no constituents.

Greedy Mapping

$$\text{Map}(X_i) = \text{argmax}_{Y_j} C_{X_i,Y_j}$$



Label-to-Label Mapping

Form a complete bipartite graph between X and Y where edge (X_i, Y_j) has weight $w_{ij} = C_{X_i,Y_j}$. Find the optimal assignment from X to Y using the Kuhn-Munkres algorithm.

Ambiguities and Problems

How many labels?

When clustering the BMM induced labels to the top D labels, what is $|D|$?
The number of POS-tags in the corpus?

What does clustering optimize?

BMM is formally justified by MDL.
Clustering is an engineering method to fit the data.

Reducing the Number of Categories

Modify BMM so $|D|$ labels are produced

Naively continue to merge produces poor results (Reichart and Rappoport)
Can we change the MDL to penalize a label size other than $|D|$?

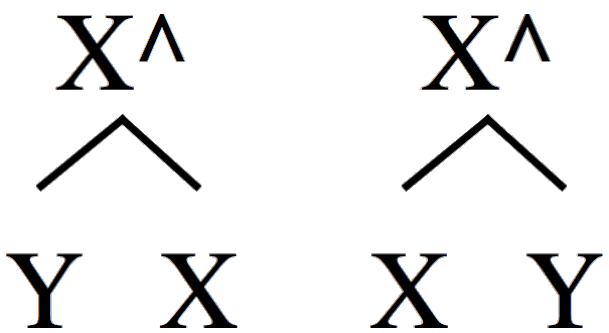
Common Cover Links for Constituent Labeling

Given POS-tags assigned to our lexical items use common cover links as the head-dependency relationship.
GIven the head-dependency relationship use X-bar theory to label constituents.

(oversimplified) X-bar Theory

Example

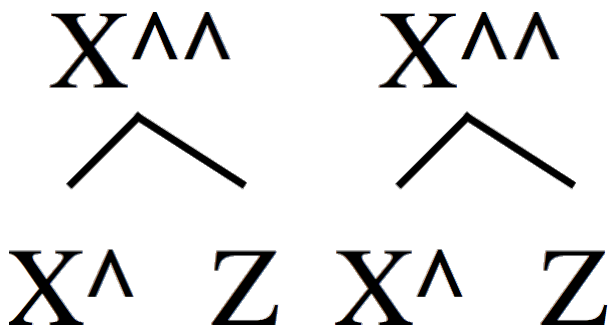
Given X is the head and Y is a complement
We elevate the head to a phrase label



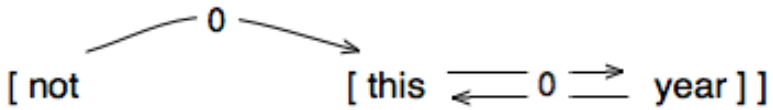
(oversimplified) X-bar Theory

Example

Given $\overline{X}$ is an X-bar type and Z is a specifier
We elevate the X-bar type to a higher phrase label



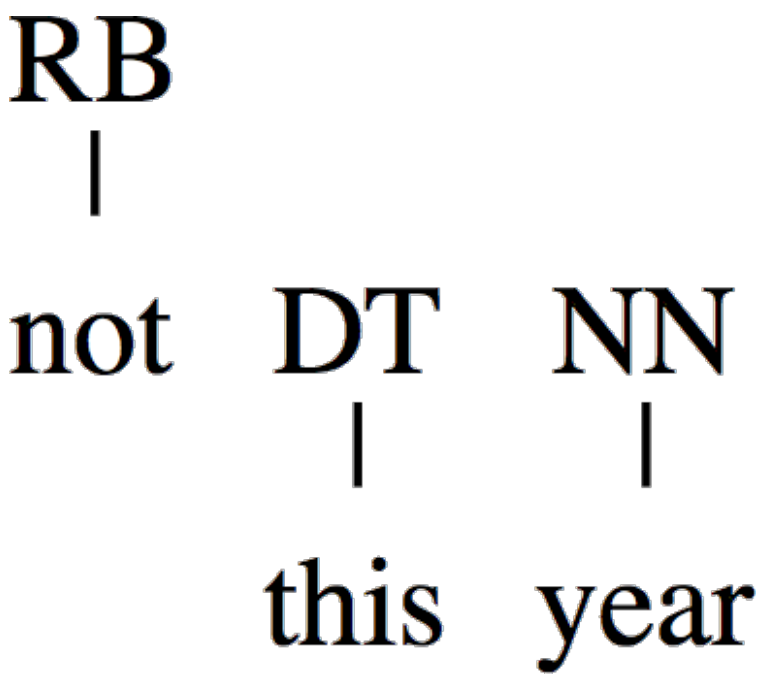
CCL to X-bar Labels



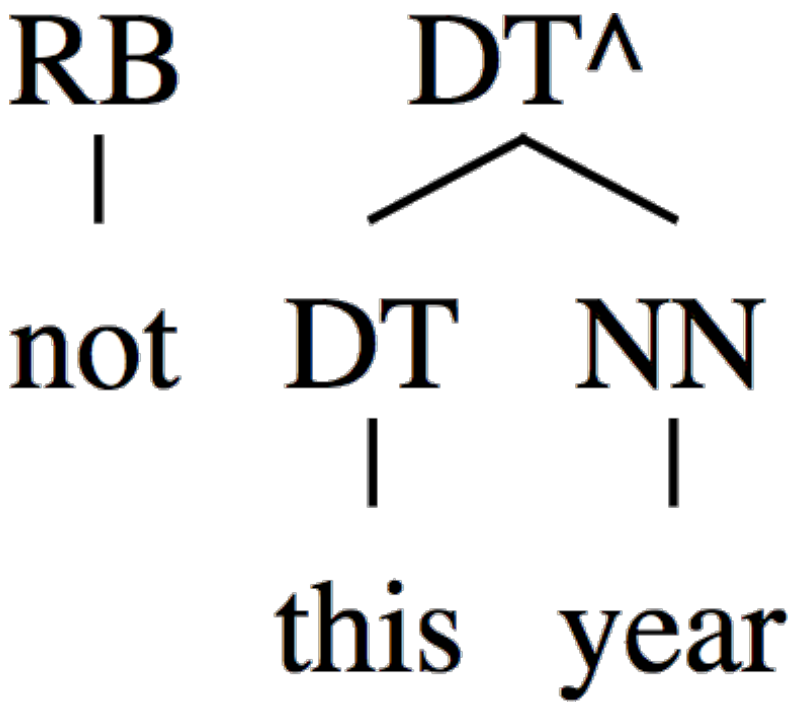
Common Cover Links for Constituent Labeling

- Given a bracketed CCL structure, take POS-tags for each word
- The POS-tag of an argument labels its head's constituent

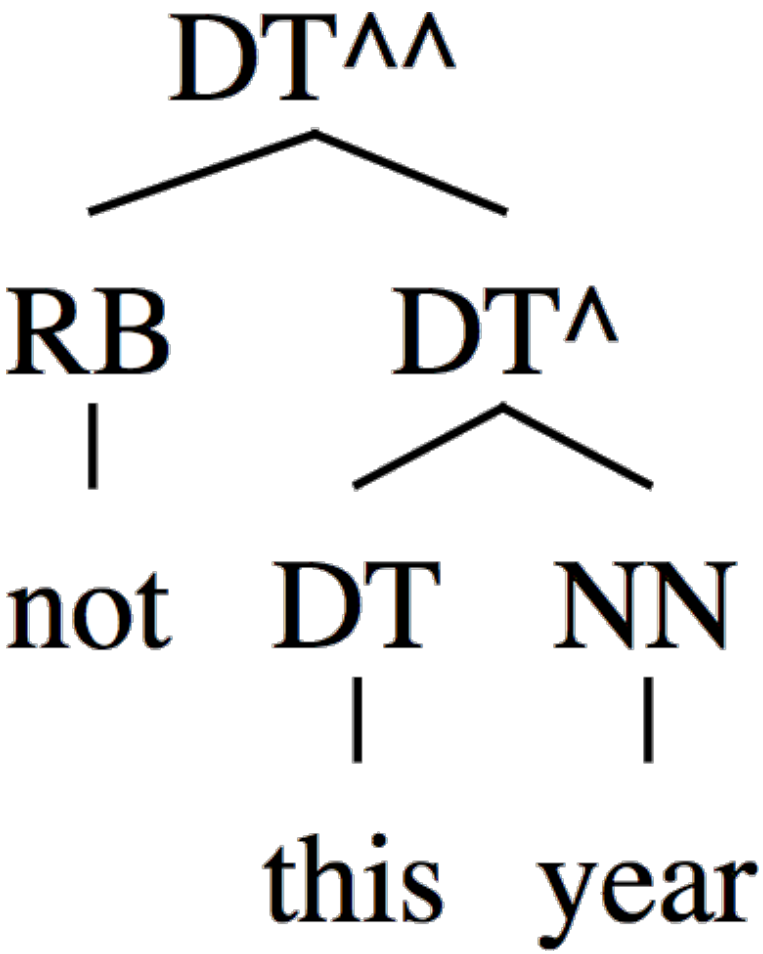
Labeling a Sentence



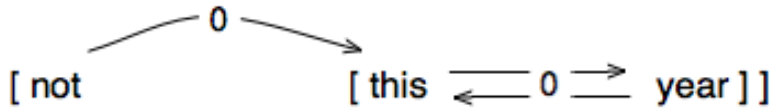
Labeling a Sentence



Labeling a Sentence



Ambiguities



DT - NN is exocentric

- We choose the left most as the head
- Worse results when choosing the right most
- A linguistically motivated heuristic?

Aren't we engineering to match the data?

Results

Pure Labeling Results

Induce POS-tags by taking the most frequent POS-tag for each word according to the gold standard WSJ10

Method	Experiment	Greedy F-Score
Reichart & Rappoport	Syntactic Clustering	**80**
	Random Clustering	67
	Random Baseline	30
Gold POS-tags	Exocentric LHS	**80**
	Exocentric RHS	76

Future Work

- Evaluate the whole bracketing, evaluate on other corpora
- Can we induce POS-tags from the CCL data?
- Can we use the BMM POS-tags?
- Is there a better way to select the head for exocentric links?

Questions?