CS 294-5: Statistical Natural Language Processing

Grammar Induction
Dan Klein

Assignment 3 Honors

Idea: Lexical Affinity Models
- Words select other words on syntactic grounds
- Congress narrowly passed the amended bill
- Idea: Link up pairs with high mutual information
  - [Yuret, 1998]: Greedy linkage
  - [Paskin, 2001]: Iterative re-estimation with EM
- Evaluation: compare linked pairs to a gold standard

Lexical Affinity Models
- Generative Model for [Paskin, 2001]
- Congress narrowly passed the amended bill
- Empirical Uniform passed .... bill

Problem: Non-Syntactic Affinity
- Mutual information between words does not necessarily indicate syntactic selection.
- Congress narrowly passed the amended bill
- Expect brushbacks but no beanballs
- A new year begins in New York

Idea: Word Classes
- Individual words like congress are entwined with semantic facts about the world.
- Syntactic classes, like NOUN and ADVERB are bleached of word-specific semantics.
- Automatic word classes more likely to look like DAYS-OF-WEEK or PERSON-NAME.
- We could build dependency models over word classes. [cf. Carroll and Charniak, 1992]
- Congress narrowly passed the amended bill

congress narrowly passed the amended bill

NOUN ADVERB VERB DET PARTICIPLE NOUN
A Word-Class Model

$$P(\text{words, classes, graph}) = P(\text{length}) \cdot P(\text{graph})$$

Problem: Word Class Models

- Issues:
  - Too simple a model – doesn’t work much better supervised
  - No representation of valence (number of arguments)

A Head-Outward Model (DMV)

- Supervised statistical parsers benefit from modeling tree distributions implicitly. [e.g., Collins, 99]
- A head-outward model with word classes and valence/adjacency:

$$P(t_b) = \prod_{\text{dist} < |t_r|}$$

Issue: Local Representations

- Distance

Results: Dependencies

- Model is re-estimated with EM
  - Cubic dynamic program run over each sentence
  - Expected counts of each modeled configuration are aggregated
- Initialization:
  - Initial parameters from simple heuristics:

$$P(a | h, \text{dir}) \propto \sum_{\text{dist}} \frac{1}{\text{dist}} \text{count}(h, a, \text{dist}, \text{dir})$$

Common Errors: Dependency

<table>
<thead>
<tr>
<th>Overproposed Dependencies</th>
<th>Underproposed Dependencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>DET $\leftarrow$ N</td>
<td>DET $\rightarrow$ N</td>
</tr>
<tr>
<td>3474</td>
<td>3079</td>
</tr>
<tr>
<td>N-PROP $\leftarrow$ N-PROP</td>
<td>N-PROP $\rightarrow$ N-PROP</td>
</tr>
<tr>
<td>2096</td>
<td>1898</td>
</tr>
<tr>
<td>NUM $\rightarrow$ NUM</td>
<td>PREP $\leftarrow$ N</td>
</tr>
<tr>
<td>760</td>
<td>838</td>
</tr>
<tr>
<td>PREP $\leftarrow$ DET</td>
<td>N $\rightarrow$ V-PRES</td>
</tr>
<tr>
<td>735</td>
<td>714</td>
</tr>
<tr>
<td>DET $\leftarrow$ N-PL</td>
<td>DET $\rightarrow$ N-PL</td>
</tr>
<tr>
<td>636</td>
<td>672</td>
</tr>
<tr>
<td>DET $\rightarrow$ PREP</td>
<td>N $\leftarrow$ PREP</td>
</tr>
<tr>
<td>627</td>
<td>669</td>
</tr>
<tr>
<td>DET $\rightarrow$ V-PAST</td>
<td>NUM $\leftarrow$ NUM</td>
</tr>
<tr>
<td>470</td>
<td>54</td>
</tr>
<tr>
<td>DET $\rightarrow$ V-PRES</td>
<td>N $\rightarrow$ V-PAST</td>
</tr>
<tr>
<td>420</td>
<td>54</td>
</tr>
</tbody>
</table>
Early Approaches: Structure Search

- Incremental grammar learning, chunking [Wolff 88, Langley 82, many others]
- Can recover synthetic grammars
- An (extremely good) result of incremental structure search:

Issues with Chunk/Merge Systems

- Hard to recover from initial choices (c.f. EM, where the issue is initial state)
- Hard to make local decisions which will interact well with each other (e.g. group verb-preposition and preposition-determiner, both wrong, and not consistent)
- Good local heuristics often don’t have a well-formed global objective that can be evaluated for the target grammar.

Idea: Learn PCFGs with EM

- Classic experiments on learning PCFGs with Expectation-Maximization [Lari and Young, 1990]

Re-estimation of PCFGs

- Basic quantity needed for re-estimation with EM:

\[
P(X_i, j, S) = \frac{\sum_{P(T)} \sum_{P(T)_{i,j=0}} P(T)}{P(T)}
\]
- Can calculate in cubic time with the Inside-Outside algorithm.
- Consider an initial grammar where all productions have equal weight:

\[
P(X_i, X_j, X_k) = \frac{1}{n^2}
\]
- Then all trees have equal probability initially.
- Therefore, after one round of EM, the posterior over trees will (in the absence of random perturbation) be approximately uniform over all trees, and symmetric over symbols.

Problem: “Uniform” Posteriors

- Tree Uniform
- Split Uniform

Problem: Model Symmetries

- Symmetries
- How does this relate to trees?
Other Approaches

- Evaluation: fraction of nodes in gold trees correctly posited in proposed trees (unlabeled recall)
- Some recent work in learning constituency:
  - [Adrians, 99] Language grammars aren’t general PCFGs
  - [Clark, 01] Mutual-information filters detect constituents, then an MDL-guided search assembles them
  - [van Zaanen, 00] Finds low edit-distance sentence pairs and extracts their differences

Right-Branching Baseline

- English trees tend to be right-branching, not balanced
- A simple (English-specific) baseline is to choose the right chain structure for each sentence

Desiderata: Practical Learnability

- To be practically learnable, models should:
  - Be as simple as possible
  - Make symmetries self-breaking whenever possible
  - Avoid hidden structures which are not directly coupled to surface phenomena

Inspiration: Distributional Clustering

- The president said that the downturn was over
- Inspiration: Distributional Clustering

Distributional Models

\[ P(S, C) = \prod_i P(c_i)P(w_i | c_i)P(w_{i-1}, w_{i+1} | c_i) \]

- Can we use distributional clustering for learning syntax? [Harris, 51]
Problem: Identifying Constituents

Distributional classes are easy to find...

... but figuring out which are constituents is hard.

A Nested Distributional Model

- We'd like a model that:
  - Ties spans to linear contexts (like distributional clustering)
  - Considers only proper tree structures (like a PCFG model)
  - Has no symmetries to break (like a dependency model)

Constituent-Context Model (CCM)

\[
P(S|T) = \prod_{(i,j) \in T} \left( \frac{P(f_{pf})}{P(f_{pf})} \right) \prod_{(i,j) \in T} \left( \frac{P(f_{is})}{P(f_{is})} \right)
\]

Results: Constituency

Results: Combined Models

<table>
<thead>
<tr>
<th>Dependency Evaluation</th>
<th>Random</th>
<th>DMV</th>
<th>CCM + DMV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised PCFG</td>
<td>45.6</td>
<td>62.7</td>
<td>64.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Constituency Evaluation</th>
<th>Random</th>
<th>CCM</th>
<th>CCM + DMV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised PCFG</td>
<td>39.4</td>
<td>81.0</td>
<td>88.0</td>
</tr>
</tbody>
</table>

- Subject-verb groups gone, modifier placement improved

Spectrum of Systematic Errors

But the worst errors are the non-systematic ones! (~25%)
How General is This?

<table>
<thead>
<tr>
<th>Language</th>
<th>Constituency Evaluation</th>
<th>Dependency Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>English (7422 sentences)</td>
<td>39.4</td>
<td>60.0</td>
</tr>
<tr>
<td>Random Baseline</td>
<td>88.0</td>
<td></td>
</tr>
<tr>
<td>CCM+DMV</td>
<td>88.0</td>
<td></td>
</tr>
<tr>
<td>German (2175 sentences)</td>
<td>49.6</td>
<td>60.0</td>
</tr>
<tr>
<td>Random Baseline</td>
<td>89.7</td>
<td></td>
</tr>
<tr>
<td>CCM+DMV</td>
<td>89.7</td>
<td></td>
</tr>
<tr>
<td>Chinese (2473 sentences)</td>
<td>35.5</td>
<td>60.0</td>
</tr>
<tr>
<td>Random Baseline</td>
<td>54.2</td>
<td></td>
</tr>
<tr>
<td>CCM+DMV</td>
<td>54.2</td>
<td></td>
</tr>
</tbody>
</table>

Most Common Errors: English

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Overproposed Constituents</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADI N</td>
<td>1022 the [general partner]</td>
</tr>
<tr>
<td>N-PROP N-PROP</td>
<td>447 the [Big Board]</td>
</tr>
<tr>
<td>DET N</td>
<td>398 [an import] order</td>
</tr>
<tr>
<td>ADI N-PL</td>
<td>294 six million [common shares]</td>
</tr>
<tr>
<td>N-PL ADV</td>
<td>164 [seats currently] are quoted</td>
</tr>
</tbody>
</table>

Most Common Errors: German

<table>
<thead>
<tr>
<th>Error Type</th>
<th>Overproposed Constituents</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADI N</td>
<td>71 der [Dalberger Hof]</td>
</tr>
<tr>
<td>DET N</td>
<td>94 [die Moderator] der Zukunft</td>
</tr>
<tr>
<td>DET ADI N</td>
<td>30 [der erhofften Meisterschaft]</td>
</tr>
<tr>
<td>CONJ N</td>
<td>71 [Sinti und Roma]</td>
</tr>
</tbody>
</table>

What’s Been Accomplished?

- **Unsupervised learning:**
  - Constituency structure
  - Dependency structure

- **Constituency recall:**
  - Random Baseline: 39.4
  - CCM + DMV: 88.0
  - Supervised PCFG: 92.8

- **Why it works:**
  - Combination of simple models
  - Representations designed for unsupervised learning