Non-Local Modeling with a Mixture of PCFGs

Slav Petrov, Leon Barrett and Dan Klein
University of California at Berkeley

CoNLL 2006
Empirical Motivation

Slav Petrov, Leon Barrett and Dan Klein

Non-Local Modeling with a Mixture of PCFGs
Verb Phrase Expansion: capture with lexicalization. [Collins 1999, Charniak 2000]
Empirical Motivation

increased 11% to # 2.5 billion
from # 2.25 billion
Empirical Motivation

Local Correlation: capture with parent annotation.
[Johnson 1998, Klein & Manning 2003]
Empirical Motivation

Non-Local Correlation.
[This work]
<table>
<thead>
<tr>
<th>Rule</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>QP $\rightarrow$ # CD CD</td>
<td>131.6</td>
</tr>
<tr>
<td>PRN $\rightarrow$ -(LRB)- ADJP -RRB</td>
<td>77.1</td>
</tr>
<tr>
<td>VP $\rightarrow$ VBD NP , PP PP</td>
<td>33.7</td>
</tr>
<tr>
<td>VP $\rightarrow$ VBD NP NP PP</td>
<td>28.4</td>
</tr>
<tr>
<td>PRN $\rightarrow$ -(LRB)- NP -RRB-</td>
<td>17.3</td>
</tr>
<tr>
<td>ADJP $\rightarrow$ QP</td>
<td>13.3</td>
</tr>
<tr>
<td>PP $\rightarrow$ IN NP ADVP</td>
<td>12.3</td>
</tr>
<tr>
<td>NP $\rightarrow$ NP PRN</td>
<td>12.3</td>
</tr>
<tr>
<td>VP $\rightarrow$ VBN PP PP PP</td>
<td>11.6</td>
</tr>
<tr>
<td>ADVP $\rightarrow$ NP RBR</td>
<td>10.1</td>
</tr>
</tbody>
</table>
## Correlations for QP $\rightarrow$ # CD CD

<table>
<thead>
<tr>
<th>Rule</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>QP $\rightarrow$ # CD CD</td>
<td>131.6</td>
</tr>
<tr>
<td>PRN $\rightarrow$ -LRB- ADJP -RRB</td>
<td>77.1</td>
</tr>
<tr>
<td>VP $\rightarrow$ VBD NP , PP PP</td>
<td>33.7</td>
</tr>
<tr>
<td>VP $\rightarrow$ VBD NP NP PP</td>
<td>28.4</td>
</tr>
<tr>
<td>PRN $\rightarrow$ -LRB- NP -RRB-</td>
<td>17.3</td>
</tr>
<tr>
<td>ADJP $\rightarrow$ QP</td>
<td>13.3</td>
</tr>
<tr>
<td>PP $\rightarrow$ IN NP ADVP</td>
<td>12.3</td>
</tr>
<tr>
<td>NP $\rightarrow$ NP PRN</td>
<td>12.3</td>
</tr>
<tr>
<td>VP $\rightarrow$ VBN PP PP PP</td>
<td>11.6</td>
</tr>
<tr>
<td>ADVP $\rightarrow$ NP RBR</td>
<td>10.1</td>
</tr>
</tbody>
</table>
## Correlations for QP → # CD CD

<table>
<thead>
<tr>
<th>Rule</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>QP → # CD CD</td>
<td>131.6</td>
</tr>
<tr>
<td>PRN → -(LRB- ADJP ) -(RRB-)</td>
<td>77.1</td>
</tr>
<tr>
<td>VP → VBD NP , PP PP</td>
<td>33.7</td>
</tr>
<tr>
<td>VP → VBD NP NP PP</td>
<td>28.4</td>
</tr>
<tr>
<td>PRN → -(LRB- NP ) -(RRB-  )</td>
<td>17.3</td>
</tr>
<tr>
<td>ADJP → QP</td>
<td>13.3</td>
</tr>
<tr>
<td>PP → IN NP ADVP</td>
<td>12.3</td>
</tr>
<tr>
<td>NP → NP PRN</td>
<td>12.3</td>
</tr>
<tr>
<td>VP → VBN PP PP PP</td>
<td>11.6</td>
</tr>
<tr>
<td>ADVP → NP RBR</td>
<td>10.1</td>
</tr>
</tbody>
</table>
Examples

Repeated formulaic structure in one grammar:
VP $\rightarrow$ VBD NP PP $\,\,\,$, S and VP $\rightarrow$ VBG NP PP PP.

$\text{hit a record in 1998, rising } 1.7\% \text{ after inflation adjustment to }$\text{ }$\$13,120$
Examples

Repeated formulaic structure in one grammar:

\[ VP \rightarrow \text{VBD NP PP , S and VP} \rightarrow \text{VBG NP PP PP} \].
Examples

Sibling effects, though not parallel structure:

\[ \text{NX} \rightarrow \text{NNS} \]

and \[ \text{NX} \rightarrow \text{NN NNS} \].
Examples

```
S
   /\         .
  /   \      |
NP   VP     |
  /\     /  |
 DT  NX   |
   /\   /|
  No NX CC |
     /\  |
    NNS or NN NNS
           |
           |
        lawyers tape recorders
```

Sibling effects, though not parallel structure:
NX → NNS
and NX → NN NNS.
A special structure for footnotes:
ROOT $\rightarrow$ X
and X $\rightarrow$ SYM.
Motivation

Model non-local correlation that can stem from:

- Dialects,
- Priming effects,
- Genre,
- Stylistic conventions.
Single Grammar

ROOT
Single Grammar

ROOT

S
Single Grammar

Non-Local Modeling with a Mixture of PCFGs
Single Grammar

ROOT
  
  S

  NP  VP

  ...

  ...

Slav Petrov, Leon Barrett and Dan Klein

Non-Local Modeling with a Mixture of PCFGs
Mixture of PCFGs

ROOT-1  ROOT-2  ROOT-3  ROOT-4
Mixture of PCFGs

ROOT-1  ROOT-2  ROOT-3  ROOT-4
Mixture of PCFGs

ROOT-1  ROOT-2  ROOT-3  ROOT-4
Mixture of PCFGs

ROOT-1    ROOT-2    ROOT-3    ROOT-4
Mixture of PCFGs

ROOT-1  ROOT-2  ROOT-3  ROOT-4
Mixture of PCFGs

ROOT-1
   \[\hspace{5mm} S-1 \hspace{5mm}\]
   \[\hspace{10mm} NP-1 \hspace{15mm} VP-1 \hspace{10mm}\]
   \[\hspace{5mm} \ldots \hspace{5mm}\]

ROOT-2
   \[\hspace{5mm} S-2 \hspace{5mm}\]
   \[\hspace{10mm} NP-2 \hspace{15mm} VP-2 \hspace{10mm}\]
   \[\hspace{5mm} \ldots \hspace{5mm}\]

ROOT-3
   \[\hspace{5mm} S-3 \hspace{5mm}\]
   \[\hspace{10mm} NP-3 \hspace{15mm} VP-3 \hspace{10mm}\]
   \[\hspace{5mm} \ldots \hspace{5mm}\]

ROOT-4
   \[\hspace{5mm} S-4 \hspace{5mm}\]
   \[\hspace{10mm} NP-4 \hspace{15mm} VP-4 \hspace{10mm}\]
   \[\hspace{5mm} \ldots \hspace{5mm}\]
Mixture of PCFGs

- Single grammar:

\[ P(T) = \prod_{X \rightarrow \alpha \in T} P(\alpha | X). \]
Mixture of PCFGs

- Single grammar:

\[ P(T) = \prod_{X \rightarrow \alpha \in T} P(\alpha | X). \]

- Single grammar from a mixture:

\[ P(T, i) = P(i) \prod_{X \rightarrow \alpha \in T} P(\alpha | X, i). \]
Mixture of PCFGs

- Single grammar:
  \[ P(T) = \prod_{X \rightarrow \alpha \in T} P(\alpha|X). \]

- Single grammar from a mixture:
  \[ P(T, i) = P(i) \prod_{X \rightarrow \alpha \in T} P(\alpha|X, i). \]

- Mixture of grammars:
  \[ P(T) = \sum_{i} P(T, i) = \sum_{i} P(i) \prod_{X \rightarrow \alpha \in T} P(\alpha|X, i). \]
Would like the *most probable parse*:

\[ P(T|S) \propto \sum_i P(i)P(T|i). \]
Inference: Parsing

- Would like the **most probable parse**:
  \[ P(T|S) \propto \sum_i P(i)P(T|i). \]

- Mixture of grammars:
  \[
  \arg\max_T \sum_i P(T, i) = \arg\max_T \sum_i P(i) \prod_{X \rightarrow \alpha \in T} P(\alpha|X, i).
  \]
Inference: Parsing

Would like the *most probable parse*:

\[ P(T|S) \propto \sum_i P(i)P(T|i). \]

Mixture of grammars:

\[
\text{argmax}_T \sum_i P(T, i) = \text{argmax}_T \sum_i P(i) \prod_{X \rightarrow \alpha \in T} P(\alpha|X, i).
\]

Computing most probable parse is NP-hard.
Would like the *most probable parse*:

$$P(T|S) \propto \sum_i P(i)P(T|i).$$

Mixture of grammars:

$$\text{argmax } \sum_T P(T,i) = \text{argmax } \sum_i P(i) \prod_{X \rightarrow \alpha \in T} P(\alpha|X,i).$$

Computing most probable parse is NP-hard.

Compute the *most probable derivation* instead.
Manually assign sentences to grammars, e.g. Brown corpus.

Alternatively, use a standard Expectation-Maximization (EM) approach.
Learning: Training

- Manually assign sentences to grammars, e.g. Brown corpus.
- Alternatively, use a standard Expectation-Maximization (EM) approach.

E-Step:
- Fix model parameters and compute the posterior distributions of the latent variables.
- Component $G$ of each sentence:

$$P(i|T) = \frac{P(T, i)}{\sum_j P(T, j)}.$$
Learning: Training

M-Step:

- Given the posterior assignments find the maximum likelihood model parameters.
- Let \( T = \{ T_1, T_2, \ldots \} \) be the training set. The M-Step updates are:
- Component prior:

\[
P(i) \leftarrow \frac{\sum_{T_k \in T} P(i|T_k)}{\sum_i \sum_{T_k \in T} P(i|T_k)} = \frac{\sum_{T_k \in T} P(i|T_k)}{k}.
\]

- Estimate rule probabilities as for a single grammar but with fractional counts.
Hierarchical Estimation

Pool common rules (e.g. NP → DT NN) in a *shared grammar* $G_s$. 
Pool common rules (e.g. NP $\rightarrow$ DT NN) in a shared grammar $G_s$.

Latent variable $L = \{s, l\}$ at each rewrite:
Hierarchical Estimation

- Pool common rules (e.g. NP $\rightarrow$ DT NN) in a *shared grammar* $G_s$.
- Latent variable $L = \{S, I\}$ at each rewrite:

$$P(\alpha | X, i) = \lambda P(\alpha | X, i, \ell = I) + (1 - \lambda) P(\alpha | X, i, \ell = S),$$
Hierarchical Estimation

- Pool common rules (e.g. NP $\rightarrow$ DT NN) in a *shared grammar* $G_s$.
- Latent variable $L = \{s, i\}$ at each rewrite:

$$ P(\alpha|X, i) = \lambda P(\alpha|X, i, \ell = i) + (1 - \lambda) P(\alpha|X, i, \ell = s), $$

- Two kinds of hidden variables: the grammar $G$ (for each sentence) and the level $L$ (for each node).
Component $G$ of each sentence as before:

$$P(i|T) = \frac{P(T, i)}{\sum_j P(T, j)}.$$ 

Hierarchy level $L$ of each rewrite:

$$P(\ell = 1|X \rightarrow \alpha, i, T) = \frac{\lambda P(\alpha|X, \ell = 1)}{\lambda P(\alpha|X, i, \ell = 1) + (1 - \lambda)P(\alpha|X, \ell = S)}.$$
M-Step

Component prior as before:

\[
P(i) \leftarrow \frac{\sum_{T_k \in T} P(i|T_k)}{\sum_i \sum_{T_k \in T} P(i|T_k)} = \frac{\sum_{T_k \in T} P(i|T_k)}{k}.
\]
M-Step

- **Component prior as before:**

\[ P(i) \leftarrow \frac{\sum_{T_k \in T} P(i | T_k)}{\sum_i \sum_{T_k \in T} P(i | T_k)} = \frac{\sum_{T_k \in T} P(i | T_k)}{k}. \]

- **Hierarchy Level:**

\[ P(l = 1) \leftarrow \frac{\sum_{T_k \in T} \sum_{X \rightarrow \alpha \in T_k} P(l = 1 | X \rightarrow \alpha)}{\sum_{T_k \in T} |T_k|}. \]
Experimental Setup

- WSJ with standard setup:
  - Section 2-21 training set,
  - Section 22 validation set,
  - Section 23 test set.
- Baseline: Markovized grammar annotated with parent and sibling information (vertical order=2, horizontal order=1 [Klein & Manning 2003]).
Slav Petrov, Leon Barrett and Dan Klein

Non-Local Modeling with a Mixture of PCFGs
Mixture model captures non-local correlations.

10% reduction in total correlation error:
- Estimate rule correlations from corpus.
- Generate trees with grammar and estimate rule correlations.
- Compute correlation difference.
Brown corpus’ genres are statistically coherent.

Assign each genre to an individual grammar (no EM training):

\[ F_1 = 79.48, \text{ LL} = -242332. \]

Initialize by genre then train with EM:

\[ F_1 = 79.37, \text{ LL} = -242100. \]

EM with a random initialization:

\[ F_1 = 79.16, \text{ LL} = -242459. \]

Model can capture variation between genres, but maximum training data likelihood does not necessarily give maximum accuracy.
"Learning Accurate, Compact, and Interpretable Tree Annotation", Petrov et al., ACL 2006:

- $F_1 = 90.2\%$
- More flexible learning framework.
- Split and merge training to keep grammar compact.
- Similar in spirit to Klein & Manning 2003 and Matsuzaki et al. 2005.
Conclusions

- Examined rule correlations that may be found in natural language corpora, discovering non-local correlations not captured by traditional models.

- A Mixture of PCFGs can represent these non-local features and gives an improvement in parsing accuracy and data likelihood.

- This improvement is modest, however, primarily because local correlations are so much stronger than non-local ones.
Thank you very much for your attention.

Questions?

{petrov, lbarrett, klein}@eecs.berkeley.edu