Sparsity in Dependency Grammar Induction

Jennifer Gillenwater\textsuperscript{1} \quad Kuzman Ganchev\textsuperscript{1} \quad João Graça\textsuperscript{2} \\
Ben Taskar\textsuperscript{1} \quad Fernando Pereira\textsuperscript{3}

\textsuperscript{1}Computer & Information Science \quad University of Pennsylvania

\textsuperscript{2}L\textsuperscript{2}F INESC-ID, Lisboa, Portugal

\textsuperscript{3}Google, Inc.

July 12, 2010
A generative dependency parsing model
A generative dependency parsing model
The ambiguity problem this model faces
- A generative dependency parsing model
- The *ambiguity* problem this model faces
- Previous attempts to reduce ambiguity
A generative dependency parsing model

The ambiguity problem this model faces

Previous attempts to reduce ambiguity

How posteriors provide a good measure of ambiguity
A generative dependency parsing model
The ambiguity problem this model faces
Previous attempts to reduce ambiguity
How posteriors provide a good measure of ambiguity
Applying posterior regularization to the likelihood objective
A generative dependency parsing model

The ambiguity problem this model faces

Previous attempts to reduce ambiguity

How posteriors provide a good measure of ambiguity

Applying posterior regularization to the likelihood objective

Success with respect to EM and parameter prior baselines
Dependency model with valence

(Klein and Manning, ACL 2004)

\[ p_\theta(x, y) = \theta_{\text{root}}(V) \]
Dependency model with valence

(Klein and Manning, ACL 2004)

\[ p_\theta(x, y) = \theta_{\text{root}}(V) \cdot \theta_{\text{stop}}(\text{nostop}|V,\text{right},\text{false}) \cdot \theta_{\text{child}}(N|V,\text{right}) \]
Dependency model with valence

(Klein and Manning, ACL 2004)

\[ p_\theta(x, y) = \theta_{\text{root}}(V) \]

\[ \cdot \theta_{\text{stop}}(\text{nostop}|V, \text{right}, \text{false}) \cdot \theta_{\text{child}}(N|V, \text{right}) \]

\[ \cdot \theta_{\text{stop}}(\text{stop}|V, \text{right}, \text{true}) \cdot \theta_{\text{stop}}(\text{nostop}|V, \text{left}, \text{false}) \cdot \theta_{\text{child}}(N|V, \text{left}) \]

\[ \ldots \]
Traditional objective optimization

- **Traditional objective**: marginal log likelihood

\[
\max_\theta \mathcal{L}(\theta) = E_X[\log p_\theta(x)] = E_X[\log \sum_y p_\theta(x, y)]
\]
Traditional objective optimization

- **Traditional objective**: marginal log likelihood

$$\max_{\theta} \mathcal{L}(\theta) = E_x[\log p_\theta(x)] = E_x[\log \sum_y p_\theta(x, y)]$$

- **Optimization method**: expectation maximization (EM)

Problem: EM may learn a very ambiguous grammar

Too many non-zero probabilities

Ex: $V \rightarrow N$ should have non-zero probability, but $V \rightarrow DET$, $V \rightarrow JJ$, $V \rightarrow PRP$, etc. should be 0
Traditional objective optimization

- **Traditional objective**: marginal log likelihood
  \[
  \max_{\theta} \mathcal{L}(\theta) = E_X[\log p_\theta(x)] = E_X[\log \sum_yp_\theta(x, y)]
  \]
  
- **Optimization method**: expectation maximization (EM)
  
- **Problem**: EM may learn a very ambiguous grammar
  - Too many non-zero probabilities
  - Ex: \texttt{V → N} should have non-zero probability, but \texttt{V → DET}, \texttt{V → JJ}, \texttt{V → PRP$}}, etc. should be 0
Previous approaches to improving performance

- Structural annealing\(^1\)

1. Smith and Eisner, ACL 2006
2. Headden et al., NAACL 2009
3. Liang et al., EMNLP 2007; Johnson et al., NIPS 2007; Cohen et al., NIPS 2008, NAACL 2009
Previous approaches to improving performance

- Structural annealing$^1$
- $\mathcal{L}(\theta')$: Model extension$^2$

---

1. Smith and Eisner, ACL 2006
2. Headden et al., NAACL 2009
3. Liang et al., EMNLP 2007; Johnson et al., NIPS 2007; Cohen et al., NIPS 2008, NAACL 2009
Previous approaches to improving performance

- **Structural annealing**
- $\mathcal{L}(\theta')$: Model extension
- $\mathcal{L}(\theta) + \log p(\theta)$: Parameter regularization
  - Tend to reduce unique # of children per parent, rather than directly reducing # of unique parent-child pairs
  - $\theta_{\text{child}}(Y|X, \text{dir}) \neq \text{posterior}(X \rightarrow Y)$

1 Smith and Eisner, ACL 2006
2 Headden et al., NAACL 2009
3 Liang et al., EMNLP 2007; Johnson et al., NIPS 2007; Cohen et al., NIPS 2008, NAACL 2009
Intuition: True # of unique parent tags for a child tag is small
Ambiguity measure using posteriors: $L_{1/\infty}$

Sparsity is working

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
</table>
Ambiguity measure using posteriors: $L_{1/\infty}$

Sparsity is working

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Z</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>V</th>
<th>V</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>N</td>
<td>&gt;</td>
</tr>
<tr>
<td>ADJ</td>
<td>N</td>
<td>&gt;</td>
</tr>
<tr>
<td>ADJ</td>
<td>N</td>
<td>&gt;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0 1 0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0 1 0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>
Ambiguity measure using posteriors: $L_{1/\infty}$

Sparsity is working

Use good grammars
Ambiguity measure using posteriors: $L_{1/\infty}$

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Ambiguity measure using posteriors: $L_{1/\infty}$

Sparsity is working

Use good grammars

$\text{sum} = 3$
Measuring ambiguity on distributions over trees

For a distribution $p_\theta(y \mid x)$ instead of gold trees:

\[
\begin{array}{ccccccccc}
\text{N} & \text{N} & \text{N} & \text{V} & \text{V} & \text{V} & \text{ADJ} & \text{ADJ} & \text{ADJ} \\
\uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\
\text{N} & > & \text{ADJ} & \text{N} & > & \text{ADJ} & \text{N} & > & \text{ADJ} \\
\end{array}
\]
Measuring ambiguity on distributions over trees

Sparsity is working.

Use good ADJ grammars.

\[
\begin{array}{c|c|c}
A & B & C \\
0 & 1 & 0 \\
\end{array}
\]

max ↓ sum = 3.3
Measuring ambiguity on distributions over trees

Sparsity is working

0 1 0

.4 .6 0
Measuring ambiguity on distributions over trees

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>V</th>
<th>ADJ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.4</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>Sparsity</td>
<td>working</td>
<td>working</td>
<td>grammars</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>V</th>
<th>ADJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.6</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.7</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>
Measuring ambiguity on distributions over trees

<table>
<thead>
<tr>
<th>N</th>
<th>N</th>
<th>V</th>
<th>V</th>
<th>ADJ</th>
<th>N</th>
<th>V</th>
<th>V</th>
<th>ADJ</th>
<th>ADJ</th>
<th>ADJ</th>
<th>ADJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>N</td>
<td>V</td>
<td>V</td>
<td>ADJ</td>
<td>N</td>
<td>V</td>
<td>V</td>
<td>ADJ</td>
<td>ADJ</td>
<td>ADJ</td>
<td>ADJ</td>
</tr>
</tbody>
</table>

Sparsity is working

Use good grammars

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.4</td>
<td>.6</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.7</td>
<td>.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.4</td>
<td>.6</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Measuring ambiguity on distributions over trees

Sparsity is working

Sparsity is working

Use good grammars

Use good grammars

\[
\text{sum} = 3.3
\]

\[
\begin{array}{ccc}
0 & 1 & 0 \\
& .4 & .6 & 0 \\
0 & .7 & .3 \\
& .4 & .6 & 0 \\
\end{array}
\]

\[
\text{max} \downarrow
\]

\[
\begin{array}{cccc}
0 & 1 & .3 & .4 & .6 & 0 & .4 & .6 & 0 \\
\end{array}
\]
Minimizing ambiguity through posterior regularization

E-Step \[ q^t(y \mid x) = \arg\min_{q(y \mid x)} KL(q \parallel p_{\theta^t}) \]
Minimizing ambiguity through posterior regularization

Apply E-step penalty $L^1/\infty$ on posteriors $q(y|x)$ to induce sparsity (Graca et al., NIPS 2007 & 2009)

\[ E-\text{Step} \quad q^t(y \mid x) = \arg \min_{q(y|x)} KL(q \parallel p_{\theta^t}) + \sigma L^1/\infty(q(y|x)) \]

\[ q(y \mid x) = \begin{array}{cccc}
D & N & V & N \\
.. & .. & .. & .. \\
q(\text{root} \rightarrow x_i) \\
\end{array} \quad \begin{array}{cccc}
\text{parent} \\
D & N & V & N \\
\text{child} \\
D & .. & .. & .. \\
N & .. & .. & .. \\
V & .. & .. & .. \\
N & .. & .. & .. \\
q(x_i \rightarrow x_j) \\
\end{array} \quad \text{Probability} \quad \begin{array}{c}
. \\
. \\
. \\
\end{array} \]
Apply E-step penalty $L_{1/\infty}$ on posteriors $q(y \mid x)$ to induce sparsity (Graca et al., NIPS 2007 & 2009)

\[
\text{E-Step} \quad q^t(y \mid x) = \arg \min_{q(y \mid x)} KL(q \parallel p_{\theta^t}) + \sigma L_{1/\infty}(q(y \mid x))
\]

$q(y \mid x)$ =

- $D$  N  V  N
- \(q(\text{root } \rightarrow x_i)\)

- $D$  $N$  $V$  $N$
- $q(x_i \rightarrow x_j)$

- $D$  $\cdot$  $\cdot$  $\cdot$  $\cdot$
- $N$  $\cdot$  $\cdot$  $\cdot$  $\cdot$
- $V$  $\cdot$  $\cdot$  $\cdot$  $\cdot$
- $N$  $\cdot$  $\cdot$  $\cdot$  $\cdot$

Probability

- $\cdot$  $\cdot$  $\cdot$
Experimental results

- English from Penn Treebank: state-of-the-art accuracy

<table>
<thead>
<tr>
<th>Learning Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≤ 10</td>
</tr>
<tr>
<td>PR ($\sigma = 140$)</td>
<td>62.1</td>
</tr>
<tr>
<td>LN families</td>
<td>59.3</td>
</tr>
<tr>
<td>SLN TieV &amp; N</td>
<td>61.3</td>
</tr>
<tr>
<td>PR ($\sigma = 140$, $\lambda = 1/3$)</td>
<td>64.4</td>
</tr>
<tr>
<td>DD ($\alpha = 1$, $\lambda$ learned)</td>
<td>65.0 ($\pm 5.7$)</td>
</tr>
</tbody>
</table>

11 other languages from CoNLL-X:
- Dirichlet prior baseline: 1.5% average gain over EM
- Posterior regularization: 6.5% average gain over EM

Come see the poster for more details
Experimental results

- English from Penn Treebank: state-of-the-art accuracy

<table>
<thead>
<tr>
<th>Learning Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≤ 10</td>
</tr>
<tr>
<td>PR ($\sigma = 140$)</td>
<td>62.1</td>
</tr>
<tr>
<td>LN families</td>
<td>59.3</td>
</tr>
<tr>
<td>SLN TieV &amp; N</td>
<td>61.3</td>
</tr>
<tr>
<td>PR ($\sigma = 140$, $\lambda = 1/3$)</td>
<td>64.4</td>
</tr>
<tr>
<td>DD ($\alpha = 1$, $\lambda$ learned)</td>
<td><strong>65.0</strong> (±5.7)</td>
</tr>
</tbody>
</table>

- 11 other languages from CoNLL-X:
  - Dirichlet prior baseline: 1.5% average gain over EM
  - Posterior regularization: 6.5% average gain over EM
Experimental results

- English from Penn Treebank: state-of-the-art accuracy

<table>
<thead>
<tr>
<th>Learning Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≤ 10</td>
</tr>
<tr>
<td>PR (σ = 140)</td>
<td>62.1</td>
</tr>
<tr>
<td>LN families</td>
<td>59.3</td>
</tr>
<tr>
<td>SLN TieV &amp; N</td>
<td>61.3</td>
</tr>
<tr>
<td>PR (σ = 140, λ = 1/3)</td>
<td>64.4</td>
</tr>
<tr>
<td>DD (α = 1, λ learned)</td>
<td>65.0 (±5.7)</td>
</tr>
</tbody>
</table>

- 11 other languages from CoNLL-X:
  - Dirichlet prior baseline: 1.5% average gain over EM
  - Posterior regularization: 6.5% average gain over EM

- Come see the poster for more details