Combining Labeled and Unlabeled Data in Statistical Parsing

DARPA Site Visit – March 11, 2002
Anoop Sarkar

anoop@linc.cis.upenn.edu
http://www.cis.upenn.edu/~anoop
Supervised Statistical Parsing on the Penn Treebank

<table>
<thead>
<tr>
<th>System</th>
<th>(\leq 40)wds LP</th>
<th>(\leq 40)wds LR</th>
<th>(\leq 100)wds LP</th>
<th>(\leq 100)wds LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current (Chiang 2000)</td>
<td>86.0</td>
<td>85.2</td>
<td>86.9</td>
<td>87.0</td>
</tr>
<tr>
<td>(Charniak 99)</td>
<td>87.7</td>
<td>87.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Collins 00)</td>
<td>90.1</td>
<td>90.1</td>
<td>89.6</td>
<td>89.5</td>
</tr>
<tr>
<td>Voting (HB99)</td>
<td>90.1</td>
<td>90.4</td>
<td>89.6</td>
<td>89.9</td>
</tr>
<tr>
<td></td>
<td>92.09</td>
<td>89.18</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Can we use less labeled data and still get reasonable performance?
- Can we use the full Treebank combined with (low-cost) unlabeled data to improve parsing?
Co-Training (Blum and Mitchell 1998; Yarowsky 1995)

- Pick two “views” of a classification problem.

- Build separate models for each of these “views” and train each model on a small set of labeled data.

- Sample an unlabeled data set and to find examples that each model independently labels with high confidence. (Nigam and Ghani 2000)

- Pick confidently labeled examples.
  (Collins and Singer 1999; Goldman and Zhou 2000); Active Learning

- Each model labels examples for the other in each iteration.
Pierre Vinken will join the board as a non-executive director
Parsing as $n$-best Tree Classification and Stapler: (Xia 2000; Srinivas 1997)

Model H1: $P(T_i \mid T_{i-2}T_{i-1}) \times P(w_i \mid T_i)$
Parsing as Tree Classification and Attachment

Model H2: $\mathcal{P}(w, T \mid \text{TOP}) \times \Pi_i \mathcal{P}(w_i, T_i \mid \eta, w, T)$
The Co-Training Algorithm

1. Input: labeled and unlabeled

2. Update cache
   - Randomly select sentences from unlabeled and refill cache
   - If cache is empty; exit

3. Train models H1 and H2 using labeled

4. Apply H1 and H2 to cache.

5. Pick most probable $n$ from H1 (stapled together) and add to labeled.

6. Pick most probable $n$ from H2 and add to labeled

7. $n = n + k$; Go to Step 2
Results

- *labeled* was set to Sections 02-06 of the Penn Treebank WSJ (9625 sentences)

- *unlabeled* was 30137 sentences (Section 07-21 of the Treebank stripped of all annotations).

- A tree dictionary of all lexicalized trees. cf. (Brill 1997)
Results

- Test set: Section 23

- Baseline Model was trained only on the labeled set:
  Labeled Bracketing Precision = 72.23% Recall = 69.12%

- After 12 iterations of Co-Training:
  Labeled Bracketing Precision = 80.02% Recall = 79.64%

- Evaluation of an unsupervised approach is directly comparable to other supervised parsers (unlike previous work).
Co-training with two parsers

- Two different probability models for adjunction: single vs. multiple adjunction
- Non-overlapping lexicalized features: \(\langle \text{join}, \text{Nov} \_29 \rangle\) vs. \(\langle \text{as}, \text{Nov} \_29 \rangle\).
Co-training with two parsers

- Trained two parsers using these two models on sections 02-21 of the Penn Treebank.

- We then performed co-training using a larger set of WSJ unlabeled text (23M words).

- Even after 12 iterations of co-training, performance did not improve significantly over the baseline of LR 85.2% and LP 86%.

- Reason: Substantial overlap between the features used in each of the probability models;
  \[ \Rightarrow \text{only 22\% of the lexicalized features were different} \]
Future Work in Exploiting Unlabeled Data for Parsing

- Co-training multiple parsers (JHU summer workshop 2002)

- Explicit search for conditionally independent features

- Exploiting voting methods: combining parsers to get a Maximum Constituents Parse (can bootstrap new rules)

- Conservatively change parameter values by exploiting the generative model