Applying Co-Training Methods to Statistical Parsing

Anoop Sarkar
http://www.cis.upenn.edu/~anoop/
anoop@linc.cis.upenn.edu
the company’s clinical trials of both its animal and human-based insulins indicated no difference in the level of hypoglycemia between users of either product
Bilexical CFG (History-based parsers)

```
S
  ..
  VP{indicated}
    VB{indicated} NP{difference} PP{in}
      indicated difference in NP

```
Bilexical CFG: VP\{indicate\} → VB\{+H:indicate\} NP\{difference\} PP\{in\}
Independence Assumptions (Collins 99)

2.23%

```
 VP
  ..
  VP
   VB NP PP
```

0.06%

```
 VP
  ..
  VP
   VB NP
   PP
```

60.8%

```
 VP
   VB NP
```

0.7%

```
 VP
   VB PP NP
```
Tree Adjoining Grammars: Different Modeling of Bilexical Dependencies

The store bought IBM last week.
Probabilistic TAGs: Substitution

\[ \sum_{t'} P(t, \eta \rightarrow t') = 1 \]
Probabilistic TAGs: Adjunction

\[ \mathcal{P}(t, \eta \to NA) + \sum_{t'} \mathcal{P}(t, \eta \to t') = 1 \]
Tree Adjoining Grammars

- Simple and well-defined model for parsing. (Schabes 92, Resnik 92, Sarkar 98)
  Performance (Chiang 2000): 86.9% LR 86.6% LP (≤ 40 words)

- Locality and independence assumptions are captured elegantly.

- Parsing can be treated in two steps (Srinivas 97):
  1. Classification: structured labels (elementary trees) are assigned to each word in the sentence.
  2. Attachment: Apply substitution or adjunction to combine the elementary trees to form the parse.
Training a Statistical Parser

- How should the parameters (e.g., rule probabilities) be chosen?

- Several alternatives:
  - EM algorithm: Inside-Outside Algorithm (Schabes 92; Hwa 98)
  - Supervised training from a Treebank (Chiang 2000)
  - Parsing as Classification. Explore new machine learning techniques.
    - Achieving higher performance when using limited amounts of annotated data.
    - Conditional independence of features in the data. Can we exploit this...
Statistical Parsing: Supervised vs. Unsupervised Methods

- “Stone soup” approaches to unsupervised learning of parsers cannot handle structurally rich parses found in the Penn Treebank. (Lafferty et al 92; Della Pietra et al 94; de Marcken 95)

- A feasible technique: Combining Labeled and Unlabeled Data
  - Active Learning: Bet on which examples are the hardest. (and annotate them) (Hwa 2000)
  - Co-Training: Bet on which examples can be handled with high confidence. (use as labeled data)
Case Study in Unsupervised Methods: POS Tagging

- POS Tagging: finding categories for words

- … the stocks \textit{rose}/V … vs. … a \textit{rose}/N bouquet …

- Tag dictionary: \textit{rose}: \textit{N}, \textit{V}
  and nothing else
Case Study: Unsupervised POS Tagging

- (Cutting et al. 92) The Xerox Tagger: used HMMs with hand-built tag dictionaries. High performance: 96% on Brown

- (Merialdo 94; Elworthy 94) used varying amounts of labeled data as seed information for training HMMs. Conclusion: HMMs do not effectively combine labeled and unlabeled data

- (Brill 97) aggressively used tag dictionaries taken from labeled data to train an unsupervised POS tagger. c.f. text classification results Performance: 95% on WSJ. Approach does not easily extend to parsing: no notion of tag dictionary.
Co-Training (Blum and Mitchell 98; Yarowsky 95)

- 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 99; 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
Recursion in Parse Trees

- Usual decomposition of parse trees:

  \[ S(\text{join}) \rightarrow \text{NP(Vinken)} \ \text{VP(join)} \]

  \[ \text{NP(Vinken)} \rightarrow \text{Pierre Vinken} \]

  \[ \text{VP(join)} \rightarrow \text{will VP(join)} \]

  \[ \text{VP(join)} \rightarrow \text{join NP(board) PP(as)} \]

  \[ \ldots \]
Parsing as Tree Classification and Attachment: (Srinivas 97; Xia 2000)

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: \( P(\text{TOP} = w, T) \times \Pi_i P(w_i, T_i \mid \eta, w, T) \)
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 (run through H2) and add to labeled.

6. Pick most probable \( n \) from H2 and add to labeled

7. \( n = n + k \); Go to Step ??
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 from *labeled* and *unlabeled*. Similar to the approach of (Brill 97)
  Novel trees were treated as unknown tree tokens

- The *cache* size was 3000 sentences.
Results

- Test set: Section 23

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

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

- Methods that combine labeled and unlabeled data provide a promising new direction towards unsupervised learning.

- Co-Training, previously used for classifiers with 2/3 labels, was extended to the complex problem of statistical parsing.

- Parsing treated as providing structured (tree) labels with attachments computed between these labels.

- Evaluation of a unsupervised method for parsing directly comparable with supervised approaches.
Current Work

- Still needs human supervision to create the tree dictionary. For small datasets, this is unavoidable.

- Another application: use a large labeled dataset. But improve performance using a much larger unlabeled dataset.

- Current expt: 1M words labeled and 23M words unlabeled. Tree dictionary is completely defined by the labeled set.

- Investigating the relationship between Co-Training and EM.
Co-Training and EM

<table>
<thead>
<tr>
<th>max output of a generative model</th>
<th>gradient descent over unlabeled</th>
<th>iterative selection from unlabeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>select new examples independently</td>
<td>EM</td>
<td>co-EM*</td>
</tr>
<tr>
<td></td>
<td>Discriminative Objective Function</td>
<td>Co-Training</td>
</tr>
</tbody>
</table>

* (Nigam and Ghani, 2000)