Active Learning for Dependency Parsing with Partial Annotation

Zhenghua Li†, Min Zhang†, Yue Zhang†, Zhanyi Liu‡, Wenliang Chen†, Hua Wu‡, Haifeng Wang‡

† Soochow University, Suzhou, China
‡ Baidu Inc., Beijing, China

zhli13@suda.edu.cn
http://hlt.suda.edu.cn/~zhli
Dependency parsing

- The goal is to build an acyclic, directed tree.

$I_1$ eat the fish with $a_6$ folk

$\text{root} \rightarrow \text{subj} \rightarrow \text{eat} \rightarrow \text{obj det}$

$\text{pmod} \rightarrow \text{the fish} \rightarrow \text{with} \rightarrow \text{obj det}$
Current challenges of dependency parsing

- Web data parsing
  - Need to annotate data for evaluation/learning
- How to better serve specific tasks?
  - The usefulness of parsing outputs is far below expectation for many natural language understanding tasks (e.g. QA, dialogue, summarization, even MT).
  - Need to annotate data for specific tasks, following specialized annotation guidelines.
- The problem of current statistic models
  - How to encode knowledge such as semantics?
- Low-resource language parsing
  - multilingual transfer
Treebanking based on full annotation (FA)

- Several drawbacks
  - Annotation is extremely difficult.
  - Inter-annotator consistency is low.
  - Some structures are unnecessary to be annotated.
Treebanking based on partial annotation (PA)

- Only annotate most uncertain words for the current model (or most concerned ones)
Active learning framework

1. **Train**
   - CRF-based Parser

2. **Task selection:** $\mathcal{U}'$
   - sentence-wise (full annotation)
   - dependency-wise (partial annotation)

3. **Manual annotation**

4. **Add $\mathcal{L}'$**

   4. **Put back $\mathcal{L}'$**
      - (partial annotation)

   Newly labeled data: $\mathcal{L}'$
Previous work on AL with partial annotation (PA)

- Word segmentation [Li et al., 2012]
- Sequence labeling [Marcheggiani and Artieres, 2014]
- Constituent parsing [Hwa (1999)]
- CCG parsing [Clark and Curran (2006)]
  - Linear models (non-probabilistic)
    - Approximate uncertainty metrics
    - Heuristic methods to learn from PA
  - No human annotation experiments, except Flannery and Mori (2015)
CRF-based dependency parsing

\[ p(d|x; w) = \frac{e^{Score(x, d; w)}}{\sum_{d' \in Y(x)} e^{Score(x, d'; w)}} \]
Marginal probability of dependencies

\[ p(h \sim m | x; w) = \sum_{d \in \mathcal{Y}(x) : h \sim m \in d} p(d | x; w) \]

- \$0 \quad I_1 \quad \text{eat}_2 \quad \text{the}_3 \quad \text{fish}_4 \quad \text{with}_5 \quad a_6 \quad \text{folk}_7

- \text{p}(2 \ 5) = 0.5
- \text{p}(4 \ 5) = 0.4
Learn from partial annotation

- Convert partial trees into forests
- Maximize the probabilities of forests

\[
p(\mathcal{F}|\mathbf{x}; \mathbf{w}) = \sum_{\mathbf{d} \in \mathcal{F}} p(\mathbf{d}|\mathbf{x}; \mathbf{w}) = \frac{\sum_{\mathbf{d} \in \mathcal{F}} e^{Score(x,d;w)}}{\sum_{\mathbf{d} \prime \in \mathcal{Y}(x)} e^{Score(x,d\prime;w)}}
\]

\[
\mathcal{L}(\mathcal{D}; \mathbf{w}) = \sum_{i=1}^{N} \log p(\mathcal{F}_i|\mathbf{x}_i; \mathbf{w})
\]
Active learning framework

1. Train
   - CRF-based Parser

2. Task selection: \( U' \)
   - sentence-wise (full annotation)
   - dependency-wise (partial annotation)

3. Manual annotation

4. Add \( L' \)

4. Put back \( L' \) (partial annotation)

Newly labeled data: \( L' \)
Uncertainty metrics for FA

• Normalized tree score

\[
Conf_i(x) = \frac{\text{Score}(x, d^*)}{n^{1.5}}
\]

• Normalized tree probability

\[
Conf_i(x) = \sqrt[n]{p(d^*|x)}
\]

• Averaged marginal probability

\[
Conf_i(x) = \frac{\sum_{h \sim m \in d^*} p(h \sim m|x)}{n}
\]
Uncertainty metrics for PA

- Marginal probability max
  \[ Conf_i(x, i) = p(h^0 \sim i \mid x) \]

- Marginal probability gap
  \[ Conf_i(x, i) = p(h^0 \sim i \mid x) - p(h^1 \sim i \mid x) \]

- Marginal probability entropy
  \[ Conf_i(x, i) = \sum_{h} p(h \sim i \mid x) \log p(h \sim i \mid x) \]
Simulation experiments settings

- Chinese
  - Penn Chinese Treebank 5.1 (CTB5)
- English
  - Penn Treebank 3 (PTB)
- Automatic POS tags
- Constituent-to-dependency conversion using Penn2Malt with default rules
FA vs. PA on CTB5-dev
Active learning framework

1. Train
   - CRF-based Parser

2. Task selection: $U'$
   - sentence-wise (full annotation)
   - dependency-wise (partial annotation)

3. Manual annotation

4. Add $L'$

4. Put back $L'$ (partial annotation)
Single vs. batch dependency-wise PA?

- Single dependency-wise PA

\[ \$_0 \ I_1 \ \text{eat}_2 \ \text{the}_3 \ \text{fish}_4 \ \text{with}_5 \ a_6 \ \text{folk}_7 \]

- Batch dependency-wise PA

\[ \$_0 \ I_1 \ \text{eat}_2 \ \text{the}_3 \ \text{fish}_4 \ \text{with}_5 \ a_6 \ \text{folk}_7 \]
Single vs. batch PA on CTB5-dev
### Result on test data

<table>
<thead>
<tr>
<th></th>
<th>Chinese</th>
<th></th>
<th>English</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#Dep labeled</td>
<td>UAS</td>
<td>#Dep labeled</td>
<td>UAS</td>
</tr>
<tr>
<td>ZPar</td>
<td>318,408</td>
<td>77.97</td>
<td>908,154</td>
<td>91.45</td>
</tr>
<tr>
<td>This parser</td>
<td>318,408</td>
<td>78.36</td>
<td>908,154</td>
<td>91.66</td>
</tr>
<tr>
<td>FA (random)</td>
<td>187,123</td>
<td>77.43</td>
<td>395,199</td>
<td>90.67</td>
</tr>
<tr>
<td>FA (best)</td>
<td>149,051</td>
<td>77.32</td>
<td>197,907</td>
<td>90.66</td>
</tr>
<tr>
<td>PA (single)</td>
<td>50,958</td>
<td>77.22</td>
<td>61,448</td>
<td>90.72</td>
</tr>
<tr>
<td>PA (batch)</td>
<td>56,389</td>
<td>77.38</td>
<td>51,016</td>
<td>90.70</td>
</tr>
</tbody>
</table>

- **Chinese:** 62.2% reduction
- **English:** 74.2% reduction
Human annotation experiment settings

- CTB7 with Stanford dependencies
- Only use 12,912 sentences of length [10,20] in CTB7-train, randomly split into two parts
  - Part 1: to train a parser
  - Part 2: to select 20% most uncertain words for each sentence.
- Randomly select 100 sentences from part 2
  - Six students
  - For each sentence, 3 random students do FA task, the left 3 students do PA task.
Annotation system


**Before annotation**

```
ROOT 据统计，在民生银行的各项贷款中有**一半**是投向**非国有**经济部门的。
```

**After annotation**

```
ROOT 据统计，在民生银行的各项贷款中有**一半**是投向**非国有**经济部门的。
```
Results of human annotation exp

<table>
<thead>
<tr>
<th></th>
<th>Time: Sec/Dep</th>
<th>Annotation accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FA</td>
<td>PA</td>
</tr>
<tr>
<td>Annotator #1</td>
<td>4.0</td>
<td>7.9</td>
</tr>
<tr>
<td>Annotator #2</td>
<td>7.5</td>
<td>16.0</td>
</tr>
<tr>
<td>Annotator #3</td>
<td>10.0</td>
<td>22.2</td>
</tr>
<tr>
<td>Annotator #4</td>
<td>5.1</td>
<td>8.7</td>
</tr>
<tr>
<td>Annotator #5</td>
<td>7.0</td>
<td>17.3</td>
</tr>
<tr>
<td>Annotator #6</td>
<td>7.0</td>
<td>10.6</td>
</tr>
<tr>
<td>Overall</td>
<td>6.7</td>
<td>13.6</td>
</tr>
</tbody>
</table>

Speed on per dependency: PA ≈ 2*FA;
Quality: PA > FA
Contributions

- We first apply a probabilistic parser to active learning (AL) for dependency parsing.
- We systematically investigate several uncertainty metrics for AL with FA and PA.
- Simulation experiments show AL with PA needs much less annotated dependencies than AL with FA (62.2% less on Chinese, 74.2% less on English).
- Human annotation experiments lead to several interesting findings.
Future plan

- Further consider representativeness of unlabeled data, besides uncertainty.
- Study which syntactic structures are more suitable for human annotation.
- Start annotating partial trees on web data for some specific task.
Thanks for your time!
Questions?

Codes, experiment settings, and data are released at [http://hlt.suda.edu.cn/~zhli](http://hlt.suda.edu.cn/~zhli) for future research.
FA vs. PA on PTB-dev
Single vs. batch PA on PTB-dev
Partial trees $\Rightarrow$ forests

$0 \quad I_1 \quad \text{saw}_2 \quad \text{Sarah}_3 \quad \text{with}_4 \quad a_5 \quad \text{telescope}_6$