OntoNotes English Treebanking – Y3

2009-02-24
## Treebanking – Year 3 (Ann Taylor)

**Newly Parsed (new NML & hyphen styles)**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Eng-from-Ara Web</td>
<td>55,279</td>
</tr>
<tr>
<td>Eng-from-Chi Web</td>
<td>41,694</td>
</tr>
<tr>
<td>P2.5</td>
<td>140,322</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>237,295</strong></td>
</tr>
</tbody>
</table>

**Style Updated (Hyphens and/or NMLs)**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BC</td>
<td>151,267</td>
<td>Hyphens</td>
</tr>
<tr>
<td>BN</td>
<td>220,174</td>
<td>Hyphens</td>
</tr>
<tr>
<td>ECTB</td>
<td>57,058</td>
<td>Hyphens</td>
</tr>
<tr>
<td>WSJ300</td>
<td>352,957</td>
<td>Hyphens &amp; NMLs</td>
</tr>
<tr>
<td>WSJ400</td>
<td>536,452</td>
<td>Hyphens &amp; NMLs</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td><strong>428,499</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>889,409</strong></td>
</tr>
</tbody>
</table>
Intra-Annnotator Treebanking Experiment

- Q: How consistent is human syntactic annotation?
- Experiment:
  - Ann Taylor selected at random 23 files from BN she annotated in Y2 (6K words total).
  - The texts were re-parsed with a parser trained only on WSJ.
    - (7 sentences were skipped, due to retokenization during first annotation pass, yielding 223 re-parsed sentences, 4200 brackets.)
  - Ann re-corrected the texts.
  - evalb was run on new texts, given old texts as gold standard.
  - Separately, Ann used sdiff to find & examine all differences
Results of evalb on all sentences

- Bracketing Recall: 98.4
- Bracketing Precision: 98.5
- Tagging Accuracy: 98.7
  - (Ann corrects wrong tags as she goes.)
Mis-tagging seems to cause mis-parsing

- No tagging inconsistencies: 182 sentences
  - No bracketing inconsistencies of these: 161 (88%)

- Tagging inconsistencies: 41 sentences
  - No bracketing inconsistencies of these: 26 (63%)
Error Analysis using *sdiff*

- 146 sentences categorized for inconsistencies
- 176 disagreements total
  - Many of these aren’t *evalb* errors
- For 52 of these, Ann not sure which was “right”
- Primary cause of disagreements (some multiple)
  - POS Tag differences: 70*
  - Dash-tag differences: 35*
  - Attachment differences: 23
  - Coindexing differences: 11*
  - Node label different: 9
  - Extra node: 8
  - Other: 8
  - Missing node: 7
  - Empty category: 5*

*: Not an *evalb* error