Toward Tree Substitution Grammars with Latent Annotations

Francis Ferraro, Benjamin Van Durme and Matt Post • Johns Hopkins University

Motivation
PCFGs may be overly permissive and make unrealistic independence assumptions. Automatically-learned latent annotations of PCFG symbols can help address this issue, and work well at the lexical level (Matsuzaki et al., 2005; Petrov et al., 2006).

Tree substitution grammars (TSGs) have an extended domain of locality and can capture long-range grammatical dependencies: They give him the report.

Automatically-Induced Latent PCFGs
Matsuzaki et al., (2005) split all categories equally and learned weights via EM. Petrov et al., (2006) used an iterative split-merge EM algorithm, which recovered many of the splits manually determined by Klein and Manning, (2003) and allowed symbols to have differing numbers of refinements.

Learning Probabilistic TSGs
DOP
Extract all subtrees from observed corpora, learned fragments can be large and not generalizable.

Bayesian Methods
A DP Prior encourages compact fragments (Cohn et al., 2009; Post et al., 2009); they are non-deterministic and can be complex to implement.

Can we combine the intuitions motivating the Bayesian approach with the simplicity of DOP?

Implementation Notes
PTSG as PCFG: Make the root of every internal depth-one subtree unique and place the entirety of the TSG weight on the root depth-one rule.

Control exponential growth: (1) use binary trees; (2) forbid multiple frontier nodes from simultaneously becoming internal nodes (“chained” couplings); and (3) allow couplings only if permitted by a constraint set $c$.

Algorithm Overview

**Split symbols in two**
- Given a grammar $\mathcal{G}$ and constraint set $c$, we couple by:
  1. Constructing a grammar $\mathcal{G}'$ from $\mathcal{G}$ and allowed couplings from $c$ $\mathcal{G}' = \mathcal{G} \cup \{Y \in c \mid X \in \mathcal{G}\}$
  2. Estimating initial $\mathcal{G}'$ fragment weights
  3. Fitting weights of $\mathcal{G}'$ via inside/outside

**Merge back some refinements**

**Couple existing fragments**

WSJ (Sect. 2-3)
We learn descriptive lexicalized and unlexicalized fragments. We learn perfective constructions (→), a four-step modal construction (‘), and potentially useful extended nominals (1).

Universal Tag Set, WSJ (Sect. 2-3)
Using a coarse “universal” part-of-speech tag set we learn lexical clusters. Linguistic constraints are respected in TSG fragments, e.g., correctly placing PRONs in accusative position.

Preterminals as “X,” WSJ (Sect. 2-3)
Recovery of the universal tag set is promising: refinements reasonably correspond with open- and closed-class distinctions, which interact syntactically.

References and Acknowledgements

We would like to thank Byung Gyu Ahn for graciously helping us analyze the Korean results.

Universal Tag

<table>
<thead>
<tr>
<th>X</th>
<th>Universal Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 two, market, brain</td>
<td>NOUN</td>
</tr>
<tr>
<td>2 %, company, year</td>
<td>NOUN</td>
</tr>
<tr>
<td>1 ‘s, said, yes</td>
<td>VERB</td>
</tr>
<tr>
<td>13 is, was, are</td>
<td>VERB</td>
</tr>
<tr>
<td>3 it, he, they</td>
<td>PRON</td>
</tr>
<tr>
<td>12 which, that, who</td>
<td>PRON</td>
</tr>
<tr>
<td>6 the, a, The</td>
<td>DET</td>
</tr>
</tbody>
</table>

Korean Treebank
Common nouns refinements yield temporally representative nouns (NNC), while NNC can be adjectively inflected, and NNC can be adjectively inflected.

<table>
<thead>
<tr>
<th>NNC</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>경우 (case), 이날 (this day), 현재 (at the moment)</td>
</tr>
<tr>
<td>1</td>
<td>국제 (international), 경제 (economy), 세계 (world)</td>
</tr>
<tr>
<td>2</td>
<td>관련 (relation), 발표 (announcement), 보도 (report)</td>
</tr>
</tbody>
</table>