Soft Syntactic Constraints for Hierarchical Phrase-based Translation Using Latent Syntactic Distributions

By Z. Huang, M. Čmejrek, and B. Zhou
Presented by Austin Matthews
The Goal

• Integrate syntax into MT without being overly constraining, but without losing linguistic guidance.
• Develop a method to determine syntactic similarity between tag sequences.
Prior Work

- Hiero, no syntax
  - Try to fit any phrase into any hole

- Zollman and Venugopal
  - Annotate phrase pairs with tree bank categories
  - Require exact match to substitute
Prior Work

• 我爬上$X_1^{NP}$. $\rightarrow$ I climb $X_1^{NP}$.
• 长城 $\rightarrow$ the Great Wall (NP)
• 楼梯 $\rightarrow$ the stairs (DT NN)

• Can’t plug in “the stairs”!
The Idea

• Extract *tag sequences* from the data
• Each X non-terminal can be annotated with a *distribution* over the tag sequences
• I am reading ...

<table>
<thead>
<tr>
<th>Tag Sequence</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP</td>
<td>0.40</td>
</tr>
<tr>
<td>DT NN</td>
<td>0.35</td>
</tr>
<tr>
<td>DT NN NN</td>
<td>0.25</td>
</tr>
</tbody>
</table>

• Here ‘give the pen’ is dominated exactly and minimally by VBP NP
The Idea

- Represent tag sequences with latent syntactic categories
- For each sequence, compute distribution over latent categories
- Can now compute similarity via dot product!
- Feature: negative log of dot product

Example:
- “NP VP” = \{0.5, 0.2, 0.3\} might mean NP VP acts as:
  - a VP 50% of the time
    - He gives me the pen.
  - a JP 20% of the time
    - It’s a time telling device.
  - A NP 30% of the time
    - Him running bothers me.
Rule Extraction

- Lemma: Two maximal phrase pairs are either disjoint, or one is fully contained within the other
- Exploit this to build a tree of maximal phrase pairs
- Binarize into a forest
- Annotate nodes with constituent labels

X: Phrase, B: Non-phrase
Inducing Latent Syntax

- For each X and B non-terminal, split them into \{2,4,8,16\} non-terminals and learn their category distribution.
- Parameter tuning done through inside-outside
Experiments

- **EN \(\rightarrow\) DE**
  - Europarl (~300k TUs)
  - Average 15 tokens/TU

- **EN \(\rightarrow\) ZH**
  - Travel domain (~500k TUs)
  - Average 6 tokens/TU

- Hiero baseline with 4-gram LM

- You will be aware from the press and television that there have been a number of bomb explosions and killings in Sri Lanka.

- Wie Sie sicher aus der Presse und dem Fernsehen wissen, gab es in Sri Lanka mehrere Bombenexplosionen mit zahlreichen Toten.

- Give the pen to me.
- 钢笔 给 我。
Experiments

- Found 8.3M rules for EN→DE and 9.7M for EN→ZH
- ~180k unique tag sequences

<table>
<thead>
<tr>
<th>EN→DE</th>
<th>Baseline</th>
<th>Syntax</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>16.26</td>
<td>17.06</td>
<td>0.80</td>
</tr>
<tr>
<td>Test</td>
<td>16.41</td>
<td>17.01</td>
<td>0.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EN→ZH</th>
<th>Baseline</th>
<th>Syntax</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>46.47</td>
<td>47.39</td>
<td>0.92</td>
</tr>
<tr>
<td>Test</td>
<td>45.45</td>
<td>45.86</td>
<td>0.41</td>
</tr>
</tbody>
</table>
Evaluating Similarity Metric

<table>
<thead>
<tr>
<th></th>
<th>Very similar</th>
<th>Not so similar</th>
<th>Very dissimilar</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT JJ NN</td>
<td>DT NN DT JJ JJ NN DT ADJP NN</td>
<td>DT JJ JJ NML NN DT JJ CC INTJ VB DT NN NN JJ</td>
<td>PP NP NN NN CD VP RB NP IN CD</td>
</tr>
<tr>
<td>VP</td>
<td>VB VB RB VB PP VB DT DT NN</td>
<td>VP PP JJ NN VB NN NN VB VB RB IN JJ</td>
<td>JJ NN TO VP JJ WHNP DT NN IN INTJ NP</td>
</tr>
<tr>
<td>ADJP</td>
<td>JJ PDT JJ RB JJ</td>
<td>ADJP JJ JJ CC ADJP VB JJ JJ ADVP WHNP JJ</td>
<td>ADJP IN NP JJ AUX RB ADJP ADJP VP</td>
</tr>
</tbody>
</table>
Discussion

• Low margin of improvement

• Could show a phrase-based baseline

• Better to collapse similar categories top-down, or to build categories bottom-up?