The UOT System: Improve String-to-Tree translation Using Head-Driven Phrase Structure Grammar and Predicate-Argument Structures

Xianchao Wu¹, Takuya Matsuzaki¹, Naoaki Okazaki¹, Yusuke Miyao¹, Jun’ichi Tsujii¹,²,³

1. The University of Tokyo 2. The University of Manchester 3. National Centre for Text Mining

OBJECTIVES
- Integrate deep syntactic information into current syntax-based SMT systems
- Extract compact translation rules guided by Predicate-Argument Structures (PASs)

BACKGROUND

CFG vs. HPSG
- Both have tree structures
- POS/phrasal tags vs. Typed Feature Structures
- + PAS: semantic dependencies among a word and its arguments

Typed Feature Structure

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Type of Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAT</td>
<td>phrasal category</td>
<td>non-terminal nodes</td>
</tr>
<tr>
<td>XCAT</td>
<td>fine-grained phrasal category</td>
<td></td>
</tr>
<tr>
<td>SCHEMA</td>
<td>name of the schema applied in the node</td>
<td></td>
</tr>
<tr>
<td>HEAD</td>
<td>pointer to the head daughter</td>
<td></td>
</tr>
<tr>
<td>SEM_HEAD</td>
<td>pointer to the semantic head daughter</td>
<td></td>
</tr>
<tr>
<td>POS</td>
<td>part-of-speech</td>
<td></td>
</tr>
<tr>
<td>TENSE</td>
<td>tense of a verb (past, present, etc.)</td>
<td></td>
</tr>
<tr>
<td>VOICE</td>
<td>voice of a verb (passive/active)</td>
<td></td>
</tr>
<tr>
<td>PRED</td>
<td>type of a predicate</td>
<td></td>
</tr>
<tr>
<td>ARG kep</td>
<td>pointer to semantic arguments (n=1,2,3,4)</td>
<td></td>
</tr>
</tbody>
</table>

Predicate-Argument Structure

ARG1  
**she**  
ARG1  
**want**  
ARG2  
**dispute**  

The UOT String-to-Tree System

Framework

English sentence parsing (Enju)

Phrasal-aligned and target-language-parsed parallel corpora

Phrase-aligned parallel corpora

Tokenized parallel corpora

Original parallel corpora

Lexical analyzing

Minimum Error Rate Training on the development sets

Minimum Error Rate Training on the (5-best)

HPSG translation rules

Binning

HPSG-based translation rules

Binning

PAS-based translation rules

PAS: minimum covering tree

Rule Extraction from PAS

Rule Configuration | BLEU (%) | Dev. | Test |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ptt_3</td>
<td>30.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ptt_4</td>
<td>30.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ptt_5</td>
<td>31.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ptt_5+shpsg</td>
<td>34.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ptt_5+hpsg</td>
<td>34.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ptt_5+pasr</td>
<td>32.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ptt_5+pasr+hpsg</td>
<td>34.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>35.38</td>
<td>2003-2008</td>
<td>2009</td>
</tr>
</tbody>
</table>

0.44 BLEU points gained by changing POS/phrasal tags to Typed Feature Structures;
1.20 BLEU points gained by appending pasr to ptt_5; and
best result gained when using both hpsg and pasr.

EXPERIMENTS

Data
- IWSLT2009 Chinese-English training data
- IWSLT2003-2007 dev. sets
- IWSLT2008/2009 test sets

Rule Sets Used
- ptt: phrase translation table, maximum phrase length = 3, 4, or 5;
- shpsg: only use CAT and POS in Typed Feature Structures to approximate CFG trees;
- hpsg: normal HPSG tree based rules, using GHKM algorithm (Galley et al., 2004); and
- pasr: compact rules extracted from PASs.

RESULTS

FUTURE DIRECTIONS
- Test the proposals on written text sets, and
- extract composed rules (Galley et al., 2006) from HPSG tree structures.