Forest-to-String SMT for Asian Language Translation: NAIST at WAT 2014

Graham Neubig
Nara Institute of Science and Technology (NAIST)
2014-10-4
Features of ASPEC

• Translation between languages with different grammatical structures

流動 プラズマ を 正確 に 測定 する ため に 画像 を 再 構成 した。

an image was reconstituted in order to measure flowing plasma correctly.

• We all know: Phrase-based MT is not enough

for the accurate measurement of plasma flow image was reconstructed.
Solution?: 2-step Translation Process

- Pre-ordering [Weblio, SAS_MT, NII, TMU, NICT]

- RBMT+Statistical Post Editing [TOSHIBA, EIWA]
This is a lot of work... :(

How do I make good Japanese-English preordering rules?!

How do I make good Japanese-Chinese preordering rules?!

What about error propagation?

What if better preordering accuracy doesn’t equal better translation accuracy?
# Evidence

<table>
<thead>
<tr>
<th>Team ID</th>
<th>Organization</th>
<th>JE</th>
<th>EJ</th>
<th>JC</th>
<th>CJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAIST (Neubig, 2014)</td>
<td>Nara Institute of Science and Technology</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>EIWA (Ehara, 2014)</td>
<td>Yamanashi Eiwa College</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Kyoto-U (Richardson et al., 2014)</td>
<td>Kyoto University</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TMU (Ohwada et al., 2014)</td>
<td>Tokyo Metropolitan University</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>BJTUNLP (Cai et al., 2014)</td>
<td>Beijing Jiaotong University</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>NII (Hoshino et al., 2014)</td>
<td>National Institute of Informatics</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SAS_MT (Wang et al., 2014)</td>
<td>SAS Research and Development Co., Ltd</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Sense (Tan and Bond, 2014)</td>
<td>Saarland University &amp; Nanyang Technological University</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>NICT (Ding et al., 2014)</td>
<td>National Institute of Information and Communication Technology</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TOSHIBA (Sonoh et al., 2014)</td>
<td>Toshiba Corporation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>WASUIPS (Yang and Lepage, 2014)</td>
<td>Waseda University</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 4: The list of participants which submitted translation results to WAT2014 and their participations to each subtasks. (*Only submitted to automatic evaluations.)
Our Solution: Tree-to-String Translation

[Liu+ 06]
Requirements for a Tree-to-String Model

Source Sentence Parser

Parallel Corpus

Alignments

Rule Extraction
Rule Scoring
Optimization

Tree-to-String Model

This is a test.

It uses data.

これはテストです。
データを使用します。
Reducing our work load.

How do I make good Japanese-English preordering rules?! 

How do I make good Japanese-Chinese preordering rules?! 

What about error propagation? 

What if better preordering accuracy doesn't equal better translation accuracy?
I saw a girl with a telescope

Forest-to-string Translation

[Mi+ 08]
Travatar Toolkit

- Forest-to-string translation toolkit
- Supports training, decoding
- Includes preprocessing scripts for parsing, etc.
- Many other features (optimization, Hiero, etc...)

Available open source!
http://phontron.com/travatar
NAIST WAT System
WAT Results

First place in all tasks!
System Elements

Travatar!
Same as [Neubig & Duh, ACL2014]

Recurrent Neural Net Language Model
Pre/post Processing (UNK splitting, transliteration)
Dictionaries
Recurrent Neural Network LM

I can eat an apple

- Vector representation $\rightarrow$ robustness
- Recurrent architecture $\rightarrow$ longer context
Pre/post processing

UNK segmentation (ja-en)

球内部 試験 管立て
球 内部 試験 管立て

Kanji Normalization (ja-zh, zh-ja)

イチョウ黄叶 臭気鑑定師
イチョウ黄葉 臭気鑑定師

Transliteration (ja-en)

Japan インテック
Japan Intekku

Dictionary addition (ja-en)

膿瘍
apostema
典型 archetype
Conclusion
Future Work

**LOSE** at next year's WAT.

(Make Travatar so easy to use that others can use it to make really good MT systems for Asian languages.)

Starting soon! Training scripts to be available: http://phontron.com/project/wat2014