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

| Team ID                  | Organization                                           | JE | EJ | JC | CJ |
|--------------------------|--------------------------------------------------------|----|----|----|----|
| NAIST (Neubig, 2014)     | Nara Institute of Science and Technology               | ✓  | ✓  | ✓  | ✓  |
| EIWA (Ehara, 2014)       | Yamanashi Eiwa College                                 | ✓  | ✓  |    | ✓  |
| Kyoto-U (Richardson et al., 2014) | Kyoto University                                    | ✓  | ✓  | ✓  |✓  |
| WEBLIO-EJ1 (Zhu, 2014)   | Weblio, Inc.                                           | ✓  | ✓  | ✓  | ✓  |
| TMU (Ohwada et al., 2014) | Tokyo Metropolitan University                          | ✓  | ✓  |    | ✓  |
| BJTUNLP (Cai et al., 2014)| Beijing Jiaotong University                            | ✓  | ✓  |    | ✓  |
| NII (Hoshino et al., 2014)| National Institute of Informatics                     | ✓  | ✓  |    | ✓  |
| SAS_MT (Wang et al., 2014)| SAS Research and Development Co., Ltd                | ✓  | ✓  | ✓  | ✓  |
| Sense (Tan and Bond, 2014)| Saarland University & Nanyang Technological University| ✓  | ✓  | ✓  |✓  |
| NICT (Ding et al., 2014) | National Institute of Information and Communication Technology | ✓  | ✓  | ✓  | ✓  |
| TOSHIBA (Sonoh et al., 2014)| Toshiba Corporation                                  | ✓  | ✓  | ✓  | ✓  |
| WASUIPS (Yang and Lepage, 2014)| Waseda University                                     | ✓  | ✓  | ✓  | ✓  |

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]

ate a meal with my friend
Requirements for a Tree-to-String Model

Source Sentence
Parser

Parallel Corpus

Alignments

This is a test.
It uses data.

Rule Extraction
Rule Scoring
Optimization

Tree-to-String Model
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
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 </s>

• 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