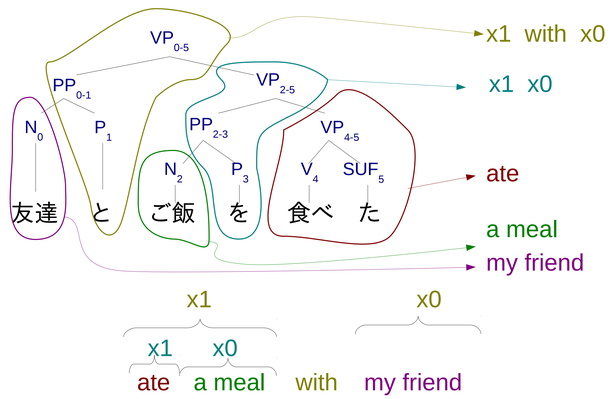
Forest-to-String SMT for Asian Language Translation: NAIST at WAT 2014

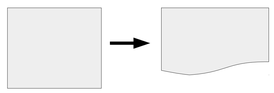
Graham Neubig
Nara Institute of Science and Technology (NAIST), Japan

Framework: F2S SMT



Data Preparation

Data Selection



- ja-zh TM: All data (672k)
- ja-en TM:
  - ASPEC first 2M
  - Optional dictionaries: EIJIRO, EDICT, Wiki
- LM: All data

Tokenization

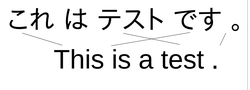
農産業 → 農 産業

- en: Stanford Tokenizer (+ split “-” and “/”)
- zh: Stanford Segmenter
- ja: KyTea

Parsing

- Parser: Egret
- en→Penn TB, zh→Penn CTB, ja→Ja. Word Dependency Corpus [Mori+14]
- For Japanese, convert w/ head rules

Alignment



- ja-zh: GIZA++
- ja-en: Nile (supervised syntactic aligner)
  Trained on KFTT aligned data

Base Model

Translation Model

- Synchronous tree substitution grammar
- 5 composed rules
- Kneser-Ney count smoothing
- Right binarization

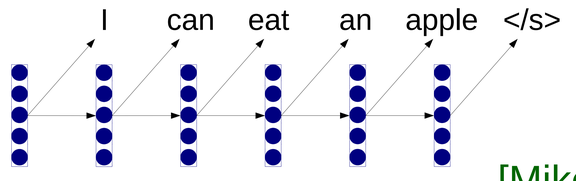
Language Model

- KenLM trained 6-gram model
  - For ja, interpolated zh-ja and en-ja data

Optimization

- Minimum error rate training
- Tested with BLEU or BLEU+RIBES

Recurrent Neural Network LM



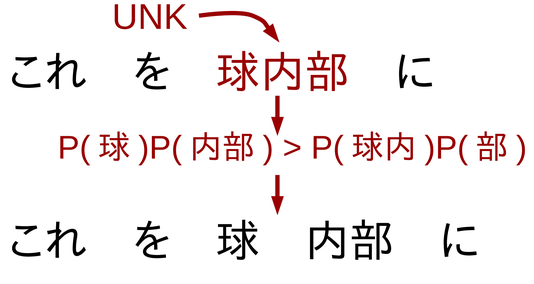
[Mikolov+ 10]

- Improve robustness, longer context
- 500 hidden nodes, 300 classes
- Trained on first 500k sentences
- 10,000-best reranking

Unknown Splitting (ja-en)

[Koehn+ 03]

- Choose split that maximizes unigram prob:



- Test time only

Transliteration/Dictionaries

ja-zh, zh-ja

- Post: Convert Simplified ↔ Japanese Kanji
  - Use Kanconvit.pm script

イチョウ黄叶→イチョウ黄葉　　臭気鉴定师→臭気鑑定師

ja-en

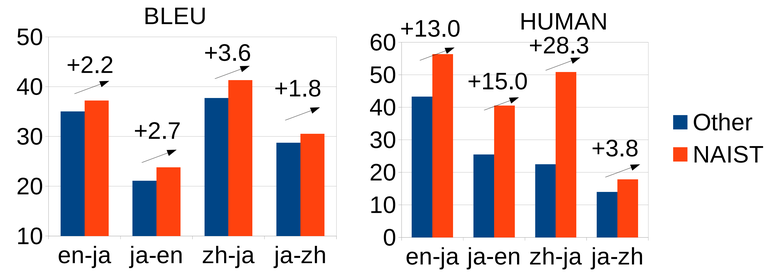
- Pre: Normalize 標題 to 表題
- Post: If word exists in dictionary, translate
  - Eijiro, Edict, Wiki Language Link
- Post: Romanize Hiragana/Katakana Words

Japan インテック→ Japan Intekku

- Post: Delete remaining Japanese words

Results

First place in all tasks!



RNNLM Helps!

BLEU	en-ja	ja-en	zh-ja	ja-zh
w/o RNN	36.50	23.76	39.82	29.27
w/ RNN	**37.21**	**24.72**	**40.61**	**29.78**

Tuning w/ RIBES hurts human eval!

Tune	BLEU			B+R		
Eval	B	R	H	B	R	H
en-ja	**37.2**	80.2	**56.3**	**37.2**	**80.7**	**51.5**
zh-ja	**41.3**	83.5	**50.8**	40.8	**83.8**	38.0
ja-zh	**30.5**	81.8	**17.8**	29.8	**83.0**	1.3

Why? Too-short hypotheses.

Scripts available!
http://phontron.com/project/wat2014