Chinese-to-Japanese Patent Machine Translation based on Syntactic Pre-ordering for WAT 2015

Katsuhito Sudoh and Masaaki Nagata, NTT Communication Science Laboratories, Japan

NTT

Overview

Dependency-based pre-ordering for Zh-Ja MT

- Patent-adapted in-house dependency parser
- Two-types of pre-ordering:
  * Rule-based, *Head Final Chinese* (Han+ 2012)
  * Data-driven, Learning to Rank (Yang+ 2012)
- Rule-based system is better, comparable to T2S

Syntactic Analysis

[***Word segmentation & POS tagging***]

- Joint sequential labeling (Suzuki+ 2012)

[***Dependency parsing (untyped)***]

- Second-order graph-based parsing

[***Semi-supervised learning***] (Suzuki+ 2009)

- Labeled: 31K sents. (news), 35K sents. (patents)
- Unlabeled: 9GB (news), 100GB (patents)

Table 1: Performance in Chinese syntactic analysis

	Word seg.	POS	Dep.
Accuracy (F0 / UAS)	0.927	0.855	0.927

References:

Han, Dan et al., Head Finalization Reordering for Chinese-to-Japanese MT, Proc. SSST-6 (2012)

Hoshino, Sho et al., Discriminative Preordering Meets Kendall's tau Maximization, Proc. ACL (2015)

Isozaki, Hideki et al, HPSG-Based Preprocessing for English-to-Japanese Translation, ACM TALIP No.11 Vol.3 (2012)

Suzuki, Jun et al, An Empirical Study of Semi-supervised Structured Conditional Models for Depenedency Parsing, Proc. EMNLP (2009)

Suzuki, Jun et al., 拡張ラグランジュ緩和を用いた同時自然言語解析法, Proc. NLP (2012) [in Japanese]

Yang, Nan et. al., A Ranking-based Approach to Word Reordering for SMT, Proc. ACL (2012)

Rule-based pre-ordering

Reordering into *head-final* order in Japanese
(En-Ja: Isozaki+ 2012, Zh-Ja: Han+ 2012)

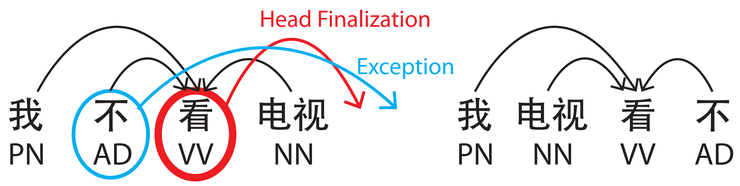
Base rule: Moving a head word *after* its modifiers

Exceptions (*placed after their head words*):

AS (*aspect particle*), SP (*sentence-final particle*)

PU (*punctuation*), CC (*coordinating conjunction*)

IJ (*interjection*), "不" (*negation*), "等" ("*etc.*")



Pros: stability, domain independence (?)
Cons: effort for rule management

Data-driven pre-ordering

Reordering by reranking a head & its modifiers
(Yang+ 2012)

- Implemented with Ranking SVM
  * Features:
    - surface/POS (head & modifier)
    - head surface/POS (h & m)
    - modifier surface/POS (head)
    - span surfaces/POSs (modifier)
    - relative position (h & m)
  * Reordering oracles are determined by ***maximizing Kendall's tau*** criterion (Hoshino+ 2015)

Pros: no special effort, target adaptability
Cons: instability, noisy auto. word alignment

SMT setup

Standard Moses Phrase-based MT

- MGIZA word alignment, g-d-f-a symal
- Kneser-Ney phrase-table score smoothing
- Word 5-gram LM with Kneser-Ney smoothing
- Distortion limit: 9 (chosen over 0,3,6,9)
- Weights chosen over 5 indep. MERT runs

Results

Comparable to T2S baseline

Rule-based is better than data-driven

Table 2: Official evaluation results

	Human	RIBES	BLEU
BL PBMT	n/a	0.781	0.382
BL T2S	***20.75***	0.814	0.394
Rule-based	16.25	***0.822***	***0.406***
Data-driven	8.00	0.812	0.399

Conclusion

Pre-ordering is a deterministic approx. of T2S
--- good in efficiency with some loss in accuracy
> *forest-based pre-ordering, pre-ordering lattice*

Rule-based pre-ordering works robustly
--- due to head-final nature in Japanese

Data-driven pre-ordering is still challenging...
--- difficulty in word alignment, non-parallelism
--- constituent or dependency structures?

Remained patent MT issues:

- Context awareness (consistency)
- Domain awareness (lexical choice)