KLE Lab POSTECH

POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

The POSTECH Statistical Machine Translation Systems for NTCIR-7 Patent Translation Task

Jin-Ji Li[0], Hwi-Dong Na[0], Hankyong Kim[*], Chang-Hu Jin[*], and Jong-Hyeok Lee[0]

[0]Department of Computer Science and Engineering,
Electrical and Computer Engineering Division,

[*]Graduate School for Information Technology,

Pohang University of Science and Technology (POSTECH),

San 31 Hyoja Dong, Pohang, 790-784, R. of Korea

E-mail: {ljj, leona, arch, hchchh, jhlee}@postech.ac.kr

Contents

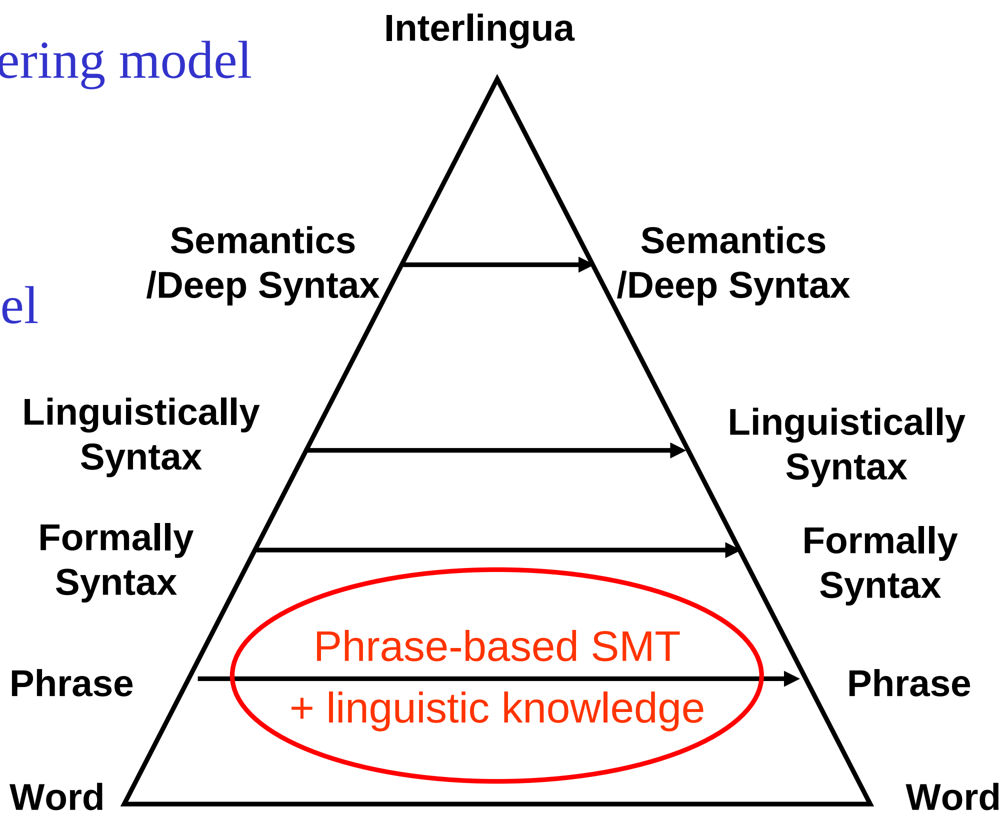
Introduction

- **State-of-the-art system: Phrase-based SMT**
  - **However, [Kevin Knight, 2007] said**
    - Translation output is ‘n-grammatical’, not grammatical
    - Re-ordering is poorly explained as ‘distortion’
- **Linguistically-distant language pairs require more sophisticated linguistic knowledge**
  - **Japanese and English**
    - Big differences in morphological & word-order typologies
- **How to effectively encode linguistic knowledge into SMT?**

Where & How to Encode Linguistic Knowledge?

In what steps of SMT system to apply?

- **Preprocessing**
  - Source word reordering model
- **Post-processing**
  - Cluster-based model



Word Reordering Model as Preprocessing

[5]

Motivation

- **Japanese and English is linguistically distant language pair**
  - SOV vs. SVO language
- **Reorder the word sequence of source language similar to target language before decoding**
  - A complement to a phrase-based SMT system which uses a relatively simple distortion model in the decoding phase

System overview

- **Parse Japanese sentences into dependency trees**
- **Apply a series of manually constructed reordering rules to each node recursively**
- **Recover the surface strings from the reconstructed dependency trees**

Word Reordering Model as Preprocessing

Ex) Reordering rules

LHS	RHS
を$_0$. (動詞-自立$_1$)	(1) 0
の$_0$. (を$_1$)	(1) 0
に$_0$. (動詞-自立$_1$)	(1) 0
の$_0$. (により$_1$)	(1) 0
の$_0$. (の$_1$)	(1) 0
の$_0$. (に$_1$)	(1) 0
の$_0$. (名詞-数$_1$)	(1) 0
の$_0$. (名詞-一般$_1$)	(1) 0
は$_0$. を$_1$. (動詞-自立$_2$)	0 (2) 1
は$_0$. に$_1$. (動詞-自立$_2$)	0 (2) 1
を$_0$. に$_1$. (動詞-自立$_2$)	(2) 0 1
に$_0$. を$_1$. (動詞-自立$_2$)	(2) 1 0
の$_0$. (で$_1$)	(1) 0
が$_0$. に$_1$. (動詞-自立$_2$)	0 (2) 1
で$_0$. (動詞-自立$_1$)	(1) 0

…… ……

Word Reordering Model as Preprocessing

[7]

- **Ex)**

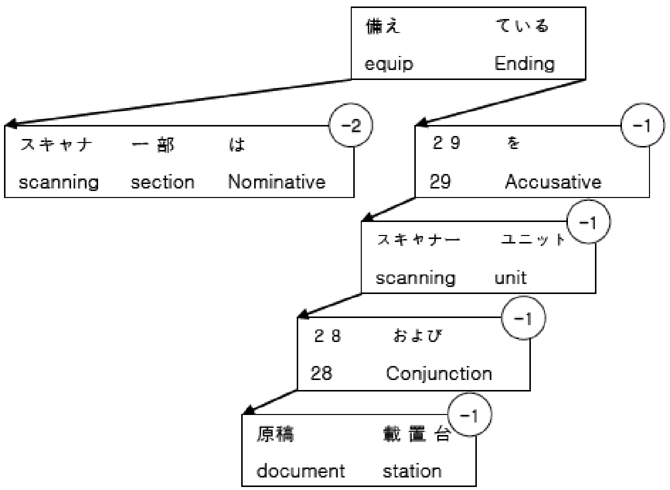


Figure 1. A dependency tree of a Japanese sentence with head-relative position information.

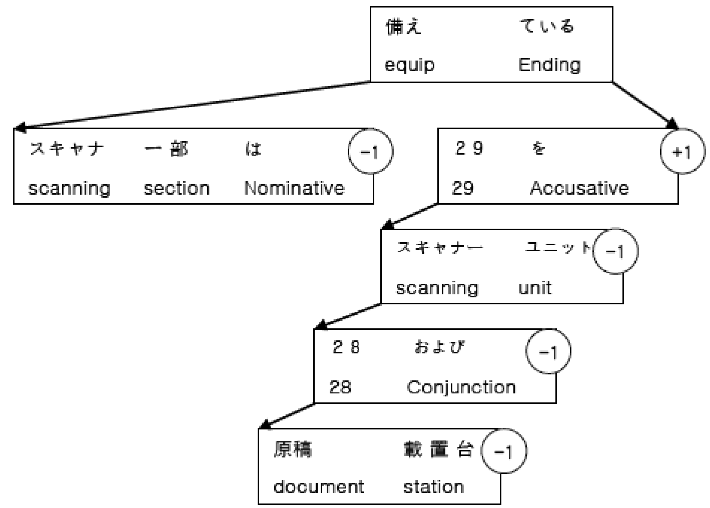


Figure 2. A dependency tree of a Japanese sentence after reordering.

Before reordering:
スキャナ/一部/は　原稿/載置台　２８/および
スキャナ/ーユニット　２９/を　備え/ている

After reordering:
スキャナ/一部/は　備え/ている　原稿/載置台/
２８/および　スキャナ/ーユニット　２９/を

Cluster-based Model

Motivation

- Usually, sentences with similar syntactic structures yield similar distributions of n-grams reflecting their word order
- Cluster-specific LM benefits SMT system

System overview

- 1. Predict clusters according to cluster types
- 2. Translate using baseline SMT system
- 3. Optimize LM integration parameters
- 4. Re-translate using general + cluster-specific LM
- 5. Select best translation result

Cluster-based Model

- **27 cluster types of syntactic patterns**
  - **Subtree structures in the source dependency trees which ignore the adjuncts as cluster types**

- **Ex) Cluster type list**

Cluster type	Freq.
を. (動詞-自立)	407,629
に. (動詞-自立)	246,188
が. (動詞-自立)	143,579
動詞-自立. (動詞-自立)	134,566
は. に. (動詞-自立)	81,434
は. を. (動詞-自立)	79,717
は. (ある)	63,646
を. に. (動詞-自立)	59,294
に. を. (動詞-自立)	53,438
と. (動詞-自立)	39,354
は. (動詞-自立)	36,164
……	……

NTCIR Corpus Profile

	Training corpus (1,172,709 sentences)	
	Chinese	**Korean**
Number of words	**30,761,076**	**28,683,697**
Number of singletons	**131,219**	**131,321**
Average length	**26.23**	**24.46**
	Development corpus (609 sentences)	
	Japanese	**Korean**
Number of words	**15,997**	**14,818**
Number of singletons	**2,697**	**2,817**
Average length	**26.27**	**24.33**
	Test corpus (1,381 sentences)	
	Japanese	**English**
Number of words	**48,278**	**44,910**
Number of singletons	**4,088**	**4,273**
Average length	**34.96**	**32.52**

Experimental Scenario

- **Corpus processing**
  - **Japanese: Cabocha tokenizer and parser**
  - **http://chasen.org/~taku/software/cabocha/**
- **English-to-Japanese SMT**
  - **Vanilla MOSES with 5-gram LM**
- **Japanese-to-English SMT**
  - **Vanilla MOSES with syntactic motivated knowledge**
    - Source word reordering model as preprocessing
    - Cluster-based model as post-processing

NTCIR7 Results

[12]

1

Method	Bleu
Baseline	24.48
First Method	24.21
Second Method	23.45

Table 4. The Bleu values when reordering models are applied.

2

General LM	Bleu
Baseline	24.48
(1 - λ) * General LM + λ * Cluster-specific LM	Bleu
λ = 0.1	24.67
λ = 0.2	24.54
λ = 0.3	24.52
λ = 0.4	24.44
λ = 0.5	24.28
λ = 0.6	24.25
λ = 0.7	23.99
λ = 0.8	23.76
λ = 0.9	23.59

Table 6. The optimized parameter when integrating general and cluster-specific LM.

3

Training corpus size	Baseline	Cluster-based
50k	21.48	22.14
100k	22.55	22.91
300k	23.46	23.74
All	24.48	24.67

Table 7. The Bleu values when the training corpus size is different.

✓This is not the official experimental results.

Conclusion & Future Work

- Source word reordering model
  - Need human evaluation to verity the effectiveness of proposed method
- Cluster-based model
  - Applied to a small size corpus, it worked better than when applied to a large size one
  - Cluster types are too simple which can cause multiple matching
- Enlightening SMT with various linguistic knowledge
  - Developing more elaborate reordering rules & applying other cluster types

Reference

- [1] Michael Collins, Philip Koehn, and Ivona Kučerová, *Clause restructuring for statistical machine translation*, In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005
- [2] Fei Xia and Michael McCord, *Improving a statistical MT system with automatically learned rewrite patterns*. In Proceedings of the 20th international Conference on Computational Linguistics, 2004
- [3] Chao Wang, Michael Collins, Philip Koehn, *Chinese Syntactic Reordering for Statistical Machine Translation* Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp.737-745, 2007
- [4] Deepa Gupta, Mauro Cettolo, and Marcello Federico. *POS-based reordering models for statistical machine translation*. In Proceedings of Machine Translation Summit XI, pp. 207-213, 2007
- [5] Kay Rottmann and Stephan Vogel. *Word reordering in statistical machine translation with a POS-based distortion model*; Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation, 2007
- [6] Yuqi Zhang, Richard Zens, and Hermann Ney: *Chunk-level reordering of source sentences with automatically learned rules for statistical machine translation*. Workshop at op at NAACL-HLT 2007 "Syntax and structure in statistical translation", pp.1-8, 2007

[6] Yuqi Zhang, Richard Zens, and Hermann Ney: *Chunk-level reordering of source*

Reference

- [7] Yuqi Zhang, Richard Zens, Hermann Ney: *Improved chunk-level reordering for statistical machine translation*. In Proceeding of International Workshop on Spoken Language Translation, 2007
- [8] Sasa Hasan, Hermann Ney. *Clustered Language Models based on Regular Expressions for SMT*, 2005, EAMT
- [9] Hirofumi Yamamoto, Eiichiro Sumita. *Bilingual Cluster Based Models for Statistical Machine Translation*, 2007, EMNLP
- [10] Matthias Eck, Stephan Vogel, Alex Waibel. *Language Model Adaptation for Statistical Machine Translation based on Information Retrieval*, 2004, LREC
- [11] Bing Zhao, Matthias Eck, Stephan Vogel. *Language Model Adaptation for Statistical Machine Translation with Structured Query Models*, 2004, COLING
- [12] Masao Utiyama and Hitoshi Isahara. A Japanese-English patent parallel corpus. MT Summit XI, 2007
- [13] Masao Utiyama, Mikio Yamamoto, Atsushi Fujii, and Takehito Utsuro. Description of Patent Parallel Corpus for NTCIR-7 Patent Translation Task. http://if-lab.slis.tsukuba.ac.jp/fujii/ntc7patmt/ppc.pdf
- [14] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop. Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2008.