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

Katsuhito Sudoh **Masaaki Nagata**
NTT Communication Science Laboratories, Japan
sudoh.katsuhito@lab.ntt.co.jp

Abstract

This paper presents our Chinese-to-Japanese patent machine translation system for WAT 2015 (Group ID: ntt) that uses syntactic pre-ordering over Chinese dependency structures. A head word and its modifier words are reordered by hand-written rules or a learning-to-rank model. Our system outperforms baseline phrase-based machine translations and competes with baseline tree-to-string machine translations.

1 Introduction

Patent documents, which well-structured written documents that describe the technical details of inventions, are expected to have almost no semantic ambiguities caused by indirect or rhetorical expressions. Due to this aspect, patent documents are good candidates for literal translation, which most machine translation (MT) approaches aim to do.

One technical challenge for patent machine translation is the complex syntactic structure of patent documents, which typically have long sentences that complicate MT reordering, especially for the word order in distant languages. Chinese and Japanese have similar word order in noun modifiers but different subject-verb-object order, requiring long distance reordering in translation. In this year's WAT evaluation campaign (Nakazawa et al., 2015), we tackle long distance reordering by syntactic pre-ordering based on Chinese dependency structures (Sudoh et al., 2014) in a Chinese-to-Japanese patent translation task.

Our system basically consists of three components: Chinese syntactic analysis (word segmentation, part-of-speech (POS) tagging, and dependency parsing) adapted to patent documents; dependency-based syntactic pre-ordering with hand-written rules or a learning-to-rank model; and a standard phrase-based statistical MT. This paper describes our system's details and discusses our evaluation results.

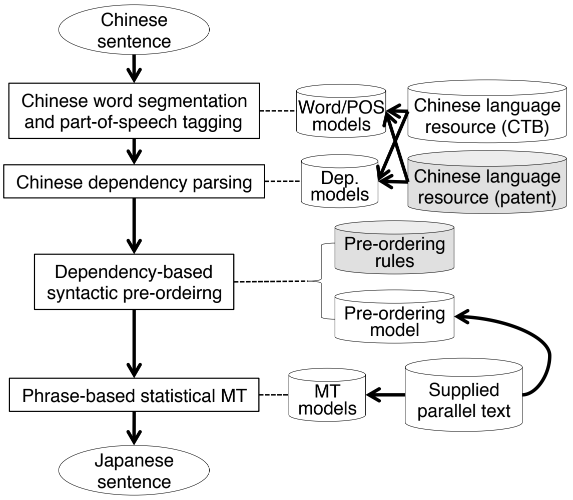
Chinese sentence
Chinese word segmentation and part-of-speech tagging
Word/POS models
Chinese language resource (CTB)
Chinese dependency parsing
Dep. models
Chinese language resource (patent)
Dependency-based syntactic pre-ordeirng
Pre-ordering rules
Pre-ordering model
Phrase-based statistical MT
MT models
Supplied parallel text
Japanese sentence

Figure 1: Brief workflow of our Chinese-to-Japanese MT system. Gray resources are in-house ones.

2 System Overview

Figure 1 shows a brief workflow of our Chinese-to-Japanese MT system. Its basic architecture is standard with syntactic pre-ordering. Input sentences are first applied to word segmentation and POS tagging, parsed into dependency trees, reordered using pre-ordering rules or a pre-ordering model, and finally translated into Japanese by a phrase-based statistical MT.

3 Chinese Syntactic Analysis: Word Segmentation, Part-of-Speech Tagging, and Dependency Parsing

Word segmentation and POS tagging are solved jointly (Suzuki et al., 2012) for better Chinese word segmentation based on the POS tag sequences. The dependency parser produces *un-*

Proceedings of the 2nd Workshop on Asian Translation (WAT2015), pages 95–98,
Kyoto, Japan, 16th October 2015.

typed dependency trees. The Chinese analysis models were trained using in-house Chinese treebanks in the patent domain (about 35,000 sentences) as well as the standard Penn Chinese Treebank dataset (Sudoh et al., 2014). The training also utilized unlabeled Chinese patent documents (about 100 G bytes) for semi-supervised training (Suzuki et al., 2009; Sudoh et al., 2014).

4 Syntactic Pre-ordering

We compared two different syntactic pre-ordering approaches, rule-based and data-driven, in this evaluation campaign.

4.1 Rule-based Pre-ordering

Rule-based pre-ordering is an intuitive approach that generates target language word order based on linguistic knowledge and analysis. For example, in an English-to-Japanese patent MT, a very simple pre-ordering rule called Head Finalization (Isozaki et al., 2012) gives very successful results.

Our pre-ordering rules are based on Head Final Chinese (Han et al., 2012) developed for Chinese HPSG trees. We modified the HPSG-oriented rules for dependency structure, but their basic actions are almost the same. One expected advantage of rule-based pre-ordering is that its stability is independent on bilingual corpora, while model-based approaches basically depend on bilingual corpora to determine *which reordering is the best*.

4.2 Data-driven Pre-ordering by Learning-to-Rank

Data-driven pre-ordering obtains the most probable reordering of a source language sentence that is *monotone* with the target language counterpart. It learns rules or models using reordering oracles over word-aligned bilingual corpora.

We used a learning-to-rank approach with Ranking SVMs (Yang et al., 2012), which reorders the head word and its modifier words in a dependency tree based on their *ranks*. The features resemble those by Yang et al. (2012); we did not use label-related ones because our dependency trees do not have labels. The reordering oracles were determined to maximize Kendall's τ over automatic word alignment in a similar manner to Hoshino et al. (2015). The only difference is the tree structure; Hoshino et al. (2015) used binary trees and just considered monotone or reverse for two child nodes of each tree node. But we use dependency trees and have to consider all the possible permutation over one head word and one or more modifier words.

5 Evaluation

5.1 Setup

We trained a word n-gram language model and two different phrase-based translation models by the above different pre-ordering approaches. We used all of the supplied Chinese-Japanese bilingual training corpora of one million sentence pairs (except for long sentences over 64 words) for the MT models: phrase tables, lexicalized reordering tables, and word 5-gram language models using standard Moses and KenLM training parameters. We applied modified Kneser-Ney phrase table smoothing with an additional phrase scoring option: `--KneserNey`. The model weights were optimized by standard Minimum Error Rate Training (MERT), but we compared five independent MERT runs and chose the best weights for the development test set. The distortion limits were also chosen from 0, 3, 6, and 9 by comparing the results of the MERT runs. We chose 9 both for the rule-based and data-driven pre-ordering.

The pre-ordering model for the data-driven method was trained by the word alignment used for the phrase table by a Ranking SVM implementation with Liblinear[1]. Its soft margin parameter `C` was chosen by the ranking accuracy on the development set.

5.2 Official Results

Table 1 shows the official evaluation results by the organizers in Pairwise Cloudsourcing Evaluation scores (Human), RIBES, and BLEU. Our rule-based system showed slightly better performance in RIBES and BLEU than the tree-to-string baseline, but the difference may not be significant. The performances of our systems were lower than the tree-to-string baseline in the Human evaluation. With respect to the difference in the pre-ordering approaches, the rule-based system outperformed the data-driven one.

5.3 Discussion

One critical concern is the difference between tree-to-string baseline and our pre-ordering systems. Syntactic pre-ordering based on child ranking/classification is a simple approximation of a

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/

Table 1: Official evaluation results in Pairwise Cloudsourcing Evaluation scores (Human), RIBES, and BLEU. RIBES and BLEU are based on MeCab Japanese word segmentation. Scores in **bold** are the best ones.

System	Human	RIBES	BLEU
Organizer PBMT	n/a	0.781	0.382
Organizer T2S	**20.75**	0.814	0.394
Ours rule-based	16.25	**0.822**	**0.406**
Ours data-driven	8.00	0.812	0.399

standard tree-to-string MT[2]. The tree-to-string MT can compare different reordering hypotheses; our pre-ordering just chooses one-best pre-ordering.

A comparison between the results by rule-based and data-driven pre-ordering systems suggests our pre-ordering rules work robustly. Even though we expected model-based pre-ordering to capture the complex reordering phenomena in the dependency structure, it gave worse results than the rule-based one. One possible reason is the noisy automatic word alignment in the bilingual corpora; using better word alignment (manual annotation or supervised word alignment) is promising to learn a consistent pre-ordering model (Hoshino et al., 2015).

5.4 Issues for Context-aware Machine Translation

We did not include any context-aware constraints or features in our system, because it translated every sentence independently. We just tried using domain-dependent translation models based on the given category information (chemistry, electricity, mechanical engineering, and physics), but they did not work effectively in our pilot test.

6 Conclusion

This paper presented our pre-ordering-based system for Chinese-to-Japanese patent MT for the WAT 2015 evaluation campaign. Our results showed that pre-ordering had similar MT performance to the tree-to-string baseline, but it was slightly worse in the human evaluation. Future work will investigate sophisticated pre-ordering methodology such as pre-ordering lattices or forest-based pre-ordering and better word alignment for data-driven pre-ordering.

Acknowledgments

We thank Dr. Jun Suzuki for his great help on the Chinese dependency parsing. The basis of our pre-ordering rules was developed in cooperation with Dr. Dan Han, whom we thank. Finally, we greatly appreciate the workshop organizers for this valuable evaluation campaign.

References

Dan Han, Katsuhito Sudoh, Xianchao Wu, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2012. Head Finalization Reordering for Chinese-to-Japanese Machine Translation. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 57–66.

Sho Hoshino, Yusuke Miyao, Katsuhito Sudoh, Katsuhiko Hayashi, and Masaaki Nagata. 2015. Discriminative Preordering Meets Kendall's τ Maximization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 139–144.

Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2012. HPSG-Based Preprocessing for English-to-Japanese Translation. *ACM Transactions on Asian Language Information Processing*, 11(3).

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2015. Overview of the 2nd Workshop on Asian Translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, Kyoto, Japan, October.

Katsuhito Sudoh, Jun Suzuki, Yasuhiro Akiba, Hajime Tsukada, and Masaaki Nagata. 2014. A English/Chinese/Korean-to-Japanese Statistical Machine Translation System for Patent Documents. In *Proceedings of the 20th Annual Meeting of the Association for Natural Language Processing*, pages 606–609. (in Japanese; 須藤,鈴木,秋葉,塚田,永田: 英中韓から日本語への特許文向け統計翻訳システム).

Jun Suzuki, Hideki Isozaki, Xavier Carreras, and Michael Collins. 2009. An Empirical Study of Semi-supervised Structured Conditional Models for Dependency Parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 551–560.

[2] Theoretically this is not true, since phrase-based MT can use phrases that have non-syntactic spans. The effect of these phenomena has not reported yet in the field of pre-ordering; they introduce more ambiguities in phrasal translations, but the ambiguities may work both positively and negatively. The use of non-syntactic spans was proposed as an extention in tree-to-string MT (Zollmann and Venugopal, 2006; Zhang et al., 2011).

Jun Suzuki, Kevin Duh, and Masaaki Nagata. 2012. Joint Natural Language Analysis using Augmented Lagrangian. In *Proceedings of the 18th Annual Meeting of the Association for Natural Language Processing*, pages 1284–1287. (in Japanese; 鈴木,Duh,永田: 拡張ラグランジュ緩和を用いた同時自然言語解析法).

Nan Yang, Mu Li, Dongdong Zhang, and Nenghai Yu. 2012. A Ranking-based Approach to Word Reordering for Statistical Machine Translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 912–920.

Hao Zhang, Licheng Fang, Peng Xu, and Xiaoyun Wu. 2011. Binarized forest to string translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 835–845.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141.