Chinese-to-Japanese Patent Machine Translation based on Syntactic Pre-ordering for WAT 2016

Katsuhito Sudoh and **Masaaki Nagata**
Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
sudoh.katsuhito@lab.ntt.co.jp

Abstract

This paper presents our Chinese-to-Japanese patent machine translation system for WAT 2016 (Group ID: `ntt`) that uses syntactic pre-ordering over Chinese dependency structures. Chinese words are reordered by a learning-to-rank model based on pairwise classification to obtain word order close to Japanese. In this year's system, two different machine translation methods are compared: traditional phrase-based statistical machine translation and recent sequence-to-sequence neural machine translation with an attention mechanism. Our pre-ordering showed a significant improvement over the phrase-based baseline, but, in contrast, it degraded the neural machine translation baseline.

1 Introduction

Patent documents, which are well-structured written texts that describe the technical details of inventions, are expected to have almost no semantic ambiguities caused by indirect or rhetorical expressions. Therefore, they 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 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 the WAT 2016 evaluation campaign (Nakazawa et al., 2016), we participated in a Chinese-to-Japanese patent translation task and tackled long distance reordering by syntactic pre-ordering based on Chinese dependency structures, as in our last year's system (Sudoh and Nagata, 2015). We also use a recent neural MT as the following MT implementation for comparison with a traditional phrase-based statistical MT.

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 the following MT component (phrase-based MT or neural MT). This paper describes our system's details and discusses our evaluation results.

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 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 POS tag sequences. The dependency parser produces *untyped* dependency trees. The

Proceedings of the 3rd Workshop on Asian Translation,
pages 211–215, Osaka, Japan, December 11-17 2016.

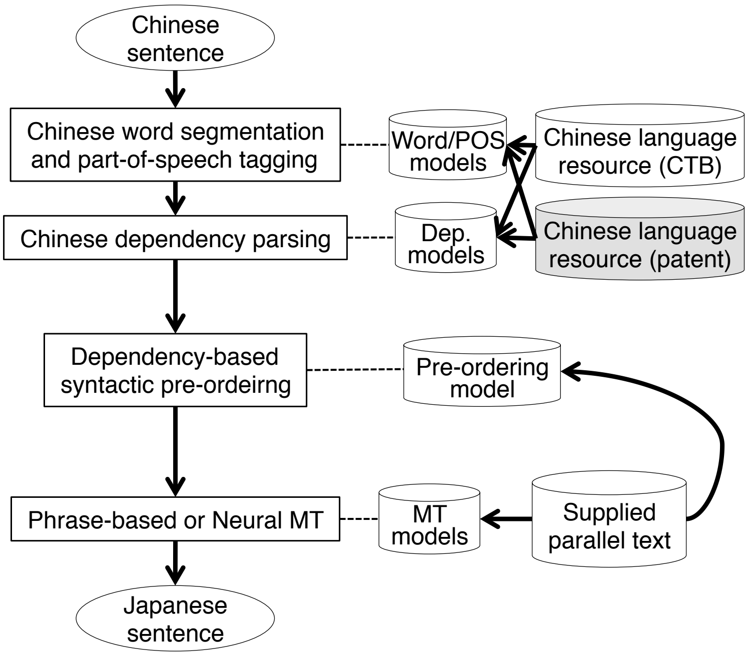

Figure 1: Brief workflow of our MT system. Gray-colored resource is an in-house one.

Chinese analysis models were trained using an in-house Chinese treebank of about 35,000 sentences in the patent domain (Sudoh et al., 2014) as well as the standard Penn Chinese Treebank dataset. 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

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 pairwise-classification-based model for pre-ordering (Jehl et al., 2014), instead of Ranking SVMs (Yang et al., 2012) that we used the last year. An advantage of pairwise classification is that we can use features defined on every node pair, while we can only use node-wise features with Ranking SVMs. We found that the pairwise-based method gave slightly better pre-ordering performance than the Ranking SVMs in our pilot test, as did Jehl et al. (2014).

We also renewed the features for this year's system. We used span-based features (word and part-of-speech sequences over dependency sub-structures) like Hoshino et al. (2015), word and part-of-speech n-grams (n=2,3,4) including head word annotations, and those described in Jehl et al. (2014). Since these features are very sparse, we chose those appearing more than twice in the training parallel data. The reordering oracles were determined to maximize Kendall's τ over automatic word alignment in a similar manner to Hoshino et al. (2015). We used the intersection of bidirectional automatic word alignment (Nakagawa, 2015). The pairwise formulation enables a simple solution to determine the oracles for which we choose a binary decision, obtaining higher Kendall's τ with and without swapping every node pair.

5 Evaluation

5.1 Pre-ordering Setup

The pre-ordering model for the data-driven method was trained over the MGIZA++ word alignment used for the phrase tables described later. We trained a logistic-regression-based binary classification model using the reordering oracles over training data with LIBLINEAR (version 2.1). Hyperparameter `c` was set to 0.01, chosen by the binary classification accuracy on the development set.

5.2 Phrase-based MT Setup

The phrase-based MT used in our system was a standard Moses-based one. We trained a word n-gram language model and phrase-based translation models with and without pre-ordering. 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 limit was 9 for both the baseline and pre-ordering conditions, chosen from 0, 3, 6, and 9 by comparing the results of the MERT runs.

5.3 Neural MT Setup

We also tried a recent neural MT for comparison with a phrase-based MT. We used a sequence-to-sequence attentional neural MT (Luong et al., 2015) implemented by the Harvard NLP group[1] with a vocabulary size of 50,000 and a 2-layer bidirectional LSTM with 500 hidden units on both the encoder/decoder[2]. The neural MT, which was word-based with the same tokenizer used in the phrase-based MT setting, did not employ recent subword-based or character-based methods. The training time of the neural MT was about two days (13 epochs with 3.5 hours/epoch) with a NVIDIA Tesla K80 GPU. The decoding employed a beam search with a beam size of five and dictionary-based unknown word mapping with the IBM-4 lexical translation table obtained by MGIZA++.

5.4 Official Results

Table 1 shows the official evaluation results by the organizers in the JPO Adequacy, the Pairwise Crowdsourcing Evaluation scores (Human), BLEU, RIBES, and AMFM. This year's data-driven pre-ordering gave competitive performance with last year's rule-based pre-ordering with a refined model and features, but the difference was not significant. The neural MT gave very surprising results; its baseline achieved 45% in BLEU and 85% in RIBES, both of which were much higher than our PBMT results and other good-scored phrase-based MT systems. The syntactic pre-ordering negatively affected the neural MT, resulting in about 1% lower BLEU and RIBES (less severe in AMFM). But the pre-ordering-based neural MT results was still the best in human evaluation.

We chose pre-ordering-based systems with PBMT and NMT for the official human evaluation. With respect to the human evaluation results, our neural MT was very competitive with the best-scored phrase-based MT system using external resources. Surprisingly, an un-tuned neural MT (even a state-of-the-art one) showed competitive performance with a highly tuned statistical pre-ordering MT. However, we have to keep in mind that the crowdsourcing evaluation was just based on win/lose counts against the organizers' baseline system and did not reflect all aspects of the translation quality.

5.5 Discussion

Syntactic pre-ordering achieved consistent improvements in phrase-based MT in many language pairs with large word order differences. Our results this year also suggest an advantage of pre-ordering in Chinese-to-Japanese phrase-based MT tasks. We expected that pre-ordering would also help a neural attentional MT because the attention mechanism would also be affected by word order problems. However, pre-ordering significantly decreased the evaluation scores. We do not have a solid answer yet, but one possible reason may be the consistency in the source language; pre-ordering reconstructs a source language sentence close to the target language word order for the effective phrase-based MT, but it may also introduce noise on source language structures that hurts neural MT. We actually found that pre-ordering

[1] `https://github.com/harvardnlp/seq2seq-attn` (We used the version of 08/12/2016.)
[2] They are the default network settings of the toolkit, except for bidirectionality.

System	JPO Adequacy	Pairwise	BLEU	RIBES	AMFM
PBMT w/o pre-ordering	n/a	n/a	0.3903	0.8057	0.7203
PBMT w/pre-ordering	n/a	39.250	**0.4075**	**0.8260**	**0.7302**
PBMT w/pre-ordering (2015/rule-based)	n/a	n/a	0.4060	0.8234	n/a
PBMT w/pre-ordering (2015/data-driven)	n/a	n/a	0.3977	0.8163	n/a
NMT w/o pre-ordering	n/a	n/a	**0.4499**	**0.8530**	**0.7522**
NMT w/pre-ordering	**3.44**	46.500	0.4347	0.8453	0.7493
JAPIO PBMT w/pre-ordering†	3.24	46.250	0.4432	0.8350	0.7512
NICT PBMT w/pre-ordering†	3.23	43.250	0.4187	0.8296	0.7399
NICT PBMT w/pre-ordering	n/a	36.750	0.4109	0.8270	0.7330

Table 1: Official evaluation results in JPO Adequacy, Pairwise Crowdsourcing Evaluation scores (Pairwise), BLEU, RIBES, and AMFM. Automatic evaluation scores are based on JUMAN Japanese word segmentation. Scores in **bold** are best in the same group. †: Systems used external resources.

Language	ppl. (dev)	ppl. (devtest)	ppl. (test)
Chinese	185.573	211.370	220.821
Pre-ordered Chinese	203.639	231.218	240.533

Table 2: Source-side test set perplexities on dev, devtest, and test sets by word 5-gram language models of Chinese and pre-ordered Chinese. The vocabulary size is 172,108.

increased the test set perplexity in the source language (Chinese) by about 10% (Table 2). Since this time we do not have human evaluation results of the baseline neural MT, we cannot evaluate the actual influence of pre-ordering in the neural MT for human understanding. This issue needs further analysis and investigation.

6 Conclusion

This paper presented our pre-ordering-based system for a Chinese-to-Japanese patent MT for the WAT 2016 evaluation campaign. Our results showed that pre-ordering worked effectively with a phrase-based MT but not with a neural MT. The neural MT surprisingly improved the translation performance without any careful tuning. Its result was competitive with a highly tuned phrase-based MT system.

Acknowledgments

We greatly appreciate the workshop organizers for this valuable evaluation campaign. We also thank the Japan Patent Office for providing its patent translation dataset.

References

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.

Laura Jehl, Adrià de Gispert, Mark Hopkins, and Bill Byrne. 2014. Source-side Preordering for Translation using Logistic Regression and Depth-first Branch-and-Bound Search. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 239–248.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September. Association for Computational Linguistics.

Tetsuji Nakagawa. 2015. Efficient top-down btg parsing for machine translation preordering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint*

Conference on Natural Language Processing (Volume 1: Long Papers), pages 208–218, Beijing, China, July. Association for Computational Linguistics.

Toshiaki Nakazawa, Hideya Mino, Chenchen Ding, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2016. Overview of the 3rd Workshop on Asian Translation. In *Proceeding of the 3rd Workshop on Asian Translation (WAT2016)*, Osaka, Japan, December.

Katsuhito Sudoh and Masaaki Nagata. 2015. Chinese-to-Japanese Patent Machine Translation based on Syntactic Pre-ordering for WAT 2015. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 95–98.

Katsuhito Sudoh, Jun Suzuki, Yasuhiro Akiba, Hajime Tsukada, and Masaaki Nagata. 2014. An 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.

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.