The SAS Statistical Machine Translation System for WAT 2014

Rui Wang, Xu Yang and Yan Gao
SAS Institute Inc
Beijing, China
rui.wang|xu.yang|yan.gao@sas.com

Abstract

This paper is a description of the techniques and experiment results by SAS Institute Inc in WAT 2014 evaluation campaign. We participate in two subtasks of WAT 2014: the Chinese to Japanese track and the English to Japanese track. Our baseline system is MOSES statistical machine translation toolkit. We propose syntactic reordering approaches for English to Japanese and Chinese to Japanese translation which transform the order of source sentence into the target-like order. In addition, we apply the segmentation tool in SAS® Text Miner to enhance the translation results. Several contrastive experiments are presented based on the automatic evaluation results.

1 Introduction

This paper describes the machine translation system employed by SAS Institute Inc in the 1st Workshop on Asian Translation. We participate in two subtasks in this year's WAT evaluation campaign:

1) Chinese to Japanese;
2) English to Japanese.

We apply MOSES toolkit as the baseline system.

The sentence structure of Japanese is different with that of English and Chinese. Japanese is typically a Subject-Object-Verb (SOV) language while Chinese and English are Subject-Verb-Object (SVO) languages, as illustrated in Figure 1. The statistic machine translation between Japanese and the SVO language is particularly difficult because of the long distance difference of word orders. We propose a simple syntactic reordering approach for Chinese to Japanese and English to Japanese SMT in order to transform Chinese and English into SVO languages. Our submission mainly focuses on using the syntactic approaches to improve effectively improve the translation results. In addition, unlike the alphabetic languages, Chinese and Japanese need to be segmented into words of characters before translation. The accuracy of segmentation highly influences the performance of translation for Chinese and Japanese. We apply the tokenization tool in SAS® Text Miner to the corpus and obtain improvement of the translation results.

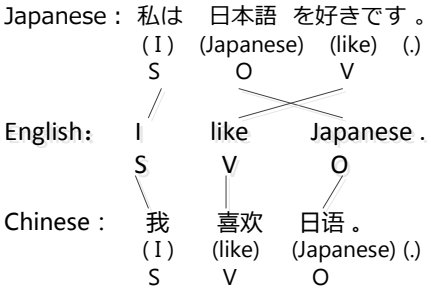


Figure 1: Word order differences of Japanese, Chinese and English.

The rest of this paper is organized as follows: in section 2 we describe the background; in section 3 we present the architecture of our system; Section 4 is the detailed description of the syntactic reordering rules we propose; experiment results are shown in section 5 and conclusion is made in section 6.

2 Background

The state of the art Statistical Machine Translation (SMT), phrase-based SMT (Koehn et al., 2003), works well on the translation between short phrases as well as on long sentences pairs with similar word orders. However the phrase-based SMT has limited capacity for long distance reordering since it does not consider the syntactic information in the translation. A great deal of recent research enhances the translation results by adding syntactic features, such syntactic-based SMT

Proceedings of the 1st Workshop on Asian Translation (WAT2014), pages 64–68,
Tokyo, Japan, 4th October 2014.

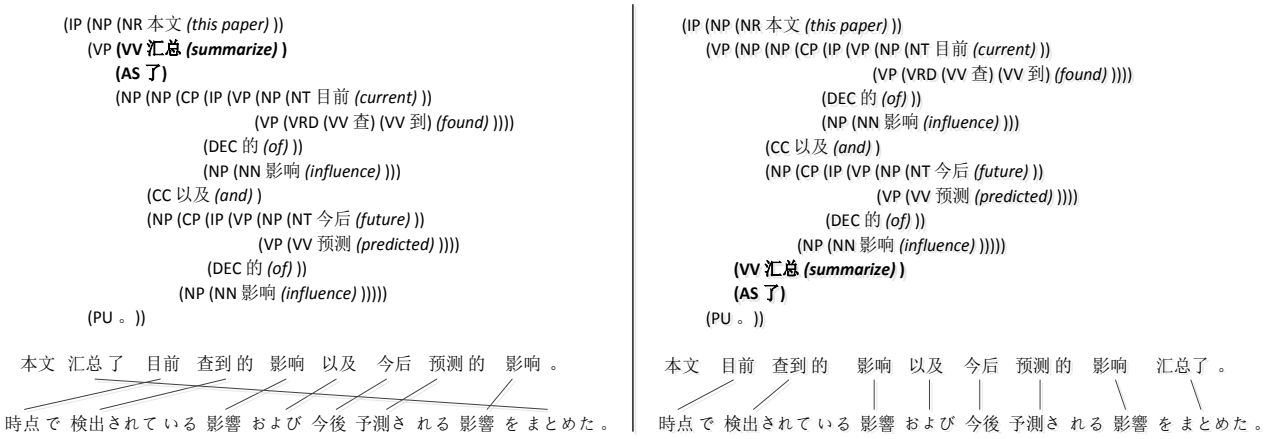
(Liu et al., 2006), or forest-based SMT (Mi et al., 2008). The language-independent methods parse the input sentences and then train reordering model from these parsed trees, such as Quirk et al. (2005), Li et al. (2007). These approaches improve the translation considerably but they are time consuming during decoding. Syntactic reordering approaches effectively improve the translation results by transforming the order of source-side language into target-like order.

Previous studies have proposed the syntactic reordering approach which is regarded as a preprocessing step on the source side language. Xia et al. (2004) automatically extract reordering rules on source side for English to French translation. Collins et al. (2005) improve the German to English translation by combining morphological and syntactical information into SMT system. Wang et al. (2007) extract the syntactic reordering rules manually on Chinese to English translation. Isozaki et al. (2010) propose a simple head finalized approach to reorder English into Japanese order based on Enju parser. We apply the head finalized method to our system and achieve improvement on the translation from English to Japanese. Dan et al. (2012) carry out a similar head finalized approach for Chinese to Japanese translation using a Chinese Enju parser (Yu et al., 2011). We propose a simple reordering methods for Chinese based on the accurate Berkeley Parser (Petrov et al., 2006).

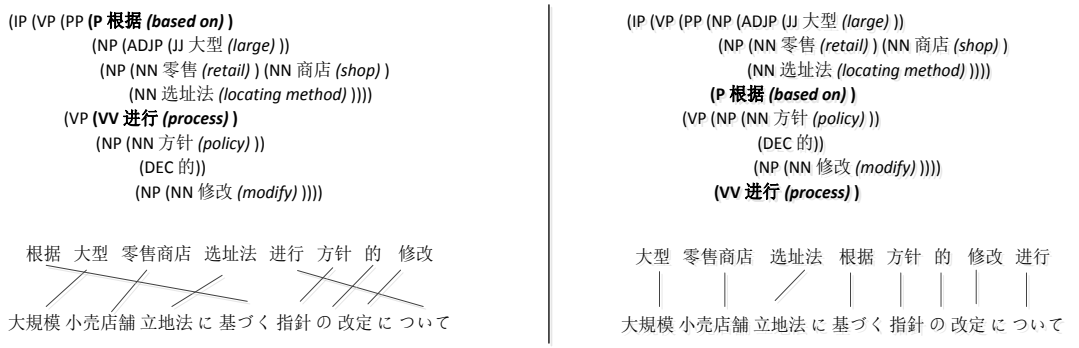
3 System Architecture

We use the open-source SMT toolkit MOSES as the baseline to train translation model and to decode the test set for submission. SRILM (Stolcke et al. 2002) toolkit is applied to train the N-gram statistical language model. All the corpus are tokenized and lowercased before training and decoding. All the submitted results are detokenized. In addition to the basic phrase-based model, we also try hierarchical model (Chiang et al., 2007) using MOSES.

We use the segementation tool in SAS® Text Miner to tokenize Japanese and Chinese corpus



(a)



(b)

for training, tuning and testing. Then we use MOSES tool to lowercase the corpus and to clean long sentences. The syntactic reordering approach is applied on source-side cleaned sentences for English and Chinese. Finally the cleaned and reordered corpus is sent to MOSES to train models and decode results.

We use Berkeley parser (Petrov et al., 2006) to parse the source side Chinese sentence into syntactic tree and develop the reorder rules based on the grammatical structure of the sentence. English sentence is parsed by Enju parser and the head finalized reorder is applied as described by Isozaki et al. (2010).

4 Syntactic Reordering Approach

4.1 Chinese to Japanese reordering

The Chinese sentence is parsed into a grammatical tree using Pen Treebank syntactic tagset as illustrated in Figure 2. The main difference between Chinese and Japanese is the position of Verb Phrase (VP) and Prepositional Phrase (PP). We develop two simple reordering rules, VP-rule and PP-rule, based on the syntactic tree.

We develop the VP-rule, 'VP (VV AS (XXX)) →VP ((XXX) VV AS)', which means to move the verb (VV) and the auxiliary word (AS) behind VV to the end of the verb phrase (VP). For example, in Figure 2 (a), 'VV 汇总 (summarize)' and 'AS 了' are moved to the end of VP. The reordered Chinese sentence has more similar order with the Japanese translation than the original one.

The Preposition (P) is always at the end the Prepositional Phrase (PP) in Japanese while in Chinese P is located at the beginning of PP. We imply the PP-rule, 'PP (P (XXX)) → PP ((XXX) P)', to move P to the end of PP. In Figure 2 (b), the preposition '根据 (based on)' is move to the end of PP. The reordered sentence reduce the long distance order differences between the source Chinese and the target Japanese.

4.2 English to Japanese reordering

For English to Japanese translation, we use the head finalization approach proposed by Isozaki et al. (2010). In Japanese, the syntactic head word is located behind the dependent words while English is the opposite. The dependency parser, Enju parser, is applied on the source side which can output syntactic head, as illustrated in Figure 3. A simple reorder rule is defined: to move the syntactic head to the end of non-head words. In Figure 3, node C3 is the head of node C2. According to the head finalized reordering rule, C3 should be moved behind the other child of C2, C4. After reordering, the English sentence has the similar order with the Japanese sentence.

The head finalized reorder rule will reorder the coordination expressions such as, 'A and B', into 'B and A'. We add a coordination exception rule to stop reordering at coordination nodes.

5 Experiments

The experiment is conducted on the training data from ASPEC, consists of approximately 3 million

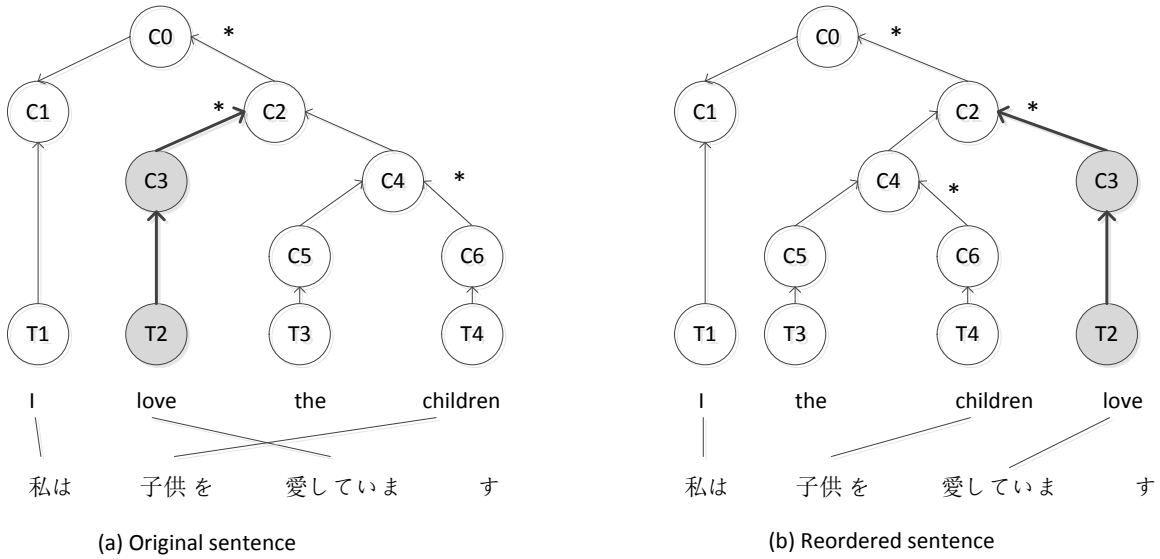


Figure 3: Example of reordering rules for English to Japanese translation. * represents the head of the node. (a) is the original sentence; (b) is the reordered sentence by head finalized rule.

Japanese-English parallel sentences and approximately 0.7 million Japanese-Chinese. BLEU and RIBES are used as evaluation metrics. On the evaluation websites, for Chinese to Japanese and English to Japanese evaluation, the results of BLEU and RIBES are measured under three Japanese segmentation tools: Juman, Kytea and Mecab. For the sake of brevity, we calculate the average score of the three segmentation tools for each evaluation metric.

5.1 Effect of the segmentation tool of SAS® Text Miner

We use phrase-based model as baseline for the Chinese to Japanese track to evaluate the effect of SAS segmentation tool. The result of baseline is provided by the WAT organizer (Toshiaki, et al. 2014). Table 1 shows that both the BLEU and RIBES are increased by using the segmentation tool of SAS® Text Miner for Chinese and Japanese corpus compared with the Juman segmentation tool for Japanese and Stanford Word Segmenter for Chinese. We conduct the other experiments using the segmentation tool of SAS® Text Miner because of its good performance.

	BLEU	**RIBES**
Baseline	34.86	0.769962
SAS segmentation	**35.31**	**0.809631**

Table 1: Effect of the segmentation tool of *SAS® Text Miner*.

5.2 Chinese to Japanese translation

	BLEU	**RIBES**
Baseline	34.86	0.769962
Baseline+VP	36.19	0.826146
Baseline+PP	36.30	0.815694
Baseline+PP+VP	36.40	0.826015
Hierarchical	36.06	0.814207
Hierarchical+PP+VP	**37.38**	**0.830909**

Table 2: Effect of reordering rules for Chinese to Japanese translation.

Table 2 shows the effect of reordering rules for Chinese to Japanese translation. The baseline is the results from the WAT organizer (Toshiaki, et al. 2014). We propose two reordering rules: VP rule and PP rule. The baseline is the phrase-based machine translation using MOSES. We apply the VP rule, PP rule and the combination of the two rules to the corpus separately and send the reordered sentences to the baseline system. Both the VP rule and PP rule have contribution to the improvement of BLEU scores. We also apply the reorder rules to the hierarchical model and gain 2.07 in BLEU scores compared with the baseline system.

5.3 English to Japanese translation

The effect of the head finalized reordering rule for English to Japanese translation is reported in Table 3. The phrase-based model is applied as baseline system. After head finalized reordering, the BLEU score is increased from 28.52 by the organizer (Toshiaki, et al. 2014) to 31.09. We gain 3.13 in BLEU score by applying the hierarchical model with pre-reordering.

	BLEU	**RIBES**
Baseline	28.52	0.690350
Baseline+reorder	31.09	0.765005
Hierarchical	31.23	0.743135
Hierarchical+reorder	**31.65**	**0.767323**

Table 3: Effect of the head finalized reordering rule for English to Japanese translation.

6 Conclusion

This paper describes the techniques and experiment results in WAT 2014 evaluation campaign submitted by SAS Institute Inc. By applying the tokenization tool in SAS® Text Miner and the syntactic reordering rules, we achieve significant improvements on Chinese to Japanese translation and English to Japanese translation. We also report some extensive experiment results to illustrate the contribution on different parts.

In the future, we intend to consider case markers in Japanese which other languages do not have. Since the form of Japanese case markers is known, we may define rules to insert particular tags to the source languages. We expect improvements of translation by adding the rules of case markers in Japanese.

Reference

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. "Statistical phrase-based translation." *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics.

Liu, Yang, Qun Liu, and Shouxun Lin. 2006. "Tree-to-string alignment template for statistical machine translation." *Proceedings of the 21st International Conference on Computational Linguistics and the*

44th annual meeting of the Association for Computational Linguistics.

Mi, Haitao, Liang Huang, and Qun Liu. 2008. "Forest-Based Translation." *ACL*.

Quirk, Chris, Arul Menezes, and Colin Cherry. 2005. "Dependency treelet translation: Syntactically informed phrasal SMT." *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.*

Li, Chi-Ho, et al. 2007. "A probabilistic approach to syntax-based reordering for statistical machine translation." *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Vol. 45. No. 1*.

Xia, Fei, and Michael McCord. 2004. "Improving a statistical MT system with automatically learned rewrite patterns." *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics.

Collins, Michael, Philipp Koehn, and Ivona Kučerová. 2005. "Clause restructuring for statistical machine translation." *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics.

Wang, Chao, Michael Collins, and Philipp Koehn. 2007. "Chinese Syntactic Reordering for Statistical Machine Translation." *EMNLP-CoNLL*.

Isozaki, Hideki, et al. 2010. "Head finalization: A simple reordering rule for sov languages." *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Association for Computational Linguistics.

Dan, Han, et al. 2012. "Head finalization reordering for Chinese-to-Japanese machine translation." *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics.

Yu, Kun, et al. 2011. "Analysis of the difficulties in Chinese deep parsing."*Proceedings of the 12th International Conference on Parsing Technologies*. Association for Computational Linguistics.

Stolcke, Andreas. 2002. "SRILM-an extensible language modeling toolkit."*INTERSPEECH*.

Chiang, David. 2007. "Hierarchical phrase-based translation." *computational linguistics* 33.2: 201-228.

Petrov, Slav, et al. 2006. "Learning accurate, compact, and interpretable tree annotation." *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Toshiaki Nakazawa, et al. 2014. "Overview of the 1st Workshop on Asian Translation". Proceedings of the 1st Workshop on Asian Translation (WAT2014)