Toshiba MT System Description for the WAT2015 Workshop

Satoshi Sonoh and Satoshi Kinoshita

Knowledge Media Laboratory, Corporate Research & Development Center,
Toshiba Corporation.
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki 212-8582, JAPAN
{satoshi.sonoo,satoshi.kinoshita}@toshiba.co.jp

Abstract

This paper provides the system description of Toshiba Machine Translation System for the 2nd Workshop on Asian Translation (WAT2015). We participated in all tasks that consist of "scientific papers subtask" and "patents subtask". We submitted statistically post edited translation (SPE) results based on our rule based translation system and SMT for each language pair. In addition, we submitted system combination results between SPE and SMT with a recurrent neural language model (RNNLM). In experimental results, the system combination achieved higher BLEU scores than single system with reranking. We also obtained improvements in Chinese translation in crowdsourcing evaluations.

1 Introduction

Recently, statistical machine translation (SMT) has been broadly developed and successfully used in the portion of practicable systems. However, it is costly to make a large volume of parallel corpora in a wide range of domains for commercial use. For this reason, we have developed rule based machine translation (RBMT) system using a monolingual corpus in the target language. For example, target word selection is possible based on co-occurrence relationship extracted from a monolingual corpus (Suzuki et al., 2005). Furthermore, we have developed a word sense disambiguation based on a monolingual corpus in the target domain, and it has been applied to Japanese-Korean and Korean-Japanese translation systems (Kumano 2013, Tanaka et al., 2014). On the other hand, open Asian parallel corpora including ASPEC [1], NTCIR PatentMT[2] and JPO Patent Corpus[3] are available for the research of machine translation systems. By using the parallel corpora, we have confirmed advantages which apply statistical post editing (SPE) to RBMT in domain adaptation (Suzuki, 2011).

In the last workshop (Nakazawa et al., 2014), we participated in Japanese-English and Japanese-Chinese tasks with SPE approach and obtained higher evaluation results than RBMT. Meanwhile, RBMT showed better performance than SPE in the direct and relative comparison (Sonoh et al., 2014). In this workshop (WAT2015), we participated in all tasks including Japanese-English (ja-en), English-Japanese (en-ja), Japanese-Chinese (ja-zh) and Chinese-Japanese (zh-ja) for "scientific paper subtask", and Chinese-Japanese (JPCzh-ja) and Korean-Japanese (JPCko-ja) for "patents subtask". Patents subtask is newly added, and its parallel corpus has 4 sections (Chemistry, Electricity, Mechanical Engineering and Physics).

In all the tasks, we submitted SPE translation results based on our RBMT and SMT. In addition, we submitted system combination results between SPE and SMT with recurrent neural language model (RNNLM; Mikolov el al., 2010).

Section 2 and 3 describe the overview of our systems and some pre/post processing. The experimental results and official results are shown in Section 4 and 5. The analysis for the official results is discussed in Section 6 and finally, Section 7 concludes this paper. As for a context-aware translation, the description was omitted because our baseline system is the same as the last workshop (see Sonoh et al., 2014).

[1] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
[2] http://research.nii.ac.jp/ntcir/permission/ntcir-10/perm-en-PatentMT.html
[3] http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/

Proceedings of the 2nd Workshop on Asian Translation (WAT2015), pages 48–53,
Kyoto, Japan, 16th October 2015.

2 Overview of Toshiba System

2.1 RBMT System

Our RBMT system is basically a transfer-based machine translation (Izuha et al., 2008). The core framework consists of morphological analysis, syntactic/semantic analysis, target word selection, structural transfer, syntactic generation and morphological generation. Furthermore, huge amount of rules as translation knowledge including word dictionaries can realize both high translation performance and flexibility of customization. As for Japanese-Korean translation, syntactic analysis and transfer are omitted because the languages are grammatically similar.

2.2 Statistical Post Editing

SPE using phrase-based SMT has been proposed and it is an efficient framework which is able to adapt translation output to target domains (Michel et al., 2007).

We first translated source sentences of training data in ASPEC and JPO Patent Corpus by RBMT. Then we trained phrase-based model between translated sentences and reference sentences using Moses toolkit (Kohen et al., 2007). In the training, we used 1M sentences for ja-en, en-ja, JPCzh-ja and JPCko-ja, 0.67M for ja-zh and zh-ja in the training data. Japanese sentences were tokenized by JUMAN[4], and Moses tokenizer for English, and Kytea (Neubig et. al, 2011) for Chinese. We also trained 5-gram language models using KenLM (Heafield et al., 2013). In tuning and decoding, we set distortion limit to 0 for JPOko-ja in consideration of grammatical similarity and 6 for other language pairs.

2.3 System Combination using RNNLM

Although both SPE and SMT are based on a statistical model from the given corpora, they generate different translation candidates because SPE has some features from RBMT. If a better system can be selected from the candidates in each translation, we can get a better translation result.

Thus, we realized a system combination between SPE and SMT as n-best reranking using a RNNLM. The n-best reranking can be achieved using both basic features and RNNLM score. In tuning, we combined 100-best candidates of both SPE and SMT for dev-set, and ran MERT tuning by adding the RNNLM score to the basic features. In decoding, we re-ranked combined candidates by product-sum of the features including RNNLM score and tuned weights.

For ja-en, en-ja, ja-zh and zh-ja, we used RNNLMs trained by the first 500k sentences in the training data of ASPEC. For JPOzh-ja and JPOko-ja, we used 500k sentences which were evenly extracted from 4 sections in JPO Patent Corpus. All RNNLMs were trained with 500 hidden layers and 50 classes by RNNLM toolkit[5].

3 Tuning RBMT and pre/post-processing

3.1 Technical Term Dictionaries

As the preparation for each task, we selected technical term dictionaries by the same principle in the last workshop (Sonoh el al., 2014). For JPOzh-ja, we used an additional patent dictionary, which is extracted from JPO Chinese-Japanese dictionary[6]. Furthermore, for JPOko-ja, we used n-gram probability dictionary, which was made from monolingual patent resources, in order to resolve word sense disambiguation (Tanaka et al., 2014).

3.2 English Word Correction

To improve translation of sentences including misspelled words in English, we applied correction processing based on an edited distance. We replaced the word considered as misspelling with a word which had the smallest edited distance in the training data. However, because SMT and SPE basically have robustness to the misspelling, we confined words to be replaced to words which remain as unknown words in SMT and SPE results.

3.3 Japanese KATAKANA Normalization

In the case where a target language is Japanese; we applied normalization of KATAKANA notation. In advance of translation, we counted the frequency of KATAKANA notation, which has fluctuations of prolonged sound mark, in the

[4] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

[5] http://www.fit.vutbr.cz/~imikolov/rnnlm/

[6] https://alaginrc.nict.go.jp/resources/jpo-info/jpo-outline.html#jpo-dic-zh

Table 1: Overall BLEU and RIBES scores for "scientific papers subtask".

System	Rerank	ja-en		en-ja		ja-zh		zh-ja	
		BLEU	RIBES	BLEU	RIBES	BLEU	RIBES	BLEU	RIBES
RBMT	No	15.31	0.677	14.78	0.685	19.51	0.767	15.39	0.723
SMT	No	17.41	0.620	25.17	0.642	28.20	**0.810**	36.34	0.815
	Yes	17.85	0.619	25.37	0.643	28.46	0.809	36.69	0.815
SPE	No	22.65	0.717	31.10	0.767	29.48	0.809	35.76	0.822
	Yes	22.92	**0.718**	31.73	**0.770**	29.49	0.809	36.06	0.823
COMB	Yes	**23.00**	0.716	**31.82**	**0.770**	**29.60**	**0.810**	**37.47**	**0.827**

Table 2: Overall BLEU and RIBES scores for "patents subtask".

System	Rerank	JPOzh-ja		JPOko-ja	
		BLEU	RIBES	BLEU	RIBES
RBMT	No	25.81	0.764	51.28	0.902
SMT	No	38.77	0.802	70.17	0.943
	Yes	39.18	0.805	**70.89**	**0.944**
SPE	No	39.01	**0.813**	68.47	0.940
	Yes	39.30	0.811	68.76	0.940
COMB	Yes	**40.23**	**0.813**	70.40	0.942

target sentences of the training data. In the translation results, KATAKANA fluctuations were replaced with those of highly-frequent notations, such as "from スクリュ to スクリュー" and "from サーバー to サーバ". By applying normalization, we got improvements of about 0.5 BLEU in RBMT.

Furthermore, we replaced the ideographic comma "、" in number expression with a normal comma "," for translation results in Japanese.

3.4 Other Post Processing

In order to reduce unknown words in SMT, we applied RBMT to SMT results. For example, in ja-zh, we translated KATAKANA words, which remain in SMT results, into Chinese or English words, if the words were found in RBMT dictionaries. Also, Hangul words in SMT results of JPOko-ja were translated into Japanese words.

4 Experimental Results

This section shows experimental results of our translation systems.

Table 1 and 2 show the overall BLEU and RIBES scores for "scientific papers subtask" and "patents subtask", respectively. COMB means results of the system combination and Rerank means results of reranking using RNNLM (100-best for SMT and SPE, 200-best for COMB). In all tasks, SPE improves translation results of RBMT on the BLEU and RIBES. In tasks except JPOko-ja, SPE achieves performance equal to or better than phrase-based SMT.

Moreover, in most tasks, Rerank improves about 0.3-0.5 BLEU score, and COMB shows better performance than other systems. In JPOko-ja, SMT, SPE and COMB show very high performances which are close to 70 BLEU, and SMT with reranking achieves the highest BLEU and RIBES scores. In ja-en, en-ja, ja-zh and zh-ja, more than half of translations selected from SPE and the others selected from SMT. In particular SPE accounted for about 80% translations in ja-en, en-ja and zh-ja. On the other hand, more than half of translations selected from SMT in JPOzh-ja and JPOko-ja. Table 3 shows the translation examples that COMB achieves better results than SPE with reranking in sentence-level BLEU.

Finally, we compared between phrase-based model and hierarchical phrase-based model. Table 4 shows comparison in ja-zh task. In all systems including SPE, hierarchical phrase-based model improves about 0.4 BLEU. We applied hierarchical phrase-based model to ja-zh only, because significant improvements were not confirmed in other language pairs.

Table 3: Translation examples indicating that COMB achieves better results than SPE in sentece-level BLEU.

ja-en	SRC	揺動時に比べて，発電量は40倍である。
	REF	In comparison with the fluctuation, the electric power generation is the 40 twice.
	SPE	Compared with the time of rocking, production of electricity is 40 times.
	COMB	Compared with the time of fluctuations, the electric power generation is 40 times.
en-ja	SRC	SiO2 films showed excellent performance even at 430℃ or less, and the memory effect of Si dot MOS capacitor was confirmed.
	REF	SiO2膜は，430℃以下でも優れた性能を示し，SiドットMOSコンデンサのメモリ効果を確認した。
	SPE	SiO2膜は430℃でも優れた性能を示す以下であり，SiドットMOSキャパシタの記憶効果が確認された。
	COMB	SiO2膜は430℃でも優れた性能を示し，以下，SiドットMOSキャパシタの記憶効果を確認した。
ja-zh	SRC	この擾乱からの回復についても考察した。
	REF	对这种干扰的恢复也进行了考察。
	SPE	考察了从该干扰恢复。
	COMB	在这种干扰的恢复也进行了考察。
zh-ja	SRC	还对在完成来所登记之前的各个环节进行了介绍。
	REF	来所登録が完了するまでの流れ等も紹介した。
	SPE	登録の完了までの各段階について紹介した。
	COMB	登録が完了するまでの各段階について紹介した。
JPO zh-ja	SRC	在固体中加入 BHT，混合物在丙酮中溶解。
	REF	BHTを固体に加え、混合物をアセトンに溶解する。
	SPE	固体BHTを加え、この混合物は、アセトンに溶解した。
	COMB	固体BHTを加え、混合物をアセトンに溶解した。
JPO ko-ja	SRC	원재료 필름(1)에서의 비접합부의 적어도 일부를 덮도록 배치되어도 좋다.
	REF	原反フィルム1における非接合部の少なくとも一部を覆うように配置されてもよい 。
	SPE	原反フィルム1での非接合部の少なくとも一部を覆うように配置されてもよい。
	COMB	原反フィルム1における非接合部の少なくとも一部を覆うように配置されてもよい。

Table 4: A Comparison of Phrase-based Model.

System	hierarchical	ja-zh	
		BLEU	RIBES
SMT	No	28.46	0.809
	Yes	29.82	0.810
SPE	No	29.49	0.809
	Yes	29.89	0.809
COMB	No	29.60	0.810
	Yes	**30.07**	**0.817**

5 Official Results

This section shows official results of our translation systems.

We basically submitted two results, one is SPE[7] and the other is the system combination between SPE and SMT. Furthermore, top two systems on the BLEU scores were evaluated by the crowdsourcing. In the crowdsourcing evaluation, pair-wise evaluation against the baseline system (phrase-based SMT) was performed by 5 evaluators, and HUMAN score was calculated

[7] In JPOko-ja, because SMT showed higher BLEU score than SPE, we submitted SMT result.

Table 5: Overall official results for "scientific papers subtask". B, R and H mean BLEU, RIBES, HUMAN, respectively. HUMAN was evaluated by 5 evaluators using crowdsorcing.

System	ja-en			en-ja			ja-zh			zh-ja		
	B	R	H	B	R	H	B	R	H	B	R	H
SPE	22.89	0.719	**25.00**	32.06	0.771	40.25	30.17	0.813	2.50	35.85	0.825	-1.00
COMB	23.00	0.716	21.25	31.82	0.770	-	30.07	0.817	**17.00**	37.47	0.827	**18.00**

Table 6: Overall official results for "patents subtask".

System	JPOzh-ja			JPOko-ja		
	B	R	H	B	R	H
SMT	-	-	-	71.01	0.944	4.50
SPE	41.12	0.822	**24.25**	-	-	-
COMB	41.82	0.821	14.50	70.51	0.942	3.00

based on majority voting (Nakazawa et al., 2014).

In WAT2015 results (Nakazawa et al., 2015), we note that Toshiba systems were ranked as one of the top three systems in human evaluation in ja-en, ja-zh and JPOzh-ja. Especially, ja-zh achieved the highest score although the BLEU score is lower than other systems. On the other hand, as for JPOko-ja, we got a comparatively high BLEU score, but were disappointed by its low HUMAN score.

Table 5 and 6 are the overall official results for each task, respectively. In ja-zh and zh-ja, COMB shows higher HUMAN score than SPE. On the other hand, SPE or SMT is higher than COMB in ja-en, JPOzh-ja and JPOko-ja. These results indicate that the system combination improves human evaluation of Chinese translation in the scientific documents, at least. We guess that the system combination between equivalent systems achieves complementary translation to improve human evaluations. For example, BLEU scores of SPE and SMT are nearly equal in ja-zh and zh-ja (shown in Table 1).

6 Discussion

On receiving the crowdsourcing results, we analyzed differences between our system and Online A, which obtained the highest HUMAN score in JPOko-ja. Table 7 shows the comparison between our system (COMB) and Online A. Here, 'Baseline' column is the HUMAN score in the result of crowdsourcing (official results) and the other was evaluated by inner evaluators. The inner evaluation was conducted excluding expressional differences as described in detail below. Although Online A achieves a very high HUMAN score to the baseline system, superior results of COMB over Online A are shown in the pair-wise evaluation.

We hypothesize that the significant difference between the crowdsourcing and the inner evaluators occurs from the evaluation of the number expressions, such as "システム(100)" and "システム 100". In the training data of JPOko-ja, a lot of brackets of numbers in the source sentences disappear in the target sentences. Thus, brackets are dropped in SPE and SMT. As for well-translated target sentences such as JPOko-ja, it is possible that evaluators in the crowdsourcing judged faithful translation as better by focusing on existence of brackets.

Table 7: The relationship between automatic evaluations and human evaluations.

	BLEU	RIBES	HUMAN		
			Baseline	COMB	Online A
COMB	70.51	0.94	3.00	-	10.75
Online A	55.05	0.91	38.75	-10.75	-

7 Conclusion

The overview of Toshiba machine translation systems, which applied the statistical post editing and the system combination with RNNLM, is described in this paper.

SPE and reranking with RNNLM achieved higher BLEU than phrase-based SMT in most language pairs. Furthermore, the system combination between SPE and SMT improved BLEU score in Japanese-English pair and Japanese-Chinese pair. In the other hand, a straightforward correlation between automatic evaluation

and human evaluation is not confirmed in our system. We need to establish the combination of multi-systems for practical use purpose, taking advantage of their characteristics and qualities.

Reference

Hirokazu Suzuki and Akira Kumano. 2005. Learning Translations from Monolingual Corpora. In *Proc. of MT Summit X*.Akira Kumano. 2013. Korean Translation System for Patent Documents. In *Japio YEAR BOOK 2013*, pages 298-301. (In Japanese)

Hiroyuki Tanaka, Satoshi Sonoh, Satoshi Kinoshita and Satoshi Kamatani. 2014. Improvement of Korean machine translation using statistical word selection. In *Proc. of the 3rd Symposium on Patent Information Processing*, pages 97-100. (In Japanese)

Hirokazu Suzuki. 2011. Automatic Post-Editing based on SMT and its selective application by Sentence-Level Automatic Quality Evaluation. In *Proc. of MT Summit XIII*.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kurohashi and Eiichiro Sumita. 2014. Overview of the 1st Workshop on Asian Translation. In *Proc. of the 1st Workshop on Asian Translation (WAT2014).*

Satoshi Sonoh, Satoshi Kinoshita, Hiroyuki Tanaka and Satoshi Kamatani. 2014. Toshiba MT System Description for the WAT2014 Workshop. In *Proc. of the 1st Workshop on Asian Translation (WAT2014).*

Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proc. of INTERSPEECH 2010* pages 1045–1048.

Tatsuya Izuha, Akira Kumano and Yuka Kuroda. 2008. Toshiba Rule-Based Translation System at NTCIR-7 PAT MT. In *Proc. of NTCIR-7 Workshop Meeting*, pages 430-434.

Michel Simard, Cyril Goutte and Pierre Isabell. 2007. Statistical Phrase-based Post-editing. In *Proc. of NAACL HLT 2007, ACL,* pages 508-515.

Philipp Kohen, Marcell Federuci, Brooke Cowan, Richard Zens, Chris Dyer, Ondrej Bojar Alexandra Constantin and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of the ACL,* pages 177-180.

Graham Neubig, Yosuke Nakata and Shinsuke Mori. 2011. Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis. In *Proc. of ACL-HLT*.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proc. of the ACL*, pages 690–696.

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