Consistent Improvement in Translation Quality of Chinese–Japanese Technical Texts by Adding Additional Quasi-parallel Training Data

Wei Yang
Graduate School of IPS
Waseda University
2-7 Hibikino, Wakamatsu-ku
Kitakyushu, Fukuoka 808-0135, Japan
kevinyoogi@akane.waseda.ja

Yves Lepage
Graduate School of IPS
Waseda University
2-7 Hibikino, Wakamatsu-ku
Kitakyushu, Fukuoka 808-0135, Japan
yves.lepage@waseda.ja

Abstract

Bilingual parallel corpora are an extremely important resource as they are typically used in data-driven machine translation. There already exist many freely available corpora for European languages, but almost none between Chinese and Japanese. The constitution of large bilingual corpora is a problem for less documented language pairs. We construct a quasi-parallel corpus automatically by using analogical associations based on certain number of parallel corpus and a small number of monolingual data. Furthermore, in SMT experiments performed on Chinese-Japanese, by adding this kind of data into the baseline training corpus, on the same test set, the evaluation scores of the translation results we obtained were significantly or slightly improved over the baseline systems.

1 Introduction

Bilingual corpora are an essential resource for current SMT. So as to enlarge such corpora, technology research has been done in extracting parallel sentences from existing non-parallel corpora. The approaches and difficulties depend on the parallelness of the given bilingual parallel corpus. Fung and Cheung (2004) give a detailed description of the types of non-parallel corpora. They proposed a completely unsupervised method for mining parallel sentences from quasi-comparable bilingual texts which include both in-topic and off-topic documents. Chu et al. (2013) proposed a novel method of classifier training and testing that simulates the real parallel sentence extraction process. They used linguistic knowledge of Chinese character features. Their approach improved in several aspects and worked well for extracting parallel sentences from quasi–comparable corpora. Their experimental results on parallel sentence extraction from quasi–comparable corpora indicated that their proposed system performs significantly better than previous studies.

There also exist some works on extracting parallel parallel sentences from comparable corpora, such as Wikipedia. Smith et al. (2010) include features which make use of the additional annotation given by Wikipedia, and features using an automatically induced lexicon model.

In this paper, we propose to construct a bilingual corpus of quasi-parallel sentences automatically. This is different from parallel or comparable or quasi-comparable corpora. A quasi-parallel corpus contains aligned sentence pairs that are translations to each other to a certain extent. The method relies on a certain number of existing parallel sentences and a small number of unaligned, unrelated, monolingual sentences. To construct the quasi-parallel corpus, analogical associations captured by analogical clusters are used. The motivation is that the construction of large bilingual corpora is a problem for less-resourced language pairs, but it is to be noticed that the monolingual data are easier to access in large amounts. The languages that we tackle in this paper are: Chinese and Japanese.

Our approach leverages Chinese and Japanese monolingual data collected from the Web by clustering and grouping these sentences using analogical associations. Our clusters can be considered as rewriting models for new sentence generation. We generate new sentences using these rewriting models starting from seed sentences from the monolingual part of the existing parallel corpus we used, and filter out dubious newly over-generated sentences. Finally, we extract newly generated sentences and assess the strength of translation relations between them based on the similarity, across languages, between the clusters they were generated from.

Proceedings of the 1st Workshop on Asian Translation (WAT2014), pages 69–76,
Tokyo, Japan, 4th October 2014.

2 Chinese and Japanese Linguistic Resources

2.1 Chinese and Japanese Parallel Sentences

The Chinese and Japanese linguistic resources we use in this paper are the ASPEC-JC[1] corpus. It is a parallel corpus consisting of Japanese scientific papers from the reference database and electronic journal site J-STAGE of the Japan Science and Technology Agency (JST) that have been translated to Chinese after receiving permission from the necessary academic associations. The parts selected were abstracts and paragraph units from the body text, as these contain the highest overall vocabulary coverage.

This corpus is designed for Machine Translation and is split as below (some statistics are given in Table 1):

- Training Data: 672,315 sentences;
- Development Data: 2,090 sentences;
- Development-Test Data: 2,148 sentences;
- Test Data: 2,107 sentences.

For new sentence generation from the training data, we extracted 103,629 Chinese-Japanese parallel sentences with less than 30 characters in length. We propose to make use of this part of data as seed sentences for new sentence generation in both languages, then deduce and construct a Chinese—Japanese quasi-parallel corpus that we will use as additional data to inflate the baseline training corpus.

2.2 Chinese and Japanese Monolingual Sentences

To generate new quasi-parallel data, we also use unrelated unaligned monolingual data. We collected monolingual Chinese and Japanese short sentences with less than 30 characters in size from the Web using an in-house Web-crawler, mainly from the following websites: “Yahoo China”, “Yahoo China News”, “douban” for Chinese and “Yahoo! JAPAN”, “Mainichi Japan” for Japanese. Table 2 gives the statistics of the cleaned 70,000 monolingual data that we used in the experiments.

3 Constructing Analogical Clusters According to Proportional Analogies

3.1 Proportional Analogies

Proportional analogies establish a structural relationship between four objects, A, B, C and D: ‘A is to B as C is to D’. An efficient algorithm for the resolution of analogical equations between strings of characters has been proposed in (Lepage, 1998).

The algorithm relies on counting numbers of occurrences of characters and computing edit distances (with only insertion and deletion as edit operations) between strings of characters ($d(A,B) = d(C,D)$ and $d(A,C) = d(B,D)$). The algorithm uses fast bit string operations and distance computation (Allison and Dix, 1986).

3.1.1 Sentential Analogies

We gather pairs of sentences that constitute proportional analogies, independently in Chinese and Japanese. For instance, the two following pairs of Japanese sentences are said to form an analogy, because the edit distance between the sentence pair on the left of ‘::’ is the same as between the sentence pair on the right side: $d(A,B) = d(C,D) = 13$ and $d(A,C) = d(B,D) = 5$, and the relation on the number of occurrences of characters, which must be valid for each character, may be illustrated as follows for the character 茶: 1 (in A) - 1 (in B) = 0 (in C) - 0 (in D). We call any such two pairs of sentences a *sentential analogy*.

紅茶が飲みたい。	:	あなたは紅茶が好きですか。	::	ビールが飲みたい。	:	あなたはビールが好きですか。
I'd like a cup of black tea.	:	*Do you like black tea?*	::	*I'd like a beer.*	:	*Do you like beer?*

3.1.2 Analogical Cluster

When several sentential analogies involve the same pairs of sentences, they form a series of analogous sentences, and they can be written on a sequence of lines where each line contains one sentence pair and where any two pairs of sentences from the sequence of lines forms a *sentential analogy*. We call such a sequence of lines an *analogical cluster*. The size of a cluster is the number of its sentential pairs. The following example in Japanese shows three possible sentential analogies and the size of the cluster is 3. English translation is given below.

[1] http://orchid.kuee.kyoto-u.ac.jp/ASPEC/

ASPEC-JC	Language	# of different sentences	size of sentences in characters			total characters	total words	voc. size
			mean	±	std.dev.			
	Chinese	668,942	46.93	±	26.62	31,462,440	18,847,514	295,580
	Japanese	666,938	59.69	±	32.05	39,987,827	23,480,703	145,074

Table 1: Statistics on the ASPEC Chinese–Japanese corpus used for training (672,315 sentences). Segmentation tools: urheen for Chinese and mecab for Japanese.

	# of different sentences (cleaned)	size of sentences in characters (mean ± std.dev.)			total characters	total words
Chinese	70,000	10.29	±	6.21	775,530	525,462
Japanese	70,000	15.06	±	6.34	1,139,588	765,085

Table 2: Statistics on the cleaned Chinese and Japanese monolingual short sentences. Segmentation tools: urheen for Chinese and mecab for Japanese.

紅茶が飲みたい。	:	あなたは紅茶が好きですか。
ビールが飲みたい。	:	あなたはビールが好きですか。
ジュースが飲みたい。	:	あなたはジュースが好きですか。
I'd like a cup of black tea.	:	*Do you like black tea?*
I'd like a beer.	:	*Do you like beer?*
I'd like some juice.	:	*Do you like juice?*

As we will see in Section 4, analogical clusters can be considered as *rewriting models*. New sentences can be generated using them.

3.2 Experiments on clusters production

In each language, independently, we also construct analogical clusters from the unrelated monolingual data. The number of unique sentences used is 70,000 for both languages. Table 3 summarizes some statistics on the clusters produced.

	Chinese	Japanese
# of different sentences	70,000	70,000
# of clusters	23,182	21,975

Table 3: Statistics on the Chinese and Japanese clusters constructed from our unrelated monolingual data independently in each language.

3.3 Determining corresponding clusters by computing similarity

The steps for determining corresponding clusters are,

- First, for each sentence pair in a cluster, we extract the change between the left and the right sides by finding the longest common subsequence (LCS) (Wagner and Fischer, 1974).

- Then, we consider the changes between the left (S_{left}) and the right (S_{right}) sides in one cluster as two sets. We perform word segmentation[2] on these changes in sets to obtain minimal sets of changes made up with words or characters.

- Finally, we compute the similarity between the left sets (S_{left}) and the right sets (S_{right}) of Chinese and Japanese clusters. To this end, we make use of the EDR dictionary[3] and word-to-word alignments (based on ASPEC-JC data using Anymalign[4]), We keep 72,610 word-to-word correspondences obtained with

[2] Segmentation toolkits: Mecab, Part-of-Speech and Morphological Analyzer: `http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html` for Japanese and Urheen, a Chinese lexical analysis toolkit (National Laboratory of Pattern Recognition, China) for Chinese.

[3] `http://www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html`

[4] `http://anymalign.limsi.fr`

Anymalign in 1 hour after filtering on both translation probabilities with a threshold of 0.3, the quality of these word-to-word correspondences is about 96%. We also use a traditional-simplified Chinese variant table[5] and Kanji-Hanzi Conversion Table[6] to translate all Japanese words into Chinese, or convert Japanese characters into simplified Chinese characters. We calculate the similarity between two Chinese and Japanese word sets according to a classical Dice formula:

$$Sim = \frac{2 \times |S_{zh} \cap S_{ja}|}{|S_{zh}| + |S_{ja}|} \quad (1)$$

Here, S_{zh} and S_{ja} denote the minimal sets of changes across the clusters (both on the left or right) in both languages (after translation and conversion). To compute the similarity between two Chinese and Japanese clusters we take the arithmetic mean on both sides, as given in formula (2):

$$Sim_{C_{zh}-C_{ja}} = \frac{1}{2}(Sim_{left} + Sim_{right}) \quad (2)$$

We set different thresholds for $Sim_{C_{\mathrm{zh}}-C_{\mathrm{ja}}}$ and check the correspondence between these extracted clusters by sampling. Where the $Sim_{C_{\mathrm{zh}}-C_{\mathrm{ja}}}$ threshold is set to 0.300, the acceptability of the correspondence between the extracted clusters reaches 78%. About 15,710 corresponding clusters were extracted ($Sim_{C_{\mathrm{zh}}-C_{\mathrm{ja}}} \geq 0.300$) by the above steps.

4 Generating New Sentences Using Analogical Associations

4.1 Generation of New Sentences

Analogy is not only a structural relationship. It is also a process (Itkonen, 2005) by which, "given two related forms and only one form, the fourth missing form is coined" (de Saussure, 1916). If the objects *A*, *B*, *C* are given, we may obtain an other unknown object *D* according to the analogical equation $A : B :: C : D$. This principle can be illustrated as follows with sentences:

$$\text{紅茶が飲みたい。} : \text{あなたは紅茶が好きですか。} :: \text{ビールが飲みたい。} : x$$

$$\Rightarrow x = \text{あなたはビールが好きですか。}$$

In this example, the solution of the analogical equation is *D* = "あなたはビールが好きですか。" (Do you like beer?). If we regard each sentence pair in a cluster as a pair $A : B$ (left to right or right to left), and any short sentence not belonging to the cluster as *C* (a *seed sentence*), the analogical equation $A : B :: C : D$ of unknown *D* can be forged. Such analogical equations allow us to produce new candidate sentences. Each sentence pair in a cluster is a potential template for the generation of new candidate sentences.

4.2 Experiments on New Sentences Generation and Filtering by N-sequences

For the generation of new sentences, we make use of the clusters we obtained from the experiments in Section 3.2 as *rewriting models*. The seed sentences as input data for new sentences generation are the unique Chinese and Japanese short sentences from the 103,629 ASPEC-JC parallel sentences (less than 30 characters). In this experiment, we generated new sentences with each pair of sentences in clusters for Chinese and Japanese respectively. Table 4 gives the statistics for new sentence generation.

To filter out invalid and grammatically incorrect sentences and keep only well-formed sentences with high fluency of expression and adequacy of meaning, we eliminate any sentence that contains an N-sequence of a given length unseen in the reference corpus. This technique to assess the quality of outputs of NLP systems has been used in previous works (Lin and Hovy, 2003; Doddington, 2002; Lepage and Denoual, 2005). In our experiment, we introduced begin/end markers to make sure that the beginning and the end of a sentence are also correct. The best quality was obtained for the values N=6 for Chinese and N=7 for Japanese with the size of reference corpus (about 1,700,000 monolingual data for both Chinese and Japanese). Quality assessment was performed by extracting a sample of 1,000 sentences randomly and checking manually by native speakers. The grammatical quality was at least 96%. This means that 96% of the Chinese and Japanese sentences may be considered as grammatically correct. For new valid sentences, we remember their corresponding seed sentences and the cluster they were generated from.

[5] `http://www.unicode.org/Public/UNIDATA/`

[6] `http://www.kishugiken.co.jp/cn/code10d.html`

		Chinese		Japanese	
Initial data	# of seed sentences	99,538		97,152	
	# of clusters	23,182		21,975	
New sentence generation	# of candidate sentences	105,038,200 Q= 29%		80,183,424 Q= 40%	
Quality assessment (filtered)	# of new valid sentences	unique	seed–new–#	unique	seed–new–#
		33,141	67,099	40,234	84,533
		Q= 96%		Q= 96%	

Table 4: Statistics on new sentence generation in Chinese and Japanese. Q is the quality of the new candidate sentences or new valid sentences after filtering.

Chinese	Japanese	Chinese–Japanese		
seed–new–#	seed–new–#	Initial parallel corpus	Corresponding clusters	Quasi-parallel corpus
67,099	84,533	103,629	15,710	35,817

Table 5: Statistics on the quasi-parallel corpus deducing.

4.3 Deducing and Acquiring Quasi-parallel Sentences

We deduce translation relations based on the initial parallel corpus and corresponding clusters between Chinese and Japanese. If the seeds of two new generated sentences in Chinese and Japanese are aligned in the initial parallel corpus, and if the clusters which they were generated from are corresponding, we suppose that these two Chinese and Japanese newly generated sentences are translations of one another to a certain extent. Table 5 gives the statistics on the quasi-parallel deducing obtained. Among the 35,817 unique Chinese–Japanese quasi-parallel sentences obtained, about 74% were found to be exact translations by manual check on a sampling of 1,000 pairs of sentences. This justifies our use of the term “quasi-parallel” for this kind of data.

5 SMT Experiments

5.1 Experimental Protocol

To assess the contribution of the generated quasi-parallel corpus, we propose to compare two SMT systems. The first one is constructed using the initial given ASPEC-JC parallel corpus. This is the baseline. The second one adds the additional quasi-parallel corpus obtained using analogical associations and analogical clusters.

Baseline: The statistics of the data used in the experiments are given in Table 6 (left). The training corpus consists of 672,315 sentences of initial Chinese–Japanese parallel corpus. The tuning set is 2,090 sentences from the ASPEC-JC.dev corpus, and 2,107 sentences also from the ASPEC-JC.test corpus were used for testing. We perform all experiments using the standard GIZA++/MOSES pipeline (Och and Ney, 2003).

Adding Additional Quasi-parallel Corpus: The statistics of the data used in this second setting are given in Table 6 (right). The training corpus is made of 708,132 (672,315 + 35,817) sentences, i.e., the combination of the initial Chinese–Japanese parallel corpus used in the baseline and the quasi-parallel corpus.

Experimental Results: Table 7 and Table 8 give the evaluation results. We use the standard metrics BLEU (Papineni et al., 2002), NIST (Doddington et al., 2000), WER (Nießen et al., 2000), TER (Snover et al., 2006) and RIBES (Isozaki et al., 2010). As Table 7 shows, significant improvement over the baseline is obtained by adding the quasi-parallel generated data based on the Moses version 1.0, and Table 8 shows a slightly improvement over the baseline is obtained by adding the quasi-parallel generated data based on the Moses version 2.1.1.

5.2 Influence of Segmentation on Translation Results

We also use Kytea[7] to segment Chinese and Japanese. Table 9 and Table 10 show the evaluation results by using Kytea as the segmentation

[7] http://www.phontron.com/kytea/index-ja.html

tools based on standard GIZA++/MOSES (different version in 1.0 and 2.1.1) pipeline. As the evaluation scores (BLEU and RIBES) shown in Table 7, Table 8, Table 9 and Table 10:

- We obtained more increase based on Moses version 1.0 than Moses version 2.1.1 by using urheen/mecab or kytea for Chinese and Japanese as the segmentation tools;
- But, based on Moses version 2.1.1 we obtained higher BLEU and RIBES than Moses version 1.0 by using two different segmentation tools;
- Based on the same Moses version, most of the BLEU and RIBES scores are higher by using urheen and mecab as the segmentation tools for Chinese and Japanese than using kytea (except ja-zh by using kytea based on Moses version 2.1.1).

5.3 Issues for Context-aware Machine Translation

Context-aware plays an important role in disambiguation and machine translation. Usually, the MT systems look at surface form only, conversational speech tends to be more concise and more context-dependent (Example1), and some ambiguities often arises due to polysemy (Example2 from our experiment results by using urheen and mecab as the segmentation tools) and homonymy.

Example1: 下次我要尝尝白的。

Reference_en: I'll try Chinese **wine** next time.

Reference_ja: 今度は**中国**のワインを試してみます。

MT output_en: Next time I'll try the **white**.

MT output_ja: 次回は私は**白**を試してみます。

Example2: 结果发现，其中昼夜均符合环境标准的地点是，平成１５年度为６３**处**（３６．２％），平成１６年度为５９**处**（３７．８％）。

Reference_ja: その結果，全地点のうち昼夜ともに環境基準地を達成したのは，平成１５年度の６３**地点**（３６．２％），平成１６年度で５９**地点**（３７．８％）であった。

MT output (google): それは、昼と夜が環境基準に沿ったものである場所が63（36.2％）が平成15年であることが判明した、平成59（37.8％）が16歳。

MT output (our baseline): その結果，その昼夜ともに環境基準の地点は，平成１５年度は６３**箇所**（３６．２％）では，平成１６年度は５９**箇所**（３７．８％）であった。

MT output (our baseline+add): その結果，その昼夜ともに環境基準の地点は，平成１５年度は６３**箇所**（３６．２％）では，平成１６年度は５９**箇所**（３７．８％）であった。

As the Example2 shows, we obtained the better and more correct translation results based on our translation systems. Correct meaning of a word or a sentence depends context information. The large training data in the same domain is also an extremely important factor in translation systems. They allow us to obtain the well-formed translation result with high fluency of expression and adequacy of meaning.

6 Conclusion

We presented a technique to automatically generate a quasi-parallel corpus to inflate the training corpus used to build an SMT system. The experimental data we use are ASPEC-JC corpus and the monolingual data were collected from the Web. We produced analogical clusters as rewriting models to generate new sentences, and filter newly over-generated sentences by the N-sequences filtering method. The grammatical quality of the valid new sentences is at least 96%. We then assess translation relations between newly generated short sentences across both languages, relying on the similarity between the clusters across languages. We automatically obtained 35,817 Chinese–Japanese sentence pairs, 74% of which were found to be exact translations. We call such sentence pairs a quasi-parallel corpus.

In SMT experiments performed on Chinese–Japanese, using the standard GIZA++/MOSES pipeline, by adding our quasi-parallel data, we were able to inflate the training data in a rewarding way. On the same test set, based on different MOSES versions and segmentation tools, all of translation scores significantly or slightly improved over the baseline systems. It should be stressed that the data that allowed us to get such improvement are not so large in quantity and not so good in quality, but we were able to control both quantity and quality so as to consistently improve translation quality.

Acknowledgments

This work was supported in part by Foreign Joint Project funds from the Kitakyushu Foundation for the Advancement of Industry, Science and Technology (FAIS).

References

Lloyd Allison and Trevor I. Dix. 1986. A bit string longest common subsequence algorithm. *Information Processing Letter*, 23:305–310.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2013. Chinese–Japanese parallel sentence extraction from quasi–comparable corpora. In *ACL 2013*, pages 34–42.

Ferdinand de Saussure. 1916. *Cours de linguistique générale*. Payot, Lausanne et Paris, [1ère éd. 1916] edition.

George R Doddington, Mark A Przybocki, Alvin F Martin, and Douglas A Reynolds. 2000. The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective. *Speech Communication*, 31(2):225–254.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Human Language Technology Conference (HLT2002)*, pages 128–132, San Diego, CA, USA. Morgan Kaufmann.

Pascale Fung and Percy Cheung. 2004. Multilevel bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *In COLING 2004*, pages 1051–1057.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952. Association for Computational Linguistics.

Esa Itkonen. 2005. *Analogy as Structure and Process: Approaches in linguistics, cognitive psychology and philosophy of science*, volume 14.

Yves Lepage and Etienne Denoual. 2005. Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation. In *IWP2005*, pages 57–64.

Yves Lepage. 1998. Solving analogies on words: An algorithm. In *Proceedings of COLING-ACL'98*, pages 728–735, Montréal, August.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL-2003)*, pages 71–78.

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

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for machine translation research. In *Proceedings of LREC '00*, pages 39–45.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL 2002*, pages 311–318.

Jason R Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *AMTA '06*, pages 223–231.

Robert A. Wagner and Michael J. Fischer. 1974. The string-to-string correction problem. *Journal of the ACM*, 21:168–173.

	Baseline	Chinese	Japanese
train	sentences	672,315	672,315
	words	18,847,514	23,480,703
	mean ± std.dev.	28.12 ± 15.20	35.05 ± 18.88

+ Quasi-parallel	Chinese	Japanese
sentences	**708,132**	**708,132**
words	19,212,187	24,512,079
mean ± std.dev.	27.13 ± 14.19	34.23 ± 17.22

	Both experiments	Chinese	Japanese
tune	sentences	2,090	2,090
	words	60,458	73,177
	mean ± std.dev.	28.93 ± 15.86	35.01 ± 18.87
test	sentences	2,107	2,107
	words	59,594	72,027
	mean ± std.dev.	28.28 ± 14.55	34.18 ± 17.43

Table 6: Statistics on the Chinese–Japanese corpus used for the training, tuning, and test sets in baseline (left) and baseline + quasi-parallel data (right). The tuning and testing sets are the same in both experiments. Segmentation tools: urheen for Chinese and Mecab for Japanese.

		BLEU	NIST	WER	TER	RIBES
zh-ja	baseline	29.10	7.5677	0.5352	0.5478	0.7801
	+ additional training data	**32.03**	**7.9741**	**0.5069**	**0.5172**	**0.7906**
ja-zh	baseline	22.98	7.0103	0.5481	0.5711	0.7893
	+ additional training data	**24.87**	**7.3208**	**0.5273**	**0.5482**	**0.8013**

Table 7: Evaluation results for Chinese–Japanese translation across two SMT systems (baseline and baseline + additional quasi-parallel data), Moses version: 1.0, segmentation tools: urheen and mecab.

		BLEU	NIST	WER	TER	RIBES
zh-ja	baseline	33.41	8.1537	0.4967	0.5061	0.7956
	+ additional training data	**33.68**	**8.1820**	**0.4955**	**0.5039**	**0.7964**
ja-zh	baseline	25.53	7.3885	0.5227	0.5427	0.8053
	+ additional training data	**25.80**	**7.4571**	**0.5176**	**0.5378**	**0.8060**

Table 8: Evaluation results for Chinese–Japanese translation across two SMT systems (baseline and baseline + additional quasi-parallel data), Moses version: 2.1.1, segmentation tools: Urheen and Mecab.

		BLEU	NIST	WER	TER	RIBES
zh-ja	baseline	28.35	7.3123	0.5667	0.5741	0.7610
	+ additional training data	**28.87**	**7.4637**	**0.5566**	**0.5615**	**0.7739**
ja-zh	baseline	22.83	6.9533	0.5633	0.5853	0.7807
	+ additional training data	**23.18**	**7.0402**	**0.5547**	**0.5778**	**0.7865**

Table 9: Evaluation results for Chinese–Japanese translation across two SMT systems (baseline and baseline + additional quasi-parallel data), Moses version:1.0, segmentation tools: Kytea.

		BLEU	NIST	WER	TER	RIBES
zh-ja	baseline	33.27	7.9579	0.5249	0.5272	0.7820
	+ additional training data	**33.56**	**8.0229**	**0.5178**	**0.5206**	**0.7849**
ja-zh	baseline	26.25	7.4931	0.5197	0.5398	0.8085
	+ additional training data	**26.52**	**7.5523**	**0.5128**	**0.5335**	**0.8105**

Table 10: Evaluation results for Chinese–Japanese translation across two SMT systems (baseline and baseline + additional quasi-parallel data), Moses version: 2.1.1, segmentation tools: Kytea.