Inflating a Training Corpus for SMT by Using Unrelated Unaligned Monolingual Data

Wei Yang and Yves Lepage

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

Abstract. To improve the translation quality of less resourced language pairs, the most natural answer is to build larger and larger aligned training data, that is to make those language pairs well resourced. But aligned data is not always easy to collect. In contrast, monolingual data are usually easier to access. In this paper we show how to leverage unrelated unaligned monolingual data to construct additional training data that varies only a little from the original training data. We measure the contribution of such additional data to translation quality. We report an experiment between Chinese and Japanese where we use 70,000 sentences of unrelated unaligned monolingual additional data in each language to construct new sentence pairs that are not perfectly aligned. We add these sentence pairs to a training corpus of 110,000 sentence pairs, and report an increase of 6 BLEU points.

Keywords: monolingual corpus, analogies, quasi-parallel corpus, machine translation.

1 Introduction

Sentence-level aligned parallel corpora are an extremely important resource as training data in statistical machine translation (SMT). The quantity of the parallel sentences is crucial, because the translation knowledge is acquired from these sentential parallel corpora [4]. The quality of the aligned parallel sentences is another important factor that impacts greatly the quality of the translation relations extracted between words or phrases between the source language and the target language. To summarize, translation quality depends on the quantity and the quality of the parallel corpus.

There exist numerous freely available bilingual or multilingual parallel corpora for language pairs that involve English, such as the Europarl parallel corpus [9]. The Europarl corpus was designed for research purposes in statistical machine translation and it has been used for multiple other research purposes, including for example word sense disambiguation. But the linguistic resources between languages like: Chinese, Japanese, Thai, Hindi or Bahasa Indonesian are relatively scarce. This does not mean that they are minority languages, as

A. Przepiórkowski and M. Ogrodniczuk (Eds.): PolTAL 2014, LNAI 8686, pp. 236–248, 2014.

all these languages have multimillion-strong speaker and writer bases and monolingual data are quite easy to collect. But bilingual sentence-aligned corpora in any of these language pairs are not so easily accessible. Manual construction of resources for such less-resourced language pairs is time consuming and costly. Researchers face many difficulties to extract parallel corpora from general texts [1] or from specialized texts like patent families [3]. For Chinese or Japanese, another important issue comes from copyright restrictions: most existing resources are not free due to copyright. For all the above reasons, we propose a novel method to combine small freely available aligned bi-corpora with approximately aligned sentence pairs generated from a reasonable size of monolingual unaligned data. We call such approximately aligned sentence pairs a quasi-parallel corpus, because it contains sentences that are translations to each other only to a certain extent. The degree of correspondence in translation is estimated by some similarity scores.

Our method to construct a quasi-parallel corpus is based on the notion of proportional analogy. This notion has already been applied to machine translation. For instance, [6] and [10] both address the problem of translating unknown words in a statistical machine translation framework and use analogy for that. Corpus-based analogical techniques have also been used to build a complete example-based machine translation system [13]. Analogy has also been used to cluster word pairs by semantic relations [2].

In this paper, we show how to cluster sentences to be used as rewriting models for new sentence generation and how to deduce translation relations between these newly generated sentences to construct a quasi-parallel corpus. We then report SMT experiments and compare a baseline system using a relatively small amount of parallel data and a system built on the baseline by adding the previous quasi-parallel corpus as additional training data. This new system performs significantly better. We also evaluate the quality of the quasi-parallel corpus in terms of language quality (grammaticality) and alignment quality (translation similarity).

2 Data Preparation for the Experiments

We perform our experiments on Chinese–Japanese which can be considered a less resourced language pair. We started by collecting parallel Chinese–Japanese data and monolingual data form the Web by using an in-house crawler.

2.1 Collecting Parallel Chinese–Japanese Data

We collected and aligned Chinese–Japanese sentences from the subtitles of movies and drama series, based on the time of the subtitles in two languages. Table 2.1 shows examples of Chinese and Japanese subtitles as translations to each other in a movie. We collected these data from the following websites: *Subscene.com* and *Opensubtitles.org*. Each text piece consists of one or two short sentences

shown on the screen nearly every second in Chinese and Japanese. These sentences are short and simple for readers to easily understand them in limited time (one second). The subtitle corpus we used here comes from about 300 subtitle files. We obtained 106,310 pairs of aligned Chinese–Japanese sentence pairs after some cleaning.

Table 1. Examples for the Chinese–Japanese subtitle short sentence pairs

[Events] Format: Start, End, Name, Text
Dialogue: 0:02:11.99,0:02:14.66,cn.sub,不停船的话就击沉你们！快停！
Dialogue: 0:02:14.66,0:02:16.73,cn.sub,被抢劫咯是海盗哦
. . . . . .
Dialogue: 0:02:11.99,0:02:14.66,jp.sub,止まらないと沈めるぞ！止まれー！
Dialogue: 0:02:14.66,0:02:16.73,jp.sub,さらわれるー 海賊だー
. . . . . .

We also downloaded the JEC Basic Sentence Data by Kyoto U. and NICT with 5,304 Chinese–Japanese sentence pairs. We extracted 1,500 pairs of sentences for tuning and testing; the statistics about all this data will be given in Sect. 5.1. The rest of the data (3,804 pairs) will be combined with the subtitle corpus as the initial parallel corpus we used in this paper. Table 2.1 shows the statistics about the Chinese–Japanese initial parallel corpus we used in the experiments. Word segmentation has been done using the following toolkits: Mecab for Japanese and Urheen for Chinese.

Table 2. Statistics on the Chinese–Japanese subtitle data combined with a part of JEC sentences. This constitutes our initial parallel corpus (110,114 sentence pairs in total: 106,310 + 3,804).

	Language	# of different sentences (cleaned)	size of sentences in characters (mean ± std.dev.)	total characters	total words
Subtitle Corpus (106,310) + JEC (3,804)	Chinese	99,251	8.68 ± 3.59	861,723	589,757
	Japanese	90,406	11.99 ± 4.36	1,084,287	647,285

2.2 Collection of Monolingual Resources

To generate new quasi-parallel data, we use unrelated unaligned monolingual data. We collected monolingual Chinese and Japanese short sentences (less than 30 characters in size) mainly from the following websites: “Yahoo China”, “Yahoo China News”, “douban” for Chinese and “Yahoo! JAPAN”, “Mainichi Japan” for Japanese. Table 3 gives the statistics of the cleaned 70,000 monolingual data we used in the experiments.

Table 3. Statistics on the cleaned Chinese and Japanese monolingual short sentences

	# 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

3 Construction of Analogical Clusters

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 [11]. 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 $d(A,C) = d(B,D)$).

Sentential Analogies. We gather pairs of sentences in Chinese and Japanese respectively, that constitute proportional analogies. For instance, the two following pairs of Japanese sentences:

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 black tea.' : *'Do you like black tea?'* :: *'I'd like a beer.'* : *'Do you like beer?'*

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: $d(A,B) = d(C,D) = 13$. The same must be true by exchanging the two sentences in the middle, B and C, so that $d(A,C) = d(B,D) = 5$. The relation on the number of occurrences of characters, which must be valid for each character, may be illustrated as follows for the character **飲**: 1 (in A) – 0 (in B) = 1 (in C) – 0 (in D). We call any such two pairs of sentences a *sentential analogy*.

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 form 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 (English translation below) shows 6 possible sentential analogies. The size of this cluster is 4. In this analogical cluster, the box shows what remains the same on the left and on the right of the cluster, and the underline shows the changes on one sentence pair of the cluster. Analogical clusters like this can be considered as *rewriting models* to generate new sentences. The technique will be described in Sect. 4.

紅茶が 飲みたい。 : あなたは 紅茶が 好きですか。
'I'd like a cup of black tea.' *'Do you like black tea?'*
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 beer.' *'Do you like beer?'*
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 some juice.' *'Do you like juice?'*
冷たいお水が 飲みたい。 : あなたは 冷たいお水が 好きですか。
'I'd like some cold water.' *'Do you like cold water?'*

3.2 Cluster Construction

In each language, independently, we construct analogical clusters from the unrelated monolingual data. The number of unique sentences used is 70,000 for both languages. Table 4 summarizes the statistics on the clusters constructed.

Table 4. Statistics on the Chinese and Japanese clusters constructed from our unrelated monolingual data independently in each language

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

3.3 Computing the Correspondence between Clusters

The core of our technique lies in the computation of the correspondence between clusters across languages. The goal of the technique described hereafter is to spot situations as the following one. Suppose that we have the two clusters in two different languages shown in Table 5. The lines in the clusters are not translations one of another. But the change between the left and the right columns are the same in both languages. In this case, we say that the two clusters correspond. The sequel of this section describes the needed computation to spot such correspondences.

Table 5. Two corresponding clusters. They do not have the same sizes and the sentences contained are not translations (i.e., here, the same). But the change between the left and the right columns ('I . . . .' → 'Do you . . . ?') is the same.

Language 1:	Language 2:
I like beer.:**Do you** like beer?	**I** study maths.:**Do you** study maths?
I like juice.:**Do you** like juice?	**I** watch movies.:**Do you** watch movies?
	I read books.:**Do you** read books?

We extract corresponding clusters by computing the similarity between the changes in them using the following steps:

- First, for each sentence pair in a cluster, we extract the change between the left and the right sides by computing their longest common subsequence (LCS) [18].
- 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 on these changes so as to obtain minimal changes expressed in sets of words or characters.
- Finally, we compute the similarity between the left sets (S_{left}) and the right sets (S_{right}) of the Chinese and Japanese clusters. To this end, we make use of the EDR dictionary[1], a traditional-simplified Chinese variant table[2] and a Kanji–Hanzi Conversion Table[3] to translate all Japanese words into Chinese, or convert Japanese characters into simplified Chinese. We calculate the similarity between two Chinese and Japanese word sets according to a classical Dice formula:

$$Sim = \frac{2 \times |S_{\text{zh}} \cap S_{\text{ja}}|}{|S_{\text{zh}}| + |S_{\text{ja}}|} \tag{1}$$

  S_{zh} and S_{ja} denote the minimal sets of changes across the clusters (both on the left or right) in both languages. The formula for computing the similarity between two Chinese and Japanese clusters is given in formula (2):

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

The previous computation shows that the correspondence between two clusters does not rely on the translation correspondences between the entire sentences contained in the clusters, but only a small part of them. For this reason, correspondences between clusters have much higher chance to be found than translation between sentences. Figure 1 gives an example of the extracted corresponding clusters and the sets (S_{left} and S_{right}) of the changes (shown in underline) in the Chinese ($S_{\text{zh}_{\text{left}}} =$ { 经典 } and $S_{\text{zh}_{\text{right}}} =$ { 很, 不错 }) and Japanese ($S_{\text{ja}_{\text{left}}} =$ { クラシック } and $S_{\text{ja}_{\text{right}}} =$ { この, は, とても, いい }) clusters respectively, the similarity between the two clusters calculated according to formula (2) is 0.833. We set different threshold for $Sim_{C_{\text{zh}}-C_{\text{ja}}}$ and check the correspondence between these extracted clusters by sampling. While the $Sim_{C_{\text{zh}}-C_{\text{ja}}}$ threshold was set to 0.300, the acceptability of the correspondence between the extracted clusters were able to achieve to 80%. About 14,578 corresponding clusters were extracted ($Sim_{C_{\text{zh}}-C_{\text{ja}}} \geq 0.300$) by the above steps.

[1] http://www2.nict.go.jp/out-promotion/techtransfer/EDR
[2] http://www.unicode.org/Public/UNIDATA/
[3] http://www.kishugiken.co.jp/cn/code10d.html

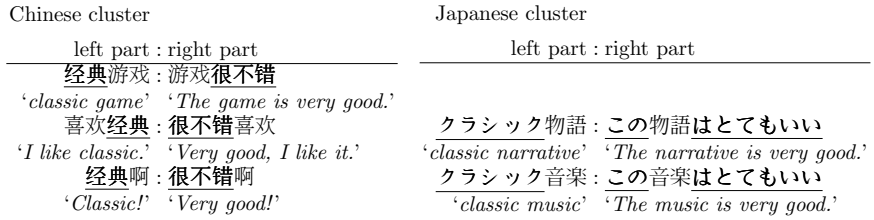

Fig. 1. Two corresponding clusters constructed based on unrelated unaligned monolingual data. Chinese on the left of the figure, Japanese on the right of the figure. In both clusters, i.e., in both languages, the word meaning 'classic' on the left part of the cluster, is replaced on the right part of the cluster, by words meaning 'very good'.

4 Generation of New Sentences Using Analogical Associations

4.1 Generation of New Sentences

Analogy is not only a structural relationship. It is also a process [8] by which, given two related forms and only one form, the fourth missing form is built [5]. If the objects A, B, C are given, we may build an other object D according to the analogical equation $A : B :: C : D$. This principle can be illustrated with sentences:

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

$$\textit{'I'd like a black tea.'} : \textit{'Do you like black tea?'} :: \textit{'I'd like a beer.'} : x \quad \Rightarrow \quad x = \textit{'Do you like beer?'}$$

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 thus a potential rewriting template for the generation of new candidate sentences.

4.2 Experiments on New Sentence Generation and Filtering

We generate new sentences in Chinese and Japanese respectively and independently by using analogical associations based on the result of the clusters constructed and the initial parallel corpus (unique Chinese and Japanese sentences, described in Sect. 2.1) as the seed sentences. We generate new candidate sentences with each sentence pair in each cluster. Then we filter these candidate sentences by using the N-sequence method, a technique to assess the quality of

outputs of NLP systems used in previous work [12]. The technique keeps the sentences where all N-sequences can be found in a reference corpus. Said the way round, we eliminate any sentence that contains an N-sequence not found in the reference corpus. We introduced begin/end markers to make sure that also the beginning and the end of a sentence are correct. We set N to 6 for Chinese and to 7 for Japanese. The size of the reference data we used are 1,059,985 for Chinese and 1,074,851 for Japanese. Table 6 shows the statistics on the new sentence generation.

Table 6. 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	
Initial data	# of seed sentences	99,251		90,406	
	# of clusters	23,182		21,975	
New sentence generation	# of candidate sentences	192,121,764 Q= 20%		50,418,891 Q= 50%	
Quality assessment (filtered)	# of new valid sentences	unique	seed–new–#	unique	seed–new–#
		34,230	105,537	142,820	191,409
		Q= 99%		Q= 99%	

Quality assessment was performed by sampling 1,000 sentences randomly and asking native speakers to check for grammaticality. The grammatical quality was at least 99%. This means that 99% of the Chinese and Japanese sentences may be considered as grammatically correct. For each valid sentence, we remember the corresponding seed sentence and the cluster identifier it was generated from. Table 6 shows the statistics of the filtering result. Table 7 shows the examples of new sentences generated from two given clusters (Fig. 1) and a pair of parallel seed sentences in Chinese and Japanese. In the Chinese example, we obtained two different valid sentences through the same cluster.

Table 7. The result of new sentence generation in Chinese and Japanese based on a pair of parallel seed sentences according to the clusters given in Fig. 1

A : B :: C_{seed} : X_{new-zh}			
经典游戏 : 游戏**很不错** 喜欢**经典** : **很不错**喜欢 **经典**啊 : **很不错**啊	::	经典电影 '*classic film*'	⇒ 电影很不错 '*The film is very good.*' 很不错电影 '*That's very good, the film.*'

A : B :: C_{seed} : X_{new-ja}			
クラシック物語 : **この**物語**はとてもいい** クラシック音楽 : **この**音楽**はとてもいい**	::	クラシック映画 '*classic film*'	⇒ この映画はとてもいい '*The film is very good.*'

4.3 Deduction of Translation Relations between New Generated 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 generated from are corresponding, we suppose that these two Chinese and Japanese newly generated sentences are translations of one another to a certain extent. In the example in Table 7, we obtain two pairs of quasi-parallel sentences: "电影很不错 : この映画はとてもいい" and "很不错电影 : この映画はとてもいい". Table 8 gives the statistics of the quasi-parallel corpus. Among the 76,151 unique Chinese–Japanese quasi-parallel sentences obtained, about 75% were found to be exact translations by manual check by sampling 1,000 pairs of sentences. This justifies our use of the term "quasi-parallel" for this kind of data.

Table 8. Statistics on the quasi-parallel corpus deducing

Chinese	Japanese	Chinese–Japanese		
seed–new–#	seed–new–#	Initial parallel corpus	Corresponding clusters	Quasi-parallel corpus
105,537	191,409	110,114	14,578	76,151

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 an initial 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 9 (left). The training corpus consists of 110,114 sentences of initial Chinese–Japanese parallel corpus. The tuning set is 500 sentences from the JEC parallel corpus, and 1,000 sentences also from the JEC corpus were used for testing. We perform all experiments using the standard GIZA++/MOSES pipeline [15].

Adding Additional Quasi-parallel Corpus. The statistics of the data used in this second setting are given in Table 9 (right). The training corpus is made of 186,265 (110,114 + 76,151) sentences, i.e., the combination of the initial Chinese–Japanese parallel corpus used in the baseline and the quasi-parallel corpus.

Table 9. 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.

	Baseline	Chinese	Japanese
train	sentences	110,114	110,114
	words	637,036	721,850
	mean ± std.dev.	5.94 ± 2.60	6.69 ± 2.94

+ Quasi-parallel	Chinese	Japanese
sentences	**186,265**	**186,265**
words	1,147,098	1,318,747
mean ± std.dev.	6.06 ± 2.61	7.16 ± 3.08

	Both experiments	Chinese	Japanese
tune	sentences	500	500
	words	3,582	5,042
	mean ± std.dev.	7.15 ± 2.86	10.12 ± 3.39
test	sentences	1,000	1,000
	words	7,285	10,126
	mean ± std.dev.	7.28 ± 2.87	10.15 ± 3.30

Experimental Results. Table 10 gives the evaluation results. We use the standard metrics BLEU [16], NIST [7], WER [14], and TER [17]. As Table 10 shows, significant improvement over the baseline is obtained by adding the quasi-parallel generated data.

Table 10. Evaluation results for Chinese–Japanese translation across two SMT systems (baseline and baseline + additional quasi-parallel data)

		BLEU	NIST	WER	TER
zh-ja	baseline	13.10	4.1732	0.7229	0.7344
	+ additional training data	**19.27**	**4.7013**	**0.6880**	**0.6933**
ja-zh	baseline	10.94	4.4028	0.7545	0.7621
	+ additional training data	**17.66**	**4.7989**	**0.7140**	**0.7214**

5.2 Analysis of the Results

We investigated the N (source length) $\times$ M (target length) distribution in phrase tables generated from the initial parallel corpus and the inflated training corpus by adding the quasi-parallel data. In Table 11 and 12, the statistics (zh→ja) show that the total number of phrase pairs generated by adding additional quasi-parallel corpus is larger than when using only the initial parallel corpus as training data. If we compare the number of entries, the number of phrase pairs (in Table 12) on the diagonal got a significant increase in the number of phrase pairs of similar length (except 1 $\times$ 1). Considering the correspondence between lengths in Chinese–Japanese translation, the increase in phrase pairs with different lengths (like 2 (zh) $\times$ 3 (ja) and 3 (zh) $\times$ 4 (ja)) is felicitous. This means that adding the additional quasi-parallel corpus for inflating the training

corpus for SMT allowed us to produce much more numerous useful alignments. Table 13 gives samples of phrase pairs showing the same Chinese phrase aligned with different Japanese phrases. We obtained additional alignments compared to the use of the limited initial parallel corpus. The additional phrase pairs in this given samples are all correct by checking manually. By increasing the size of the training corpus by adding quasi-parallel data, we got more good alignment informations between Chinese and Japanese.

Table 11. Distribution of phrase pairs in the phrase translation table of GIZA++ (baseline zh→ja)

		Target language							
		1-grams	2-grams	3-grams	4-grams	5-grams	6-grams	7-grams	total
Source language	1-grams	23,789	35,494	25,069	13,670	6,568	2,982	1,257	108,829
	2-grams	34,865	52,596	38,612	22,429	12,300	6,413	3,113	170,328
	3-grams	18,904	33,116	39,633	29,465	19,262	11,881	6,617	158,878
	4-grams	8,097	15,948	24,779	28,160	23,629	17,628	11,495	129,736
	5-grams	3,235	7,020	12,656	18,532	21,166	19,277	15,072	96,958
	6-grams	1,195	2,860	6,027	10,405	14,537	16,470	15,245	66,739
	7-grams	466	1,223	2,615	5,196	8,395	11,003	12,239	41,137
	total	90,551	148,257	149,391	127,857	105,857	85,654	65,038	772,605

Table 12. Distribution of phrase pairs in the phrase translation table (baseline + additional training data: zh→ja). Compare with Table 13, increase in entry numbers in boldface.

		Target language							
		1-grams	2-grams	3-grams	4-grams	5-grams	6-grams	7-grams	total
Source language	1-grams	23,752	**38,758**	**29,920**	**18,025**	9,671	4,777	2,211	127,114
	2-grams	38,997	**56,814**	**44,647**	**28,149**	16,733	9,539	4,985	199,864
	3-grams	23,240	38,360	**45,596**	**35,724**	25,148	16,653	9,977	194,698
	4-grams	10,954	19,398	29,343	**33,124**	28,991	22,801	16,078	160,689
	5-grams	4,779	9,143	15,540	21,864	25,515	24,242	20,052	121,135
	6-grams	1,858	3,787	7,475	12,683	17,799	20,577	19,996	84,175
	7-grams	765	1,577	3,357	6,531	10,335	13,969	15,787	52,321
	total	104,345	167,837	175,878	156,100	134,192	112,558	89,086	**939,996**

6 Conclusion

Phrase-based statistical machine translation (PB-SMT) relies on the availability of parallel data to extract and align phrases and to estimate various features like translation probabilities. The quality of translation depends on the quantity of the available data and on its quality. The conventional approach for improving

Table 13. Samples of phrase alignments in zh→ja phrase table. The same Chinese phrase and corresponding Japanese phrases.

Baseline (Initial parallel corpus)	只能这样了 これ で 行く しか ない
	只能这样了 それ しか ない ん だ よ
	只能这样了 やる しか ない だろ
Adding quasi-parallel data	只能这样了 これ で 行く しか なかっ た
	只能这样了 これ しか ない
	只能这样了 それ しか ない ん だ
	只能这样了 やる しか ない
	只能这样了 やる しか なかっ た

performance of PB-SMT systems is to increase the size of the parallel corpus by adding new training data, usually extracted from comparable corpora.

We followed a slightly different path. Firstly, we chose to add more data that may be not so well aligned. We called such data quasi parallel data. Secondly, the quantity of data that we added was not so important as we expanded the training data by only two thirds (110,000 to 186,000). The main point in our view is the original method to generate the new quasi parallel data. These were obtained by structuring unaligned unrelated monolingual data according to analogical associations. These analogical associations are used as rewriting models to produce new sentences.

In experiments performed on Chinese–Japanese, by adding this kind of quasi-parallel data, no so large in quantity and not so good in quality (as translation is concerned), we were able to inflate the translation table in a rewarding way. On the same test set, the translation quality significantly increased over the baseline systems (more than 6 BLEU points).

The explanation why the method works may be as follows. The quasi-parallel data that we produced reflect changes, i.e., linguistic variations, that are attested in monolingual unrelated texts. This may greatly help to better cover the linguistic variations that may appear in a test set which, by definition, is unknown in advance. We claim that the analogical associations helped to capture and to reproduce in an efficient way these linguistic variations. We believe that, if we had used data from the same domain, our results may have been less convincing. Our method worked with a relatively small training corpus, and one may well think that, of course, its performance will decrease if really larger training data were used in the baseline.

An other interesting point in the method used is that the productivity of analogy allowed us to generate a reasonable quantity of new sentences. But it is also possible to use very simple techniques (attested n-gram method) to easily limit over-generation and to ensure the grammaticality of sentences. In our experiments, it was easy to reach grammaticality for 99% of the sentences generated. All this resulted in a significant increase in the number of entries in the phrase tables produced by the standard GIZA++/MOSES pipeline, especially for the lengths of phrases relevant for the languages at hand.

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

1. Abdul Rauf, S., Schwenk, H.: Parallel sentence generation from comparable corpora for improved SMT. Machine Translation 25(4), 341–375 (2011)
2. Biçici, E., Yuret, D.: Clustering word pairs to answer analogy questions. In: Proceedings of TAINN 2006, pp. 277–284 (2006)
3. Bin, L., Tao, J., Kapo, C., Benjamin, K.: T.: Building a large English-Chinese parallel corpus from comparable patents and its experimental application to SMT. In: Proceedings of LREC 2010, pp. 42–49 (2010)
4. Chu, C., Nakazawa, T., Kurohashi, S.: Chinese–Japanese parallel sentence extraction from quasi–comparable corpora. In: ACL 2013, pp. 34–42 (2013)
5. De Saussure, F.: Cours de linguistique générale. Payot, Paris (1916, 1995)
6. Denoual, E.: Analogical translation of unknown words in a statistical machine translation framework. In: MT Summit XI, Copenhagen pp. 10–14 (2007)
7. Doddington, G.R., Przybocki, M.A., Martin, A.F., Reynolds, D.A.: The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective. Speech Communication 31(2), 225–254 (2000)
8. Itkonen, E.: Analogy as Structure and Process: Approaches in linguistics, cognitive psychology and philosophy of science 14 (2005)
9. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of MT Summit, pp. 79–86 (2005)
10. Langlais, P., Patry, A.: Translating unknown words by analogical learning. In: EMNLP-CoNLL, pp. 877–886 (2007)
11. Lepage, Y.: Solving analogies on words: An algorithm. In: Proceedings of COLING-ACL 1998, Montréal, pp. 728–735 (August 1998)
12. Lepage, Y., Denoual, E.: Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation. In: IWP 2005, pp. 57–64 (2005)
13. Lepage, Y., Denoual, E.: Purest ever example-based machine translation: detailed presentation and assessment. Machine Translation 19, 251–282 (2005)
14. Nießen, S., Och, F.J., Leusch, G., Ney, H.: An evaluation tool for machine translation: Fast evaluation for machine translation research. In: Proceedings of LREC 2000, pp. 39–45 (2000)
15. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), 19–51 (2003)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: ACL 2002, pp. 311–318 (2002)
17. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: AMTA 2006, pp. 223–231 (2006)
18. Wagner, R.A., Fischer, M.J.: The string-to-string correction problem. Journal of the ACM 21, 168–173 (1974)