Regular Paper

# Inflating a Small Parallel Corpus into a Large Quasi-parallel Corpus Using Monolingual Data for Chinese–Japanese Machine Translation

Wei Yang[1,a)] Hanfei Shen[1,†1,b)] Yves Lepage[1,c)]

Received: April 7, 2016, Accepted: October 4, 2016

**Abstract:** Increasing the size of parallel corpora for less-resourced language pairs is essential for machine translation (MT). To address the shortage of parallel corpora between Chinese and Japanese, we propose a method to construct a quasi-parallel corpus by inflating a small amount of Chinese–Japanese corpus, so as to improve statistical machine translation (SMT) quality. We generate new sentences using analogical associations based on large amounts of monolingual data and a small amount of parallel data. We filter over-generated sentences using two filtering methods: one based on BLEU and the second one based on N-sequences. We add the obtained aligned quasi-parallel corpus to a small parallel Chinese–Japanese corpus and perform SMT experiments. We obtain significant improvements over a baseline system.

**Keywords:** quasi-parallel corpus, analogies, clustering, filtering, BLEU, machine translation

## 1. Introduction

Sentence-level aligned parallel corpora are an essential resource in corpus-based MT like SMT. The quantity and the quality of the parallel sentences are two important factors that strongly impact the translation quality. In SMT systems, the translation knowledge is acquired from these parallel sentences. Consequently, the quantity and the quality of the translation relations extracted between words or phrases between the source language and the target language depends on the quantity and the quality of the parallel sentences.

There already exist numerous freely available bilingual or multilingual corpora for European languages. For instance, the Europarl parallel corpus [1] is a collection of parallel text from the proceedings of the European Parliament. It includes versions in 21 European languages. The aligned multilingual JRC-Acquis corpus [2] also funded by the European Union, contains resources in 21 European languages.

Currently, there are almost no Chinese–Japanese parallel corpora publicly freely available on all domains for users and researchers. Some research institutions have tried to construct Chinese–Japanese bilingual parallel corpora, for instance, the basic traveler's expression corpus (BTEC) in Japanese, English, and Chinese has been constructed by the Advanced Telecommunications Research Institute International (ATR). A speech recognition engine was developed based on this corpus [3]. The National Institute of Information and Communications Technology (NICT) in Japan created a Japanese–Chinese corpus of 38,383 sentences by selecting Japanese sentences from the Mainichi Newspaper and translating them manually into Chinese [4]. Harbin Institute of Technology in China (HIT) constructed the Olympic Oriented Chinese–English–Japanese Trilingual Corpus [5] from a Chinese–English parallel corpus collection by adding Japanese translations. This initiative was intended for the development of natural language processing (NLP) for the Olympic Games in Beijing in 2008. The resource consists of 54,043 sentence pairs. Most of the above corpora are not released or freely available, due to copyright problems.

In the last two years, two parallel corpora were released in the domain of scientific papers and patents. They are provided under the condition of participating in the open evaluation campaign Workshop on Asian Translation (WAT) [*1]. The first parallel corpus is the Asian Scientific Paper Excerpt Corpus (ASPEC) [*2]. It contains 680,000 Japanese–Chinese parallel sentences extracted from scientific papers. It was built within the frame of a four year project of translating Japanese scientific papers from the literature database and electronic journal site J-STAGE of JST into Chinese after receiving permission from the necessary academic associations [6]. The second parallel corpus provided for WAT is the JPO corpus [*3], created jointly, based on an agreement between the Japan Patent Office (JPO) and NICT. This corpus consists of a Chinese–Japanese and a Korean–Japanese patent description corpus of one million parallel sentences in science and technology divided into four sections. As already mentioned above, for the collection of Chinese–Japanese parallel corpora, an important is-

[1] Graduate School of Information, Production and Systems, Waseda University, Kitakyushu, Fukuoka 808–0135, Japan
[†1] Presently with NEC Corporation
[a)] kevinyoogi@akane.waseda.jp
[b)] h-shen@bk.jp.nec.com
[c)] yves.lepage@waseda.jp

[*1] http://orchid.kuee.kyoto-u.ac.jp/WAT/WAT2014/index.html and http://orchid.kuee.kyoto-u.ac.jp/WAT/
[*2] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/
[*3] http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/

sue arises from copyright restrictions. Most existing resources are not freely available due to copyright restrictions.

It is worth noticing that the data contained in the mentioned corpora above are translated from one language into another language manually in the frame of long term projects (e.g., the ASPEC corpus) or extracted from the existing article level aligned text via sentence alignment (e.g., the Europarl and JPO corpora). There are also some works for parallel corpora construction by collaborative manner (e.g., the Tatoeba project) [*4] or crowdsourcing translation [7]. Automatic extraction or construction of parallel corpus in different domains is research that is indispensable for improving SMT performance, especially for this less-resourced language pair. In general, researchers face many difficulties in extracting or constructing parallel corpora from general texts or from specialized texts like patent families.

In this paper, we propose a different way to construct a quasi-parallel Chinese–Japanese corpus by leveraging a small amount of parallel data and large amounts of unrelated monolingual data and using analogical associations. In our research, a *quasi-parallel corpus* contains sentences that are translations of each other to a certain extent as estimated by certain similarity scores. To do this, we construct analogical clusters from Chinese and Japanese monolingual data. These clusters group sentences with the same exchanges. They can be used as *rewriting models* for the generation of new sentences. We also compute the similarity between Chinese and Japanese clusters across languages so as to find corresponding clusters with the similar exchanges. We then generate new sentences in Chinese and Japanese independently using corresponding clusters and an existing Chinese–Japanese parallel corpus. We filter the newly generated candidate sentences to retain the ones which are more fluent in expression and more appropriate in meaning. Finally, we deduce the translation relations between these Chinese and Japanese filtered newly generated sentences based on the translation relations in existing parallel corpus and the correspondence between clusters.

This paper is structured as follows, Section 2 reviews related works. In Section 3, we present the overview of our proposed method. In Section 4, we describe the first two steps of our proposed method, i.e., clustering and generation of new sentences. Section 5 describes filtering techniques for obtaining the final Chinese–Japanese quasi-parallel corpus. Section 6 shows the experimental data used, the experimental settings and the evaluation results, as well as some analysis of the results. Conclusion and future work are given in Section 7.

## 2. Related Work

In recent years, there have been several approaches developed for obtaining parallel sentences or fragments from non-parallel data [8], [9], such as comparable data [8], [10], [11], [12] and quasi-comparable data [13] to make contributions to SMT. *Parallel corpora* contain parallel sentences, i.e., sentences which are translations of each other. The term *comparable corpora* refers to texts in two languages that are similar in meaning or expressions, but are not exact translations. *Quasi-comparable corpora* that contain more disparate very-non-parallel bilingual documents that could either be on the same topic (in-topic) or not (out-topic) [13], are more available than *comparable corpora*. In *quasi-comparable corpora*, there are few or no parallel sentences [14]. In Ref. [8], they extract parallel sentences from non-parallel corpora by starting with a relatively small parallel corpus and large Chinese, Arabic, and English non-parallel newspaper corpora. They train a maximum entropy classifier to determine which sentences may be aligned. They aim at improving the performance of an SMT system for less-resourced language pairs. Similarly, we also start with an existing small parallel corpus, but combine it with large amounts of monolingual data to construct a *quasi-parallel corpus*. In the method in Ref. [8], the final sentences come from the monolingual corpora. In our method, the final sentences are created by similarity with sentences in the parallel corpus.

Paraphrase generation is another way to make a contribution to SMT. It aims at reducing out-of-vocabulary words and acquiring paraphrases of unknown phrases to increase the model coverage [15]. Some of the previous work showed that word lattices constructed to express input sentences in different ways are helpful for obtaining better translation quality [16]. A syntax-based algorithm to automatically build word lattices that are used as finite state automata (FSA) to represent paraphrases is described in Ref. [17]. FSAs extract paraphrase pairs and generate new, unseen sentences that contain the same meaning as the input sentences. In our work, we also generate unseen, new sentence pairs (i.e., they do not come from given parallel sentences). However, FSAs are replaced by the resolution of analogical equations to produce new sentences.

Research is growing on analogical learning for NLP applications. In Ref. [18], they show how to retrieve all analogies for a given word (i.e., a sequence of letters) in a very fast way, so as to allow the application of analogy to practical tasks. In Ref. [19], they present a theoretical generalization of analogies between sequences of letters. They show how to extend elementary analogies between letters of the alphabet to sequences of letters (e.g., $a : b :: c : d$ and $a : \varepsilon :: a : \varepsilon$ imply $aaa : bb :: cca : dd$) based on an edit distance given in Refs. [20], [21]. In Ref. [22], they use proportional analogies to translate sentences in an example-based machine translation. Translation of unknown words by analogy has also been proposed in Refs. [23], [24]. In Ref. [25], they present the basic steps of analogical learning and a definition of formal analogical relationships suitable for learning large datasets in NLP, and use this approach in morphological analysis tasks. Different from these works, in our research, we propose to cluster monolingual Chinese and Japanese short sentences respectively using analogical associations. This allows us to obtain *rewriting models* that can produce new sentences by solving analogical equations.

In Ref. [26], they introduce the basic idea of automatic MT evaluation method by using N-gram co-occurrence statistics. And in Ref. [27], they describe a framework by using N-gram co-occurrence statistics as an automatic evaluation of NLP applications. To cut down on over-generation, we use filtering by seen N-sequences [28] or using BLEU [29] to keep only those newly

[*4] http://tatoeba.org/eng/

generated sentences which are acceptable in fluency of expression and in adequacy of meaning.

In Ref. [30], the same proposed method as the one used here is used for constructing a Chinese–Japanese quasi-parallel corpus based on a scientific corpus (ASPEC). A quasi-parallel corpus is constructed based on the short sentence pairs in ASPEC corpus with less than 30 characters. These newly generated quasi-parallel sentences are used in addition to a part or the entirety of the ASPEC corpus to train SMT systems. Significant improvements are obtained compared with baseline systems. The difference with the work presented here lies in the type of seed sentences. The experimental data used in the present paper are more general, and do not belong to the same genre. The present paper demonstrates the generality of the method.

To compare with Ref. [31], we test a new filtering method based on BLEU in addition to the N-sequence filtering method. In the present paper we perform experiments with 7 sets of quasi-parallel corpora as additional training data added to the baseline training data. In Ref. [31] there was only one quasi-parallel corpus constructed as additional data. In the present paper we perform a range of extensive experiments and analyze more evaluation results. We also provide more insight as to reasons for the success of the method by giving and analyzing statistics concerning the phrase pairs used in decoding.

## 3. Overview of the Proposed Method

In this section, we present our proposed method to construct a Chinese–Japanese quasi-parallel corpus by using analogical associations. The overview of our method is given in **Fig. 1**. The procedure in our method has four steps:

(1) Construction of analogical clusters.

In this step, we cluster large amounts of short sentences collected from the Web in both Chinese and Japanese independently. These clusters are groups of sentence pairs with the same exchanges. We find corresponding Chinese and Japanese clusters with similar exchanges by computing the similarity. Such corresponding clusters can be considered as *rewriting models* that allow us to generate new sentences.

(2) Generation of new sentences.

In this step, we generate new sentences using these *rewriting models* from an existing small amount of Chinese–Japanese parallel sentences, called *seed sentences*.

(3) Filtering over-generated sentences.

In this step, we filter out dubious newly generated sentences and keep only the well-formed sentences using BLEU and N-sequence methods.

(4) Deduction of translation relations.

In this step, finally, we deduce translation relations between the filtered new sentences and construct a quasi-parallel corpus based on the existing parallel corpus and the corresponding clusters. Adding such quasi-parallel corpora to the training data leads to improvements in translation quality.

## 4. Clustering and Generation of New Sentence

In this section, we describe the first two steps of our proposed method: construction of analogical clusters and generation of new sentences.

### 4.1 Construction of Analogical Clusters

(1) Sentential analogies:

References [32], [33] and [34] gave different definitions of proportional analogies. The common notion is that proportional analogies establish a structural relationship between four objects, $A$, $B$, $C$ and $D$. It is written $A : B :: C : D$ ('$A$ is to $B$ as $C$ is to $D$').

Analogies can be classified as being semantical or formal. An

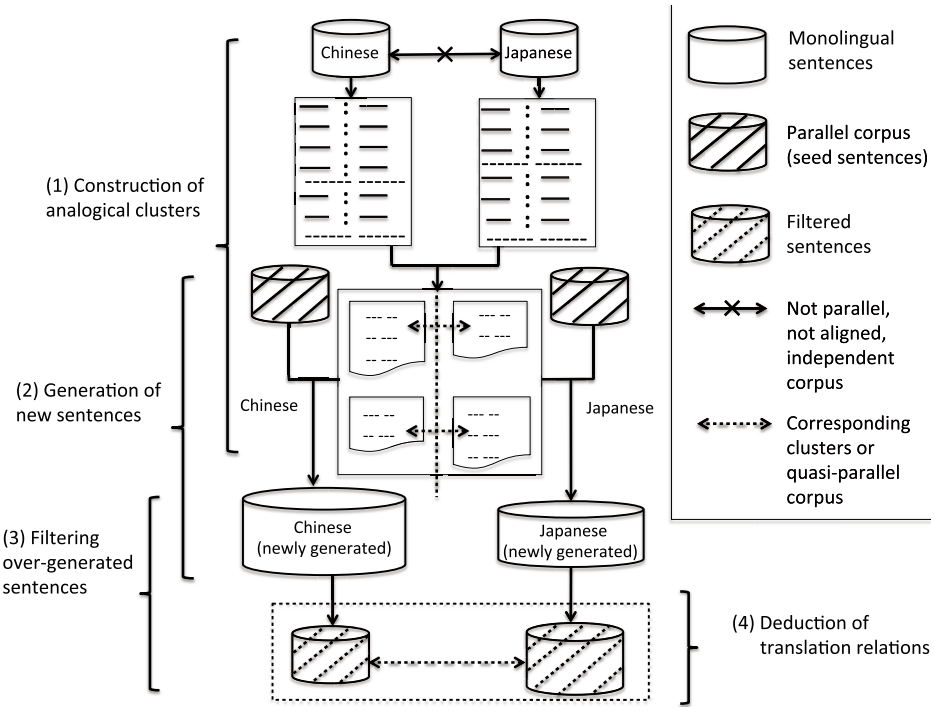



**Fig. 1** Overview of the proposed method: construction of a Chinese–Japanese quasi-parallel corpus.

example of semantic analogy is:

*traffic : street :: water : riverbed*

For such semantic analogy, Ref. [35] gives a definition of verbal analogies based on high relational similarity.

On the other hand, an example of formal analogy is:

*walk : walked :: work : worked*

We use the same notion to cluster sentences. In *sentential analogies*, the changes between the first and second sentences are the same as between the third and fourth sentences, as in

*I like music. : Do you go to concert? :: I like classical music. : Do you go to classical concert?*

An efficient algorithm for the resolution of analogical equations between strings of characters has been proposed by Ref. [20]. The algorithm relies on counting numbers of occurrences of characters and computing edit distances (with only insertion and deletion 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 [36].

In our research, we group pairs of sentences that constitute proportional analogies in Chinese and Japanese respectively. For instance, the following two pairs of Japanese sentences are said to form a *sentential 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: $d(A,B) = d(C,D) = 6$ and $d(A,C) = d(B,D) = 8$. The equality which deals with the number of occurrences of characters, which must be valid for each character is met. It may be illustrated for the character 迷: 1 (in $A$) − 1 (in $B$) = 0 (in $C$) − 0 (in $D$). An interpretation of the analogy is that the word '本当に' (really) is substituted for 'とても' (very).

本当に迷惑です。 : とても迷惑です。 :: 本当に困っています。 : とても困っています。
*It's really annoying.* : *It's very annoying.* :: *I'm really troubled.* : *I'm very troubled.*

(2) Analogical clusters:

When several sentential analogies involve the same pairs of sentences, they form a series of analogous sentences. They can be written on a sequence of lines where each line contains one sentence pair and any two pairs of sentences form a sentential analogy. We call this an *analogical cluster*. The size of a cluster is the number of its sentential pairs (=lines). The clusters contain at least 2 pairs of sentences. **Figure 2** and **Figure 3** show two examples of clusters in Japanese.

(3) Determining corresponding clusters:

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) [37].
- Then, we consider the changes (see $L_{zh} : R_{zh}$ and $L_{ja} : R_{ja}$ in **Fig. 4**) between the left ($S_{left}$) and the right ($S_{right}$) sides in one cluster as two sets. We perform word segmentation 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 [*5], a traditional-simplified Chinese variant table [*6] and a Kanji-Hanzi Conversion Table [*7] 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:

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

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

$$\mathrm{Sim}_{C_{zh}-C_{ja}} = \frac{1}{2}(\mathrm{Sim}_{left} + \mathrm{Sim}_{right}) \tag{2}$$

Application on the example given in Fig. 4:
(knowing クラシック=经典, とても=很 and いい=不错)

$$\begin{aligned}\mathrm{Sim}_{C_{zh}-C_{ja}} &= \frac{1}{2}\left(\frac{2 \times |\{\text{クラシック=经典}\}|}{|\{\text{经典}\}| + |\{\text{クラシック}\}|} + \frac{2 \times |\{\text{とても=很, いい=不错}\}|}{|\{\text{很, 不错}\}| + |\{\text{この, は, とても, いい}\}|}\right) \\ &= \frac{1}{2}\left(\frac{2 \times 1}{1+1} + \frac{2 \times 2}{2+4}\right) \\ &= \frac{1}{2}\left(1 + \frac{2}{3}\right) \\ &= 0.833\end{aligned}$$

表示されません : 表示されなくなりました
*'Is not displayed'* *'No longer able to be displayed'*

ブログが投稿できません : ブログが投稿できなくなりました
*'Cannot post on blog'* *'No longer able to post on blog'*

記事の編集ができません : 記事の編集ができなくなりました
*'Cannot edit the article'* *'No longer able to edit the article'*

⋮ : ⋮

**Fig. 3** An example of an analogical cluster in Japanese exhibiting the exchange of "ません" with "なくなりました".

本当に迷惑です。 : とても迷惑です。
*'It's really annoying.'* *'It's very annoying.'*
本当に困っています。 : とても困っています。
*'I'm really troubled.'* *'I'm very troubled.'*
本当に迷惑しています。 : とても迷惑しています。
*'I'm really in trouble.'* *'I'm in a deep trouble.'*
⋮ : ⋮

**Fig. 2** An example of an analogical cluster in Japanese exhibiting the exchange of "本当に" with "とても".

[*5] The EDR Electronic Dictionary: National Institute of Information and Communication Technology (NICT). URL: http://www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html
[*6] http://www.unicode.org/Public/UNIDATA/
[*7] http://www.kishugiken.co.jp/cn/code10d.html

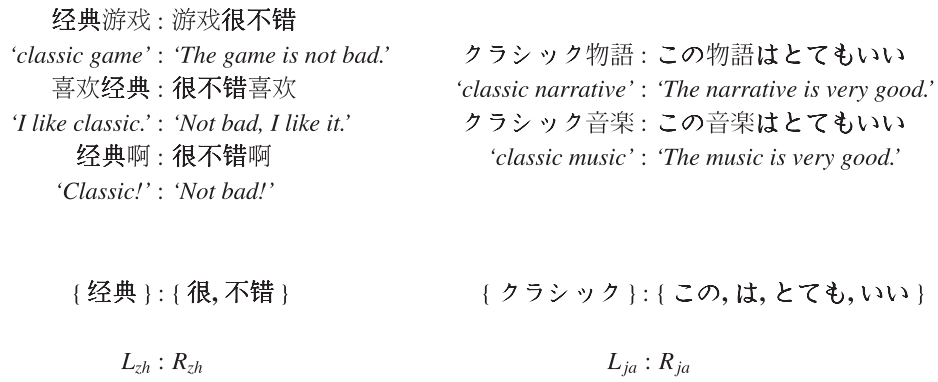



**Fig. 4** An example of a real case of changes between the left and the right sides in Chinese ($L_{zh} : R_{zh}$) and Japanese clusters ($L_{ja} : R_{ja}$). The characters/words in bold face show the changes between the left and right sides of each sentence pair in the clusters and the minimal sets of changes in Chinese or Japanese cluster after segmentation. Note that the sentences in Japanese are not translations of the sentences in Chinese.

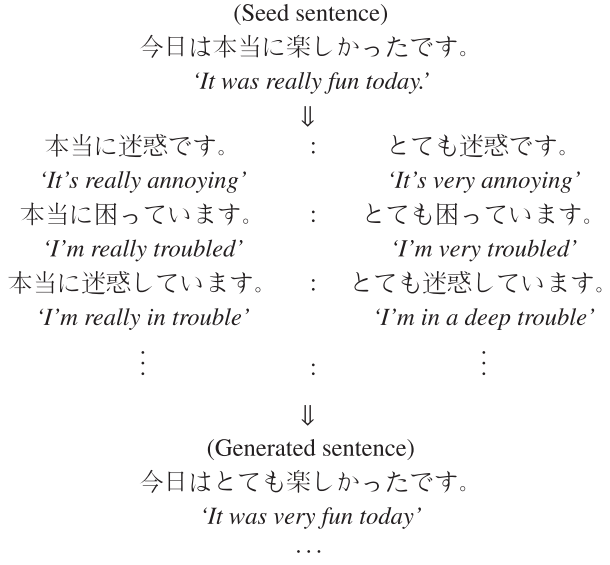



**Fig. 5** An example of sentence generation result (valid sentence).

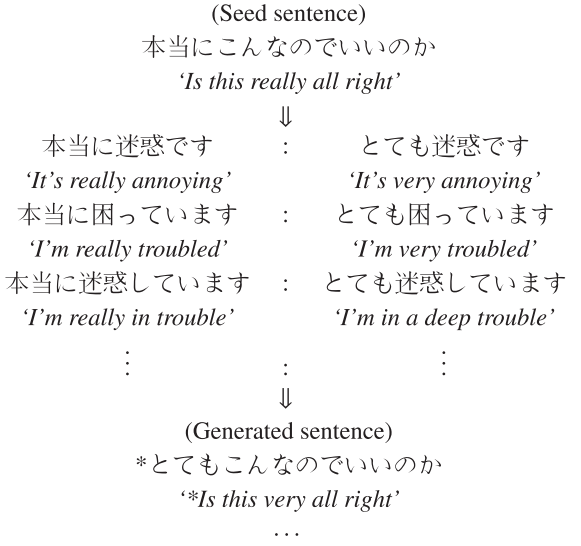



**Fig. 6** An example of sentence generation result (invalid sentence).

Such corresponding clusters can be considered as *rewriting models* that can be used to generate new sentences. The larger the size of a cluster, the more productive it is.

### 4.2 Generation of New Sentences

Analogy is also a process [38] by which, given two related forms and only one form, the fourth missing form is coined [39]. In our work, in a sentential analogy $A : B :: C : D$, a cluster provides $A$ and $B$ and we use a seed sentence $C$ to generate a new candidate sentence $D$. The generated $D$ should satisfy the conditions given above on edit distance and number of occurrences of characters. This can be illustrated with the following example:

本当に迷惑です。 : とても迷惑です。 ::
今日は本当に楽しかったです。 : $x$
⇒ $x$ = 今日はとても楽しかったです。

In this example, the solution of the analogical equation is $D$ = 今日はとても楽しかったです。 *'It was very fun today.'* It should be said that there may exist no solution to an analogical equation, so that a new candidate is not coined each time.

**Figure 5** and **Figure 6** are two examples of sentence generation in Japanese. In the case of Fig. 6, we generate a sentence which is not valid in meaning for a native speaker. To eliminate invalid over-generated sentences and keep only well-formed sentences, a filtering step is needed.

## 5. Filtering Techniques for Quasi-parallel Corpus Construction

To filter out semantically or grammatically invalid sentences and keep only well-formed sentences, we make use of a BLEU-based filtering method and an N-sequence filtering method.

### 5.1 BLEU Based Filtering Method

BLEU is the main evaluation metric for automatic MT [29]. It compares a candidate sentence output from an MT system to possibly refer sentences. The formula of BLEU we use is as follows:

$$\text{BLEU} = BP \times \sqrt{\prod_{n=1}^{N} p_n} \tag{3}$$

$p_n$ stands for modified n-gram precision. It is the core of the calculation of BLEU. $p_n$ calculates the precision from 1-gram to 4-gram. Different from the normal N-gram precision, modified N-gram precision counts N-grams in the references and clips the

count of the same number of N-grams in the candidate sentence to give a lower score to repeated words or phrases. The geometric average of the $p_n$ is computed as a global score. In addition, in order to lessen the advantage given to short candidates by this global score, it is multiplied by a brevity penalty (*BP*) depending on the length of the candidate and reference sentences.

In the sequel we consider applying BLEU as a filtering method for our work on construction of quasi-parallel corpora. However, the calculation of BLEU is very time consuming, because the quantity of candidate sentences to be filtered is usually very large. A possible solution to this problem is to reduce the size of the reference set used for each candidate sentence. For each candidate sentence, we use a set of reference sentences, and calculate its BLEU score relative to this reference set. By setting a threshold, we will be able to keep candidate sentences with higher BLEU scores and discard any sentence with lower scores. So we propose three steps:

( 1 ) Group seed sentences by similarity;
( 2 ) Build small reference sets for each seed group;
( 3 ) Calculate BLEU score.

The information of the generated sentence consists of the associated seed sentence and the cluster the sentence was generated from. In the first step, to reduce the time for construction of reference sets, we group the seed sentences actually used to generate new sentences by computing their Dice similarity. We make several seed groups in this way.

In the second step, we construct a small reference set for each seed group. We propose and apply a weighting method to weight references by N-grams, and only make use of references with higher weights. The formula is given as follows:

$$\text{R-weight}(f_i) = \frac{\sum_{p\in\hat{T}\cap\hat{f}_i}(-\log c(p)\times|p|)}{\sum_{p\in\hat{T}}(-\log c(p)\times|p|)} \times \frac{\left|\hat{T}\cap\hat{f}_i\right|}{\left|\hat{T}\right|} \times \frac{\left|\hat{T}\cap\hat{f}_i\right|}{\left|\hat{f}_i\right|} \tag{4}$$

$f_i$ is a line in the reference corpus.

$\hat{T}$ is the n-gram representation of the seed group.

$\hat{f}_i$ is the set of N-grams contained in the line of the reference corpus.

$p$ represents an N-gram, $|p|$ is the length of the N-gram.

$-\log c(p)$ is proportional to the self-information of an N-gram.

We extract all 1-grams to 4-grams in each seed group and use these N-grams to compute the weight of each reference sentence. We only take the reference sentences with higher weights to construct a small reference set, so that each seed group will have a specific corresponding reference set.

In the third step, for each candidate sentence, we search the seed sentence used in seed groups. If the seed is found in any seed group, we calculate the BLEU score of the candidate sentence against the corresponding reference set. Only sentences with BLEU scores higher than some given threshold will be kept in this step.

### 5.2 N-sequence Filtering Method

We consider that a generated sentence should be valid if almost all of its sequences of N characters are attested in a reference corpus. The number of non-attested strings that can be tolerated is called the tolerance. In other words, any sentence containing a higher number of non-attested N-sequences of characters than the tolerance will be discarded.

In this paper, we thoroughly test several values of N and tolerance to assess the quality of the sentences kept. Since the experiments are time consuming, we developed a method which makes use of the *shortest absent substring* to output all the filtering results we expect at the same time so as to reduce the overall experiment time. The algorithm is based on the computation of *shortest absent substrings* computed on a representation of the reference corpus into a suffix array [40], [41], [42].

The *shortest absent substring* of a string is the shortest substring that cannot be found in a reference text or corpus. Necessarily, if an N-gram contains one or several *shortest absent substrings*, this N-gram is an absent substring itself.

For example in the sentence “とてもいいのか”, suppose that the 2-gram “て も” and the 1-gram “か” are *shortest absent substrings*. This means that we cannot find “て も” and “か” but can find “て”, “も”, “と て”, “も い” in the reference corpus. By definition, any N-gram which contains “か” or “て も” is also an absent substring. This will be the case for “の か” and “と て も い”.

### 5.3 Differences between the BLEU-based Filtering Method and the N-sequence Filtering Method

The common feature of the proposed BLEU-based method and the N-sequence filtering method is that both of them are based on the precision of N-grams in candidate sentences. The primary difference between these two methods is the length of the N-grams used. The N-grams used in the N-sequence filtering method are relatively long, e.g., 6 characters for Chinese and 7 characters for Japanese in our experiments. However, longer N-grams usually cause a low recall (smaller than 10%) of the valid sentences. The positive aspect is that a very high precision of 99% can be reached.

The purpose of using BLEU as a filtering method is to increase the recall. BLEU uses N-grams from 1 to 4 in length which are relatively shorter than the N-grams used in the N-sequence filtering method. Therefore, we consider that BLEU may help reach a higher precision when keeping sentences with higher scores, and at the same time a reasonable recall by using shorter N-grams. It seems natural to think that the sentences with higher BLEU scores should induce a positive effect on the evaluation results of our SMT systems.

The BLEU method may seem “very ad hoc”, but the kept sentences are not so much similar (or copying) with the seed sentences. Because it just keeps new sentences with a BLEU score higher than a given threshold, there may be many sequences of words which did not appear in the reference set (selected based on seed sentences).

The BLEU method and the N-sequence method use different reference corpora in each language in the filtering steps. Especially for the N-sequence method, we just use the additional monolingual corpus which is not related to seed sentences. It is not an “ad hoc” method because the comparison with the refer-

ence corpus is not based on the seed sentences, but based on all additional reference sentences; many sequences of words do not exist in the seed sentences.

### 5.4 Deduction of Translation Relations

Relying on the similarity of the correspondence between the clusters across languages and the translation relations between the seed sentences, we deduce the translation relations between filtered newly generated sentences. A Chinese sentence and a Japanese sentence are considered translations of one another to a certain extent if they satisfy the following two conditions:

- Their seed sentences are aligned in the parallel corpus;
- They were generated from corresponding clusters.

## 6. Experiments and Evaluation Results

### 6.1 Data Preparation

Chinese and Japanese subtitles sites of movies and TV series have been collected from the Web *Subscene.com* and *Opensubtitles.org* using an in-house Web-crawler and aligned. After cleaning, 106,310 pairs of parallel Chinese–Japanese sentences were obtained.

To build our baseline SMT system, 500 and 1,000 sentence pairs from JEC Basic Sentence Data [*8] were extracted as tuning and testing data. The rest of the 3,804 pairs of sentences were combined with the subtitle corpus and constitute the training data with 110,114 sentence pairs. **Table 1** shows the statistics on the data preparation.

To construct a quasi-parallel corpus, we prepared unaligned unrelated monolingual sentences in each language to construct analogical clusters (**Table 2**). Monolingual resources are collected mainly from the following website: "douban" [*9], "Yahoo China" [*10], and "Yahoo China News" [*11] for Chinese, and "Yahoo! Japan" [*12], "Rakuten Japan" [*13], and "The Mainichi Japan" [*14] for Japanese.

The monolingual part of the training data for the baseline system is also used as the initial data for construction of quasi-parallel corpus. We extract unique Chinese and Japanese sentences from the initial parallel corpus. These sentences are used as seed sentences in the generation of new sentences. The sizes of the monolingual sentences used as the reference data are 1,059,985 for Chinese and 1,074,851 for Japanese.

**Table 1** Statistics on the Chinese–Japanese corpus used for the training, tuning, and test sets in baseline system. The tuning and testing sets are the same in all SMT 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 |
| 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 |

[*8] JEC Basic Sentence Data: http://nlp.ist.i.kyoto-u.ac.jp by Kurohashi-Kawahara Lab., Kyoto University. Released in 2011.
[*9] douban: http://www.douban.com
[*10] Yahoo China: http://cn.yahoo.com Closed in 2013.
[*11] Yahoo China News: http://news.cn.yahoo.com Closed in 2013.
[*12] Yahoo Japan: http://www.yahoo.co.jp/
[*13] Rakuten Japan: http://www.rakuten.co.jp/
[*14] The Mainichi Japan: http://www.mainichi.co.jp/

### 6.2 Experimental Setting

The segmentation toolkits that we use in all experiments are Urheen for Chinese (zh) and Mecab for Japanese (ja) [*15]. We perform all SMT experiments using the standard GIZA++/MOSES pipeline [43] with the default options. Tuning was performed by minimum error rate training [44] using 500 tuning sentence pairs. We trained 5-gram language models on the target part of the training data using the SRILM toolkit [45].

### 6.3 Experiments for Cluster Construction and New Sentence Generation

**Table 3** shows the details of the monolingual data and seed sentences we used and the results of clusters construction and new sentences generation. About 14,578 corresponding clusters were extracted ($\mathrm{Sim}_{C_{zh}-C_{ja}} \geqslant 0.300$) by the steps described in Section 4.1. We checked the quality of the newly generated sentences manually. More than half of the generated sentences were found to be grammatically invalid. This is indicated in the last row of Table 3, where $Q$ stands for the grammatical quality as evaluated by extracting 1,000 sentences randomly and checking them manually.

### 6.4 Experiments for Filtering and Quasi-parallel Corpus Construction for SMT System

#### 6.4.1 Filtering by BLEU Method

We performed BLEU-based filtering experiments with the same candidate sentences as those described in Table 3. We grouped the seed sentences by similarity using the Dice coefficient between sets of words. We extracted all 1-grams to 4-grams in each seed group to weight the references, and only selected 100 reference sentences with highest weight to build reference sets. Each reference set corresponds to a seed group.

After having obtained seed groups (**Table 4**) and corresponding reference sets for each seed group, for each candidate sentence, its seed sentence is identified among the possible seed groups and the BLEU scores of the candidate sentence against the corresponding reference sets are computed. In our experiments, we set several thresholds to check the filtering results (**Table 5**).

Firstly, we kept 1,793,541 Chinese sentences and 1,062,751 Japanese sentences. The highest BLEU scores reached 81 in Japanese and 46 in Chinese. We found that the candidate sentences with high scores are very similar to the seed sentences they are generated from. Most of them only add several characters. Generally the sentences generated from the same seed share a same BLEU score. The reason is that the references we used in BLEU calculation are extracted by seed sentences. For the candidate sentence, small changes in the generation will make it

[*15] Urheen, a Chinese lexical analysis toolkit (Chinese Academy of Sciences, Institute of Automation, CASIA); Mecab, part-of-speech and morphological analyzer: http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.

**Table 2** Statistics on the unaligned Chinese and Japanese monolingual short sentences for construction of analogical clusters.

| | # 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 3** Results of clustering and generation of new sentences.

| | | Chinese | Japanese |
|---|---|---|---|
| Initial data | # of monolingual sentences | 70,000 | 70,000 |
| | # of seed sentences | 99,251 | 90,406 |
| | # of clusters | 23,182 | 21,975 |
| New sentence generation | # of candidate sentences | 221,447,016 $Q$= 20% | 75,278,961 $Q$= 50% |

**Table 4** Statistics of grouping seed sentences.

| | Chinese | Japanese |
|---|---|---|
| # of seeds | 99,251 | 90,406 |
| # of seed groups | 600 | 300 |
| Size of groups | 165 | 301 |

**Table 5** Filtering results by using BLEU-based filtering method.

| Threshold | Chinese | | Japanese | |
|---|---|---|---|---|
| (BLEU%) | # of sentence | $Q_{zh}$ | # of sentence | $Q_{ja}$ |
| > 10 | 13,469 | 95% | 653 | 90% |
| > 5 | 13,570 | 85% | 1,192 | 88% |
| > 4 | 1,471,080 | 75% | 26,164 | 70% |
| > 1 | 1,793,541 | 70% | 1,062,751 | 60% |
| Total | 221,447,016 | 20% | 75,278,961 | 50% |

similar to the seed sentence, and lead to a higher BLEU score in filtering. On average, scores of Chinese sentences are higher than the scores of Japanese sentences, because more Chinese reference sets were built than in Japanese.

Finally, we checked and found that there are about 500,000 unique Japanese filtered sentences after filtering by the BLEU method (threshold > 1 BLEU%). Thus, we only kept the 500,000 unique Japanese filtered sentences with their corresponding seed sentences in higher BLEU scores. The same size of filtered sentences in Chinese with higher BLEU scores are also extracted. Deducing translation relationships allowed us to construct a quasi-parallel corpus of 353,729 sentence pairs. We added the new corpus into the baseline and evaluated it. **Table 6** shows the results.

The BLEU based filtering method increases the baseline system by only 0.8 BLEU points. We reduce the size of the reference corpora and only used grouped seed sentences to weight the reference sentences. It was observed that most of the BLEU scores obtained in the filtering step are around 1, which is close to the improvement obtained in SMT.

#### 6.4.2 Filtering by N-sequence Method

To determine the most appropriate N which can keep the largest number of well-formed sentences to be added to the training corpus, we performed a series of filtering experiments using the N-sequence method with different values of N and tolerance. **Table 7** shows the results for N equal to 4 to 9 and tolerance equal to 0 and 1 in Chinese and Japanese.

We assessed the quality of filtered sentences manually by selecting 1,000 sentences randomly and checked their grammatical quality. With a tolerance of 0, the quality of sentences increased when N increases. We obtained the highest grammatical quality of 99% when N equals 6 characters in Chinese and 7 characters in Japanese with a tolerance of 0. This means that 99% of the sentences kept are grammatically correct. Also, with a tolerance of 0, the quality of sentences with a larger value of N than 6 characters in Chinese and 7 characters in Japanese was kept between 98% and 99%.

The quality decreases in the same value of N when the tolerance increases. Because sentences with a tolerance of 1 may contain an N-gram that cannot be found in the reference corpus, noise creeps into sentences. For that reason, the quality of Chinese kept sentences with N = 6 and the tolerance = 1 decreases down to 89%.

Using these results, we selected 4 sets of filtered sentences in the two languages with high quality obtained using a tolerance of 0. We also selected 3 similar sets with a tolerance of 1. This makes 7 quasi-parallel corpus in total. In each corpus, N (ja) equals N (zh)+1 so as to make the number of filtered sentences comparable. **Table 8** describes the quasi-parallel corpora constructed.

For the 7 quasi-parallel corpora, we added each of them as additional data to our initial Chinese–Japanese training data to perform Chinese-to-Japanese SMT experiments. We recomputed translation tables (training), tuned the system, performed translation of the same test set and calculated the BLEU scores. Table 8 shows the results for each of the SMT systems. All the BLEU scores of the SMT systems with additional data are 1 to 6 points higher than the baseline system. The highest score is obtained when N (zh) = 6 and N (ja) = 7 with a tolerance of 0. Quasi-parallel corpora with a tolerance of 0 contain less noise, and the BLEU scores increase when the size of additional data becomes larger. Therefore, even if the size of quasi-parallel corpora adding data with a tolerance of 1 is much larger than data with a tolerance of 0, because of the noise, it cannot improve the translation results effectively.

Table 8 also shows the BLEU scores obtained on the tuning set when the parameters are optimized on this same tuning set for each SMT system. We vary the filtering parameters. The best system obtained by considering the scores on the tuning set is ob-

**Table 6** Comparison of the baseline SMT system and an SMT system with additional quasi-parallel data output by BLEU-based filtering. The figure in bold characters (13.89) shows a significant improvement with a p-value < 0.01.

| | # of lines (zh) | $Q_{zh}$ | # of lines (ja) | $Q_{ja}$ | Quasi | BLEU |
|---|---|---|---|---|---|---|
| Baseline | 110,114 | - | 110,114 | - | - | 13.10 |
| BLEU filtering | 500,000 | 81% | 500,000 | 65% | 343,729 | **13.89** |

**Table 7** Filtering results by using the N-sequence filtering method in different Ns and tolerances.

| N | Chinese | | | | Japanese | | | |
|---|---|---|---|---|---|---|---|---|
| All | 221,447,016 ($Q$ = 20%) | | | | 75,278,961 ($Q$ = 50%) | | | |
| | Tolerance = 0 | $Q$ | Tolerance = 1 | $Q$ | Tolerance = 0 | $Q$ | Tolerance = 1 | $Q$ |
| 4 | 1,848,254 | 83% | 9,063,117 | 74% | 2,252,589 | 80% | 7,295,155 | 75% |
| 5 | 244,495 | 90% | 1,187,362 | 79% | 474,072 | 89% | 1,668,322 | 77% |
| 6 | 105,537 | 99% | 369,625 | 89% | 312,557 | 92% | 981,429 | 81% |
| 7 | 89,728 | 98% | 237,159 | 87% | 192,124 | 99% | 572,616 | 85% |
| 8 | 86,523 | 98% | 198,077 | 83% | 117,133 | 98% | 286,587 | 88% |
| 9 | 85,690 | 99% | 174,849 | 87% | 98,136 | 99% | 192,586 | 90% |

**Table 8** Construction results of a quasi-parallel corpus by using N-sequence filtering and the evaluation results for Chinese–Japanese baseline system and baseline + additional quasi-parallel systems. The figures in bold characters show a significant improvement with a p-value < 0.01.

| Chinese | | | | Japanese | | | | Quasi-parallel corpus | BLEU% | BLEU% |
|---|---|---|---|---|---|---|---|---|---|---|
| N | Tolerance | Size | $Q_{zh}$ | N | Tolerance | Size | $Q_{ja}$ | # of sentence pairs | (tuning) | (test) |
| 8 | 1 | 198,077 | 83% | 9 | 1 | 192,586 | 90% | 120,338 | 16.97 | **14.31** |
| 7 | 1 | 237,159 | 87% | 8 | 1 | 286,587 | 88% | 163,043 | 17.49 | **14.54** |
| 5 | 0 | 244,495 | 90% | 6 | 0 | 312,557 | 92% | 193,561 | 17.52 | **14.82** |
| 8 | 0 | 86,523 | 98% | 9 | 0 | 98,136 | 99% | 28,733 | 17.91 | **15.70** |
| 6 | 1 | 369,625 | 89% | 7 | 1 | 572,616 | 85% | 276,999 | 18.04 | **15.99** |
| 7 | 0 | 89,728 | 98% | 8 | 0 | 117,133 | 98% | 37,067 | 18.49 | **16.37** |
| 6 | 0 | 105,537 | 99% | 7 | 0 | 192,124 | 99% | 76,151 | 21.18 | **19.27** |
| 6<br>- | 0<br>- | 105,537<br>500,000 | 99%<br>81% | 7<br>- | 0<br>- | 192,124<br>500,000 | 99%<br>65% | 76,151+343,729 | 23.94 | **20.35** |
| | | | | | | | | baseline | 16.08 | 13.10 |

tained for N (zh) = 6 and N (ja) = 7 with a tolerance of 0. We then evaluate this best system with these parameters on a test set. We verify that the score obtained on the test set with these parameters is the best score by evaluating all other systems on the same test set. We confirm that the best configuration obtained by tuning leads to the best score on the test set (see Table 8).

We also build an SMT system based on the quasi-parallel corpora obtained by combining the BLEU and the N-sequence filtering methods. This arrangement yielded even greater improvement (Table 8): a more lenient filtering method (more sentences remain after filtering) is boosting the performance of the more drastic filtering method (less sentences kept). This is shown by a relatively higher than expected increase in translation accuracy as measured by BLEU, as 7.25 > 0.79 + 6.17 = 6.96.

### 6.5 Analysis of the Results

We investigated the $N$ (source length) × $M$ (target length) distribution in phrase tables (used during testing) generated from the initial parallel corpus and the inflated training corpus by adding the constructed quasi-parallel data (filtered with N (zh) = 6 and N (ja) = 7, Tolerance is 0). In **Table 9** and **Table 10**, the statistics (zh→ja) show that the total number of phrase pairs used by adding additional quasi-parallel corpus is larger than when using only the initial parallel corpus as training data, especially for 1-4 grams in both languages. If we compare the number of entries, the number of phrase pairs (in Table 10) on the diagonal got a significant increase in the number of phrase pairs of similar length. Considering the correspondence between lengths in Chinese–Japanese translation, the increase in phrase pairs with different lengths (like 1 (zh) × 2 (ja), 2 (zh) × 3 (ja) and 3 (zh) × 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 potentially useful alignments.

**Table 11** illustrates the fact that new translation candidates have been added for an existing phrase, and that new phrase pairs have also been added. The fact that these additional phrases are reasonable is indicated by the improvements in BLEU scores. **Table 12** illustrates changes in lexical weights and translation probabilities for the same Chinese phrase. More accurate phrase alignments may be extracted by adding additional quasi-parallel corpus. We also believe that we improved its features by adding quasi-parallel data.

## 7. Conclusion and Future Work

We presented a different way for automatic acquisition of rewriting models for the construction of a quasi-parallel corpus. The reason for constructing quasi-parallel corpora to be added to training data in SMT, is to extract new additional translation knowledge from unrelated unaligned monolingual data. Quasi-parallel corpora are used as additional training data to train SMT systems and in this way improves translation quality. The experimental data we use are collected from Websites with open

**Table 9** Distribution of phrase pairs used during testing in Chinese-to-Japanese SMT experiment (baseline).

| | | Target = Japanese | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1-gram | 2-gram | 3-gram | 4-gram | 5-gram | 6-gram | 7-gram | total |
| Source = Chinese | 1-gram | 10,833 | 21,570 | 16,142 | 9,042 | 4,360 | 1,899 | 780 | 64,626 |
| | 2-gram | 3,318 | 5,938 | 4,911 | 2,789 | 1,402 | 678 | 289 | 19,325 |
| | 3-gram | 217 | 400 | 426 | 288 | 168 | 81 | 32 | 1,612 |
| | 4-gram | 14 | 29 | 33 | 37 | 33 | 16 | 10 | 172 |
| | 5-gram | 1 | 3 | 4 | 7 | 8 | 10 | 10 | 43 |
| | 6-gram | 0 | 0 | 3 | 3 | 5 | 6 | 8 | 25 |
| | 7-gram | 0 | 0 | 1 | 2 | 2 | 2 | 1 | 8 |
| | total | 14,383 | 27,940 | 21,520 | 12,168 | 5,978 | 2,692 | 1,130 | 85,811 |

**Table 10** Distribution of phrase pairs used during testing in Chinese-to-Japanese SMT experiment (baseline + quasi-parallel). The bold numbers show the increased numbers of $N$ (Chinese) × $M$ (Japanese)-grams (less than 4-gram) in the phrase table, and the total number of $N$ (Chinese) × $M$ (Japanese)-grams, which increased compared with the baseline system.

| | | Target = Japanese | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1-gram | 2-gram | 3-gram | 4-gram | 5-gram | 6-gram | 7-gram | total |
| Source = Chinese | 1-gram | **11,337** | **24,456** | **20,285** | **12,559** | 6,655 | 3,197 | 1,492 | 79,981 |
| | 2-gram | **3,971** | **7,081** | **6,422** | **4,131** | 2,370 | 1,213 | 565 | 25,753 |
| | 3-gram | **316** | **541** | **604** | **463** | 332 | 209 | 132 | 2,597 |
| | 4-gram | **36** | **50** | **82** | **109** | 88 | 70 | 58 | 493 |
| | 5-gram | 4 | 11 | 24 | 39 | 46 | 44 | 50 | 218 |
| | 6-gram | 2 | 2 | 6 | 17 | 21 | 24 | 45 | 117 |
| | 7-gram | 0 | 0 | 1 | 7 | 11 | 16 | 24 | 59 |
| | total | **15,666** | **32,141** | **27,424** | **17,325** | 9,523 | 4,773 | 2,366 | **109,218** |

**Table 11** Samples of phrase alignments in zh→ja phrase table. Same Chinese phrase and corresponding Japanese phrases in baseline phrase table and baseline + quasi-parallel phrase table.

| | zh | | ja | |
|---|---|---|---|---|
| Baseline | 只 能 这样 了 | ||| | これ で 行く しか ない | ||| |
| | *'it can only be this way'* | ||| | *'no choice but to go'* | ||| |
| | 只 能 这样 了 | ||| | それ しか ない ん だ よ | ||| |
| | *'it can only be this way'* | ||| | *'that's all I have to do'* | ||| |
| | 只 能 这样 了 | ||| | やる しか ない だろ | ||| |
| | *'it can only be this way'* | ||| | *'I only have to do it'* (presumption) | ||| |
| Additional | 只 能 这样 了 | ||| | これ で 行く しか なかっ た (new) | ||| |
| | *'it can only be this way'* | ||| | *'no choice but to go'* (past form) | ||| |
| | 只 能 这样 了 | ||| | これ しか ない (new) | ||| |
| | *'it can only be this way'* | ||| | *'nothing but this'* | ||| |
| | 只 能 这样 了 | ||| | それ しか ない ん だ (new) | ||| |
| | *'it can only be this way'* | ||| | *'that's all I have to do'* (without modal particle) | ||| |
| | 只 能 这样 了 | ||| | やる しか ない (new) | ||| |
| | *'it can only be this way'* | ||| | *'I only have to do it'* | ||| |
| | 只 能 这样 了 | ||| | やる しか なかっ た (new) | ||| |
| | *'it can only be this way'* | ||| | *'I only have to do it'* (past form) | ||| |
| | 主要 画 游戏 (new) | ||| | 主 に ゲーム (new) | ||| |
| | *'mainly draw games'* | ||| | *'mainly games'* | ||| |

**Table 12** Samples of the same phrase alignments in Chinese and Japanese with different translation probabilities (TP) and lexical weights (LW) in baseline phrase table and baseline + quasi-parallel phrase table.

| | zh (s) | | ja (t) | | TP(s\|t) | LW(s\|t) | TP(t\|s) | LW(t\|s) |
|---|---|---|---|---|---|---|---|---|
| in baseline phrase table | 実际上 | ||| | 実は | ||| | 0.148 | 0.103 | 0.310 | 0.138 |
| | *'actually'* | ||| | *'actually'* | ||| | - | - | - | - |
| | 実际上 | ||| | 実際 | ||| | 0.314 | 0.303 | 0.379 | 0.288 |
| | *'actually'* | ||| | *'actually'* (saying in different way) | ||| | - | - | - | - |
| | 実际上 | ||| | 実際 に | ||| | 0.335 | 0.152 | 0.139 | 0.043 |
| | *'actually'* | ||| | *'actually'* (saying in different way) | ||| | - | - | - | - |
| | 実际上 | ||| | 実際 に は | ||| | 0.089 | 0.101 | 0.003 | 0.002 |
| | *'actually'* | ||| | *'actually'* (saying in different way) | ||| | - | - | - | - |
| in baseline + additional phrase table | 实际上 | ||| | 実は | ||| | 0.182 | 0.060 | 0.290 | 0.073 |
| | 实际上 | ||| | 実際 | ||| | 0.397 | 0.313 | 0.362 | 0.230 |
| | 实际上 | ||| | 実際 に | ||| | 0.053 | 0.157 | 0.012 | 0.034 |
| | 实际上 | ||| | 実際 に は | ||| | 0.867 | 0.105 | 0.188 | 0.005 |

licences [*16,*17] with the concern of avoiding any copyright problem. We produced all possible analogical clusters as rewriting models for generating new sentences, then filtered newly overgenerated sentences by a BLEU-based method and N-sequence method. We envisage the release of the quasi-parallel corpus constructed in our experiments.

We improved the computational efficiency of the basic N-sequence filtering method so that we could add a new parameter, tolerance, as an attempt at relaxing the constraint. We performed a series of filtering experiments with different values of N and tolerance. The algorithm could save processing time when we use more than 2 different values of N and tolerance. To make use of shorter N-grams, we proposed a new filtering method based on BLEU. Facing the problem of time, we applied a weighting method to decrease the size of the reference corpus and used similarity computation to group seed sentences to reduce the processing time.

We conducted a series of experiments and constructed several quasi-parallel corpora using different filtering results and added them to a baseline SMT system. We obtained increases of 0.8 BLEU point with the BLEU filtering method and 1 to 6 BLEU points in experiments using the N-sequence filtering method. We are able to conclude that better sentence quality and larger sizes of additional quasi-parallel corpora lead to higher scores in translation evaluation. We also combined quasi-parallel corpora obtained by using the BLEU and the N-sequence filtering methods as additional training data to train an SMT system. In this way we achieved an even better improvement than expected in translation accuracy as measured by BLEU.

**Acknowledgments** This work was supported in part by a JSPS Grant, Number 15K00317 (Kakenhi C), entitled Language productivity: efficient extraction of productive analogical clusters and their evaluation using statistical machine translation.

## References

[1] Koehn, P.: Europarl: A parallel corpus for statistical machine translation, *MT summit*, Vol.5, pp.79–86 (2005).

[2] Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. and Varga, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, arXiv preprint cs/0609058 (2006).

[3] Sakti, S., Vu, T., Finch, A., Paul, M., Maia, R., Sakai, S., Hayashi, T., Kimura, N., Ashikari, Y., Sumita, E., et al.: NICT/ATR Asian spoken language translation system for multi-party travel conversation, *Proc. TCAST Workshop*, pp.26–30 (2009).

[4] Zhang, Y., Uchimoto, K., Ma, Q. and Isahara, H.: Building an Annotated Japanese-Chinese Parallel Corpus–A Part of NICT Multilingual Corpora, *Proc. 10th Machine Translation Summit* (*MT Summit X*), pp.71–78 (2005).

[5] Yang, M., Jiang, H., Zhao, T. and Li, S.: Construct trilingual parallel corpus on demand, *Chinese Spoken Language Processing*, pp.760–767, Springer (2006).

[6] Nakazawa, T., Mino, H., Goto, I., Kurohashi, S. and Sumita, E.: Overview of the 1st Workshop on Asian Translation, *Proc. 1st Workshop on Asian Translation* (*WAT2014*), pp.1–19 (2014).

[7] Zaidan, O.F. and Callison-Burch, C.: Crowdsourcing translation: Professional quality from non-professionals, *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp.1220–1229, Association for Computational Linguistics (2011).

[8] Munteanu, D.S. and Marcu, D.: Improving machine translation performance by exploiting non-parallel corpora, *Computational Linguistics*, Vol.31, No.4, pp.477–504 (2005).

*16 Subscence.com: https://subscene.com/site/legal-information

*17 Opensubtitles.org: http://www.opensubtitles.org/ja/disclaimer

[9] Munteanu, D.S. and Marcu, D.: Extracting parallel sub-sentential fragments from non-parallel corpora, *Proc. 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pp.81–88, Association for Computational Linguistics (2006).

[10] Bin, L., Jiang, T., Chow, K. and Benjamin, K.T.: Building a large English-Chinese parallel corpus from comparable patents and its experimental application to SMT, *Proc. 3rd Workshop on Building and Using Comparable Corpora*, pp.42–49 (2010).

[11] Smith, J.R., Quirk, C. and Toutanova, K.: Extracting parallel sentences from comparable corpora using document level alignment, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp.403–411, Association for Computational Linguistics (2010).

[12] Chu, C., Nakazawa, T. and Kurohashi, S.: Integrated Parallel Sentence and Fragment Extraction from Comparable Corpora: A Case Study on Chinese–Japanese Wikipedia, *ACM Trans. Asian and Low-Resource Language Information Processing*, Vol.15, No.2, p.10 (2015).

[13] Fung, P. and Cheung, P.: Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus, *Proc. 20th International Conference on Computational Linguistics*, p.1051, Association for Computational Linguistics (2004).

[14] Chu, C., Nakazawa, T. and Kurohashi, S.: Accurate Parallel Fragment Extraction from Quasi-Comparable Corpora using Alignment Model and Translation Lexicon, *IJCNLP*, pp.1144–1150 (2013).

[15] Jiang, J., Du, J. and Way, A.: Incorporating source-language paraphrases into phrase-based SMT with confusion networks, *Proc. 5th Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp.31–40, Association for Computational Linguistics (2011).

[16] Onishi, T., Utiyama, M. and Sumita, E.: Paraphrase lattice for statistical machine translation, *IEICE Trans. Inf. Syst.*, Vol.94, No.6, pp.1299–1305 (2011).

[17] Pang, B., Knight, K. and Marcu, D.: Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences, *Proc. 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp.102–109, Association for Computational Linguistics (2003).

[18] Langlais, P. and Yvon, F.: Scaling up Analogical Learning, *Coling 2008: Companion volume: Posters*, pp.51–54 (2008).

[19] Delhay, A. and Miclet, L.: Analogical Equations in Sequences: Definition and Resolution, *Lecture Notes in Computer Science*, Vol.3264, pp.127–138 (2004).

[20] Lepage, Y.: Solving analogies on words: An algorithm, *Proc. COLING-ACL '98*, pp.728–735 (1998).

[21] Pirrelli, V. and Yvon, F.: Analogy in the lexicon: a probe into analogy-based machine learning of language, *Proc. 6th International Symposium on Human Communication*, Santiago de Cuba, Cuba (1999).

[22] Lepage, Y. and Denoual, E.: Purest ever example-based machine translation: Detailed presentation and assessment, *Machine Translation*, Vol.19, pp.251–282 (2005).

[23] Langlais, P. and Patry, A.: Translating Unknown Words by Analogical Learning, *EMNLP-CoNLL*, pp.877–886 (2007).

[24] Silva, J., Coheur, L., Costa, Â. and Trancoso, I.: Dealing with unknown words in statistical machine translation, *Proc. 8th International Conference on Language Resources and Evaluation* (*LREC '12*), pp.3977–3981 (2012).

[25] Stroppa, N. and Yvon, F.: An analogical learner for morphological analysis, *Proc. 9th Conference on Computational Natural Language Learning* (*CoNLL 2005*), pp.120–127, Ann Arbor, MI (2005).

[26] Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, *Proc. Human Language Technology Conference* (*HLT2002*), pp.128–132 (2002).

[27] Soricut, R. and Brill, E.: A unified framework for automatic evaluation using N-gram co-occurrence statistics, *Proc. 42nd Annual Meeting on Association for Computational Linguistics*, p.613, Association for Computational Linguistics (2004).

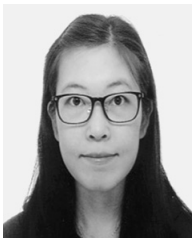
[28] Lepage, Y. and Denoual, E.: Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation, *Proc. 3rd Int. Workshop on Paraphrasing*, pp.57–64 (2005).

[29] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: A method for automatic evaluation of machine translation, *ACL 2002*, pp.311–318 (2002).

[30] Yang, W., Zhao, Z. and Lepage, Y.: Inflating Training Data for Statistical Machine Translation using Unaligned Monolingual Data, *The Association for Natural Language Processing*, pp.1016–1019 (2015).

[31] Yang, W. and Lepage, Y.: Inflating a training corpus for SMT by using unrelated unaligned monolingual data, *International Conference on Natural Language Processing*, pp.236–248, Springer (2014).

[32] Gentner, D.: Structure-Mapping: A Theoretical Framework for Analogy, *Cognitive Science*, Vol.7, No.2, pp.155–170 (1983).
[33] Lepage, Y.: Analogy and formal languages, *Electronic Notes in Theoretical Computer Science*, Vol.53, pp.180–191 (2004).
[34] Yvon, F., Stroppa, N., Delhay, A. and Miclet, L.: Solving analogical equations on words, *Rapport interne D*, Vol.5 (2004).
[35] Turney, P.D.: Similarity of semantic relations, *Computational Linguistics*, Vol.32, No.3, pp.379–416 (2006).
[36] Allison, L. and Dix, T.I.: A bit-string longest-common-subsequence algorithm, *Inf. Process. Lett.*, Vol.23, No.5, pp.305–310 (1986).
[37] Wagner, R.A. and Fischer, M.J.: The string-to-string correction problem, *Journal of the ACM* (*JACM*), Vol.21, No.1, pp.168–173 (1974).
[38] Itkonen, E.: *Analogy as Structure and Process: Approaches in linguistics, cognitive psychology and philosophy of science*, Vol.14 (2005).
[39] de Saussure, F.: *Cours de linguistique générale*, Payot, Lausanne et Paris, [1ère éd. 1916] edition (1995).
[40] Nagao, M. and Mori, S.: A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese, *Proc. 15th Conference on Computational Linguistics-Volume 1*, pp.611–615, Association for Computational Linguistics (1994).
[41] Yamamoto, M. and Church, K.W.: Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus, *Computational Linguistics*, Vol.27, No.1, pp.1–30 (2001).
[42] Kärkkäinen, J. and Sanders, P.: Simple linear work suffix array construction, *Automata, Languages and Programming*, pp.943–955, Springer (2003).
[43] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: Open source toolkit for statistical machine translation, *Proc. 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (*ACL 2007*), pp.177–180, Association for Computational Linguistics (2007).
[44] Och, F.J.: Minimum error rate training in statistical machine translation, *Proc. 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pp.160–167, Association for Computational Linguistics (2003).
[45] Stolcke, A. et al.: SRILM–An extensible language modeling toolkit, *Proc. ICSLP*, Vol.2, pp.901–904 (2002).



**Wei Yang** received her Master Degree in 2012 from Waseda University, Graduate School of Information, Production and Systems. During her Master Course, her research interests in combining several automatic techniques to build Chinese-Japanese lexicon from freely available resources and make them free available for users and researchers in Natural Language Processing and Machine Translation. She is currently a Ph.D. candidate at the Waseda University, Graduate School of Information, Production and Systems. Her research interests are in Natural Language Processing, Machine Translation, especially between Chinese and Japanese.

**Hanfei Shen** received her Master Degree in 2015 from Waseda University, Graduate School of Information, Production and Systems. During her Master Course, she worked on improving translation accuracy of statistical machine translation between Chinese and Japanese based on a small-scale parallel corpus using analogical associations and filtering methods. She started to work at NEC as a system engineer from 2015.

**Yves Lepage** received his Ph.D. degree from GETA, Grenoble university, France. He worked for ATR labs, Japan, as an invited researcher and a senior researcher until 2006. He joined Waseda University, Graduate School of Information, Production and Systems in April 2010. He is a member of the Information Processing Society of Japan, the Japanese Natural Language Processing Association, and the French Natural Language Processing Association, ATALA. He was editor-in-chief of the French journal on Natural Language Processing, TAL, from 2008 to 2016.