Integrated Parallel Sentence and Fragment Extraction from Comparable Corpora: A Case Study on Chinese–Japanese Wikipedia

CHENHUI CHU, Kyoto University, Japan Society for the Promotion of Science Research Fellow
TOSHIAKI NAKAZAWA, Japan Science and Technology Agency
SADAO KUROHASHI, Kyoto University

Parallel corpora are crucial for statistical machine translation (SMT), however they are quite scarce for most language pairs and domains. As comparable corpora are far more available, many studies have been conducted to extract either parallel sentences or fragments from them for SMT. In this article, we propose an integrated system to extract both parallel sentences and fragments from comparable corpora. We first apply parallel sentence extraction to identify parallel sentences from comparable sentences. We then extract parallel fragments from the comparable sentences. Parallel sentence extraction is based on a parallel sentence candidate filter and classifier for parallel sentence identification. We improve it by proposing a novel filtering strategy and three novel feature sets for classification. Previous studies have found it difficult to accurately extract parallel fragments from comparable sentences. We propose an accurate parallel fragment extraction method that uses an alignment model to locate the parallel fragment candidates and uses an accurate lexicon-based filter to identify the truly parallel fragments. A case study on the Chinese–Japanese Wikipedia indicates that our proposed methods outperform previously proposed methods, and the parallel data extracted by our system significantly improves SMT performance.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Machine translation

General Terms: Languages, Experimentation, Performance

Additional Key Words and Phrases: Integrated system, parallel sentence, parallel fragment, comparable corpora

ACM Reference Format:
Chu, C., Nakazawa, T., and Kurohashi, S. 2014. Integrated Parallel Sentence and Fragment Extraction from Comparable Corpora: A Case Study on Chinese–Japanese Wikipedia. *ACM Trans. Asian Lang. Inform. Process.* V, N, Article A (January YYYY), 22 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

In statistical machine translation (SMT) [Brown et al. 1993; Och and Ney 2003; Koehn 2010], because translation knowledge is acquired from parallel corpora (sentence-aligned bilingual texts), the quality and quantity of parallel corpora are crucial. However, currently, high quality parallel corpora of sufficient size are only available for a few language pairs such as languages paired with English and several European language pairs. Moreover, even for these language pairs, the available domains are

This work is supported by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for JSPS Fellows.

Author addresses: C. Chu (corresponding author), and S. Kurohashi, Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan; T. Nakazawa, Japan Science and Technology Agency, Honcho, Kawaguchi-shi, Saitama 332-0012, Japan; email: chu@nlp.ist.i.kyoto-u.ac.jp, nakazawa@pa.jst.jp, kuro@i.kyoto-u.ac.jp.

DOI:http://dx.doi.org/10.1145/0000000.0000000

limited. For the rest, comprising the majority of language pairs and domains, only few or no parallel corpora are available. This scarceness of parallel corpora has become the main bottleneck for SMT.

Comparable corpora are a set of monolingual corpora that describe roughly the same topic in different languages, but are not exact translation equivalents of each other. Exploiting comparable corpora for SMT is the key to addressing the scarceness of parallel corpora. The reason for this is that comparable corpora are far more available than parallel corpora, and there is a large amount of parallel data contained in the comparable texts.

Previous studies proposed extracting either parallel sentences [Munteanu and Marcu 2005; Smith et al. 2010; Ştefănescu and Ion 2013; Chu et al. 2014] or fragments [Munteanu and Marcu 2006; Quirk et al. 2007; Aker et al. 2012; Chu et al. 2013b] from comparable corpora based on the comparability of the corpora. The assumption in previous studies is that in comparable corpora with high comparability, there are many parallel sentences, and thus previous studies only focus on parallel sentence extraction from this kind of corpora [Munteanu and Marcu 2005; Smith et al. 2010; Ştefănescu and Ion 2013; Chu et al. 2014]. Howerver, in comparable corpora with low comparability, there are few or no parallel sentences, only parallel fragments in comparable sentences, and thus parallel fragment extraction is more appropriate [Munteanu and Marcu 2006; Quirk et al. 2007; Aker et al. 2012; Chu et al. 2013b; Gupta et al. 2013].[1]

One important fact that most previous studies ignore is that there could be both parallel sentences and fragments in many comparable corpora.[2] Wikipedia is one typical example of such comparable corpora. In Wikipedia, articles in different languages on the same topic are manually aligned via interlanguage links by the authors, making it a valuable multilingual comparable corpus. However, these aligned articles have various degrees of comparability. Some Wikipedia authors translate the article from one language to another, which produces parallel sentences in these article pairs. Other authors write the aligned articles by themselves, thus causing the article pairs to contain few or no parallel sentences but many parallel fragments. Moreover, even the translated article pairs may later diverge because of independent edits in either language, and both parallel sentences and fragments can exist in these article pairs. Figure 1 shows an example of Chinese–Japanese comparable texts describing a French city “Sète” from Wikipedia, in which both parallel sentences and fragments are contained. Because both parallel sentences and fragments are helpful for SMT, we believe that it is better to extract both of them instead of only focusing on one.

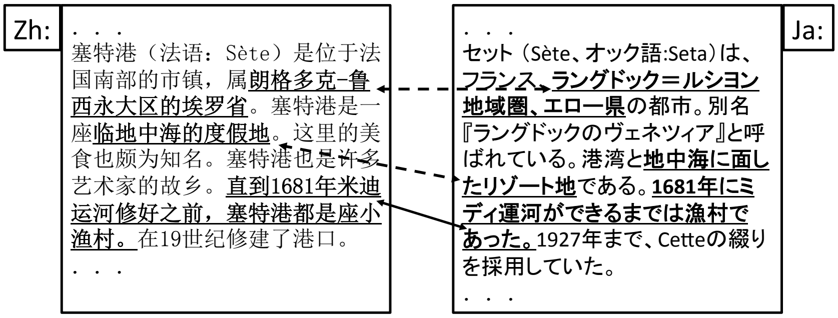
In this work, we exploit the Chinese–Japanese Wikipedia as a case study. We propose an integrated system to extract both parallel sentences and fragments from the Chinese–Japanese Wikipedia for SMT. A special characteristic of the Chinese–Japanese languages is that they share common Chinese characters[3] [Chu et al. 2013a], and we exploit them for both parallel sentence and fragment extraction. The integrated system consists of two major components:

— Parallel sentence extraction: This follows the method of our previous study [Chu et al. 2014], and is used to identify parallel sentences from comparable sentences. In our previous study, we only focused on extracting parallel sentences from Chinese–Japanese Wikipedia, while in this study, we further extract parallel fragments from comparable sentences. Our parallel sentence extraction method is inspired by

[1] Ştefănescu et al. [2012] conducted experiments on different levels of comparability, however they stayed on the study of parallel sentence extraction from these levels.

[2] Although [Munteanu and Marcu 2005; 2006; Gupta et al. 2013] were aware of this possibility, none of them provided an integrated framework that addresses both problems.

[3] Common Chinese characters can be seen as cognates (words or languages that have the same origin).



. . .
塞特港（法语：Sète）是位于法国南部的市镇，属朗格多克−鲁西永大区的埃罗省。塞特港是一座临地中海的度假地。这里的美食也颇为知名。塞特港也是许多艺术家的故乡。直到1681年米迪运河修好之前，塞特港都是座小渔村。在19世纪修建了港口。
. . .

. . .
セット（Sète、オック語:Seta）は、フランス、ラングドック＝ルシヨン地域圏、エロー県の都市。別名『ラングドックのヴェネツィア』と呼ばれている。港湾と地中海に面したリゾート地である。1681年にミディ運河ができるまでは漁村であった。1927年まで、Cetteの綴りを採用していた。
. . .

Fig. 1. Example of Chinese–Japanese comparable texts describing the French city “Sète” from Wikipedia (parallel sentences are linked with solid lines, and parallel fragments are linked with dashed lines).

[Munteanu and Marcu 2005], and mainly consists of a parallel sentence candidate filter and classifier for parallel sentence identification. We further developed it in the following two aspects in our previous study [Chu et al. 2014]:

  - Using common Chinese characters for the filter to solve the domain-dependent problem caused by the lack of an open domain dictionary.
  - Improving the classifier by introducing Chinese character features together with two other novel feature sets.

  The identification of parallel sentences from comparable sentences is based on the classification probability given by the classifier, and we empirically determine the classification probability thresholds for parallel and comparable sentences in our experiments.

- Parallel fragment extraction: This procedure follows that of our previous study [Chu et al. 2013b], and is used to extract parallel fragments from comparable sentences. In our previous study, we proposed an accurate parallel fragment extraction method. We located parallel fragment candidates using an alignment model, and used an accurate lexicon-based filter to identify the truly parallel ones. In this study, we further extend it by using common Chinese characters for the lexicon-based filter to improve its coverage. Our previous study only focused on extracting parallel fragments from a very-non-parallel scientific comparable corpus, while in this study we extract the parallel fragments from comparable sentences in the Chinese–Japanese Wikipedia.

Experimental results on the Chinese–Japanese Wikipedia show that both of our proposed parallel sentence and fragment extraction methods significantly outperform previous studies, and the integrated extraction of parallel sentences and fragments significantly improves SMT performance. Our system is language independent, except for the use of common Chinese characters, however, a similar idea can be applied to other language pairs that share cognates. Our system also can be applied to comparable corpora other than Wikipedia.

2. RELATED WORK

As no previous studies extract both parallel sentences and fragments from comparable corpora in an integrated framework, in this section we describe the related work of parallel sentence and fragment extraction separately.

2.1. Parallel Sentence Extraction

As parallel sentences tend to appear in similar article pairs, many studies first conduct article alignment from comparable corpora and then identify the parallel sentences from the aligned article pairs. Cross-lingual information retrieval technology is commonly used for article alignment [Utiyama and Isahara 2003; Fung and Cheung 2004; Munteanu and Marcu 2005; Gahbiche-Braham et al. 2011]. Large-scale article alignment from the Web also has been studied [Nie et al. 1999; Resnik and Smith 2003; Zhang et al. 2006; Fung et al. 2010; Uszkoreit et al. 2010]. This study extracts parallel sentences from Wikipedia. Wikipedia is a special type of comparable corpora because article alignment is established via interlanguage links. Approaches without article alignment have also been proposed [Tillmann 2009; Abdul-Rauf and Schwenk 2011; Ştefǎnescu et al. 2012; Ling et al. 2013]. These studies directly retrieve candidate sentence pairs and select the parallel sentences using various filtering methods.

Parallel sentence identification methods can be classified into two different approaches: classification [Munteanu and Marcu 2005; Tillmann 2009; Smith et al. 2010; Bharadwaj and Varma 2011; Ştefǎnescu et al. 2012] and translation similarity measures [Utiyama and Isahara 2003; Fung and Cheung 2004; Fung et al. 2010; Abdul-Rauf and Schwenk 2011]. Similar features such as word overlap and sentence length based features are used in both of these approaches. We believe that a machine learning approach can be more discriminative with respect to the features, thus we adopt a classification approach with novel features sets.

Most previous studies use supervised or semi-supervised methods that require external resources in addition to the comparable corpora. These studies differ in their use of a manually created seed dictionary [Utiyama and Isahara 2003; Fung and Cheung 2004; Adafre and de Rijke 2006; Lu et al. 2010], a seed parallel corpus [Zhao and Vogel 2002; Munteanu and Marcu 2005; Tillmann 2009; Smith et al. 2010; Gahbiche-Braham et al. 2011; Abdul-Rauf and Schwenk 2011; Ştefǎnescu et al. 2012; Ştefǎnescu and Ion 2013; Ling et al. 2013], or link structure and meta data in Wikipedia [Bharadwaj and Varma 2011]. This study uses a seed parallel corpus. An unsupervised method has also been proposed [Do et al. 2010], however their method suffers from high computational complexity.

Previous studies extract parallel sentences from various types of comparable corpora, such as bilingual news articles [Zhao and Vogel 2002; Utiyama and Isahara 2003; Munteanu and Marcu 2005; Tillmann 2009; Do et al. 2010; Gahbiche-Braham et al. 2011; Abdul-Rauf and Schwenk 2011], patent data [Utiyama and Isahara 2007; Lu et al. 2010], social media [Ling et al. 2013], and the Web [Nie et al. 1999; Resnik and Smith 2003; Zhang et al. 2006; Ishisaka et al. 2009; Jiang et al. 2009; Fung et al. 2010; Hong et al. 2010]. However, few studies have been conducted to extract parallel sentences from Wikipedia [Adafre and de Rijke 2006; Smith et al. 2010; Bharadwaj and Varma 2011; Ştefǎnescu and Ion 2013]. Previous studies are interested in language pairs between English and other languages such as German or Spanish. We focus on Chinese–Japanese, where parallel corpora are very scarce.

2.2. Parallel Fragment Extraction

[Munteanu and Marcu 2006] was the first attempt to extract parallel fragments from comparable sentences. They extracted sub-sentential parallel fragments using a Log-Likelihood-Ratio (LLR) lexicon estimated on a seed parallel corpus and a smoothing filter. They showed the effectiveness of fragment extraction for SMT. Their method has a drawback in that they do not locate the source and target fragments simultaneously, which cannot guarantee that the extracted fragments are translations of each other. We solve this problem by using an alignment model to locate the source and target fragments simultaneously.

Quirk et al. [2007] introduced two generative alignment models to extract parallel fragments from comparable sentences. However, the extracted fragments slightly decrease SMT performance when they are appended to in-domain training data. We believe that this is because the comparable sentences are quite noisy, and hence the

Table I. Examples of common Chinese characters

Meaning	world	freeze	two
TC	世 (U+4E16)	凍 (U+51CD)	兩 (U+5169)
SC	世 (U+4E16)	冻(U+51BB)	两(U+4E24)
kanji	世 (U+4E16)	凍 (U+51CD)	両 (U+4E21)

Note: TC denotes Traditional Chinese and SC denotes Simplified Chinese.

alignment models cannot accurately extract parallel fragments. To solve this problem, we only use alignment models for parallel fragment candidate detection, and use an accurate lexicon-based filter to guarantee the accuracy of the extracted parallel fragments.

In addition to the above studies, there are some other efforts. Hewavitharana and Vogel [2011] proposed a method that calculates both the inside and outside probabilities for fragments in a comparable sentence pair, and show that the context of the sentence helps fragment extraction. Riesa and Marcu [2012] used a syntax-based alignment model to extract parallel fragments from noisy parallel data. Gupta et al. [2013] translated a source fragment with an existing SMT system, and identified the target fragment by calculating the similarity between the translated source and target fragments. Fu et al. [2013] proposed a method that is based on hierarchical phrase-based force decoding. Afli et al. [2013] attempted to extract parallel fragments from multimodal comparable corpora. Supervised methods have also been proposed for parallel fragment extraction [Aker et al. 2012]. Zhang and Zong [2013] went a step further in that they not only extracted parallel fragments, but also estimated translation probabilities for the extracted fragments to construct a translation model. Our study differs from these in that it focuses on the task of accurately extracting parallel fragments and the best approach for achieving it.

3. COMMON CHINESE CHARACTERS

In contrast to some other language pairs, Chinese and Japanese share Chinese characters. In Chinese, the Chinese characters are called hanzi, while in Japanese they are called kanji. Hanzi can be divided into two groups, Simplified Chinese (used in mainland China and Singapore) and Traditional Chinese (used in Taiwan, Hong Kong, and Macau). The number of strokes needed to write characters has been largely reduced in Simplified Chinese, and the shapes may be different from those in Traditional Chinese. Because kanji characters originated from ancient China, many common Chinese characters exist between hanzi and kanji. We previously created a Chinese character mapping table between Traditional Chinese, Simplified Chinese, and Japanese [Chu et al. 2013a].[4] Table I gives some examples of common Chinese characters from that mapping table along with their Unicode.

Because Chinese characters contain significant semantic information and common Chinese characters share the same meaning, they can be valuable linguistic clues for many Chinese–Japanese natural language processing tasks. Many studies have exploited common Chinese characters. Tan et al. [1995] used the occurrence of identical common Chinese characters in Chinese and Japanese (e.g., "world" in Table I) in automatic sentence alignment task for document-level aligned text. Goh et al. [2005] detected common Chinese characters where kanji are identical to Traditional Chinese, but different from Simplified Chinese (e.g., "freeze" in Table I). Using a Chinese encoding converter[5] that can convert Traditional Chinese into Simplified Chinese, they

[4] http://nlp.ist.i.kyoto-u.ac.jp/member/chu/pubdb/LREC2012/kanji_mapping_table.txt
[5] http://www.mandarintools.com/zhcode.html

built a Japanese–Simplified Chinese dictionary, partly using the direct conversion of Japanese into Chinese for Japanese kanji words. We previously made use of the Unihan database[6] to detect common Chinese characters that are visual variants of each other (e.g., "two" in Table I) and show the effectiveness of common Chinese characters in Chinese–Japanese phrase alignment and Chinese word segmentation optimization for Chinese–Japanese SMT [Chu et al. 2013a].

We previously investigated the coverage of common Chinese characters on a scientific paper abstract parallel corpus, and showed that over 45% of Chinese hanzi and 75% of Japanese kanji are common Chinese characters [Chu et al. 2013a]. This phenomenon also happens in the case of parallel fragments. Therefore, common Chinese characters can be powerful linguistic clues to identify both parallel sentences and parallel fragments. In this study, we exploit common Chinese characters in both parallel sentence and fragment extraction.

4. PARALLEL SENTENCE AND FRAGMENT EXTRACTION SYSTEM

This study extracts parallel sentences and fragments from the Chinese–Japanese Wikipedia. The overview of our parallel sentence and fragment extraction system is presented in Figure 2. We first align articles on the same topic in the Chinese and Japanese Wikipedia via the interlanguage links ((1) in Figure 2). Next, we generate all possible sentence pairs using the Cartesian product from the aligned articles and discard the pairs that do not pass a filter that reduces the candidate pairs by keeping more reliable sentences ((2) in Figure 2).[7] Next, we use a classifier trained on a small number of parallel sentences from a seed parallel corpus to classify the parallel sentence candidates into parallel and comparable sentences based on the classification probability[8] given by the classifier ((3) in Figure 2). As the noise in comparable sentences will decrease the SMT performance, we further apply parallel fragment extraction. We use two steps to accurately extract parallel fragments. We first detect parallel fragment candidates using alignment models ((4) in Figure 2). We then filter the candidates using probabilistic translation lexicons to produce accurate results ((5) in Figure 2).

Steps (2), (3), (4), and (5) in Figure 2 form the four main components of our system, further described in Sections 4.1, 4.2, 4.3, and 4.4 in detail.

4.1. Parallel Sentence Candidate Filtering

A parallel sentence candidate filter is necessary because it can remove most of the noise introduced by the simple Cartesian product sentence generator and reduce the computational cost of parallel sentence and fragment identification. Previous studies use a filter with sentence length ratio and dictionary-based word overlap conditions [Munteanu and Marcu 2005]. Although the sentence length ratio condition is domain independent, the word overlap condition is not.[9] Wikipedia is an open domain database, thus using a domain dependent condition for filtering may decrease the performance of our system. In the scenario where an open domain dictionary is unavail-

[6] http://unicode.org/charts/unihan.html

[7] In Wikipedia, because article alignment has been established, the Cartesian product with a filter works just well. However, for comparable corpora where article alignment is not available, it is necessary to use cross-lingual information retrieval to retrieve candidate sentence pairs [Tillmann 2009; Abdul-Rauf and Schwenk 2011; Ştefănescu et al. 2012; Ling et al. 2013] or perform article alignment beforehand [Utiyama and Isahara 2003; Fung and Cheung 2004; Munteanu and Marcu 2005].

[8] The classification probability thresholds for parallel and comparable sentences were empirically determined in our experiments.

[9] The dictionary is automatically generated using a word alignment tool from a seed parallel corpus, which is domain specific.

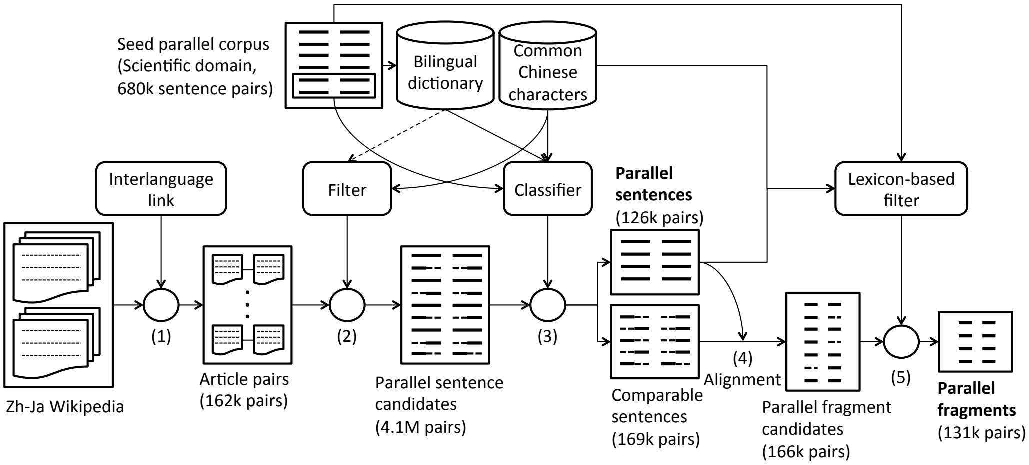


Fig. 2. Parallel sentence and fragment extraction system (the amounts of data for each step in our experiments are indicated).

able, we must search for alternatives that are robust against domain diversity and can effectively filter noise.

Because common Chinese characters are domain independent and an effective way to filter the noise introduced by the simple Cartesian product sentence generator, here we propose using them for the filter. We compared four different filtering strategies: dictionary-based word overlap (Word), common Chinese character overlap (CCO), and their logical combinations. We define them as follows:

— Word filter: uses a dictionary-based word overlap.
— CCO filter: uses a common Chinese character overlap.
— Word and CCO filter: uses the logical conjunction of the word and common Chinese character overlaps.
— Word or CCO filter: uses the logical disjunction of the word and common Chinese character overlaps.

The common Chinese character overlap is calculated based on the Chinese character mapping table in [Chu et al. 2013a]. In our experiments, we used a 1-gram common Chinese character overlap with a threshold of 0.1 for Chinese and 0.3 for Japanese. Note that a same sentence length ratio threshold is used as an additional filtering condition for all four filters. In our experiments, we set the sentence length ratio threshold to two. We compare the performance of the different filtering strategies in Section 5.2.2.

4.2. Parallel Sentence Identification by Classification

Because the parallel and comparable sentences are determined by the classifier, it is the core component of the extraction system. In this section, we first describe the training and testing process, and then introduce the features we use for the classifier.

4.2.1. Training and Testing. We use a support vector machine classifier [Chang and Lin 2011]. Training and testing instances for the classifier are created following the method of [Munteanu and Marcu 2005]. We use a small number of parallel sentences from a seed parallel corpus as positive instances. Negative instances are generated by the Cartesian product of the positive instances excluding the original positive instances, and they are filtered by the same filtering method used in Section 4.1. More-

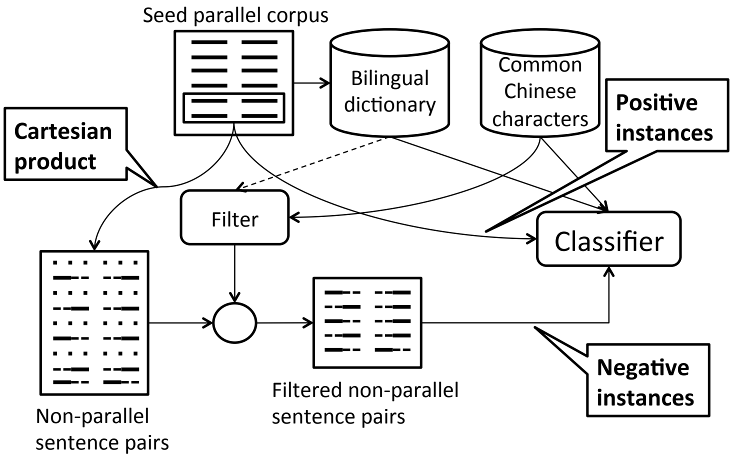


Fig. 3. Parallel sentence classifier.

over, we randomly discard some negative instances for training when necessary[10] to guarantee that the ratio of negative to positive instances is less than five for the performance of the classifier. Figure 3 illustrates this process.

4.2.2. Features. In this study, we reuse the features proposed in previous studies (we call these the basic features), and propose three novel feature sets, namely Chinese character (CC) features, Non-CC word features, and content word features.
Basic Features. The basic features were proposed in [Munteanu and Marcu 2005]:

- Sentence length, length difference, and length ratio.
- Word overlap: the percentage of words on each side that have a translation on the other side (according to the dictionary).
- Alignment features:
  - Percentage and number of words that have no connection on each side.
  - Top three largest fertilities.
  - Length of the longest contiguous connected span.
  - Length of the longest unconnected substring.

The alignment features[11] are extracted from the alignment results of the parallel and non-parallel sentences used as instances for the classifier. Note that alignment features may be unreliable when the quantity of non-parallel sentences is significantly larger than the parallel sentences.
CC Features. We use the example of a Chinese–Japanese parallel sentence presented in Figure 4 to explain the CC features in detail using the following features:

- Number of Chinese characters on each side (Zh: 18, Ja: 14).
- Percentage of characters that are Chinese characters on each side (Zh: $18/20 = 90\%$, Ja: $14/32 = 43\%$).
- Ratio of Chinese characters on both sides ($18/14 = 128\%$).

[10] Note that we keep all negative instances for testing.
[11] We do not give the detailed information of the alignment features such as the definitions of fertility, connected span and unconnected substring etc. in this article, as they are proposed in [Munteanu and Marcu 2005], we recommend the interested readers to refer to the original paper.

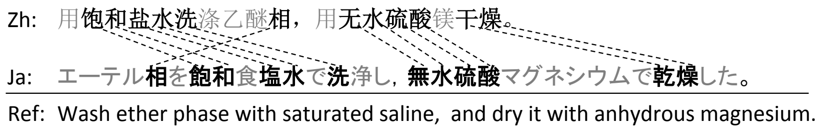

Zh: 用饱和盐水洗涤乙醚相，用无水硫酸镁干燥。
Ja: エーテル相を飽和食塩水で洗浄し，無水硫酸マグネシウムで乾燥した。
Ref: Wash ether phase with saturated saline, and dry it with anhydrous magnesium.

Fig. 4. Example of common Chinese characters (in bold and linked with dotted lines) in a Chinese–Japanese parallel sentence pair.

- Number of n-gram common Chinese characters (1-gram: 12, 2-gram: 6, 3-gram: 2, 4-gram: 1).
- Percentage of n-gram Chinese characters that are n-gram common Chinese characters on each side (Zh: 1-gram: $12/18 = 66\%$, 2-gram: $6/16 = 37\%$, 3-gram: $2/14 = 14\%$, 4-gram: $1/12 = 8\%$; Ja: 1-gram: $12/14 = 85\%$, 2-gram: $6/9 = 66\%$, 3-gram=: $2/5 = 40\%$, 4-gram: $1/3 = 33\%$).

The n-gram common Chinese characters are detected using the Chinese character mapping table in [Chu et al. 2013a].

Non-CC Word Features. Chinese–Japanese parallel sentences often contain alignable words that do not consist of Chinese characters, such as foreign words and numbers, which we call Non-Chinese character (Non-CC) words. Note that we do not count Japanese kana as Non-CC words. Non-CC words can be helpful clues to identify parallel sentences. We use the following features:

- Number of Non-CC words on each side.
- Percentage of words that are Non-CC words on each side.
- Ratio of Non-CC words on both sides.
- Number of the same Non-CC words.
- Percentage of the Non-CC words that are the same on each side.

Content Word Features. The word overlap feature proposed in [Munteanu and Marcu 2005] has the problem that function words and content words are handled in the same way. Function words often have a translation on the other side, thus erroneous parallel sentence pairs with a few content word translations are often produced by the classifier. Therefore, we add the following content word features:

- Percentage of words that are content words on each side.
- Percentage of content words on each side that have a translation on the other side (according to the dictionary).

We determine a word as a content or function word using predefined part-of-speech (POS) tag sets of function words for Chinese and Japanese accordingly.[12]

4.3. Parallel Fragment Candidate Detection

Figure 5 shows an example of comparable sentences extracted by our system from Chinese–Japanese comparable corpora. The alignment results are computed by IBM models [Brown et al. 1993] with symmetrization heuristics [Koehn et al. 2007]. We notice that the truly parallel fragments "Princeton advanced research institute" and

[12] For Chinese, they are AS, BA, CC, CS, DEC, DEG, DER, DEV, DT, IJ, LB, LC, MSP, P, PN, PU, SB, SP, VC and VE in Penn Chinese Treebank (CTB) standard [Xia et al. 2000]. For Japanese, they are 接頭辞 (conjunction), 接尾辞 (suffix), 助詞 (particle), 助動詞 (auxiliary verb), 判定詞 (copula), 指示詞 (demonstrative), 特殊:句点 (special:period), 特殊:読点 (special:comma), 特殊:空白 (special:blank), 名詞:形式名詞 (noun:formal noun) and 名詞:副詞的名詞 (noun:adverbial noun) in JUMAN [Kurohashi et al. 1994].

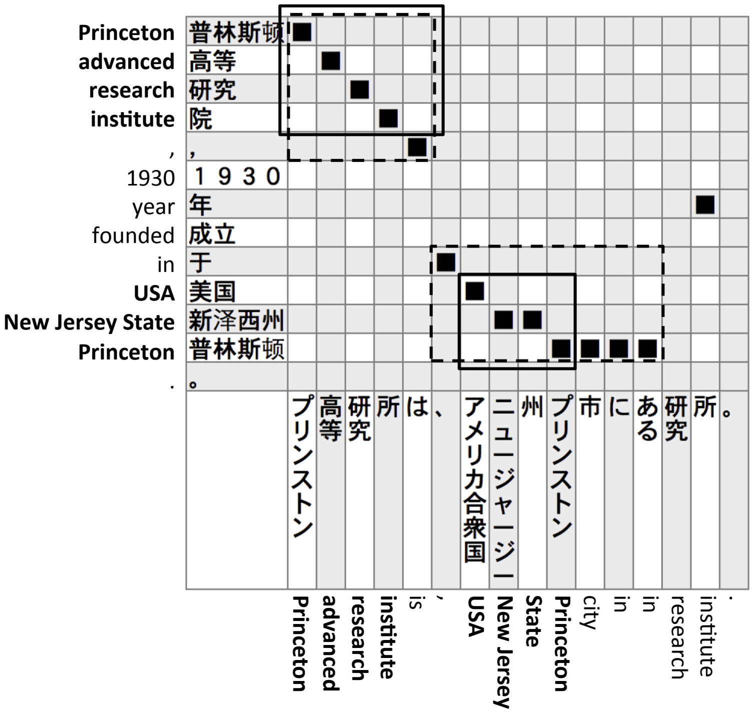


Fig. 5. Example of comparable sentences with alignment results computed by IBM models (parallel fragment candidates are in dashed rectangles, parallel fragments are in solid-border rectangles).

"USA New Jersey State Princeton" are aligned, although there are some incorrectly aligned word pairs. We believe that this kind of alignment information can be helpful for fragment extraction. However, we need to develop a method to separate the truly parallel fragments from the aligned fragments. Therefore, we propose a two-step parallel fragment extraction method: In the first step, we detect parallel fragment candidates using alignment models; in the second step, we apply a lexicon-based filter to produce accurate fragments. In this section, we describe the parallel fragment candidate detection method.

For alignment, we use the parallel sentences together with the comparable sentences, which can help improve the alignment accuracy for the comparable sentences. We treat the longest spans that have monotonic and non-null alignment as parallel fragment candidates. The reason we only consider monotonic ones is that, based on our observation, the ordering of alignment models on comparable sentences is unreliable. Quirk et al. [2007] also produced monotonic alignments in their generative model. Monotonic alignments are not sufficient for many language pairs. In the future, we plan to develop a method to deal with this problem. The non-null constraint can limit us from extracting incorrect fragments. Similar to previous studies, we are interested in fragment pairs with size ≥ 3. Taking the comparable sentences in Figure 5 as an example, we extract the fragments in dashed rectangles as parallel fragment candidates.

4.4. Lexicon-based Filter

The parallel fragment candidates cannot be used directly because many of them are still noisy, as shown in Figure 5. To produce accurate results, we use a lexicon-based filter. We filter a candidate parallel fragment pair with probabilistic translation lexicons. The lexicon pair can be extracted from a seed parallel corpus. However, as described in Section 4.1, the lexicons extracted from a domain specific seed parallel corpus are domain dependent. Fortunately, we already have the parallel sentences extracted by our system from Wikipedia that are domain independent. Therefore, we append the extracted parallel sentences to a seed parallel corpus to generate the lexicons (called hereafter the combined parallel corpus).[13] Different lexicons may have different filtering effects. Here, we compare the following three types of lexicon:

— IBM Model 1: The first lexicon we use is the IBM Model 1 lexicon, obtained by running GIZA++[14] that implements the sequential word-based statistical alignment model of the IBM models on the combined parallel corpus.
— LLR: The second lexicon we use is the LLR lexicon. Munteanu and Marcu [2006] showed that the LLR lexicon performs better than the IBM Model 1 lexicon for parallel fragment extraction. One advantage of the LLR lexicon is that it can produce both positive and negative associations. Munteanu and Marcu [2006] developed a smoothing filter that applies this advantage. We extracted the LLR lexicon from the automatically word-aligned combined parallel corpus using the same method as [Munteanu and Marcu 2006].
— SampLEX: The last lexicon we use is the SampLEX lexicon. Vulić and Moens [2012] proposed an associative approach for lexicon extraction from parallel corpora that relies on the paradigm of data reduction. They extract translation pairs from many smaller sub-corpora that are randomly sampled from the original corpus, based on some frequency-based criteria of similarity. They showed that their method outperforms IBM Model 1 and other associative methods such as LLR in terms of precision. We extracted the SampLEX lexicon from the combined parallel corpus using the same method as [Vulić and Moens 2012].

To gain new knowledge that does not exist in the lexicon, we apply a smoothing filter similar to [Munteanu and Marcu 2006]. For each aligned word pair in the fragment candidates, we score the words in both directions according to the extracted lexicon. If the aligned word pair exists in the lexicon, we use the corresponding translation probabilities as the scores. For the LLR lexicon, we use both positive and negative association values. If the aligned word pair does not exist in the lexicon, we set the scores in both directions to -1. There is the one exception when the aligned words are the same, which can happen for numbers, punctuation, abbreviations, etc. In this case, we set the scores to 1 without considering the existence of the word pair in the lexicon. Note that in Chinese–Japanese, aligned words can consist of the same common Chinese characters. We make use of our Chinese character mapping table [Chu et al. 2013a] to detect these word pairs. For these word pairs, we also set the scores to 1, and we discuss the effect of this in Section 5.3. After this process, we obtain *initial scores* for the words in the fragment candidates in both directions.

We then apply an averaging filter to the *initial scores* to obtain *filtered scores* in both directions. The averaging filter sets the score of one word to the average score of several words around it. We believe that the words with initial positive scores are reliable because they satisfy two strong constraints, namely their alignment according to the alignment models and existence in the lexicon. Therefore, unlike [Munteanu and Marcu 2006], we only apply the averaging filter to the words with negative scores. Moreover, we add the constraint that we only filter a word when both its immediately

[13] The extracted parallel sentences also can be used for bootstrapping following [Masuichi et al. 2000; Fung and Cheung 2004; Munteanu and Marcu 2005], however this is not the focus of this study.
[14] http://code.google.com/p/giza-pp

preceding and following words have positive scores, which further guarantees accuracy. For the number of words used for averaging, we used five (two preceding words and two following words). The heuristics presented here produced good results on a development set.

Finally, we extract parallel fragments according to the *filtered scores*. We extract word aligned fragment pairs with continuous positive scores in both directions. Fragments with less than three words may be produced in this process and we discard them, as done in previous studies.

5. EXPERIMENTS

Parallel sentence and fragment extraction and translation experiments were conducted on Chinese–Japanese data. In all our experiments, we preprocessed the data by segmenting and POS tagging Chinese and Japanese sentences using a tool proposed in our previous study [Chu et al. 2013a] and JUMAN [Kurohashi et al. 1994], respectively.

In this section, we first describe the data used in our experiments. Next, we conduct parallel sentence extraction and translation experiments, which is treated as the baseline in our experiments. We then perform parallel fragment extraction experiments. Finally, we conduct translation experiments using both the parallel sentences and fragments to show the effectiveness of our proposed integrated system.

5.1. Data

The seed parallel corpus we used is the Chinese–Japanese section of the Asian Scientific Paper Excerpt Corpus (ASPEC).[15] This corpus is a scientific domain corpus provided by the Japan Science and Technology Agency (JST)[16] and the National Institute of Information and Communications Technology (NICT).[17] It was created by the Japanese project “Development and Research of Chinese–Japanese Natural Language Processing Technology,” and contains $680k$ sentences ($18.2M$ Chinese and $21.8M$ Japanese tokens, respectively).

In addition, we downloaded the Chinese[18] (2012/09/21) and Japanese[19] (2012/09/16) Wikipedia database dumps. We used an open-source Python script[20] to extract and clean the text from the dumps. Because the Chinese dump is a mixture of Traditional and Simplified Chinese, we converted all Traditional Chinese to Simplified Chinese using a conversion table published by Wikipedia.[21] We aligned the articles on the same topics in Chinese and Japanese via the interlanguage links, obtaining $162k$ article pairs ($2.1M$ Chinese and $3.5M$ Japanese sentences, respectively).

5.2. Parallel Sentence Extraction and Translation Experiments

We evaluated the classification accuracy and conducted extraction and translation experiments to verify the effectiveness of our proposed parallel sentence extraction method. We also investigated the effect on different classification probability thresholds for parallel sentence identification.

5.2.1. Classification Accuracy Evaluation. We evaluated classification accuracy using two distinct sets of 5k parallel sentences from the seed parallel corpus for training and

[15] http://lotus.kuee.kyoto-u.ac.jp/ASPEC
[16] http://www.jst.go.jp
[17] http://www.nict.go.jp
[18] http://dumps.wikimedia.org/zhwiki
[19] http://dumps.wikimedia.org/jawiki
[20] http://code.google.com/p/recommend-2011/source/browse/Ass4/WikiExtractor.py
[21] http://svn.wikimedia.org/svnroot/mediawiki/branches/REL1_12/phase3/includes/ZhConversion.php

Table II. Classification results

Features	Precision	Recall	F-measure
Munteanu+, 2005	96.65	83.56	89.63
+CC	97.05	93.52	95.25
+Non-CC	97.38	93.64	95.47
+Content	**98.34**	**95.94**	**97.12**

testing, respectively. For the support vector machine classifier, we used the LIBSVM toolkit [Chang and Lin 2011][22] with 5-fold cross-validation and a radial basis function kernel. In this section and Section 5.2.2, we report the results for a classification probability threshold of 0.9, namely, we treat the sentence pairs with classification probability ≥ 0.9 as parallel sentences. We address the effect of different thresholds in Section 5.2.3. We used the word alignment tool GIZA++ to generate a dictionary from the seed parallel corpus and calculate the alignment features. For the dictionary, we kept the top five translations with translation probabilities higher than 0.1 for each source word.[23] Word overlap was calculated based on that dictionary. We report the results using word overlap filtering for easier comparison to previous studies. The word overlap threshold was set to 0.25. We compared the following feature settings:

- Munteanu+, 2005: the basic features proposed in [Munteanu and Marcu 2005] only
- +CC: adding the CC features
- +Non-CC: adding the Non-CC word features
- +Content: adding the content word features

We evaluated the performance of classification by computing the precision, recall, and F-measure, defined as:

$$precision = 100 \times \frac{classified\_well}{classified\_parallel}, \tag{1}$$

$$recall = 100 \times \frac{classified\_well}{true\_parallel}, \tag{2}$$

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{3}$$

where $classified\_well$ is the number of pairs that the classifier correctly identified as parallel, $classified\_parallel$ is the number of pairs that the classifier identified as parallel, and $true\_parallel$ is the number of actual parallel pairs in the test set. Note that we only used the top result identified as parallel by the classifier for evaluation.

Classification results are shown in Table II. We can see that the Chinese character features can significantly improve the accuracy compared to "Munteanu+, 2005." Our proposed Non-CC word and content word overlap features further improve the accuracy.

5.2.2. Extraction and Translation Experiments. We extracted parallel sentences from Wikipedia and evaluated the Chinese-to-Japanese SMT performance using the extracted sentences as training data. For decoding, we used the state-of-the-art phrase-based SMT toolkit Moses [Koehn et al. 2007] with the default options, except for the distortion limit ($6 \rightarrow 20$). We trained a 5-gram language model on the Japanese Wikipedia ($10.7M$ sentences) using the SRILM toolkit[24] with interpolated Kneser-Ney

[22] http://www.csie.ntu.edu.tw/˜cjlin/libsvm
[23] Note that the dictionary might contain noisy translation pairs and further cleaning them might be helpful for our task [Aker et al. 2014], however, we leave it as future work.
[24] http://www.speech.sri.com/projects/srilm

Table III. Parallel sentence extraction and translation results

Features	Filter	# Sentences	BLEU-4	OOV
Seed			25.42	9.11%
Munteanu+, 2005	Word	122,569	35.18	4.56%
+CC	Word	146,797	36.27†	3.82%
+Non-CC	Word	161,046	36.79†	3.68%
+Content	Word	164,993	37.39†‡	3.80%
+Content	CCO	126,811	**37.82**†‡	3.71%
+Content	Word and CCO	80,598	36.14	4.72%
+Content	Word or CCO	184,103	36.41†	3.56%

Note: "†" and "‡" denote that the result is significantly better than "Munteanu+, 2005" and "+CC" respectively at $p < 0.05$.

discounting. For tuning and testing, we used two distinct sets of 198 parallel sentences. These sentences were randomly selected from the sentence pairs extracted from Wikipedia by our system with different methods, and the erroneous parallel sentences were manually discarded[25] because the tuning and testing sets for SMT require truly parallel sentences. Note that for training, we kept all the sentences extracted by different methods except for the sentences duplicated in the tuning and testing sets. Tuning was performed by minimum error rate training [Och 2003], and it was re-run for every experiment. The other settings were the same as the ones used in the classification experiments described in Section 5.2.1.

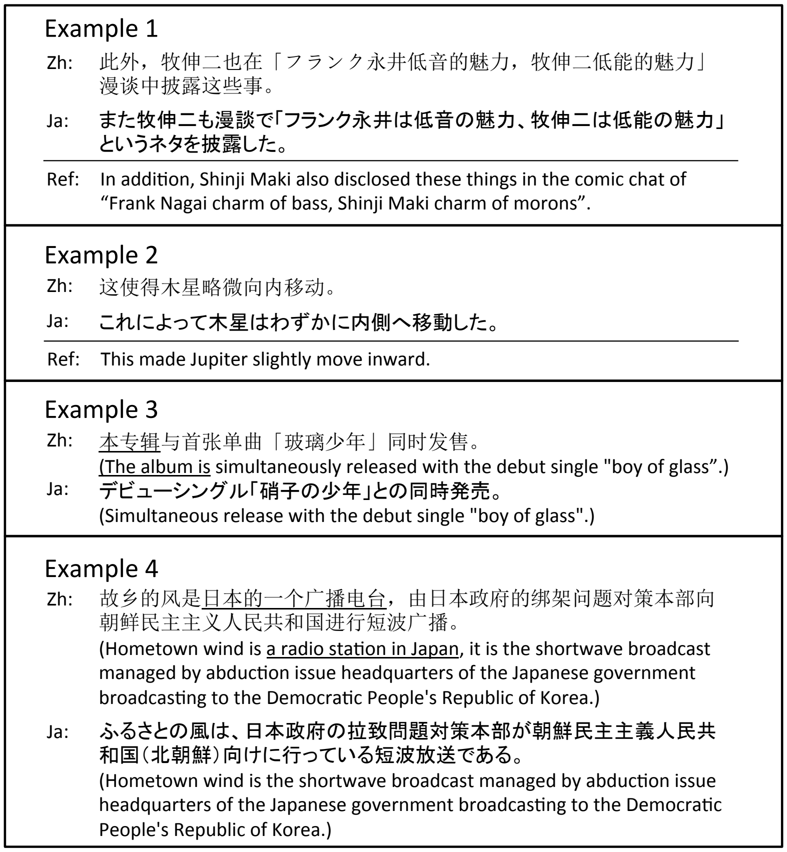
Parallel sentence extraction and translation results using different methods are shown in Table III. We report the Chinese-to-Japanese translation results on the test set using the BLEU-4 score [Papineni et al. 2002]. "Munteanu+, 2005," "+CC," "+Non-CC," and "+Content" denote the different features described in Section 5.2.1. "Word," "CCO," "Word and CCO," and "Word or CCO" denote the four different filtering strategies described in Section 4.1. "# Sentences" denotes hereafter the number of sentences extracted by different methods after discarding the sentences duplicated in the tuning and testing sets, which were used as training data for SMT. For comparison, we also conducted translation experiments using the seed parallel corpus as training data, denoted as "Seed." The significance test was performed using the bootstrap resampling method proposed by Koehn [2004].

We can see that the Seed system does not perform well. The reason for this is that the Seed system is trained on a seed parallel corpus that is a scientific domain corpus. This differs from the tuning and testing sets that are open domain data extracted from Wikipedia, leading to a high out of vocabulary (OOV) word rate. The systems trained on the parallel sentences extracted from Wikipedia perform better than Seed. This is because they consist of the same domain data as the tuning and testing sets, and the OOV word rate is significantly lower than Seed.

Compared to Munteanu+, 2005, our proposed CC, Non-CC word, and content word features improve SMT performance significantly. One reason for this is that our proposed features can improve the recall of the classifier, which extracts more parallel sentences and hence causes the OOV word rate to be lower than Munteanu+, 2005. The other reason is that our proposed features improve the quality of the extracted sentences.

The CCO filter shows better performance than the Word filter, indicating that for open domain data such as Wikipedia, using common Chinese characters for filtering is more effective than a domain specific dictionary. The Word and CCO filter decreases the performance because the number of extracted sentences decreases significantly,

[25] To get the 396 sentences for tuning and testing, 404 sentences were manually discarded.



Example 1

Zh: 此外，牧伸二也在「フランク永井低音的魅力，牧伸二低能的魅力」漫谈中披露这些事。

Ja: また牧伸二も漫談で「フランク永井は低音の魅力、牧伸二は低能の魅力」というネタを披露した。

Ref: In addition, Shinji Maki also disclosed these things in the comic chat of “Frank Nagai charm of bass, Shinji Maki charm of morons”.

Example 2

Zh: 这使得木星略微向内移动。

Ja: これによって木星はわずかに内側へ移動した。

Ref: This made Jupiter slightly move inward.

Example 3

Zh: 本专辑与首张单曲「玻璃少年」同时发售。
(The album is simultaneously released with the debut single "boy of glass”.)

Ja: デビューシングル「硝子の少年」との同時発売。
(Simultaneous release with the debut single "boy of glass".)

Example 4

Zh: 故乡的风是日本的一个广播电台，由日本政府的绑架问题对策本部向朝鲜民主主义人民共和国进行短波广播。
(Hometown wind is a radio station in Japan, it is the shortwave broadcast managed by abduction issue headquarters of the Japanese government broadcasting to the Democratic People's Republic of Korea.)

Ja: ふるさとの風は、日本政府の拉致問題対策本部が朝鮮民主主義人民共和国（北朝鮮）向けに行っている短波放送である。
(Hometown wind is the shortwave broadcast managed by abduction issue headquarters of the Japanese government broadcasting to the Democratic People's Republic of Korea.)

Fig. 6. Examples of some extracted parallel sentences (noisy parts are underlined).

leading to a higher OOV word rate. The Word or CCO filter also shows poor performance, and we suspect the reason is the increase of erroneous parallel sentence pairs.

For the best performing method, +Content with CCO filter, we manually estimated 100 sentence pairs that were randomly selected from the extracted sentences. We found that 64% of them are actual translation equivalents, while the other erroneous parallel sentences only contain a small amount of noise. Based on our analysis, the majority of errors occur when one sentence in a sentence pair contains a small amount of extra information that does not exist in the other sentence. These sentence pairs are extracted because most parts are parallel and the classifier gives them relatively high scores. Figure 6 shows some examples of the extracted parallel sentences including some noisy sentence pairs. Because SMT models are robust to this kind of noise, the noisy sentence pairs can also be used to improve SMT performance. For these sentence pairs, it is not necessary to further apply parallel fragment extraction.

5.2.3. Effect on Classification Probability Threshold. The classifier is used to identify the parallel sentences from comparable sentences in our system, and the classification probability threshold is the criterion. In this section, we investigate the effect of using different thresholds for parallel sentence identification.

In our experiments, we compared the effects of different thresholds from 0.1 to 0.9 in intervals of 0.1, and treated the sentences pairs with classification probability greater

Table IV. Translation results for different thresholds

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
# Sentences	296,204	247,469	220,712	20,2038	187,957	173,827	160,980	146,562	126,811
BLEU-4	36.92	36.42	36.98	37.19	37.29	37.27	37.15	37.27	**37.82**

Table V. Parallel fragment extraction results

Method	# fragments	# fragments w/o CCC	Avg size (zh/ja)	Accuracy
Munteanu+, 2006	153,919		16.76/17.70	(6%)
IBM model 1	140,077	137,053	4.20/4.66	72%
LLR	131,509	129,477	4.18/4.63	82%
SampLEX	100,727	95,537	3.85/4.12	82%

Note: The accuracy was manually evaluated on 100 fragments randomly selected from the fragments extracted using different lexicons based on the number of exact matches. Furthermore, "w/o CCC" denotes the results that did not use common Chinese characters for the lexicon-based filter described in Section 4.4.

than or equal to the threshold as parallel sentences. Sentence extraction was performed using the best performing method +Content with CCO filter, described in Section 5.2.2. We conducted Chinese-to-Japanese translation experiments using the parallel sentences extracted using different thresholds as training data. The other settings were the same as the ones used in the translation experiments described in Section 5.2.2.

Table IV shows the translation results for different thresholds. We can see that threshold 0.9 shows the best performance. When the threshold is lowered, although more sentence pairs are extracted, the SMT performance decreases. The reason for this is that the additional sentences extracted by lowering the threshold are comparable sentences that contain noise, negatively affecting the SMT. We aim to extract the parallel fragments from these comparable sentences to further improve SMT.

In the following section, we treated the sentence pairs with "$0.1 \leq$ classification probability < 0.9" as comparable sentences[26], obtaining $169k$ sentences. We performed parallel fragment extraction from these comparable sentences. We also used the parallel sentences that were extracted with threshold 0.9 to assist the parallel fragment extraction, obtaining $126k$ sentences.[27] The SMT system trained on these parallel sentences is treated as the baseline system in Section 5.4.

5.3. Parallel Fragment Extraction Experiments

In our experiments, we compared our proposed fragment extraction method with [Munteanu and Marcu 2006]. For our proposed method, we applied word alignment using GIZA++ on the comparable sentences together with the parallel sentences described in Section 5.2.3 for parallel fragment candidate detection. For the lexicon-based filter, different lexicons may have different effects. Therefore, we compared the IBM Model 1, LLR, and SampLEX lexicons, which were all generated from the combined parallel corpus that appends the parallel sentences described in Section 5.2.3 to the seed parallel corpus described in Section 5.1.

The fragment extraction results are shown in Table V. We can see that the average size of the fragments (i.e., the number of words in the fragments) extracted by [Munteanu and Marcu 2006] is unusually long, which is also reported in [Quirk et al. 2007]. Our proposed method extracts shorter fragments. The IBM model 1 and LLR

[26] We did not extract parallel fragments from the sentences pairs with a classification probability of less than 0.1, because these sentences pairs are too noisy and rarely contain parallel fragments.

[27] Note that the sentences duplicated in the tuning and testing sets have been discarded.

Table VI. Examples of some fragment pairs extracted by our proposed method using LLR lexicon for the lexicon-based filter

ID	Zh fragment	Ja fragment
1	第 73 装甲掷弹兵团 (73rd Armored Grenadier Regiment)	第７３装甲擲弾兵連隊 (73rd Armored Grenadier Regiment)
2	银幕投影系统 (screen projection system)	スクリーン投影システム (screen projection system)
3	为成人 杂志 (are adult magazines)	は成人向け雑誌 (are adult magazines)
4	１９９７年世界女子手球锦标赛为 (Women's World Handball Championship 1997 is)	１９９７年世界女子ハンドボール選手権は (Women's World Handball Championship 1997 is)
5	氦开始聚变 (Helium begins fusion)	ヘリウム が核 (Helium is Nucleus)
6	日本 福岛县岩濑 (Japan Fukushima Prefecture Iwase)	、福島県岩瀬 (, Fukushima Prefecture Iwase)
7	和 学术 参考 书 (and academic reference books)	や参考書 (and reference books)
8	上将 军衔。 (general rank .)	上将 に就任。 (general inauguration .)

Note: Noisy parts are underlined.

lexicons extract more fragments than SampLEX, and the average size is slightly larger. The reason for this is that SampLEX generates a smaller lexicon compared to IBM model 1 and LLR. Common Chinese characters help to extract more fragments, especially when we use a smaller lexicon (i.e., SampLEX).

To evaluate accuracy, we randomly selected 100 fragments extracted using different lexicons. A more reliable way to evaluate the accuracy is creating a much larger test set that contains a representative sample of data points (i.e. fragments) under scrutiny, and evaluating the precision, recall and F-measure like [Hewavitharana and Vogel 2011], however, we leave it as future work. We manually evaluated the accuracy based on the number of exact matches. As we evaluated the accuracy manually, the statistical significance could not be evaluated. Note that the exact match criterion has a bias against [Munteanu and Marcu 2006], because their method extracts sub-sentential fragments that are quite long. We found that only six of the fragments extracted by "Munteanu+, 2006" were exact matches, while for the remainder, only partial matches are contained in long fragments. Our proposed method can extract significantly more exactly matched fragments, while the remainders are partial matches. As to the effects of different lexicons, LLR and SampLEX outperform the IBM Model 1 lexicon. We think the reason is the same as the one reported in previous studies: that the LLR and SampLEX lexicons are more precise than the IBM Model 1 lexicon.

We also analyzed the noisy fragment pairs extracted by our proposed method. We found that these noisy pairs are extracted because the lexicon-based filter fails to filter the incorrectly aligned word pairs in the parallel fragment candidates. Most filtering failures are caused by the noisy translation lexicon, and score smoothing also can lead to some failures. Moreover, some filtering failures occur because of both reasons. Table VI shows examples of fragment pairs extracted by our proposed method using LLR lexicon for the lexicon-based filter. In examples 5 and 6, the noisy parts "开始 (begin)" and "が (a case particle)," "聚变(fusion)" and "核 (nuclear)," and "日本 (Japan)" and "、(,)" are extracted because they are incorrectly aligned by the alignment model and exist in the translation lexicon. In example 7, "学术(academic)" and "参考 (reference)" are incorrectly aligned, but they do not exist in the translation lexicon, thus the initial score of this word pair is −1. However, after smoothing, the score becomes positive, and thus this noisy pair is extracted. In example 8, "军衔(rank)" and "就任 (inauguration)"

Table VII. Parallel sentence and fragment integrated translation results

Method	BLEU-4	OOV
Baseline	37.82	3.71%
+Comparable sentences	36.92	2.55%
+Munteanu+, 2006	37.16	3.16%
+IBM model 1	$38.48^{\dagger\ddagger}$	3.68%
+LLR	$\mathbf{38.98}^{\dagger\ddagger *}$	3.68%
+SampLEX	$38.06^{\dagger\ddagger}$	3.68%

Note: "†", "‡" and "*" denote the result is significantly better than "+Munteanu+, 2006", "+Comparable sentences" and "Baseline" respectively at $p < 0.05$.

is a noisy translation lexicon pair and incorrectly aligned. Furthermore, "军衔(rank)" and "に (a case particle)" are also incorrectly aligned, but they do not exist in the translation lexicon. However, after smoothing the score becomes positive, causing this noisy fragment pair.

Based on this analysis, we think that to further improve the accuracy, first, a more efficient alignment model should be used for parallel fragment candidate detection to decrease the number of incorrectly aligned word pairs. Second, the effectiveness of the lexicon-based filter should be further improved. Using a more accurate translation lexicon is the key to improving the lexicon-based filter because the effectiveness of smoothing also highly depends on the accuracy of the translation lexicon. Further cleaning the noisy translation pairs is a possible way to achieve this [Aker et al. 2014], however, we leave it as future work.

5.4. Parallel Sentence and Fragment Integrated Translation Experiments

We conducted Chinese-to-Japanese parallel sentence and fragment integrated translation experiments by appending the extracted fragments to a baseline system. The baseline system used the parallel sentences described in Section 5.2.3 as SMT training data. The other settings were the same as the ones used in the translation experiments described in Section 5.2.2.

We report the translation results on the test set using BLEU-4 [Papineni et al. 2002]. The results of the Chinese-to-Japanese translation experiments are shown in Table VII. For comparison, we also show the translation results of the baseline system (labeled "Baseline") and the system that appends the extracted comparable sentences to the baseline system (labeled "+Comparable sentences"). The significance test was performed using the bootstrap resampling method proposed by Koehn [2004]. We can see that appending the extracted comparable sentences and fragments extracted by [Munteanu and Marcu 2006] has a negative impact on translation quality. Our proposed method outperforms the Baseline, +Comparable sentences, and Munteanu+, 2006 methods, indicating the effectiveness of our proposed integrated extraction system and our proposed method for extracting useful parallel fragments for SMT.

We compared the phrase tables produced by different methods to investigate the reason for the different SMT performances. We found that all methods increased the size of the phrase table, meaning that new phrases are acquired from the extracted data. The sizes are larger for the Comparable sentences and Munteanu+, 2006 methods than they are for our proposed method because these methods extract more data, leading to lower OOV word rates. However, the noise contained in the data extracted by the Comparable sentences and Munteanu+, 2006 methods produces many noisy phrase pairs, which may decrease the SMT performance. Our proposed method extracts accurate parallel fragments, leading to correct new phrases. The LLR lexicon shows the best performance because it extracts more accurate fragments than IBM model 1 and extracts both more and larger parallel fragments than SampLEX.

The parallel sentences described in Section 5.2.3, the parallel fragments extracted by the best method of LLR, and the tuning and testing sets used in the translation experiments are available at http://lotus.kuee.kyoto-u.ac.jp/~chu/wiki_zh_ja/data.tar.gz.

6. CONCLUSION

Extracting parallel data from comparable corpora is an effective way to solve the scarceness of parallel corpora that SMT suffers. Previous studies extract either parallel sentences or fragments from comparable corpora. In this article, we proposed an integrated system to extract both parallel sentences and fragments from comparable corpora to improve SMT. We first applied parallel sentence extraction to identify parallel sentences from comparable sentences. We then extracted parallel fragments from the comparable sentences. Moreover, we proposed novel methods to improve the parallel sentence and fragment extraction components in our system. Experiments conducted on the Chinese–Japanese Wikipedia verified the effectiveness of our proposed system and methods.

As future work, because our study showed that common Chinese characters are helpful for both Chinese–Japanese parallel sentence and fragment extraction, we plan to apply a similar idea to other language pairs by using cognates. Moreover, in this article we only conducted experiments on Wikipedia. Our proposed system is expected to work well on other comparable corpora where both parallel sentences and fragments trend to appear, such as bilingual news articles, social media, and the Web. We plan to do experiments on these comparable corpora to construct a large parallel corpus for various domains.

REFERENCES

Sadaf Abdul-Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved SMT. *Machine Translation* 25, 4 (2011), 341–375.

Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding similar sentences across multiple languages in Wikipedia. In *Proceedings of the workshop on NEW TEXT Wikis and blogs and other dynamic text sources*. 62–69.

Haithem Afli, Loïc Barrault, and Holger Schwenk. 2013. Multimodal Comparable Corpora as Resources for Extracting Parallel Data: Parallel Phrases Extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Nagoya, Japan, 286–292. http://www.aclweb.org/anthology/I13-1033

Ahmet Aker, Yang Feng, and Robert Gaizauskas. 2012. Automatic Bilingual Phrase Extraction from Comparable Corpora. In *Proceedings of COLING 2012: Posters*. The COLING 2012 Organizing Committee, Mumbai, India, 23–32. http://www.aclweb.org/anthology/C12-2003

Ahmet Aker, Monica Paramita, Marcis Pinnis, and Robert Gaizauskas. 2014. Bilingual dictionaries for all EU languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (26-31), Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Reykjavik, Iceland, 2839–2845. http://www.lrec-conf.org/proceedings/lrec2014/pdf/803_Paper.pdf ACL Anthology Identifier: L14-1623.

Rohit G. Bharadwaj and Vasudeva Varma. 2011. Language Independent Identification of Parallel Sentences Using Wikipedia. In *Proceedings of the 20th International Conference Companion on World Wide Web (WWW '11)*. ACM, New York, NY, USA, 11–12. DOI: http://dx.doi.org/10.1145/1963192.1963199

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Association for Computational Linguistics* 19, 2 (1993), 263–312.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27. Issue 3.

Chenhui Chu, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2013a. Chinese–Japanese Machine Translation Exploiting Chinese Characters. *ACM Transactions on*

Asian Language Information Processing (TALIP) 12, 4, Article 16 (Oct. 2013), 25 pages. DOI:http://dx.doi.org/10.1145/2523057.2523059

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2013b. Accurate Parallel Fragment Extraction from Quasi–Comparable Corpora using Alignment Model and Translation Lexicon. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Nagoya, Japan, 1144–1150. http://www.aclweb.org/anthology/I13-1163

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Constructing a Chinese–Japanese Parallel Corpus from Wikipedia. In *Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland, 642–647.

Dan Ştefănescu and Radu Ion. 2013. Parallel-Wiki: A collection of parallel sentences extracted from Wikipedia. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013)*. Samos, Greece, 117–128.

Dan Ştefănescu, Radu Ion, and Sabine Hunsicker. 2012. Hybrid Parallel Sentence Mining from Comparable Corpora. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012)*. Trento, Italy, 137–144.

Thi Ngoc Diep Do, Laurent Besacier, and Eric Castelli. 2010. A Fully Unsupervised Approach for Mining Parallel Data from Comparable Corpora. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation (EAMT 2010)*. St Raphael, France.

Xiaoyin Fu, Wei Wei, Shixiang Lu, Zhenbiao Chen, and Bo Xu. 2013. Phrase-based Parallel Fragments Extraction from Comparable Corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Nagoya, Japan, 972–976. http://www.aclweb.org/anthology/I13-1129

Pascale Fung and Percy Cheung. 2004. Multi-level Bootstrapping For Extracting Parallel Sentences From a Quasi-Comparable Corpus. In *Proceedings of Coling 2004*. COLING, Geneva, Switzerland, 1051–1057.

Pascale Fung, Emmanuel Prochasson, and Simon Shi. 2010. Trillions of Comparable Documents. In *3rd workshop on Building and Using Comparable Corpora (BUCC'10), Language Resource and Evaluation Conference (LREC'10)*. Malta, 26–34.

Souhir Gahbiche-Braham, Hélène Bonneau-Maynard, and François Yvon. 2011. Two Ways to Use a Noisy Parallel News Corpus for Improving Statistical Machine Translation. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*. Association for Computational Linguistics, Portland, Oregon, 44–51. http://www.aclweb.org/anthology/W11-1207

Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto. 2005. Building a Japanese-Chinese dictionary using Kanji/Hanzi conversion. In *Proceedings of the International Joint Conference on Natural Language Processing*. 670–681. http://www.aclweb.org/anthology/I/I05/I05-1059.pdf

Rajdeep Gupta, Santanu Pal, and Sivaji Bandyopadhyay. 2013. Improving MT System Using Extracted Parallel Fragments of Text from Comparable Corpora. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*. Association for Computational Linguistics, Sofia, Bulgaria, 69–76. http://www.aclweb.org/anthology/W13-2509

Sanjika Hewavitharana and Stephan Vogel. 2011. Extracting Parallel Phrases from Comparable Data. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*. Association for Computational Linguistics, Portland, Oregon, 61–68. http://www.aclweb.org/anthology/W11-1209

Gumwon Hong, Chi-Ho Li, Ming Zhou, and Hae-Chang Rim. 2010. An Empirical Study on Web Mining of Parallel Data. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Coling 2010 Organizing Committee, Beijing, China, 474–482. http://www.aclweb.org/anthology/C10-1054

Tatsuya Ishisaka, Masao Utiyama, Eiichiro Sumita, and Kazuhide Yamamoto. 2009. Development of a Japanese-English Software Manual Parallel Corpus. In *MT Summit*.

Long Jiang, Shiquan Yang, Ming Zhou, Xiaohua Liu, and Qingsheng Zhu. 2009. Mining Bilingual Data from the Web with Adaptively Learnt Patterns. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, Suntec, Singapore, 870–878. http://www.aclweb.org/anthology/P/P09/P09-1098

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP 2004*, Dekang Lin and Dekai Wu (Eds.). Association for Computational Linguistics, Barcelona, Spain, 388–395.

Philipp Koehn. 2010. *Statistical Machine Translation* (1st ed.). Cambridge University Press, New York, NY, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Association for Computational Linguistics, Prague, Czech Republic, 177–180. http://www.aclweb.org/anthology/P/P07/P07-2045

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language*. 22–28.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as Parallel Corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 176–186. http://www.aclweb.org/anthology/P13-1018

Bin Lu, Tao Jiang, Kapo Chow, and Benjamin K. Tsou. 2010. Building a Large English-Chinese Parallel Corpus from Comparable Patents and its Experimental Application to SMT. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*. Valletta, Malta, 42–49.

Hiroshi Masuichi, Raymond Flournoy, Stefan Kaufmann, and Stanley Peters. 2000. A Bootstrapping Method for Extracting Bilingual Text Pairs. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2 (COLING '00)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1066–1070. DOI:http://dx.doi.org/10.3115/992730.992806

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics* 31, 4 (December 2005), 477–504.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sydney, Australia, 81–88. DOI:http://dx.doi.org/10.3115/1220175.1220186

Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. 1999. Cross-language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*. ACM, New York, NY, USA, 74–81. DOI:http://dx.doi.org/10.1145/312624.312656

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sapporo, Japan, 160–167. DOI:http://dx.doi.org/10.3115/1075096.1075117

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29, 1 (March 2003), 19–51. DOI:http://dx.doi.org/10.1162/089120103321337421

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. DOI:http://dx.doi.org/10.3115/1073083.1073135

Chris Quirk, Raghavendra Udupa U, and Arul Menezes. 2007. Generative Models of Noisy Translations with Applications to Parallel Fragment Extraction. In *In Proceedings of MT Summit XI, European Association for Machine Translation*.

Philip Resnik and Noah A. Smith. 2003. The Web As a Parallel Corpus. *Comput. Linguist.* 29, 3 (Sept. 2003), 349–380. DOI:http://dx.doi.org/10.1162/089120103322711578

Jason Riesa and Daniel Marcu. 2012. Automatic Parallel Fragment Extraction from Noisy Data. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, 538–542. http://www.aclweb.org/anthology/N12-1061

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Los Angeles, California, 403–411. http://www.aclweb.org/anthology/N10-1063

Chew Lim Tan and Makoto Nagao. 1995. Automatic alignment of Japanese-Chinese bilingual texts. *IEICE Transactions on Information and Systems* E78-D, 1 (1995), 68–76.

Christoph Tillmann. 2009. A Beam-Search Extraction Algorithm for Comparable Data. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, Suntec, Singapore, 225–228. http://www.aclweb.org/anthology/P/P09/P09-2057

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large Scale Parallel Document Mining for Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Coling 2010 Organizing Committee, Beijing, China, 1101–1109. http://www.aclweb.org/anthology/C10-1124

Masao Utiyama and Hitoshi Isahara. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sapporo, Japan, 72–79. DOI:http://dx.doi.org/10.3115/1075096.1075106

Masao Utiyama and Hitoshi Isahara. 2007. A Japanese-English Patent Parallel Corpus. In *Proceedings of MT summit XI*. 475–482.

Ivan Vulić and Marie-Francine Moens. 2012. Sub-corpora Sampling with an Application to Bilingual Lexicon Extraction. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, 2721–2738. http://www.aclweb.org/anthology/C12-1166

Fei Xia, Martha Palmerand Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu dong Chiou, and Shizhe Huang. 2000. Developing guidelines and ensuring consistency for Chinese text annotation. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*. Athens, Greece.

Jiajun Zhang and Chengqing Zong. 2013. Learning a Phrase-based Translation Model from Monolingual Data with Application to Domain Adaptation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 1425–1434. http://www.aclweb.org/anthology/P13-1140

Ying Zhang, Ke Wu, Jianfeng Gao, and Phil Vines. 2006. Automatic Acquisition of Chinese-English Parallel Corpus from the Web.. In *ECIR* (2006-04-03) *(Lecture Notes in Computer Science)*, Mounia Lalmas, Andy MacFarlane, Stefan M. Rü ger, Anastasios Tombros, Theodora Tsikrika, and Alexei Yavlinsky (Eds.), Vol. 3936. Springer, 420–431. http://dblp.uni-trier.de/db/conf/ecir/ecir2006.html#ZhangWGV06

Bing Zhao and Stephan Vogel. 2002. Adaptive Parallel Sentences Mining from Web Bilingual News Collections. In *Proceedings of the 2002 IEEE International Conference on Data Mining*. IEEE Computer Society, Maebashi City, Japan, 745–748.