Chinese–Japanese Parallel Sentence Extraction from Quasi–Comparable Corpora

Chenhui Chu, Toshiaki Nakazawa, Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, 606-8501, Japan
{chu,nakazawa}@nlp.ist.i.kyoto-u.ac.jp kuro@i.kyoto-u.ac.jp

Abstract
Parallel sentences are crucial for statistical machine translation (SMT). However, they are quite scarce for most language pairs, such as Chinese–Japanese. Many studies have been conducted on extracting parallel sentences from noisy parallel or comparable corpora. We extract Chinese–Japanese parallel sentences from quasi–comparable corpora, which are available in far larger quantities. The task is significantly more difficult than the extraction from noisy parallel or comparable corpora. We extend a previous study that treats parallel sentence identification as a binary classification problem. Previous method of classifier training by the Cartesian product is not practical, because it differs from the real process of parallel sentence extraction. We propose a novel classifier training method that simulates the real sentence extraction process. Furthermore, we use linguistic knowledge of Chinese character features. Experimental results on quasi–comparable corpora indicate that our proposed approach performs significantly better than the previous study.

1 Introduction
In statistical machine translation (SMT) (Brown et al., 1993; Koehn et al., 2007), the quality and quantity of the parallel sentences are crucial, because translation knowledge is acquired from a sentence–level aligned parallel corpus. However, except for a few language pairs, such as English–French, English–Arabic and English–Chinese, parallel corpora remain a scarce resource. The cost of manual construction for parallel corpora is high. As non–parallel corpora are far more available, constructing parallel corpora from non–parallel corpora is an attractive research field.

Non–parallel corpora include various levels of comparability: noisy parallel, comparable and quasi–comparable. Noisy parallel corpora contain non–aligned sentences that are nevertheless mostly bilingual translations of the same document, comparable corpora contain non–sentence–aligned, non–translated bilingual documents that are topic–aligned, while quasi–comparable corpora contain far more disparate very–non–parallel bilingual documents that could either be on the same topic (in–topic) or not (out–topic) (Fung and Cheung, 2004). Most studies focus on extracting parallel sentences from noisy parallel corpora or comparable corpora, such as bilingual news articles (Zhao and Vogel, 2002; Utiyama and Isahara, 2003; Munteanu and Marcu, 2005; Tillmann, 2009; Abdul-Rauf and Schwenk, 2011), patent data (Utiyama and Isahara, 2007; Lu et al., 2010) and Wikipedia (Adafre and de Rijke, 2006; Smith et al., 2010). Few studies have been conducted on quasi–comparable corpora. Quasi–comparable corpora are available in far larger quantities than noisy parallel or comparable corpora, while the parallel sentence extraction task is significantly more difficult.

While most studies are interested in language pairs between English and other languages, we focus on Chinese–Japanese, where parallel corpora are very scarce. This study extracts Chinese–Japanese parallel sentences from quasi–comparable corpora. We adopt a system proposed by Munteanu and Marcu (2005), which is for parallel sentence extraction from comparable corpora. We extend the system in several aspects to make it even suitable for quasi–comparable corpora. The core component of the system is a classifier which can identify parallel sentences from non–parallel sentences. Previous method of classifier training by the Cartesian product is not practical, because it differs from the real process of parallel sentence extraction. We propose a novel
method of classifier training and testing that sim-
ulates the real sentence extraction process, which
 Guarantees the quality of the extracted sentences.
 Since Chinese characters are used both in Chi-
 nese and Japanese, they can be powerful linguistic
 clues to identify parallel sentences. Therefore, we
 use Chinese character features, which significantly
 improve the accuracy of the classifier. We con-
duct parallel sentence extraction experiments on
 quasi–comparable corpora, and evaluate the qual-
ity of the extracted sentences from the perspective
 of MT performance. Experimental results show
 that our proposed system performs significantly
 better than the previous study.

2 Parallel Sentence Extraction System

The overview of our parallel sentence extraction
system is presented in Figure 1. Source sentences
are translated to target language using a SMT sys-
tem (1). We retrieve the top N documents from tar-
get language corpora with a information retrieval
(IR) framework, using the translated sentences as
queries (2). For each source sentence, we treat
 all target sentences in the retrieved documents as
candidates. Then, we pass the candidate sentence
pairs through a sentence ratio filter and a word–
overlap–based filter based on a probabilistic dic-
tionary, to reduce the candidates keeping more re-
liable sentences (3). Finally, a classifier trained on
a small number of parallel sentences, is used to
identify the parallel sentences from the candidates
(4). A parallel corpus is needed to train the SMT
system, generate the probabilistic dictionary and
train the classifier.

Our system is inspired by Munteanu and Marcu
(2005), however, there are several differences. The
first difference is query generation. Munteanu and
Marcu (2005) generate queries by taking the top
N translations of each source word according to
the probabilistic dictionary. This method is im-
precise due to the noise in the dictionary. In-
stead, we adopt a method proposed by Abdul–
Rauf and Schwenk (2011). We translate the source
sentences to target language with a SMT system
trained on the parallel corpus. Then use the trans-
lated sentences as queries. This method can gen-
erate more precise queries, because phrase–based
MT is better than word–based translation.

Another difference is that we do not conduct
document matching. The reason is that docu-
ments on the same topic may not exist in quasi–
comparable corpora. Instead, we retrieve the top
N documents for each source sentence. In com-
parable corpora, it is reasonable to only use the
best target sentence in the retrieved documents as
candidates (Abdul-Rauf and Schwenk, 2011). In
quasi–comparable corpora, it is important to fur-
ther guarantee the recall. Therefore, we keep all
target sentences in the retrieved documents as can-
didates.

Our system also differs by the way of classi-
fier training and testing, which is described in Sec-
tion 3 in detail.

3 Binary Classification of Parallel
Sentence Identification

Parallel sentence identification from non–parallel
sentences can be seen as a binary classification
problem (Munteanu and Marcu, 2005; Tillmann,
2009; Smith et al., 2010; Ștefănescu et al., 2012).
Since the quality of the extracted sentences is determined by the accuracy of the classifier, the classifier becomes the core component of the extraction system. In this section, we first describe the training and testing process, then introduce the features we use for the classifier.

3.1 Training and Testing

Munteanu and Marcu (2005) propose a method of creating training and test instances for the classifier. They use a small number of parallel sentences as positive instances, and generate non–parallel sentences from the parallel sentences as negative instances. They generate all the sentence pairs except the original parallel sentence pairs in the Cartesian product, and discard the pairs that do not fulfill the condition of a sentence ratio filter and a word–overlap–based filter. Furthermore, they randomly discard some of the non–parallel sentences when necessary, to guarantee the ratio of negative to positive instances smaller than five for the performance of the classifier.

Creating instances by using the Cartesian product is not practical, because it differs from the real process of parallel sentence extraction. Here, we propose a novel method of classifier training and testing that simulates the real parallel sentence extraction process. For training, we first select 5k parallel sentences from a parallel corpus. Then translate the source side of the selected sentences to target language with a SMT system trained on the parallel corpus excluding the selected parallel sentences. We retrieve the top N documents from the target language side of the parallel corpus, using the translated sentences as queries. For each source sentence, we consider all target sentences in the retrieved documents as candidates. Finally, we pass the candidate sentence pairs through a sentence ratio filter and a word–overlap–based filter, and get the training instances. We treat the sentence pairs that exist in the original 5k parallel sentences as positive instances, while the remainder as negative instances. Note that positive instances may be less than 5k, because some of the parallel sentences do not pass the IR framework and the filters. For the negative instances, we also randomly discard some of them when necessary, to guarantee the ratio of negative to positive instances smaller than five. Test instances are generated by another 5k parallel sentences from the parallel corpus using the same method.

There are several merits of the proposed method. It can guarantee the quality of the extracted sentences, because of the similarity between the real sentence extraction process. Also, features from the IR results can be used to further improve the accuracy of the classifier. The proposed method can be evaluated not only on the test sentences that passed the IR framework and the filters, but also on all the test sentences, which is similar to the evaluation for the real extraction process. However, there is a limitation of our method that a both sentence–level and document–level aligned parallel corpus is needed.

3.2 Features

3.2.1 Basic Features

The following features are the basic features we use for the classifier, which are proposed by Munteanu and Marcu (2005):

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

Alignment features 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 parallel sentences.

3.2.2 Chinese Character Features

Different from 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
Using saturated saline to wash the ether phase, and dry it with anhydrous magnesium.

### Table 1: Examples of common Chinese characters

<table>
<thead>
<tr>
<th>Character</th>
<th>TC (U+96EA)</th>
<th>SC (U+96EA)</th>
<th>Kanji (U+96EA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>snow</td>
<td>爱 (U+611B)</td>
<td>爱 (U+611B)</td>
<td>丸 (U+767A)</td>
</tr>
<tr>
<td>love</td>
<td>爱 (U+611B)</td>
<td>爱 (U+611B)</td>
<td>丸 (U+767A)</td>
</tr>
<tr>
<td>begin</td>
<td>爱 (U+611B)</td>
<td>爱 (U+611B)</td>
<td>丸 (U+767A)</td>
</tr>
</tbody>
</table>

Since Chinese characters contain significant semantic information, and common Chinese characters share the same meaning, they can be valuable linguistic clues for many Chinese–Japanese NLP 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. “snow” in Table 1) 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. “love” in Table 1). Using a Chinese encoding converter that can convert Traditional Chinese into Simplified Chinese, they built a Japanese–Simplified Chinese dictionary partly using direct conversion of Japanese into Chinese for Japanese Kanji words. Chu et al. (2011) made use of the Unihan database to detect common Chinese characters which are visual variants of each other (e.g. “begin” in Table 1), and proved the effectiveness of common Chinese characters in Chinese–Japanese phrase alignment. Chu et al. (2012a) exploited common Chinese characters in Chinese word segmentation optimization, which improved the translation performance.

In this study, we exploit common Chinese characters in parallel sentence extraction. Chu et al. (2011) investigated the coverage of common Chinese characters on a scientific paper abstract parallel corpus, and showed that over 45% Chinese Hanzi and 75% Japanese Kanji are common Chinese characters. Therefore, common Chinese characters can be powerful linguistic clues to identify parallel sentences.

We make use of the Chinese character mapping table created by Chu et al. (2012b) to detect common Chinese characters. Following features are used. We use an example of Chinese–Japanese parallel sentence presented in Figure 2 to explain the features in detail, where common Chinese characters are in bold and linked with dotted lines.

- Number of Chinese characters on each side (Zh: 18, Ja: 14).
- Percentage of Chinese characters out of all characters on each side (Zh: 18/20=90%, Ja: 14/32=43%).
- Ratio of Chinese character numbers on both sides (18/14=128%).
- Number of n-gram common Chinese characters (1-gram: 12, 2-gram: 6, 3-gram: 2, 4-gram: 1).
- Percentage of n-gram common Chinese characters out of all n-gram 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%).

Note that Chinese character features are only applicable to Chinese–Japanese. However, since Chinese and Japanese character information is a kind of cognates (words or languages which have the same origin), the similar idea can be applied to other language pairs by using cognates. Cognates among European languages have been shown effective in word alignments (Kondrak et al., 2003). We also can use cognates for parallel sentence extraction.

---

2. [http://unicode.org/charts/unihan.html](http://unicode.org/charts/unihan.html)
3.3 Rank Feature

One merit of our classifier training and testing method is that features from the IR results can be used. Here, we use the ranks of the retrieved documents returned by the IR framework as feature.

4 Experiments

We conducted classification and translation experiments to evaluate the effectiveness of our proposed parallel sentence extraction system.

4.1 Data

4.1.1 Parallel Corpus

The parallel corpus we used is a scientific paper abstract corpus provided by JST\(^3\) and NICT\(^4\). This corpus was created by the Japanese project “Development and Research of Chinese–Japanese Natural Language Processing Technology”, containing various domains such as chemistry, physics, biology and agriculture etc. This corpus is aligned in both sentence–level and document–level, containing 680k sentences and 100k articles.

4.1.2 Quasi–Comparable Corpora

The quasi–comparable corpora we used are scientific paper abstracts collected from academic websites. The Chinese corpora were collected from CNKI\(^5\), containing 420k sentences and 90k articles. The Japanese corpora were collected from CiNii\(^6\) web portal, containing 5M sentences and 880k articles. Note that since the paper abstracts in these two websites were written by Chinese and Japanese researchers respectively through different periods, documents on the same topic may not exist in the collected corpora. We investigated the domains of the Chinese and Japanese corpora in detail. We found that most documents in the Chinese corpora belong to the domain of chemistry. While the Japanese corpora contain various domains such as chemistry, physics, biology and computer science etc. However, the domain information is unannotated in both corpora.

4.2 Classification Experiments

We conducted experiments to evaluate the accuracy of the proposed method of classification, using different 5k parallel sentences from the parallel corpus as training and test data.

4.2.1 Settings

- Probabilistic dictionary: We took the top 5 translations with translation probability larger than 0.1 created from the parallel corpus.
- IR tool: Indri\(^7\) with the top 10 results.
- Segmenter: For Chinese, we used a segmenter optimized for Chinese–Japanese SMT (Chu et al., 2012a). For Japanese, we used JUMAN (Kurohashi et al., 1994).
- Alignment: GIZA++\(^8\).
- SMT: We used the state–of–the–art phrase–based SMT toolkit Moses (Koehn et al., 2007) with default options, except for the distortion limit (6 → 20).
- Classifier: LIBSVM\(^9\) with 5–fold cross–validation and radial basis function (RBF) kernel.
  - Sentence ratio filter threshold: 2.
  - Word–overlap–based filter threshold: 0.25.
  - Classifier probability threshold: 0.5.

4.2.2 Evaluation

We evaluate the performance of classification by computing precision, recall and F–value, defined as:

\[
\text{precision} = 100 \times \frac{\text{classified well}}{\text{classified parallel}},
\]

\[
\text{recall} = 100 \times \frac{\text{classified well}}{\text{true parallel}},
\]

\[
F \text{– value} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.
\]

Where \text{classified well} is the number of pairs that the classifier correctly identified as parallel, \text{classified parallel} is the number of pairs that the classifier identified as parallel, \text{true parallel} is the number of real parallel pairs in the test set. Note that we only use the top 1 result identified as parallel by the classifier for evaluation.

\(^{\text{3http://www.jst.go.jp}}\)
\(^{\text{4http://www.nict.go.jp}}\)
\(^{\text{5http://www.cnki.net}}\)
\(^{\text{6http://ci.nii.ac.jp}}\)
\(^{\text{7http://www.lemurproject.org/indri}}\)
\(^{\text{8http://code.google.com/p/giza-pp}}\)
\(^{\text{9http://www.csie.ntu.edu.tw/~cjlin/libsvm}}\)
4.2.3 Results

We conducted classification experiments, comparing the following three experimental settings:

- +Chinese character: Add the Chinese character features.
- +Rank: Further add the rank feature.

Results evaluated for the test sentences that passed the IR framework and the filters, and all the test sentences are shown in Table 2. We can see that the Chinese character features can significantly improve the accuracy. The accuracy can be further improved by the rank feature.

4.3 Translation Experiments

We extracted parallel sentences from the quasi-comparable corpora, and evaluated Chinese-to-Japanese MT performance by appending the extracted sentences to two baseline settings.

4.3.1 Settings

- Baseline: Using all the 680k parallel sentences in the parallel corpus as training data (containing 11k sentences of chemistry domain).
- Tuning: Using another 368 sentences of chemistry domain.
- Test: Using another 367 sentences of chemistry domain.
- Language model: 5-gram LM trained on the Japanese side of the parallel corpus (680k sentences) using SRILM toolkit\(^\text{10}\).
- Classifier probability threshold: 0.6.

\(^{10}\text{http://www.speech.sri.com/projects/srilm}\)

The reason we evaluate on chemistry domain is the one we described in Section 4.1.2 that most documents in the Chinese corpora belong to the domain of chemistry. We keep all the sentence pairs rather than the top 1 result (used in the classification evaluation) identified as parallel by the classifier. The other settings are the same as the ones used in the classification experiments.

4.3.2 Results

Numbers of extracted sentences using different classifiers are shown in Table 3, where

- Munteanu+ 2005 (Cartesian): Classifier trained using the Cartesian product, and only using the features proposed by Munteanu and Marcu (2005).
- Munteanu+ 2005 (Proposed): Classifier trained using the proposed method, and only using the features proposed by Munteanu and Marcu (2005).
- +Chinese character (Proposed): Add the Chinese character features.
- +Rank (Proposed): Further add the rank feature.

We can see that the extracted number is significantly decreased by the proposed method compared to the Cartesian product, which may indicate the quality improvement of the extracted sentences. Adding more features further decreases the number.

We conducted Chinese-to-Japanese translation experiments by appending the extracted sentences to the baseline. BLEU–4 scores for experiments are shown in Table 4. We can see that our proposed method of classifier training performs better than the Cartesian product. Adding the Chinese character features and rank feature further improves the translation performance significantly.
Example 1
Zh: 最后最后最后最后，本文说明了说明了说明了说明了光学算符的物理意义的物理意义的物理意义的物理意义。
(Finally, this article explains the physical meaning of the optical operator.)
Ja: 最後に化学ポテンシャルの物理的意味について簡単に説明した。
(Finally, briefly explain the physical meaning of the chemical potential.)

Example 2
Zh: 発光分光分析法の検出限界。 発光分光分析法の検出限界。 発光分光分析法の検出限界。 発光分光分析方法の検出限界。
(Detection limit of emission spectral analysis method. by photoelectric photometry.)
Ja: 光電測光法による発光分光分析方法の検出限界。 発光分光分析方法の検出限界。 発光分光分析法の検出限界。 発光分光分析法の検出限界。
(Detection limit of emission spectral analysis method by photoelectric photometry.)

Figure 3: Examples of extracted sentences (parallel subsentential fragments are in bold).

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>38.64</td>
</tr>
<tr>
<td>Munteanu+ 2005 (Cartesian)</td>
<td>38.10</td>
</tr>
<tr>
<td>Munteanu+ 2005 (Proposed)</td>
<td>38.54</td>
</tr>
<tr>
<td>+Chinese character (Proposed)</td>
<td>38.87</td>
</tr>
<tr>
<td>+Rank (Proposed)</td>
<td>39.47</td>
</tr>
</tbody>
</table>

Table 4: BLEU scores for Chinese–to–Japanese translation experiments (“†” and “*” denotes the result is better than “Baseline” significantly at p < 0.01.

4.3.3 Discussion
The translation results indicate that compared to the previous study, our proposed method can extract sentences with better qualities. However, when we investigated the extracted sentences, we found that most of the extracted sentences are not sentence–level parallel. Instead, they contain many parallel subsentential fragments. Figure 3 presents two examples of sentence pairs extracted by “+Rank (Proposed)”, where parallel subsentential fragments are in bold. We investigated the alignment results of the extracted sentences. We found that most of the parallel subsentential fragments were correctly aligned with the help of the parallel sentences in the baseline system. Therefore, translation performance was improved by appending the extracted sentences. However, it also led to many wrong alignments among the non–parallel fragments which are harmful to translation. In the future, we plan to further extract these parallel subsentential fragments, which can be more effective for SMT (Munteanu and Marcu, 2006).

5 Related Work
As parallel sentences trend to appear in similar document pairs, many studies first conduct document matching, then identify the parallel sentences from the matched document pairs (Utiyama and Isahara, 2003; Fung and Cheung, 2004; Munteanu and Marcu, 2005). Approaches without document matching also have been proposed (Tillmann, 2009; Abdul-Rauf and Schwenk, 2011; Ştefănescu et al., 2012). These studies directly retrieve candidate sentence pairs, and select the parallel sentences using some filtering methods. We adopt a moderate strategy, which retrieves candidate documents for sentences.

The way of parallel sentence identification can be specified with two different approaches: binary classification (Munteanu and Marcu, 2005; Tillmann, 2009; Smith et al., 2010; Ştefănescu et al., 2012) and translation similarity measures (Utiyama and Isahara, 2003; Fung and Cheung, 2004; Abdul-Rauf and Schwenk, 2011). We adopt the binary classification approach with a novel classifier training and testing method and Chinese character features.

Few studies have been conducted for extracting parallel sentences from quasi–comparable corpora. We are aware of only two previous efforts. Fung and Cheung (2004) proposed a multi-level bootstrapping approach. Wu and Fung (2005) exploited generic bracketing Inversion Transduction Grammars (ITG) for this task. Our approach differs from the previous studies that we extend the approach for comparable corpora in several aspects to make it work well for quasi–comparable corpora.

6 Conclusion and Future Work
In this paper, we proposed a novel method of classifier training and testing that simulates the real parallel sentence extraction process. Furthermore, we used linguistic knowledge of Chinese character features. Experimental results of parallel sentence extraction from quasi–comparable corpora indicated that our proposed system performs significantly better than the previous study.
Our approach can be improved in several aspects. One is bootstrapping, which has been proven effective in some related works (Fung and Cheung, 2004; Munteanu and Marcu, 2005). In our system, bootstrapping can be done not only for extension of the probabilistic dictionary, but also for improvement of the SMT system used to translate the source language to target language for query generation. Moreover, as parallel sentences rarely exist in quasi–comparable corpora, we plan to extend our system to parallel subsentential fragment extraction. Our study showed that Chinese character features are helpful for Chinese–Japanese parallel sentence extraction. We plan to apply the similar idea to other language pairs by using cognates.

References


