Integrated Parallel Data Extraction from Comparable Corpora for Statistical Machine Translation

Chenhui CHU

February 2015

Abstract

Machine translation (MT), as a high level application of natural language processing (NLP), is a powerful tool to improve the efficiency and reduce the cost of translation. Over the last decade or two, statistical machine translation (SMT) has been the main approach in both the research community and the commercial sector. In SMT, translation knowledge is automatically acquired from parallel corpora (sentence-aligned bilingual texts), making the rapid development of MT systems for different language pairs and domains possible once parallel corpora are available. Because of the high dependence on parallel corpora, the quality and quantity of parallel corpora are crucial for SMT. However, except for a few language pairs and some specialized domains, high quality parallel corpora of sufficient size remain a scarce resource. 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. The main focus of this thesis is extracting the parallel data from comparable corpora to improve SMT. There are three types of parallel data in comparable corpora: bilingual lexicons, parallel sentences and parallel fragments. In this thesis, we propose novel approaches to extract these three types of parallel data from comparable corpora in an integrated framework. In addition, we exploit linguistic knowledge of common Chinese characters for Chinese-Japanese parallel data extraction as a case study. Bilingual lexicon

extraction (BLE) is used for parallel sentence extraction and improving SMT accuracy. The extracted parallel sentences and fragments are used as training data for SMT. Experiments verify the effectiveness of our proposed approaches for the scarceness of parallel corpora that SMT suffers.

In Chapter 1, we first introduce the mechanism of SMT and the scarceness of parallel corpora. Next, we review the literature of exploiting comparable corpora for SMT. Finally, we briefly describe our approaches and contributions in exploiting comparable corpora for SMT.

In Chapter 2, we propose a method for automatically constructing a more complete resource of common Chinese characters for the Chinese-Japanese language pair using freely available resources. In addition, we propose an approach exploiting common Chinese characters in Chinese word segmentation for SMT. Common Chinese characters are used for parallel sentence (Chapter 4) and fragment extraction (Chapter 5). The optimized segmenter is used throughout this thesis work.

In Chapter 3, we present an iterative BLE system that is based on a novel combination of topic model and context based methods, which are the two main categories of methods that have been proposed for BLE from comparable corpora in the literature. Our system does not rely on any prior knowledge and the performance can be iteratively improved. Experiments conducted on Chinese-English, Japanese-English and Chinese-Japanese Wikipedia data verify the effectiveness of our proposed method.

In Chapter 4, we present a robust parallel sentence extraction system for constructing a Chinese-Japanese parallel corpus from Wikipedia. The system mainly consists of a parallel sentence candidate filter and a classifier for parallel sentence identification. We improve the system by using common Chinese characters for filtering and three novel feature sets for classification. Experiments show that our system performs significantly better than the previous studies for both accuracy in parallel sentence extraction and SMT performance. We further apply the bilingual lexicons extracted in Chapter 3 for parallel sentence extraction.

In Chapter 5, an accurate parallel fragment extraction system is proposed. In many types of comparable corpora, there are parallel fragments existing in

comparable sentences that are also helpful for SMT. We propose a system that uses an alignment model to locate the parallel fragment candidates, and uses an accurate lexicon-based filter to identify the truly parallel ones. We further use common Chinese characters for the lexicon-based filter to improve its coverage. Experiments conducted on Chinese-Japanese comparable corpora indicate that our system can accurately extract parallel fragments. In addition, we show that parallel sentences and fragments can be integrally extracted from some types of comparable corpora.

In Chapter 6, BLE together with paraphrases is proposed for the *accuracy problem* of SMT. The translation pairs and their feature scores in the translation model of SMT can be inaccurate, because of the quality of the unsupervised methods used for translation model learning. Estimating comparable features from comparable corpora with BLE has been proposed for the *accuracy problem* of SMT. However, BLE suffers from the data sparseness, which makes the comparable features inaccurate. We propose using paraphrases to addressing this. Paraphrases are used to smooth the vectors used in comparable feature estimation with BLE. Experiments conducted on Chinese-English SMT verify the effectiveness of our proposed method.

In Chapter 7, we provide concluding remarks and summaries of this thesis, and outline the possible directions for future work.

Acknowledgments

First and foremost, I would like to express my sincere appreciation to Professor Sadao Kurohashi for guiding me into machine translation research, and supervising this thesis and whole my thesis work. He continuously encouraged me throughout this work, and guided me in the right direction. I am so grateful for everything he has taught me, whether it is about technical topics such as how to write readable code or natural language processing, or about the basics of being a researcher such as the way to do research or give a good academic presentation. Moreover, he gave me various supports in daily life during my study at Kyoto University.

I am also very grateful to Professor Toru Ishida and Professor Tatsuya Kawahara for agreeing to serve as members of my doctoral committee, and for their valuable advice and comments about my thesis work.

I would also like to express my gratitude to Dr. Toshiaki Nakazawa, who was in a way the second advisor for most of my thesis work. I have learned a large amount of my knowledge about machine translation from him. He also gave me a lot of suggestive advice, and helped me many times when my research did not go smoothly.

Thank you to Professor Daisuke Kawahara for his help in understanding and using NLP tools such as morphological analyzers, parsers and information retrieval systems. Thank you to Dr. Tomohide Shibata for helping me get familiar to the computational environment in our laboratory, and the machine learning seminars he hosted that help me understand the foundations. I would also like to thank Dr. Isao Goto who is a researcher at NHK, and for many enlightening discussions regarding to the research topic of Chinese analysis for machine translation.

I owe a great deal to all previous and current members in my laboratory. Especially, I would like to thank Dr. Fabien Cromieres, Dr. Yugo Murawaki (currently at Kyushu University), Dr. Jun Harashima (currently at Cookpad), Dr. Masatsugu Hangyo (currently at Weather News), Ben Humpheys, Gongye Jin, Mo Shen, John Richardson. Thank you for all your help and advice. Moreover, I would also like to thank Yuko Ashihara who is the secretary in my laboratory, and for helping me make the administrative matters run smoothly.

Thank you to the people in the research community who kindly helped me by taking the time to answer my questions. They are Dr. Graham Neubig, Professor Chris Callison-Burch, Dr. Alexandre Klementiev, Dr. Ann Irvine, Dr. Juri Ganitkevitch, Professor Lior Wolf, Dr. Tomas Mikolov and Professor Peter Turney.

There are also two organizations that I must thank to, without their supports I could not smoothly finish this work. The first is Hattori International Scholarship Foundation.[1] Hattori International Scholarship Foundation gave me economic supports for the first year of my doctoral course. The second is the Japan Society for the Promotion of Science (JSPS).[2] JSPS gave me economic supports through the "JSPS Research Fellowship for Young Scientists" program for the rest time of my doctoral course.

Finally, I would like to thank my family and friends for their endlessly encouraging and understanding during and before this thesis work.

[1] http://www.hattori-zaidan.or.jp
[2] http://www.jsps.go.jp/english/

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this global era, the demand for translation is rapidly growing in various scenes, and it is impossible to translate everything manually. Machine translation (MT), as a powerful tool to improve the efficiency and reduce the cost of translation, is quite important to promote globalization.

MT is a high level application of natural language processing (NLP), and it has a long history. In the literature, two main approaches have been proposed, namely rule-based and statistical approaches. In the early days of MT research, rule-based MT is the main research direction. In rule-based MT, all the translation rules are written by linguists manually, and then encoded into the MT system. However, because language is too rich and complex, it is impossible to fully analyze and distill it into a set of rules.

Motivated by the development of data-driven statistical approaches in many other NLP problems, MT research turns to a new direction, namely statistical machine translation (SMT) [17, 100, 71]. In SMT, translation knowledge is automatically learned from parallel corpora, making it possible to rapidly develop MT systems for different language pairs and domains once parallel corpora are available. Over the last decade or two, SMT has been the main approach. Nowadays, most MT research is conducted based on this approach. Moreover, the major online translation systems such as Google Translate,[1] Microsoft Bing Translate[2]

[1] https://translate.google.com/

[2] http://www.bing.com/translator/

and Baidu Online Translate,[3] are primarily using this approach.

In SMT, because translation knowledge is acquired from parallel corpora, the quality and quantity of parallel corpora are crucial. However, except for a few language pairs and some specialized domains, high quality parallel corpora of sufficient size remain a scarce resource. 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.

In this chapter, firstly we give a brief introduction about the mechanism of SMT. We then explain the scarceness of parallel corpora that SMT suffers. Next, we describe comparable corpora, and review the literature of exploiting comparable corpora for SMT. Finally, we present our approaches and contributions in exploiting comparable corpora for SMT, and give an outline of this thesis.

1.1 Statistical Machine Translation

The mechanism of SMT can be expressed using the noisy channel model [117]. Given a source sentence $\mathbf{f}$, we want to find the best target sentence translation $\hat{\mathbf{e}}$ that maximizes the conditional probability $p(\mathbf{e}|\mathbf{f})$, where $\mathbf{e}$ is a target sentence. As it is hard to build one complete model, we apply the Bayes rule and decompose it into two sub models:

$$\begin{aligned} \hat{\mathbf{e}} &= \text{argmax}_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) \\ &= \text{argmax}_{\mathbf{e}} \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{p(\mathbf{f})} \\ &= \text{argmax}_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})p(\mathbf{e}) \end{aligned} \tag{1.1}$$

where $p(\mathbf{f}|\mathbf{e})$ is called the translation model, and $p(\mathbf{e})$ is called the language model. The translation model denotes the probability that the source sentence $\mathbf{f}$ is gen-

[3] http://translate.baidu.com

erated when the target sentence $\mathbf{e}$ is given. In SMT, the translation model is trained on parallel corpora in an unsupervised way, whose quality correlates to the quality and quantity of parallel corpora. Different translation models such as word-based models [17], phrase-based models [74] and syntax-based models [45] have been proposed in the literature. The language model denotes the fluency of the target sentence $\mathbf{e}$. It is trained on monolingual corpora, and the n-gram language model is commonly used in SMT.

Och and Ney [99] generalized the noisy channel model of SMT to a log-linear model. The log-linear model can be expressed using the following equation:

$$\hat{\mathbf{e}} = \text{argmax}_{\mathbf{e}} \Big\{ \exp \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}, \mathbf{f}) \Big\} \tag{1.2}$$

where $h_m(\mathbf{e}, \mathbf{f})$ denotes a feature function, and λ_m is its corresponding weight. The noisy channel model can be seen as a special case of the log-linear model when we have the following two feature functions:

$$h_1(\mathbf{e}, \mathbf{f}) = \log p(\mathbf{f}|\mathbf{e}) \tag{1.3}$$

$$h_2(\mathbf{e}, \mathbf{f}) = \log p(\mathbf{e}) \tag{1.4}$$

and set their weights $\lambda_1 = \lambda_2 = 1$. Compared to the noisy channel model, there are two main advantages of the log-linear model:

- It weights different model components, which may improve MT performance.
- It allows including additional model components in the form of feature function.

1.2 Scarceness of Parallel Corpora

With the spread of the web, monolingual corpora become easy to obtain. However, parallel corpora remain a scarce resource. Parallel corpora are sentence-aligned bilingual texts. Figure 1.1 shows an example of a Chinese-Japanese parallel corpus.

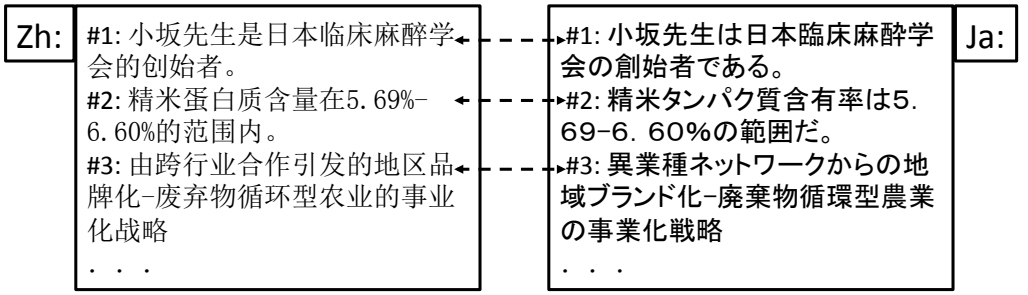

Figure 1.1: Example of a Chinese-Japanese parallel corpus.

Because of the importance of parallel corpora in SMT, various efforts have been made for collecting and constructing parallel corpora. Most existing parallel corpora are collected from manually translated multilingual data such as the United Nation official documents [38], the proceedings of the European Parliament [70], the European Union (EU) legal documents [121], the EU Bookshop [118], the patent family [132, 85] and movie subtitles [149], however such data is very limited and creating such data is very expensive and time-consuming. Parallel corpora also can be constructed by collecting parallel sentences from the Web [125], however the Web can be very noisy, which leads to noisy sentence pairs. Recently, studies have been conducted on constructing parallel corpora via crowdsourcing [147, 106], however these studies have found that it is very difficult to control the quality. Moreover, parallel corpora can be constructed in a collaborative manner such as the Tatoeba project,[4] however how to motivate people to collaborate is a difficult issue.

In addition to the limitations of previous studies, there are several other reasons for the scarceness of parallel corpora:

- Richness of languages. There are about 7,000 languages in the world. The number of possible language pairs equals to the square of the number of languages. Obviously, it is difficult to construct parallel corpora for every language pair, especially for the low resource language pairs.

- Domain diversity. To improve SMT performance, the translation systems

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

should be specialized on particular domains. Constructing parallel corpora in every domain is not an easy task, even for the language pairs that have rich resources.

- Evolution of languages. Languages are evolving over time. Therefore, the SMT training corpora should be updated on a regular basis. Again, this is difficult in the case of parallel corpora.

Table 1.1 shows a list of multilingual parallel corpora that are available online.[5] This list is collected with the help of [71], the website[6] of the state-of-the-art phrase-based SMT toolkit Moses [72] and the open parallel corpus (OPUS) [126] website.[7] We can see that currently 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 limited. For the rest, comprising the majority of language pairs and domains, only few or no parallel corpora are available. Taking Chinese-Japanese as an example, the only available parallel corpus is a scientific domain corpus, containing 680k sentences.

The scarceness of parallel corpora can lead to two main problems of SMT:

- The *coverage problem*. The scarceness of parallel corpora makes the coverage of the translation model low, which leads to high out of vocabulary (OOV) word rates when conducting translation [18]. Even we have parallel corpora in sufficient size in one domain, the *coverage problem* occurs when the domain shifts. Irvine et al. [62] showed that SMT performance decreases significantly when using a system trained on one domain to translate texts in different domains.

- The *accuracy problem*. As described in Section 1.1 the translation model in SMT is automatically learned form parallel corpora in an unsupervised way, and the quality of the unsupervised method used for translation model

[5] There are also several bilingual parallel corpora for particular language pairs, and we do not list them up in Table 1.1.

[6] http://www.statmt.org/moses/?n=Moses.LinksToCorpora

[7] http://opus.lingfil.uu.se

Corpora	Language	Domain	Size (# sentences)
ASPEC[a]	Ja-En, Zh-Ja	science	3M, 680k
LDC[b]	Zh-En, Ar-En	news	100M, 100M (words)
IWSLT BTEC[c]	Asian-En	travel	700k
NTCIR PatentMT[d]	Zh-En, Ja-En	patent	1M, 3M
Multi-UN[e]	7 languages	politics	11M
OpenOffice[f]	8 languages	office	620k
Microtopia[g]	En-5, Zh-9	social media	500k, 1M
WMT news commentary[h]	5 European-En	news	800k
ECB[i]	19 European	banking	30M
EMEA[j]	22 European	medicines	26M
Europarl[k]	21 European	politics	30M
JRC-Acquis[l]	22 European	laws	1B (words)
EUbookshop[m]	48 European	book	173M
TED[n]	33 languages	subtitles	100k
OpenSubtitles[o]	59 languages	subtitles	630M
Tatoeba[p]	129 languages	example	3M
...			

[a] http://orchid.kuee.kyoto-u.ac.jp/ASPEC/
[b] https://www.ldc.upenn.edu
[c] http://iwslt2010.fbk.eu
[d] http://ntcir.nii.ac.jp/PatentMT/
[e] http://www.euromatrixplus.net/multi-un/
[f] http://opus.lingfil.uu.se/OpenOffice3.php/
[g] http://www.cs.cmu.edu/~lingwang/microtopia/
[h] http://www.statmt.org/wmt13/translation-task.html#download
[i] http://opus.lingfil.uu.se/ECB.php
[j] http://opus.lingfil.uu.se/EMEA.php
[k] http://www.statmt.org/europarl/
[l] http://langtech.jrc.it/JRC-Acquis.html
[m] http://opus.lingfil.uu.se/EUbookshop.php
[n] http://www.ted.com/about/programs-initiatives/ted-open-translation-project
[o] http://www.opensubtitles.org/
[p] http://tatoeba.org/eng/

Table 1.1: List of multilingual parallel corpora that are available online (the sizes of the LDC and JRC-Acquis corpora are estimated on words, because the numbers of sentences in these two corpora are not published).

learning always correlates with the amount of parallel corpora. Therefore, the scarceness of parallel corpora can lead to inaccurate translation models [60], naming that the translation pairs and their feature scores in the translation model can be inaccurate.

1.3 Comparable Corpora

Comparable corpora are a set of monolingual corpora describing roughly the same topic in different languages, which are not exact translation equivalents of each other. We believe that exploiting comparable corpora is an effective way to addressing the scarceness of parallel corpora for SMT. Firstly, when using parallel corpora one bilingual corpus is required for each language pair. In contrast, when using comparable corpora one monolingual corpus per language suffices, and monolingual corpora are easy to obtain. Secondly, comparable corpora are far more available for various domains than parallel corpora, such as Wikipedia, patent documents, news articles and academic papers. Thirdly, there are a large amount of parallel data in comparable corpora, such as bilingual lexicons, parallel sentences and parallel fragments. Comparable corpora have various granularities. In comparable corpora with high comparability, there are many parallel sentences. While in comparable corpora with low comparability, there are few parallel sentences. However, there could be bilingual lexicons and parallel fragments. Moreover, there could be bilingual lexicons, parallel sentences and fragments in one comparable corpus, in which comparable texts with different comparabilities are contained (e.g., Wikipedia). Figure 1.2 shows an example of aligned Chinese-Japanese comparable texts describing a French city "Sète" from Wikipedia.[8] We can see that there are three types of parallel data: bilingual lexicons, parallel sentences and parallel fragments in the comparable texts.

Compared to parallel corpora, research on comparable corpora for SMT is still at an earlier stage. However, the history is not short, which has been ongoing for almost 20 years. As this is a very challenging and important problem, research

[8]In Wikipedia, articles in different languages on the same topic are manually aligned via interlanguage links by the authors.

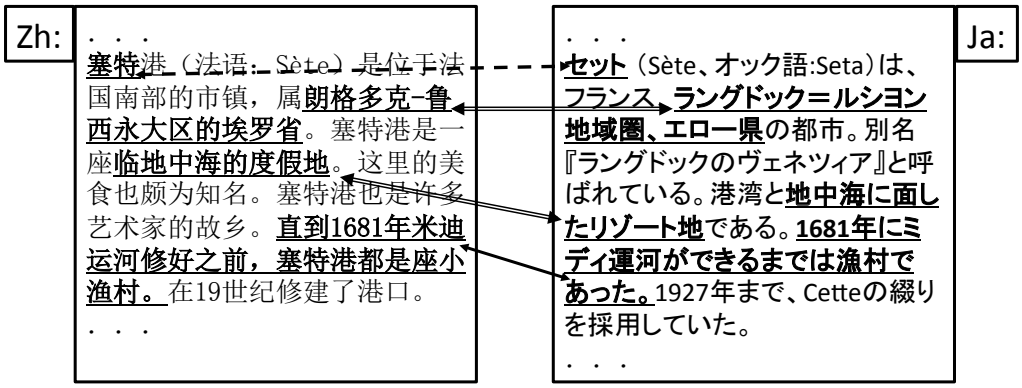

Figure 1.2: Example of comparable texts describing the French city "Sète" from Wikipedia (bilingual lexicons are linked with dashed lines, parallel sentences are linked with solid lines, and parallel fragments are linked with double lines).

interest has steadily increased. Previous studies on comparable corpora mainly focus on the following directions.

- Parallel data extraction.
  - Bilingual lexicon extraction (BLE). From the early work of [110, 40], BLE has the longest history in exploiting comparable corpora for SMT. The main goal of BLE is the construction of bilingual dictionaries, which are important for both SMT and cross-lingual information retrieval (CLIR) [105]. BLE from comparable corpora is based on the distributional hypothesis [54], stating that words with similar meaning appear in similar distributions across languages. Contextual similarity is mostly used in BLE.
  - Parallel sentence extraction. It identifies parallel sentences from comparable corpora, and automatically constructs parallel corpora for SMT. Parallel sentences can be identified based on classification [93] or using some translation similarity measures [131]. Similar features such as word overlap and sentence length based features are used in both of these two approaches.
  - Parallel fragment extraction. Although there are few or no parallel

sentences in comparable corpora with low comparability, there could be parallel fragments in comparable sentences. Parallel fragments are also helpful for SMT. Parallel fragment extraction mainly relies on bilingual lexicons [94] or alignment models [109].

- Translation model improvement.
  - As described in Section 1.2, the scarceness of parallel corpora can lead to the *coverage problem* of SMT. BLE can be used to addressing this problem, which mines translations for the unknown words or phrases in the translation model from comparable corpora [34]. Parallel fragment extraction also has been used for this problem, which extracts fragments from comparable corpora to construct a new translation model [148].
  - As described in Section 1.2, the scarceness of parallel corpora also can lead to the *accuracy problem* of SMT. BLE can be used to addressing this problem, which estimates comparable features from comparable corpora for the translation pairs in the translation model [67].
- Language model adaptation. It retrieves target side comparable sentences [151] or documents [120] for a source sentence or document from comparable corpora using CLIR, and trains a specific language model on the retrieved data. This adapted language model is used when translating the source sentence or document, which is helpful for generating target resemble translations and thus improve MT performance.

Besides the increase of research interest on comparable corpora, there are also a considerable number of research projects (such as ACCURAT[9] and TTC.[10]) that devote fully or partly using comparable corpora for SMT. Moreover, the workshop series on "Building and Using Comparable Corpora" (BUCC)[11] is now in its seventh year, and publish a related book.[12]

[9] http://www.accurat-project.eu/
[10] http://www.ttc-project.eu/
[11] http://comparable.limsi.fr/bucc2014/
[12] http://www.springer.com/computer/ai/book/978-3-642-20127-1

1.4 Overview of Our Approach

The main focus of this thesis is exploiting comparable corpora to addressing the scarceness of parallel corpora. We propose an integrated framework for this. The overview of our approach is presented in Figure 1.3. As initial, we have comparable corpora and a small seed parallel corpus. We first generate a bilingual dictionary from the seed parallel corpus. As the coverage of this dictionary is low, we further extract bilingual lexicons from comparable corpora ((1) in Figure 1.3) and combine them with the generated dictionary. Using the combined dictionary, we can apply CLIR [105] to generate parallel sentence candidates from comparable corpora.[13] Next, we apply parallel sentence extraction that can classify the parallel sentence candidates into parallel and comparable sentences ((2) in Figure 1.3). We then apply parallel fragment extraction to extract parallel fragments from the comparable sentences ((3) in Figure 1.3). The combined dictionary can be used for both parallel sentence and fragment extraction. Moreover, the extracted parallel sentences can be used to support parallel fragment extraction. The extracted parallel sentences and fragments are used as training data for SMT, which are helpful to addressing both the *coverage and accuracy problems* of SMT caused by the scarceness of parallel corpora described in Section 1.2. Also, they can be appended to the seed parallel corpus for bootstrapping. Finally, we apply BLE to further improve the accuracy of SMT ((4) in Figure 1.3).

The framework is language independent, and can be further improved using language specific knowledge. In this thesis work, we further exploit linguistic knowledge for the Chinese-Japanese language pair as a case study. A special characteristic of the Chinese-Japanese languages is that they share common Chinese characters[14] [26]. Because common Chinese characters share the same meaning, they can be valuable linguistic clues for Chinese-Japanese parallel data extraction. In this work, we use common Chinese characters for both Chinese-Japanese parallel sentence and fragment extraction ((5) in Figure 1.3).

[13] For comparable corpora that article alignment has been manually established such as Wikipedia, CLIR is not required for parallel sentence candidate generation.

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

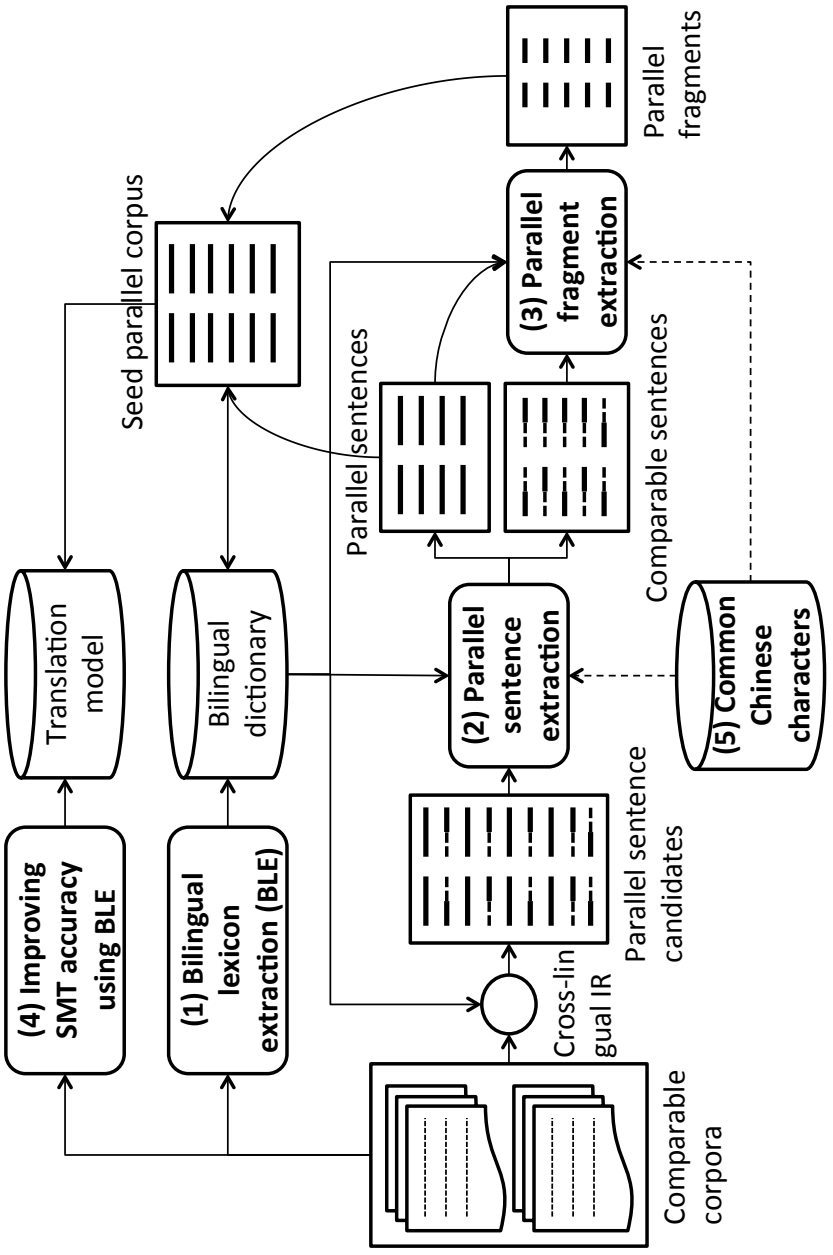

Figure 1.3: Overview of our approach.

The main contributions of this thesis can be summarized as follows:

- We propose an integrated parallel data extraction framework for SMT. Although different approaches have been proposed for extracting bilingual lexicons, parallel sentences and parallel fragments, and the *accuracy problem* of SMT using BLE as we described in Section 1.3, all the previous studies treat them as individual tasks. This is the first study that proposes an integrated framework, in which all the tasks are closely connected and can benefit each other.

- We propose novel approaches to extract bilingual lexicons, parallel sentences and parallel fragments from comparable corpora respectively, and a novel approach for using BLE to addressing the *accuracy problem* of SMT.

- We show that common Chinese characters are helpful for Chinese-Japanese parallel data extraction. Although common Chinese characters are specific for the Chinese-Japanese language pair, however, a similar idea can be applied to other language pairs that share cognates.

1.5 Outline of This Thesis

The rest of this thesis is structured as follows.

In Chapter 2, we propose a method for automatically creating a Chinese character mapping table using freely available resources, and construct a more complete resource containing common Chinese characters for the Chinese-Japanese language pair. In addition, we propose an approach exploiting common Chinese characters in Chinese word segmentation for SMT. The mapping table is used for parallel sentence (Chapter 4) and fragment extraction (Chapter 5). The optimized segmenter is used throughout this thesis work.

In Chapter 3, an iterative BLE system with topical and contextual knowledge is presented. In the literature, two main categories of methods have been proposed for BLE from comparable corpora, namely topic model and context based methods. We present a BLE system that is based on a novel combination of these

two methods in an iterative process. Our system does not rely on any prior knowledge and the performance can be iteratively improved. Experiments conducted on Chinese-English, Japanese-English and Chinese-Japanese Wikipedia data show that our proposed method significantly outperforms the previous studies.

In Chapter 4, a robust parallel sentence extraction system is presented. The system is inspired by previous studies that mainly consist of a parallel sentence candidate filter and a classifier for parallel sentence identification. We improve the system by using common Chinese characters for filtering and classification. Experiments show that our system performs significantly better than the previous studies for both accuracy in parallel sentence extraction and SMT performance. Using the system, we construct a Chinese-Japanese parallel corpus with more than 126k highly accurate parallel sentences from Wikipedia. We further apply the bilingual lexicons extracted in Chapter 3 for parallel sentence extraction.

In Chapter 5, we propose an accurate parallel fragment extraction system using alignment model and bilingual lexicon. Previous studies have found it difficult to accurately extract parallel fragments from comparable sentences. To addressing this, we propose an accurate parallel fragment extraction system that uses an alignment model to locate the parallel fragment candidates, and uses an accurate lexicon-based filter to identify the truly parallel ones. We further use common Chinese characters for the lexicon-based filter to improve its coverage. Experimental results on Chinese-Japanese comparable corpora indicate that our system can accurately extract parallel fragments. In addition, we show that parallel sentences and fragments can be integrally extracted from some types of comparable corpora.

In Chapter 6, we propose using BLE together with paraphrases to addressing the *accuracy problem* of SMT. Previous studies propose estimating comparable features for the translation pairs in the translation model from comparable corpora, to improve the accuracy of the translation model. Comparable feature estimation is based on BLE technology. However, BLE suffers from the data sparseness, which makes the comparable features inaccurate. We propose using paraphrases to addressing this. Paraphrases are used to smooth the vectors used in comparable feature estimation with BLE. In this way, we improve the quality

of comparable features, which can improve the accuracy of the translation model thus improve SMT performance. Experiments conducted on Chinese-English SMT verify the effectiveness of our proposed method.

In Chapter 7, we summarize this thesis, and remark on possible future directions of this work.

Chapter 2

Common Chinese Characters

Differing from other language pairs, Chinese and Japanese share Chinese characters. In Chinese, 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 in Hanzi and Kanji.

Because Chinese characters contain a significant amount of semantic information, and common Chinese characters share the same meaning, they can be valuable linguistic clues in many Chinese-Japanese natural language processing (NLP) tasks. Many studies have exploited common Chinese characters. For example, Tan et al. [124] used the occurrence of identical common Chinese characters in Chinese and Japanese in an automatic sentence alignment task. Goh et al. [50] detected common Chinese characters where Kanji are identical to Traditional Chinese, but differ from Simplified Chinese. Using a Chinese encoding converter[1] 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. Huang et al. [59] examined and analyzed the semantic relations between Chinese and Japanese at a word level

[1] http://www.mandarintools.com/zhcode.html

	C1	C2	C3	C4	C5	C6
Meaning	snow	love	country	begin	octopus	included
Kanji	雪	愛	国	発	鱚	込
Traditional Chinese	雪	愛	國	發	鱚	N/A
Simplified Chinese	雪	爱	国	发	N/A	N/A

Table 2.1: Examples of Chinese characters ("C" denotes Category, which is described in Section 2.1.4).

based on a common Chinese character mapping. They used a small list of 125 visual variational pairs of manually matched common Chinese characters.

However, the resources for common Chinese characters used in these previous studies are not complete. In this chapter, we propose a method for automatically creating a Chinese character mapping table for Japanese, Traditional Chinese, and Simplified Chinese using freely available resources, with the aim of constructing a more complete resource containing common Chinese characters.

In addition, we point out two main problems in Chinese word segmentation for Chinese-Japanese SMT, namely, unknown words and word segmentation granularity. In Chinese-Japanese SMT, parallel sentences contain equivalent meanings in each language, and we assume that common Chinese characters appear in the sentences. Therefore, we propose an approach exploiting common Chinese characters to solve these problems. Experimental results show that our proposed approaches improve SMT performance significantly.

2.1 Chinese Character Mapping Table

Table 2.1 gives some examples of Chinese characters in Japanese, Traditional Chinese, and Simplified Chinese, from which we can see that the relation between Kanji and Hanzi is quite complicated.

Because Kanji characters originated from ancient China, most Kanji have fully corresponding Chinese characters in Hanzi. In fact, despite Japanese having continued to evolve and change because its adoption of Chinese characters, the visual

forms of the Chinese characters have retained a certain level of similarity; many Kanji are identical to Hanzi (e.g., "雪 (snow)" in Table 2.1), some Kanji are identical to Traditional Chinese characters but differ from Simplified Chinese ones (e.g., "愛 (love)" in Table 2.1), while others are identical to Simplified Chinese characters but differ from Traditional Chinese ones (e.g., "国 (country)" in Table 2.1). There are also some visual variations in Kanji that have corresponding Chinese characters in Hanzi although the shapes differ from those in Hanzi (e.g., "発 (begin)" in Table 2.1). However, there are some Kanji that do not have fully corresponding Chinese characters in Hanzi. Some Kanji only have corresponding Traditional Chinese characters (e.g., "鱆 (octopus)" in Table 2.1), because they were not simplified into Simplified Chinese. Moreover, there are some Chinese characters that originated in Japan namely, Kokuji, which means that these national characters may have no corresponding Chinese characters in Hanzi (e.g., "込 (included)" in Table 2.1).

What makes the relation even more complicated is that a single Kanji form may correspond to multiple Hanzi forms. Also, a single Simplified Chinese form may correspond to multiple Traditional Chinese forms, and vice versa.

Focusing on the relation between Kanji and Hanzi, we present a method for automatically creating a Chinese character mapping table for Japanese, Traditional Chinese, and Simplified Chinese using freely available resources [27]. Common Chinese characters shared in Chinese and Japanese can be found in the mapping table. Because Chinese characters contain significant semantic information, this mapping table could be very useful in Chinese-Japanese MT.

2.1.1 Related Work

Hantology [23] is a character-based Chinese language resource, which has adopted the Suggested Upper Merged Ontology (SUMO) [97] for a systematic and theoretical study of Chinese characters. Hantology represents orthographic forms, the evolution of script, pronunciation, senses, lexicalization, as well as variants for different Chinese characters. However, the variants in Hantology are limited to Chinese Hanzi.

Chou et al. [24] extended the architecture of Hantology to Japanese Kanji, and

included links between Chinese Hanzi and Japanese Kanji, thereby providing a platform for systematically analyzing variations in Kanji. However, a detailed analysis of variants of Kanji has not been presented. Moreover, because the current version of Hantology only contains 2,100 Chinese characters, whereas our mapping table includes all 6,355 JIS Kanji, it is difficult to create a mapping table between Kanji and Hanzi, based on Hantology, and which is as complete as our proposed method.

2.1.2 Kanji and Hanzi Character Sets

The character set in use for Kanji is JIS Kanji code, whereas for Hanzi, there are several, of which we have selected Big5 for Traditional Chinese and GB2312 for Simplified Chinese, both of which are widely used.

- For JIS Kanji code, JIS X 0208 is a widely used character set specified as the Japanese Industrial Standard, containing 6,879 graphic characters, including 6,355 Kanji and 524 non-Kanji. The mapping table is for the 6,355 Kanji characters, that is, JIS Kanji, in JIS X 0208.

- Big5 is the most commonly used character set for Traditional Chinese in Taiwan, Hong Kong, and Macau, and was defined by the "Institute for Information Industry" in Taiwan. There are 13,060 Traditional Chinese characters in Big5.

- GB2312 is the main official character set of the People's Republic of China for Simplified Chinese characters, and is widely used in mainland China and Singapore. GB2312 contains 6,763 Simplified Chinese characters.

2.1.3 Related Freely Available Resources

- Unihan database[2] is the repository for the Unicode Consortium's collective knowledge regarding the CJK (Chinese-Japanese-Korean) Unified Ideographs

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

Traditional Chinese	故	說	錢	沖, 衝	干, 幹, 乾	...
Simplified Chinese	故	说	钱	冲	干	...

Table 2.2: Hanzi converter standard conversion table.

contained in the Unicode Standard[3]. The database consists of a number of fields containing data for each Chinese character in the Unicode Standard. These fields are grouped into categories according to their purpose, including "mappings," "readings," "dictionary indices," "radical stroke counts," and "variants." The "mappings" and "variants" categories contain information regarding the relation between Kanji and Hanzi.

- The Chinese encoding converter[4] is a open source system that converts Traditional Chinese into Simplified Chinese. The Hanzi converter standard conversion table, a resource used by the converter, contains 6,740 corresponding Traditional Chinese and Simplified Chinese character pairs. It can be downloaded from the website. Table 2.2 depicts a portion of the table.

- Kanconvit[5] is a publicly available tool for Kanji-Simplified Chinese conversion. It uses 1,159 visual variational Kanji-Simplified Chinese character pairs extracted from a Kanji, Traditional Chinese, and Simplified Chinese mapping table, containing 3,506 one-to-one mappings. Table 2.3 depicts a portion of this table.

2.1.4 Construction Method

Based on the relation between Kanji and Hanzi, we define the following seven categories for Kanji.

[3]The Unicode Standard is a character coding system for the consistent encoding, representation and handling of text expressed in most of the world's writing systems. The latest version of the Unicode Standard is 6.1.0.

[4]http://www.mandarintools.com/zhcode.html

[5]http://kanconvit.ta2o.net/

Kanji	安	詞	会	広	壱	潟	...
Traditional Chinese	安	詞	會	廣	壹	瀉	...
Simplified Chinese	安	词	会	广	壹	泻	...

Table 2.3: Kanconvit mapping table.

- Category 1: identical to Hanzi
- Category 2: identical to Traditional Chinese, but different from Simplified Chinese
- Category 3: identical to Simplified Chinese, but different from Traditional Chinese
- Category 4: visual variations
- Category 5: with a corresponding Traditional Chinese character only
- Category 6: no corresponding Hanzi
- Others: does not belong to the above categories

We create a Chinese character mapping table for Japanese, Traditional Chinese, and Simplified Chinese by classifying JIS Kanji into these seven categories and automatically finding the corresponding Traditional Chinese and Simplified Chinese characters using the resources introduced in Section 2.1.3. The method involves two steps:

- Step 1: extraction
- Step 2: categorization and construction

In Step 1, we extract the JIS Kanji, Big5 Traditional Chinese, and GB2312 Simplified Chinese from the Unihan database. These Chinese characters are collected in the "mappings" category, which contains mappings between Unicode and other encoded character sets for Chinese characters. JIS Kanji are obtained from the "kIRG_JSource J0" field, Big5 Traditional Chinese from the "kBigFive" field, and GB2312 Simplified Chinese from the "kIRG_GSource G0" field.

In Step 2, we categorize the JIS Kanji and construct a mapping table. We automatically check every character in the JIS Kanji as follows. If the Kanji exists in both Big5 and GB2312, it belongs to Category 1. If the Kanji exists only in Big5, we check whether a corresponding Simplified Chinese character can be found; if so, it belongs to Category 2, otherwise, it belongs to Category 5. If the Kanji exists only in GB2312, we check whether a corresponding Traditional Chinese character can be found; if so, it belongs to Category 3. If the Kanji exists in neither Big5 nor GB2312, we check whether corresponding Hanzi can be found; if a fully corresponding Chinese character exists in Hanzi, it belongs to Category 4, else if only a corresponding Traditional Chinese character exists, it belongs to Category 5, else if no corresponding Chinese character exists in Hanzi, it belongs to Category 6, otherwise, it belongs to Others.

To find the corresponding Hanzi, we search Traditional Chinese and Simplified Chinese variants, as well as other variants for all Kanji. This search is carried out using the "variants" category in the Unihan database, in which there are five fields: "kTraditionalVariant" corresponding to Traditional Chinese variants, "kSimplifiedVariant" corresponding to Simplified Chinese variants, and "kZVariant," "kSemanticVariant," and "kSpecializedSemanticVariants" corresponding to the other variants. In addition, we also use the Hanzi converter standard conversion table and Kanconvit mapping table. Note that the resources in the Hanzi converter standard conversion table can only be used for the Traditional Chinese and Simplified Chinese variants search, whereas the Kanconvit mapping table can also be used for the other variants search.

2.1.5 Details of the Mapping Table

The format for Kanji in Categories 1, 2, 3, and 4 in the mapping table is as follows:

- Kanji[TAB]Traditional Chinese[TAB]Simplified Chinese[RET]

If multiple Hanzi forms exist for a single Kanji, we separate them with ",." Table 2.4 shows some examples of multiple Hanzi forms. The formats for Kanji in Categories 5 and 6 are as follows:

- Category 5: Kanji[TAB]Traditional Chinese[TAB]N/A[RET]

Kanji	弁	伝	鯰	働	...
Traditional Chinese	弁, 瓣,辦, 辯, 辮, 辨	傳, 伝	鯰	動, 仂	...
Simplified Chinese	弁, 瓣,办,辩,辫, 辨	传	鲶,鮎	动, 仂	...

Table 2.4: Examples of multiple Hanzi forms.

	C1	C2	C3	C4	C5	C6	Others
Unihan	3141	1815	177	533	384	289	16
+Han	3141	1843	177	542	347	289	16
+Kan	3141	1847	177	550	342	282	16

Table 2.5: Resource statistics ("Han" denotes the Hanzi converter standard conversion table, while "Kan" denotes the Kanconvit mapping table).

- Category 6: Kanji[TAB]N/A[TAB]N/A[RET]

Table 2.5 gives some statistics of the Chinese character mapping table we created for Japanese, Traditional Chinese, and Simplified Chinese. Here, "Others" are the Kanji that have a corresponding Simplified Chinese character only. There are corresponding Traditional Chinese characters for these Kanji, but they were not collected in Big5 Traditional Chinese. Kanji "鮃 (bastard halibut)" is one of such examples. Compared with using only the Unihan database, incorporating the Hanzi converter standard conversion and Kanconvit mapping tables can improve the completeness of the mapping table. Tables 2.6 and 2.7 give some examples of additional Chinese character mappings found using the Hanzi converter standard conversion table and Kanconvit mapping table, respectively.

2.1.6 Completeness Evaluation

To show the completeness of the mapping table we created, we used a resource from Wiktionary[6], which is a wiki project aimed at producing a free-content multilingual dictionary. In the Japanese version of Wiktionary, there is a Kanji category that provides a great deal of information about Kanji, such as variants, origins,

[6]http://www.wiktionary.org/

Kanji	祇	託	浄	畭	...
Traditional Chinese	祇, 只,衹, 隻, 祇	託,侂, 托	淨,凈	畬	...
Simplified Chinese	祇, 只	托	凈	畬	...

Table 2.6: Examples of additional mappings found using the Hanzi converter standard conversion table.

Kanji	雰	艶	対	県	挿	...
Traditional Chinese	氛, 雰	豔, 艷	對	縣	插	...
Simplified Chinese	氛	艳	对	县	插	...

Table 2.7: Examples of additional mappings found using the Kanconvit mapping table.

meanings, pronunciation, idioms, Kanji in Chinese and Korean, and codes. We are interested in the variants part. Figure 2.1 gives an example of Kanji "広" from the Japanese Wiktionary, in which the variants part, containing the Traditional Chinese variant "廣," Simplified Chinese variant "广," and other variant "廎" of Kanji "広," is enclosed by a rectangle.

We downloaded the Japanese Wiktionary database dump data[7] (2012-Jan-31) and extracted the variants for JIS Kanji. We then constructed a mapping table based on the Wiktionary using the method described in Section 2.1.4, the only difference being that for the Traditional Chinese, Simplified Chinese, and other variants search, we used the variants extracted from the Japanese Wiktionary.

To evaluate the completeness of the mapping table created using the proposed method, we compared the statistics thereof with those of Wiktionary. Table 2.8 shows the completeness comparison between the proposed method and Wiktionary. We can see that the proposed method creates a more complete mapping table than Wiktionary. Table 2.9 gives some examples of Chinese character mappings found by the proposed method, but which do not exist in the current version of Wiktionary.

Furthermore, we carried out an experiment by combining the mapping table

[7]http://dumps.wikimedia.org/jawiktionary/

Figure 2.1: Example of Kanji "広" from Japanese Wiktionary.

	C1	C2	C3	C4	C5	C6	Others
Proposed	3141	1847	177	550	342	282	16
Wiktionary	3141	1781	172	503	412	316	30
Combination	3141	1867	178	579	325	249	16

Table 2.8: Completeness comparison between proposed method and Wiktionary.

we created with Wiktionary. The results in Table 2.8 show that Wiktionary can be used as a supplementary resource to further improve the completeness of the mapping table. Table 2.10 gives some examples of Chinese character mappings contained in Wiktionary, but which were not found by the proposed method.

2.1.7 Coverage of Common Chinese Characters

We investigated the coverage of common Chinese characters on a Simplified Chinese-Japanese corpus, namely, the Chinese-Japanese section of the Asian Scientific

Kanji	尨	荔	値	幇	咲	...
Traditional Chinese	尨, 龍	荔	值	幫	笑	...
Simplified Chinese	龙	荔	值	帮	笑	...

Table 2.9: Examples of mappings that do not exist in Wiktionary.

Kanji	冴	扨	疂	滝	愼	...
Traditional Chinese	冴, 洗	扠, 叉	疊	灑	慎	...
Simplified Chinese	冴	叉	叠	洒	慎	...

Table 2.10: Examples of mappings not found by the proposed method.

Paper Excerpt Corpus (ASPEC).[8] This corpus is a scientific domain corpus provided by the Japan Science and Technology Agency (JST)[9] and the National Institute of Information and Communications Technology (NICT).[10] It was created by the Japanese project "Development and Research of Chinese-Japanese Natural Language Processing Technology." Some statistics of this corpus are given in Table 2.11.

We measured the coverage in terms of both characters and words under two different experimental conditions:

- Identical: only exactly the same Chinese characters.

- +Common: perform Kanji to Hanzi conversion for common Chinese characters using the Chinese character mapping table constructed as described in Section 2.1.

Table 2.12 presents the coverage results for common Chinese characters. If we use all the resources available, we can find corresponding Hanzi characters for over 76% of the Kanji characters.

[8] http://lotus.kuee.kyoto-u.ac.jp/ASPEC
[9] http://www.jst.go.jp
[10] http://www.nict.go.jp

	Ja	Zh
# of sentences	680k	
# of words	21.8M	18.2M
# of Chinese characters	14.0M	24.2M
average sentence length	32.9	22.7

Table 2.11: Statistics of Chinese-Japanese corpus.

	character		word	
	Ja	Zh	Ja	Zh
Identical	52.41%	30.48%	26.27%	32.09%
+Common	76.66%	44.58%	32.84%	39.46%

Table 2.12: Coverage of common Chinese characters.

2.2 Exploiting in Chinese Word Segmentation Optimization

As there are no explicit word boundary markers in Chinese, word segmentation is considered an important first step in MT. Studies have shown that an MT system with Chinese word segmentation outperforms those treating each Chinese character as a single word, while the quality of Chinese word segmentation affects MT performance [145, 21]. It has been found that besides segmentation accuracy, segmentation consistency and granularity of Chinese words are also important for MT [21]. Moreover, optimal Chinese word segmentation for MT is dependent on the other language, and therefore, a bilingual approach is necessary [86].

Most studies have focused on language pairs containing Chinese and another language with white spaces between words (e.g., English). Our focus is on Chinese-Japanese MT, where segmentation is needed on both sides. Segmentation for Japanese successfully achieves an F-score of nearly 99% [76], while that for Chinese is still about 95% [141]. Therefore, we only do word segmentation optimization in Chinese, and use the Japanese segmentation results directly.

Similar to previous works, we also consider the following two Chinese word

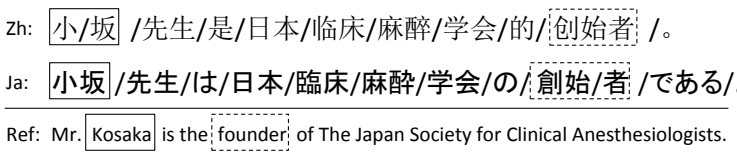

Figure 2.2: Example of Chinese word segmentation problems in Chinese-Japanese MT.

segmentation problems to be important for Chinese-Japanese MT. The first problem relates to unknown words, which cause major difficulties for Chinese segmenters and affect segmentation accuracy and consistency. Consider, for example, "Kosaka" shown in Figure 2.2, which is a proper noun in Japanese. Because "Kosaka" is a unknown word for a Chinese segmenter, it is mistakenly segmented into two tokens, whereas the Japanese word segmentation result is correct.

The second problem is word segmentation granularity. Most Chinese segmenters adopt the famous Penn Chinese Treebank (CTB) standard [144], while most Japanese segmenters adopt a shorter unit standard. Therefore, the segmentation unit in Chinese may be longer than that in Japanese even for the same concept. This can increase the number of 1-to-n alignments making the word alignment task more difficult. Taking "founder" in Figure 2.2 as an example, the Chinese segmenter recognizes it as one token, while the Japanese segmenter splits it into two tokens because of the different word segmentation standards.

To solve the above problems, we proposed an approach based on a bilingual perspective that exploits common Chinese characters shared between Chinese and Japanese in Chinese word segmentation optimization for MT [25]. In this approach, Chinese entries are extracted from a parallel training corpus based on common Chinese characters to augment the system dictionary of a Chinese segmenter. In addition, the granularity of the training data for the Chinese segmenter is adjusted to that of the Japanese one by means of extracted Chinese entries.

2.2.1 Related Work

Exploiting lexicons from external resources [104, 21] is one way of dealing with the unknown word problem. However, the external lexicons may not be very efficient

for a specific domain. Some studies [145, 86] have used the method of learning a domain specific dictionary from the character-based alignment results of a parallel training corpus, which separate each Chinese character, and consider consecutive Chinese characters as a lexicon in n-to-1 alignment results. Our proposed method differs from these studies in that we obtain a domain specific dictionary by extracting Chinese lexicons directly from a segmented parallel training corpus, making word alignment unnecessary.

The goal of our proposed short unit transformation method is to form the segmentation results of Chinese and Japanese into a 1-to-1 mapping, which can improve alignment accuracy and MT performance. Bai et al. [10] proposed a method for learning affix rules from an aligned Chinese-English bilingual terminology bank to adjust Chinese word segmentation in the parallel corpus directly with the aim of achieving the same goal. Our proposed method does not adjust Chinese word segmentation directly. Instead, we utilize the extracted Chinese lexicons to transform the annotated training data of a Chinese segmenter into a short unit standard, and perform segmentation using the retrained Chinese segmenter.

Wang et al. [142] also proposed a short unit transformation method. The proposed method is based on transfer rules and a transfer database. The transfer rules are extracted from alignment results of annotated Chinese and segmented Japanese training data, while the transfer database is constructed using external lexicons and is manually modified. Our proposed method learns transfer knowledge based on common Chinese characters. Moreover, no external lexicons or manual work is required.

2.2.2 Chinese Entry Extraction

Chinese entries are extracted from a parallel training corpus through the following steps.

- Step 1: Segment Japanese sentences in the parallel training corpus.
- Step 2: Convert Japanese tokens consisting only of Kanji [11] into Simplified Chinese using the Chinese character mapping table created in Section 2.1.

[11] Japanese has several other kinds of character types apart from Kanji.

- Step 3: Extract the converted tokens as Chinese entries if they exist in the corresponding Chinese sentence.

For example, "小坂 (Kosaka)," "先生 (Mr.)," "日本 (Japan)," "临床 (clinical)," "麻醉 (anesthesia)," "学会 (society)," "创始 (found)," and "者 (person)" in Figure 2.2 would be extracted. Note that although "临床 ↔ 臨床 (clinical)," "麻醉 ↔ 麻酔 (anesthesia)," and "创始 ↔ 創始 (found)" are not identical, because "临↔ 臨 (arrive)," "醉 ↔ 酔 (drunk)," and "创↔ 創 (create)" are common Chinese characters, "臨床 (clinical)" is converted into "临床 (clinical)," "麻酔 (anesthesia)" is converted into "麻醉 (anesthesia)," and "創始 (found)" is converted into "创始 (found)" in Step 2.

2.2.3 Chinese Entry Incorporation

Several studies have shown that using a system dictionary is helpful for Chinese word segmentation [84, 141]. Therefore, we used a corpus-based Chinese word segmentation and POS tagging tool with a system dictionary and incorporated the extracted entries into the system dictionary. The extracted entries are not only effective for the unknown word problem, but also useful in solving the word segmentation granularity problem.

However, setting POS tags for the extracted entries is problematic. To solve this problem, we created a POS tag mapping table between Chinese and Japanese by hand. For Chinese, we used the POS tagset used in CTB, which is also used in our Chinese segmenter. For Japanese, we used the POS tagset defined in the morphological analyzer JUMAN [77]. JUMAN uses a POS tagset containing sub POS tags. For example, the POS tag "名詞 (noun)" contains sub POS tags such as "普通名詞 (common noun)," "固有名詞 (proper noun)," "時相名詞 (temporal noun)," and so on. Table 2.13 shows a part of the Chinese-Japanese POS tag mapping table we created, with the sub POS tags of JUMAN given within square brackets.

POS tags for the extracted Chinese entries are assigned by converting the POS tags of Japanese tokens assigned by JUMAN into POS tags of CTB. Note that not all POS tags of JUMAN can be converted into POS tags of CTB, and vice versa. Those that cannot be converted are not incorporated into the system dictionary.

JUMAN	CTB
副詞 (adverb)	AD
接続詞 (conjunction)	CC
名詞 (noun) [数詞 (numeral noun)]	CD
未定義語 (undefined word) [アルファベット (alphabet)]	FW
感動詞 (interjection)	IJ
接尾辞 (suffix) [名詞性名詞助数辞 (measure word suffix)]	M
名詞 (noun) [普通名詞 (common noun) / サ変名詞 (sahen noun) / 形式名詞 (formal noun) / 副詞的名詞 (adverbial noun)] / 接尾辞 (suffix) [名詞性名詞接尾辞 (noun suffix) / 名詞性特殊接尾辞 (special noun suffix)]	NN
名詞 (noun) [固有名詞 (proper noun) / 地名 (place name) / 人名 (person name) / 組織名 (organization name)]	NR
名詞 (noun) [時相名詞 (temporal noun)]	NT
特殊 (special word)	PU
形容詞 (adjective)	VA
動詞 (verb) / 名詞 (noun) [サ変名詞 (sahen noun)]	VV

Table 2.13: Chinese-Japanese POS tag mapping table.

2.2.4 Short Unit Transformation

Bai et al. [10] showed that adjusting Chinese word segmentation to create a token 1-to-1 mapping as far as possible between parallel sentences can improve alignment accuracy, which is crucial for corpus-based MT. Wang et al. [142] proposed a short unit standard for Chinese word segmentation that is more similar to the Japanese word segmentation standard, and which can reduce the number of 1-to-n alignments and improve MT performance.

We previously proposed a method for transforming the annotated training data of the Chinese segmenter into the Japanese word segmentation standard using the extracted Chinese entries, and then used the transformed data to train the Chinese segmenter [25]. Because the extracted entries are derived from Japanese

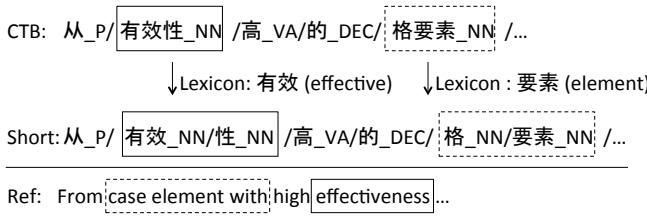

Figure 2.3: Example of previous short unit transformation.

word segmentation results, they follow the Japanese word segmentation standard. Therefore, we utilize these entries in short unit transformation. We use the Chinese entries extracted in Section 2.2.2 and modify every token in the training data for the Chinese segmenter. If the token is longer than a extracted entry, it is simply split. Figure 2.3 gives an example of this process, where "有效 (effective)" and "要素 (element)" are both extracted entries. Because "有效性 (effectiveness)" is longer than "有效 (effective)," it is split into "有效 (effective)" and "性 (a noun suffix)," and because "格要素 (case element)" is longer than "要素 (element)," it is split into "格 (case)" and "要素 (element)." For POS tags, the originally annotated one is retained for the split tokens.

Although this method works well in most cases, it suffers from the problem of transformation ambiguity. For example, for a long token like "留学生 (student studying abroad)" in the annotated training data, entries "留学 (study abroad)" and "学生 (student)" are extracted from the parallel training corpus. In this case, our previous method randomly chose one entry for transformation. Therefore, "留学生 (student studying abroad)" could be split into "留 (stay)"and "学生 (student)," which is incorrect. To solve this problem, we improved the transformation method by utilizing both short unit information and extracted entries. Short unit information is short unit transformation information extracted from the parallel training corpus. Short unit information extraction is similar to the Chinese entry extraction described in Section 2.2.2, and includes the following steps.

- Step 1: Segment both Chinese and Japanese sentences in the parallel training corpus.
- Step 2: Convert Japanese tokens consisting of only Kanji into Simplified

Chinese using the Chinese character mapping table we created in Section 2.1.

- Step 3: Extract the converted tokens composed of consecutive tokens in the segmented Chinese sentence and the corresponding Chinese tokens.

For example, "创始者 (founder)→创始 (found) / 者 (person)" in Figure 2.2 is extracted as short unit information.

In the improved transformation method, we modify the tokens in the training data using the following processes in order.

1. If the token itself exists in the extracted entries, keep it.
2. If the token can be transferred using short unit information, transfer it according to the short unit information.
3. If the token can be split using extracted entries, transfer it according to the extracted entries.
4. Otherwise, keep it.

Following [25], we do not use extracted entries that are composed of only one Chinese character, because these entries may lead to undesirable transformation results. Taking the Chinese character "歌 (song)" as an example, "歌 (song)" can be used as a single word, but we can also use "歌 (song)" to construct other words by combining it with other Chinese characters, such as "歌颂(praise)," "诗歌 (poem)," and so on. Obviously, splitting "歌颂(praise)" into "歌 (song)" and "颂(eulogy)," or splitting "诗歌 (poem)" into "诗(poem)" and "歌 (song)" is undesirable. We do not use extracted number entries either, as these can also lead to undesirable transformation. For example, using "十八 (18)" to split "二百九十八 (298)" into "二百九 (290)" and "十八 (18)" is obviously incorrect. Moreover, there are a few consecutive tokens in the training data that can be combined into a single extracted entry; however, we do not consider these patterns.

Figure 2.4 gives an example of our improved transformation method. In this example, because "地中海 (Mediterranean)" also exists in the extracted entries, it is not changed, even though there is an extracted entry "地中 (in earth)." The

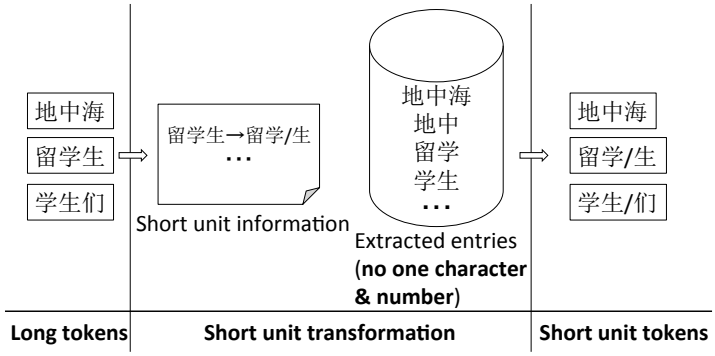

Figure 2.4: Example of improved short unit transformation.

long token "留学生 (student studying abroad)" can be transferred using short unit information, so it is transferred into "留学 (study abroad)" and "生 (student)." Meanwhile, the long token "学生们(students)" can be split into "学生 (student)" and "们(plural for student)" using the extracted entry "学生 (student)."

We record the extracted entries and short unit information used for transformation with the corresponding Japanese tokens, and store them as transforming word dictionary. This dictionary contains 16k entries, and will be helpful for word alignment.

2.2.5 Experiments

We conducted Chinese-Japanese translation experiments to show the effectiveness of exploiting common Chinese characters in Chinese word segmentation optimization.

The parallel training corpus we used was the same as that used in Section 2.1.7. We further used CTB 7 (LDC2010T07)[12] to train the Chinese segmenter. Training data, containing 31,131 sentences, was created from CTB 7 using the same method described in [141]. The segmenter used for Chinese was an in-house corpus-based word segmentation and POS tagging tool with a system dictionary. Weights for the entries in the system dictionary were automatically learned from the training data using an averaged structured perceptron [31]. For Japanese, we used JU-

[12]http://www.ldc.upenn.edu/

	Set 1	Set 2	Set 3	Set 4	Set 5	Total
# sentences	255	336	391	395	393	1770
# words	6.6K	8.7K	10.0K	11.7K	16.6K	53.6K
# Chinese characters	8.6K	10.7K	12.9K	15.8K	22.1K	70.1K
average sentence length	44.9	47.0	45.4	52.2	74.1	53.58

Table 2.14: Statistics of test sets containing Chinese sentences ("Total" denotes the combined statistics for the five test sets).

	Set 1	Set 2	Set 3	Set 4	Set 5	Total
# sentences	255	336	391	395	393	1770
# words	8.0K	11.0K	12.7K	14.4K	20.1K	66.2K
# Chinese characters	5.1K	6.3K	7.7K	9.0K	13.0K	41.1K
average sentence length	55.6	57.6	56.6	66.2	90.4	66.3

Table 2.15: Statistics of test sets containing Japanese sentences.

MAN [77]. For decoding, we used the state-of-the-art phrase-based SMT toolkit Moses [72] with the default options, except for the distortion limit ($6 \rightarrow 20$). Tuning was performed by minimum error rate training [98] using a further 500 development sentence pairs and it was re-run for every experiment. We trained word-based 5 gram language model on the target side of the training data using SRILM toolkit [122][13] with interpolated Kneser-Ney discounting. We translated 5 test sets from the same domain as the parallel training corpus. The statistics of the test sets of Chinese and Japanese sentences are given in Tables 2.14 and 2.15, respectively. Note that none of the sentences in the test sets are included in the parallel training corpus.

We carried out Chinese-Japanese translation experiments, comparing the following three experimental settings:

- Baseline: Using only entries extracted from the Chinese annotated corpus as the system dictionary for the Chinese segmenter.

[13] http://www.speech.sri.com/projects/srilm

BLEU	Set 1	Set 2	Set 3	Set 4	Set 5	Total
Baseline	51.03	48.98	40.52	29.20	26.08	36.64
Chu+ 2012	**52.83**	51.13	41.57	31.01	**28.82**	38.59*
Optimized	52.55	51.88	41.62	30.69	28.43	38.52*
+Dictionary	52.73	**52.21**	**42.02**	**31.19**	28.72	**38.86***

Table 2.16: Results of Chinese-to-Japanese translation experiments ("*" denotes the "Total" result is better than "Baseline" significantly at $p < 0.01$).

BLEU	Set 1	Set 2	Set 3	Set 4	Set 5	Total
Baseline	42.26	42.47	35.60	26.70	27.92	33.31
Chu+ 2012	42.89	43.27	34.95	27.80	28.82	33.90*
Optimized	43.06	44.04	35.53	28.00	**29.04**	34.30*†
+Dictionary	**43.17**	**44.78**	**36.34**	**28.10**	28.89	**34.53***‡

Table 2.17: Results of Japanese-to-Chinese translation experiments ("*" denotes the "Total" result is better than "Baseline" significantly at $p < 0.01$, "†" and "‡" denotes the "Total" result is better than "Chu+ 2012" significantly at $p < 0.05$ and $p < 0.01$ respectively).

- Optimized: Incorporating the Chinese entries extracted in Section 2.2.2 into the system dictionary and training the Chinese segmenter on the short unit training data transformed in Section 2.2.4.
- +Dictionary: Appending the transforming word dictionary stored in Section 2.2.4 to the parallel training corpus.

The translations were evaluated using BLEU-4 [102] calculated on words. For Japanese-to-Chinese translation, we re-segmented the translations using the optimized Chinese segmenter. Tables 2.16 and 2.17 give the BLEU scores for Chinese-to-Japanese and Japanese-to-Chinese translation, respectively. For comparison, we also list the optimized results of [25], which are denoted as "Chu+ 2012." The results show that our proposed approach can improve MT performance. We notice that compared with [25], the improvement in the current short unit transfor-

Input: 本论文中，提议考虑现存实现方式的功能适应性决定对策目标的保密基本设计法。

Baseline (BLEU=49.38)
Segmented: 本/论文/中/，/提议/考虑/现存/实现/方式/的/ 功能 / **适应性** /决定/对策/目标/的/保密/基本/设计法/。
Output: 本/論文/では/，/提案/する/ 適応/的 /対策/を/決定/する/セキュリティ/基本/設計/法/を/考える/現存/の/実現/方式/の/ 機能 /を/目標/と/して/いる/.

Segmentation Optimization (BLEU=56.33)
Segmented: 本/论文/中/，/提议/考虑/现存/实现/方式/的/ 功能 / **适应/性** /决定/对策/目标/的/保密/基本/设计/法/。
Output: 本/論文/では/，/提案/する/考え/現存/の/実現/方式/の/ 機能/的 / 適応/性 /を/決定/する/対策/目標/の/セキュリティ/基本/設計/法/を/提案/する/.

Reference
本/論文/では/，/対策/目標/を/現存/の/実現/方式/の/ 機能/的 / 適合/性 /も/考慮/して/決定/する/セキュリティ/基本/設計/法/を/提案/する/.
(In this paper, we propose a basic security design method also consider functional suitability of the existing implementation method for determining countermeasures target.)

Figure 2.5: Example of translation improvement.

mation method further improved the Japanese-to-Chinese translation. However, it had no effect on the Chinese-to-Japanese translation. Appending the transforming word dictionary further improved the translation performance. Similar to [25], the improvement in Japanese-to-Chinese translation compared with that in Chinese-to-Japanese translation is not that significant. We believe the reason for this is the input sentence. For Chinese-to-Japanese translation, the segmentation of input Chinese sentences is optimized, whereas for Japanese-to-Chinese translation, our proposed approach does not change the segmentation results of the input Japanese sentences.

2.2.6 Discussion

Short Unit Effectiveness

Experimental results indicate that our proposed approach can improve MT performance significantly. We present an example to show the effectiveness of optimized short unit segmentation results. Figure 2.5 gives an example of Chinese-to-Japanese translation improvement using optimized short unit segmentation results

compared with the baseline. The difference between the short unit and baseline is whether "适应性 (suitability)" is split in Chinese, whereas the Japanese segmenter always splits it. By splitting it, the short unit improves word alignment and phrase extraction, which eventually affects the decoding process. In decoding, the short unit treats "功能适应性 (functional suitability)" as one phrase, while the baseline separates it, leading to a undesirable translation result.

Short Unit Transformation Problems

Although we have improved the short unit transformation method, there are still some transformation problems. One problem is incorrect transformation. For example, there is a long token "不好意思 (sorry)" and an extracted entry "好意 (favor)," and therefore, the long token is transferred into "不 (not)," "好意 (favor)," and "思 (think)," which is obviously undesirable. Our current method cannot deal with such cases, making this one of the future works in this study.

Another problem is POS tag assignment for the transformed short unit tokens. Our proposed method simply keeps the original annotated POS tag of the long token for the transformed short unit tokens, which works well in most cases. However, there are also some exceptions. For example, there is a long token "被实验者 (test subject)" in the annotated training data, and an entry "实验(test)" extracted from the parallel training corpus, so the long token is split into "被 (be)," "实验(test)," and "者 (person)." As the POS tag for the original long token is NN, the POS tags for the transformed short unit tokens are all set to NN, which is undesirable for "被 (be)." The correct POS tag for "被 (be)" should be LB. An external dictionary would be helpful in solving this problem. Furthermore, the transformed short unit tokens may have more than one possible POS tag. All these problems will be dealt with in future work.

2.3 Summary of This Chapter

Common Chinese characters can be very helpful in Chinese-Japanese MT. In this article, we proposed a method for creating a Chinese character mapping table automatically for Japanese, Traditional Chinese, and Simplified Chinese using freely

available resources, and constructed a more complete resource of common Chinese characters than the existing ones. We exploited common Chinese characters in Chinese word segmentation optimization. Experimental results show that our proposed approaches can improve MT performance significantly, thus verifying the effectiveness of using common Chinese characters in Chinese-Japanese MT.

In the remainder of this thesis, we further exploit common Chinese characters in two aspects. The mapping table is exploited in parallel sentence (Chapter 4) and fragment extraction (Chapter 5). As shown in our previous work, the optimized segmenter can also improve Chinese-English MT [28]. Therefore, we use it for Chinese segmentation in all the tasks in this thesis.

Chapter 3

Bilingual Lexicon Extraction

Bilingual lexicons are important for many bilingual natural language processing (NLP) tasks, such as statistical machine translation (SMT) [17, 100, 71] and dictionary based cross-language information retrieval (CLIR) [105]. Because manual construction of bilingual lexicons is expensive and time-consuming, automatic construction is desirable. Mining bilingual lexicons from parallel corpora is a possible method. However, it is only feasible for a few language pairs and domains, because parallel corpora remain a scarce resource. As comparable corpora are far more widely available than parallel corpora, extracting bilingual lexicons from comparable corpora is an attractive research field.

In the literature, two main categories of methods have been proposed for bilingual lexicon extraction (BLE) from comparable corpora, namely topic model based method (TMBM) [135] and context based method (CBM) [111]. Both methods are based on the distributional hypothesis [54], stating that words with similar meaning have similar distributions across languages. TMBM measures the similarity of two words on cross-lingual topical distributions, while CBM measures the similarity on contextual distributions across languages.

In this chapter, we present a BLE system that is based on a novel combination of TMBM and CBM. The motivation is that a combination of these two methods can exploit both topical and contextual knowledge to measure the distributional similarity of two words, making bilingual lexicon extraction more reliable and accurate than only using one knowledge source. The key points for the combination

are as follows:

- TMBM can extract bilingual lexicons from comparable corpora without any prior knowledge. The extracted lexicons are semantically related and provide comprehensible and useful contextual information in the target language for the source word [135]. Therefore, it is effective to use the lexicons extracted by TMBM as a seed dictionary, which is required for CBM.
- The lexicons extracted by CBM can be combined with the lexicons extracted by TMBM to further improve the accuracy.
- The combined lexicons again can be used as the seed dictionary for CBM. Therefore the accuracy of the lexicons can be iteratively improved.

Our system not only maintains the advantage of TMBM that does not require any prior knowledge, but also can iteratively improve the accuracy of BLE through combination CBM. To the best of our knowledge, this is the first study that iteratively exploits both topical and contextual knowledge for bilingual lexicon extraction. Experimental results on Chinese-English, Japanese-English and Japanese-Chinese Wikipedia data show that our proposed method performs significantly better than the method only using topical knowledge [135].

3.1 Related Work

3.1.1 Topic Model Based Methods

TMBM uses the distributional hypothesis on topics, stating that two words are potential translation candidates if they are often present in the same cross-lingual topics and not observed in other cross-lingual topics [135]. It trains a Bilingual Latent Dirichlet Allocation (BiLDA) topic model on document-aligned comparable corpora, and identifies word translations relying on word-topic distributions from the trained topic model. This method is attractive because it does not require any prior knowledge.

Vulić et al. [135] first proposed this method. Later, Vulić and Moens [136] extended this method to detect highly confident word translations by a symmetrization process and the one-to-one constraints, and demonstrated a way to

build a high quality seed dictionary using both BiLDA and cognates. Liu et al. [83] developed this method by converting document-aligned comparable corpora into a parallel topic-aligned corpus using BiLDA topic models, and identify word translations with the help of word alignment. Richardson et al. [115] exploited this method in the task of transliteration. Vulić and Moens [138] improved this method by using BiLDA to learn the semantic word responses of words, and identify word translations using the semantic word response vectors.

Our study differs from previous studies in using a novel combination of TMBM and CBM. Vulić and Moens [139] also proposed a combination method that obtains an initial seed dictionary with a variant of TMBM, and iteratively increases the size of the seed dictionary using only CBM. Our study differs from [139] in producing an initial seed dictionary for all the source words in the vocabulary with TMBM, and iteratively improving the quality using a combination of TMBM and CBM. We show that the combination outperforms both TMBM and CBM. In addition, Vulić and Moens [139] compared the effect of the size of the initial seed dictionary and showed that using all bilingual lexicons obtained by the TMBM showed the best or comparable to the best performing method, which is similar to our method that iterates using a seed dictionary for all the source words.

3.1.2 Context Based Methods

From the pioneering work of [110, 40], various studies have been conducted on CBM for extracting bilingual lexicons from comparable corpora. CBM is based on the distributional hypothesis on context, stating that words with similar meaning appear in similar contexts across languages. It usually consists of three steps: context vector modeling, vector similarity calculation and translation identification that treats a candidate with higher similarity score as a more confident translation. Gaussier et al. [49] presented a geometric view of this process. Previous studies use different definitions of context, such as window-based context [40, 111, 73, 52, 107, 123], sentence-based context [43] and syntax-based context [48, 146, 108]. To quantify the strength of the association between a word and its context word, different association measures have been used, such as log-likelihood-ratio (LLR) [111], term frequency - inverse document frequency (TF-

IDF) [43] and pointwise mutual information (PMI) [7]. Previous studies also use different measures to compute the similarity between the vectors, such as cosine similarity [43, 48, 107, 123], Euclidean distance [40, 146], city-block metric [111] and Spearman rank order [73]. Laroche and Langlais [78] conducted a systematic study of using different association and similarity measures for CBM.

To further improve the performance of CBM, various efforts have been made. These efforts include enhancing the corpus comparability of comparable corpora [79, 80], re-ranking the translation candidates acquired by CBM [53], and using large-scale background knowledge from Wikipedia [15]. Also, CBM suffers from the data sparseness problem especially for the low frequency words, smoothing [103, 9, 55] and prediction [56] technologies have been proposed for this problem.

Basically, CBM requires a seed dictionary to project the source vector onto the vector space of the target language, which is one of the main concerns of this study. In previous studies, a seed dictionary is usually manually created [111, 48], and sometimes complemented by bilingual lexicons extracted from a parallel corpus [43, 123] or the Web [107]. In addition, some studies try to create a seed dictionary using cognates [73, 52], however this cannot be applied to distant language pairs that do not share cognates, such as Chinese-English and Japanese-English. In the case that a word in the seed dictionary has several polysemous translations, word sense disambiguation is necessary [16]. There are also some studies that do not require a seed dictionary [110, 40, 146]. However, these studies show lower accuracy compared to the conventional methods using a seed dictionary.

Our study differs from previous studies in using a seed dictionary automatically acquired without any prior knowledge, which is learned from comparable corpora in an unsupervised way.

3.1.3 Other Methods

Besides TMBM and CBM, other methods also have been proposed for BLE recently. One method is decipherment [112, 36]. They treat the source text as a cipher for the target text, and treat BLE as a decipherment task for a word substitution cipher. Decipherment is solved using Bayesian technologies. This method does not require a seed dictionary. Another method is using deep learning for

BLE [90]. They first learn monolingual word representations using a neural network architecture [89]. Then they learn a linear projection between the source and target word representations using a small bilingual dictionary. Finally, they identify the translations of source words by computing the similarity between the projected source and target word representations.

3.2 Proposed Method

The overview of our proposed BLE system is presented in Figure 3.1. We first apply TMBM to obtain bilingual lexicons from comparable corpora, which we call topical bilingual lexicons. The topical bilingual lexicons contain a list of translation candidates for a source word w_i^S, where a target word w_j^T in the list has a topical similarity score $Sim_{Topic}(w_i^S, w_j^T)$. Then using the topical bilingual lexicons as an initial seed dictionary, we apply CBM to obtain bilingual lexicons, which we call contextual bilingual lexicons. The contextual bilingual lexicons also contain a list of translation candidates for a source word, where each candidate has a contextual similarity score $Sim_{Context}(w_i^S, w_j^T)$. We then combine the topical bilingual lexicons with the contextual bilingual lexicons to obtain combined bilingual lexicons. The combination is done by calculating a combined similarity score $Sim_{Comb}(w_i^S, w_j^T)$ using the $Sim_{Topic}(w_i^S, w_j^T)$ and $Sim_{Context}(w_i^S, w_j^T)$ scores. After combination, the quality of the lexicons can be higher, namely the correct translation in the candidate list is assigned a high score and ranked higher. Therefore, we iteratively use the combined bilingual lexicons as the seed dictionary for CBM and conduct combination, to improve the contextual bilingual lexicons and further improve the combined bilingual lexicons.

Our system not only maintains the advantage of TMBM that does not require any prior knowledge, but also can iteratively improve the accuracy by a novel combination with CBM. Details of TMBM, CBM and combination method will be described in Section 3.2.1, 3.2.2 and 3.2.3 respectively.

3.2.1 Topic Model Based Method

In this section, we describe TMBM to calculate the topical similarity score Sim_{Topic} (w_i^S, w_j^T). We first train a BiLDA topic model presented in [91], which is an ex-

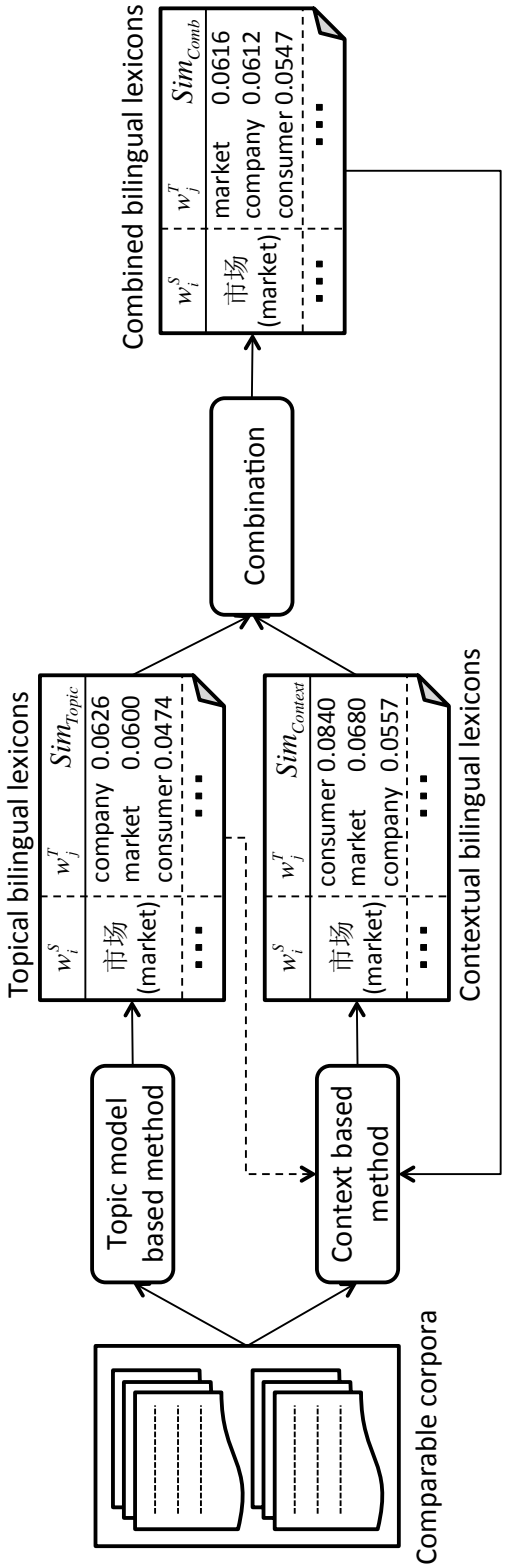

Figure 3.1: Bilingual lexicon extraction system.

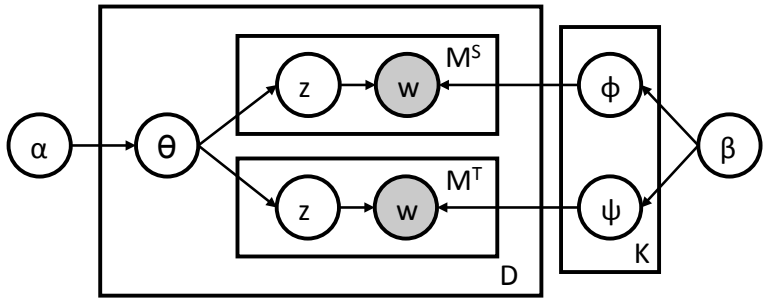

Figure 3.2: BiLDA topic model.

tension of the standard LDA model [14]. Figure 3.2 shows the plate model for BiLDA, with D document pairs, K topics and hyper-parameters α, β. Topics for each document are sampled from a single variable θ, which contains the topic distribution and is language-independent. Words of the two languages are sampled from θ in conjugation with the word-topic distributions ϕ (for source language S) and ψ (for target language T).

Once the BiLDA topic model is trained and the associated word-topic distributions are obtained for both source and target corpora, we can calculate the similarity of word-topic distributions to identify word translations. For similarity calculation, we use the *TI+Cue* measure presented in [135], which shows the best performance for identifying word translations in their study. *TI+Cue* measure is a linear combination of the *TI* and *Cue* measures, defined as follows:

$$Sim_{TI+Cue}(w_i^S, w_j^T) = \lambda Sim_{TI}(w_i^S, w_j^T) + (1-\lambda) Sim_{Cue}(w_i^S, w_j^T) \qquad (3.1)$$

TI and *Cue* measures interpret and exploit the word-topic distributions in different ways, thus combining the two leads to better results.

The *TI* measure is the similarity calculated from source and target word vectors constructed over a shared space of cross-lingual topics. Each dimension of the vectors is a term frequency - inverse topic frequency score (*TF-ITF*). *TF-ITF* score is computed in a word-topic space, which is similar to *TF-IDF* score that is computed in a word-document space. *TF* measures the importance of a word w_i within a particular topic z_k, while *ITF* of a word w_i measures the importance of w_i across all topics. Let $n_k^{(w_i)}$ be the number of times the word w_i is associated

with the topic z_k, W denotes the vocabulary and K denotes the number of topics, then

$$TF_{i,k} = \frac{n_k^{(w_i)}}{\sum_{w_j \in W} n_k^{(w_j)}} \tag{3.2}$$

$$ITF_i = \log \frac{K}{1 + |\{k : n_k^{(w_i)} > 0\}|} \tag{3.3}$$

TF-ITF score is the product of $TF_{i,k}$ and ITF_i. Then, the *TI* measure is obtained by calculating the cosine similarity of the K dimensional source and target vectors. Let S^i be the source vector for a source word w_i^S, T^j be the target vector for a target word w_j^T, then cosine similarity is defined as follows:

$$Cos(w_i^S, w_j^T) = \frac{\sum_{k=1}^{K} S_k^i \times T_k^j}{\sqrt{\sum_{k=1}^{K} (S_k^i)^2} \times \sqrt{\sum_{k=1}^{K} (T_k^j)^2}} \tag{3.4}$$

The *Cue* measure is the probability $P(w_j^T | w_i^S)$, where w_j^T and w_i^S are linked via the shared topic space, defined as:

$$P(w_j^T | w_i^S) = \sum_{k=1}^{K} \psi_{k,j} \frac{\phi_{k,i}}{Norm_\phi} \tag{3.5}$$

where

$$\phi_{k,i} = \frac{n_k^{(w_i)} + \beta}{\sum_{w_j \in W} n_k^{(w_j)} + W\beta} \tag{3.6}$$

$\psi_{k,j}$ is defined in the similar way, and $Norm_\phi$ denotes the normalization factor given by $Norm_\phi = \sum_{k=1}^{K} \phi_{k,i}$ for a word w_i.

3.2.2 Context Based Method

In this section, we describe CBM to calculate the contextual similarity score $Sim_{Context}(w_i^S, w_j^T)$. We use window-based context, and leave the comparison of using different definitions of context as future work. Given a word, we count all its immediate context words, with a window size of 4 (2 preceding words and 2 following words). We build a context by collecting the counts in a bag of words fashion, namely we do not distinguish the positions that the context words appear in. The number of dimensions of the constructed vector is equal to the vocabulary

size. We further reweight each component in the vector by multiplying by the *IDF* score following [48], which is defined as follows:

$$IDF(t, D) = \log \frac{|D|}{1 + |\{d \in D : t \in d\}|} \tag{3.7}$$

where $|D|$ is the total number of documents in the corpus, and $|\{d \in D : t \in d\}|$ denotes number of documents where the term t appears. We model the source and target vectors using the method described above, and project the source vector onto the vector space of the target language using a seed dictionary. The similarity of the vectors is computed using cosine similarity (Equation 3.4).

As initial, we use the topical bilingual lexicons extracted in Section 3.2.1 as seed dictionary. Note that the topical bilingual lexicons are noisy especially for the rare words [136]. However, because they provide comprehensible and useful contextual information in the target language for the source word [135], it is effective to use the lexicons as a seed dictionary for CBM.

Once contextual bilingual lexicons are extracted, we combine them with the topical bilingual lexicons. After combination, the quality of the lexicons will be improved. Therefore, we further use the combined lexicons as seed dictionary for CBM, which will produce better contextual bilingual lexicons. Again, we combine the better contextual bilingual lexicons to the topical bilingual lexicons. By repeating these steps, both the contextual bilingual lexicons and the combined bilingual lexicons will be iteratively improved.

Applying CBM and combination one time is defined as one iteration. At iteration 1, the topical bilingual lexicons are used as seed dictionary for CBM. From the second iteration, the combined lexicons are used as seed dictionary. In all iterations, we produce a seed dictionary for all the source words in the vocabulary, and use the top 1 candidate to project the source context vector to the target language. We stop the iteration when the predefined number of iterations have been done.

3.2.3 Combination

TMBM measures the distributional similarity of two words on cross-lingual topics, while CBM measures the distributional similarity on contexts across languages. A

combination of these two methods can exploit both topical and contextual knowledge to measure the distributional similarity, making bilingual lexicon extraction more reliable and accurate. Here we use a linear combination for the two methods to calculate a combined similarity score, defined as follows:

$$Sim_{Comb}(w_i^S, w_j^T) = \gamma Sim_{Topic}(w_i^S, w_j^T) + (1-\gamma) Sim_{Context}(w_i^S, w_j^T) \quad (3.8)$$

To reduce computational complexity, we only keep the Top-N translation candidates for a source word during all the steps in our system. We first produce a Top-N candidate list for a source word using TMBM. Then we apply CBM to calculate the similarity only for the candidates in the list. Finally, we conduct combination. Therefore, the combination process is a kind of re-ranking of the candidates produced by TMBM. Note that both $Sim_{Topic}(w_i^S, w_j^T)$ and $Sim_{Context}(w_i^S, w_j^T)$ are normalized before combination, where the normalization is given by:

$$Sim_{Norm}(w_i^S, w_j^T) = \frac{Sim(w_i^S, w_j^T)}{\sum_{n=1}^{N} Sim(w_i^S, w_n^T)} \quad (3.9)$$

where N is the number of translation candidates for a source word.

3.3 Experiments

We evaluated our proposed method on Chinese-English, Japanese-English and Japanese-Chinese Wikipedia data. For people who want to reproduce the results reported in this chapter, we released a software that contains all the required codes and data at `http://lotus.kuee.kyoto-u.ac.jp/~chu/code/iBiLexExtractor.tgz`.

Note that Wikipedia is a special type of comparable corpora, because article alignment is manually established. In Wikipedia, articles describing the same topic in different languages are manually linked by the authors. These links are usually called interlanguage links. Figure 3.3 shows an example of interlanguage links in Wikipedia. For many other types of comparable corpora, it is necessary to perform article alignment as an initial step. Many methods have been proposed for article alignment in the literature, such as IR-based [131, 93], feature-based [134]

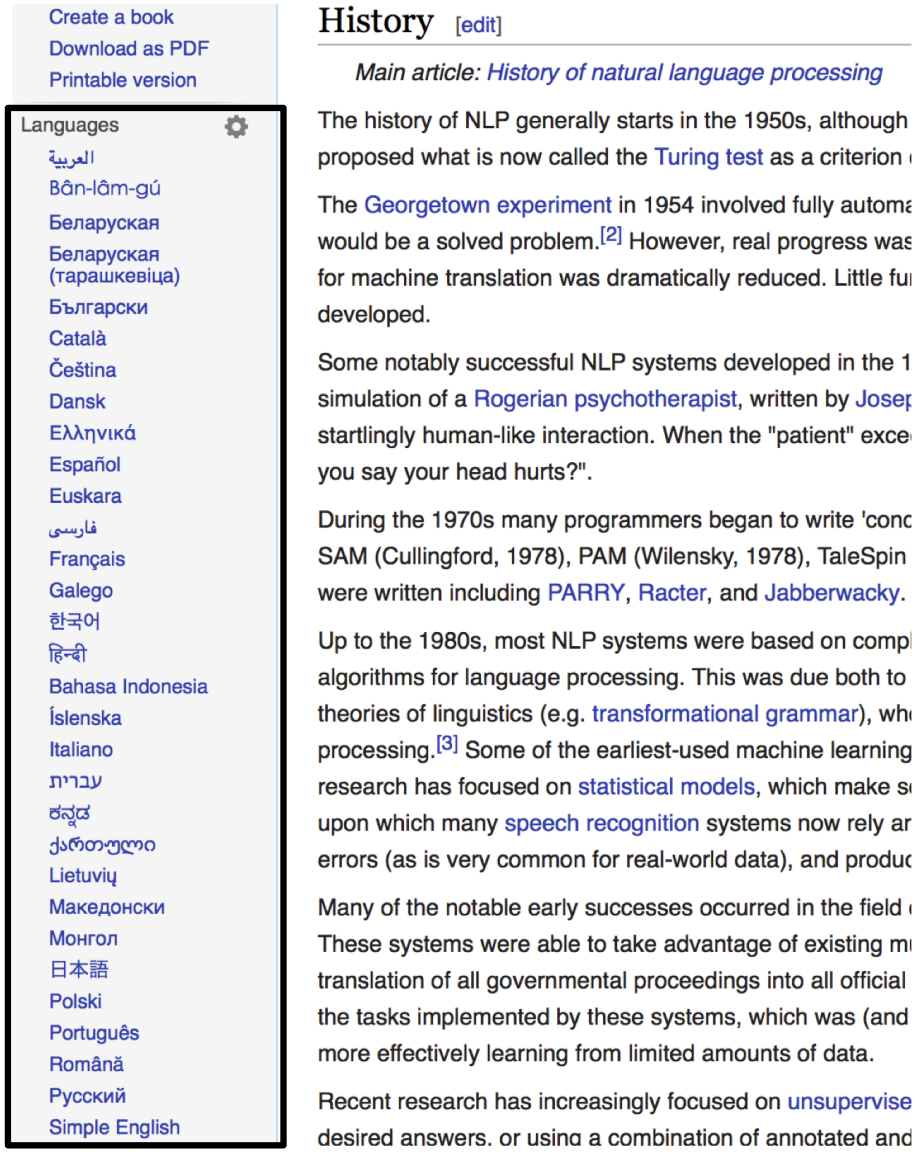

Figure 3.3: Interlanguage links (in rectangles) in Wikipedia.

and topic-based [153] methods. After article alignment, our proposed method can be applied to any type of comparable corpora.

3.3.1 Data

We created the experimental data according to the following steps. We downloaded Chinese[1] (2012/09/21), Japanese[2] (2012/09/16) and English[3] (2012/10/01) Wikipedia database dumps. We used an open-source Python script[4] 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.[5] We aligned the articles on the same topic in Chinese-English, Japanese-English and Japanese-Chinese Wikipedia via the interlanguage links. From the aligned articles, we selected $10k$ Chinese-English, Japanese-English and Japanese-Chinese pairs as our training corpora. For Japanese-Chinese, we also conducted experiments using all the aligned articles, containing $162k$ pairs. There are two reasons for further using all the aligned articles for Japanese-Chinese. Firstly, it is helpful to investigate the effect of the size of the training data for our proposed method. Secondly, we use the extracted bilingual lexicons to assist parallel sentence extraction that is conducted on all the aligned articles, which will be described in Chapter 4.

We preprocessed the Chinese and Japanese corpora using a tool proposed by Chu et al. [25] and JUMAN [77] respectively for segmentation and Part-of-Speech (POS) tagging. The English corpora were POS tagged using Lookahead POS Tagger [129]. To reduce data sparsity, we kept only lemmatized noun forms. The vocabularies of the Chinese-English data contain 112,682 Chinese and 179,058 English nouns. The vocabularies of the Japanese-English data contain 47,911 Japanese and 188,480 English nouns. The vocabularies of the Japanese-Chinese data contain 51,823 Japanese and 114,256 Chinese nouns for the $10k$ article pairs,

[1] http://dumps.wikimedia.org/zhwiki
[2] http://dumps.wikimedia.org/jawiki
[3] http://dumps.wikimedia.org/enwiki
[4] http://code.google.com/p/recommend-2011/source/browse/Ass4/WikiExtractor.py
[5] http://svn.wikimedia.org/svnroot/mediawiki/branches/REL1_12/phase3/includes/ZhConversion.php

104,461 Japanese and 772,433 Chinese nouns for all the article pairs. The vocabulary size of Japanese is smaller than that of Chinese and English, because we kept only common, sahen and proper nouns; place, person and organization names among all sub POS tags of noun in JUMAN.

3.3.2 Experimental Settings

For BiLDA topic model training, we used the implementation PolyLDA++ by Richardson et al. [115].[6] We set the hyper-parameters $\alpha = 50/K, \beta = 0.01$ following Vulić et al. [135], where K denotes the number of topics. We trained the BiLDA topic model using Gibbs sampling with $1,000$ iterations. For the combined *TI+Cue* method, we used the toolkit BLETM obtained from Vulić et al. [135],[7] where we set the linear interpolation parameter $\lambda = 0.1$ following their study. For our proposed method, we empirically set the linear interpolation parameter $\gamma = 0.8$,[8] and conducted 20 iterations.[9]

3.3.3 Evaluation Criterion

We manually created Chinese-English, Japanese-English and Japanese-Chinese test sets for the most 1,000 frequent source nouns[10] in the experimental data with the help of Google Translate.[11] For each source noun, if the correct translations are given by Google Translate we used them, otherwise we manually translated it. Note that some source nouns could have multiple translations, and we tried

[6] https://bitbucket.org/trickytoforget/polylda

[7] http://people.cs.kuleuven.be/~ivan.vulic/software/BLETMv1.0wExamples.zip

[8] Because we did not have a held-out data set, we determined γ based on the Chinese-English test set, we compared the effects of different γ from 0.1 to 0.9 in intervals of 0.1, and 0.8 showed the best performance. We applied the same parameter for the Japanese-English and Japanese-Chinese tasks. For sure, it is better to determine all the parameters using held-out data, however, we leave it as future work.

[9] This iteration number was also empirically determined on the Chinese-English test set. Based on the experimental results (see Figure 3.4), the accuracy of our proposed method greatly improves in the first few iterations, and after that the performance becomes stable. We believe that the accuracy would not be improved in further iterations, therefore we stopped at iteration 20.

[10] For Japanese-Chinese, the test sets were created for the most frequent 1,000 Japanese nouns that are limited to the sub POS tags listed in Section 3.3.1 in all the article pairs.

[11] http://translate.google.com

to give all the translations based on the best of our knowledge. However, the test sets could be still incomplete, namely some translations of the source words might be not registered. Following [135], we evaluated the accuracy using the following two metrics:

- Precision@1: Percentage of words where the top 1 word from the list of translation candidates is the correct one.
- Mean Reciprocal Rank (MRR) [133]: Let w be a source word, $rank_w$ denotes the rank of its correct translation within the list of translation candidates, V denotes the set of words used for evaluation. Then MRR is defined as:

$$MRR = \frac{1}{|V|} \sum_{w \in V} \frac{1}{rank_w} \tag{3.10}$$

  We only used the top 20 candidates from the ranked list for calculating MRR. Note that for some source words, the correct translation might be not included in the top 20 candidate list. In this case, we assume $rank_w$ to be infinity, and thus $\frac{1}{rank_w}$ is 0. We did not discard these source words for calculating MRR, namely V is always 1,000. Moreover, if a source word has multiple translations in the test set and more than two of them are included in the candidate list, we used the most highly ranked translation for calculating MRR.

3.3.4 Results

The results for the Chinese-English, Japanese-English and Japanese-Chinese test sets are shown in Figure 3.4, where "Topic" denotes the lexicons extracted only using TMBM described in Section 3.2.1, "Context" denotes the lexicons extracted only using CBM method described in Section 3.2.2, "Combination" denotes the lexicons after applying the combination method described in Section 3.2.3, "K" denotes the number of topics, "N" denotes the number of translation candidates for a word we compared in our experiments, "10k" and "all" denote using 10k and all the article pairs as training data respectively. For Chinese-English and Japanese-English and the $10k$ Japanese-Chinese data, we tried $K = 200$, $K =$

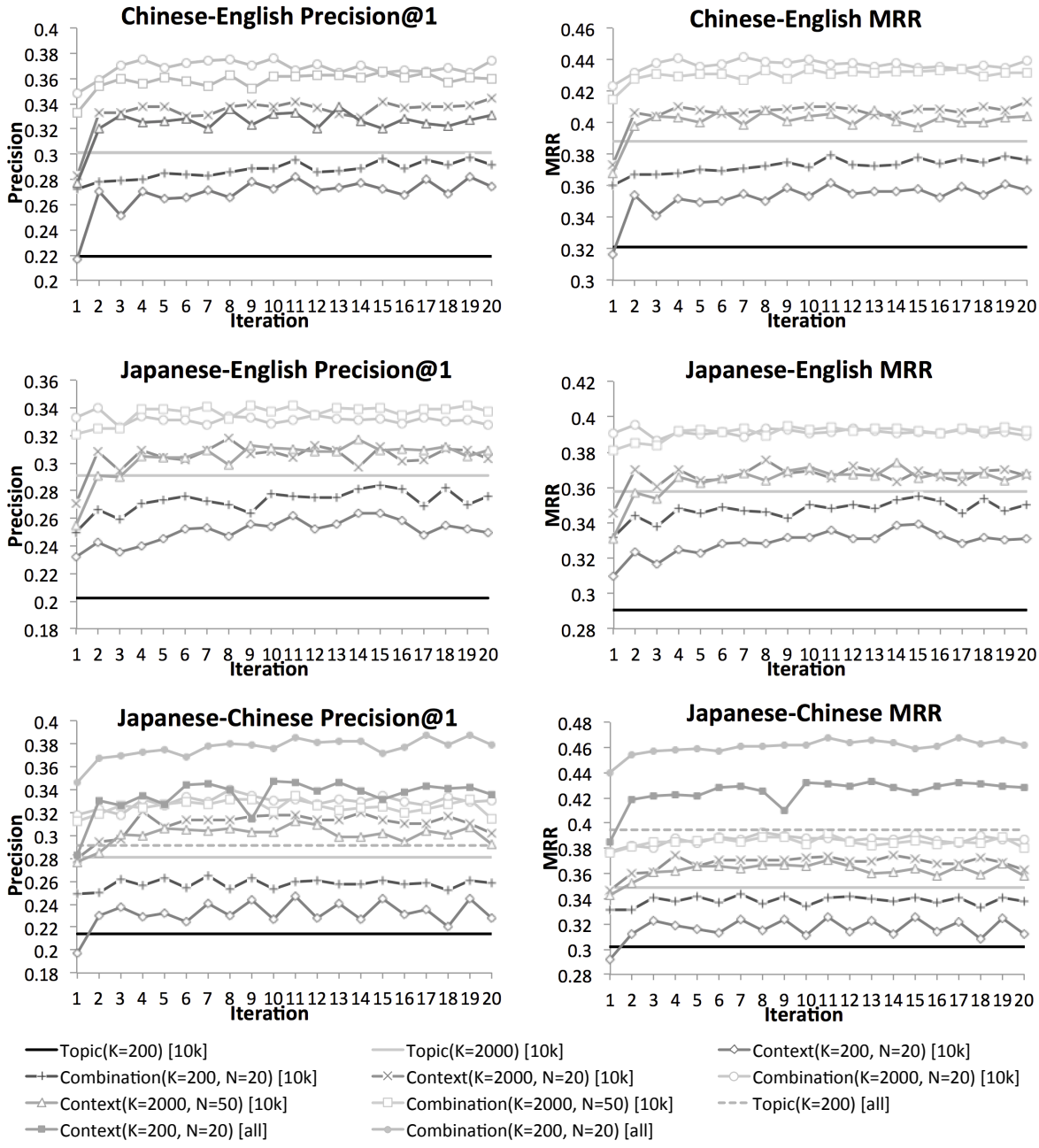

Figure 3.4: Results for Chinese-English, Japanese-English and Japanese-Chinese on the test sets.

2000[12] and $N = 20$, $N = 50$,[13] while for the Japanese-Chinese setting that uses all the articles, we only tried $K = 200$[14] and $N = 20$.

In general, we can see that our proposed method can significantly improve the accuracy in both Precision@1 and MRR metrics compared to "Topic." "Context" outperforms "Topic," which verifies the effectiveness of using the lexicons extracted by TMBM as seed dictionary for CBM. "Combination" performs better than both "Topic" and "Context," which verifies the effectiveness of using both topical and contextual knowledge for BLE. Moreover, iteration can further improve the accuracy, especially in the first few iterations. Detailed analysis for the results will be given in Section 3.4.

3.4 Discussion

Why are Our "Topic" Scores Lower Than Vulić et al. [135]?

The "Topic" scores are lower than the ones reported in [135], which are over 0.6 when $K = 2000$. The main reason is that the experimental data we used is much more sparse. Our vocabulary size is from tens of thousands to hundreds of thousands (see Section 3.3.1), while in [135] it is only several thousands (7,160 Italian and 9,166 English nouns). Moreover, the number of article pairs we used for training is less than [135] except for Japanese-Chinese.

Another reason is the evaluation method. It may underestimate simply because of the incompleteness of our test set (e.g. our system successfully finds the correct translation "vehicle" for the Chinese word "车," but our test set only contains "car" as the correct translation). We investigated the words with their top 1 translation incorrect according to our evaluation method. Based on our in-

[12] Vulić et al. [135] empirically studied the effect of the number of topics K on the performance of TMBM. In our experiments, we compared $2,000$ topics that showed the best performance in [135], to a small number of topics 200.

[13] Because 20 is the number of candidates that we used to calculate MRR, we did not try a number smaller than 20. On the other hand, because increasing it to 50 showed worse performance in our experiments, we believe that further increasing N to a number larger than 20 is not helpful.

[14] The reason for this is that $2,000$ is not scalable for this large data set.

vestigation, nearly 30% of them were undervalued because the correct translation given by our system is not included in our test set.

How Does the Proposed Method Perform on Different Language Pairs?

Our proposed method is language-independent, which is also indicated by the experimental results on three different language pairs of Chinese-English, Japanese-English and Japanese-Chinese. In Figure 3.4, we can see that although the "Topic" scores and the absolute values of improvement by our proposed method on Chinese-English, Japanese-English and Japanese-Chinese are different because of the different characteristics of the data, the improvement curves are similar.

How Many Iterations are Required?

In our experiments, we conducted 20 iterations. The accuracy improves significantly in the first few iterations, and after that the performance becomes stable (see Figure 3.4). We suspect the reason is that there is an upper bound for our proposed method. After several iterations, the performance nearly reaches that upper bound, making it difficult to be further improved, thus the performance becomes stable. The iteration number at which the performance becomes stable depends on the particular experimental settings. Therefore, we may conclude that several iterations are enough to achieve a significant improvement, and the performance at each respective iteration depends heavily on the experimental settings.

How Does the Number of Topics Affect the Performance?

According to [135], the number of topics can significantly affect the performance of the "Topic" system. In our experiments, we compared 2,000 topics that show the best performance in [135], to a small number of topics 200 for Chinese-English and Japanese-English. Similar to [135], using 2,000 topics is significantly better than 200 topics for the "Topic" lexicons.

For the affect on the improvement by our proposed method, the improvements over "Topic" are smaller on 2,000 topics than the ones on 200 topics for both "Context" and "Combination." We suspect the reason is that the absolute

values of improvement on the seed dictionary cannot lead to the same level of improvement for CBM. At iteration 1, the improvement of the "Topic" scores cannot fully reflect on the "Context" scores. Thus, the "Context" scores are lower than the "Topic" scores for 2,000 topics, while they are similar to or higher than the "Topic" scores for 200 topics (see Figure 3.4). The performance at iteration 1 impacts the overall improvement performance for the future iterations.

How Does the Number of Candidates Affect the Performance?

In the Chinese-English and Japanese-English experiments, we measured the difference in using 20 and 50 translation candidates for each word. The results show that using more candidates slightly decreases the performance (see Figure 3.4). Although using more candidates may increase the percentage of words where the correct translation is contained within the top N word list of translation candidates (Precision@N), it also leads to more noisy pairs. According to our investigation on Precision@N of the two settings, the difference is quite small. For Chinese-English: Precision@20=0.5620, Precision@50=0.5780, while for Japanese-English: Precision@20=0.4930, Precision@50=0.5030. Therefore, we suspect the decrease is because the negative effect outweighs the positive. Furthermore, using more candidates will increase the computational complexity. Therefore, we believe a small number of candidates such as 20 is appropriate for our proposed method.

How Does the Size of the Training Data Affect the Performance?

In our experiments, we compared two different sizes of Japanese-Chinese training data, i.e., using $10k$ and all the article pairs. In Figure 3.4, we can see that the "Topic" scores of using all the article pairs is much higher than that of using the $10k$ pairs regardless of the number of topics used, which indicates that using more training data can improve the accuracy of TMBM. As for our proposed method, the improvements over "Topic" for "Context" are larger when using all the article pairs than the ones on the $10k$ pairs, indicating that using more training data also can improve the effectiveness of our proposed method.

Candidate	Sim_{Topic}	$Sim_{Context}$	Sim_{Comb}
research	0.0530	**0.2176**	**0.0859**
scientist	0.0525	0.1163	0.0653
science	**0.0558**	0.0761	0.0599
theory	0.0509	0.0879	0.0583
journal	0.0501	0.0793	0.0559

Table 3.1: Improved example of "研究 (research)," where Sim_{Topic} scores are similar, while $Sim_{Context}$ scores are distinguishable.

What Kind of Lexicons are Improved?

Although TMBM has the advantage of finding topic related translations, it lacks of the ability to distinguish candidates that have highly similar word-topic distributions to the source word. This weakness can be solved with CBM.

Table 3.1 shows an improved example of the Chinese word "研究 (research)." All the candidates identified by "Topic" are strongly related to the topic of academia. The differences among the Sim_{Topic} scores are quite small, because of the high similarities of the word-topic distributions between these candidates and the source word, and "Topic" fails to find the correct translation. However, the differences in contextual similarities between the candidates and the source word are quite explicit. With the help of $Sim_{Context}$ scores, our proposed method finds the correct translation. Based on our investigation on the improved lexicons, most improvements belong to this type, where the Sim_{Topic} scores are similar, while the $Sim_{Context}$ scores are easy to distinguish.

Table 3.2 shows an improved example of the Japanese word "施設 (facility)." The Sim_{Topic} scores are similar to the ones in the example of Table 3.1 that are not quite distinguishable, and "Topic" fails to find the correct translation. The difference is that CBM also fails to find the correct translation, and the top 2 $Sim_{Context}$ scores are quite similar. The combination of the two methods successfully finds the correct translation, although this could be by chance. Based on our investigation, a small number of improvements belong to this type, where both Sim_{Topic} and $Sim_{Context}$ scores are not distinguishable.

Candidate	Sim_{Topic}	$Sim_{Context}$	Sim_{Comb}
facility	0.0561	0.1127	**0.0674**
center	0.0525	**0.1135**	0.0647
building	0.0568	0.0933	0.0641
landmark	**0.0571**	0.0578	0.0572
plan	0.0460	0.1007	0.0570

Table 3.2: Improved example of "施設 (facility)," where both Sim_{Topic} and $Sim_{Context}$ scores are not distinguishable.

What Kind of Errors are Made?

As described above, for nearly half of the words in the test sets, the correct translation is not included in the top N candidate list produced by TMBM. We investigated these words and found several types of errors. The majority of errors are caused by unsuccessful identification despite topic alignment being correct (e.g. Japanese word "選手 (player)" is translated as "team"). Some errors are caused by unsuccessful topic alignment between the source and target words (e.g. Japanese word "設置 (establishment)" is translated as "kumagaya" which is a Japanese city name). There are also errors caused by words that do not clearly fit into one topic (e.g. Chinese word "爵士 (jazz/sir)" may belong to either a musical or social topic). The remaining errors are due to English compound nouns. There are several pairs that contain English compound nouns in our test sets (e.g. "香港 (Hong Kong)" in Chinese-English, and "ソ連 (Soviet Union)" in Japanese-English). Currently, our system cannot deal with compound nouns, and we leave it as future work for this study.

There are still some errors for words with their correct translation included in the top N candidate list produced by TMBM, although our proposed method significantly improves the accuracy. Based on our investigation, most errors happen in the case that either the "Topic" or "Context" gives a significantly lower score to the correct translation than the scores given to the incorrect translations, while the other gives the highest or almost highest score to the correct translation. In this case, a simple linear combination of the two scores is not discriminative

enough, and incorporating both scores as features in a machine learning way may be more effective.

3.5 Summary of This Chapter

In this chapter, we presented a BLE system exploiting both topical and contextual knowledge. Our system is based on a novel combination of TMBM and CBM, which does not rely on any prior knowledge and can be iteratively improved. Experiments conducted on Chinese-English, Japanese-English and Japanese-Chinese Wikipedia data verified the effectiveness of our system for BLE from comparable corpora.

Our system can be improved in several aspects. Firstly, the scalability of TMBM is one drawback of our system, which may be solved by the method presented in [83]. Secondly, different definitions of context should be compared for CBM. Thirdly, currently our system cannot deal with compound words, for which compositional [92] and classification approaches have been proposed [5]. Fourthly, our system does not pay special attention to rare words, and smoothing and [103, 55, 56] classification approaches may be considered for this. Fifthly, polysemy should be handled by our system, an aspect often neglected in related studies. Finally, additional experiments should be conducted on other comparable corpora rather than Wikipedia, where article alignment is required beforehand.

Chapter 4

Parallel Sentence Extraction

In statistical machine translation (SMT) [17, 100, 71], because translation knowledge is acquired from parallel corpora, the quality and quantity of parallel corpora are crucial. However, as described in Section 1.2, parallel corpora remain a scarce resource. As comparable corpora are far more available, automatic construction of parallel corpora from comparable corpora is an attractive research field.

Many studies have been conducted on constructing parallel corpora from comparable corpora, such as bilingual news articles [152, 131, 93, 127, 35, 1], patent data [132, 85] and social media [82]. The Web also can be seen as large comparable corpora, and many studies have been conducted for constructing parallel corpora from it [114, 65, 58]. Recently, some researchers try to construct parallel corpora from Wikipedia [2, 119, 32].

While most previous studies are interested in language pairs between English and other languages, we focus on Chinese-Japanese, where parallel corpora are very scarce. In this chapter, we describe our efforts to improve a parallel sentence extraction system for constructing a Chinese-Japanese parallel corpus from Wikipedia. The system is inspired by [93], which mainly consists of a parallel sentence candidate filter and a classifier for parallel sentence identification. The main contributions of this chapter are in two aspects:

- Using common Chinese characters described in Chapter 2 for the filter to addressing 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.

Experiments show that our system performs significantly better than the previous studies for both accuracy in sentence extraction and SMT performance. Using the system, we construct a Chinese-Japanese parallel corpus with more than 126k highly accurate parallel sentences from Wikipedia. In addition, we apply bootstrapping and the bilingual lexicons extracted in Chapter 3 for parallel sentence extraction, which further improve the performances.

4.1 Related Work

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 [131, 41, 93, 44]. Large-scale article alignment from the Web also has been studied [96, 114, 150, 42, 130]. 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 [127, 1, 33, 82, 29]. 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 [93, 127, 119, 13, 33] and translation similarity measures [131, 41, 42, 1]. 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 [131, 41, 2, 85] or a seed parallel corpus [152, 93, 127, 119, 44, 1, 33, 32, 82], or link structure and meta data in Wikipedia [13]. This study uses a seed parallel corpus. An unsupervised method

has also been proposed [35], however their method suffers from high computational complexity.

Bootstrapping has been proven effective in some related works [88, 41, 93]. These studies update the bilingual dictionary required for the parallel sentence extraction system, by generating new bilingual lexicons from the extracted parallel sentences. The updated dictionary has a higher coverage that can improve the performance of the parallel sentence extraction system. Bootstrapping can be applied to our system in a similar way.

Bilingual lexicon extraction (BLE) has been used for parallel sentence extraction [119]. They extract bilingual lexicons from aligned Wikipedia articles based on a supervised method. Then they use the extracted lexicons for parallel sentence extraction. One drawback of their method is that manually created language specific training data is required to achieve satisfactory results, which is difficult to obtain. This study differs from [119] in using an unsupervised BLE method described in Chapter 3, which does not require manual efforts.

Previous studies extract parallel sentences from various types of comparable corpora, such as bilingual news articles [152, 131, 93, 127, 35, 44, 1], patent data [132, 85], social media [82], and the Web [96, 114, 150, 64, 65, 42, 58]. However, few studies have been conducted to extract parallel sentences from Wikipedia [2, 119, 13, 32]. 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.

4.2 Chinese-Japanese Wikipedia

Wikipedia[1] is a free, collaborative and multilingual encyclopedia. Chinese and Japanese Wikipedia are in the top 20 language editions of Wikipedia, with more than 740k and 887k articles respectively (24th December 2013).

As parallel sentences tend to appear in similar article pairs, article alignment is the first step for extracting parallel sentences from comparable corpora in many previous studies. A special characteristic of Wikipedia is that article alignment is

[1] http://en.wikipedia.org/wiki/Wikipedia

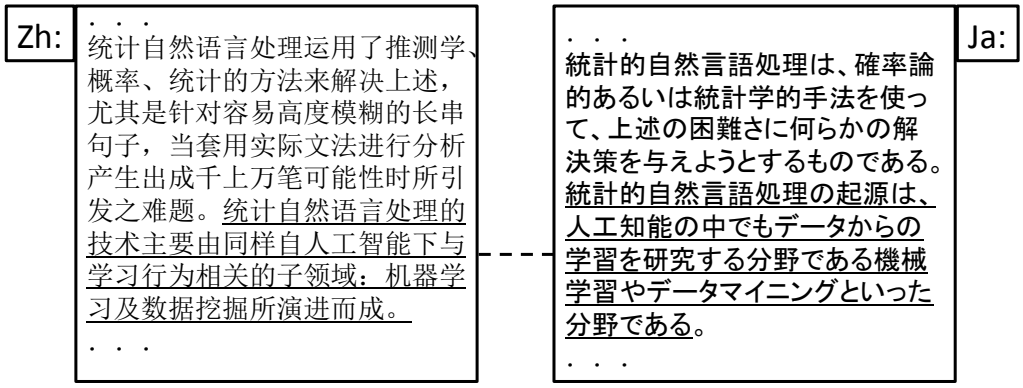

Figure 4.1: Example of aligned Chinese and Japanese article pairs via interlanguage links from Wikipedia, both describe the topic of "statistical natural language processing" (parallel sentences are linked with dashed lines).

established via interlanguage links. Because parallel sentences tend to appear in these linked article pairs, Wikipedia can be a valuable resource for constructing parallel corpora. Figure 4.1 shows an example of aligned article pairs via interlanguage links from Chinese and Japanese Wikipedia, where there are parallel sentences. Our task is to identify the parallel sentences from the aligned article pairs.

4.3 Parallel Sentence Extraction System

The overview of our parallel sentence extraction system is presented in Figure 4.2. We first align articles on the same topic in the Chinese and Japanese Wikipedia via the interlanguage links ((1) in Figure 4.2). 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 4.2).[2] Finally, we use a classifier trained

[2]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 [127, 1, 33, 82] or perform article alignment beforehand [131, 41, 93].

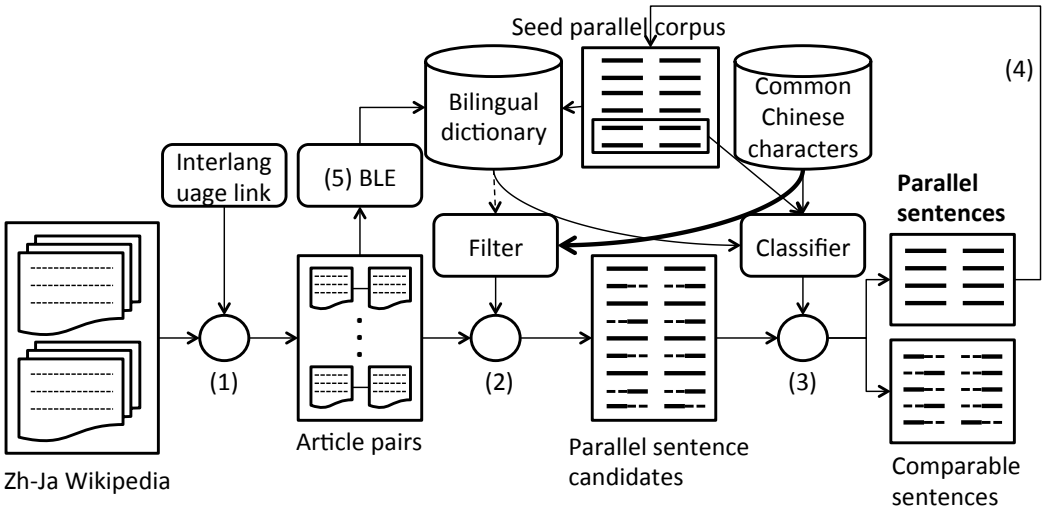

Figure 4.2: Parallel sentence extraction system.

on a small number of parallel sentences from a seed parallel corpus to classify the parallel sentence candidates into parallel and comparable sentences ((3) in Figure 4.2). Once the parallel sentences are extracted, they can be appended to the seed parallel corpus for bootstrapping ((4) in Figure 4.2). Moreover, our system applies BLE described in Chapter 3 for parallel sentence extraction ((5) in Figure 4.2).

Our system differs from previous studies in the strategy of the filter and the features used for the classifier, which will be described in Section 4.3.1 and Section 4.3.2 in detail.

4.3.1 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 [93]. Although the sentence length ratio condition is domain

independent, the word overlap condition is not.[3] 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 unavailable, 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 [27]. 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 4.4.3.

4.3.2 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

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

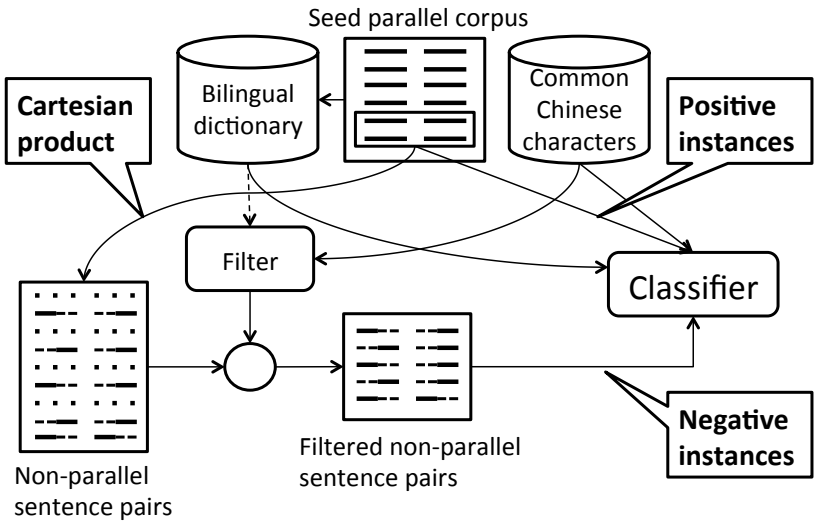

Figure 4.3: Parallel sentence classifier.

the training and testing process, and then introduce the features we use for the classifier.

Training and Testing

We use a support vector machine classifier [20]. Training and testing instances for the classifier are created following the method of [93]. 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 4.3.1. Moreover, we randomly discard some negative instances for training when necessary[4] to guarantee that the ratio of negative to positive instances is less than five for the performance of the classifier. Figure 4.3 illustrates this process.

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

[4]Note that we keep all negative instances for testing.

(CC) features, Non-CC word features, and content word features.

Basic Features. The basic features were proposed in [93]:

- Sentence length, length difference, and length ratio.[5]
- 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[6] 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 4.4 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 ($18/14 = 128\%$).
- Number of n-gram common Chinese characters (1-gram: 12, 2-gram: 6, 3-gram: 2, 4-gram: 1).

[5] In our experiments, sentence length was calculated based on the number of words in a sentence.

[6] 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 [93], we recommend the interested readers to refer to the original paper.

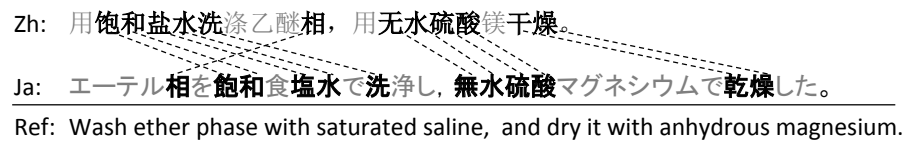

Figure 4.4: Example of common Chinese characters (in bold and linked with dotted lines) in a Chinese-Japanese parallel sentence pair.

- 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 [27]. Note that Chinese character features are only applicable to Chinese-Japanese. However, because common Chinese characters can be seen as cognates, the similar idea can be applied to other language pairs sharing cognates. Cognates among European languages have been shown effective in word alignments [75] and parallel fragment extraction [4]. We also can use cognates for parallel sentence extraction, however we leave it as future work.

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 [93] 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.[7]

4.4 Experiments

We evaluated classification accuracy, and conducted extraction, translation, bootstrapping and BLE based experiments to verify the effectiveness of our proposed parallel sentence extraction system. In all our experiments, we preprocessed the data by segmenting and POS tagging Chinese and Japanese sentences using a tool proposed by Chu et al. [25] and JUMAN [77], respectively.

4.4.1 Data

The seed parallel corpus we used is the Chinese-Japanese section of the Asian Scientific Paper Excerpt Corpus (ASPEC).[8] This corpus is a scientific domain corpus provided by the Japan Science and Technology Agency (JST)[9] and the National Institute of Information and Communications Technology (NICT).[10] It

[7]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 [144]. 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 [77].

[8]http://lotus.kuee.kyoto-u.ac.jp/ASPEC

[9]http://www.jst.go.jp

[10]http://www.nict.go.jp

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[11] (2012/09/21) and Japanese[12] (2012/09/16) Wikipedia database dumps. We used an open-source Python script[13] 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.[14] 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).

4.4.2 Classification Accuracy Evaluation

We evaluated classification accuracy using two distinct sets of 5k parallel sentences from the seed parallel corpus for training and testing, respectively. For the support vector machine classifier, we used the LIBSVM toolkit [20][15] with 5-fold cross-validation and a radial basis function kernel. In this section and Section 4.4.3, 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 4.4.4. We used the word alignment tool GIZA++[16] 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 larger than 0.1 for each source word following [93].[17] Word overlap was calculated based on that dictionary. We report the results using word overlap filtering, for easier comparison to previous studies.

[11] http://dumps.wikimedia.org/zhwiki

[12] http://dumps.wikimedia.org/jawiki

[13] http://code.google.com/p/recommend-2011/source/browse/Ass4/WikiExtractor.py

[14] http://svn.wikimedia.org/svnroot/mediawiki/branches/REL1_12/phase3/includes/ZhConversion.php

[15] http://www.csie.ntu.edu.tw/~cjlin/libsvm

[16] http://code.google.com/p/giza-pp

[17] Note that the dictionary might contain noisy translation pairs and further cleaning them might be helpful for our task [6], however, we leave it as future work.

The word overlap threshold was set to 0.25. We compared the following feature settings:

- Munteanu+, 2005: the basic features proposed in [93] 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:

$$precision = 100 \times \frac{classified\_well}{classified\_parallel}, \tag{4.1}$$

$$recall = 100 \times \frac{classified\_well}{true\_parallel}, \tag{4.2}$$

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{4.3}$$

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 4.1. 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.

4.4.3 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 [72] 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

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**

Table 4.1: Classification results.

SRILM toolkit [122][18] with interpolated Kneser-Ney discounting.[19] For tuning and testing, we used two distinct sets of 198 parallel sentences with 1 reference. 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[20] 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 [98], 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 4.4.2.

Parallel sentence extraction and translation results using different methods are shown in Table 4.2. We report the Chinese-to-Japanese translation results on the test set using the BLEU-4 score [102]. "Munteanu+, 2005," "+CC," "+Non-CC," and "+Content" denote the different features described in Section 4.4.2. "Word," "CCO," "Word and CCO," and "Word or CCO" denote the four different filtering strategies described in Section 4.3.1. "# 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 [69].

[18] http://www.speech.sri.com/projects/srilm

[19] Note that the Japanese sentences in the tuning and testing sets were not discarded from the data used for training the language model.

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

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

Table 4.2: Parallel sentence extraction and translation results ("†" and "‡" denote that the result is significantly better than "Munteanu+ 2005" and "+CC" respectively at $p < 0.05$).

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, 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 4.5 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.

The parallel sentences extracted by the best performing method, +Content with CCO filter, and the tuning and testing sets used in the translation experiments are available at `http://lotus.kuee.kyoto-u.ac.jp/~chu/resource/wiki_zh_ja.tgz`.

4.4.4 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 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 4.4.3. 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 4.4.3.

Table 4.3 shows the translation results for different thresholds. We can see

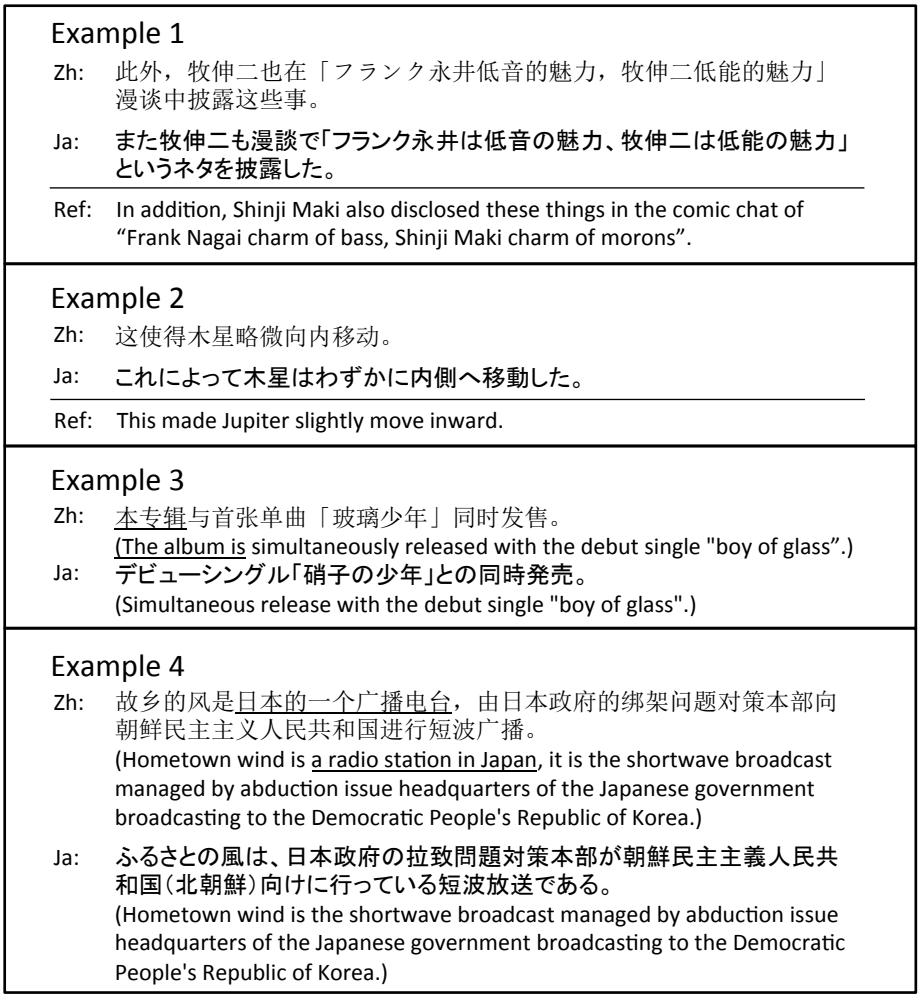

Figure 4.5: Examples of some extracted parallel sentences (noisy parts are underlined).

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 4.3: Translation results for different thresholds.

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. Chapter 5 describes our proposed method to extract the parallel fragments from these comparable sentences to further improve SMT.

4.4.5 Bootstrapping Experiments

We further conducted bootstrapping experiments following [88, 41, 93]. We first appended the extracted parallel sentences to the seed parallel corpus. Then we generated a new bilingual dictionary from the combined corpus. Finally, we used the new bilingual dictionary for parallel sentence extraction. Bootstrapping experiments were conducted based on the best performing method, +Content with CCO filter, described in Section 4.4.3. Experimental settings were the same as the ones used in the extraction and translation experiments described in Section 4.4.3. We iterated until there were no further improvements in MT performance on the tuning set.

Bootstrapping sentence extraction and translation results using different methods are shown in Table 4.4. "Seed" is the same one described in Section 4.4.3 that uses the seed parallel corpus as training data. "+Content with CCO filter" denotes the best performing method described in Section 4.4.3. "Iteration" denotes different iterations using our bootstrapping method. The number after "Iteration" denotes iteration number. "# dictionary entries" denotes hereafter the number of dictionary entries for different methods.

We can see that the number of entries of the generated bilingual dictionary increases by bootstrapping. The increased entries are newly generated from the

Method	# dictionary entries	# sentences	BLEU-4	OOV
Seed			25.42	9.11
+Content with CCO filter	204,254	126,811	37.82	3.71
Iteration 1	274,496	164,403	37.99	3.40
Iteration 2	292,186	167,310	**38.71**	3.38

Table 4.4: Bootstrapping sentence extraction and translation results.

parallel sentences extracted from Wikipedia in the earlier iteration, which are helpful for extracting more parallel sentences. More parallel sentences lead to lower OOV word rates, and improve MT performance.

4.4.6 Bilingual Lexicon Extraction Based Experiments

We further conducted BLE based experiments. In the BLE based experiments, we directly extracted bilingual lexicons from the aligned Chinese-Japanese Wikipedia articles, and used the extracted lexicons for parallel sentence extraction. The details of BLE were described in Section 3.3. We compared two different settings for seed parallel corpus based lexicon generation to investigate the affect of the initial dictionary size on the BLE based experiments:

- Baseline ($10k$): used the $10k$ parallel sentences from the seed parallel corpus, which were used as the training and testing data for the classifier in Section 4.4.2.

- Baseline ($680k$): used all the parallel sentences ($680k$) in the seed parallel corpus.

For the generated lexicons, we kept the top five translations with translation probability larger than 0.1 for each source word. For the bilingual lexicon extraction based experiments, we used the Japanese-Chinese bilingual lexicons extracted at iteration 2 of our proposed method (i.e., combination at iteration 2) shown in Figure 3.4, which show significant improvement over the previous iteration and the Topic method. We empirically kept the bilingual lexicon extraction results for

Method	# dictionary entries	# sentences	BLEU-4	OOV
Seed (10k)			16.59	23.18
Baseline (10k)	32,607	57,681	33.62†	5.83
Baseline (10k) + lexicon	87,523	94,931	35.30†‡	4.93
Seed (680k)			25.42	9.11
Baseline (680k)	204,254	126,811	37.82†	3.71
Baseline (680k) + lexicon	258,124	152,511	**37.97**†	3.38

Table 4.5: Bilingual lexicon extraction based parallel sentence extraction and translation results ("†" and "‡" denote the result is significantly better than "Seed" and "Baseline" respectively at $p < 0.01$).

the source (Japanese) words whose frequencies are not smaller than 100 and the top three candidates for each source word, containing about $56k$ lexicons. The reason for only using the highly frequent lexicons is that the extraction results are noisy for the words with low frequencies. We combined the lexicons generated from the parallel sentences in the seed parallel corpus with the extracted bilingual lexicons, further obtaining following two dictionary settings:

- Baseline ($10k$) + lexicon: combined the Baseline ($10k$) dictionary with the extracted bilingual lexicons.
- Baseline ($680k$) + lexicon: combined the Baseline ($680k$) dictionary with the extracted bilingual lexicons.

The word overlap features were calculated based on the above four different dictionary settings, obtaining four classifiers that estimate the word overlap features using different dictionaries while the other settings are the same. BLE based experiments were also conducted based on the best performing method, +Content with CCO filter, described in Section 4.4.3. Other experimental settings were the same as the ones used in the extraction and translation experiments described in Section 4.4.3.

Parallel sentence extraction and translation results using different methods are shown in Table 4.5. For comparison, we also conducted translation experiments

using the parallel sentences from the seed parallel corpus that were used for lexicon generation, as SMT training data (labeled "Seed ($10k$)" and "Seed ($680k$)").

Naturally, using more parallel sentences from the seed parallel corpus can improve the performances of both the Seed and Baseline systems. The Baseline + lexicon systems outperform the Baseline systems. The reason for this is that combining the extracted bilingual lexicons to the Baseline dictionaries can help to extract more parallel sentences, leading to lower OOV word rates and thus higher SMT performances. The improvement on Baseline ($10k$) is larger than that of Baseline ($680k$), indicating that the extracted bilingual lexicons are more helpful in the case that we only have a small seed parallel corpus for lexicon generation. Baseline ($680k$) + lexicon does not show significant difference over Baseline ($680k$). We suspect the reason for this is the ratio of the number of the extracted lexicons to the number of lexicons in the Baseline dictionary is much smaller than that of Baseline ($10k$) + lexicon to Baseline ($10k$), which also leads to a smaller ratio of newly extracted sentences that does not lead to a significant difference on MT.

Focusing on the difference of the number of dictionary entries between the Baseline and Baseline + lexicon systems, in the case of Baseline (10k) the difference is $87,523 - 32,607 = 54,916$, and it is $258,124 - 204,254 = 53,870$ in the case of Baseline (680k). Because the number of extracted lexicons that we combined to the Baseline system is $56k$, we can know that there are only a few overlaps between the extracted lexicons and Baseline dictionary, even we use all the parallel sentences (680k) in the seed parallel corpus for lexicon generation. The reason for this is the domain difference between the seed parallel corpus and Wikipedia. Because our proposed method can extract in-domain lexicons from comparable corpora, it does not require any in-domain seed parallel corpus, which is another advantage of our proposed method.

Figure 4.6 shows some examples of sentences additionally extracted by combining the extracted bilingual lexicons to Baseline ($10k$). The Baseline system cannot extract these sentence pairs, because of the low word overlap between them based on the Baseline generated dictionary. Combining the extracted bilingual lexicons increases the word overlap, making these sentences been extracted. Based on our

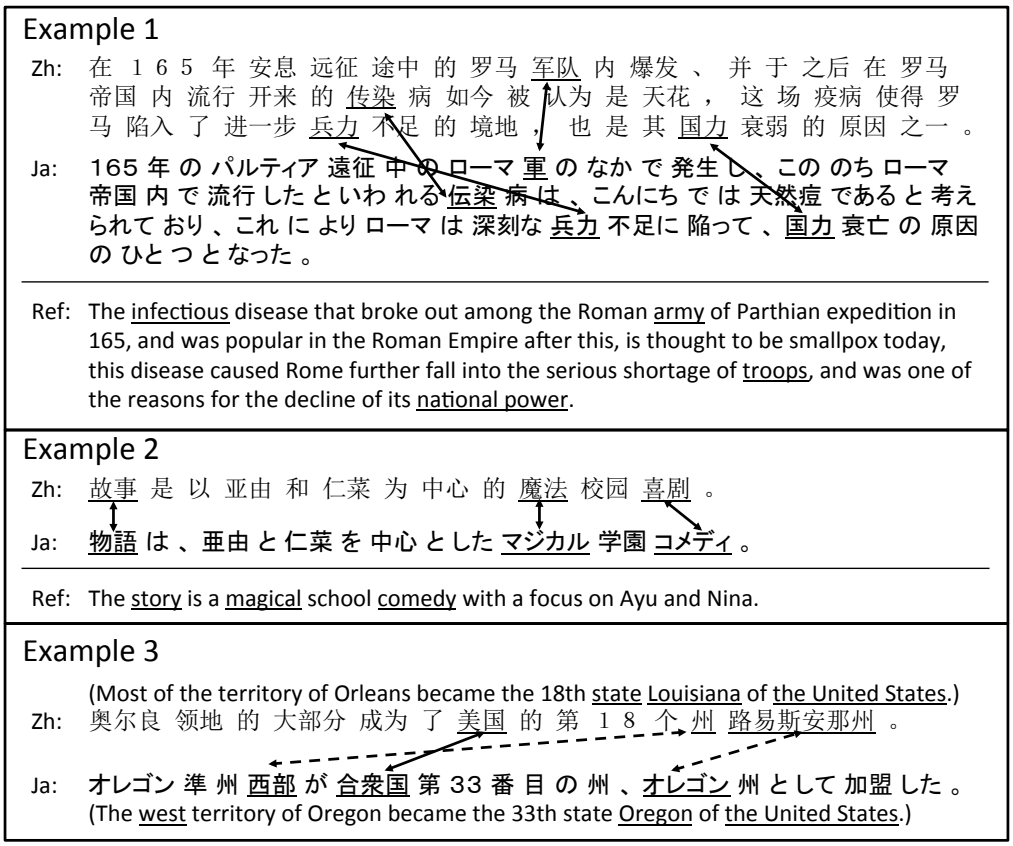

Figure 4.6: Examples of sentences additionally extracted by combining the extracted bilingual lexicons to the Baseline (example 1 and 2 are truly parallel sentences, while example 3 is an erroneous parallel sentence pair). The lexicon pairs that do not exist in the Baseline generated dictionary but extracted by our bilingual lexicon extraction method are linked (correct lexicon pairs are linked with solid lines, incorrect lexicon pairs are linked with dashed lines).

investigation, about 2/3 of the additionally extracted sentences are truly parallel sentences. The rest erroneous parallel sentences are extracted because of the noise contained in the extracted bilingual lexicons. Example 3 in Figure 4.6 shows an erroneous parallel sentence pair, it is extracted because of the noise lexicons "州 (state), 西部 (west)" and "路易斯安那州 (Louisiana), オレゴン (Oregon)." One possible solution to address this problem is further discarding these noisy lexicon pairs by setting a more strict filtering threshold, however, it might decrease the coverage of the lexicon.

4.5 Summary of This Chapter

In this chapter, we improved a parallel sentence extraction system by using the common Chinese characters for filtering, and three novel feature sets for classification. The Experimental results on Wikipedia showed that our proposed methods are more effective than the previous studies. In addition, we conducted bootstrapping and BLE based experiments, which can further improve the performance of our system.

Our study showed that Chinese characters are significantly helpful for Chinese-Japanese parallel sentence extraction. As future work, we plan to apply the similar idea to other language pairs by using cognates. Moreover, in this chapter we only conducted experiments on Wikipedia. Our proposed system is expected to work well on other comparable corpora, such as bilingual news articles, patent data and social media. We plan to do experiments on these comparable corpora to construct a large parallel corpus for various domains.

Chapter 5

Parallel Fragment Extraction

In statistical machine translation (SMT) [17, 100, 71], because translation knowledge is acquired from parallel data, the quality and quantity of parallel data are crucial. However, as described in Section 1.2, parallel data remains a scarce resource. As non-parallel corpora are far more available, extracting parallel data 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) [41]. Many studies focus on extracting parallel sentences from noisy parallel corpora or comparable corpora, such as bilingual news articles [152, 131, 93, 127, 1], patent data [132, 85] and social media [82]. Studies have also been conducted on quasi-comparable corpora [94, 109]. Although quasi-comparable corpora are available in far larger quantities than noisy parallel or comparable corpora, there are few or no parallel sentences. However, there could be parallel fragments in comparable sentences that are also helpful for SMT.

One important fact that most previous studies ignore is that there could be

both parallel sentences and fragments in many comparable corpora. [1] 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 in comparable sentences. 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 1.2 in Chapter 1 shows an example of comparable texts 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 for this type of comparable corpora.

The fragments in quasi-comparable corpora and Wikipedia have one common point that they both exist in comparable sentences. Previous studies have found it difficult to accurately extract parallel fragments from comparable sentences. Some studies extract parallel fragments relying on a probabilistic bilingual lexicon estimated on a seed parallel corpus. They locate the source and target fragments independently, making the extracted fragments unreliable [94]. Some studies develop alignment models for comparable sentences to extract parallel fragments [109]. Because the comparable sentences are quite noisy, the extracted fragments are not accurate.

In this chapter, we propose an accurate parallel fragment extraction system. We locate parallel fragment candidates using an alignment model, and use an accurate lexicon-based filter to identify the truly parallel ones. We further use common Chinese characters for the lexicon-based filter to improve its coverage. Experiments are conducted on both Chinese-Japanese quasi-comparable corpora and Wikipedia. The experimental results show that our proposed method signif-

[1] Although [93, 94, 51] were aware of this possibility, none of them provided an integrated framework that addresses both problems.

icantly outperforms a state-of-the-art approach, which indicate the effectiveness of our parallel fragment extraction system. Moreover, we investigate the factors that may affect the performance of our system in detail.

5.1 Related Work

[94] 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 address this problem by using an alignment model to locate the source and target fragments simultaneously.

Quirk et al. [109] 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 alignment models cannot accurately extract parallel fragments. To addressing 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 [57] 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 [116] used a syntax-based alignment model to extract parallel fragments from noisy parallel data. Gupta et al. [51] 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. [39] proposed a method that is based on hierarchical phrase-based force decoding. Afli et al. [3] attempted to extract parallel fragments from multimodal comparable corpora. Supervised

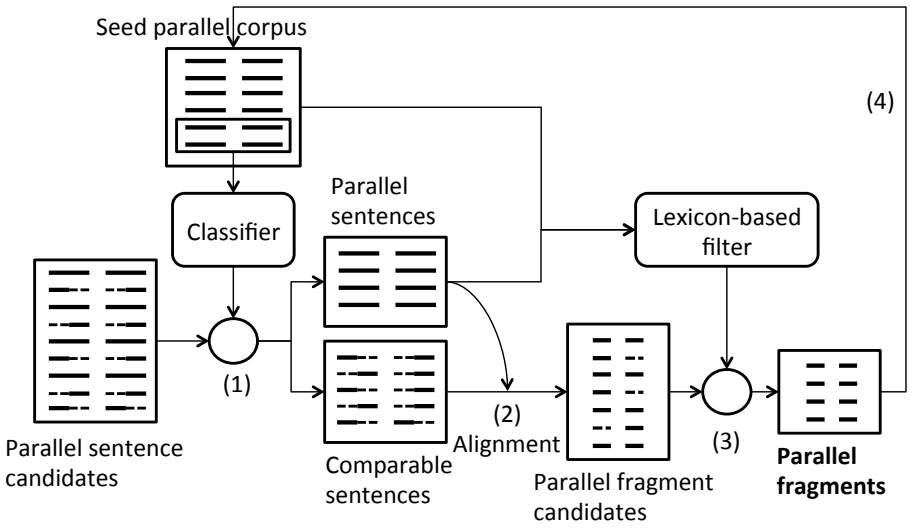

Figure 5.1: Parallel fragment extraction system.

methods have also been proposed for parallel fragment extraction [4]. Zhang and Zong [148] 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.

There are also some studies that try to extract parallel sentences from quasi-comparable corpora. Fung and Cheung [41] proposed a multi-level bootstrapping approach for parallel sentence extraction from quasi-comparable corpora. Wu and Fung [143] exploited generic bracketing inversion transduction grammars (ITG) for this task. Chu et al. [29] used a classification approach. As there are few or no parallel sentences in quasi-comparable corpora, these studies only can extract comparable sentences that contain parallel fragments.

5.2 Proposed Method

5.2.1 System Overview

Figure 5.1 shows an overview of our parallel fragment extraction system. Similar to parallel sentence extraction, we first generate parallel sentence candidates,

and apply 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 ((1) in Figure 5.1) (Refer to Chapter 4 for the details of parallel sentence candidate generation and classification).[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 ((2) in Figure 5.1). We then filter the candidates using probabilistic bilingual lexicons to produce accurate results ((3) in Figure 5.1). Similar to parallel sentence extraction, the extracted parallel fragments can be appended to the seed parallel corpus for bootstrapping ((4) in Figure 5.1). We will present the details of our proposed method in the following sections.

5.2.2 A Brief Example

Figure 5.2 shows an example of comparable sentences extracted by our system from a Chinese-Japanese quasi-comparable corpus. The alignment results are computed by IBM models [17]. We notice that the truly parallel fragments "lead ion selective electrode" and "potentiometric titration method" 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.

5.2.3 Parallel Fragment Candidate Detection

In our experiments, we tried the bidirectional IBM models [17] with symmetrization heuristics [72], and a Bayesian subtree alignment model [95] for parallel fragment candidate detection. The generative alignment models proposed by Quirk et al. [109] that are designed for comparable sentences may be more efficient,

[2]Note that in comparable corpora where article alignment has not been established, the process of candidate generation might rely on cross-lingual information retrieval (CLIR) [29]. Moreover, in the case of quasi-comparable corpora where there are few parallel sentences, we might get comparable sentences only after classification.

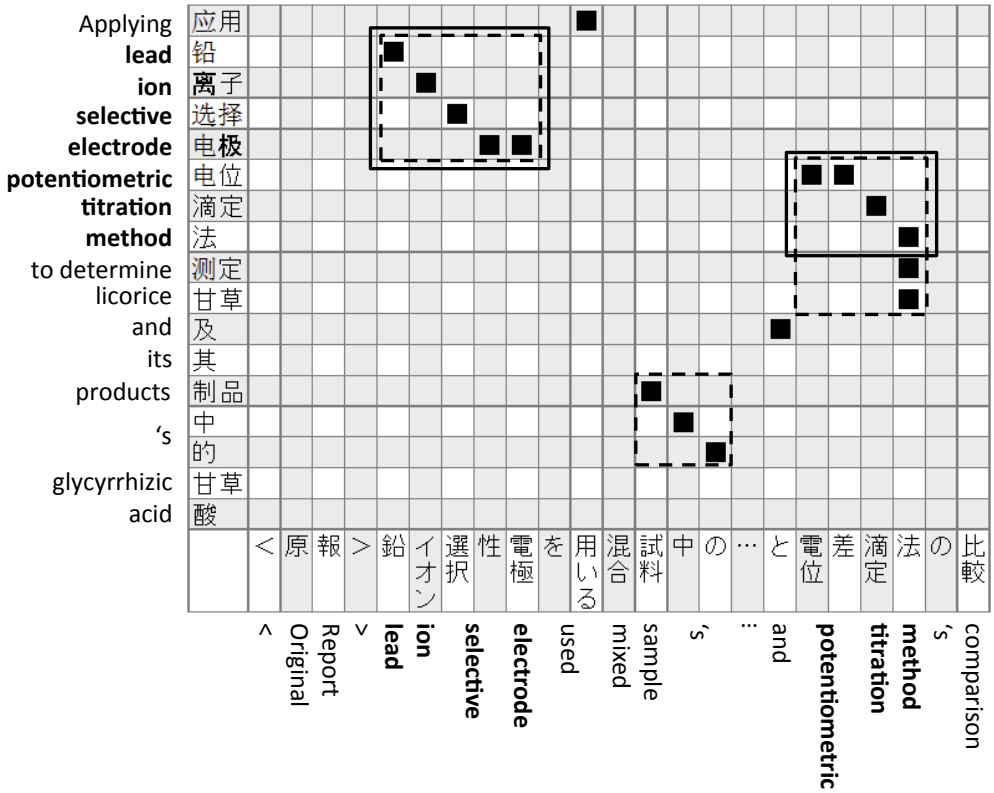

Figure 5.2: Example of comparable sentences with alignment results computed by IBM models (parallel fragment candidates are in dashed rectangles, parallel fragments are in solid-border rectangles).

however we leave this for future work. For alignment, we use both the extracted parallel and comparable sentences, which can help improve the alignment accuracy for the comparable sentences.[3]

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. [109] 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 5.2 as an example, we extract the fragments in dashed rectangles as parallel fragment candidates.

5.2.4 Lexicon-Based Filter

The parallel fragment candidates cannot be used directly, because many of them are still noisy as shown in Figure 5.2. To produce accurate results, we use a lexicon-based filter. We filter a candidate parallel fragment pair with probabilistic bilingual lexicons. The lexicon-pair may be extracted from a seed parallel corpus, or from comparable corpora using some state-of-the-art bilingual lexicon extraction (BLE) approaches such as our proposed BLE approach described in Chapter 3. Furthermore, the parallel sentences extracted by our system can also be used for lexicon generation. In this study, we append the extracted parallel sentences to a seed parallel corpus to generate the lexicons (called hereafter the combined parallel corpus).[4] Different lexicons may have different filtering effects. Here, we compare three types of lexicon.

- IBM Model 1: The first lexicon we use is the IBM Model 1 lexicon, obtained by running GIZA++[5] that implements the sequential word-based statistical

[3]Note that in the case of quasi-comparable corpora, parallel sentences might not be available.

[4]Note that in quasi-comparable corpora, because parallel sentences might not be available, we generate the lexicons only from a seed parallel corpus in this case.

[5]http://code.google.com/p/giza-pp

alignment model of the IBM models on the combined parallel corpus.

- LLR: The second lexicon we use is the LLR lexicon. Munteanu and Marcu [94] 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 [94] 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 [94].

- SampLEX: The last lexicon we use is the SampLEX lexicon. Vulić and Moens [137] 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 [137].

To gain new knowledge that does not exist in the lexicon, we apply a smoothing filter similar to [94]. 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 [27] 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 5.3.1. 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 [94], 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 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.

5.3 Experiments

In our experiments, we compared our proposed fragment extraction method with [94]. We manually evaluated the accuracy of the extracted fragments. We used the extracted fragments as additional SMT training data, and evaluated the effectiveness of the fragments for SMT. For people who want to reproduce the results reported in this chapter, we released a software that contains all the required codes at `http://lotus.kuee.kyoto-u.ac.jp/~chu/code/FragExtractor.tar.gz`.

We conducted experiments on two types of comparable corpora. One is quasi-comparable corpora, where only parallel fragments exist. The other is Wikipedia, where both parallel sentences and fragments exist, thus integrated extraction is desirable. Our experiments were conducted on Chinese-Japanese data. In all our experiments, we preprocessed the data by segmenting Chinese and Japanese sentences using a segmenter proposed by Chu et al. [25] and JUMAN [77] respectively.

5.3.1 Experiments on Quasi-comparable corpora

In this section, we describe the experiments conducted on a Chinese-Japanese quasi-comparable corpus. We investigated the effect of different settings for our proposed method. Moreover, we conducted bootstrapping experiments.

Data

Seed Parallel Corpus. The seed parallel corpus we used is the Chinese-Japanese part of Asian Scientific Paper Excerpt Corpus (ASPEC),[6] which was also used in Chapter 4. This corpus was provided by JST[7] and NICT.[8] It was created by the Japanese project "Development and Research of Chinese-Japanese Natural Language Processing Technology," containing 680k sentences (18.2M Chinese and 21.8M Japanese tokens respectively). This corpus contains various scientific domains such as chemistry, physics, biology and agriculture.

Quasi-Comparable Corpus. The quasi-comparable corpus we used is scientific paper abstracts collected from academic websites. The Chinese side of the corpus was collected from CNKI,[9] containing 420k sentences and 90k articles. The Japanese side of corpus was collected from CiNii[10] web portal, containing 5M sentences and 880k articles. Most articles in the Chinese side of the corpus belong to the domain of chemistry, while the Japanese side of the corpus contains various domains such as chemistry, physics and biology. Note that because the articles in these two websites were written by Chinese and Japanese researchers respectively, the collected corpus is very-non-parallel. In addition, article alignment has not been established for this corpus.

Extraction Experiments

We first applied comparable sentence extraction from the quasi-comparable corpus using a system proposed by Chu et al. [29]. This system is originally proposed for extracting parallel sentences from quasi-comparable corpora. Because there are

[6] http://orchid.kuee.kyoto-u.ac.jp/ASPEC
[7] http://www.jst.go.jp
[8] http://www.nict.go.jp
[9] http://www.cnki.net
[10] http://ci.nii.ac.jp

few or no parallel sentences in quasi-comparable corpora, this system only can extract comparable sentences that contain parallel fragments. In general, the system of [29] is similar to the parallel sentence extraction system proposed in Chapter 4, with an additional component that uses CLIR for candidate sentence generation. In detail, we first translated the Chinese sentences in the quasi-comparable corpus to Japanese with a SMT system trained on the seed parallel corpus. Then we used the translated Japanese sentences as queries for information retrieval. We retrieved the top 10 Japanese documents for each Chinese sentence using Indri,[11] and used all sentences in the Japanese documents as sentence candidates. Next, we identified the comparable sentences from the candidates using a classifier trained on 5k parallel sentences from the seed parallel corpus. We treated the sentence pairs with classification probability ≥ 0.5 as comparable sentences, obtaining 30k chemistry domain sentences.

We then applied fragment extraction on the extracted comparable sentences. For our proposed method, different alignment models may have different effects for parallel fragment candidate detection. Therefore, we compared the following two alignment models:

- GIZA++: It implements the sequential word-based statistical alignment model of IBM models.
- Nakazawa+: It is a Bayesian subtree alignment model [95]. Nakazawa and Kurohashi [95] showed that it performs better than IBM models especially for distant language pairs such as Japanese-English. Because this alignment model is dependency tree-based, we used the Chinese dependency analyzer CNP [22], while the Japanese dependency analyzer was KNP [66]. After alignment, we converted the subtree alignment results to word sequences for our proposed method.

Moreover, external parallel data might be helpful for the alignment models to detect parallel fragment candidates from comparable sentences. Therefore, we compared two different settings to investigate the influence of external parallel data for alignment to our proposed method:

[11] http://www.lemurproject.org/indri

Method	# fragments	Average size (zh/ja)	Accuracy
Munteanu+, 2006	28.4k	20.36/21.39	(1%)
Only (IBM)	18.9k	4.03/4.14	80%
Only (LLR)	18.3k	4.00/4.14	**89%**
Only (SampLEX)	18.4k	3.96/4.05	87%
External (IBM)	28.7k	4.18/4.33	81%
External (LLR)	26.9k	4.17/4.33	85%
External (SampLEX)	28.0k	4.11/4.23	82%

Table 5.1: Fragment extraction results on quasi-comparable corpora using "GIZA++" for parallel fragment candidate detection (accuracy was manually evaluated on 100 fragments randomly selected from the fragments extracted by different methods, based on the number of exact matches).

- Only: Only use the extracted comparable sentences.
- External: Use a small number of external parallel sentences together with the comparable sentences (In our experiment, we used chemistry domain data of the seed parallel corpus, containing 11k sentences).

We also compared IBM Model 1 (labeled "IBM"), LLR and SampLEX lexicon for the lexicon-based filter. All lexicons were extracted from the seed parallel corpus.

Table 5.1 and 5.2 show the results for fragment extraction using "GIZA++" and "Nakazawa+" for parallel fragment candidate detection respectively. We can see that the average size of the fragments (i.e., the number of words in the fragments) extracted by "Munteanu+, 2006" [94] is unusually long, which is also reported in [109]. Our proposed method extracts shorter fragments. The number of extracted fragments and the average size are similar among the three lexicons when using the same alignment setting. Using the external parallel data for alignment extracts more fragments than only using the comparable sentences, and the average size is slightly larger. We think the reason is that the external parallel data is helpful to improve the recall of alignment for the parallel fragments in the comparable sentences, thus more parallel fragments will be detected. Compared

Method	# fragments	Average size (zh/ja)
Munteanu+, 2006	28.4k	20.36/21.39
Only (IBM)	13.8k	3.85/4.13
Only (LLR)	13.3k	3.87/4.12
Only (SampLEX)	13.5k	3.81/4.06
External (IBM)	16.8k	3.87/4.13
External (LLR)	16.0k	3.88/4.13
External (SampLEX)	16.4k	3.84/4.09

Table 5.2: Fragment extraction results on quasi-comparable corpora using "Nakazawa+" for parallel fragment candidate detection.

to "GIZA++," "Nakazawa+" produces shorter and less fragments. We think the reason for this is that the monotonic and non-null constraints used for parallel fragment candidate detection are much harder for a subtree alignment model to satisfy, thus shorter and less fragment candidates are detected.

To evaluate accuracy, we randomly selected 100 fragments extracted by different methods using "GIZA++" for parallel fragment candidate detection.[12] We manually evaluated the accuracy based on the number of exact matches. Note that the exact match criterion has a bias against "Munteanu+, 2006" [94], because their method extracts sub-sentential fragments that are quite long. We found that only one of the fragments extracted by "Munteanu+, 2006" was exact match, while for the remainder only partial matches are contained in long fragments. The accuracy of our proposed method is over 80%, while the remainder 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 accurate than the IBM Model 1 lexicon. Also, the LLR lexicon performs slightly better than the SampLEX lexicon in this experiment. The accuracy of only using the comparable sentences for alignment are better than using the external par-

[12] A more reliable way to evaluate the accuracy might be creating a test set, and evaluating the precision, recall and F-measure like [57], however, we leave it as future work.

allel data, except for the IBM Model 1 lexicon. We think the reason is that the external parallel data may have a bad effect on the precision of alignment for the parallel fragments in the comparable sentences.

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 bilingual lexicon, and score smoothing also can lead to some failures. Moreover, some filtering failures occur because of both reasons. Table 5.3 shows examples of some fragment pairs extracted by our proposed method of "Only (LLR)" using "GIZA++" for parallel fragment candidate detection. In example 5 and 6, the noisy parts "了 (a past tense marker)" and "を (a case particle)," and "扫描 (scanning)" and "型 (type)" are extracted because they are incorrect aligned by the alignment model and they exist in the bilingual lexicon. In example 7, "粉末 (powder)" and "X線 (x-ray)" is incorrectly aligned, but they do not exist in the bilingual 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, "证明 (prove)" and "から (from)" is a noisy bilingual lexicon pair and incorrectly aligned. Furthermore, "了 (past tense marker)" and "本 (this)" are also incorrectly aligned, but they do not exist in the bilingual 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 bilingual lexicon is the key to improving the lexicon-based filter because the effectiveness of smoothing also highly depends on the accuracy of the bilingual lexicon. Further cleaning the noisy translation pairs is a possible way to achieve this [6], however, we leave it as future work.

ID	Zh fragment	Ja fragment
1	直接甲醇燃料电池 (Direct methanol fuel cell)	直接メタノール燃料電池 (Direct methanol fuel cell)
2	X射线光电子能谱（ＸＰＳ）(X-ray photoelectron spectroscopy (XPS))	X線光電子分光法（ＸＰＳ）(X-ray photoelectron spectroscopy (XPS))
3	（ＯＨ）２４（Ｈ２Ｏ）１２］	（ＯＨ）２４（Ｈ２Ｏ）１２］
4	的原生质体融合 (protoplast fusion of)	のプロトプラスト融合 (protoplast fusion of)
5	分子动力学（ＭＤ）模拟了 (molecular dynamics (MD) simulated)	分子動力学（ＭＤ）シミュレーション を (molecular dynamics (MD) simulation)
6	扫描电子显微镜（ＳＥＭ、） 透射电子显微镜（ＴＥＭ） (scanning electron microscopy (SEM), transmission electron microscopy (TEM))	型電子顕微鏡（ＳＥＭ）， 透過型電子顕微鏡（ＴＥＭ） (type electron microscopy (SEM), transmission electron microscopy (TEM))
7	X射线粉末衍射 (X-ray powder diffraction)	X線回折 分析 (X-ray diffraction analysis)
8	证明了本算法的 (proved the algorithm)	から本アルゴリズムの (from the algorithm)

Table 5.3: Examples of some fragment pairs extracted by our proposed method of “Only (LLR)” from quasi-comparable corpora using “GIZA++” for parallel fragment candidate detection (noisy parts are underlined).

System	GIZA++	Nakazawa+
Baseline	38.64	
+Sentences	39.16	
+Munteanu+, 2006	38.87	
+Only (IBM)	38.86	38.96
+Only (LLR)	$39.27^{\dagger}$	39.17
+Only (SampLEX)	$39.28^{\dagger}$	39.28
+External (IBM)	$39.63^{\ddagger *}$	$\mathbf{39.88}^{\ddagger *+}$
+External (LLR)	39.22	$39.35^{\dagger}$
+External (SampLEX)	$39.40^{\dagger}$	$39.42^{\dagger}$

Table 5.4: BLEU-4 scores for Chinese-to-Japanese translation experiments ("†" and "‡" denote the result is better than "Baseline" significantly at $p < 0.05$ and $p < 0.01$ respectively, "∗" and "+" denotes the result is significantly better than "+Munteanu+, 2006" and "+Sentences" respectively at $p < 0.05$).

Translation Experiments

We conducted Chinese-to-Japanese translation experiments by appending the extracted fragments to a baseline system. For comparison, we also conducted translation experiments by appending the extracted comparable sentences (labeled "+Sentences"). For decoding, we used the state-of-the-art phrase-based SMT toolkit Moses [72] with default options, except for the distortion limit (6→20). The baseline system used the seed parallel corpus (680k sentences). We used another 368 and 367 sentences from the chemistry domain for tuning and testing respectively. We trained a 5-gram language model on the Japanese side of the parallel seed corpus using the SRILM toolkit [122][13] with interpolated Kneser-Ney discounting. Tuning was performed by minimum error rate training (MERT) [98], and it was re-run for every experiment.

We report the translation results on the test set using BLEU-4 [102]. Table 5.4 shows the results of the Chinese-to-Japanese translation experiments. The significance test was performed using the bootstrap resampling method proposed

[13]http://www.speech.sri.com/projects/srilm

by Koehn [69]. We can see that appending the extracted comparable sentences have a positive effect on translation quality. Adding the fragments extracted by "Munteanu+, 2006" [94] has a negative impact, compared to appending the sentences. Our proposed method outperforms both the Baseline, +Sentences, and Munteanu+, 2006 methods, indicating the effectiveness of our proposed method for extracting useful parallel fragments for SMT.

We compared the phrase tables produced by different methods to investigate the reason for different the SMT performances. We found that all methods increased the size of the phrase table, meaning that new phrases are acquired from the extracted data. However, the noise contained in the data extracted by the +Sentences and Munteanu+, 2006 methods produce many noisy phrase pairs, which may decrease MT performance. Our proposed method extracts accurate parallel fragments, which lead to correct new phrases. Among all the settings of our proposed method, the +External (IBM) method shows the best performance, no matter which alignment model is used. The reason for this is that it extracts more correct parallel fragments than the other settings, thus more new phrase pairs are produced. Although the GIZA++ method extracts more parallel fragments than the Nakazawa+ method, they show similar MT performance. We think the reason for this is that the fragments extracted by the Nakazawa+ are more accurate than the GIZA++, because the Nakazawa+ performs better than the GIZA++ for word alignment [95].

Surprisingly, the translation performance after appending the fragments extracted by our proposed method only using the comparable sentences for alignment shows comparable results when using LLR and SampLEX lexicon for filtering, compared to the ones using the external parallel data for alignment. We think the reason is that the extracted fragments not only can produce new phrases, but also can improve the quality of phrase pairs extracted from the original parallel corpus. Because the fragments extracted only using the comparable sentences are more accurate than the ones using the external parallel data, they are more helpful to extract good phrase pairs from the original parallel corpus. This result indicates that external parallel data is not indispensable for the alignment model of our proposed method.

Method	# fragments	Average size (zh/ja)	BLEU-4
Baseline			38.64
Only (LLR)	18.3k	4.00/4.14	39.27
Only (LLR) Iteration 1	18.5k	4.03/4.16	**39.68**
Only (LLR) Iteration 2	18.5k	4.03/4.16	39.13
External (IBM)	28.7k	4.18/4.33	39.63
External (IBM) Iteration 1	29.3k	4.21/4.38	39.58
External (IBM) Iteration 2	29.5k	4.22/4.38	39.39

Table 5.5: Bootstrapping fragment extraction and translation results.

Bootstrapping Experiments

We further conducted bootstrapping experiments. We first appended the extracted parallel fragments to the seed parallel corpus. Then we generated new bilingual lexicons from the combined corpus. Finally, we used the new bilingual lexicons for the lexicon-based filter to extract parallel fragments. Bootstrapping experiments were conducted based on the most accurate method Only (LLR) and the method +External (IBM) that shows the best MT performance. We only conducted bootstrapping experiments for the methods that use the GIZA++ for parallel fragment candidate detection. Experimental settings were the same as the ones used in the extraction and translation experiments. We iterated until there were no further improvements in MT performance on the tuning set.

Bootstrapping fragment extraction and translation results using different methods are shown in Table 5.5. "Baseline" is the same one described in Section 5.3.1 that uses the seed parallel corpus as training data. "Iteration" denotes different iterations using our bootstrapping method. The number after "Iteration" denotes iteration number.

We can see that by bootstrapping, both the number of the extracted fragments and their average size slightly increase for both of the two methods. We think the main reason for this is the quality improvement of the generated bilingual lexicons by bootstrapping. The extracted fragments not only can produce new bilingual lexicons, but also can improve the quality of bilingual lexicons generated

from the seed parallel corpus. Because the amount of the extracted fragments are relatively small compared to the seed parallel corpus, the extracted fragments only can lead to a small increase of the number of the bilingual lexicons. However, the quality of the generated bilingual lexicons can be improved to some extent, which leads to more and longer fragments being extracted. For the Only (LLR) method, the MT performance is slightly improved by bootstrapping. However, no MT improvement is shown for the External (IBM) method. We think the reason is that the fragments extracted by the Only (LLR) method are more accurate the ones extracted by the External (IBM) method. Therefore, they are more helpful to improve the quality of the bilingual lexicons, which leads better parallel fragments that can improve the MT performance.

5.3.2 Experiments on Wikipedia

In this section, we describe the parallel sentence and fragment integrated extraction and translation experiments conducted on the Chinese-Japanese Wikipedia data. Experiments were conducted based on the results described in Section 4.4.4.

Data

We treated the sentence pairs with "$0.1 \leq$ classification probability < 0.9" described in Section 4.4.4 as comparable sentences,[14] 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.[15] The SMT system trained on these parallel sentences is treated as the baseline system in Section 5.3.1.

Extraction Experiments

Based on the investigation of different settings for our proposed method in Section 5.3.1, we chose the following settings for the extraction experiments:

- Parallel fragment candidate detection: We applied word alignment using

[14] 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.

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

Method	# fragments	# fragments w/o CCC	Avg size (zh/ja)	Accuracy
Munteanu+, 2006	153,919		16.76/17.70	(6%)
IBM	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%

Table 5.6: Parallel fragment extraction results on Wikipedia (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 5.2.4).

GIZA++ on the comparable sentences together with the parallel sentences described in Section 5.3.2.

- Lexicon-based filter: We compared the IBM Model 1, LLR, and SampLEX lexicons, which were all generated from a combined parallel corpus that appends the parallel sentences described in Section 5.3.2 to the seed parallel corpus described in Section 5.3.1.

In this experiment, we also investigated the effectiveness of using common Chinese characters for the lexicon-based filter.

The fragment extraction results are shown in Table 5.6. We can see that in general the results are similar to the ones reported in Section 5.3.1. Our proposed method extracts shorter fragments than [94]. The accuracy of our proposed method is significantly better than that of [94], and LLR and SampLEX outperform the IBM Model 1 lexicon. One difference is that the IBM model 1 and LLR lexicons extract significantly more fragments than SampLEX on the Wikipedia data, and the average size is slightly larger. We suspect the reason for this might be that the SampLEX algorithm [137] does not perform well on the combined corpus, and thus the generated lexicon is much smaller 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).

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 .)

Table 5.7: Examples of some fragment pairs extracted by our proposed method from Wikipedia using LLR lexicon for the lexicon-based filter (noisy parts are underlined).

We also analyzed the noisy fragment pairs extracted by our proposed method on the Wikipedia data, and found that these noisy pairs are extracted because of the same reasons as we discussed in Section 5.3.1. Table 5.7 shows examples of fragment pairs extracted by our proposed method using LLR lexicon for the lexicon-based filter on the Wikipedia data.

Translation Experiments

We conducted Chinese-to-Japanese parallel sentence and fragment integrated translation experiments by appending the extracted fragments to a baseline system.

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

Table 5.8: Parallel sentence and fragment integrated translation results ("†", "‡" and "*" denote the result is significantly better than "+Munteanu+, 2006", "+Sentences" and "Baseline" respectively at $p < 0.05$).

The baseline system used the parallel sentences described in Section 5.3.2 as SMT training data. The other settings were the same as the ones used in the translation experiments described in Section 4.4.3.

We report the translation results on the test set using BLEU-4 [102]. The results of the Chinese-to-Japanese translation experiments are shown in Table 5.8. 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 "+Sentences"). The significance test was performed using the bootstrap resampling method proposed by Koehn [69]. The translation results are similar to the ones reported in Section 5.3.1. Appending the extracted comparable sentences and fragments extracted by [94] has a negative impact on translation quality. Our proposed method outperforms the Baseline, +Sentences, and Munteanu+, 2006 methods, indicating the effectiveness of our proposed integrated extraction method and our proposed method for extracting useful parallel fragments for SMT. Different from the results in Section 5.3.1, the LLR lexicon shows the best performance on the Wikipedia data. We suspect the reason for this is that it extracts significantly more accurate fragments than IBM model 1 and extracts both more and larger parallel fragments than SampLEX.

5.4 Summary of This Chapter

In this chapter, we proposed an accurate parallel fragment extraction system using alignment model together with bilingual lexicon. Experiments conducted on both Chinese-Japanese quasi-comparable corpora and Wikipedia showed that our proposed method significantly outperforms a state-of-the-art approach and improves MT performance.

Our system can be improved in several aspects. Firstly, we only used alignment models designed for parallel sentences to detect parallel fragment candidates, alignment models such as the ones proposed by Quirk et al. [109] that are designed for comparable sentences could be more effective. Secondly, although we used some state-of-the-art bilingual lexicons for the lexicon-based filter, there is still some noise and we plan to develop a more accurate bilingual lexicon extraction method. Thirdly, currently our proposed method cannot deal with ordering, an alignment model that is effective for ordering even on comparable sentences should be developed. Fourthly, currently our system only can extract fragments consisting of word sequences, we plan to extend the system to extract fragments including syntax subtrees, which are also very useful for SMT. Finally, although our proposed method is designed to be language and domain independent, the effectiveness for other language pairs and domains needs to be verified.

Chapter 6

Improving SMT Accuracy Using Bilingual Lexicon Extraction with Paraphrases

In statistical machine translation (SMT) [17, 100, 71], the translation model is automatically learned form parallel corpora in an unsupervised way. The translation model contains translation pairs with their features scores. SMT suffers from the *accuracy problem* that the translation model may be inaccurate, meaning that the translation pairs and their features scores may be inaccurate. The *accuracy problem* is caused by the quality of the unsupervised method used for translation model learning, which always correlates with the amount of parallel corpora. Increasing the amount of parallel corpora is a possible way to improve the accuracy, however parallel corpora remain a scarce resource for most language pairs and domains.[1] Accuracy also can be improved by filtering out the noisy translation pairs from the translation model, however meanwhile we may lose some good translation pairs, thus the coverage of the translation model may decrease. A good solution to improve the accuracy while keeping the coverage is estimating new features for the translation pairs from comparable corpora (which

[1]Scarceness of parallel corpora also leads to the low coverage of the translation model (which we call the *coverage problem* of SMT), however we do not tackle this in this chapter.

we call comparable features), to make the translation model more discriminative thus more accurate.

Previous studies use bilingual lexicon extraction (BLE) technology to estimate comparable features [67, 60]. They extend traditional BLE that estimates similarity for bilingual word pairs on comparable corpora, to translation pairs in the translation model of SMT. The similarity scores of the translation pairs are used as comparable features. These comparable features are combined with the original features used in SMT, which can provide additional information to distinguish good and bad translation pairs. A major problem of previous studies is that they do not deal with the data sparseness problem that BLE suffers from. BLE uses vector representations for word pairs to compare the similarity between them. Data sparseness makes the vector representations sparse (e.g., the vector of a low frequent word tends to have many zero entries), thus they do not always reliably represent the meanings of words. Therefore, the similarity of word pairs can be inaccurate. Smoothing technology has been proposed to address the data sparseness problem for BLE. Pekar et al. [103] smoothed the vectors of words with their distributional nearest neighbors, however distributional nearest neighbors can have different meanings and thus introduce noise. Andrade et al. [9] used synonym sets in WordNet to smooth the vectors of words, however WordNet is not available for every language. More importantly, both studies work for words, which are not suitable for comparable feature estimation. The reason is that translation pairs can also be phrases [74] or syntactic rules [45], depending on what kind of SMT models we use.

In this chapter, we propose using paraphrases to address the data sparseness problem of BLE for comparable feature estimation. A paraphrase is a restatement of the meaning of a word, phrase or syntactic rule, therefore it is suitable for the data sparseness problem. We generate paraphrases from the parallel corpus used for translation model learning. Then, we use the paraphrases to smooth the vectors of the translation pairs in the translation model for comparable feature estimation. Smoothing is done by learning vectors that combine the vectors of the original translation pairs with the vectors of their paraphrases. The smoothed vectors can overcome the data sparseness problem, making the vectors more accu-

rately represent the meanings of the translation pairs. In this way, we improve the quality of comparable features, which can improve the accuracy of the translation model thus improve SMT performance.

We conduct experiments on Chinese-English Phrase-based SMT (PBSMT) [74].[2] Experimental results show that our proposed method can improve SMT performance, compared to the previous studies that estimate comparable features without dealing with the data sparseness problem of BLE [67, 60]. The results verify the effectiveness of using BLE together with paraphrases for the *accuracy problem* of SMT.

6.1 Related Work

6.1.1 Bilingual Lexicon Extraction for SMT

From the pioneering work of [110], BLE from comparable corpora has been studied for a long time. BLE is based on the distributional hypothesis [54], stating that words with similar meaning have similar distributions across languages. Contextual similarity [110], topical similarity [135] and temporal similarity [68] can be important clues for BLE. Orthographic similarity may also be used for BLE for some similar language pairs [73]. Moreover, some studies try to use the combinations of different similarities for BLE [61, 30]. To address the data sparseness problem of BLE, smoothing technology has been proposed [103, 9].

BLE can be used to address the *accuracy problem* of SMT, which estimates comparable features for the translation pairs in the translation model [67]. BLE also can be used to address the *coverage problem* of SMT, which mines translations for the unknown words or phrases in the translation model from comparable corpora [34, 63]. Moreover, studies have been conducted to address the *accuracy and coverage problems* of SMT simultaneously with BLE [60].

Our study focuses on addressing the *accuracy problem* of SMT with BLE. We use paraphrases to address the data sparseness problem of BLE for comparable feature estimation, which makes the comparable features more accurate.

[2]Our proposed method can also be applied to other language pairs and SMT models.

6.1.2 Paraphrases for SMT

Many methods have been proposed to use paraphrases for SMT, mainly for the *coverage problem*. One method is paraphrasing unknown words or phrases in the translation model [18, 113, 87]. Another method is constructing a paraphrase lattice for the tuning and testing data, and performing lattice decoding [37, 12]. Paraphrases also can be incorporated as additional training data, which may improve both coverage and accuracy of SMT [101].

Previous studies require external data in addition to the parallel corpus used for SMT for paraphrase generation to make their methods effective. These paraphrases can be generated from external parallel corpora [18, 37], or monolingual corpora based on distributional similarity [87, 113, 101, 12].

Our study differs from previous studies in using paraphrases for smoothing the vectors of BLE, which is used for comparable feature estimation that can improve the accuracy of SMT. Another difference is that our proposed method is effective when only using the paraphrases generated from the parallel corpus used for SMT, while previous studies require external data for paraphrase generation.

6.2 Accuracy Problem of Phrase-based SMT

In this study, we conduct experiments on PBSMT [74]. Here, we give a brief overview of PBSMT, and explain the *accuracy problem* of PBSMT.

In PBSMT, the translation model is represented as a phrase table, containing phrase pairs together with their feature scores.[3] The phrase pairs are extracted based on unsupervised word alignments, whose quality always correlates with the amount of the parallel corpus. Inverse and direct phrase translation probabilities $\phi(f|e)$ and $\phi(e|f)$, inverse and direct lexical weighting $lex(f|e)$ and $lex(e|f)$ are used as features for the phrase table. Phrase translation probabilities are calculated via maximum likelihood estimation, which counts how often a source phrase f is aligned to target phrase e in the parallel corpus, and vise versa. Lexical weighting is the average word translation probability calculated using internal word alignments of a phrase pair, which is used to smooth the overestimation of

[3]Note that in PBSMT, the definition of a phrase also includes a single word.

f	e	$\phi(f\|e)$	$lex(f\|e)$	$\phi(e\|f)$	$lex(e\|f)$	Alignment
失业 人数	**unemployment figures**	0.3	0.0037	0.0769	0.0018	0-0 1-1
失业 人数	**number of unemployed**	0.1333	0.0188	0.1025	0.0041	1-0 1-1 0-2
失业 人数	. unemployment was	0.3333	0.0015	0.0256	6.8e-06	0-1 1-1 1-2
失业 人数	unemployment and bringing	1	0.0029	0.0256	5.4e-07	0-0 1-0

Table 6.1: Example of the *accuracy problem* in PBSMT (The correct translations are in bold).

the phrase translation probabilities. Other typical features such as the reordering model features and the n-gram language model features are also used in PBSMT. These features are combined in a log linear model, and their weights are tuned using a small size of parallel sentences. During decoding, these features together with their tuned weights are used to produce new translations.

One problem of PBSMT is that the phrase pairs and their feature scores in the phrase table may be inaccurate. One reason for this is the quality of the word alignment. Another reason is that the translation probabilities of rare word and phrase pairs tend to be grossly overestimated. Sparseness of the parallel corpus leads to word alignment errors and overestimations, which result in inaccurate phrase pairs and feature scores. Table 6.1 shows an example of phrase pairs and feature scores taken from the phrase table constructed in our experiments (See Section 6.4 for the details of the experiments), which contains inaccurate phrase pairs. The correct translations of "失业 (unemployment) 人数 (number of people)" are in bold. The incorrect phrase pairs are extracted because "人数 (number of people)" is incorrectly aligned to "unemployment," and their feature scores are incorrect. We cannot simply filter out these incorrect phrase pairs, because we may lose some good phrase pairs, thus the coverage of the phrase table may decrease.

6.3 Proposed Method

Figure 6.1 shows an overview of our proposed method. We construct a phrase table from a parallel corpus following [74]. Because this phrase table may be

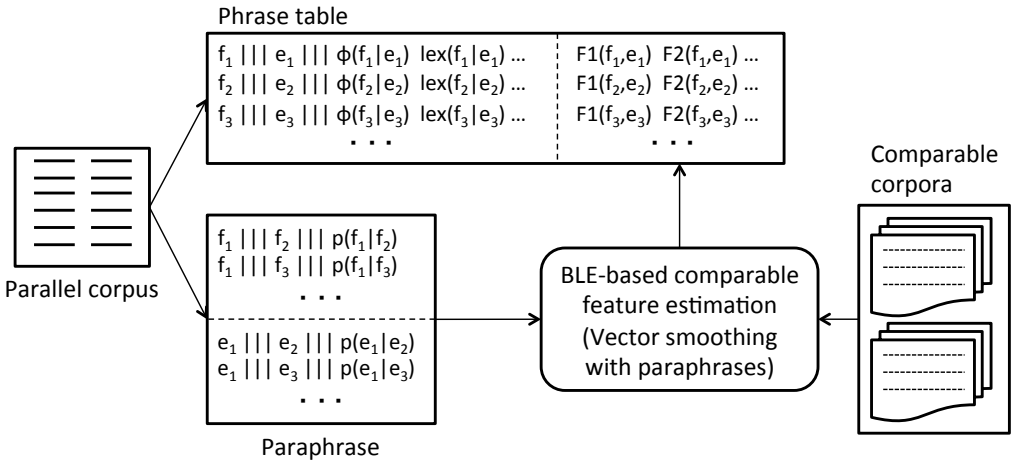

Figure 6.1: Overview of our proposed method.

inaccurate, we estimate comparable features from comparable corpora following [67, 60]. These comparable features are appended to the original phrase table, to address the *accuracy problem* of PBSMT. Comparable feature estimation is based on BLE, which suffers from the data sparseness problem. We propose using paraphrases to address this problem. We generate phrasal level paraphrases for both the source and target language from the parallel corpus. Then we use the paraphrases to smooth the vectors of the source and target phrases used for comparable feature estimation respectively. Smoothing is done by learning a vector that combines the original vector of a phrase with the vectors of its paraphrases. The smoothed vectors can represent the meanings of phrase pairs more accurately. Finally, we compute the similarity of phrase pairs based on the smoothed source and target vectors. In this way, we improve the quality of comparable features, which can improve the accuracy of the phrase table thus improve SMT performance.

Details of paraphrase generation, comparable feature estimation and vector smoothing with paraphrases will be described in Section 6.3.1, 6.3.2 and 6.3.3 respectively.

6.3.1 Paraphrase Generation

In this study, we generate both source and target phrasal level paraphrases from the parallel corpus used for SMT[4] through bilingual pivoting [11]. The idea of this method is that if two source phrases f_1 and f_2 are translated to the same target phrase e, we can assume that f_1 and f_2 are a paraphrase pair. Probability of this paraphrase pair can be assigned by marginalizing over all shared target translations e in the parallel corpus, defined as follows:

$$p(f_1|f_2) = \sum_e \phi(f_1|e)\phi(e|f_2) \tag{6.1}$$

where, $\phi(f_1|e)$ and $\phi(e|f_2)$ are phrase translation probability. Target paraphrases can be generated in a similar way.

Note that word alignment errors can also lead to incorrect paraphrase generation. For example, "unemployment figures" and "unemployment and bringing" in Table 6.1 might be generated as a paraphrase pair. However, this kind of noisy pairs can be easily pruned according to their low probabilities.

6.3.2 Comparable Feature Estimation

Following [67, 60], we estimate contextual, topical and temporal similarities as comparable features. However, we do not use orthographic similarity as comparable feature, because we experiment on Chinese-English, which is not an orthographically similar language pair.

Besides phrasal features, we also estimate lexical features following [67, 60]. The lexical features are the average similarity scores of word pairs over all possible word alignments across two phrases. They are used to smooth the phrasal features, like the lexical weighting in PBSMT. However, they only can slightly alleviate the sparseness of phrasal features, because individual words also suffer from the data sparseness problem.

In the following sections, we describe the methods to estimate contextual, topical and temporal features in detail.

[4]Paraphrases also can be generated from external parallel corpora and monolingual corpora, however we leave it as future work.

Contextual feature

Contextual feature is the contextual similarity of a phrase pair. Contextual similarity is based on the distributional hypothesis on context, stating that phrases with similar meaning appear in similar contexts across languages. From the pioneering work of [110], contextual similarity has been used for BLE for a long time.

In the literature, different definitions of context have been proposed for BLE, such as window-based context, sentence-based context and syntax-based context. In this study, we use window-based context, and leave the comparison of using different definitions of context as future work. Given a phrase, we count all its immediate context words, with a window size of 4 (2 preceding words and 2 following words). We build a context by collecting the counts in a bag of words fashion, namely we do not distinguish the positions that the context words appear in. The number of dimensions of the constructed vector is equal to the vocabulary size. We further reweight each component in the vector by multiplying by the *IDF* score following [48, 30], which is defined as follows:

$$IDF(t, D) = \log \frac{|D|}{1 + |\{d \in D : t \in d\}|} \tag{6.2}$$

where $|D|$ is the total number of documents in the corpus, and $|\{d \in D : t \in d\}|$ denotes number of documents where the term t appears.[5] We model the source and target vectors using the method described above, and project the source vector onto the vector space of the target language using a seed dictionary. The contextual similarity of the phrase pair is the similarity of the vectors, which is computed using cosine similarity defined as follows:

$$Cos(f, e) = \frac{\sum_{k=1}^{K} F_k \times E_k}{\sqrt{\sum_{k=1}^{K} (F_k)^2} \times \sqrt{\sum_{k=1}^{K} (E_k)^2}} \tag{6.3}$$

where f and e are the source and target phrases, F and E are the projected source vector and target vector, K is the number of dimensions of the vectors.

[5] Because there are no document bounds in the corpus we used to estimate contextual feature, we treated every 100 sentences as one document.

Topical feature

Topical feature is the topical similarity of a phrase pair. Topical similarity uses the distributional hypothesis on topics, stating that two phrases are potential translation candidates if they are often present in the same cross-lingual topics and not observed in other cross-lingual topics [135]. Vulić et al. [135] proposed using bilingual topic model based method to estimate topical similarity. However, this method is not scalable for large data sets.

In this study, we estimate topical feature in a scalable way following [67]. We treat an article pair aligned by interlanguage links in Wikipedia as a topic aligned pair. For a phrase pair, we build source and target topical occurrence vectors by counting their occurrences in its corresponding language articles. The number of dimensions of the constructed vector is equal to the number of aligned article pairs, and each dimension is the number of times that the phrase appears in the corresponding article. The similarity of the phrase pair is computed as the similarity of the source and target vectors using cosine similarity (Equation 6.3).

Temporal feature

Temporal feature is the temporal similarity of a phrase pair. The intuition of temporal similarity is that news stories across languages tend to discuss the same world events on the same day, and the occurrences of a translated phrase pair over time tend to spike on the same dates [68, 67].

We estimate temporal feature following [68, 67]. For a phrase pair, we build source and target temporal occurrence vectors by counting their occurrences in equally sized temporal bins, which are sorted from the set of time-stamped documents in the comparable corpus. We set the window size of a bin to 1 day. Therefore the number of dimensions of the constructed vector is equal to the number of days spanned by the corpus, and each dimension is the number of times that the phrase appears in the corresponding bin. The similarity of the phrase pair is computed as the similarity of the source and target vectors using cosine similarity (Equation 6.3).

Phrase	Paraphrase
tampered	being tampered
an appropriation	appropriation
11th	11th .
so many years	many years
first thing	first thing that
mass media ,	media ,

Table 6.2: Examples of overlaps between a phrase and its paraphrase.

6.3.3 Vector Smoothing with Paraphrases

Data sparseness results in sparse representations of the vectors, therefore the similarity of the phrase pair can be inaccurate. We propose using paraphrases to smooth both the source and target vectors, to deal with the data sparseness problem. After smoothing, the vectors can more accurately represent the phrases. We compute the similarity of the phrase pair based on the smoothed source and target vectors, and use it as comparable features for PBSMT.

One problem of using paraphrases for smoothing is that a phrase and its paraphrase may overlap. Table 6.2 shows some examples of overlaps between a phrase and its paraphrase generated from the parallel corpus we use. The vector of the overlapped paraphrase contains overlapped information of the vector of the original phrase. Therefore, it is necessary to consider overlap when using paraphrases for vector smoothing.

There are three types of vectors (context, topical and temporal occurrence vectors) need to be smoothed. The method for smoothing context vector is different from topical and temporal occurrence vectors, because the components in context vector are different. Topical and temporal occurrence vectors can be smoothed using the same method, because the components of both vectors are occurrence information. The following sections describe the methods to smooth the context vector, and topical and temporal occurrence vectors respectively.

Context Vector Smoothing

We smooth the context vector of a phrase x with the following equation:

$$X' = \frac{f(x)}{f(x)+\sum_{j=1}^{n} f(x_j)} \cdot X + \sum_{i=1}^{n} \frac{f(x_i)}{f(x)+\sum_{j=1}^{n} f(x_j)} \cdot p(x_i|x) \cdot \begin{cases} X_i \backslash X & (x \subset x_i) \\ X_i - X & (x \supset x_i) \\ X_i & (otherwise) \end{cases} \tag{6.4}$$

where X' is the smoothed context vector, X is the context vector of x, n is the number of paraphrases that x has, X_i is the context vector of paraphrase x_i, $p(x_i|x)$ is the probability that x_i is a paraphrase of x. $f(x)$ is the frequency of x in the corpus, and $\frac{f(x)}{f(x)+\sum_{j=1}^{n} f(x_j)}$ is the frequency weight for x. Frequency weight is also used for the paraphrases in a similar way. The frequency weight is proposed by Andrade et al. [9] when using synonyms to smooth the context vector of a word. They show that using the frequency information of words as weights performs better than simple summation of the vectors. For the overlap problem between x and x_i, we do the following:

- If $x \subset x_i$ namely x is contained in x_i, we use the context words that exist in X_i but do not exist in X for smoothing, which is $X_i \backslash X$;
- If $x \supset x_i$ namely x contains x_i, we remove the overlapped contextual information between X_i and X for smoothing, which is $X_i - X$;
- Otherwise, we use X_i for smoothing.

Topical and Temporal Occurrence Vectors Smoothing

We smooth the topical and temporal occurrence vectors of a phrase x with the following equation:

$$X' = X + \sum_{i=1}^{n} p(x_i|x) \cdot \begin{cases} 0 & (x \subset x_i) \\ X_i - X & (x \supset x_i) \\ X_i & (otherwise) \end{cases} \tag{6.5}$$

where X' is the smoothed occurrence vector, X is the occurrence vector of x, n is the number of paraphrases that x has, X_i is the occurrence vector of paraphrase

Context (before smoothing)	<rising: 2.37, economic: 0, recession: 3.94 $\cdots$ >
Context (after smoothing)	<rising: 0.03, economic: 0.06, recession: 0.04 $\cdots$ >
Topical (before smoothing)	<Topic1: 0, Topic2: 1, Topic3: 0 $\cdots$ >
Topical (after smoothing)	<Topic1: 0.12, Topic2: 1.27, Topic3: 0.05 $\cdots$ >
Temporal (before smoothing)	<Date1: 1, Date2: 0, Date3: 6 $\cdots$ >
Temporal (after smoothing)	<Date1: 1.25, Date2: 0.08, Date3: 6.38 $\cdots$ >

Table 6.3: Examples of the three types of vectors for the phrase "unemployment figures" before and after smoothing.

x_i, $p(x_i|x)$ is the probability that x_i is a paraphrase of x. For the overlap problem between x and x_i, we do the following:

- If $x \subset x_i$, we do not use X_i for smoothing, because X already contains the occurrence information in X_i;
- If $x \supset x_i$, we remove the overlapped occurrence information between X_i and X for smoothing, which is $X_i - X$;
- Otherwise, we use X_i for smoothing.

Examples of the three types of vectors before and after smoothing are shown in Table 6.3.

6.4 Experiments

In our experiments, we compared our proposed method with [67]. We estimated comparable features from comparable corpora using the method of [67] and our proposed method respectively. We appended the comparable features to the phrase table, and evaluated the two methods in the perspective of SMT performance. We conducted experiments on Chinese-English data. In all our experiments, we preprocessed the data by segmenting Chinese sentences using a segmenter proposed by Chu et al. [25], and tokenizing English sentences.

	NIST	Gigaword	Wikipedia
# Zh articles	N/A	3.6M	248k
# En articles	N/A	4.3M	248k
# Zh sentences	991k	42.6M	2.8M
# En sentences	991k	56.9M	10.1M
# Zh tokens	26.1M	1.1B	70.5M
# En tokens	27.2M	1.3B	240.5M

Table 6.4: Statistics of the comparable data used for comparable feature estimation.

6.4.1 Experimental Settings

SMT Settings

We conducted Chinese-to-English translation experiments. The parallel corpus we used is from Chinese-English NIST open MT.[6] The "NIST" column of Table 6.4 shows the statistics of this parallel corpus. For decoding, we used the state-of-the-art PBSMT toolkit Moses [72] with default options, except for the phrase length limit (7→3) following [67]. We trained a 5-gram language model on the English side of the parallel corpus using the SRILM toolkit [122][7] with interpolated Kneser-Ney discounting, and used it for all the experiments. We used NIST open MT 2002 and 2003 data sets for tuning and testing, containing 878 and 919 sentence pairs respectively. Note that both MT 2002 and 2003 data sets contain 4 references for each Chinese sentence. Tuning was performed by minimum error rate training (MERT) [98], and it was re-run for every experiment.

Comparable Feature Estimation Settings

Table 6.4 shows the statistics of the comparable data used for comparable feature estimation. The contextual feature was estimated on the parallel corpus. We

[6]LDC2007T02, LDC2002T01, LDC2003T17, LDC2004T07, HK News part of LDC2004T08, LDC2005T10 and LDC2006T04

[7]http://www.speech.sri.com/projects/srilm

# Phrase pairs	4,886,067
# Zh phrases	45,905
# En phrases	2,078,230
# Zh unigrams	6,719
Avg # translations	509.1
# Zh bigrams	23,029
Avg # translations	56.7
# Zh trigrams	16,157
Avg # translations	9.8

Table 6.5: Statistics of the filtered phrase table.

treated the two sides of the parallel corpus as independent monolingual corpora, following [52, 67]. Contextual feature estimation requires a seed dictionary. The seed dictionary we used is NIST Chinese-English translation lexicon Version 3.0,[8] containing 82k entries. The temporal feature was estimated on Chinese[9] and English[10] Gigaword version 5.0. We used the afp, cna and xin sections with date range 1994/05-2010/12 of the corpora. The topical feature was estimated on Chinese and English Wikipedia data. We downloaded Chinese[11] (2012/09/21) and English[12] (2012/10/01) Wikipedia database dumps. We used an open-source Python script[13] to extract and clean the text from the dumps. We aligned the articles on the same topic in Chinese-English Wikipedia via the interlanguage links.

We estimated comparable features for the unique phrase pairs used for tuning and testing. These phrase pairs were extracted from the entire phrase table constructed from the parallel corpus, by checking all the source phrases in the tuning and testing data sets. We call these phrase pairs the filtered phrase table. Table

[8] LDC2002L27
[9] LDC2011T13
[10] LDC2011T07
[11] http://dumps.wikimedia.org/zhwiki
[12] http://dumps.wikimedia.org/enwiki
[13] http://code.google.com/p/recommend-2011/source/browse/Ass4/WikiExtractor.py

	Zh	En
# Phrases&words	46,112	2,090,345
# Phrases&words w/ paraphrases	26,718	455,099
# Unigrams w/ paraphrases	6,273	46,191
Avg # paraphrases	39.8	21.6
# Bigrams w/ paraphrases	15,026	223,299
Avg # paraphrases	34.6	17.7
# Trigrams w/ paraphrases	5,419	185,609
Avg # paraphrases	20.0	14.9

Table 6.6: Statistics the generated paraphrases for the phrases and individual words inside the phrases in the filtered phrase table.

6.5 shows the statistics of the filtered phrase table. We can see that each Chinese phrase has a large number of translations on average especially for the lower order n-gram phrases, which can indicate the inaccuracy of the filtered phrase table.

Our proposed method requires paraphrases for vector smoothing. We used Joshua [47] to generate both Chinese and English paraphrases from the parallel corpus. We kept the paraphrase pairs that satisfy $\log p(x_1|x_2) > -7$ and $\log p(x_2|x_1) > -7$ [14] for smoothing, where $p(x_1|x_2)$ is the probability that x_1 is a paraphrase of x_2, and $p(x_2|x_1)$ is the probability that x_2 is a paraphrase of x_1. Table 6.6 shows the statistics of the paraphrase generation results for the Chinese and English phrases, and individual words inside the phrases in the filtered phrase table.

Note that, for some phrase pairs, their comparable feature scores may be 0, because of data sparseness. In that case, we set their comparable features to a small positive number of $1e-07$.

6.4.2 Results

We report results on the test set using case-insensitive BLEU-4 score and four references. Table 6.7 shows the results of Chinese-to-English translation exper-

[14] We also tried other pruning thresholds, and this threshold showed the best performance in the preliminary experiments.

System	+Contextual	+Topical	+Temporal	+All
Baseline	45.45			
Klementiev+	43.69	45.72	45.05	45.92
Proposed	45.56‡	46.10†‡	46.00†‡	**46.26**†

Table 6.7: BLEU-4 scores for Chinese-to-English translation experiments ("†" and "‡" denote that the result is significantly better than "Baseline" at $p < 0.01$ and "Klementiev+" at $p < 0.05$ respectively)

iments. "Baseline" denotes the baseline system that does not use comparable features. "Klementiev+" denotes the system that appends the comparable features estimated following [67] to the phrase table. "Proposed" denotes the system that uses the comparable features estimated by our proposed method. "+Contextual," "+Topical" and "+Temporal" denote the systems that append contextual, topical and temporal features respectively. "+All" denotes the system that appends all the three types of features. The significance test was performed using the bootstrap resampling method proposed by Koehn [69].

We can see that "Klementiev+" does not always outperform "Baseline." The reason for this is that the comparable features estimated by [67] are inaccurate. "Proposed" performs significantly better than both "Baseline" and "Klementiev+." The reason for this is that "Proposed" deals with the data sparseness problem of BLE for comparable feature estimation, making the features more accurate thus improve the SMT performance. As for different comparable features of "Proposed," "+Contextual," "+Topical" and "+Temporal" are all helpful, and combining them can be more effective. The results verify the effectiveness of our proposed method for the *accuracy problem* of PBSMT.

We also investigated the comparable features estimated by the method of [67] and our proposed method. Based on our investigation, most comparable features estimated by our proposed method are more accurate than the ones estimated by the method of [67]. Here, we give an example of the comparable feature scores estimated for the phrase pairs shown in Table 6.1. Table 6.8 shows the comparable feature scores estimated by the method of [67] (above the bold line) and our

f	e	*con*	*con_lex*	*top*	*top_lex*	*tem*	*tem_lex*
失业 人数	**unemployment figures**	1.4e-06	0.0408	1e-07	0.2061	0.1942	0.6832
失业 人数	**number of unemployed**	0.0144	0.0299	1e-07	0.1675	0.0236	0.6277
失业 人数	. unemployment was	0.0107	0.0701	1e-07	0.1908	0.0709	0.6981
失业 人数	unemployment and bringing	1e-07	0.0603	1e-07	0.1730	1e-07	0.6898
失业 人数	**unemployment figures**	0.0749	0.0806	0.5434	0.2629	0.4307	0.7033
失业 人数	**number of unemployed**	0.0522	0.1053	0.1907	0.2235	0.5983	0.7240
失业 人数	. unemployment was	0.0050	0.1206	0.0117	0.2336	0.0967	0.7094
失业 人数	unemployment and bringing	5.1e-05	0.0904	1e-07	0.2034	0.0073	0.7003

Table 6.8: Examples of comparable feature scores estimated by the method of [67] (above the bold line) and our proposed method (below the bold line) for the phrase pairs shown in Table 6.1 ("con," "top" and "tem" denote phrasal contextual, topical and temporal features respectively, "con_lex," "top_lex" and "tem_lex" denote lexical contextual, topical and temporal features respectively).

proposed method (below the bold line). We can see that the method of [67] suffers from the data sparseness problem. Many of the feature scores are $1e-07$, and many of the feature scores for the correct translations ("unemployment figures" and "number of unemployed") are lower than the incorrect ones (". unemployment was" and "unemployment and bringing"). Our proposed method addresses the data sparseness problem by using paraphrases for vector smoothing. We can see that, after smoothing the feature scores can more accurately distinguish the good translations from the bad ones.

6.5 Summary of This Chapter

In this chapter, we proposed using BLE together with paraphrases to address the *accuracy problem* of SMT. The translation pairs and their feature scores in the translation model of SMT can be inaccurate, because of the quality of the unsupervised methods used for translation model learning. Estimating comparable features from comparable corpora with BLE has been proposed for the *accuracy*

problem of SMT. However, BLE suffers from the data sparseness problem, which makes the comparable features inaccurate. We proposed using paraphrases to address this problem. Paraphrases were used to smooth the vectors used in comparable feature estimation with BLE. Experiments conducted on Chinese-English PBSMT verified the effectiveness of our proposed method.

As future work, firstly we plan to generate paraphrases from external parallel corpora and monolingual corpora, where as in this study we used the paraphrases generated from the parallel corpus used for SMT. Secondly, in this study we estimated contextual features from the parallel corpus, however in the future we plan to estimate it from comparable corpora. Finally, because our proposed method should be language independent and can be applied to other SMT models, we plan to conduct experiments on other language pairs and SMT models to verify this.

Chapter 7

Conclusion

The scarceness of parallel corpora is the main bottleneck of SMT. In this thesis, we exploited comparable corpora to addressing this. We proposed novel approaches to extract bilingual lexicons, parallel sentences and parallel fragments from comparable corpora in an integrated framework. In addition, we exploited linguistic knowledge of common Chinese characters for Chinese-Japanese parallel data extraction as a case study. Bilingual lexicon extraction (BLE) was used for parallel sentence extraction and addressing the *accuracy problem* of SMT. The extracted parallel sentences and fragments were used as additional training data for SMT. Experiments showed that our proposed approaches are effective for the scarceness of parallel corpora that SMT suffers.

7.1 Summary

In Chapter 2, we proposed a method for constructing a more complete resource of common Chinese characters using freely available resources. In addition, we exploited common Chinese characters in Chinese word segmentation for SMT. Common Chinese characters were used for parallel sentence (Chapter 4) and fragment extraction (Chapter 5). The optimized segmenter was used throughout this thesis work.

In Chapter 3, we presented a BLE system exploiting both topical and contextual knowledge. Our system is based on a novel combination of topic model

and context based methods, which is fully unsupervised and can be iteratively improved. Experiments conducted on Chinese-English, Japanese-English and Chinese-Japanese Wikipedia data verified the effectiveness of our system for BLE from comparable corpora.

In Chapter 4, we presented a robust parallel sentence extraction system consisting of a parallel sentence candidate filter and a binary classifier for parallel sentence identification. We improved the system by using the common Chinese characters for the filter, and three novel feature sets for the classifier. Experiments conducted on Chinese-Japanese Wikipedia showed that our proposed methods are more effective than the previous studies. We further applied the bilingual lexicons extracted in Chapter 3 for parallel sentence extraction.

In Chapter 5, we proposed an accurate parallel fragment extraction system using alignment model together with bilingual lexicon. Common Chinese characters were also used to improve the coverage of the system. Experiments conducted on Chinese-Japanese quasi-comparable corpora and Wikipedia showed that our proposed our system can accurately extract parallel fragments, and the extracted parallel fragments can improve SMT performance.

In Chapter 6, we proposed using BLE together with paraphrases for the *accuracy problem* of SMT. Estimating comparable features from comparable corpora with BLE has been proposed for the *accuracy problem* of SMT. However, BLE suffers from the data sparseness, which makes the comparable features inaccurate. We proposed using paraphrases to addressing this. Experiments conducted on Chinese-English SMT verified the effectiveness of our proposed method.

The main problem of exploiting comparable corpora is that they are noisy, making the extracted parallel data noisy. In this thesis, we proposed many approaches to addressing this problem, and verified the effectiveness of them. In our integrated framework, BLE is the key to addressing the noisy problem, because it is the fundamental part to accurately extract the parallel data of larger unit. How to improve the robustness on noisy data is a common problem in many artificial intelligence research fields. We believe that this thesis work can benefit the research in other fields that also conducts on noisy data such as speech recognition and computer vision.

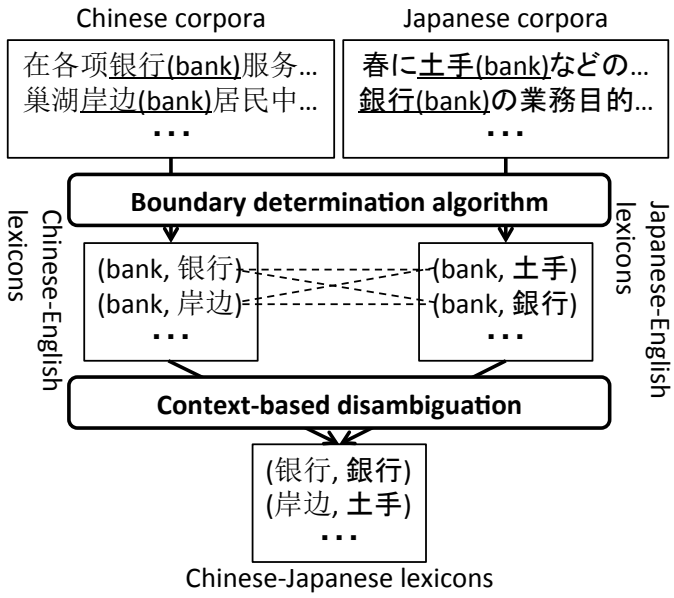

Figure 7.1: Bilingual lexicon extraction from monolingual corpora.

7.2 Future Work

Our proposed approaches have improved the state-of-the-art performance for parallel data extraction from comparable corpora. However, the sizes of the extracted bilingual lexicons, parallel sentences and fragments in this thesis are relatively small, and the language pairs and domains of them are limited. Aiming to extract large-scale parallel data for various language pairs and domains, following directions can be considered.

7.2.1 Bilingual Lexicon Extraction from Monolingual Corpora

Because parallel sentence and fragment extraction systems usually highly rely on bilingual dictionaries, constructing large-scale dictionaries for various language pairs and domains automatically is crucial for making large-scale parallel data extraction available. In Chapter 4, we have already shown the effectiveness of using the bilingual lexicons extracted from comparable corpora for parallel sentence extraction, however the bilingual lexicons used in the experiments are not

large-scale. To automatically construct large-scale dictionaries form comparable corpora, our proposed BLE system described in Chapter 3 needs to be further improved following Section 3.5.

BLE from comparable corpora may have its limitation in domain diversity and coverage. As monolingual corpora are more available than comparable corpora, we consider extracting bilingual lexicons from monolingual corpora. For languages paired with English, we may directly extract bilingual lexicons from monolingual corpora. Taking Chinese-English and Japanese-English as an example. In both Chinese and Japanese corpora, there are many expressions that have English translations in special formats (e.g. parenthetical translation). Using these clues and efficient word boundary determination algorithms such as the ones proposed in [19, 81], both Chinese-English and Japanese-English lexicons can be extracted. For the languages pairs without English, we may use English as a pivot to construct bilingual lexicons. Taking Chinese-Japanese as an example. Once Chinese-English and Japanese-English lexicons are extracted by the above method, we can construct Chinese-Japanese lexicons via English. However, many ambiguous pairs may be produced by pivoting. These ambiguous pairs can be removed using the context of the lexicons [128]. Figure 7.1 shows an example of this process.

We plan to construct large bilingual dictionaries for various domains by combining the bilingual lexicons extracted from comparable and monolingual corpora. By combination, we can further filter out some noisy pairs, and thus improve the precision of the lexicons.

7.2.2 Unsupervised Parallel Data Extraction

The motivation of exploiting comparable corpora for SMT is to addressing the scarceness of parallel corpora. However, most previous studies of parallel data extraction from comparable corpora are supervised or semi-supervised that rely on existing parallel data. BLE from comparable corpora usually relies on a manually created seed dictionary, and parallel sentence and fragment extraction relies on either a manually created seed dictionary or a seed parallel corpus. Obviously, this kind of parallel data is not available for many language pairs and domains.

Unsupervised methods are the key for large-scale parallel data extraction for these language pairs and domains.

Unsupervised parallel data extraction is still a challenging research area. Recently, some unsupervised BLE methods have been proposed, such as the topical model based method [135], the decipherment based method [112, 36] and our proposed method described in Chapter 3. However, challenges for unsupervised BLE still remain such as scalability, compound words, rare words and polysemy. Some of the challenges are in common with the supervised and semi-supervised BLE. For parallel sentence and fragment extraction, few studies have been conducted on unsupervised methods. The only unsupervised parallel sentence extraction method that we are aware is [35]. However their method suffers from high computational complexity. The only unsupervised parallel fragment extraction method that we are aware is [109]. However their method cannot accurately extract parallel fragments. Therefore, we plan to develop more efficient unsupervised methods for parallel data extraction from comparable corpora.

7.2.3 Paraphrases Based Extraction

Parallel data is the equivalent of a word, phrase or sentence in two languages. A paraphrase is a restatement of the meaning of a word, phrase or sentence. Parallel data extraction compares the similarity of a word, phrase or sentence pair bilingually, while paraphrase extraction compares the similarity of a word, phrase or sentence pair monolingually. The tasks of parallel data extraction and paraphrase extraction are highly comparable, and many methods are in common between these two tasks. For example, Wang and Callison-Burch [140] directly applied the method for parallel fragment extraction from comparable corpora proposed in [94], to paraphrase fragment extraction from monolingual comparable corpora. Therefore, it is natural to consider using paraphrases for parallel data extraction.

As described in Section 6.1.2, paraphrases have been used to addressing the *coverage problem* of SMT in many previous studies. In Chapter 6, we also have proposed a method of using paraphrases to addressing the *accuracy problem* of SMT. However, few studies have been conducted on using paraphrases for parallel

data extraction from comparable corpora. The only study that we are aware is [9]. They use lexical level paraphrases namely synonyms to improve the accuracy of BLE from comparable corpora. We believe that the accuracy of parallel sentence and fragment extraction also could be further improved by paraphrases. Moreover, not only the accuracy but also the coverage of parallel data extraction could be improved by paraphrases. Naive ideas such as improving the coverage of the bilingual dictionaries by paraphrasing could be a possible approach for large-scale parallel data extraction. Recently, large paraphrase databases for many languages such as the multilingual paraphrase database [46] have become available to acquire, making it much easier to try paraphrases based large-scale parallel data extraction for various language pairs.

On the other hand, parallel data also can be used for paraphrase extraction. For example, Andrade et al. [8] improved the accuracy of synonym extraction with bilingual lexicons. This direction can be a possible extension for this work.

Bibliography

[1] S. Abdul-Rauf and H. Schwenk. Parallel sentence generation from comparable corpora for improved smt. *Machine Translation*, 25(4):341–375, 2011.

[2] S. F. Adafre and M. de Rijke. Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the workshop on NEW TEXT Wikis and blogs and other dynamic text sources*, pages 62–69, 2006.

[3] H. Afli, L. Barrault, and H. Schwenk. Multimodal comparable corpora as resources for extracting parallel data: Parallel phrases extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 286–292, 2013.

[4] A. Aker, Y. Feng, and R. Gaizauskas. Automatic bilingual phrase extraction from comparable corpora. In *Proceedings of COLING 2012: Posters*, pages 23–32, 2012.

[5] A. Aker, M. Paramita, and R. Gaizauskas. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 402–411, 2013.

[6] A. Aker, M. Paramita, M. Pinnis, and R. Gaizauskas. Bilingual dictionaries for all eu languages. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2839–2845, 2014. ACL Anthology Identifier: L14-1623.

[7] D. Andrade, T. Nasukawa, and J. Tsujii. Robust measurement and comparison of context similarity for finding translation pairs. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 19–27, 2010.

[8] D. Andrade, M. Tsuchida, T. Onishi, and K. Ishikawa. Synonym acquisition using bilingual comparable corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1077–1081, 2013.

[9] D. Andrade, M. Tsuchida, T. Onishi, and K. Ishikawa. Translation acquisition using synonym sets. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 655–660, 2013.

[10] M.-H. Bai, K.-J. Chen, and J. S.Chang. Improving word alignment by adjusting Chinese word segmentation. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 249–256, 2008.

[11] C. Bannard and C. Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, 2005.

[12] K. Bar and N. Dershowitz. Inferring paraphrases for a highly inflected language from a monolingual corpus. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing2014)*, pages 8404:2:245–256, 2014.

[13] R. G. Bharadwaj and V. Varma. Language independent identification of parallel sentences using wikipedia. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 11–12, 2011.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[15] D. Bouamor, A. Popescu, N. Semmar, and P. Zweigenbaum. Building specialized bilingual lexicons using large scale background knowledge. In *Pro-*

ceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 479–489, 2013.

[16] D. Bouamor, N. Semmar, and P. Zweigenbaum. Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 759–764, 2013.

[17] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Association for Computational Linguistics*, 19(2):263–312, 1993.

[18] C. Callison-Burch, P. Koehn, and M. Osborne. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, 2006.

[19] G. Cao, J. Gao, and J.-Y. Nie. A system to mine large-scale bilingual dictionaries from monolingual web pages. In *Proceedings of MT Summit 2007*, pages 57–64, 2007.

[20] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[21] P.-C. Chang, M. Galley, and C. D. Manning. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, 2008.

[22] W. Chen, D. Kawahara, K. Uchimoto, Y. Zhang, and H. Isahara. Dependency parsing with short dependency relation in unlabeled data. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pages 88–94, 2008.

[23] Y.-M. Chou and C.-R. Huang. Hantology: A linguistic resource for Chinese language processing and studying. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 587–590, 2006.

[24] Y.-M. Chou, C.-R. Huang, and J.-F. Hong. The extended architecture of Hantology for kanji. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1693–1696, 2008.

[25] C. Chu, T. Nakazawa, D. Kawahara, and S. Kurohashi. Exploiting shared Chinese characters in Chinese word segmentation optimization for Chinese-Japanese machine translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012)*, pages 35–42, 2012.

[26] C. Chu, T. Nakazawa, D. Kawahara, and S. Kurohashi. Chinese-japanese machine translation exploiting chinese characters. *ACM Transactions on Asian Language Information Processing (TALIP)*, 12(4):16:1–16:25, 2013.

[27] C. Chu, T. Nakazawa, and S. Kurohashi. Chinese characters mapping table of Japanese, Traditional Chinese and Simplified Chinese. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC 2012)*, pages 2149–2152, 2012.

[28] C. Chu, T. Nakazawa, and S. Kurohashi. Ebmt system of kyoto university in olympics task at iwslt 2012. In *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT 2012)*, pages 96–101, Hong Kong, China, December 2012.

[29] C. Chu, T. Nakazawa, and S. Kurohashi. Chinese–japanese parallel sentence extraction from quasi–comparable corpora. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 34–42, 2013.

[30] C. Chu, T. Nakazawa, and S. Kurohashi. Iterative bilingual lexicon extraction from comparable corpora with topical and contextual knowledge. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing2014)*, pages 8404:2:296–309, 2014.

[31] M. Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the*

2002 Conference on Empirical Methods in Natural Language Processing, pages 1–8, 2002.

[32] D. Ştefănescu and R. Ion. 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)*, pages 117–128, 2013.

[33] D. Ştefănescu, R. Ion, and S. Hunsicker. Hybrid parallel sentence mining from comparable corpora. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012)*, pages 137–144, 2012.

[34] H. Daume III and J. Jagarlamudi. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 407–412, 2011.

[35] T. N. D. Do, L. Besacier, and E. Castelli. 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)*, 2010.

[36] Q. Dou and K. Knight. Large scale decipherment for out-of-domain machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 266–275, 2012.

[37] J. Du, J. Jiang, and A. Way. Facilitating translation using source language paraphrase lattices. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 420–429, 2010.

[38] A. Eisele and Y. Chen. Multiun: A multilingual corpus from united nation documents. In D. Tapias, M. Rosner, S. Piperidis, J. Odjik, J. Mariani, B. Maegaard, K. Choukri, and N. C. C. Chair), editors, *Proceedings of the*

Seventh conference on International Language Resources and Evaluation, pages 2868–2872, 2010.

[39] X. Fu, W. Wei, S. Lu, Z. Chen, and B. Xu. Phrase-based parallel fragments extraction from comparable corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 972–976, 2013.

[40] P. Fung. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, pages 173–183, 1995.

[41] P. Fung and P. Cheung. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of Coling 2004*, pages 1051–1057, 2004.

[42] P. Fung, E. Prochasson, and S. Shi. Trillions of comparable documents. In *3rd workshop on Building and Using Comparable Corpora (BUCC'10), Language Resource and Evaluation Conference (LREC'10)*, pages 26–34, 2010.

[43] P. Fung and L. Y. Yee. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 414–420, 1998.

[44] S. Gahbiche-Braham, H. Bonneau-Maynard, and F. Yvon. 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*, pages 44–51, 2011.

[45] M. Galley, M. Hopkins, K. Knight, and D. Marcu. What's in a translation rule? In D. M. Susan Dumais and S. Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 273–280, 2004.

[46] J. Ganitkevitch and C. Callison-Burch. The multilingual paraphrase database. In *The 9th edition of the Language Resources and Evaluation Conference*, pages 4276–4283, 2014.

[47] J. Ganitkevitch, Y. Cao, J. Weese, M. Post, and C. Callison-Burch. Joshua 4.0: Packing, pro, and paraphrases. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 283–291, 2012.

[48] N. Garera, C. Callison-Burch, and D. Yarowsky. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 129–137, 2009.

[49] E. Gaussier, J. Renders, I. Matveeva, C. Goutte, and H. Dejean. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 526–533, 2004.

[50] C.-L. Goh, M. Asahara, and Y. Matsumoto. Building a Japanese-Chinese dictionary using kanji/hanzi conversion. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 670–681, 2005.

[51] R. Gupta, S. Pal, and S. Bandyopadhyay. Improving mt system using extracted parallel fragments of text from comparable corpora. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 69–76, 2013.

[52] A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, 2008.

[53] R. Harastani, B. Daille, and E. Morin. Ranking translation candidates acquired from comparable corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 401–409, 2013.

[54] Z. S. Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

[55] A. HAZEM and E. MORIN. A comparison of smoothing techniques for bilingual lexicon extraction from comparable corpora. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 24–33, 2013.

[56] A. Hazem and E. Morin. Word co-occurrence counts prediction for bilingual terminology extraction from comparable corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1392–1400, 2013.

[57] S. Hewavitharana and S. Vogel. Extracting parallel phrases from comparable data. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 61–68, 2011.

[58] G. Hong, C.-H. Li, M. Zhou, and H.-C. Rim. An empirical study on web mining of parallel data. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 474–482, 2010.

[59] C.-R. Huang, Y.-M. Chou, C. Hotani, S.-Y. Chen, and W.-Y. Lin. Multilingual conceptual access to lexicon based on shared orthography: An ontology-driven study of Chinese and Japanese. In *Coling 2008: Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*, pages 47–54, 2008.

[60] A. Irvine and C. Callison-Burch. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 262–270, 2013.

[61] A. Irvine and C. Callison-Burch. Supervised bilingual lexicon induction with multiple monolingual signals. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–523, 2013.

[62] A. Irvine, J. Morgan, M. Carpuat, H. D. III, and D. Munteanu. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics (TACL)*, 1:429–440, 2013.

[63] A. Irvine, C. Quirk, and H. Daumé III. Monolingual marginal matching for translation model adaptation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1077–1088, 2013.

[64] T. Ishisaka, M. Utiyama, E. Sumita, and K. Yamamoto. Development of a japanese-english software manual parallel corpus. In *MT Summit*, 2009.

[65] L. Jiang, S. Yang, M. Zhou, X. Liu, and Q. Zhu. 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*, pages 870–878, 2009.

[66] D. Kawahara and S. Kurohashi. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In *Proceedings of Human Language Technology Conference of the NAACL, Main Conference*, pages 176–183, 2006.

[67] A. Klementiev, A. Irvine, C. Callison-Burch, and D. Yarowsky. Toward statistical machine translation without parallel corpora. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 130–140, 2012.

[68] A. Klementiev and D. Roth. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 817–824, 2006.

[69] P. Koehn. Statistical significance tests for machine translation evaluation. In D. Lin and D. Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, 2004.

[70] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, pages 79–86, 2005.

[71] P. Koehn. *Statistical Machine Translation*. Cambridge University Press, 1st edition, 2010.

[72] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 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*, pages 177–180, 2007.

[73] P. Koehn and K. Knight. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 9–16, 2002.

[74] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, 2003.

[75] G. Kondrak, D. Marcu, and K. Knight. Cognates can improve statistical translation models. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 46–48, 2003.

[76] T. Kudo, K. Yamamoto, and Y. Matsumoto. Applying conditional random fields to Japanese morphological analysis. In D. Lin and D. Wu, editors, *Proceedings of EMNLP 2004*, pages 230–237, 2004.

[77] S. Kurohashi, T. Nakamura, Y. Matsumoto, and M. Nagao. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language*, pages 22–28, 1994.

[78] A. Laroche and P. Langlais. Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 617–625, 2010.

[79] B. Li and E. Gaussier. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 644–652, 2010.

[80] B. Li, E. Gaussier, and A. Aizawa. Clustering comparable corpora for bilingual lexicon extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 473–478, 2011.

[81] D. Lin, S. Zhao, B. Van Durme, and M. Paşca. Mining parenthetical translations from the web by word alignment. In *Proceedings of ACL-08: HLT*, pages 994–1002, 2008.

[82] W. Ling, G. Xiang, C. Dyer, A. Black, and I. Trancoso. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 176–186, 2013.

[83] X. Liu, K. Duh, and Y. Matsumoto. Topic models + word alignment = a flexible framework for extracting bilingual dictionary from comparable corpus. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 212–221, 2013.

[84] J. K. Low, H. Tou Ng, and W. Guo. A maximum entropy approach to Chinese word segmentation. In *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing (SIGHAN05)*, pages 161–164, 2005.

[85] B. Lu, T. Jiang, K. Chow, and B. K. Tsou. 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*, pages 42–49, 2010.

[86] Y. Ma and A. Way. Bilingually motivated domain-adapted word segmentation for statistical machine translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 549–557, 2009.

[87] Y. Marton, C. Callison-Burch, and P. Resnik. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390, 2009.

[88] H. Masuichi, R. Flournoy, S. Kaufmann, and S. Peters. A bootstrapping method for extracting bilingual text pairs. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, pages 1066–1070, 2000.

[89] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[90] T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013.

[91] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889, 2009.

[92] E. Morin and B. Daille. Revising the compositional method for terminology acquisition from comparable corpora. In *Proceedings of COLING 2012*, pages 1797–1810, 2012.

[93] D. S. Munteanu and D. Marcu. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504, 2005.

[94] D. S. Munteanu and D. Marcu. 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*, pages 81–88, 2006.

[95] T. Nakazawa and S. Kurohashi. Bayesian subtree alignment model based on dependency trees. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 794–802, 2011.

[96] J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand. 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, pages 74–81, 1999.

[97] I. Niles and A. Pease. Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, pages 2–9, 2001.

[98] F. J. Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, 2003.

[99] F. J. Och and H. Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, 2002.

[100] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, Mar. 2003.

[101] S. Pal, P. Lohar, and S. K. Naskar. Role of paraphrases in pb-smt. In *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing2014)*, pages 8404:2:245–256, 2014.

[102] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[103] V. Pekar, R. Mitkov, D. Blagoev, and A. Mulloni. Finding translations for low-frequency words in comparable corpora. *Machine Translation*, 20(4):247–266, 2006.

[104] F. Peng, F. Feng, and A. McCallum. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of Coling 2004*, pages 562–568, 2004.

[105] A. Pirkola, T. Hedlund, H. Keskustalo, and K. Jä rvelin. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval*, 4:209–230, 2001.

[106] M. Post, C. Callison-Burch, and M. Osborne. Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, 2012.

[107] E. Prochasson and P. Fung. Rare word translation extraction from aligned comparable documents. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1327–1335, 2011.

[108] L. Qian, H. Wang, G. Zhou, and Q. Zhu. Bilingual lexicon construction from comparable corpora via dependency mapping. In *Proceedings of COLING 2012*, pages 2275–2290, 2012.

[109] C. Quirk, R. U. U, and A. Menezes. Generative models of noisy translations with applications to parallel fragment extraction. In *In Proceedings of MT Summit XI, European Association for Machine Translation*, 2007.

[110] R. Rapp. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322, 1995.

[111] R. Rapp. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 519–526, 1999.

[112] S. Ravi and K. Knight. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, 2011.

[113] M. Razmara, M. Siahbani, R. Haffari, and A. Sarkar. Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115, 2013.

[114] P. Resnik and N. A. Smith. The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380, 2003.

[115] J. Richardson, T. Nakazawa, and S. Kurohashi. Robust transliteration mining from comparable corpora with bilingual topic models. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 261–269, 2013.

[116] J. Riesa and D. Marcu. 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*, pages 538–542, 2012.

[117] C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[118] R. Skadiņš, J. Tiedemann, R. Rozis, and D. Deksne. Billions of parallel words for free: Building and using the eu bookshop corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1850–1855, 2014.

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

[120] M. Snover, B. Dorr, and R. Schwartz. Language and translation model adaptation using comparable corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 857–866, 2008.

[121] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, and D. Tufiş . The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 2142–2147, 2006.

[122] A. Stolcke. Srilm – an extensible language modeling toolkit. In *Proceedings of ICSLP*, volume 2, pages 901–904, 2002.

[123] A. Tamura, T. Watanabe, and E. Sumita. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 24–36, 2012.

[124] C. L. Tan and M. Nagao. Automatic alignment of Japanese-Chinese bilingual texts. *IEICE Transactions on Information and Systems*, E78-D(1):68–76, 1995.

[125] L. Tian, D. F. Wong, L. S. Chao, P. Quaresma, F. Oliveira, and L. Yi. Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1837–1842, 2014.

[126] J. Tiedemann. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, 2012.

[127] C. Tillmann. A beam-search extraction algorithm for comparable data. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 225–228, 2009.

[128] T. Tsunakawa, Y. Yamamoto, and H. Kaji. Improving calculation of contextual similarity for constructing a bilingual dictionary via a third language. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1057–1061, 2013.

[129] Y. Tsuruoka, Y. Miyao, and J. Kazama. Learning with lookahead: Can history-based models rival globally optimized models? In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 238–246, 2011.

[130] J. Uszkoreit, J. Ponte, A. Popat, and M. Dubiner. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1101–1109, 2010.

[131] M. Utiyama and H. Isahara. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79, 2003.

[132] M. Utiyama and H. Isahara. A japanese-english patent parallel corpus. In *Proceedings of MT summit XI*, pages 475–482, 2007.

[133] E. M. Voorhees. The TREC-8 question answering track report. In *Proceedings of the Eighth TExt Retrieval Conference (TREC-8)*, pages 77–82, 1999.

[134] T. Vu, A. T. Aw, and M. Zhang. Feature-based method for document alignment in comparable news corpora. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 843–851, 2009.

[135] I. Vulić, W. De Smet, and M.-F. Moens. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 479–484, 2011.

[136] I. Vulić and M.-F. Moens. Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 449–459, 2012.

[137] I. Vulić and M.-F. Moens. Sub-corpora sampling with an application to bilingual lexicon extraction. In *Proceedings of COLING 2012*, pages 2721–2738, 2012.

[138] I. Vulić and M.-F. Moens. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–116, 2013.

[139] I. Vulić and M.-F. Moens. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Con-*

ference on Empirical Methods in Natural Language Processing, pages 1613–1624, 2013.

[140] R. Wang and C. Callison-Burch. Paraphrase fragment extraction from monolingual comparable corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 52–60, 2011.

[141] Y. Wang, J. Kazama, Y. Tsuruoka, W. Chen, Y. Zhang, and K. Torisawa. Improving word segmentationChinese word segmentation and POS tagging with semi-supervised methods using large auto-analyzed data. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 309–317, 2011.

[142] Y. Wang, K. Uchimoto, J. Kazama, C. Kruengkrai, and K. Torisawa. Adapting Chinese word segmentation for machine translation based on short units. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, 2010.

[143] D. Wu and P. Fung. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *IJCNLP*, pages 257–268, 2005.

[144] F. Xia, M. P. N. Xue, M. E. Okurowski, J. Kovarik, F. dong Chiou, and S. Huang. Developing guidelines and ensuring consistency for Chinese text annotation. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, 2000.

[145] J. Xu, R. Zens, and H. Ney. Do we need Chinese word segmentation for statistical machine translation? In O. Streiter and Q. Lu, editors, *ACL SIGHAN Workshop 2004*, pages 122–128, 2004.

[146] K. Yu and J. Tsujii. Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chap-*

ter of the Association for Computational Linguistics, Companion Volume: Short Papers, pages 121–124, 2009.

[147] O. F. Zaidan and C. Callison-Burch. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, 2011.

[148] J. Zhang and C. Zong. 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)*, pages 1425–1434, 2013.

[149] S. Zhang, W. Ling, and C. Dyer. Dual subtitles as parallel corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1869–1874, 2014.

[150] Y. Zhang, K. Wu, J. Gao, and P. Vines. Automatic acquisition of chinese-english parallel corpus from the web. In M. Lalmas, A. MacFarlane, S. M. Rü ger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, *ECIR*, volume 3936 of *Lecture Notes in Computer Science*, pages 420–431, 2006.

[151] B. Zhao, M. Eck, and S. Vogel. Language model adaptation for statistical machine translation via structured query models. In *Proceedings of Coling 2004*, pages 411–417, 2004.

[152] B. Zhao and S. Vogel. Adaptive parallel sentences mining from web bilingual news collections. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, pages 745–748, 2002.

[153] Z. Zhu, M. Li, L. Chen, and Z. Yang. Building comparable corpora based on bilingual lda model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 278–282, 2013.

List of Major Publications

[1] Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi. Japanese–Chinese Phrase Alignment Using Common Chinese Characters Information. In *Proceedings of MT Summit XIII*, pages 475–482, 2011.

[2] Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi. Chinese Characters Mapping Table of Japanese, Traditional Chinese and Simplified Chinese. In *Proceedings of the 8th Conference on International Language Resources and Evaluation (LREC2012)*, pages 2149–2152, 2012.

[3] Chenhui Chu, Toshiaki Nakazawa, Daisuke Kawahara and Sadao Kurohashi. Exploiting Shared Chinese Characters in Chinese Word Segmentation Optimization for Chinese–Japanese Machine Translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT2012)*, pages 35–42, 2012.

[4] Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi. Chinese–Japanese Parallel Sentence Extraction from Quasi–Comparable Corpora. In *Proceedings of the 6th Workshop on Building and Using Comparable Corpora (BUCC2013)*, pages 34–42, 2013.

[5] Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi. Accurate Parallel Fragment Extraction from Quasi–Comparable Corpora using Alignment Model and Translation Lexicon. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP2013)*, pages 1144–1150, 2013.

[6] Chenhui Chu, Toshiaki Nakazawa, Daisuke Kawahara and Sadao Kurohashi. Chinese–Japanese Machine Translation Exploiting Chinese Characters. ACM Transactions on Asian Language Information Processing (TALIP), 12(4):16:1–16:25, 2013.

[7] Chenhui Chu, Toshiaki Nakazawa, Sadao Kurohashi. Iterative Bilingual Lexicon Extraction from Comparable Corpora with Topical and Contextual Knowledge. In *Proceedings of the 15th International Conference on*

Intelligent Text Processing and Computational Linguistics (CICLing2014), Springer Lecture Notes in Computer Science (LNCS) 8404(II):296–309, 2014. **(Best Student Paper Award)**

[8] Chenhui Chu, Toshiaki Nakazawa, Sadao Kurohashi. Constructing a Chinese–Japanese Parallel Corpus from Wikipedia. In *Proceedings of the 9th Conference on International Language Resources and Evaluation (LREC2014)*, pages 642–647, 2014.

[9] Chenhui Chu, Toshiaki Nakazawa, Sadao Kurohashi. Improving Statistical Machine Translation Accuracy Using Bilingual Lexicon Extraction with Paraphrases. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing (PACLIC2014)*, pages 262–271, 2014.

[10] Chenhui Chu, Toshiaki Nakazawa, Sadao Kurohashi. Integrated Parallel Sentence and Fragment Extraction from Comparable Corpora: A Case Study on Chinese-Japanese Wikipedia. ACM Transactions on Asian Language Information Processing (TALIP). (Conditionally accepted)

[11] Chenhui Chu, Toshiaki Nakazawa, Sadao Kurohashi. Parallel Sentence Extraction Based on Unsupervised Bilingual Lexicon Extraction from Comparable Corpora. Journal of Natural Language Processing (JNLP). (Conditionally accepted)

List of Other Publications

[1] Chenhui Chu and Keiji Shinzato. Chinese–Japanese Search Query Translation System. In *Proceedings of 4th Rakuten R&D Symposium*, 2011.

[2] Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi. Japanese–Chinese Phrase Alignment Exploiting Shared Chinese Characters. In *Proceedings of the 18th Annual Meeting of the Association for Natural Language Processing (NLP2012)*, pages 143–146, 2012.

[3] Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi. EBMT System of Kyoto University in OLYMPICS Task at IWSLT 2012. In *Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT2012)*, pages 96–101, 2012.

[4] Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi. Iterative Bilingual Lexicon Extraction from Comparable Corpora Using Topic Model and Context Based Methods. In *Proceedings of the 20th Annual Meeting of the Association for Natural Language Processing (NLP2014)*, pages 729–732, 2014.

[5] Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi. Chinese-Japanese Parallel Sentence Extraction from Quasi-Comparable and Comparable Corpora. Invited Chapter in the Book of *Using Comparable Corpora for Under-Resourced Areas of Machine Translation*, 17 pages, 2015. (to appear)

[6] Chenhui Chu, Raj Dabre, Toshiaki Nakazawa and Sadao Kurohashi. Large-scale Japanese-Chinese Scientific Dictionary Construction via Pivot-based Statistical Machine Translation. In *Proceedings of the 21th Annual Meeting of the Association for Natural Language Processing (NLP2015)*, 4 pages, 2015. (to appear)