Waseda University Doctoral Dissertation

A study of translation equivalences at various levels of granularity for Chinese–Japanese technical translation

Wei YANG

Graduate School of Information, Production and Systems
Waseda University

July 2017

Abstract

China and Japan are producing a large amount of scientific articles and patents in their respective languages. The World Intellectual Property Organization (WIPO) Indicators show that China was the first country for patent applications in 2013. Japan was the first country for patent grants in 2013. Much of current scientific development in China or Japan is not readily available to non-Chinese or non-Japanese speaking researchers. Making Chinese patents and scientific texts available in Japanese, and Japanese patents and scientific texts in Chinese is a key issue for increasing economical development in Asia.

The crucial process of translation of technical texts should be helped by the use of machine translation (MT) as recognized by the Japan Science and Technology Agency (JST) project and State Intellectual Property Office (SIPO) project in China. Sentence-level aligned parallel corpora are an essential resource for data-driven statistical machine translation (SMT). For Chinese–Japanese SMT, the first problem is that there are almost no Chinese–Japanese parallel corpora publicly freely available in any domain. Another problem is that Chinese and Japanese do not have typographic boundaries in their writing systems. Thus, for machine translation, word segmentation (tokenization), i.e., breaking sentences down into individual characters or words (tokens) is normally treated as the first step of pre-processing in natural language processing (NLP). But different segmentation conventions lead to different segmentation results at different levels of granularity, such as segment a sentence into characters, words or chunks between Chinese and Japanese that lead to inconsistencies in alignment that negatively affect the accuracy of translation. In Chinese–Japanese technical machine translation, a corpus may contain large amounts of domain-specific terms in words or multi-word expressions. Reasonable word segmentation and multi-word alignment in terms is an important processing task for technical machine translation in order to keep higher translation accuracy.

When investigating the translation equivalences between Chinese and Japanese, we have to notice that in the Chinese and Japanese writing systems, there exist translation equivalences at character and word level. Indeed, Chinese and Japanese share a large amount of characters and words with the same or similar meaning. Most of the Japanese kanji ideograms were original created in ancient China and a large amount of words written in kanji created in Japan were re-imported back to China to be widely used. Even nowadays, there still constantly exist creation of words in Japan that come into China. These characteristics should be helpful in construction or acquisition of different types of data for the less-resourced language pair Chinese–Japanese, for instance, lexica (word level), bilingual term alignments (multi-word or phrase level) and approximately parallel corpus (sentence level). These bilingual data should be very helpful to improve translation accuracy of statistical machine translation.

We firstly introduce the basic knowledge and background of the research: machine translation and different types of machine translation, less-resourced languages and language pairs including the language pair address in our work: Chinese–Japanese, parallel corpora and non-parallel corpora including quasi-parallel corpora, moreover, we present an investigation on the word segmentation and granularity for Chinese and Japanese. The research and experiments are given for our subsequent work on Chinese and Japanese.

Bilingual dictionaries are very useful for several types of machine translation. For SMT, a dictionary or a lexicon can increase lexical coverage and the quality of phrase alignment. Dictionaries can be added into an existing training corpus, or they can be used in the decoding process directly without changing the translation model. For addressing the problem of scarcity of bilingual lexica, we construct a bilingual lexicon by combining several automatic techniques on several freely available resources. We try to increase the quality of the lexicon with several methods, including kanji-hanzi conversion method. We obtain a Chinese–Japanese lexicon with more than 45,000 entries, 85% of which have correct translation correspondence (40% increase in accuracy). 83% of the entries in this lexicon are not included in a reference dictionary.

Improving the quality of word and phrase alignment in a phrase-based statistical machine translation (PB-SMT) system could lead to improvements in machine translation performance. Multi-word terms in technical translation need to be translated as one word to avoid being translated using incorrect word-to-word alignments. Consequently, our ultimate goal is to enforce the proper translation of multi-word terms between Chinese and Japanese. We extract bilingual multi-word to multi-word or single-word to multi-word terms from an existing training corpus and re-tokenize the training corpus with these extracted bilingual terms. Finally, we train a translation model using this re-tokenized training corpus. We combine several statistical methods and the kanji-hanzi conversion method. We obtained better results in bilingual term extraction with 90%+ and in statistically significant improvement evaluation results of SMT with an increase of 1 to 2 BLEU points.

Exploiting existing parallel corpora and monolingual corpora using analogical associations is our approach for addressing the problem of scarcity of parallel corpora for the Chinese–Japanese language pair. Monolingual data is easier to access than bilingual data for Chinese and Japanese. Each of these two languages is a well-resourced language. We propose a method that generates large amounts of new sentences from a small amount of parallel data and certain number of analogical rewriting models which are built from a large amount of monolingual data. We proposed two methods (BLEU and N-sequence) for filtering the over-generated sentences. The N-sequence method allows us to keep sentences with 99% in grammatically correct accuracy. A quasi-parallel corpus is constructed based on the similarity of the clusters across two languages for new sentence generation and the translation relations between the parallel corpus used for new sentence generation. By adding the constructed quasi-parallel corpus into an existing training corpus, we obtain 0.27 to 6 BLEU points of statistically significant improvement over the baseline system in several experiments.

In our work, we also combine several proposed methods and works on translation equivalences at various levels of granularity for Chinese–Japanese to improve technical machine translation accuracy. We obtain a statistically significant improvement of 1.8 BLEU point using bilingual multi-word term extraction and re-tokenization methods, the result of quasi-parallel corpus constructed from monolingual data using analogical associations, and the result of a lexicon constructed using several technologies and free resources.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor, Professor Yves Lepage for giving me a chance to join the EBMT/NLP laboratory, and guiding me into machine translation and natural language processing research. Without his continuous guidance, support and encourage throughout my master course and doctoral course, I could not finish my work and this doctoral dissertation smoothly. I am also thankful for everything he has taught me, whether it is about how to program for solving the problems in data analysis and natural language processing, or about how to become a researcher and how to do research such as the scientific way to write a paper or give a good presentation. Thanks to him for giving me an interesting and promising research topic that it is about the construction of the data at variant levels of granularity between Chinese and Japanese. Moreover, he gave me various chances and exercises during my doctoral course, such as filling a post of teacher assistant and research assistant.

I am also very grateful to Professor Makoto Tsubokawa and Professor Jiro Tanaka for their generous donation of time in reviewing and giving valuable comments on the doctoral dissertation.

I would like to thank Professor Yujie Zhang (currently at Beijing Jiaotong University) for giving me suggestive advice and helpful comments in the conference or research communication.

I would also like to thank my doctoral dissertation committees: Professor Yves Lepage, Professor Takayuki Furuzuki, and Professor Mizuho Iwaihara for their valuable advice and comments on this work.

Many thanks to all previous and current members in the laboratory. Especially, I would like to thank Dr. Juan Luo, Jing Sun, Jin Matsuoka, Hao Wang, Jinghui Yan, Hanfei Shen (currently at NEC), Mengru Gao, Tomohiro Kawabe (currently at Canon), Hiroki Mouri (currently at Hitachi), Tatsuya Kimura (currently at Hitachi), Bin Shan, Kun Chen, Kevin Patrick Yancey, Vivatchai Kaveeta, Rashel Putraruddy, Weiqing Tian to name a few. Thanks for their helping and many enlightening discussions, that makes me finishing the research smoothly.

Finally, I would like to give a special thank to my family and friends for their endless support, encouraging and understanding during my master and doctoral work and through in my life.

Contents

List of Figures

List of Tables

Publications Related to the Thesis

Most of the research described in this dissertation has been published in journal papers and conference papers.

- Chapter 2
  - Wei Yang and Yves Lepage. Combining Several Automatic Techniques to Build a Chinese–Japanese Lexicon from Freely Available Resources. In *Proceedings of the 18th Yearly Conference of the Japanese Association for Natural Language Processing (言語処理学会第18回年次大会 NLP2012)*, pp. 747–750, Hiroshima, March 2012.
- Chapter 3
  - Wei Yang and Yves Lepage. Improving Automatic Chinese–Japanese Patent Translation using Bilingual Term Extraction. *IEEJ Transactions on Electrical and Electronic Engineering*, Vol.13, No.1, January 2018. (to appear)
  - Wei Yang and Yves Lepage. Bilingual Multi-Word Term Tokenization for Chinese–Japanese Patent Translation. In *Proceedings of the 23th Yearly Conference of the Japanese Association for Natural Language Processing (言語処理学会第23回年次大会 NLP2017)*, pp. 855–858, Tsukuba, March 13-17, 2017.
  - Wei Yang and Yves Lepage. Improving Patent Translation Using Bilingual Term Extraction and Re-tokenization for Chinese–Japanese. In *Proceedings of the 3rd Workshop on Asian Translation (WAT 2016) co-located with COLING 2016*, pp. 194–202, Osaka, Japan, December 11-17, 2016.
  - Wei Yang, Jinghui Yan and Yves Lepage. Extraction of Bilingual Technical Terms for Chinese–Japanese Patent Translation. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT*

2016) Student Research Workshop (SRW), pp. 81–87, San Diego, California, June 12-17, 2016.

- Chapter 4
  - Wei Yang, Hanfei Shen and Yves Lepage. Inflating a Small Parallel Corpus into a Large Quasi-parallel Corpus Using Monolingual Data for Chinese–Japanese Machine Translation. *Journal of Information Processing*, Vol.25, pp. 88–99, January 2017.
  - Wei Yang, Hao Wang and Yves Lepage. Deduction of Translation Relations between New Short Sentences in Chinese and Japanese Using Analogical Associations. *International Journal of Advanced Intelligence (IJAI)*, Vol.6, No.1, pp. 13–34, December 2014.
  - Wei Yang and Yves Lepage. Consistent Improvement in Translation Quality of Chinese–Japanese Technical Texts by Adding Additional Quasi-parallel Training Data. In *Proceedings of the 1st Workshop on Asian Translation (WAT 2014)*, pp. 69–76, October 4, 2014.
  - Wei Yang and Yves Lepage. Inflating a Training Corpus for SMT by Using Unrelated Unaligned Monolingual Data. In *Proceedings of the 9th International Conference on NLP (PolTAL 2014) LNAI 8686*, pp. 236–248, September 17-19, 2014.

Chapter 1

Introduction

This chapter introduces the background and related work for our research. We introduce the types of machine translation (MT) and especially describe the phrase-based statistical machine translation (PB-SMT) used in our work. We also describe what are less-resourced languages and language pairs. The parallel corpora and the related work on construction of one of the less-resourced language pairs: Chinese–Japanese are also introduced in this chapter. We also introduce the granularity and word segmentation which is necessary as pre-processing for Chinese and Japanese. In this chapter, we give a big picture and the overview of our work, as well as the outline and research contributions of this dissertation.

The structure of this chapter is as follows.

- Section 1.1 introduces the basic knowledge and the overview of the machine translation system. We especially give the introduction for phrase-based statistical machine translation used in our research.
- In Section 1.2, we introduce the less-resourced languages and language pairs.
- Section 1.3 describes the parallel corpora which are important used in phrase-based statistical machine translation. We also give some related work on construction of Chinese–Japanese parallel corpus.
- Section 1.4 introduces the granularity and word segmentation for Chinese and Japanese.
- Section 1.5 and Section 1.6 describe the overview of our research and the contribution of our research in this dissertation.

1.1 Statistical Machine Translation

Machine translation (MT) is a specific task of natural language processing (NLP). It is used to automatically translate speech or text from one natural language to another natural language using translation system. Basically, there exist five kinds (methods) of machine translation: rule-based machine translation (RBMT), example-based machine translation (EBMT), statistical machine translation (SMT), hybrid machine translation (HMT) and neural machine translation (NMT).

In rule-based machine translation (RBMT) (Hutchins and Somers, 1992), (Sánchez-Cartagena et al., 2011), there are intermediate states between the source language and the final target translation. Firstly, a source language is analyzed to build a source language intermediate representation, based on grammar rules and linguistic analysis using, for instance, lemma, part-of-speech (POS) tagging, syntactic analysis, etc; secondly, the source language intermediate representation is transferred to a target language intermediate representation; finally, a final translation in the target language is generated from the target intermediate representation.

Example-based machine translation (EBMT) as one of the methods of machine translation was firstly proposed by Nagao (1984). The basic ideas of EBMT is translation of sentences by analogy and translation of a sentence based on learned translation knowledge of portions (sub-sentential components) of a sentence.

In statistical machine translation (SMT) (Weaver, 1955), (Brown et al., 1990), (Brown et al., 1993), (Koehn, 2010), different from the RBMT and EBMT, SMT a system is built using several statistical models based on bilingual or monolingual corpora. SMT is the most widely used and studied machine translation method in recent 10 years. In SMT, machine-readable parallel corpora are a crucial resource to acquire the translation knowledge from aligned parallel data. Thus, the scarcity of parallel corpora is one of the problems for SMT.

There are several kinds of translation models used in SMT, for instance, word-based model (translation based on words) (Och and Ney, 2003); phrase-based model (Koehn et al., 2003) (translation based on any sequences of words (phrases)), i.e., single word and multi-word; syntax-based model (Yamada and Knight, 2001) which differ from the word-based model and phrase-based model as they are based on syntactic units, i.e., parse trees of sentences; and hierarchical phrase-based model (Chiang, 2005) which combines the phrase-based and syntax-based translation models.

The phrase-based translation model usually contains a phrase translation table (phrase table) and a configuration file for translation (decoding or testing). Phrase translation tables contain word or multi-word alignments (phrase entries) with: the probability of the source phrase ($\bar{f}$) knowing the target phrase ($\bar{e}$), the probability of the target phrase knowing the source phrase, the lexical weighting (lexical translation probability) of the source phrase knowing the target phrase, the lexical weighting of the target phrase knowing the source phrase, and a phrase penalty (set to 2.718). They are denoted as: $\phi(\bar{f}|\bar{e})$, $lex(\bar{f}|\bar{e})$, $\phi(\bar{e}|\bar{f})$, $lex(\bar{e}|\bar{f})$ (Koehn et al., 2003). The translation probabilities ($\phi(\bar{f}|\bar{e})$ and $\phi(\bar{e}|\bar{f})$) are estimated based on the relative frequency given in Formula 1.1. The lexical weightings ($lex(\bar{f}|\bar{e})$ and $lex(\bar{e}|\bar{f})$) are estimated based on alignments (a) between the words ($w(e_i|f_j)$ or $w(f_i|e_j)$) contained in the phrase as given in Formula 1.2.

$$\phi(\bar{f}|\bar{e}) = \frac{count(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} count(\bar{e}, \bar{f})} \tag{1.1}$$

$$lex(\bar{e}|\bar{f}, a) = \prod_{i=1}^{length(\bar{e})} \frac{1}{|\{j|(i,j) \in a\}|} \sum_{\forall(i,j)\in a} w(e_i|f_j) \tag{1.2}$$

Table 1.1 shows an example of a Chinese–Japanese translation table obtained based on a parallel training corpus in the technical domain.

Table 1.1: An example of a Chinese–Japanese phrase translation table obtained from a parallel corpus in the technical domain.

Source language	**Target language**	**Feature scores**				
Chinese	Japanese	$\phi(\bar{f}|\bar{e})$	$lex(\bar{f}|\bar{e})$	$\phi(\bar{e}|\bar{f})$	$lex(\bar{e}|\bar{f})$	*penalty*
程序	プログラム	0.409967	0.364362	0.273312	0.532784	2.718
气孔	気孔	0.128052	0.438749	0.128052	0.410469	2.718
将 电缆	ケーブル	0.000337	0.014059	0.137377	0.929329	2.718
油墨 是 指	インク と は	0.128052	0.198724	0.128052	0.673249	2.718
将 监视	は、	3.36618e-06	2.95662e-06	0.045792	0.014704	2.718
将 监视	は 、 モニタ	0.027476	2.95662e-06	0.045792	2.33791e-06	2.718
将 监视	は 、 監視	0.013738	0.011537	0.045792	0.034063	2.718
能够 高效率	肺 気腫	0.019929	1.000000	0.006643	1.000000	2.718

There are two main components in creating a statistical machine translation (SMT) system: the training pipeline and the decoding. The training pipeline mainly contains in training the translation model (here we use the phrase-based translation model mentioned above), training the language model (LM) and tuning (weighting) the different statistical models. The overview of creating a SMT system is given in Figure 1.1.

A language model is trained on a monolingual corpus in the target language. The language model is one of the important part in training a SMT translation system to ensure fluency of the outputs in the target language. In our experiments, we use SRILM (Stolcke et al., 2002) and KenLM (Heafield, 2011).

The tuning step is used for determining the weighting parameters for the different statistical models to produce the best possible translations of the test set in the source language. In our experiments, we use minimum error rate training (MERT) (Och, 2003) in the tuning step.

The decoding step is used to find and output the translation of the source language with the highest scores according to the translation model. It can also output different types of information, for instance the trace of the corresponding source phrase used in decoding into the target phrase. This will be used in Chapter 3 and Chapter 4. For translation, it considers the length of the sentences, the word order in the languages and the fluency of the sentences.

The main metric used in our experiments for automatically evaluating the translation outputs is BLEU (Papineni et al., 2002) method. The basic idea of the BLEU metric is counting the number of n-grams (sequence of word) in the translation output against the reference in the target language. The n-grams used in our experiments are 1-gram to 4-gram. We also check the statistical significance of two SMT systems by p-value (Koehn, 2010). If the p-value is less than 5% (0.05) or even 1% (0.01), it means that there is less than 5% or 1% chance that the difference in two scores obtained by two different systems is due to accidental fluctuation of two equal systems. In other words, the difference of the two systems has 95% or 99% statistical significance with p-values of $p < 0.05$ or $p < 0.01$.

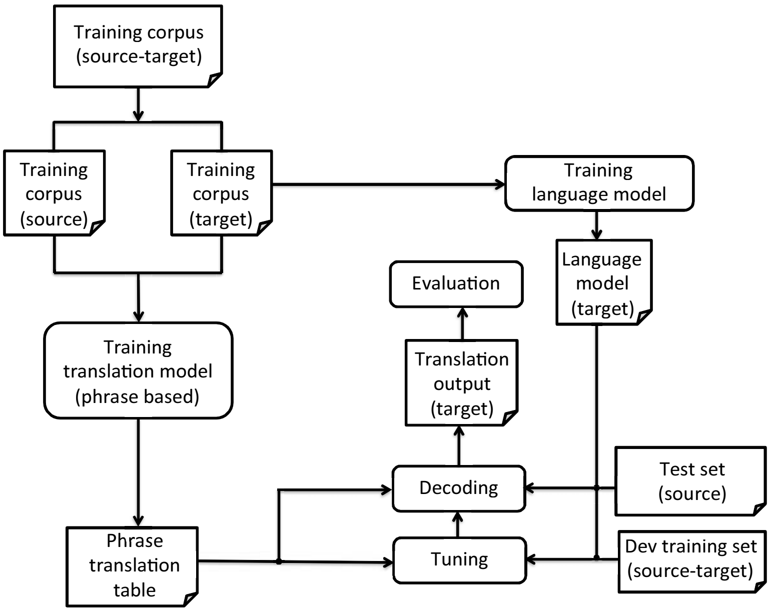

FIGURE 1.1: The overview of producing a phrase-based SMT system from training data.

Here we also introduce two other kinds of machine translation. Hybrid machine translation (HMT) builds machine translation systems which combine multiple machine translation approaches. For instance, combining statistical and rule-based translation methods. There are two ways of doing this: translating text using RBMT and then adjust or correct the output of the translation using SMT; the rules used in RBMT are used as pre-processing or post-processing to guide the SMT system or post-process the translation output by SMT.

In neural machine translation (NMT)[1], different from SMT, deep learning is done using neural network technology. In the last two years, statistical machine translation is gradually fading out in favor of neural machine translation. Google translate supports over 100 languages. In November 2016, Google[2] has switched to a neural machine translation engine for 8 languages between English (to and from) and Chinese, French, German, Japanese, Korean, Portuguese, Spanish and Turkish. Microsoft Translator live and Skype Translator released 10 languages for speech translation[3] (in November, 2016). An open source "OpenNMT" has been released by the Harvard NLP group[4] (in March, 2017).

In our work, we focus on the statistical machine translation based on phrase-based translation model. All SMT experiments in our research are performed by using a state-of-the-art phrase-based SMT, an open-resource toolkit: Moses (Koehn et al., 2007) with GIZA++ (Och and Ney, 2000, 2003) for word alignment.

1.2 Less-resourced Languages and Language Pairs

The term "less-resourced language" or "under-resourced language" was introduced by Krauwer (2003) and mentioned in (Scannell, 2007) and (Besacier et al., 2014). It refers to a language with some or all of the following aspects in the implementation of Human Language Technologies (HLT) (speech recognition and machine translation for instance): lack of independent writing system or orthography, limited resource on the Web, lack of linguist knowledge, lack of electronic (machine-readable) resources, such as bilingual electronic dictionaries, monolingual corpus, bilingual corpora, lack of part-of-speech and morphological analyzers, parsers, pronunciation database, etc.

Linguistic resources between languages like: Chinese, Japanese, Thai, Hindi or Bahasa Indonesian are relatively scarce. This does not mean that they are minority languages or less-resourced languages, as all these languages have several million speakers and writers and monolingual data is quite easy to collect. The language pairs (not the languages) in this case are called less-resourced language pairs.

In the following section, we introduce extremely important resources used in SMT: parallel corpora, and some related work on construction of resources for the less-resourced language pair: Chinese–Japanese.

[1] https://en.wikipedia.org/wiki/Neural_machine_translation

[2] https://en.wikipedia.org/wiki/Google_Neural_Machine_Translation

[3] https://blogs.msdn.microsoft.com

[4] http://opennmt.net

1.3 Parallel Corpora and Non-parallel Corpora

Parallel corpora are parallel texts aligned at the sentence level. Parallel sentences are an extremely important resource in current data-driven Natural Language Processing (NLP). Especially, they are a prerequisite for training corpus-based MT, like statistical machine translation (SMT). Figure 1.2 shows an example of a parallel corpus between Chinese and Japanese from ASPEC corpus[5] in technical domain.

Chinese	**Japanese**
使用的形容词对是如下所示的２０种。	用いた形容詞対は以下に示す２０種類である。
本研究中，把办公室内的知识共享支持作为研究的对象。	本研究ではオフィス内の知識共有支援を研究の対象とした。
还通过探索有关内脏的新型消化酵素，调查了各脏器抽出物的情况。	内蔵に関し新規な消化酵素探索で各臓器抽出物を調べた。
介绍了在促进新能源的利用和技术研发方面，生物质能与微栅的研究。	新エネルギー利用の促進·技術開発では，バイオマスとマイクログリッドの研究を紹介した。

FIGURE 1.2: An example of a Chinese–Japanese parallel corpus in technical domain.

Corresponding to the concept of parallel corpora, there are several types of non-parallel corpora. For instance, comparable corpus and quasi-comparable corpus. Comparable corpora are texts in two languages that express similar contents on the same topic, but are not exact sentence-aligned translations of each other. Figure 1.3 shows an example of a comparable corpus between Chinese and Japanese from Wikipedia[6] [7]. Quasi-comparable corpora include more disparate very-non-parallel bilingual documents that could either be on the same topic or not. In our research, we construct a quasi-parallel corpus which contains sentences that are translations of each other to a certain extent as estimated by certain similarity scores (see Figure 1.4).

Chinese	**Japanese**
消化酶（英语：Digestive enzymes）是将聚合的高分子降解为他们的构建单元的酶类，以促进他们被身体吸收。消化酶类可在动物（及人）的消化管内找到，在那里帮助食物的消化，他们也存在于细胞中，特别是在其溶酶体中发挥作用，以维护细胞中的残留物。消化酶类多种多样，他们存在于：由唾腺分泌的唾液之中、由胃内壁细胞分泌的胃液之中、由胰腺外分泌细胞分泌的胰液之中以及在肠（大与小）胃分泌物之中。	消化酵素（しょうかこうそ）は、消化に使われる酵素のことである。分解される栄養素によって炭水化物分解酵素、タンパク質分解酵素、脂肪分解酵素などに分けられる。生物が食物を分解するために産生するほかは、食品加工、胃腸薬、洗剤として使用される。海外ではサプリメントとしての利用も一般化している。

FIGURE 1.3: An example of a Chinese–Japanese comparable corpus which describes ‘digestive enzymes’ from Wikipedia.

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

[6] https://zh.wikipedia.org/wiki/消化酶

[7] https://ja.wikipedia.org/wiki/消化酵素

Chinese	**Japanese**	Sim_1	Sim_2
在那里帮助食物的消化	消化に使われる酵素のことである	0.978	0.667
把办公室内的知识支持作为研究的对象。	オフィス内の知識サプリメントを研究の対象とした。	0.731	0.552
还通过探索有关内脏的新型消化酵素，调查了各脏器抽出物。	内蔵に関し新規な消化酵素探索で各臓器抽出物を調べた。	0.821	0.333
介绍了在促进新原油储备的利用和技术研发方面的研究。	新原油備蓄利用の促進·技術開発を紹介した。	0.715	0.311

Figure 1.4: An example of a Chinese–Japanese quasi-parallel corpus.

The quantity and the quality of the parallel sentences are two important factors that strongly impact translation quality. In SMT systems, the translation knowledge is acquired from these parallel sentences. Consequently, the quantity and the quality of the translation relations extracted between words or phrases between the source language and the target language depend on the quantity and the quality of the parallel sentences.

There already exist numerous freely available bilingual or multilingual corpora for European languages. For instance, the Europarl parallel corpus (Koehn, 2005) is a collection of parallel text from the proceedings of the European Parliament. It includes versions in 21 European languages. The aligned multilingual JRC-Acquis corpus (Steinberger et al., 2006) also funded by the European Union, contains resources in 21 European languages.

Currently, there are almost no Chinese–Japanese parallel corpora publicly freely available on all domains for users and researchers. Some research institutions have tried to construct Chinese–Japanese bilingual parallel corpora, for instance, the basic traveler's expression corpus (BTEC) in Japanese, English, and Chinese has been constructed by the Advanced Telecommunications Research Institute International (ATR). It was then extended to over 20 languages. A speech recognition engine was developed based on this corpus (Sakti et al., 2009). The National Institute of Information and Communications Technology (NICT) in Japan created a Japanese–Chinese corpus of 38,383 sentences by selecting Japanese sentences from the Mainichi Newspaper and translating them manually into Chinese (Zhang et al., 2005). Harbin Institute of Technology in China (HIT) constructed the Olympic Oriented Chinese–English–Japanese Trilingual Corpus (Yang et al., 2006) from a Chinese–English parallel corpus collection by adding Japanese translations. This initiative was intended for the development of natural language processing (NLP) for the Olympic Games in Beijing in 2008. The resource consists of 54,043 sentence pairs. Most of the above corpora are not released or freely available, due to copyright problems.

In the last two years, two parallel corpora were released in the domain of scientific papers and patents. They are provided under the condition of participating in the open evaluation campaign Workshop on Asian Translation (WAT)[8] [9]. The first parallel corpus is the Asian Scientific Paper Excerpt Corpus (ASPEC)[10]. It contains 680,000 Japanese–Chinese parallel sentences extracted from scientific papers. It was built within the frame of a four-year project of translating Japanese scientific papers from the literature database and electronic journal site J-STAGE of JST into Chinese after receiving permission from the relevant academic associations (Nakazawa et al., 2014). The second parallel corpus provided for WAT is the JPO corpus[11], created jointly, based on an agreement between the Japan Patent Office (JPO) and NICT. This corpus consists of a Chinese–Japanese and a Korean–Japanese patent description corpus of one million parallel sentences in science and technology divided into four sections. As already mentioned above, for the collection of Chinese–Japanese parallel corpora, an important issue arises from copyright restrictions. Most existing resources are not freely available due to copyright restrictions.

It is worth noticing that the data contained in the mentioned corpora above are translated from one language into another language manually in the frame of long term projects (e.g., the ASPEC corpus) or extracted from the existing article level aligned text via sentence alignment (e.g., the Europarl and JPO corpora). There are also some works for parallel corpora construction by collaborative manner (e.g., the Tatoeba project)[12] or crowdsourcing translation (Zaidan and Callison-Burch, 2011). Automatic extraction or construction of parallel corpus in different domains is research that is indispensable for improving SMT performance, especially for the less-resourced language pair addressed here: Chinese–Japanese. In general, researchers face many difficulties in extracting or constructing parallel corpora from general texts or from specialized texts like patent families.

In our work, In Chapter 4, we propose a different way to construct a quasi-parallel Chinese–Japanese corpus by leveraging a small amount of parallel data and large amounts of unrelated monolingual data and using analogical associations.

1.4 Granularity and Word Segmentation

Many Asian languages like Chinese and Japanese do not have typographic boundaries in their writing systems. Word segmentation (tokenization), i.e., breaking sentences down into individual words (tokens), is normally treated as the first step of preprocessing for natural language processing. For Chinese and Japanese, different rules and segmentation standards lead to different segmentation results at different levels of granularity.

[8] `http://orchid.kuee.kyoto-u.ac.jp/WAT/WAT2014/index.html`
[9] `http://orchid.kuee.kyoto-u.ac.jp/WAT/`
[10] `http://lotus.kuee.kyoto-u.ac.jp/ASPEC/`
[11] `http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/`
[12] `http://tatoeba.org/eng/`

In tokenization for practical tasks, like machine translation, there should be a balance in granularity to keep consistency between Chinese and Japanese. Because choosing the reasonable granularity of tokenization for each pair of sentences is a critical task for both word/phrase alignment and translation accuracy.

In (Zhao et al., 2013), an investigation in the relationship between the choice of segmentation strategy and the improvement of MT is conducted. It is concluded that the segmented corpora and the dictionary that the segmenter relies on is crucial and affects MT performance. Thus, they improve word segmentation for Chinese MT by directly optimizing the dictionary for word segmentation. In (Chang et al., 2008), it is demonstrated that the way different segmentation strategies affect MT is still poorly understood: optimizing segmentation for an existing segmentation standard does not always yield better MT performance. It is found that segmentation granularity and consistency of Chinese word segmentation are very important. They optimize segmentation granularity and improve segmentation consistency using an external lexicon and proper noun features to improve Chinese–English translation accuracy.

In (Bai et al., 2008), they improve word alignment (obtain more 1-to-1 mapping tokens) by adjusting Chinese word segmentation in Chinese–English MT. So as to significantly improve the performance of word alignment, two methods are used: learning affix rules from a Chinese–English bilingual terminology bank and using the concept of impurity measure motivated by a decision tree. In (Wu and Wang, 2004), to improve word alignment on a small-scale domain-specific bilingual corpus, they combine the use of word alignment based on a large-scale corpus in the general domain and a small-scale corpus in a specific domain. They improve the domain-specific word alignment by combing these two statistical word alignment models.

In our work, we shall improve translation accuracy for patent SMT by adjusting the granularity and keeping consistency between Chinese and Japanese training corpus using extracted bilingual terms but without the use of any additional lexicon, bilingual terminology bank or additional corpus.

In (Xu et al., 2004), an investigation in word segmentation is performed in relation to Chinese translation quality. They perform Chinese word segmentation (only one side of parallel corpus) using a self-trained domain-specific Chinese dictionary from an existing Chinese–English training corpus. In (Ma and Way, 2009), the correspondence between source and target languages in Chinese and English (bilingually motivated segmentation process) is examined. They make use of bilingual corpora and statistical word alignment techniques. Compared with (Xu et al., 2004), they focus on more specific domain translation tasks and avoid the use of monolingual segmenters in order to improve the segmentation across different domains. In our work, we shall use a similar idea. We shall adjust Chinese and Japanese tokenization at the same time for technical training corpus in different domains for SMT, but only around bilingual multi-word terms.

In Chinese–Japanese technical machine translation, a corpus may contain large amounts of domain-specific terms in words or multi-word expressions. This brings up the question of tokenization. In some fields, it may be harmful to tokenize some words in patents or scientific texts, as they may just be parts of more complex single units like terms. But it is uneasy to control the tokenization of multi-word terms, as it may happen that multi-word terms are blindly segmented into several single words in one language but are not segmented in the other language. It may also happen that some of the multi-word terms have different levels of granularity due to different conventions of segmentation for different languages.

Figure 1.5 shows examples of segmentation results for a Chinese–Japanese sentence pair in the chemical domain based on different segmentation tools. We use Standford (Tseng et al., 2005), Urheen (Wang et al., 2010a) and KyTea (Neubig et al., 2011) for Chinese segmentation, Juman[13], Mecab (Kudo, 2005) and KyTea (Neubig et al., 2011) for Japanese segmentation. For the Stanford Chinese segmentation tool, there are two models with two different segmentation standards used in segmentation: the Chinese Penn Treebank (ctb) standard[14] and the Peking University (pku) standard[15].

Chinese sentence	这是因为水与异氰酸酯基反应，以形成脲键。
Japanese sentence	これは、水とイソシアネート基が反応することで、ウレア結合が生じるためである。
Meaning	‘This is because of the reaction between water and isocyanate groups for forming urea bonds.’
by Stanford (ctb)	这/是/因为/水/与/异氰/酸酯基/反应/，/以/形成/脲键/。
by Stanford (pku)	这/是/因为/水/与/异氰/酸酯/基/反应/，/以/形成/脲键/。
by Urheen	这/是/因为/水/与/异氰/酸酯基/反应/，/以/形成/脲键/。
by KyTea	这/是/因为/水/与/异/氰酸/酯/基/反应/，/以/形成/脲/键/。
by Juman	これ/は/、/水/と/イソシアネート/基/が/反応/する/こと/で/、/ウレア/結合/が/生じる/ため/である/。
by Mecab	これ/は/、/水/と/イソシアネート/基/が/反応/する/こと/で/、/ウレア/結合/が/生じる/ため/で/ある/。
by KyTea	これ/は/、/水/と/イソシアネート/基/が/反応/する/こと/で/、/ウレア/結合/が/生じ/る/ため/で/あ/る/。

Figure 1.5: Segmentation results for a Chinese–Japanese sentence pair using different segmentation tools.

From these examples, we see that different segmentation tools used in Chinese and Japanese lead to different segmentation results with different numbers of tokens for terms across languages. In other words, there are inconsistencies of segmentation results in different levels of granularity between Chinese and Japanese terms. For instance, the term 异氰酸酯基 ‘isocyanate group’ in Chinese is segmented as 异/氰酸/酯/基 (4 tokens) by KyTea, but it is segmented as イソシアネート/基 (2 tokens) by any segmentation tool used in Japanese.

Even for the same language (Chinese), different segmentation tools lead to different levels of granularity in terms. For instance, the term 异氰酸酯基 in Chinese is segmented as 异氰/酸酯/基 by Stanford (pku), 异氰/酸酯基 by Stanford (ctb)/Urheen and

[13] `http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN`

[14] `http://www.cis.upenn.edu/~chinese/segguide.3rd.ch.pdf`

[15] `http://sighan.cs.uchicago.edu/bakeoff2005/data/pku_spec.pdf`

异/氰酸/酯/基 by KyTea. The same segmentation tools based on different models with different segmentation standards also may produce different levels of granularity. For instance, the term 异氰酸酯基 in Chinese is segmented as 异氰/酸酯基 by Stanford (ctb) and 异氰/酸酯/基 by Stanford (pku).

In these examples, whatever segmentation tool is used in Chinese or Japanese, actually, they do not have any correspondence in word-to-word alignments for terms. Another case in these examples is that some terms are single-word terms in one language but multi-word terms in another language. For instance, the single-word term 脲键 'urea bond' (one token) segmented by Stanford (ctb)/Stanford (pku)/Urheen in Chinese and the multi-word terms ウレア/結合 (2 tokens) segmented by any tool in Japanese. In SMT, these cases lead to inconsistencies in word/phrase alignment and negatively affect the accuracy of translation. Thus, for keeping the direct and exact translations between Chinese and Japanese terms, re-tokenization (re-segmentation) centered around terms is necessary. For the above examples, the expected tokenization (segmentation) results for the two pairs of terms across languages are: 异氰酸酯基 (Chinese) to イソシアネート基 (Japanese) and 脲键 (Chinese) to ウレア結合 (Japanese).

1.5 Big Picture and Overview of Our Approach

We propose an integrated framework for our research in this dissertation. The overview of our approach is presented in a big picture given in Figure 1.6. The main focus of this dissertation is to exploit several freely available linguistic resources to address the scarcity of linguistic data and technical term translation in specific domains between Chinese and Japanese.

Firstly, for solving the problem of scarcity of bilingual lexica between Chinese and Japanese, we construct a Chinese–Japanese word-to-word lexicon based on freely available Chinese–English and Japanese–English lexica ((Chapter 2) in Figure 1.6). Using the constructed lexicon, we may solve the unknown word problem directly in the decoding process of SMT.

Secondly, for solving the problem of the translation of terms in technical domains, we extract bilingual multi-word terms from an existing training corpus. These extracted terms are used in adjusting and balancing the tokenization between Chinese and Japanese technical data ((Chapter 3) in Figure 1.6).

Thirdly, for solving the problem of scarcity of bilingual corpora between Chinese and Japanese, we construct a quasi-parallel corpus of sentences using analogical associations based on freely available Chinese and Japanese monolingual data. This kind of quasi-parallel sentences are used as additional training data in SMT. They help in acquiring more potential useful translation knowledge from the inflated training corpus ((Chapter 4) in Figure 1.6).

Finally, we combine and apply all above approaches to further improve the translation accuracy of Chinese–Japanese SMT ((Section 4.5) in Figure 1.6). We make use of the kanji-hanzi conversion method through our work in this dissertation, because in the Chinese and Japanese writing systems, there exist large amounts of characters shared with the same meaning. They can be considered as a linguistic clue to align words or multi-word expressions. Table 1.2 shows the overview of our work in each chapter for solving different problems in technical translation between Chinese and Japanese.

Table 1.2: The overview of our work in the dissertation.

Chapter	**granularity**	**Equivalence**	**Problem addressed**
Chapter 2	characters/words	lexical graphical meaning	scarcity of open bilingual lexicon
Chapter 3	terms	technical	+ translation of terms in technical domain due to inconsistency in segmentation + scarcity of open term banks
Chapter 4	short sentences	similar translation	scarcity of open bilingual corpora

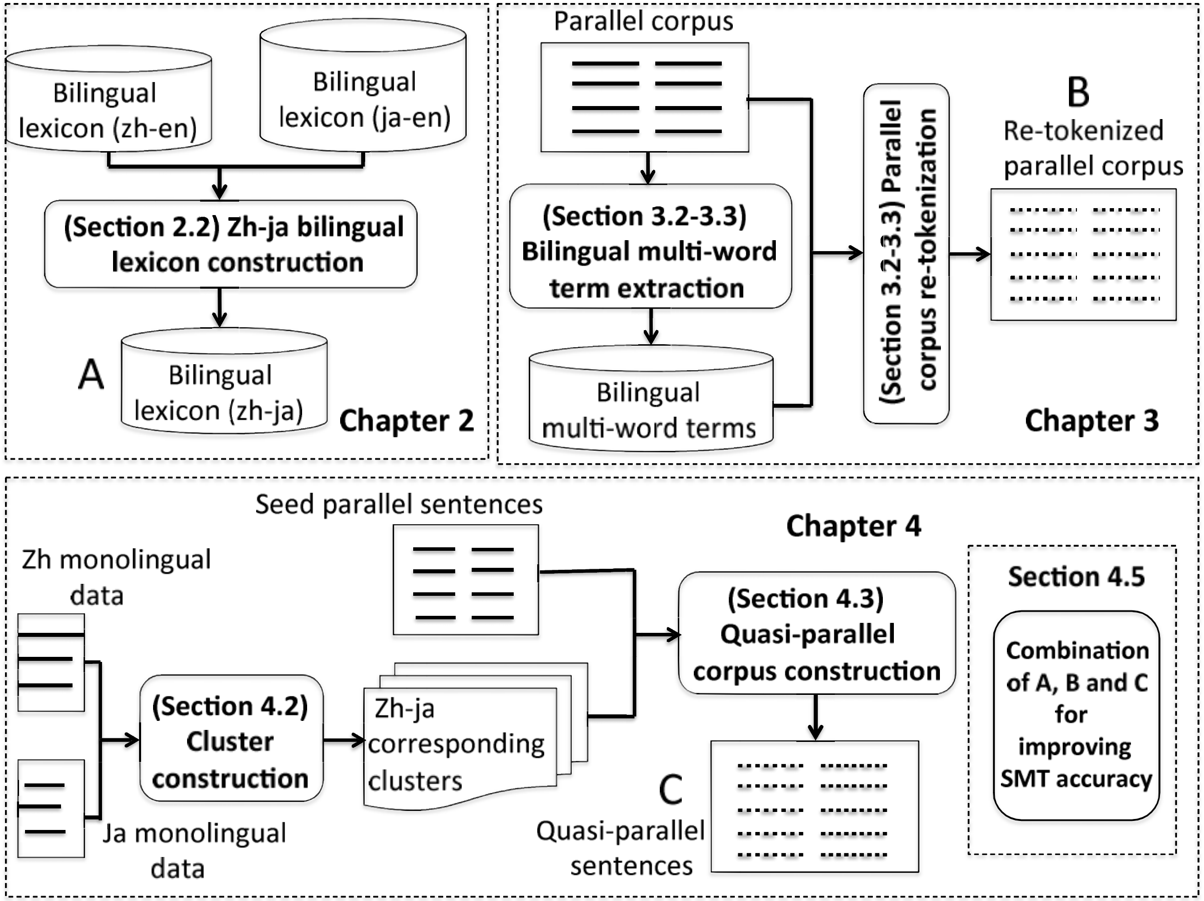
Bilingual lexicon (zh-en)
Bilingual lexicon (ja-en)
(Section 2.2) Zh-ja bilingual lexicon construction
A
Bilingual lexicon (zh-ja)
Chapter 2
Parallel corpus
(Section 3.2-3.3) Bilingual multi-word term extraction
Bilingual multi-word terms
(Section 3.2-3.3) Parallel corpus re-tokenization
B
Re-tokenized parallel corpus
Chapter 3
Seed parallel sentences
Chapter 4
Zh monolingual data
Ja monolingual data
(Section 4.2) Cluster construction
Zh-ja corresponding clusters
(Section 4.3) Quasi-parallel corpus construction
C
Quasi-parallel sentences
Section 4.5
Combination of A, B and C for improving SMT accuracy

Figure 1.6: Big picture and overview of our work in this dissertation.

1.6 Outline of This Dissertation and Research Contributions

The organization of the dissertation is as follows.

In Chapter 1, we describe the background and the basic knowledge of the research: statistical machine translation (SMT), parallel corpora and non-parallel corpora, as well as less-resourced languages and language pairs. We also introduce the basic notions in Chinese–Japanese technical translation, different levels of granularity and segmentation. Moreover, we give the overview of our approach and present the contributions of this dissertation.

In Chapter 2, we propose a method to construct a Chinese–Japanese lexicon by combining several automatic techniques on several freely available resources. The basic technique used is the classical pivot language technique. To improve the quality of the resource built, we combine three additional different techniques: one time inverse consultation; Japanese kanji to Chinese hanzi character conversion; expansion through a Chinese synonym table. The Chinese–Japanese lexicon built consists of more than 45,000 Chinese-Japanese word pairs with an accuracy of 85%. This work can be used to enrich the Chinese–Japanese training corpus to train a translation model or directly used in decoding of statistical machine translation.

The main contributions of the work described in this chapter are: 1) Combine several automatic methods to construct a Chinese–Japanese bilingual lexicon 2) Propose an approach using kanji-hanzi conversion and a Chinese synonyms table to improve the quality and quantity of the resource built. 3) In this work, all of the resources used are free available. In addition, the kanji-hanzi conversion method is used in bilingual multi-word term extraction (Chapter 3) and quasi-parallel corpus construction (Chapter 4).

Chapter 3 describes a method to improve Chinese–Japanese statistical machine translation of patents by re-tokenizing the training corpus with aligned bilingual multi-word terms. We try two experimental protocols to make use of the extracted bilingual terms in Chinese–Japanese SMT experiments and find the better one. We obtain a high quality of correspondence with 93% in bilingual term extraction and a significant improvement of 1.5 BLEU score in a translation experiment. We also consider the terms for which one side is a single-word term and the other side is a multi-word term. We combine using the kanji-hanzi conversion method, and obtain even better results in BLEU with 2 point improvement.

The main contributions of the work described in this chapter are: 1) We automatically extract multi-word terms from monolingual corpora using statistical and linguistic filtering methods. We propose an automatic alignment method to identify corresponding terms. The most promising bilingual multi-word terms are extracted by setting some threshold on translation probabilities and further filtering. 2) We also use kanji (Japanese)–hanzi (Chinese) character conversion to confirm and extract more promising bilingual multi-word terms. 3) This work is helpful for improving the translation accuracy for Chinese–Japanese patent or scientific corpora by adjusting and balancing the granularity of segmentation results around terms. 4) This work improves the performance of SMT not only on small-scale scientific training sets, but also for large-scale training sets.

In Chapter 4, we propose a method to construct a quasi-parallel corpus by using analogical associations based on large amounts of monolingual data and a small amount of parallel data, so as to improve Chinese–Japanese statistical machine translation quality. These quasi-parallel corpora are added to an existing training corpus to address the shortage of parallel corpora between Chinese and Japanese. The best result that we obtain is a very significant improvement of 6 BLEU points over a Chinese–Japanese baseline system.

The main contributions of the work described in this chapter are: 1) Construct a quasi-parallel corpus with freely and easily accessible monolingual data and an existing small number of parallel corpus. 2) Generated new quasi-parallel corpus using analogical associations. 3) Filtered over-generated sentences using two filtering methods, the N-sequence method and the BLEU method, independently or combined.

In Chapter 4, we also combine several proposed techniques described in previous chapters and this chapter in statistical machine translation. The evaluation result in BLEU (1.8 point statistically significant improvement) shows that the combination of our proposed methods and obtained results are helpful and effective for improving translation accuracy in technical domain.

In Chapter 5, we summarize and conclude this dissertation. Future directions are also presented. In summary, the main points of this dissertation and research work are as follows:

To solve the problem of the translation of large amounts of terms (due to different segmentation granularity) in Chinese–Japanese statistical machine translation in specific domains, adjusting the granularity of segmentation results of these terms in training corpus is necessary. To solve the translation quality problem of less-resourced language pairs, like we address here with Chinese–Japanese, the most natural answer is to build larger and larger aligned training data, that is to make those language pairs well-resourced.

Thus, we propose several approaches to improve Chinese–Japanese statistical machine translation accuracy with lexicon, bilingual multi-word terms and quasi-parallel corpus. We make use of these different automatically constructed or extracted resources as additional data in phrase-based statistical machine translation systems and improved translation accuracy.

In our work, we provide experimental results that show that it is possible to obtain additional quasi-parallel corpus using monolingual data and an existing small number of training data. We are also able to extract bilingual multi-word terms from an existing training corpus and make use of them in adjusting and balancing the segmentation results on the training corpus. The originality of this work is the possibility and efficiency of the proposed techniques in constructing new data and re-tokenizing training corpus for statistical machine translation. The characteristics of our methods are to make use of existing data as much as possible; adjusting tokenization for Chinese–Japanese technical corpus for both languages at the same time, i.e., not only for one side of a parallel corpus, or do not consider the technical term translation in some previous work.

Chapter 2

Chinese and Japanese Characters

Because the Chinese and Japanese writing systems share a large amount of characters with same or similar meanings, they can be considered as a linguistic clue to align words or multi-word expressions. For addressing the problem of the scarcity of open, machine-readable Chinese–Japanese lexicon, we make use of freely available Chinese–English and Japanese–English dictionaries to construct a Chinese–Japanese lexicon. We make use of it to translate words in the test data, primarily to solve the problem of unknown words that cannot be successfully translated by an existing translation system.

This chapter[1] focuses on a kanji-hanzi conversion method. Also in this chapter, we show how to combine several automatic techniques to build a Chinese–Japanese lexicon from freely available resources. This lexicon can be used in statistical machine translation.

The structure of this chapter is as follows.

- Section 2.1 introduces the relationship between written Chinese and Japanese. We also introduce some freely available resources and tools for kanji-hanzi conversion.
- Section 2.2 presents the basic method to generate a Chinese–Japanese bilingual lexicon via English as the third language by joining two bilingual lexical resources for Chinese–English and Japanese–English. We describe a first additional method: one time inverse consultation (Tanaka and Umemura, 1994). Compared with the classical joining approach, it increases the accuracy. The second additional method: consists in using kanji-hanzi conversion and comparison between Chinese words and Japanese words; third and a last improvement is achieved by expansion through Chinese synonym table. By combining these three additional methods, we increased the number of translation candidates and the accuracy of the resulting Chinese–Japanese lexicon.
- We summarize this chapter in Section 2.3.

[1] Related to (Yang and Lepage, 2012)

2.1 Relationship between Written Chinese and Japanese

2.1.1 Hanzi and Kanji Ideograms

There exist a large amounts of variants of Sinitic languages or dialects, for instance, Putonghua (standard Mandarin), Cantonese, Hokkien, Chaozhou, Hakka, etc. For standard writing systems, they adopt in Chinese characters (hanzi). There are two types of writing systems for Chinese: simplified and traditional Chinese. Thus, there are two styles of characters (hanzi) used: simplified characters and traditional characters. We can easily imagine that simplified Chinese characters are just a simplified version of traditional Chinese characters. The number of strokes used in simplified Chinese characters is reduced in comparison with the traditional Chinese characters. Somewhat, it makes writing and remembering easier for people learning Chinese. Today, simplified Chinese characters are used in mainland China and Singapore, and traditional Chinese characters are used in Taiwan, Hong Kong and Macau. Table 2.1 shows examples of the characteristics between simplified Chinese and traditional Chinese in different cases:

- There are Chinese hanzi characters used in simplified Chinese but not in traditional Chinese, and characters used in traditional Chinese but not in simplified Chinese. These characters need to be converted by mapping them. For instance, 爱 'love' is only used in simplified Chinese and 愛 'love' is only used in traditional Chinese.
- There are Chinese hanzi which are used in both simplified Chinese and traditional Chinese without any change. For instance, hanzi 初 which means first or early is both used in simplified Chinese and traditional Chinese.
- A Chinese hanzi is mapped to itself or to another character in simplified to traditional conversion depending on context. For instance, the hanzi character 斗 has two traditional Chinese mapping characters: itself 斗 (with the meaning of a measure for grain) and 鬥 (with the meaning of fight). Conversely, when converting from traditional Chinese to simplified Chinese, two hanzi are mapped to one only.
- Similar but different from the case above, there are also a one-to-many mappings in simplified to traditional Chinese conversion but the multiple mapping in traditional Chinese are different depending on context but not including the simplified character itself. For instance, the simplified character 脏 has two different corresponding traditional characters: 臟 when it means viscera and 髒 when it means dirty.
- A Chinese hanzi used as both simplified Chinese and traditional Chinese with different meanings. For instance, hanzi 苧 used in traditional Chinese, it is pronounced /zhù/ and means a kind of nettle. Its simplified form is 苎. But when it is used as a simplified Chinese, it is pronounced /níng/ and means limonene, its traditional Chinese is 薴.

Table 2.1: Comparison of simplified Chinese and traditional Chinese.

Simplified Chinese	Traditional Chinese	Meaning of the Chinese hanzi in English
爱	愛	love
败	敗	fail, lose
变	變	change
财	財	wealth, property
彼	彼	that, those
比	比	compare
初	初	first, early
丰	丰/豐	abundant / good-looking
后	后/後	queen / later
斗	斗/鬥	a measure for grain / fight
脏	臟/髒	viscera / dirty
苎	苧	a kind of nettle
苧	薴	limonene

From the examples given above, we understand that the relations between simplified and traditional Chinese characters are diverse. We now turn to the investigation of the relationship between written Chinese (hanzi) and Japanese (kanji).

Most of the kanji ideograms were original created in ancient China. Many hanzi and kanji ideograms shared and look similar in Chinese and Japanese. Most of these characters express the same meaning (Table 2.2), although the number of hanzi used in Chinese is larger than the number of kanji used in Japanese. Of course, some words made up of hanzi and kanji which are shared and look similar in Chinese and Japanese do not have the same meanings in both languages. Table 2.2 shows some examples of such cases.

Table 2.2: Comparison of simplified Chinese, traditional Chinese and Japanese in character and word granularity.

Chinese hanzi (simplified)	Chinese hanzi (traditional)	Meaning	Japanese kanji	Meaning
基	基	basic	基	basic
数	數	number	数	number
剂	劑	agent	剤	agent
中央	中央	news	中央	center
构造	構造	construction	構造	construction
新闻	新聞	**news**	新聞	**newspaper**
结构	結構	**structure**	結構	**well, fine**

Some changes and developments from hanzi to kanji characters occurred when they were transmitted to Japan. Some kanji, named kokuji (the number of kokuji: about 30 characters) were also created in Japan. They do not exist in the original Chinese writing system. Some of the kokujis also have its corresponding hanzi to show up (see Table 2.3).

Table 2.3: Comparison of simplified Chinese, traditional Chinese and some of the Japanese kokujis.

Chinese hanzi (simplified)	Chinese hanzi (traditional)	Meaning	Japanese kokuji	Meaning
卡	卡	calorie, clip, checkpost	峠	mountain pass
田	田	cultivated field, cropland	畑	cultivated field, cropland
枥	櫪	manger	栃	horse chestnut
神	神	deity, god	榊	evergreen tree used in a Shinto ritual

During the Meji period, Japan began to adopt Western culture, technology, medical science, economy, philosophy and so on. A large amount of words written in kanji were created in Japan during that period and were re-imported back to China to be widely used. Even nowadays, there still constantly exists creation of words in Japan that come into China. Instead of creating translations from Japanese to Chinese, it is preferred to share the same meaning across the two languages by using the same hanzi/kanji. Such examples of words created in Japan and adopted in China are: 哲学 'philosophy', 理性 'rational', 感性 'perceptual', 意識 'consciousness', 科学 'science', 物理 'physics', 化学 'chemistry', 分子 'molecule', 原子 'atomatom', 時間 'time', 空間 'space', 理論 'theory', 文学 'literature', 美術 'art', 主観 'subjectivity', 客観 'objectivity', 写真 'photo', 料理 'deal with', 達人 'talent', 物語 'story', 違和感 'something wrong'.

2.1.2 Freely Available Resources and Tools

The Unihan database[2] is a database for the Unicode Consortium's collective knowledge for the Chinese–Japanese–Korean (CJK) Unified Ideographs contained in the Unicode Standard.

The Unihan database contains a number of categories with different properties for Han ideographs in the Unicode Standard, for instance, readings, structural analyses, definitions and so on.

There are several fields in the Unihan database for each category. These fields are divided according to the purpose they fulfill for ideographs. For instance, the "Unihan_Variants" category includes traditional-to-simplified variation, simplified-to-traditional variation and so on.

In our work, we make use of "ksimplified_variant" (SimplifiedVariant) in the "Unihan_Variants" category to convert traditional Chinese characters into simplified Chinese characters. This field only contains characters used in traditional Chinese, not simplified Chinese. It allows us to obtain the equivalence in characters between Chinese and Japanese in cases where the Japanese character is the same as the traditional Chinese character. There are 3,662 SimplifiedVariant pairs.

[2] http://www.unicode.org/Public/UNIDATA/

Langconv[3] is a tool for simplified-traditional conversion between Chinese and Japanese. The database used by Langconv comes from Wikipedia. It provides simplified Chinese to traditional Chinese conversion and traditional Chinese to simplified Chinese conversion. In our work, we make use of this database for traditional Chinese to simplified Chinese conversion. This database does not only include the conversion between characters, but also includes the conversion between words. For instance, 权限 'permission': 許可權; 接口 'interface' : 介面; 便携式 'portable': 攜帶型. It contains a Wiki traditional-simplified conversion database, consisting of about 3,000 traditional to simplified conversion pairs.

The Hanzi-kanji Conversion Table[4] is a conversion table between simplified Chinese hanzi and Japanese kanji. We use a hanzi-kanji conversion table which consists of 2,236 simplified hanzi-kanji pairs in characters.

The Chinese encoding converter[5] is an open source tool that converts between traditional Chinese and simplified Chinese. The conversion database contains simplified-traditional pairs in Chinese, including the characters which are the same in both simplified and traditional Chinese. It contains 6,740 corresponding simplified-traditional Chinese pairs in characters.

Kanconvit[6] is a freely available tool for conversion between Japanese kanji and simplified Chinese. There are 1,158 pairs of kanji-hanzi conversion data used in this tool.

In our work, we use the first three sources or tools (Unihan database, Langconv and Hanzi-kanji Conversion Table) in our method for kanji-hanzi conversion. Because the SimplifiedVariant filed in Unihan database only contains the traditional-simplified conversion only for those characters are different used in traditional and simplified Chinese. The Hanzi-kanji Conversion Table contains more hanzi-kanji conversion pairs compare with Kanconvit.

2.1.3 Discussion

Kanji is one of the three character sets (the other two sets are katakana and hiragana) used in Japanese. It is normally used to write content words, such as nouns, adjective stems and verb stems. Hanzi in Chinese and Kanji in Japanese can be considered as semantic clues to connect Chinese and Japanese, because they are ideographic characters that partly describe their meanings. Although, some of the kanjis (hanzis) have been modified and developed after they were imported from China to fit the Japanese language. We also find corresponding hanzi in the Chinese language which share the same meaning. All these relations and characteristics between Chinese and Japanese

[3] http://code.google.com/p/advanced-langconv/source/browse/trunk/langconv/?r=7

[4] https://www.kishugiken.co.jp/cn/code10d.html

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

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

characters allow us to construct a Chinese–Japanese lexicon (Chapter 2), extract bilingual Chinese–Japanese terms (Chapter 3) and compute similarity between two sets of changes between Chinese and Japanese sentences (Chapter 4).

2.2 Free Resource-based Lexicon Construction

2.2.1 Related Work

In the development of machine translation and cross-language information retrieval systems, it is necessary to construct machine-readable bilingual dictionaries from one language to another, but the cost is enormous from the viewpoint of labor and time.

However, even if bilingual dictionaries do not directly exist for a particular source language and a particular target language, the possibility is high that bilingual dictionaries exist into an identical third language, particularly English nowadays. In other words, it is conceivable that a bilingual dictionary between Chinese and Japanese be built through a third language, like English.

In addition, Japanese kanji are similar to Chinese hanzi. We propose, relying on the similarities between kanji and hanzi, to compare the Unicode of Chinese words with that of Japanese words in kanji-hanzi conversion.

The existence of a large amount of characters sharing the same meaning in the Chinese and Japanese writing systems can be considered as a linguistic clue to align words or multi-word expressions. Many studies have exploited common Chinese and Japanese characters. In (Goh et al., 2005), they build a Japanese–Simplified Chinese dictionary consisting of kanjis which are identical to traditional Chinese and associate the corresponding simplified Chinese character to it. In (Tan and Nagao, 1995), they use the occurrence of identical common Chinese characters in Chinese–Japanese in the sentence alignment task.

In our work, different from previous work, we constructed a Chinese–Japanese lexicon by combining several automatic techniques on several freely available resources. The basic technique used is the classical pivot language technique. To improve the quality of the resource built, we combined three additional different techniques: one time inverse consultation; Japanese kanji to Chinese hanzi character conversion; expansion through Chinese synonyms table. We used the Unihan database, Langconv tranditional-simplified conversion data (tool), Hanzi-kanji Conversion Table and Chinese synonym table as resources.

2.2.2 Construction with a Classical Pivot Language Technique

In this section, first we will describe the Chinese–English and Japanese–English dictionaries we use, and then how we join them via English as the pivot language.

The XDXF dictionary[7] is a project to unite all existing open dictionaries and provide both users and developers with universal XML-based format, convertible to and from other popular dictionary formats. The Chinese–English XDXF dictionary consists of 43,433 articles, each article consists of three main components: (1) both traditional Chinese and simplified Chinese; (2) pronunciation in pinyin; (3) English translations. The Japanese–English XDXF dictionary we use consists of 108,473 articles. Some articles consist of the pronunciation in katakana, especially for those Japanese words made up of kanji only. From these two dictionaries we extract simplified Chinese and Japanese words only with their corresponding English translations. We make use of these two generated lexica as our experimental primary resources. After eliminating duplicate lines, we obtained a Chinese–English and Japanese–English lexica consisting of 43,389 and 105,182 entries respectively.

In a first step, we proceed as follows:

- Firstly, convert Chinese–English and Japanese–English dictionaries into lexicon resources.
- Secondly, output the phrase translation tables (using Anymalign (Lardilleux and Lepage, 2009)) corresponding to Chinese–English and Japanese–English lexica by computing translation probabilities.
- Thirdly, perform a join of the two phrase translation tables through English as the pivot language and compute probabilities to get a Chinese–Japanese phrase translation table. Here the join is the same as the algebraic operation in relational databases.

Hereafter, zh, en, and ja denote terms in the source language Chinese, in the pivot language English, and in the target language Japanese respectively. For the translation pairs (zh, en) and (en, ja), the translation probabilities $P(en|zh)$ and $P(ja|en)$ are computed using the maximum likelihood estimation from the co-occurrence frequencies that are consistent with the word alignment in the phrase translation table:

$$P(en|zh) = \frac{P(zh, en)}{P(zh)} = \frac{C(en \leftrightarrow zh)}{C(zh)} \tag{2.1}$$

$$P(ja|en) = \frac{P(en, ja)}{P(en)} = \frac{C(ja \leftrightarrow en)}{C(en)} \tag{2.2}$$

[7] `http://xdxf.revdanica.com/down/`

In the equations, $C(x)$ denotes the number of occurences of the word or phrase x in the lexicon, and $C(x \leftrightarrow y)$ is the number of co-occurrences of x and y in the lexicon. In theory we calculate the direct translation probabilities between the source language Chinese and target language Japanese by the following equation (e.g., for the probability of the target language Japanese knowing the source language Chinese):

$$P(ja|zh) = \sum_{all\ pivot\ en} P(ja|en) \times P(en|zh) \tag{2.3}$$

One of the characteristics of using this approach is that we obtain all possible alignments as a result in a phrase translation table. We can normalize translation probabilities by discarding any translation pair with both translation probabilities less than a threshold. The threshold we used was 0.05 for both $P(zh|ja)$ and $P(ja|zh)$. We obtained a Chinese–Japanese lexicon consisting of 119,203 pairs.

We extract ten samples with 100 translation alignments randomly and check manually in an existing bilingual dictionary. We calculate the accuracy of the result by p-value using Student's t-test:

$$T = \frac{(X - H0)}{S} \times \sqrt{n-1} \tag{2.4}$$

With a null hypothesis of 45%, and an experimental result of 42.6%, the p-value is 0.06, above the usual 0.05. We conclude that there is not enough evidence to state that the overall translation quality is higher than 45%. We infer that the quality lies below 45% of correct entries, or is even worse.

2.2.3 Using One Time Inverse Consultation

In previous work, Tanaka and Umemura (1994) used the inverse consultation method through English as a pivot language to improve a Japanese–French lexicon built using the join method. We also rely on one time inverse consultation to find suitable equivalents for our Chinese–Japanese lexicon. We proceed as follows: first look up English translations of a Chinese word, then look up Japanese translations of these English translations; for each Japanese translation, look up how many English translations shared with the original Chinese word. The more matches there are, the better the Japanese translation candidate is. Figure 2.1 illustrates one time inverse consultation between Chinese and Japanese.

To measure the quality of a Japanese translation candidate, a similarity score is calculated according to a classical Dice coefficient formula:

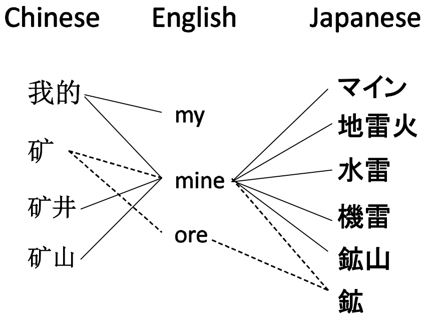

FIGURE 2.1: Sample of one time inverse consultation between Chinese and Japanese using English as a pivot language.

$$SimilarityScore = \frac{2 \times |E(C) \cap E(J)|}{|E(C)| + |E(J)|} \tag{2.5}$$

Here $E(C)$ and $E(J)$ are the sets of English translations for the Chinese word and the Japanese word respectively.

Due to the relatively small sizes of the two lexica we use, a similarity score equal to one does not necessarily mean that a translation pair is correct. As shown in Figure 2.1, the similarity scores of "矿–鉱", "矿井–マイン" and "矿山-水雷" are all equal to one, but only "矿–鉱" is a correct translation pair. Using this method may lead to generate many irrelevant translation candidates.

In our experiment we obtained 33,297 translation candidates. A same p-value evaluation of the results showed an accuracy of 76%. Compared with standard classical pivot technique, the number of translation candidates was reduced, but the accuracy was increased. However, the problem of distinguishing ambiguous words was not solved completely.

2.2.4 Increasing Translation Candidates by Using Kanji-Hanzi Conversion

Figure 2.1 leads to the observation that some translation pairs can be directly retrieved or reinforced by looking at the similarity between hanzi and kanji. In this figure, the pair "矿山–鉱山" is supported by the kanji-hanzi conversion of the first element "矿–鉱".

Consequently, we propose to convert Japanese words made up of only Japanese kanji into simplified Chinese characters through kanji-hanzi conversion. By doing so, we generate a ja'–ja file automatically where each line consists in the converted Japanese word (simplified Chinese) and the original Japanese word. In this way, we avoid the difficult problem of converting Chinese simplified characters back to Japanese kanji (Goh et al., 2005). By comparing ja' with the Chinese entries in the Chinese–English lexicon we can select more reliable Chinese–Japanese translation pairs.

Below, we explain how we combine three sources of data for our conversion experiments so as to maximize the result quality and quantity.

We also describe a method based on the use of a Chinese synonym table[8] to increase the candidates for those different words which share a similar meaning in Chinese and Japanese after kanji-hanzi conversion.

We combined three different sources of data to maximize our conversion results. Table 2.4 shows the relationships between Chinese (traditional and simplified) and Japanese. The Japanese words made up of kanji in the parts "All same" and "TC different" (Traditional Chinese different) could compare the Unicode with Chinese directly without any conversion; the characters in "SC different" (Simplified Chinese different) become comparable by traditional Chinese to simplified Chinese conversion; for the "All different" and "Ja different" parts we propose to utilize Hanzi-kanji Conversion Table (簡体字と日本漢字対照表[9]) to make them comparable with Chinese.

Table 2.4: Relationship between Chinese hanzi and Japanese kanji.

Relationship	All same		TC different		SC different		All different		Ja different	
	word	center	conuntry	learn	struct	wind	value	fight	multiplication	flame
Japanese	世界	中央	中国	数学	構造	風	価値	戦闘	乗法	火焔
T Chinese	世界	中央	中國	數學	構造	風	價值	戰鬥	乘法	火焰
S Chinese	世界	中央	中国	数学	构造	风	价值	战斗	乘法	火焰

The first source of data we used is the Unihan database[10]. In particular we use the correspondence relation SimplifiedVariant in the "Unihan_Variants" of the Unihan database. There are 3,662 SimplifiedVariant pairs. Using them, we could check translation pairs between Japanese words (made up of kanji) and simplified Chinese words (made up of hanzi) in the following way. For each Japanese character, consider it as a traditional Chinese character, and look up for its corresponding simplified Chinese character through the SimplifiedVariant relation and replace it. If this simplified Chinese word (converted Japanese word) is the one in the Chinese-English lexicon, confirm the translation pair.

The second source of data we used is that of the Langconv traditional-simplified Conversion[11] data (tool). It contains a Wiki traditional-simplified conversion database, consisting of about 3,000 traditional to simplified conversion pairs. We perform similar experiments as above to confirm Chinese–Japanese translation word pairs.

[8] `http://ishare.iask.sina.com.cn/f/21267706.html`

[9] `http://www.kishugiken.co.jp/cn/code10d.html`

[10] `http://www.unicode.org/Public/UNIDATA/`

[11] `http://code.google.com/p/advanced-langconv/source/browse/trunk/langconv/?r=7`

The third source of data we used concerns the case where the characters in Japanese are neither found in the traditional Chinese nor simplified Chinese character sets. For this case, we use a Hanzi-kanji Conversion Table which consists of 2,236 simplified hanzi and kanji pairs. We use this table same as described above for the two previous sources of data.

Table 2.5 shows the result of kanji-hanzi conversion using these three sources of data. There exist about 62,852 Japanese entries made up of kanji only from the Japanese–English lexicon, 36,590 Japanese words were converted successfully. For all Japanese words we confirm their simplified Chinese translations: 8,137 translation pairs were confirmed. The accuracy is 98.5%, which shows that the method is quite efficient.

Table 2.5: Result of kanji-hanzi conversion and Chinese (zh)–Japanese (ja) lexicon construction.

Method	Successful conversion	Zh–ja lexicon	Accuracy
(a) Unihan database	27,929 (44.4 %)	6,856	98.0 %
(b) Langconv	28,153 (44.8 %)	6,877	98.5 %
(c) Hanzi-kanji Conoversion Table	36,035 (57.3 %)	8,012	98.5 %
Combining results (a) + (b) + (c)	36,590 (58.2 %)	8,137	98.5 %

2.2.5 Expansion through Chinese Synonym Table

The last method we use to improve the quality of our Chinese–Japanese lexicon is to use a Chinese synonym table to extract more translation candidates for words in Chinese and Japanese that share similar meaning. Again, this applies for Japanese words consisting only of kanji, after conversion into simplified Chinese characters.

The source of data we used consists of 17,170 Chinese synonym pairs. For each Chinese word found in the synonym table, we checked whether the Japanese word (converted) appears as a synonym in the table. This allows to confirm translation pairs. Using this method, we obtained a Chinese–Japanese lexicon consisting of 3,952 pairs. The accuracy was shown to reach 98.5%, which shows the efficiency of this method.

2.3 Summary of This Chapter

In this chapter, we combined different methods and different sources of data to construct a Chinese–Japanese lexicon. We basically joined two bilingual lexica sharing a pivot language, English. The accuracy of the resulting Chinese–Japanese lexicon was improved by using three additional methods:

1. one time inverse consultation through the pivot language, English;
2. Japanese kanji to Chinese hanzi character conversion, using three different sources of data;
3. expansion through a Chinese synonym table.

The combination of these three additional methods produced the final translation candidates of our resultant lexicon. In total, we obtained 45,386 translation pairs. Among the 12,089 candidate pairs obtained using kanji-hanzi conversion (8,137) and Chinese synonym table (3,952), 1,399 already existed in the result obtained by one time inverse consultation method. The kanji-hanzi conversion and Chinese synonyms table thus added 10,690 (12,089 - 1,399) candidates of very high quality. The three additional methods allowed us to increase the quality of our Chinese–Japanese lexicon from less that 45% to 85%. Table 2.6 shows an excerpt of our final lexicon. A comparison with a reference dictionary (the EDR dictionary constructed by the National Institute of Information and Communications Technology (NICT)), shows that our techniques are very efficient as 83% of the word pairs in the lexicon were not present in the EDR dictionary.

Table 2.6: An excerpt of the final Chinese–Japanese lexicon with indications on the origin of the word pairs and a final human assessment.

Chinese	Japanese	English meaning	classical joining	one time inverse	hanzi/kanji	synonyms	human assessment
中央	中央	center	○	1.000	○		√
矿山	水雷	diggings / torpedo	○	1.000			×
构造	構造	construction	○	0.667	○		√
战果	戦果	results of battle	○	1.000	○		√
春季	ばね	spring season / bounce	○	0.333			×
新闻	新聞	news / newspaper			○		×
不景气	不景気	bad times			○		√
乘法	乗法	multiplication			○		√
古典音乐	クラシック音楽	classical music	○	1.000			√
作法	行動方針	course of action	○	0.400			√
美国人	アメリカ人	American person	○	0.667			√
核电站	原子力発電所	nuclear power plant	○	1.000			√
空白	空欄	blank space	○	1.000			√
恭贺新禧	あけおめ	Happy New Year	○	1.000			√
去年	昨年	last year	○	1.000			√
南部	南部	southern part	○	1.000	○		√
探	訪ねる	to visit	○	0.400			√
讲话	話す	to speak	○	0.400			√
丛林	森林	forest	○	1.000		○	√
中心	中央	center	○	0.667		○	√
不时	時時	frequently				○	√
生词	新語	new word				○	√

Chapter 3

Monolingual and Bilingual Term Extraction for Re-tokenization in SMT

Because Chinese and Japanese do not have typographic boundaries (words are not separated by white spaces) like English in their writing systems, word segmentation or tokenization is used to break sentences down into individual words or tokens. Word segmentation is normally treated as a preprocessing step for machine translation. This allows us to obtain word-to-word or multi-word to multi-word translation relations in SMT. In Chinese–Japanese technical machine translation, it is uneasy to control the tokenization of multi-word technical terms. It may lead to erroneous isolated word-to-word translation relations in an inappropriate order or position. Thus, we propose to re-tokenize and group bilingual multi-word terms together to increase the translation probability of multi-word term to multi-word term translation. We propose to do this without the use of any additional lexicon, bilingual terminology bank or additional corpus.

This chapter[1] focuses on the study of re-tokenization of parallel data with multi-word terms for solving the problem of different segmentation constraints leading to different levels of granularity in Chinese and in Japanese. We present an automatic method to extract bilingual multi-word terms (or single-word to multi-word terms) from the training parallel corpus to re-tokenize parallel corpora. We perform experiments with an SMT system.

The structure of this chapter is as follows.

[1] Related to (Yang and Lepage, 2018), (Yang and Lepage, 2017), (Yang and Lepage, 2016), and (Yang et al., 2016)

- Section 3.1 introduces and identifies the problem of different segmentation between Chinese and Japanese in technical parallel corpora. It also reviews related works.
- Section 3.2 describes the extraction of Chinese–Japanese bilingual multi-word terms (multi-word to multi-word terms) using the C-value method and the sampling-based alignment method, and how to use these terms in two different SMT protocols.
- Section 3.3 presents how to further filter the extracted bilingual multi-word terms to obtain better accuracy in translation correspondence. Especially it shows how to extract single-word to multi-word terms to increase the number of bilingual terms for re-tokenizing the training data, so as to obtain even better results in translation accuracy.
- Finally, Section 3.4 summarizes this chapter.

3.1 Related Work

In (Wang et al., 2010b), a short unit transformation method for adapting Chinese word segmentation for MT based on transfer rules is described. The rules are obtained from alignment results and from a database constructed using additional lexica. In the research by Li et al. (2012), they improved a Chinese-to-Japanese patent translation system by using English as a pivot language for three different purposes: corpus enrichment, sentence pivot translation and phrase pivot translation. In our work, we propose a way to improve Chinese-to-Japanese phrase-based statistical machine translation (PB-SMT) quality based on the re-tokenization of a bilingual patent corpus with extracted bilingual aligned terms, without exploiting extra bilingual data, nor using a third language.

There exist previous work on extracting scientific or technical terms in different languages and different domains for applications like information retrieval, text categorization and also for machine translation. As an important milestone in terminology extraction, Frantzi et al. (2000) describe a combination of linguistic filtering and statistical measure, called C-value/NC-value, for the automatic extraction of multi-word terms from English scientific or technical texts. As an application for estimating the similarity of scientific papers, Milios et al. (2003) show how to extract English terms in the computer science and the medical domains using the C-value/NC-value extraction method. They make use of these terms to estimate the similarity of scientific papers in a vector space model. Lossio-Ventura et al. (2013) extracted and ranked English and French biomedical terms from free texts by using linguistic patterns, the C-value and keyword extraction measures. They showed that an appropriate harmonic mean of the C-value with keyword extraction measures offers better results in precision, either for the extraction of single-word or multi-words terms. Hadni et al. (2014) extracted

Arabic multi-word terms in the environment domain using a linguistic and a statistical approach. They incorporated contextual information as an association measure for unithood and termhood. From these previous works, we can see that the C-value is commonly used as a domain-independent method for single-word or multi-word term extraction. As for language independence, it was shown in (Mima and Ananiadou, 2001) that the C-value/NC-value method is an efficient domain-independent multi-word term extraction technique not only in English but in Japanese as well. In our work, we only focus on multi-word term extraction using the C-value method, because our ultimate gaol is that making use of these extract monolingual or bilingual multi-word terms in re-tokenization of Chinese and Japanese technical corpora.

Some pieces of work recognize monolingual or bilingual terms by considering compound words and their constituents. In (Nakagawa et al., 2004), they automatically extract Chinese terms from Web pages based on compound word productivity. They basically focus on how many words or characters adjoin the word or character under consideration to form compound words. They also take into account the frequency of terms. In (Fan et al., 2009), Chinese–Japanese multi-word terms are extracted by re-segmenting a Chinese and Japanese bi-corpus and combining multi-word terms as one token (glue them as one word) based on the extracted monolingual terms. The word alignments containing terms are smoothed by computing the associations between pairs of bilingual term candidates. They add the extracted bilingual terms to the phrase tables and compare translation accuracy with a baseline system. Different from their work, we focus on improving translation accuracy by re-tokenizing the training corpus with extracted bilingual multi-word terms (that we align using no-space word separators), i.e., improving technical translation quality by changing and balancing the granularity of the training data in Chinese and Japanese based on bilingual multi-word terms.

In (Mima et al., 1998), English–Japanese multi-word terms are recognized by the C-value method and by an example-based approach. In the example-based framework, translation example pairs describe the correspondence between source language expressions and target language expressions. They compute the semantic distance of the translation of terms extracted from a corpus in one language by C-value and terms extracted from another language using the same method. For translation of terms, they adopt the Transfer-Driven Machine Translation (TDMT) (Furuse and Iida, 1996) mechanism. In TDMT, source and target language expressions are expressed by patterns at various linguistic levels, which efficiently represent meaningful units for linguistic analysis and transfer.

In this chapter, we similarly consider monolingual multi-word terms, extracted from a Chinese and Japanese corpus using the C-value method as one token for processing. Different from the example-based approach used in (Mima et al., 1998), we then use the sampling-based alignment method (Lardilleux and Lepage, 2009) to align multi-word terms. We filter the aligned bilingual candidate terms by setting thresholds on translation probabilities and further filtering by taking the component of the terms and the ratio of the lengths in words between bilingual candidate terms into consideration.

Additionally, we use free available simplified–traditional character conversion data and a freely available hanzi–kanji conversion table between Chinese and Japanese characters to confirm or extract more promising bilingual multi-word terms. Because in the Chinese and Japanese writing systems, there exists a large amount of characters which share the same meaning, they can be considered as a linguistic clue to align words or multi-word expressions. Many studies have exploited common Chinese and Japanese characters. In (Goh et al., 2005), they build a Japanese–Simplified Chinese dictionary consisting of kanjis which are identical to traditional Chinese and associate the simplified Chinese character to it. In (Tan and Nagao, 1995), they use the occurrence of identical common Chinese characters in Chinese–Japanese in the sentence alignment task. In (Chu et al., 2012) and (Chu et al., 2013a), they construct a mapping table of Japanese, traditional Chinese and simplified Chinese using several freely available resources. In their work, they make use of the mapping table for adjusting Chinese segmentation results according to Japanese segmentation based on characters shared between Chinese and Japanese. In our work, we focus on terms and patent translation. We change and adjust the segmentation for terms in Chinese and Japanese at the same time (not only for Chinese) for improving SMT. We do not only consider the segmentation for the terms which made up of hanzi/kanji, but also take the terms which made up of katakana in Japanese into consideration.

In this chapter, we adopt the C-value method with its linguistic and statistical components to extract monolingual multi-word terms in Chinese and Japanese independently from a training corpus used to build a machine translation system. We first re-tokenize the Chinese–Japanese training corpus with these extracted monolingual multi-word terms to enforce these terms to be considered as one token (aligned with non-space separators). We then apply an alignment technique, the sampling-based alignment method (Lardilleux and Lepage, 2009), on this re-tokenized Chinese–Japanese training corpus to extract aligned candidate terms. The best aligned candidate terms (filtered bilingual multi-word terms) are finally kept by setting thresholds on translation probabilities (Chapter 3.2).

We perform SMT experiments using the Chinese–Japanese experimental data re-tokenized again using the filtered bilingual multi-word terms.

In a first experimental protocol, we re-tokenize the training, tuning and test data with the filtered bilingual multi-word terms before building an SMT system. The corpus used for learning the language model (LM) is the target language part of the re-tokenized training corpus.

In a second experimental protocol, we only re-tokenize the training corpus with the filtered bilingual multi-word terms. In the phrase tables, we segment these multi-word terms into words back before performing tuning and decoding. Another difference with the first experimental protocol is that the corpus for learning the language model is the original unchanged target language part of the training corpus.

We compare the two above different translation systems with a baseline system. We obtain a significant improvement in translation accuracy as evaluated by BLEU (Papineni et al., 2002) with the second experimental protocol.

We further filter the extracted bilingual multi-word terms by considering the components of the bilingual multi-word terms in characters as well as the ratio of their lengths in words (Chapter 3.3.1). We make use of kanji (Japanese)–hanzi (Chinese) character conversion to confirm and extract more promising bilingual terms (Chapter 3.3.1). We then consider the case in bilingual term extraction that some terms are single-word terms in one language but multi-word terms in another language (Chapter 3.3.4). Our pre-processing on terms has the effect of solving the problem of different segmentation conventions across languages.

Figure 3.1 gives examples of Chinese–Japanese patent sentences which are tokenized at different levels of granularity based on different segmentation tools. For instance, the multi-word term 钽阳/极体 'tantalum anode body' in Chinese is in translation relation with the multi-word タンタル/陽極/ボディ in Japanese, but actually, there is not any correspondence in word-to-word alignments (Case 1). Similar examples are 异氰/酸酯基 'isocyanate group' in Chinese and イソシアネート/基 in Japanese, or 放射线/量 'radiation dose' in Chinese and 放射/線量 in Japanese. Another case is that some terms are single-word terms in one language but multi-word terms in another language. For instance, the single-word term 肺气肿 'emphysema' in Chinese and the multi-word term 肺/気腫 in Japanese, 控制器 'controller' in Chinese and コント/ローラ in Japanese, or 缺氧 'oxygen deficiency' in Chinese and 酸素/欠乏 in Japanese. Actually, this also can be divided into two cases (Case 2 and Case 3): the first case (Case 2) is when the single-word to multi-word terms are made up of hanzi/kanji and they share all characters after kanji-hanzi conversion; the second case (Case 3) is when the single-word to multi-word terms have no shared (or partial shared) characters after kanji-hanzi conversion, even if they are made up of hanzi/kanji. For keeping the direct and exact translations between Chinese and Japanese terms, we intend to re-tokenize Chinese–Japanese parallel sentences around the bilingual multi-word terms. We intend to solve the segmentation problem of these three cases in the following sections.

- Case 1: 钽阳␣极体 ↔ タンタル␣陽極␣ボディ
  异氰␣酸酯基 ↔ イソシアネート␣基
  放射线␣量 ↔ 放射␣線量
- Case 2: 半导体层 ↔ 半導体␣層

肺气肿 ↔ 肺␣ 気腫

- Case 3: 控制器 ↔ コント␣ ローラ
  脲键 ↔ ウレア␣ 結合
  氧化物 ↔ 酸化␣ 物
  缺氧 ↔ 酸素␣ 欠乏

Chinese	该/钽阳/极体/通常/是/烧结/的/。
Japanese	タンタル/陽極/ボディ/は/、/ 通常/、/ 焼結/さ/れて/いる/。
Meaning	‘Normally the tantalum anode body is sintered.’
Chinese	贴片/52/-/58/也/通过/导线/连接/到/系统/控制器/30/。
Japanese	パッチ/52/～/58/は/、/また/、/電線/に/よって/システム/コント/ローラ/30/に/接続/さ/れる/。
Meaning	‘Patches 52-58 are also connected to the system controller 30 by wires.’
Chinese	在/第一/热/处理/之后/，/氧化物/半导体层/变成/缺氧/的/氧化物/半导体/，/即/，/电阻率/变得/更低/。
Japanese	酸化/物/半導体/層/は/、/第/1/の/加熱/処理/後/に/酸素/欠乏/型/と/なり/、/低/抵抗/化/する/。
Meaning	‘The oxide semiconductor layer becomes an oxygen-deficient type after the first heat treatment, namely, the resistivity becomes lower.’
Chinese	这/是/因为/水/与/异氰/酸酯基/反应/，/以/形成/脲键/。
Japanese	これ/は/、/水/と/イソシアネート/基/が/反応/する/こと/で/、/ウレア/結合/が/生じる/ため/である/。
Meaning	‘This is because of the reaction between water and isocyanate groups for forming urea bonds.’
Chinese	在/检测/出/的/放射线/量/小于/阈值/的/情况/下/，/为/否定/判断/，/从而/进入/到/步骤/110/。
Japanese	検知/した/放射/線量/が/閾値/未満/である/場合/は/、/否定/さ/れて/ステップ/110/へ/進む/。
Meaning	‘In the case where the radiation dose detected is less than the threshold, it is considered as the negative judgment, then go to step 110.’
Chinese	因而/，/在/本/实施/方式/中/，/能够/高效率/地/进行/关于/肺气肿/的/图像/诊断/的/支援/。
Japanese	従って/、/本/実施/形態/で/は/、/肺/気腫/に/関する/画像/診断/の/支援/を/効率/良く/行なう/こと/が/できる/。
Meaning	‘Thus, in this embodiment, the support on the image diagnosis of emphysema can be performed efficiently.’

Figure 3.1: Examples of segmentation in Chinese–Japanese patents. Technical terms in different languages are tokenized at different levels of granularity. The segmentation tools used here are the Stanford parser for Chinese and Juman for Japanese. The words given in the box are multi-word terms or single-word terms in Chinese or Japanese.

3.2 Bilingual Multi-Word Term Extraction for Different SMT Protocols

This section describes our proposed method that uses the C-value method (Frantzi et al., 2000) combined with the sampling-based alignment method (Lardilleux and Lepage, 2009) for the extraction of Chinese–Japanese bilingual multi-word terms and how to use these terms in different SMT protocols.

3.2.1 C-value based Monolingual Multi-Word Term Extraction

The C-value is a commonly used automatic domain-independent method for multi-word term extraction. This method has two main parts: a linguistic part and a statistical part. The linguistic part constrains the type of terms extracted relying on part-of-speech tagging, linguistic filters, stopword list, etc. The statistical part provides a termhood measure called C-value. The larger this value, the higher the probability for an extracted candidate term to actually be a term. The advantage of the C-value method is that it can compute multi-word terms made up of complex structures even when these structures have a low frequency. In our experiments, we monolingually extract multi-word terms which contain a sequence of nouns or adjectives followed by a noun in both Chinese and Japanese.

This linguistic pattern can be written as follows using a regular expression[2]:

(Adjective | Noun)$^+$ Noun

The statistical component, the measure of termhood, called the C-value, is given by the following formula:

$$\text{C-value}(a) = \begin{cases} \log_2 |a| \times \ f(a) & \text{if } a \text{ is not nested,} \\ \log_2 |a| \times \big(f(a) - \frac{1}{|T_a|} \sum_{b \in T_a} f(b)\big) & \text{otherwise} \end{cases} \tag{3.1}$$

where a is the candidate string. |a| is the length of a. $f(a)$ is its frequency of occurrence in the corpus. T_a is the set of extracted candidate terms that contain a. $|T_a|$ is the number of these candidate terms.

In our experiments, we follow the basic steps of the C-value approach to extract monolingual multi-word terms from the monolingual part of the existing Chinese–Japanese training corpus. Firstly, we tag each word in the Chinese and the Japanese corpus respectively; then, we compute and extract multi-word terms based on the linguistic pattern and the formula given above for each language. The stopword list is used to

[2] Pattern for Chinese: $(JJ|NN)^+ NN$, pattern for Japanese: (形容詞 | 名詞)$^+$ 名詞. ‘JJ’ and ‘形容詞’ are codes for adjectives, ‘NN’ and ‘名詞’ are codes for nouns in the Chinese and the Japanese taggers that we use.

avoid extracting infelicitous sequences of words. Our stopword list consists of 240 function words (including numbers, letters, punctuations etc.)

Then, we re-tokenize the training corpus with these extracted monolingual multi-word terms by enforcing these terms to be considered as one token. Technically, we just replace each space inside a multi-word term by a non-space word separator, so that each multi-word term is considered as one token. In the figures and tables of this document, the non-space word separator is notes by ␣ .

The segmenter and part-of-speech tagger that we use are the Stanford parser[3] for Chinese and Juman[4] for Japanese. Figure 3.2 shows examples of Chinese and Japanese monolingual multi-word term candidates extracted based on this approach. In some cases (Case A), we may obtain the same number of multi-word terms in Chinese and Japanese respectively; in some other cases (Case B), we obtain different numbers of multi-word terms in the two languages; we also may extract a certain number of multi-word terms in one language, but fail to extract any term in the other language (Case C). Figure 3.3 shows examples of re-tokenized Chinese–Japanese sentences with monolingual multi-word terms in Chinese and Japanese respectively, as extracted by this approach.

3.2.2 Sampling-based Bilingual Multi-Word Term Extraction

Bilingual multi-word terms are multi-word term to multi-word term alignments, i.e., we only want to extract corresponding terms which are multi-word terms at the same time in both languages. We extract them by performing word-to-word, or, better said, token-to-token alignment on the Chinese–Japanese training corpus re-tokenized as described in the previous section (see Figure 3.3). For that, we use the open source implementation of the sampling-based alignment method, Anymalign (Lardilleux and Lepage, 2009)[5].

We filter these aligned multi-word candidate terms by setting some threshold P on each of the direct and inverse translation probabilities ($0 < P \leq 1$). The translation probabilities $P(t|s)$ and $P(t|s)$ are computed as prior probabilities from the co-occurrence frequencies consistent with the word alignment in the translation table:

$$P(ja|zh) = \frac{P(zh, ja)}{P(zh)} = \frac{C(ja \leftrightarrow zh)}{C(zh)} \tag{3.2}$$

$$P(zh|ja) = \frac{P(ja, zh)}{P(ja)} = \frac{C(zh \leftrightarrow ja)}{C(ja)} \tag{3.3}$$

[3] http://nlp.stanford.edu/software/segmenter.shtml

[4] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

[5] Anymalign (https://anymalign.limsi.fr) is a phrase-to-phrase alignment tool, but the use of option -N 1 limits its functionality to word-to-word alignment. Technically, we identify the multi-word term to multi-word term alignments by spotting the non-space word separators inserted inside multi-word terms in place of spaces.

Case A	Chinese	图♯NN 21♯CD 是♯VC 表示♯VV **硬质♯NN** **碳♯NN** **皮膜♯NN** 的♯DEG **接触♯NN** **电阻♯NN** 的♯DEG 图表♯NN 。♯PU	硬质 碳 皮膜 ‘diamond-like carbon’ 接触 电阻 ‘contact resistance’
	Japanese	図/名詞 21/特殊 は/助詞 、/特殊 **硬質/名詞** **炭素/名詞** **皮/名詞** **膜/名詞** の/助詞 **接触/名詞** **抵抗/名詞** を/助詞 示す/動詞 グラフ/名詞 である/判定詞 。/特殊	硬質 炭素 皮膜 ‘diamond-like carbon’ 接触 抵抗 ‘contact resistance’
	Meaning	‘Figure 21 shows the graph for the contact resistance of the diamond-like carbon.’	
Case B	Chinese	在♯P **栅极♯JJ** **电阻♯NN** 7♯CD 的♯DEG 两♯CD 端♯NN ,♯PU 层间♯AD 绝缘♯VV 膜♯NN 12♯CD 被♯SB 刻蚀♯VV ,♯PU 埋入♯VV 钨等♯JJ 的♯DEG **接触♯NN** **电极♯NN** 6♯CD 。♯PU	栅极 电阻 ‘gate resistance’ 接触 电极 ‘contact electrode’
	Japanese	**ゲート/名詞** **抵抗/名詞** 7/特殊 の/助詞 両端/名詞 で/助詞 層/名詞 間/接尾辞 **絶縁/名詞** **膜/名詞** 12/特殊 が/助詞 エッチング/名詞 さ/動詞 れ/接尾辞 、/特殊 タングステン/名詞 等/接尾辞 の/助詞 **コンタクト/名詞** **電極/名詞** 6/特殊 が/助詞 埋め/動詞 込ま/動詞 れて/接尾辞 いる/接尾辞 。/特殊	ゲート 抵抗 ‘gate resistance’ 絶縁 膜 ‘dielectric film’ コンタクト 電極 ‘contact electrode’
	Meaning	‘A layer dielectric film 12 is etched at both ends of the gate resistor 7, the contact electrode 6 of tungsten is embedded.’	
Case C	Chinese	当♯P 在♯P 脑瘤♯NN 组织♯VV 的♯DEC 测定♯VV 中将♯NN 预定♯VV 的♯DEC 时间段♯NN 设定♯VV 为♯VC 大约♯AD 5♯CD 分钟♯M 的♯DEC 时候♯NN ,♯PU 得到♯VV 充分♯JJ 的♯DEG 结果♯NN 。♯PU	(no mono-terms extracted in Chinese)
	Japanese	この/指示詞 所要/名詞 時間/名詞 は/助詞 、/特殊 **脳/名詞** **腫瘍/名詞** **組織/名詞** の/助詞 測定/名詞 に/助詞 おいて/動詞 は/助詞 5/特殊 分/接尾辞 程度/接尾辞 で/助詞 **十分な/形容詞** **結果/名詞** が/助詞 得/動詞 られた/接尾辞 。/特殊	脳 腫瘍 組織 ‘brain tumor tissue’ 十分な 結果 ‘satisfactory results’
	Meaning	‘In the measurement of brain tumor tissue, the satisfactory results are obtained while the scheduled time is set as about 5 minutes.’	

Figure 3.2: Examples of different cases of Chinese and Japanese monolingual multi-word terms extracted using the C-value and the linguistic pattern: (Adjective | Noun)$^+$ Noun. See Footnote 2 for POS codes.

In the equations, $C(x)$ denotes the number of occurences of the word or phrase x in the monolingual re-tokenized part of the training corpus, and $C(x \leftrightarrow y)$ is the number of co-occurrences of x and y in the re-tokenized parallel training corpus.

Table 3.1 shows examples from the results of bilingual multi-word term extraction.

3.2.3 Using Bilingual Multi-Word Terms in SMT

We propose two protocols to use these extracted bilingual multi-word terms in SMT experiments. We compare these two protocols with a standard baseline system.

图 21 是 表 示 **硬质␣ 碳␣ 皮膜** 的 **接触␣ 电阻** 的 图表 。	図 21 は 、 **硬質␣ 炭素␣ 皮␣ 膜** の **接触␣ 抵抗** を 示す グラフ である 。
在 **栅极␣ 电阻** 7 的 两 端 , 层 间 绝 缘 膜 12 被 刻 蚀 , 埋 入 钨 等 的 **接触␣ 电极** 6 。	**ゲート␣ 抵抗** 7 の 両 端 で 層 間 **絶縁␣ 膜** 12 が エ ッ チ ン グ さ れ 、 タ ン グ ス テ ン 等 の **コンタクト␣ 電極** 6 が 埋 め 込 まれて いる 。
当 在 脑瘤 组织 的 测定 中将 预定 的 时间段 设定 为 大约 5 分钟 的 时候 , 得到 充分 的 结果 。	こ の 所 要 時 間 は 、 **脳␣ 腫瘍␣ 組織** の 測 定 に お い て は 5 分 程 度 で **十分な␣ 結果** が 得 られた 。

FIGURE 3.3: Examples of Chinese *(left)* and Japanese *(right)* sentences re-tokenized using the extracted monolingual multi-word terms. The boxes in plain line show the monolingual multi-word terms which correspond across languages. The boxes in dashed line show the monolingual multi-word terms which are re-tokenized in their language only.

TABLE 3.1: Examples of bilingual multi-word terms extracted and then filtered by our method: firstly, pairs with only one word on any side are rejected (non-space word separator: ␣), then pairs of multi-word terms where one of the translation probabilities is below the threshold (0.6 here) are rejected. The but last column indicates the bilingual multi-word term pairs which are kept. The last column shows which extracted multi-word term pairs were considered correct or not by manual inspection.

Chinese	Japanese	Meaning	$P(t\|s)$	$P(s\|t)$	Kept	Good match
硬质␣ 碳␣ 皮膜	硬質␣ 炭素␣ 皮␣ 膜	'diamond-like carbon'	1.000	1.000	yes	yes
接触␣ 电阻	接触␣ 抵抗	'contact resistance'	0.920	0.973	yes	yes
栅极␣ 电阻	ゲート␣ 抵抗	'grid resistance'	1.000	1.000	yes	yes
接触␣ 电极	コンタクト␣ 電極	'contact electrode'	0.946	0.972	yes	yes
核酸	核␣ 酸	'nucleic acid'	0.974	0.956	no	yes
极板	極␣ 板	'electrode plate'	0.992	1.000	no	yes
废␣ 热	廃␣ 熱	'waste heat'	0.844	0.241	no	yes
变速␣ 机	変速␣ 機	'variable-speed motor'	1.000	0.006	no	yes
芯片␣ 级␣ 控制␣ 手机␣ 模块	チップ␣ レベル	–	1.000	1.000	yes	no
铃音␣ 服务器	マルチメディア␣ リングー␣ バック␣ トーン␣ サーバ	–	1.000	1.000	yes	no
激振␣ 电极	主に␣ 形成	–	0.862	0.982	yes	no

The first protocol (System re-tok-all) is as follows. We train the translation model on the training corpus re-tokenized with the best bilingual multi-word terms. The language model is learnt on the target part of the re-tokenized training corpus. The tuning and test sets used are also re-tokenized with the same bilingual multi-word terms as used for re-tokenizing the training data.

The second protocol (System re-tok-train-only) is as follows. We also train the translation model only on the training corpus re-tokenized with the best bilingual multi-word terms. But the language model is learnt based on the original, un-re-tokenized, target language part of the training corpus. For consistency, we remove the non-space word separators from the phrase tables before performing tuning and decoding.

3.2.4 Experiments and Results

3.2.4.1 Chinese and Japanese Data Used

In this section, we describe the results for the monolingual and the bilingual multi-word term extraction from the training data. We give the technical settings and describe the tools used for the baseline SMT system and in the two other protocols. We also give the results of the evaluation of the systems and compare them.

The Chinese–Japanese parallel sentences used in our experiments are randomly extracted from the Chinese–Japanese JPO Patent Corpus (JPC)[6]. This corpus consists of about 1 million parallel sentences with four sections (Chemistry, Electricity, Mechanical engineering, and Physics). It is already divided into training, tuning and test sets: 1 million sentences, 4,000 sentences and 2,000 sentences respectively. For our experiments, we randomly extract 100,000 parallel sentences from the training part, 500 parallel sentences from the tuning part, and 1,000 from the test part. The sentences have a length of approximately 23 words in Chinese and 30 words in Japanese. Table 3.2 shows basic statistics on our data sets.

Table 3.2: Statistics on our experimental data sets (after tokenizing and lowercasing). Here 'mean ± std.dev.' gives the average length of the sentences in words.

	Baseline	Chinese	Japanese
train	sentences	100,000	100,000
	words	2,314,922	2,975,479
	length in words (avg. ± std.dev.)	23.29 ± 11.69	29.93 ± 13.94
tune	sentences	500	500
	words	14,251	17,904
	length in words (avg. ± std.dev.)	28.61 ± 21.88	35.94 ± 25.07
test	sentences	1,000	1,000
	words	27,267	34,292
	length in words (avg. ± std.dev.)	27.34 ± 15.59	34.38 ± 18.78

3.2.4.2 Monolingual and Bilingual Multi-Word Term Extraction

We apply the method described in Section 3.2.1 to independently extract monolingual multi-word terms from the 100,000 sentences of the training data of our Chinese–Japanese parallel corpus. We independently obtain 81,618 multi-word terms in Chinese

[6] `http://lotus.kuee.kyoto-u.ac.jp/WAT/patent/index.html`

and 93,105 in Japanese. We manually checked the precision of the extracted monolingual multi-word terms for Chinese and Japanese by sampling 1,000 monolingual terms. The precision was 95% in both languages. The extracted monolingual multi-word terms were ranked by decreasing order of C-values. For keeping the balance between monolingual term extraction in different languages, we re-tokenize the training corpus with the same number of Chinese and Japanese monolingual multi-word terms respectively. These terms are the first 80,000 monolingual multi-word terms with the highest C-values in each language.

We then extract bilingual multi-word terms from the Chinese–Japanese training corpus re-tokenized using these 80,000 monolingual multi-word terms, by following the method described in Section 3.2.2. We measured the number of bilingual multi-word terms extracted from the re-tokenized training corpus of 100,000 sentence pairs by the sampling-based alignment method which meet the constraint of having both translation probabilities above a given threshold. The second column in Table 3.3 shows this number when the threshold varies[7]. In addition, we manually checked the correspondence between these bilingual multi-word terms. The percentage of good matches was roughly estimated to be over 70 % when the threshold becomes greater than 0.4.

Table 3.3: Results of bilingual multi-word extraction and evaluation results for Chinese-to-Japanese translation with the two proposed protocols (Systems re-tok-all and re-tok-train-only) for different thresholds on the translation probabilities. The score of the baseline is given on line 4. The best BLEU score obtained (33.32) is for the System re-tok-train-only with a threshold of 0.6 (boldfaced score). BLEU scores marked with * are significantly better than the score of the Baseline system at p-value < 0.01, except for threshold ≥ 0.4 at p-value < 0.05.

Thresholds	♯ of bilingual multi-word terms (filtered by thresholds)	BLEU (System re-tok-all)	BLEU (System re-tok-train-only)
≥ 0.0	52,785 (35 %)	32.08±1.07	32.44±1.07
≥ 0.1	31,795 (52 %)	31.88±1.10	32.23±1.18
≥ 0.2	27,916 (58 %)	32.42±1.14	32.00±1.16
Baseline	-	32.35±1.15	32.35±1.15
≥ 0.3	25,404 (63 %)	31.85±1.08	33.08±1.12*
≥ 0.4	23,515 (72 %)	31.45±1.13	32.77±1.15*
≥ 0.5	21,846 (76 %)	32.11±1.12	33.02±1.14*
$\geq$ 0.6	20,248 (78 %)	32.68±1.13	**33.32±1.15***
≥ 0.7	18,759 (79 %)	32.61±1.12	32.85±1.19*
≥ 0.8	17,311 (79 %)	32.34±1.15	33.25±1.06*
≥ 0.9	15,464 (80 %)	32.16±1.11	33.20±1.15*

[7] We tried to extract bilingual multi-word terms using GIZA++. We obtained two times less multi-word to multi-word alignments (23,085 without any filtering) compared with the sampling-based alignment method (52,785, $P \geq 0.0$ without any filtering). The sampling-based alignment method is more efficient than GIZA++.

3.2.4.3 SMT Systems

- Baseline: no re-tokenization

We train a standard baseline system using the GIZA++/MOSES pipeline (Koehn et al., 2007). We train the Chinese-to-Japanese translation model with the training parallel corpus described in Table 3.2. The monolingual part in the target language (Japanese) is used to learn a language model using KenLM (Heafield, 2011) in word-based 5-grams. The development data with 500 parallel sentences is used for tuning by minimum error rate training (MERT) (Och, 2003). For decoding, we use the default options of Moses, the distortion limit is set to 20.

- System re-tok-all: re-tokenization of all the data

Different from the baseline SMT system, here we make use of bilingual multi-word terms. We re-tokenize the training, tuning and test data with the bilingual multi-word terms by enforcing them to be considered as one token. Each multi-word term is re-tokenized with non-space word separators.

The bilingual multi-word terms used for re-tokenizing the training data are all the bilingual terms extracted with both translation probabilities above a given threshold. The tuning data is re-tokenized with the bilingual multi-word terms used during re-tokenization of the training data. The test data in Chinese is re-tokenized with the monolingual part of the bilingual multi-word terms used during re-tokenization of the training and tuning data. The data for learning the language model is the target language (Japanese) part of the re-tokenized training data. Of course, we remove the non-space word separators after decoding before the evaluation process.

Table 3.3 (column 3) shows the evaluation results for our Chinese-to-Japanese SMT in BLEU scores (Papineni et al., 2002). We did not obtain significant difference in BLEU in comparison with the baseline system, except for BLEU scores which are significant lower than those of the baseline system when the thresholds are $P \geq 0.1$ and $P \geq 0.4$.

- System re-tok-train-only: retokenization of the training data for translation models only

Because re-tokenization of all of the data did not lead to improvement, we decide to only re-tokenize the Chinese–Japanese training parallel corpus. We train the Chinese–Japanese translation models, i.e., the phrase tables using this re-tokenized data. We remove the non-space word separators from the phrase tables, and then train further models, and perform tuning and decoding. The data for tuning, for learning the language model, as well as the test data, are the same as in the baseline system.

Table 3.3 (column 4) shows the evaluation of the results of Chinese-to-Japanese translation in BLEU scores based on the procedure for this system. Compared with the baseline system and the System re-tok-all, we obtained significantly better results in BLEU scores for thresholds equal to or greater than 0.3, while the scores for lower thresholds are similar to and not significantly different from the score of the baseline system. This shows that this protocol at least does not hurt and may be beneficial when applied with any value for the threshold.

3.2.5 Analysis of the Results and Discussion

We further compare the best system: System re-tok-train-only (threshold of 0.6), with the baseline system and the system: System re-tok-all (also $P \geq 0.6$).

An inspection of the BLEU scores obtained during tuning for the System re-tok-train-only and for the System re-tok-all reveals that the former ones are all higher than the scores of the latter ones at each step of the tuning (Table 3.4). This proves that the second protocol is immediately profitable for improvement of BLEU scores. Figure 3.4 gives examples of translation results for one sentence from the tuning set.

We conducted manual inspection of the Chinese N-grams × Japanese M-grams distribution in the reduced phrase table, i.e., the phrase table which contains only those potentially useful phrase pairs for the translation of the test set. Table 3.5 and Table 3.6 show that the total number of potentially useful phrase pairs with the re-tokenized training corpus is larger than that in the baseline system. This simply shows that multi-word terms extracted from the training set are potentially useful for the translation of the test set, which was precisely the goal of our work on extracting bilingual multi-word terms.

Table 3.4: Comparison of BLEU scores in tuning for Systems re-tokenize all data and System re-tokenize training only ($P \geq 0.6$).

	System re-tokenize all data	System re-tokenize training only
run1.moses.ini	–not-estimated–	–not-estimated–
run2.moses.ini	34.0215	35.1813
run3.moses.ini	33.9925	34.9997
run4.moses.ini	34.5491	35.8575
run5.moses.ini	34.6492	35.7729
run6.moses.ini	34.4521	35.8732
run7.moses.ini	34.7065	36.1072
run8.moses.ini	35.1249	36.2294
run9.moses.ini	35.2961	36.4267
run10.moses.ini	35.1621	36.4747
run11.moses.ini	35.3492	36.4911
run12.moses.ini	35.3836	-

3.2.6 Summary

In this section, we proposed an approach to improve translation accuracy in statistical machine translation of Chinese–Japanese patents by re-tokenizing the parallel training

System	BLEU (in tuning)	Translation results (Japanese)	Reference (Japanese)
System re-tok-all (non-space word separators ␣ kept even in evaluation)	49.11	また 、 この 構造 に 代えて 、 図 7 (b) に 示す ように 、 予め 設定 さ れた 優先 順位 を 使用 する こと に より 、 優先 順位 の 高い レーダ␣ アンテナ に 取得 した 速度␣ データ レーダ␣ アンテナ 。	また 、 この 構成 に 代えて 、 図 7 (b) に 示す ように 、 予め レーダ␣ アンテナ の 優先 順位 を 設定 して おき 、 優先 順位 の 高い レーダ␣ アンテナ が 取得 した 速度␣ データ を 優先 的に 用いる こと も できる 。
System re-tok-all	51.38	また 、 この 構造 に 代えて 、 図 7 (b) に 示す ように 、 予め 設定 さ れた 優先 順位 を 使用 する こと に より 、 優先 順位 の 高い レーダ アンテナ に 取得 した 速度 データ レーダ アンテナ 。	また 、 この 構成 に 代えて 、 図 7 (b) に 示す ように 、 予め レーダ アンテナ の 優先 順位 を 設定 して おき 、 優先 順位 の 高い レーダ アンテナ が 取得 した 速度 データ を 優先 的に 用いる こと も できる 。(Cell A)
Baseline	54.73	また 、 この 構造 に 代えて 、 図 7 (b) に 示す ように 、 予め 設定 さ れて レーダ アンテナ の 優先 順位 を 優先 的に は 、 優先 順位 の 高い レーダ アンテナ 取得 さ れた 速度 データ である 。	same as Cell A
System re-tok-train-only	55.70	また 、 この 構造 に 代えて 、 図 7 (b) に 示す ように 、 予め 設定 レーダ アンテナ の 優先 順位 を 使用 して 、 優先 順位 の 高い レーダ アンテナ に 優先 的 に 取得 さ れた 速度 データ である 。	same as Cell A

Figure 3.4: Examples of translation results in the last iteration of the tuning. A higher BLEU score (55.70) is obtained by the System re-tok-train-only in comparison with the System re-tok-all (49.11). An obvious reason for this is that we aligned multi-word terms with non-space word separators (␣) which are thus considered as one token, hence the computation of BLEU mechanically leads to lower scores. When removing the non-space word separators, we mechanically obtain a higher BLEU score (51.38), but still lower than the BLEU score obtained by the System re-tok-train-only and the baseline system. Here, $P \geq 0.6$.

corpus with bilingual multi-word terms. We did not use any other additional corpus or terminological lexicon. We extracted monolingual multi-word terms from the monolingual parts of the Chinese–Japanese training corpus independently by using a simple linguistic pattern and the C-value score. We considered each monolingual multi-word term as one token. We re-tokenized the training corpus with these monolingual multi-word terms and performed a token-to-token alignment. Bilingual multi-word terms were extracted among the corresponding monolingual multi-word terms obtained by alignment. We also set a threshold on translation probabilities in both directions. An investigation of the results of our experiments indicate that the bilingual multi-word terms extracted have over 70 % precision (good match) for threshold values over 0.4. We obtained the highest percentage with 80% with a threshold of 0.9.

We proposed two experimental protocols for using the extracted bilingual multi-word terms in SMT experiments. The first protocol re-tokenized all of the data with the

Table 3.5: Distribution of the reduced phrase table (System re-tok-train-only) of a C-value/sampling-based alignment term extraction method based on GIZA++/Moses 2.1.1. The bold face numbers showing the increased N (Chinese) × M (Japanese)-grams (less than 4-grams) in the reduced phrase table, and the total number of N (Chinese) × M (Japanese)-grams, which increased in comparison with the baseline system.

		Target = Japanese									
		1-gram	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	8-gram	9-gram	total
Source = Chinese	1-gram	29986	**86874**	**79132**	**49514**	27936	14843	7767	149	15	296218
	2-gram	**14201**	39342	**42833**	**27865**	15746	8292	4293	103	14	152690
	3-gram	**1492**	**3997**	7985	**7244**	4627	2528	1290	65	3	29231
	4-gram	**186**	**434**	1106	2099	1896	1310	691	23	0	7745
	5-gram	27	49	163	388	659	556	392	12	0	2246
	6-gram	2	6	14	60	114	180	170	10	1	557
	7-gram	0	0	4	4	22	48	72	6	1	157
	8-gram	0	0	0	0	0	0	0	1	0	1
	9-gram	0	0	0	0	0	0	0	0	1	1
	total	45894	130702	131237	87174	51000	27757	14675	369	35	**488846**

Table 3.6: Distribution of the reduced phrase table of the baseline system based on GIZA++/Moses 2.1.1.

		Target = Japanese									
		1-gram	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	8-gram	9-gram	total
Source = Chinese	1-gram	32320	84308	71713	42518	22831	11726	6035	0	0	271451
	2-gram	13570	39534	41775	25628	13703	6922	3518	0	0	144650
	3-gram	1384	3906	8067	7117	4276	2238	1093	0	0	28081
	4-gram	163	413	1124	2124	1853	1248	614	0	0	7539
	5-gram	27	50	154	386	658	562	360	0	0	2197
	6-gram	6	9	13	59	116	181	164	0	0	548
	7-gram	1	1	3	5	20	50	73	0	0	153
	8-gram	0	0	0	0	0	0	0	0	0	0
	9-gram	0	0	0	0	0	0	0	0	0	0
	total	47471	128221	122849	77837	43457	22927	11857	0	0	454619

bilingual multi-word terms. The second protocol only re-tokenized the training data to produce the phrase tables of the SMT system. The first protocol did not lead to improvements in translation accuracy compared with the baseline system. The second protocol led to statistically significant improvements for thresholds equal to or greater than 0.3.

In this section, for bilingual term extraction, we considered only the case (Case 1 in page 35) where multi-word terms can be found in both languages at the same time so that they can be aligned, e.g., (Chinese) 栅极␣ 电阻 / (Japanese) ゲート␣ 抵抗 'grid resistance' or (Chinese) 硬质␣ 碳␣ 皮膜 / (Japanese) 硬質␣ 炭素␣ 皮␣ 膜 'diamond-like carbon'. One can interpret our work as a way to change the granularity of multi-word terms and enforce a different segmentation in the case of such aligned terms.

Manual inspection of the data allowed us to spot the fact that, for Chinese and Japanese, many cases of bilingual terms are single-word terms in one language but multi-word terms in another language. These cases are mainly due to different segmentation results in Chinese and Japanese. There are three cases of these kind of bilingual terms:

- the first case (Case 2 in page 35) is when these bilingual terms share common hanzi and kanji after kanji-hanzi conversion. E.g., (Chinese) 极板 / (Japanese) 極␣板 'electrode plate' or (Chinese)癌细胞 / (Japanese) 癌␣細胞 'cancer cell';
- the second case (Case 3 in page 35) is when only one side is recognized as multi-word term, but the Japanese part is made up of katakana or a combination of kanji and katakana, e.g., (Chinese) 碳纳米管 / (Japanese) カーボン␣ナノチューブ 'carbon nano tube', or (Chinese) 逆变␣器/ (Japanese) インバータ 'inverter' or still (Chinese) 乙酸乙酯 / (Japanese) 酢酸␣エチル 'ethyl acetate';
- the third case (Case 3 in page 35) is when a single-word corresponds to a multi-word term made up of hanzi/kanji not shared (without any corresponding character) or partial shared, e.g., (Chinese) 缺氧 / (Japanese) 酸素␣欠乏 'oxygen deficit' or (Chinese) 供油路 / (Japanese) 給␣油路 'feed oil passage'.

In the following section, we intend to leverage previous results on bilingual term extraction to address these cases and make use of kanji-hanzi conversion to improve the extraction of bilingual terms. We expect further improvements in translation results.

3.3 Expending Bilingual Terms by Single-Word to Multi-Word Term Extraction

This section describes the work on expending extraction of bilingual terms by considering alignments of single-word to multi-word terms. We use kanji-hanzi conversion combined with the further filtering results based on the results obtained by the sampling-based alignment method. We also try to ignore the constraints on the components of terms (make up of hanzi/kanji). We just focus on single-word to multi-word term extraction. We present evaluation results for SMT experiments based only on the second experimental protocol (System re-tok-train-only) described in Section 3.2.3.

3.3.1 Kanji-Hanzi Conversion based Method

Table 3.7 shows some bilingual multi-word terms that we extracted by setting a threshold P of 0.6 based on the method proposed in Section 3.2.2. It is possible that some incorrect alignments are extracted. Such examples appear on the last three lines in this table.

TABLE 3.7: Extraction of bilingual aligned multi-word terms in both languages at the same time by setting a threshold of 0.6. Yes and no in the but last column show the multi-word term (either on both sides or on one side only) alignments that are kept or excluded. Yes and no in the last column show whether the extracted (or not extracted) term pairs are correct or incorrect alignments by manual check.

Chinese	Japanese	Meaning	$P(t\|s)$	$P(s\|t)$	Kept	Good match
葡萄糖␣浓度	グルコース␣濃度	‘glucose concentration’	0.962	0.891	yes	yes
血糖␣正常␣水平	正常␣血糖␣レベル	‘normal blood glucose level’	1.000	1.000	yes	yes
心脏␣周期	心臓␣周期	‘cardiac cycle’	1.000	1.000	yes	yes
心收缩␣期	心␣収縮␣期	‘systole’	1.000	0.833	yes	yes
加热␣烹饪	加熱␣調理	‘cooking’	1.000	0.815	yes	yes
油脂␣组成␣物	油脂␣組成␣物	‘fat composition’	1.000	1.000	yes	yes
脂肪␣酸酯	脂肪␣酸␣エステル	‘fatty acid ester’	1.000	0.983	yes	yes
植物␣油脂	植物␣油脂	‘vegetable oil and fat’	1.000	1.000	yes	yes
糖尿病	糖尿␣病	‘diabetes’	1.000	0.667	no	yes
肺␣癌	肺癌	‘lung cancer’	1.000	1.000	no	yes
杀生␣物剂	殺生␣物␣剤	‘biocide’	0.600	0.107	no	yes
官能␣基	官能␣基	‘functional group’	0.250	0.009	no	yes
糖尿病␣小鼠␣中肾␣小管␣上皮␣细胞	上皮␣細胞	-	1.000	1.000	yes	no
上述␣液体状	前記␣アルカリ␣活性␣結合␣材	-	1.000	1.000	yes	no
上述靶␣蛋白	種々の␣上記	-	1.000	1.000	yes	no

To improve the results, we further filter these extracted bilingual multi-word terms by comparing the lengths in words of the Chinese (Japanese) part to its corresponding Japanese (Chinese) part. We investigate the relation between the ratio of the lengths in words between Chinese and Japanese multi-word terms and the precision of the extracted bilingual multi-word terms. We set the ratio of the lengths to 1.0, 1.5, 2.0 and 2.5. The precision of the kept bilingual multi-word terms in each ratio is checked by sampling 100 bilingual multi-word terms. On the bilingual multi-word term extraction results obtained by setting P=0.6, the precisions for each ratio are 94%, 92%, 90% and 80%. Because the precision of the extracted bilingual multi-word terms decreases rapidly when the ratio tends to 2.5, we set the ratio of the lengths in both directions to a maximum value of 2.0 to keep precision and recall high at the same time. This means that we exclude aligned multi-word terms with a Chinese (resp. Japanese) part more than twice as long as the Japanese (resp. Chinese) part. Another filtering constraint is to filter out alignments which contain hiragana on the Japanese side. This constraint results from an investigation of the distribution of the components in Japanese by which we found that multi-word terms made up of “kanji + hiragana” or “kanji + hiragana + katakana” have lower chance to be aligned with Chinese multi-word terms (see Table 3.8).

Table 3.7 leads to the observation that some correctly aligned bilingual terms cannot be extracted by using the methods we described in Section 3.2.2. Such examples of terms are given in Table 3.9. All such examples are cases where the terms in Japanese (or in

Table 3.8: Distribution of the components for multi-word terms in Japanese (52,785 bilingual multi-word terms obtained by setting threshold P with 0).

Components for multi-word terms in Japanese	Sample	♯ of these terms
all kanji	心␣収縮␣期	28,978 (55%)
kanji/katakana + katakana	正常␣血糖␣レベル ホスト␣システム	19,913 (37.7%)
kanji + hiragana	様々な␣分野	3,377 (6.3%)
kanji + hiragana + katakana	好適な␣重力␣ミキサー	517 (1%)

Chinese) are not multi-word terms, or cases discarded by the threshold P on translation probabilities. Such aligned terms can be retrieved by taking the similarity between hanzi and kanji into consideration. For instance, in Table 3.9, the pair "(Chinese) 添加剂/ (Japanese) 添加␣剤" 'additive' is supported by the kanji-hanzi conversion of the last element "剂–剤" 'agent'. On the contrary, "(Chinese) 官能␣基/ (Japanese) 官能␣基" 'functional group' can be extracted without kanji-hanzi conversion.

Table 3.9: Examples of discarded bilingual aligned multi-word terms (either on both sides or on one side only) by setting threshold $P \geq 0.6$.

Cases	Chinese	Japanese
One side is multi-word terms	糖尿病 肺癌 添加剂 水␣蒸气	糖尿␣病 肺␣癌 添加␣剤 水蒸気
Probability (P) is lower than threshold	杀生␣物剂 官能␣基	殺生␣物␣剤 官能␣基

Consequently, we keep the alignments where either one side is a multi-word term after token-to-token alignment. We convert Japanese words made up of Japanese kanji only into simplified Chinese characters through kanji-hanzi conversion. By doing so, we generate a Zh–Ja–Converted-Ja file automatically where each line consists in the Chinese term, the original Japanese term and the converted Japanese term (simplified Chinese term). In this way, by comparing Converted-Ja with the Chinese term (Zh), if a converted Japanese term is equal to its corresponding Chinese term for each character, we can extract more reliable Chinese–Japanese bilingual aligned terms.

Table 3.10 shows all possible cases of correspondence between traditional/simplified Chinese characters and Japanese characters.

- The Japanese words made up of kanji in the columns "All same" and "TC different" (Traditional Chinese different) could be compared with Chinese directly without any conversion;

Table 3.10: Correspondence between Chinese and Japanese characters.

Relationship	All same	TC different	SC different	All different	Ja different
Meaning	basic	number	intestines	agent	collect
Japanese	基	数	腸	剤	収
T Chinese	基	數	腸	劑	收
S Chinese	基	数	肠	剂	收

- The Japanese characters in “SC different” (Simplified Chinese different) become comparable when Japanese (traditional Chinese) is converted to simplified Chinese;
- For the “All different” and “Ja different” parts we propose to utilize hanzi-kanji conversion table to make them comparable with simplified Chinese.

We combined three different freely available sources of data to maximize our conversion results. The first source of data we used is the Unihan database[8]. In particular we used the correspondence relation SimplifiedVariant in the Unihan_Variant of the Unihan database. The second source of data we used is the Langconv traditional-simplified conversion data[9]. It contains a database for traditional-simplified characters. The third source of data we used concerns the case where the characters in Japanese are proper to Japanese. For this case, we used a hanzi-kanji conversion table, provided in the resource 簡体字と日本漢字対照表[10] which consists pairs of simplified hanzi and kanji. Table 3.11 shows the results of extracting bilingual multi-word terms by kanji-hanzi conversion using these three sources of data.

[8] http://www.unicode.org/Public/UNIDATA/

[9] http://code.google.com/p/advanced-langconv/source/browse/trunk/langconv/?r=7

[10] http://www.kishugiken.co.jp/cn/code10d.html

TABLE 3.11: Extraction of bilingual multi-word terms using kanji-hanzi conversion.

	Zh	Ja	Converted-Ja	Meaning	Human assessment
Without any Conversion	官能␣基	官能␣基	官能␣基	‘functional group’	yes
	肺癌	肺␣癌	肺␣癌	‘lung cancer’	yes
	免疫原	免疫␣原	免疫␣原	‘immunogen’	yes
	透析液	透析␣液	透析␣液	‘dialyzate’	yes
	数␣密度	数␣密度	数␣密度	‘number density’	yes
By Traditional-Simplified Conversion	脉管	脈␣管	脉␣管	‘vessel’	yes
	肠壁	腸␣壁	肠␣壁	‘intestinal wall’	yes
	高温␣杀菌	高温␣殺菌	高温␣杀菌	‘high temperature sterilization’	yes
	放射线␣源	放射␣線␣源	放射␣线␣源	‘radiation source’	yes
By Hanzi-kanji Conversion Table	乘员␣保护␣方法	乗員␣保護␣方法	乘␣保护␣方法	‘occupant protection method’	yes
	心收缩␣期	心␣収縮␣期	心␣收缩␣期	‘systole’	yes
	废热␣回收	廃␣熱␣回収	废␣热␣回收	‘waste heat recovery’	yes
	肺气肿	肺␣気腫	肺␣气肿	‘pulmonary emphysema’	yes
	添加剂	添加␣剤	添加␣剂	‘additive’	yes
	肝脏␣再生␣作用	肝臓␣再生␣作用	肝脏␣再生␣作用	‘liver regeneration action’	yes

3.3.2 Experiments and Results

Table 3.12 and Table 3.13 show two experimental settings: Table 3.12 gives the same setting as given in Section 3.2.4 for Table 3.2 for bilingual multi-word term extraction and SMT experiments. Table 3.13 gives the setting which the tuning set is different from the setting given in Table 3.12. We propose to compare the BLEU scores of SMT experiments base on these two experimental settings.

Table 3.12: Statistics on our experimental data sets (after tokenizing and lowercasing). Here 'mean ± std.dev.' gives the average length of the sentences in words. Tuning set = 500 lines.

	Baseline	Chinese	Japanese
train	sentences (lines)	100,000	100,000
	words	2,314,922	2,975,479
	length in words (avg. ± std.dev.)	23.29 ± 11.69	29.93 ± 13.94
tune	sentences (lines)	**500**	**500**
	words	14,251	17,904
	length in words (avg. ± std.dev.)	28.61 ± 21.88	35.94 ± 25.07
test	sentences (lines)	1,000	1,000
	words	27,267	34,292
	length in words (avg. ± std.dev.)	27.34 ± 15.59	34.38 ± 18.78

Table 3.13: Statistics on our experimental data sets (after tokenizing and lowercasing). Here 'mean ± std.dev.' gives the average length of the sentences in words. Tuning set = 1000 lines.

	Baseline	Chinese	Japanese
train	sentences (lines)	100,000	100,000
	words	2,314,922	2,975,479
	length in words (avg. ± std.dev.)	23.29 ± 11.69	29.93 ± 13.94
tune	sentences (lines)	**1,000**	**1,000**
	words	28,203	35,452
	length in words (avg. ± std.dev.)	28.31 ± 17.52	35.61 ± 20.78
test	sentences (lines)	1,000	1,000
	words	27,267	34,292
	length in words (avg. ± std.dev.)	27.34 ± 15.59	34.38 ± 18.78

We extracted 4,591 bilingual multi-word terms (100% good match) from 309,406 phrase alignments obtained by word-to-word (token-to-token) alignment from the Chinese–Japanese re-tokenized training corpus using kanji-hanzi conversion as described in Section 3.3.1. The numbers of extracted multi-word terms using kanji-hanzi conversion combined with further filtering by constraints are given in Table 3.14 and Table 3.15 (column (a + b + c)). The percentage of good match terms is over 80%, when the threshold is greater than 0.2. We obtained the highest percentage with 93% for a threshold of 0.9 by combining kanji-hanzi conversion and further filtering methods.

We build several Chinese–Japanese training corpora re-tokenized with:

- several thresholds P for filtering (Table 3.15 (a)) (the same as the results shown in Table 3.3)
- further filtering by constraints with several thresholds combined with kanji-hanzi conversion results (Table 3.15 (a +b + c))

We train several Chinese-to-Japanese SMT systems using the standard GIZA++/MOSES pipeline (Koehn et al., 2007). The Japanese corpus without re-tokenization is used to train a language model using KenLM (Heafield, 2011). After removing non-space word separators from the phrase table, we tune and test. In all experiments, the same data sets are used, the only difference being whether the training data is re-tokenized or not with bilingual multi-word terms. Table 3.14 presents the results of the evaluation in Chinese-to-Japanese translation in BLEU scores (Papineni et al., 2002) with a tuning set of 500 lines. Table 3.15 shows the results for a tuning set of 1,000 lines.

TABLE 3.14: Evaluation results in BLEU (tuning = 500 lines) for Chinese-to-Japanese translation based on a re-tokenized training corpus using different thresholds (a), constraints on the ratio of lengths and on the components (b) and kanji-hanzi conversion (c).

	Filtering by thresholds P (a)					Filtering by thresholds P (a) + ratio of lengths + components (b) + kanji-hanzi conversion (c)						
Thresholds P	♯ of bilingual multi-word terms (a)		Good match	BLEU	p-value	♯ of bilingual multi-word terms (a + b)	Good match	♯ of bilingual multi-word terms (a + b + c)		Good match	BLEU	p-value
	Type	Token	(type)					Type	Token	(type)		
$\geq$ 0.0	52,785	136,235	35%	32.44±1.07	> 0.05	48,239	63%	49,474	150,145	70%	33.19±1.09	< 0.01
$\geq$ 0.1	31,795	114,287	52%	32.23±1.18	> 0.05	29,050	68%	30,516	131,997	78%	33.09±1.06	< 0.01
$\geq$ 0.2	27,916	107,790	58%	32.00±1.16	> 0.05	25,562	75%	27,146	126,795	83%	33.12±1.06	< 0.01
Baseline	-	-	-	**32.35±1.15**	-	-	-	-	-	-	**32.35±1.15**	-
$\geq$ 0.3	25,404	102,447	63%	33.08±1.12	< 0.01	23,321	78%	25,006	122,531	83%	33.25±1.14	< 0.01
$\geq$ 0.4	23,515	97,488	72%	32.77±1.15	< 0.05	21,644	80%	23,424	118,674	84%	33.31±1.14	< 0.01
$\geq$ 0.5	21,846	93,143	76%	33.02±1.14	< 0.01	20,134	85%	22,000	115,145	88%	33.23±1.19	< 0.01
$\geq$ **0.6**	**20,248**	84,967	78%	**33.32±1.15**	< 0.01	**18,691**	88%	**20,679**	108,130	89%	**33.75±1.08**	< 0.01
$\geq$ 0.7	18,759	84,908	79%	32.85±1.19	< 0.01	17,340	88%	19,460	104,298	90%	33.41±1.11	< 0.01
$\geq$ 0.8	17,311	71,048	79%	33.25±1.06	< 0.01	16,001	89%	18,265	99,314	90%	33.38±1.14	< 0.01
$\geq$ 0.9	15,464	59,567	80%	33.20±1.15	< 0.01	14,284	92%	16,814	90,660	93%	33.43±1.15	< 0.01

Table 3.15: Same as Table 3.14, but for a tuning set of 1,000 lines is different.

	Filtering by thresholds P (a)					Filtering by thresholds P (a) + the ratio of lengths + the components (b) + kanji-hanzi conversion (c)						
Thresholds P	♯ of bilingual multi-word terms (a)		Good match	BLEU	p-value	♯ of bilingual multi-word terms (a + b)	Good match	♯ of bilingual multi-word terms (a + b + c)		Good match	BLEU	p-value
	Type	Token	(type)					Type	Token	(type)		
$\geq$ 0.0	52,785	136,235	35%	32.63±1.15	> 0.05	48,239	63%	49,474	150,145	70%	33.15±1.09	< 0.01
$\geq$ 0.1	31,795	114,287	52%	32.76±1.18	> 0.05	29,050	68%	30,516	131,997	78%	33.10±1.15	< 0.01
$\geq$ 0.2	27,916	107,790	58%	32.57±1.10	> 0.05	25,562	75%	27,146	126,795	83%	33.05±1.11	< 0.01
Baseline	-	-	-	**32.38±1.16**	-	-	-	-	-	-	**32.38±1.16**	-
$\geq$ 0.3	25,404	102,447	63%	33.07±1.13	< 0.01	23,321	78%	25,006	122,531	83%	33.21±1.10	< 0.01
$\geq$ 0.4	23,515	97,488	72%	32.92±1.13	< 0.01	21,644	80%	23,424	118,674	84%	33.29±1.10	< 0.01
$\geq$ 0.5	21,846	93,143	76%	33.05±1.11	< 0.01	20,134	85%	22,000	115,145	88%	33.38±1.12	< 0.01
$\geq$ **0.6**	**20,248**	84,967	78%	**33.61±1.17**	< 0.01	**18,691**	88%	**20,679**	108,130	89%	**33.93±1.12**	< 0.01
$\geq$ 0.7	18,759	84,908	79%	32.92±1.18	< 0.01	17,340	88%	19,460	104,298	90%	33.43±1.13	< 0.01
$\geq$ 0.8	17,311	71,048	79%	33.34±1.14	< 0.01	16,001	89%	18,265	99,314	90%	33.41±1.14	< 0.01
$\geq$ 0.9	15,464	59,567	80%	33.47±1.14	< 0.01	14,284	92%	16,814	90,660	93%	33.52±1.13	< 0.01

To participate in the evaluation campaign of the 3rd Workshop on Asian Translation (WAT 2017)[11], we test 2,000 sentences based on this best SMT system (tuning = 1,000 lines and (a + b + c) with threshold of 0.6 in Table 3.15) and the baseline system (tuning = 1,000 lines). We obtain a significant increased in BLEU score: 33.61 compared with 32.29 for the baseline system (p-value < 0.01).

3.3.3 Analysis of the Results and Discussion

We compare each BLEU score obtained with a tuning set of 1,000 lines with the BLEU scores obtained with a tuning set of 500 lines in each comparable SMT experiments. The similar results were obtained with a tuning set of 1,000 lines compared with a tuning set of 500 lines.

When comparing the BLEU scores with the baseline system (tuning set of 1,000 lines) in Table 3.15,

- for the training corpus re-tokenized with the results of several thresholds P for filtering (a), we obtain significant improvements as soon as the threshold on translation probabilities becomes greater than 0.3. A statistically significant improvement of 1.2 BLEU point (p-value of 0.001) is observed when the threshold is greater than 0.6. In the case of 0.6, the training corpus contains 20,248 re-tokenized bilingual multi-word terms.
- for the training corpus re-tokenized with further filtering combined with kanji-hanzi conversion results (a + b + c), we obtain significant improvements for all thresholds. We obtain a 1.5 BLEU point (threshold of 0.6) improvement over the baseline system. In this case, 20,679 re-tokenized terms are used. It also improves by 0.3 BLEU point in comparison with the case where the bilingual terms are filtered only by thresholds (a).

We also compare a system (tuning of 1,000 lines) based on a re-tokenized training corpus with further filtering results for a threshold of 0.6 combined with kanji-hanzi conversion results with a baseline system. We investigate the N (Chinese) × M (Japanese)-gram distribution in the phrase tables potentially used in translation. These phrase tables only contain the potentially useful phrase pairs which have some chance to be used in the translation of the test set (before translation). MOSES discards all entries which do not appear in the test set. In Tables 3.16 and 3.17, the statistics (Chinese→Japanese) show that the total number of potentially useful phrase pairs used in translation based on the re-tokenized corpus is larger than that used in the baseline system. We compare the number of entries, the number of phrase pairs shows a significant increase in comparison with the baseline system.

Table 3.16: Distribution of N (Chinese) × M (Japanese)-gram entries in the phrase table potentially used in testing using a C-value/sampling-based + kanji-hanzi conversion method (threshold with 0.6). The bold face numbers show the N (Chinese) × M (Japanese)-grams which increased in comparison to the baseline system.

		Target = Japanese							
		1-gram	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	total
Source = Chinese	1-gram	30074	**86392**	**78101**	**48643**	**27069**	**14255**	**7461**	291996
	2-gram	**14063**	39245	**42304**	**27707**	**15587**	**8214**	**4306**	151427
	3-gram	**1484**	**4021**	8052	**7256**	**4674**	**2576**	**1307**	29370
	4-gram	**172**	**430**	1109	2117	**1869**	**1308**	**685**	7690
	5-gram	23	46	163	378	**667**	**566**	**377**	2220
	6-gram	4	7	12	57	106	183	164	533
	7-gram	0	0	1	2	19	42	73	137
	total	45820	130141	129742	86160	49991	27144	14373	**483373**

Table 3.17: Distribution of N (Chinese) × M (Japanese)-gram in the phrase table for the baseline system potentially used in testing.

		Target = Japanese							
		1-gram	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	total
Source = Chinese	1-gram	32320	84308	71713	42518	22831	11726	6035	271451
	2-gram	13570	39534	41775	25628	13703	6922	3518	144650
	3-gram	1384	3906	8067	7117	4276	2238	1093	28081
	4-gram	163	413	1124	2124	1853	1248	614	7539
	5-gram	27	50	154	386	658	562	360	2197
	6-gram	6	9	13	59	116	181	164	548
	7-gram	1	1	3	5	20	50	73	153
	total	47471	128221	122849	77837	43457	22927	11857	454619

We also investigate the distribution of the phrases actually used during the translation of the test set by inspection of traces of translation. Tables 3.18 and 3.19 show the distribution of phrases used during testing in our proposed method (monolingual term extraction by C-value, bilingual terms aligned by the sampling-based alignment method + kanji-hanzi conversion bilingual multi-word term extraction method for re-tokenizing training corpus) and in the baseline system. From these tables, we see that more uni-grams and bi-grams are actually used in Chinese with our method than with the baseline system. These uni-grams or bi-grams were translated into 1-gram to 7-gram phrases in Japanese. The improved translation accuracy (Table 3.14 and Table 3.15) and the analysis of the increase of potentially used and actually used phrase pairs are respectively the effect and the cause of the impact of our method of re-tokenizing the training corpus with bilingual aligned terms.

Figure 3.5 gives an example of improvement in Chinese-to-Japanese translation, thanks to our method. Re-tokenizing the training corpus with bilingual terms gave a better translation accuracy (BLEU=65.74) of the test sentence in this example. Re-tokenizing

[11] `http://lotus.kuee.kyoto-u.ac.jp/WAT/index.html`

Table 3.18: Distribution of phrases used during testing in a system that uses: a C-value/sampling-based + hanzi/kanji conversion bilingual multi-word term extraction method for re-tokenizing the training corpus (threshold with 0.6). The bold face numbers show the increased N (Chinese) × M (Japanese)-grams actually used in decoding in our SMT experiment.

		Target = Japanese							
		1-gram	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	total
Source = Chinese	1-gram	**9364**	**2252**	**534**	**190**	**64**	**12**	3	**12419**
	2-gram	**616**	**2725**	**1001**	407	176	**94**	**38**	**5057**
	3-gram	62	253	393	218	119	**59**	**27**	1131
	4-gram	**6**	16	35	64	56	25	14	216
	5-gram	4	1	3	10	22	13	7	60
	6-gram	0	0	**2**	**2**	1	11	4	20
	7-gram	0	0	0	0	1	0	3	4
	total	10052	5247	1968	891	439	214	96	**18907**

Table 3.19: Distribution of phrases used during testing in the baseline system.

		Target = Japanese							
		1-gram	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	total
Source = Chinese	1-gram	9144	2086	478	183	40	9	3	11943
	2-gram	615	2593	980	414	184	84	30	4900
	3-gram	69	259	439	237	127	57	22	1210
	4-gram	3	25	53	94	67	31	22	295
	5-gram	5	1	4	16	30	17	16	89
	6-gram	0	0	1	1	3	16	11	32
	7-gram	0	0	0	0	2	3	6	11
	total	9836	4964	1955	945	453	217	110	18480

and grouping the bilingual multi-word term together increased the probability of multi-word term to multi-word term translation, i.e., “定向 控制 模块” to “指向 性 制御 モジュール” ‘directivity control module’. This prevents the erroneous 1-to-1 gram translation of isolated source words, like “定向” ‘orientation’ to “ように する ことが できる” ‘can become like that’. Figure 3.6 gives another example of improvement in Chinese-to-Japanese translation, i.e., “免疫 检测 方法” to “免疫 測定 方法” ‘immunoassay’ in this example. Our proposed method prevents the separated 1-to-1 or 2-to-2 gram translation of isolated source words in inappropriate order or position, like “免疫” to “免疫” ‘immunity’ and “检测 方法” to “測定 方法” ‘measuring method’. In these examples, re-tokenization of the training corpus with extracted bilingual aligned terms induced a direct and exact translation.

Test sentence (Chinese):	如图(0) 5(1) 中(2) 所(3) 示(4) ，(5) 定向(6) 发射(7) 序列(8) 536(9) 被(10) 发送(11) 到(12)
	定向(13) 控制(14) 模块(15) 516(16) 。(17)

Baseline (BLEU=53.71):	図 5 に 示す ように |0-4| 、 定向 |5-6| 制御 |14-14| モジュール |15-15| 516 |16-16| 送信 |7-7| シーケンス |8-8| 536 |9-9| 送信 さ れる |10-12| ように する こと が できる |13-13| 。 |17-17|
Re-tokenizing training corpus with bilingual aligned terms (**BLEU=65.74**):	図 5 に 示す |0-3| ように 、 |4-5| 配向 |6-6| 送信 |7-7| シーケンス |8-8| 536 |9-9| 指向 性 制御 モジュール |13-15| 516 |16-16| に 送信 さ れる |10-12| 。 |17-17|
Reference (Japanese):	図 5 に 示す ように 、 指向 性 シーケンス 536 は 、 指向 性 制御 モジュール 516 に 送信 さ れる 。

FIGURE 3.5: Example of Chinese-to-Japanese technical translation (Example-1). The letters in parentheses show the translation correspondence.

Test sentence (Chinese):

作为(0) 测定(1) 被(2) 检液(3) 中(4) 的(5) 特定(6) 成分(7) 的(8) 方法(9) ，(10) 存在(11) 许多(12)
利用(13) 了(14) 抗原(15) 抗体(16) 反应(17) 的(18) 免疫(19) 检测(20) 方法(21) 。(22)

Baseline (BLEU=15.92):

測定 |1-1| は |2-2| 、 |10-11| 多く の |12-12| 方法 |9-9| と して |0-0| は |13-13| 、 |14-14|
抗原 抗体 |15-16| 反応 の |17-18| 免疫 |19-19| 検液 |3-3| 内 の |4-5| 特定 の |6-6| 成分 |7-7| の |8-8|
測定 |20-20| 方法 |21-21| 。 |22-22|

Re-tokenizing training corpus with bilingual aligned terms (**BLEU=25.54**):

測定 |1-1| が |2-2| 液 |3-3| 内 の |4-5| 特定 の |6-6| 成分 の |7-8| 方法 |9-9| と して |0-0| 、 |10-11|
抗原 抗体 反応 さ せ |15-17| の |18-18| 免疫 測定 方法 |19-21| に ついて は 多数 の |12-12|
利用 |13-13| さ れて いる |14-14| 。 |22-22|

Reference (Japanese):

被 検 液 中 の 特定 成分 を 測定 する 方法 と して 、 抗原 抗体 反応 を 利用 した
免疫 測定 方法 が 数多く 存在 する 。

FIGURE 3.6: Example of Chinese-to-Japanese technical translation (Example-2). The letters in parentheses show the translation correspondence.

To summarize these three sections (Section 3.3.1, 3.3.2 and 3.3.3), we used kanji-hanzi conversion to extract bilingual multi-word terms and single-word to multi-word terms which could not be extracted using thresholds in bilingual multi-word extraction. This also allowed us to extract bilingual aligned terms made up of hanzi/kanji that were recognized in one language as a multi-word term but not in the other language. By using kanji-hanzi conversion, more reliable bilingual aligned terms could be retrieved or reinforced thanks to the similarity between hanzi and kanji. We still did not use any other additional corpus or lexicon in this work. The results of our experiments indicate that the combination of the bilingual multi-word terms extracted have over 80% precision for a threshold of 0.2. We obtained the highest precision with 93% for a threshold of 0.9. Re-tokenizing the parallel training corpus with these terms led to statistically significant improvements in BLEU scores for each threshold: 1.5 BLEU point (p-value of 0.001) improvements over the baseline system (threshold of 0.6, tuning of 1,000 lines).

3.3.4 Bilingual Terms which not Share any Hanzi/Kanji

This section describes the work on the extraction of single-word to multi-word terms which do not share any hanzi/kanji character or partly share some characters, so as to increase the quantity of the extracted bilingual terms between Chinese and Japanese. This work is justified by the fact that some single-word to multi-word bilingual terms cannot be extracted by using the kanji-hanzi conversion method. We present evaluation results of SMT experiments based on the second experimental protocol (System re-tok-train-only) as described in Section 3.2.3 and compare them with the results obtained in previous sections.

Table 3.20 shows the extraction of bilingual aligned multi-word terms using kanji-hanzi conversion by setting a threshold of 0.6. It leads to the observation that some correctly aligned bilingual terms made up of hanzi/kanji on both sides do not share any (or all) character (or characters), or that only one side is a multi-word term but the components of the terms in Japanese are made up of katakana or kanji+katakana, so that they cannot be extracted by our proposed methods described in previous sections (Section 3.2 and Section 3.3.1).

In this section, we propose to extract these kind of bilingual terms based on the results obtained by the sampling-based alignment method without considering the components (hanzi/kanji). Table 3.21 shows the constrains for further filtering these kind of extracted bilingual terms. Similar to the filtering constraints adopted for filtering multi-word to multi-word terms, we also consider the ratio of the lengths in words for the Chinese (Japanese) part to the corresponding Japanese (Chinese) part and the components of the Japanese part.

TABLE 3.20: Extraction of bilingual aligned multi-word terms using kanji-hanzi conversion with a threshold of 0.6. Yes and no in the but last column indicate whether the bilingual multi-word term (or one side is multi-word term) alignment is kept or excluded. Yes and no in the last column indicate whether the extracted multi-word term pairs (or single-word to multi-word pairs) where judged to be correct alignments by manual check.

	Chinese	Japanese	Meaning	$P(t\|s)$	$P(s\|t)$	Kept	Good match
Sec. 3.2	主机␣系统	ホスト␣システム	‘host system’	1.000	1.000	yes	yes
	命令␣信息	コマント␣情報	‘command information’	1.000	0.671	yes	yes
	顶盖␣主体	キャップ␣本体	‘cap body’	1.000	0.833	yes	yes
	冷却␣层	冷却␣層	‘cooling layer’	1.000	0.951	yes	yes
	薄␣膜片	薄膜␣シート	‘filmcoated tablets’	1.000	1.000	yes	yes
	肺气肿	肺␣気腫	‘pulmonary emphysema’	0.818	0.900	yes	yes
Sec. 3.3.1	激振␣电极	主に␣形成	-	0.862	0.982	no	no
	芯片␣级␣控制␣手机␣模块	チップ␣レベル	-	1.000	1.000	no	no
	废␣热	廃␣熱	‘waste heat’	0.844	0.241	yes	yes
	变速␣机	変速␣機	‘variable-speed motor’	1.000	0.006	yes	yes
	壁部	壁␣部	‘wall part’	0.948	0.678	yes	yes
	核酸	核␣酸	‘nucleic acid’	0.974	0.956	yes	yes
	极板	極␣板	‘electrode plate’	0.992	1.000	yes	yes
	薄␣膜	薄膜	‘thin film’	0.198	0.058	yes	yes
	贵␣金属	貴金属	‘noble metal’	0.990	0.985	yes	yes
Sec. 3.3.4	缺氧	酸素␣欠乏	‘oxygen deficit’	0.957	0.984	no	yes
	供油路	給油␣路	‘feed oil passage’	1.000	1.000	no	yes
	输入␣输出	入出力	‘in-out’	0.952	0.811	no	yes
	制动液	ブレーキ␣液	‘brake fluid’	0.985	0.902	no	yes
	甲醛	ホルム␣アルデヒド	‘formaldehyde’	0.997	0.910	no	yes
	存储器␣控制器	メモリコントローラ	‘memory controller’	0.969	0.918	no	yes
	枢轴␣板	ピボットプレート	‘pivot plate’	0.977	1.000	no	yes
	切换␣步骤	Handover	-	1.000	1.000	no	no
	亭	キオスク␣端末	-	1.000	1.000	no	no
	飞行物	前記␣飛行␣体	-	1.000	1.000	no	no
	总␣能量␣消耗量	総計	-	1.000	1.000	no	no

TABLE 3.21: Constraints for filtering the extracted single-word to multi-word terms. We found that most of the filtered out alignments are not good match (95% of these alignments are not correct). ○ shows the constraints used in further filtering.

Filtering constraints	Component filtering	Filtering based on the length			
	hiragana, digits, letters in zh or ja	one character in either one side	len(zh)/len(ja) > 2 and len(ja)/len(zh) > 2	len(ja)/len(zh) >= 3/1	multi-word to one len(zh) > len(ja)
examples	显影␣辊↔近い␣位置 电源␣管理部↔完全に␣ストップ r-o↔逆␣浸透 切换␣步骤↔Handover 堵塞␣试验↔塞がり 0.1-50↔基板␣厚み	亭↔キオスク␣端末 渡↔キャリア␣走行␣層 撑↔テンプル␣装置 接␣装置↔継 伞的␣结构↔傘 管制␣员↔官	芯片␣级␣控制␣手机␣模块↔チップ␣レベル 糖尿病␣小鼠␣中肾␣小管␣上皮␣细胞↔上皮␣細胞 铃音␣服务器↔マルチメディア␣リング␣バック␣トーン␣サーバ 上述␣液体状↔前記␣アルカリ␣活性␣結合␣材 输送␣方向␣下游␣侧␣端␣部↔ローダ␣ホッパ 直接␣操作␣图像␣形成␣装置↔トラブル␣状況	飞行物↔前記␣飛行␣体 调芯↔調␣芯␣実装␣器 转导↔形␣質␣導入 频分↔直交␣波␣周波 装载部↔半導体␣チップ␣搭載␣部 可反转↔反転␣可能␣ローラ	重复␣购买␣协议␣申请↔役務 三线态␣能量␣传递↔三重 总␣能量␣消耗量↔総計 技术␣思想↔思想 曝气␣搅拌机↔曝気 无及␣强弱↔強弱
multi-word to multi-word term	○	-	○	-	-
single-word to multi-word term	○	○	-	○	○

3.3.5 Experiments and Results

Table 3.22 shows the experimental setting for considering single-word to multi-word term extraction and SMT experiments. This experimental setting is the same as shown by Table 3.2 in Section 3.2.4, and Table 3.12 in Section 3.3.2. The justification for doing experiments with a tuning set of 500 lines is that we may obtain similar results in less time (saving time in tuning).

Table 3.22: Statistics on our experimental data sets (after tokenizing and lowercasing). Here 'mean ± std.dev.' gives the average length of the sentences in words. Tuning set of 500 lines.

	Baseline	Chinese	Japanese
train	sentences	100,000	100,000
	words	2,314,922	2,975,479
	length in words (avg. ± std.dev.)	23.29 ± 11.69	29.93 ± 13.94
tune	sentences	**500**	**500**
	words	14,251	17,904
	length in words (avg. ± std.dev.)	28.61 ± 21.88	35.94 ± 25.07
test	sentences	1,000	1,000
	words	27,267	34,292
	length in words (avg. ± std.dev.)	27.34 ± 15.59	34.38 ± 18.78

The basic steps are as follows:

- Extract single-word to multi-word candidate terms from the results obtained by the sampling-based alignment method;
- Filter the extracted bilingual candidate terms by the constraints given in Table 3.21;
- Combine the results obtained by multi-word to multi-word term extraction (further filtering results), kanji-hanzi conversion based extraction and single-word to multi-word term extraction without hanzi/kanji component constraints;
- Re-tokenize the training corpus with the combined bilingual terms;
- Perform SMT experiments with re-tokenized training data and compare the results with the baseline and previous results.

Table 3.23 shows the evaluation results for Chinese-to-Japanese translation based on the training corpus re-tokenized using the combined techniques for extraction of bilingual terms.

Table 3.23: Evaluation results in BLEU for Chinese-to-Japanese translation based on re-tokenized training corpus using different thresholds (a), the ratio of the lengths + the components (b) and kanji-hanzi conversion (c) + single-word to multi-word terms + filtering by constrains (d).

	Filtering by thresholds (a)			Filtering by thresholds (a) + the ratio of the lengths + the components (b) + kanji-hanzi conversion (c)				a + b + c + d				
Thresholds	♯ of bilingual multi-word terms (a)	BLEU	p-value	♯ of bilingual multi-word terms (a + b)	♯ of bilingual multi-word terms (a + b + c)	BLEU	p-value	♯ of one side multi-word terms	♯ of filtered one side multi-word terms (d)	♯ of combination of multi-word terms (a + b + c + d)	BLEU	p-value
$\geq$ 0.0	52,785 (35%)	32.44±1.07	> 0.05	48,239 (63%)	49,474 (70%)	33.19±1.09	< 0.05	72,428 (2%)	27,116 (40%)	75,425 (64%)	32.55	> 0.05
$\geq$ 0.1	31,795 (52%)	32.23±1.18	> 0.05	29,050 (68%)	30,516 (78%)	33.09±1.06	< 0.05	18,395 (7%)	7,570 (55%)	37,059 (78%)	33.36	< 0.01
$\geq$ 0.2	27,916 (58%)	32.00±1.16	> 0.05	25,562 (75%)	27,146 (83%)	33.12±1.06	< 0.05	14,179 (12%)	6,031 (62%)	32,224 (85%)	33.20	< 0.01
Baseline	- -	**32.35±1.15**	-	- -	- -	**32.35±1.15**	-	-	-	-	-	-
$\geq$ 0.3	25,404 (63%)	33.08±1.12	< 0.01	23,321 (78%)	25,006 (83%)	33.25±1.14	< 0.01	11,849 (15%)	5,161 (70%)	29,280 (90%)	33.41	< 0.01
$\geq$ 0.4	23,515 (72%)	32.77±1.15	< 0.05	21,644 (80%)	23,424 (84%)	33.31±1.14	< 0.01	10,259 (17%)	4,537 (76%)	27,125 (90%)	33.37	< 0.01
$\geq$ 0.5	21,846 (76%)	33.02±1.14	< 0.01	20,134 (85%)	22,000 (88%)	33.23±1.19	< 0.01	9,069 (17%)	4,050 (76%)	25,270 (90%)	33.63	< 0.01
$\geq$ **0.6**	**20,248 (78%)**	**33.32±1.15**	< 0.01	18,691 (88%)	**20,679 (89%)**	**33.75±1.08**	< 0.01	7,875 (30%)	3,575 (76%)	23,522 (93%)	**34.27**	< 0.01
$\geq$ 0.7	18,759 (79%)	32.85±1.19	< 0.01	17,340 (88%)	19,460 (90%)	33.41±1.11	< 0.01	6,900 (30%)	3,088 (80%)	21,874 (93%)	33.90	< 0.01
$\geq$ 0.8	17,311 (79%)	33.25±1.06	< 0.01	16,001 (89%)	18,265 (90%)	33.38±1.14	< 0.01	6,026 (30%)	2,726 (80%)	20,318 (93%)	33.85	< 0.01
$\geq$ 0.9	15,464 (80%)	33.20±1.15	< 0.01	14,284 (92%)	16,814 (93%)	33.43±1.15	< 0.01	5,062 (30%)	2,275 (82%)	18,484 (95%)	33.75	< 0.01

3.3.6 Analysis of the Results and Discussion

Compared with the baseline system and the results obtained in previous sections in Table 3.23,

- for the training corpus re-tokenized with combined term extraction results (a + b + c + d), we obtain significant improvements in all thresholds expect we keep all extracted terms (P greater than 0). We obtain nearly 2 BLEU point (threshold of 0.6) improvement in comparison with the baseline system. In this case, 23,522 extracted terms are used.
- We obtain an improvement of 0.6 BLEU point (threshold of 0.6) in comparison with the case where bilingual terms are extracted by further filtering and using kanji-hanzi conversion (a + b + c). We also obtain an improvement of 1 BLEU point (threshold of 0.6) in comparison with the case where bilingual terms are filtered only by thresholds (a).

We also compare a system (tuning of 500 lines) based on a re-tokenized training corpus with combined term extraction results (a + b + c + d) with a threshold of 0.6 to a baseline system. We investigate the N (Chinese) × M (Japanese)-gram distribution in the phrase tables potentially used in translation. In Tables 3.24, the statistics (Chinese→Japanese) show that the total number of potentially useful phrase pairs used in translation based on the combined term extraction results (a + b + c + d) is larger than that used in the baseline system (see Table 3.17). The number of phrase pairs show a significant increase in comparison with the baseline system. It is also larger than that in the system based on further filtering results with the same threshold combined with kanji-hanzi conversion results (see Table 3.16).

Table 3.24: Distribution of N (Chinese) × M (Japanese)-gram entries in the phrase table potentially used in testing based on: multi-word to multi-word terms + single-word to multi-word terms extraction (threshold with 0.6). The bold face numbers show the total number of N (Chinese) × M (Japanese)-grams which increased in comparison with the baseline system.

		Target = Japanese							
		1-gram	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	total
Source = Chinese	1-gram	29675	**86657**	**78804**	**49416**	**28062**	**14982**	**7822**	295418
	2-gram	**14001**	39174	**42711**	**27954**	**15810**	**8348**	**4350**	152348
	3-gram	**1452**	**3921**	7988	**7234**	**4711**	**2572**	**1290**	29168
	4-gram	**178**	**444**	1110	**2147**	**1884**	**1345**	**712**	7820
	5-gram	23	46	172	**388**	**672**	554	**380**	2235
	6-gram	4	7	**15**	58	106	**183**	164	537
	7-gram	1	2	1	6	20	43	73	146
	total	45334	130251	130801	87203	51265	28027	14791	**487672**

We also investigate the distribution of the phrases actually used during the translation of the test set. Table 3.25 shows the distribution of phrases used during testing with the combined term extraction results (a + b + c + d) with a threshold of 0.6. More 1×1-grams, 1×2-grams and 1×3-grams are actually used in Chinese–Japanese phrase alignment than in the baseline system (see Table 3.19) and than in the system with further filtering results on the threshold of 0.6 combined with kanji-hanzi conversion (a + b + c) (see Table 3.18). This analysis of the increase in potentially used and actually used phrase pairs explains the improved translation accuracy (Table 3.23). It makes the impact of our method of re-tokenizing the training corpus with bilingual single-word to multi-word terms.

Table 3.25: Distribution of phrases used during testing based on: multi-word to multi-word terms + single-word to multi-word terms extraction for re-tokenizing training corpus (threshold with 0.6). The bold face numbers show the increased N (Chinese) × M (Japanese)-grams actually used in decoding of SMT experiment.

		Target = Japanese							
		1-gram	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	total
Source = Chinese	1-gram	**10079**	**2968**	**658**	150	34	3	2	**13894**
	2-gram	375	2538	967	314	126	32	15	4367
	3-gram	38	212	408	215	75	26	8	982
	4-gram	3	17	41	**104**	59	25	9	258
	5-gram	5	1	**11**	10	25	**21**	11	84
	6-gram	0	0	1	**2**	3	10	8	24
	7-gram	0	0	0	0	1	1	4	6
	total	10500	5736	2086	795	322	118	58	**19615**

3.4 Summary of This Chapter

We presented an approach to improve the performance of Chinese–Japanese patent machine translation by re-tokenizing the parallel training corpus with extracted bilingual aligned terms. We extracted multi-word terms monolingually from each monolingual part of the corpus by using the C-value method. We re-tokenized each extracted multi-word terms as one token in their monolingual parts of the corpus. We then used the sampling-based alignment method to align the re-tokenized parallel corpus and only kept the aligned bilingual multi-word terms by setting different thresholds on translation probabilities in both directions. We also used kanji-hanzi conversion to extract bilingual aligned terms which could not be extracted using thresholds only. This allowed us to extract more bilingual terms made up of hanzi/kanji that were recognized in one language as a multi-word term but not in the other language. By using kanji-hanzi conversion, more reliable bilingual aligned terms could be retrieved or reinforced thanks to the similarity between hanzi and kanji. We even considered the case of single-word to multi-word aligned terms which do not share or partly share hanzi/kanji characters. We did not use any other additional corpus, lexicon or pivot language in our work for improving segmentation accuracy and translation accuracy. The results of our experiments indicate that the combination of the bilingual aligned terms extracted have over 80% precision for a threshold of 0.2. We obtained the highest precision with 95% for a threshold of 0.9. Re-tokenizing the parallel training corpus with these terms led to statistically significant improvements in BLEU scores for each threshold: about 2 BLEU point improvement (p-value of 0.001) in comparison with the baseline system (threshold of 0.6).

Chapter 4

Quasi-parallel Data Construction

For data-driven machine translation, e.g., statistical machine translation, parallel corpora (parallel sentences) are crucial for training translation systems. To increase the translation accuracy in SMT, the most natural answer is to make the training data larger and larger. However, open linguistic resources between Chinese and Japanese are relatively scarce and most existing resources are not freely available due to copyright restrictions. For solving the scarcity problem of open bilingual corpora between Chinese and Japanese, we propose to make use of well-resourced monolingual data to generate new sentences between Chinese and Japanese (quasi-parallel corpora).

This chapter[1] focuses on the study of construction of quasi-parallel corpora using analogical associations based on freely available monolingual data and an existing parallel training corpora. We present experiments and results in constructing quasi-parallel corpora and adding this kind of quasi-parallel corpora as additional training data for training SMT systems.

The structure of this chapter is as follows.

- Section 4.1 reviews related works and identifies the problem of scarcity of bilingual corpora between Chinese and Japanese.
- In Section 4.2, we present the overview of our proposed method. We propose a method to generate new candidate sentences using analogical associations. In addition, we propose two filtering techniques to increase the grammatical accuracy of the newly generated sentences. Deduction of translation relations between filtered sentences in Chinese and Japanese is also described in this section.
- Section 4.3 shows the experimental data used, the experimental settings and the evaluation results for clustering, new sentence generation, filtering, quasi-parallel corpus construction and SMT experiments, as well as some analysis of the results.

[1] Related to (Yang and Lepage, 2017), (Yang et al., 2014), (Yang and Lepage, 2014a) and (Yang and Lepage, 2014b)

- Section 4.4 describes the experiments and results in translating scientific and technical corpus (ASPEC corpus) using our proposed method. We also investigate the influence of segmentation on translation results.

- Section 4.5 describes the experiments and the results that make use of the proposed techniques described in previous chapters and this chapter.

- Finally, we summarize this chapter in Section 4.6.

4.1 Related Work

In recent years, there have been several approaches developed for obtaining parallel sentences or fragments from non-parallel data (Munteanu and Marcu, 2005), (Munteanu and Marcu, 2006), such as comparable data (Munteanu and Marcu, 2005), (Bin et al., 2010), (Smith et al., 2010), (Chu et al., 2015) and quasi-comparable data (Fung and Cheung, 2004) to make contributions to SMT. *Parallel corpora* contain parallel sentences, i.e., sentences which are translations of each other. The term *comparable corpora* refers to texts in two languages that are similar in meaning or expressions, but are not exact translations. *Quasi-comparable corpora* that contain more disparate very-non-parallel bilingual documents that could either be on the same topic (in-topic) or not (out-topic) (Fung and Cheung, 2004), are more available than *comparable corpora*. In *quasi-comparable corpora*, there are few or no parallel sentences (Chu et al., 2013b). In (Munteanu and Marcu, 2005), they extract parallel sentences from non-parallel corpora by starting with a relatively small parallel corpus and large Chinese, Arabic, and English non-parallel newspaper corpora. They train a maximum entropy classifier to determine which sentences may be aligned. They aim at improving the performance of an SMT system for less-resourced language pairs. Similarly, we also start with an existing small parallel corpus, but combine it with large amounts of monolingual data to construct a *quasi-parallel corpus*. In the method in (Munteanu and Marcu, 2005), the final sentences come from the monolingual corpora. In our method, the final sentences are created by similarity with sentences in the parallel corpus.

Paraphrase generation is another way to make a contribution to SMT. This aims at reducing out-of-vocabulary words and acquiring paraphrases of unknown phrases to increase the model coverage (Jiang et al., 2011). Some of the previous work showed that word lattices constructed to express input sentences in different ways are helpful for obtaining better translation quality (Onishi et al., 2011). A syntax-based algorithm to automatically build word lattices that are used as finite state automata (FSA) to represent paraphrases is described in (Pang et al., 2003). FSAs extract paraphrase pairs and generate new, unseen sentences that contain the same meaning as the input sentences. In our work, we also generate unseen, new sentence pairs (i.e., they do not come from given parallel sentences). However, FSAs are replaced by the resolution of analogical equations to produce new sentences.

Research is growing on analogical learning for NLP applications. In (Langlais and Yvon, 2008), they show how to retrieve all analogies for a given word (i.e., a sequence of letters) in a very fast way, so as to allow the application of analogy to practical tasks. In (Delhay and Miclet, 2004), they present a theoretical generalization of analogies between sequences of letters. They show how to extend elementary analogies between letters of the alphabet to sequences of letters (e.g., $a : b :: c : d$ and $a : \varepsilon :: a : \varepsilon$ imply $aaa : bb :: cca : dd$) based on an edit distance given in (Lepage, 1998), (Pirrelli and Yvon, 1999). In (Lepage and Denoual, 2005b), they use proportional analogies to translate sentences in an example-based machine translation. Translation of unknown words by analogy has also been proposed in (Langlais and Patry, 2007), (Silva et al., 2012). In (Stroppa and Yvon, 2005), they present the basic steps of analogical learning and a definition of formal analogical relationships suitable for learning large datasets in NLP, and use this approach in morphological analysis tasks. Different from these works, in our research, we propose to cluster monolingual Chinese and Japanese short sentences respectively using analogical associations. This allows us to obtain *rewriting models* that can produce new sentences by solving analogical equations.

In (Doddington, 2002), the basic idea of automatic MT evaluation method is introduced by using N-gram co-occurrence statistics. And in (Soricut and Brill, 2004), they describe a framework by using N-gram co-occurrence statistics as an automatic evaluation of NLP applications. To cut down on over-generation, we use filtering by seen N-sequences (Lepage and Denoual, 2005a) or using BLEU (Papineni et al., 2002) to keep only those newly generated sentences which are acceptable in fluency of expression and in adequacy of meaning.

4.2 Overview of the Proposed Method

In this section, we present our proposed method to construct a Chinese–Japanese quasi-parallel corpus by using analogical associations. The overview of our method is given in Figure 4.1. The procedure in our method has four steps:

(1) Construction of analogical clusters.

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

(2) Generation of new sentences.

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

(3) Filtering over-generated sentences.

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

(4) Deduction of translation relations.

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

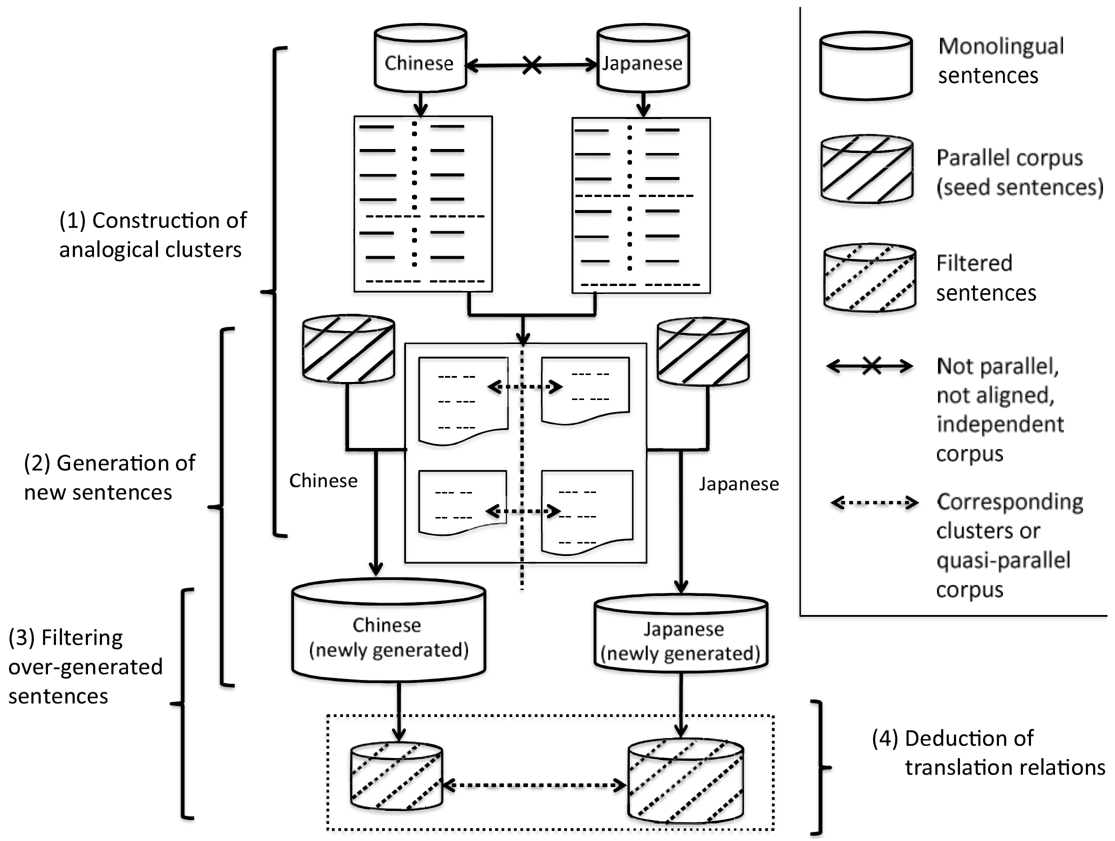
Chinese
Japanese
(1) Construction of analogical clusters
(2) Generation of new sentences
Chinese
Japanese
(3) Filtering over-generated sentences
Chinese (newly generated)
Japanese (newly generated)
(4) Deduction of translation relations
Monolingual sentences
Parallel corpus (seed sentences)
Filtered sentences
Not parallel, not aligned, independent corpus
Corresponding clusters or quasi-parallel corpus

Figure 4.1: Overview of the proposed method: construction of a Chinese–Japanese quasi-parallel corpus.

4.2.1 Clustering and Generation of New Sentences

4.2.1.1 Construction of Analogical Clusters

(1) Sentential analogies:

Gentner (1983), Lepage (2004) and Yvon et al. (2004) gave different definitions of proportional analogies. The common notion is that proportional analogies establish a structural relationship between four objects, A, B, C and D. It is written $A : B :: C : D$ ('A is to B as C is to D').

Analogies can be classified as being semantical or formal. Examples of semantic analogy are:

hand : glove :: foot : shoe

traffic : street :: water : riverbed

For such semantic analogy, Turney (2006) gives a definition of verbal analogies based on high relational similarity.

On the other hand, examples of formal analogy are:

walk : walked :: work : worked

to create : creator :: to translate : translator

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

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

An efficient algorithm for the resolution of analogical equations between strings of characters has been proposed by Lepage (1998). The algorithm relies on counting numbers of occurrences of characters and computing edit distances between strings of characters ($d\,(A, B) = d\,(C, D)$ and $d\,(A, C) = d\,(B, D)$). It is given by Formula 4.1. where $|A|_a$ stands for the number of occurrences of character a in string A and $d\,(A, B)$ stands for the edit distance between strings A and B with only insertion and deletion as edit operations. As B and C may be exchanged in an analogy, the constraint on edit distance has also to be verified for $A : C :: B : D$, i.e., $d\,(A, C) = d\,(B, D)$. The algorithm uses fast bit string operations and distance computation: $d(A,B) = |A| + |B| - 2 \times s(A,B)$ (Allison and Dix, 1986).

$$A : B :: C : D \Rightarrow \begin{cases} |A|_a - |B|_a = |C|_a - |D|_a, \forall a \\ \quad d(A,B) = d(C,D) \\ \quad d(A,C) = d(B,D) \end{cases} \tag{4.1}$$

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

本当に迷惑です。 : とても迷惑です。 :: 本当に困っています。 : とても困っています。

It's really annoying. : *It's very annoying.* :: *I'm really troubled.* : *I'm very troubled.*

(2) Analogical clusters:

When several sentential analogies involve the same pairs of sentences, they form a series of analogous sentences. They can be written on a sequence of lines where each line contains one sentence pair and any two pairs of sentences form a sentential analogy. We call this an *analogical cluster*. The size of a cluster is the number of its sentential pairs (=lines). The clusters contain at least 2 pairs of sentences. Figures 4.2 and 4.3 show two examples of clusters in Japanese. For each example, there are three possible sentential analogies.

We give more examples of constructed analogical clusters and investigate the contents of these clusters. The sentence pairs on the left and right are not necessarily paraphrases, i.e., they do not necessarily have the similar meaning. Because different clusters illustrate different linguistic or semantic features, the same sentence may appear in different clusters. For instance, the Japanese sentence 改善お願いします。 /kaizen onegai shimasu/ 'Improve it, please.' appears on the right in the cluster in Figure 4.4 (indicated with a '▲'). The linguistic interpretation of this cluster is that the noun 検討 /kentou/ 'investigate' is exchanged with 改善 /kaizen/ 'improve' in similar situational and

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

Figure 4.2: An example of an analogical cluster in Japanese exhibiting the exchange of " 本当に " /hontoni/ with " とても " /totemo/.

表示されません	:	表示されなくなりました
'Is not displayed'		*'No longer able to be displayed'*
ブログが投稿できません	:	ブログが投稿できなくなりました
'Cannot post on blog'		*'No longer able to post on blog'*
記事の編集ができません	:	記事の編集ができなくなりました
'Cannot edit the article'		*'No longer able to edit the article'*
⋮	:	⋮

Figure 4.3: An example of an analogical cluster in Japanese exhibiting the exchange of " ません " /masen/ with " なくなりました " /naku narimashita/.

structural contexts in different meanings. The same sentence also appears in the cluster in Figure 4.5. This cluster shows the insertion of the degree adverbial よろしく /yoroshiku/. In terms of linguistic features, it lies between a neutral and a more polite form of expression.

We also found that the position of the changes in a cluster is not necessarily exactly the same. As shown in Figure 4.6, obviously the position of insertion of the Chinese adverbial 非常 /fēicháng/ 'very much' in sentence pairs 1, 2 and 4 is different from that observed in the sentence pairs 3 and 5.

From the point of view of the size of the clusters, the largest cluster for Chinese in our experiment contains 240 pairs of sentences. The interpretation of this cluster is the insertion of the Chinese degree adverbial 很 /hěn/ 'very'. It is similar to the clusters we give in Figure 4.6. The largest cluster for Japanese contains 192 pairs of sentences. This cluster exhibits similar phenomena as the cluster shown in Figure 4.5. It lies between a neutral and a more polite form of speaking or expresses a solemn decision by adding the auxiliary verb です /desu/.

The next largest and the third largest clusters for Chinese contain 125 and 121 sentence pairs respectively, they both show the insertion model of Chinese word 的 /de/. They were separated into two clusters due to the difference of distances between the pairs of sentences on the left and right. They also reflect different linguistic phenomena. The next largest cluster (the distance is 5) shows subject-predicate phrase change to nominal endocentric phrase (Figure 4.7). The third largest clusters (the distance is 1) reflect some kind of usage of the word 的: (1) make a word or phrase into an adjective; (2) change a word or phrase into a demonstrative pronoun; (3) express the relationship between words (4) as the auxiliary word, that in the end of the sentence to strengthen a affirmative tone. All these usages are shown in Figure 4.8.

検討願います。	:	改善願います。
検討をお願いします。	:	▲改善をお願いします。
検討よろしくおねがいします。	:	改善よろしくおねがいします。
検討よろしくお願いします。	:	改善よろしくお願いします。
ご検討ください。	:	ご改善ください。
検討お願いします 。	:	改善お願いします。

Figure 4.4: A Japanese cluster that illustrates the substitution of the verb "検討" /kentou/ 'investigate' with "改善" /kaizen/ 'improve'.

▲改善お願いします。	:	改善よろしくお願いします。
復旧作業お願いします。	:	復旧作業よろしくお願いします。
復旧をお願いします。	:	復旧をよろしくお願いします。
ご確認お願いし ます 。	:	ご確認よろしくお願いします。

Figure 4.5: A Japanese cluster that illustrates the possible insertion of the adverbial "よろしく" /yoroshiku/.

操作方便	:	操作非常方便
效果不错	:	效果非常不错
值得推荐	:	非常值得推荐
孩子喜欢	:	孩子非常喜欢
值得称赞	:	非常值得称赞
......		......

Figure 4.6: A Chinese cluster that illustrates the insertion of the adverbial "非常" /fēicháng/ 'very much'.

A manual inspection of the other larger Japanese clusters obtained shows that the clusters illustrate a range of linguistic phenomena:

- Orthographical variations, mainly for Japanese with writing in kanji vs kana (e.g., 下さい /kudasai/ vs ください /kudasai/, they both mean 'please'.);
- Exchange of place names, people names etc. (e.g., 秋田 /Akita/ and 福島 /Fukushima/.);

画面漂亮	:	漂亮的画面
游戏很好玩	:	很好玩的游戏
故事有趣	:	有趣的故事
软件非常好用	:	非常好用的软件
节奏欢快	:	欢快的节奏
……		……

FIGURE 4.7: A Chinese cluster that illustrates the insertion of the word “的” /de/, shows the subject-predicate phrase change to nominal endocentric phrase.

挺简单	:	挺简单的	(1)
没声音	:	没声音的	(1)
其他	:	其他的	(2)
唱歌	:	唱歌的	(2)
他评论	:	他的评论	(3)
明明是免费	:	明明是免费的	(4)
……		……	

FIGURE 4.8: A Chinese cluster that illustrates the insertion of the word “的” /de/ reflect different linguistic phenomenas.

- Some clusters contain dozens of pairs of sentences that illustrate the exchange of digits (e.g., 8月18日生まれ /hachi gatsu jyuhachi nichi umare/ ‘Born on August 18th.’ and 8月28日生まれ /hachi gatsu nijyuhachi nichi umare/ ‘Born on August 28th.’.);

- Change of attributive or adverbial to other expressions (e.g., change adverbial 超 /chou/ to とても /totemo/, they both mean ‘very much’.);

- etc.

(3) Determining corresponding clusters:

The steps for determining corresponding clusters are:

- First, for each sentence pair in a cluster, we extract the change between the left and the right sides by finding the longest common subsequence (LCS) (Wagner and Fischer, 1974).

- Then, we consider the changes (see $L_{zh} : R_{zh}$ and $L_{ja} : R_{ja}$ in Figure 4.9) between the left (S_{left}) and the right (S_{right}) sides in one cluster as two sets. We perform word segmentation on these changes in sets to obtain minimal sets of changes made up with words or characters.

- Finally, we compute the similarity between the left sets (S_{left}) and the right sets (S_{right}) of Chinese and Japanese clusters. To this end, we make use of the EDR

经典游戏	:	游戏**很不错**
‘classic game’	:	*‘The game is not bad.’*
喜欢**经典**	:	**很不错**喜欢
‘I like classic.’	:	*‘Not bad, I like it.’*
经典啊	:	**很不错**啊
‘Classic!’	:	*‘Not bad!’*

クラシック物語	:	**この**物語**はとてもいい**
‘classic narrative’	:	*‘The narrative is very good.’*
クラシック音楽	:	**この**音楽**はとてもいい**
‘classic music’	:	*‘The music is very good.’*

{ 经典 } : { 很, 不错 } { クラシック } : { この, は, とても, いい }

L_{zh} : R_{zh} L_{ja} : R_{ja}

Figure 4.9: An example of a real case of changes between the left and the right sides in Chinese (L_{zh} : R_{zh}) and Japanese clusters (L_{ja} : R_{ja}). The characters/words in bold face show the changes between the left and right sides of each sentence pair in the clusters and the minimal sets of changes in Chinese or Japanese cluster after segmentation. Note that the sentences in Japanese are not translations of the sentences in Chinese.

dictionary[2], Unihan database[3] and a Kanji-hanzi Conversion Table[4] to translate all Japanese words into Chinese, or convert Japanese characters into simplified Chinese. We calculate the similarity between two Chinese and Japanese word sets according to a classical Dice formula:

$$\text{Sim} = \frac{2 \times |S_{zh} \cap S_{ja}|}{|S_{zh}| + |S_{ja}|} \tag{4.2}$$

S_{zh} and S_{ja} denote the minimal sets of changes across the clusters (both on the left or right) in both languages (after translation and conversion). The formula for computing the similarity between two Chinese and Japanese clusters is given in equation (4.3):

$$\text{Sim}_{C_{zh}-C_{ja}} = \frac{1}{2}(\text{Sim}_{left} + \text{Sim}_{right}) \tag{4.3}$$

Application on the example given in Figure 4.9:
(knowing クラシック=经典, とても=很 and いい=不错)

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

[2] The EDR Electronic Dictionary: National Institute of Information and Communication Technology (NICT). URL: `http://www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html`

[3] `http://www.unicode.org/Public/UNIDATA/`

[4] `http://www.kishugiken.co.jp/cn/code10d.html`

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

4.2.1.2 Generation of New Sentences

Analogy is also a process (Itkonen, 2005) by which, given two related forms and only one form, the fourth missing form is coined (de Saussure, 1995). In our work, in a sentential analogy $A : B :: C : D$, a cluster provides A and B (left to right or right to left) and we use a seed sentence C to generate a new candidate sentence D. The generated D should satisfy the conditions given above on edit distance and number of occurrences of characters. This can be illustrated with the following example:

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

In this example, the solution of the analogical equation is $D =$ 今日はとても楽しかったです。 *'It was very fun today.'*. We generated new sentences with each pair of sentences in clusters for Chinese and Japanese respectively. Because each sentence pair in a cluster is thus a potential rewriting template for the generation of new candidate sentences. It should be said that there may exist no solution to an analogical equation, so that a new candidate is not coined each time.

Figures 4.10 and 4.11 are two examples of sentence generation in Japanese. In the case of Figure 4.11, we generate a sentence which is not valid in meaning for a native speaker.

From the generated candidate sentences point of view, there are some characteristics in new sentences generation process. Figure 4.12 and 4.13 give an example for new sentence generation with some seed sentences and a series of clusters in Chinese. Several important points that characterize the method are listed below:

- One seed sentence may produce different candidate sentences according to different clusters, because different clusters illustrate different linguistic features. For instance, the seed sentence 食物很不错。 'Food is very good.' in Figure 4.12 produced 7 different candidate sentences according to all clusters given in Figure 4.12.
- One seed sentence may produce different candidate sentences even for the same cluster as different sentence pairs (templates) are used. Figure 4.13 illustrates this situation. For instance, the seed sentence 这个女孩长得美。 'The girl looks beautiful.' yielded two different candidate sentences using a clusters (identifier '3') which is examplified by '非常感谢 : 非常感谢作者' and '希望尽快解决 : 希望作者尽快解决';

(Seed sentence)
今日は本当に楽しかったです。
‘It was really fun today.’
⇓

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

⇓
(Generated sentence)
今日はとても楽しかったです。
‘It was very fun today’
…

Figure 4.10: An example of sentence generation result (valid sentence).

(Seed sentence)
本当にこんなのでいいのか
‘Is this really all right’
⇓

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

⇓
(Generated sentence)
*とてもこんなのでいいのか
*‘*Is this very all right’*
…

Figure 4.11: An example of sentence generation result (invalid sentence).

- Different seed sentences may produce different candidate sentences according to the same cluster depending on the direction of the rewriting model, from left to right or right to left. For instance, the seed sentences 经典电影 ‘classic movie’ and 食物很不错。 ‘Food is very good.’ produced four different candidate sentences according to a cluster (identifier ‘2’) which represents the exchange of “经典” ‘classic’ with “很不错” ‘very good’ in one direction and the exchange of “很不错” ‘very good’ with “经典” ‘classic’ in the other direction.

- Different seed sentences may produce the same candidate sentence when passed to different clusters. For instance, the seed sentences 经典电影 ‘classic movie’ and 糟糕电影 ‘bad movie’ produced the same sentence 电影很不错 ‘the movie is very good’ when passed to a cluster (identifier ‘2’) which model is to change “经典” into “很不错” and a cluster (identifier ‘4’) which models the exchange of “糟糕” ‘bad’ to “很不错” ‘very good’.

Seed short sentences	Chinese clusters			Cluster identifier
必须感谢 经典电影 这个女孩长得美。 糟糕电影 食物很不错。	画面很美	:	画面很不错	
	挺美的	:	挺不错的	
	故事很美	:	故事很不错	
	真美	:	真不错	
	美	:	不错	
	美图	:	图不错	
	....		....	1
	经典啊	:	很不错啊	
	经典	:	很不错	
	经典故事	:	故事很不错	
	喜欢经典	:	很不错喜欢	
	经典游戏	:	游戏很不错	
	....		....	2
	非常感谢	:	非常感谢作者	
	希望尽快解决	:	希望作者尽快解决	
	....		....	3
	糟糕啊	:	很不错啊	
	糟糕游戏	:	游戏很不错	
	....		....	4

FIGURE 4.12: Examples of some seed sentences and a series of clusters in Chinese used for new sentence generation.

Seed short sentences		Newly generated sentences	Cluster identifier	Freq.
必须感谢	:	必须感谢作者	3	1
必须感谢	:	作者必须感谢	3	1
经典电影	:	很不错电影	2	4
经典电影	:	电影很不错	2	3
这个女孩长得美。	:	这个女孩长得不错。	1	4
这个女孩长得美。	:	这个女孩长得美作者。	3	1
这个女孩长得美。	:	作者这个女孩长得美。	3	1
糟糕电影	:	很不错电影	4	4
糟糕电影	:	电影很不错	4	3
食物很不错。	:	食物经典。	2	2
食物很不错。	:	食物。经典	2	1
食物很不错。	:	食物很美。	1	5
食物很不错。	:	美食物很。	1	1
食物很不错。	:	食物很不错作者。	3	1
食物很不错。	:	作者食物很不错。	3	1
食物很不错。	:	食物糟糕。	4	1

FIGURE 4.13: The result of newly generated sentences according to the seed sentences and clusters in Table 4.12. The frequencies shows the times a candidate sentence has been generated using all possible clusters.

During the generation of candidate sentences, many invalid (e.g., “食物。经典”) and grammatically incorrect (e.g., “这个女孩长得美作者” and “とてもこんなのていいのか”) sentences are produced. To filter out these sentences and keep only well-formed sentences (e.g., “这个女孩长得不错。”), a filtering step is needed so as to ensure fluency of expression and adequacy of meaning.

4.2.2 Filtering Techniques for Quasi-parallel Corpus Construction

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

4.2.2.1 BLEU-based Filtering Method

BLEU is the main evaluation metric for automatic MT Papineni et al. (2002). It compares a candidate sentence output from an MT system to possibly refer sentences. The formula of BLEU we use is as follows:

$$\text{BLEU} = BP \times \sqrt[n]{\prod_{n=1}^{N} p_n} \tag{4.4}$$

p_n stands for modified n-gram precision. It is the core of the calculation of BLEU. p_n calculates the precision from 1-gram to 4-gram. Different from the normal N-gram precision, modified N-gram precision counts N-grams in the references and clips the count of the same number of N-grams in the candidate sentence to give a lower score to repeated words or phrases. The geometric average of the p_n is computed as a global score. In addition, in order to lessen the advantage given to short candidates by this global score, it is multiplied by a brevity penalty (BP) depending on the length of the candidate and reference sentences.

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

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

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

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

$$\text{R-weight}(f_i) = \frac{\sum_{p \in \hat{T} \cap \hat{f}_i} (-\text{log}c(p) \times |p|)}{\sum_{p \in \hat{T}} (-\text{log}c(p) \times |p|)} \times \frac{\left|\hat{T} \cap \hat{f}_i\right|}{\left|\hat{T}\right|} \times \frac{\left|\hat{T} \cap \hat{f}_i\right|}{\left|\hat{f}_i\right|} \tag{4.5}$$

f_i is a line in the reference corpus.

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

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

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

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

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

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

4.2.2.2 N-sequence Filtering Method

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

We thoroughly test several values of N and tolerance to assess the quality of the sentences kept. In English, the beginning and the end of sentences are well defined by the use of capital letters and the full stop. So as to reproduce similar conditions in Chinese and Japanese, we introduced begin/end markers to make sure that at least the beginning and the end of a sentence is correct. Since the experiments are time consuming, we developed a method which makes use of the *shortest absent substring* to output all the filtering results we expect at the same time so as to reduce the overall experiment time. The algorithm is based on the computation of *shortest absent substrings* computed on a representation of the reference corpus into a suffix array (Nagao and Mori, 1994), (Yamamoto and Church, 2001), (Kärkkäinen and Sanders, 2003).

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

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

4.2.2.3 Differences between the BLEU-based Filtering Method and the N-sequence Filtering Method

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

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

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

The BLEU-based method and the N-sequence filtering method use different reference corpora in each language in the filtering steps. Especially for the N-sequence filtering method, we just use the additional monolingual corpus which is not related to seed sentences. It is not an “ad hoc” method because the comparison with the reference corpus is not based on the seed sentences, but based on all additional reference sentences; many sequences of words do not exist in the seed sentences.

4.2.3 Deduction of Translation Relations

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

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

4.3 Experiments for Quasi-parallel Sentence Construction

4.3.1 Data Preparation

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

To build our baseline SMT system, 500 and 1,000 sentence pairs from JEC Basic Sentence Data[5] were extracted as tuning and testing data. The rest of the 3,804 pairs of sentences were combined with the subtitle corpus and constitute the training data with 110,114 sentence pairs. Table 4.13 shows the statistics on the data preparation.

To construct a quasi-parallel corpus, we prepared unaligned unrelated monolingual sentences in each language to construct analogical clusters (Table 4.2). Monolingual resources are collected mainly from the following website: “douban”[6], “Yahoo China”[7],

[5] JEC Basic Sentence Data: `http://nlp.ist.i.kyoto-u.ac.jp` by Kurohashi-Kawahara Lab., Kyoto University. Released in 2011.

[6] douban: `http://www.douban.com`

[7] Yahoo China: `http://cn.yahoo.com` Closed in 2013.

Table 4.1: Statistics on the Chinese–Japanese corpus used for the training, tuning, and test sets in baseline system. The tuning and testing sets are the same in all SMT experiments.

	Baseline	Chinese	Japanese
train	sentences	110,114	110,114
	words	637,036	721,850
	mean ± std.dev.	5.94 ± 2.60	6.69 ± 2.94
tune	sentences	500	500
	words	3,582	5,042
	mean ± std.dev.	7.15 ± 2.86	10.12 ± 3.39
test	sentences	1,000	1,000
	words	7,285	10,126
	mean ± std.dev.	7.28 ± 2.87	10.15 ± 3.30

Table 4.2: Statistics on the unaligned Chinese and Japanese monolingual short sentences for construction of analogical clusters.

	# of different sentences (cleaned)	size of sentences in characters (mean ± std.dev.)	total characters	total words
Chinese	70,000	10.29 ± 6.21	775,530	525,462
Japanese	70,000	15.06 ± 6.34	1,139,588	765,085

and “Yahoo China News”[8] for Chinese, and “Yahoo! Japan”[9], “Rakuten Japan”[10] and “The Mainichi Japan”[11] for Japanese.

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

[8] Yahoo China News: `http://news.cn.yahoo.com` Closed in 2013.

[9] Yahoo Japan: `http://www.yahoo.co.jp/`

[10] Rakuten Japan: `http://www.rakuten.co.jp/`

[11] The Mainichi Japan: `http://www.mainichi.co.jp/`

4.3.2 Experimental Setting

The segmentation toolkits that we use in all experiments are Urheen for Chinese (zh) and Mecab for Japanese (ja)[12]. We perform all SMT experiments using the standard GIZA++/MOSES pipeline (Koehn et al., 2007) with the default options. Tuning was performed by minimum error rate training (Och, 2003) using 500 tuning sentence pairs. We trained 5-gram language models on the target part of the training data using the SRILM toolkit (Stolcke et al., 2002).

4.3.3 Cluster Construction and New Sentence Generation

Table 4.3 shows the details of the monolingual data and seed sentences we used and the results of clusters construction and new sentences generation. About 14,578 corresponding clusters were extracted ($\text{Sim}_{C_{zh}-C_{ja}} \geqslant 0.3$) by the steps described in Section 4.1. We checked the quality of the newly generated sentences manually. More than half of the generated sentences were found to be grammatically invalid. This is indicated in the last row of Table 4.3, where Q stands for the grammatical quality as evaluated by extracting 1,000 sentences randomly and checking them manually.

Table 4.3: Results of clustering and generation of new sentences.

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

4.3.4 Filtering and Quasi-parallel Corpus Construction for SMT System

We performed BLEU-based filtering experiments with the same candidate sentences as those described in Table 4.3. We grouped the seed sentences by similarity using the Dice coefficient between sets of words. We extracted all 1-grams to 4-grams in each seed group to weight the references, and only selected 100 reference sentences with highest weight to build reference sets. Each reference set corresponds to a seed group.

After having obtained seed groups (Table 4.4) and corresponding reference sets for each seed group, for each candidate sentence, its seed sentence is identified among the possible

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

seed groups and the BLEU scores of the candidate sentence against the corresponding reference sets are computed. In our experiments, we set several thresholds to check the filtering results (Table 4.5).

Table 4.4: Statistics of grouping seed sentences.

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

Table 4.5: Filtering results by using BLEU-based filtering method.

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

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

Finally, we checked and found that there are about 500,000 unique Japanese filtered sentences after filtering by the BLEU method (threshold > 1 BLEU%). Thus, we only kept the 500,000 unique Japanese filtered sentences with their corresponding seed sentences in higher BLEU scores. The same size of filtered sentences in Chinese with higher BLEU scores are also extracted. Deducing translation relationships allowed us to construct a quasi-parallel corpus of 353,729 sentence pairs. We added the new corpus into the baseline and evaluated it. Table 4.6 shows the results.

The BLEU based filtering method increases the baseline system by only 0.8 BLEU point. We reduce the size of the reference corpora and only used grouped seed sentences to

Table 4.6: Comparison of the baseline SMT system and an SMT system with additional quasi-parallel data output by BLEU-based filtering. The figure in bold characters (13.89) shows a significant improvement with a p-value < 0.01.

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

weight the reference sentences. It was observed that most of the BLEU scores obtained in the filtering step are around 1, which is close to the improvement obtained in SMT.

To determine the most appropriate N which can keep the largest number of well-formed sentences to be added to the training corpus, we performed a series of filtering experiments using the N-sequence method with different values of N and tolerance. Table 4.7 shows the results for N equal to 4 to 9 and tolerance equal to 0 and 1 in Chinese and Japanese.

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

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

Table 4.7: Filtering results by using the N-sequence filtering method in different Ns and tolerances.

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

Using these results, we selected 4 sets of filtered sentences in the two languages with high quality obtained using a tolerance of 0. We also selected 3 similar sets with a tolerance of 1. This makes 7 quasi-parallel corpus in total. In each corpus, N (ja) equals N (zh)+1 so as to make the number of filtered sentences comparable. Table 4.8 describes the quasi-parallel corpora constructed.

Table 4.8: Construction results of a quasi-parallel corpus by using N-sequence filtering and the evaluation results for Chinese–Japanese baseline system and baseline + additional quasi-parallel systems. The figures in bold characters show a significant improvement with a p-value < 0.01.

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

For the 7 quasi-parallel corpora, we added each of them as additional data to our initial Chinese–Japanese training data to perform Chinese-to-Japanese SMT experiments. We recomputed translation tables (training), tuned the system, performed translation of the same test set and calculated the BLEU scores. Table 4.8 shows the results for each of the SMT systems. All the BLEU scores of the SMT systems with additional data are 1 to 6 points higher than the baseline system. The highest score is obtained when N (zh) = 6 and N (ja) = 7 with a tolerance of 0. Quasi-parallel corpora with a tolerance of 0 contain less noise, and the BLEU scores increase when the size of additional data becomes larger. Therefore, even if the size of quasi-parallel corpora adding data with a tolerance of 1 is much larger than data with a tolerance of 0, because of the noise, it cannot improve the translation results effectively.

Table 4.8 also shows the BLEU scores obtained on the tuning set when the parameters are optimized on this same tuning set for each SMT system. We vary the filtering parameters. The best system obtained by considering the scores on the tuning set is obtained for N (zh) = 6 and N (ja) = 7 with a tolerance of 0. We then evaluate this best system with these parameters on a test set. We verify that the score obtained on the test set with these parameters is the best score by evaluating all other systems on the same test set. We confirm that the best configuration obtained by tuning leads to the best score on the test set (see Table 4.8).

We also build an SMT system based on the quasi-parallel corpora obtained by combining the BLEU and the N-sequence filtering methods. This arrangement yielded even greater improvement (Table 4.8): a more lenient filtering method (more sentences remain after filtering) is boosting the performance of the more drastic filtering method (less sentences kept). This is shown by a relatively higher than expected increase in translation accuracy as measured by BLEU, as 7.25 > 0.79 + 6.17 = 6.96.

4.3.5 Analysis of the Results and Discussion

We investigated the N (source length) × M (target length) distribution in phrase tables (used during testing) generated from the initial parallel corpus and the inflated training corpus by adding the constructed quasi-parallel data (filtered with N (zh) = 6 and N (ja) =7, Tolerance is 0). In Table 4.9 and Table 4.10, the statistics (zh→ja) show that the total number of phrase pairs used by adding additional quasi-parallel corpus is larger than when using only the initial parallel corpus as training data, especially for 1-4 grams in both languages. If we compare the number of entries, the number of phrase pairs (in Table 4.10) on the diagonal got a significant increase in the number of phrase pairs of similar length. Considering the correspondence between lengths in Chinese–Japanese translation, the increase in phrase pairs with different lengths (like 1 (zh) × 2 (ja), 2 (zh) × 3 (ja) and 3 (zh) × 4 (ja)) is felicitous. This means that adding the additional quasi-parallel corpus for inflating the training corpus for SMT allowed us to produce much more numerous potentially useful alignments.

Table 4.9: Distribution of phrase pairs used during testing in Chinese-to-Japanese SMT experiment (baseline).

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

Table 4.11 illustrates the fact that new translation candidates have been added for an existing phrase, and that new phrase pairs have also been added. The fact that these additional phrases are reasonable is indicated by the improvements in BLEU scores. Table 4.12 illustrates changes in lexical weights and translation probabilities for the same Chinese phrase. More accurate phrase alignments may be extracted by adding additional quasi-parallel corpus. We also believe that we improved its features by adding quasi-parallel data.

Table 4.10: Distribution of phrase pairs used during testing in Chinese-to-Japanese SMT experiment (baseline + quasi-parallel). The bold numbers show the increased numbers of N (Chinese) × M (Japanese)-grams (less than 4-gram) in the phrase table, and the total number of N (Chinese) × M (Japanese)-grams, which increased compared with the baseline system.

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

Table 4.11: Samples of phrase alignments in zh→ja phrase table. Same Chinese phrase and corresponding Japanese phrases in baseline phrase table and baseline + quasi-parallel phrase table.

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

4.4 Experiments on Technical Translation

4.4.1 Experiments and Results

To assess the contribution of our proposed methods on technical machine translation, we propose to compare two SMT systems. The first one is constructed using the initial given ASPEC-JC parallel corpus. This is the baseline. The second one adds the additional

Table 4.12: Samples of the same phrase alignments in Chinese and Japanese with different translation probabilities (TP) and lexical weights (LW) in baseline phrase table and baseline + quasi-parallel phrase table.

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

quasi-parallel corpus obtained using analogical associations based on the constructed clusters and a part of ASPEC corpus which less than 30 characters in length.

Baseline: The statistics of the data used in the experiments are given in Table 4.13 (left). The training corpus consists of 672,315 sentences of initial Chinese–Japanese parallel corpus. The tuning set is 2,090 sentences from the ASPEC-JC.dev corpus, and 2,107 sentences also from the ASPEC-JC.test corpus were used for testing. We perform all experiments using the standard GIZA++/MOSES pipeline (Koehn et al., 2007).

Table 4.13: Statistics on the Chinese–Japanese corpus used for the training, tuning, and test sets in baseline (left) and baseline + quasi-parallel data (right). The tuning and testing sets are the same in both experiments. Segmentation tools: Urheen for Chinese and Mecab for Japanese.

	Baseline	Chinese	Japanese
train	sentences	672,315	672,315
	words	18,847,514	23,480,703
	mean ± std.dev.	28.12 ± 15.20	35.05 ± 18.88

+ Quasi-parallel	Chinese	Japanese
sentences	**708,132**	**708,132**
words	19,212,187	24,512,079
mean ± std.dev.	27.13 ± 14.19	34.23 ± 17.22

	Both experiments	Chinese	Japanese
tune	sentences	2,090	2,090
	words	60,458	73,177
	mean ± std.dev.	28.93 ± 15.86	35.01 ± 18.87
test	sentences	2,107	2,107
	words	59,594	72,027
	mean ± std.dev.	28.28 ± 14.55	34.18 ± 17.43

Experiments on New Sentence Generation and Filtering by N-sequences: For the generation of new sentences, we make use of the clusters obtained from the experimental results shown in Table 4.3 as *rewriting models*. Different from the extraction of corresponding clusters, we make use of additional resource for computing the similarity between Chinese and Japanese clusters. They are word-to-word alignments based on ASPEC-JC data using Anymalign[13]. We keep 72,610 word-to-word correspondences (use option -N) obtained with Anymalign in 1 hour after filtering on both translation probabilities with a threshold of 0.3, the quality of these word-to-word correspondences is about 96%. We set different thresholds for $Sim_{C_{\mathrm{zh}}-C_{\mathrm{ja}}}$ and check the correspondence between these extracted clusters by sampling. Where the $Sim_{C_{\mathrm{zh}}-C_{\mathrm{ja}}}$ threshold is set to 0.3, the acceptability of the correspondence between the extracted clusters reaches 78%. About 15,710 corresponding clusters were extracted ($Sim_{C_{\mathrm{zh}}-C_{\mathrm{ja}}} \geq 0.3$) by steps.

The seed sentences as input data for new sentences generation are the unique Chinese and Japanese short sentences from the 103,629 ASPEC-JC parallel sentences (less than 30 characters). In this experiment, we generated new sentences with each pair of sentences in clusters for Chinese and Japanese respectively. Table 4.14 gives the statistics for new sentence generation.

Table 4.14: Statistics on new sentence generation in Chinese and Japanese. Q is the quality of the new candidate sentences or new valid sentences after filtering.

		Chinese		Japanese	
Initial data	# of seed sentences	99,538		97,152	
	# of clusters	23,182		21,975	
New sentence generation	# of candidate sentences	105,038,200 Q= 29%		80,183,424 Q= 40%	
Quality assessment (filtered)	# of new valid sentences	unique	seed–new–#	unique	seed–new–#
		33,141	67,099	40,234	84,533
		Q= 96%		Q= 96%	

To filter out invalid and grammatically incorrect sentences and keep only well-formed sentences, we eliminate any sentence that contains an N-sequence of given length (N = 6 for Chinese and N = 7 for Japanese, tolerance = 0) of a given length unseen in the reference corpus, because the best quality was obtained for the values N = 6 for Chinese and N = 7 (tolerance = 0) for Japanese in previous experiments. We use the size of reference corpus with 1,700,000 monolingual data for both Chinese and Japanese. Quality assessment was performed by extracting a sample of 1,000 sentences randomly and checking manually by native speakers. The grammatical quality was at least 96%. This means that 96% of the Chinese and Japanese sentences may be considered as grammatically correct. For new valid sentences, we remember their corresponding seed sentences and the cluster they were generated from.

[13] `http://anymalign.limsi.fr`

We deduce translation relations based on the initial parallel corpus and corresponding clusters between Chinese and Japanese. Table 4.15 gives the statistics on the quasi-parallel deducing obtained. Among the 35,817 unique Chinese–Japanese quasi-parallel sentences obtained, about 74% were found to be exact translations by manual check on a sampling of 1,000 pairs of sentences.

Table 4.15: Statistics on the quasi-parallel corpus deducing.

Chinese	Japanese	Chinese–Japanese		
seed–new–#	seed–new–#	Initial parallel corpus	Corresponding clusters	Quasi-parallel corpus
67,099	84,533	103,629	15,710	35,817

Adding Additional Quasi-parallel Corpus: The statistics of the data used in this second setting are given in Table 4.13 (right). The training corpus is made of 708,132 (672,315 + 35,817) lines of sentences, i.e., the combination of the initial Chinese–Japanese parallel corpus used in the baseline and the quasi-parallel corpus.

Experimental Results: Tables 4.16 and 4.17 give the evaluation results. We use the standard metrics BLEU (Papineni et al., 2002), NIST (Doddington, 2002), WER (Nießen et al., 2000), TER (Snover et al., 2006) and RIBES (Isozaki et al., 2010). As Table 4.16 shows, significant improvement over the baseline is obtained by adding the quasi-parallel generated data based on the Moses version 1.0, and Table 4.17 show that a slight improvement over the baseline is obtained by adding the quasi-parallel generated data based on the Moses version 2.1.1.

4.4.2 Influence of Segmentation on Translation Results

We also use KyTea[14] to segment Chinese and Japanese. Table 4.18 and Table 4.19 show the evaluation results by using KyTea as the segmentation tools based on standard GIZA++/MOSES (different version in 1.0 and 2.1.1) pipeline. As the evaluation scores (BLEU and RIBES) shown in Table 4.16, Table 4.17, Table 4.18 and Table 4.19:

- We obtained more increase based on Moses version 1.0 than Moses version 2.1.1 by using Urheen/Mecab or kyTea for Chinese and Japanese as the segmentation tools;
- But, based on Moses version 2.1.1 we obtained higher BLEU and RIBES scores than with Moses version 1.0 by using two different segmentation tools;

[14] `http://www.phontron.com/kytea/index-ja.html`

- Based on the same Moses version, most of the BLEU and RIBES scores are higher by using Urheen and Mecab as the segmentation tools for Chinese and Japanese than using KyTea (except Japanese-to-Chinese by using KyTea based on Moses version 2.1.1).

Table 4.16: Evaluation results for Chinese–Japanese translation across two SMT systems (baseline and baseline + additional quasi-parallel data), Moses version: 1.0, segmentation tools: Urheen and Mecab.

		BLEU	NIST	WER	TER	RIBES
zh-ja	baseline	29.10	7.5677	0.5352	0.5478	0.7801
	+ additional training data	**32.03**	**7.9741**	**0.5069**	**0.5172**	**0.7906**
ja-zh	baseline	22.98	7.0103	0.5481	0.5711	0.7893
	+ additional training data	**24.87**	**7.3208**	**0.5273**	**0.5482**	**0.8013**

Table 4.17: Same as Table 4.16 with Moses version: 2.1.1, segmentation tools: Urheen and Mecab.

		BLEU	NIST	WER	TER	RIBES
zh-ja	baseline	33.41	8.1537	0.4967	0.5061	0.7956
	+ additional training data	**33.68**	**8.1820**	**0.4955**	**0.5039**	**0.7964**
ja-zh	baseline	25.53	7.3885	0.5227	0.5427	0.8053
	+ additional training data	**25.80**	**7.4571**	**0.5176**	**0.5378**	**0.8060**

Table 4.18: Same as Table 4.16 with Moses version:1.0, segmentation tools: KyTea.

		BLEU	NIST	WER	TER	RIBES
zh-ja	baseline	28.35	7.3123	0.5667	0.5741	0.7610
	+ additional training data	**28.87**	**7.4637**	**0.5566**	**0.5615**	**0.7739**
ja-zh	baseline	22.83	6.9533	0.5633	0.5853	0.7807
	+ additional training data	**23.18**	**7.0402**	**0.5547**	**0.5778**	**0.7865**

Table 4.19: Same as Table 4.16 with Moses version: 2.1.1, segmentation tools: KyTea.

		BLEU	NIST	WER	TER	RIBES
zh-ja	baseline	33.27	7.9579	0.5249	0.5272	0.7820
	+ additional training data	**33.56**	**8.0229**	**0.5178**	**0.5206**	**0.7849**
ja-zh	baseline	26.25	7.4931	0.5197	0.5398	0.8085
	+ additional training data	**26.52**	**7.5523**	**0.5128**	**0.5335**	**0.8105**

4.5 Combination of Proposed Techniques in Technical SMT

This section describes the experiments and the results that make use of the proposed techniques and results described in previous chapters and sections. We aim at assessing the translation accuracy in technical statistical machine translation (SMT) when combining the proposed methods and results. These methods and works are: construction of a Chinese–Japanese lexicon combining several automatic techniques based on several freely available resources (Chapter 2); re-tokenization of Chinese–Japanese training corpus in SMT with extracted bilingual terms using the C-value and sampling-based alignment methods, as well as kanji-hanzi conversion method (Chapter 3); construction of a Chinese–Japanese quasi-parallel corpus using analogical associations for inflating an existing training corpus to train the translation model (Chapter 4).

4.5.1 Data Used and Overview of the Experiments

The experimental data used in this section are the same as in Section 4.4.1 given in Table 4.13, but the segmentation tools are different. The segmentation tools used here are Stanford (ctb) for Chinese and Juman for Japanese.

Table 4.20: Statistics on the Chinese–Japanese corpus used for the training, tuning, and test sets in baseline (upper left) and baseline + Quasi-parallel data (upper right). The tuning and testing sets are the same in both experiments. Segmentation tools: Stanford (ctb) for Chinese and Juman for Japanese.

	Baseline	Chinese	Japanese
train	sentences	672,315	672,315
	words	18,208,123	22,322,141
	mean ± std.dev.	27.16 ± 14.59	33.32 ± 17.68

+ Quasi-parallel	Chinese	Japanese
sentences	**708,132**	**708,132**
words	18,516,044	22,679,058
mean ± std.dev.	26.66 ± 14.73	32.75 ± 17.87

	Both experiments	Chinese	Japanese
tune	sentences	2,090	2,090
	words	59,279	70,250
	mean ± std.dev.	28.36 ± 15.58	33.61 ± 17.95
test	sentences	2,107	2,107
	words	58,318	69,246
	mean ± std.dev.	27.68 ± 14.13	32.86 ± 16.71

The procedure of our experiments is given as follows:

- Extract monolingual terms using the C-value method from Chinese and Japanese parts of the training corpus, respectively;
- Re-tokenize the Chinese and Japanese training corpus respectively with the extracted monolingual terms;
- Extract Chinese–Japanese bilingual terms combining a sampling-based alignment method and kanji-hanzi conversion method;
- Re-tokenize the inflated Chinese–Japanese training corpus (baseline + Quasi-parallel corpus) with the extracted bilingual terms;

- Perform a baseline SMT experiment with the data given in Table 4.20;
- Perform a SMT experiment with the re-tokenized inflated Chinese–Japanese training data, additionally, we also make use of the lexicon constructed in Chapter 2 in SMT decoding process.

4.5.2 Extraction of Monolingual and Bilingual Terms

In this section, we firstly extract monolingual terms using the C-value method from 672,315 Chinese and Japanese parts of training corpus respectively. 568,974 monolingual terms are extracted from Chinese part and 510,792 terms are extracted for Japanese part. For keeping the balance between monolingual term extraction in different languages, we re-tokenize the training corpus with the same number of Chinese and Japanese monolingual multi-word terms respectively. These terms are the first 510,000 monolingual multi-word terms with the highest C-values in each language.

We then extract bilingual terms from the re-tokenized Chinese–Japanese training corpus. We obtained 75,350 bilingual multi-word terms using sampling-based alignment method and filtering by setting translation probabilities in both directions with 0.6. Because P = 0.6 as the threshold for bilingual multi-word extraction allowed us obtained the best system and translation accuracy in Section 3.2.4.2 (see Table 3.3), Section 3.3.2 (see Table 3.14 and Table 3.15) and Section 3.3.5 (see Table 3.23). We further filter these bilingual multi-word terms by considering the ratio of the lengths in words between Chinese and Japanese, and the components of the terms in Japanese (see Section 3.3.1). We finally obtained 63,687 Chinese–Japanese bilingual multi-word to multi-word terms. The percentage of the good bilingual multi-word term matches is 85%.

Following the proposed method that extracts bilingual terms using kanji-hanzi conversion, we extracted 27,233 bilingual terms based on the sampling-based alignment result. The percentage of the good bilingual term matches is 100%. We also take single-word to multi-word term extraction into consideration. We extract 7,401 single-word to multi-word terms after several filtering constraints. The percentage of the good bilingual single-word to multi-word term matches is 80%.

Finally, we re-tokenize the Chinese–Japanese training corpus with these extracted bilingual terms combining several proposed methods. The total number of the bilingual terms used in re-tokenization of Chinese–Japanese training corpus is 63,687 + 27,233 + 7,401 = 98,321 (unique terms: 83,593, because there is an intersection between multi-word to multi-word term extraction and kanji-hanzi conversion based bilingual term extraction; another intersection between kanji-hanzi conversion based bilingual term extraction and single-word to multi-word term extraction). The percentage of the good bilingual term matches (total) is 90%.

4.5.3 SMT Experiments

To assess the contribution of our proposed methods described in previous sections and chapters on technical statistical machine translation, we propose to compare two SMT systems. The first one is constructed using the initial given ASPEC-JC parallel corpus. This is the baseline. The second one adds the additional quasi-parallel corpus obtained using analogical associations into the baseline training data, then re-tokenize the inflated training corpus with the extracted bilingual terms.

Baseline: The statistics of the training data used in the experiments are given in Table 4.20 (upper left). The training corpus consists of 672,315 lines of sentences of initial ASPEC Chinese–Japanese parallel corpus. The tuning set is 2,090 sentences from the ASPEC-JC.dev corpus, and 2,107 sentences also from the ASPEC-JC.test corpus were used for testing. We perform the experiment using the standard GIZA++/MOSES (2.1.1) pipeline (Koehn et al., 2007). The BLEU score obtained is 34.45.

((Baseline + Quasi-parallel) → Re-tokenization) + Lexicon: The statistics of the training data used in the experiments are given in Table 4.20 (upper right). The training corpus consists of 708,132 lines sentences of initial Chinese–Japanese training corpus (672,315 lines) + Quasi-parallel corpus (35,817 lines). The number of bilingual terms used in re-tokenization of the training corpus is 83,593. The tuning and test sets are the same as used in baseline system. The same as the baseline system, we perform the experiment using the standard GIZA++/MOSES (2.1.1) pipeline (Koehn et al., 2007). In this experiment, we also make use of the lexicon constructed in Chapter 2 in decoding process. The BLEU score obtained in this system is 36.24. Compare with the baseline system (BLEU = 34.45), we obtained 1.8 BLEU point with p-value < 0.01. We can conclude that the improvement of the translation accuracy is statistically significant in comparison with the baseline system.

Lexicon used in decoding: For SMT decoding process, sometimes we want to bring some external knowledge to the decoder. For instance, we have a Chinese–Japanese lexicon for translation of certain words or phrases. We would like to make use of these translations in decoding without changing the translation model. We firstly change the format of the test set by adding the translation of words or phrases appeared in the lexicon. We can also provide probabilities for these translation candidates. Multiple translation existing in the lexicon can be separated by two bars (||) with multiple probabilities. In decoding process, this gives the decoder a choice to use either translations from the lexicon or from the translation model. The specified translation based on lexicon will compute with all the phrase table choices and ultimately the choice will be made by the language model.

Analysis of the Results and Discussion: We investigate the N (Chinese) × M (Japanese)-gram distribution in the phrase tables potentially used in translation. In

将/高压/<n translation=“二酸化炭素” prob=“1.0”>二氧化碳</n>/气体/吹入/污水/污泥/,/通过/在/减压/时/破坏/污泥......

现在/的/输出地/限定/在/以/<n translation=“ナイジェリア” prob=“1.0”>尼日利亚</n>/等/非洲/各/国/为/首位......

并且/<n translation=“二酸化炭素” prob=“1.0”>二氧化碳</n>/加热泵/已经/在/轿车/空调/等/领域/被/实用化/。

在/缺乏/视/触觉/信息/的/环境/下/,/人类/还/能/<n translation=“イージー||易い” prob=“0.5||0.5”>轻而易举</n>/地......

Figure 4.14: Examples of XML-specified translation in the test set (Chinese).

Tables 4.21 and 4.22, the statistics (Chinese-to-Japanese) show that the total number of potentially useful phrase pairs used in translation based on re-tokenized inflated corpus + lexicon is larger than that used in the baseline system. We compare the number of entries, the number of phrase pairs have a significant increase compare with the baseline system. This simply shows that using additional constructed lexicon, adding a quasi-parallel corpus and re-tokenization of training corpus with multi-word terms extracted from the initial training set are potentially useful for the translation of the test set, which was precisely the goal of our work on improving technical statistical machine translation combine using several proposed methods at various levels of granularity (characters, words, terms and short sentences).

Table 4.21: Distribution of N (Chinese) × M (Japanese)-gram entries in the phrase table potentially used in testing in system ((Baseline + Quasi-parallel) → Re-tokenization) + Lexicon. The bold face numbers showing the increased N (Chinese) × M (Japanese)-grams in the phrase table, and the total number of N (Chinese) × M (Japanese)-grams, which increased compared with the baseline system.

		Target = Japanese							
		1-gram	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	total
Source = Chinese	1-gram	**146446**	**533230**	**533292**	**326908**	169989	84820	43291	1837976
	2-gram	**96198**	**304434**	**367151**	**243041**	125086	58958	27628	1222496
	3-gram	**14184**	**39688**	**69469**	**66077**	40450	19833	9223	258924
	4-gram	**1233**	**2965**	**6568**	**11004**	10771	6755	3217	42513
	5-gram	189	336	651	1260	2525	3141	2631	10733
	6-gram	52	79	115	191	392	1147	1587	3563
	7-gram	11	12	17	41	49	138	403	671
	total	258313	880744	977263	648522	349262	174792	87980	**3376876**

4.6 Summary of This Chapter

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

TABLE 4.22: Distribution of N (Chinese) × M (Japanese)-gram entries in the phrase table potentially used in testing in the baseline system.

		Target = Japanese							
		1-gram	2-gram	3-gram	4-gram	5-gram	6-gram	7-gram	total
Source = Chinese	1-gram	157840	507416	478267	281186	141623	69797	35716	1671845
	2-gram	93941	299749	356214	222944	109483	49557	22739	1154627
	3-gram	13271	38298	69565	64717	37542	17447	7796	248636
	4-gram	1112	2766	6396	11114	10678	6345	2863	41274
	5-gram	178	308	638	1292	2588	3078	2435	10517
	6-gram	46	78	113	176	419	1160	1543	3535
	7-gram	10	12	21	32	47	161	433	716
	total	266398	848627	911214	581461	302380	147545	73525	3131150

licences[15] [16] with the concern of avoiding any copyright problem. We produced all possible analogical clusters as rewriting models for generating new sentences, then filter newly over-generated sentences by a BLEU-based method and N-sequence method.

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

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

The same proposed method as the one used here is also used for constructing a Chinese–Japanese quasi-parallel corpus based on a scientific corpus (ASPEC-JC). We produced analogical clusters as rewriting models to generate new sentences. A quasi-parallel corpus is constructed based on the short sentence pairs in ASPEC corpus with less than 30 characters. We filter newly over-generated sentences by the N-sequences filtering method. The grammatical quality of the valid new sentences is at least 96%. We

[15] Subscence.com: `https://subscene.com/site/legal-information`

[16] Opensubtitles.org: `http://www.opensubtitles.org/ja/disclaimer`

then assess translation relations between newly generated short sentences across both languages, relying on the similarity between the clusters across languages. We automatically obtained 35,817 Chinese–Japanese sentence pairs, 74% of which were found to be exact translations. In SMT experiments performed on Chinese–Japanese, using the standard GIZA++/MOSES pipeline, by adding our quasi-parallel data, we were able to inflate the training data in a rewarding way. On the same test set, based on different MOSES versions and segmentation tools, all of translation scores significantly or slightly improved over the baseline systems. It should be stressed that the data that allowed us to get such improvement are not so large in quantity and not so good in quality, but we were able to control both quantity and quality so as to consistently improve translation quality.

In our SMT experiments, whatever the experimental data used are more general or in some special technical domain, significant improvements are obtained compared with baseline systems. The experimental results demonstrate the generality of the method.

In this chapter, we also presented experiments on statistical machine translation on technical domains by combining several proposed methods and results described in previous chapters and this chapter. We made use of the lexicon constructed in Chapter 2 as the external knowledge using in decoding process of SMT experiment. This allows us to provide more choice for the translation of test sentences. We extracted monolingual and bilingual terms from an existing baseline training corpus based on the methods described in Chapter 3. We aimed at re-tokenizing the inflated training corpus (obtained in Chapter 4) with these extracted bilingual terms. The combination of these works leaded to statistically significant improvements in translation accuracy which evaluated by BLEU scores. We obtained about 1.8 BLEU point (at p-value < 0.01) improvements in comparison with a baseline system.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The main focus of this dissertation was to exploit several freely available linguist resources to address the scarcity of linguistic data and exploit technical terms for technical translation in specific domains between Chinese and Japanese. The conclusions of the dissertation are as follows.

Chapter 1 described the background and the basic knowledge of the research. It gave an overview of the approach adopted in the thesis and presented the contributions of the thesis.

Chapter 2 addressed the problem of scarcity of bilingual lexica between Chinese and Japanese. Extracting bilingual lexica from Chinese–Japanese parallel or comparable corpora has been proposed in previous works (Rapp, 1999), (Vulić et al., 2011). However, the scarcity of parallel corpora and the parallelism of comparable corpora are still problems. We proposed a method to construct a Chinese–Japanese lexicon by combining several automatic techniques on several freely available resources. Our method elaborates on the classical pivot language technique. With this method the quality lies below 45% of correct entries in our experiments. To improve the quality, we proposed to combine three additional techniques: one time inverse consultation (76% of correct entries); Japanese kanji to Chinese hanzi character conversion (98.5% of correct entries); expansion through a Chinese synonym table (98.5% of correct entries). The three additional methods allowed us to increase the quality of the Chinese–Japanese lexicon from less than 45% to 85% and get 45,386 entries in total. By comparison with a reference dictionary, 83% of word pairs in our lexicon do not appear in a large reference dictionary, the EDR dictionary constructed by NICT (about 300,000 entries). We made use of our kanji-hanzi conversion method through out our work, because there exist a large amount of characters shared with the same meaning in the Chinese and Japanese writing systems. Our results show that they can be safely used as clues to align words or multi-word expressions.

Chapter 3 addressed the problem of the scarcity of digitalized terminological banks between Chinese and Japanese. The identification and translation of terms in patents and scientific texts is of course crucial in technical translation. We proposed a method to improve Chinese–Japanese technical translation of patents and scientific texts by re-tokenizing the training corpus with aligned bilingual multi-word terms. We extracted bilingual multi-word terms from the training corpus. These extracted terms are used in adjusting and balancing the tokenization between Chinese and Japanese technical data. We proposed two experimental protocols to make use of the extracted bilingual terms in Chinese–Japanese statistical machine translation (SMT) experiments so as to select the better one. We obtained a quality of correspondence of 80% in bilingual multi-word term extraction and a significant improvement of 1 BLEU score ($p < 0.01$) in translation accuracy. We combined using the kanji-hanzi conversion method (Chapter 2), and obtained better results in correspondence of bilingual terms (93%) and BLEU with 1.5 BLEU point improvement ($p < 0.01$). We also considered the cases where one side is a single-word term and the other side is a multi-word term without hanzi/kanji constraints. We obtained even better results with 95% in correspondence of terms and 2 BLEU point improvement ($p < 0.01$) in translation accuracy. Our pre-processing on terms has the effect of reducing the problem of different segmentation conventions across languages.

Chapter 4 addressed the problem of scarcity of bilingual corpora between Chinese and Japanese. In SMT, the translation knowledge is acquired from the parallel sentences. Consequently, the quantity and the quality of the translation relations extracted between words or phrases between two languages depend on the quantity and the quality of the parallel sentences. We proposed a method to construct a quasi-parallel corpus by using analogical associations based on large amounts of monolingual data and a small amount of parallel data, so as to improve Chinese–Japanese SMT quality. We generated large amounts of new candidate sentences using analogical associations. We filtered over-generated sentences using two filtering methods: one based on BLEU (used in Chapter 3 for evaluation) and the second one based on N-sequences. We also combined these two filtering methods. The N-sequence method allowed us to keep sentences which may be considered grammatically correct in 99% of the cases. The constructed quasi-parallel corpora were added to the existing training corpus to address the shortage of parallel corpora between Chinese and Japanese. The best result that we obtained is a very significant improvement of 6 BLEU points ($p < 0.01$) over a Chinese–Japanese baseline system. This kind of quasi-parallel sentences used as additional training data in SMT helps in acquiring more potential useful translation knowledge from the inflated training corpus. We also combined all proposed techniques and results described in previous chapters and this chapter. Firstly we made use of the quasi-parallel data obtained in Chapter 4 as additional training data. This quasi-parallel corpus was constructed using analogical associations based on a small number of parallel corpus and amounts of monolingual data. We then re-tokenized this inflated training corpus with bilingual terms extracted from baseline training data based on methods proposed in Chapter 3. Finally, we combined using the lexicon constructed in Chapter 2 in the decoding step to enforce the translation of word in the test set. The translation system based on these data was compared with a baseline system. We obtained a statistically significant improvement of 1.8 BLEU point with p-value less than 0.01.

5.2 Future Work

In this dissertation, we described our proposed approaches and presented the experiments and results that improved the state-of-the-art performance of statistical machine translation in technical domains.

There still remain many challenges and scalability problems in automatically extracting or constructing parallel data for improving technical machine translation, not only for statistical machine translation (SMT) but also for neural machine translation (NMT). We give some possible directions hereafter.

5.2.1 A Large-scale Bilingual Lexicon Construction in Different Domains

Bilingual lexicon extraction is crucial for machine translation or information retrieval. For instance, it can be used for solving the unknown word problem or the drop of words in machine translation; it can also be used as clues for extracting parallel data from parallel or comparable articles. In Chapter 2, we have shown a combination method for Chinese–Japanese lexicon construction using several freely available resources. We also made use of this lexicon in the decoding process of SMT experiment for specifying the translation of words existing in the lexicon to solve the problem of unknown word and increase the coverage in translated words. Our construction method is a novel combination method of using one time inverse consultation, kanji-hanzi conversion and synonyms. It is fully automatic and allowed us obtain more promising word pairs. There exist some possible further directions for future work.

Compared with the EDR dictionary (about 300,000 entries), although our method is fully automatic and there exist 83% word pairs in our lexicon which do not exist in the EDR dictionary, the size (the number of word pairs) of our lexicon is only one sixth in comparison with that of the EDR dictionary. From this point of view, exploiting large-sized Chinese–English and Japanese–English freely available dictionaries for bilingual lexicon construction using the same method should help to increase the number of entries. From the domain point of view, exploiting Chinese–English and Japanese–English dictionaries or lexica in technical domains could also contribute, for instance, the use of 英語日本語電気専門用語辞書 (English–Japanese Electric Terminology Lexicon)[1] and 电气专业名词中英对照表 (Chinese–English Electric Terminology Mapping Table)[2]. From the type of data point of view, our proposed method made use of Chinese–English and Japanese–English *bilingual* lexica to construct a Chinese–Japanese lexicon, but Chinese and Japanese *monolingual* data is sufficient for kanji-hanzi conversion method and synonym method. Chinese and Japanese monolingual data is obviously more easily to access than bilingual data. Of course, it is also possible to extract Chinese–English and Japanese–English lexica from Chinese–English and Japanese–English parallel corpora, and then make use of this kind of lexica to construct a Chinese–Japanese lexicon.

5.2.2 Bilingual Term Extraction by Changing POS Tagging

In our work, we extracted bilingual terms from an existing parallel corpus so as to re-tokenize technical terms in the existing training corpus. The motivation was to address the segmentation discrepancies in using different segmentation tools or standards that may lead to different segmentation results at different levels of granularity, especially for terms in technical machine translation. Different from the previous work, we did not use any additional lexicon or corpus or intermediate languages. We re-tokenized the training corpus for training the translation model with the bilingual terms extracted from the same training corpus. These bilingual terms are extracted based a POS (part-of-speech) tagged corpus using linguistic pattern ((Adjective | Noun)$^{+}$ Noun) and statistical computation.

We investigated the result of the extraction of these terms and found that there remain some promising monolingual or bilingual terms which were not extracted due to the limitation in POS or different segmentation, or different POS standards between Chinese and Japanese. In the POS result of ASPEC-J corpus, there exist 14,026 words made up of katakana in Japanese which were tagged as “未定義語” by Juman, e.g., バッファ ‘puffer’, セル ‘cell’, or キャッピング ‘capping’ etc. There exist 1,863 characters or words made up of kanji in Japanese which were tagged as “未定義語” by Juman, e.g., 圧 ‘press’, 固 ‘firm’ or 噛 ‘biting’ and etc. There also exist some Japanese characters (tokens) made up of kanji which were tagged as “接尾辞” (47 characters) or “接頭辞” (176 characters), e.g., 率 (‘rate’, “接尾辞”), 副 (‘side’, “接頭辞”) or 高 (‘high’, “接頭辞”).

We propose to tag all Japanese words tagged “未定義語” which are made up of katakana with “名詞”, and change the “未定義語”, “接尾辞” and “接頭辞” which made up of

[1] http://www.geocities.jp/ps_dictionary/a.htm

[2] https://wenku.baidu.com/view/f29f281658fafab069dc02fc.html

kanji and can be found in Chinese extracted monolingual terms into “名詞”. This gives us a very high chance to obtain the bilingual terms such as the following ones:

- “未定義語”: 激光♯NN 熔凝♯NN 图案♯NN ↔ レーザ/名詞 固/未定義語 結/名詞 パターン/名詞
  封盖♯NN 位置♯NN ↔ キャッピング/未定義語 位置/名詞
  试验♯NN 电池♯NN ↔ 試験/名詞 セル /未定義語
  气体♯NN 压力♯NN ↔ ガス/名詞 圧/未定義語
  载置部♯NN ↔ 載/未定義語 置/未定義語 部/名詞
  气室♯NN ↔ パッファ/未定義語 室/名詞
  筐体♯NN ↔ 筐/未定義語 体/名詞
  啮合♯NN ↔ 噛/未定義語 合/名詞
- “接頭辞”: 副反应♯NN ↔ 副/接頭辞 反応/名詞
  高纯度♯NN ↔ 高/接頭辞 純度/名詞
  高精度♯JJ 技术♯NN ↔ 高/接頭辞 精度/名詞 技法/名詞
  预处理♯NN ↔ 前/接頭辞 処理/名詞
  高分子♯NN ↔ 高/接頭辞 分子/名詞
  副♯JJ 生成物♯NN ↔ 副/接頭辞 生成/名詞 物/名詞
- “接尾辞”: 细胞♯NN 遗传学♯NN ↔ 細胞/名詞 遺伝/名詞 学/接尾辞
  磷化铟♯NN ↔ リン /名詞 化/接尾辞 インジウム/名詞
  平滑化♯NN ↔ 平滑/形容詞 化/接尾辞

We would obtain more promising bilingual terms with high quality for re-tokenizing the training corpus before training translation models for SMT systems.

It can also be used in neural machine translation (NMT) for solving the problem of rare words or terms. Such words are simply left out in translation in current NMT systems. As pre-processing, we can make use of these extracted bilingual terms in the source language and replace these terms in test sentences with specific markers which are known to the NMT system. After translation using an NMT system, post-processing these specific markers consists in replacing them with the corresponding translation in the target language.

5.2.3 Faster Production of Quasi-parallel Corpus using Chunks

In our work, we proposed an original way to construct a bilingual corpus for inflating an existing parallel corpus and improving the SMT translation accuracy. The constructed bilingual data are quasi-parallel short sentences. In our work, the length was limited to up to 30 characters. The newly generated sentences come from analogical associations contained in analogical clusters. In our work, we generated new sentences using each ratio (i.e., each sentence pair as a rewriting model) in each cluster. This allowed us to obtain different newly generated sentences even for one cluster. However this generating process was very much time consuming. To accelerate the method, to test a sampling method, i.e., we propose to generate new sentences according to sampled ratios in each cluster. Comparison with the results with the results obtained using all ratios will show whether they are comparable or not. If yes, this should be a way to obtain a similar number of newly generated sentences in a less time.

Another direction for future work is to use chunks or sequences of chunks instead of short sentences. This is a granularity between words and short sentences. Chunks can be computed as fragments delimited by markers. Such markers are を、に、は、が、と in Japanese and 的, 和, 了 in Chinese. We already started work on the automatic construction of a quasi-parallel corpus using analogical associations between chunks in Chinese and Japanese scientific texts. Preliminary results show that chunks might not be the right unit, but sequences of chunks seem more promising.

In this dissertation, we investigated the translation equivalences at different levels of granularity for improving Chinese–Japanese phrase-based statistical machine translation accuracy in technical domain. Word segmentation and segmentation consistency for languages without typographic boundaries are crucial, not only for statistical machine translation, but also for neural machine translation. Some previous work perform SMT or NMT experiments without word segmentation for Chinese or European languages (Xu et al., 2004), (Ling et al., 2015), but they do not always obtain better translation results in comparison with word-based translation systems. Previous work (Chang et al., 2008), (Chu et al., 2013a) and our work show that consistency in segmentation and granularity is more important for improving the translation accuracy of machine translation. Thus, how to automatically adjust the corresponding segmentation to keep granularity consistency across languages (not only for Chinese and Japanese, but also for some European languages) in any domain should be a direction of future work for improving the accuracy of natural language processing in general and the accuracy of machine translation in particular.

Bibliography

Aker, A., M. L. Paramita, and R. J. Gaizauskas
2013. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Pp. 402–411.

Allison, L. and T. I. Dix
1986. A bit-string longest-common-subsequence algorithm. *Inf. Process. Lett.*, 23(5):305–310.

Bai, M.-H., K.-J. Chen, and J. S. Chang
2008. Improving word alignment by adjusting Chinese word segmentation. In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP 2008)*, Pp. 249–256.

Besacier, L., E. Barnard, A. Karpov, and T. Schultz
2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Bin, L., T. Jiang, K. Chow, and B. K. Tsou
2010. Building a large English–Chinese parallel corpus from comparable patents and its experimental application to SMT. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, Pp. 42–49.

Brown, P. F., J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin
1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Brown, P. F., V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer
1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Chang, P.-C., M. Galley, and C. D. Manning
2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT 2008)*, Pp. 224–232. Association for Computational Linguistics.

Chiang, D.
2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 2005)*, Pp. 263–270. Association for Computational Linguistics.

Chu, C., T. Nakazawa, D. Kawahara, and S. Kurohashi
2013a. Chinese–Japanese machine translation exploiting Chinese characters. *ACM Transactions on Asian Language Information Processing (TALIP)*, 12(4):16.

Chu, C., T. Nakazawa, and S. Kurohashi
2012. Chinese characters mapping table of Japanese, traditional Chinese and simplified Chinese. In *Proceedings of International Conference on Language Resources and Evaluation (LREC 2012)*, Pp. 2149–2152.

Chu, C., T. Nakazawa, and S. Kurohashi
2013b. Accurate parallel fragment extraction from quasi-comparable corpora using alignment model and translation lexicon. In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP 2013)*, Pp. 1144–1150.

Chu, C., T. Nakazawa, and S. Kurohashi
2015. Integrated parallel sentence and fragment extraction from comparable corpora: A case study on Chinese–Japanese wikipedia. *ACM Trans. on Asian and Low-Resource Language Information Processing*, 15(2):10.

de Saussure, F.
1995. *Cours de linguistique générale*, [1ère éd. 1916] edition. Lausanne et Paris: Payot.

Delhay, A. and L. Miclet
2004. Analogical equations in sequences: Definition and resolution. *Lecture Notes in Computer Science*, 3264:127–138.

Do, T. N. D., L. Besacier, and E. Castelli
2010. A fully unsupervised approach for mining parallel data from comparable corpora. In *Proceedings of European Conference on Machine Translation (EAMT 2010).*

Doddington, G.
2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of Human Language Technology Conference (HLT 2002)*, Pp. 128–132.

Doddington, G. R., M. A. Przybocki, A. F. Martin, and D. A. Reynolds
2000. The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective. *Speech Communication*, 31(2):225–254.

Fan, X., N. Shimizu, and H. Nakagawa
2009. Automatic extraction of bilingual terms from a Chinese–Japanese parallel corpus. In *Proceedings of the 3rd International Universal Communication Symposium*, Pp. 41–45. ACM.

Frantzi, K., S. Ananiadou, and H. Mima
2000. Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Frantzi, K. T., S. Ananiadou, and J. Tsujii
1998. The C-value/NC-value method of automatic recognition for multi-word terms. In *Proceedings of International Conference on Theory and Practice of Digital Libraries*, Pp. 585–604. Springer.

Fung, P. and P. Cheung
2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, P. 1051. Association for Computational Linguistics.

Furuse, O. and H. Iida
1996. Incremental translation utilizing constituent boundary patterns. In *Proceedings of the 16th conference on Computational linguistics (COLING 1996)*, volume 1, Pp. 412–417. Association for Computational Linguistics.

Gentner, D.
1983. Structure-Mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Gentner, D., K. J. Holyoak, and B. N. Kokinov
2001. *The analogical mind: Perspectives from cognitive science*. MIT press.

Goh, C.-L., M. Asahara, and Y. Matsumoto
2005. Building a Japanese–Chinese dictionary using kanji/hanzi conversion. In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP 2005)*, Pp. 670–681. Springer.

Hadni, M., A. Lachkar, and S. A. Ouatik
2014. Multi-word term extraction based on a new hybrid approach for Arabic. In *Computer Science & Information Technology*, D. Nagamalai et al., eds., volume 4, Pp. 109–120.

Heafield, K.
2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, Pp. 187–197, Edinburgh, UK. Association for Computational Linguistics.

Higashinaka, R., N. Kobayashi, T. Hirano, C. Miyazaki, T. Meguro, T. Makino, and Y. Matsuo
2016. Syntactic filtering and content-based retrieval of twitter sentences for the generation of system utterances in dialogue systems. In *Situated Dialog in Speech-Based Human-Computer Interaction*, Pp. 15–26. Springer.

Holyoak, K. J. and J. E. Hummel
2001. Toward an understanding of analogy within a biological symbol system. *In: The Analogical mind: Perspectives from cognitive science*, Pp. 23–58.

Hu, X., H. Wang, and H. Wu
2007. Using RBMT systems to produce bilingual corpus for SMT. In *Proceedings of Conference on Empirical Methods in Natural Language Processing Conference on Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Pp. 287–295.

Huang, M., B. Ye, Y. Wang, H. Chen, J. Cheng, and X. Zhu
2014. New word detection for sentiment analysis. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Pp. 531–541.

Hutchins, W. J. and H. L. Somers
1992. *An introduction to machine translation*, volume 362. Academic Press London.

Isozaki, H., T. Hirao, K. Duh, K. Sudoh, and H. Tsukada
2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, Pp. 944–952. Association for Computational Linguistics.

Itkonen, E.
2005. *Analogy as Structure and Process: Approaches in linguistics, cognitive psychology and philosophy of science*, volume 14.

Jiang, J., J. Du, and A. Way
2011. Incorporating source-language paraphrases into phrase-based SMT with confusion networks. In *Proceedings of the 5th Workshop on Syntax, Semantics and Structure in Statistical Translation*, Pp. 31–40. Association for Computational Linguistics.

Jin, Y. and Z. Liu
2010. Improving Chinese–English patent machine translation using sentence segmentation. In *Proceedings of Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on*, Pp. 1–6. IEEE.

Kärkkäinen, J. and P. Sanders
2003. Simple linear work suffix array construction. In *Automata, Languages and Programming*, Pp. 943–955. Springer.

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

Koehn, P.
2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of (MT Summit X)*, volume 5, Pp. 79–86.

Koehn, P.
2010. *Statistical machine translation*. Cambridge University Press.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, and et al.
2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions (ACL 2007)*, Pp. 177–180. Association for Computational Linguistics.

Koehn, P., F. J. Och, and D. Marcu
2003. Statistical phrase-based translation. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003)*, Pp. 48–54.

Korkontzelos, I., I. P. Klapaftis, and S. Manandhar
2008. Reviewing and evaluating automatic term recognition techniques. In *Advances in Natural Language Processing*, Pp. 248–259. Springer.

Krauwer, S.
2003. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In *Proceedings of International Workshop SPEECH AND COMPUTER (SPECOM 2003)*, Pp. 8–15.

Kudo, T.
2005. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. net/.*

Langlais, P. and A. Patry
2007. Translating unknown words by analogical learning. In *Proceedings of Conference on Empirical Methods in Natural Language Processing Conference on Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Pp. 877–886.

Langlais, P. and F. Yvon
2008. Scaling up analogical learning. In *Proceedings of International Conference on Computational Linguistics (COLING 2008)*, volume Posters, Pp. 51–54.

Lardilleux, A. and Y. Lepage
2009. Sampling-based multilingual alignment. In *Proceedings of the 7th Conference on Recent Advances in Natural Language Processing (RANLP 2009)*, Pp. 214–218.

Lavallée, J.-F. and P. Langlais
2009. Morphological acquisition by formal analogy. In *Proceedings of Morpho Challenge 2009*, Corfu, Greece.

Lepage, Y.
1998. Solving analogies on words: An algorithm. In *Proceedings of International Conference on Computational Linguistics (COLING 1998)*, Pp. 728–735.

Lepage, Y.
2004. Analogy and formal languages. *Electronic Notes in Theoretical Computer Science*, 53:180–191.

Lepage, Y. and E. Denoual
2005a. Automatic generation of paraphrases to be used as translation references in objective evaluation measures of machine translation. In *Proceedings of the 3rd International Workshop on Paraphrasing*, Pp. 57–64.

Lepage, Y. and E. Denoual
2005b. Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, 19(3-4):251–282.

Li, X., Y. Meng, and H. Yu
2012. Improving Chinese-to-Japanese patent translation using English as pivot language. In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation (PACLIC 2012)*, Pp. 117–126.

Lin, C.-Y. and E. Hovy
2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT 2003)*, Pp. 71–78.

Ling, W., I. Trancoso, C. Dyer, and A. W. Black
2015. Character-based neural machine translation. *arXiv preprint arXiv:1511.04586.*

Lossio-Ventura, J. A., C. Jonquet, M. Roche, and M. Teisseire
2013. Combining C-value and keyword extraction methods for biomedical terms extraction. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine (LBM 2013)*, Tokyo, Japan.

Lu, B. and B. K. Tsou
2009. Towards bilingual term extraction in comparable patents. In *Proceedings of*

Pacific Asia Conference on Language, Information and Computation (PACLIC 2009), Pp. 755–762.

Lü, Y., J. Huang, and Q. Liu
2007. Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of Conference on Empirical Methods in Natural Language Processing Conference on Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Pp. 343–350.

Ma, Y. and A. Way
2009. Bilingually motivated domain-adapted word segmentation for statistical machine translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, Pp. 549–557. Association for Computational Linguistics.

Michael, P., A. Finch, and E. Sumita
2011. Integration of multiple bilingually-trained segmentation schemes into statistical machine translation. *IEICE transactions on information and systems*, 94(3):690–697.

Miclet, L. and A. Delhay
2003. Analogy on sequences: a definition and an algorithm. *technical report inria-00071610*, 1.

Milios, E., Y. Zhang, B. He, and L. Dong
2003. Automatic term extraction and document similarity in special text corpora. In *Proceedings of the 6th Conference of the Pacific Association for Computational Linguistics (PACLIC 2003)*, Pp. 275–284, Seoul, Korea. Association for Computational Linguistics.

Mima, H. and S. Ananiadou
2001. An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese. *Terminology*, 6(2):175–194.

Mima, H., K. Frantzi, and S. Ananiadou
1998. The C-value/Example-based approach to the automatic recognition of multi-word terms for cross-language terminolog. In *Proceedings of PRICAI, Joint Workshop on Cross Language Issues in Artificial Intelligence and Issues of Cross Cultural Communication*, Pp. 10–21.

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

Munteanu, D. S. and D. Marcu
2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics (COLING*

2006) and the 44th annual meeting of the Association for Computational Linguistics (ACL 2006), Pp. 81–88. Association for Computational Linguistics.

Na, S.-H. and H. T. Ng
2011. Enriching document representation via translation for improved monolingual information retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, Pp. 853–862. ACM.

Nagao, M.
1984. A framework of a mechanical translation between Japanese and English by analogy principle. *Artificial and human intelligence*, Pp. 351–354.

Nagao, M. and S. Mori
1994. A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese. In *Proceedings of the 15th Conference on Computational linguistics (COLING 1994)*, volume 1, Pp. 611–615. Association for Computational Linguistics.

Nakagawa, H., H. Kojima, and A. Maeda
2004. Chinese term extraction from web pages based on compound word productivity. In *Proceedings of the Third SIGHAN Workshop on Chinese Language Processing, ACL*. Citeseer.

Nakazawa, T., H. Mino, I. Goto, S. Kurohashi, and E. Sumita
2014. Overview of the 1st workshop on Asian translation. In *Proceedings of the 1st Workshop on Asian Translation (WAT 2014)*, Pp. 1–19.

Neubig, G., Y. Nakata, and S. Mori
2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics Human Language Technologies (ACL-HTL 2011)*, volume 2, short papers, Pp. 529–533.

Nießen, S., F. J. Och, G. Leusch, and H. Ney
2000. An evaluation tool for machine translation: Fast evaluation for machine translation research. In *Proceedings of International Conference on Language Resources and Evaluation (LREC 2000)*, Pp. 39–45.

Och, F. J.
2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL 2003)*, volume 1, Pp. 160–167. Association for Computational Linguistics.

Och, F. J. and H. Ney
2000. Improved statistical alignment models. In *Proceedings of the 38th Annual*

Meeting on Association for Computational Linguistics (ACL 2000), Pp. 440–447. Association for Computational Linguistics.

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

Onishi, T., M. Utiyama, and E. Sumita
2011. Paraphrase lattice for statistical machine translation. *IEICE Trans. Inf. Syst.*, 94(6):1299–1305.

Pang, B., K. Knight, and D. Marcu
2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT 2003)*, volume 1, Pp. 102–109. Association for Computational Linguistics.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu
2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of Annual Meeting on Association for Computational Linguistics (ACL 2002)*, Pp. 311–318.

Pirrelli, V. and F. Yvon
1999. Analogy in the lexicon: A probe into analogy-based machine learning of language. In *Proceedings of the 6th International Symposium on Human Communication, Santiago de Cuba, Cuba.*

Pitler, E., A. Louis, and A. Nenkova
2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics (ACL 2010)*, Pp. 544–554. Association for Computational Linguistics.

Rapp, R.
1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL 1999)*, Pp. 519–526. Association for Computational Linguistics.

Rauf, S. A. and H. Schwenk
2011. Parallel sentence generation from comparable corpora for improved SMT. *Machine translation*, 25(4):341–375.

Sakti, S., T. Vu, A. Finch, M. Paul, R. Maia, S. Sakai, T. Hayashi, N. Kimura, Y. Ashikari, E. Sumita, and et al.
2009. NICT/ATR Asian spoken language translation system for multi-party travel conversation. In *Proceedings of TCAST Workshop*, Pp. 26–30.

Sánchez-Cartagena, V. M., F. Sánchez-Martínez, J. A. Pérez-Ortiz, et al.
2011. Enriching a statistical machine translation system trained on small parallel corpora with rule-based bilingual phrases. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Pp. 90–96.

Scannell, K. P.
2007. The crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, Pp. 5–15.

Servan, C. and S. Petitrenaud
2012. Calculation of phrase probabilities for statistical machine translation by using belief functions. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*.

Silva, J., L. Coheur, Â. Costa, and I. Trancoso
2012. Dealing with unknown words in statistical machine translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Pp. 3977–3981.

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

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul
2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA 2006)*, Pp. 223–231.

Soricut, R. and E. Brill
2004. A unified framework for automatic evaluation using n-gram co-occurrence statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*, P. 613. Association for Computational Linguistics.

Specia, L., M. Turchi, N. Cancedda, M. Dymetman, and N. Cristianini
2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Conference of the European Association for Machine Translation (EAMT 2009)*, Pp. 28–37.

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

Stolcke, A. et al.
2002. SRILM–an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing (ICSLP 2002)*, volume 2, Pp. 257–286.

Stroppa, N. and F. Yvon
2005. An analogical learner for morphological analysis. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005)*, Pp. 120–127, Ann Arbor, MI.

Stroppa, N. and F. Yvon
2006. Formal models of analogical proportions.

Tan, C. L. and M. Nagao
1995. Automatic alignment of Japanese–Chinese bilingual texts. *IEICE Transactions on Information and Systems*, 78(1):68–76.

Tan, L. and S. Pal
2014. Manawi: Using multi-word expressions and named entities to improve machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Pp. 201–206.

Tan, L., J. van Genabith, and F. Bond
2015. Passive and pervasive use of a bilingual dictionary in statistical machine translation. *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP 2015)*, P. 30.

Tanaka, K. and K. Umemura
1994. Construction of a bulingual dictionary intermediated by a construction of a bulingual dictionary intermediated by a third language. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994)*, Pp. 297–303.

Tseng, H., P. Chang, G. Andrew, D. Jurafsky, and C. Manning
2005. A conditional random field word segmenter. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, volume 168–171.

Turney, P. D.
2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Turney, P. D. and M. L. Littman
2005. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1–3):251–278.

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

Vivaldi, J. and H. Rodríguez
2007. Evaluation of terms and term extraction systems: A practical approach. *Terminology*, 13(2):225–248.

Vulić, I., W. De Smet, and M.-F. Moens
2011. 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 (NAACL-HLT 2011)*, volume 2, short papers, Pp. 479–484. Association for Computational Linguistics.

Wagner, R. A. and M. J. Fischer
1974. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173.

Wang, D.
2009. Chinese to English automatic patent machine translation at SIPO. *World Patent Information*, 31(2):137–139.

Wang, H., W. Yang, and Y. Lepage
2014. Improved Chinese–Japanese phrase-based MT quality using an extended quasi-parallel corpus. In *Proceedings of Progress in Informatics and Computing (PIC 2014)*, Pp. 6–10. IEEE.

Wang, K., C. Zong, and K.-Y. Su
2010a. A character-based joint model for Chinese word segmentation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Pp. 1173–1181.

Wang, Y., K. Uchimoto, J. Kazama, C. Kruengkrai, and K. Torisawa
2010b. Adapting Chinese word segmentation for machine translation based on short units. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Pp. 1758–1764.

Weaver, W.
1955. Translation. *Machine translation of languages*, 14:15–23.

Wu, H. and H. Wang
2004. Improving domain-specific word alignment with a general bilingual corpus. In *Proceedings of Conference of the Association for Machine Translation in the Americas*, Pp. 262–271. Springer.

Xu, J., R. Zens, and H. Ney
2004. Do we need Chinese word segmentation for statistical machine translation. In *Proceedings of the Third SIGHAN Workshop on Chinese Language Learning*, Pp. 122–128.

Yamada, K. and K. Knight
2001. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL 2001)*, Pp. 523–530. Association for Computational Linguistics.

Yamamoto, M. and K. W. Church
2001. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. *Computational Linguistics*, 27(1):1–30.

Yang, M., H. Jiang, T. Zhao, and S. Li
2006. Construct trilingual parallel corpus on demand. In *Proceedings of Chinese Spoken Language Processing*, Pp. 760–767. Springer.

Yang, W. and Y. Lepage
2012. Combining several automatic techniques to build a Chinese–Japanese lexicon from freely available resources. In *Proceedings of the 18th Yearly Conference of the Japanese Association for Natural Language Processing*, Pp. 747–750.

Yang, W. and Y. Lepage
2014a. Consistent improvement in translation quality of Chinese–Japanese technical texts by adding additional quasi-parallel training data. In *Proceedings of the 1st Workshop on Asian Translation (WAT 2014)*, Pp. 69–76.

Yang, W. and Y. Lepage
2014b. Inflating a training corpus for SMT by using unrelated unaligned monolingual data. In *Proceedings of International Conference on Natural Language Processing*, Pp. 236–248. Springer.

Yang, W. and Y. Lepage
2016. Improving patent translation using bilingual term extraction and re-tokenization for Chinese–Japanese. In *Proceedings of the 3th Workshop on Asian Translaion (WAT 2016) co-located with COLING 2016*, Pp. 194–202.

Yang, W. and Y. Lepage
2017. Bilingual multi-word term tokenization for Chinese–Japanese patent translation. In *Proceedings of the 23th Yearly Conference of the Japanese Association for Natural Language Processing*, Pp. 855–858.

Yang, W., H. Shen, and Y. Lepage
2017. Inflating a small parallel corpus into a large quasi-parallel corpus using monolingual data for Chinese–Japanese machine translation. *Journal of Information Processing*, 25:88–99.

Yang, W., H. Wang, and Y. Lepage
2014. Deduction of translation relations between new short sentences in Chinese and Japanese using analogical associations. *International Journal of Advanced Intelligence (IJAI)*, 6(1):13–34.

Yang, W., J. Yan, and Y. Lepage
2016. Extraction of bilingual technical terms for Chinese–Japanese patent translation.

In *Proceedings of Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*, Pp. 81–87.

Yang, W., Z. Zhao, and Y. Lepage
2015. Inflating training data for statistical machine translation using unaligned monolingual data. In *Proceedings of The Association for Natural Language Processing*, Pp. 1016–1019.

Yvon, F.
1999. Pronouncing unknown words using multi-dimensional analogies. In *Proceedings of EUROSPEECH.*

Yvon, F., N. Stroppa, A. Delhay, and L. Miclet
2004. Solving analogical equations on words. *Rapport interne D*, 5.

Zaidan, O. F. and C. Callison-Burch
2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2011)*, volume 1, Pp. 1220–1229. Association for Computational Linguistics.

Zechner, K.
1996. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In *Proceedings of the 16th conference on Computational linguistics (COLING 1996)*, volume 2, Pp. 986–989. Association for Computational Linguistics.

Zeng, X., L. S. Chao, D. F. Wong, I. Trancoso, and L. Tian
2014. Toward better Chinese word segmentation for SMT via bilingual constraints. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Pp. 1360–1369.

Zhang, J. and C. Zong
2013. Learning a phrase-based translation model from monolingual data with application to domain adaptation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Pp. 1425–1434.

Zhang, Y., K. Uchimoto, Q. Ma, and H. Isahara
2005. Building an annotated Japanese–Chinese parallel corpus–A part of NICT multilingual corpora. In *Proceedings of the 10th Machine Translation Summit (MT Summit X)*, Pp. 71–78.

Zhao, H., M. Utiyama, E. Sumita, and B.-L. Lu
2013. An empirical study on word segmentation for Chinese machine translation. In *Proceedings of Computational Linguistics and Intelligent Text Processing*, Pp. 248–263. Springer.

List of Publications

Journals

- Wei Yang and Yves Lepage. Improving Automatic Chinese–Japanese Patent Translation using Bilingual Term Extraction. *IEEJ Transactions on Electrical and Electronic Engineering*, Vol.13, No.1, January 2018. (to appear)

- Wei Yang, Hanfei Shen and Yves Lepage. Inflating a Small Parallel Corpus into a Large Quasi-parallel Corpus Using Monolingual Data for Chinese–Japanese Machine Translation. *Journal of Information Processing*, Vol.25, pp. 88–99, January 2017.

- Wei Yang, Hao Wang and Yves Lepage. Deduction of Translation Relations between New Short Sentences in Chinese and Japanese Using Analogical Associations. *International Journal of Advanced Intelligence (IJAI)*, Vol.6, No.1, pp.13–34, December 2014.

International Conferences with Reviewing Committee

- Wei Yang and Yves Lepage. Improving Patent Translation Using Bilingual Term Extraction and Re-tokenization for Chinese–Japanese. In *Proceedings of the 3rd Workshop on Asian Translation (WAT 2016) co-located with COLING 2016*, pp. 194–202, Osaka, Japan, December 11-17, 2016.

- Wei Yang, Jinghui Yan and Yves Lepage. Extraction of Bilingual Technical Terms for Chinese–Japanese Patent Translation. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016) Student Research Workshop (SRW)*, pp. 81–87, San Diego, California, June 12-17, 2016.

- Wei Yang, Zhongwen Zhao, Baosong Yang and Yves Lepage. Sampling-based Alignment and Hierarchical Sub-sentential Alignment in Chinese–Japanese Translation of Patents. In *Proceedings of the 2nd Workshop on Asian Translation (WAT 2015)*, pp. 87–94, October 16, 2015.

- Wei Yang and Yves Lepage. Consistent Improvement in Translation Quality of Chinese–Japanese Technical Texts by Adding Additional Quasi-parallel Training Data. In *Proceedings of the 1st Workshop on Asian Translation (WAT 2014)*, pp. 69–76, October 4, 2014.

- Wei Yang and Yves Lepage. Inflating a Training Corpus for SMT by Using Unrelated Unaligned Monolingual Data. In *Proceedings of the 9th International Conference on NLP (PolTAL 2014) LNAI 8686*, pp. 236–248, September 17-19, 2014.

- Hao Wang, Wei Yang and Yves Lepage. Improved Chinese–Japanese Phrase-based MT Quality Using an Extended Quasi-parallel Corpus. In *Proceedings of 2014 IEEE International Conference on Progress in Information and Computing*, pp. 6–10, May 16-18, 2014.

- Wei Yang, Hao Wang and Yves Lepage. Automatic Acquisition of Rewriting Models for the Generation of Chinese–Japanese Quasi-parallel Corpus. In *Proceedings of the 6th Language and Technology Conference (LTC 2013)*, pp. 409–413, Poznań, Poland, December 2013.

- Wei Yang, Hao Wang and Yves Lepage. Using Analogical Associations to Acquire Chinese–Japanese Quasi-parallel Sentences. In *Proceedings of the 10th International Symposium on Natural Language Processing (SNLP 2013)*, pp. 86–93, October 28-30, 2013.

Conferences without Reviewing Committee

- Wei Yang and Yves Lepage. Bilingual Multi-Word Term Tokenization for Chinese–Japanese Patent Translation. In *Proceedings of the 23th Yearly Conference of the Japanese Association for Natural Language Processing (言語処理学会第23回年次大会 NLP2017)*, pp. 855–858, Tsukuba, March 13-17, 2017.

- Wei Yang and Mengru Gao and Yves Lepage. Production of Analogical Clusters between Marker-based Chunks in Chinese and Japanese. *10th International collaboration Symposium on Information, Production and Systems (ISIPS 2016)*, pp. 238–241, Kitakyushu, Fukuoka, Japan, November 2016.

- Wei Yang, Zhongwen Zhao and Yves Lepage. Inflating Training Data for Statistical Machine Translation Using Unaligned Monolingual Data. In *Proceedings of the 21th Yearly Conference of the Japanese Association for Natural Language Processing (言語処理学会第21回年次大会 NLP2015)*, pp. 1016–1019, March 2015.

- Wei Yang and Yves Lepage. Extending a Training Corpus for SMT by Using Analogical Associations based on Unrelated Monolingual Data. *8th International collaboration Symposium on Information, Production and Systems (ISIPS 2014)*, Kitakyushu, Fukuoka, Japan, November 2014. (no pagination)

- Hanfei Shen, Wei Yang and Yves Lepage. Filtering Techniques of Construction of a Quasi-parallel Corpus for Chinese–Japanese SMT System. *8th International collaboration Symposium on Information, Production and Systems (ISIPS 2014)*, Kitakyushu, Fukuoka, Japan, November 2014. (no pagination)

- Hao Wang, Wei Yang and Yves Lepage. Sentence Generation by Analogy: Towards the Construction of A Quasi-parallel Corpus for Chinese–Japanese. In *Proceedings of the 20th Yearly Conference of the Japanese Association for Natural Language Processing (言語処理学会第20回年次大会 NLP2014)*, pp. 900–903, March 2014.

- Wei Yang, Hao Wang and Yves Lepage. Using Analogical Association to Acquire Chinese–Japanese Quasi-parallel Sentences. *7th International collaboration Symposium on Information, Production and Systems (ISIPS 2013)*, Kitakyushu, Fukuoka, Japan, November 2013. (no pagination)

- Wei Yang, Hao Wang and Yves Lepage. Using Analogical Associations to Acquire Chinese–Japanese Quasi-parallel Sentences. *the International Workshop on Machine Vision for Industrial Innovation (MVII2013)*, p. 138, Kitakyushu, Fukuoka, Japan, October 2013.

- Wei Yang, Jiajia Xie and Yves Lepage. Structuring Sentential Data for Less-Documented Language Pairs: Chinese–Japanese. *6th International collaboration Symposium on Information, Production and Systems (ISIPS 2012)*, Kitakyushu, Fukuoka, Japan, November 2012. (no pagination)

- Wei Yang and Yves Lepage. Combining Several Automatic Techniques to Build a Chinese–Japanese Lexicon from Freely Available Resources. In *Proceedings of the 18th Yearly Conference of the Japanese Association for Natural Language Processing (言語処理学会第18回年次大会 NLP2012)*, pp. 747–750, Hiroshima, March 2012.