Improving Patent Translation using Bilingual Term Extraction and Re-tokenization for Chinese–Japanese

Wei Yang
IPS, Waseda University
2-7 Hibikino, Wakamatsu Kitakyushu
Fukuoka,, Japan
kevinyoogi@akene.waseda.jp

Yves Lepage
IPS, Waseda University
2-7 Hibikino, Wakamatsu Kitakyushu
Fukuoka,, Japan
yves.lepage@waseda.jp

Abstract

Unlike European languages, many Asian languages like Chinese and Japanese do not have typographic boundaries in written system. Word segmentation (tokenization) that break sentences down into individual words (tokens) is normally treated as the first step for machine translation (MT). For Chinese and Japanese, different rules and segmentation tools lead different segmentation results in different level of granularity between Chinese and Japanese. To improve the translation accuracy, we adjust and balance the granularity of segmentation results around terms for Chinese–Japanese patent corpus for training translation model. In this paper, we describe a statistical machine translation (SMT) system which is built on re-tokenized Chinese–Japanese patent training corpus using extracted bilingual multi-word terms.

1 Introduction

China and Japan are producing a large amount of patents in their respective languages. Making Chinese patents available in Japanese, and Japanese patents in Chinese is an important task for increasing economical development in Asia and international world. The translation of patents is a key issue that should be helped by the use of SMT.

Word segmentation is normally treated as the first step for SMT between Chinese and Japanese. Patents contain large amounts of domain-specific terms in words or multi-word expressions. This brings up the question of word segmentation: we may not want to tokenize terms in specific domains in patents. But we cannot control the tokenization of the multi-word terms: a large number of multi-word terms are always segmented into several single-word terms in one language but may not be segmented in another language, or some of the multi-word terms in two languages have different levels of granularity in segmentation because of different conventions of segmentation in different languages.

The related work by Chang et al. (2008) shows that segmentation granularity of Chinese word segmentation affects the translation accuracy and that it is very important for MT. In (Chu et al., 2013), for improving the translation accuracy of scientific papers, they make use of a constructed 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 segmentation and translation. To improve SMT translation accuracy, we change and adjust the segmentation for terms using extracted bilingual multi-word terms for both languages (not only for Chinese or Japanese).

Frantzi et al. (2000) describes a combination of linguistic and statistical methods (C-value/NC-value) for the automatic extraction of multi-word terms from English corpora. In (Mima and Ananiadou, 2001), it is showed that the C-/NC-value method is an efficient domain-independent multi-word term recognition not only in English but in Japanese as well. In this paper, we adopt the C-value method to extract monolingual multi-word terms in Chinese and Japanese, and combine it with the sampling-based alignment method (Lardilleux and Lepage, 2009) and kanji-hanzi conversion method for bilingual multi-word term extraction. We build SMT systems based on re-tokenized Chinese–Japanese patent training corpus using the extracted bilingual multi-word terms.

Proceedings of the 3rd Workshop on Asian Translation,
pages 194–202, Osaka, Japan, December 11-17 2016.

Language	Sentence
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	‘Tantalum anode body are usually 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 1: Examples of Chinese–Japanese patent segmentation. Terms in different languages are tokenized at different levels of granularity. Segmentation tools used are Stanford for Chinese and Juman for Japanese. The words given in the box are the multi-word terms or single-word terms in Chinese or Japanese. The words in the same color have corresponding translation relations between two languages.

2 Word Segmentation for Chinese–Japanese Patent Corpus

Figure 1 gives the examples for 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 has a translation relation with the multi-word タンタル/陽極/ボディ in Japanese, but actually, they do not have any correspondence in word-to-word alignments. Similar examples are given as 异氰/酸酯基 (‘isocyanate group’) in Chinese and イソシアネート/基 in Japanese, 放射线/量 (‘radiation dose’) in Chinese and 放射/線量 in Japanese. Another case is that some terms are multi-word terms in one language but single-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. For keeping the direct and exact translations between Chinese and Japanese terms, we intend to re-tokenize Chinese–Japanese parallel sentences center around bilingual multi-word terms. As such, correspondence and meaning of terms come into focus when adjusting word tokenization granularity.

To do this, we extract bilingual multi-word terms from an existing Chinese–Japanese training corpus, then we build SMT systems based on the re-tokenized training corpus using these extracted bilingual multi-word terms by enforcing them to be considered as one token.

3 Chinese–Japanese Bilingual Multi-word Term Extraction

In this section, we describe a bilingual multi-word term extraction method used in our work. We combine using C-value for monolingual multi-word extraction with the sampling-based alignment method and kanji-hanzi conversion method for bilingual multi-word term extraction.

3.1 Monolingual Multi-word Term Extraction

The C-value is an automatic domain-independent method, commonly used for multi-word term extraction. This method has two main parts: linguistic part and statistical part. The linguistic part gives the type of multi-word terms extracted relying on part-of-speech tagging, linguistic filters, stop list, etc. The statistical part measure a termhood to a candidate string, and output a list of candidate terms with decreasing order of C-value. In our experiments, we extract multi-word terms which contain a sequence of nouns or adjectives followed by a noun in both Chinese and Japanese. This linguistic pattern[1] can be written as follows using a regular expression: $(Adjective|Noun)^+ \, Noun$. The segmenter and part-of-speech tagger that we use are the Stanford parser[2] for Chinese and Juman[3] for Japanese.

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

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

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

In our experiments, we follow the basic steps of the C-value approach to extract Chinese and Japanese monolingual multi-word terms respectively from the existing Chinese–Japanese training corpus. We firstly tag each word in the Chinese and the Japanese corpus respectively; we then extract multi-word terms based on the linguistic pattern and the formula given above for each language. The stop list is used to avoid extracting infelicitous sequences of words consists of 240 function words (including numbers, letters and punctuations etc.). Examples of term candidates in Chinese and Japanese extracted are shown in Table 1. We then re-tokenize such candidate terms in the Chinese–Japanese training corpus by enforcing them to be considered as one token. Each candidate multi-word term is aligned with markers.

3.2 Bilingual Multi-word Term Extraction

We extract bilingual multi-word terms based on re-tokenized Chinese–Japanese training corpus (with extracted monolingual muti-word terms) with two methods: one is using the sampling-based alignment method, another one is taking kanji-hanzi conversion into consideration.

3.2.1 Using Sampling-based Method

To extract bilingual aligned multi-word terms, we use the open source implementation of the sampling-based alignment method, Anymalign (Lardilleux and Lepage, 2009), to perform word-to-word alignment (token-to-token alignment)[4] from the above monolingual terms based re-tokenized Chinese–Japanese training corpus. We recognize the multi-word term to multi-word term alignments between Chinese and Japanese by using the markers. We then filter these aligned multi-word candidate terms by setting some threshold P for the translation probabilities in both directions.

Table 2 shows some bilingual multi-word terms that we extracted by setting a threshold P with 0.6. It is possible that some incorrect alignments are extracted. Such examples appear on the alignments with $*$. To improve the precision (good match) of the results, we further filter these extracted bilingual multi-word terms (obtained by setting threshold P) by computing the ratio of the lengths in words between the Chinese (Japanese) part and its corresponding Japanese (Chinese) part.

We set the ratio of the length in words between two languages with 1.0, 1.5, 2.0 and 2.5. The precision of the kept bilingual multi-word terms in each ratio is assessed by sampling 100 bilingual multi-word terms. On the bilingual multi-word term extraction results obtained by setting P=0.6, the precisions

[1] 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.

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

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

[4] This is done by the option -N 1 on the command line. Experiments were also done with GIZA++, the sampling-based alignment method is more efficient than GIZA++.

Chinese or Japanese sentences	Extracted monolingual terms
Chinese: 在♯P 糖尿病♯NN ,♯PU 更♯AD 具体♯VA 地♯DEV 1♯CD 型♯NN 或♯CC 2♯CD **型♯NN 糖尿病♯NN** 患者♯NN 的♯DEC 情况♯NN 中♯LC ,♯PU 本♯DT 发明♯NN 的♯DEC 药物♯NN 容许♯VV 血液♯NN **葡萄糖♯NN 浓度♯NN** 更♯AD 有效♯VA 地♯DEV 适应♯VV 于♯P **血糖♯NN 正常♯JJ 水平♯NN** 。 ♯PU **Japanese:** 本/接頭辞 発明/名詞 の/助詞 薬剤/名詞 は/助詞 、/特殊 **糖尿/名詞 病/名詞** 、/特殊 より/副詞 詳細に/形容詞 は/助詞 1/特殊 型/接尾辞 又は/助詞 2/特殊 型/接尾辞 **糖尿/名詞 病/名詞** の/助詞 患者/名詞 の/助詞 場合/名詞 に/助詞 血/名詞 中/接尾辞 **グルコース/名詞 濃度/名詞** を/助詞 **正常/形容詞 血糖/名詞 レベル/名詞** まで/助詞 より/助詞 効果/名詞 的に/接尾辞 適合/名詞 さ/動詞 せる/接尾辞 こと/名詞 を/助詞 可能に/形容詞 する/接尾辞 。 /特殊 English meaning: 'In diabetes, more particularly, type 1 or 2 diabetes cases, the drug of the present invention allows the blood glucose concentration more effectively adapt to normal blood glucose levels.'	型 糖尿病 'the type of diabetes' 葡萄糖 浓度 'glucose concentration' 血糖 正常 水平 'normal blood glucose level' 糖尿 病 'diabetes' グルコース 濃度 'glucose concentration' 正常 血糖 レベル 'normal blood glucose level'
Chinese: 在♯P 该♯DT 方法♯NN 中♯LC ,♯PU 能够♯VV 得到♯VV 从♯P **心脏♯NN 周期♯NN** 内♯LC 的♯DEG **心收缩♯NN 期♯NN** 到♯VV 心♯NN 舒张♯VV 期♯NN 之间♯LC 的♯DEG 血液♯NN 移动♯VV 的♯DEC 1♯CD 个♯M 以上♯LC 的♯DEG 图像♯NN 。 ♯PU **Japanese:** この/指示詞 方法/名詞 に/助詞 おいて/動詞 は/助詞 、/特殊 **心臓/名詞 周期/名詞** 内/接尾辞 の/助詞 **心/名詞 収縮/名詞 期/名詞** から/助詞 **心/名詞 拡張/名詞 期/名詞** まで/助詞 の/助詞 間/名詞 の/助詞 **血液/名詞 移動/名詞** の/助詞 1/名詞 枚/接尾辞 以上/接尾辞 の/助詞 画像/名詞 が/助詞 得/動詞 られる/接尾辞 。 /特殊 English meaning: 'In this method, we can obtain more than one images of blood moving from systole of cardiac cycle to diastole.'	心脏 周期 'cardiac cycle' 心收缩 期 'systole' 心臓 周期 'cardiac cycle' 心 収縮 期 'systole' 心 拡張 期 'diastole' 血液 移動 'blood moving'

Table 1: Examples of multi-word term extracted using C-value, based on the linguistic pattern: $(Adjective|Noun)^{+}\ Noun$.

for each ratio are 94%, 92%, 90% and 80%. It is obvious that the precision of the extracted bilingual multi-word terms decreases rapidly when the ratio tends to 2.5, thus 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. Another filtering constraint is to filter out alignments of the Japanese part which contains hiragana. 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).

3.2.2 Using Kanji-hanzi Conversion Method

Table 2 leads to the observation that some correctly aligned bilingual terms cannot be extracted by using the methods we described in Section 3.2.1. Such examples of terms are given in Table 2 with ×. Such examples are the multi-word terms on one side (Chinese or Japanese) are not multi-word terms in another side (Japanese or Chinese), or filtered by setting a threshold on translation probabilities. Kanji-hanzi

Extract or not	Correct or not	Chinese	Japanese	Meaning	$P(t\mid s)$	$P(s\mid t)$
◯	√	葡萄糖␣浓度	グルコース␣濃度	'glucose concentration'	0.962121	0.891228
◯	√	血糖␣正常␣水平	正常␣血糖␣レベル	'normal blood glucose level'	1.000000	1.000000
◯	√	心脏␣周期	心臓␣周期	'cardiac cycle'	1.000000	1.000000
◯	√	心收缩␣期	心␣収縮␣期	'systole'	1.000000	0.833333
◯	√	脂肪␣酸酯	脂肪␣酸␣エステル	'fatty acid ester'	1.000000	0.983333
◯	∗	糖尿病␣小鼠␣中肾␣小管␣上皮␣细胞	上皮␣細胞	-	1.000000	1.000000
◯	∗	上述␣液体状	前記␣アルカリ␣活性␣結合␣材	-	1.000000	1.000000
◯	∗	上述靶␣蛋白	種々の␣上記	-	1.000000	1.000000
×	√	糖尿病	糖尿␣病	'diabetes'	1.000000	0.666667
×	√	肺癌	肺␣癌	'lung cancer'	1.000000	1.000000
×	√	杀生␣物剂	殺生␣物␣剤	'biocide'	0.600000	0.107143
×	√	官能␣基	官能␣基	'functional group'	0.250000	0.009231
×	√	废␣热	廃␣熱	'waste heat'	0.844444	0.240506

Table 2: Extraction of Chinese–Japanese bilingual multi-word terms by setting a threshold P with 0.6 for both directions. ◯ and × show the bilingual multi-word term alignment that are kept or excluded. √ and ∗ show the extracted multi-word terms are correct or incorrect alignments by human assessment.

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%)

Table 3: Distribution of the components for multi-word terms in Japanese (52,785 bilingual multi-word terms obtained by setting threshold P with 0).

conversion method can be used to extract this kind of bilingual multi-word terms.

We keep the alignments where either one side is a multi-word term; we convert Japanese words only made up of Japanese kanji 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). We compare Converted-Ja with the Zh, if a converted Japanese term is equal to its corresponding Chinese term in each character, we keep this pair of bilingual term. In this way, we can extract more reliable Chinese–Japanese bilingual aligned multi-word terms.

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[5]. In particular we used the correspondence relation SimplifiedVariant in the Unihan Mapping Data of the Unihan database. The second source of data we used is the Langconv Traditional-Simplified Conversion[6] data. It contains a database for traditional-simplified character. 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 mapping table, provided in the resource 簡体字と日本漢字対照表[7] which consists of simplified hanzi and kanji pairs. Table 4 shows the results of extracted bilingual multi-word terms by kanji-hanzi conversion using these three sources of data.

	Zh	Ja	Converted-Ja	Meaning	Human assessment
Without any Conversion	官能__基	官能__基	官能__基	‘functional group’	√
	肺癌	肺__癌	肺__癌	‘lung cancer’	√
	免疫原	免疫__原	免疫__原	‘immunogen’	√
By Traditional-Simplified Conversion	脉管	**脈**__管	脉__管	‘vessel’	√
	高温__杀菌	高温__**殺**菌	高温__杀菌	‘high temperature sterilization’	√
	放射线__源	放射__**線**__源	放射__线__源	‘radiation source’	√
By hanzi-kanji Mapping Table	心收缩__期	心__**収縮**__期	心__收缩__期	‘systole’	√
	废热__回收	**廃**__**熱**__回**収**	废__热__回收	‘waste heat recovery’	√
	肺气肿	肺__**気腫**	肺__气肿	‘pulmonary emphysema’	√
	添加剂	添加__**剤**	添加__剂	‘additive’	√
	肝脏__再生__作用	肝**臓**__再生__作用	肝脏__再生__作用	‘liver regeneration action’	√

Table 4: Extraction of bilingual Chinese–Japanese multi-word terms using kanji-hanzi conversion.

3.3 Bilingual Multi-word Terms Used in SMT

We re-tokenize the Chinese–Japanese training parallel corpus with the further filtered bilingual multi-word terms (by ratio of the lengths in words and components of the terms) combine with the extraction results by kanji-hanzi conversion. Each pair of bilingual multi-word terms are re-tokenized as one token and aligned with markers. In the procedure for building SMT systems, we training the Chinese–Japanese translation models on the re-tokenized training corpus. A language model is trained with the Japanese corpus without re-tokenizing annotation. We then remove the markers from the phrase tables before perform tuning and decoding in SMT experiments.

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

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

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

4 Experiments and Results

4.1 Chinese–Japanese Experimental Data Used

The Chinese–Japanese parallel sentences used in our experiments are randomly extracted from the Chinese–Japanese JPO Patent Corpus (JPC)[8]. JPC 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, 1,000 parallel sentences from the tuning part, and 1,000 from the test part. Table 5 shows basic statistics on our data sets.

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

Table 5: Statistics on our experimental data sets (after tokenizing and lowercasing). Here 'mean ± std.dev' gives the average length of the sentences in words.

4.2 Monolingual and Bilingual Multi-word Term Extraction

We extract 81,618 monolingual multi-word terms for Chinese and 93,105 for Japanese respectively based on the 100,000 lines of training corpus as indicated in Table 5. The precision was 95% in both languages. For keeping the balance between monolingual term used for re-tokenization in both languages, we re-tokenize the training corpus in each language with the same number of Chinese and Japanese monolingual multi-word terms. They are the first 80,000 monolingual multi-word terms with higher C-value in both languages.

Table 6 gives the number of bilingual multi-word terms obtained for different thresholds P (translation probabilities) from the re-tokenized (with extracted monolingual multi-word terms) 100,000 lines of training corpus (given in column (a)). Table 6 also gives the results of filtering with the constraints on the ratio of lengths in words between Chinese and Japanese terms and filtering out Japanese terms containing hiragana (given in column (a + b)). We extracted 4,591 bilingual multi-word terms (100% good match) from 309,406 phrase alignments obtained by word-to-word alignment from Chinese–Japanese training corpus using kanji-hanzi conversion. The number of the extracted multi-word terms using kanji-hanzi conversion combined with further filtering by constraints are given in Table 6 (column (a + b + c)).

5 Translation Accuracy in BLEU and Result Analysis

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

- several thresholds P for filtering (Table 6 (a))
- further filtering with several thresholds combined with kanji-hanzi conversion results (Table 6 (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-tokenizing is used to train a language model using KenLM (Heafield, 2011). After removing markers from the phrase table, we tune and test.

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

	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)	BLEU	p-value	♯ of bilingual multi-word terms (a + b)	♯ of bilingual multi-word terms (a + b + c)	BLEU	p-value
$\geq$ 0.0	52,785 (35%)	32.63	$>$ 0.05	48,239 (63%)	49,474 (70%)	33.15	$<$ 0.01
$\geq$ 0.1	31,795 (52%)	32.76	$>$ 0.05	29,050 (68%)	30,516 (78%)	33.10	$<$ 0.01
$\geq$ 0.2	27,916 (58%)	32.57	$>$ 0.05	25,562 (75%)	27,146 (83%)	33.05	$<$ 0.01
Baseline	-	32.38	-	-	-	32.38	-
$\geq$ 0.3	25,404 (63%)	33.07	$<$ 0.01	23,321 (78%)	25,006 (83%)	33.21	$<$ 0.01
$\geq$ 0.4	23,515 (72%)	32.92	$<$ 0.01	21,644 (80%)	23,424 (84%)	33.29	$<$ 0.01
$\geq$ 0.5	21,846 (76%)	33.05	$<$ 0.01	20,134 (85%)	22,000 (88%)	33.38	$<$ 0.01
$\geq$ 0.6	**20,248 (78%)**	**33.61**	$<$ 0.01	18,691 (88%)	**20,679 (89%)**	**33.93**	$<$ 0.01
$\geq$ 0.7	18,759 (79%)	32.92	$<$ 0.01	17,340 (88%)	19,460 (90%)	33.43	$<$ 0.01
$\geq$ 0.8	17,311 (79%)	33.34	$<$ 0.01	16,001 (89%)	18,265 (90%)	33.41	$<$ 0.01
$\geq$ 0.9	15,464 (80%)	33.47	$<$ 0.01	14,284 (92%)	16,814 (93%)	33.52	$<$ 0.01

Table 6: Evaluation results in BLEU for Chinese to Japanese translation based on re-tokenized training corpus using different thresholds (a); based on combination of the ratio of lengths + the components (b) with kanji-hanzi conversion (c).

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 6 shows the evaluation of the results of Chinese-to-Japanese translation in BLEU scores (Papineni et al., 2002). Compared with the baseline system, for the training corpus re-tokenized with further filtering combined with kanji-hanzi conversion results (a +b + c), we obtain significant improvements in all thresholds. We obtain 1.55 BLEU point (threshold of 0.6) improvements compare with the baseline system. In this case, 20,679 re-tokenized terms are used. It is also improve 0.3 BLEU point comparing with the case of the bilingual terms are filtered only by thresholds (a). We then test 2,000 sentences based on this best SMT system and the baseline system. We obtain a significant BLEU score with 33.61 compare with the baseline system 32.29 (p-value $<$ 0.01).

Figure 2 gives an example of improvement in Chinese-to-Japanese translation. Thanks to our method, re-tokenizing the training corpus with bilingual multi-word terms gave a better translation accuracy (BLEU=15.92) of the test sentence given in this example. Re-tokenizing and grouping the bilingual multi-word term together increased the probability of multi-word term to multi-word term translation, i.e., "免疫 测定 方法" to "免疫 測定 方法" ('immunoassay') in this example. This 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 "测定 方法" to "測定 方法" ('measuring method'). In this example, re-tokenization of the training corpus with extracted bilingual multi-word terms induced a direct and exact translation.

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) 测定(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 multi-word 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 2: Example of Chinese-to-Japanese translation improvement. The numbers in the parentheses show the position of the word in the test sentence. The numbers in the vertical lines show for the translation result (Japanese), the position of the n-gram used in the test sentence (Chinese).

6 Conclusion and Future Work

In this paper, we described a Chinese–Japanese SMT system for the translation of patents built on a training corpus re-tokenized using automatically extracted bilingual multi-word terms.

We extracted monolingual multi-word terms from each part of the Chinese–Japanese training corpus by using the C-value method. For extraction of bilingual multi-word terms, we firstly re-tokenized the training corpus with these extracted monolingual multi-word terms for each language. 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 multi-word terms which could not be extracted using thresholds or only one side is multi-word terms. We did not use any other additional corpus or lexicon in our work.

Re-tokenizing the parallel training corpus with the results of the combination of the extracted bilingual multi-word terms led to statistically significant improvements in BLEU scores for each threshold. We then test 2,000 sentences based on the SMT system with the highest BLEU score (threshold of 0.6). We also obtained a significant improvement in BLEU score compare with the baseline system.

In this work, we limited ourselves to the cases where multi-word terms could be found in both languages at the same time, e.g., 血糖__正常__水平 (Chinese) 正常__血糖__レヘル (Japanese) ('normal blood glucose level'), and the case where multi-word terms made up of hanzi/kanji are recognized in one of the languages, but not in the other language. e.g. 癌细胞 (Chinese) 癌__細胞 (Japanese) ('cancer cell') or 低__压 (Chinese) 低圧 (Japanese) ('low tension').

Manual inspection of the data allowed us to identify a third case. It is the case where 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, or the Japanese part is made up of kanji but they do not share the same characters with Chinese after kanji-hanzi conversion. Such a case is, e.g., 碳纳米管 (Chinese) カーボン__ナノチューブ (Japanese) ('carbon nano tube') and 控制器 (Chinese) コント__ローラ (Japanese) ('controller'), or 逆变__器 (Chinese) インバータ (Japanese) ('inverter') or still 乙酸乙酯 (Chinese) 酢酸__エチル (Japanese) ('ethyl acetate') and 尿键 (Chinese) ウレア__結合 (Japanese) ('urea bond') , or 氧化物 (Chinese) 酸化__物 (Japanese) ('oxide') and 缺氧 (Chinese) 酸素__欠乏 (Japanese) ('oxygen deficit'). In a future work, we intend to address this third case and expect further improvements in translation results.

Acknowledgements

References

Pi-Chuan Chang, Michel Galley, and Christopher D Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the third workshop on statistical machine translation*, pages 224–232. Association for Computational Linguistics.

Chenhui Chu, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2013. Chinese-Japanese machine translation exploiting Chinese characters. *ACM Transactions on Asian Language Information Processing (TALIP)*, 12(4):16.

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

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, and et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL 2007)*, pages 177–180. Association for Computational Linguistics.

Adrien Lardilleux and Yves Lepage. 2009. Sampling-based multilingual alignment. In *Recent Advances in Natural Language Processing*, pages 214–218.

Hideki Mima and Sophia 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL 2002*, pages 311–318.