IEEJ Transactions on Electronics, Information and Systems
Vol.131 No.1 pp.1–9 DOI: 10.1541/ieejeiss.131.1

Paper

Improving Automatic Chinese–Japanese Patent Translation using Bilingual Term Extraction

Wei Yang*a) Non-member, Yves Lepage* Non-member

(Manuscript received xxx 00, 2016, revised xxx 00, 2016)

The identification of terms in scientific and patent documents is a crucial issue for applications like information retrieval, text categorization and also for machine translation (MT). This paper describes a method to improve Chinese–Japanese statistical machine translation (SMT) of patents by re-tokenizing the training corpus with aligned bilingual multi-word terms. We automatically extract multi-word terms from monolingual corpora by combining statistical and linguistic filtering methods. An automatic alignment method is used to identify corresponding terms. The most promising bilingual multi-word terms are extracted by setting some threshold on translation probabilities and further filtering by considering the components of the bilingual multi-word terms in characters as well as the ratio of their lengths in words. We also use kanji (Japanese)–hanzi (Chinese) character conversion to confirm and extract more promising bilingual multi-word terms. We obtain a high quality of correspondence with 93% in bilingual term extraction and a significant improvement of 1.5 BLEU score in translation experiment.

Keywords: Term extraction, monolingual term, bilingual term, alignment, statistical machine translation

1. Introduction

In today's Japan, with the concern of discovering new growth strategy, preserving social security and introducing tax integrated reform, specific goals in highly technical domains like innovative pharmaceuticals, medical equipment etc. in conjunction with very specialized goals such as achieving a healthy longevity society, a considerable number of researchers and developers are creating, writing, consulting and reading a large amount of documents in new fields such as "life innovation" or "green innovation". An important factor for the success of these innovations is international exchange, with the aim of contributing to economic growth by strengthening international competitiveness.

China and Japan are producing a large amount of scientific journals 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 scientists. Additionally, China and Japan are more efficient at converting research and development dollars into patents than the U.S. or the European countries. 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 and international world.

Terms are scientific or technical words or expressions with specific scientific or technical meaning in specific domains. Their usage is generally limited to description for scientific or technical documents. Monolingual or bilingual term extraction is an important task for information retrieval, text categorization, clustering and also for machine translation. Reference [1] describes a combination of linguistic and statistical methods (C-value/NC-value) for the automatic extraction of multi-word terms from English corpora. As an application, in Ref. [2], they extract English terms in the computer science and medical domains using a C-value/NC-value extraction method, and made use of these terms to estimate the similarity of papers in the computer science corpus using the Vector Space Model (VSM). Compared with the word-based approach, the term-based representation retrieval use a much smaller size of vocabulary. As for language independence, in Ref. [3], 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 bilingual multi-word term extraction to improve Chinese–Japanese patent translation by re-tokenizing the training corpus with extracted bilingual multi-word terms.

Some pieces of work recognize monolingual or bilingual terms by considering compound words and their constituents. In Ref. [4], 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 Ref. [5], 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 extracted monolingual terms. The word alignments containing terms are smoothed by computing the associations between pairs of bilingual term candidates. They add

a) Correspondence to: kevinyoogi@akane.waseda.jp
* Graduate School of IPS, Waseda University,
2-7 Hibikino, Wakamatsu Kitakyushu Fukuoka, 808-0135, Japan

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 markers), i.e., improving patent translation quality by changing and balancing the granularity of the training data in Chinese and Japanese based on bilingual multi-word terms. In Ref. [6], English–Japanese multi-word terms are recognized by C-value 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) [7] 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 paper, we similarly consider monolingual multi-word terms extracted from a Chinese and Japanese corpus using C-value as one token for processing. Different from the example-based approach used in Ref. [6], we then use the sampling-based alignment method [8] 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 a simplified–traditional character conversion data and a free available hanzi–kanji mapping 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 exist 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 Ref. [9], 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 Ref. [10], they use the occurrence of identical common Chinese characters in Chinese–Japanese in the sentence alignment task. In Ref. [11][12], 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.

The paper is organized as follows: in Section 2, we describe our proposed methods to extract Chinese–Japanese bilingual multi-word terms using the C-value and the sampling-based alignment method, and using a kanji-hanzi conversion based extraction method. Section 3 introduces the experimental data sets used in our experiments and gives an analysis of the experimental results. In Section 4, we give the conclusion and discuss future directions.

2. Chinese–Japanese Bilingual Multi-word Term Extraction

In this section, we present our bilingual multi-word term extraction method that uses C-value [1] combined with the sampling-based alignment method [8]. In addition, we describe our improved Chinese–Japanese aligned term extraction method which use kanji-hanzi conversion based on freely available resources.

2.1 Monolingual Multi-word Term Extraction Using C-value

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, stop 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 be a real term. 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 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[††] for Chinese and Juman[†††] for Japanese.

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

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

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 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 stop list is used to avoid extracting infelicitous sequences of words. Our stop list consists of 240 function words (including numbers, letters and punctuations etc.) Examples of term candidates in Chinese and Japanese extracted are shown in Tab. 1. We re-tokenize such candidate terms in the corpus by enforcing them to be considered as one token (see Tab. 2). Each candidate multi-word term is re-tokenized with markers.

[†] 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.

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

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

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

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	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'
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.'	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	心脏　周期 'cardiac cycle' 心收缩　期 'systole'
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.'	心臓　周期 'cardiac cycle' 心　収縮　期 'systole' 心　拡張　期 'diastole' 血液　移動 'blood moving'
Chinese: 于本/NR 发明/VV 的/DEC **加热/JJ 烹饪/NN** 用/P **油脂/NN 组成/NN 物/NN** 中/LC ,/PU 除了/P 上述/JJ 聚/VV 甘油/NN **脂肪/NN 酸酯/NN** 以外/LC ,/PU 亦/AD 可/VV 包含/VV **植物/NN 油脂/NN** 。/PU	加热　烹饪 'cooking' 油脂　组成　物 'fat composition' 脂肪　酸酯 'fatty acid ester' 植物　油脂 'vegetable oil and fat'
Japanese: 本/接頭辞 発明/名詞 の/助詞 **加熱/名詞 調理/名詞** 用/接尾辞 **油脂/名詞 組成/名詞 物/名詞** に/助詞 は/助詞 、/特殊 上記/名詞 ポリグリセリン/未定義語 **脂肪/名詞 酸/名詞 エステル/名詞** の/助詞 ほか/名詞 、/特殊 **植物/名詞 油脂/名詞** を/助詞 含む/動詞 こと/名詞 が/助詞 できる/動詞 。/特殊 English meaning: 'The fat composition for cooking of this invention, in additional to the above-mentioned polyglycerol fatty acid ester, it may also contain vegetable oil and fat.'	加熱　調理 'cooking' 油脂　組成　物 'fat composition' 脂肪　酸　エステル 'fatty acid ester' 植物　油脂 'vegetable oil and fat'

Table 2. Examples of re-tokenized parallel sentences with extracted monolingual multi-word terms in Chinese and Japanese respectively.

Chinese sentence	Japanese sentence
在 糖尿病 ， 更 具体 地 1 型 或 2 **型__糖尿病** 患者 的 情况 中 ， 本 发明 的 药物 容许 血液 **葡萄糖__浓度** 更 有效 地 适应 于 **血糖__正常__水平** 。	本 発明 の 薬剤 は 、 **糖尿__病** 、 より 詳細に は 1 型 又は 2 型 **糖尿__病** の 患者 の 場合 に 血 中 **グルコース__濃度** を **正常__血糖__レベル** まで より 効果 的に 適合 さ せる こと を 可能に する 。
在 该 方法 中 ， 能够 得到 从 **心脏__周期** 内 的 **心收缩__期** 到 心 舒张 期 之间 的 血液 移动 的 1 个 以上 的 图像 。	この 方法 に おいて は 、 **心臓__周期** 内 の **心__収縮__期** から **心__拡張__期** まで の 間 の **血液__移動** の 1 枚 以上 の 画像 が 得 られる 。
于本 发明 的 **加热__烹饪** 用 **油脂__组成__物** 中 ， 除了 上述 聚 甘油 **脂肪__酸酯** 以外 ， 亦 可 包含 **植物__油脂** 。	本 発明 の **加熱__調理** 用 **油脂__組成__物** に は 、 上記 ポリグリセリン **脂肪__酸__エステル** の ほか 、 **植物__油脂** を 含む こと が できる 。

2.2 Bilingual Multi-word Term Extraction

2.2.1 Using Sampling-based Method To extract bilingual aligned multi-word terms, we use the open source implementation of the sampling-based alignment method, Anymalign [8], to perform word-to-word alignment[††††] or, better said, token-to-token alignment from the above monolingual terms based re-tokenized Chinese–Japanese training corpus. Among the aligned words or tokens, we find the

[††††] This is done by the option -N 1 on the command line.

Table 3. Extraction of bilingual aligned multi-word terms in both languages at the same time by setting a threshold of 0.6. ○ and × show the bilingual multi-word term or one side is multi-word term alignment that are kept or excluded. √ and ∗ show the extracted or cannot be extracted multi-word term pairs are correct or incorrect alignments by checking manually.

Extract or not	Correct or not	Chinese	Japanese	Meaning	$P(t\|s)$	$P(s\|t)$
○	√	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
○	√	加热__烹饪	加熱__調理	'cooking'	1.000000	0.814815
○	√	油脂__组成__物	油脂__組成__物	'fat composition'	1.000000	1.000000
○	√	脂肪__酸酯	脂肪__酸__エステル	'fatty acid ester'	1.000000	0.983333
○	√	植物__油脂	植物__油脂	'vegetable oil and fat'	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
○	∗	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

multi-word term to multi-word term alignments between Chinese and Japanese by using the markers. We filter these aligned multi-word candidate terms by setting some threshold P for the translation probabilities in both directions. The translation probabilities $P(t|s)$ and $P(t|s)$ are computed using the maximum possibility estimation from the co-occurrence frequencies that are consistent with the word alignment in the translation table:

$$P(ja|zh) = \frac{P(zh, ja)}{P(zh)} = \frac{C(ja \leftrightarrow zh)}{C(zh)} \quad (2)$$

$$P(zh|ja) = \frac{P(ja, zh)}{P(ja)} = \frac{C(zh \leftrightarrow ja)}{C(ja)} \quad (3)$$

In the equations, $C(x)$ denotes the number of occurences of the word or phrases x in the re-tokenized Chinese–Japanese training corpus, and $C(x \leftrightarrow y)$ is the number of co-occurrences of x and y in the re-tokenized Chinese–Japanese training corpus.

Table 3 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 last three lines in Table 3.

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 length with 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 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 Tab. 4).

Table 4. 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%)

2.2.2 Using kanji-hanzi Conversion Method Table 3 leads to the observation that some correctly aligned bilingual terms cannot be extracted by using the methods we described in Sec. 2.2.1. Such examples of terms are given in Tab. 5.

Table 5. Examples of discarded bilingual aligned multi-word terms by setting threshold P.

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

All such examples are cases where the terms in Japanese (or in 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 5, the pair "添加剂 (Chinese) 添加__剤 (Japanese)" ('additive') is supported by the kanji-hanzi conversion of the last element "剂-剤" ('agent').

Consequently, we keep the alignments where either one side is a multi-word term after token-to-token alignment, we

Table 6. 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

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). In this way, by comparing Converted-Ja with the Chinese term (Zh), if a converted Japanese term is equal to its corresponding Chinese term in each character, we can extract more reliable Chinese–Japanese bilingual aligned multi-word terms.

Table 6 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 compare with Chinese directly without any conversion;
- The Japanese 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 mapping 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[†]. 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[††] 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[†††] which consists of simplified hanzi and kanji pairs. Table 7 shows the results of extracted bilingual multi-word terms by kanji-hanzi conversion using these three sources of data.

2.3 Bilingual Multi-word Terms Used in SMT We combine the further filtered results with the kanji-hanzi conversion results to maximize the bilingual multi-word term extraction. In the procedure for building SMT systems, we re-tokenize the Chinese–Japanese training parallel corpus with the extracted bilingual multi-word terms by enforcing them to be considered as one token. We then train the Chinese–Japanese translation models on the re-tokenized training parallel corpus. A language model is trained with the Japanese corpus without re-tokenizing annotation. We remove the markers from the phrase tables before performing tuning and decoding in SMT experiments. We compare such systems with a standard baseline system.

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

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

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

3. Experiments and Results

3.1 Chinese and Japanese Data Used The Chinese–Japanese parallel sentences used in our experiments are randomly extracted from the Chinese–Japanese JPO Patent Corpus (JPC)[††††]. 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, 1,000 parallel sentences from the tuning part, and 1,000 from the test part. The scenario in our work is to deal with a limited corpus, we assume there is no larger corpus. Table 8 shows basic statistics on our data sets.

Section 3.2 will describe how many monolingual and bilingual multi-word terms will be extracted from the training part of these data. Section 3.3 and 3.4 will show the translation accuracy of SMT systems based on these data and compare the results with the baseline system.

3.2 Monolingual and Bilingual Multi-word Term Extraction We extract monolingual multi-word terms from a Chinese–Japanese training corpus of 100,000 lines as indicated in Tab. 8. We extract 81,618 monolingual multi-word terms for Chinese and 93,105 for Japanese respectively based on the linguistic pattern and C-value computation given in Sec. 2.1. We manually checked the precision of the extracted monolingual multi-word terms for Chinese and Japanese by sampling 1000 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. They are the first 80,000 monolingual multi-word terms with higher C-value in both languages.

Following the description given in Sec. 2.2.1, Table 9 gives the number of bilingual multi-word terms obtained for different thresholds P (translation probabilities), given in column (a) from the re-tokenized (with extracted monolingual multi-word terms) 100,000 lines of training corpus[†5]. We randomly extract 100 bilingual multi-word terms from each result and check the correspondence manually. The percentage of the good match terms is over 70%, when the threshold is greater than 0.4. Table 9 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

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

[†5] 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++.

Table 7. 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'	√
	肺癌	肺__癌	肺__癌	'lung cancer'	√
	免疫原	免疫__原	免疫__原	'immunogen'	√
	透析液	透析__液	透析__液	'dialyzate'	√
	数__密度	数__密度	数__密度	'number density'	√
By Traditional-Simplified Conversion	脉管	脈__管	脉__管	'vessel'	√
	肠壁	腸__壁	肠__壁	'intestinal wall'	√
	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	乘员__保护__方法	乗員__保護__方法	乘__保护__方法	'occupant protection method'	√
	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 8. 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 (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

by word-to-word alignment from Chinese–Japanese training corpus using kanji-hanzi conversion as described in Sec. 2.2.2. The number of the extracted multi-word terms using kanji-hanzi conversion combined with further filtering by constraints are given in Tab. 9 (column (a + b + c)). The percentage of the good match terms is over 80%, when the threshold is greater than 0.2. We obtained the highest percentage with 93% with threshold of 0.9 by combining of kanji-hanzi conversion and further filtering methods.

3.3 Translation Accuracy in BLEU We build several Chinese–Japanese training corpora re-tokenized with:

- several thresholds P for filtering (Tab. 9 (a))
- further filtering with several thresholds combined with kanji-hanzi conversion results (Tab. 9 (a +b + c))

We train several Chinese to Japanese SMT systems using the standard GIZA++/MOSES pipeline [13]. The Japanese corpus without re-tokenizing is used to train a language model using KenLM [14]. After removing markers 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 9 shows the evaluation of the results of Chinese to Japanese translation in BLEU scores [15]. Compared with the baseline system,

- 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 in all thresholds. We obtain 1.5 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).

3.4 Analysis of the Results We also compare a system based on a re-tokenized training corpus with further filtering results based on threshold of 0.6 combined with kanji-hanzi conversion results with a baseline system. We investigate the N (Chinese) $\times$ 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 translating). MOSES discards all entries which do not appear in the test set. In Tables 10 and 11, 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 have a significant increase compare with the baseline system.

We also investigate the distribution of the phrases actually used during the translation of the test set using traces of translation. Tables 12 and 13 show the distribution of phrases used during testing based on 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 the baseline system. From these tables, we can 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 9) and the analysis of the increase of potentially used and actually used phrase pairs reflect the impact of our method of re-tokenizing the training corpus with bilingual multi-word terms.

Figure 1 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=53.74) of the test sentence given in this

Table 9. Evaluation results in BLEU for Chinese to Japanese translation based on re-tokenized training corpus using different thresholds (a), the ratio of lengths + the components (b) and kanji-hanzi conversion (c).

Thresholds P	Filtering by thresholds P (a)					Filtering by thresholds P (a) + the ratio of lengths + the components (b) + kanji-hanzi conversion (c)						
	♯ 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)		
≥ 0.0	52,785	136,235	35%	32.63±1.15	0.095	48,239	63%	49,474	150,145	70%	33.15±1.09	0.001
≥ 0.1	31,795	114,287	52%	32.76±1.18	0.074	29,050	68%	30,516	131,997	78%	33.10±1.15	0.002
≥ 0.2	27,916	107,790	58%	32.57±1.10	0.142	25,562	75%	27,146	126,795	83%	33.05±1.11	0.001
Baseline	-	-	-	**32.38±1.16**	-	-	-	-	-	-	**32.38±1.16**	-
≥ 0.3	25,404	102,447	63%	33.07±1.13	0.002	23,321	78%	25,006	122,531	83%	33.21±1.10	0.001
≥ 0.4	23,515	97,488	72%	32.92±1.13	0.007	21,644	80%	23,424	118,674	84%	33.29±1.10	0.001
≥ 0.5	21,846	93,143	76%	33.05±1.11	0.001	20,134	85%	22,000	115,145	88%	33.38±1.12	0.001
≥ **0.6**	**20,248**	84,967	78%	**33.61±1.17**	0.001	18,691	88%	**20,679**	108,130	89%	**33.93±1.12**	0.001
≥ 0.7	18,759	84,908	79%	32.92±1.18	0.002	17,340	88%	19,460	104,298	90%	33.43±1.13	0.001
≥ 0.8	17,311	71,048	79%	33.34±1.14	0.001	16,001	89%	18,265	99,314	90%	33.41±1.14	0.001
≥ 0.9	15,464	59,567	80%	33.47±1.14	0.001	14,284	92%	16,814	90,660	93%	33.52±1.13	0.001

Table 10. 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 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	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 11. Distribution of the phrase table used 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	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

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., “定向 控制 模块” to “指向 性 制御 モジュール” (‘directivity control module’) in this example. 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’). In this example, re-tokenization of the training corpus with extracted bilingual multi-word terms induced a direct and exact translation.

4. Conclusion and Future Work

We presented an approach to improve the performance of Chinese–Japanese patent machine translation by re-tokenizing parallel training corpus with extracted bilingual multi-word 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 a token in their monolingual part 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 multi-word terms which could not be extracted using thresholds. This allowed us to extract multi-word 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 multi-word terms could be retrieved or reinforced thanks to the similarity between hanzi and kanji. We did not use any other additional corpus or lexicon in our 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. We obtained 1.5 BLEU point (p-value of 0.001) improvements compare with the baseline system (threshold of 0.6).

Table 12. Distribution of phrases used during testing based on: a C-value/sampling-based + hanzi/kanji conversion bilingual multi-word term extraction method for re-tokenizing training corpus (threshold with 0.6). The bold face numbers showing 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	**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 13. Distribution of phrases used during testing based on: 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

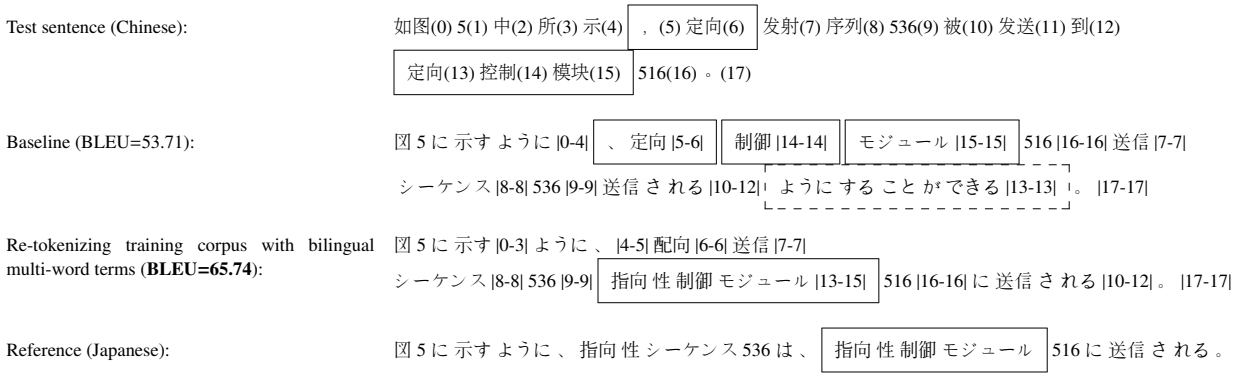

Fig. 1. Example of Chinese-to-Japanese translation improvement. The numbers in the parentheses show the position of the word in the test sentence. The numbers in the vertical lines show for the translation result (Japanese), the position of the n-gram used in the test sentence (Chinese).

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., 葡萄糖__浓度 (Chinese) グルコース__濃度 (Japanese) ('glucose concentration'), and the case where multi-word terms made up of kanzi/kanji are identified in one of the languages, but not in the other language. The second case mainly is due to different segmentation results in Chinese and Japanese, 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. 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'), 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'). In a future work, we intend to address this third case and expect further improvements in translation results.

References

(1) K. Frantzi, S. Ananiadou and H. Mima: "Automatic recognition of multi-word terms: the C-value/NC-value method", *International Journal on Digital Libraries*, Vol.3, No.2, pp.115–130 (2000)

(2) E. Milios, Y. Zhang, B. He and L Dong: "Automatic term extraction and document similarity in special text corpora", PACLING, pp.275–284 (2003)

(3) H. Mima and S. Ananiadou: "An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese", *Terminology*, Vol.6, No.2, pp.175–194 (2001)

(4) H. Nakagawa, H. Kojima and A. Maeda: "Chinese term extraction from web pages based on compound word productivity", SIGHAN Workshop on Chinese Language Processing, ACL, (2004)

(5) X. R. Fan, N. Shimizu and H. Nakagawa: "Automatic extraction of bilingual terms from a Chinese-Japanese parallel corpus", the 3rd. IUCS, pp.41–45 (2009)

(6) H. Mima, K. Frantzi and S. Ananiadou: "The C-value/Example-based Approach to the Automatic Recognition of Multi-Word Terms for Cross-Language Terminology", PRICAI, Joint Workshop on Cross Language Issues in Artificial Intelligence and Issues of Cross Cultural Communication,

pp.10–21 (1998)
(7) O. Furuse and H. Iida: "Incremental translation utilizing constituent boundary patterns", COLING'96, pp.412–417 (1996)
(8) A. Lardilleux and Y. Lepage: "Sampling-based multilingual alignment", RANLP, pp.214–218 (2009)
(9) C. L. Goh, M. Asahara, and Y. Matsumoto: "Building a Japanese-Chinese dictionary using kanji/hanzi conversion", IJCNLP, pp.670—681 (2005)
(10) C. L. Tan and M. Nagao: "Automatic alignment of Japanese-Chinese bilingual texts", *IEICE Transactions on Information and Systems*, E78-D(1):68—76 (1995)
(11) C. Chu, T. Nakazawa and S. Kurohashi: "Chinese Characters Mapping Table of Japanese, Traditional Chinese and Simplified Chinese", LREC2012, pp.2149-2152 (2012)
(12) C. Chu, T. Nakazawa, D. Kawahara and S. Kurohashi: "Chinese-Japanese machine translation exploiting Chinese characters" *ACM Transactions on Asian Language Information Processing (TALIP)*, 12-4:6 (2013)
(13) P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, and et al: "Moses: Open source toolkit for statistical machine translation", the 45th Annual Meeting of the ACL, pp.177–180 (2007)
(14) K. Heafield: "KenLM: Faster and smaller language model queries", the Sixth Workshop on SMT, pp.187–197 (2011)
(15) P. Kishore, R. Salim, W. Todd and W. J. Zhu: "BLEU: a method for automatic evaluation of machine translation", ACL, pp.311–318 (2002)

Wei Yang (Non-member) She received the Master Degree in 2012 from Waseda University, graduate school of Information, Production and Systems. During her Master Course, her research interests in combining several automatic techniques to build Chinese-Japanese lexicon from freely available resources and make them free available for users and researchers in Natural Language Processing and Machine Translation. She is currently a Ph.D. candidate at the Waseda University, graduate school of Information, Production and Systems. Her research interests are in Natural Language Processing, Machine Translation, especially between Chinese and Japanese.

Yves Lepage (Non-member) He received his Ph.D. degrees from Grenoble university, France, in GETA. He worked for ATR labs, Japan, as an invited researcher and a senior researcher until 2006. He joined Waseda University, graduate school of Information, Production and Systems in April 2010. His research interests are in Natural Language Processing, Machine Translation, and in particular Example-Based Machine Translation. He is a member of the Japanese Natural Language Processing Association. He is a member of the French Natural Language Processing Association, ATALA, and editor-in-chief of the French journal on Natural Language Processing, TAL.