Extraction of Bilingual Technical Terms for Chinese–Japanese Patent Translation

Wei Yang and **Jinghui Yan** and **Yves Lepage**
Graduate School of IPS, Waseda University,
2-7 Hibikino, Wakamatsu Kitakyushu Fukuoka, 808-0135, Japan
{kevinyoogi@akane,jess256@suou}.waseda.jp, yves.lepage@waseda.jp

Abstract

The translation of patents or scientific papers is a key issue that should be helped by the use of statistical machine translation (SMT). In this paper, we propose a method to improve Chinese–Japanese patent SMT by premarking 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. We use the sampling-based alignment method to identify aligned terms and set some threshold on translation probabilities to select the most promising bilingual multi-word terms. We pre-mark a Chinese–Japanese training corpus with such selected aligned bilingual multi-word terms. We obtain the performance of over 70% precision in bilingual term extraction and a significant improvement of BLEU scores in our experiments on a Chinese–Japanese patent parallel corpus.

1 Introduction

China and Japan are producing a large amount of scientific journals and patents in their respective languages. The World Intellectual Property Organization (WIPO) Indicators[1] 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[2]. Making Chinese patents available in Japanese, and Japanese patents available and in Chinese is a key issue for increased economical development in Asia.

In recent years, Chinese–Japanese machine translation of patents or scientific papers has made rapid progress with the large quantities of parallel corpora provided by the organizers of the Workshop on Asian Translation (WAT)[3] [4]. In the "patents subtask" of WAT 2015, in (Sonoh and Kinoshita, 2015), a Chinese to Japanese translation system is described that achieves higher BLEU scores by combination of results between Statistical Post Editing (SPE) based on their rule-based translation system and SMT system equipped with a recurrent neural language model (RNNLM).

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. Still, the availability of patent bilingual corpora between Chinese and Japanese in certain domains is a problem.

In this paper, we propose a simpler way to improve Chinese to Japanese phrase-based machine translation quality based on a small size of available bilingual patent corpus, without exploiting extra bilingual data, or using a third language, with

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

[2] http://www.ipwatchdog.com/2013/04/04/chinas-great-leap-forward-in-patents/id=38625/

[3] http://orchid.kuee.kyoto-u.ac.jp/WAT/WAT2014/index.html

[4] http://orchid.kuee.kyoto-u.ac.jp/WAT/

Proceedings of NAACL-HLT 2016, pages 81–87,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

no complex approach. Patents or scientific papers contain large amounts of domain-specific terms in words or multi-word expressions. Monolingual or bilingual term extraction is an important task for the fields of information retrieval, text categorization, clustering, machine translation, etc. There exist work on monolingual or bilingual term extraction in different languages. In (Kang et al., 2009), multi-word terms in Chinese in the information technology (IT) domain and the medicine domain are extracted based on the integration of Web information and termhood estimation. Frantzi et al. (2000) describes a combination of linguistic and statistical information method (C-value/NC-value) for the automatic extraction of multi-word terms from English corpora. In (Mima and Ananiadou, 2001), it was 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.

Some work consider the case of bilingual term extraction. In (Fan et al., 2009), Chinese–Japanese multi-word terms are extracted by re-segmenting the Chinese and Japanese bi-corpus and combining multi-word terms as one single word based on extracted monolingual terms. The word alignments containing terms are smoothed by computing the associations between pairs of bilingual term candidates.

In this paper, we propose a method to extract Chinese–Japanese bilingual multi-word terms by extracting Chinese and Japanese monolingual multi-word terms using a linguistic and statistical technique (C-value) (Frantzi et al., 2000) and the sampling-based alignment method (Lardilleux and Lepage, 2009) for bilingual multi-word term alignment. We filter the aligned candidate terms by setting thresholds on translation probabilities. We perform experiments on the Chinese–Japanese JPO patent corpus of WAT 2015. We pre-mark the extracted bilingual terms in the Chinese–Japanese training corpus of an SMT system. We compare the translation system which uses our proposed method with a baseline system. We obtain a significant improvement in translation accuracy as evaluated by BLEU (Papineni et al., 2002).

The paper is organized as follows: in Section 2, we introduce the experimental data sets used in our experiments. Section 3 gives our proposed method to extract Chinese–Japanese bilingual multi-word terms using the C-value and the sampling-based alignment method. In Section 4, we describe our experiments and their results based on the data introduced in Section 2, and an analysis of the experimental results. Section 5 gives the conclusion and discusses future directions.

2 Chinese and Japanese Data Used

The Chinese–Japanese parallel sentences used in this paper are randomly extracted from the Chinese–Japanese JPO Patent Corpus (JPC)[5]. 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. Table 1 shows the basic statistics on our experimental data sets.

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

Table 1: Statistics on our experimental data sets (after tokenizing and lowercasing). Here 'mean ± std.dev.' gives the average length of the sentences in words.

In Section 3, monolingual and bilingual multi-word terms will be extracted from the training data. In Section 4, these data (train, tune and test) will be used in the baseline SMT system.

3 Bilingual Multi-word Term Extraction

This section presents our bilingual multi-word term extraction method that uses C-value (Frantzi et al., 2000) combined with the sampling-based alignment

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

method (Lardilleux and Lepage, 2009). We also describe how we use these extracted bilingual multi-word terms in SMT experiments.

3.1 Monolingual Multi-word Term Extraction Using the C-value Approach

The C-value is a commonly used domain-independent method for multi-word term extraction. This method has a linguistic part and a statistical part. The linguistic part constrains the type of terms extracted. In our experiments, we extract multi-word terms which contain a sequence of nouns or adjectives followed by a noun for both Chinese and Japanese. This linguistic pattern can be written as follows using a regular expression[6]:

$(Adjective|Noun)^{+}\ Noun$

The segmenter and part-of-speech tagger that we use are the Stanford parser[7] for Chinese and Juman[8] for Japanese. Examples of outputs are shown in Table 2.

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

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

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 Chinese–Japanese training corpus. Then, we mark the extracted monolingual multi-word terms in the corpus by enforcing them to be considered as one token (aligned with markers).

3.2 Bilingual Multi-word Term Extraction Using Sampling-based Method

To extract bilingual multi-word terms, we use the open source implementation of the sampling-based approach, Anymalign (Lardilleux and Lepage, 2009), to perform phrase alignment from the above marked Chinese–Japanese training corpus. We filter out any alignment ($N \times M$-grams) that is greater than 1, to obtain only word-to-word alignments[9]. In our experiments, we identify the multi-word term to multi-word term alignments between Chinese and Japanese by using the markers. We filter the aligned multi-word candidate terms by setting some threshold P for both translation probabilities of term alignments ($0<P\leq1$).

3.3 Bilingual Multi-word Terms Used in SMT Experiments

We train the Chinese–Japanese translation models on the training parallel pre-marked corpus with the extracted filtered aligned bilingual multi-word terms. A language model is trained with the original Japanese corpus without pre-marking annotation. We remove the markers from obtained phrase tables before performing tuning and decoding processes. We compare such a systems with a standard baseline system.

4 Experiments and Results

We extract monolingual multi-word terms from a Chinese–Japanese training corpus of 100,000 lines as indicated in Table 1 (Section 2). Table 3 shows the number of monolingual multi-word terms extracted in Chinese and Japanese respectively using C-value and the linguistic pattern given in Section 3.1. The extracted monolingual multi-word terms were ranked by decreasing order of C-values. We mark the training corpus with the same size 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.

Follow the description given in Section 3.2. Table 4 gives the number of bilingual multi-word terms obtained for different thresholds from the marked 100,000 training corpus. We randomly extract 100 bilingual multi-word terms respectively and roughly

[6] 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 annotated corpora that we use.

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

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

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

Chinese or Japanese sentences	Extracted monolingual terms
Chinese: 完全♯AD 布置♯VV 在♯P **环形♯JJ 间隙♯NN** 中♯LC 。 ♯PU	环形　间隙
Japanese: 完全に/形容詞 この/指示詞 **環状/名詞 隙間/名詞** 内/接尾辞 に/助詞 配置/名詞 さ/動詞 れる/接尾辞 。 /特殊	環状　隙間
English meaning: 　　‘Completely arranged in the annular gap.’	‘annular gap’

Table 2: Examples of outputs on the tags used based on the linguistic pattern.

Language	♯ of Multi-word terms
Chinese	81,618
Japanese	93,105

Table 3: The number of monolingual multi-word terms extracted from Chinese–Japanese training corpus using C-value.

Extract or not	Chinese	Japanese
○	控制__电路 ‘*control circuit*’	制御__回路
×	核酸 ‘*nucleic acid*’	核__酸
×	粘接剂 ‘*adhesive*’	接着__剤
○	信息__处理__装置 ‘*information-processing device*’	情報__処理__装置
○	发光__二极管__元件 ‘*light emitting diode element*’	発光__ダイオード__素子
○	压力__传感器 ‘*pressure sensor*’	圧力__センサ
×	存储器__控制器 ‘*memory controller*’	メモリコントローラ
×	枢轴__板 ‘*pivot plate*’	ピボットプレート

Table 5: Extraction of bilingual multi-word terms in both languages at the same time. ○ and × show the bilingual multi-word term alignment that are kept or excluded.

check how well they correspond/match manually. The precision (good match) of the extracted bilingual multi-word terms is over 70%, while threshold becomes greater than 0.4. Table 5 shows sample of bilingual multi-word terms we extracted.

4.1 Translation Accuracy in BLEU

We pre-mark the original Chinese–Japanese training corpus with the extracted bilingual multi-word terms filtering by several thresholds (Table 4) and train several Chinese to Japanese SMT systems using the standard GIZA++/MOSES pipeline (Koehn et al., 2007). The un-pre-marked (original) Japanese corpus is used to train a language model using KenLM (Heafield, 2011). 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 pre-marked or not with bilingual multi-word terms filtered by a given threshold. Table 4 shows the evaluation of the results of Chinese to Japanese translation in BLEU scores (Papineni et al., 2002). Compared with the baseline system, we obtain significant improvements as soon as the threshold becomes greater than 0.3. A statistically significant improvement of one BLEU point (p-value is 0.001) is observed when the threshold is greater than 0.6. In that case, the training corpus is pre-marked with roughly 20,000 bilingual multi-word terms.

4.2 Analysis of the Content of Phrase Tables

We further compare a system based on a pre-marked training corpus using bilingual multi-word terms (threshold of 0.6) with a baseline system. We investigate the N (Chinese) × M (Japanese)-grams distribution in the reduced phrase tables[10] used in translation. In Tables 6 and 7, the statistics (Chinese→Japanese) show that the total number of potentially useful phrase pairs used in translation with the pre-marked corpus is larger than that of the baseline system. Considering the correspondence between lengths in Chinese–Japanese patent translation, we compare the number of entries, the number of phrase pairs with different lengths (like 2 (zh) × 1 (ja), 2 (zh) × 3 (ja), 2 (zh) × 4 (ja) and 3 (zh) × 4 (ja)) and observe a significant increase for these categories.

We also investigate the number of phrase alignments which the Chinese source language part containing multi-word terms in the reduced phrase table obtained when pre-marking the training corpus. There exists 8,940 phrase alignments in this case. A sample is shown in Table 8. Compared with the reduced phrase table used in the baseline system, there exist 2,503 additional phrase alignments. They con-

[10] The phrase table only contains the potentially useful phrase alignments used in the translation of the test set.

Thresholds $P(t\|s)$ and $P(s\|t)$	$\sharp$ of bilingual multi-word terms	Good match	BLEU	p-value
$\geq$ 0.0	52,785	35%	32.44±1.07	0.197
$\geq$ 0.1	31,795	52%	32.23±1.18	0.062
$\geq$ 0.2	27,916	58%	32.00±1.16	0.072
Baseline	**Baseline**	**Baseline**	**32.35±1.15**	**Baseline**
$\geq$ 0.3	25,404	63%	33.08±1.12	0.004
$\geq$ 0.4	23,515	72%	32.77±1.15	0.027
$\geq$ 0.5	21,846	76%	33.02±1.14	0.007
$\geq$ **0.6**	**20,248**	78%	**33.32±1.15**	0.001
$\geq$ 0.7	18,759	79%	32.85±1.19	0.006
$\geq$ 0.8	17,311	79%	33.25±1.06	0.001
$\geq$ 0.9	15,464	80%	33.20±1.15	0.002

Table 4: Evaluation results in BLEU for Chinese to Japanese translation based on pre-marked training corpus with bilingual multi-word terms using different thresholds, tools used are Giza++/Moses 2.1.1, KenLM.

Chinese		Japanese
n型 半 导__体层	|||	n 型
pdu__大小	|||	pdu サイズ
新__数据	|||	新しい__データ
白__平衡	|||	ホワイト__バランス
x__射线	|||	x線
x__射线	|||	x線 が
所 需 的 构成__要素	|||	に 必要な 構成__要素
个 碳__原子 的	|||	個 の 炭素__原子 の
接触__孔	|||	コンタクト__ホール
将 反应__混合物	|||	反応 混合__物 を
在 玻璃__基板	|||	ガラス__基板
的 视频__信号	|||	ビデオ__信号
负 极活性__物质	|||	負__極__物質
控制__部	|||	制御__部
旋转__量	|||	回転__量
新__数据	|||	新しい__データ
白__平衡	|||	ホワイト__バランス

Table 8: Sample of phrase alignments for which the source language part (Chinese) contains multi-word terms in the reduced phrase table. We show multi-word terms as one token in the phrase table aligned with markers.

tain multi-word terms that did not exist in the reduced phrase table of the baseline system. Table 9 shows examples of more potentially useful phrase alignments obtained with our proposed method.

5 Conclusion and Future Work

We presented an approach to improve Chinese–Japanese patent machine translation performance by pre-marking the parallel training corpus with bilingual multi-word terms. We extracted monolingual multi-word terms from each monolingual part of a corpus by using the C-value method. We used the sampling-based alignment method to align the marked parallel corpus with monolingual multi-word terms and only kept the aligned bilingual multi-word terms by setting thresholds in both directions. We did not use any other additional corpus or lexicon. The results of our experiments indicate that the bilingual multi-word terms extracted have over 70% precision (the thresholds $P \geq 0.4$). Pre-marking the parallel training corpus with these terms led to statistically significant improvements in BLEU scores (the thresholds $P \geq 0.3$).

In this work, we considered only the case where multi-word terms can be found in both languages at the same time, e.g., 半导体__芯片 (zh) 半導体__チップ (ja) ‘semiconductor chip’. However, we found many cases where a multi-word term is recognized in one of the languages, while the other side is not recognized as a multi-word term, although they may be correct translation candidates. This mainly is due to different segmentation results in Chinese and Japanese. E.g., 压缩机 (Chinese) 圧縮__機 (Japanese) ‘compressor’, and 流程图 (Chinese) フロー__チャート (Japanese) ‘flow chart’. In a future work, we thus intend to address this issue and expect further improvements in translation results. We also intend to do experiments with our proposed method using a larger size of experimental training data.

References

Xiaorong Fan, Nobuyuki Shimizu, and Hiroshi Nakagawa. 2009. Automatic extraction of bilingual terms from a Chinese-Japanese parallel corpus. In *Proceed-*

ings of the 3rd International Universal Communication Symposium, pages 41–45. ACM.

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.

Wei Kang, Zhifang Sui, and Yao Liu. 2009. Research on automatic Chinese multi-word term extraction based on integration of Web information and term component. In *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*, volume 3, pages 267–270. IET.

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.

Xianhua Li, Yao Meng, and Hao Yu. 2012. Improving Chinese-to-Japanese patent translation using English as pivot language. In *26th Pacific Asia Conference on Language, Information and Computation (PACLIC 26)*, pages 117–126.

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.

Satoshi Sonoh and Satoshi Kinoshita. 2015. Toshiba MT system description for the WAT2015 workshop. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 48–53.

		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 6: Distribution of the reduced phrase table 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 compared 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	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	1	0	0	0	0	0	0	0	0	0
	total	47471	128221	122849	77837	43457	22927	11857	0	0	454619

Table 7: Distribution of the reduced phrase table of baseline system based on GIZA++/Moses 2.1.1.

<table>
<tr><th>Baseline</th><th>English meaning</th></tr>
<tr><td>传输 信道 的 ||| 伝送 路 の
位置 信息 ||| 位置
位置 信息 ||| 位置 の 情報
位置 信息 ||| 位置 は</td><td>transmission channel ||| transmission channel
location information ||| location
location information ||| location information
location information ||| location (with a auxiliary word ‘は’)</td></tr>
<tr><td><b>Pre-marked training corpus</b></td><td></td></tr>
<tr><td>传输 信道 的 ||| 伝送 路 の
<b>传输 信道 的 ||| 、 伝送 チャネル の</b>
位置 信息 ||| 位置
位置 信息 ||| 位置 の 情報
位置 信息 ||| 位置 は
<b>位置 信息 ||| 位置 の 情報 である</b>
<b>位置 信息 ||| その 位置 情報</b>
<b>位置 信息 ||| 位置 情報 は</b></td><td>transmission channel ||| transmission channel
<b>transmission channel ||| , transmission channel of (another way of saying ‘transmission channel’ in Japanese)</b>
location information ||| location
location information ||| location information
location information ||| location (with an auxiliary word ‘は’)
<b>location information ||| location information (with an auxiliary ‘である’)</b>
<b>location information ||| that location information</b>
<b>location information ||| location information (with an auxiliary word ‘は’)</b></td></tr>
</table>

Table 9: Samples of phrase alignments in reduced Chinese→Japanese phrase tables. Alignments given in bold face are additional phrase alignments compared with the baseline systems.