Collecting Bilingual Technical Terms from Patent Families of Character-Segmented Chinese Sentences and Morpheme-Segmented Japanese Sentences

Zi Long
Takehito Utsuro
Grad. Sch. Sys. & Inf. Eng., University of Tsukuba, Tsukuba, 305-8573, Japan

Tomoharu Mitsuhashi
Japan Patent Information Organization, 4-1-7, Tokyo, Koto-ku, Tokyo, 135-0016, Japan

Mikio Yamamoto
Grad. Sch. Sys. & Inf. Eng., University of Tsukuba, Tsukuba, 305-8573, Japan

Abstract

In manual translation of patent documents, a technical term bilingual lexicon is inevitable for a translator to efficiently translate patent documents. Dong et al. (2015) proposed a method of generating bilingual technical term lexicon from morpheme-segmented parallel patent sentences. The proposed method estimates Japanese-Chinese translation of technical terms using the phrase translation table of a statistical machine translation model. The procedure of generating bilingual technical term lexicon consists of the following four steps: (1) extracting Japanese technical terms from Japanese side of parallel patent sentences, (2) collecting all the sentences that contain the extracted Japanese term, (3) generating Chinese translation of the Japanese technical term referring to the phrase translation table of a statistical machine translation model, and (4) applying the Support Vector Machines (SVMs) to the task of identifying bilingual technical terms. In this paper, we segment the Chinese sentences into characters instead of segmenting them into morphemes as in Dong et al. (2015), and represent Japanese-Chinese patent families in terms of character-segmented Chinese sentences and morpheme-segmented Japanese sentences. Then, to those Japanese-Chinese patent families, we apply the framework (Dong et al., 2015) of identifying bilingual technical terms. As a result, we achieve the performance of over 90% precision with the condition of more than or equal to 60% recall.

1 Introduction

For both high quality machine and human translation, a large scale and high quality bilingual lexicon is the most important key resource. Since manual compilation of bilingual lexicon requires plenty of time and huge manual labor, in the research area of knowledge acquisition from natural language text, automatic bilingual lexicon compilation have been studied. Techniques invented so far include translation term pair acquisition based on statistical co-occurrence measure from parallel sentences (Matsumoto and Utsuro, 2000), translation term pair acquisition from comparable corpora (Fung and Yee, 1998; Bouamor et al., 2013; Morin and Hazem,

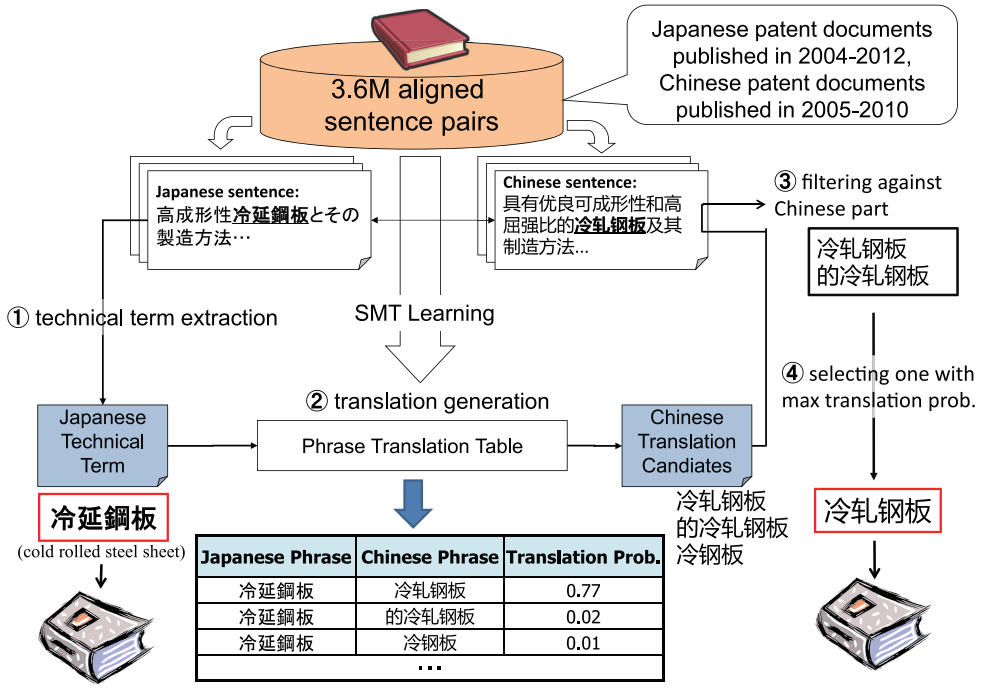

Figure 1: The Procedure of Translation Estimation of a Technical Term using a Phrase Translation Table and a Parallel Sentence Pair

2014), compositional translation generation based on an existing bilingual lexicon for human use (Tonoike et al., 2006), translation term pair acquisition by collecting partially bilingual texts through the search engine (Huang et al., 2005; Lin et al., 2008), and translation term pair acquisition from multilingual resources included in Wikipedia (Erdmann et al., 2009).

Among those efforts of acquiring bilingual lexicon from text, Dong et al. (2015) proposed to acquire Japanese-Chinese technical term translation lexicon from the phrase translation tables, which are trained by a phrase-based statistical machine translation (SMT) model with parallel sentences automatically extracted from patent families. One of the major advantages of the proposed approach is that the resource we utilize in this approach is Japanese-Chinese patent families, which continue to be published every year. The procedure of generating bilingual technical term lexicon consists of the following four steps: (1) extracting Japanese technical terms from Japanese side of parallel patent sentences, (2) collecting all the sentences that contain the extracted Japanese term, (3) generating Chinese translation of the Japanese technical term referring to the phrase translation table of a statistical machine translation model, and (4) applying the Support Vector Machines (SVMs) (Vapnik, 1998) to the task of identifying bilingual technical terms. In this paper, we segment the Chinese sentences into characters instead of segmenting them into morphemes as in (Dong et al., 2015), and represent Japanese-Chinese patent families in terms of character-segmented Chinese sentences and morpheme-segmented Japanese sentences. Then, to those Japanese-Chinese patent families, we apply the framework (Dong et al., 2015) of identifying bilingual technical terms. As a result, we achieve the performance of over 90% precision with the condition of more than or equal to 60% recall.

2 Japanese-Chinese Parallel Patent Documents

Japanese-Chinese parallel patent documents are collected from the Japanese patent documents published by the Japanese Patent Office (JPO) in 2004-2012 and the Chinese patent documents published by State Intellectual Property Office of the People's Republic of China (SIPO) in 2005-2010. From them, we extract 312,492 patent families, and the method of Utiyama and Isahara (2007) is applied[1] to the text of those patent families, and Japanese and Chinese sentences are aligned. In this paper, we use 3.6M parallel patent sentences with the highest scores of sentence alignment[2].

3 Phrase Translation Table of an SMT Model

As a toolkit of a phrase-based SMT model, we use Moses Koehn et al. (2007) and apply it to the whole 3.6M parallel patent sentences. Before applying Moses, Japanese sentences are segmented into a sequence of morphemes by the Japanese morphological analyzer MeCab[3] with the morpheme lexicon IPAdic[4], while Chinese sentences are segmented by characters. In this procedure of Chinese segmentation, a consecutive sequence of numbers as well as a consecutive sequence of alphabetical characters are segmented into a token.

As the result of applying Moses[5], we have a phrase translation table in the direction of Japanese to Chinese translation, consisting of 268M translation pairs with 194M Japanese phrases with Japanese to Chinese phrase translation probabilities $P(p_C \mid p_J)$ of translating a Japanese phrase p_J into a Chinese phrase p_C. For each Japanese phrase, those multiple translation candidates in the phrase translation table are ranked in descending order of Japanese to Chinese phrase translation probabilities.

4 Translation Estimation of a Technical Term using a Phrase Translation Table and a Parallel Sentence Pair

Figure 1 shows the procedure of estimating Chinese translation of a Japanese technical term using a phrase translation table and a parallel sentence pair. The phrase translation table is first referred to when identifying a bilingual technical term pair, given a parallel sentence pair $\langle S_J, S_C \rangle$ and a Japanese technical term t_J. Given a parallel sentence pair $\langle S_J, S_C \rangle$ containing a Japanese technical term t_J, Chinese translation candidates collected from the phrase translation table are matched against the Chinese sentence S_C of the parallel sentence pair. Among those found in S_C, $\hat{t}_C$ with the largest translation probability $P(t_C \mid t_J)$ is selected and the bilingual technical term pair $\langle t_J, \hat{t}_C \rangle$ is identified.

5 Translation Estimation by SVM using Features extracted from Multiple Parallel Patent Sentences

5.1 Selecting Japanese Technical Terms for Evaluation

When selecting Japanese technical terms for evaluation, we first extract 1.2M noun phrases from the 3.6M parallel patent sentences[6]. Next, we divide the set of all the Japanese noun phrases

[1] Here, we used a Japanese-Chinese translation lexicon consisting of about 170,000 Chinese head words.

[2] The 3.6M parallel patent sentences used in this paper are the same as those used in (Dong et al., 2015)

[3] `http://mecab.sourceforge.net/`

[4] `http://sourceforge.jp/projects/ipadic/`

[5] We set the upper bound of the numbers of the morphemes of Japanese phrases as well as the characters of Chinese phrases as 15.

[6] Those noun phrases are extracted by simply concatenating a sequence of morphemes whose parts-of-speech are either noun, prefix, suffix, unknown word, number, or alphabet. Here, morphemes sequence starting with certain

Table 1: Rates of Positive Examples (# of positive examples / # of positive and negative examples) for Each Frequency Range of Japanese Technical Term Frequency (jf) and Japanese-Chinese Co-occurrence Frequency (jcf)

	jf=1	2≦jf≦5	6≦jf≦10	11≦jf≦15	16≦jf≦20	21≦jf≦30	31≦jf≦50	51≦jf≦100	101≦jf≦200	201≦jf≦500	501≦jf≦1,000	1,001≦jf≦10,000	10,001≦jf	Total
jcf=1	37/62 =59.7%	20/26 =77.0%	12/19 =63.2%	20/33 =60.7%	21/37 =56.8%	11/20 =55.1%	24/40 =60.0%	19/49 =38.8%	25/70 =35.8%	22/76 =29.0%	36/121 =29.8%	8/69 =11.6%	1/28 =3.6%	256/650 =39.4%
2≦jcf≦5		55/58 =94.9%	24/28 =85.8%	12/15 =80.0%	16/24 =66.7%	14/23 =60.9%	22/32 =68.8%	21/42 =50.0%	28/55 =51.0%	23/55 =41.9%	22/86 =25.6%	34/79 =43.1%	1/28 =3.6%	272/525 =51.9%
6≦jcf≦10			41/43 =95.4%	16/16 =100.0%	14/16 =87.5%	14/16 =87.5%	16/18 =88.9%	9/15 =60.0%	15/25 =60.0%	12/19 =63.2%	23/39 =59.0%	14/31 =45.2%	1/17 =5.9%	175/255 =68.7%
11≦jcf≦15				38/40 =95.0%	13/14 =92.9%	9/9 =100.0%	6/7 =85.8%	4/6 =66.7%	8/8 =100.0%	10/15 =66.7%	7/10 =70.0%	6/17 =35.3%	1/6 =16.7%	102/132 =77.3%
16≦jcf≦20					31/31 =100.0%	7/7 =100.0%	8/8 =100.0%	1/2 =50.0%	7/8 =87.5%	4/6 =66.7%	8/9 =88.9%	7/11 =63.7%	1/4 =25.1%	74/86 =86.1%
21≦jcf≦30						32/32 =100.0%	6/8 =75.0%	3/4 =75.0%	6/10 =60.0%	5/7 =71.5%	6/9 =66.7%	2/8 =25.1%	1/5 =20.1%	61/83 =73.5%
31≦jcf≦50							25/26 =96.2%	15/17 =88.3%	7/9 =77.8%	7/11 =63.7%	8/10 =80.0%	3/5 =60.0%	3/13 =23.1%	68/91 =74.8%
51≦jcf≦100								24/25 =96.0%	21/21 =100.0%	7/9 =77.8%	8/8 =100.0%	6/6 =100.0%	4/14 =28.6%	70/83 =84.4%
101≦jcf≦200									25/25 =100.0%	17/18 =94.5%	5/6 =83.4%	7/8 =87.5%	4/8 =50.0%	58/65 =89.3%
201≦jcf≦500										23/24 =95.9%	10/10 =100.0%	6/6 =100.0%	3/4 =75.0%	42/44 =95.5%
501≦jcf≦1,000											31/31 =100.0%	5/5 =100.0%	1/1 =100.0%	37/37 =100.0%
1,001≦jcf≦10,000												24/24 =100.0%	6/7 =85.8%	30/31 =96.8%
10,001≦jcf													10/10 =100.0%	10/10 =100.0%
Total	37/62 =59.7%	75/84 =89.3%	77/90 =85.6%	86/104 =82.7%	95/122 =77.9%	87/107 =81.4%	107/139 =77.0%	96/160 =60.0%	142/231 =61.5%	130/240 =54.2%	164/339 =48.4%	122/269 =45.4%	37/145 =25.6%	1255/2092 =60.0%

Table 2: Numbers of Positive / Negative Examples in the Reference Set

positive	negative	total
1,255	837	2,092

Table 3: Positive Examples for Low / Middle / High Frequency Range of Japanese Technical Terms Frequency (jf) and Japanese-Chinese Co-occurrence Frequency (jcf)

	low frequency range ($1 \leqq jf \leqq 15$)	middle frequency range ($16 \leqq jf \leqq 100$)	high frequency range ($101 \leqq jf$)
low frequency range ($1 \leqq jcf \leqq 15$)	<水車/羽根/型/発電/装置, 水/轮/机/叶/片/型/发/电/装/置> (water wheel impeller blade type electric power generating apparatus) (jf=1, jcf=1)	<リンパ/球/増殖, 淋/巴/细/胞 /增/殖> (lymphocyte proliferation) (jf=21, jcf=10)	<スクリーン/装置, 筛/装/置> (screen device) (jf=104, jcf=14)
middle frequency range ($16 \leqq jcf \leqq 100$)		<高圧/水素/ガス, 高/压/氢/气> (high pressure hydrogen gas) (jf=37, jcf=37)	<窒素/化合/物, 氮/化/合/物> (nitorgen compound) (jf=112, jcf=97)
high frequency range ($101 \leqq jcf$)			<反応/混合/物, 反/应/混/合/物> (reactant mixture) (jf=30,620, jcf=29,860)

into 13 frequency ranges that are shown in Table 1 according to the frequency within the whole parallel patent sentences. Then, from each frequency range, we randomly select 90 Japanese noun phrases and manually judge whether each of the randomly selected Japanese noun phrases is appropriate as a technical term to be used in the evaluation of Chinese translation estimation. As a result, we select 578 Japanese technical terms for evaluation[7]. Among those manually excluded are such as those just fragments of longer technical terms as well as general noun phrases.

5.2 Developing a Reference Set of Bilingual Technical Terms

For each t_J of the 578 Japanese technical terms selected in the previous section, we first collect all the parallel sentence pairs $\langle S_J^i, S_C^i \rangle$ ($i = 1, 2, \ldots, n$) containing the given Japanese technical term t_J. From each of those parallel sentence pairs $\langle S_J^i, S_C^i \rangle$ ($i = 1, 2, \ldots, n$), at most one bilingual technical term pair $\langle t_J, \hat{t}_C \rangle$ is extracted according to the translation estimation procedure presented in Section 4. Thus, for an input Japanese technical term t_J, we obtain zero or more Japanese-Chinese technical term translation pairs $\langle t_J, \hat{t}_C^j \rangle$ ($j = 1, 2, \ldots, m\ (\leq n)$). In total, for the 578 Japanese technical terms selected in the previous section, we obtain 2,092 candidates of Japanese-Chinese technical term translation pairs. For each of the 2,092 candidates of technical term translation pairs, we manually judge whether it is correct technical term translation pair or not. Finally, as shown in Table 2, we obtain 1,255 correct translation pairs as positive examples, and the remaining 837 erroneous ones as negative examples[8] .We use the

prefixes or ending with certain suffixes are not appropriate as Japanese technical terms and are excluded. Those which include symbols or numbers are also excluded.

[7] Those 578 Japanese technical terms for evaluation are exactly the same as those used in (Dong et al., 2015).

[8] The Japanese-Chinese phrase translation table that is applied in the procedure of translation estimation is trained with patent families consisting of character-segmented Chinese sentences and morpheme-segmented Japanese sentences, and is different from that used in (Dong et al., 2015). As a result, we obtained a set of candidates of Japanese-

Table 4: Features for Identifying Bilingual Technical Terms by SVM

class	feature	definition
monolingual features	f_1: frequency of Japanese term	the ID (1∼13) of the frequency range of the Japanese technical term
	f_2: frequency of Chinese term	the ID (1∼13) of the frequency range of the Chinese technical term
bilingual features	f_3: translation probability	the translation probability $\mathrm{P}(t_C \mid t_J)$
	f_4: rank of Chinese translation candidates (descending order)	the rank of t_C with respect to the descending order of the conditional translation probability $\mathrm{P}(t_C \mid t_J)$
	f_5: co-occurrence frequency of bilingual technical term pairs	the ID (1∼13) of the co-occurrence frequency range of the Japanese-Chinese technical term pairs
	f_6: difference of the frequency of Japanese technical term and the co-occurrence frequency of bilingual technical term pairs	returns 1 if the difference of the frequency of the Japanese technical term and the co-occurrence frequency of bilingual technical terms is less than or equal to the upper bound (we use 105 as this upper bound in this paper), while returns 0 otherwise.
	f_7: number of Chinese translation candidates	the number of Chinese translation candidates for the Japanese technical term t_J
	f_8: rate of parallel sentences where phrase alignment is consistent with word alignments	$f_8 = \frac{\text{the number of parallel sentences where the phrase alignment is consistent with word alignments}}{\text{co-occurrence frequency of the Japanese-Chinese technical term pair}}$
	f_9: translation probability of compositional translation generation	translation probability when generating the Chinese translation candidate compositionally from constituents of the Japanese technical term

Table 5: Evaluation Results (%)

		precision	recall	F-measure
baseline		60.0	100	75.0
SVM	maximum precision	**93.9**	59.0	72.5
	maximum F-measure	80.6	87.2	83.7

set of those positive / negative examples as the reference set of Japanese-Chinese technical term translation pairs in the evaluation of this paper.

In Table 1, we also show the numbers of positive / negative examples for each pair of the 13 frequency ranges of Japanese technical term frequency (jf) and Japanese-Chinese co-occurrence frequency (jcf). Furthermore, Table 3 lists positive examples of Japanese-Chinese technical term translation pairs for each pair of low / middle / high frequency ranges of Japanese technical term frequency (jf) and Japanese-Chinese co-occurrence frequency (jcf).

5.3 Procedure of Applying SVM

In the training and testing of the classifier for identifying bilingual technical terms, we first divide the reference set of 2,092 bilingual technical terms into 10 subsets. Here, Japanese-Chinese bilingual technical term pairs which are generated from an identical Japanese term are collected into one subset, but are not separated into more than one subsets.

As a tool for learning SVMs, we use TinySVM[9]. As the kernel function, we use the polynomial (2nd order) kernel. In the training of SVMs, we use 8 subsets out of the whole 10 subsets. In the tuning of SVMs classifier, we regard the distance from the separating hyperplane to each test instance as a confidence measure and tune the lower bound of the confidence with one of the remaining two subsets. We consider tuning instances satisfying the confidence measure over a certain lower bound only as positive samples. Here, we tune the lower bound in two ways: i.e, for maximizing precision while keeping recall more than or equal to 60% [10], and for maximizing F-measure. In the testing, we test the trained classifier against another one of the remaining two subsets, where we return test instances satisfying the confidence measure over the lower bound only as positive samples. Finally, we repeat this procedure of training / tuning / testing 10 times, and average the 10 results of test performance.

5.4 Features

Table 4 lists all the features used for training and testing of SVMs for identifying Japanese-Chinese technical term translation pairs. Features are roughly divided into two types: those of the first type f_1 and f_2 are monolingual features, while those of the second type $f_2, \cdots, f_9$ are bilingual features which represent various characteristics of the input bilingual technical term pairs.

Chinese technical term translation pairs, which is different from that used in (Dong et al., 2015). While over 95% of those correct technical term translation pairs are the same as those used in (Dong et al., 2015), only about 55% of erroneous ones are the same as those used in (Dong et al., 2015).

[9] `http://chasen.org/~taku/software/TinySVM`

[10] In the situation of a real application of the technique of compiling a bilingual technical term lexicon, it is recommended to prefer precision rather than F-measure as the evaluation criterion. This is mainly because those who are working on lexicon compilation just judge whether the output bilingual technical term pairs are correct or not and keep the positive ones while ignore the negative ones, instead of finding out the appropriate translations for all of the negative cases.

Among the monolingual features are the frequency of the Japanese term (f_1) and the frequency of the Chinese term (f_2), where their feature values are represented as IDs of the 13 frequency ranges.

Among the bilingual features are the translation probability (f_3), rank of the Chinese translation candidates (f_4), co-occurrence frequency of the bilingual technical term pairs (f_5), the difference of the frequency of Japanese technical term and the co-occurrence frequency of bilingual technical term pairs (f_6)[11], the number of Chinese translation candidates (f_7), rate of parallel sentences where phrase alignment is consistent with word alignments (f_8), and the translation probability when generating the Chinese translation candidates compositionally from constituents of the Japanese technical term (f_9)[12].

The following briefly describes why we employ those features introduced in this section. First, we observed that each term of a bilingual technical term pair tends to be a correct translation of each other when their frequencies are close to each other. Also, since we apply the polynomial (2nd order) kernel as the kernel function of SVMs, we simply introduce primitive features such as frequency of Japanese terms (f_1), frequency of Chinese terms (f_2), and co-occurrence frequency of bilingual technical term pairs (f_5), as well as the difference of those frequencies (f_6). In addition to that, they are correct translation of each other if they have a high translation probability and/or are ranked highly in the SMT phrase translation table. Thus, we use those information directly as the features of f_3 and f_4. Furthermore, we define a feature for the translation probability of compositional translation generation using the phrase translation table (f_9). We also employ the number of translation candidates as another feature (f_7), since a term tends to be a technical term if the number of its translation candidates is small. Finally, we employ the rate of parallel sentences where phrase alignment is consistent with word alignments as a feature (f_8), since this rate tends to be large in the case of correct translation pairs.

5.5 Evaluation Results

Table 5 shows the evaluation results for a baseline as well as for SVMs. As the baseline, we simply judge all of the input Japanese-Chinese technical term pairs as correct translation, which is exactly the same procedure as shown in Figure 1. In the tuning of the lower bound of the confidence measure, when maximizing precision, we achieve almost 94% precision while keeping recall almost 60% with the test data. When maximizing F-measure, we achieve almost 84% F-measure with around 80% precision and 87% recall.

Table 6 also shows the evaluation results for each pair of the 13 frequency ranges of Japanese technical term frequency (jf) and Japanese-Chinese co-occurrence frequency (jcf). As shown in the table, in most pairs of the 13 frequency ranges of Japanese technical term frequency and Japanese-Chinese co-occurrence frequency, we achieve around 90% or higher precision. It is also obvious by comparing those evaluation results with the rates of positive examples for pairs of the 13 frequency ranges of Japanese technical term frequency (jf) and Japanese-Chinese co-occurrence frequency (jcf) in Table 1 that, we have lower recalls for certain frequency range pairs when the rates of positive examples are lower for those frequency range pairs[13].

[11] The upper bound 105 shown in Table 4 is used following the result of a preliminary experiment.

[12] Patent families are one of the largest parallel sentences resource which contain lots of technical term pairs, and their number grows year by year. As the result of using patent families as knowledge source for solving the task of extracting bilingual technical term pairs, some of the features studied in this paper, such as co-occurrence frequency of the bilingual technical term pairs (f_5) and the number of Chinese translation candidates (f_7) and so on, happen to be effective only in the case of using patent families as knowledge source.

[13] It is also interesting to note that the higher the frequencies of the Japanese technical terms are, the more the variety of translation candidates is, the more the rates of negative examples are, and finally the lower the recall is. On the contrary, the higher the co-occurrence frequencies of the Japanese-Chinese technical term pairs are, the more reliably

Table 6: Evaluation Results (Precision / Recall / F-measure) (%) for Each Frequency Range of Japanese Technical Term Frequency (jf) and Japanese-Chinese Co-occurrence Frequency (jcf)

	jf= 1	2≦jf ≦5	6≦jf ≦10	11≦jf ≦15	16≦jf ≦20	21≦jf ≦30	31≦jf ≦50	51≦jf ≦100	101≦jf ≦200	201≦jf ≦500	501≦jf ≦1,000	1,001≦jf ≦10,000	10,001≦jf	Total
jcf=1	100/27.1 /42.6	100/25.1 /40.0	100/25.1 /40.0	100/35.0 /51.9	100/23.9 /38.5	50.0/9.1 /15.4	100/4.2 /8.0	33.4/5.3 /9.1	0/0/0	0/0/0	0/0/0	0/0/0	0/0/0	89.2/12.9 /22.6
2≦jcf ≦5		97.9/83.7 /90.2	93.4/58.4 /71.8	85.8/50.0 /63.2	90.0/56.3 /69.3	77.8/50.0 /60.9	100/18.2 /30.8	85.8/28.6 /42.9	75.0/10.8 /18.8	0/0/0	0/0/0	0/0/0	0/0/0	91.4/35.0 /50.6
6≦jcf ≦10			100/95.2 /97.5	100/93.8 /96.8	100/85.8 /92.4	92.9/92.9 /92.9	92.4/75.0 /82.8	80.0/44.5 /57.2	75.0/20.1 /31.6	100/16.7 /28.6	100/21.8 /35.8	100/35.8 /52.7	0/0/0	96.5/62.9 /76.2
11≦jcf ≦15				100/97.4 /98.7	92.9/100 /96.3	100/100 /100	100/66.7 /80.0	50.0/50.0 /50.0	100/62.5 /77.0	85.8/60.0 /70.6	83.4/71.5 /77.0	71.5/83.4 /77.0	0/0/0	91.5/84.4 /87.8
16≦jcf ≦20					100/96.8 /98.4	100/100 /100	100/87.5 /93.4	50.0/100 /66.7	87.5/100 /93.4	80.0/100 /88.9	100/100 /100	83.4/71.5 /77.0	0/0/0	94.6/93.3 /93.9
21≦jcf ≦30						100/96.9 /98.5	75.0/100 /85.8	75.0/100 /85.8	66.7/100 /80.0	100/80.0 /88.9	85.8/100 /92.4	66.7/100 /80.0	100/100 /100	88.1/96.8 /92.2
31≦jcf ≦50							100/96.0 /98.0	87.5/93.4 /90.4	100/85.8 /92.4	71.5/71.5 /71.5	80.0/100 /88.9	66.7/66.7 /66.7	50.0/33.4 /40.0	88.3/88.3 /88.3
51≦jcf ≦100								95.9/95.9 /95.9	100/95.3 /97.6	75.0/85.8 /80.0	100/100 /100	100/100 /100	100/50.0 /66.7	95.6/92.9 /94.3
101≦jcf ≦200									100/96.0 /98.0	100/100 /100	100/100 /100	100/85.8 /92.4	100/50.0 /66.7	100/93.2 /96.5
201≦jcf ≦500										100/87.0 /93.1	100/90.0 /94.8	100/100 /100	100/66.7 /80.0	100/88.1 /93.7
501≦jcf ≦1,000											100/96.8 /98.4	100/100 /100	100/100 /100	100/97.3 /98.7
1,001≦jcf ≦10,000												100/100 /100	100/50.0 /66.7	100/90.0 /94.8
10,001≦jcf													100/100 /100	100/100 /100
Total	100/27.1 /42.6	98.1/68.0 /80.4	98.3/72.8 /83.6	98.5/75.6 /85.6	97.2/72.7 /83.2	94.5/78.2 /85.6	95.1/54.3 /69.1	83.1/56.3 /67.1	91.4/52.2 /66.4	91.5/49.3 /64.1	94.4/51.3 /66.5	93.0/54.1 /68.4	91.7/59.5 /72.2	93.9/59.0 /72.5

Table 7: Examples of Judgement by SVM

(a) Examples of Correct Judgement by SVM

Japanese technical term for evaluation	Chinese translation candidate	feature f_1	feature f_3	feature f_4	feature f_5	feature f_7	feature f_9	reference judgement	judgement by SVM
置換/基 (substituent)	取/代/基 (substituent)	1,000≦jf ≦10,000	0.8	1	1,000≦jcf ≦10,000	6	0.12	correct translation	correct translation
気/液/分離/器 (gas-liquid separator)	气/液/反/应/器 (gas-liquid reactor)	1,000≦jf ≦10,000	0.0008	14	jcf=1	14	0	translation error	translation error

(b) Examples of Erroneous Judgement by SVM

Japanese technical term for evaluation	Chinese translation candidate	feature f_1	feature f_3	feature f_4	feature f_5	feature f_7	feature f_9	reference judgement	judgement by SVM
カバー/絶縁/層 (cover insulating layer)	盖/绝/缘/层 (substring of word "cover" + insulating layer)	501≦jf ≦1,000	0.05	3	21≦jcf ≦30	10	0.57	translation error	correct translation
駆動/回路 (drive circuit)	驱/动/器/电/路 (drive circuit)	10,001≦jcf	0.006	3	101≦jcf ≦200	5	0	correct translation	translation error

Next, Table 7 shows examples of correct and erroneous SVMs' judgments. As shown in Table 7(a), a Japanese-Chinese technical term pair ⟨"置換/基", "取/代/基" ⟩ are correctly judged by SVM, mainly because its translation probability in the phrase translation table (f_3) is high and the rank of the Chinese translation candidate (f_4) is 1. In addition, its translation probability of compositional translation generation (f_9) are relatively high, and its number of Chinese translation candidates (f_7) are relatively small. Compared with this result of correct translation by the framework of this paper based on Chinese sentences segmented by characters, on the other hand, the framework of Dong et al. (2015) based on Chinese sentences segmented by morphemes translates the Japanese technical term "置換/基" not only into the correct Chinese translation "取代/基", but also into an erroneous Chinese translation "取代/基如" (which means "substituent such as"). This is mainly because of the error in Chinese morphological analysis where the two Chinese characters "基"and "如" are concatenated into one morpheme, and then, SVM trained with the phrase translation table with Chinese sentences segmented by morphemes can not judge "取代/基如" as an erroneous translation. Also, another Japanese-Chinese technical term pair ⟨"気/液/分離/器", "气/液/反/应/器"⟩ is correctly judged by SVM to be a translation error, mainly because its values of f_3 and f_9 are 0 or quite small, while those of f_4 and f_7 are fairly large.

Table 7(b) shows erroneous judgements by SVM. The first bilingual technical term pair ⟨"カバー/絶縁/層", "盖/绝/缘/层"⟩ is a translation error because the Chinese character "盖" is a substring of the Chinese word "覆盖" (which means "cover"), while the correct translation should be ⟨"カバー/絶縁/層", "覆/盖/绝/缘/层"⟩. In the framework based on Chinese sentences segmented by characters, however, both of these two bilingual technical term pair are judged as correct translations. The erroneous bilingual technical term pair ⟨"カバー/絶縁/層", "盖/绝/缘/层"⟩ is judged to be a correct translation mainly because its values of f_3 and f_9 are not quite small, while the rank of f_4 is relatively high and the value of f_7 is relatively small. Compared with this result of erroneous translation by the framework of this paper based on Chinese

the bilingual technical term pairs are to be correct translation, and then the higher the recall is.

sentences segmented by characters, on the other hand, in the framework of Dong et al. (2015) based on Chinese sentences segmented by morphemes, SVM judges only "覆盖/绝缘层" as the correct translation of the Japanese technical term "カバー/絶縁/層", while it judges "盖/绝缘层" as the erroneous translation of "カバー/絶縁/層". This is mainly because, for the erroneous translation pair ⟨"カバー/絶縁/層", "盖/绝缘层"⟩, the value of f_9 is low.

The second bilingual technical term pair, ⟨"駆動/回路", "驱/动/器/电/路"⟩, on the other hand, is a correct translation according to the reference judgement. However, this technical term pair is judged by SVM to be a translation error mainly because the value of the translation probability of compositional translation generation f_9 is 0. In this example, although the Japanese-Chinese constituent phrase translation pair ⟨"駆動", "驱/动/器"⟩ exists in the phrase translation table, its translation probability is below the pre-determined lower bound [14].

6 Related Work

Among related works on acquiring bilingual lexicon from text, Itagaki et al. (2007) focused on automatic validation of translation pairs available in the phrase translation table trained by an SMT model. Itagaki et al. (2007) especially studied to apply a Gaussian mixture model based classifier to the task of automatic validation of translation pairs available in the phrase translation table. Yasuda and Sumita (2013) also studied to extract bilingual terms from comparable patents, where, they first extract parallel sentences from patent families, and then extract bilingual terms from parallel sentences. Yasuda and Sumita (2013) especially studied to exploit kanji character similarity between Japanese and Chinese languages in the task of extracting Japanese-Chinese bilingual term pairs. It is also reported that two types of SMT phrase translation tables are integrated in this task. Haque et al. (2014a) also presented a bilingual terminology extraction method using the phrase translation table trained by a phrase-based SMT. One of the major differences of our approach and those proposed in Itagaki et al. (2007) , Yasuda and Sumita (2013) and Haque et al. (2014b) is that we apply the SVM-based classifier learning framework to the task of identifying bilingual technical term pairs from parallel patent sentences, where we examine various features extracted from parallel patent sentences themselves as well as the phrase translation table of a statistical machine translation model trained with those parallel patent sentences.

Lu and Tsou (2009) also studied to extract English-Chinese bilingual terms from patent families, where they first extract parallel sentences from patent families, and then extract bilingual terms from parallel sentences based on an SVM classifier. One of the major differences of our approach and that proposed in Lu and Tsou (2009) is that our features studied in this paper are much finer-grained and cover wider range of information that are available from parallel patent sentences themselves as well as the phrase translation table of a statistical machine translation model trained with those parallel patent sentences.

Morishita et al. (2008) studied to acquire Japanese-English technical term translation lexicon from the phrase translation tables, which are trained by a phrase-based SMT model with parallel sentences automatically extracted from patent families. The approach taken in Morishita et al. (2008) is based on integrating the phrase translation table and compositional translation generation based on an existing bilingual lexicon for human use. This approach is quite effective in the case of language pairs such as Japanese and English, where an existing bilingual lexicon for human use is widely available. However, this is not always the case in the case of other language pairs such as Japanese and Chinese. Compared with the approach of Morishita et al. (2008), our approach is advantageous in that we concentrate on utilizing information that are available from patent families, but not rely on information source other than patent fami-

[14] In this paper, we introduce a lower bound of the translation probability of constituent phrase translation pairs in the procedure of compositional translation generation of the feature f_9, and set the lower bound as 0.005.

lies. Also, compared with the features of SVM examined in Morishita et al. (2008), those we employed in this paper cover much wider range of information that are available from patent families. Especially, we concentrate more on utilizing features that are based on statistics of all the parallel sentences of the patent families rather than a single parallel sentence pair. In our proposed framework, we introduce the feature of the number of Chinese translation candidates (f_7), which was not examined in Morishita et al. (2008). We also use the rate of parallel sentences where phrase alignment is consistent with word alignments as a feature (f_8), while Morishita et al. (2008) used a binary feature which judges for each parallel sentence pair whether a phrase alignment is consistent with word alignments. Finally, we use the feature of the translation probability of compositional translation generation, which, through a preliminary evaluation, is proved to perform better than the binary feature of compositional translation generation employed in Morishita et al. (2008).

7 Conclusion

In this paper, we segment the Chinese sentences into characters instead of segmenting them into morphemes as in (Dong et al., 2015), and represent Japanese-Chinese patent families in terms of character-segmented Chinese sentences and morpheme-segmented Japanese sentences. Then, to those Japanese-Chinese patent families, we apply the framework (Dong et al., 2015) of identifying bilingual technical terms. As a result, we achieve the performance of over 90% precision with the condition of more than or equal to 60% recall.

As a future work, in order to avoid errors caused by character-based segmentation of Chinese sentences as discussed in section 5.5, we plan to integrate two types of a Japanese-Chinese phrase translation table, where Chinese sentences are segmented not only by characters, but also by morphemes. Another future work is to integrate phonetic level (Xu et al., 2006) as well as character level (Chu et al., 2013) correspondences between Japanese and Chinese within our feature framework such as the one of the translation probability of compositional translation generation (f_9). As the phonetic level correspondences, introducing the framework of transliteration based on Katakana-Pinyin correspondence (Xu et al., 2006) is expected to improve the performance. As the character level correspondences, introducing the correspondence based on shared Chinese characters between Japanese and Chinese languages is expected to improve the performance.

References

Bouamor, D., Semmar, N., and Zweigenbaum, P. (2013). Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In *Proc. 51st ACL*, pages 759–764.

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

Dong, L., Long, Z., Utsuro, T., Mitsuhashi, T., and Yamamoto, M. (2015). Collecting bilingual technical terms from Japanese-Chinese patent families by SVM. In *Proc. PACLING*, pages 71–79.

Erdmann, M., Nakayama, K., Hara, T., and Nishio, S. (2009). Improving the extraction of bilingual terminology from Wikipedia. *ACM Transactions on Multimedia Computing, Communications and Applications*, 5(4):31:1–31:17.

Fung, P. and Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proc. 17th COLING and 36th ACL*, pages 414–420.

Haque, R., Penkale, S., and Way, A. (2014a). Bilingual termbank creation via log-likelihood comparison and phrase-based statistical machine translation. In *Proc. 4th Computerm*, pages 42–51.

Haque, R., Penkale, S., and Way, A. (2014b). Bilingual termbank creation via log-likelihood comparison and phrase-based statistical machine translation. In *Proc. 4th Computerm*, pages 42–51.

Huang, F., Zhang, Y., and Vogel, S. (2005). Mining key phrase translations from Web corpora. In *Proc. HLT/EMNLP*, pages 483–490.

Itagaki, M., Aikawa, T., and He, X. (2007). Automatic validation of terminology translation consistency with statistical method. In *Proc. MT Summit XI*, pages 269–274.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pages 177–180.

Lin, D., Zhao, S., Van Durme, B., and Paşca, M. (2008). Mining parenthetical translations from the Web by word alignment. In *Proc. 46th ACL: HLT*, pages 994–1002.

Lu, B. and Tsou, B. K. (2009). Towards bilingual term extraction in comparable patents. In *Proc. 23rd PACLIC*, pages 755–762.

Matsumoto, Y. and Utsuro, T. (2000). Lexical knowledge acquisition. In Dale, R., Moisl, H., and Somers, H., editors, *Handbook of Natural Language Processing*, chapter 24, pages 563–610. Marcel Dekker Inc.

Morin, E. and Hazem, A. (2014). Looking at unbalanced specialized comparable corpora for bilingual lexicon extraction. In *Proc. 52nd ACL*, pages 1284–1293.

Morishita, Y., Utsuro, T., and Yamamoto, M. (2008). Integrating a phrase-based SMT model and a bilingual lexicon for human in semi-automatic acquisition of technical term translation lexicon. In *Proc. 8th AMTA*, pages 153–162.

Tonoike, M., Kida, M., Takagi, T., Sasaki, Y., Utsuro, T., and Sato, S. (2006). A comparative study on compositional translation estimation using a domain/topic-specific corpus collected from the web. In *Proc. 2nd Intl. Workshop on Web as Corpus*, pages 11–18.

Utiyama, M. and Isahara, H. (2007). A Japanese-English patent parallel corpus. In *Proc. MT Summit XI*, pages 475–482.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.

Xu, L., Fujii, A., and Ishikawa, T. (2006). Modeling impression in probabilistic transliteration into Chinese. In *Proc. 2006 EMNLP*, pages 242–249.

Yasuda, K. and Sumita, E. (2013). Building a bilingual dictionary from a Japanese-Chinese patent corpus. In *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *LNCS*, pages 276–284. Springer.