Evaluating Features for Identifying Japanese-Chinese Bilingual Synonymous Technical Terms from Patent Families

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 the process of translating patent documents, a bilingual lexicon of technical terms is inevitable knowledge source. It is important to develop techniques of acquiring technical term translation equivalent pairs automatically from parallel patent documents. We take an approach of utilizing the phrase table of a state-of-the-art phrase-based statistical machine translation model. First, we collect candidates of synonymous translation equivalent pairs from parallel patent sentences. Then, we apply the Support Vector Machines (SVMs) to the task of identifying bilingual synonymous technical terms. This paper especially focuses on the issue of examining the effectiveness of each feature and identifies the minimum number of features that perform as comparatively well as the optimal set of features. Finally, we achieve the performance of over 90% precision with the condition of more than or equal to 25% 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), 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), and translation term pair acquisition from comparable corpora (Fung and Yee, 1998; Aker et al., 2013; Kontonatsios et al., 2014; Rapp and Sharoff, 2014).

Among those efforts of acquiring bilingual lexicon from text, Morishita et al. (2008) studied to acquire Japanese-English technical term translation lexicon from phrase tables, which are trained by a phrase-based SMT model with parallel sentences automatically extracted from parallel patent documents. Furthermore, based on the achievement above, Liang et al. (2011a) studied the issue of identifying Japanese-English synonymous translation equivalent pairs in the task of acquiring Japanese-English technical term translation equivalent pairs. Based on the technique and the results of identifying Japanese-English synonymous translation equivalent pairs in Liang et al. (2011a), Long et al. (2014) next studied how to identify Japanese-Chinese synonymous translation equivalent pairs from Japanese-Chinese patent families.

In the task of identifying Japanese-Chinese synonymous translation equivalent pairs from Japanese-Chinese patent families (Figure 1) studied in Long et al. (2014), this paper modifies some of the features studied in Long et al. (2014) and further focuses on the issue of examining the effectiveness of each feature. This paper especially identifies the minimum number of features that perform as comparatively well as the optimal set of features, where the most effective feature is discovered to be the rate of intersection in translation by the phrase table. Based on the evaluation results, we finally achieve the performance of over 90% precision with the condition of more than or equal to 25% recall.

2 Japanese-Chinese Parallel Patent Documents

Japanese-Chinese parallel patent documents are collected from the Japanese patent documents

Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 52–61,
Beijing, China, July 30, 2015.

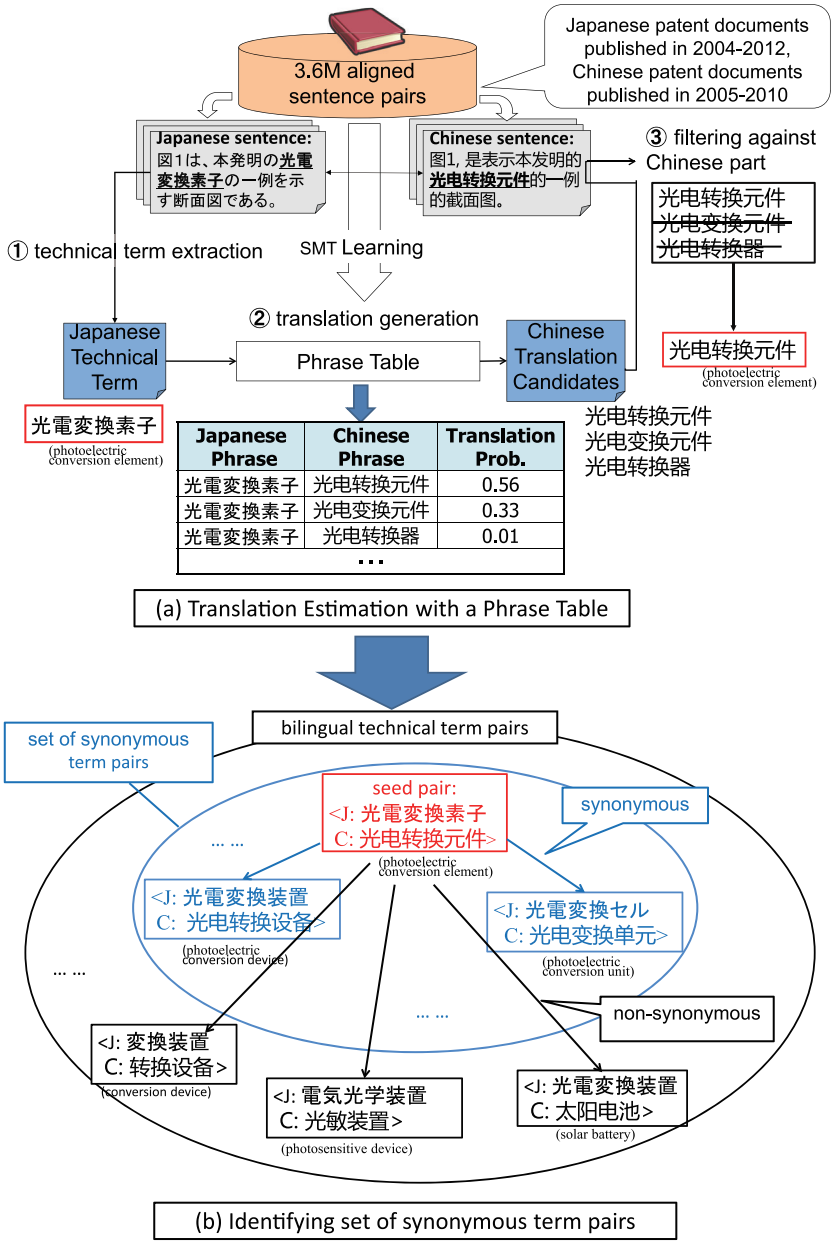

Figure 1: Framework of Identifying Japanese-Chinese Bilingual Synonymous Technical Terms from Patent Families

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 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]. For Chinese sentences, we examine two types of segmentation,

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

[2]The maximum score of the method of Utiyama and Isahara (2007) is set to be 1.0, while the lower bound of its score is about 0.152 with the 3.6M parallel patent sentences.

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

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

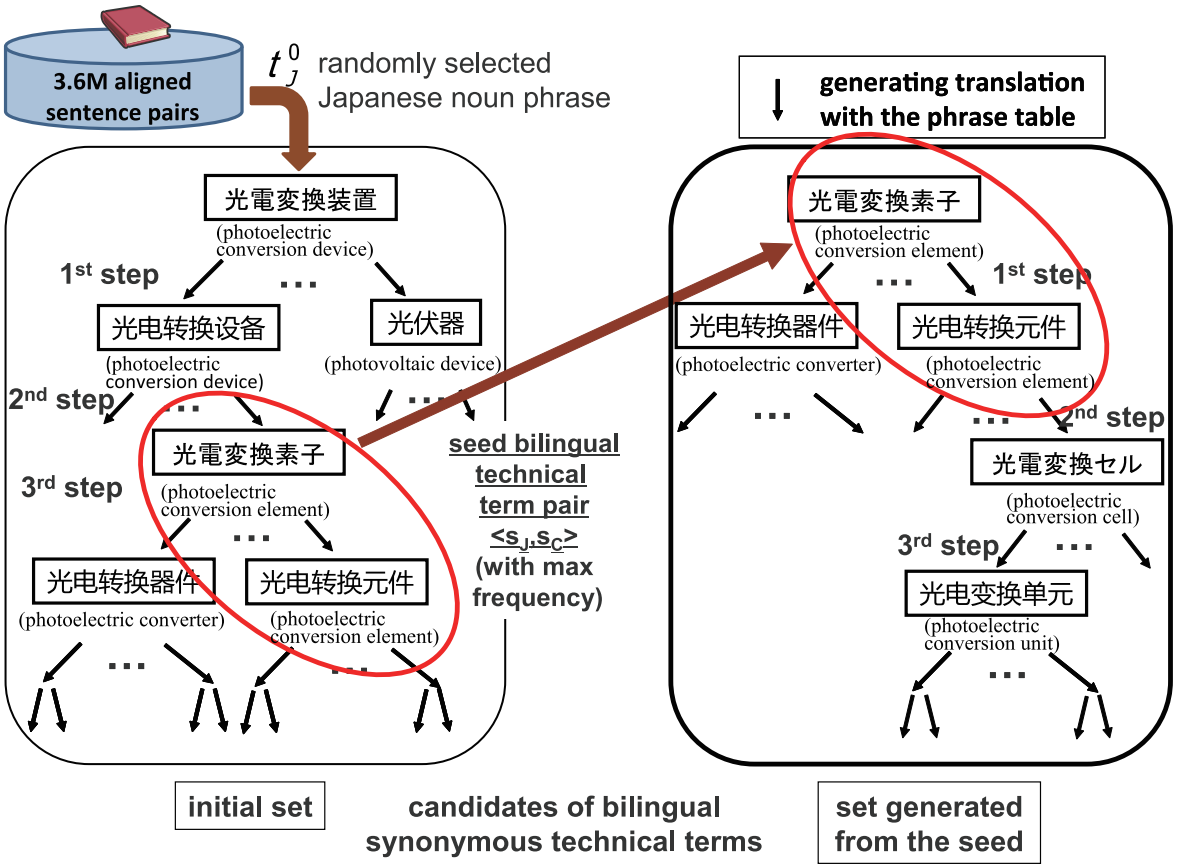

Figure 2: Developing a Reference Set of Bilingual Synonymous Technical Terms

i.e., segmentation by characters[5] and segmentation by morphemes[6].

As the result of applying Moses, we have a phrase table in the direction of Japanese to Chinese translation, and another one in the opposite direction of Chinese to Japanese translation. In the direction of Japanese to Chinese translation, when Chinese side of parallel sentences are segmented by morphemes, we finally obtain 108M translation pairs with 75M unique 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. When Chinese sentences are segmented by characters, on the other hand, we obtain 274M translation pairs with 197M unique Japanese phrases. For each Japanese phrase, those multiple translation candidates in the phrase table are ranked in descending order of Japanese to Chinese phrase translation probabilities. In the similar way, in the phrase table in the opposite direction of Chinese to Japanese translation, for each Chinese phrase, multiple Japanese translation candidates are ranked in descending order of Chinese to Japanese phrase translation probabilities.

Those two phrase tables are then 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, or a Chinese technical term t_C. In the direction of Japanese to Chinese, as shown in Figure 1 (a), given a parallel sentence pair $\langle S_J, S_C \rangle$ containing a Japanese technical term t_J, Chinese translation candidates collected from the Japanese to Chinese phrase 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. Similarly, in the opposite direction of Chinese to Japanese, given a parallel sentence pair $\langle S_J, S_C \rangle$ containing a Chinese technical term t_C, the Chinese to Japanese

[5]A consecutive sequence of numbers as well as a consecutive sequence of alphabetical characters are segmented into a token.

[6]Chinese sentences are segmented into a sequence of morphemes by the Chinese morphological analyzer Stanford Word Segment (Tseng et al., 2005) trained with Chinese Penn Treebank.

phrase table is referred to when identifying a bilingual technical term pair.

4 Developing a Reference Set of Bilingual Synonymous Technical Terms

When developing a reference set of bilingual synonymous technical terms (detailed procedure to be found in Long et al. (2014)), as illustrated in Figure 2, starting from a seed bilingual term pair $s_{JC} = \langle s_J, s_C \rangle$, we repeat the translation estimation procedure of the previous section in both Japanese-Chinese direction and Chinese-Japanese direction six times in total, and generate the set $CBP(s_J)$ of candidates of bilingual synonymous technical term pairs. Then, we manually divide the set $CBP(s_J)$ into $SBP(s_{JC})$, those of which are synonymous with s_{JC}, and the remaining $NSBP(s_{JC})$. As in Table 1, we collect 114 seeds, where the number of bilingual technical terms included in $SBP(s_{JC})$ in total for all of the 114 seed bilingual technical term pairs is around 2,300 to 2,400, which amounts to around 21 per seed on average[7]. As shown in Figure 1 (b), to all of those bilingual term pairs, the procedure of identifying the synonymous sets is applied.

5 Identifying Bilingual Synonymous Technical Terms by Machine Learning

In this section, we apply the Support Vector Machines (SVMs) (Vapnik, 1998) to the task of identifying bilingual synonymous technical terms. In this paper, we model the task of identifying bilingual synonymous technical terms by the SVMs as that of judging whether or not the input bilingual term pair $\langle t_J, t_C \rangle$ is synonymous with the seed bilingual technical term pair $s_{JC} = \langle s_J, s_C \rangle$.

5.1 The Procedure

First, let CBP be the union of the sets $CBP(s_J)$ of candidates of bilingual synonymous technical term pairs for all of the 114 seed bilingual technical term pairs. In the training and testing of the classifier for identifying bilingual synonymous technical terms, we first divide the set of 114 seed bilingual technical term pairs into 10 subsets. Here, for each i-th subset ($i = 1, \ldots, 10$), we construct the union CBP_i of the sets $CBP(s_J)$ of candidates of bilingual synonymous technical term pairs, where $CBP_1, \ldots, CBP_{10}$ are 10 disjoint subsets[8] of CBP.

As a tool for learning SVMs, we use TinySVM (`http://chasen.org/~taku/software/TinySVM/`). As the kernel function, we use the polynomial (1st order) kernel[9]. In the testing of a SVMs classifier, we regard the distance from the separating hyperplane to each test instance as a confidence measure, and return test instances satisfying confidence measures over a certain lower bound only as positive samples (i.e., synonymous with the seed). In the training of SVMs, we use 8 subsets out of the whole 10 subsets $CBP_1, \ldots, CBP_{10}$. Then, we tune the lower bound of the confidence measure with one of the remaining two subsets. With this subset, we also tune the parameter of TinySVM for trade-off between training error and margin. Finally, we test the trained classifier against another one of the remaining two subsets. We repeat this procedure of training / tuning / testing 10 times, and average the 10 results of test performance.

5.2 Features

Table 2 lists all the features used for training and testing of SVMs for identifying bilingual synonymous technical terms. Features are roughly divided into two types: those of the first type $f_1, \ldots, f_6$ simply represent various characteristics of the input bilingual technical term $\langle t_J, t_C \rangle$, while those of the second type $f_7, \ldots, f_{17}$ represent relation of the input bilingual technical term $\langle t_J, t_C \rangle$ and the seed bilingual technical term pair $s_{JC} = \langle s_J, s_C \rangle$

Among the features of the first type are the frequency (f_1), ranks of terms with respect to the conditional translation probabilities (f_2 and f_3), length of terms (f_4 and f_5), and the number of times repeating the procedure of generating translation with the phrase tables until generating input terms t_J and t_C from the Japanese seed term s_J (f_6).

Among the features of the second type are identity of monolingual terms (f_7 and f_8), edit distance of monolingual terms (f_9), character bigram sim-

[7] We manually generate the reference set by discarding the bilingual pairs which are judged as not synonymous with the seed pair. The procedure of generating the whole reference sets took about 30 hours, i.e., about 3 seconds for judging a bilingual term pair on average.

[8] Here, we divide the set of 114 seed bilingual technical term pairs into 10 subsets so that the numbers of positive (i.e., synonymous with the seed) / negative (i.e., not synonymous with the seed) samples in each CBP_i ($i = 1, \ldots, 10$) are comparative among the 10 subsets.

[9] We compare the performance of the 1st order and 2nd order kernels, where we have almost comparative performance.

Table 1: Number of Bilingual Technical Terms: Candidates and Reference of Synonyms

(a) With the Phrase Table based on Chinese Sentences Segmented by Morphemes

		# of bilingual technical terms for the total 114 seeds		average per seed	
Candidates of Synonyms $\bigcup_{s_J} CBP(s_J)$	included only in the set (a)	12,640	24,621	110.9	216.0
	included in the intersection of the sets (a) and (b)	11,981		105.1	
Reference of Synonyms $\bigcup_{s_{JC}} SBP(s_{JC})$	included only in the set (a)	228	2,473	2.0	21.7
	included in the intersection of the sets (a) and (b)	2,245		19.7	

(b) With the Phrase Table based on Chinese Sentences Segmented by Characters

		# of bilingual technical terms for the total 114 seeds		average per seed	
Candidates of Synonyms $\bigcup_{s_J} CBP(s_J)$	included only in the set (b)	6,358	17,478	55.8	153.3
	included in the intersection of the sets (a) and (b)	11,120		97.5	
Reference of Synonyms $\bigcup_{s_{JC}} SBP(s_{JC})$	included only in the set (b)	287	2,318	2.5	20.3
	included in the intersection of the sets (a) and (b)	2,031		17.8	

Table 4: Pairs of Features having No Significant Difference (5% Significance Level) with Maximum Precision Features and their Evaluation Results (%)

(a) Chinese sentences are segmented by morphemes

feature	precision	recall	f-measure
$f_{15} + f_{16}$	85.6	25.4	39.2
$f_9 + f_{16}$	86.8	24.9	38.7
$f_{13} + f_{14} + f_{16}$	86.8	24.8	38.6

(b) Chinese sentences are segmented by characters

feature	precision	recall	f-measure
$f_9 + f_{15}$	87.4	25.4	39.3

ilarity of monolingual terms (f_{10}), rate of identical morphemes (in Japanese, f_{11}) / characters (in Chinese, f_{12}), string subsumption and variants for Japanese (f_{13}), identical stem for Chinese (f_{14}), rate of intersection in translation by the phrase table (f_{15}), rate of intersection in translation by the phrase table for the substrings not common between the seed and a term (f_{16}), and translation by the phrase tables (f_{17}).

As we discuss in the next section, among all of those features, f_{15} and f_{16}, which utilize the rate of intersection in translation by the phrase table, are the most effective, where we add f_{16} in this paper to those studied in Long et al. (2014).

5.3 Evaluating the Effectiveness of Features

Table 3 shows the evaluation results for a baseline as well as for SVMs. As the baseline, we simply judge the input bilingual term pair $\langle t_J, t_C \rangle$ as synonymous with the seed bilingual technical term pair $s_{JC} = \langle s_J, s_C \rangle$ when t_J and s_J are identical, or, t_C and s_C are identical. When training / testing a SVMs classifier, we tune the lower bound of the confidence measure of the distance from the separating hyperplane in two ways: i.e., for maximizing precision and for maximizing F-measure. As shown in Table 3, when we use the set of features which maximize precision, we achieve higher precisions of 89.0% and 90.4% for morpheme-based segmentation and character-based segmentation, respectively, compared with when we use all of the proposed features (86.5% and 89.0%) with the condition of more than or equal to 40% F-measure[10]. The sets of features which maximize precision are $f_{1\sim6} + f_{9\sim16}$ for morpheme-based

[10]Out of 655 (for morpheme-based segmentation) / 605 (for character-based segmentation) pairs which are correctly judged as synonymous with the seed pair by SVM , 197 (30.1%) / 161 (26.6%) are not judged as synonymous by the baseline method, i.e., neither the Japanese term nor the Chinese term is identical to that of the seed pair. On the other hand, out of 986 (for morpheme-based segmentation) / 927 (for character-based segmentation) pairs which are correctly judged as synonymous by the baseline method, 458 (46.5%) / 444 (47.9%) are judged as synonymous with the seed pair by SVM, while the rests are not judged as synonymous by SVM.

Table 2: Features for Identifying Bilingual Synonymous Technical Terms by Machine Learning

class	feature	definition (where X denotes J or C, and $\langle s_J, s_C \rangle$ denotes the seed bilingual technical term pair)
features for bilingual technical terms $\langle t_J, t_C \rangle$	f_1: frequency	log of the frequency of $\langle t_J, t_C \rangle$ within the whole parallel patent sentences
	f_2: rank of the Chinese term	given t_J, log of the rank of t_C with respect to the descending order of the conditional translation probability $\mathrm{P}(t_C \mid t_J)$
	f_3: rank of the Japanese term	given t_C, log of the rank of t_J with respect to the descending order of the conditional translation probability $\mathrm{P}(t_J \mid t_C)$
	f_4: number of Japanese characters	number of characters in t_J
	f_5: number of Chinese characters	number of characters in t_C
	f_6: number of times generating translation by applying the phrase tables	the number of times repeating the procedure of generating translation by applying the phrase tables until generating t_C or t_J from s_J, as in $s_C \rightarrow \cdots \rightarrow t_J \rightarrow t_C$, or, $s_J \rightarrow \cdots \rightarrow t_C \rightarrow t_J$
features for the relation of bilingual technical terms $\langle t_J, t_C \rangle$ and the seed $\langle s_J, s_C \rangle$	f_7: identity of Japanese terms	returns 1 when $t_J = s_J$
	f_8: identity of Chinese terms	returns 1 when $t_C = s_C$
	f_9: edit distance similarity of monolingual terms	$f_9(t_X, s_X) = 1 - \frac{ED(t_X, s_X)}{\max(\lvert t_X \rvert, \lvert s_X \rvert)}$ (where ED is the edit distance of t_X and s_X, and $\mid t \mid$ denotes the number of characters of t.)
	f_{10}: character bigram similarity of monolingual terms	$f_{10}(t_X, s_X) = \frac{\lvert bigram(t_X) \cap bigram(s_X) \rvert}{\max(\lvert t_X \rvert, \lvert s_X \rvert) - 1}$ (where $bigram(t)$ is the set of character bigrams of the term t.)
	f_{11}: rate of identical morphemes (for Japanese terms)	$f_{11}(t_J, s_J) = \frac{\lvert const(t_J) \cap const(s_J) \rvert}{\max(\lvert const(t_J) \rvert, \lvert const(s_J) \rvert)}$ (where $const(t)$ is the set of morphemes in the Japanese term t.)
	f_{12}: rate of identical characters (for Chinese terms)	$f_{11}(t_C, s_C) = \frac{\lvert const(t_C) \cap const(s_C) \rvert}{\max(\lvert const(t_C) \rvert, \lvert const(s_C) \rvert)}$ (where $const(t)$ is the set of Characters in the Chinese term t.)
	f_{13}: subsumption relation of strings / variants relation of surface forms (for Japanese terms)	returns 1 when the difference of t_J and s_J is only in their suffixes, or only whether or not having the prolonged sound "ー", or only in their hiragana parts.
	f_{14}: identical stem (for Chinese terms)	returns 1 when the difference of t_C and s_C is only whether or not having the word "性" which is not the prefix or suffix.
	f_{15}: rate of intersection in translation by the phrase table	$f_{15}(t_X, s_X) = \frac{\lvert trans(t_X) \cap trans(s_X) \rvert}{\max(\lvert trans(t_X) \rvert, \lvert trans(s_X) \rvert)}$ (where $trans(t)$ is the set of translation of term t from the phrase table.)
	f_{16}: rate of intersection in translation by the phrase table (for the substrings not common between t_X and s_X)	Suppose that $x_t^1, \ldots, x_t^m$ and $x_s^1, \ldots, x_s^n$ are the substrings which are not common between t_X and s_X. Here, we find l $(= \min(m, n))$ pairs of one-to-one mappings between x_t^i $(i = 1, \ldots, m)$ and x_s^j $(j = 1, \ldots, n)$ which maximize the product of the rates $f_{15}(x_t^i, x_s^j)$ of intersection in translation by the phrase table and return this product.
	f_{17}: translation by the phrase table	returns 1 when s_J can be generated by translating t_C with the phrase table, or, s_C can be generated by translating t_J with the phrase table.

segmentation and $f_{2,3} + f_{6\sim9} + f_{11,12,15,16}$ for character-based segmentation, respectively. However, their differences are not significant (5% significance level). Next, we evaluate the effect of each single feature as well as combinations of small number of features, where, among those results, Table 4 shows pairs of features each of which achieves a precision with no significant difference (5% significance level) with the set of features having the maximum precision. It is obvious that features f_{15} and f_{16}, which utilize the rate of intersection in translation by the phrase table, are the most effective. Also, when we remove features f_{15} and f_{16} from all the features, precisions are significantly damaged (5% significance level) to 78.5% and 79.4% for morpheme-based and character-based segmentations, respectively. The reason why these features are the most effective among other features is that they directly measure the degree of being synonymous within one language with respect to the rate of intersection of translations into the other language, while other features just measure the character-based or morpheme-based similarity within one language.

We further compare the performance of the proposed features with those studied in Tsunakawa and Tsujii (2008), where we modify the features of Tsunakawa and Tsujii (2008) as shown in Table 5, and then evaluate those modified features. As we compare the performance of the proposed features and the modified features of Tsunakawa and Tsujii (2008) in Table 3, it is clear that the pro-

Table 3: Evaluation Results (%)

		segmented by morphemes			segmented by characters		
		precision	recall	f-measure	precision	recall	f-measure
baseline (t_J and s_J are identical, or, t_C and s_C are identical.)		71.4	40.0	51.3	74.0	40.1	52.0
SVM (all features)	maximum precision	**86.5**	26.5	40.5	**89.0**	26.1	40.4
	maximum f-measure	64.3	64.1	**64.2**	63.5	65.3	**64.4**
SVM (features with maximum precision)	maximum precision	**89.0** ($f_{1\sim6} + f_{9\sim16}$)	23.9	37.7	**90.4** ($f_{2,3} + f_{6\sim9} + f_{11,12,15,16}$)	25.5	40.4
SVM (features in Tsunakawa and Tsujii (2008))	maximum precision	72.6	26.1	38.4	74.4	36.7	49.2
	maximum f-measure	71.0	54.7	61.5	72.7	53.7	61.8

Table 5: Features for Identifying Bilingual Synonymous Technical Terms by Tsunakawa and Tsujii (2008)

class	features	definition
basical features	h_{1J}, h_{1C}: agreement of the first characters	returns 1 when the fisrt characters of t_X and s_X match.
	h_{2J}, h_{2C}: edit distance of similarity of monolingual terms	the same as f_9
	h_{3J}, h_{3C}: character of bigram similarity of monolingual terms	the same as f_{10}
	h_{4J}, h_{4C}: agreement of word substring	return the count that substrings of t_X match s_X. (Here, Tsunakawa and Tsujii (2008) count not only the common substrings but also substrings in known synonymous relation between t_X and s_X. However, in our work, we have no lexicon available for synonymous relation. So, we utilize only the count of common substrings.)
	h_{5J}, h_{5C}: translation by the phrase table	the same as f_{17}. (Here, instead of the phrase table, Tsunakawa and Tsujii (2008) utilize a bilingual lexicon and consider the existence of bilingual lexical items as features.)
	h_6: identical stem for Chinese terms	the same as f_{14} (Although Tsunakawa and Tsujii (2008) define this feature as examining the acronym relation of English terms, we modify this feature as examining the difference of the Chinese terms as the Chinese word "的".)
	h_7: subsumption relation of strings / variants relation of surface forms for Japanese terms	the same as f_{13} (Although Tsunanakwa and Tsujii (2008) examine only the katakana variant, we additionally examine the difference of suffixes and variants of hiragana parts.)
combinatorial feature	$h_{1J} \wedge h_{1C}$	—
	$\sqrt{h_{2J} \cdot h_{2C}}$	—
	$\sqrt{h_{3J} \cdot h_{3C}}$	—
	$h_{5J} \wedge h_{5C}$	—
	$h_6 \cdot h_{2J}$	—
	$h_7 \cdot h_{2C}$	—

posed features outperform the modified features of Tsunakawa and Tsujii (2008).

Next, Table 6 shows examples of improvement by SVM compared with the baseline. As shown in Table 6 (a), the relation between input bilingual term pairs and seed bilingual term pairs is correctly judged as "synonym", while judgement by the baseline is "not synonym" since neither the Chinese terms nor the Japanese terms are identical. In our proposed features, f_{17} contributes to the correct judgement, where it returns 1 because of the existence of the translation pairs ⟨"ガラス転移温度","绝缘件" ⟩ and ⟨"ガラス転移点","绝热体"⟩ in the phrase table. In the case of another example shown in Table 6 (b), on the other hand, the proposed method correctly judges as "not synonym" by SVM compared with the baseline, where both the edit distance similarity

Table 6: Examples of Improvement in Identifying Bilingual Synonymous Technical Terms by SVM

Baseline:	Judge the input bilingual term pair $\langle t_J, t_C \rangle$ as synonymous with the seed bilingual term pair $\langle s_J, s_C \rangle$ when t_J and s_J are identical, or, t_C and s_C are identical.
SVM:	Maximize precision by tuning the lower bound of the confidence measure of the distance from the separating hyperplane (Chinese sentences are segmented by morphemes).

(a) Correct Judgement as "Synonym" only by SVM

seed <s_J, s_C>	bilingual term pair <t_J, t_C>	reference judgement	judgement by baseline	judgement by SVM
<ガラス転移温度, 玻璃化转变温度> (glass transition temperature)	<ガラス転移点, 玻璃态转化温度> (glass transition temperature)	synonym	not synonym	synonym

(b) Correct Judgement as "Not Synonym" only by SVM

seed <s_J, s_C>	bilingual term pair <t_J, t_C>	reference judgement	judgement by baseline	judgement by SVM
<集電装置, 集电器> (current collector)	<コレクト(collector), 集电器(current collector)>	not synonym	synonym	not synonym

(f_9) and the character bigram similarity (f_{10}) between the Japanese terms "集電装置" and "コレクト" are 0 ($f_9(\langle t_J, t_C \rangle, \langle s_J, s_C \rangle) = 0$ and $f_{10}(\langle t_J, t_C \rangle, \langle s_J, s_C \rangle) = 0$).

Finally, Table 7 shows examples of erroneous judgements by SVM. As shown in Table 7 (a), since erroneous translation pairs ⟨"断热体","绝缘件"⟩ and ⟨"インシュレーター","绝热体"⟩ exist in the phrase table, both f_{17} (both of the translations pairs $\langle s_J, t_C \rangle$ and $\langle t_J, s_C \rangle$ exist in the phrase table) and f_{17} (either the translation pair $\langle s_J, t_C \rangle$ or $\langle t_J, s_C \rangle$ exist in the phrase table) return 1, resulting in erroneous judgement.

Another example is shown in Table 7 (b), where the proposed method returns erroneous judgement as "not synonym". In this case, since the translation pair ⟨"成膜チャンバー","成膜室"⟩ only exists in the phrase table, f_{17} (either the translation pair $\langle s_J, t_C \rangle$ or $\langle t_J, s_C \rangle$ exist in the phrase table) returns 1, while f_{17} (both of translations pairs $\langle s_J, t_C \rangle$ and $\langle t_J, s_C \rangle$ exist in the phrase table) returns 0. Furthermore, even though Chinese words "成膜" and "膜成形" are synonymous, their character bigram similarity is computed as 0, since they have opposite character orderings.

6 Related Work

Among related works on acquiring bilingual lexicon from text, Lu and Tsou (2009) and Yasuda and Sumita (2013) studied to extract bilingual terms from comparable patents, where, they first extract parallel sentences from comparable patents, and then extract bilingual terms from parallel sentences. Those studies differ from this paper in that those studies did not address the issue of acquiring bilingual synonymous technical terms. Tsunakawa and Tsujii (2008) is mostly related to our study, in that they also proposed to apply machine learning technique to the task of identifying bilingual synonymous technical terms. However, Tsunakawa and Tsujii (2008) studied the issue of identifying bilingual synonymous technical terms only within manually compiled bilingual technical term lexicon and thus are quite limited in its applicability. Our approach, on the other hand, is quite advantageous in that we start from parallel patent documents which continue to be published every year and then, that we can generate candidates of bilingual synonymous technical terms automatically. Furthermore, as we show in the previous section, the features proposed in this paper outperform that of Tsunakawa and Tsujii (2008).

7 Conclusion

In the task of acquiring Japanese-Chinese technical term translation equivalent pairs from parallel patent documents, this paper studied the issue of identifying synonymous translation equivalent pairs. This paper especially focused on the issue of examining the effectiveness of each feature and identified the minimum number of features that perform as comparatively well as the optimal set of features. One of the most important future work is definitely to improve recall. To do this, we plan to apply the semi-automatic framework (Liang et al., 2011b) which have been invented in the task of identifying Japanese-English synonymous translation equivalent pairs and have been proven to be effective in improving recall. Another important future work is to train the SVM of identifying bilingual synonymous technical pairs with a set

Table 7: Examples of Errors in Identifying Bilingual Synonymous Technical Terms By the Proposed Method

(a) Incorrect Judgement as "Synonym" by SVM

seed $<s_J, s_C>$	bilingual term pair $<t_J, t_C>$	for Japanese		for Chinese		feature f_{17} (by translating with the phrase table, both s_J and s_C can be generated from t_C and t_J, respectively)	feature f_{17} (by translating with the phrase table, s_J or s_C can be generated from t_C or t_J, respectively)	reference judgement	judgement by SVM
		feature f_9	feature f_{10}	feature f_9	feature f_{10}				
<断熱体, 绝热体> (heat insulator)	<インシュレータ, 绝缘件> (insulator)	0	0	0.33	0	1	1	not synonym	synonym

(a) Incorrect Judgement as "Not Synonym" by SVM

seed $<s_J, s_C>$	bilingual term pair $<t_J, t_C>$	for Japanese		for Chinese		feature f_{17} (by translating with the phrase table, both s_J and s_C can be generated from t_C and t_J, respectively)	feature f_{17} (by translating with the phrase table, s_J or s_C can be generated from t_C or t_J, respectively)	reference judgement	judgement by SVM
		feature f_9	feature f_{10}	feature f_9	feature f_{10}				
<成膜室, 成膜室> (film desposition chamber)	<成膜チャンバー, 膜成形室> (film desposition chamber)	0.29	0.17	0.5	0	0	1	synonym	not synonym

of patent families, and then to evaluate the trained SVM against parallel patent sentences and phrase tables extracted from another set of patent families.

References

A. Aker, M. Paramita, and R. Gaizauskas. 2013. Extracting bilingual terminologies from comparable corpora. In *Proc. 51st ACL*, pages 402–411.

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

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

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

G. Kontonatsios, I. Korkontzelos, J. Tsujii, and S. Ananiadou. 2014. Using random forest classifier to compile bilingual dictionaries of technical terms from comparable corpora. In *Proc. 14th EACL*, pages 111–116.

B. Liang, T. Utsuro, and M. Yamamoto. 2011a. Identifying bilingual synonymous technical terms from phrase tables and parallel patent sentences. *Procedia - Social and Behavioral Sciences*, 27:50–60.

B. Liang, T. Utsuro, and M. Yamamoto. 2011b. Semi-automatic identification of bilingual synonymous technical terms from phrase tables and parallel patent sentences. In *Proc. 25th PACLIC*, pages 196–205.

Z. Long, L. Dong, T. Utsuro, T. Mitsuhashi, and M. Yamamoto. 2014. Identifying Japanese-Chinese bilingual synonymous technical terms from patent families. In *Proc. 7th BUCC*, pages 49–54.

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

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

Y. Morishita, T. Utsuro, and M. Yamamoto. 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.

R. Rapp and S. Sharoff. 2014. Extracting multiword translations from aligned comparable documents. In *Proc. 3rd Workshop on Hybrid Approaches to Translation*, pages 83–91.

M. Tonoike, M. Kida, T. Takagi, Y. Sasaki, T. Utsuro, and S. Sato. 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.

H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. 2005. A conditional random field word segmenter for Sighan bakeoff 2005. In *Proc. 4th SIGHAN Workshop on Chinese Language Processing*, pages 168–171.

T. Tsunakawa and J. Tsujii. 2008. Bilingual synonym identification with spelling variations. In *Proc. 3rd IJCNLP*, pages 457–464.

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

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

K. Yasuda and E. Sumita. 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.