Semi-Automatic Identification of Bilingual Synonymous Technical Terms from Phrase Tables and Parallel Patent Sentences

Bing Liang[a], Takehito Utsuro[b], and Mikio Yamamoto[b]

[a] Graduate School of Systems and Information Engineering,
University of Tsukuba, Tsukuba, 305-8573, JAPAN
[b] Faculty of Engineering, Information and Systems,
University of Tsukuba, Tsukuba, 305-8573, JAPAN

Abstract. In the research field of machine translation of patent documents, the issue of acquiring technical term translation equivalent pairs automatically from parallel patent documents is one of those most important. We take an approach of utilizing the phrase table of a state-of-the-art phrase-based statistical machine translation model. In this task, we consider situations where a technical term is observed in many parallel patent sentences and is translated into many translation equivalents. We apply SVM to the task of identifying synonymous translation equivalent pairs and achieve almost 98% precision and over 40% F-measure. Then, in order to improve recall, we introduce a semi-automatic framework, where we employ the strategy of selecting more than one seeds for each set of candidates bilingual synonymous term pairs. By manually judging whether each pair of two seeds is synonymous or not, we achieve over 95% precision and 50% recall.

Keywords: Bilingual Lexicon Acquisition, Synonyms, Phrase Tables

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), compositional translation generation based on an existing bilingual lexicon for human use (Tonoike *et al.*, 2006), and translation term pair acquisition by collecting partially bilingual texts through the search engine (Huang *et al.*, 2005).

Among those efforts of acquiring bilingual lexicon from text, Morishita *et al.* (2008) studied to acquire technical term translation lexicon from phrase tables, which are trained by a phrase-based statistical machine translation model with parallel sentences automatically extracted from parallel patent documents. Recently, we further studied to require the acquired technical term translation equivalents to be consistent with word alignment in parallel sentences and achieved 91.9% precision with almost 70% recall. This technique has been actually adopted by a Japanese organization which is responsible for translating Japanese patent applications published by the Japanese Patent Office (JPO) into English, where it has been utilized in the process of semi-automatically compiling bilingual technical term lexicon from parallel patent sentences. In this process, persons who are working on compiling bilingual technical term lexicon judge whether to accept or not candidates of bilingual technical term pairs presented by the system.

Based on the achievement so far, in this paper, we consider situations where a technical term is observed in many parallel patent sentences and is translated into many translation equivalents. More specifically, in the task of acquiring technical term translation equivalent pairs, this paper studies the issue of identifying synonymous translation equivalent pairs. First, we collect candidates of synonymous translation equivalent pairs from parallel patent sentences. Then, we analyze features for identifying synonymous translation equivalent pairs. Finally, we apply the Support Vector Machines (SVMs) (Vapnik, 1998) to the task of identifying bilingual synonymous technical terms, and achieve the performance of almost 98% precision and over 40% F-measure. Then, in order to improve recall, we introduce a semi-automatic framework, where we employ the strategy of selecting more than one seeds for each set of candidates bilingual synonymous term pairs. By manually judging whether each pair of two seeds is synonymous or not, we achieve over 95% precision and 50% recall.

2 Japanese-English Parallel Patent Documents

In the NTCIR-7 workshop, the Japanese-English patent translation task is organized (Fujii *et al.*, 2008), where parallel patent documents and sentences are provided by the organizer. Those parallel patent documents are collected from the 10 years of unexamined Japanese patent applications published by the Japanese Patent Office (JPO) and the 10 years patent grant data published by the U.S. Patent & Trademark Office (USPTO) in 1993-2000. The numbers of documents are approximately 3,500,000 for Japanese and 1,300,000 for English. Because the USPTO documents consist of only patent that have been granted, the number of these documents is smaller than that of the JPO documents. From these document sets, patent families are automatically extracted and the fields of "Background of the Invention" and "Detailed Description of the Preferred Embodiments" are selected. This is because the text of those fields is usually translated on a sentence-by-sentence basis. Then, the method of Utiyama and Isahara (2007) is applied to the text of those fields, and Japanese and English sentences are aligned.

3 Phrase Table of an SMT Model

As a toolkit of a phrase-based statistical machine translation model, we use Moses (Koehn *et al.*, 2007) and apply it to the whole 1.8M parallel patent sentences. In Moses, first, word alignment of parallel sentences are obtained by GIZA++ (Och and Ney, 2003) in both translation directions and then the two alignments are symmetrised. Next, any phrase pair that is consistent with word alignment is collected into the phrase table and a phrase translation probability is assigned to each pair. More specifically, we construct a phrase table in the direction of Japanese to English translation, and another one in the opposite direction of English to Japanese translation. In the direction of Japanese to English translation, we finally obtain 76M translation pairs with 33M unique Japanese phrases, i.e., 2.29 English translations per Japanese phrase on average, with Japanese to English phrase translation probabilities $P(p_E \mid p_J)$ of translating a Japanese phrase p_J into an English phrase p_E. For each Japanese phrase, those multiple translation candidates in the phrase table are ranked in descending order of Japanese to English phrase translation probabilities. In the similar way, in the phrase table in the opposite direction of English to Japanese translation, for each English phrase, multiple Japanese translation candidates are ranked in descending order of English 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 $\langle S_J, S_E \rangle$ and a Japanese technical term t_J, or an English technical term t_E. In the direction of Japanese to English, given a parallel sentence pair $\langle S_J, S_E \rangle$ containing a Japanese technical term t_J, the Japanese to English phrase table is referred to when identifying a bilingual technical term pair. From the Japanese to English phrase table, candidates of translating t_J into English which are consistent with word alignment are collected. Then, those English

translation candidates are matched against the English sentence S_E of the parallel sentence pair, and those which are not found in S_E are filtered out. Finally, among the remaining translation candidates, $\hat{t}_E$ with the largest translation probability $P(t_E \mid t_J)$ is selected and the bilingual technical term pair $\langle t_J, \hat{t}_E \rangle$ is identified. The precision of identifying bilingual technical term pair here is 91.9%. Similarly, in the opposite direction of English to Japanese, given a parallel sentence pair $\langle S_J, S_E \rangle$ containing an English technical term t_E, the English to Japanese phrase table is referred to when identifying a bilingual technical term pair.

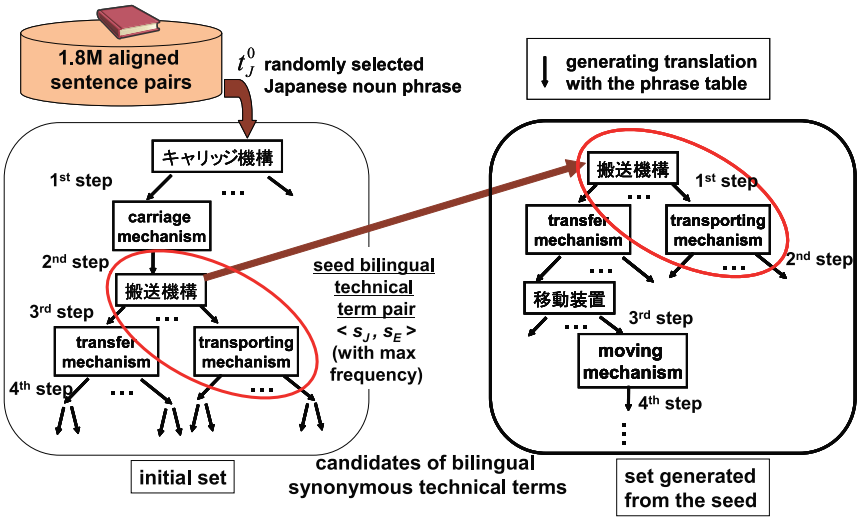

Figure 1: Developing a Reference Set of Bilingual Synonymous Technical Terms

4 Developing a Reference Set of Bilingual Synonymous Technical Terms

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

	# of bilingual technical terms for the total 134 seeds	average per seed
Candidates of Synonyms	$\left\lvert \bigcup_{s_J} CBP(s_J) \right\rvert$ = 22,473	167.7
Reference of Synonyms	$\left\lvert \bigcup_{s_{JE}} SBP(s_{JE}) \right\rvert$ = 1,680	12.5

The following describes the procedure of developing a reference set of bilingual synonymous technical terms from the whole 1.8M parallel patent sentences and the Japanese to English / English to Japanese phrase tables. Figure 1 illustrates the whole procedure.

1. First, a initial Japanese noun phrase t_J^0 is randomly selected from the Japanese part of the 1.8M parallel patent sentences.
2. Then, to the initial Japanese noun phrase t_J^0, the following "Iteration: Generating Candidates Bilingual Synonymous Term Pairs" is applied, where the iteration is repeated steps of translation generation from the 1.8M parallel patent sentences and the Japanese to English / English to Japanese phrase tables[1]. Next, the initial set $CBP(t_J^0)$ of candidate bilingual

[1] The number of iteration 6 here is based on our preliminary evaluation, and is decided so that most synonymous bilingual technical terms are generated from the initial Japanese phrase t_J^0, while the number of candidates other than true synonyms is minimized. Throughout those steps, we simply avoid duplicate generation of terms.

synonymous term pairs is generated as in the left half of Figure 1.

Iteration: Generating Candidates of Bilingual Synonymous Term Pairs

1st step Given the input Japanese term t'_J, collect all the parallel sentence pairs which contain t'_J from the 1.8M parallel patent sentences. Next, from each parallel sentence pair, t'_J is translated into English according to the procedure in the previous section, referring to the Japanese to English phrase table. Then, all the bilingual term pairs $\langle t'_J, t^i_E \rangle$ are collected into the initial set $CBP(t'_J)$ of candidates bilingual synonymous term pairs[2].

2nd step Similarly, for each English term t_E in $CBP(t'_J)$, collect all the parallel sentence pairs which contain t_E from the 1.8M parallel patent sentences, and translate t_E into Japanese, referring to the English to Japanese phrase table. Then, all the bilingual term pairs $\langle t^i_J, t_E \rangle$ are added to $CBP(t'_J)$.

3rd step Similarly, for each Japanese term t_J in $CBP(t'_J)$, collect all the parallel sentence pairs which contain t_J from the 1.8M parallel patent sentences, and translate t_J into English, referring to the Japanese to English phrase table. Then, all the bilingual term pairs $\langle t_J, t^i_E \rangle$ are added to $CBP(t'_J)$.

4th step Repeat the procedure of the "2nd step".

5th step Repeat the procedure of the "3rd step".

6th step Repeat the procedure of the "2nd step".

After the candidate generation iteration, we restrict the set $CBP(t^0_J)$ as having more than or equal to 10 members (i.e., $\mid CBP(t^0_J) \mid \geq 10$). In the evaluation of this paper, out of 4,000 randomly selected initial Japanese noun phrases and corresponding initial sets $CBP(t^0_J)$, about 350 sets satisfy the lower bound of the number of members.

3. Next, out of the members of the initial set $CBP(t^0_J)$ of candidates bilingual synonymous term pairs for the initial Japanese noun phrase t^0_J, we select the seed bilingual term pair $s_{JE} = \langle s_J, s_E \rangle$ as below:

   First, in order to distinguish technical terms and general terms and to select bilingual technical term pairs as seeds, we assume the candidates of seeds to satisfy at least one of the following requirements:

   (a) The co-occurring frequency of the bilingual term pair in the 1.8M parallel patent sentences is less than 500.

   (b) The character length of the Japanese term is more than two when it contains kanji (Chinese characters) or hiragana (Japanese characters). The Japanese term consists of more than one morpheme when all of its characters are katakana (Japanese characters for foreign words).

   (c) The English term consists of more than one word.

   Then, we manually examine the bilingual term pair with the largest co-occurring frequency in the 1.8M parallel patent sentences. If the one with the largest co-occurring frequency is appropriate as a pair of technical terms, we select it as seed. Otherwise, we manually examine all the members of the initial set $CBP(t^0_J)$ and select the most appropriate pair as seed. If the initial set $CBP(t^0_J)$ does not include any pair of bilingual technical terms, we discard the set $CBP(t^0_J)$ at this step.

   In the evaluation of this paper, out of all the initial sets $CBP(t^0_J)$, for about 29% of the initial sets, we keep the bilingual term pair with the largest co-occurring frequency as seed, for about 14% of them, we manually select as seed the pair other than the one with the largest

[2] Throughout the steps from the "1st" to the "6th", we only keep bilingual term pairs which satisfy the lower bound 6 as well as the upper bound 800 of the co-occurring frequency in the 1.8M parallel patent sentences.

co-occurring frequency, and for the remaining 57%, we discard the initial sets $CBP(t_J^0)$. It took about 5.5 minutes on average to manually examine all the members of each initial set $CBP(t_J^0)$.

4. To the Japanese technical term s_J of the seed bilingual technical term pair $s_{JE} = \langle s_J, s_E \rangle$, "Iteration: Generating Candidates Bilingual Synonymous Term Pairs" is applied. As the result of this iteration, the set $CBP(s_J)$ of candidates of bilingual synonymous technical term pairs is generated as in the right half of Figure 1. Here, we again restrict the set $CBP(s_J)$ as having more than or equal to 10 members (i.e., $\mid CBP(s_J) \mid \geq 10$). In the evaluation of this paper, about 90% of the sets $CBP(s_J)$ satisfy the lower bound of the number of members. Finally, we have 134 seed bilingual technical term pairs, where the number of bilingual technical terms in total and their average are shown in Table 1.
5. Finally, for each seed bilingual technical term pair $s_{JE} = \langle s_J, s_E \rangle$, we manually divide the set $CBP(s_J)$ of candidates of bilingual synonymous technical term pairs into $SBP(s_{JE})$, those of which are synonymous with s_{JE}, and the remaining $NSBP(s_{JE})$. As in Table 1, the number of bilingual technical terms included in $SBP(s_{JE})$ in total for all of the 134 seed bilingual technical term pairs is 1,680, which amounts to 12.5 per seed on average.

5 Automatic Identification of Bilingual Synonymous Technical Terms by Machine Learning

In this section, we apply the SVMs to the task of identifying bilingual synonymous technical terms, which we originally proposed in Liang *et al.* (2011).

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 134 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 134 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[3] 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 (2nd order) kernel. 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 (henceforth named as the *development set*). 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 (henceforth named as the *evaluation set*). 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 $\langle t_J, t_E \rangle$,

[3] Here, we divide the set of 134 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.

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

class	feature	definition (where X denotes J or E, and $\langle s_J, s_E \rangle$ denotes the seed bilingual technical term pair)
features for bilingual technical terms $\langle t_J, t_E \rangle$	f_1: frequency	log of the frequency of $\langle t_J, t_E \rangle$ within the whole parallel patent sentences
	f_2: rank of the Japanese term	given t_E, log of the rank of t_J with respect to the descending order of the conditional translation probability $\mathrm{P}(t_J \mid t_E)$
	f_3: rank of the English term	given t_J, log of the rank of t_E with respect to the descending order of the conditional translation probability $\mathrm{P}(t_E \mid t_J)$
	f_4: number of Japanese characters	number of characters in t_J
	f_5: number of English words	number of words in t_E
	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_E or t_J from s_J, as in $s_J \rightarrow \cdots \rightarrow t_J \rightarrow t_E$, or, $s_J \rightarrow \cdots \rightarrow t_E \rightarrow t_J$
features for the relation of bilingual technical terms $\langle t_J, t_E \rangle$ and the seed $\langle s_J, s_E \rangle$	f_7: identity of Japanese terms	returns 1 when $t_J = s_J$
	f_8: identity of English terms	returns 1 when $t_E = s_E$
	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) / words (for English)	$f_{11}(t_X, s_X) = \frac{\lvert const(t_X) \cap const(s_X) \rvert}{\max(\lvert const(t_X) \rvert, \lvert const(s_X) \rvert)}$ (where $const(t)$ is the set of morphemes (for Japanese) / words (for English) in the term t.)
	f_{12}: 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 "—", or only in their hiragana parts.
	f_{13}: identical stem (for English terms)	returns 1 when the numbers of constituent words of t_E and s_E are the same, and their corresponding constituents have the same stem.
	f_{14}: hyphen / space (for English terms)	returns 1 when the difference of t_E and s_E is only whether having hyphen or space.
	f_{15}: compositional translation with an existing bilingual lexicon	returns 1 when s_J can be compositionally generated by translating constituents of t_E with an existing bilingual lexicon, or, s_E can be compositionally generated by translating constituents of t_J with an existing bilingual lexicon (Tonoike *et al.*, 2006).
	f_{16}: translation by the phrase table	returns 1 when s_J can be generated by translating t_E with the phrase table, or, s_E can be generated by translating t_J with the phrase table.

while those of the second type $f_7, \ldots, f_{16}$ represent relation of the input bilingual technical term $\langle t_J, t_E \rangle$ and the seed bilingual technical term pair $s_{JE} = \langle s_J, s_E \rangle$.

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_E 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 similarity of monolingual terms (f_{10}), rate of identical morphemes / words (f_{11}), string subsumption and variants for Japanese (f_{12}), identical stems for English (f_{13}), hyphen / space of English terms (f_{14}), compositional translation with an existing bilingual lexicon[4] (f_{15}), and translation by the phrase tables (f_{16}).

5.3 Evaluation Results

Table 3: Evaluation Results of Automatic Identification of Bilingual Synonymous Technical Terms (%)

		Precision	Recall	F-measure
Baseline (t_J and s_J are identical, or, t_E and s_E are identical.)		67.0	54.3	60.8
SVM	Maximum Precision	**97.5**	28.7	43.9
	Maximum F-measure	73.5	68.1	**70.5**

Table 4: Examples of Improvement in Automatic Identification of Bilingual Synonymous Technical Terms by SVM

(a) Correctly Judging as "Synonym" only by SVM

seed $\langle s_J, s_E \rangle$	$\langle t_J, t_E \rangle$	Reference	Baseline	SVM (Maximum Precision)
〈ホールド回路, hold circuit〉	〈保持回路, holding circuit〉	synonym	not synonym	synonym

(b) Correctly Judging as "Not Synonym" only by SVM

seed $\langle s_J, s_E \rangle$	$\langle t_J, t_E \rangle$	Reference	Baseline	SVM (Maximum Precision)
〈転写器, transfer unit〉	〈搬送ユニット, transfer unit〉	not synonym	synonym	not synonym

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 $\langle t_J, t_E \rangle$ as synonymous with the seed bilingual technical term pair $s_{JE} = \langle s_J, s_E \rangle$ when t_J and s_J are identical, or, t_E and s_E 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. When maximizing precision, we achieve almost 98% precision where F-measure is over 40%. When maximizing F-measure, we achieve over 70% F-measure with over 73% precision and over 68% recall.

Table 4 also show examples of improving the baseline by SVMs.

Table 4 (a) shows the case of correctly judging as "synonym" only by the proposed method. Here, the baseline judges as "not synonym", since neither t_J and s_J nor t_E and s_E are identical. With the proposed method, on the other hand, f_{13} returns 1 since "holding" and "hold" have the same stem. Also, f_{16} returns 1 since, by the phrase tables, "ホールド回路" can be generated by translating "holding circuit", and "保持回路" can be generated by translating "hold circuit".

Table 4 (b) shows the case of correctly judging as "not synonym" only by the proposed method. Here, the baseline judges as "synonym", since t_E and s_E are identical. With the proposed method,

[4] As the existing Japanese-English bilingual lexicon, Eijiro (`http://www.eijiro.jp/`, Ver.79, with 1.6M translation pairs, is used.

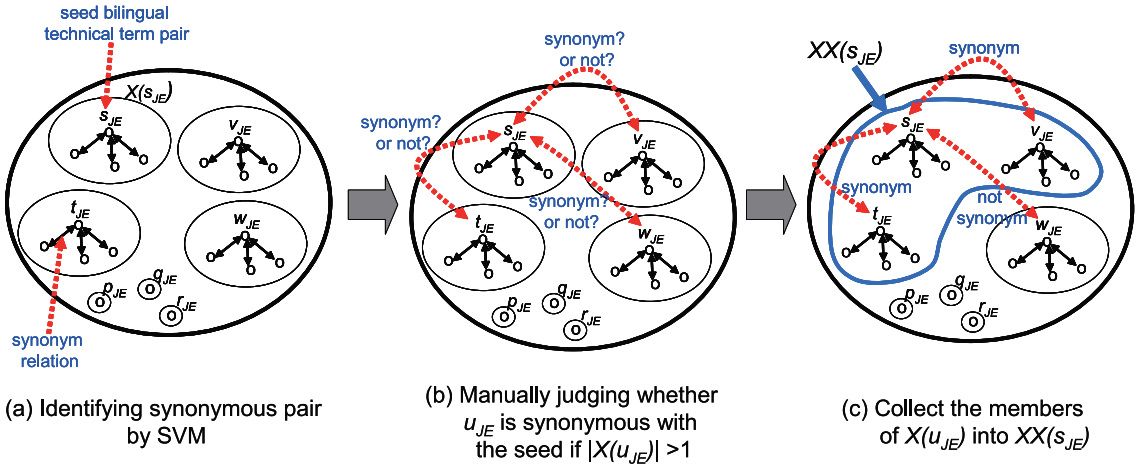

Figure 2: The Procedure of Semi-Automatic Transitive Identification of Bilingual Technical Terms

on the other hand, both edit distance similarity f_9 and character bigram similarity f_{10} return 0 for the Japanese terms "転写器" and "搬送ユニット". Also, f_{15} returns 0 since, by compositional translation with an existing bilingual lexicon, "転写器" cannot be generated by translating "transfer unit", nor "搬送ユニット" cannot be generated by translating "transfer unit".

6 Semi-Automatic Approach to Transitive Identification of Bilingual Synonymous Technical Terms

Evaluation results of the previous section is satisfactory in terms of precision of identifying bilingual synonymous technical terms. However, its recall is relatively low, which needs to be improved. In this section, we allow semi-automatic approach to the task of identifying bilingual synonymous technical terms. In this approach, we assume that the SVM classifier trained by the procedure of section 5.1 is tuned so that it can achieve high precision against relatively easier instances, while it judges relatively harder instances as not synonymous, resulting in relatively low recall. Based on this assumption, we design a manual process which is responsible for examining whether pairs of bilingual technical terms judged by the SVM classifier as not synonymous are not actually synonymous.

6.1 The Procedure

The detailed procedure of the semi-automatic approach is presented in this section.

1st step Suppose that we are given the SVM classifier trained by the procedure of section 5.1. Then, for a set $CBP(s_J)$ of candidates of bilingual synonymous technical terms, apply the SVM classifier to every pair $u_{JE} = \langle u_J, u_E \rangle$ and $v_{JE} = \langle v_J, v_E \rangle$ of the members of $CBP(s_J)$.

2nd step Next, for each member $u_{JE} = \langle u_J, u_E \rangle$ of $CBP(s_J)$, collect u_{JE} itself and other member $v_{JE} = \langle v_J, v_E \rangle$ of $CBP(s_J)$ judged as synonymous with u_{JE} by SVM into a set $X(u_{JE})$ (Figure 2 (a))[5].

$$X(u_{JE}) = \left\{ v_{JE} = \langle v_J, v_E \rangle \ (\in CBP(s_J)) \;\middle|\; v_{JE} = u_{JE}, \text{ or, } v_{JE} \text{ is judged as synonymous with } u_{JE} \text{ by SVM.} \right\}$$

[5] Here, if both u^1_{JE} and u^2_{JE} are judged as synonymous with v_{JE}, but u^1_{JE} and u^2_{JE} are judged as not synonymous, we simply include both u^1_{JE} and u^2_{JE} in $X(v_{JE})$.

3rd step For each bilingual technical term pair $u_{JE} = \langle u_J, u_E \rangle$ ($\in CBP(s_J)$), if $|X(u_{JE})| > 1$ holds[6], then manually examine whether u_{JE} is synonymous with the seed bilingual technical term pair s_{JE}, i.e., whether $u_{JE} \in SBP(s_{JE})$ holds (Figure 2 (b)). If so, then collect the members of $X(u_{JE})$ into $XX(s_{JE})$ (Figure 2 (c)).

Finally, the set $XX(s_{JE})$ can be regarded as the final output of the procedure for semi-automatic transitive identification of bilingual technical terms that are judged as synonymous with the seed bilingual technical term pair s_{JE}, and can be denoted as the following formula:

$$XX(s_{JE}) = \bigcup_{u_{JE} \in SBP(s_{JE})} X(u_{JE})$$

6.2 Evaluation Results

In the evaluation, for each of the 134 seed bilingual technical term pairs, we evaluate the precision, recall, and F-measure of the set $XX(s_{JE})$. As in the case of the procedure of section 5.1, with the development set, we tune the lower bound of the confidence measure as well as the parameter of TinySVM for trade-off between training error and margin, so that we can control the precision of the set $XX(s_{JE})$ as over 80%, 85%, 90%, and 95%. Evaluation results against the evaluation set are shown in Table 5. Here, we achieve over 95% precision with more than 50% recall, and over 90% precision with almost 70% recall. As can be clearly seen from these results, by simply transitively merging the results of identifying bilingual synonymous technical terms by SVM, we can improve the recall of bilingual synonym identification task[7].

Table 5: Evaluation Results of Transitive Identification of Bilingual Synonymous Technical Terms (%)

Requirement for Precision against the Development Set	Precision	Recall	F-measure
$> 80\%$	81.3	89.9	85.1
$> 85\%$	86.9	80.9	83.4
$> 90\%$	91.3	69.1	78.2
$> 95\%$	95.2	53.1	67.9

7 Related Works

Among related works on acquiring bilingual lexicon from text, Itagaki *et al.* (2007) focused on automatic validation of translation pairs available in the phrase table learned by a statistical machine translation model, where their study differs with this paper in that Itagaki *et al.* (2007) did not study the issue of synonymous bilingual 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 synonymous bilingual technical terms and that the features of machine learning studied in Tsunakawa and Tsujii (2008) are closely related those studied in this paper. However, Tsunakawa and Tsujii (2008) studied the issue of identifying synonymous bilingual technical terms only within manually compiled bilingual technical term lexicon and thus are quite limited in its applicability. Our study in this paper, 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 synonymous bilingual technical terms automatically.

[6] This condition means that at least one technical term pair v_{JE} is judged as synonymous with u_{JE} by SVM. Otherwise, we skip the process of manually examining the pair v_{JE}.

[7] In this evaluation, we simply measure the average number of the members u_{JE} of $SBP(s_{JE})$ for each of which $\left|X(u_{JE})\right| = 1$ holds. This number represents that, for how many members of $SBP(s_{JE})$, we can actually skip examining whether u_{JE} is synonymous with the seed bilingual technical term pair s_{JE}. Out of the average 12.5 members per seed, in Table 5, the numbers are 0.9 for "$>$ 80%", 1.4 for "$>$ 85%", 2.2 for "$>$ 90%", and 4.0 for "$>$ 95%", respectively.

Our study in this paper is also different from previous works on identifying synonyms based on bilingual and monolingual resources (e.g. Lin and Zhao (2003)) in that we learn synonymous bilingual technical terms from phrase tables of a phrase-based statistical machine translation model trained with very large parallel sentences.

In Liang *et al.* (2011), we proposed the framework of applying machine learning technique to the task of identifying bilingual synonymous technical terms in the process of acquiring technical term translation equivalent pairs from parallel patent documents. The major drawback of the framework of Liang *et al.* (2011) is in its low recall when preferring precision as over 90%. The framework proposed in Liang *et al.* (2011) is also employed in the first half of this paper, where the major contribution of this paper is in showing that the approach of manually merging synonym candidate sets identified by SVM in the first half of this paper is quite effective in improving low recall reported in Liang *et al.* (2011).

8 Conclusion

In the task of acquiring technical term translation equivalent pairs, this paper studied the issue of identifying synonymous translation equivalent pairs. We applied the SVMs to this task and achieved the performance of almost 98% precision and over 40% F-measure. Then, in order to improve recall, we simply introduced a semi-automatic framework, where we employed the strategy of selecting more than one seeds for each set of candidates bilingual synonymous term pairs. By manually judging whether each pair of two seeds is synonymous or not, we achieved over 95% precision and 50% recall. We are planning to incorporate the results of judgment by SVM when judging whether each pair of two seeds is synonymous or not.

References

Fujii, A., M. Utiyama, M. Yamamoto, and T. Utsuro. 2008. Toward the evaluation of machine translation using patent information. In *Proc. 8th AMTA*, pp. 97–106.

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

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

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

Koehn, P., 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*, pp. 177–180.

Liang, B., T. Utsuro, and M. Yamamoto. 2011. Identifying bilingual synonymous technical terms from phrase tables and parallel patent sentences. In *Proc. 12th PACLING,* #7.

Lin, D. and S. Zhao. 2003. Identifying synonyms among distributionally similar words. In *Proc. 18th IJCAI*, pp. 1492–1493.

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

Morishita, Y., 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*, pp. 153–162.

Och, F. J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 19–51.

Tonoike, M., 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*, pp. 11–18.

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

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

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