Compositional Translation of Technical Terms by Integrating Patent Families as a Parallel Corpus and a Comparable Corpus

Itsuki Toyota Zi Long Lijuan Dong
Grad. Sch. Sys. & Inf. Eng.,
University of Tsukuba,
Tsukuba, 305-8573, JAPAN

Takehito Utsuro Mikio Yamamoto
Fclty of Eng., Inf.& Sys.,
University of Tsukuba,
Tsukuba, 305-8573, JAPAN

Abstract

In the previous methods of generating bilingual lexicon from parallel patent sentences extracted from patent families, the portion from which parallel patent sentences are extracted is about 30% out of the whole "Background" and "Embodiment" parts and about 70% are not used. Considering this situation, this paper proposes to generate bilingual lexicon for technical terms not only from the 30% but also from the remaining 70% out of the whole "Background" and "Embodiment" parts. The proposed method employs the compositional translation estimation technique utilizing the remaining 70% as a comparable corpus for validating translation candidates. As the bilingual constituent lexicons in compositional translation, we use an existing bilingual lexicon as well as the phrase translation table trained with the parallel patent sentences extracted from the 30%. Finally, we show that about 3,600 technical term translation pairs can be acquired from 1,000 patent families.

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 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), transliteration (Knight and Graehl, 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 (2008) studied to acquire technical term translation lexicon from the phrase translation table, which are trained by a phrase-based statistical machine translation model with parallel sentences automatically extracted from patent families. 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. According to our personal communication with the organization, under a certain amount of budget for the labor of judging the correctness of bilingual technical term pairs suggested by the system, the organization collected about 500,000 bilingual technical term pairs per year. The orga-

Proceedings of the 5th Workshop on Patent Translation, Nice, September 2, 2013. Yokoyama, S., ed.

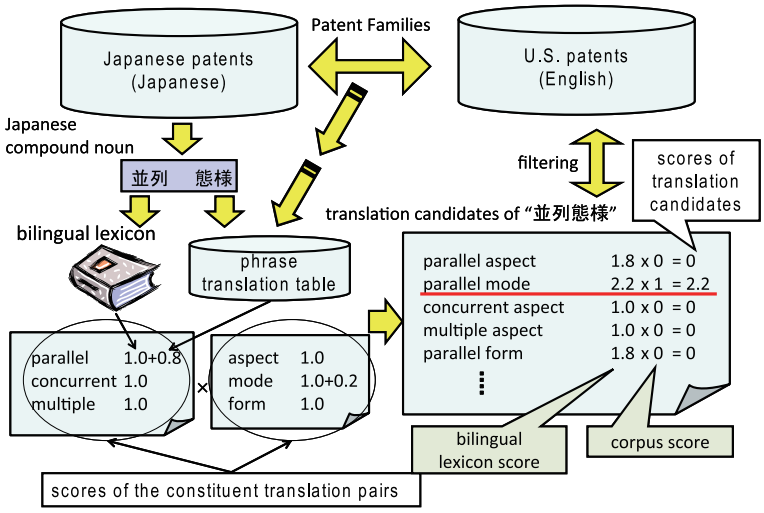

Figure 1: Proposed Framework of Compositional Translation Estimation for the Japanese Technical Term "並列態様" *(parallel mode)*

nization is also working on the task of compiling a Japanese-Chinese bilingual technical term lexicon from Japanese-Chinese patent families, where they claim that, under a certain amount of budget, they are able to compile 1,000,000 bilingual technical term pairs per year.

In Morishita (2008), the portion from which parallel patent sentences are extracted is composed of the parts of "Background" and "Embodiment". However, this portion is about 30% out of the whole "Background" and "Embodiment" parts and about 70% are not used. Considering this situation, this paper proposes to generate bilingual lexicon for technical terms not only from the 30% but also from the remaining 70% out of the whole "Background" and "Embodiment" parts. As shown in Figure 1, the proposed method employs the compositional translation estimation technique utilizing the remaining 70% as a comparable corpus for selecting translation candidates that actually appear in the target language side of the comparable corpus. As the bilingual constituent lexicons, the compositional translation procedure uses an existing bilingual lexicon as well as the phrase translation table trained with the parallel patent sentences extracted from the 30%. Through the experimental evaluation, we show that about 3,600 technical term translation pairs can be acquired from 1,000 patent families.

2 Related Work

Lu and Tsou (2009) and Yasuda and Sumita (2013) studied to extract bilingual terms from comparable patents, where, as we studied in Morishita (2008), they first extract parallel sentences from comparable patents, and then extract bilingual terms from parallel sentences. As we discussed in section 1, in this paper, we concentrate on generating bilingual lexicon for technical terms not only from the parallel patent sentences extracted from patent families, but also from the remaining parts of patent families.

Liang et al. (2011) considered situations where a technical term is observed in many parallel patent sentences and is translated into many translation equivalents. They then studied the issue of identifying synonymous translation equivalent pairs. The technique proposed in this paper can be easily integrated into the achievement presented in Liang et al. (2011) in the task of identifying synonymous translation equivalent pairs.

The task of translation term pair acquisition from comparable corpora (e.g., (Fung and Yee, 1998)) has been well studied, where most of those works rely on measuring contextual similarity of translation term pair candidates across two languages. Compared with those techniques, our proposed method relies on the compositional translation approach utilizing patent families. Patent families can be regarded as a partially parallel and partially comparable corpus, where a relatively large portion of technical terms are compositionally translated across two languages, and in those cases, translation candidates can be easily detected without introducing contextual similarity.

3 Japanese-English Patent Families

In the NTCIR-7 workshop, the Japanese-English patent translation task is organized (Fujii et al., 2008), where patent families and sentences are provided by the organizer. Those patent families 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 Uchiyama and Isahara (2007) is applied to the text of those fields, and Japanese and English sentences are aligned (about 1.8M sentences in total).

4 Compositional Translation of Technical Terms

As the procedure of compositional translation of technical terms, translation candidates of a term are compositionally generated by concatenating the translation of the constituents of the term (Tonoike et al., 2006)[1] [2].

4.1 Bilingual Constituents Lexicons

First, the following sections describe the bilingual lexicons we use for translating constituents of technical terms, where Table 1 shows the numbers of entries and translation pairs in those lexicons.

[1]Tonoike et. al (2006) studied how to compositionally translate technical terms using an existing bilingual lexicon as well as bilingual constituent lexicons constructed from the constituents collected from the existing bilingual lexicon. Compared to Tonoike et. al (2006), this paper proposes how to optimally incorporate constituent translation pairs collected from the phrase translation table trained with the parallel patent sentences introduced in section 3 into the procedure of compositional translation.

[2]As "constituents", we do not consider "syntactic constituents", but simply consider a word or a sequence of two or more consecutive words.

4.1.1 A Bilingual Lexicon (Eijiro) and its Constituent Lexicons

As an existing Japanese-English translation lexicon for human use, we use Eijiro (`http://www.eijiro.jp/`, We merged two versions Ver.79 and Ver. 131.).

We also compiled bilingual constituents lexicons from the translation pairs of Eijiro. Here, we first collect translation pairs whose English terms and Japanese terms consist of two constituents into another lexicon P_2. We compile the "bilingual constituents lexicon (prefix)" from the first constituents of the translation pairs in P_2 and compile the "bilingual constituents lexicon (suffix)" from their second constituents[3].

4.1.2 Phrase Translation Table of an SMT Model

As a toolkit of a phrase-based statistical machine translation model, we use Moses (Koehn and others, 2007) and apply it to the whole 1.8M parallel patent sentences described in section 3. 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 translation table and a phrase translation probability is assigned to each pair (Koehn et al., 2003). 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 translation table are ranked in descending order of Japanese to English phrase translation probabilities.

4.2 Score of Translation Candidates

This section gives the definition of the score of a translation candidate in compositional translation.

First, let y_S be a technical term whose translation is to be estimated. We assume that y_S is de-

[3]Tonoike et. al (2006) reported that those two bilingual constituent lexicons compiled from the translation pairs of Eijiro improved the coverage of compositional translation from 49% up to 69%.

Table 1: Numbers of Entries and Translation Pairs in Lexicons

lexicon	# of entries		# of translation pairs
	English	Japanese	
Eijiro	1,631,099	1,847,945	2,244,117
bilingual constituents lexicon (prefix) B_P	47,554	41,810	129,420
bilingual constituents lexicon (suffix) B_S	24,696	23,025	82,087
phrase translation table	33,845,218	33,130,728	76,118,632

composed into their constituents as below:

$$y_S = s_1, s_2, \cdots, s_n \quad (1)$$

where each s_i is a single word or a sequence of words. For y_S, we denote a generated translation candidate as y_T:

$$y_T = t_1, t_2, \cdots, t_n \quad (2)$$

where each t_i is a translation of s_i, and is also a single word or a sequence of words independently of s_i. Then the translation pair $\langle y_S, y_T \rangle$ is represented as follows[4].

$$\langle y_S, y_T \rangle = \langle s_1, t_1 \rangle, \langle s_2, t_2 \rangle, \cdots, \langle s_n, t_n \rangle \quad (3)$$

The score of a generated translation candidate y_T is defined as the product of a bilingual lexicon score and a corpus score as follows.

$$\prod_{i=1}^{n} q(\langle s_i, t_i \rangle) \cdot Q_{corpus}(y_T) \quad (4)$$

The bilingual lexicon score $\prod_{i=1}^{n} q(\langle s_i, t_i \rangle)$ is represented as the product of the score $q(\langle s_i, t_i \rangle)$ of a constituent translation pair $\langle s_i, t_i \rangle$, while the corpus score is denoted as $Q_{corpus}(y_T)$. Here, the bilingual lexicon score measures the appropriateness of the translation of each constituent pair $\langle s_i, t_i \rangle$ referring to bilingual lexicons provided as a resource for term translation, while the corpus score measures the appropriateness of the translation candidate y_T based on the occurrence of y_T in a given target language corpus.

More specifically, when the technical term y_S of the source language is decomposed into a sequence of constituents, the variation of the constituent sequence could be more than one. Then, this situation could lead to the case where a translation candidate y_T can be generated from more than one variations of the constituent sequence $s_1, s_2, \cdots, s_n$ of y_S. Considering such a situation, the overall score $Q(y_S, y_T)$ of the translation pair $\langle y_S, y_T \rangle$ is denoted as the sum of the score for each variation of the constituent sequence $s_1, s_2, \cdots, s_n$ of y_S.

$$Q(y_S, y_T) = \sum_{y_S = s_1, s_2, \ldots, s_n} \prod_{i=1}^{n} q(\langle s_i, t_i \rangle) \cdot Q_{corpus}(y_T) \quad (5)$$

4.2.1 Bilingual Lexicon Score

The bilingual lexicon score $q(\langle s, t \rangle)$ of a constituent translation pair $\langle s, t \rangle$ is defined as the sum of the score q_{man} for the pairs included in Eijiro, B_P, or B_S, as well as the score q_{smt} for those included in the phrase translation table:

$$q(\langle s, t \rangle) = q_{man}(\langle s, t \rangle) + q_{smt}(\langle s, t \rangle)$$

$$q_{man}(\langle s, t \rangle) = \begin{cases} 1 & \text{(if } \langle s, t \rangle \text{ in Eijiro, or } B_P \text{, or } B_S\text{)} \\ 0 & \text{(otherwise)} \end{cases}$$

$$q_{smt}(\langle s, t \rangle) = \begin{cases} P(t|s) & \text{(if } \langle s, t \rangle \text{ in the phrase translation table and } P(t|s) \geq p_0\text{)} \\ 0 & \text{(otherwise)} \end{cases}$$

In this definition, When the pair $\langle s, t \rangle$ is in Eijiro, B_P, or B_S, the score $q_{man}(\langle s, t \rangle)$ is defined as 1, while it is defined as 0 otherwise[5]. When the pair $\langle s, t \rangle$ is in the phrase translation table, on the other hand, we introduce the lower bound p_0 of

[4] Those bilingual constituents lexicons we introduced in section 4.1 have both single word entries and compound word entries. Thus, each constituent translation pair $\langle s_i, t_i \rangle$ could be not only one word to one word, but also one word to multi words, or multi words to multi words.

[5] In Tonoike et. al (2006), the score $q_{man}(\langle s, t \rangle)$ is defined to be a function of the number of constituents in s and t when the pair $\langle s, t \rangle$ is included in Eijiro, while it is defined to be a function of the frequency of the pair $\langle s, t \rangle$ in Eijiro when the pair is included in B_P or B_S. However, in our preliminary tuning phase, this definition achieves almost the same performance than the one we present in this paper. Thus, we prefer a simpler definition of q_{man} in this paper.

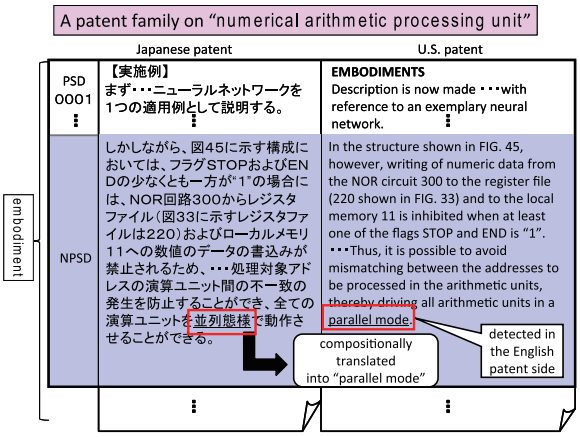

Figure 2: An Example of "Embodiment" Part with No Parallel Sentences Extracted

the translation probability. In this definition, when the translation probability $P(t|s)$ is more than or equal to the lower bound p_0 ($P(t|s) \geq p_0$), then the score $q_{smt}(\langle s, t\rangle)$ is defined as $P(t|s)$, while it is defined as 0 otherwise. In the evaluation in section 6, the parameter p_0 is optimized with a tuning data set other than the evaluation set.

4.2.2 Corpus Score

The corpus score measures whether the translation candidate y_T does appear in a given target language corpus:

$$Q_{corpus}(y_T) = \begin{cases} 1 & y_T \text{ occurs in the corpus of the target language} \\ 0 & y_T \text{ does not occur in the corpus of the target language} \end{cases} \quad (6)$$

5 Translation Estimation with the Part of No Parallel Sentences Extracted as a Comparable Corpus

This section describes how to estimate translation of technical terms using the part of patent families from which no parallel sentences are extracted, regarding it as a comparable corpus.

First, as we denote below, the Japanese part D_J of a Japanese-English patent family consists of the "Background of the Invention" part B_J, the "Detailed Description of the Preferred Embodiments" part M_J, and the rest N_J. B_J and M_J are then decomposed into the part PSD_J from which parallel sentences are extracted, and that $NPSD_J$ from which parallel sentences are NOT extracted. Similarly, the English part D_E of a Japanese-English patent family consists of the "Background of the Invention" part B_E, the "Detailed Description of the Preferred Embodiments" part M_E, and the rest N_E. B_E and M_E are then decomposed into the part PSD_E from which parallel sentences are extracted, and that $NPSD_E$ from which parallel sentences are NOT extracted. Figure 2 shows an example of "Embodiments" part, along with its PSD part and $NPSD$ part.

$$\begin{aligned} D_J &= \langle B_J, M_J, N_J\rangle \\ B_J \cup M_J &= \langle PSD_J, NPSD_J\rangle \\ D_E &= \langle B_E, M_E, N_E\rangle \\ B_E \cup M_E &= \langle PSD_E, NPSD_E\rangle \end{aligned}$$

In this paper, we extract a Japanese technical term t_J to translate into English from $NPSD_J$. This is mainly because we assume that Japanese technical terms appearing in PSD_J are expected to be translated into English by referring to the phrase translation table trained with parallel sentences extracted from PSD_J and PSD_E.

Then, considering the "Background" part B_E and the "Embodiment" part M_E in the English side as the target language corpus, we apply the compositional translation procedure of section 4 to t_J and collect the candidates of English translation which have the positive score $Q(t_J, t_E)$ into the set $TranCand\Big(t_J, B_E \cup M_E\Big)$:[6]

$$\begin{aligned} &TranCand\Big(t_J, B_E \cup M_E\Big) \\ &= \Big\{ t_E \in B_E \cup M_E \;\Big|\; t_J \text{ is compositionally translated into } t_E \text{ by the procedure of section 4 and (equation (5)) } Q(t_J,t_E) > 0 \Big\} \end{aligned}$$

Finally, out of the set $TranCand(t_J,\ B_E \cup M_E)$ of the translation candidates, we have t_E with the maximum score by the following function

[6] As the target language corpus, we also evaluate the part $NPSD_E$ (of B_E and M_E) from which parallel sentences are NOT extracted. However, in this case, we had a lower rate of correctly matching the translation candidates in the target language corpus. From this result, we prefer to have B_E and M_E as the target language corpus.

Table 2: Classification of the Japanese Compound Nouns in the 1,000 Japan-US Patent Families

(1) for the whole 61,133 Japanese noun phrases

Categories		Bilingual Constituent Lexicons		
		Eijiro ONLY	phrase translation table ONLY	Eijiro AND phrase translation table
(a)	Its English translation listed in Eijiro appears in the target language corpus	5,449 (8.9%)		
(b)	Included in the phrase translation table as one of the Japanese entries	32,516 (53.2%)		
(c)	Its compositional English translation (by the proposed method) appears in the target language corpus	4,004 (6.6%) (set E)	14,310 (23.4%) (set P, when maximizing $\|P\|$ ($p_0 = 0$))	14,575 (23.8%) (set EP, when maximizing $\|EP\|$ ($p_0 = 0$))
(d)	An English translation can be generated by Eijiro or compositional translation (by the proposed method), which does not appear in the target language corpus	397 (0.6%)	993 (1.6%)	1,041 (1.7%)
(e)	No English translation can be generated by Eijiro nor compositional translation (by the proposed method)	18,767 (30.7%)	7,865 (12.9%)	7,552 (12.4%)
total		61,133 (100%)		

(2) the set of whole 61,133 Japanese noun phrases $-$ the set (a) $-$ the set (b) $-$ the set E

Categories		Bilingual Constituent Lexicons	
		phrase translation table ONLY	Eijiro AND phrase translation table
(c)	Its compositional English translation (by the proposed method) appears in the target language corpus	10,375 (17.0%) (set $P - (E \cap P)$)	10,571 (17.3%) (set $EP - (E \cap EP)$)

$TranCand(t_J,\ B_E \cup M_E)$.

$$\mathrm{CompoTrans}_{\max}(t_J, B_E \cup M_E) = \underset{t_E \in TranCand(t_J,\ B_E \cup M_E)}{\arg\max} Q(t_J, t_E)$$

6 Evaluation

In order to evaluate the proposed method, we compare the following three cases:

(i) *Eijiro ONLY* $\cdots$ As bilingual constituents lexicons, Eijiro and its constituent lexicons are employed.

(ii) *Phrase translation table ONLY* $\cdots$ As bilingual constituents lexicons, the phrase translation table is employed.

(iii) *Eijiro AND phrase translation table* $\cdots$ As bilingual constituents lexicons, Eijiro and its constituent lexicons as well as the phrase translation table are employed.

First, we pick up 1,000 patent families, from which we extract 61,133 Japanese noun phrases. Then, we apply the compositional translation procedure of section 4 to those 61,133 Japanese noun phrases, and classify them into the following five categories (as shown in Table 2-(1)):

(a) The Japanese noun phrase is included in Eijiro as one of the Japanese entries, and its English translation appears in the target language corpus.

(b) The Japanese noun phrase is not in (a), and is included in the phrase translation table as one of the Japanese entries.

(c) The Japanese noun phrase is not in (a) nor (b), and by applying the proposed method of compositional translation to it, its English translation appears in the target language corpus.

(d) The Japanese noun phrase is not in (a), (b),

Table 3: Result of Evaluating Compositional Translation and Estimated Numbers of Bilingual Technical Term Translation Pairs to be acquired by the Proposed Method (per 1,000 Patent Families)

(1) for each case of bilingual constituent lexicons in compositional translation

	Bilingual Constituent Lexicons		
	Eijiro ONLY	phrase translation table ONLY	Eijiro AND phrase translation table
Evaluation Sets	$E' \subset E$, $\|E'\| = 93$	$P' \subset P - (E \cap P)$, $\|P'\| = 224$	$EP' \subset EP - (E \cap EP)$, $\|EP'\| = 230$
recall (%) precision (%) F-measure (%)	97.8 97.8 97.8	30.1 / 88.3 / 44.9 ($p_0 = 0.07$, when maximizing precision with recall $> 20\%$)	32.6 / 93.8 / 48.4 ($p_0 = 0.15$, when maximizing precision with recall $> 30\%$)
estimated numbers of term translation pairs	1,957 ($= 4,004 \times 0.5 \times 0.978$) (for the set E, $\|E\| = 4,004$)	1,561 ($= 10,375 \times 0.5 \times 0.301$) (for the set $P - (E \cap P)$, $\|P - (E \cap P)\| = 10,375$)	1,723 ($= 10,571 \times 0.5 \times 0.326$) (for the set $EP - (E \cap EP)$, $\|EP - (E \cap EP)\| = 10,571$)

(2) for the whole 61,133 Japanese noun phrases

	translation estimation for the set E with Eijiro ONLY + translation estimation for the set $P - (E \cap P)$ with phrase translation table ONLY	translation estimation for the set E with Eijiro ONLY + translation estimation for the set $EP - (E \cap EP)$ with Eijiro AND phrase translation table
estimated numbers of term translation pairs	3,518 (= 1,957+1,561)	3,680 (= 1,957+1,723)

nor (c), and from it, an English translation can be generated by Eijiro or by the proposed method of compositional translation, while the English translation does not appear in the target language corpus.

(e) The Japanese noun phrase is not in (a), (b), (c), nor (d), and from it, no English translation can be generated by Eijiro nor by the proposed method of compositional translation, simply because one or more constituents of the Japanese noun phrase can not be found in any constituent lexicons.

As in Table 2-(1), the number of the Japanese noun phrases of category (c) is 4,004 when *Eijiro ONLY* (denoted as the "set E"). The number is 14,310 when *phrase translation table ONLY* and the lower bound p_0 of the translation probability is equal to 0 (denoted as the "set P"), which becomes about 3.5 times larger. Furthermore, the number is 14,575 when *Eijiro AND phrase translation table* and the lower bound p_0 of the translation probability is equal to 0 (denoted as the "set EP"), which then becomes about 3.6 times larger compared with the set E.

Next, Table 3 shows the results of measuring recall / precision / F-measure of the proposed method, where we compare the three cases of bilingual constituent lexicons. First, we construct evaluation sets E', P', and EP' from the sets E, $P - (E \cap P)$, and $EP - (E \cap EP) = EP - E$, respectively[7]. Since we can mostly correctly estimate translation of the Japanese compound nouns within the set E when *Eijiro ONLY*, we exclude those members of E from the evaluation sets P' and EP'. Second, with tuning data sets other than those evaluation sets P' and EP', we optimize the

[7]We examined the sets E, $P - (E \cap P)$, and $EP - (E \cap EP) = EP - E$ in advance, and found that only 50% of their members are Japanese technical terms, while the remaining 50% consist of general compound nouns other than technical terms, terms with errors in segmentation of morphemes, and those not translated in the English patent side in the patent family. Thus, we construct the evaluation sets E', P', and EP' only from the Japanese technical terms portion of E, $P - (E \cap P)$, and $EP - (E \cap EP)$, i.e., 50% of them.

lower bound p_0 of the translation probability individually for both P' and EP'. Requiring that the recall is to be around 20∼30%, while the precision is to be around 80∼90%, we have the lower bounds p_0 as 0.07 for P' and as 0.15 for EP'

As shown in Table 3-(1), for the evaluation set E', we achieve high recall / precision / F-measure (97.8%), and the estimated number of technical term translation pairs to be acquired is more than 1,900[8]. This result is very impressive compared with the relatively low recalls when incorporating the phrase translation table as a bilingual constituent lexicon (30.1% for the set P' and 32.6% for the set EP'). This is simply because we restrict translation pairs within the phrase translation table by introducing the lower bounds p_0 of the translation probability. Consequently, we achieve the precisions to be around 80∼90% and satisfy the requirement of the procedure of manual judgement on accepting / ignoring the candidates. The estimated number of technical term translation pairs to be acquired is more than 1,500 for the evaluation set P' and is more than 1,700 for EP'. In total, for the set EP, we can acquire more than 3,600 novel technical term translation pairs per 1,000 patent families. Note that, in this procedure, acceptance rate of the manual judgement is over 95%, which is reasonably high.

7 Conclusion

This paper proposed to generate bilingual lexicon for technical terms not only from the parallel patent sentences extracted from patent families, but also from the remaining parts of patent families. The proposed method employed the compositional translation estimation technique utilizing the remaining parts as a comparable corpus for validating translation candidates. As the bilingual constituent lexicons in compositional translation, we used an existing bilingual lexicon as well as the phrase translation table trained with the parallel patent sentences extracted from the patent families. Finally, we showed that about 3,600 technical term translation pairs can be acquired from 1,000 patent families. Future works include applying an SMT technique straightforwardly to the task of technical term translation and comparing its performance with the compositional translation technique presented in this paper. We believe that the proposed framework of validating translation candidates is also effective with an SMT technique.

[8]Here, we suppose that we manually judge whether the translation candidates provided by the proposed method is correct or not and accept the correct ones while ignore the incorrect ones. We also assume that we can automatically or manually select Japanese technical terms (50%) from the whole set of compound nouns.

References

Fujii, A., M. Utiyama, M. Yamamoto, and T. Utsuro. 2008. Toward the evaluation of machine translation using patent information. In *Proc. 8th AMTA*, pages 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*, pages 414–420.

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

Knight, K. and J. Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.

Koehn, P. et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. 45th ACL, Companion Volume*, pages 177–180.

Koehn, P., F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. HLT-NAACL*, pages 127–133.

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

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

Matsumoto, Y. and T. Utsuro. 2000. Lexical knowledge acquisition. In Dale, R., H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, chapter 24, pages 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*, pages 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*, pages 11–18.

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

Yasuda, K. 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.