Identifying Japanese-Chinese Bilingual Synonymous Technical Terms from Patent Families

Zi Long[†] **Lijuan Dong**[†] **Takehito Utsuro**[†] **Tomoharu Mitsuhashi**[‡] **Mikio Yamamoto**[†]

†Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, 305-8573, Japan
‡ Japan Patent Information Organization, 4-1-7, Toyo, Koto-ku, Tokyo, 135-0016, Japan

Abstract

In the task of acquiring Japanese-Chinese technical term translation equivalent pairs from parallel patent documents, this paper considers situations where a technical term is observed in many parallel patent sentences and is translated into many translation equivalents and studies the issue of identifying synonymous translation equivalent pairs. 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, and achieve the performance of over 85% precision and over 60% F-measure. We further examine two types of segmentation of Chinese sentences, i.e., by characters and by morphemes, and integrate those two types of segmentation in the form of the intersection of SVM judgments, which achieved over 90% precision.

Keywords: synonymous technical terms, patent families, technical term translation

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 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. In more recent studies, they 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. Furthermore, based on the achievement above, Liang et al. (2011a) considered 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 Japanese-English technical term translation equivalent pairs, Liang et al. (2011a) studied the issue of identifying Japanese-English synonymous translation equivalent pairs. First, they collect candidates of synonymous translation equivalent pairs from parallel patent sentences. Then, they apply the Support Vector Machines (SVMs) (Vapnik, 1998) to the task of identifying bilingual synonymous technical terms.

Based on the technique and the results of identifying Japanese-English synonymous translation equivalent pairs in Liang et al. (2011a), we aim at identifying Japanese-Chinese synonymous translation equivalent pairs from Japanese-Chinese patent families. We especially examine two types of segmentation of Chinese sentences, namely, by characters and by morphemes. Although both types of segmentation achieved almost similar performance around 95∼97% (in recall / precision / f-measure) in the task of acquiring Japanese-Chinese technical term translation pairs, they have different types of errors. Also in the task of identifying Japanese-Chinese synonymous technical terms, both types of segmentation achieved almost similar performance, while they have different types of errors. Thus, we integrate those two types of segmentation in the form of the intersection of SVM judgments, and show that this achieves over 90% precision.

2. Japanese-Chinese Parallel Patent Documents

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

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[2] with the

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

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

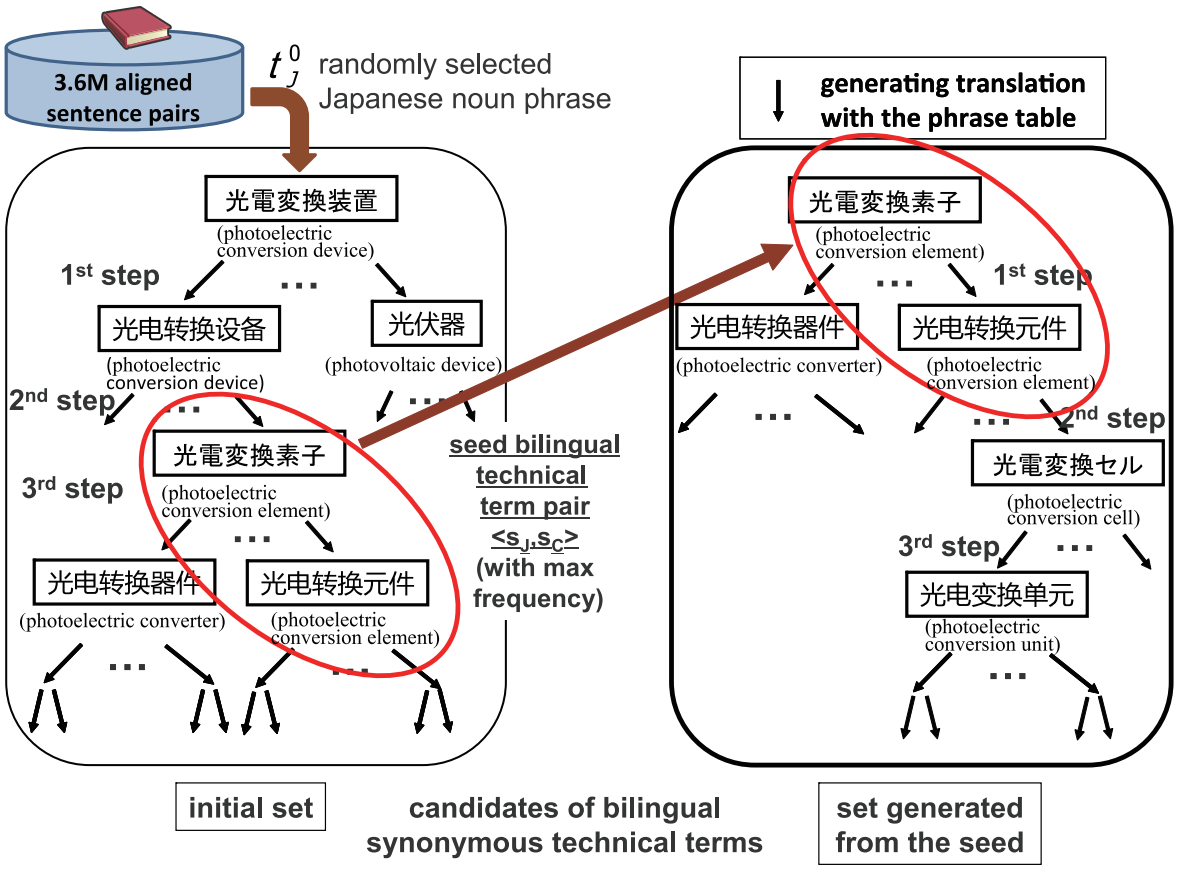


Figure 1: Developing a Reference Set of Bilingual Synonymous Technical Terms

morpheme lexicon IPAdic[3]. For Chinese sentences, we examine two types of segmentation, i.e., segmentation by characters[4] and segmentation by morphemes[5].

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, we finally obtain 108M (Chinese sentences segmented by morphemes) / 274M (Chinese sentences segmented by characters) translation pairs with 75M / 197M 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. 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, 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 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 Liang et al. (2011a)), 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 six times and generate the set $CBP(s_J)$ of candidates of bilingual synonymous technical term pairs. Figure 1 illustrates the whole procedure.

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 bilin-

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

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

[5] 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.

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

(a) 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 (a)	8,816	22,563	77.3	197.92
	included in the intersection of the sets (a) and (b)	13,747		120.6	
Reference of Synonyms $\bigcup_{s_{JC}} SBP(s_{JC})$	included only in the set (a)	309	2,496	2.7	21.9
	included in the intersection of the sets (a) and (b)	2,187		19.2	

(b) 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 (b)	14,161	28,948	124.2	253.9
	included in the intersection of the sets (a) and (b)	14,787		129.7	
Reference of Synonyms $\bigcup_{s_{JC}} SBP(s_{JC})$	included only in the set (b)	180	2,604	1.6	22.8
	included in the intersection of the sets (a) and (b)	2,424		21.3	

gual technical term pairs is around 2,500 to 2,600, which amounts to around 22 per seed on average. It can be also seen from Table 1 that although about 90% of reference of synonymous technical terms are shared by the two types of segmentation (by characters and by morphemes), only about 40% to 50% of candidates of synonymous technical terms are shared by the two types of segmentation.

5. Identifying Bilingual Synonymous Technical Terms by Machine Learning

In this section, we apply the SVMs 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[6] 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[7]. 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 $f_7, \ldots, f_{16}$ represent relation of the input bilingual technical term $\langle t_J, t_C \rangle$ and the

[6] 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.

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

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 haing the word "的" 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}: 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.

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).

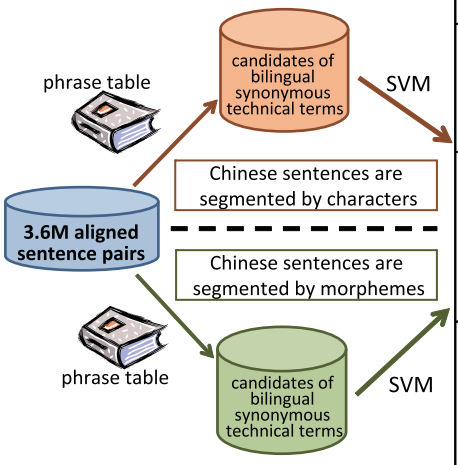
Among the features of the second type are identity of monolingual terms (f_7 and f_8), edit distance of monolingual terms (f_9), character bigram similarity of monolingual terms (f_{10}), rate of identical morphemes (in Japanese, f_{11}) / characters (in Chinese, f_{12}), string subsumption and variants for Japanese (f_{13}), identical stem for Chinese (f_{14}), rate of intersection in translation by the phrase table (f_{15}), and translation by the phrase tables (f_{16}).

5.3. Evaluation Results

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. When maximizing precision, we achieve almost 87% precision where F-measure is over 40%. When maximizing F-measure, we achieve over 60% F-measure with around 71% precision and over 52% recall. As shown in Figure 2, the two types of segmentation of Chinese sentences, namely, by characters and by morphemes, tend to have different types of errors. So, we integrate those two types of segmentation in the form of the intersection of

Table 3: Evaluation Results (%)

		segmented by characters			segmented by morphemes			intersection		
		precision	recall	f-measure	precision	recall	f-measure	precision	recall	f-measure
baseline (t_J and s_J are identical, or, t_C and s_C are identical.)		71.5	39.4	50.8	69.1	40.0	50.7	77.3	33.1	46.3
SVM	maximum precision	**86.9**	26.0	40.0	84.3	24.5	38.0	**90.0**	25.1	39.2
	maximum f-measure	71.0	52.8	60.6	68.6	54.4	**60.7**	—	—	—



<t_J,t_C>	seed <s_J,s_C>	judgment by SVM: segmentation by characters	judgment by SVM: segmentation by morphemes	intersection	reference
<印刷/回路/基板, 印/刷/电/路/基/板 or 印刷/电路/基板> (printed circuit board)	<プリント/回路/基板, 印/刷/电/路/板 or 印刷/电路/板> (printed circuit board)	synonym	synonym	synonym	synonym
<電磁/駆動/装置 (electromagnetic driving device) , 电/池/驱/动/装/置 or 电池/驱动/装置 (battery operated device)>	<電磁/駆動/装置, 电/磁/驱/动/装/置 or 电磁/驱动/装置> (electromagnetic driving device)	synonym	not synonym	not synonym	not synonym
<流体/導管 (fluid conduit), 流/体/传/导 or 流体/传导 (fluid transmission)>	<流体/導管, 流/体/管/道 or 流体/管道> (fluid conduit)	not synonym	synonym	not synonym	not synonym
... ...					

Figure 2: Evaluating Intersection of Judgments by SVM based on Character/Morpheme based Segmentation of Chinese Sentences

SVM judgments, where, for both types of segmentation, we tune the lower bound of the confidence measure of the distance from the separating hyperplane. We maximize precision while keeping recall over 25% with held-out data, and this achieves over 90% precision as shown in Table 3.

6. Related Work

Among related works on acquiring bilingual lexicon from text, Itagaki et al. (2007) focused on automatic validation of translation pairs available in the phrase table trained by an SMT model. Lu and Tsou (2009) and Yasuda and Sumita (2013) also 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.

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 bilingual synonymous technical terms from phrase tables of a phrase-based SMT model trained with very large parallel sentences. Also in the context of SMT between Japanese and Chinese, Sun and Lepage (2012) pointed out that character-based segmentation of sentences contributed to improving machine translation performance compared to morpheme-based segmentation of sentences.

7. Conclusion

In the task of acquiring Japanese-Chinese technical term translation equivalent pairs from parallel patent documents, this paper considered situations where a technical term is observed in many parallel patent sentences and is translated into many translation equivalents and studied the is-

sue of identifying synonymous translation equivalent pairs. We especially examined two types of segmentation of Chinese sentences, i.e., by characters and by morphemes, and integrated those two types of segmentation in the form of the intersection of SVM judgments, which achieved over 90% precision. One of the most important future works 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. We plan to examine whether this semi-automatic framework is also effective in the task of identifying Japanese-Chinese synonymous translation equivalent pairs.

8. References

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.

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

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.

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.

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

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.

J. Sun and Y. Lepage. 2012. Can word segmentation be considered harmful for statistical machine translation tasks between Japanese and Chinese? In *Proc. 26th PACLIC*, pages 351–360.

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.