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. 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 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.

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 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 Ishihara (2007) is applied 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 with the

†Here, we used a Japanese-Chinese translation lexicon consisting of about 170,000 Chinese head words.
‡http://mecab.sourceforge.net/
Figure 1: 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 \(s \equiv (s_J, s_C)\), 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 \(SBP(s_J, s_C)\), those of which are synonymous with \(s_J\), and the remaining \(NSBP(s_J, s_C)\). As in Table 1, we collect 114 seeds, where the number of bilingual technical terms included in \(SBP(s_J, s_C)\) is at least 108M in total for all of the 114 seed bilin-
The number of bilingual 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 \( (t_j, t_C) \) is synonymous with the seed bilingual technical term pair \( s_{JC} = (s_j, s_C) \).

#### 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 \( (t_j, t_C) \), while those of the second type \( f_7, \ldots, f_{16} \) represent relation of the input bilingual technical term \( (t_j, t_C) \) and the 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.

---

\(^6\)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 comparatively equal.

\(^7\)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

<table>
<thead>
<tr>
<th>class</th>
<th>feature</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>features for bilingual technical terms ( (t_J, t_C) )</td>
<td>( f_1 ): frequency</td>
<td>\log of the frequency of ( \langle t_J, t_C \rangle ) within the whole parallel patent sentences</td>
</tr>
<tr>
<td>features for bilingual technical terms ( (t_J, t_C) )</td>
<td>( f_2 ): rank of the Chinese term</td>
<td>\text{given } t_J, \log \text{ of the rank of } t_C \text{ with respect to the descending order of the conditional translation probability } P(t_C \mid t_J)</td>
</tr>
<tr>
<td>features for bilingual technical terms ( (t_J, t_C) )</td>
<td>( f_3 ): rank of the Japanese term</td>
<td>\text{given } t_C, \log \text{ of the rank of } t_J \text{ with respect to the descending order of the conditional translation probability } P(t_J \mid t_C)</td>
</tr>
<tr>
<td>features for bilingual technical terms ( (t_J, t_C) )</td>
<td>( f_4 ): number of Japanese characters</td>
<td>\text{number of characters in } t_J</td>
</tr>
<tr>
<td>features for bilingual technical terms ( (t_J, t_C) )</td>
<td>( f_5 ): number of Chinese characters</td>
<td>\text{number of characters in } t_C</td>
</tr>
<tr>
<td>features for the relation of bilingual technical terms ( (t_J, t_C) ) and the seed ( (s_J, s_C) )</td>
<td>( f_6 ): number of times generating translation by applying the phrase tables</td>
<td>\text{the number of times repeating the procedure of generating translation by applying the phrase tables until generating } t_C \text{ or } t_J \text{ from } s_J, \text{ as in } s_C \rightarrow \cdots \rightarrow t_J \rightarrow t_C, \text{ or, } s_J \rightarrow \cdots \rightarrow t_C \rightarrow t_J</td>
</tr>
<tr>
<td>features for the relation of bilingual technical terms ( (t_J, t_C) ) and the seed ( (s_J, s_C) )</td>
<td>( f_7 ): identity of Japanese terms</td>
<td>\text{return } 1 \text{ when } t_J = s_J</td>
</tr>
<tr>
<td>features for the relation of bilingual technical terms ( (t_J, t_C) ) and the seed ( (s_J, s_C) )</td>
<td>( f_8 ): identity of Chinese terms</td>
<td>\text{return } 1 \text{ when } t_C = s_C</td>
</tr>
<tr>
<td>features for the relation of bilingual technical terms ( (t_J, t_C) ) and the seed ( (s_J, s_C) )</td>
<td>( f_9 ): edit distance similarity of monolingual terms</td>
<td>\text{(f_9(t_X, s_X) = 1 - \frac{ED(t_X, s_X)}{\max</td>
</tr>
<tr>
<td>features for the relation of bilingual technical terms ( (t_J, t_C) ) and the seed ( (s_J, s_C) )</td>
<td>( f_{10} ): character bigram similarity of monolingual terms</td>
<td>\text{(f_{10}(t_X, s_X) = \frac{\text{bigram}(t_X) \cap \text{bigram}(s_X)}{\text{max}(\text{bigram}(t_X), \text{bigram}(s_X))} - 1) (where (\text{bigram}(t)) is the set of character bigrams of the term (t).)</td>
</tr>
<tr>
<td>features for the relation of bilingual technical terms ( (t_J, t_C) ) and the seed ( (s_J, s_C) )</td>
<td>( f_{11} ): rate of identical morphemes (for Japanese terms)</td>
<td>\text{(f_{11}(t_J, s_J) = \frac{\text{const}(t_J) \cap \text{const}(s_J)}{\text{max}(\text{const}(t_J), \text{const}(s_J))}) (where (\text{const}(t)) is the set of morphemes in the Japanese term (t).)</td>
</tr>
<tr>
<td>features for the relation of bilingual technical terms ( (t_J, t_C) ) and the seed ( (s_J, s_C) )</td>
<td>( f_{12} ): rate of identical characters (for Chinese terms)</td>
<td>\text{(f_{12}(t_C, s_C) = \frac{\text{const}(t_C) \cap \text{const}(s_C)}{\text{max}(\text{const}(t_C), \text{const}(s_C))}) (where (\text{const}(t)) is the set of Characters in the Chinese term (t).)</td>
</tr>
<tr>
<td>features for the relation of bilingual technical terms ( (t_J, t_C) ) and the seed ( (s_J, s_C) )</td>
<td>( f_{13} ): subsegment relation of strings / variants relation of surface forms (for Japanese terms)</td>
<td>\text{return } 1 \text{ when the difference of } t_J \text{ and } s_J \text{ is only in their suffixes, or only whether or not having the prolonged sound “〜” or, only in their hiragana parts.}</td>
</tr>
<tr>
<td>features for the relation of bilingual technical terms ( (t_J, t_C) ) and the seed ( (s_J, s_C) )</td>
<td>( f_{14} ): identical stem (for Chinese terms)</td>
<td>\text{return } 1 \text{ when the difference of } t_C \text{ and } s_C \text{ is only whether or not having the word “的” which is not the prefix or suffix.}</td>
</tr>
<tr>
<td>features for the relation of bilingual technical terms ( (t_J, t_C) ) and the seed ( (s_J, s_C) )</td>
<td>( f_{15} ): rate of intersection in translation by the phrase table</td>
<td>\text{(f_{15}(t_X, s_X) = \frac{\text{trans}(t_X) \cap \text{trans}(s_X)}{\text{max}(\text{trans}(t_X), \text{trans}(s_X))}) ( where (\text{trans}(t)) is the set of translation of term } t \text{ from the phrase table. )}</td>
</tr>
<tr>
<td>features for the relation of bilingual technical terms ( (t_J, t_C) ) and the seed ( (s_J, s_C) )</td>
<td>( f_{16} ): translation by the phrase table</td>
<td>\text{return } 1 \text{ when } s_J \text{ can be generated by translating } t_E \text{ with the phrase table, or, } s_E \text{ can be generated by translating } t_J \text{ with the phrase table.}</td>
</tr>
</tbody>
</table>

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 subsegmentation 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 table (\( f_{16} \)).

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 \( \nu \) 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 (%)

<table>
<thead>
<tr>
<th></th>
<th>segmented by characters</th>
<th>segmented by morphemes</th>
<th>intersection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>precision</td>
<td>recall</td>
<td>f-measure</td>
</tr>
<tr>
<td>baseline (t_j and s_j are identical, or, t_C and s_C are identical.)</td>
<td>71.5</td>
<td>39.4</td>
<td>50.8</td>
</tr>
<tr>
<td>SVM</td>
<td>maximum precision</td>
<td>86.9</td>
<td>26.0</td>
</tr>
<tr>
<td>SVM</td>
<td>maximum f-measure</td>
<td>71.0</td>
<td>52.8</td>
</tr>
</tbody>
</table>

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


