Translation Estimation for Technical Terms using Corpus collected from the Web

Masatsugu Tonoike†, Mitsuhiro Kida†, Toshihiro Takagi†, Yasuhiro Sasaki†, and Satoshi Sato‡
†Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan
(tonoike,kida,takagi,sasaki,utsuro)@pine.kuee.kyoto-u.ac.jp
‡Graduate School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603 JAPAN
ssato@nuee.nagoya-u.ac.jp

Abstract

In the task of estimating bilingual term correspondences of technical terms, it is usually quite difficult to find an existing corpus for the domain of such technical terms. In this paper, we take an approach of collecting the corpus for the domain of such technical terms from the Web. This paper proposes a method of compositional translation estimation for technical terms, and through experimental evaluation, shows that domain/topic specific corpus contributes to improving the performance of compositional translation estimation.

Keywords “language translation”, “corpora and corpus-based language processing”, “electronic dictionary, thesaurus and ontology”

1 Introduction

This paper studies issues on compiling a bilingual lexicon for technical terms. So far, several techniques of estimating bilingual term correspondences from parallel/comparable corpus have been studied (Matsumoto and Utsuro, 2000). However, there are a limited number of parallel/comparable corpora that are available for the purpose of estimating bilingual term correspondences. Therefore, even if one wants to apply those existing techniques to the task of estimating bilingual term correspondences of technical terms, it is usually quite difficult to find an existing corpus for the domain of such technical terms.

Considering such a situation, we take an approach of collecting a corpus for the domain of such technical terms from the Web. In this approach, in order to compile bilingual lexicon for technical terms, the following two issues have to be addressed: collecting technical terms to be listed as headwords of bilingual lexicon, and estimating translation of those technical terms. Among those two issues, this paper focuses on the second issue of translation estimation of technical terms, and proposes a method for translation estimation for technical terms using domain/topic specific corpus collected from the Web.

More specifically, the overall framework of compiling bilingual lexicon from the Web can be illustrated as in Figure 1. Suppose that we have sample terms of a specific domain/topic, technical terms to
be listed as headwords of bilingual lexicon are collected from the Web by the related term collection method of (Sato and Sasaki, 2003). Those collected technical terms can be divided into two subsets according to whether or not the term can be translated with an existing bilingual lexicon (i.e. the subset $X_S$ of terms whose translation are known, and the subset $Y_S$ of terms whose translation are unknown). For those terms of the subset $Y_S$ which can not be translated with an existing bilingual lexicon, bilingual term correspondences are estimated using domain/topic specific corpus. Here, in order to collect domain/topic specific corpus from the Web, we use the set $X_T$ of the translation of the terms in the set $X_S$.

As a method of translation estimation for technical terms, we propose a compositional translation estimation technique. Compositional translation estimation of a term can be done through the process of compositionally generating translation candidates of the term by concatenating the translation of the constituents of the term. Here, those translation candidates are validated using domain/topic specific corpus.

In order to assess the applicability of the compositional translation estimation technique, we randomly pick up 667 Japanese and English technical term translation pairs of 10 domains from existing technical term bilingual lexicons. We then manually examine their compositionality, and find out that 88% of them are actually compositional, which is a very encouraging result. Based on this assessment, this paper proposes a method of compositional translation estimation for technical terms, and through experimental evaluation, shows that domain/topic specific corpus contributes to improving the performance of compositional translation estimation.

3 Compositional Translation Estimation for Technical Terms

3.1 Overview
An example of compositional translation estimation for the Japanese technical term “応用行動分析” is shown in Figure 2. First, the Japanese technical term “応用行動分析” is decomposed into its constituents by consulting an existing bilingual lexicon and retrieving Japanese headwords.3 In this case, the result of this decomposition can be given as in the cases “a” and “b” in Figure 2. Then, each constituent is translated into the target language. A confidence score is assigned to the translation of each constituent. Finally, translation candidates are generated by concatenating the translation of those constituents without changing word order. The confidence score of translation candidates are defined as the product of the confidence scores of each constituent. Here, when validating those translation candidates using domain/topic specific corpus, those which are not observed in the corpus are not regarded as candidates.

3.2 Compiling Bilingual Constituents Lexicon
This section describes how to compile bilingual constituents lexicons from the translation pairs of the

---

1http://www.goo.ne.jp/
2http://www.altavista.com/
3Here, as an existing bilingual lexicon, we use “Eijiro”(http://www.alc.co.jp/) and bilingual constituents lexicons compiled from the translation pairs of “Eijiro” (details to be described in the next section)
existing bilingual lexicon “Eijiro”. The underlying idea of augmenting the existing bilingual lexicon with bilingual constituents lexicons is illustrated with the example of Figure 3. Suppose that the existing bilingual lexicon does not include a translation pair “applied : 応用”, while it includes many compound translation pairs with the first English word as “applied” and the first Japanese word “応用”. In such a case, we align those translation pairs and estimate a bilingual constituent translation pair, which is to be collected into bilingual constituents lexicon.

More specifically, from the existing bilingual lexicon, we first collect translation pairs whose English terms and Japanese terms consist of two constituents into another lexicon $P_2$. We compile “bilingual constituents lexicon (prefix)” from the first constituents of the translation pairs in $P_2$ and compile “bilingual constituents lexicon (suffix)” from their second constituents. The numbers of entries in each language and those of translation pairs in those lexicons are shown in Table 1.

In the result of our assessment, only 27% of the 667 translation pairs mentioned in Section 1 can be compositionally generated using “Eijiro”, while the rate increases up to 49% using both “Eijiro” and “bilingual constituents lexicons”.\(^4\)

\(^4\)Japanese entries are supposed to be segmented into a sequence of words by morphological analyzer JUMAN (http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html)

\(^3\)In our rough estimation, the upper bound of this rate is about 80%. Improvement from 49% to 80% could be achieved by extending the bilingual constituents lexicons and by introducing constituent reordering rules with prepositions into the process of compositional translation candidate generation.

Figure 3: Example of Estimating Bilingual Constituents Translation Pair (Prefix)

Table 1: Numbers of Entries and Translation Pairs in Lexicons

<table>
<thead>
<tr>
<th>lexicon</th>
<th># of entries</th>
<th># of translation pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eijiro</td>
<td>1,292,117</td>
<td>1,228,750, 1,671,230</td>
</tr>
<tr>
<td>$P_2$</td>
<td>232,716</td>
<td>200,633, 258,211</td>
</tr>
<tr>
<td>$B_P$</td>
<td>38,353</td>
<td>38,546, 112,586</td>
</tr>
<tr>
<td>$B_S$</td>
<td>22,281</td>
<td>20,627, 71,429</td>
</tr>
</tbody>
</table>

Eijiro : existing bilingual lexicon

$P_2$ : entries of Eijiro with two constituents

$B_P$ : bilingual constituents lexicon (prefix)

$B_S$ : bilingual constituents lexicon (suffix)

3.3 Score of Translation Pairs in Lexicon

This section introduces the confidence score of translation pairs in various lexicons presented in the previous section. Here, we suppose that the translation pair $\langle s, t \rangle$ of the terms $s$ and $t$ is used when estimating translation from the language of the term $s$ to that of the term $t$. First, in this paper, we assume that translation pairs follow certain preference rules and can be ordered as below:

1. Translation pairs $\langle s, t \rangle$ in the existing bilingual lexicon “Eijiro”, where the term $s$ consists of two or more constituents.
2. Translation pairs in bilingual constituents lexicons whose frequencies in $P_2$ are high.
3. Translation pairs $\langle s, t \rangle$ in the existing bilingual lexicon “Eijiro”, where the term $s$ consists of exactly one constituent.
4. Translation pairs in bilingual constituents lexicons whose frequencies in $P_2$ are not high.

As the definition of the confidence score $q(\langle s, t \rangle)$ of the translation pair $\langle s, t \rangle$, in this paper, we use the following function.

$$q(\langle s, t \rangle) = \begin{cases} 1 & \text{if } \langle s, t \rangle \text{ is a translation pair} \\ 0 & \text{otherwise} \end{cases}$$
following:

\[ q(⟨s, t⟩) = \begin{cases} 
10^{\text{compo}(s) - 1} & (⟨s, t⟩ \text{ in the existing lexicon}) \\
\log_{10} f(⟨s, t⟩) & (⟨s, t⟩ \text{ in the bilingual constituents lexicon}) 
\end{cases} \]

where \(\text{compo}(s)\) denotes the word (in English) or morpheme (in Japanese) count of \(s\), and \(f(⟨s, t⟩)\) the frequency of \(⟨s, t⟩\) in \(P_2\).

Note that the score \(q(⟨s, t⟩)\) of a translation pair \(⟨s, t⟩\) in either bilingual constituents lexicon whose \(f(⟨s, t⟩)\) is 1 becomes 0.

For example, the score of a translation pair (“behavior analysis”, “行動分析”) in the existing bilingual lexicon Eijiro is 10 (\(= 10^{2-1}\)). That of a translation pair (“applied”, “応用”) in the existing bilingual lexicon Eijiro is 1 (\(= 10^{1-1}\)). That of a translation pair (“applied”, “応用”) whose frequency as the first constituent in \(P_2\) is 40 is 1.6 (\(= \log_{10} 40\)) in the bilingual constituents lexicon (prefix). That of a translation pair (“applied”, “使用”) whose frequency as the first constituent in \(P_2\) is 2 is 0.3 (\(= \log_{10} 2\)) in the bilingual constituents lexicon (prefix).

### 3.4 Score of Translation Candidates

Suppose that a translation candidate \(y_k\) is generated from translation pairs \(⟨s_1, t_1⟩, \ldots, ⟨s_n, t_n⟩\) as \(y_k = t_1, \ldots, t_n\). Here, in this paper, we define the confidence score of \(y_k\) as the product of the confidence scores of the constituent translation pairs \(⟨s_1, t_1⟩, \ldots, ⟨s_n, t_n⟩\).

\[
Q(y_k) = \prod_{i=1}^{n} q(⟨s_i, t_i⟩)
\]

If a translation candidate is generated from more than one sequence of translation pairs, the score of the translation candidate is defined as the sum of the score of each sequence. For example, let us consider the score of Japanese translation candidate “応用行動分析” of English term “applied behavior analysis”. Suppose that this translation candidate “応用行動分析” can be generated from two sequences of translation pairs. One is the sequence of a translation pair (“applied”, “応用”) (with the score 1.6) and a translation pair (“behavior analysis”, “行動分析”) (with the score 10), where the score of “応用行動分析” can be calculated as 16 \((= 1.6 \times 10)\). The other is the sequence of a translation pair (“applied”, “応用”) (with the score 1.6), a translation pair (“behavior”, “行動”) (with the score 1), and a translation pair (“analysis”, “分析”) (with the score 1), where the score of “応用行動分析” can be calculated as 1.6 \((= 1.6 \times 1 \times 1)\). Finally, the total score of “応用行動分析” is calculated as the sum 17.6 \((= 16 + 1.6)\).

### 4 Translation Estimation using Domain/topic Specific Corpus

It is not clear whether translation candidates which are generated by the method described in Section 3 are valid as English or Japanese terms, and it is not also clear whether they belong to the domain/topic. So using domain/topic specific corpus collected by the method described in Section 2, we examine whether the translation candidates are valid as English or Japanese terms and whether they belong to the domain/topic. In our validation method, given a ranked list of translation candidates, each translation candidate is checked whether it is observed in the corpus, and one which is not observed in the corpus is removed from the list.

As an example, we illustrate the case of estimating Japanese translation of an English technical term “Euler function”. Figure 4 gives the list of the Japanese translation candidates for the English technical term “Euler function” generated by the method described in Section 3. The translation candidates are listed in descending order of the scores. The correct Japanese translation “オイラー関数” is ranked as the second highest. On the other hand, Figure 5 shows the result of validation of the translation candidates using “discrete mathematics” domain corpus. Translation candidates that are not observed in the corpus are removed from the list and only the correct Japanese translation “オイラー関数” remains in the list.
5 Experiments and Evaluation

5.1 Translation Pairs for Evaluation

In our experimental evaluation, within the framework of compiling bilingual lexicon for technical terms, we evaluate the translation estimation part which is indicated with bold line in Figure 6. In the evaluation of this paper, we simply skip the evaluation of the process of collecting technical terms to be listed as headwords of bilingual lexicon. In order to evaluate the translation estimation part, from ten categories of existing Japanese-English technical term dictionaries listed in Table 2, translation pairs \((s, t)\) of two types are randomly picked up. Translation pairs of the first type are those which can be found in the existing bilingual lexicon “Eijiro”, which are collected into the set \(X_{ST}\). Translation pairs of the second type are those for which neither \(s\) nor \(t\) can be found in the existing bilingual lexicon “Eijiro”, which are collected into the set \(Y_{ST}\). The set of terms \(X_T\) taken from \(X_{ST}\), that are found in the side of the language \(T\), is used for collecting domain/topic specific corpus from the Web, as described in Section 1. Here, we suppose that translation estimation evaluation is to be done against the set \(Y_S\) of terms of the language \(S\) taken from \(Y_{ST}\). For each of the ten categories, Table 2 shows the sizes of \(X_{ST}\) and \(Y_{ST}\), respectively.

5.2 Translation Estimation for Technical Terms with/without Domain-specific Corpus

Without domain specific corpus, the correct rate of first ranked translation candidate is 19% on the average (both from English to Japanese and from Japanese to English). The rate of including correct candidate within top 10 is 40% from English to Japanese and 43% from Japanese to English on the average. With domain specific corpus, on the average, the correct rate of first ranked translation candidate improved by 10% from English to Japanese and by 7% from Japanese to English. However, the rate of including correct candidate within top 10 de-
Table 2: Number of Translation Pairs for Evaluation

| dictionaries         | categories               | $|X|_{ST}$ | $|Y|_{ST}$ |
|----------------------|--------------------------|---------|-----------|
| McGraw-Hill          | Electromagnetics         | 73      | 33        |
|                      | Electrical engineering   | 72      | 45        |
|                      | Optics                   | 71      | 31        |
| Iwanami              | Programming language     | 61      | 29        |
|                      | Programming              | 63      | 29        |
| Dictionary of Computer| (Computer)               | 100     | 100       |
| Dictionary of 250,000 medical terms | Anatomical Terms Disease | 100     | 100       |
|                      | Chemicals and Drugs      | 100     | 100       |
|                      | Physical Science and Statistics | 100 | 100 |
| Total                |                          | 840     | 667       |

McGraw-Hill: Dictionary of Scientific and Technical Terms
Iwanami: Encyclopedic Dictionary of Computer Science

Increased by 3% from English to Japanese, and by 4% from Japanese to English. This is because correct translation does not exist in the corpus for 20% of the 667 translation pairs for evaluation.

For about 40% of the 667 translation pairs for evaluation, correct translation does exist in the corpus and can be generated through the compositional translation estimation process. For those 40% translation pairs, Figure 7 compares the correct rate of first ranked translation pairs between with/without domain-specific corpus. The correct rates increase by 26~36% with domain-specific corpus. This result supports the claim that domain/topic specific corpus is effective in translation estimation of technical terms.

6 Conclusion

This paper proposed a method of compositional translation estimation for technical terms, and through experimental evaluation, showed that domain/topic specific corpus contributes to improving the performance of compositional translation estimation.

As related works, (Cao and Li, 2002; Fujii and Ishikawa, 2001) also proposed techniques of compositional estimation of bilingual term correspondences. One of the major differences of the techniques of (Cao and Li, 2002; Fujii and Ishikawa, 2001) and the one proposed in this paper is that we concentrated on the translation estimation of technical terms. One of the other differences is the type of corpus used to validate generated translation candidates. (Cao and Li, 2002) use the whole Web as a corpus, and (Fujii and Ishikawa, 2001) use corpus of the collection of the technical papers, each of which is published by one of the 65 Japanese associations for various technical domains. Both can be regarded as domain/topic independent corpora. On the other hand, our method use a domain/topic specific corpus corrected from the Web. If we use a domain/topic independent corpus such as the Web to validate generated translation candidates, we might select translation candidates which are used in domains/topics other than the target domains/topics. One of the merits of using a domain/topic specific corpus is that we can select only translation candidates which are used in the target domain/topic.

As a future work, we are planning to introduce a mechanism of re-ranking translation candidates based on frequencies of technical terms in domain/topic specific corpus.
References


