Effect of Domain-Specific Corpus in Compositional Translation Estimation for Technical Terms

Masatsugu Tonoike†, Mitsuhiro Kida†,Toshihiro Takagi†, Yasuhiro Sasaki†,
Takehito Utsuro† and Satoshi Sato‡

†Graduate School of Informatics,
Kyoto University
Yoshida-Honmachi, Sakyo-ku,
Kyoto 606-8501 Japan

‡Graduate School of Engineering,
Nagoya University
Furo-cho, Chikusa-ku,
Nagoya 464-8603 JAPAN

(tonoike,kida,takagi,sasaki,
utsuro)@pine.kuee.kyoto-u.ac.jp

ssato@nuee.nagoya-u.ac.jp

Abstract

This paper studies issues on compiling a bilingual lexicon for technical terms. 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 a corpus for the domain of such technical terms from the Web. As a method of translation estimation for technical terms, we propose a compositional translation estimation technique. Through experimental evaluation, we show that the domain/topic specific corpus contributes to improving the performance of the compositional translation estimation.

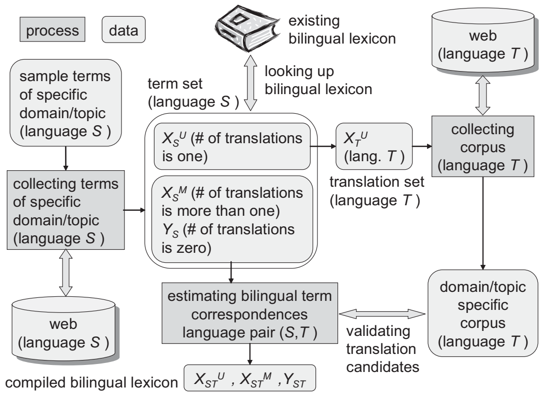


Figure 1: Compilation of a Domain/Topic Specific Bilingual Lexicon

1 Introduction

This paper studies issues on compiling a bilingual lexicon for technical terms. So far, several techniques of estimating bilingual term correspondences from a parallel/comparable corpus have been studied (Matsumoto and Utsuro, 2000). For example, in the case of estimation from comparable corpora, (Fung and Yee, 1998; Rapp, 1999) proposed standard techniques of estimating bilingual term correspondences from comparable corpora. In their techniques, contextual similarity between a source language term and its translation candidate is measured across the languages, and all the translation candidates are re-ranked according to the contextual similarities. However, there are 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 a bilingual lexicon for technical terms, the following two issues have to be addressed: collecting technical terms to be listed as the headwords of a 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 a domain/topic specific corpus collected from the Web.

More specifically, the overall framework of compiling a 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 the headwords of a 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 three subsets according to the number of translation candidates they have in an existing bilingual lexicon, i.e., the subset X_S^U of terms for which the number of translations in the existing bilingual lexicon is one, the subset X_S^M of terms for which the number of translations is more than one, and the subset Y_S of terms which are not found in the existing bilingual lexicon. (Henceforth, the union $X_S^U \cup X_S^M$ is denoted as X_S.) The translation estimation task here is to estimate translations for the terms of X_S^M and Y_S. For the terms of X_S^M, it is required to select an appropriate translation from the translation candidates found in the existing bilingual lexicon. For example, as a translation of the Japanese technical term "レジスタ", which belongs to the *logic circuit* field, the term "register" should be selected but not the term "regista" of the *football* field. On the other hand, for the terms of Y_S, it is required to generate and validate translation candidates. In this paper, for the above two tasks, we use a domain/topic specific corpus. Each term of X_S^U has the only one translation in the existing bilingual lexicon. The set of the translations of terms of X_S^U is denoted as X_T^U. Then, the domain/topic specific corpus is collected from the Web using the terms in the set X_T^U. A new bilingual lexicon is compiled from the result of translation estimation for the terms of X_S^M and Y_S, as well as the translation pairs which consist of the terms of X_S^U and their translations found in the existing bilingual lexicon.

For each term of X_S^M, from the translation candidates found in the existing bilingual lexicon, we select the one which appears most frequently in the domain/topic specific corpus. The experimental result of this translation selection process is described in Section 5.2.

As a method of translation generation/validation 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 the 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 the domain/topic specific corpus contributes to improving the performance of compositional translation estimation.

2 Collecting a Domain/Topic Specific Corpus

When collecting a domain/topic specific corpus of the language T, for each technical term x_T^U in the set X_T^U, we collect the top 100 pages with search engine queries including x_T^U. Our search engine queries are designed so that documents which describe the technical term x_T^U is to be ranked high. For example, an online glossary is one of such documents. Note that queries in English and those in Japanese do not correspond. When collecting a Japanese corpus, the search engine "goo"[1] is used. Specific queries used here are phrases with topic-marking postpositional particles such as "x_T^U とは", "x_T^U という", "x_T^U は", and an adnominal phrase "x_T^U の", and "x_T^U". When collecting a English corpus, the search engine "Yahoo!"[2] is used. Specific queries used here are "x_T^U AND what's", "x_T^U AND glossary", and "x_T^U".

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

[1] http://www.goo.ne.jp/

[2] http://www.yahoo.com/

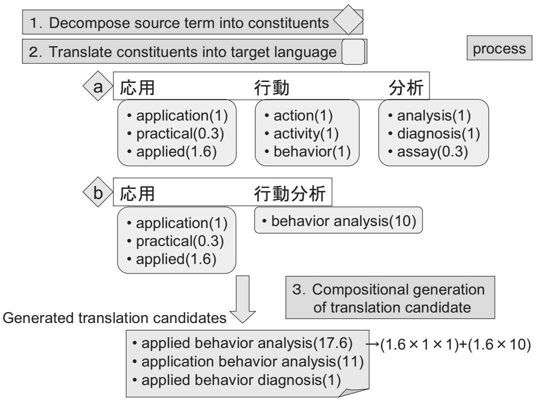

1. Decompose source term into constituents
2. Translate constituents into target language
process
a
応用
行動
分析
• application(1)
• practical(0.3)
• applied(1.6)
• action(1)
• activity(1)
• behavior(1)
• analysis(1)
• diagnosis(1)
• assay(0.3)
b
応用
行動分析
• application(1)
• practical(0.3)
• applied(1.6)
• behavior analysis(10)
3. Compositional generation of translation candidate
Generated translation candidates
• applied behavior analysis(17.6)
• application behavior analysis(11)
• applied behavior diagnosis(1)
→(1.6 × 1 × 1)+(1.6 × 10)

Figure 2: Compositional Translation Estimation for the Japanese Technical Term “応用行動分析”

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 the domain/topic specific corpus, those which are not observed in the corpus are not regarded as candidates.

3.2 Compiling Bilingual Constituents Lexicons

This section describes how to compile bilingual constituents lexicons from the translation pairs of the 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 the 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”.[4] In such a case, we align those translation pairs and estimate a bilingual constituent translation pair, which is to be collected into a bilingual constituents lexicon.

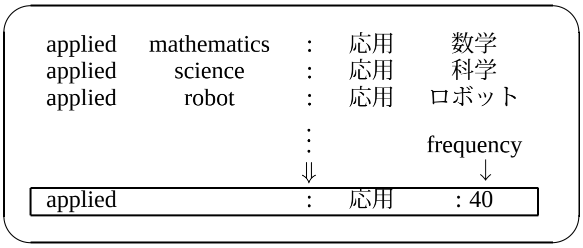

applied mathematics : 応用 数学
applied science : 応用 科学
applied robot : 応用 ロボット
:
frequency
⇓
↓
applied : 応用 : 40

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

Table 1: Numbers of Entries and Translation Pairs in the Lexicons

lexicon	# of entries		# of translation pairs
	English	Japanese	
Eijiro	1,292,117	1,228,750	1,671,230
P_2	232,716	200,633	258,211
B_P	38,353	38,546	112,586
B_S	22,281	20,627	71,429

Eijiro : existing bilingual lexicon
P_2 : entries of Eijiro with two constituents in both languages
B_P : bilingual constituents lexicon (prefix)
B_S : bilingual constituents lexicon (suffix)

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”.[5]

[3] Here, 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).

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

[5] 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.

3.3 Score of Translation Pairs in the Lexicons

This section introduces a confidence score of translation pairs in the various lexicons presented in the previous section. Here, we suppose that the translation pair $\langle s, t\rangle$ of 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 the 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 the bilingual constituents lexicons whose frequencies in P_2 are not high.

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

$$q(\langle s,t\rangle) = \begin{cases} 10^{(compo(s)-1)} & (\langle s,t\rangle \text{ in Eijiro}) \\ \log_{10} f_p(\langle s,t\rangle) & (\langle s,t\rangle \text{ in } B_P) \\ \log_{10} f_s(\langle s,t\rangle) & (\langle s,t\rangle \text{ in } B_S) \end{cases} \tag{1}$$

where $compo(s)$ denotes the word (in English) or morpheme (in Japanese) count of s, $f_p(\langle s,t\rangle)$ the frequency of $\langle s,t\rangle$ as the first constituent in P_2, and $f_s(\langle s,t\rangle)$ the frequency of $\langle s,t\rangle$ as the second constituent in P_2.[6]

3.4 Score of Translation Candidates

Suppose that a translation candidate y_t is generated from translation pairs $\langle s_1,t_1\rangle,\cdots,\langle s_n,t_n\rangle$ by concatenating $t_1,\cdots,t_n$ as $y_t = t_1\cdots t_n$. Here, in this paper, we define the confidence score of y_t as the product of the confidence scores of the constituent translation pairs $\langle s_1,t_1\rangle,\cdots,\langle s_n,t_n\rangle$.

$$Q(y_t) = \prod_{i=1}^{n} q(\langle s_i,t_i\rangle) \tag{2}$$

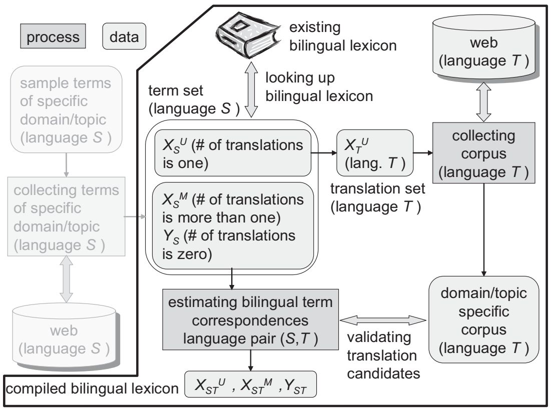


Figure 4: Experimental Evaluation of Translation Estimation for Technical Terms with/without the Domain/Topic Specific Corpus (taken from Figure 1)

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.

4 Translation Candidate Validation using a 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 a 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.

5 Experiments and Evaluation

5.1 Translation Pairs for Evaluation

In our experimental evaluation, within the framework of compiling a bilingual lexicon for technical terms, we evaluate the translation estimation part which is indicated with bold line in Fig-

[6] It is necessary to empirically examine whether this definition of the confidence score is optimal or not. However, according to our rough qualitative examination, the results of the confidence scoring seem stable when without a domain/topic specific corpus, even with minor tuning by incorporating certain parameters into the score.

Table 2: Number of Translation Pairs for Evaluation

dictionaries	categories	$\|X_S\|$	$\|Y_S\|$	S = English			S = Japanese		
				$\|X_S^U\|$	$\|X_S^M\|$	$C(S)$	$\|X_S^U\|$	$\|X_S^M\|$	$C(S)$
McGraw-Hill	Electromagnetics	58	33	36	22	82%	32	26	76%
	Electrical engineering	52	45	34	18	67%	25	27	64%
	Optics	54	31	42	12	65%	22	32	65%
Iwanami	Programming language	55	29	37	18	86%	38	17	100%
	Programming	53	29	29	24	86%	29	24	79%
Dictionary of Computer	(Computer)	100	100	91	9	46%	69	31	56%
Dictionary of 250,000 medical terms	Anatomical Terms	100	100	91	9	86%	33	67	39%
	Disease	100	100	91	9	74%	53	47	51%
	Chemicals and Drugs	100	100	94	6	58%	74	26	51%
	Physical Science and Statistics	100	100	88	12	64%	58	42	55%
Total		772	667	633	139	68%	433	339	57%

McGraw-Hill : Dictionary of Scientific and Technical Terms
Iwanami : Encyclopedic Dictionary of Computer Science
$C(S)$: for Y_S, the rate of including correct translations within the collected domain/topic specific corpus

ure 4. In the evaluation of this paper, we simply skip the evaluation of the process of collecting technical terms to be listed as the headwords of a bilingual lexicon. In order to evaluate the translation estimation part, from ten categories of existing Japanese-English technical term dictionaries listed in Table 2, terms are randomly picked up for each of the set X_S^U, X_S^M, and Y_S. (Here, as the terms of Y_S, these which consist of the only one word or morpheme are excluded.) As described in Section 1, the terms of X_T^U (the set of the translations for the terms of X_S^U) is used for collecting a domain/topic specific corpus from the Web. Translation estimation evaluation is to be done against the set X_S^M and Y_S. For each of the ten categories, Table 2 shows the sizes of X_S^U, X_S^M and Y_S, and for Y_S, the rate of including correct translation within the collected domain/topic specific corpus, respectively.

5.2 Translation Selection from Existing Bilingual Lexicon

For the terms of X_S^M, the selected translations are judged by a human. The correct rates are 69% from English to Japanese on the average and 75% from Japanese to English on the average.

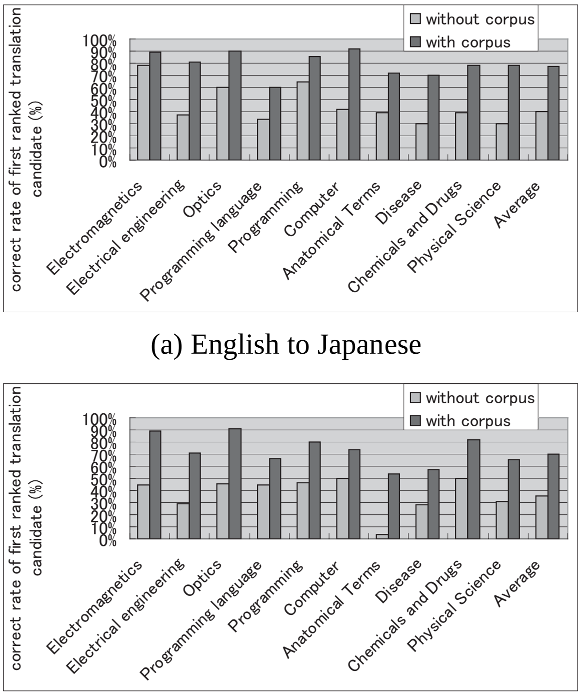
5.3 Compositional Translation Estimation for Technical Terms without the Domain/Topic Specific Corpus

Without the domain specific corpus, the correct rate of the 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. The rate of compositionally generating correct translation using both Eijiro and the bilingual constituents lexicons ($n = \infty$) is about 50% on the average (both from English to Japanese and from Japanese to English).

5.4 Compositional Translation Estimation for Technical Terms with the Domain/Topic Specific Corpus

With domain specific corpus, on the average, the correct rate of the first ranked translation candidate improved by 8% from English to Japanese and by 2% from Japanese to English. However, the rate of including correct candidate within top 10 decreased by 7% from English to Japanese, and by 14% from Japanese to English. This is because correct translation does not exist in the corpus for 32% (from English to Japanese) or 43% (from Japanese to English) of the 667 translation pairs for evaluation.

For about 35% (from English to Japanese) or 30% (from Japanese to English) 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 35% or 30% translation pairs, Figure 5 compares the correct rate of the first ranked translation pairs between with/without the domain/topic specific corpus. The correct rates increase by 34∼37% with the domain/topic specific corpus. This result supports the claim that the do-



(a) English to Japanese

(b) Japanese to English

Figure 5: Evaluation against the Translation Pairs whose Correct Translation Exist in the Corpus and can be Generated Compositionally

main/topic specific corpus is effective in translation estimation of technical terms.

6 Related Works

As a related work, (Fujii and Ishikawa, 2001) proposed a technique of compositional estimation of bilingual term correspondences for the purpose of cross-language information retrieval. In (Fujii and Ishikawa, 2001), a bilingual constituents lexicon is compiled from the translation pairs included in an existing bilingual lexicon in the same way as our proposed method. One of the major differences of the technique of (Fujii and Ishikawa, 2001) and the one proposed in this paper is that in (Fujii and Ishikawa, 2001), instead of the domain/topic specific corpus, they use a corpus of the collection of the technical papers, each of which is published by one of the 65 Japanese associations for various technical domains. Another important difference is that in (Fujii and Ishikawa, 2001), they evaluate only the performance of cross-language information retrieval but not that of translation estimation.

(Cao and Li, 2002) proposed a method of compositional translation estimation for compounds. In the proposed method of (Cao and Li, 2002), translation candidates of a term are compositionally generated by concatenating the translation of the constituents of the term and are re-ranked by measuring contextual similarity against the source language term. One of the major differences of the technique of (Cao and Li, 2002) and the one proposed in this paper is that in (Cao and Li, 2002), they do not use the domain/topic specific corpus.

7 Conclusion

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

Future works include the followings: first, in order to improve the proposed method with respect to its coverage, for example, it is desirable to extend the bilingual constituents lexicons and to introduce constituent reordering rules with prepositions into the process of compositional translation candidate generation. Second, we are planning to introduce a mechanism of re-ranking translation candidates based on the frequencies of technical terms in the domain/topic specific corpus.

References

Y. Cao and H. Li. 2002. Base noun phrase translation using Web data and the EM algorithm. In *Proc. 19th COLING*, pages 127–133.

Atsushi Fujii and Tetsuya Ishikawa. 2001. Japanese/english cross-language information retrieval: Exploration of query translation and transliteration. *Computers and the Humanities*, 35(4):389–420.

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.

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.

R. Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proc. 37th ACL*, pages 519–526.

S. Sato and Y. Sasaki. 2003. Automatic collection of related terms from the web. In *Proc. 41st ACL*, pages 121–124.