Translation Estimation for Technical Terms using Corpus collected from the Web

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
(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.

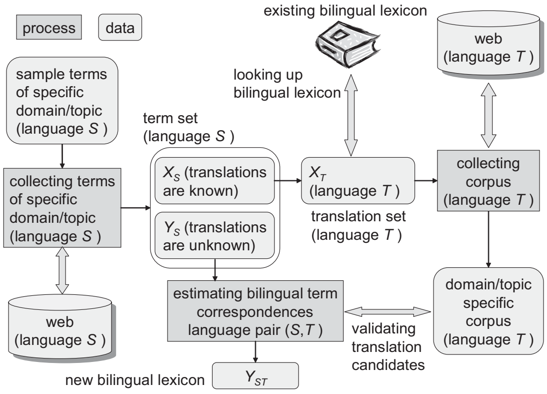

process
data
existing bilingual lexicon
web (language T)
sample terms of specific domain/topic (language S)
looking up bilingual lexicon
term set (language S)
X_S (translations are known)
Y_S (translations are unknown)
X_T (language T)
translation set (language T)
collecting corpus (language T)
collecting terms of specific domain/topic (language S)
web (language S)
estimating bilingual term correspondences language pair (S,T)
validating translation candidates
domain/topic specific corpus (language T)
new bilingual lexicon
Y_{ST}

Figure 1: Compilation of Domain/Topic Specific Bilingual Lexicon

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.

2 Collecting Domain/Topic Specific Corpus

As introduced in the previous section, X_S denotes the set of terms which can be translated with an existing bilingual lexicon, and X_T the set of the translation of the terms in the set X_S. When collecting domain/topic specific corpus of the language T, for each technical term x_T in the set X_T, we collect the top 100 pages with search engine queries including x_T. Our search engine queries are designed so that documents which describe the technical term x_t is to be ranked high. For example, an online glossary is one of such documents. Note that the queries in English and those in Japanese do not correspond. When collecting Japanese corpus, search engine “goo”[1] is used. The specific queries used here are phrases with topic-marking postpositional particles such as “x_T とは”, “x_T という”, “x_T は”, and an adnominal phrase “x_T の”, and “x_T”. When collecting English corpus, search engine “AltaVista”[2] is used. The specific queries used here are “x_T”, “x_T AND what’s”, and “x_T AND glossary” “x_T”.

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

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

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

[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)

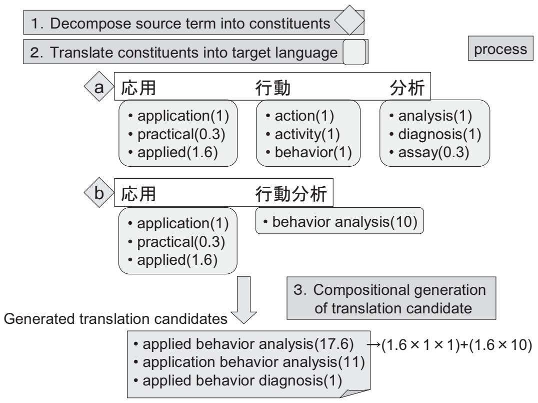

1. Decompose source term into constituents
2. Translate constituents into target language
process
a
応用
行動
分析
• application(1)
• practical(0.3)
• applied(1.6)
• action(1)
• activity(1)
• behavior(1)
• analysis(1)
• diagnosis(1)
• assay(0.3)
b
応用
行動分析
• application(1)
• practical(0.3)
• applied(1.6)
• 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”

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”.[4] 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”.[5]

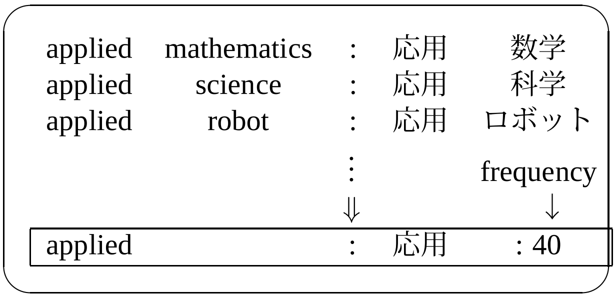

applied mathematics : 応用 数学
applied science : 応用 科学
applied robot : 応用 ロボット
:
frequency
⇓
↓
applied : 応用 : 40

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

Table 1: Numbers of Entries and Translation Pairs in 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)

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

[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)

[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.

following:

$$q(\langle s,t\rangle) = \begin{cases} 10^{(compo(s)-1)} & (\langle s,t\rangle \text{ in the existing lexicon}) \\ \log_{10} f(\langle s,t\rangle) & (\langle s,t\rangle \text{ in the bilingual constituents lexicon}) \end{cases} \tag{1}$$

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

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

For example, the score of a translation pair $\langle$“behavior analysis”, “行動分析”$\rangle$ in the existing bilingual lexicon Eijiro is 10 ($= 10^{2-1}$). That of a translation pair $\langle$“applied”, “応用の”$\rangle$ in the existing bilingual lexicon Eijiro is 1 ($= 10^{1-1}$). That of a translation pair $\langle$“applied”, “応用”$\rangle$ 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 $\langle$“applied”, “使用”$\rangle$ 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_t is generated from translation pairs $\langle s_1,t_1\rangle,\cdots,\langle s_n,t_n\rangle$ 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 $\langle s_1,t_1\rangle,\cdots,\langle s_n,t_n\rangle$.

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

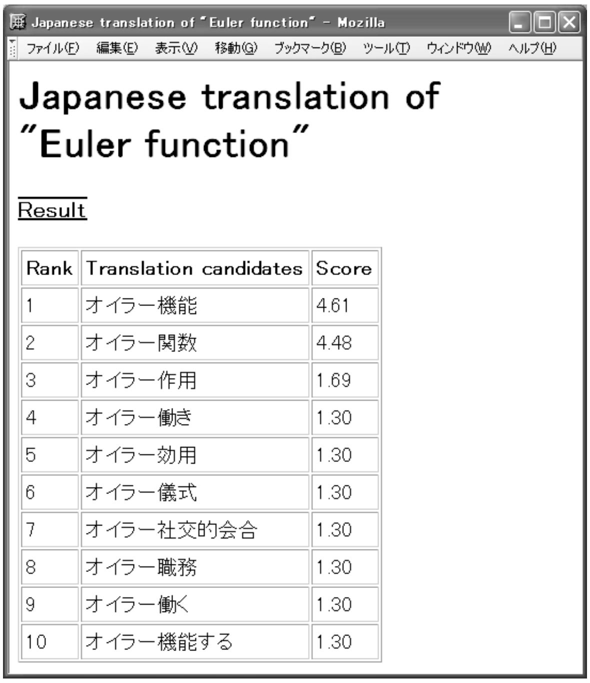
If a translation candidate is generated from more than one sequence of translation pairs, the score of the translation candidate is defined as the sum of the score of each sequence. For example, let us consider the score of Japanese translation candidate “応用行動分析” of English term “applied behavior analysis”. 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 $\langle$“applied”, “応用” $\rangle$ (with the score 1.6) and a translation pair $\langle$“behavior analysis”, “行動 分析”$\rangle$ (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 $\langle$“applied”, “応用”$\rangle$ (with the score 1.6), a translation pair $\langle$“behavior”, “行動”$\rangle$ (with the score 1) and a translation pair $\langle$“analysis”, “分析”$\rangle$ (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.

[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 confidence scoring seem stable when without domain/topic specific corpus, even with minor tuning by incorporating certain parameters into the score.


Japanese translation of "Euler function" - Mozilla
ファイル(F) 編集(E) 表示(V) 移動(G) ブックマーク(B) ツール(T) ウィンドウ(W) ヘルプ(H)

Japanese translation of "Euler function"

Result

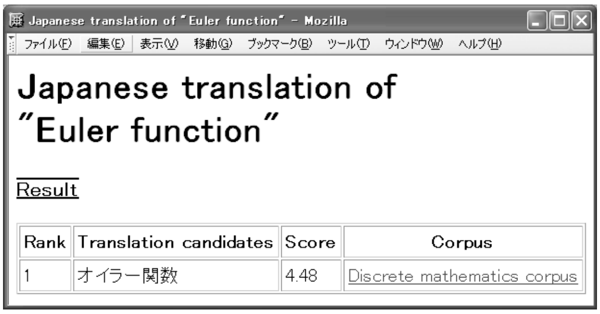
Rank	Translation candidates	Score
1	オイラー機能	4.61
2	オイラー関数	4.48
3	オイラー作用	1.69
4	オイラー働き	1.30
5	オイラー効用	1.30
6	オイラー儀式	1.30
7	オイラー社交的会合	1.30
8	オイラー職務	1.30
9	オイラー働く	1.30
10	オイラー機能する	1.30

Figure 4: Translation Estimation for the Technical Term "Euler function" without Domain/Topic Specific Corpus

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 $\langle s, t \rangle$ 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.


Japanese translation of "Euler function" - Mozilla
ファイル(F) 編集(E) 表示(V) 移動(G) ブックマーク(B) ツール(T) ウィンドウ(W) ヘルプ(H)

Japanese translation of "Euler function"

Result

Rank	Translation candidates	Score	Corpus
1	オイラー関数	4.48	Discrete mathematics corpus

Figure 5: Translation Estimation for the Technical Term "Euler function" using "Discrete Mathematics" Domain Corpus

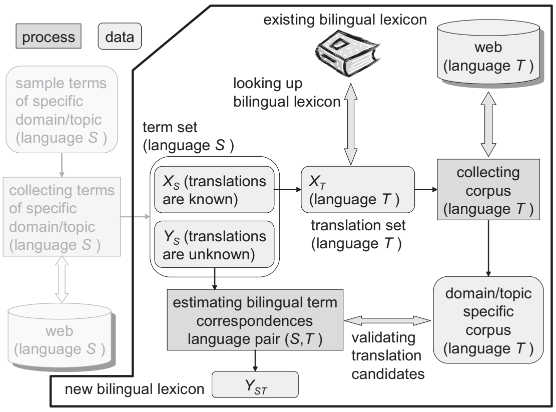

process
data
existing bilingual lexicon
web (language T)
looking up bilingual lexicon
sample terms of specific domain/topic (language S)
term set (language S)
X_S (translations are known)
X_T (language T)
collecting corpus (language T)
Y_S (translations are unknown)
translation set (language T)
collecting terms of specific domain/topic (language S)
estimating bilingual term correspondences language pair (S,T)
validating translation candidates
domain/topic specific corpus (language T)
web (language S)
new bilingual lexicon
Y_{ST}

Figure 6: Experimental Evaluation of Translation Estimation for Technical Terms with/without Domain-specific Corpus (taken from Figure 1)

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-

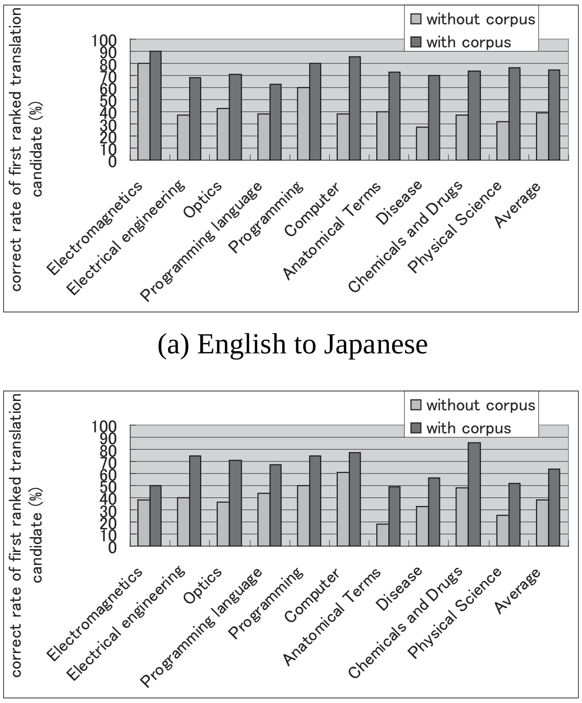
Table 2: Number of Translation Pairs for Evaluation

dictionaries	categories	$\lvert X_{ST} \rvert$	$\lvert Y_{ST} \rvert$
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	100	100
	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

creased 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.



(a) English to Japanese

(b) Japanese to English

Figure 7: Evaluation against the Translation Pairs whose Correct Translation Exist in the Corpus and can be Generated Compositionally

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

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.

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.

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