

A Comparative Study on Compositional Translation Estimation using a Domain/Topic-Specific Corpus collected from the Web

Masatsugu Tonoike[†], Mitsuhiro Kida[†], Toshihiro Takagi[†], Yasuhiro Sasaki[†],
Takehito Utsuro^{††}, Satoshi Sato^{†††}

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

^{††}Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1, Tennodai, Tsukuba, 305-8573, Japan

^{†††}Graduate School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

Abstract

This paper studies issues related to the compilation of a bilingual lexicon for technical terms. In the task of estimating bilingual term correspondences of technical terms, it is usually rather difficult to find an existing corpus for the domain of such technical terms. In this paper, we adopt 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 employ a compositional translation estimation technique. This paper focuses on quantitatively comparing variations of the components in the scoring functions of compositional translation estimation. Through experimental evaluation, we show that the domain/topic-specific corpus contributes toward improving the performance of the compositional translation estimation.

1 Introduction

This paper studies issues related to the compilation of a bilingual lexicon for technical terms. Thus 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 their 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 rather difficult to find an existing corpus for the domain of such technical terms.

On the other hand, compositional translation estimation techniques that use a monolingual corpus (Fujii and Ishikawa, 2001; Tanaka and Baldwin, 2003) are more practical. It is because collecting a monolingual corpus is less expensive than collecting a parallel/comparable corpus. Translation candidates of a term can be compositionally generated by concatenating the translation of the constituents of the term. Here, the generated 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.

But still, it is expensive to collect a domain/topic-specific corpus. Here, we adopt an approach of using the Web, since documents of various domains/topics are available on the Web. When validating translation candidates using the Web, roughly speaking, there exist the following two approaches. In the first approach, translation candidates are validated through the search engine (Cao and Li, 2002). In the second approach, a domain/topic-specific corpus is collected from the Web in advance and fixed

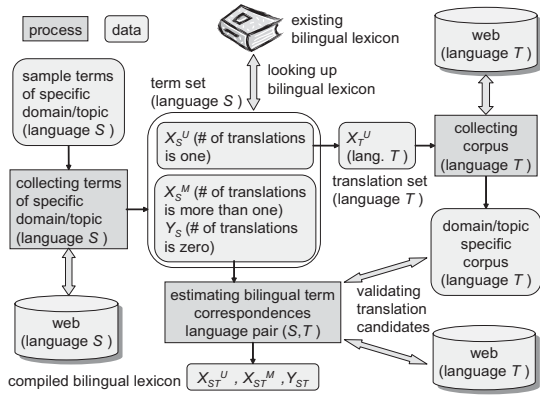


Figure 1: Compilation of a Domain/Topic-Specific Bilingual Lexicon using the Web

before translation estimation, then generated translation candidates are validated against the domain/topic-specific corpus (Tonoike et al., 2005). The first approach is preferable in terms of coverage, while the second is preferable in terms of computational efficiency. This paper mainly focuses on quantitatively comparing the two approaches in terms of coverage and precision of compositional translation estimation.

More specifically, in compositional translation estimation, we decompose the scoring function of a translation candidate into two components: bilingual lexicon score and corpus score. In this paper, we examine variants for those components and define 9 types of scoring functions in total. Regarding the above mentioned two approaches to validating translation candidates using the Web, the experimental result shows that the second approach outperforms the first when the correct translation does exist in the corpus. Furthermore, we examine the methods that combine two scoring functions based on their agreement. The experimental result shows that it is quite possible to achieve precision much higher than those of single scoring functions.

2 Overall framework

The overall framework of compiling a bilingual lexicon from the Web is illustrated as in Figure 1. Suppose that we have sample terms of a specific domain/topic, then the technical terms that are 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). These collected technical terms can be divided into three

subsets depending on the number of translation candidates present 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 that are not found in the existing bilingual lexicon (henceforth, the union $X_S^U \cup X_S^M$ will be denoted as X_S). Here, the translation estimation task here is to estimate translations for the terms of the subsets X_S^M and Y_S . A new bilingual lexicon is compiled from the result of the translation estimation for the terms of the subsets X_S^M and Y_S as well as the translation pairs that consist of the terms of the subset X_S^U and their translations found in the existing bilingual lexicon.

For the terms of the subset X_S^M , it is required that an appropriate translation is selected from among 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* domain, the term “register” should be selected but not the term “registra” of the *football* domain. On the other hand, for the terms of Y_S , it is required that the translation candidates are generated and validated. In this paper, out of the above two tasks, we focus on the latter of translation candidate generation and validation using the Web. As we introduced in the previous section, here we experimentally compare the two approaches to validating translation candidates. The first approach directly uses the search engine, while the second uses the domain/topic-specific corpus, which is collected in advance from the Web. Here, in the second approach, we use the term of X_S^U , which has only one translation in the existing bilingual lexicon. The set of translations of the terms of the subset X_S^U is denoted as X_T^U . Then, in the second approach, the domain/topic-specific corpus is collected from the Web using the terms of the set X_T^U .

3 Compositional Translation Estimation for Technical Terms

3.1 Overview

An example of compositional translation estimation for the Japanese technical term “応用行動分析” is illustrated 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 head-

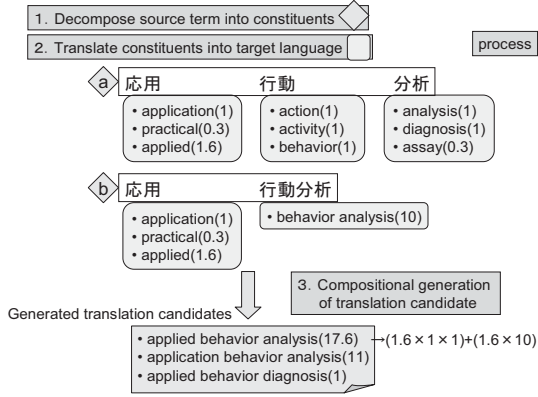


Figure 2: Compositional Translation Estimation for the Japanese Technical Term “応用行動分析”

words.¹ 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 according to word ordering rules considering prepositional phrase construction.

3.2 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 obtained from search engine queries that include the term x_T^U . Our search engine queries are designed such that documents that describe the technical term x_T^U are ranked high. For example, an online glossary is one such document. When collecting a Japanese corpus, the search engine “goo”² is used. The specific queries that are used in this search engine 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 .”

3.3 Translation Estimation

3.3.1 Compiling Bilingual Constituents Lexicons

This section describes how to compile bilingual constituents lexicons from the translation pairs of

¹Here, 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).

²<http://www.goo.ne.jp/>

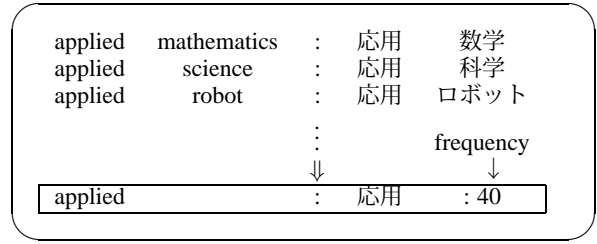


Figure 3: Example of Estimating Bilingual Constituents Translation Pair (Prefix)

the existing bilingual lexicon Eijiro. The underlying idea of augmenting the existing bilingual lexicon with bilingual constituents lexicons is illustrated in 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 “applied” and the first Japanese word “応用.”³ 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.

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 the “bilingual constituents lexicon (prefix)” from the first constituents of the translation pairs in P_2 and compile the “bilingual constituents lexicon (suffix)” from their second constituents. The number of entries in each language and those of the translation pairs in these lexicons are shown in Table 1.

The result of our assessment reveals that only 48% of the 667 translation pairs mentioned in Section 1 can be compositionally generated by using Eijiro, while the rate increases up to 69% using both Eijiro and “bilingual constituents lexicons.”⁴

3.3.2 Score of Translation Candidates

This section gives the definition of the scores of a translation candidate in compositional translation estimation.

First, let y_s be a technical term whose translation is to be estimated. We assume that y_s is de-

³Japanese 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>).

⁴In our rough estimation, the upper bound of this rate is approximately 80%. An improvement from 69% to 80% could be achieved by extending the bilingual constituents lexicons.

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	217,861	186,823	235,979
B_P	37,090	34,048	95,568
B_S	20,315	19,345	62,419
B	48,000	42,796	147,848

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)
 B : bilingual constituents lexicon (merged)

composed into their constituents as below:

$$y_s = s_1, s_2, \dots, s_n \quad (1)$$

where each s_i is a single word or a sequence of words.⁵ For y_s , we denote a generated translation candidate as y_t .

$$y_t = t_1, t_2, \dots, t_n \quad (2)$$

where each t_i is a translation of s_i . Then the translation pair $\langle y_s, y_t \rangle$ is represented as follows.

$$\langle y_s, y_t \rangle = \langle s_1, t_1 \rangle, \langle s_2, t_2 \rangle, \dots, \langle s_n, t_n \rangle \quad (3)$$

The score of a generated translation candidate is defined as the product of a bilingual lexicon score and a corpus score as follows.

$$Q(y_s, y_t) = Q_{dict}(y_s, y_t) \cdot Q_{corpus}(y_t) \quad (4)$$

Bilingual lexicon score measures appropriateness of correspondence of y_s and y_t . Corpus score measures appropriateness of the translation candidate y_t based on the target language corpus. 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.

Bilingual Lexicon Score

In this paper, we compare two types of bilingual lexicon scores. Both scores are defined as the product of scores of translation pairs included in the lexicons presented in the previous section as follows.

⁵Eijiro has both single word entries and compound word entries.

• Frequency-Length

$$Q_{dict}(y_s, y_t) = \prod_{i=1}^n q(\langle s_i, t_i \rangle) \quad (5)$$

The first type of bilingual lexicon scores is referred to as ‘‘Frequency-Length.’’ This score is based on the length of translation pairs and the frequencies of translation pairs in the bilingual constituent lexicons (prefix, suffix) B_P, B_S in Table 1. In this paper, we first assume that the translation pairs follow certain preference rules and that they 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 $\langle s, t \rangle$, we use the following:

$$q(\langle s, t \rangle) = \begin{cases} 10^{(compo(s)-1)} & (\langle s, t \rangle \text{ in Eijiro}) \\ \log_{10} f_p(\langle s, t \rangle) & (\langle s, t \rangle \text{ in } B_P) \\ \log_{10} f_s(\langle s, t \rangle) & (\langle s, t \rangle \text{ in } B_S) \end{cases} \quad (6)$$

, where $compo(s)$ denotes the word count of s , $f_p(\langle s, t \rangle)$ represents the frequency of $\langle s, t \rangle$ as the first constituent in P_2 , and $f_s(\langle s, t \rangle)$ represents the frequency of $\langle s, t \rangle$ as the second constituent in P_2 .

• Probability

$$Q_{dict}(y_s, y_t) = \prod_{i=1}^n P(s_i|t_i) \quad (7)$$

The second type of bilingual lexicon scores is referred to as ‘‘Probability.’’ This score is calculated as the product of the conditional probabilities $P(s_i|t_i)$. $P(s|t)$ is calculated using bilingual lexicons in Table 1.

$$P(s|t) = \frac{f_{prob}(\langle s, t \rangle)}{\sum_{s_j} f_{prob}(\langle s_j, t \rangle)} \quad (8)$$

Table 2: 9 Scoring Functions of Translation Candidates and their Components

score ID	bilingual lexicon score		corpus score			corpus	
	freq-length	probability	probability	frequency	occurrence	off-line	on-line (search engine)
A		prune/final	prune/final			o	
B		prune/final		prune/final		o	
C	prune/final		prune/final			o	
D	prune/final				prune	o	
E	prune/final						
F	prune/final			final	prune	o	
G	prune/final			prune/final		o	
H	prune/final			final		o	
I	prune/final			final			o

$f_{prob}(\langle s, t \rangle)$ denotes the frequency of the translation pair $\langle s, t \rangle$ in the bilingual lexicons as follows:

$$f_{prob}(\langle s, t \rangle) = \begin{cases} 10 & (\langle s, t \rangle \text{ in Eijiro}) \\ f_B(\langle s, t \rangle) & (\langle s, t \rangle \text{ in } B) \end{cases} \quad (9)$$

Note that the frequency of a translation pair in Eijiro is regarded as 10^6 and $f_B(\langle s, t \rangle)$ denotes the frequency of the translation pair $\langle s, t \rangle$ in the bilingual constituent lexicon B .

Corpus Score

We evaluate three types of corpus scores as follows.

- Probability: the occurrence probability of y_t estimated by the following bi-gram model

$$Q_{corpus}(y_t) = P(t_1) \cdot \prod_{i=1}^n P(t_{i+1}|t_i) \quad (10)$$

- Frequency: the frequency of a translation candidate in a target language corpus

$$Q_{corpus}(y_t) = freq(y_t) \quad (11)$$

- Occurrence: whether a translation candidate occurs in a target language corpus or not

$$Q_{corpus}(y_t) = \begin{cases} 1 & y_t \text{ occurs in a corpus} \\ 0 & y_t \text{ does not occur} \\ & \text{in a corpus} \end{cases} \quad (12)$$

⁶It is necessary to empirically examine whether or not the definition of the frequency of a translation pair in Eijiro is appropriate.

Variation of the total scoring functions

As shown in Table 2, in this paper, we examine the 9 combinations of the bilingual lexicon scores and the corpus scores. In the table, ‘prune’ indicates that the score is used for ranking and pruning sub-sequences of generated translation candidates in the course of generating translation candidates using a dynamic programming algorithm. ‘Final’ indicates that the score is used for ranking the final outputs of generating translation candidates. In the column ‘corpus’, ‘off-line’ indicates that a domain/topic-specific corpus is collected from the Web in advance and then generated translation candidates are validated against this corpus. ‘On-line’ indicates that translation candidates are directly validated through the search engine.

Roughly speaking, the scoring function ‘A’ corresponds to a variant of the model proposed by (Fujii and Ishikawa, 2001). The scoring function ‘D’ is a variant of the model proposed by (Tonoike et al., 2005) and ‘E’ corresponds to the bilingual lexicon score of the scoring function ‘D’. The scoring function ‘I’ is intended to evaluate the approach proposed in (Cao and Li, 2002).

3.3.3 Combining Two Scoring Functions based on their Agreement

In this section, we examine the method that combines two scoring functions based on their agreement. The two scoring functions are selected out of the 9 functions introduced in the previous section. In this method, first, confidence of translation candidates of a technical term are measured by the two scoring functions. Then, if the first ranked translation candidates of both scoring functions agree, this method outputs the agreed translation candidate. The purpose of introducing this method is to prefer precision to recall.

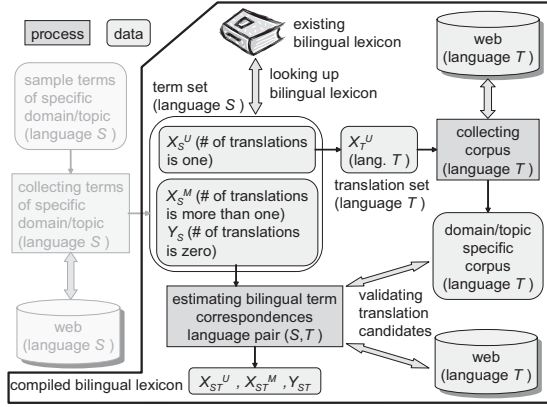


Figure 4: Experimental Evaluation of Translation Estimation for Technical Terms with/without the Domain/Topic-Specific Corpus (taken from Figure 1)

4 Experiments and Evaluation

4.1 Translation Pairs for Evaluation

In our experimental evaluation, within the framework of compiling a bilingual lexicon for technical terms, we evaluate the translation estimation portion that is indicated by the bold line in Figure 4. In this paper, we simply omit 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 portion, terms are randomly selected from the 10 categories of existing Japanese-English technical term dictionaries listed in Table 3, for each of the subsets X_S^U and Y_S (here, the terms of Y_S that consist of only one word or morpheme are excluded). As described in Section 1, the terms of the set X_T^U (the set of translations for the terms of the subset X_S^U) is used for collecting a domain/topic-specific corpus from the Web. As shown in Table 3, size of the collected corpora is 48MB on the average. Translation estimation evaluation is to be conducted for the subset Y_S . For each of the 10 categories, Table 3 shows the sizes of the subsets X_S^U and Y_S , and the rate of including correct translation within the collected domain/topic-specific corpus for Y_S . In the following, we show the evaluation results with the source language S as English and the target language T as Japanese.

4.2 Evaluation of single scoring functions

This section gives the results of evaluating single scoring functions A ~ I listed in Table 2.

Table 4 shows three types of experimental re-

sults. The column ‘the whole set Y_S ’ shows the results against the whole set Y_S . The column ‘generatable’ shows the results against the translation pairs in Y_S that can be generated through the compositional translation estimation process. 69% of the terms in ‘the whole set Y_S ’ belongs to the set ‘generatable’. The column ‘gene.-exist’ shows the result against the source terms whose correct translations do exist in the corpus and that can be generated through the compositional translation estimation process. 50% of the terms in ‘the whole set Y_S ’ belongs to the set ‘gene.-exist’. The column ‘top 1’ shows the correct rate of the first ranked translation candidate. The column ‘top 10’ shows the rate of including the correct candidate within top 10.

First, in order to evaluate the effectiveness of the approach of validating translation candidates by using a target language corpus, we compare the scoring functions ‘D’ and ‘E’. The difference between them is whether or not they use a corpus score. The results for the whole set Y_S show that using a corpus score, the precision improves from 33.9% to 43.0%. This result supports the effectiveness of the approach of validating translation candidates using a target language corpus.

As can be seen from these results for the whole set Y_S , the correct rate of the scoring function ‘I’ that directly uses the web search engine in the calculation of its corpus score is higher than those of other scoring functions that use the collected domain/topic-specific corpus. This is because, for the whole set Y_S , the rate of including correct translation within the collected domain/topic-specific corpus is 72% on the average, which is not very high. On the other hand, the results of the column ‘gene.-exist’ show that if the correct translation does exist in the corpus, most of the scoring functions other than ‘I’ can achieve precisions higher than that of the scoring function ‘I’. This result supports the effectiveness of the approach of collecting a domain/topic-specific corpus from the Web in advance and then validating generated translation candidates against this corpus.

4.3 Evaluation of combining two scoring functions based on their agreement

The result of evaluating the method that combines two scoring functions based on their agreement is shown in Table 5. This result indicates that combinations of scoring functions with ‘off-line’/‘on-

Table 3: Number of Translation Pairs for Evaluation (S =English)

dictionaries	categories	$ Y_S $	$ X_S^U $	corpus size	$C(S)$
McGraw-Hill	Electromagnetics	33	36	28MB	85%
	Electrical engineering	45	34	21MB	71%
	Optics	31	42	37MB	65%
Iwanami	Programming language	29	37	34MB	93%
	Programming	29	29	33MB	97%
Dictionary of Computer	(Computer)	100	91	67MB	51%
Dictionary of 250,000 medical terms	Anatomical Terms	100	91	73MB	86%
	Disease	100	91	83MB	77%
	Chemicals and Drugs	100	94	54MB	60%
	Physical Science and Statistics	100	88	56MB	68%
Total		667	633	482MB	72%

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

Table 4: Result of Evaluating single Scoring Functions

ID	the whole set Y_S (667 terms~100%)		generatable (458 terms~69%)		gene.-exist (333 terms~50%)	
	top 1	top 10	top 1	top 10	top 1	top 10
A	43.8%	52.9%	63.8%	77.1%	82.0%	98.5%
B	42.9%	50.7%	62.4%	73.8%	83.8%	99.4%
C	43.0%	58.0%	62.7%	84.5%	75.1%	94.6%
D	43.0%	47.4%	62.7%	69.0%	85.9%	94.6%
E	33.9%	57.3%	49.3%	83.4%	51.1%	84.1%
F	40.2%	47.4%	58.5%	69.0%	80.2%	94.6%
G	39.1%	46.8%	57.0%	68.1%	78.1%	93.4%
H	43.8%	57.3%	63.8%	83.4%	73.6%	84.1%
I	49.8%	57.3%	72.5%	83.4%	74.8%	84.1%

Table 5: Result of combining two scoring functions based on their agreement

corpus	combination	precision	recall	$F_{\beta=1}$
off-line/ on-line	A & I	88.0%	27.6%	0.420
	D & I	86.0%	29.5%	0.440
	F & I	85.1%	29.1%	0.434
	H & I	58.7%	37.5%	0.457
off-line/ off-line	A & H	86.0%	30.4%	0.450
	F & H	80.6%	33.7%	0.476
	D & H	80.4%	32.7%	0.465
	A & D	79.0%	32.1%	0.456
	A & F	74.6%	33.0%	0.457
	D & F	68.2%	35.7%	0.469

line' corpus tend to achieve higher precisions than those with 'off-line'/'off-line' corpus. This result also shows that it is quite possible to achieve high precisions even by combining scoring functions with 'off-line'/'off-line' corpus (the pair 'A' and 'H'). Here, the two scoring functions 'A' and 'H' are the one with frequency-based scoring functions and that with probability-based scoring functions, and hence, have quite different nature in the design of their scoring functions.

5 Related Works

As a related work, (Fujii and Ishikawa, 2001) proposed a technique for compositional estimation of bilingual term correspondences for the purpose of cross-language information retrieval. One of the major differences between the technique of (Fujii and Ishikawa, 2001) and the one proposed in this paper is that in (Fujii and Ishikawa, 2001), instead of a domain/topic-specific corpus, they use a corpus containing the collection of technical papers, each of which is published by one of the 65 Japanese associations for various technical domains. Another significant difference is that in (Fujii and Ishikawa, 2001), they evaluate only the performance of the cross-language information retrieval and not that of translation estimation.

(Cao and Li, 2002) also proposed a method of compositional translation estimation for compounds. In the method of (Cao and Li, 2002), the translation candidates of a term are compositionally generated by concatenating the translation of the constituents of the term and are validated directly through the search engine. In this paper, we evaluate the approach proposed in (Cao and Li, 2002) by introducing a total scoring function

that is based on validating translation candidates directly through the search engine.

6 Conclusion

This paper studied issues related to the compilation a bilingual lexicon for technical terms. In the task of estimating bilingual term correspondences of technical terms, it is usually rather difficult to find an existing corpus for the domain of such technical terms. In this paper, we adopt 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 employed a compositional translation estimation technique. This paper focused on quantitatively comparing variations of the components in the scoring functions of compositional translation estimation. Through experimental evaluation, we showed that the domain/topic specific corpus contributes to improving the performance of the compositional translation estimation.

Future work includes complementally integrating the proposed framework of compositional translation estimation using the Web with other translation estimation techniques. One of them is that based on collecting partially bilingual texts through the search engine (Nagata and others, 2001; Huang et al., 2005). Another technique which seems to be useful is that of transliteration of names (Knight and Graehl, 1998; Oh and Choi, 2005).

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.
- A. Fujii and T. 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.
- F. Huang, Y. Zhang, and S. Vogel. 2005. Mining key phrase translations from web corpora. In *Proc. HLT/EMNLP*, pages 483–490.
- K. Knight and J. Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599–612.
- 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.
- M. Nagata et al. 2001. Using the Web as a bilingual dictionary. In *Proc. ACL-2001 Workshop on Data-driven Methods in Machine Translation*, pages 95–102.
- J. Oh and K. Choi. 2005. Automatic extraction of english-korean translations for constituents of technical terms. In *Proc. 2nd IJCNLP*, pages 450–461.
- 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.
- T. Tanaka and T. Baldwin. 2003. Translation selection for japanese-english noun-noun compounds. In *Proc. Machine Translation Summit IX*, pages 378–85.
- M. Tonoike, M. Kida, T. Takagi, Y. Sasaki, T. Utsuro, and S. Sato. 2005. Effect of domain-specific corpus in compositional translation estimation for technical terms. In *Proc. 2nd IJCNLP, Companion Volume*, pages 116–121.