Integrating Cross-Lingually Relevant News Articles and Monolingual Web Documents in Bilingual Lexicon Acquisition

Takehito Utsuro† and **Kohei Hino**‡ and **Mitsuhiro Kida**†
Seiichi Nakagawa‡ and **Satoshi Sato**†
†Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, 606-8501, Japan
‡Department of Information and Computer Sciences, Toyohashi University of Technology
Tenpaku-cho, Toyohashi, 441–8580, Japan

Abstract

In the framework of bilingual lexicon acquisition from cross-lingually relevant news articles on the Web, it is relatively harder to reliably estimate bilingual term correspondences for low frequency terms. Considering such a situation, this paper proposes to complementarily use much larger monolingual Web documents collected by search engines, as a resource for reliably re-estimating bilingual term correspondences. We experimentally show that, using a sufficient number of monolingual Web documents, it is quite possible to have reliable estimate of bilingual term correspondences for those low frequency terms.

1 Introduction

Translation knowledge acquisition from parallel/comparative corpora is one of the most important research topics of corpus-based MT. This is because it is necessary for an MT system to (semi-)automatically increase its translation knowledge in order for it to be used in the real world situation. One limitation of the corpus-based translation knowledge acquisition approach is that the techniques of translation knowledge acquisition heavily rely on availability of parallel/comparative corpora. However, the sizes as well as the domain of existing parallel/comparative corpora are limited, while it is very expensive to manually collect parallel/comparative corpora. Therefore, it is quite important to overcome this resource scarcity bottleneck in corpus-based translation knowledge acquisition research.

In order to solve this problem, we proposed an approach of taking bilingual news articles on Web news sites as a source for translation knowledge acquisition (Utsuro et al., 2003). In the case of Web news sites in Japan, Japanese as well as English news articles are updated everyday. Although most of those bilingual news articles are not parallel even if they are from the same site, certain portion of those bilingual news articles share their contents or at least report quite relevant topics. This characteristic is quite important for the purpose of translation knowledge acquisition. Utsuro et al. (2003) showed that it is possible to acquire translation knowledge of domain specific named entities, event expressions, and collocational expressions from the collection of bilingual news articles on Web news sites.

Based on the results of our previous study, this paper further examines the correlation of term frequency and the reliability of bilingual term correspondences estimated from bilingual news articles. We show that, for high frequency terms, it is relatively easier to reliably estimate bilingual term correspondences. However, for low frequency terms, it is relatively harder to reliably estimate bilingual term correspondences. Low frequency problem of this type often happens when a sufficient number of bilingual news articles are not available at hand.

Considering such a situation, this paper then proposes to complementarily use much larger monolingual Web documents collected by search engines, as a resource for reliably re-estimating bilingual term correspondences. Those collected monolingual Web documents are regarded as comparable corpora. Here, a standard technique of estimating bilingual term correspondences from comparable corpora is employed. In the evaluation, we show that, using a sufficient number of monolingual Web documents, it is relatively easier to have reliable estimate of bilingual term correspondences. As one of the most remarkable experimental evaluation results, we further show that, for the terms which appear infrequently in news articles, the accuracy of re-estimating bilingual term correspondences does actually improve.

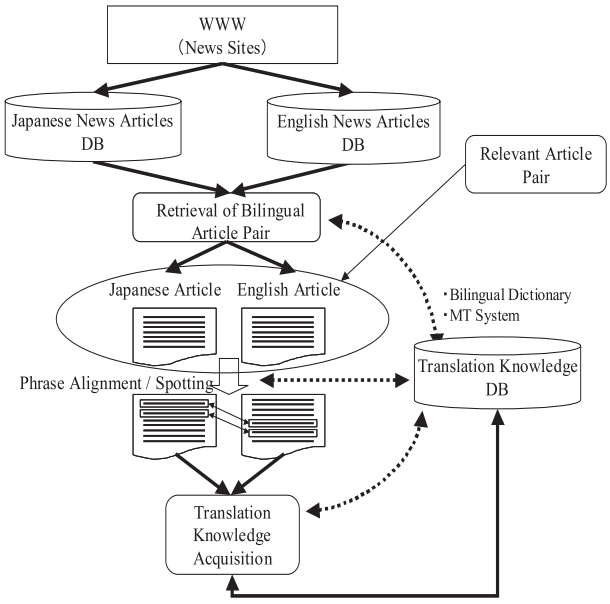

Figure 1: Translation Knowledge Acquisition from Web News Sites: Overview

2 Estimating Bilingual Term Correspondences from Cross-Lingually Relevant News Articles

2.1 Overview

Figure 1 illustrates the overview of our framework of translation knowledge acquisition from Web news sites. First, pairs of Japanese and English news articles which report identical contents or at least closely related contents are retrieved. In this cross-lingual retrieval process, translation knowledge such as a bilingual dictionary and an MT software is used for measuring similarity of Japanese and English articles across languages. Then, by applying previously studied techniques of translation knowledge acquisition from parallel/comparative corpora, translation knowledge such as bilingual term correspondences are acquired.

2.2 Cross-Language Retrieval of Relevant News Articles

This section gives the overview of our framework of cross-language retrieval of relevant news articles from Web news sites (Utsuro et al., 2003). First, from Web news sites, both Japanese and English news articles within certain range of dates are retrieved. Let d_J and d_E denote one of the retrieved Japanese and English articles, respectively. Then, each English article d_E is translated into a Japanese document d_J^{MT} by some commercial MT software[1]. Each Japanese article d_J as well as the Japanese translation d_J^{MT} of each English article are next segmented into word sequences, and word frequency vectors $v(d_J)$ and $v(d_J^{MT})$ are generated. Then, cosine similarities between $v(d_J)$ and $v(d_J^{MT})$ are calculated[2] and pairs of articles d_J and d_E which satisfy certain criterion are considered as candidates for *cross-lingually relevant* article pairs.

As we describe in section 4.1, on Web news sites in Japan, the number of articles updated per day is far greater (about 4 times) in Japanese than in English. Thus, it is much easier to find cross-lingually relevant articles for each *English* query article than for each *Japanese* query article. Considering this fact, we estimate bilingual term correspondences from the results of cross-lingually retrieving relevant *Japanese* articles with *English query* articles. For each English query article d_E^i and its Japanese translation d_J^{MTi}, the set D_J^i of Japanese articles that are within certain range of dates and are with cosine similarities higher than or equal to a certain lower bound L_d is constructed:

$$D_J^i = \left\{ d_J \mid \cos(v(d_J^{MTi}), v(d_J)) \geq L_d \right\} \quad (1)$$

2.3 Estimating Bilingual Term Correspondences with Pseudo-Parallel Corpus

This section describes the technique we apply to the task of estimating bilingual term correspondences from cross-lingually relevant news texts. Here, we regard cross-lingually relevant news texts as a pseudo-parallel corpus, to which standard techniques of estimating bilingual term correspondences from parallel corpora can be applied[3].

[1] In this query translation process, we compared an MT software with a bilingual lexicon. CLIR with query translation by an MT software performed much better than that by a bilingual lexicon. In the case of news articles on Web news sites, it is relatively easier to find articles in the other language which report closely related contents, with just a few days difference of report dates. In such a case, exact query translation by an MT software is suitable, because exact translation is expected to easily match the closely related articles in the other language. As we mention in section 3.3, this is opposite to the situation of monolingual Web documents, where it is much less expected to find closely related documents in the other language.

[2] It is also quite possible to employ weights other than word frequencies such as $tf \cdot idf$ and similarity measures other than cosine measure such as dice or Jaccard coefficients.

[3] We also applied another technique based on contextual vector similarities (Utsuro et al., 2003), which

First, we concatenate constituent Japanese articles of D_J^i into one article $D_J'^i$, and regard the article pair d_E^i and $D_J'^i$ as a pseudo-parallel sentence pair. Next, we collect such pseudo-parallel sentence pairs and construct a pseudo-parallel corpus PPC_{EJ} of English and Japanese articles:

$$PPC_{EJ} = \left\{ \langle d_E^i, D_J'^i \rangle \mid D_J^i \neq \emptyset \right\}$$

Then, we apply standard techniques of estimating bilingual term correspondences from parallel corpora (Matsumoto and Utsuro, 2000) to this pseudo-parallel corpus PPC_{EJ}. First, from a pseudo-parallel sentence pair d_E^i and $D_J'^i$, we extract monolingual (possibly compound[4]) term pair t_E and t_J:

$$\langle t_E, t_J \rangle \text{ s.t. } \exists d_E^i \exists d_J,\ t_E \text{ in } d_E^i,\ t_J \text{ in } d_J, \quad \cos(v(d_J^{MTi}), v(d_J)) \geq L_d \tag{2}$$

Then, based on the contingency table of co-occurrence document frequencies of t_E and t_J below, we estimate bilingual term correspondences according to the statistical measures such as the mutual information, the ϕ^2 statistic, the dice coefficient, and the log-likelihood ratio.

	t_J	$\neg t_J$
t_E	$df(t_E, t_J) = a$	$df(t_E, \neg t_J) = b$
$\neg t_E$	$df(\neg t_E, t_J) = c$	$df(\neg t_E, \neg t_J) = d$

We compare the performance of those four measures, where the ϕ^2 statistic and the log-likelihood ratio perform best, the dice coefficient the second best, and the mutual information the worst. In section 4.3, we show results with the ϕ^2 statistic as the bilingual term correspondence $corr_{EJ}(t_E, t_J)$:

$$\phi^2(t_E, t_J) = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

3 Re-estimating Bilingual Term Correspondences using Monolingual Web Documents

3.1 Overview

This section illustrates the overview of the process of re-estimating bilingual term correspondences using monolingual Web documents collected by search engines. Figure 2 gives its rough idea.

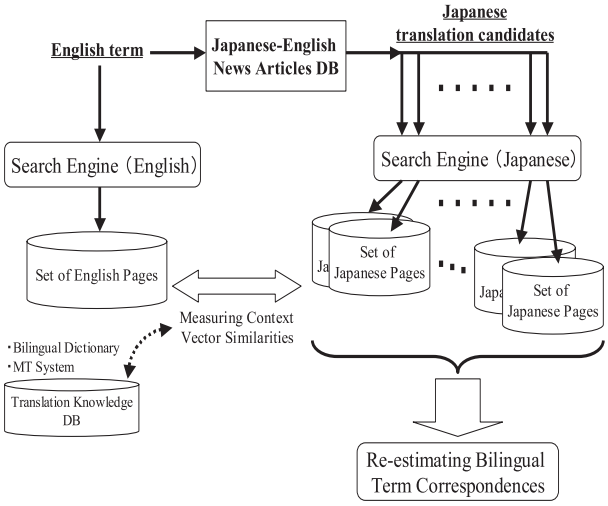

Figure 2: Re-estimating Bilingual Term Correspondences using Monolingual Web Documents: Overview

Suppose that we have an English term, and that the problem to solve here is to find its Japanese translation. As we described in the previous section and in Figure 1, with a cross-lingually relevant Japanese and English news articles database, we can have a certain number of Japanese translation candidates for the target English term. Here, for high frequency terms, it is relatively easier to have reliable ranking of those Japanese translation candidates. However, for low frequency terms, having reliable ranking of those Japanese translation candidates is difficult. Especially, low frequency problem of this type often happens when we do not have large enough language resources (in this case, cross-lingually relevant news articles).

Considering such a situation, re-estimation of bilingual term correspondences proceeds as follows, using much larger monolingual Web documents sets that are easily accessible through search engines. First, English pages which contain the target English term are collected through an English search engine. In the similar way, for each Japanese term in the Japanese translation candidates, Japanese pages which contain the Japanese term are collected through a Japanese search engine. Then, texts contained in those English and Japanese pages are extracted and are regarded as comparable corpora. Here, a standard technique of estimating bilingual term correspondences from comparable corpora (e.g., Fung and Yee (1998) and Rapp (1999)) is employed. Contextual similarity between the target English term and

has been well studied in the context of bilingual lexicon acquisition from comparable corpora. In this method, we regard cross-lingually relevant texts as a comparable corpus, where bilingual term correspondences are estimated in terms of contextual similarities across languages. This technique is less effective than the one we describe here (Utsuro et al., 2003).

[4] In the evaluation of this paper, we restrict English and Japanese terms t_E and t_J to be up to five words long.

the Japanese translation candidate is measured across languages, and all the Japanese translation candidates are re-ranked according to the contextual similarities.

3.2 Filtering by Hits of Search Engines

Before re-estimating bilingual term correspondences using monolingual Web documents, we assume there exists certain correlation between hits of the English term t_E and the Japanese term t_J returned by search engines. Depending on the hits $h(t_E)$ of t_E, we restrict the hits $h(t_J)$ of t_J to be within the range of a lower bound h_L and an upper bound h_U:

$$h_L < h(t_J) \leq h_U$$

As search engines, we used AltaVista (`http://www. altavista.com/` for English, and goo (`http://www.goo.ne.jp/`) for Japanese. With a development data set consisting of translation pairs of an English term and a Japanese term, we manually constructed the following rules for determining the lower bound h_L and the upper bound h_U:

1. $0 < h(t_E) \leq 100$
   $h_L = 0$, $h_U = 10,000 \times h(t_E)$
2. $100 < h(t_E) \leq 20,000$
   $h_L = 0.05 \times h(t_E)$, $h_U = 1,000,000$
3. $20,000 < h(t_E)$
   $h_L = 1,000$, $h_U = 50 \times h(t_E)$

In the experimental evaluation of Section 4.4, the initial set of Japanese translation candidates consists of 50 terms for each English term, which are then reduced to on the average 24.8 terms with this filtering.

3.3 Re-estimating Bilingual Term Correspondences based on Contextual Similarity

This section describes how to re-estimate bilingual term correspondences using monolingual Web documents collected by search engines.

For an English term t_E and a Japanese term t_J, let $D(t_E)$ and $D(t_J)$ be the sets of documents returned by search engines with queries t_E and t_J, respectively. Then, for the English term t_E, translated contextual vector $cv_{trJ}(t_E)$ is constructed as below: each English *sentence* s_E which contains t_E is translated into Japanese sentence s_J^{tr}, then the term frequency vectors[5] $v(s_J^{tr})$ of Japanese translation s_J^{tr} are summed up into the translated contextual vector $cv_{trJ}(t_E)$:

$$cv_{trJ}(t_E) = \sum_{\forall s_E \text{ in } D(t_E) \text{ s.t. } t_E \text{ in } s_E} v(s_J^{tr})$$

The contextual vector $cv(t_J)$ for the Japanese term t_J is also constructed by summing up the term frequency vectors $v(s_J)$ of each Japanese *sentence* s_J which contains t_J:

$$cv(t_J) = \sum_{\forall s_J \text{ in } D(t_J) \text{ s.t. } t_J \text{ in } s_J} v(s_J)$$

In the translation of English sentences into Japanese, we evaluated an MT software and a bilingual lexicon in terms of the performance of re-estimation of bilingual term correspondences. Unlike the situation of cross-lingually relevant news articles mentioned in Section 2.2, translation by a bilingual lexicon is more effective for monolingual Web documents. In the case of monolingual Web documents, it is much less expected to find closely related documents in the other language. In such cases, multiple translation rather than exact translation by an MT software is suitable. In Section 4.4, we show evaluation results with translation by a bilingual lexicon[6].

Finally, bilingual term correspondence $corr_{EJ}(t_E, t_J)$ is estimated in terms of cosine measure $\cos(cv_{trJ}(t_E), cv(t_J))$ between contextual vectors $cv_{trJ}(t_E)$ and $cv(t_J)$.

4 Experimental Evaluation

4.1 Japanese-English Relevant News Articles on Web News Sites

We collected Japanese and English news articles from a Web news site. Table 1 shows the total number of collected articles and the range of dates of those articles represented as the number of days. Table 1 also shows the number of articles updated in one day, and the average article size. The number of Japanese articles updated in one day are far greater (about 4 times) than that of English articles.

Table 1: Statistics of # of Days, Articles, and Article Sizes

	total # of days	total # of articles	average # of articles per day	average article size (bytes)
Eng	935	23064	24.7	3228.9
Jap	941	96688	102.8	837.7

[5] In the term frequency vectores, compound terms are restricted to be up to five words long both for English and Japanese.

[6] Eijiro Ver.37, 850,000 entries, `http://homepage3.nifty.com/edp/`.

Table 2: # of Japanese/English Articles Pairs with Similarity Values above Lower Bounds

Lower Bound L_d of Articles' Sim	w/o	0.3	0.4	0.5
Difference of Dates (days)	CLIR	≤ 2		
# of English Articles	23064	6073	2392	701
# of Japanese Articles	96688	12367	3444	882
# of English-Japanese Article Pairs	—	16507	3840	918

Next, for several lower bounds L_d of the similarity between English and Japanese articles, Table 2 shows the numbers of English and Japanese articles as well as article pairs which satisfy the similarity lower bound. Here, the difference of dates of English and Japanese articles is within two days, with which it is guaranteed that, if exist, closely related articles in the other language can be discovered (see Utsuro et al. (2003) for details). Note that it can happen that one article has similarity values above the lower bound against more than one articles in the other language.

According to our previous study (Utsuro et al., 2003), cross-lingually relevant news articles are available in the direction of English-to-Japanese retrieval for more than half of the retrieval query English articles. Furthermore, with the similarity lower bound $L_d = 0.3$, precision and recall of cross-language retrieval are around 30% and 60%, respectively. Therefore, with the similarity lower bound $L_d = 0.3$, at least 1,800 ($\approx 6,073 \times 0.5 \times 0.6$) English articles have relevant Japanese articles in the results of cross-language retrieval. Based on this analysis, the next section gives evaluation results with the similarity lower bound $L_d = 0.3$.

4.2 English Term List for Evaluation

For the evaluation of this paper, we first manually select target English terms and their reference Japanese translation, and examine whether reference bilingual term correspondences can be estimated by the methods presented in Sections 2 and 3. Target English terms are selected by the following procedure.

First, from the whole English articles of Table 1, any sequence of more than one words whose frequency is more than or equal to 10 is enumerated. This enumeration is easily implemented and efficiently computed by employing the technique of PrefixSpan (Pei et al., 2001). Here, certain portion of those word sequences are appropriate as compound terms, while the rest are some fragments of a compound term, or concatenation of those fragments. In order to automatically select candidates for correct compound terms, we parse those word sequences by Charniak parser[7], and collect noun phrases which consist of adjectives, nouns, and present/past participles. For each of those word sequences, the ϕ^2 statistic against Japanese translation candidates is calculated, then those word sequences are sorted in descending order of their ϕ^2 statistic. Finally, among top 3,000 candidates for compound terms, 100 English compound terms are randomly selected for the evaluation of this paper. Selected 100 terms satisfy the following condition: those English terms can be correctly translated neither by the MT software used in Section 2.2, nor by the bilingual lexicon used in Section 3.3.

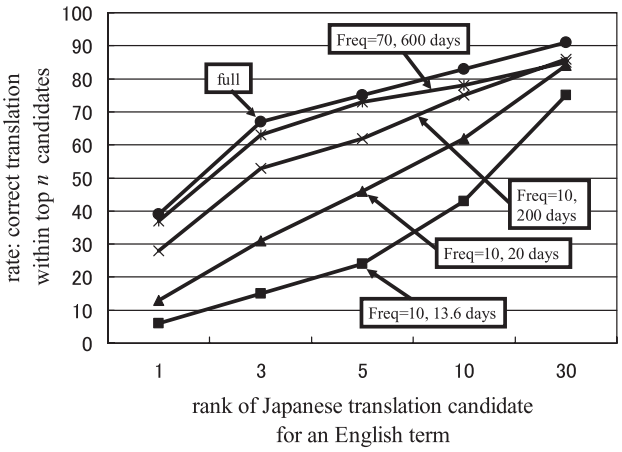

Figure 3: Accuracy of Estimating Bilingual Term Correspondences with News Articles

4.3 Estimating Bilingual Term Correspondences with News Articles

For the 100 English terms selected in the previous section, Japanese translation candidates which satisfy the condition of the formula (2) in Section 2.3 are collected, and are ranked according to the ϕ^2 statistic. Figure 3 plots the rate of reference Japanese translation being within top n candidates. In the figure, the plot labeled as "full" is the result with the whole articles in Table 1. In this case, the accuracy of the top ranked Japanese translation candidate is about 40%, and the rate of reference Japanese translation within top five candidates is about 75%.

[7] http://www.cs.brown.edu/people/ec/

Table 3: Statistics of Average Document Frequencies and Number of Days

Data Set	Document Frequencies of target English Term		# of Days	
	$df(t_E)$	$df(t_E, t_J)$	Eng	Jap
freq=10, 13.6 days	14.9	9.1	13.6	21.9
freq=10, 20 days	14.9	9.1	21.0	78.7
freq=10, 200 days	14.9	9.1	200	581
freq=70, 600 days	37.4	24.9	600	872
full	53.9	35.6	935	941

On the other hand, other plots labeled as "Freq=x, y days" are the results when the number of the news articles is reduced, which are simulations for estimating bilingual term correspondences for low frequency terms. Here, the label "Freq=x, y days" indicates that news articles used for ϕ^2 statistic estimation is restricted to certain portion of the whole news articles so that the following condition be satisfied: i) co-occurrence document frequency of a target English term and its reference Japanese translation is fixed to be x,[8] ii) the number of days be greater than or equal to y. For each news articles data set, Table 3 shows document frequencies $df(t_E)$ of a target English term t_E, co-occurrence document frequencies $df(t_E, t_J)$ of t_E and its reference Japanese translation t_J, and the numbers of days for English as well as Japanese articles. Those numbers are all averaged over the 100 English terms. The number of days for Japanese articles could be at maximum five times larger than that for English articles, because relevant Japanese articles are retrieved against a query English article from the dates of differences within two days (details are in Sections 2.2 and 4.1).

As can be seen from the plots of Figure 3, the smaller the news articles data set, the lower the plot is. Especially, in the case of the smallest news articles data set, it is clear that reliable ranking of Japanese translation candidates is difficult. This is because it is not easy to discriminate the reference Japanese translation and the other candidates with statistics obtained from such a small news articles data set.

4.4 Re-estimating Bilingual Term Correspondences with Monolingual Web Documents

For the 100 target English terms evaluated in the previous section, this section describes the result of applying the technique presented in Section 3.3, i.e., re-estimating bilingual term correspondences with monolingual Web documents. For each of the 100 target English terms, bilingual term correspondences are re-estimated against candidates of Japanese translation ranked within top 50 according to the ϕ^2 statistic. Here, as a simulation for terms that are infrequent in news articles, 50 candidate terms for Japanese translation are collected from the smallest data set labeled as "Freq=10, 13.6 days". As mentioned in Section 3.2, those 50 candidates are reduced to on the average 24.8 terms with the filtering by hits of search engines. For each of an English term t_E and a Japanese term t_J, 100 monolingual documents are collected by search engines[9] [10].

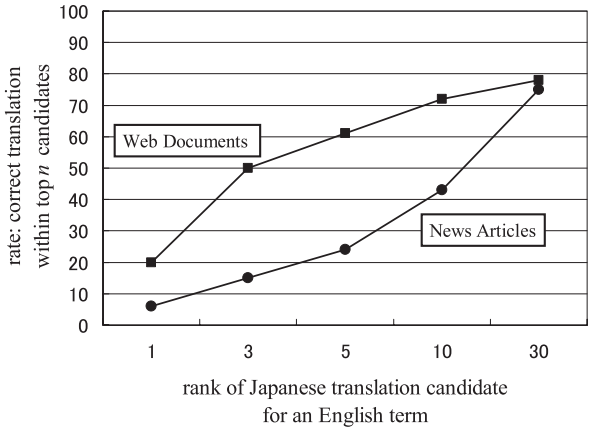

Figure 4: Accuracy of Re-estimating Bilingual Term Correspondences with Monolingual Web Documents

Figure 4 compares the plots of re-estimation with monolingual Web documents and estimation by news articles (data set "Freq=10, 13.6

[8] When the co-occurrence document frequency of t_E and t_J in the whole news articles is less than x, all the co-occurring dates are included.

[9] In the result of our preliminary evaluation, accuracy of re-estimating bilingual term correspondences did not improve even if more than 100 documents were used.

[10] Alternatively, as the monolingual documents from which contextual vectors are constructed, we evaluated each of the short passages listed in the summary pages returned by search engines, instead of the whole documents of the URLs listed in the summary pages. The difference of the performance of bilingual term correspondence estimation is little, while the computational cost can reduced to almost 5%.

days"). It is clear from this result that monolingual Web documents contribute to improving the accuracy of estimating bilingual term correspondences for low frequency terms.

One of the major reasons for this improvement is that topics of monolingual Web documents collected through search engines are much more diverse than those of news articles. Such diverse topics help discriminate correct and incorrect Japanese translation candidates. For example, suppose that the target English term t_E is "special anti-terrorism law" and its reference Japanese translation is "テロ対策特別措置法". In the news articles we used for evaluation, most articles in which t_E or t_J appear have *"dispatch of Self-Defense Force for reconstruction of Iraq"* as their topics. Here, Japanese translation candidates other than "テロ対策特別措置法" that are highly ranked according to the ϕ^2 statistic are: e.g., "衆院解散 (dissolution of the House of Representatives)" and "イラク復興支援 (assistance for reconstruction of Iraq)", which frequently appear in the topic of *"dispatch of Self-Defense Force for reconstruction of Iraq"*.

On the other hand, in the case of monolingual Web documents collected through search engines, it can be expected that topics of documents may vary according to the query terms. In the case of the example above, the major topic is *"dispatch of Self-Defense Force for reconstruction of Iraq"* for both of reference terms t_E and t_J, while major topics for other Japanese translation candidates are: *"issues on Japanese Diet"* for "衆院解散 (dissolution of the House of Representatives)" and *"issues on reconstruction of Iraq, not only in Japan, but all over the world"* for "イラク復興支援 (assistance for reconstruction of Iraq)". Those topics of incorrect Japanese translation candidates are different from that of the target English term t_E, and their contextual vector similarities against the target English term t_E are relatively low compared with the reference Japanese translation t_J. Consequently, the reference Japanese translation t_J is re-ranked higher compared with the ranking based on news articles.

5 Related Works

In large scale experimental evaluation of bilingual term correspondence estimation from comparable corpora, it is difficult to estimate bilingual term correspondences against every possible pair of terms due to its computational complexity. Previous works on bilingual term correspondence estimation from comparable corpora controlled experimental evaluation in various ways in order to reduce this computational complexity. For example, Rapp (1999) filtered out bilingual term pairs with low monolingual frequencies (those below 100 times), while Fung and Yee (1998) restricted candidate bilingual term pairs to be pairs of the most frequent 118 unknown words. Cao and Li (2002) restricted candidate bilingual compound term pairs by consulting a seed bilingual lexicon and requiring their constituent words to be translation of each other across languages. On the other hand, in the framework of bilingual term correspondences estimation of this paper, the computational complexity of enumerating translation candidates can be easily avoided with the help of cross-language retrieval of relevant news texts. Furthermore, unlike Cao and Li (2002), bilingual term correspondences for compound terms are not restricted to compositional translation.

6 Conclusion

In the framework of bilingual lexicon acquisition from cross-lingually relevant news articles on the Web, it has been relatively harder to reliably estimate bilingual term correspondences for low frequency terms. This paper proposed to complementarily use much larger monolingual Web documents collected by search engines, as a resource for reliably re-estimating bilingual term correspondences. We showed that, for the terms which appear infrequently in news articles, the accuracy of re-estimating bilingual term correspondences actually improved.

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.

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.

J. Pei, J. Han, B. Mortazavi-Asl, and H. Pinto. 2001. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proc. Inter. Conf. Data Mining*, pages 215–224.

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

T. Utsuro, T. Horiuchi, T. Hamamoto, K. Hino, and T. Nakayama. 2003. Effect of cross-language IR in bilingual lexicon acquisition from comparable corpora. In *Proc. 10th EACL*, pages 355–362.