Semi-automatic Compilation of Bilingual Lexicon Entries from Cross-Lingually Relevant News Articles on WWW News Sites

Takehito Utsuro, Takashi Horiuchi,
Yasunobu Chiba, and Takeshi Hamamoto

Department of Information and Computer Sciences,
Toyohashi University of Technology
Tenpaku-cho, Toyohashi 441-8580, Japan
{utsuro,takashi,chiba,hamamo}@cl.ics.tut.ac.jp

Abstract. For the purpose of overcoming resource scarcity bottleneck in corpus-based translation knowledge acquisition research, this paper takes an approach of semi-automatically acquiring domain specific translation knowledge from the collection of bilingual news articles on WWW news sites. This paper presents results of applying standard co-occurrence frequency based techniques of estimating bilingual term correspondences from parallel corpora to relevant article pairs automatically collected from WWW news sites. The experimental evaluation results are very encouraging and it is proved that many useful bilingual term correspondences can be efficiently discovered with little human intervention from relevant article pairs on WWW news sites.

1 Introduction

Translation knowledge acquisition from parallel/comparative corpora [4] 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, this paper focuses on bilingual news articles on WWW news sites as a source for translation knowledge acquisition. In the case of WWW 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

S.D. Richardson (Ed.): AMTA 2002, LNAI 2499, pp. 165–176, 2002.

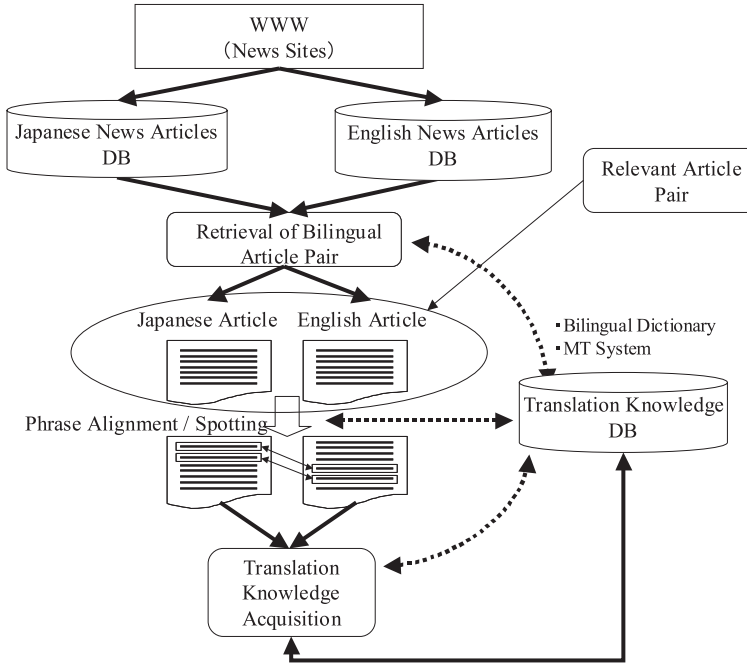


Fig. 1. Translation Knowledge Acquisition from WWW News Sites: Overview

their contents or at least report quite relevant topics. Based on this observation, we take an approach of semi-automatically acquiring translation knowledge of domain specific named entities, event expressions, and collocational functional expressions from the collection of bilingual news articles on WWW news sites.

Figure 1 illustrates the overview of our framework of translation knowledge acquisition from WWW news sites. First, pairs of Japanese and English news articles which report identical contents or at least closely related contents are retrieved. (Hereafter, we call pairs of bilingual news articles which report identical contents as *"identical"* pair, and those which report closely related contents (e.g., a pair of a crime report and the arrest of its suspect) as *"relevant"* pair.) Then, by applying term/phrase alignment techniques to Japanese and English news articles, various kinds of translation knowledge are acquired. In the process of translation knowledge acquisition, we allow human intervention if necessary. Especially, we aim at developing user interface facilities for efficient semi-automatic acquisition of translation knowledge, where previously studied techniques of translation knowledge acquisition from parallel/comparative corpora [4] are integrated in an optimal fashion.

Within this framework of translation knowledge acquisition from WWW news sites, this paper studies issues regarding cross-language retrieval and collection of *"identical"*/ *"relevant"* article pairs. We also present results of apply-

ing standard co-occurrence frequency based techniques of estimating bilingual term correspondences from parallel corpora [4] to those automatically collected *"identical"*/ *"relevant"* article pairs. The experimental evaluation results are very encouraging and it is proved that many useful bilingual term correspondences can be efficiently discovered with little human intervention from relevant article pairs on WWW news sites. Details of those evaluation results are presented.

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 WWW news sites. First, from WWW 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 each Japanese translation d_J^{MT} of the English articles are next segmented into word sequences by the Japanese morphological analyzer CHASEN (`http://chasen.aist-nara.ac.jp/`), and word frequency vectors v_J and v_J^{MT} are generated[2]. Then, cosine similarities between v_J and v_J^{MT} are calculated[3] and pairs of articles d_J and d_E (d_J^{MT}) which satisfy certain criterion are considered as candidates for *"identical"* or *"relevant"* article pairs.

3 Acquisition of Bilingual Term Correspondences from Relevant News Articles

3.1 Estimating Bilingual Term Correspondences

This section briefly describes the method of estimating bilingual term correspondences from the results of retrieving cross-lingually relevant English and Japanese news articles. As will be described in section 4.1, on WWW news sites in Japan, the number of articles updated per day is far greater (5∼30 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.

[1] As the commercial MT software, we chose English-Japanese Japanese-English Translation Software HONYAKUDAMASHII for Linux/BSD, OMRON SOFTWARE Co., Ltd, which is the one with slightly better performance than the others.

[2] After removing the most frequent 26 *hiragana* functional expressions as stop words, word frequency vectors are generated only from nouns and verbs.

[3] It is also quite possible to translate Japanese news articles into English and to calculate similarities of word frequency vectors in English side.

For an English query article d^i_E, let D^i_J denote the set of Japanese articles with cosine similarities higher than or equal to a certain lower bound L_d:

$$D^i_J = \left\{ d_J \mid \cos(d^i_E, d_J) \geq L_d \right\}$$

Then, we concatenate constituent Japanese articles of D^i_J into one article D'^i_J, and construct a pseudo-parallel corpus PPC_{EJ} of English and Japanese articles:

$$PPC_{EJ} = \left\{ \langle d^i_E, D'^i_J \rangle \mid D^i_J \neq \emptyset \right\}$$

Next, we apply standard techniques of estimating bilingual term correspondences from parallel corpora [4] to this pseudo-parallel corpus PPC_{EJ}. First, we extract monolingual (possibly compound) terms t_E and t_J which satisfy requirements on frequency lower bound and the upper bound of the number of constituent words. Then, based on the contingency table of co-occurrence 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 [4].

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

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:

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

3.2 Semi-automatic Acquisition of Bilingual Term Correspondences

This section describes the method of semi-automatic acquisition of bilingual term correspondences from the results of estimating bilingual term correspondences. Since our source of compiling bilingual lexicon entries is not clean parallel corpus, but artificially generated noisy pseudo-parallel corpus, it is difficult to compile bilingual lexicon entries full-automatically. In order to reduce the amount of human intervention necessary for selecting correctly estimated bilingual term correspondences, we divide the whole set of estimated bilingual term correspondences into subsets according to the following two criteria. First, we divide the whole set of estimated bilingual term correspondences into subsets, where each subset consists of English and Japanese term pairs which have a common English term. Next, we define the relation $t \succeq t'$ between two terms t and t' as t being identical with t' or the term t' constituting a part of the compound term

t. Then, for each English term t_E, only when any other English term t'_E does not satisfy the relation $t'_E \succeq t_E$, we construct the set $TP(t_E)$ of English and Japanese term pairs which have t_E or its sub-sequence term in the English side and satisfy the requirements on (co-occurrence) frequencies and term length in their constituent words as below:

$$TP(t_E) = \Big\{ \langle t'_E, t_J \rangle \mid t_E \succeq t'_E, freq(t_E) \geq L_f^E, freq(t_J) \geq L_f^J, \\ freq(t_E, t_J) \geq L_f^{EJ}, length(t_E) \leq U_l^E, length(t_J) \leq U_l^J \Big\}$$

We call the shared English term t_E of the set $TP(t_E)$ as *index*.

Next, all the sets $TP(t_E^1), \ldots, TP(t_E^m)$ are sorted in descending order of the maximum value $\hat{\phi^2}(TP(t_E))$ of ϕ^2 statistic of their constituent term pairs:

$$\hat{\phi^2}(TP(t_E)) = \max_{\langle t_E, t_J \rangle \in TP(t_E)} \phi^2(t_E, t_J)$$

Then, each set $TP(t_E^i)$ is examined by hand according to whether or not it includes correct bilingual term correspondences. Finally, we evaluate the following rate of containing correct bilingual term correspondences:

$$\text{rate of containing correct bilingual term correspondences} = \frac{\Big| \Big\{ TP(t_E) \mid \text{correct bilingual term correspondence } \langle t_E, t_J \rangle \in TP(t_E) \Big\} \Big|}{\Big| \Big\{ TP(t_E) \mid TP(t_E) \neq \emptyset \Big\} \Big|} \quad (1)$$

3.3 Example

Figure 2 illustrates the underlying idea of semi-automatic selection of correct bilingual term correspondences, with the help of browsing cross-lingually relevant article pairs. Suppose that an English compound term "Tokyo District Court" is chosen as the *index term* t_E. The figure lists the term pairs t_E and t_J with high values of ϕ^2 statistic in descending order, together with $freq(t_E)$, $freq(t_J)$, $freq(t_E, t_J)$, and $\phi^2(t_E, t_J)$. In this case, t_J with the highest value of ϕ^2 statistic is the correct Japanese translation of "Tokyo District Court". A human operator can select an arbitrary pair of English and Japanese terms t_E and t_J and then browse an English and Japanese article pair d_E and d_J, each of which contains t_E and t_J, respectively, and satisfies the similarity requirement $\cos(d_E, d_J) \geq L_d$. When the human operator browses such an article pair d_E and d_J, titles of English articles which contain t_E are first listed, and then, for each of the English articles, titles of Japanese articles which contain t_J and satisfy the similarity requirement are listed. Browsing through the title list as well as the body texts of the English and Japanese article pairs, the human operator can easily judge whether the selected term pair t_E and t_J is actually correct translation of each other. Even when the selected term pair is not correct translation, it is usually quite easy for the human operator to discover true

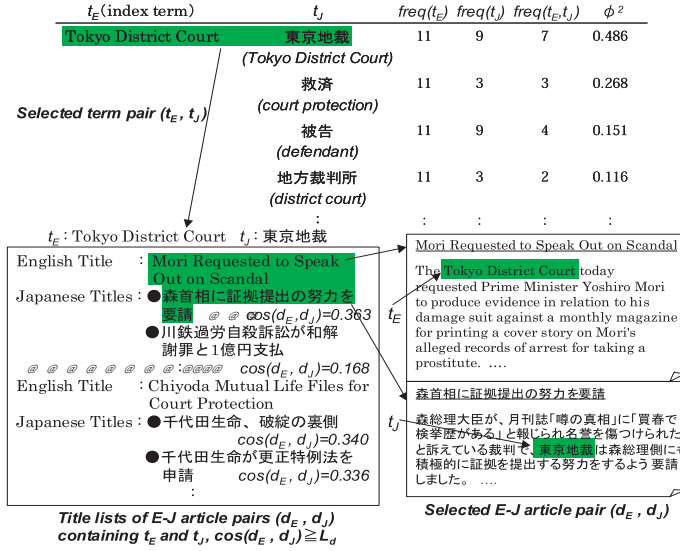


Fig. 2. Example of Semi-Automatic Selection of Bilingual Term Correspondences with Browsing Cross-Lingually Relevant Article Pairs

term correspondence if the selected article pair reports closely related contents. Otherwise, the human operator can quickly switch the article pair to the one which reports closely related contents.

4 Experimental Evaluation

4.1 Japanese-English Relevant News Articles on WWW News Sites

Table 1. Total # of Days, Total/Average # of Articles / Average Article Size / # of Reference Article Pairs for CLIR Evaluation

Site	Total # of Days		Total # of Articles		Average # of Articles per Day		Average Article Size (bytes)		# of Reference Article Pairs for CLIR Evaluation	
	Eng	Jap	Eng	Jap	Eng	Jap	Eng	Jap	Identical	Relevant
A	562	578	607	21349	1.1	36.9	1087.3	759.9	24	33
B	162	168	2910	14854	18.0	88.4	3135.5	836.4	28	82
C	162	166	3435	16166	21.2	97.4	3228.9	837.7	28	31

We collected Japanese and English news articles from three WWW news sites A, B, and C. 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 (5∼30 times)

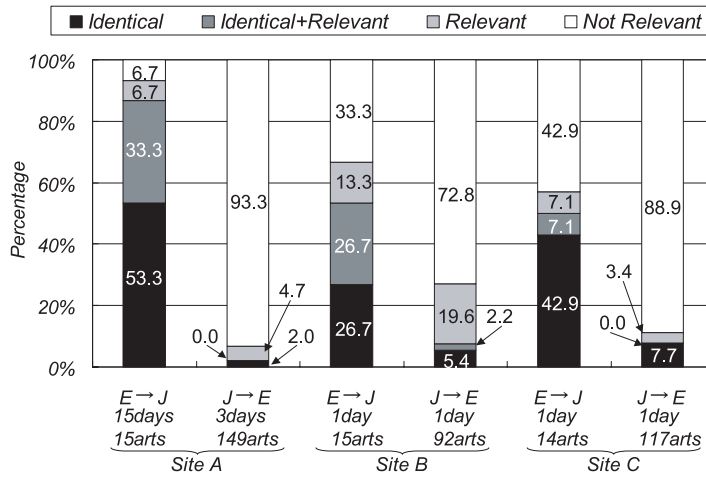


Fig. 3. Availability of Cross-Lingually *"Identical"*/ *"Relevant"* Articles

than that of English articles. In addition to that, the table gives the numbers of reference *"identical"*/ *"relevant"* article pairs manually collected for the evaluation of cross-language retrieval of relevant news articles. This evaluation result will be presented in the next section. In the case of those reference article pairs, the difference of dates between *"identical"* article pairs is less than ± 5 days, and that between *"relevant"* article pairs is around ± 10 days.

Next, Figure 3 shows rates of whether cross-lingually *"identical"* or *"relevant"* articles are available or not for each retrieval query article, where the following counts are recorded and their distributions are shown in the figure: i) the number of queries for which at least one *"identical"* article is available, but not any *"relevant"* article, ii) the number of queries for which at least one *"identical"* article and one *"relevant"* article are available, iii) the number of queries for which at least one *"relevant"* article is available, but not any *"identical"* article, iv) the number of queries for which neither *"identical"* nor *"relevant"* article is available. As can be clearly seen from these results, since the number of Japanese articles are far greater than that of English articles, the availability rate in Japanese-to-English retrieval is much lower than that in English-to-Japanese retrieval. The availability rate (either *"identical"* or *"relevant"*) in Japanese-to-English retrieval is around 10∼30%, while in English-to-Japanese retrieval, that for *"identical"* articles is more than 50%, and that for either *"identical"* or *"relevant"* increases by around 10% and more. These results guarantee that cross-lingually *"identical"* news articles are available in the direction of English-to-Japanese retrieval for more than half of the retrieval query English articles.

4.2 Cross-Language Retrieval of Relevant News Articles

Next, we evaluate the performance of cross-language retrieval of *"identical"* / *"relevant"* reference article pairs given in Table 1. In the direction of English to Japanese cross-language retrieval, precision/recall rates of the reference

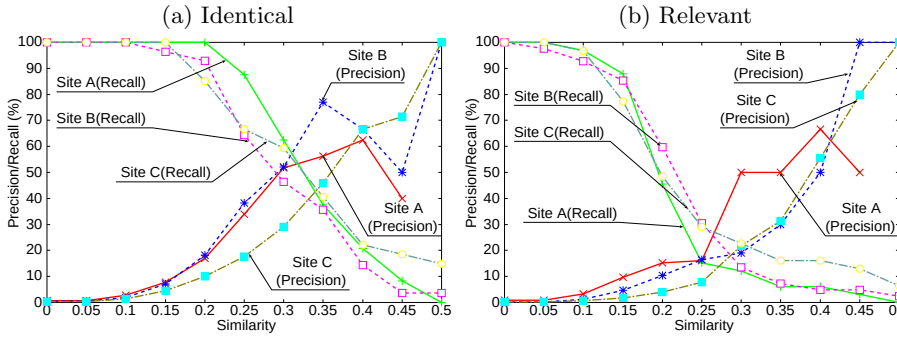


Fig. 4. Precision/Recall of Cross-Language Retrieval of Relevant News Articles (Article Similarity $\geq L_d$)

"identical"/ "relevant" articles against those with the similarity values above the lower bound L_d are measured, and their curves against the changes of L_d are shown in Figure 4. The difference of dates of English and Japanese articles is given as the maximum range of dates, with which all the cross-lingually *"identical"/ "relevant"* articles can be discovered (less than $\pm$ 5 days for the *"identical"* article pairs and around $\pm$ 10 days for the *"relevant"* article pairs). Let DP_{ref} denote the set of reference article pairs within the range of dates, the precise definitions of the precision and recall rates of this task are given below:

$$\text{precision} = \frac{|\{d_J \mid \exists d_E, \langle d_E, d_J\rangle \in DP_{ref}, \cos(d_E, d_J) \geq L_d\}|}{|\{d_J \mid \exists d_E \exists d'_J, \langle d_E, d'_J\rangle \in DP_{ref}, \cos(d_E, d_J) \geq L_d\}|}$$

$$\text{recall} = \frac{|\{d_J \mid \exists d_E, \langle d_E, d_J\rangle \in DP_{ref}, \cos(d_E, d_J) \geq L_d\}|}{|\{d_J \mid \exists d_E, \langle d_E, d_J\rangle \in DP_{ref}\}|}$$

In the case of *"identical"* article pairs, Japanese articles with the similarity values above 0.4 have precision of around 40% or more[4].

4.3 Semi-automatic Acquisition of Bilingual Term Correspondences from Relevant News Articles

In this section, we evaluate our framework of semi-automatic acquisition of bilingual term correspondences from relevant news articles. For the news sites A, B, and C, and for several lower bounds L_d of the similarity between English and Japanese articles, Table 2 shows the numbers of English and Japanese articles which satisfy the similarity lower bound[5]. Then, under the conditions

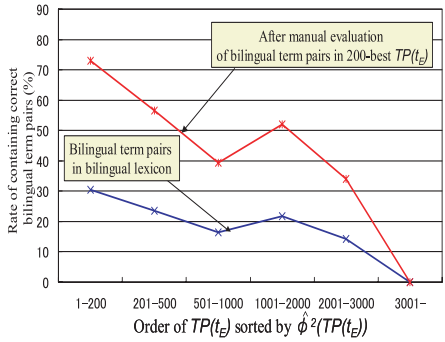
[4] We are now working on examining usefulness of additional clues such as titles, pronunciation of foreign names, and numerical expressions, and furthermore on incorporating them for the purpose of improving the performance of cross-language retrieval.

[5] It can happen that one Japanese article is retrieved by more than one English query articles. In such cases, the occurrence of Japanese articles is duplicated.

Table 2. Numbers of Japanese/English Articles Pairs with Similarity Values above the Lower Bounds

Site	A				B		C	
Lower Bound L_d of Articles' Sim	0.25	0.3	0.4	0.5	0.4	0.5	0.4	0.5
Difference of Dates (days)	± 4				± 3		± 2	
# of English Articles	473	362	190	74	415	92	453	144
# of Japanese Articles	1990	1128	377	101	631	127	725	185

(a) Rates of Containing Correct Bilingual Term Pairs (Site A, $L_d = 0.3$)



(b) Ranks of 146 Correct Bilingual Term Pairs within 200-best $TP(t_E)$, Sorted by ϕ^2 (Site A, $L_d = 0.3$)

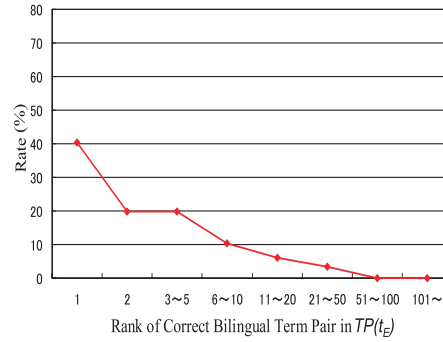


Fig. 5. Evaluation Results using Bilingual Term Pairs in a Bilingual Lexicon / by Manual Evaluation

$L_f^E = L_f^J = 3, L_f^{EJ} = 2, U_l^E = U_l^J = 5$ (the difference of dates of English and Japanese articles is given as the maximum range of dates, with which all the cross-lingually *"identical"* articles can be discovered), the sets $TP(t_E)$ are constructed and the "rate of containing correct bilingual term correspondences" in the equation (1) (section 3.2) is evaluated.

For the site A with the similarity lower bound $L_d = 0.3$, the rates of containing correct bilingual term pairs taken from an existing bilingual lexicon (Eijiro Ver.37, 850,000 entries, `http://member.nifty.ne.jp/eijiro/`) are shown in Figure 5 (a) as "Bilingual term pairs in bilingual lexicon". This result supports the usefulness of ϕ^2 statistic in this task, since the rate of containing correct bilingual term pairs tends to decrease as the order of $TP(t_E)$ sorted by $\hat{\phi}^2(TP(t_E))$ becomes lower. Furthermore, topmost 200 $TP(t_E)$ according to the ϕ^2 statistic $\hat{\phi}^2(TP(t_E))$ are examined by hand and 146 bilingual term pairs contained in the topmost 200 $TP(t_E)$ are judged as correct. This manual evaluation result indicates that, compared with the bilingual term pairs found in the existing bilingual lexicon, about 1.4 times those found in the existing bilingual lexicon can be acquired from the topmost 200 $TP(t_E)$. Figure 5 (a) also shows the estimated plot of "After manual evaluation of bilingual term pairs in 200-best $TP(t_E)$", which is the rate of containing correct bilingual term pairs taken from the existing bilingual lexicon, multiplied by the ratio of about 2.4 (i.e., 146 of

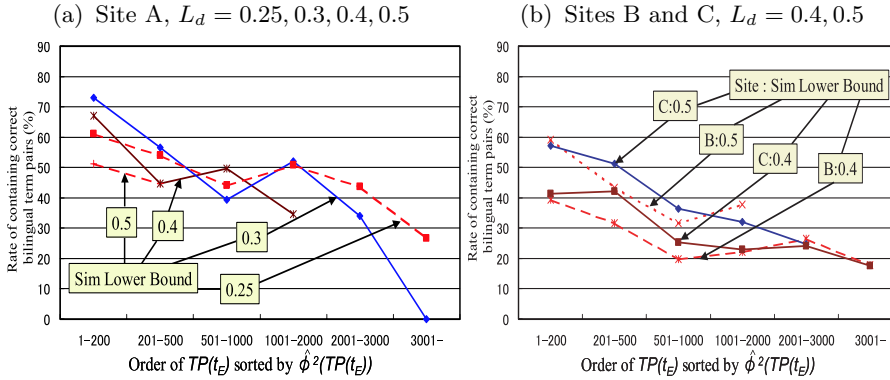


Fig. 6. Rates of Containing Correct Bilingual Term Pairs

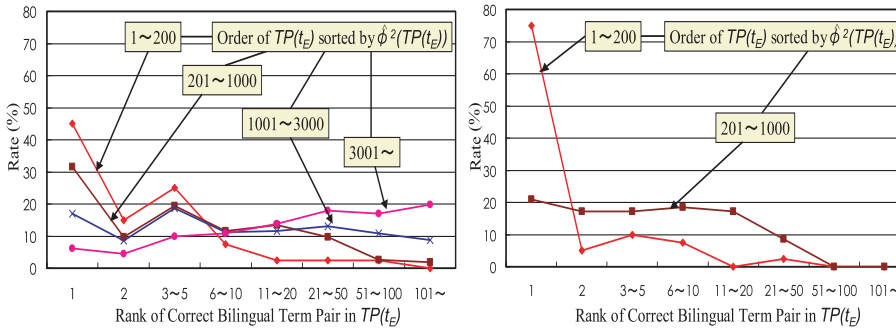


Fig. 7. Ranks of Correct Bilingual Term Pairs within a $TP(t_E)$, Sorted by ϕ^2 (Site A, Bilingual Term Pairs taken from a Bilingual Lexicon)

those judged as correct by manual evaluation/61 of those found in the existing bilingual lexicon).

Next, for the similarity lower bound $L_d = 0.25, 0.3, 0.4, 0.4$ (site A) and $L_d = 0.4, 0.5$ (sites B and C), estimated plots of rates of containing correct bilingual term pairs judged as correct by manual evaluation (i.e., those for correct bilingual term pairs taken from the existing bilingual lexicon, multiplied by the ratio of about 2.4) are shown in Figure 6. As can be seen from these results, the lower the similarity lower bound L_d is, the more the number of articles retrieved is and the more the number of candidate bilingual term pairs is, which is indicated by the difference of the lengths of those plots. For the site A, and for the sites B and C when the similarity lower bound $L_d = 0.5$, rates of containing correct bilingual term pairs are over 40% within the top 500 $TP(t_E)$. These rates are high enough for efficient human intervention in semi-automatic compilation of bilingual lexicon entries. Furthermore, the rates of containing correct bilingual term pairs are comparable among those three sites, even though the availability rates of cross-lingually *"identical"*/ *"relevant"* articles are much lower for the sites

B and C than for the site A (Figure 3). This result is very encouraging because news sites with less availability rates of cross-lingually *"identical"/"relevant"* articles are still very useful in our framework and it proves the effectiveness of our approach[6].

Finally, we evaluate the rank of correct bilingual term correspondences within each set $TP(t_E)$, sorted by ϕ^2 statistic. Within a set $TP(t_E)$, estimated Japanese term translation t_J are sorted by $\phi^2(t_E, t_J)$, and the ranks of correct Japanese translation of t_E are recorded. For the site A with the similarity lower bound $L_d = 0.3$, Figure 5 (b) shows the distribution of the ranks of correctly estimated Japanese terms for the 146 bilingual term pairs, which are contained in the topmost 200 $TP(t_E)$ and judged as correct. This result indicates that about 90% of those correct bilingual term pairs are included within the 10-best candidates in each $TP(t_E)$. For the site A with the similarity lower bound $L_d = 0.3, 0.5$, Figure 7 also shows this distribution for the correct bilingual term pairs taken from the existing bilingual lexicon. These results also support the usefulness of ϕ^2 statistic in this task, since the relative orders of correct bilingual term pairs tend to become lower as the order of $TP(t_E)$ sorted by $\hat{\phi^2}(TP(t_E))$ becomes lower. The criterion of the ϕ^2 statistic can be regarded as quite effective in reducing the amount of human intervention necessary for selecting correctly estimated bilingual term correspondences[7]. Furthermore, comparing the results of Figure 7 (a) and (b), the relative orders of correct bilingual term pairs become significantly higher when the similarity lower bound L_d is high. This result claims that the efficiency of semi-automatic acquisition of bilingual term pairs greatly depends on the accuracy of retrieving cross-lingually relevant news articles.

5 Related Works

Previously studied techniques of estimating bilingual word correspondences from non-parallel corpora (e.g., [1]) are based on the idea that semantically similar words appear in similar contexts. In those techniques, frequency information of contextual words co-occurring in the monolingual text is stored and their similarity is measured across languages. One of the most important difference between our approach and those techniques for translation knowledge acquisition from non-parallel corpora is that, we estimate bilingual term correspondences after selecting relevant article pairs, while in the latter techniques, co-occurrence frequency information is collected from the whole monolingual text. One of the

[6] We also evaluate Japanese used as the language of the *index* term of each set TP and compare the "rate of containing correct bilingual term correspondences" with those with English *index* terms. Since the number of Japanese articles is far greater than that of English articles, this rate with Japanese *index* terms becomes lower for the similarity lower bounds $L_d \leq 0.4$.

[7] It is also very important to note that the results of this paper can be easily improved by employing more sophisticated techniques of estimating bilingual compound term correspondences from parallel corpora (e.g., [2]), especially in the performance of selecting appropriate monolingual compound terms in each language.

major contribution of our work to the community of translation knowledge acquisition from parallel/comparable corpora is that: we showed even with standard techniques of estimating bilingual term correspondences from *parallel* corpora, many useful bilingual term correspondences can be efficiently discovered with little human intervention from relevant article pairs on WWW news sites. Furthermore, our results can be improved by incorporating those techniques based on co-occurrence frequency information in monolingual text (e.g., [1]), which are robust against noisy parallel corpora like those used in our work.

Related works on automatic document alignment between two languages include [3], which, in the context of cross-language information retrieval (CLIR) research, proposed to apply bootstrapping technique to an existing corpus-based CLIR approach for the task of extracting bilingual text pairs. Previous works on automatic document alignment mainly focused on the performance of automatic document alignment. Another type of related works include an approach of collecting partially bilingual texts from WWW [5]. One advantage of this approach is that it is applicable to various domains that infrequently become topics of news articles, although there might exist the case that the quality of translation by non-natives is possibly low. On the other hand, one of the advantages of our approach of employing bilingual news articles on WWW news sites as a source for translation knowledge acquisition is that high translation quality is guaranteed and articles of up-to-date topics are updated everyday.

6 Conclusion

Within the framework of translation knowledge acquisition from WWW news sites, this paper presented results of applying standard co-occurrence frequency based techniques of estimating bilingual term correspondences from parallel corpora to relevant article pairs automatically collected from WWW news sites. The experimental evaluation results were very encouraging and it was proved that many useful bilingual term correspondences can be efficiently discovered with little human intervention from relevant article pairs on WWW news sites.

References

1. Fung, P. and Yee, L. Y.: An IR Approach for Translating New Words from Nonparallel, Comparable Texts, *Proc. 17th COLING and 36th ACL* (1998) 414–420 175, 176
2. Haruno, M., Ikehara, S. and Yamazaki, T.: Learning Bilingual Collocations by Word-Level Sorting, *Proc. 16th COLING* (1996) 525–530 175
3. Masuichi, H., Flournoy, R., Kaufmann, S. and Peters, S.: A Bootstrapping Method for Extracting Bilingual Text Pairs, *Proc. 18th COLING* (2000) 1066–1070 176
4. Matsumoto, Y. and Utsuro, T.: Lexical Knowledge Acquisition, Dale, R., Moisl, H. and Somers, H. (eds.), *Handbook of Natural Language Processing*, chapter 24, Marcel Dekker Inc. (2000) 563–610 165, 166, 167, 168, 168
5. Nagata, M., Saito, T. and Suzuki, K.: Using the Web as a Bilingual Dictionary, *Proc. Workshop on Data-driven Methods in Machine Translation* (2001) 95–102 176