Effect of Cross-Language IR in Bilingual Lexicon Acquisition from Comparable Corpora

Takehito Utsuro
Graduate School of Informatics,
Kyoto University
Sakyo-ku, Kyoto, 606-8501, Japan
utsuro@i.kyoto-u.ac.jp

Takashi Horiuchi and **Kohei Hino**
Takeshi Hamamoto and **Takeaki Nakayama**
Dpt. Information and Computer Sciences,
Toyohashi University of Technology
Tenpaku-cho, Toyohashi, 441–8580, Japan

Abstract

Within the framework of translation knowledge acquisition from WWW news sites, this paper studies issues on the effect of cross-language retrieval of relevant texts in bilingual lexicon acquisition from comparable corpora. We experimentally show that it is quite effective to reduce the candidate bilingual term pairs against which bilingual term correspondences are estimated, in terms of both computational complexity and the performance of precise estimation of bilingual term correspondences.

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, 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 their contents or at least report quite relevant topics. Based on this observation, we take an approach of acquiring translation knowledge of domain specific named entities, event expressions, and collocational expressions from the collection of bilingual news articles on WWW news sites (Utsuro and others, 2002).

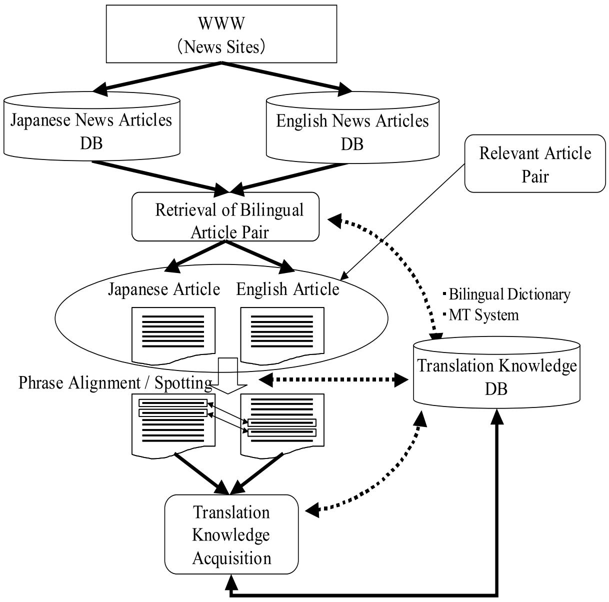


Figure 1: Translation Knowledge Acquisition from WWW News Sites: Overview

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 previously studied techniques of translation knowledge acquisition from parallel/comparative corpora, various kinds of translation knowledge are acquired.

Within this framework of translation knowledge acquisition from WWW news sites, this paper studies issues on the effect of cross-language retrieval of relevant texts in bilingual lexicon acquisition from comparable corpora. First, we show that, due to its computational complexity, it is difficult to straightforwardly apply previously studied techniques of bilingual term correspondence estimation from comparable corpora, especially in the case of large scale evaluation such as those presented in this paper. Then, we show that, with the help of cross-language retrieval of relevant texts, this computational difficulty can be easily avoided by reducing the candidate bilingual term pairs against which bilingual term correspondences are estimated. It is also experimentally shown that candidate reduction with the help of cross-language retrieval of relevant texts is quite effective in improving the performance of precise estimation of bilingual term correspondences.

2 Acquisition of Bilingual Term Correspondences from Comparable Corpora

Previously studied techniques of estimating bilingual term correspondences from comparable corpora are mostly based on the idea that semantically similar words appear in similar contexts (Fung, 1995; Rapp, 1995; Kaji and Aizono, 1996; Tanaka and Iwasaki, 1996; Fung and Yee, 1998; Rapp, 1999; Tanaka, 2002). In those techniques, frequency information of contextual words co-occurring in the monolingual text is stored and their similarity is measured across languages.

The following gives a rough formalization of the previous approaches to acquiring bilingual term correspondences from comparable corpora. Suppose that CC_E and CC_J denote an English corpus and a Japanese corpus, respectively, and that they can be considered as comparable corpora. Then, in the previous approaches, for each English term t_E in CC_E and each Japanese term t_J in CC_J, occurrences of surrounding words are recorded in the form of some vector $cv(t_E, CC_E)$ and $cv(t_J, CC_J)$, respectively[1]. In previous works, as weights of these contextual vectors, word frequencies or modified weights such as $tf \cdot idf$ are used. Finally, for every pair of an English term t_E and a Japanese term t_J, bilingual term correspondence $corr_{EJ}(t_E, t_J)$ is estimated in terms of a certain similarity measure $sim(cv(t_E, CC_E), cv(t_J, CC_J))$ between contextual vectors $cv(t_E, CC_E)$ and $cv(t_J, CC_J)$:

$$corr_{EJ}(t_E, t_J) \quad \equiv \quad sim_{EJ}(cv(t_E, CC_E), cv(t_J, CC_J))$$

Here, in the modeling of contextual similarities across languages, earlier works such as Fung (1995), Rapp (1995), and Tanaka and Iwasaki (1996) studied to measure the similarities of contextual co-occurrence patterns across languages without the help of any existing bilingual lexicons. On the other hand, later works such as Kaji and Aizono (1996), Fung and Yee (1998), Rapp (1999), and Tanaka (2002) studied to exploit existing bilingual lexicons as initial seed for modeling of contextual similarities across languages. As the similarity measure $sim(cv(t_E, CC_E), cv(t_J, CC_J))$ between contextual vectors $cv(t_E, CC_E)$ and $cv(t_J, CC_J)$, measures such as cosine measure, dice coefficient, and Jaccard coefficient are used.

3 Acquisition of Bilingual Term Correspondences from Cross-Lingually Relevant Texts

3.1 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 (Utsuro and others, 2002). 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[2]. Each Japanese article

[1] In most previous works, surrounding words that are considered as contexts of a term are those that co-occur in the same sentence, or in a window of a few words.

[2] In this query translation process, we also evaluated simply consulting a bilingual lexicon instead of employing an MT software. As reported in Collier and others (1998), the precision of simple word by word query translation with a bilingual lexicon is much lower than that with an MT software. Since we prefer precision rather than recall in our experiments, in this paper, we show results with query translation by an MT software.

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[3] and pairs of articles d_J and d_E which satisfy certain criterion are considered as candidates for *"identical"* or *"relevant"* article pairs.

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. For each English query article d_E^i in CC_E and its Japanese translation d_J^{MTi}, the set D_J^i of Japanese articles with cosine similarities higher than or equal to a certain lower bound L_d is constructed:

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

3.2 Estimating Bilingual Term Correspondences

This section describes the techniques we apply to the task of estimating bilingual term correspondences from cross-lingually relevant texts. Here, we compare several techniques in order to evaluate the effect of cross-language retrieval of relevant texts in the performance of acquiring bilingual term correspondences from comparable corpora. In the first technique, we regard cross-lingually relevant texts as a pseudo-parallel corpus, where standard techniques of estimating bilingual term correspondences from parallel corpora are employed. In the second technique, we regard cross-lingually relevant texts as a comparable corpus, where bilingual term correspondences are estimated in terms of contextual similarities across languages. In this second approach, we further evaluate the effect of cross-language retrieval of relevant texts by comparing the cases with/without reducing candidates of bilingual term pairs with the help of cross-lingually relevant text pairs.

3.2.1 Estimation based on Pseudo-Parallel Corpus

Here, we describe how to estimate bilingual term correspondences from cross-lingually relevant texts by regarding them as a pseudo-parallel corpus. 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) term pair t_E and t_J:

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

where those term pairs are possibly required to satisfy frequency lower bounds 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.

	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 as the bilingual term correspondence $corr_{EJ}(t_E, t_J)$:

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

3.2.2 Estimation based on Contextual Similarity

Next, we describe how to estimate bilingual term correspondences from cross-lingually relevant texts by regarding them as a comparable corpus. Here, when selecting the candidates of bilingual term pairs against which bilingual term correspondences are estimated, we evaluate two approaches. In the first approach, as described in section 2 for the case of acquisition from comparable corpora, for every pair of an English term and a Japanese term, bilingual term correspondence is

[3] 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. We are planning to evaluate those alternatives in cross-language retrieval of relevant news articles.

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

	Total # of Days		Total # of of Articles		Average # of Articles per Day		Average Article Size (bytes)	
Site	Eng	Jap	Eng	Jap	Eng	Jap	Eng	Jap
A	562	578	607	21349	1.1	36.9	1087.3	759.9
B	162	168	2910	14854	18.0	88.4	3135.5	836.4
C	162	166	3435	16166	21.2	97.4	3228.9	837.7

estimated. In the second approach, on the other hand, as described in the previous section for the case of acquisition from (pseudo-) parallel corpora, the candidates of bilingual term pairs are selected from a pseudo-parallel sentence pair d_E^i and $D_J^{\prime i}$ as in the formula (2). In this second approach, we intend to evaluate the effect of cross-language retrieval of relevant texts in the performance of acquiring bilingual term correspondences from comparable corpora, i.e., in reducing useless bilingual term pairs and in increasing the estimated confidence of useful bilingual term pairs.

More specifically, first, a reduced but cross-lingually more relevant comparable corpus is constructed from the result of cross-language retrieval of relevant news articles in section 3.1. Referring to the definition of the set D_J^i of relevant Japanese articles in the equation (1), the reduced English corpus RC_E is constructed by collecting English query articles each of which has at least one relevant Japanese article:

$$RC_E \quad = \quad \left\{ d_E^i \in CC_E \mid D_J^i \neq \emptyset \right\}$$

Next, the reduced Japanese corpus RC_J that is cross-lingually relevant to RC_E is constructed by collecting those relevant Japanese articles:

$$RC_J \quad = \quad \bigcup_{d_E^i \in RC_E} D_J^i$$

Then, for each English term t_E in RC_E and each Japanese term t_J in RC_J, occurrences of surrounding words are recorded in the form of some vector $cv(t_E, RC_E)$ and $cv(t_J, RC_J)$, respectively[4]. Here, more precisely, the contextual vector $cv(t_E, RC_E)$ of an English term t_E is constructed by summing up the word frequency vector $v(s_J^{MTi})$ of Japanese translation s_J^{MTi} of each English *sentence* s_E^i which contains t_E:

$$cv(t_E, RC_E) \quad = \sum_{\forall s_E^i \text{ in } RC_E \text{ s.t. } t_E \in s_E^i} v(s_J^{MTi})$$

Finally, bilingual term correspondence $corr_{EJ}(t_E, t_J)$ is estimated in terms of a certain similarity measure sim_{EJ} between contextual vectors $cv(t_E, RC_E)$ and $cv(t_J, RC_J)$:

$$corr_{EJ}(t_E, t_J) \quad \equiv \quad sim_{EJ}(cv(t_E, RC_E), cv(t_J, RC_J))$$

In the experimental evaluation, we show results with cosine measure as the similarity measure $sim_{EJ}(cv(t_E, RC_E), cv(t_J, RC_J))$. Here, when selecting the candidates of bilingual term pairs, we compare the two approaches mentioned above.

4 Experimental Evaluation

4.1 Japanese-English Relevant News Articles on WWW News Sites

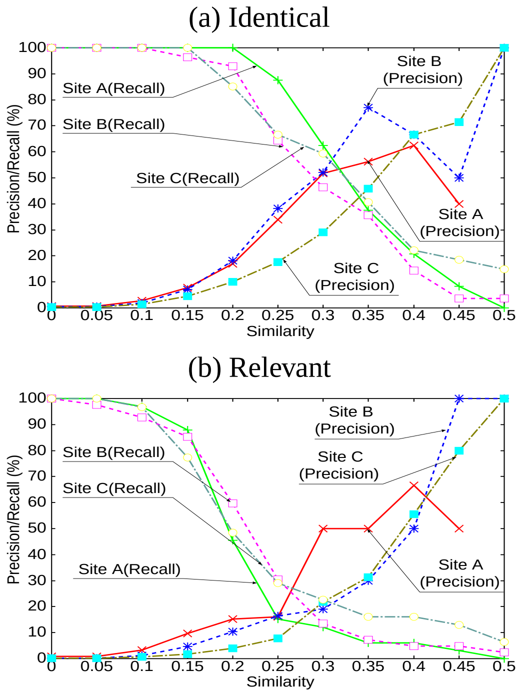
We collected Japanese and English news articles from three WWW news sites A, B, and C. Table 1 shows the total number of collected articles and the range of dates of those articles represented as the number of days. Table 1 also shows the number of articles updated in one day, and the average article size. The number of Japanese articles updated in one day are far greater (5∼30 times) than that of English articles. Then, for each of the three sites and for each of the two classes *"identical"*/*"relevant"*, we manually collected 50 (i.e., $50 \times 3 \times 2 = 300$ in total) reference article pairs for the evaluation of cross-language retrieval of relevant news articles[5]. This evaluation result will be presented in the next section.

4.2 Cross-Language Retrieval of Relevant News Articles

We evaluate the performance of cross-language retrieval of *"identical"* / *"relevant"* reference article pairs (Utsuro and others, 2002). In the direction of English to Japanese cross-language retrieval, precision/recall rates of the reference

[4] In the experimental evaluation, we show results where surrounding words that are considered as contexts of a term are those that co-occur in the same *sentence*. We also experimentally evaluated weights of vectors other than word frequencies such as $tf \cdot idf$, where its performance is quite similar to that of word frequency vectors.

[5] 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. We also examined the rates of whether at least one cross-lingually *"identical"* article is available for each retrieval query article (Utsuro and others, 2002). 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.


(a) Identical

(b) Relevant

Figure 2: Precision/Recall of Cross-Language IR of Relevant News Articles (Article Sim $\geq L_d$)

"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 2. 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 (here, $\cos(d_E, d_J) \equiv \cos(v(d_J^{MT}), v(d_J)))$:

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

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

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

4.3 Estimation of Bilingual Term Correspondences

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 (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). In the evaluation of estimating bilingual term correspondences, 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. We construct the set $TP(t_E)$ of English and Japanese term pairs which have t_E in the English side and satisfy the requirements on (co-occurrence) frequencies and term length in their constituent words as below:

$$TP(t_E) = \Big\{\langle t_E, t_J\rangle \mid freq(t_E) \geq L_f^E, freq(t_J) \geq L_f^J,$$
$$freq(t_E, t_J) \geq L_f^{EJ}, length(t_E) \leq U_l^E, length(t_J) \leq U_l^J\Big\}$$

(In the following, we show results under the conditions $L_f^E = L_f^J = 3, L_f^{EJ} = 2, U_l^E = U_l^J = 5$). 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 of the bilingual term correspondence $corr_{EJ}(t_E, t_J)$ among their constituent term pairs. We denote this maximum value as $corr_{EJ}(TP(t_E))$:

$$corr_{EJ}(TP(t_E)) = \max_{\langle t_E, t_J\rangle \in TP(t_E)} corr_{EJ}(t_E, t_J)$$

4.3.1 Numbers of Bilingual Term Pairs

First, for the site A with the similarity lower bound $L_d = 0.3$, topmost 200 $TP(t_E)$ according to the maximum bilingual term correspondence $corr_{EJ}(TP(t_E))$ are examined by hand and 146 bilingual term pairs contained in the topmost 200 $TP(t_E)$ are judged as correct. We compared those 146 bilingual term pairs with an existing bilingual lexicon (Eijiro Ver.37, 850,000 entries, `http://member.nifty.ne.jp/eijiro/`), where 86 of them (almost 60%) are not included in the existing bilingual lexicon. This manual evaluation result indicates that it is quite possible to extend a large scale existing bilingual lexicon such as the one used in our evaluation.

Next, Table 3 lists the numbers of English and Japanese monolingual terms, those of candidate term pairs against which bilingual term correspondences are estimated, and those of term pairs found in the existing bilingual lexicon. The rows with "(without CLIR)" show statistics for the whole comparable corpus CC_E and CC_J. The rows with "L_d (with CLIR)" show lower bounds of article similarities and statistics for the cross-lingually relevant English corpus RC_E and Japanese corpus RC_J, that are reduced from the whole comparable corpus CC_E and CC_J. The columns with "reduced" show statistics when the candidate bilingual term pairs are selected from a pseudo-parallel sentence pair as in the formula (2). The columns with "full" shows statistics when

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	0.3	0.4	0.5	0.4	0.5	0.4	0.5
Difference of Dates (days)	± 4			± 3		± 2	
# of English Articles	362	190	74	415	92	453	144
# of Japanese Articles	1128	377	101	631	127	725	185

Table 3: Numbers of Japanese/English Terms and Bilingual Term Pairs

Site			# of Monolingual Terms		Candidate Term Pairs			Term Pairs Found in an Existing Bilingual Lexicon		
					# of Term Pairs		rate (full/ reduced)	# of Term Pairs		rate (full/ reduced)
			English	Japanese	reduced	full		reduced	full	
A	L_d (with CLIR)	0.5	780	737	52435	574860	11.0	141	285	2.0
		0.4	2684	3231	427889	8672004	20.3	543	1467	2.7
		0.3	5463	8119	1639714	44354097	27.1	1298	3492	2.7
	without CLIR		9265	65324	—	605226860	—	—	n/a	—
B	L_d (with CLIR)	0.5	2468	2158	494544	5325944	10.8	507	1206	2.4
		0.4	11968	8658	4074980	103618944	25.4	2155	n/a	—
	without CLIR		97998	71638	—	7020380724	—	—	n/a	—
C	L_d (with CLIR)	0.5	3760	2612	638089	9821120	15.4	753	1860	2.5
		0.4	13200	9433	4367775	124515600	28.5	2353	n/a	—
	without CLIR		119071	82055	—	9770370905	—	—	n/a	—

full: every term pair, *reduced*: term pairs found in a pseudo-parallel sentence pair, *n/a*: due to time complexity,

the candidate bilingual term pairs are every pair of an English term found in RC_E or CC_E and a Japanese term found in RC_J or CC_J. For the moment, several numbers are unavailable (marked with "n/a") due to time complexity[6].

It is very important to compare the column "rate (full/reduced)" for the numbers of candidate term pairs with that for the numbers of term pairs found in the existing bilingual lexicon. The candidate term pairs can be *reduced* to about 3.5∼10% of their original sizes with the help of a pseudo-parallel sentence pair, while about 37∼50% of the correct bilingual term pairs found in the existing bilingual lexicon are preserved. Therefore, candidate reduction with the help of a pseudo-parallel sentence pair is quite effective in removing useless term pairs while preserving useful ones. This result clearly supports our claim on the usefulness of cross-language retrieval of relevant texts in acquisition of bilingual term correspondences.

4.3.2 Rates of Containing Correct Bilingual Term Pairs

Next, we evaluate the following rate of containing correct bilingual term correspondences:

$$\begin{array}{c}\text{rate of}\\ \text{correct}\\ \text{bilingual}\\ \text{term}\\ \text{correspon-}\\ \text{dences}\end{array} = \frac{\left|\left\{TP(t_E) \mid \text{correct bilingual term correspondence } \langle t_E, t_J\rangle \in TP(t_E)\right\}\right|}{\left|\left\{TP(t_E) \mid TP(t_E) \neq \emptyset\right\}\right|}$$

where the correctness of the estimated bilingual term correspondences is judged against the existing bilingual lexicon. For the site A with the similarity lower bound $L_d = 0.4$, Figure 3 plots the changes in this rate against the order of $TP(t_E)$ sorted by $corr_{EJ}(TP(t_E))$ (we have similar results with other similarity lower bounds L_d and for other sites B and C). In the figure, "pseudo-parallel with CLIR" indicates the plot for estimating bilingual term correspondence based on the pseudo-parallel corpus technique described in section 3.2.1. "Contextual similarity with CLIR" indicates the plots for estimation based on contextual similarity described in section 3.2.2, where in "reduced", the candidates of bilingual term pairs are selected from a pseudo-parallel sentence pair

[6] The computational complexity of bilingual term correspondence estimation based on contextual similarity in comparable corpora (sections 2 and 3.2.2) is much more than that based on pseudo-parallel corpus (section 3.2.1). The whole process of estimating bilingual term correspondences for "without CLIR" (i.e., from the whole comparable corpus CC_E and CC_J by the technique described in section 2), for the site A, would take about 6 days on a PentiumIV 1.9GHz processor. For the sites B and C, $L_d = 0.4$, it would take 3 ∼ 6 days for the processes for "with CLIR: full" (i.e., when the candidates of bilingual term pairs are every pair of an English term found in RC_E and a Japanese term found in RC_J) to complete. Furthermore, in the case of such large scale experiments as ours (e.g., for the sites B and C), where frequency lower bounds are very low and compound terms are assumed to be up to five words long, it would take more than half a year for the processes for "without CLIR" to complete, unless with careful implementation.

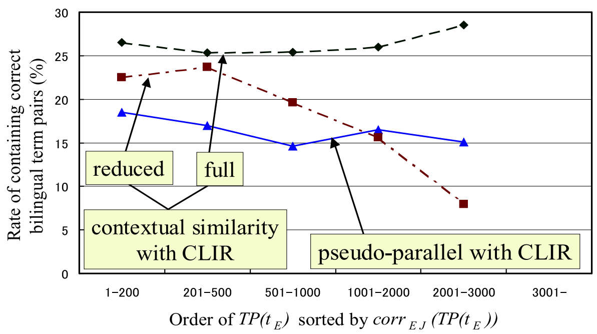


Figure 3: Rates of Containing Correct Bilingual Term Pairs (Site A, $L_d = 0.4$)

as in the formula (2), while, in "full", the candidates are every pair of an English term found in RC_E and a Japanese term found in RC_J.

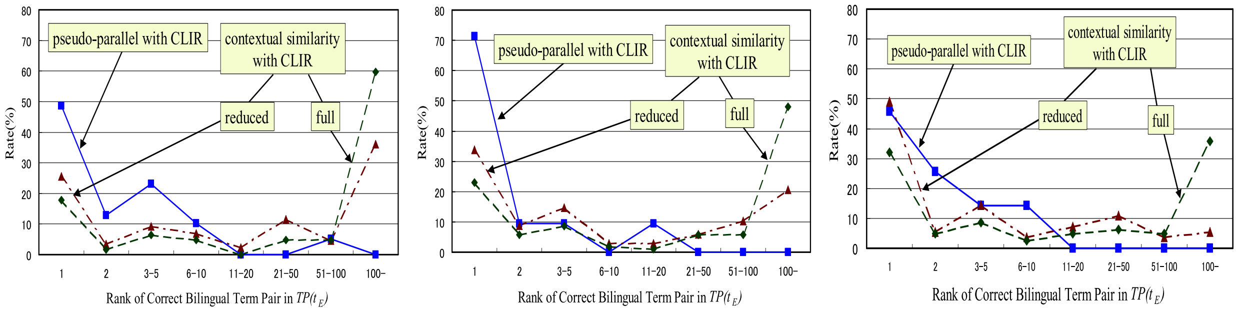
For both "pseudo-parallel with CLIR" and "contextual similarity with CLIR: reduced", the number of bilingual term pairs found in the existing bilingual lexicon corresponds to the one in the column with "reduced" in Table 3 (i.e., 543), while, for "contextual similarity with CLIR: full", that number corresponds to the one in the column with "full" in Table 3 (i.e., 1467). The differences of the rates in Figure 3 correspond to the difference of these numbers (i.e., 1467 and 543). However, it is very important to note that, for both "pseudo-parallel with CLIR" and "contextual similarity with CLIR: reduced", the rate of containing correct bilingual term pairs tends to decrease as the order of $TP(t_E)$ sorted by $corr_{EJ}(TP(t_E))$ becomes lower. This tendency indicates that the estimated values of bilingual term correspondences have positive correlations with the correctness of bilingual term pairs, which supports the usefulness of the estimated bilingual term correspondences. For "contextual similarity with CLIR: full", on the other hand, the rate of containing correct bilingual term pairs seems to be constant and thus the estimated values of bilingual term correspondences do not seem useful. This result again supports our claim on the usefulness of cross-language retrieval of relevant texts in acquisition of bilingual term correspondences.

4.3.3 Ranks of Correct Bilingual Term Pairs

Finally, we evaluate the rank of correct bilingual term correspondences within each set $TP(t_E)$, sorted by the estimated bilingual term correspondence $corr_{EJ}(t_E, t_J)$. Within a set $TP(t_E)$, estimated Japanese term translation t_J are sorted by $corr_{EJ}(t_E, t_J)$, and the ranks of correct Japanese translation of t_E are recorded. For the site A with the similarity lower bounds $L_d = 0.3, 0.4, 0.5$, Figure 4 shows this distribution for the correct bilingual term pairs, which are contained in the topmost 200 $TP(t_E)$ and are found in the existing bilingual lexicon (we have similar results for other sites B and C). Here, we compare this distribution among "pseudo-parallel with CLIR", "contextual similarity with CLIR: reduced", and "contextual similarity with CLIR: full".

For all the similarity lower bounds L_d, "pseudo-parallel with CLIR" performs best, where about 85∼90% of correct bilingual term pairs are included within the 5-best candidates in each $TP(t_E)$, and about 90∼100% are included within the 10-best. Here, it is important to note that bilingual term correspondence estimation by "pseudo-parallel with CLIR" has another advantage over that by "contextual similarity with CLIR: reduced/full" in terms of computational complexity. Also note that the performance of "pseudo-parallel with CLIR" is affected little by the similarity lower bounds L_d. On the other hand, for "contextual similarity with CLIR: reduced/full", the performance becomes worse as the similarity lower bound L_d becomes smaller and the cross-lingually relevant English/Japanese corpus RC_E and RC_J becomes noisier. More specifically, for "full", as the similarity lower bound L_d becomes smaller, more and more correct bilingual term pairs become outside of the 100-best candidates[7]. For "reduced", the rate of correct bilingual term pairs included within the 5-best candidates decreases from 70 to 40%, and that within the 10-best decreases from 73 to 45%, as the similarity lower bound L_d becomes smaller. Furthermore, "reduced" outperforms "full" and their performance gap seems to become larger as the similarity lower bound L_d becomes larger. To summarize those results, candidate reduction with the help of a pseudo-parallel sentence pair is quite effective also in the precise estimation of bilingual

[7] We manually examined all of those bilingual term pairs that are judged as *"correct"* against the existing bilingual lexicon. We confirmed that most of those outside of the 100-best candidates are not translation of each other in the cross-lingually relevant text pairs.



(a) $L_d = 0.3$

(b) $L_d = 0.4$

(c) $L_d = 0.5$

Figure 4: Ranks of Correct Bilingual Term Pairs within a $TP(t_E)$ (Site A, topmost 200 $TP(t_E)$)

term correspondences. This result again clearly supports our claim on the usefulness of cross-language retrieval of relevant texts in acquisition of bilingual term correspondences.

5 Related Works

As we showed in section 4.3.1, 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. Tanaka (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. In this paper, on the other hand, we showed in section 4.3.1 that, due to its computational complexity, it is difficult to straightforwardly apply previously studied techniques of bilingual term correspondence estimation from comparable corpora, especially in the case of large scale evaluation such as those presented in this paper. Then, we showed that this computational difficulty can be easily avoided with the help of cross-language retrieval of relevant texts without harming the performance of precisely estimating bilingual term correspondences.

6 Conclusion

Within the framework of translation knowledge acquisition from WWW news sites, we studied issues on the effect of cross-language retrieval of relevant texts in bilingual lexicon acquisition from comparable corpora. We showed that it is quite effective to reduce the candidate bilingual term pairs against which bilingual term correspondences are estimated, in terms of both computational complexity and the performance of precise estimation of bilingual term correspondences.

References

N. Collier et al. 1998. Machine translation vs. dictionary term translation — a comparison for English-Japanese news article alignment. In *Proc. 17th COLING and 36th ACL*, pages 263–267.

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.

P. Fung. 1995. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proc. 3rd WVLC*, pages 173–183.

H. Kaji and T. Aizono. 1996. Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *Proc. 16th COLING*, pages 23–28.

Y. Matsumoto and T. Utsuro. 2000. Lexical knowledge acquisition. In R. Dale, H. Moisl, and H. Somers, editors, Handbook of Natural Language Processing, chapter 24, pages 563–610. Marcel Dekker Inc.

R. Rapp. 1995. Identifying word translations in non-parallel texts. In *Proc. 33rd ACL*, pages 320–322.

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

K. Tanaka and H. Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora. In *Proc. 16th COLING*, pages 580–585.

T. Tanaka. 2002. Measuring the similarity between compound nouns in different languages using non-parallel corpora. In *Proc. 19th COLING*, pages 981–987.

T. Utsuro et al. 2002. Semi-automatic compilation of bilingual lexicon entries from cross-lingually relevant news articles on WWW news sites. In *Machine Translation: From Research to Real Users*, Lecture Notes in Artificial Intelligence: Vol. 2499, pages 165–176. Springer.