Translation Knowledge Acquisition from Cross-Lingually Relevant News Articles

Takehito Utsuro
Department of Information and Computer Sciences,
Toyohashi University of Technology
Tenpaku-cho, Toyohashi 441-8580, Japan
utsuro@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. After briefly reviewing previous works on translation knowledge acquisition from both parallel and non-parallel corpora, we discuss major advantages of taking an approach of translation knowledge acquisition from cross-lingually relevant article pairs automatically collected by CLIR techniques. Then, as a case study, we present results of applying standard co-occurrence frequency based techniques of estimating bilingual term correspondences 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 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 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 [Utsuro02]. 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/non-parallel corpora are integrated in an optimal fashion.

The rest of the paper is organized as follows. First, we give a brief review of previous works on translation knowledge acquisition from both parallel and non-parallel corpora. Based on the discussions of the review, we examine major advantages of taking an approach of translation knowledge acquisition from cross-lingually relevant article pairs automatically collected by CLIR techniques. Then, as a case study of translation knowledge acquisition in our framework, we further study issues regarding cross-language retrieval and collection of *"identical"*/ *"relevant"* article pairs. We also present results of applying standard co-occurrence frequency based techniques of estimating bilingual term correspondences from parallel corpora to those automatically collected *"identical"*/ *"relevant"* article

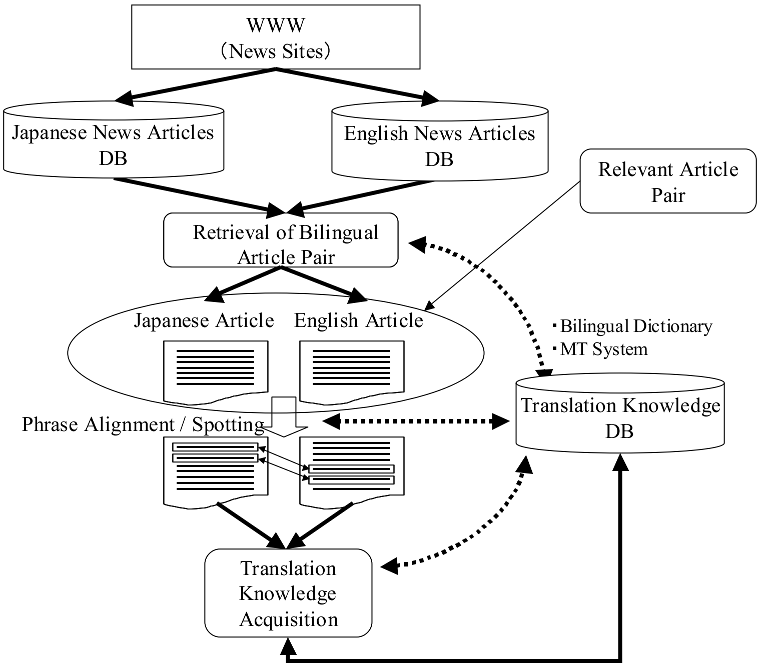


Figure 1: Translation Knowledge Acquisition from WWW News Sites: Overview

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 Translation Knowledge Acquisition from Parallel/Non-parallel Corpora: A Brief Review

In general, research issues and techniques employed in previous works on translation knowledge acquisition from parallel/non-parallel corpora vary according to the source of translation knowledge acquisition: i.e., *parallel* or *non-parallel* corpora. The followings briefly mention research issues for each approach and review several previous works.

2.1 Translation Knowledge Acquisition from Parallel Corpora

Parallel corpora are relatively clean knowledge source for acquiring translation knowledge such as term correspondences and translation rules, although they are harder to obtain than non-parallel corpora. When (relatively) noise free parallel corpora are available, it is somehow easier to find sentence-level/phrase-level/word-level correspondences of parallel texts across languages. Research issues related to translation knowledge acquisition from parallel corpora can be roughly categorized into the following two types: i) aligning sentences/phrases/words within parallel text, and, ii) acquisition of translation knowledge from sentence-aligned parallel text.

2.1.1 Aligning Sentences/Phrases/Words within Parallel Text

Some of earlier works on parallel text alignment studied techniques for aligning parallel text at sentence level. For instance, [Gale93, Chen93, Utsuro94] studied sentence alignment techniques based on dynamic programming, using sentence length and lexical mapping information across languages. [Kay93, Haruno96b] applied iterative refinement algorithms to sentence level alignment tasks. Or, assuming parallel text being aligned at sentence level, some of other works concentrated on chunk-level alignment of parallel sentences (e.g., [Le00]), or syntactic structure level alignment of parallel sentences (e.g., [Utsuro92, Kaji92, Matsumoto93, Wu97]). Another stream of efforts toward finding sentence-level/phrase-level/word-level correspondences of parallel texts is along the line of statistical machine translation models, which were initially proposed by [Brown90, Brown93] and, more recently, have been intensively studied by several research groups (e.g., [Germann01, Och02]). Some other researchers also invented word-level alignment techniques for noisy parallel texts (e.g. [Melamed97]). Detailed introductory descriptions regarding issues mentioned in this section can be found in [Manning99, Wu00].

2.1.2 Translation Knowledge Acquisition

One of well studied techniques of learning translation knowledge such as translation probabilities from parallel text is the one based on statistical machine translation models [Brown90, Brown93]. Another well studied techniques of estimating correspondences of words and compound terms across two languages belong to those based on the contingency table of co-occurrence frequencies across two languages. Let t_X and t_Y denote (possibly compound) terms of the language X and the language Y, respectively. Then, consulting the contingency table of co-occurrence frequencies of t_X and t_Y below, bilingual term correspondences can be estimated according to the statistical measures such as the mutual information, the ϕ^2 statistic, the dice coefficient, the log-likelihood ratio, and also certain types of their extensions (e.g, [Gale91, Kumano94, Haruno96a, Smadja96, Kitamura96, Melamed00]).

	t_Y	$\neg t_Y$
t_X	$freq(t_X, t_Y)$	$freq(t_X, \neg t_Y)$
$\neg t_X$	$freq(\neg t_X, t_Y)$	$freq(\neg t_X, \neg t_Y)$

[Matsumoto97] also proposed a method for acquiring translation rules of machine translation systems from the results of syntactic structure level alignment [Matsumoto93] of parallel sentences. Detailed introductory descriptions regarding issues mentioned in this section can be found in [Matsumoto00].

2.2 Translation Knowledge Acquisition from Non-parallel Corpora

Although parallel corpora are relatively clean knowledge source and it is somehow easier to acquire translation knowledge from them, one limitation of the approach of acquiring translation knowledge from parallel corpora is that it heavily relies on availability of parallel corpora. Since it is very expensive to manually collect parallel corpora, clean parallel corpora are less available compared with noisier parallel corpora or non-parallel corpora of related contents. Therefore, considering this resource scarcity bottleneck in translation knowledge acquisition from parallel corpora, research activities for acquiring translation knowledge from non-parallel corpora are also very popular.

Translation knowledge acquisition from non-parallel corpora is relatively harder compared with that from parallel corpora. Although most of the previous works on translation knowledge acquisition from non-parallel corpora considered acquisition from related non-parallel corpora (i.e., comparable corpora), some of them studied acquisition from unrelated bilingual corpora. Research issues related to translation knowledge acquisition from non-parallel corpora can be roughly categorized into the following two types: i) finding parallel document pairs from the collection of related bilingual documents set, and, ii) acquisition of translation knowledge such as (possibly compound) bilingual term correspondences from related/unrelated non-parallel bilingual corpora.

2.2.1 Finding Parallel Document Pairs

One of well studied techniques of finding parallel document pairs is based on measuring document similarities across languages by considering cross-language similarities of words, that are obtained by employing cross-language information retrieval models or exploiting existing machine translation systems/bilingual lexicons. [Takahashi97, Xu99] studied to exploit anchor expressions such as numerical expressions and names. [Collier98] compared the performance of finding parallel document pairs between measures of document similarity based on machine translation systems and those based on bilingual lexicons. [Matsumoto02] studied to exploit machine translation systems and company name bilingual lexicon. In the context of cross-language information retrieval (CLIR) research, [Masuichi00] proposed to apply bootstrapping technique to an existing corpus-based CLIR approach for the task of extracting bilingual text pairs. [Hasan01] proposed a method for aligning Chinese and Japanese documents where an MT software as well as several statistical measures are employed for document similarity calculation. Some of them evaluated their techniques of finding parallel document pairs against existing bilingual newspaper articles [Takahashi97, Collier98, Matsumoto02], while others evaluated against bilingual newspaper articles available on WWW news sites [Xu99, Hasan01].

Another type of previous works on finding parallel document pairs is an approach of collecting parallel document URLs from WWW by examining clues in the URLs and the structures of the HTML source texts (e.g., [Resnik99, Nie99]).

2.2.2 Translation Knowledge Acquisition

Previously studied techniques of estimating bilingual term correspondences from non-parallel corpora do not rely on the process of finding parallel document pairs. They are mostly based on the idea that semantically similar words appear in similar contexts [Kaji96, Tanaka96, Fung98, Rapp99, Chiao02, Tanaka02]. In those techniques, frequency information of contextual words co-occurring in the monolingual text is stored and their similarity is measured across languages. In the modeling of contextual similarities across languages, earlier works such as [Fung95, Rapp95, Tanaka96] 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 [Kaji96, Fung98, Rapp99, Cao02, Chiao02, Tanaka02] studied to exploit existing bilingual lexicons as initial seed for obtaining translation candidates of constituent words of compound terms and/or for modeling of contextual similarities across languages. As for some of the earlier works cited above, detailed introductory descriptions can be found in [Matsumoto00].

In the context of resolving target translation ambiguities or estimating word translation probabilities, [Dagan94, Koehn00, Nakagawa01] proposed first to obtain translation candidates by consulting initial seed bilingual lexicon and then to resolve target translation ambiguities or to estimate word translation probabilities based on certain types of statistics measured against monolingual target corpora.

[Nagata01] proposed another technique of acquiring bilingual term correspondences by collecting partially bilingual texts from WWW with Internet search engines. [Cao02] also studied to filter out inappropriate translation candidates by exploiting Internet search engines.

3 A Case Study: Semi-automatic Compilation of Bilingual Lexicon Entries from Cross-Lingually Relevant News Articles on WWW News Sites

3.1 Motivation

This section discusses our major motivations of taking an approach of translation knowledge acquisition from cross-lingually relevant article pairs automatically collected by CLIR techniques.

First of all, as can be seen from the discussions of section 2.2, research efforts of finding parallel document pairs and those of translation knowledge acquisition from non-parallel corpora are activities that are not related to each other so far, and, to the best of our knowledge, there have existed no research activities which aim at integrating the two techniques in the context of translation knowledge acquisition. One of the major advantages of our approach is that it becomes possible to improve the performance of translation knowledge acquisition by integrating CLIR techniques and translation knowledge acquisition techniques in an optimal fashion. It is expected that the translation knowledge acquisition process benefits from the results of restricting relevant article pairs by the CLIR techniques. It is obvious that it is inevitable to restrict relevant article pairs by the CLIR techniques when applying techniques for acquiring translation knowledge from parallel corpora. However, even when applying techniques for acquiring translation knowledge from non-parallel corpora, it might happen that the translation knowledge acquisition process again benefits from the CLIR techniques by restricting the articles from which contextual vectors are extracted. Furthermore, in the opposite direction, it is also expected that the CLIR process benefits from the results of translation knowledge acquisition.

Compared with our approach of employing bilingual news articles on WWW news sites as a source for translation knowledge acquisition, the techniques for finding parallel document pairs from general WWW texts [Resnik99, Nie99] and those for acquiring translation knowledge from partially bilingual texts on the WWW [Nagata01] or based on collocations in the monolingual texts on the WWW [Cao02] have one advantage: i.e., 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 is that high translation quality is guaranteed and articles of up-to-date topics are updated everyday.

Another related work is that for semi-automatic acquisition of translation knowledge from parallel corpora [Dagan97]. Compared with this work, we aim at developing user interface facilities for efficient semi-automatic acquisition of translation knowledge from one kind of non-parallel corpora, i.e., bilingual news articles on WWW news sites. Our approach suffers from less resource scarcity bottleneck because it is far easier to collect knowledge source for acquiring translation knowledge.

The results of the case study reported in the rest of this paper are preliminary in that full range of techniques applicable in our framework have not been examined for the moment, both for the CLIR process and for the translation knowledge acquisition process. When calculating similarities of articles across languages, so far we have just finished evaluating similarities based on existing machine translation systems, but not those based on anchor expressions such as numerical expressions and those based on existing bilingual lexicons. As for the translation knowledge acquisition process, so far we have just finished evaluating techniques for acquisition from parallel corpora, but not those for acquisition from non-parallel corpora. Furthermore, as the techniques for translation knowledge acquisition from parallel corpora, we have just finished evaluating those for acquiring word level translation correspondences, but not those for acquiring compound term level translation correspondences. We are now intensively working those extensions, and their results will be reported in the very near future.

3.2 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.

3.3 Acquisition of Bilingual Term Correspondences from Relevant News Articles

3.3.1 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 3.4.1, 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 an English query article d_E^i, let D_J^i denote the set of Japanese articles with cosine similarities higher than or equal to a certain lower bound L_d:

$$D_J^i \quad = \quad \Big\{ d_J \mid \cos(d_E^i, d_J) \geq L_d \Big\}$$

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

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

Next, we apply standard techniques of estimating bilingual term correspondences from parallel corpora [Matsumoto00] 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 [Matsumoto00].

	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 3.4.3, we show results with the ϕ^2 statistic:

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

[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.

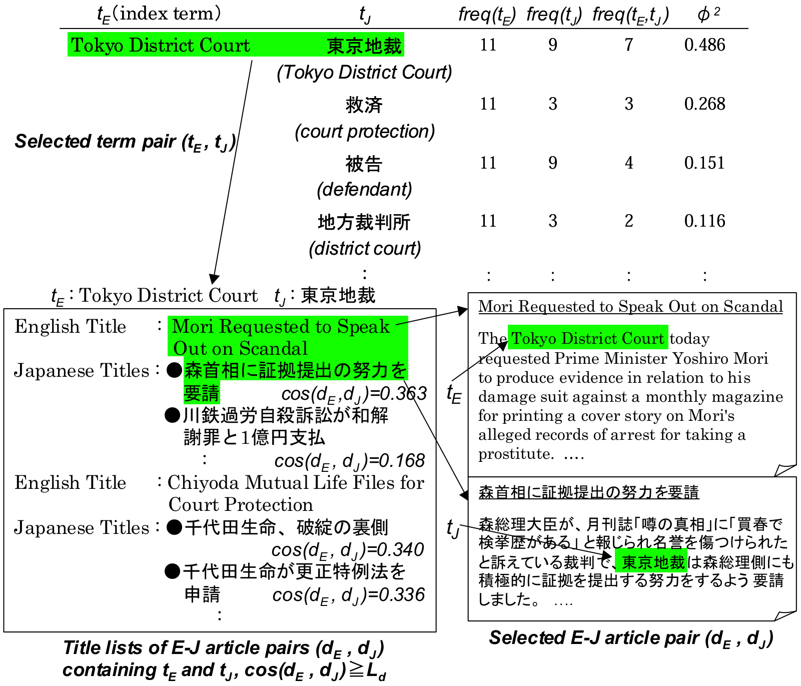

t_E(index term)
t_J
$freq(t_E)$
$freq(t_J)$
$freq(t_E,t_J)$
ϕ^2
Tokyo District Court
東京地裁
(Tokyo District Court)
11
9
7
0.486
救済
(court protection)
11
3
3
0.268
被告
(defendant)
11
9
4
0.151
地方裁判所
(district court)
11
3
2
0.116
Selected term pair (t_E, t_J)
t_E : Tokyo District Court t_J : 東京地裁
English Title : Mori Requested to Speak Out on Scandal
Japanese Titles : ●森首相に証拠提出の努力を要請 $cos(d_E,d_J)$=0.363
●川鉄過労自殺訴訟が和解謝罪と1億円支払
$cos(d_E, d_J)$=0.168
English Title : Chiyoda Mutual Life Files for Court Protection
Japanese Titles : ●千代田生命、破綻の裏側 $cos(d_E, d_J)$=0.340
●千代田生命が更正特例法を申請 $cos(d_E, d_J)$=0.336
Title lists of E-J article pairs (d_E , d_J) containing t_E and t_J, $cos(d_E , d_J) \geqq L_d$
Mori Requested to Speak Out on Scandal
The Tokyo District Court today requested Prime Minister Yoshiro Mori to produce evidence in relation to his damage suit against a monthly magazine for printing a cover story on Mori's alleged records of arrest for taking a prostitute. ....
t_E
森首相に証拠提出の努力を要請
森総理大臣が、月刊誌「噂の真相」に「買春で検挙歴がある」と報じられ名誉を傷つけられたと訴えている裁判で、東京地裁は森総理側にも積極的に証拠を提出する努力をするよう 要請しました。 ....
t_J
Selected E-J article pair (d_E , d_J)

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

3.3.2 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:

$$\begin{aligned} TP(t_E) \quad = \quad & \Big\{ \langle t'_E, t_J \rangle \mid t_E \succeq t'_E, freq(t_E) \geq L^E_f, freq(t_J) \geq L^J_f, \\ & freq(t_E, t_J) \geq L^{EJ}_f, length(t_E) \leq U^E_l, length(t_J) \leq U^J_l \Big\} \end{aligned}$$

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

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

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

Then, each set $TP(t^i_E)$ 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:

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

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

	Total # of Days		Total # of Articles		Average # of Articles per Day		Average Article Size (bytes)		# of Reference Article Pairs for CLIR Evaluation	
Site	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

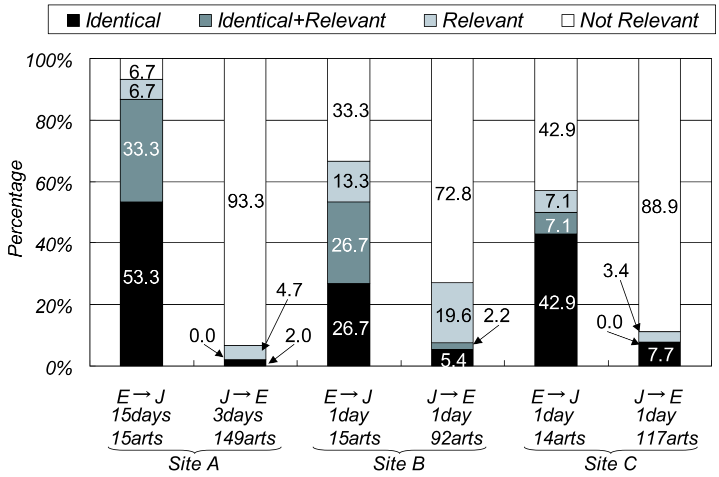


Figure 3: Availability of Cross-Lingually *"Identical"*/ *"Relevant"* Articles

3.3.3 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 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.

3.4 Experimental Evaluation

3.4.1 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. 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

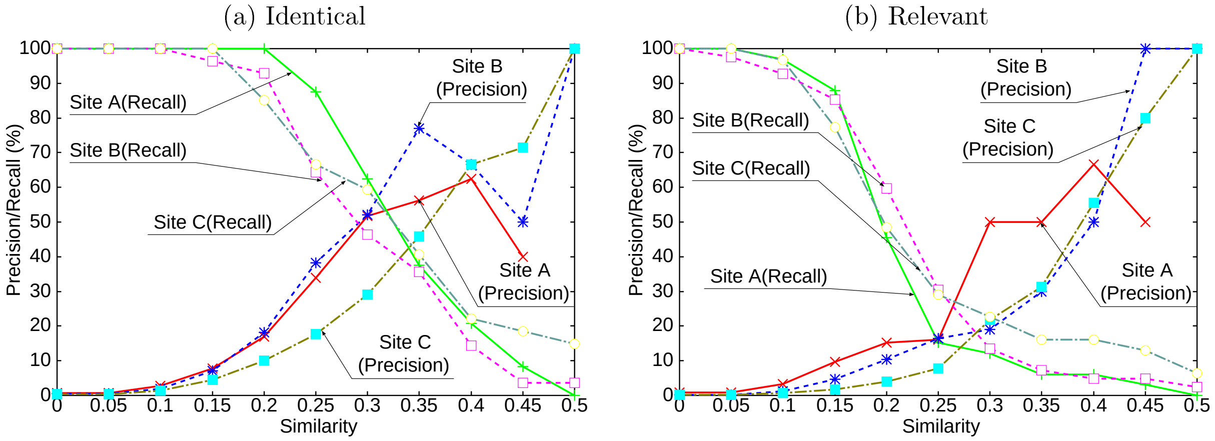


Figure 4: Precision/Recall of Cross-Language Retrieval of Relevant News Articles (Article Similarity ≥ L_d)

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

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.

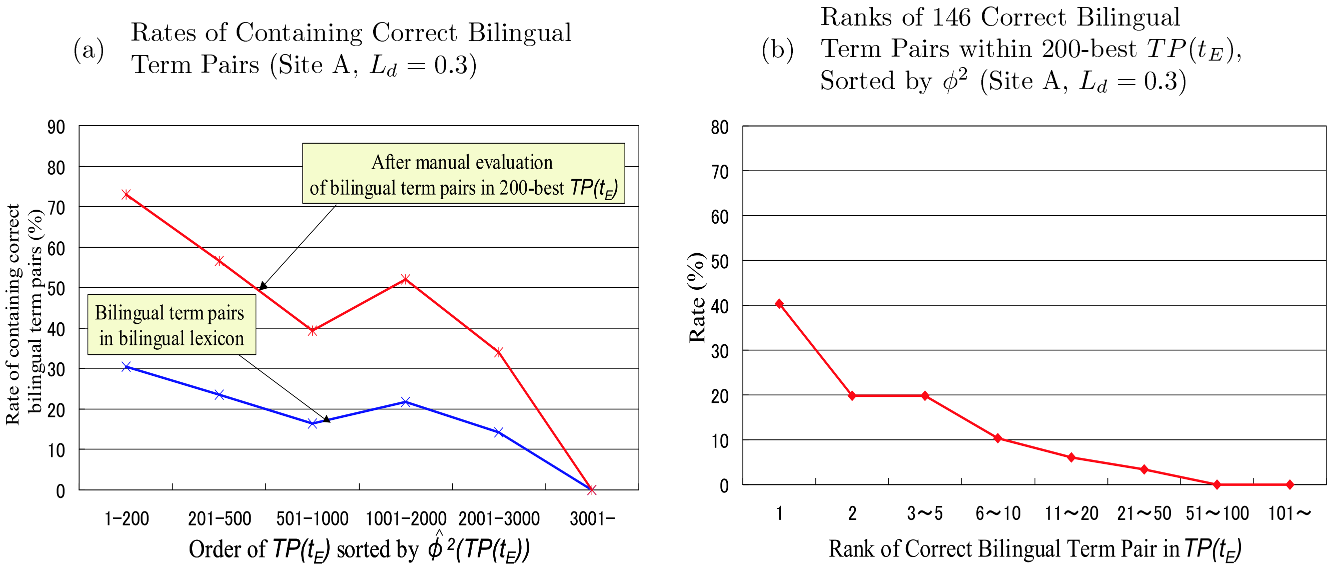
3.4.2 Cross-Language Retrieval of Relevant News Articles

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 *"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 ± 5 days for the *"identical"* article pairs and around ± 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:

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

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

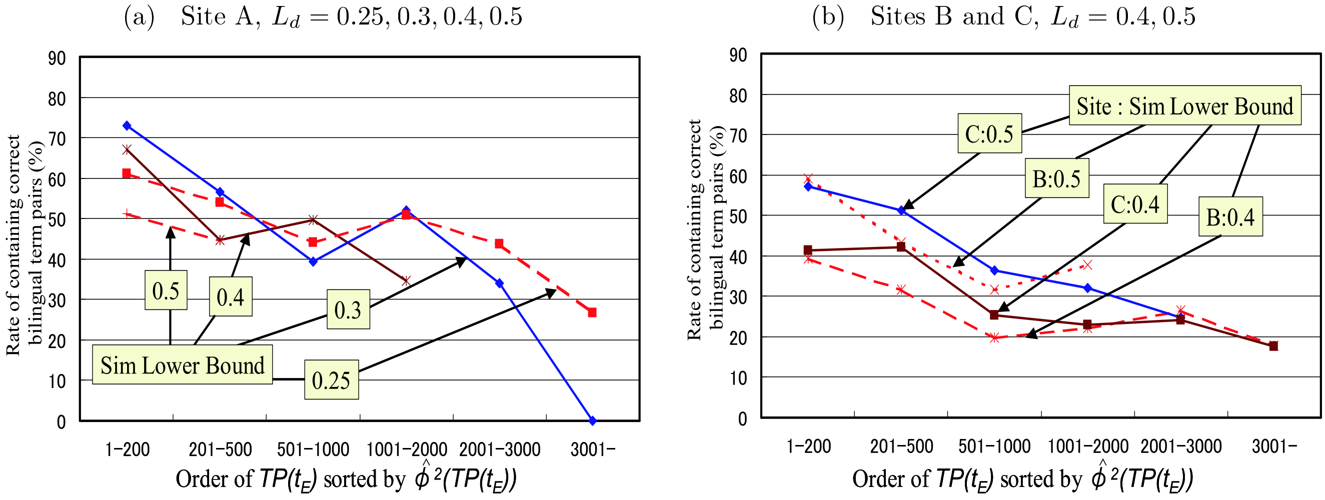
[4] We are now working on examining usefulness of additional clues such as titles, pronunciation of foreign names, and numerical expressions, and furthermore on incorporating them for the purpose of improving the performance of cross-language retrieval.



(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, $L_d = 0.3$)

Figure 5: Evaluation Results using Bilingual Term Pairs in a Bilingual Lexicon / by Manual Evaluation



(a) Site A, $L_d = 0.25, 0.3, 0.4, 0.5$

(b) Sites B and C, $L_d = 0.4, 0.5$

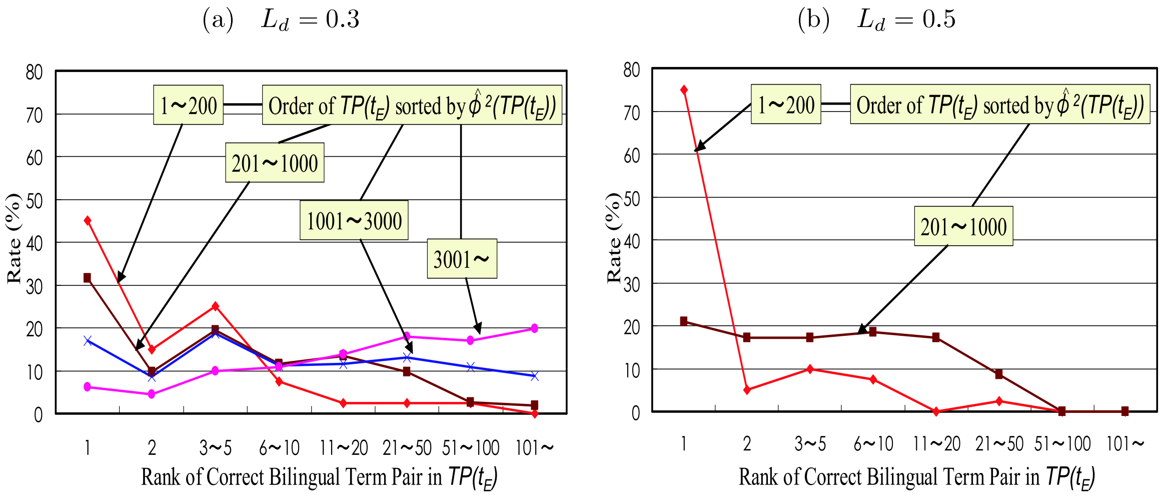
Figure 6: Rates of Containing Correct Bilingual Term Pairs

3.4.3 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 $L_f^E = L_f^J = 3, L_f^{EJ} = 2, U_l^E = U_l^J = 5$ (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 3.3.2) 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

[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.


(a) $L_d = 0.3$

(b) $L_d = 0.5$

Figure 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)

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 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. 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.

[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 $L_d \leq 0.4$.

4 Conclusion

For the purpose of overcoming resource scarcity bottleneck in corpus-based translation knowledge acquisition research, this paper took an approach of semi-automatically acquiring domain specific translation knowledge from the collection of bilingual news articles on WWW news sites. After briefly reviewing previous works on translation knowledge acquisition from both parallel and non-parallel corpora, we discussed major advantages of taking an approach of translation knowledge acquisition from cross-lingually relevant article pairs automatically collected by CLIR techniques. Then, as a case study, we presented results of applying standard co-occurrence frequency based techniques of estimating bilingual term correspondences 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

[Brown90] Brown, P. F., Cocke, J., Pietra, S. A., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L. and Roosin, P. S.: A Statistical Approach to Machine Translation, *Computational Linguistics*, Vol. 16, No. 2, pp. 79–85 (1990).

[Brown93] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J. and Mercer, R. L.: The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, Vol. 19, No. 2, pp. 263–311 (1993).

[Cao02] Cao, Y. and Li, H.: Base Noun Phrase Translation Using Web Data and the EM Algorithm, *Proc. 19th COLING*, pp. 127–133 (2002).

[Chen93] Chen, S. F.: Aligning Sentences in Bilingual Corpora Using Lexical Information, *Proc. 31st ACL*, pp. 9–16 (1993).

[Chiao02] Chiao, Y.-C. and Zweigenbaum, P.: Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora, *Proc. 19th COLING*, pp. 1208–1212 (2002).

[Collier98] Collier, N., Hirakawa, H. and Kumano, A.: Machine Translation vs. Dictionary Term Translation — A Comparison for English-Japanese News Article Alignment, *Proc. 17th COLING and 36th ACL*, pp. 263–267 (1998).

[Dagan94] Dagan, I. and Itai, A.: Word Sense Disambiguation Using a Second Language Monolingual Corpus, *Computational Linguistics*, Vol. 20, No. 4, pp. 563–596 (1994).

[Dagan97] Dagan, I. and Church, K.: *Termight*: Coordinating Humans and Machines in Bilingual Terminology Acquisition, *Machine Translation*, Vol. 12, No. 1/2, pp. 89–107 (1997).

[Fung95] Fung, P.: Compiling Bilingual Lexicon Entries From a Non-Parallel English-Chinese Corpus, *Proc. 3rd WVLC*, pp. 173–183 (1995).

[Fung98] Fung, P. and Yee, L. Y.: An IR Approach for Translating New Words from Nonparallel, Comparable Texts, *Proc. 17th COLING and 36th ACL*, pp. 414–420 (1998).

[Gale91] Gale, W. and Church, K.: Identifying Word Correspondences in Parallel Texts, *Proc. 4th DARPA Speech and Natural Language Workshop*, pp. 152–157 (1991).

[Gale93] Gale, W. A. and Church, K. W.: A Program for Aligning Sentences in Bilingual Corpora, *Computational Linguistics*, Vol. 19, No. 1, pp. 75–102 (1993).

[Germann01] Germann, U., M.Jahr, , Knight, K., Marcu, D. and Yamada, K.: Fast Decoding and Optimal Decoding for Machine Translation, *Proc. 39th ACL*, pp. 228–235 (2001).

[Haruno96a] Haruno, M., Ikehara, S. and Yamazaki, T.: Learning Bilingual Collocations by Word-Level Sorting, *Proc. 16th COLING*, pp. 525–530 (1996).

[Haruno96b] Haruno, M. and Yamazaki, T.: High-performance Bilingual Text Alignment using Statistical and Dictionary Information, *Proc. 34th ACL*, pp. 131–138 (1996).

[Hasan01] Hasan, M. M. and Matsumoto, Y.: Multilingual Document Alignment — A Study with Chinese and Japanese, *Proc. 6th NLPRS*, pp. 617–623 (2001).

[Kaji92] Kaji, H., Kida, Y. and Morimoto, Y.: Learning Translation Templates from Bilingual Text, *Proc. 14th COLING*, pp. 672–678 (1992).

[Kaji96] Kaji, H. and Aizono, T.: Extracting Word Correspondences from Bilingual Corpora Based on Word Co-occurrence Information, *Proc. 16th COLING*, pp. 23–28 (1996).

[Kay93] Kay, M. and Röscheisen, M.: Text-Translation Alignment, *Computational Linguistics*, Vol. 19, No. 1, pp. 121–142 (1993).

[Kitamura96] Kitamura, M. and Matsumoto, Y.: Automatic Extraction of Word Sequence Correspondences in Parallel Corpora, *Proc. 4th WVLC*, pp. 79–87 (1996).

[Koehn00] Koehn, P. and Knight, K.: Estimating Word Translation Probabilities from Unrelated Monolingual Corpora Using the EM Algorithm, *Proceedings of the 17th AAAI*, pp. 711–715 (2000).

[Kumano94] Kumano, A. and Hirakawa, H.: Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistical Information, *Proc. 15th COLING*, pp. 76–81 (1994).

[Le00] Le, S., Youbing, J., Lin, D. and Yufang, S.: Word Alignment of English-Chinese Bilingual Corpus Based on Chunks, *Proc. 2000 EMNLP and VLC*, pp. 110–116 (2000).

[Manning99] Manning, C. D. and Schütze, H.: Statistical Alignment and Machine Translation, *Foundations of Statistical Natural Language Processing*, chapter 13, pp. 463–494, The MIT Press (1999).

[Masuichi00] Masuichi, H., Flournoy, R., Kaufmann, S. and Peters, S.: A Bootstrapping Method for Extracting Bilingual Text Pairs, *Proc. 18th COLING*, pp. 1066–1070 (2000).

[Matsumoto93] Matsumoto, Y., Ishimoto, H. and Utsuro, T.: Structural Matching of Parallel Texts, *Proc. 31st ACL*, pp. 23 – 30 (1993).

[Matsumoto97] Matsumoto, Y. and Kitamura, M.: Acquisition of Translation Rules from Parallel Corpora, Mitkov, R. and Nicolov, N. (eds.), *Recent Advances in Natural Language Processing: Selected Papers from RANLP'95*, pp. 405–416, John Benjamins (1997).

[Matsumoto00] Matsumoto, Y. and Utsuro, T.: Lexical Knowledge Acquisition, Dale, R., Moisl, H. and Somers, H. (eds.), *Handbook of Natural Language Processing*, chapter 24, pp. 563–610, Marcel Dekker Inc. (2000).

[Matsumoto02] Matsumoto, K. and Tanaka, H.: Automatic Alignment of Japanese and English Newspaper Articles using an MT System and a Bilingual Company Name Dictionary, *Proc. 3rd LREC*, Vol. 2, pp. 480–484 (2002).

[Melamed97] Melamed, I. D.: A Portable Algorithm for Mapping Bitext Correspondences, *Proc. 35th ACL and 8th EACL*, pp. 305–312 (1997).

[Melamed00] Melamed, I. D.: Models of Translational Equivalence among Words, *Computational Linguistics*, Vol. 26, No. 2, pp. 221–249 (2000).

[Nagata01] Nagata, M., Saito, T. and Suzuki, K.: Using the Web as a Bilingual Dictionary, *Proc. Workshop on Data-driven Methods in Machine Translation*, pp. 95–102 (2001).

[Nakagawa01] Nakagawa, H.: Disambiguation of Single Noun Translations Extracted from Bilingual Comparable Corpora, *Terminology*, Vol. 7, No. 1, pp. 63–83 (2001).

[Nie99] Nie, J.-Y., et al.: Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web, *Proc. 22nd SIGIR*, pp. 74–81 (1999).

[Och02] Och, F. J. and Ney, H.: Discriminative Training and Maximum Entropy Models for Statistical Machine Translation, *Proc. 40th ACL*, pp. 295–302 (2002).

[Rapp95] Rapp, R.: Identifying Word Translations in Non-Parallel Texts, *Proc. 33rd ACL*, pp. 320–322 (1995).

[Rapp99] Rapp, R.: Automatic Identification of Word Translations from Unrelated English and German Corpora, *Proc. 37th ACL*, pp. 519–526 (1999).

[Resnik99] Resnik, P.: Mining the Web for Bilingual Text, *Proc. 37th ACL*, pp. 527–534 (1999).

[Smadja96] Smadja, F., McKeown, K. R. and Hatzivassiloglou, V.: Translating Collocations for Bilingual Lexicons: A Statistical Approach, *Computational Linguistics*, Vol. 22, No. 1, pp. 1–38 (1996).

[Takahashi97] Takahashi, Y., Shirai, S. and Bond, F.: A Method of Automatically Aligning Japanese and English Newspaper Articles, *Proc. 4th NLPRS*, pp. 657–660 (1997).

[Tanaka96] Tanaka, K. and Iwasaki, H.: Extraction of Lexical Translations from Non-Aligned Corpora, *Proc. 16th COLING*, pp. 580–585 (1996).

[Tanaka02] Tanaka, T.: Measuring the Similarity between Compound Nouns in Different Languages Using Non-Parallel Corpora, *Proc. 19th COLING*, pp. 981–987 (2002).

[Utsuro92] Utsuro, T., Matsumoto, Y. and Nagao, M.: Lexical Knowledge Acquisition from Bilingual Corpora, *Proc. 14th COLING*, pp. 581–587 (1992).

[Utsuro94] Utsuro, T., Ikeda, H., Yamane, M., Matsumoto, Y. and Nagao, M.: Bilingual Text Matching using Bilingual Dictionary and Statistics, *Proc. 15th COLING*, pp. 1076–1082 (1994).

[Utsuro02] Utsuro, T., Horiuchi, T., Chiba, Y. and Hamamoto, T.: Semi-automatic Compilation of Bilingual Lexicon Entries from Cross-Lingually Relevant News Articles on WWW News Sites, *Proc. 5th AMTA* (2002).

[Wu97] Wu, D.: Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora, *Computational Linguistics*, Vol. 23, No. 3, pp. 377–403 (1997).

[Wu00] Wu, D.: Alignment, Dale, R., Moisl, H. and Somers, H. (eds.), *Handbook of Natural Language Processing*, chapter 18, pp. 415–458, Marcel Dekker Inc. (2000).

[Xu99] Xu, D. and Tan, C. L.: Alignment and Matching of Bilingual English-Chinese News Texts, *Machine Translation*, Vol. 14, pp. 1–33 (1999).