Adapted Seed Lexicon and Combined Bidirectional Similarity Measures for Translation Equivalent Extraction from Comparable Corpora

Hiroyuki Kaji
Central Research Laboratory, Hitachi, Ltd.
1-280 Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan
kaji@crl.hitachi.co.jp

Abstract

An improved method for extracting translation equivalents from bilingual comparable corpora according to contextual similarity was developed. This method has two main features. First, a seed bilingual lexicon—which is used to bridge contexts in different languages—is adapted to the corpora from which translation equivalents are to be extracted. Second, the contextual similarity is evaluated by using a combination of similarity measures defined in opposite directions. An experiment using *Wall Street Journal* and *Nihon Keizai Shimbun* corpora, together with the EDR bilingual dictionary, demonstrated that the method effectively improves the coverage of a bilingual lexicon; the accuracy of lists of candidate translation equivalents for frequently occurring unknown words was around 30%.

1. Introduction

Wide-coverage bilingual lexicons are essential in machine translation and cross-language information retrieval; therefore, automatic extraction of translation equivalents from corpora has been an important research issue over the last decade. Technologies for extracting translation equivalents from parallel corpora have been established (Gale and Church 1991; Kupiec 1993; Dagan, et al. 1993; Fung 1995; Kitamura and Matsumoto 1996; Melamed 1997). However, the availability of large parallel corpora is extremely limited. Methods for extracting translation equivalents from a pair of weakly comparable corpora, i.e., corpora of the same domain in different languages, are therefore required.

Rapp (1995) demonstrated the possibility of extracting translation equivalents from comparable corpora; the underlying assumption is that a word and its translation occur in similar contexts. Subsequently, several researchers developed a method of evaluating the similarity between contexts of words in different languages with the assistance of a seed bilingual lexicon. However, it has not yet been proved practicable. Kaji and Aizono (1996) demonstrated the effectiveness of the method on pairs consisting of a document and its translation, but not on comparable corpora in general definition. Fung and McKeown (1997) first applied the method to comparable corpora. However, their experiment was done under an impractical setting; namely, candidate translation equivalents were beforehand restricted to a small set of manually selected words. Note that many words other than manually selected ones can have similar contexts as a target word[1]. Fung and Yee (1998) proposed an improved method that takes into account the reliability of seed pairs of translation equivalents, but it was not evaluated quantitatively. Rapp (1999) achieved relatively high extraction precision.

[1] In this paper, "target word" is used to indicate the word whose translation equivalent is to be extracted. It does not indicate a translation equivalent of a word.

However, the evaluation was done for common German words, such as “Brot” (bread) and “Musik” (music), which are already included in ordinary lexicons. It is unlikely that equal precision would be achieved for words not included in a seed bilingual lexicon.

Other methods are of course applicable to comparable corpora. For example, translation equivalents of compound words can be extracted according to the correspondence between their constituent words (Nakagawa 2001). Moreover, when a large number of pairs of comparable documents are available, the frequency of co-occurrence in a pair of comparable documents can be used to extract translation equivalents (Utsuro, et al. 2003). However, only the above-mentioned method based on contextual similarity seems capable of extracting translation equivalents of unrestricted types of words from weakly comparable corpora. The author has therefore improved this method in two ways as described in the following section.

2. Proposed Method

2.1. Outline

The contextual-similarity-based method generally consists of the following steps. First, words in two languages are characterized by context vectors, i.e., weighted vectors consisting of associated words or co-occurring words. Then, the context vectors in one language are translated into the other language by consulting a seed bilingual lexicon, and similarity between the context vectors characterizing different-language words is calculated. Finally, pairs of words with high similarity are selected.

Table 1 lists the top 20 associated words for the English word “GOP” (abbreviation of “Grand Old Party”) and those for the Japanese word “共和党<*KYOUWA-TOU*>,” which means “Republican Party.” It is clear from the table that seven out of the top 20 associated words of “共和党<*KYOUWA-TOU*>” have English translations included in the top 100 associated words of “GOP.” Thus, the context vector characterizing “共和党<*KYOUWA-TOU*>” has a relatively high similarity with that characterizing “GOP.” Accordingly, “共和党<*KYOUWA-TOU*>” is likely to be selected as a translation equivalent for “GOP.”

The proposed method, an overview of which is given in Figure 1, has two novel features: one is to adapt a seed bilingual lexicon to comparable corpora from which translation equivalents are to be extracted, and the other is to combine two similarities that are calculated in opposite directions and normalized.

One of the crucial issues regarding the contextual-similarity-based method is how correctly context vectors are translated. A seed bilingual lexicon usually suggests more than one translation equivalent for each associated word, and it is not trivial to select the appropriate ones from among them. Methods used by the previous works, e.g., weighting translation equivalents according to the order in a manually compiled list of translation equivalents (Fung and Yee 1998) and using the first translation equivalent in a manually compiled list (Rapp 1999), are obviously deficient; therefore, a method for adapting a seed bilingual lexicon to comparable corpora automatically was developed. Under the assumption that relevance of a translation equivalent of an entry word to comparable corpora correlates with how many associated words of the entry word suggest the translation

Table 1: Example lists of associated words for "GOP" and "共和党<*KYOUWA-TOU*>"

#	Top 20 associated words of "GOP" (mutual information)	#	Top 20 associated words of "共和党<*KYOUWA-TOU*>" (mutual information)	Translation equivalents included in top 100 associated words of "GOP" [rank]
1	tax cut (2.91)	1	バージニア<*BAAJINIA*> (8.44)	
2	stopgap (2.90)	2	民主<*MINSHU*> (8.38)	
3	last night (2.70)	3	上院<*JOUIN*> (8.26)	Senate [9]
4	rider (2.58)	4	中間層<*CHUUKAN-SOU*> (8.15)	
5	Sen. (2.45)	5	バージニア州<*BAAJINIA-SHUU*> (8.11)	
6	agenda (2.41)	6	テネシー州<*TENESHII-SHUU*> (8.08)	
7	House (2.35)	7	下院議員<*KAIN-GIIN*> (8.01)	Rep. [11]
8	amendment (2.34)	8	下院<*KAIN*> (8.00)	House [7]
9	Senate (2.33)	9	地滑り<*JISUBERI*> (7.93)	
10	welfare (2.32)	10	上院議員<*JOUIN-GIIN*> (7.83)	Sen. [5]
11	Rep. (2.30)	11	議席<*GISEKI*> (7.82)	seat [52]
12	veto (2.30)	12	ケネディ<*KENEDII*> (7.79)	
13	freshman (2.27)	13	過半数<*KAHANSUU*> (7.76)	majority [51]
14	appropriation (2.27)	14	アリゾナ<*ARIZONA*> (7.73)	
15	nomination (2.13)	15	ミシガン<*MISHIGAN*> (7.70)	
16	budget (2.08)	16	敗北<*HAIBOKU*> (7.67)	defeat [45]
17	vote (2.07)	17	ハイチ<*HAICHI*> (7.57)	
18	White House (1.98)	18	マサチューセッツ州<*MASACHUUSETTSU-SHUU*> (7.52)	
19	voter (1.98)	19	落選<*RAKUSEN*> (7.44)	
20	incumbent (1.97)	20	マサチューセッツ<*MASACHUUSETTSU*> (7.43)	

Note: This example is taken from the experiment described in Subsection 3.2.

equivalent, the developed adaptation method selects translation equivalents relevant to comparable corpora for each entry word.

Another crucial issue is definition of similarity between context vectors. Conventional similarity measures are affected by spurious translation equivalents, i.e., non translation equivalents that occur in the similar contexts as a target word. For example, the target word "GOP" is likely to have high similarity with spurious translation equivalents such as "民主党<*MINSHU-TOU*>" (Democratic party), "議会<*GIKAI*>" (Congress), and "選挙<*SENKYO*>" (election). Conventional similarity measures also encounter a difficulty when translation equivalents for a target word are not in a corpus from which translation equivalents are to be extracted. Note that such cases frequently occur when weakly comparable corpora are used. It is difficult to judge whether a

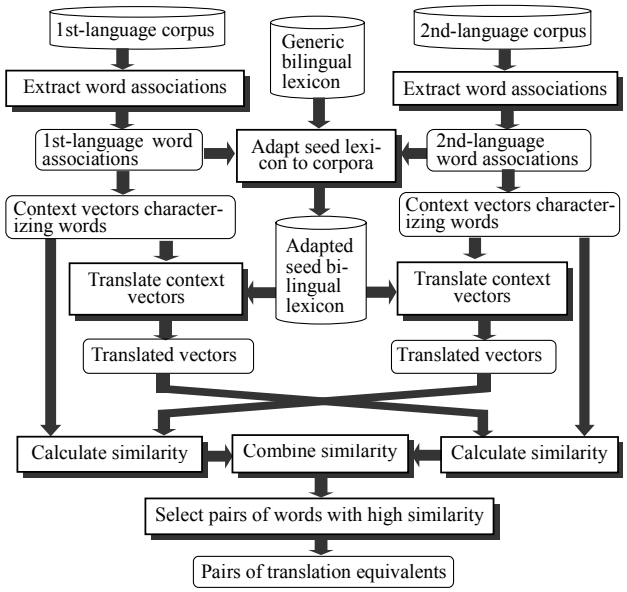

Figure 1: Proposed method for extracting translation equivalents

corpus contains translation equivalents for a target word. To overcome these difficulties, combined use of bidirectional normalized cosine coefficients was devised. It is based on the assumption that a pair of translation equivalents shows high similarity no matter which one is considered as the basis for comparison, while a pair of non translation equivalents does not.

2.2. Adaptation of Seed Bilingual Lexicon to Corpora

2.2.1. Algorithm

For each entry word in a generic bilingual lexicon, corpus-relevant translation equivalents are selected as follows.

(1) Calculate pairwise correlation between translation equivalents and associated words of the entry word.
(2) Assign each associated word to the translation equivalent with the highest correlation.
(3) Calculate the corpus relevancy of each translation equivalent, i.e., the proportion of associated words assigned to each translation equivalent.
(4) Select translation equivalents whose corpus relevancy exceeds a preset threshold.

This bilingual-lexicon adaptation method is based on the assumption that each associated word of an entry word suggests a specific sense of the entry word, in other words, specific translation equivalents of the entry word. The first step of the above-described procedure uses the sense-vs.-clue correlation algorithm originally developed for word-sense disambiguation using bilingual comparable corpora (Kaji and Morimoto 2002). Under the assumption that senses of a word are defined as sets of synonymous translation equivalents, the algorithm calculates a correlation matrix of senses vs. clues (i.e., associated words of the word in question) iteratively. It is used here with a set of translation equivalents instead of a set of senses, resulting in a correlation matrix of translation equivalents vs. associated words. The second step of the procedure may be problematic, since an associated word often suggests two or more translation equivalents that represent the same sense. However, it is difficult to separate translation equivalents suggested by an associated word from others. Each associated word is therefore assigned to the translation equivalent it suggests most strongly. See (Kaji 2004) for the detail of the bilingual-lexicon adaptation method.

2.2.2. Example

Table 2 lists example translation equivalents selected by the developed adaptation method, where

Table 2: Excerpt from EDR bilingual dictionary adapted to *WSJ* and *Nikkei* corpora

Entry word	Translation equivalents selected by proposed method	cf. Translation equivalents in descending order of frequency *)
amendment	修正<*SHUUSEI*>, 改正<*KAISEI*>	改善<*KAIZEN*> (improvement), 変更<*HENKOU*> (change), 改正, 修正, 改定<*KAITEI*> (revision)
Rep.	下院議員<*KAIN-GIIN*>	共和国<*KYOUWA-KOKU*> (Republic), 下院議員
freshman	新顔<*SHINGAO*>, 一年生<*ICHINEN-SEI*>	初心者<*SHOSHIN-SHA*> (beginner), 一年生, 新顔, 新入生<*SHIN'NYUU-SEI*>, フレッシュマン<*FURESSHU-MAN*>
budget	予算案<*YOSAN-AN*>, 予算額<*YOSAN-GAKU*>, 予算<*YOSAN*>	予算, 予算案, 中身<*NAKAMI*> (content), 集まり<*ATSUMARI*> (collection), 財布<*SAIFU*> (purse)
vote	採決<*SAIKETSU*>, 投票<*TOUHYOU*>, 投票権<*TOUHYOU-KEN*>, 決議<*KETSUGI*>	入札<*NYUUSATSU*> (bid), 投票, 決議, 採決, 有権者<*YUUKEN-SHA*> (voter)

*) Underlined translation equivalents seem more appropriate for words given in parentheses than for the entry words.

the EDR bilingual dictionary is used together with *Wall Street Journal* (*WSJ*) and *Nihon Keizai Shimbun* (*Nikkei*) corpora. The entry words are some of the associated words of "GOP" shown in Table 1. Under the threshold for corpus relevancy set to 0.05, the translation equivalents are listed in descending order of corpus relevancy. Lists of up to five translation equivalents in descending order of frequency are also given for comparison. These results clearly demonstrate the necessity and effectiveness of adapting a seed bilingual lexicon. For example, "修正<*SHUUSEI*>" is selected as the first translation equivalent for an entry word "amendment," because many associated words such as "Senate," "vote," and "Republican" have the highest correlation with it. In addition, "改正<*KAISEI*>" is selected as the second translation equivalent for "amendment," because many associated words such as "law," "legislation," and "rule" have the highest correlation with it. These translation equivalents are most appropriate for "amendment," which usually means a written change to a law or document.

2.3. Combination of Bidirectional Normalized Similarity Measures

2.3.1. Similarity measure

In the following, context vectors characterizing first-language word x and second-language word y are denoted as $\boldsymbol{a}(x) = (a_1(x), a_2(x), \ldots, a_{m(x)}(x))$ and $\boldsymbol{b}(y) = (b_1(y), b_2(y), \ldots, b_{n(y)}(y))$, respectively. That is, $m(x)$ is the number of associated words of x, and $a_i(x)$ is the mutual information between x and its i-th associated word x_i. Likewise, $n(y)$ is the number of associated words of y, and $b_j(y)$ is the mutual information between y and its j-th associated word y_j.

First, $\boldsymbol{b}(y)$ is translated into a first-language vector, denoted as $\boldsymbol{a}'(y) = (a'_1(y), a'_2(y), \ldots, a'_{m(x)}(y))$. That is,

$$a'_i(y) = \max_{j=1,2,\cdots,n(y)} \delta_{i,j} \cdot b_j(y) \quad (i = 1,2,\cdots,m(x)), \qquad [1]$$

where $\delta_{i,j}$=1 if y_j is a translation of x_i; otherwise, $\delta_{i,j}$=0. All associated words of y cannot be translated into associated words of x. Associated words of y that cannot be translated into associated words of x result in a residual second-language vector, denoted as $\boldsymbol{b}'(y) = (b'_1(y), b'_2(y), \ldots, b'_{n(y)}(y))$. That is,

$$b'_j(y) = \begin{cases} b_j(y) & \cdots \quad \sum_{i=1}^{m(x)} \delta_{i,j} = 0 \\ 0 & \cdots \quad \text{otherwise} \end{cases} \qquad (j = 1,2,\cdots,n(y)). \qquad [2]$$

Thus, $\boldsymbol{b}(y)$ is converted into $\boldsymbol{a}'(y)::\boldsymbol{b}'(y)$, i.e., a concatenation of translated vector $\boldsymbol{a}'(y)$ and residual vector $\boldsymbol{b}'(y)$. Likewise, $\boldsymbol{a}(x)$ is converted into $\boldsymbol{b}'(x)::\boldsymbol{a}'(x)$, i.e., a concatenation of translated vector $\boldsymbol{b}'(x)$ and residual vector $\boldsymbol{a}'(x)$.

Next, normalized similarity of second-language word y_0 to first-language word x_0 is defined as

$$Sim_{x_0}(y_0) = cos(\boldsymbol{a}(x_0) :: \boldsymbol{0}_{n(y_0)}, \boldsymbol{a}'(y_0) :: \boldsymbol{b}'(y_0)) \Big/ \max_{y \in T(x_0)} \{cos(\boldsymbol{a}(x_0) :: \boldsymbol{0}_{n(y)}, \boldsymbol{a}'(y) :: \boldsymbol{b}'(y))\}, \qquad [3]$$

where $\mathbf{0}_{n(y)}$ is an $n(y)$-dimensional zero vector and $T(x_0)$ is a set consisting of all candidate translation equivalents for x_0. Likewise, normalized similarity of first-language word x_0 to second-language word y_0 is defined as

$$Sim_{y_0}(x_0) = cos(\boldsymbol{b}(y_0) :: \mathbf{0}_{m(x_0)}, \boldsymbol{b}'(x_0) :: \boldsymbol{a}'(x_0)) \Big/ \max_{x \in T(y_0)} \{cos(\boldsymbol{b}(y_0) :: \mathbf{0}_{m(x)}, \boldsymbol{b}'(x) :: \boldsymbol{a}'(x))\}, \qquad [4]$$

where $\mathbf{0}_{m(x)}$ is an $m(x)$-dimensional zero vector and $T(y_0)$ is a set consisting of all candidate translation equivalents for y_0. Note that $Sim_{x_0}(y_0)$ is equal to 1 if and only if y_0 is most similar to x_0, and $Sim_{y_0}(x_0)$ is equal to 1 if and only if x_0 is most similar to y_0.

Finally, similarity between first-language word x_0 and second-language word y_0 is defined as the harmonic mean of bidirectional normalized similarities, that is,

$$Sim(x_0, y_0) = 2 \cdot Sim_{x_0}(y_0) \cdot Sim_{y_0}(x_0) \big/ (Sim_{x_0}(y_0) + Sim_{y_0}(x_0)). \qquad [5]$$

This definition is used only when y_0 is included in the top M words in descending order of normalized similarity to x_0, and vice versa. In other cases, $Sim(x_0, y_0)$ is defined as zero. Parameter M, which limits the numbers of similar words in both directions, was determined to be 100 experimentally. Note that the combination of bidirectional normalized similarities rarely ranks a candidate translation equivalent tenth or higher, when the similarity in either direction ranks it 101st or lower.

2.3.2. Example

Table 3 lists example candidate translation equivalents ranked according to the combination of bidirectional normalized similarities. For target word "GOP," the translation equivalent "共和党 <*KYOUWA-TOU*>" is successfully ranked first. It is also ranked first according to the normalized similarity to the target word (Equation [3]). For target word "stock price," the correct translation equivalent "株価<*KABUKA*>" is ranked second, while it is ranked ninth according to the normalized similarity to the target word. This exemplifies that the combination of bidirectional normalized similarities often ranks correct translation equivalents higher than the normalized similarity in either direction. For target word "Rochester," the combination of bidirectional normalized similarities produces no results, while the normalized similarity to the target word results in a list that includes "コダック<*KODAKKU*>" (Kodak), "光学機器<*KOUGAKU-KIKI*>" (optical instrument), and others but not the correct translation equivalent.

Table 3: Example ranked candidate translation equivalents

(a) Target word "GOP"

		(Sim.)
1	共和党 (Republican Party)	0.925
2	議会 (Congress)	0.873
3	上下両院 (Upper and Lower Houses)	0.846
4	中間選挙 (off-year election)	0.846
5	医療保険制度改革 (medical security system reform)	0.845
6	財政均衡 (financial balance)	0.841
7	民主 (democracy)	0.821
8	民主党 (Democratic Party)	0.820
9	上院 (Senate)	0.819
10	選挙 (election)	0.812

(b) Target word "stock price"

		(Sim.)
1	株 (stock)	0.758
2	株価 (stock price)	0.754
3	株価指数 (stock price index)	0.749
4	総合株価指数 (composite stock price index)	0.745
5	債券相場 (bond market prices)	0.734
6	ロンドン株式相場 (London stock quotations)	0.733
7	株式 (stock)	0.733
8	米大手証券 (U.S. major security firm)	0.729
9	銘柄 (brand)	0.725
10	優良銘柄 (blue chip)	0.722

(c) Target word "Rochester"

		(Sim.)
1	-	-

Note: These examples are taken from the experiment described in Subsection 3.2.

As shown by this example, a target word often has zero-similarity with all candidate translation equivalents, which suggests that the corpus does not contain translation equivalents of the target word.

3. Experiments

3.1. Comparison of Proposed Method with Alternatives

An English corpus consisting of *WSJ* articles (July 1994 to December 1995; 189 MB) and a Japanese corpus consisting of *Nikkei* articles (December 1993 to November 1994; 275 MB) were used as the comparable corpora. The experiments focused on nouns, including compound nouns defined simply by part-of-speech sequence patterns. A window of 12 content words to either side was used to count co-occurrence frequencies. Pairs of nouns with mutual information larger than zero were then extracted; the threshold for mutual information was set low to cope with the weak comparability between the corpora. Both English and Japanese nouns are thus characterized by context vectors consisting of nouns weighted with mutual information.

A generic seed lexicon was constructed by collecting pairs of nouns that are translations of one another from the EDR English-to-Japanese and Japanese-to-English dictionaries. The resulting lexicon consists of 633,000 pairs of 269,000 English nouns and 276,000 Japanese nouns[2]. Test target words and their correct translation equivalents were determined as follows. First, all pairs consisting of English and Japanese nouns that meet the following condition were collected: the Japanese noun is the only translation equivalent of the English noun according to the above-mentioned generic seed lexicon. This is because test target words should have similar characteristics as the words not contained in the bilingual lexicon, which are target words in a practical setting. Next, the pairs of English and Japanese nouns that do not meet the following condition were filtered out: the English and Japanese nouns are included in the top 5000 in descending order of frequency of occurrence in the *WSJ* and *Nikkei* corpora, respectively. This resulted in a total of 121 English target words, each having one and only one correct Japanese translation equivalent.

In addition to the proposed method, three alternative methods shown in the following table were used to output lists of top-*K* candidate translation equivalents for each test target word.

Context vector translation \ Similarity measure	Equation [5]	Equation [3]
Use translation equivalents given by the adapted seed bilingual lexicon.	Proposed method	Alternative [B]
Use up to five most-frequent translation equivalents for each entry word.	Alternative [C]	Alternative [A]

The alternative method [A], which is comparable to the previous methods, provides a baseline.

Recall and precision of each method were calculated for K=1, 2,…, 25. The recall is the proportion of test target words whose output lists contained the correct translation equivalents. The precision is the proportion of output translation equivalents that were correct ones; it is calculated by neglecting

[2] Although the seed lexicon was very large, the proposed method would also perform well with a moderate-sized seed lexicon. It has been proved experimentally that the sense-vs.-clue correlation algorithm, which plays a key role in the seed-lexicon adaptation, works well with an incomplete-coverage lexicon.

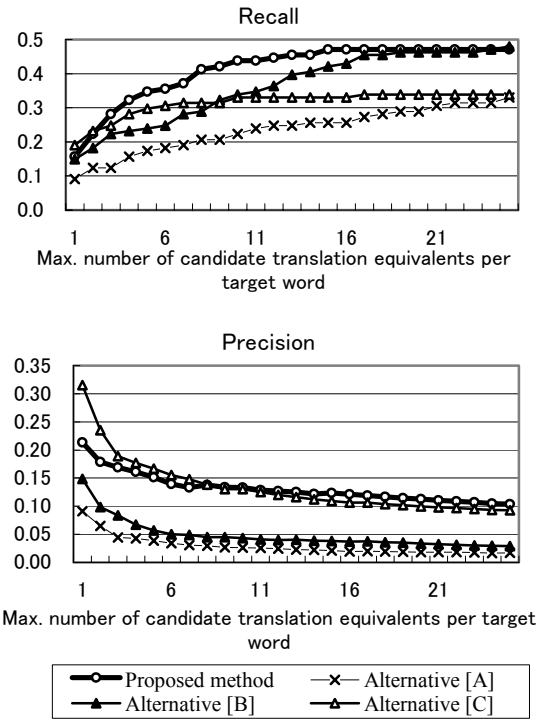

Figure 2: Recall and precision in cases of proposed method and alternatives

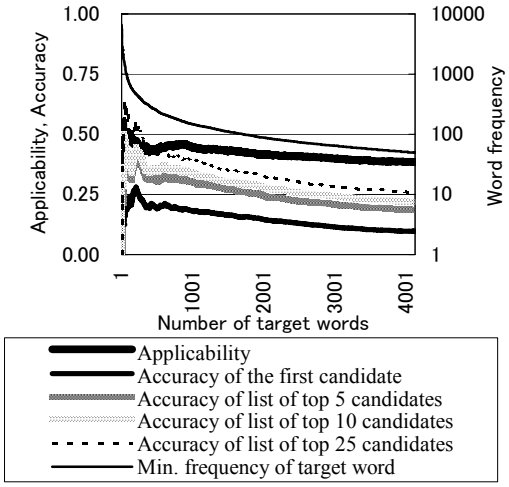

Figure 3: Applicability and accuracy

output translation equivalents ranked after the correct one, because users can skip them in a practical setting. Figure 2 shows how the recall and precision change with *K* in the cases of the proposed method and the alternative methods.

Comparing [B] to [A] reveals the effect of the adaptation of a seed bilingual lexicon. The adapted seed bilingual lexicon significantly improves the recall, although it improves the precision just slightly. This supports the assumption that the adapted seed bilingual lexicon enables context vectors to be translated correctly. Comparing [C] to [A] reveals the effect of the combination of bidirectional normalized similarities. The combined similarities significantly improve the precision, and they also improve the recall when *K* is small. This supports the assumption that the combined similarities give higher ranks to correct translation equivalents compared to the similarity in each direction. Figure 2 shows that these two effects are superimposed in the proposed method.

3.2. Evaluation under a Practical Setting

The proposed method was evaluated under a practical setting by using the same corpora and seed lexicon as described in the preceding subsection. The task was to find Japanese translations for English target words that occur frequently in the *WSJ* corpus but are not included in the EDR bilingual dictionary. Unlike the target words, candidate translation equivalents were not restricted to those not included in the EDR bilingual dictionary[3]. Although both the *WSJ* and *Nikkei* corpora consist mainly of financial and political articles, domestic news in respective countries makes up a

[3] This is because translation equivalents of unknown words may be known words. For example, "共和党 <*KYOUWA-TOU*>," a translation equivalent of an unknown word "GOP," is included in the EDR bilingual dictionary as a translation equivalent of "Republican Party."

majority. Their comparability is therefore very weak, and the existence of correct translation equivalents in the *Nikkei* corpus is not assured. The task was thus much tougher than those adopted by the previous works.

Table 4: Example translation equivalents extracted from *WSJ* and *Nikkei* corpora

Target word (Freq.)	Rank	Translation (Freq.)
Internet (1823)	4	インターネット<*INTAA-NETTO*> (592)
Sony (714)	16	ソニー<*SONII*> (1622)
European Union (529)	1	EU (2344)
budget deficit (321)	1	財政赤字<*ZAISEI-AKAJI*> (646)
Toy (268)	1	がん具<*GANGU*> (196)
Harvard (253)	5	ハーバード大学 <*HAABAADO-DAIGAKU*> (30)
World Trade Organization (227)	1	WTO (695)
American Airline (196)	15	アメリカン航空 <*AMERIKAN-KOUKUU*> (52)
World War II (183)	1	第二次大戦<*DAI-NIJI-TAISEN*> (227)
Hewlett-Packard (170)	1	HP (111)
business leader (157)	1	経済人<*KEIZAI-JIN*> (366)
Luxembourg (148)	1	ルクセンブルク <*RUKUSENBURUKU*> (126)
Gulf War (125)	1	湾岸戦争<*WANGAN-SENSOU*> (308)
privatizations (111)	1	民営化<*MIN'EI-KA*> (775)
Alzheimer disease (105)	2	アルツハイマー病 <*ARUTSUHAIMAA-BYOU*> (26)
electric vehicle (80)	1	電気自動車<*DENKI-JIDOUSHA*> (271)
assault weapon (67)	9	銃器<*JUUKI*> (29)
Japanese car (61)	5	日本車<*NIHON-SHA*> (335)
future price (56)	16	先物相場<*SAKIMONO-SOUBA*> (217)
Rabin (54)	21	ラビン<*RABIN*> (293)

Effectiveness of the method was evaluated by calculating applicability and accuracy. The applicability is the proportion of target words for which lists consisting of one or more candidate translation equivalents were output, and the accuracy is the proportion of output lists that included correct translation equivalents. Figure 3 shows how the applicability and accuracy change with *N*, the number of target words, where the target words are ordered in descending order of occurrence frequency. The accuracy was calculated in the cases of outputting the top 1, 5, 10, and 25 candidates. For example, for *N*=1716 (i.e., target words with occurrence frequencies not less than 100), the applicability was 42.4% and the accuracies of the lists of the top 1, 5, 10, and 25 candidates were 15.7%, 26.1%, 29.5%, and 34.2%, respectively.

The low applicability is not a shortcoming of the method, but it merely reflects the fact that the *Nikkei* does not contain translations for all words the *WSJ* contains. The method produced no output for target words such as "Eli Lilly & Co.," "third-quarter profit," "cyclicals," and "American Banker Association." These results seem quite reasonable, since the *Nikkei* is unlikely to contain Japanese translation equivalents of these English words; no output is more desirable than a list consisting of incorrect candidates. The low accuracy is also due in part to the weak comparability between the *WSJ* and *Nikkei* corpora. It is therefore necessary to improve the capability to judge whether a corpus contains translation equivalents for a target word. However, since the target words are all unknown words, an accuracy of around 30% would be acceptable and still useful.

Table 4 shows that useful pairs of translation equivalents, including technical terms and proper nouns, were extracted. Although the method generally performs better for frequently occurring words than for infrequently occurring words, it does not require a very high correlation between the frequencies of target words and those of their translation equivalents. The essential factor affecting the performance is how well the topics in which a target word appears and those in which its translation equivalent appears overlap. For example, the relatively low rank (16th) of "ソニー<*SONII*>" as a translation equivalent of "Sony" is due mainly to a much wider variety of *Nikkei* articles related to "ソニー<*SONII*>" compared to that of *WSJ* articles related to "Sony."

4. Conclusion

An improved method for extracting translation equivalents from bilingual comparable corpora according to contextual similarity was developed. It has two main features resulting in the improved performance. First, the seed bilingual lexicon is adapted to the corpora from which translation equivalents are to be extracted; the adapted seed bilingual lexicon improves the accuracy of translating context vectors. Second, the contextual similarity is evaluated by using a combination of bidirectional normalized similarity measures; the combined similarity measures usually rank correct translation equivalents higher than any single one does and, in addition, they make it possible to judge whether a corpus contains translation equivalents for a target word. An experiment using *Wall Street Journal* and *Nihon Keizai Shimbun* corpora together with the EDR bilingual dictionary demonstrated that the developed method is useful for improving the coverage of a bilingual lexicon.

Acknowledgments: This research was supported by the New Energy and Industrial Technology Development Organization of Japan (NEDO).

5. References

Dagan, Ido, Kenneth W. Church, and William A. Gale. 1993. Robust bilingual word alignment for machine-aided translation. In *Proc. Workshop on Very Large Corpora*, pp. 1-8.

Fung, Pascale. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proc. 33rd Annual Meeting of the ACL*, pp. 236-243.

Fung, Pascale and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proc. 5th Annual Workshop on Very Large Corpora*, pp. 192-202.

Fung, Pascale and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. 36th Annual Meeting of the ACL / 17th COLING*, pp. 414-420.

Gale, William A. and Kenneth W. Church. 1991. Identifying word correspondences in parallel texts. In *Proc. 4th DARPA Speech and Natural Language Workshop*, pp. 152-157.

Kaji, Hiroyuki and Toshiko Aizono. 1996. Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *Proc. 16th COLING*, pp. 23-28.

Kaji, Hiroyuki and Yasutsugu Morimoto. 2002. Unsupervised word sense disambiguation using bilingual comparable corpora. In *Proc. 19th COLING*, pp. 411-417.

Kaji, Hiroyuki. 2004. Bilingual-dictionary adaptation to domains. In *Proc. 20th COLING*.

Kitamura, Mihoko and Yuji Matsumoto. 1996. Automatic extraction of word sequence correspondences in parallel corpora. In *Proc. 4th Workshop on Very Large Corpora*, pp. 79-87.

Kupiec, Julian. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proc. 31st Annual Meeting of the ACL*, pp. 17-22.

Melamed, I. Dan. 1997. A word-for-word model of translational equivalence. In *Proc. 35th Annual Meeting of the ACL / 8th Conference of the EACL*, pp. 490-497.

Nakagawa, Hiroshi. 2001. Disambiguating of compound noun translations extracted from bilingual comparable corpora. In *Proc. 6th Natural Language Processing Pacific-Rim Symposium*, pp. 67-74.

Rapp, Reinhard. 1995. Identifying word translations in non-parallel texts. In *Proc. 33rd Annual Meeting of the ACL*, pp. 320-322.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proc. 37th Annual Meeting of the ACL*, pp. 519-526.

Utsuro, Takehito, Takashi Horiuchi, Kohei Hino, Takeshi Hamamoto, and Takeaki Nakayama. 2003. Effect of cross-language IR in bilingual lexicon acquisition from comparable corpora, In *Proc. 10th Conference of the EACL*, pp. 355-362.