Extracting Word Correspondences from Bilingual Corpora Based on Word Co-occurrence Information

Hiroyuki Kaji and Toshiko Aizono
Central Research Laboratory, Hitachi Ltd.
1-280, Higashi-koigakubo, Kokubunji-shi, Tokyo 185, Japan
{kaji, aizono}@crl.hitachi.co.jp

ABSTRACT

A new method has been developed for extracting word correspondences from a bilingual corpus. First, the co-occurrence information for each word in both languages is extracted from the corpus. Then, the correlations between the co-occurrence features of the words are calculated pairwisely with the assistance of a basic word bilingual dictionary. Finally, the pairs of words with the highest correlations are output selectively. This method is applicable to rather small, unaligned corpora; it can extract correspondences between compound words as well as simple words. An experiment using bilingual patent-specification corpora achieved 28% recall and 76% precision; this demonstrates that the method effectively reduces the cost of bilingual dictionary augmentation.

1 Introduction

Bilingual dictionaries are essential components for machine translation systems. One of the major problems with bilingual dictionaries is that they are expensive to build, since a huge number of terms are used in a variety of fields. Computer support is thus needed to reduce the cost of dictionary building.

With the growing volume of text available in electronic form, a number of methods have been proposed for extracting word correspondences from bilingual corpora automatically. These methods can be divided into those taking a statistical approach (Gale & Church 1991a; Kupiec 1993; Dagan et al. 1993; Inoue & Nogaito 1993; Fung 1995) and those taking a linguistic approach (Yamamoto & Sakamoto 1993; Kumano & Hirakawa 1994; Ishimoto & Nagao 1994). The statistical approach utilizes the occurrence frequencies and locations of words in a parallel corpus to calculate the pairwise correlations between the words in the two languages. The linguistic approach primarily extracts correspondences between compound words by consulting a bilingual dictionary of simple words.

These proposed methods for extracting word correspondences from bilingual corpora have the following drawbacks. First, most of them assume that the input corpora are aligned sentence by sentence, which reduces their applicability remarkably. Although a number of automatic sentence alignment methods have been proposed (Brown et al. 1991; Gale & Church 1991b; Kay & Roscheisen 1993; Chen 1993), they are not very reliable for real noisy bilingual texts. Second, the statistical methods usually require a very large corpus as their input. However, it is not easy to obtain a very large corpus. Third, the linguistic methods are restricted to extracting correspondences between compound words.

We have developed an extraction method that is free from the above drawbacks. In Sec. 2 we describe the basic idea of our method and give an overview. In Sec. 3 we describe the technical details, and in Sec. 4 we describe an experiment using patent-specification texts. In Sec. 5 we make a remark on the effectiveness of the proposed method, and discuss directions for improvement.

2 Overview of Proposed Method

The finding underlying our proposed method is as follows. In a bilingual corpus, a pair of words corresponding to each other generally accompany the same context, although expressed in the two different languages. If we calculate the pairwise correlations between the contexts in which the words occur, a correponding pair of words will show a high correlation. Although one occurrence of a word may not give a sufficient context to characterize the word, accumulating all the contexts in which the word occurs throughout the text allows the word to be distinguished from the other words in the same language text.

Figure 1 shows how two words are associated through their contexts, each expressed in its respective language. We use the set of words co-occurring with word w, which we refer to as the co-occurrence set of w, to concisely represent the accumulated contexts characterizing the word. To associate two co-occurrence sets whose elements are words in different languages, we consult a bilingual dictionary and extract the possible word correspondences between them. The point is that even if the pair of words to be associated is missing in the bilingual dictionary, their co-occurrence sets can be associated through the bilingual dictionary. Of cource, some of the correspondences between the co-occurrence sets may be also missing in the bilingual dictionary. Nevertheless, the co-occurrence sets can be still associated, owing to the other correspondences between them that are contained in the bilingual dictionary.

Our proposed method (Fig. 2) is based on the above

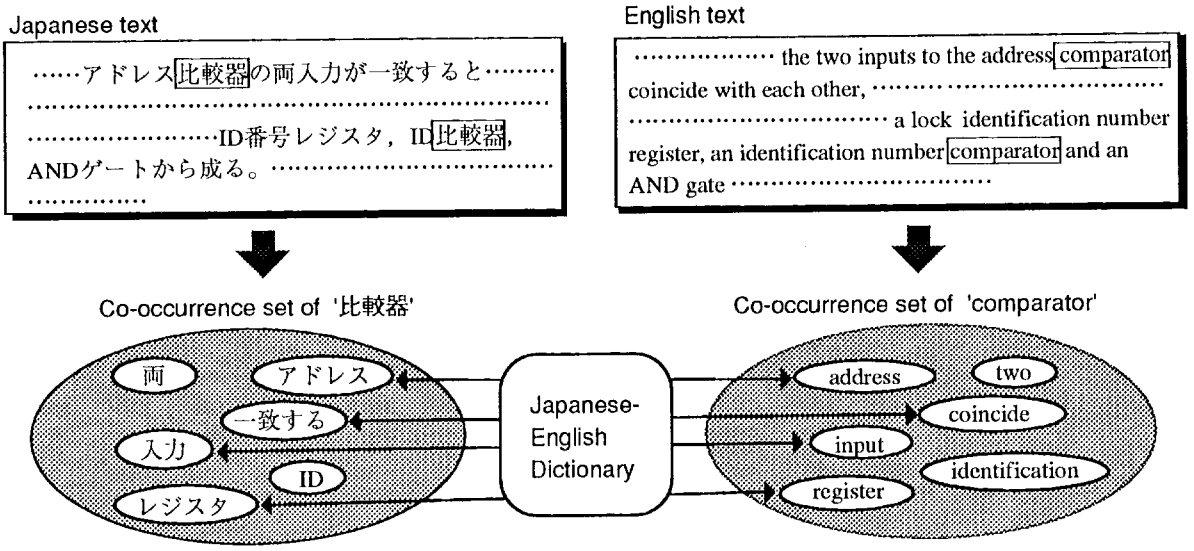


Fig. 1 Associating words through contexts.

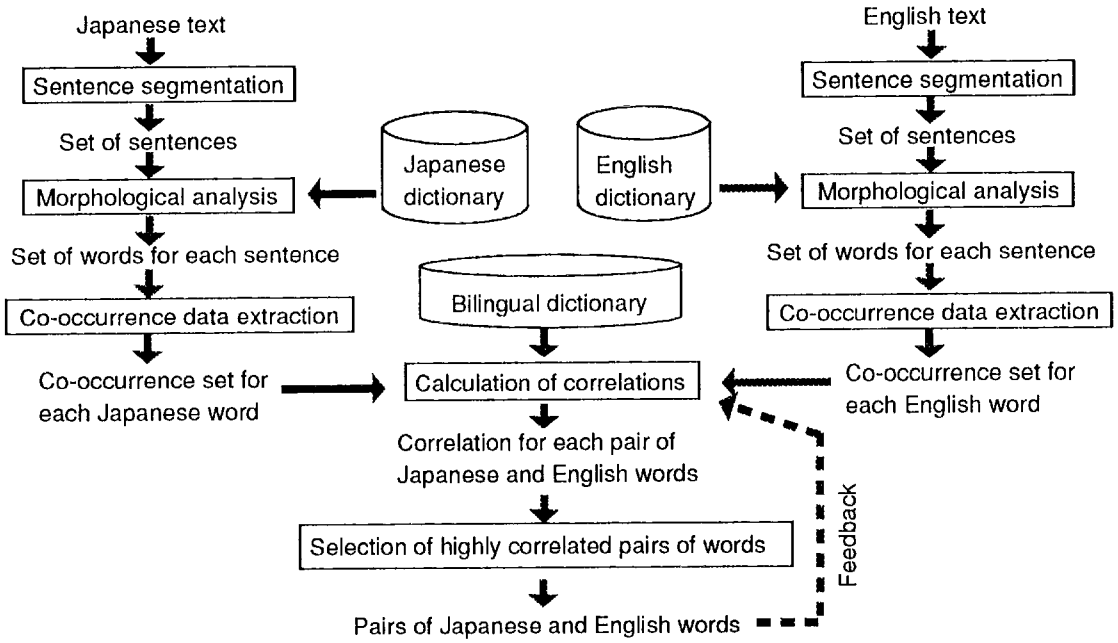


Fig. 2 Method for extracting word correspondences.

idea. While the examples shown here are for Japanese and English, the method is applicable to any pair of languages. The method is divided into three parts: Japanese text processing, English text processing, and bilingual processing. The Japanese text processing is composed of sentence segmentation, morphological analysis, and co-occurrence data extraction. It extracts a co-occurrence set for each word from a Japanese text. Likewise, the English text processing extracts a co-occurrence set for each word from an Engish text. The bilingual processing then calculates the pairwise correlations between the co-occurrence sets for Japanese words and those for English words, and selects the pairs of words with the highest correlations.

3 Technical Details

3.1 Extraction of words from text

Natural language texts are composed of two types of words: content words and function words. The target of extraction can usually be restricted to the correspondences between content words, which are characterized by both dominance in number and straightforwardness. Additionally, the function words are useless as elements of co-occurrence sets, since they do not indicate specific contexts. Therefore, we extract only the content words from the texts in both languages.

The content words are divided into simple words and

compound words. The former are extracted by dictionary look up and morphological analysis. To extract the latter, we are describing a set of rules or patterns. So far, we have only addressed nominal compounds (simple noun phrases), whose patterns are given below. Here, N, A, and NP stand for noun, adjective, and simple noun phrase, respectively. N+ stands for a string of one or more Ns.

- Japanese nominal compounds: NP := N N+
- English nominal compounds: NP := N N+ | A N+

The nominal compounds are extracted from the morphological analysis results by pattern matching. Here, an NP included in a larger NP is rejected, since only self-contained NPs qualify as nominal compounds. One exception is an English NP starting with a noun that is included in an NP starting with an adjective, because the case of an adjective modifying a nominal compound is just as likely as the case of an adjective being a part of a nominal compound.

3.2 Extraction of co-occurrence data

Definitions of 'co-occurrence' include syntactic co-occurrence, co-occurrence in a *k*-word window, co-occurrence in a sentence, and co-occurrence in a document. We use co-occurrence in a sentence, in which a pair of words occurring within the same sentence is regarded as a co-occurrence. While co-occurrence in a *k*-word window may produce better results when a sentence in one language corresponds to a sequence of two or more shorter sentences in the other language, it is difficult to determine an appropriate value of *k* because word order differs considerably between Japanese and English.

The relations between a compound word and its constituent words are not, strictly speaking, co-occurrence relations. Moreover, if we treated them in the same manner as co-occurrence relations, it would cause some confusion. Suppose that compound word w is composed of two simple words, w' and w''. If we included both w' and w'' in the co-occurrence set of w, and vice versa, the differences between the co-occurrence set of w and those of w' and w'' would decrease. Therefore, we exclude the constituent words from the co-occurrence set of a compound word and vice versa.

As mentioned in Section 2, the co-occurrence sets of a word are accumulated. This is not a mere union operation, but a union operation accompanied by frequency counting. The resultant co-occurrence set is expressed as

$$C(w) = \{ w_i / f_i \mid i = 1, \cdots, n \},$$

which shows that word w_i co-occurs with word w f_i times.

3.3 Calculation of correlations between words

We define correlation $R(jw, ew)$ between Japanese word jw and English word ew as follows.

$$R(jw, ew) = | C(jw) \cap C(ew) | \ / \ \{ | C(jw) | + | C(ew) | - | C(jw) \cap C(ew) | \}.$$

Here, $C(jw) = \{ jw_i / f_i \mid i = 1, \cdots, m \}$ and $C(ew) = \{ ew_j / g_j \mid j = 1, \cdots, n \}$ are the co-occurrence sets of jw and ew, respectively. $C(jw) \cap C(ew) = \{ (jw_i, ew_j) / h_{i,j} \mid i = 1, \cdots, m; j = 1, \cdots, n \}$ is the intersection of $C(jw)$ and $C(ew)$, whose elements are pairs of a Japanese word and an English word with their frequency. $| \cdot |$ means the sum of frequencies of all elements.

Generating intersection $C(jw) \cap C(ew)$ from $C(jw)$ and $C(ew)$ is not easy because the procedure of pairing jw_i ($\in C(jw)$) and ew_j ($\in C(ew)$) is nondeterministic. A pair of words cannot be determined independently of the other possible pairs. To reduce processing time, we calculate $| C(jw) \cap C(ew) |$ approximately, as illustrated in Fig. 3. For example, the English-based approximate calculation is done as follows. First, Japanese co-occurrence set C(jw) is transformed into pseudo co-occurrence set $C_p(jw)$ by consulting bilingual dictionary D, which is a set of pairs of words:

$$C_p(jw) = \{ ew_j / f'_j \mid j = 1, \cdots, n \},$$

where $f'_j = \sum_{w_i \in C(jw) \ \& \ (jw_i, ew_j) \in D} f_i$.

The intersection of pseudo co-occurrence set $C_p(jw)$ and English co-occurrence set $C(ew)$ is then generated:

$$C_p(jw) \cap C(ew) = \{ ew_j / \min\{f'_j, g_j\} \mid j = 1, \cdots, n \}.$$

Finally, $| C_p(jw) \cap C(ew) |$ is calculated as the approximate value of $| C(jw) \cap C(ew) |$:

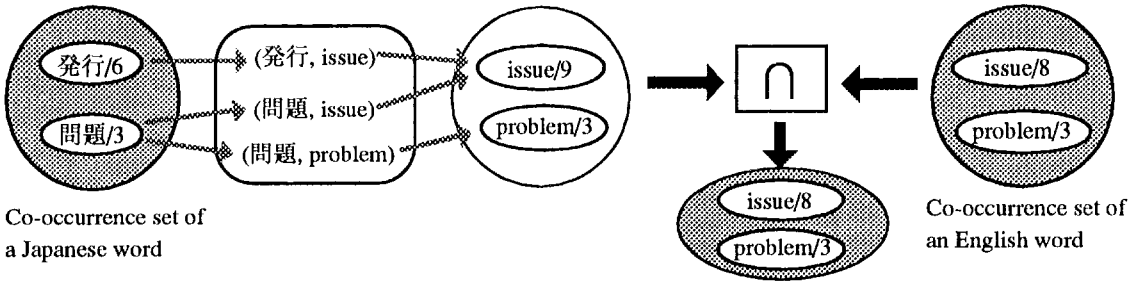
$$| C_p(jw) \cap C(ew) | = \sum_j \min\{f'_j, g_j\}.$$

This approximate calculation is likely to result in an overestimated correlation when there is ambiguity in pairing jw_i ($\in C(jw)$) and ew_j ($\in C(ew)$), as occurs in Fig. 3(a). Figure 3(a) shows that the number of elements in the intersection exceeds that in the Japanese co-occurrence set. The English-based and Japanese-based approximate calculations therefore do not always coincide with each other. While selecting the minimum of the two approximate values is safer, it does not guarantee a precise value. Since ambiguity in associating co-occurrence sets does not occur too often, and considering the need for efficiency, we execute either of the two approximate calculations rather than make a precise calculation.

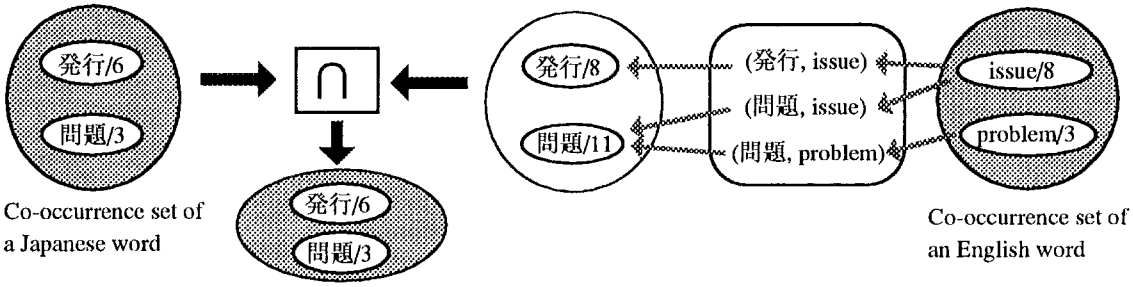
To increase the reliability of the correlation values, we remove the useless words from the co-occurrence sets before calculating the correlations. The useless Japanese word is jw such that $\{ ew \mid (jw, ew) \in D \} \cap \{ ew \mid ew \in T_E \} = \phi$ (T_E is the input English text), and the useless English word is ew such that $\{ jw \mid (jw, ew) \in D \} \cap \{ jw \mid jw \in T_J \} = \phi$ (T_J is the input Japanese text). These words do not contribute to the word-pair correlations.

3.4 Selection of pairs of words with high correlation

The absolute values of the correlations are not significant because they are sensitive to the numbers of words in the co-occurrence sets, which vary considerably from word to word. However, their relative values are significant when either a Japanese or an English word is fixed. We take the strategy of selecting the mutually best-matched pairs having no highly probable competitors. We call (jw,



(a) English-based approximate calculation



(b) Japanese-based approximate calculation

Fig. 3 Approximate calculation of correlation.

ew), a pair of a Japanese word and an English word, the mutually best-matched pair when

$R(jw, ew) > R(jw, ew')$ for any $ew' (\neq ew)$ and

$R(jw, ew) > R(jw', ew)$ for any $jw' (\neq jw)$.

When for a mutually best-matched pair (*jw*, *ew*), there exists either ew' such that

$R(jw, ew') > \alpha \cdot R(jw, ew)$ and $(jw, ew') \in D$

or jw' such that

$R(jw', ew) > \alpha \cdot R(jw, ew)$ and $(jw', ew) \in D$,

we call (*jw*, *ew'*) or (*jw'*, *ew*) a highly probable competitor. Here, α is a predetermined constant ($0 \leq \alpha < 1$), and D is the bilingual dictionary.

3.5 Feedback of extracted pairs of words

Obviously, the performance of the proposed method depends upon the coverage of the bilingual dictionary over the corpus. The coverage is the proportion of the word correspondences in the corpus that are already contained in the bilingual dictionary. Generally speaking, the wider the coverage, the more reliable the correlation values. Accordingly, the feedback of extracted pairs will probably improve performance, even though some of them are erroneous. In Fig. 2, the feedback is represented by dotted line.

4 Experiment and Results

We implemented our proposed method on a workstation and carried out an experiment using patent-specification documents in Japanese and English and a bilingual dictionary for a machine translation system. The dictionary contains approximately 60,000 Japanese entry words, each having several English translations. The quantitative profile of the sample patent documents is shown in Table 1(a).

We executed the word correspondence extraction program for each document. Parameter α in the selection of pairs of words was assumed to be 0. This means that the output pairs were limited as much as possible. Both results before and after feedback were obtained to evaluate the effect of feedback. The extracted pairs of words were divided into two groups: those which are already contained in the bilingual dictionary and those which are not yet contained in the bilingual dictionary.[1] The former are insignificant from the practical point of view. However, they are signficant in evaluating the effectiveness of the proposed correlation measure because the dictionary information regarding a particular pair of words does not contribute to the correlation between the pair itself. Accordingly, we evaluated two cases: Case A - the already known pairs of words are included - and Case B - the already known pairs of words are excluded.

A good way to evaluate word correspondence extraction methods is to measure their recall and precision. These measures are defined as follows. The recall is the proportion of all word correspondences in a

[1] We neglected the reference numbers peculiar to the patent documents because their correpondences are trivial. The underlined numerals in the following pair of sentences is an example of a reference number: ……アドレス比較器504の両入力が一致すると…／…… the two inputs to address comparator 504 coincide with ….

Table 1 Experimental profile and results.

(a) Profile of sample patent documents

	Document #	I	II	III	IV	V	Total
Japanese text	Number of content words * [a]	1,322	2,089	8,023	3,846	2,449	17,729
	Number of sentences [b]	90	120	686	230	178	1,304
	Average sentence length [a]/[b]	14.7	17.4	11.7	16.7	13.8	13.6
	Number of content words ** [c]	202	273	719	392	524	2,110
	(Number of content words whose translations are unknown) [d]	(39)	(43)	(146)	(51)	(97)	(376)
	Number of candidate compound words [e]	62	97	395	251	288	1,093
English text	Number of content words * [a']	1,463	2,055	9,561	4,326	2,872	20,277
	Number of sentences [b']	94	143	704	236	178	1,355
	Average sentence length [a']/[b']	15.6	14.4	13.6	18.3	16.1	15.0

* one count per occurrence ** one count per word

(b) Results of Case A

	Document #	I	II	III	IV	V	Total
Before feedback	Number of pairs of words extracted [f1]	78	123	366	247	203	1,017
	Number of correct pairs extracted [g1]	69	115	322	212	172	890
	Pseudo-recall [g1]/([c]+[e])	0.261	0.311	0.289	0.330	0.212	0.278
	Precision [g1]/[f1]	0.885	0.935	0.880	0.858	0.847	0.875
After feedback	Number of pairs of words extracted [f2]	83	135	400	257	231	1,106
	Number of correct pairs extracted [g2]	75	125	355	220	198	973
	Pseudo-recall [g2]/([c]+[e])	0.284	0.338	0.319	0.342	0.244	0.304
	Precision [g2]/[f2]	0.904	0.926	0.888	0.856	0.857	0.880

(c) Results of Case B

	Document #	I	II	III	IV	V	Total
Before feedback	Number of pairs of words extracted [h1]	31	53	190	131	100	505
	Number of correct pairs extracted [i1]	22	45	146	96	69	378
	Pseudo-recall [i1]/([d]+[e])	0.218	0.321	0.270	0.318	0.179	0.257
	Precision [i1]/[h1]	0.710	0.849	0.768	0.733	0.690	0.749
After feedback	Number of pairs of words extracted [h2]	31	60	202	140	111	544
	Number of correct pairs extracted [i2]	23	50	157	103	78	411
	Pseudo-recall [i2]/([d]+[e])	0.228	0.357	0.290	0.341	0.203	0.280
	Precision [i2]/[h2]	0.742	0.833	0.777	0.736	0.703	0.756

bilingual corpus that are actually extracted. The precision is the proportion of extracted word correspondences that are actually correct. While the precision is rather easy to calculate, the recall is difficult to calculate because it is a time-consuming task to manually identify all the word correspondences in the bilingual corpus. Therefore, instead of calculating the recall according to its de..nition, we make a rough estimation using the ratio of the number of correct pairs of words extracted to the number of words in either the Japanese or English text. We call this the pseudo-recall. The pseudo-recall indicates the lowest limit of the recall since a word in the Japanese text does not always have a straightforward counterpart in the English text, and vice versa.

Table 2 Examples of extracted word correspondences.

Type	Example
(S, S)	(排気, pumping) (引き続き, subsequently)
(S, C)	(液面, liquid level) (薄膜, thin film)
(C, S)	(気化器, vaporizer) (接続口, connector)
(C, C)	(ガス供給機構, gas supplier) (高周波加熱, radio frequency heating)

S: simple word, C: compound word

Tables 1(b) and (c) show the pseudo-recall and the precision in Cases A and B, respectively. In Case A, the pseudo-recall and precision before feedback were 27.8% and 87.5% respectively, and those after feedback were 30.4% and 88.0%. In Case B, the pseudo-recall and precision before feedback were 25.7% and 74.9%

respectively, and those after feedback were 28.0% and 75.6%.

The experiment confirmed that the proposed method can extract not only compound word correspondences but also simple word correspondences from a small corpus. Examples of word correspondences extracted from a patent document are shown in Table 2. The comparison of results before and after feedback supported the effectiveness of using feedback. That is, feedback increases recall while preserving precision. We also ascertained that repeating the feedback one more time did not result in significant improvement.

5 Discussion

The experiment shows that the proposed method is effective in reducing the cost of bilingual dictionary augmentation. The recall of the method is not high. Furthermore, it cannot extract more than one correspondence for a word. Still, the method is effective because it can extract from a small corpus. Bilingual documents should be handled separately. Even if a correspondence pair of words fails to be extracted from one bilingual document, it may be extracted from another bilingual document, where it occurs prevailingly.

The following are directions for further improvement.
(1) Refinement of nominal compound extraction procedure:
The simplified procedure described in Sec. 3.1 often causes omission (a nominal compound is not extracted) and noise (an inappropriate word string is extracted). These are major causes of errors in word correspondence extraction; refining the nominal compound extraction procedure will considerably improve recall and precision.
(2) Use of symbol/numeral correspondences:
In the present implementation, the correspondences of symbols and numerals are not used in calculating the correlation because the bilingual dictionary does not contain them. However, they have the potential of increasing the reliablilty of the correlation values. A character-string-matching routine to identify the correspondences of symbols/numerals should thus be added to the correlation calculation module.
(3) Use of the constituent word information of compound words:
The key idea of our method is to associate a pair of words through their co-occurrence information with the assistance of a bilingual dictionary. In contrast, that of the previous linguistic methods is to associate a pair of compound words through their constituent word information with the assistance of a bilingual dictionary. These two are not incompatible. Combining them would surely increase the recall and precision for compound word correspondences.

6 Conclusion

We have developed a new method for extracting word correspondences from bilingual corpora. The essence of the method is to calculate correlations between words based on their co-occurrence information with the assistance of a basic word bilingual dictionary. This method is applicable to rather small, unaligned corpora; it can extract correspondences between not only simple words but also between compound words. In an experiment with patent corpora, 28.0% pseudo-recall and 75.6% precision were achieved.

Acknowledgments: We would like to thank Dr. Michiharu Nakamura, Dr. Testuo Yokoyama and Dr. Hiromichi Fujisawa for their constant support and encouragement.

References

Brown, P. F., et al.1991. Aligning Sentences in Parallel Corpora. Proc. of the 29th Annual Meeting of the ACL, pp. 169-176.

Chen, S. F. 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information. Proc. of the 31st Annual Meeting of the ACL, pp. 9-16.

Dagan, I., et al. 1993. Robust Bilingual Word Alignment for Machine Aided Translation. Proc. of Workshop on Very Large Corpora, pp. 1-8.

Fung, P. 1995. A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. Proc. of the 33rd Annual Meeting of the ACL, pp. 236-243.

Gale, W. A. and K. W. Church. 1991a. Identifying Word Correspondences in Parallel Texts. Proc. of the 4th DARPA Speech and Natural Language Workshop, pp. 152-157.

Gale, W. A. and K. W. Church. 1991b. A Program for Aligning Sentences in Bilingual Corpora. Proc. of the 29th Annual Meeting of the ACL, pp. 177-184.

Inoue, N. and I. Nogaito. 1993. Automatic Construction of the Japanese-English Dictionary from Bilingual Text. Technical Report of IEICE, NLC93-39 (in Japanese).

Ishimoto, H and M. Nagao. 1994. Automatic Construction of a Bilingual Dictionary of Technical Terms from Parallel Texts. Technical Report of IPSJ, NL-102-11 (in Japanese).

Kay, M. and M. Roscheisen. 1993. Text-Translation Alignment. Computational Linguistics, Vol. 19, No. 1, pp. 121-142.

Kumano, A. and H. Hirakawa. 1994. Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistical Information. Proc. of COLING'94, pp. 76-81.

Kupiec, J. 1993. An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. Proc. of the 31st Annual Meeting of the ACL, pp. 17-22.

Yamamoto, Y. and M. Sakamoto. 1993. Extraction of Technical Term Bilingual Dictionary from Bilingual Corpus. Technical Report of IPSJ, NL-94-12 (in Japanese).