Systems and Computers in Japan, Vol. 37, No. 13, 2006
Translated from Denshi Joho Tsushin Gakkai Ronbunshi, Vol. J88-D-II, No. 4, April 2005, pp. 757–768

Automatic Extraction of Bilingual Word Pairs from Parallel Corpora with Various Languages Using Learning for Adjacent Information

Hiroshi Echizen-ya,[1] Kenji Araki,[2] and Yoshio Momouchi[1]

[1]Department of Electronics and Information Engineering, Hokkai-Gakuen University, Sapporo, 064-0926 Japan

[2]Division of Electronics and Information Engineering, Hokkaido University, Sapporo, 060-0814 Japan

SUMMARY

This paper presents a learning method using adjacent information as the method to extract bilingual word pairs efficiently from parallel corpora with various languages for which language resources are insufficient. In our method, information about correspondence between source language words and target language words is acquired automatically using the word strings that adjoin bilingual word pairs. That acquired information is used to solve the ambiguity problem of correspondence between source language words and target language words in various bilingual sentence pairs. First, the system using our method automatically acquires templates as information that indicates correspondence between source language words and target language words. The templates are based on word strings that adjoin the bilingual word pairs. Moreover, the system using our method efficiently extracts bilingual word pairs from bilingual sentence pairs using the acquired templates. Evaluation experiments showed that the system using our method extracted bilingual word pairs from parallel corpora with five kinds of languages. Results show that the total extraction rate was 60.1%. The total extraction rate was better by 8.0 percentage points compared to that obtained using a system based only on the Dice coefficient without our method. Those results confirm the effectiveness of our method. © 2006 Wiley Periodicals, Inc. Syst Comp Jpn, 37(13): 40–53, 2006; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/scj.20534

Key words: bilingual word pairs; adjacent information; learning; various languages; similarity measure.

1. Introduction

Parallel corpora are useful not only as a linguistic resource that is accessed using a computer. They are used as teaching materials of language learning. Therefore, many books and WWW sites show examples (e.g., English–Japanese, French–Japanese sentences) that appear as instructional materials. For language learning using such examples, providing the correspondence between words in sentences is important for users. That is, the first step of language learning is to understand the equivalents of various words in sentences. However, most example-based teaching materials do not supply word-level information clearly. For that reason, it is difficult for a user to obtain the word-level information reliably from example-based teaching materials alone. In such a case, it is very effective to obtain information of automatic word-level extraction of

Contract grant sponsor: Partially supported by grants from the High-Tech Research Center of Hokkai-Gakuen University.

bilingual word pairs from parallel corpora. In this paper, we demonstrate a method for efficient automatic extraction of bilingual word pairs from parallel corpora with various languages, including languages for which language resources, such as linguistic analyze tools, are insufficient.

Methods to extract bilingual word pairs automatically from a parallel corpus have been proposed. In automatic extraction of bilingual word pairs from a parallel corpus, the effectiveness of the Dice coefficient has been reported [1, 2]. The Dice coefficient is determined from a function that calculates a similarity value. However, it includes the sparse data problem. For example, the system tries to extract the bilingual word pairs for “book” using the Dice coefficient [4, 5] by function (1) as (Your book is on the table.; *teburu ni anata no hon ga ari masu.*[*])

$$Dice\,(X,Y) = \frac{2 \times f_{xy}}{f_x + f_y} \qquad (1)$$

In that equation, f_x and f_y represent the frequencies at which words X and Y appear independently; f_{xy} is the frequency at which words X and Y appear in bilingual sentence pairs simultaneously. Herein, when “book,” “*hon*,” and “*teburu*” appear only in (Your book is on the table.; *teburu ni anata no hon ga ari masu.*), the similarity value for “book” and “*hon*” is 1.0 (f_{xy}: 1, f_x: 1, f_y: 1), and the similarity value between “book” and “*teburu*” is also 1.0. Therefore, the system cannot infer a correct bilingual word pair for “book.” Such a problem of ambiguity in the correspondence between words is a common problem of systems based on similarity measures [2, 6].

To solve this problem, our method automatically acquires information about correspondences between source language words and target language words using the word strings that adjoin the bilingual word pairs. The acquired information includes two kinds of information. One is information about correspondence between the source language word strings and the target word strings in bilingual sentence pairs. The other is information about the correspondence between the source language words that adjoin the source language word strings and the target language words that adjoin the target word strings. Using such information, our method can extract bilingual word pairs by solving the problem of ambiguity in the similarity measure. For example, in the bilingual sentence pair (Your book is on the table.; *teburu ni anata no hon ga ari masu.*), our method acquires the information that “your” corresponds to “*anata no*” in Japanese, and the information that the equivalents of source language words that adjoin the right side of “your” exist on the right side of “*anata no*” in Japanese. Using the acquired information, our method can extract only a correct bilingual word pair (book; *hon*) from the bilingual sentence pair (Your book is on the table.; *teburu ni anata no hon ga ari masu.*).

Moreover, such adjacent information is acquired automatically during learning [7, 8] without analytical knowledge. As a result, our method can deal with various languages without modifying the system. In this paper, we propose a method for automatic extraction of bilingual word pairs using learning for adjacent information to extract bilingual word pairs efficiently from parallel corpora with various languages. Learning for adjacent information indicates a process that acquires the adjacent information automatically and extracts bilingual word pairs using the acquired adjacent information. In evaluation experiments, the automatic extraction of bilingual word pairs is performed by the system using our method for five parallel corpora: English–Japanese, French–Japanese, German–Japanese, Shanghai-Chinese–Japanese, and Ainu–Japanese. As a result, the extraction rates in the respective parallel corpora were 56.7% to 62.9%. The overall extraction rate was 60.1%. Moreover, for the system based only on the Dice coefficient without our method, the extraction rates in the respective parallel corpora were 47.9% to 54.9%, and the overall extraction rate was 52.1%. Those results indicate that the total extraction rate improved 8.0 percentage points using our method. Therefore, we confirmed the effectiveness of our method.

2. Outline of the System Using Our Method

The system using our method can extract bilingual word pairs without depending on specific languages whenever bilingual sentence pairs with morphological information are obtained. That is, our method requires no modification of the system even if the parallel corpus is replaced with a parallel corpus of other language. Herein, bilingual sentence pairs with morphological information are word-segmented sentences in an agglutinative language. Figure 1 shows the process flow of the system using our method.

We describe the process flow between the input of a source language word[*] and the extraction of a bilingual word pair for the source language word. Our method has two constituents. One is a process that extracts bilingual word pairs using learning for adjacent information. The

[*] In this paper, “ ” indicates the position of word segmentation. This process is performed using Japanese morphological analysis system “ChaSen” [3]. Italics denote Japanese pronunciations.

[*] In the source language words of bilingual word pairs, the number of words is always greater than 1. Therefore, “word strings” is a more illustrative expression. However, to distinguish between source language words and the word strings of adjacent information, “word” is used in this paper.

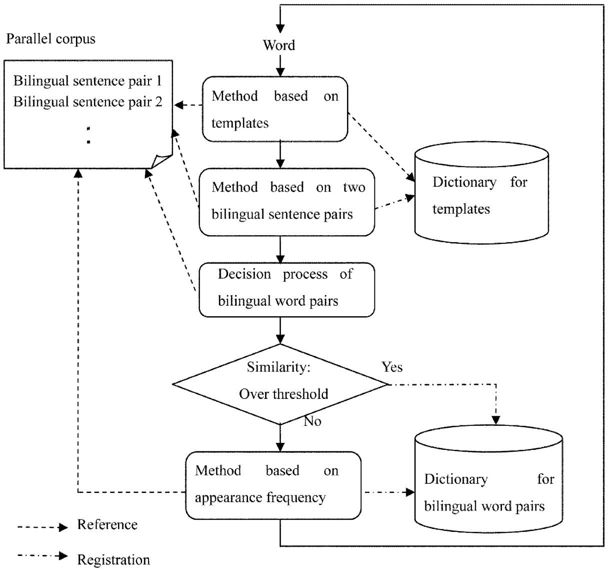

Parallel corpus
Bilingual sentence pair 1
Bilingual sentence pair 2
:
Word
Method based on templates
Method based on two bilingual sentence pairs
Dictionary for templates
Decision process of bilingual word pairs
Similarity: Over threshold
Yes
No
Method based on appearance frequency
Dictionary for bilingual word pairs
Reference
Registration

Fig. 1. Process flow.

other is a process that extracts bilingual word pairs using only the appearance frequency without learning for adjacent information. Moreover, learning for adjacent information has three constituents: the method based on templates, the method based on two bilingual sentence pairs, and the decision process of bilingual word pairs. In the method based on templates, the bilingual word pairs are extracted from bilingual sentence pairs beforehand using the acquired templates. In this paper, templates indicate bilingual knowledge that includes adjacent information to extract bilingual word pairs. In the method based on two bilingual sentence pairs, new templates and bilingual word pairs are obtained using two bilingual sentence pairs. In all obtained bilingual word pairs and templates, the similarity values are assigned, and all templates are registered into the dictionary for templates. The similarity values are obtained by referring to a parallel corpus. In the decision process of bilingual word pairs, the system chooses the most suitable bilingual word pairs when several candidates of bilingual word pairs exist. The system registers the chosen bilingual word pairs into the dictionary for bilingual word pairs when the similarity values of chosen bilingual word pairs are greater than a threshold value.

Furthermore, when the similarity values of chosen bilingual word pairs are not greater than that threshold, or when no bilingual word pairs are extracted through learning for adjacent information, the system extracts bilingual word pairs using the method based on appearance frequency. That is, in that case, the effective adjacent information was not obtained using learning for adjacent information. In the method based on appearance frequency, the system extracts bilingual word pairs using only the appearance frequency without learning of adjacent information, and registers the extracted bilingual word pairs into the dictionary for bilingual word pairs.

3. Learning for Adjacent Information

3.1. Method based on two bilingual sentence pairs

We next describe the extraction process of bilingual word pairs and the acquisition process of templates using the method based on two bilingual sentence pairs. Our method uses the same character strings (i.e., common parts) between two bilingual sentence pairs. Using common parts, the system determines the word strings that are used as templates. Moreover, in the obtained bilingual word pairs and templates, the similarity values are assigned using the Dice coefficient to represent their reliability.

3.1.1. Extraction process of bilingual word pairs

Details of the extraction process of bilingual word pairs by the method based on two bilingual sentence pairs are:

(1) From a parallel corpus, the system selects bilingual sentence pairs for which source language words of bilingual word pairs exist.

(2) The system chooses the bilingual sentence pairs that have word strings that match the word strings that adjoin the source language words in the source language sentences of the bilingual sentence pairs that are selected by process (1). In that case, the chosen bilingual sentence pairs must have common parts with the target language sentences of the bilingual sentence pairs selected by process (1).

(3) The system performs the following processes for target language sentences of two bilingual sentence pairs:

i) The system extracts parts from the words that adjoin the left side of the common parts to the words at the beginning of target language sentences using common parts at the beginning of target language sentences.

ii) The system extracts parts from the words that adjoin the right side of the common parts to the words at the end of target language sentences using the common parts at the end of target language sentences.

iii) The system extracts the parts between the two common parts using all combinations of the two common parts when the number of the common parts is greater than 2.

(4) The system checks the part-of-speech of the parts extracted in process (3). Consequently, the system removes

the extracted parts that are not nouns, verbs, adjectives, adverbs, conjunctions, noun phrases without postpositional particles, and verb phrases without postpositional particles.

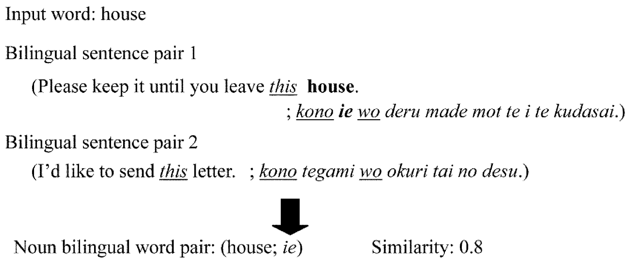
(5) The system calculates the similarity values between the source language words and the target words extracted by function (2); the system obtains the pairs of source language words and the extracted target words as bilingual word pairs.

In that equation, f_S represents the frequency at which word W_S appears independently, f_T is the frequency at which

$$sim\,(W_S, W_T) = \frac{2 \times f_{ST}}{f_S + f_T} \qquad (2)$$

word W_T appears independently, and f_{ST} is the frequency at which words W_S and W_T appear in bilingual sentence pairs simultaneously. Kitamura and Matsumoto [2] propose a Dice coefficient that is modified to reflect the frequency of bilingual word pair appearance. However, in this paper, we do not use the modified Dice coefficient because it is not effective when many bilingual word pairs, for which the frequencies are very low, appear in a parallel corpus. Moreover, in process (4), the part-of-speech is obtained using the Japanese morphological analysis system "ChaSen."

Figure 2 shows an example of extraction of English–Japanese bilingual word pairs using the method based on two bilingual sentence pairs. In Fig. 2, (house; *ie*) as a bilingual word pair for the source language word "house" is obtained using the method based on two bilingual sentence pairs. The system selects bilingual sentence pair 1, for which the source language word is "house," and chooses the bilingual sentence pair 2 that has "this," which adjoins "house," as the common part in the source language sentence, and has "*kono*" and "*wo*" as the common parts in the target language sentence. Therefore, "*ie*" between the two common parts "*kono*" and "*wo*" is extracted from the bilingual sentence pair 1 process by iii) of (3). Moreover, the similarity value between the source language word "house" and the extracted part "*ie*" is calculated using function (2) by process (5). The system determines only one bilingual word pair using the decision process of bilingual word pairs described in Section 3.3 when several bilingual word pairs are extracted.


Input word: house

Bilingual sentence pair 1

(Please keep it until you leave *this* **house**.
; *kono* ***ie*** *wo deru made mot te i te kudasai.*)

Bilingual sentence pair 2

(I'd like to send *this* letter. ; *kono tegami wo okuri tai no desu.*)

Noun bilingual word pair: (house; *ie*) Similarity: 0.8

Fig. 2. An example of extraction of bilingual word pair using method based on two bilingual sentence pairs.

3.1.2. Acquisition of templates

Details of the acquisition process of templates using the extracted bilingual word pairs and the word strings that adjoin them in the method based on two bilingual sentence pairs are:

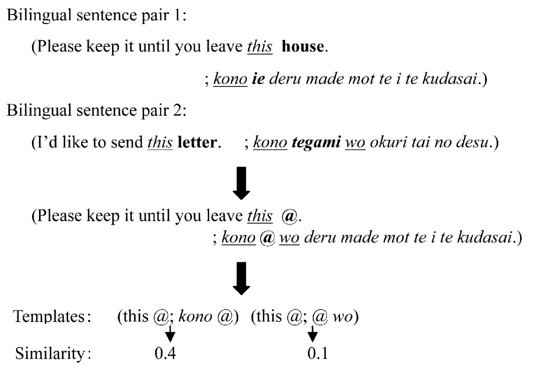
(1) The system replaces the bilingual word pairs with variables in the bilingual sentence pairs for which source language words exist.

(2) The system obtains templates by combining the pairs of the common parts and variables in source language sentences and the pairs of the common parts and variables in target language sentences.

(3) The system calculates the similarity values between the common parts in source language sentences and the common parts in target language sentences using function (2), and registers the templates into the dictionary for templates.

Figure 3 shows an example of acquisition of English–Japanese templates using the method based on two bilingual sentence pairs. Figure 3 shows the acquisition of (this; *kono* @) and (this; @ *wo*) as templates. The system replaces "house" and "*ie*" with "@" by process (1). The system extracts the pair of "this" and "*kono*" and the pair of "house" and "*ie*" by process (2). As a result, "this @" is extracted from the source language sentence of bilingual sentence pair 1, and "*kono* @" and "@ *wo*" are extracted from the target language sentence of bilingual sentence pair 1. Moreover, (this @; *kono* @) and (this; @ *wo*) are obtained by combining "this @" and "*kono* @," "@ *wo*," respectively.

In template (this @; @ *wo*), "this" does not correspond to "*wo*," which is a Japanese postpositional particle. In Japanese, the postpositional particle, auxiliary verb, "suru" that corresponds to "do" in English and "*aru*" that


Bilingual sentence pair 1:

(Please keep it until you leave *this* **house**.
; *kono* ***ie*** *deru made mot te i te kudasai.*)

Bilingual sentence pair 2:

(I'd like to send *this* **letter**. ; *kono* ***tegami*** *wo okuri tai no desu.*)

(Please keep it until you leave *this* **@**.
; *kono* **@** *wo deru made mot te i te kudasai.*)

Templates: (this @; *kono* @) (this @; @ *wo*)

Similarity: 0.4 0.1

Fig. 3. Example of acquisition of templates using the method based on two bilingual sentence pairs.

corresponds to "be" in English often appear; they have no correspondence words in general. Therefore, the system acquires numerous erroneous templates when they become the common parts. Fortunately, such erroneous templates have low similarity values. For example, in the template (this @; @ *wo*), "*wo*" often appears in target language sentences of bilingual sentence pairs. However, "this" and "*wo*" do not appear simultaneously because "this" does not correspond to "*wo*." As a result, f_T is higher than either f_{ST} or f_S, and the similarity value between "this" and "*wo*" is low. That is, the similarity value is low when the word strings of the source language do not correspond to the word strings of the target language in templates.

On the other hand, in the template (this @; *kono* @), not only "*kono*," but also "*kore*" is an equivalent of "this." Moreover, "*kono*" does not exist in (this evening; *konban*) although "this" exists. That is, the similarity values are not always high even when the word strings of the source language correspond to the word strings of the target language in templates. In our method, the similarity values of correct templates are higher than the similarity values of erroneous templates, but it cannot be said that the similarity values of correct templates are always high in themselves. The system can determine correct bilingual word pairs by distinguishing correct templates from erroneous templates even when several bilingual word pairs have equal similarity values. In the acquired templates, the source language part is called the word string of the source language, and the target language part is called the word string of the target language.

In our method, (this @; *kono* @) and (this; @ *wo*) are acquired as templates by separating "*kono*" and "*wo*," not (this @; *kono* @ *wo*). To apply the template (this @; *kono* @ *wo*), "*kono*" and "*wo*" always appear simultaneously in Japanese sentences. In contrast, templates like (this @; *kono* @) and (this; @ *wo*) are applicable to many more bilingual sentence pairs than are templates like (this @; *kono* @ *wo*).

3.2. Method based on templates

We describe the automatic extraction process of bilingual word pairs using the acquired templates in the method based on templates. Using templates, our method can extract bilingual word pairs by solving the ambiguity problem of correspondence between source language words and target language words. Details of the extraction process are:

(1) The system selects bilingual sentence pairs for which source language words exist.

(2) The system compares selected bilingual sentence pairs with templates in the dictionary for templates. Therefore, the system selects templates for which the source language parts are the same as the parts that adjoin the source language words in the source language sentences. It also selects the target language parts that have the same parts in the target language sentences.

(3) The system chooses the templates based on the following conditions between the source language sentences of the bilingual sentence pairs and the source language parts of templates.

i) The source language words must adjoin the right side of the common parts when the variables adjoin the right side of the source language parts of templates.

ii) The source language words must adjoin the left side of the common parts when the variables adjoin the left side of the source language parts of templates.

(4) The system performs the following process using the templates that satisfy the above condition to target language sentences of bilingual sentence pairs.

i) When the variables adjoin the right side in the target language parts of selected templates, the system extracts the parts (i.e., nouns, verbs, adjectives, adverbs, conjunctions, noun phrases without postpositional particles, and verb phrases without postpositional particles) that adjoin the right side of the common parts from the target language sentences.

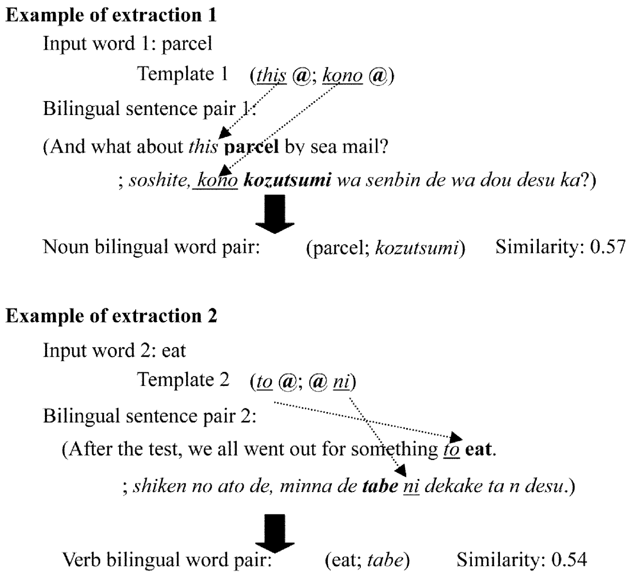
ii) When the variables adjoin the left side in the target language parts of selected templates, the system extracts the parts (i.e., nouns, verbs, adjectives, adverbs, conjunctions, noun phrases without postpositional particles, and verb phrases without postpositional particles) that adjoin the left side of the common parts from the target language sentences.

(5) The system determines the target language words of bilingual word pairs using the parts extracted by process (4). It then uses function (2) to calculate the similarity values between the source language words and the target language words.

(6) The system repeats process (4) to process (5) for other selected templates.

Figure 4 shows examples of extraction of English–Japanese bilingual word pairs using the method based on templates. In the example of extraction 1, the bilingual word pair (parcel; *kozutsumi*) was obtained using the template (this @; *kono* @) that was acquired in Fig. 3. Moreover, the bilingual word pair (eat; *tabe*[*]) was obtained using the template (to @; @ *ni*) in the example of extraction 2. In the

[*] Before the bilingual word pair (eat; *tabe*) is registered into the dictionary for bilingual word pairs, "*tabe*" is changed to "*taberu*" because "*tabe*" is the conjugated form of "*taberu*." Therefore, the bilingual word pair (eat; *taberu*) is registered into the dictionary for bilingual word pairs. Fundamentally, the extracted words are registered without modification.



Example of extraction 1

Input word 1: parcel

Template 1 (<u>this</u> @; <u>kono</u> @)

Bilingual sentence pair 1:

(And what about *this* **parcel** by sea mail?

; *soshite, <u>kono</u>* ***kozutsumi*** *wa senbin de wa dou desu ka*?)

Noun bilingual word pair: (parcel; *kozutsumi*) Similarity: 0.57

Example of extraction 2

Input word 2: eat

Template 2 (<u>to</u> @; @ <u>ni</u>)

Bilingual sentence pair 2:

(After the test, we all went out for something <u>to</u> **eat**.

; *shiken no ato de, minna de* ***tabe*** <u>*ni*</u> *dekake ta n desu*.)

Verb bilingual word pair: (eat; *tabe*) Similarity: 0.54

Fig. 4. Examples of extracted bilingual word pairs using the method based on templates.

example of extraction 1, "this" and "*kono*" in (this @; *kono* @) exist in bilingual sentence pair 1. Therefore, the noun word "*kozutsumi*," which adjoins the right side of "*kono*," is extracted from the target language sentence of the bilingual sentence pair 1 by i) of process (4). In the example of extraction 2, the verb word "*tabe*" of the left side of "*ni*" is extracted from the target language sentence of the bilingual sentence pair 2 by ii) of process (4).

In our method, templates have information to cope with the different word order between the source language and target language, even when the grammatical structure of the source language differs from the grammatical structure of the target language. Therefore, the system can extract correct bilingual word pairs. The system determines only one bilingual word pair by the decision process of bilingual word pairs described in Section 3.3 when several bilingual word pairs are extracted by the method based on templates.

3.3. Decision process of bilingual word pairs

The bilingual word pairs are ranked using their similarity values when several bilingual word pairs are obtained using the method based on two bilingual sentence pairs and the method based on templates. Consequently, only bilingual word pairs that are ranked at the top are registered into the dictionary for bilingual word pairs. Details of this process are the following.

(1) The system selects the bilingual word pairs with the highest similarity values.

(2) The system selects the bilingual word pairs that are extracted using the templates with the highest similarity values when several bilingual word pairs with the same similarity values exist by process (1).

(3) The system selects bilingual word pairs that appear for the first time in a parallel corpus when several bilingual word pairs exist by processes (1) and (2).

Through that ranking, the bilingual word pairs that are ranked at the top are registered into the dictionary for bilingual word pairs only when their similarity values are greater than the threshold.

3.4. Repetition of extraction of bilingual word pairs using templates

In the extraction process of bilingual word pairs presented in Section 2, templates are acquired every time that the source language words are input. Therefore, our method engenders the problem that the number of acquired templates might change when the input order of source language words changes. To solve such a problem, after the input of all source language words is finished, the system performs the method based on templates again using all templates in the dictionary for templates. Figure 5 shows the outline of repetition of extraction processes of bilingual word pairs using templates. All existing source language words are registered into the dictionary for bilingual word pairs when this process is performed. That is, the bilingual word pairs for all source language words are obtained using Section 2. The reason is described in Section 5.3.

In Fig. 5, the method based on templates is performed every time the source language words in the dictionary for bilingual word pairs are input. The system chooses only one bilingual word pair when several bilingual word pairs are obtained in the decision process of bilingual word pairs. In the decision process of bilingual word pairs, the most suitable bilingual word pairs are selected using their similarity values and the similarity values of templates, as described in Section 3.3. Next, the system compares the similarity values of selected bilingual word pairs with a threshold value. The system does not use the selected bilingual word pairs when the similarity values are not greater than that threshold. When the similarity values of the selected bilingual word pairs are greater than that threshold, the system performs the following process: i) The system compares the similarity values of the selected bilingual word pairs with the similarity values of the existing bilingual word pairs when the existing bilingual word pairs in the dictionary for the bilingual word pairs are obtained using learning for adjacent information. It then

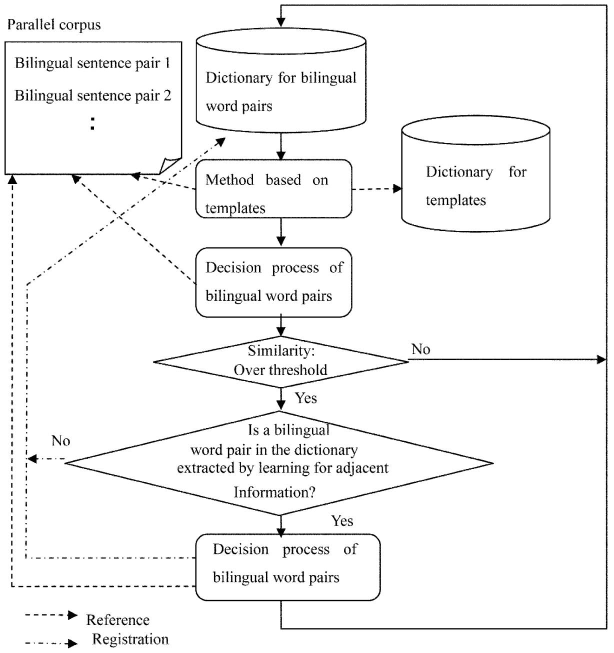

Parallel corpus
Bilingual sentence pair 1
Bilingual sentence pair 2
:
Dictionary for bilingual word pairs
Method based on templates
Dictionary for templates
Decision process of bilingual word pairs
Similarity: Over threshold
No
Yes
Is a bilingual word pair in the dictionary extracted by learning for adjacent Information?
No
Yes
Decision process of bilingual word pairs
Reference
Registration

Fig. 5. Process flow in repeated extraction of bilingual word pairs.

registers the bilingual word pairs with highest similarity values. ii) The system replaces the selected bilingual word pairs with the existing bilingual word pairs in the dictionary for bilingual word pairs when the existing bilingual word pairs are obtained by the method based on appearance frequency using the Dice coefficient. That is, the bilingual word pairs that were obtained using learning for adjacent information take precedence over the bilingual word pairs obtained using the method based on appearance frequency when the similarity values of bilingual word pairs are greater than a threshold, just as in the process of Section 2.

As one example of this process, the system can extract bilingual word pairs for "house" and "parcel" in Figs. 2 and 4 independent of the input order. In the process of Section 2, the system cannot extract the bilingual word pair (parcel; *kozutsumi*) because the template (this @; *kono* @) is not acquired when "parcel" is input as the first source language word. The templates (this @; *kono* @) and (this @; @ *wo*) are acquired by inputting "house." Therefore, only (house; *ie*) is obtained as the bilingual word pair. However, by the process of Fig. 5, the bilingual word pair (parcel; *kozutsumi*) is also obtained using the template (this @; *kono* @) for "parcel." In that case, (parcel; *kozutsumi*) is registered into the dictionary for the bilingual word pairs when the similarity value of (parcel; *kozutsumi*) is greater than a threshold.

4. Method Based on Appearance Frequency

The system extracts bilingual word pairs using only the appearance frequency without learning for adjacent information when the similarity values of the bilingual word pairs selected by the decision process of bilingual word pairs are not greater than a threshold or when no bilingual word pairs are extracted through learning for adjacent information. Details of this process are:

(1) The system selects bilingual sentence pairs for which source language words exist.

(2) The system extracts all nouns, verbs, adjectives, adverbs, conjunctions, noun phrases without postpositional particles, and verb phrases without postpositional particles from the target language sentences of the selected bilingual sentence pairs.

(3) The system uses function (2) to calculate the similarity values between the source language words and all words extracted from target language sentences. It then selects bilingual word pairs with the highest similarity values. In that case, when there are several bilingual word pairs with equal similarity values, the system selects the bilingual word pairs that appear for the first time in a parallel corpus.

Figure 6 shows an example of extraction of bilingual word pairs using the method based on appearance frequency. In Fig. 6, "*ie*," "*deru*," "*mot*," "*i*," and "*kudasai*" are shown to be extracted from the target language sentence of the bilingual sentence pair as candidates of equivalents of the source language word "house." The similarity values between the source language word "house" and all extracted target words are calculated using function (2). As a result, (house; *ie*) is selected as the bilingual word pair and is registered into the dictionary for bilingual word pairs when all similarity values are equal because (house; *ie*) appears for the first time in a bilingual sentence pair. In the method based on appearance frequency, the extracted bilingual word pairs have the highest similarity values among

Input word: house

Bilingual sentence pair

(Please keep it until you leave this **house**.
; *konoie wo deru made mot te i te kudasai.*)

⬇

Candidate bilingual word pairs:
(house; *ie*), (house; *deru*), (house; *mot*), (house; *i*), (house; *kudasai*)

Fig. 6. Example of extraction of a bilingual word pair using the method based on appearance frequency.

all candidates. Therefore, they are registered into the dictionary for bilingual word pairs independent of a threshold.

5. Experiments for Performance Evaluation

5.1. Experimental procedure

Parallel corpora used as experimental data are of five kinds: English–Japanese, French–Japanese, German–Japanese, Shanghai-Chinese–Japanese, and Ainu–Japanese parallel corpora. All bilingual sentence pairs in five kinds of parallel corpora are printed in books [14–18]. Among those five languages, the agglutinative languages that require the word segmentation are Japanese and Shanghai-Chinese. In Japanese, word segmentation is performed using the Japanese morphological "ChaSen" analysis system. Shanghai-Chinese sentences in the book [17] are sentences for which word segmentation was performed beforehand. Table 1 shows details of the parallel corpora. In Table 1, the target language is Japanese.

Extraction of bilingual word pairs is performed using the system based on Figs. 1 and 5 for each parallel corpus in Table 1. All source language words are words that correspond to nouns, verbs, adjectives, adverbs, and conjunctions printed in the glossaries of the respective books [14–18]. In all parallel corpora, the initial conditions of the dictionary for bilingual word pairs and the dictionary for templates are empty. Moreover, the values between 0.1 and 1.0 (i.e., 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0) are used as threshold values.

5.2. Evaluation standards

The extraction bilingual word pairs were evaluated by the first author using the equivalents of the source language words in the books [14–18]. After that, the extraction rate was calculated using the following function (3):

$$\textit{Extraction rate } (\%) = \frac{\textit{Number of extracted correct bilingual word pairs}}{\textit{Number of all source language words}} \times 100 \qquad (3)$$

5.3. Experimental results

To confirm the effectiveness of our method, we performed evaluation experiments through comparison with a system based on the Dice coefficient. Table 2 shows extraction rates of the system based on the Dice coefficient and the system using our method. The system based on the Dice coefficient is defined here as a system using only the method based on appearance frequency described in Section 4 without learning for adjacent information. The number of bilingual word pairs means the number of correct bilingual word pairs in the parallel corpus. Moreover, "A" indicates the number of bilingual word pairs that were extracted using our method. Their bilingual word pairs were not extracted in the system based on the Dice coefficient. "B" indicates the number of the bilingual word pairs that were not extracted using our method. Their bilingual word pairs were extracted in the system based on the Dice coefficient. That is, "B" shows the additional effect of our method. In Table 2, the extraction rates are values that are obtained when 0.5 is used as the threshold. Details of the determination of the best threshold are described in Section 5.4.2.

Table 3 shows examples of bilingual word pairs that are extracted by the system using our method when 0.5 was used as the threshold. In Table 3, the correct bilingual word pairs correspond to "A" in Table 2 and the erroneous bilingual word pairs correspond to "B" in Table 2. Moreover, the values indicate similarity values.

In our method, the precision values achieved in function (4) are equal to the extraction rates because the bilingual word pairs for all source language words are obtained. The system using our method extracts bilingual word pairs using the method based on appearance frequency when bilingual word pairs are not extracted using learning for adjacent information. In the method based on appearance

Table 1. Details of parallel corpora

Source language	Number of bilingual sentence pairs	Average number of words in source language sentences	Average number of words in target language sentences
English	393	7.4	8.6
French	399	6.8	9.0
German	377	6.2	8.5
Sh.-Chinese	392	6.4	8.9
Ainu	233	7.8	9.3
Total	1,794	6.8	8.8

Table 2. Results of evaluation experiments

Source language	Dice coefficient	Our method	Number of bilingual word pairs	A	B
English	49.7%	58.0%	169	18	4
French	47.9%	56.7%	240	27	6
German	53.3%	61.0%	195	20	5
Sh.-Chinese	54.9%	62.9%	264	30	9
Ainu	54.0%	61.5%	213	25	7
Total	52.1%	60.1%	1,081	120	31

Table 3. Examples of bilingual word pairs extracted by learning for adjacent information

Source language	Correct bilingual word pairs	Erroneous bilingual word pairs	
		Bilingual word pairs	Equivalents
English	(cereal; *shirial*) 1.0	(curtains; *atarashii* [new]) 0.67	curtains
	(boarding house; *geshuku*) 1.0	(interesting; *soto* [outside]) 0.67	interesting
French	(monuments; *kinen kenzou butsu* [monuments]) 1.0	(surtout; *kankei* [relation]) 1.0	specially
	(cherche; *sagashi* [search]) 0.67	(petit; *tokoro* [place]) 0.67	small
German	(nämlich; *tsumari* [after all]) 0.67	(Wege; *hashi* [bridge]) 1.0	lane
	(das Foto; *syashin* [photograph]) 1.0	(Neues; *shinbun* [newspaper]) 0.67	new event
Sh.-Chinese	(hhobae; *taikin shi* [leave office]) 1.0	(zonvae; *gotiso shi* [treat]) 1.0	lunch
	(dhuzhakha; *syanhai gani* [shanghai crab]) 1.0	(yhiavae; *sabisu* [service]) 0.67	dinner
Ainu	(ekupa; *kuwae* [take something in one's mouth]) 1.0	(ape; *fut* [fall]) 1.0	rain
	(set; *nedoko* [bed]) 1.0	(tunasno; *oki* [get up]) 0.67	early

frequency described in Section 4, bilingual word pairs are obtainable whenever nouns, verbs, adjectives, adverbs, conjunctions, noun phrases without postpositional particles, or verb phrases without postpositional particles exist in the target language sentences of bilingual sentence pairs for which source language words exist. In the experiments, nouns, verbs, adjectives, adverbs, conjunctions, noun phrases without postpositional particles, or verb phrases without postpositional particles existed in the target language sentences of all bilingual sentence pairs for which the source language words exist. Therefore, the bilingual word pairs for all source language words were obtained using the method based on appearance frequency, even when the bilingual word pairs were not obtained in the learning for adjacent information. As a result, the number of all source language words in function (3) is equal to the number of extracted word pairs in function (4); the precisions are also equal to the extraction rates.

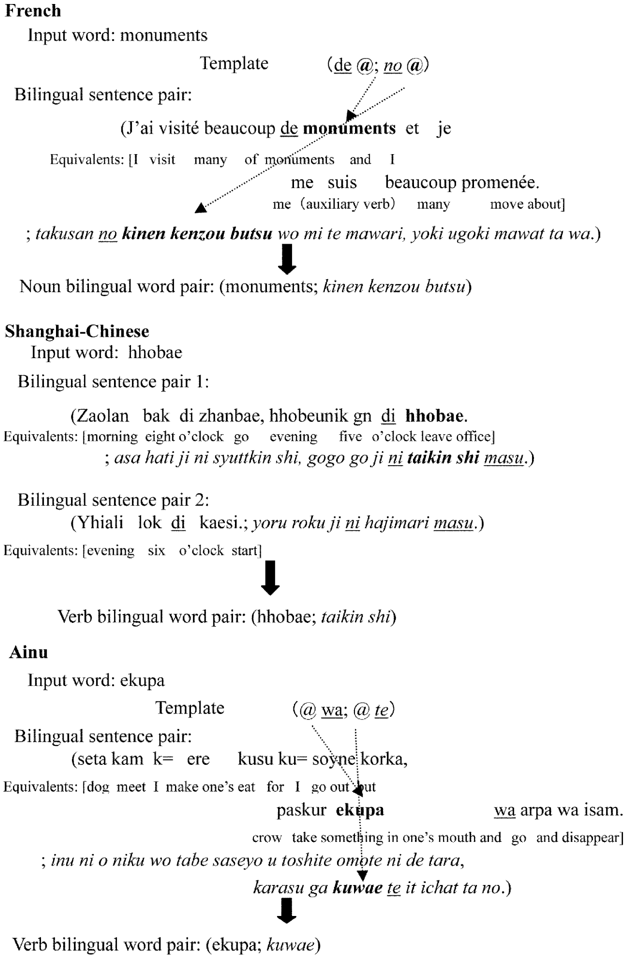
$$\textit{Precision}\ (\%) = \frac{\underline{\textit{Number of extracted correct bilingual word pairs}}}{\textit{Number of extracted bilingual word pairs}} \times 100 \tag{4}$$

5.4. Discussion

5.4.1. Effectiveness of our method

As presented in Table 2, the total extraction rate in the system using our method improved 8.0 percentage points compared with that of the system based on the Dice coefficient. In the parallel corpus, the extraction rates improved 7.5 to 8.8 percentage points. This fact indicates that our method is effective for all parallel corpora used in the experiments, independent of the language used. Figure 7 shows examples of extraction of French–Japanese bilingual word pair (monuments; *kinen kenzou butsu*), Shanghai-Chinese–Japanese bilingual word pair (hhobae; *taikin shi*), and Ainu–Japanese bilingual word pair (ekupa; *kuwae*) in Table 3. The processes shown in Figs. 2 to 4 were also obtained through experimentation.

Moreover, the bilingual word pairs that correspond to "B" in Table 2 were correctly obtained in the system based on the Dice coefficient. However, in the system using our method, they were not obtained because the erroneous bilingual word pairs were obtained using the erroneous adjacent information. The similarity values of some erroneous bilingual word pairs were higher than the threshold. Figure 8 shows examples of extraction of a German–Japanese bilingual word pair (Wege; *hashi*), and a Shanghai-Chinese–Japanese bilingual word pair (zonvae; *gotisou shi*) that are listed in Table 3. The extraction of an erroneous bilingual word pair (Wege; *hashi*) is based on the use of an erroneous template (@ und; @ *no*): "und" corresponds to "*ya*" in Japanese, not "*no*." The extraction of an erroneous bilingual word pair (zonvae; *gotisou shi*) is based on the use of erroneous adjacent information. That is, "`non qik`" does not correspond to "*wo*" and "*masu*" in Japanese. In

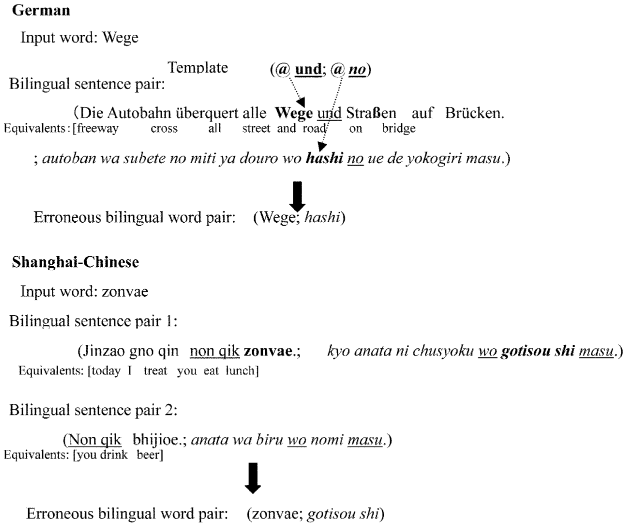


French
Input word: monuments
Template (de @; no @)
Bilingual sentence pair:
(J'ai visité beaucoup de **monuments** et je
Equivalents: [I visit many of monuments and I
me suis beaucoup promenée.
me (auxiliary verb) many move about]
; *takusan no **kinen kenzou butsu** wo mi te mawari, yoki ugoki mawat ta wa.*)
Noun bilingual word pair: (monuments; *kinen kenzou butsu*)

Shanghai-Chinese
Input word: hhobae
Bilingual sentence pair 1:
(Zaolan bak di zhanbae, hhobeunik gn di **hhobae**.
Equivalents: [morning eight o'clock go evening five o'clock leave office]
; *asa hati ji ni syuttkin shi, gogo go ji ni **taikin shi** masu.*)
Bilingual sentence pair 2:
(Yhiali lok di kaesi.; *yoru roku ji ni hajimari masu.*)
Equivalents: [evening six o'clock start]
Verb bilingual word pair: (hhobae; *taikin shi*)

Ainu
Input word: ekupa
Template (@ wa; @ te)
Bilingual sentence pair:
(seta kam k= ere kusu ku= soyne korka,
Equivalents: [dog meet I make one's eat for I go out but
paskur **ekupa** wa arpa wa isam.
crow take something in one's mouth and go and disappear]
; *inu ni o niku wo tabe saseyo u toshite omote ni de tara,*
*karasu ga **kuwae** te it ichat ta no.*)
Verb bilingual word pair: (ekupa; *kuwae*)

Fig. 7. Examples of extraction of correct bilingual word pairs.



German
Input word: Wege
Template (@ **und**; @ ***no***)
Bilingual sentence pair:
(Die Autobahn überquert alle **Wege** und Straßen auf Brücken.
Equivalents: [freeway cross all street and road on bridge
; *autoban wa subete no miti ya douro wo **hashi** no ue de yokogiri masu.*)
Erroneous bilingual word pair: (Wege; *hashi*)

Shanghai-Chinese
Input word: zonvae
Bilingual sentence pair 1:
(Jinzao gno qin non qik **zonvae**.; *kyo anata ni chusyoku wo **gotisou shi** masu.*)
Equivalents: [today I treat you eat lunch]
Bilingual sentence pair 2:
(Non qik bhijioe.; *anata wa biru wo nomi masu.*)
Equivalents: [you drink beer]
Erroneous bilingual word pair: (zonvae; *gotisou shi*)

Fig. 8. Examples of extraction of erroneous bilingual word pairs.

addition, "non" corresponds to "*anata* [you]" in Japanese and "qik" corresponds to "*taberu* [eat]" or "*nomu* [drink]" in Japanese. Moreover, (Wege; *hashi*) and (zonvae; *gotisou shi*) were registered into the dictionary for bilingual word pairs because their similarity values are 1.0. That is, the appearance frequencies of "Wege," "*hashi*," "zonvae," and "*gotisou shi*" are only 1.

The sparse data problem becomes serious in the method for extraction of bilingual word pairs using appearance frequency when the proportion of the low-frequency bilingual word pairs is high in a parallel corpus. The reason is that the accuracy of the similarity values is low in low-frequency bilingual word pairs, which is described in Section 1 as the problem of the Dice coefficient. Therefore, we investigated the effectiveness of our method in such a case. That is, we investigated the relationship between the appearance frequency and extraction rates of bilingual word pairs. Table 4 shows the investigation results. The ratio of bilingual word pairs for which the frequency is 1 was 63% among all bilingual word pairs in parallel corpora. Therefore, we investigated the extraction rates of bilingual word pairs for which the frequencies were 1 and more than 2.

Table 4 shows that the total extraction rate of bilingual word pairs for which the frequency is 1 improves by 11.0 percentage points using our method. In these evaluation experiments, the frequency is 1 for many bilingual word pairs for which the system based on the Dice coefficient cannot determine correct bilingual word pairs. Therefore, this result indicates that our method is effective when a system cannot determine correct bilingual word pairs using only similarity values, such as the system based on the Dice coefficient: the extraction rate of bilingual word pairs that have a frequency of 1 improved greatly.

Table 5 shows extraction rates and precision values of bilingual word pairs extracted through learning for adjacent information (i.e., the method based on templates, the method based on two bilingual sentence pairs, and the decision process of bilingual word pairs). Table 5 indicates different extraction rates in respective parallel corpora depending on the ratio of bilingual word pairs for which the frequency is 1. That is, the extraction rates are low because

Table 4. Details of extraction rates

Source	Dice coefficient		Our method	
language	frequency: 1	frequency: over 2	frequency: 1	frequency: over 2
English	35.7%	77.2%	46.4%	80.7%
French	37.5%	76.6%	49.4%	76.6%
German	39.5%	75.0%	51.3%	76.3%
Sh.-Chinese	40.0%	79.8%	49.1%	85.9%
Ainu	45.0%	63.5%	56.9%	66.3%
Total	39.4%	73.8%	50.4%	76.8%

the ratios of bilingual word pairs for which the frequency is 1 are large: the system cannot acquire adjacent information that uses word strings for which the frequency is greater than 2. In contrast, the extraction rates are high because the ratios of bilingual word pairs for which the frequency is 1 are small: the system can acquire adjacent information easily.

Moreover, the extraction rate in Table 5 is lower than the extraction rate of the system based on the Dice coefficient shown in Table 2. However, our system is not worse than the system based on the Dice coefficient because it extracts bilingual word pairs by determining only one equivalent for source language words. In contrast, the system based on the Dice coefficient determines a target language word that appears in a parallel corpus for the first time as the equivalent, as described in Section 4, when several equivalents with equal similarity values are obtained. Therefore, the system based on the Dice coefficient arbitrarily determines the equivalent for the source language word without a mode of authority such as similarity values. These experimental results indicated a ratio of 38.4% for the system based on the Dice coefficient: that is, the ratio of the bilingual word pairs that were extracted by selecting a target language word that appears in a parallel corpus for the first time to all extracted bilingual word pairs. The system using learning for adjacent information can extract the bilingual word pairs with authority. Furthermore, in Table 5, all precision values are between 77.0% and 88.0%; the overall precision is 82.5%. This fact means that the system using learning for adjacent information can extract the bilingual word pairs with high quality.

5.4.2. Decision of the most suitable threshold

We determined the most suitable threshold investigating the relationship between extraction rates and the threshold. Figure 9 shows the total extraction rate and the extraction rates in each parallel corpus when the thresholds are increased every 0.1.

Table 5. Extraction rates and precisions of bilingual word pairs extracted through learning for adjacent information

Source language	Extraction rate	Precision	Ratio of bilingual word pairs for which the frequency is 1
English	37.9%	85.3%	66.3%
French	35.4%	79.4%	73.3%
German	41.5%	85.3%	61.0%
Sh.-Chinese	44.3%	77.0%	62.5%
Ainu	51.6%	88.0%	51.2%
Total	42.3%	82.5%	63.0%

In the total extraction rate of Fig. 9, the extraction rate is highest when the threshold is 0.5. In each parallel corpus, the extraction rates of English and French are highest when the threshold is not 0.5. However, the differences between the highest extraction rates and extraction rates with a threshold of 0.5 are 0.6 and 1.6 points, respectively, for English and French. These different values are extremely small. Many correct bilingual word pairs that are extracted through learning for adjacent information are not registered when a large threshold value is used because their similarity values are lower than the threshold. Moreover, the method based on appearance frequency was incapable of extracting many bilingual word pairs for which the similarity value is lower than the threshold. As a result, the extraction rate is low. On the other hand, the extraction rate becomes low when a small value is used as a threshold because the similarity values of many erroneous bilingual word pairs extracted through learning for adjacent information are higher than the threshold. Therefore, the median value of 0.5 is the most suitable threshold to evaluate the extracted bilingual word pairs.

5.4.3. Problems of our method

In our method, bilingual word pairs cannot be extracted when the frequency of the word strings that adjoin the bilingual word pairs is only 1. For example, in the word string "XYZ," the system cannot extract the word "Y" when the respective frequencies of the words "X" and "Z" are only 1. That is, the words "X" and "Z" cannot become the adjacent information because they do not appear as the common parts, which would show the frequency as greater than 2. To solve such a problem, it is effective to use the bilingual word pairs extracted through learning for adjacent information. That is, the system uses the words "X" and "Z" as the adjacent information when the words "X" and "Z" already exist in the dictionary for bilingual word pairs.

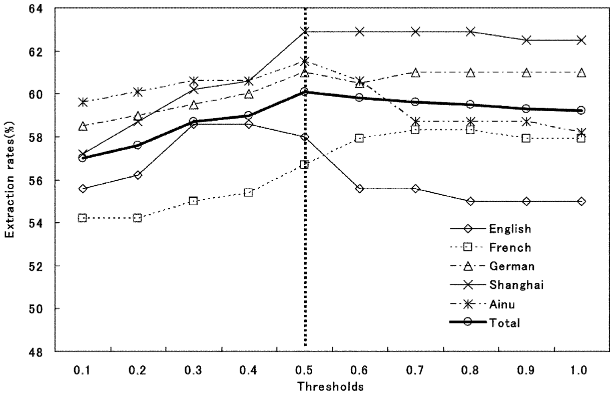


Fig. 9. Relationship between extraction rates and thresholds.

Moreover, in this evaluation experiment, a process to address word ambiguity is not performed. That is, the system determines only one bilingual word pair as correct bilingual word pairs, not all correct bilingual word pairs, even if several correct bilingual word pairs for one source language word are extracted. However, in this experimental result, the ratio of the case in which several correct bilingual word pairs exist to one source language word was 1.7% (18/1081). Therefore, the problem of word ambiguity was not serious. We will consider a new method to extract all correct bilingual word pairs by solving this problem of the word ambiguity.

In this study, the proposed system processes sentences for which word segmentation is performed for an agglutinative language. However, word segmentation is impossible in some cases; that is, when morphological analysis tools cannot be obtained. In such cases, it is effective to use the method for word segmentation presented by Wang and colleagues [9] as a method that does not depend on specific languages. This method is applicable to sentences with various languages because it performs word segmentation using only string characters of sentences.

In long bilingual sentence pairs, it is difficult to determine the correspondence between source language words and target language words because the number of words increases. For such a problem, it is effective to use syntax analysis tools for target language words. The system can divide the long sentences into phrases or clauses using the syntax analysis tools. In addition, the system limits the search scope using the bilingual word pairs obtained using learning for adjacent information, not only templates.

6. Comparison with Related Work

Tsuji and colleagues [6] proposed a method to determine the equivalents of Japanese nouns that are described in Katakana and to Romanize those words using translation rules. However, this method is inapplicable to various languages, and cannot process bilingual word pairs aside from nouns. Sato and Saito [10] proposed a method that extracts Japanese–English bilingual pairs of noun phrases and verb phrases using a support vector machine. However, this method requires a large-scale parallel corpus as a training corpus to obtain the learning model. In contrast, our method can extract bilingual word pairs efficiently using all given parallel corpora without any distinction between the training corpus and the test corpus. It is effective to solve the sparse data problem using learning for adjacent information.

Moreover, a method that automatically acquires templates has been proposed. McTait [11] proposed the method that acquires translation patterns by replacing common parts or different parts with variables. However, their translation patterns are global translation knowledge, which is effective only for whole sentences. Consequently, it is difficult to apply the translation patterns to local parts of sentences. Moreover, this method requires similarity of bilingual sentence pairs to acquire translation patterns. On the other hand, Ref. 8 proposed a method that acquires translation patterns; it specifically addresses local parts of sentences. However, this method requires bilingual word pairs that are extracted from similar bilingual sentence pairs *a priori*; it uses them to extract bilingual word pairs efficiently by focusing on the local parts of sentences. In contrast, our method can extract bilingual word pairs only from bilingual sentence pairs that are similar in local parts.

Kaji and Aizono [12] proposed a method that extracts bilingual word pairs using sets of co-occurrence words for bilingual word pairs as a method that uses words near bilingual word pairs. This method determines similarity based on the number and frequency of words that co-occur with bilingual word pairs. Therefore, this method can extract low-frequency bilingual word pairs when many words co-occur with bilingual word pairs. Moreover, Tanaka and Iwasaki [13] proposed a method that extracts bilingual word pairs by formulizing co-occurring information that is translated from the source language into the target language as a translation matrix to resolve ambiguity of the translational relation. However, these methods that use co-occurring information depend strongly on the frequency of co-occurring words. Therefore, they are insufficient in terms of efficient extraction of bilingual word pairs.

In contrast, our method merely requires the word strings that correspond to the adjacent information as the co-occurrence words. It can extract bilingual word pairs even when the frequency of the pairs of their word strings and the bilingual word pairs is only 1. For those reasons, our method can extract bilingual word pairs efficiently. For example, in extraction example 1 of Fig. 4, the system requires only the word string “this,” which corresponds to the adjacent information as the co-occurrence word to extract the bilingual word pair for “parcel.” Moreover, the system can extract the bilingual word pair for “parcel” even when the frequency of the pair of the co-occurrence word “this” and the source language word “parcel” is only 1. However, our method is insufficient to address data sparseness problems because the accuracy of word strings that correspond to adjacent information depends on similarity based on frequency information. We will consider a method that can obtain high-accuracy adjacent information without depending strongly on the frequency.

7. Conclusion

In this paper, we proposed a method for automatic extraction of bilingual word pairs using learning for adjacent information. It efficiently extracts bilingual word pairs from parallel corpora of various languages. Through this method, adjacent information, which is effective in solving the sparse data problem, is acquired automatically only from parallel corpora during the learning term. Therefore, our method can process various languages without modification of the system. It can extract bilingual word pairs with various languages by changing merely the parallel corpus. Experimental results using five kinds of parallel corpora, for which the respective source languages are English, French, German, Shanghai-Chinese, and Ainu and where the target language is Japanese, show that the extraction rate in the system using our method was 60.1%. This value is more than 8.0 percentage points higher than that of a system based on the Dice coefficient. This result indicates the effectiveness of our method.

In the future, we must resolve the ambiguity problem. That is, the system must select all correct bilingual word pairs when several correct bilingual word pairs for the same source language words are obtained. Moreover, we plan to adapt our method to a multilingual machine translation system.

Acknowledgment. This work was partially supported by grants from the High-Tech Research Center of Hokkai-Gakuen University.

REFERENCES

1. Smadja F, McKeown KR, Hatzivassiloglou V. Translation collocations for bilingual lexicons: A statistical approach. Computational Linguistics 1996;22:1–38.
2. Kitamura M, Matsumoto Y. Automatic extraction of translation patterns in parallel corpora. Inf Process Soc Japan 1997;38:727–736. (in Japanese)
3. Matsumoto Y, Kitauchi A, Yamashita T, Hirano Y, Matsuda H, Takaoka K, Asahara M. Japanese morphological analysis system ChaSen version 2.2.9 manual. Nara Institute of Science and Technology; 2002. (in Japanese)
4. Kay M, Röscheisen M. Text-translation alignment. Computational Linguistics 1993;19:121–142.
5. Manning CD, Schütze H. Foundations of statistical natural language processing. MIT Press; 1999.
6. Tsuji K, Yoshikane F, Kageura K. Low-frequency words in bilingual corpora—A step towards automatic extraction of bilingual word pairs. Tech Rep IEICE 2000;NLC2000-16. (in Japanese)
7. Echizen-ya H, Araki K, Momouchi Y, Tochinai K. Application of genetic algorithms for example-based machine translation method using inductive learning and its effectiveness. Inf Process Soc Japan 1996;37:1565–1579. (in Japanese)
8. Echizen-ya H, Araki K, Momouchi Y, Tochinai K. Machine translation using recursive chain-type learning based on translation examples. Trans IEICE 2002;J85-D-II:1840–1852. (in Japanese)
9. Wang Z, Araki K, Tochinai K. Application for Chinese and performance evaluation of word segmentation method using inductive learning. Trans IEICE 2002;J85-D-II:56–65. (in Japanese)
10. Sato K, Saito H. Extracting word sequence correspondences based on support vector machines. J Nat Language Process 2003;10:109–124. (in Japanese)
11. McTait K. Memory-based translation using translation patterns. Proc 4th Annual CLUK Colloquium, p 43–52, Sheffield, England, 2001.
12. Kaji H, Aizono T. Extracting word correspondences from bilingual corpora based on word co-occurrence information. Proc Coling'96, p 23–28, Copenhagen.
13. Tanaka K, Iwasaki H. Extraction of lexical translation from non-aligned corpora. Proc Coling'96, p 580–585, Copenhagen.
14. Harukawa Y, Snelling J. Express: English. Hakusui-sha; 1998. (in Japanese)
15. Chikushi F. Express: French. Hakusui-sha; 2001. (in Japanese)
16. Oshio T. Express: German. Hakusui-sha; 2004. (in Japanese)
17. Emoto H, Han G. Express: Shanghai. Hakusui-sha; 2004. (in Japanese)
18. Nakagawa H, Nakamoto M. Express: Ainu. Hakusui-sha; 2004. (in Japanese)

AUTHORS (from left to right)

Hiroshi Echizen-ya received his B.E. and M.E. degrees from Hokkai-Gakuen University in 1991 and 1996. From 1996 to 1998, he was a doctoral course student at Hokkaido University. Currently, he is a research associate on the Faculty of Engineering at Hokkai-Gakuen University. His research interests include natural language processing and machine translation. He is a member of ACL, IEEE, IPSJ, JSAI, and NLP.

Kenji Araki received his B.E. and Ph.D. degrees from Hokkaido University in 1982 and 1988. He was an associate professor there from 1991 to 1998, and a professor in 1998 at Hokkai-Gakuen University. He then joined Hokkaido University as an associate professor. He is currently a professor at the Graduate School of Information Science and Technology. His research interests include natural language processing, morphological analysis, machine translation, and speech dialogue processing. He is a member of ACL, IEEE, AAAI, IPSJ, and NLP.

Yoshio Momouchi received his B.E. and D.Eng. degrees from Hokkaido University in 1965 and 1973. He was a research associate from 1973 to 1984, a lecturer from 1984 to 1986, and an associate professor from 1986 to 1988 in the Division of Information Engineering at Hokkaido University. In 1988, he joined the Department of Electronics and Information Engineering at Hokkai-Gakuen University as a professor. His research interests include understanding and generation of natural language. He is a member of IPSJ, NLP, MLSJ, JSAI, JCSS, and ACL.