Available online at www.sciencedirect.com
SCIENCE DIRECT
Information Processing and Management 42 (2006) 1294–1315
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

# Automatic extraction of bilingual word pairs using inductive chain learning in various languages ☆

Hiroshi Echizen-ya [a,*], Kenji Araki [b], Yoshio Momouchi [a]

[a] Department of Electronics and Information Engineering, Hokkai-Gakuen University, South-26 West-11, Chuo-ku, Sapporo 064-0926, Japan
[b] Graduate School of Information Science and Technology, Hokkaido University, Kita-14, Nishi-9, Kita-ku, Sapporo 060-0814, Japan

Received 31 July 2005; received in revised form 23 November 2005; accepted 30 November 2005
Available online 23 January 2006

**Abstract**

In this paper, we propose a new learning method for extracting bilingual word pairs from parallel corpora in various languages. In cross-language information retrieval, the system must deal with various languages. Therefore, automatic extraction of bilingual word pairs from parallel corpora with various languages is important. However, previous works based on statistical methods are insufficient because of the sparse data problem. Our learning method automatically acquires rules, which are effective to solve the sparse data problem, only from parallel corpora without any prior preparation of a bilingual resource (e.g., a bilingual dictionary, a machine translation system). We call this learning method Inductive Chain Learning (ICL). Moreover, the system using ICL can extract bilingual word pairs even from bilingual sentence pairs for which the grammatical structures of the source language differ from the grammatical structures of the target language because the acquired rules have the information to cope with the different word orders of source language and target language in local parts of bilingual sentence pairs. Evaluation experiments demonstrated that the recalls of systems based on several statistical approaches were improved through the use of ICL.


*Keywords:* Learning method; Bilingual word pairs; Various languages; Sparse data problem; Parallel corpora; Statistical approach

## 1. Introduction

### 1.1. Sparse data problem in parallel corpora

In the field of Cross-Language Information Retrieval (CLIR) (Chen & Gey, 2004; Fujii & Ishikawa, 2001; Kishida et al., 2004; Xu & Weischedel, 2003), bilingual word pairs—the pairs of Source Language (SL) words and Target Language (TL) words—are essential. However, manual extraction by humans of bilingual word

☆ Note: An earlier version of this paper was presented at the NLDB2005 Conference.

* Corresponding author. Tel.: +81 11 841 1161x7863; fax: +81 11 551 2951.
*E-mail addresses:* echi@eli.hokkai-s-u.ac.jp (H. Echizen-ya), araki@media.eng.hokudai.ac.jp (K. Araki), momouchi@eli.hokkai-s-u.ac.jp (Y. Momouchi).

0306-4573/$ - see front matter 
doi:10.1016/j.ipm.2005.11.004

Table 1
A contingency matrix

| | $W_T$ | $\neg W_T$ | |
|---|---|---|---|
| $W_S$ | $a$ | $b$ | |
| $\neg W_S$ | $c$ | $d$ | |
| | | | $n$ |

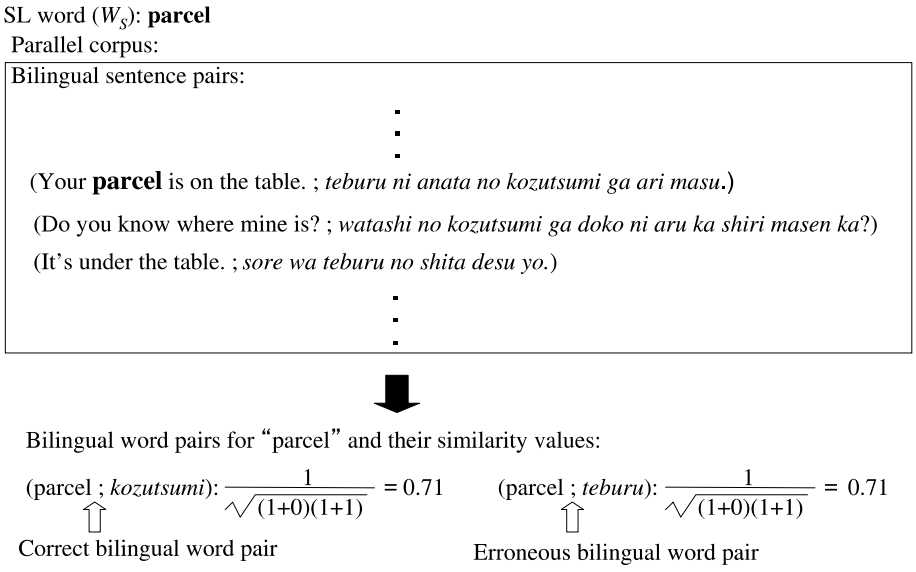


Fig. 1. An example of sparse data problem by the system based on cosine.

pairs in various languages is costly. For that reason, automatic extraction of bilingual word pairs from parallel corpora in various languages is important to make the method feasible. Statistical approaches are effective to extract bilingual word pairs from parallel corpora with various languages (Kay & Röscheisen, 1993; Manning & Schütze, 1999; Melamed, 2001; Sadat, Yoshikawa, & Uemura, 2003; Smadja, McKeown, & Hatzivassiloglou, 1996; Veronis, 2000) because they are language independent. However, they are insufficient because of sparse data problem. For example, the system uses cosine (Manning & Schütze, 1999) to extract bilingual word pairs from a parallel corpus. The cosine is the representative similarity measure in the statistical approaches. The cosine is defined as

$$\text{cosine}(W_S, W_T) = \frac{a}{\sqrt{(a+b)(a+c)}} \tag{1}$$

Table 1 shows the parameters used in function (1): $W_S$ is an SL word and $W_T$ is a TL word in a parallel corpus. The number of pieces in which both $W_S$ and $W_T$ were found in each bilingual sentence pair is represented as '$a$'; '$b$' is the number of pieces in which only $W_S$ was found in each bilingual sentence pair; and '$c$' is the number of pieces in which only $W_T$ was found in each bilingual sentence pair. In addition, '$n$' represents the total number of words in a parallel corpus; '$d$' denotes the values of '$n - (a + b + c)$'.

Fig. 1 shows an example of the sparse data problem by the system based on cosine.

In Fig. 1, the system based on cosine cannot extract only bilingual word pair (parcel;*kozutsumi*[1]) because the similarity value between "parcel" and "*kozutsumi*" becomes 0.71, and the similarity value between "parcel" and "*teburu*" also becomes 0.71 by the cosine function (1). This problem becomes very serious when the

[1] Italics indicate Japanese pronunciation. Space (i.e. ' ') in Japanese sentences are inserted after each morpheme because Japanese is an agglutinative language. This process is automatically performed using the Japanese morphological analysis system "ChaSen" (Matsumoto et al., 2000). Grammatical structure of Japanese is SOV.

system must deal with various languages because large-scale parallel corpora may be unobtainable in various languages. The frequencies of many bilingual word pairs are same and low when large-scale parallel corpora cannot be obtained. Therefore, it is difficult to automatically extract the bilingual word pairs by the system based on cosine. That is, the system based on cosine has the sparse data problem, and this obstacle is common among similarity measures, not only cosine.

### 1.2. Motivation

We propose a new learning method for solving the sparse data problem in automatic extraction of bilingual word pairs in various languages. In our learning method, the system limits the search scope for the determination of equivalents by focusing on local parts in bilingual sentence pairs. Fig. 2 shows an example of extraction of a bilingual word pair using our idea.

In Fig. 2, the rule (your @;*anata no* @) has the information that "your" corresponds to "*anata no*" in Japanese. Moreover, it has the information that the equivalents of words that adjoin the right side of "your" exist on the right side of "*anata no*" because variable "@" adjoins the right sides of "your" and "*anata no*", respectively. The variable "@" corresponds to a word. Using the rule (your @;*anata no* @), the system can extract only (parcel;*kozutsumi*) from parallel corpus. That is, the system can decrease the number of candidates of equivalents for SL words focusing on local parts in bilingual sentence pairs by the rules. Moreover, from the perspective of learning (Echizen-ya, Araki, Momouchi, & Tochinai, 2002; Echizen-ya, Araki, & Momouchi, 2005a, 2005b), all rules for extracting bilingual word pairs are automatically acquired only from parallel corpus without any bilingual resource. We call this new learning method Inductive Chain Learning (ICL), and the rule acquired by ICL is called the ICL rule. In our method, a chain reaction is caused for the acquisition of ICL rules and the extraction of bilingual word pairs. The main advantages of ICL are the following three:

(1) The system using ICL requires no bilingual resource (e.g., a bilingual dictionary, a machine translation system) beforehand to solve the sparse data problem. All ICL rules are acquired automatically solely from the parallel corpora.
(2) The system using ICL is effective for parallel corpora with various languages for which the grammatical structures of SL differ from the grammatical structures of TL (i.e., English–Japanese, not English–French, English–German) through the use of acquired ICL rules. The ICL rules have the information to cope with the different word orders of SL and TL in local parts of bilingual sentence pairs.
(3) The system using ICL can extract bilingual word pairs even when the frequency of the bilingual word pairs is only 1 in a parallel corpus. For example, in Fig. 2, when the ICL rule (your @;*anata no* @) exists, the system using ICL can extract (parcel;*kozutsumi*) even when the frequency of the pair of "parcel" and "*kozutsumi*" is only 1. This fact indicates that the system using ICL is effective to extract low-frequency bilingual word pairs which have the sparse data problem.

Evaluation experiments indicated that the system using ICL could extract not only high-frequency bilingual word pairs but also low-frequency bilingual word pairs from parallel corpora with various languages.

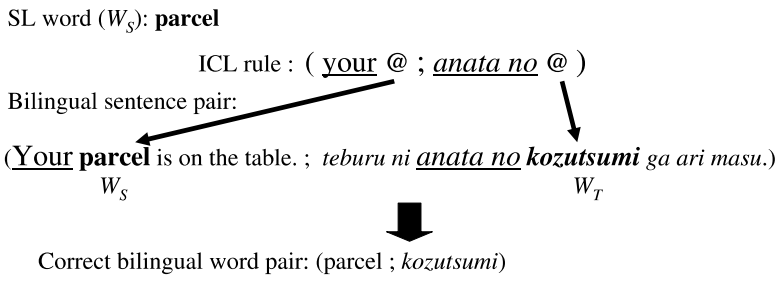


Fig. 2. An example of extraction of bilingual word pair using our idea.

### 1.3. Related work

Several methods based on co-occurrence words are proposed (Fung, 1995, 1998; Kaji & Aizono, 1996; Rapp, 1995, 1999; Tanaka & Iwasaki, 1996). These methods use the co-occurrence words to bilingual word pairs in bilingual sentence pairs. However, they depend on a pre-existing bilingual dictionary. Therefore, it is difficult to extract bilingual word pairs with various languages when a sufficient bilingual dictionary does not exist beforehand.

Methods based on large-scale bilingual resources are also proposed. For instance, Kumano and Hirakawa (1994) proposed a method that utilizes both statistical information and linguistic information to obtain corresponding words or phrases in a parallel corpus (Kumano & Hirakawa, 1994). To cite another example, Utsuro, Hino, and Kida (2004) proposed a method that acquires low-frequency bilingual terms using bilingual dictionaries and machine translation systems to measure similarity (Utsuro et al., 2004). However, it is difficult for that system to deal with various languages because of the use of large-scale bilingual resources.

Regarding methods for acquisition templates, Güvenir and Cicekli (1998) and McTait (2001) proposed methods that acquire templates by replacing common parts with variables or replacing different parts with variables in bilingual sentence pairs (Güvenir & Cicekli, 1998; McTait, 2001). However, such methods require many similar bilingual sentence pairs to extract sufficient templates because they do not focus on local parts in bilingual sentence pairs.

The method based on frequency of bilingual word pairs, K-vec (Fung & Church, 1994; Pedersen & Varma, 2003), is unable to extract low-frequency bilingual word pairs. That method is based on the fact that two words occur an almost equal number of times if they are translations of each other. However, the algorithm is applied only to bilingual word pairs for which the frequency is greater than three, except one and two.

Moreover, statistical word-alignment methods (Ahrenberg, Andersson, & Merkel, 1998; Brown, Della Pietra, Della Pietra, & Mercer, 1993; Dagan, Church, & Gale, 1993; Hiemstra, de Jong, & Kraaij, 1997; Macklovitch & Hannan, 1996; Nießen & Ney, 2004; Och & Ney, 2003; Vogel, Ney, & Tillmann, 1996) have been proposed, but they are also insufficient. That is, they cannot also extract low-frequency bilingual word pairs even though they are language-independent. Yamada and Knight (2001) proposed a phrase-based alignment method (Yamada & Knight, 2001). This method requires syntactical analysis systems for both SL sentences and TL sentences. Consequently, it is difficult to deal with languages for which syntactical analysis systems are not sufficiently obtainable.

Sentence-alignment methods based on word-alignment have also been proposed. Chen (1993) proposed an algorithm for sentence alignment that uses lexical information (Chen, 1993). Gale and Church (1993) proposed a method for aligning sentences based on a simple statistical model of character lengths (Gale & Church, 1993). Zhao, Zechner, Vogel, and Waibel (2003) proposed a method for automatically optimizing the alignment scores of such a bilingual sentence alignment program (Zhao et al., 2003). This method is based on the statistical translation models presented by Brown et al. (1993). However, these methods, which are based on the appearance frequency of words, cannot extract low-frequency bilingual word pairs because of the data sparseness problem. Moreover, Matsumoto, Ishimoto, and Utsuro (1993) Collier, Ono, and Hirakawa (1998) proposed a method that can extract bilingual word pairs from parallel corpora with languages for which the grammatical structures of SL differ from the grammatical structures of TL (Matsumoto et al., 1993; Collier et al., 1998). These methods depend on a pre-existing bilingual dictionary. Therefore, it is difficult to obtain bilingual word pairs with various languages when the bilingual resource is insufficient.

For information extraction from semi-structured documents such as HTML and XML, many systems based on templates (Appelt, Hobbs, Bear, Israel, & Tyson, 1993; Hsu & Yih, 1997; Kushmerick, Weld, & Doorenbos, 1997; Lee & Bui, 2000) have been examined. For acquisition of templates, the template detection algorithms using frequency information (Bar-Yossef & Rajagopalan, 2002; Hirokawa, Itoh, & Miyahara, 2003; Yamada, Ikeda, & Hirokawa, 2002) have been put forth. Crescenzi, Mecca, and Merialdo (2001) proposed a method that generates a wrapper using similarities and differences in HTML pages (Crescenzi et al., 2001). However, in web sites, it is comparatively easy to generate templates because HTML and XML are semi-structured documents. In contrast, natural language sentences include various linguistic phenomena. Furthermore, the system must determine word correspondence among different languages in the extraction

of bilingual word pairs. Our method is very effective because it can automatically acquire ICL rules as templates for natural language sentences that are not semi-structured documents.

Our system can extract bilingual word pairs with various frequencies without requiring any bilingual resource beforehand. It is able to do so through the use of ICL rules that are acquired automatically only from a parallel corpus.

## 2. Outline

Fig. 3 shows an outline of a system using ICL. In Fig. 3, ICL is shown as its four constituent processes: the process based on two bilingual sentence pairs; the process based on SL words; the process based on ICL rules; and the determination process of bilingual word pairs. These processes must be executed sequentially even though each process is independent.

First, the system acquires ICL rules by performing the process based on two bilingual sentence pairs. This process uses only a parallel corpus, and it is executed only one time to a parallel corpus. Therefore, the system performs this process first. Similarity values in all acquired ICL rules are assigned using the cosine function (1), and the acquired ICL rules are registered to an ICL rule dictionary. Then the user inputs SL words of bilingual word pairs. In the process based on SL words, the system obtains ICL rules and bilingual word pairs for SL words using bilingual sentence pairs for which SL words exist. This process uses SL words and a parallel corpus. It is executed every time SL words are inputted. Similarity values between SL words and TL words in all extracted bilingual word pairs are also assigned using the cosine function (1). The extracted bilingual word pairs are used as input data in the determination process of bilingual word pairs. The acquired ICL rules are registered to the ICL rule dictionary. All ICL rules are acquired in the process based on two bilingual sentence pairs and the process based on SL words. The system must acquire ICL rules to the greatest extent possible because it extracts bilingual word pairs using the acquired ICL rules. The process based on two bilingual sentence pairs is effective for acquiring many ICL rules because ICL rules are acquired from various bilingual sentence pairs in a parallel corpus. In contrast, in the process based on SL words, ICL rules are acquired only from the bilingual sentence pairs for which SL words exist.

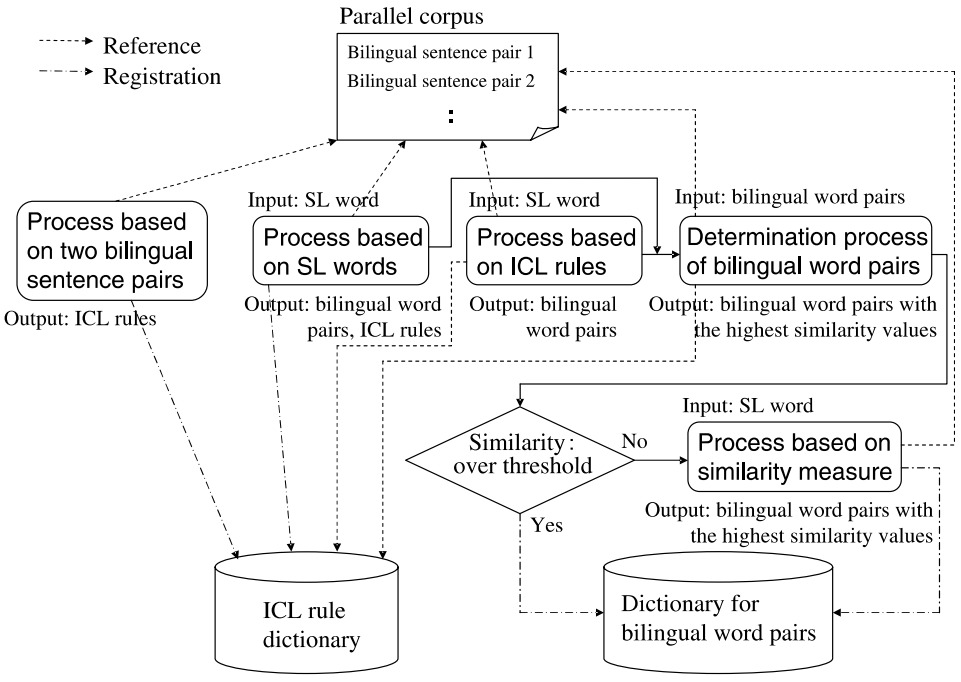


Fig. 3. Outline of a system using ICL.

In the process based on ICL rules, using the acquired ICL rules, the system extracts bilingual word pairs for SL words from bilingual sentence pairs for which SL words exist. This process uses SL words, a parallel corpus and ICL rules, and it is executed every time SL words are inputted. The extracted bilingual word pairs are also used as input data in the determination process of bilingual word pairs. In the determination process of bilingual word pairs, the system determines the most suitable bilingual word pairs among the bilingual word pairs extracted by the process based on SL words and the process based on ICL rules. In that case, the system selects the bilingual word pairs with the highest similarity values as the most suitable bilingual word pairs. This process uses the extracted bilingual word pairs, ICL rules and a parallel corpus. It is executed every time the bilingual word pairs for SL words are obtained in the process based on SL words and the process based on ICL rules.

Moreover, the system compares the similarity values of the bilingual word pairs chosen by the determination process of bilingual word pairs with a threshold value. Consequently, the system registers the chosen bilingual word pairs to the dictionary for bilingual word pairs when their respective similarity values are greater than the threshold value. On the other hand, when the similarity values of the chosen bilingual word pairs are not greater than the threshold or when no bilingual word pairs are obtained in the above ICL, i.e., when ICL cannot acquire sufficient ICL rules, the system extracts bilingual word pairs from all bilingual sentence pairs for which SL words exist using one similarity measure by the process based on similarity measure. This process uses SL words and a parallel corpus. It is executed every time SL words are inputted. In this study, we use the cosine, Dice coefficient (Manning & Schütze, 1999), Log-Likelihood Ratio (LLR) (Dunning, 1993), and Yates' $\chi^2$ (Hisamitsu & Niwa, 2001) as similarity measures. In this process, the bilingual word pairs with the highest similarity values are chosen. They are then registered to the dictionary for bilingual word pairs.

## 3. Processes

### 3.1. Process based on two bilingual sentence pairs

The system first performs the process based on two bilingual sentence pairs to acquire many ICL rules. This process requires only a parallel corpus; it is performed only one time to the parallel corpus. In this process, the system obtains ICL rules using common parts and different parts between two bilingual sentence pairs in a parallel corpus. This fact indicates that the system does not require bilingual resources to acquire ICL rules. Therefore, the system using ICL can extract bilingual word pairs between various languages. The determination of common parts and different parts is based on character strings of bilingual sentence pairs. Therefore, it is simple word matching. The use of such a technique realizes a learning system that can acquire knowledge from various data dynamically without using pre-existing statistical knowledge. The system possesses high learning ability when it can extract bilingual word pairs of various languages without requiring any bilingual resources. The ICL is the method that imparts learning ability to the system. The details of the process based on two bilingual sentence pairs are the following:

P1-(1) The system selects two bilingual sentence pairs that have SLCPs and TLCPs from a parallel corpus.[2]
P1-(2) The system determines SLDPs using SLCPs, and extracts only SLDPs for which the number of words is less than three[3] from SL sentences of two selected bilingual sentence pairs.

[2] In this paper, the respective common parts between SL sentences of two bilingual sentence pairs are called $SLCP_{1,\ldots,NSLCP}$; the respective common parts between TL sentences of two bilingual sentence pairs are called $TLCP_{1,\ldots,NTLCP}$. Here, NSLCP is the number of SLCPs; NTLCP is the number of TLCPs. On the other hand, the respective different parts between SL sentences of two bilingual sentence pairs are called $SLDP^{i=1,2}_{m=1,2,3,\ldots}$; the respective different parts between TL sentences of two bilingual sentence pairs are called $TLDP^{i=1,2}_{m=1,2,3,\ldots}$.

[3] We cannot use part-of-speech (POS) data when morphological analysis systems do not exist (e.g., an Ainu-language morphological analysis system has not been developed). In that case, the system extracts SLDPs using the number of words in this study. That is, SLDPs with numerous words are unsuitable as words. In this study, the number of SL words in all bilingual word pairs that were used as experimental data was less than 3.

P1-(3) The system determines TLDPs using TLCPs, and extracts only TLDPs that correspond to independent words[4] (i.e., noun, verb, adjective, adverb, or conjunction) from TL sentences of two selected bilingual sentence pairs.

P1-(4) The system acquires ICL rules using extracted SLDPs and TLDPs only when the number of extracted SLDPs is equal to the number of extracted TLDPs in each bilingual sentence pair;[5] it returns to P1-(1) to acquire other ICL rules.

The system determines SLCPs, SLDPs, TLCPs and TLDPs using only character strings of two bilingual sentences without part-of-speech information or syntactical information. Therefore, their units (i.e., SLCPs, SLDPs, TLCPs and TLDPs) are various. In SLCPs and TLCPs, the minimum unit is a word. When the common words appear continuously between two bilingual sentence pairs, the unit becomes a word string. In SLDPs and TLDPs, the unit of most extracted SLDPs is a word because the system extracts only SLDPs for which the number of words is less than 3 by P1-(2). Also, the unit of all extracted TLDPs is a word because the system extracts only TLDPs that correspond to independent words by P1-(3).

Furthermore, in P1-(1), the selection process of two bilingual sentence pairs that have SLCPs is performed using combinations of all bilingual sentence pairs in a parallel corpus. However, the selection process of two bilingual sentence pairs that have TLCPs is performed using the combination of only bilingual sentence pairs that have SLCPs, not the combination of all bilingual sentence pairs. Therefore, the comparative process for TL sentences is limited. Moreover, the process based on two bilingual sentence pairs that includes P1-(1) is executed to a parallel corpus only one time, not every time SL words are inputted, as in other processes shown as examples in Fig. 3.

Fig. 4 gives an example of extraction of SLDPs and TLDPs in the process based on two bilingual sentence pairs. First, the system selects two bilingual sentence pairs that have SLCPs (the common parts between two SL sentences) and TLCPs (the common parts between two TL sentences) by P1-(1). As shown in Fig. 4, the system selects bilingual sentence pairs 1 and 2 from a parallel corpus because "This" and "is" exist in two SL sentences, and "*kono*" and "*wa*" exist in two TL sentences. Therefore, "This" and "is" respectively become $SLCP_1$ and $SLCP_2$ in SL sentences. In TL sentences, "*kono*" and "*wa*" become $TLCP_1$ and $TLCP_2$, respectively. The system then determines SLDPs (the different parts between two SL sentences) using SLCPs, and extracts SLDPs from two SL sentences by P1-(2). In the SL sentence of bilingual sentence pair 1 of Fig. 4, "room" and "the room for my children to study" are determined as $SLDP_1^1$ and $SLDP_2^1$, respectively, because "room" exists between $SLCP_1$ ("This") and $SLCP_2$ ("is"), and "the room for my children to study" adjoins the right side of $SLCP_2$ ("is"). In the SL sentence of bilingual sentence pair 2, "game" and "very popular in England" are determined as $SLDP_1^2$ and $SLDP_2^2$, respectively, because "game" exists between $SLCP_1$ ("This") and $SLCP_2$ ("is"), and "very popular in England" adjoins the right side of $SLCP_2$ ("is"). However, only $SLDP_1^1$ ("room") and $SLDP_1^2$ ("game"), for which the number of words is less than 3, are extracted as SLDPs from the SL sentences of bilingual sentence pairs 1 and 2. The $SLDP_2^1$ ("the room for my children to study") and $SLDP_2^2$ ("very popular in England") are not extracted as SLDPs because the respective numbers of words in them are greater than 3.

Fig. 5 portrays the extraction algorithm of SLDPs. The processes of lines 4 and 10 of Fig. 5 represent the extraction process of SLDPs that adjoin the left side of $SLCP_1$. The processes of lines 11 and 24 represent the extraction process of SLDPs between two SLCPs. In lines 13 and 15, ${}_{NSLCP}C_2$ indicates $\frac{NSLCP!}{2!(NSLCP-2)!}$. That is, it obtains the number of combinations based on two SLCPs. Here, NSLCP is the number of SLCPs. Moreover, the processes of lines 25 and 29 represent the extraction process of SLDPs that adjoin the right side of $SLCP_{NSLCP}$. In that case, the number of words in all extracted SLDPs must be less than 3.

Moreover, the system extracts TLDPs (the different parts between two TL sentences) from two TL sentences by P1-(3). The system determines TLDPs using TLCPs same as the extraction process of SLDPs. In the TL sentence of bilingual sentence pair 1 of Fig. 4, "*heya*" and "*kodomo tachi no benkyo beya desu*" are

[4] In this paper, Japanese is used as TL. Therefore, POS information is obtained using the Japanese morphological analysis system "ChaSen" (Matsumoto et al., 2000).

[5] When the number of extracted SLDPs is not equal to the number of extracted TLDPs, this fact indicates that the system cannot determine the word correspondence in SL sentences and TL sentences of the bilingual sentence pairs. Therefore, in that case, the system does not acquire ICL rules. Such a condition is effective to prevent the acquisition of erroneous ICL rules, even though it is still incomplete.

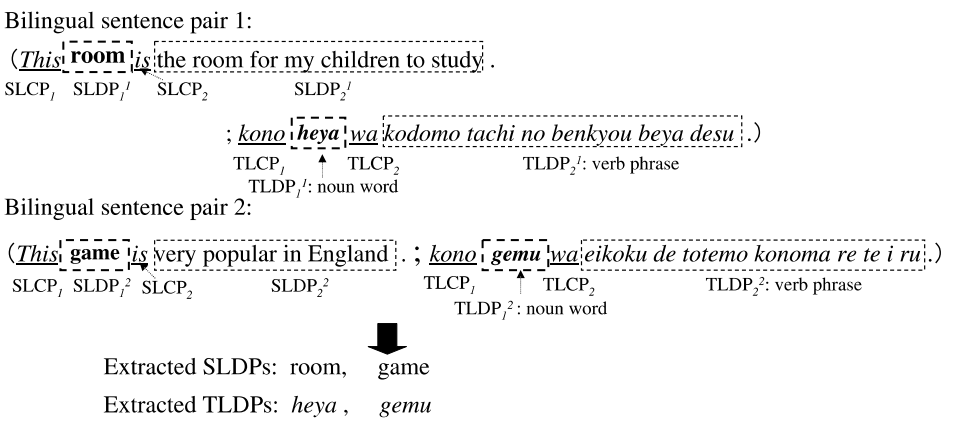


Fig. 4. An example of extraction of SLDPs and TLDPs in the process based on two bilingual sentence pairs.

```
1:  Input: Two SL sentences that have common parts
2:    i = 1
3:    while i ≦ 2
4:        m = 1
5:        if SLDP_m^i adjoins the left side of SLCP_1 then
6:            if Number of words in SLDP_m^i ≦ 3 then
7:                Extraction of SLDP_m^i (i.e., the part from word at the beginning
                  of SL sentence to word that adjoins the left side of SLCP_1)
8:            end
9:            m = m + 1
10:       end
11:       if NSLCP ≧ 2 then
12:           n = 1
13:           while n < NSLCP C2
14:               s = n + 1
15:               while s ≦ NSLCP C2
16:                   if Number of words in SLDP_m^i ≦ 3 then
17:                       Extraction of SLDP_m^i ( i.e., the part between SLCP_n and SLCP_s)
18:                   end
19:                   s = s + 1
20:                   m = m + 1
21:               end
22:               n = n + 1
23:           end
24:       end
25:       if SLDP_m^i adjoins the right side of SLCP_NSLCP then
26:           if Number of words in SLDP_m^i ≦ 3 then
27:               Extraction of SLDP_m^i (i.e., the part from word that adjoins the
                  right side of SLCP_NSLCP to word at the end of SL sentence)
28:           end
29:       end
30:       i = i + 1
31:   end
32: Output: SLDPs for which number of words are under 3
```

Fig. 5. The extraction algorithm of SLDPs in the process based on two bilingual sentence pairs.

determined as $TLDP_1^1$ and $TLDP_2^1$, respectively. In the TL sentence of bilingual sentence pair 2, "*gemu*" and "*eikoku de totemo konoma re te i ru*" are determined as $TLDP_1^2$ and $TLDP_2^2$, respectively. However, $TLDP_2^1$ ("*kodomo tachi no benkyo beya desu*") and $TLDP_2^2$ ("*eikoku de totemo konoma re te i ru*") correspond to verb phrases, not independent words. Therefore, only $TLDP_1^1$ ("*heya*") and $TLDP_1^2$ ("*gemu*"), which correspond to nouns, are extracted as TLDPs from the TL sentences of bilingual sentence pairs 1 and 2.

Consequently, in bilingual sentence pair 1, the number of extracted SLDPs is equal to the number of extracted TLDPs because the number of extracted SLDPs is 1 (“room”) and the number of extracted TLDPs is also 1 (“*heya*”). Therefore, the system performs the acquisition process of ICL rules to bilingual sentence pair 1 in the following process of P2. In bilingual sentence pair 2, the number of extracted SLDPs is equal to the number of extracted TLDPs because the number of extracted SLDPs is 1 (“game”) and the number of extracted TLDPs is also 1 (“*gemu*”). Therefore, the system performs the acquisition process of ICL rules to bilingual sentence pair 2.

Fig. 6 gives the extraction algorithm of TLDPs. The processes of lines 4 and 10 of Fig. 6 represent the extraction process of TLDPs that adjoins the left side of $TLCP_1$. The processes of lines 11 and 24 represent the extraction process of TLDPs between two TLCPs. Moreover, the processes of lines 25 and 29 represent the extraction process of TLDPs that adjoin the right side of $TLCP_{NTLCP}$. In that case, all extracted TLDPs must correspond to independent words.

The system then generates ICL rules using SLCPs, TLCPs, SLDPs and TLDPs by P1-(4). Details of generation processes of ICL rules are the following:

P2-(1) The system replaces the extracted SLDPs and TLDPs with variables in bilingual sentence pairs.
P2-(2) The system extracts the pairs of each SLCP and variable, and the pairs of each TLCP and variable.
P2-(3) The system generates ICL rules by combining the extracted pairs of SLCPs and variables with the extracted pairs of TLCPs and variables.
P2-(4) The system calculates the similarity values between SLCPs and TLCPs in the generated ICL rules. Here, the cosine function (1) is used with all bilingual sentence pairs in a parallel corpus; it registers ICL rules to the ICL rule dictionary.

```
1:  Input: Two TL sentences that have common parts
2:    i = 1
3:    while i ≦ 2
4:        m = 1
5:        if TLDP_m^i adjoins the left side of TLCP_1 then
6:            if TLDP_m^i corresponds to independent word then
7:                Extraction of TLDP_m^i (i.e., the part from word at the beginning
                  of TL sentence to word that adjoins the left side of TLCP_1)
8:            end
9:            m = m + 1
10:       end
11:       if NTLCP ≧ 2 then
12:           n = 1
13:           while n < NTLCP C2
14:               s = n + 1
15:               while s ≦ NTLCP C2
16:                   if TLDP_m^i corresponds to independent word then
17:                       Extraction of TLDP_m^i (i.e., the part between TLCP_n and TLCP_s)
18:                   end
19:                   s = s + 1
20:                   m = m + 1
21:               end
22:               n = n + 1
23:           end
24:       end
25:       if TLDP_m^i adjoins the right side of TLCP_NTLCP then
26:           if TLDP_m^i corresponds to independent word then
27:               Extraction of TLDP_m^i (i.e., the part from word that adjoins the
                  right side of TLCP_NTLCP to word at the end of TL sentence)
28:           end
29:       end
30:       i = i + 1
31:   end
32: Output: TLDPs that correspond to independent words
```

Fig. 6. The extraction algorithm of TLDPs in the process based on two bilingual sentence pairs.

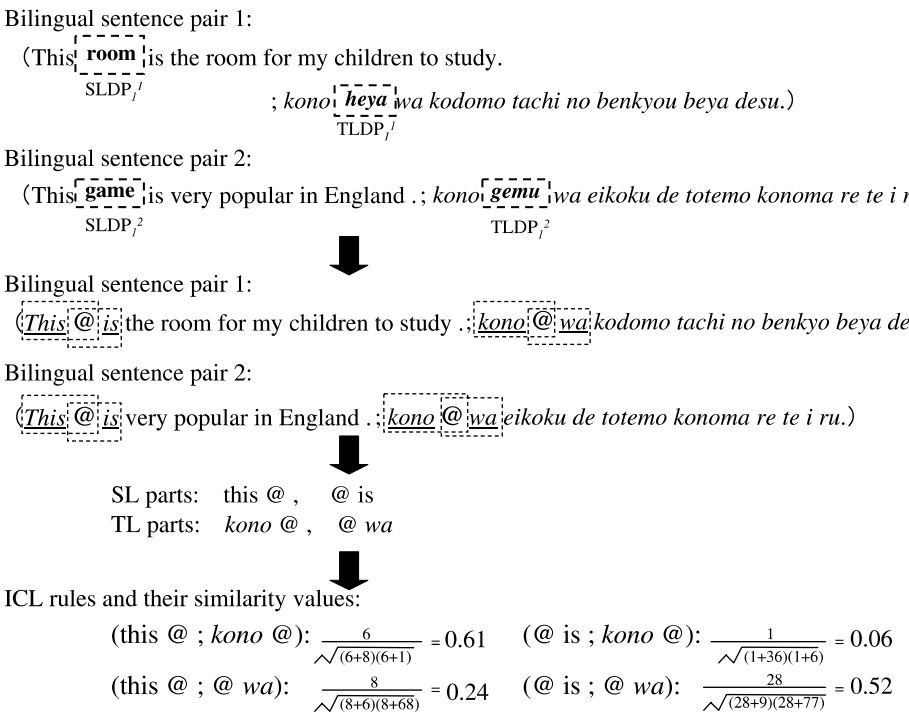


Fig. 7. An example of generation of ICL rules in the process based on two bilingual sentence pairs.

Fig. 7 gives generation examples of ICL rules using this process. In Fig. 7, the system generates ICL rules from bilingual sentence pairs 1 and 2 shown in Fig. 4. First, the system replaces the extracted SLDPs (''room'', ''game'') and TLDPs (''*heya*'', ''*gemu*'') with the variable ''@'' in bilingual sentence pairs 1 and 2 by P2-(1). Consequently, the system extracts the pairs of each SLCP and variable (''this @'', ''@ is''), and the pairs of each TLCP and variable (''*kono* @'', ''@ *wa*'') by P2-(2). Based on the combinations of all extracted pairs, (this @;*kono* @), (@ is;*kono* @), (this @;@ *wa*) and (@ is;@ *wa*) are generated as ICL rules by P2-(3). Moreover, the similarity values between SLCPs and TLCPs in the generated ICL rules are calculated using the cosine function (1) by P2-(4). The similarity values of (this @;*kono* @) and (@ is;@ *wa*) are higher than those of (@ is;*kono* @) and (this @;@ *wa*) because ''this'' corresponds to ''*kono*'', not ''*wa*'' in Japanese, and ''is'' corresponds to ''*wa*'', not ''*kono*'' in Japanese. In ICL rules, the parts extracted from SL sentences are called SL parts; the parts extracted from TL sentences are called TL parts.

### 3.2. Process based on SL words

The system performs the process based on SL words to acquire ICL rules from the bilingual sentence pairs for which SL words exist in a parallel corpus. At the same time, the system extracts the bilingual word pairs for SL words. This process requires SL words and a parallel corpus, and it is performed every time SL words are inputted. The details of the process based on SL words are the following:

P3-(1) The system selects bilingual sentence pairs for which SL words exist from a parallel corpus. Moreover, the system chooses the bilingual sentence pairs that have SLCPs (the common parts between two SL sentences) and TLCPs (the common parts between two TL sentences) as the bilingual sentence pairs with SL words. In that case, SLCPs must adjoin SL words in SL sentences.

P3-(2) The system determines $TLDP_{m=1,2,3,\ldots}$ (the different parts between two TL sentences) using TLCPs, and extracts only TLDPs that correspond to independent words from TL sentences of bilingual sentence pairs for which SL words exist.

P3-(3) The system obtains bilingual word pairs by combining SL words with the extracted TLDPs.
P3-(4) The system acquires ICL rules. The details of this process are the following:
(i) The system replaces SL words and the extracted TLDPs with variables in the bilingual sentence pairs for which SL words exist.
(ii) The system extracts the pairs of each SLCP and variable, and the pairs of each TLCP and variable.
(iii) The system generates ICL rules by combining the extracted pairs of SLCPs and variables with the extracted pairs of TLCPs and variables.
(iv) The system calculates the similarity values between SLCPs and TLCPs in the acquired ICL rules. Here, the cosine function (1) is used with all bilingual sentence pairs in a parallel corpus; it registers ICL rules to the ICL rule dictionary.

In P3-(1), the selection process of the bilingual sentence pairs for which SL words exist is performed by searching the bilingual sentence pairs for which SL words exist among all bilingual sentence pairs in the parallel corpus. The selection process of the bilingual sentence pairs that have SLCPs is performed using the combinations of the bilingual sentence pairs that SL words exist and other bilingual sentence pairs, not the combinations of all bilingual sentence pairs in the parallel corpus. Moreover, the selection process of bilingual sentence pairs that have TLCPs is performed using the combinations of the bilingual sentence pairs for which SL words exist and the bilingual sentence pairs that have SLCPs.

Fig. 8 gives an example of acquisition of ICL rules and bilingual word pairs in the process based on SL words. First, the system selects the bilingual sentence pairs for which SL words exist and the bilingual sentence

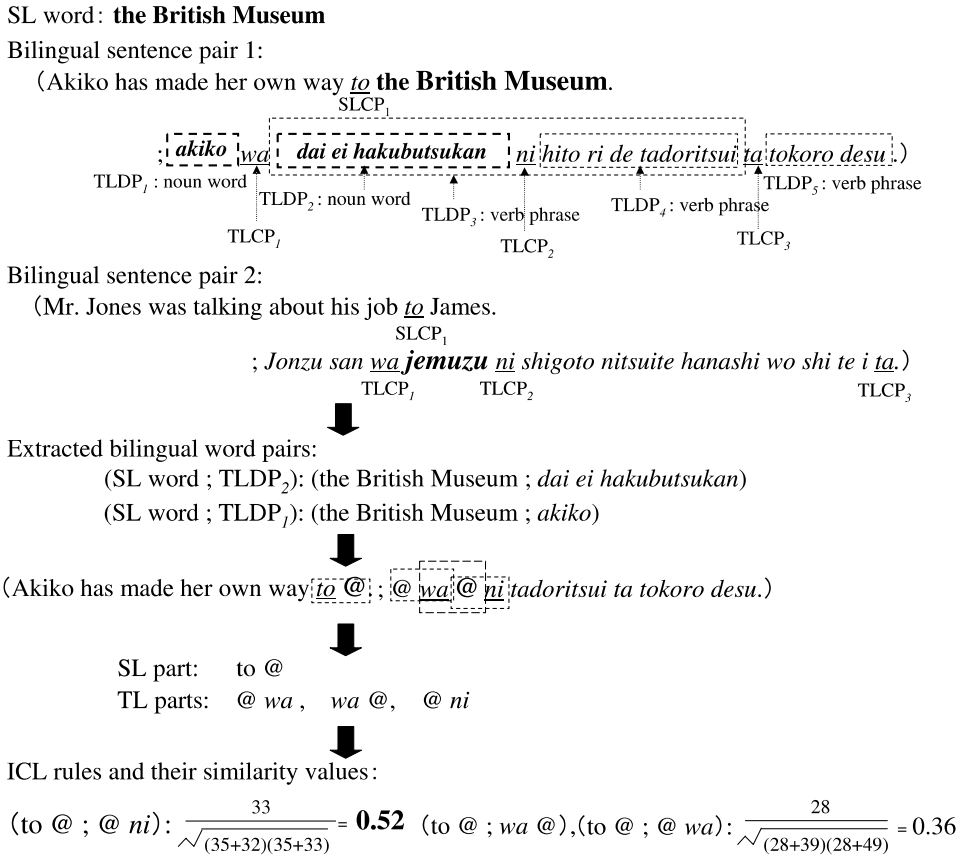


Fig. 8. An example of acquisition of ICL rules and bilingual word pairs in the process based on SL words.

pairs that have SLCPs and TLCPs as bilingual sentence pairs for which SL words exist by P3-(1). In Fig. 8, the system selects bilingual sentence pair 1 for which the SL word "the British Museum" exists. Moreover, the system chooses bilingual sentence pair 2 that has "to" as SLCP, and "*wa*", "*ni*" and "*ta*" as TLCPs. Therefore, "to" becomes $SLCP_1$ in SL sentences. In TL sentences, "*wa*", "*ni*" and "*ta*" become $TLCP_1$, $TLCP_2$ and $TLCP_3$, respectively. The system then determines TLDPs using TLCPs, and extracts TLDPs from the bilingual sentence pairs for which SL words exist by P3-(2). In the TL sentence of bilingual sentence pair 1 of Fig. 8, "*akiko*", "*dai ei hakubutsukan*", "*dai ei hakubutsukan ni hito ri de tadoritsui*", "*hito ri de tadoritsui*" and "*tokoro desu*" are determined as $TLDP_1$, $TLDP_2$, $TLDP_3$, $TLDP_4$ and $TLDP_5$, respectively, because "*akiko*" adjoins the left side of $TLCP_1$ ("*wa*"), "*dai ei hakubutsukan*" exists between $TLCP_1$ ("*wa*") and $TLCP_2$ ("*ni*"), "*dai ei hakubutsukan ni hito ri de tadoritsui*" exist between $TLCP_1$ ("*wa*") and $TLCP_3$ ("*ta*"), "*hito ri de tadoritsui*" exist between $TLCP_2$ ("*ni*") and $TLCP_3$ ("*ta*"), and "*tokoro desu*" adjoins the right side of $TLCP_3$ ("*ta*").

Fig. 9 gives the extraction algorithm of TLDPs in P3-(2). The processes of lines 2 and 8 of Fig. 9 represent the extraction process of TLDPs that adjoin the left side of $TLCP_1$. The processes of lines 9 and 22 represent the extraction process of TLDPs between two TLCPs. The processes of lines 23 and 27 represent the extraction process of TLDPs that adjoin the right side of $TLCP_{NTLCP}$. In that case, all extracted TLDPs must correspond to independent words.

Among the five TLDPs in Fig. 8, only $TLDP_1$ ("*akiko*") and $TLDP_2$ ("*dai ei hakubutsukan*") are extracted because they are independent words. The other three TLDPs are verb phrases, not independent words. The system obtains bilingual word pairs by combining SL words with the extracted TLDPs by P3-(3). In Fig. 8, (the British Museum;*dai ei hakubutsukan*) and (the British Museum;*akiko*) are obtained as the bilingual word pairs by combining the SL word ("the British Museum") with two extracted TLDPs ("*akiko*", "*dai ei hakubutsukan*"). Moreover, the system generates ICL rules by P3-(4). In Fig. 8, the system obtains (to

```
1:  Input: TL sentence of bilingual sentence pair for which SL word exists
2:      m = 1
3:      if TLDP_m adjoins the left side of TLCP_1 then
4:          If TLDP_m corresponds to independent word then
5:              Extraction of TLDP_m (i.e., the part from word at the beginning
                of TL sentence to word that adjoins the left side of TLCP_1)
6:          end
7:          m = m + 1
8:      end
9:      if NTLCP ≧ 2 then
10:         n = 1
11:         while n < NTLCP C2
12:             s = n + 1
13:             while s ≦ NTLCP C2
14:                 if TLDP_m corresponds to independent word then
15:                     Extraction of TLDP_m (i.e., the part between TLCP_n and TLCP_s)
16:                 end
17:                 s = s + 1
18:                 m = m + 1
19:             end
20:             n = n + 1
21:         end
22:     end
23:     if TLDP_m adjoins the right side of TLCP_NTLCP then
24:         if TLDP_m corresponds to independent word then
25:             Extraction of TLDP_m (i.e., the part from word that adjoins the
                right side of TLCP_NTLCP to word at the end of TL sentence)
26:         end
27:     end
28: Output: TLDPs that correspond to independent words
```

Fig. 9. The extraction algorithm of TLDPs in the process based on SL words.

@;@ *ni*), (to @;*wa* @) and (to @;@ *wa*) as ICL rules. The similarity value of (to @;@ *ni*) is higher than the similarity values of (to @;*wa* @) and (to @;@ *wa*) because "to" corresponds to "*ni*", not "*wa*" in Japanese.

### 3.3. Process based on ICL rules

In the process based on ICL rules, using the acquired ICL rules, the system extracts bilingual word pairs from bilingual sentence pairs in a parallel corpus. The system limits the search scope for determination of equivalents in TL sentences by focusing on the local parts of bilingual sentence pairs. Moreover, the system can efficiently extract bilingual word pairs using ICL rules because ICL rules have information about the word order between SL and TL. This process needs SL words, a parallel corpus and ICL rules, and it is performed every time SL words are inputted. In addition, this process is based on the position of variables in ICL rules.

Fig. 10 shows examples of extraction of bilingual word pairs in the process based on ICL rules. In example 1 of Fig. 10, (fish;*sakana*) is extracted as the noun bilingual word pair using (this @;*kono* @) acquired in Fig. 7. First, the system selects bilingual sentence pair 1 that the SL word 1 "fish" exists from a parallel corpus. Moreover, the system selects ICL rule 1 (this @;*kono* @) from the ICL rule dictionary. The SL part "this" of ICL rule 1 exists in the SL sentence of bilingual sentence pair 1. In addition, the variable "@" corresponds to SL word 1 "fish" because both variable "@" and the SL word 1 "fish" adjoin the right side of SLCP ("this"). Therefore, the system extracts "*sakana*", which adjoins the right side of TLCP ("*kono*"), from the TL sentence of bilingual sentence pair 1 as the independent word that corresponds to the variable "@" in the TL part of ICL rule 1. In this manner, the system can obtain (fish;*sakana*) as the noun bilingual word pair. It then calculates the similarity value between "fish" and "*sakana*." In example 2 of Fig. 10, (get;*noru*) and (get;*densha*) are extracted using ICL rule 2 (to @;@ *ni*), which was acquired in Fig. 8. In this example, the system extracts "*noru*" and "*densha*" that adjoin the left side of TLCP ("*ni*") from the TL sentence of bilingual sentence pair 2 as the independent words that correspond to variable "@" in the TL part of ICL rule 2.

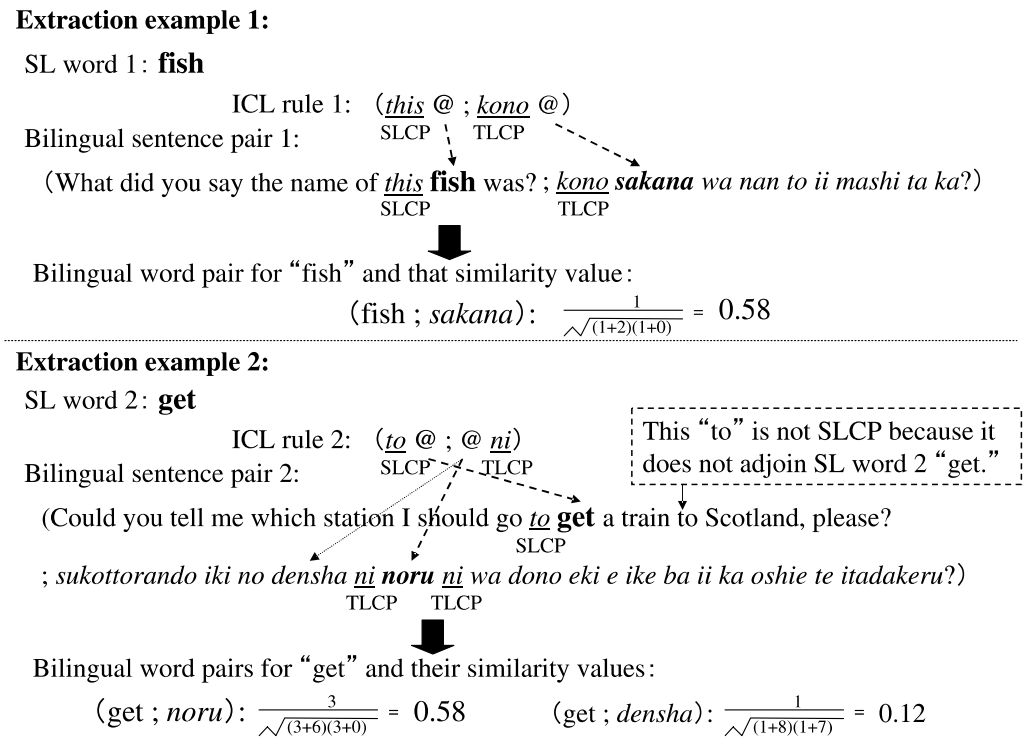


Fig. 10. Examples of extraction of bilingual word pairs in the process based on ICL rules.

```
1:  Input: SL word
2:      while Selection of bilingual sentence pair for which SL word exists, and selection
            of ICL rule that has SLCP and TLCP as the selected bilingual sentence pair
3:          if Variable adjoins the right side of TLCP in TL part of ICL rule then
4:              i = 0
5:              while i < NTLCP
6:                  Extraction of independent word that adjoins the right side of TLCP_i
                    in TL sentence
7:                  i = i + 1
8:              end
9:          end
10:         if Variable adjoins the left side of TLCP in TL part of ICL rule then
11:             i = 0
12:             while i < NTLCP
13:                 Extraction of independent word that adjoins the left side of TLCP_i
                    in TL sentence
14:                 i = i + 1
15:             end
16:         end
17:     end
18:     Extraction of bilingual word pairs by combining SL word with each TL word
19:     Calculation of similarity value between SL word and each extracted TL words. Here,
        the cosine function (1) is used with all bilingual sentence pairs in a parallel corpus
20: Output: Bilingual word pairs
```

Fig. 11. The extraction algorithm of bilingual word pairs in the process based on ICL rules.

Fig. 11 gives the extraction algorithm of bilingual word pairs using ICL rules. The processes of lines 3 and 9 of Fig. 11 represent the extraction process of bilingual word pairs using ICL rules for which the variables adjoin the right side of TLCP in TL parts. By this process, the system extracts (fish;*sakana*) using (this @;*kono* @) in Fig. 10. Also, the processes of lines 10 and 16 represent the extraction process of bilingual word pairs using ICL rules for which the variables adjoin the left side of TLCP in TL parts. By this process, the system extracts (get;*noru*) and (get;*densha*) using (to @;@ *ni*) in Fig. 10.

Moreover, the system calculates the similarity values between SL words and the independent words extracted from TL sentences using the cosine function (1) by line 19 of Fig. 11. In example 2 of Fig. 10, the similarity value between "get" and "*noru*" is higher than the similarity value between "get" and "*densha*" because "get" corresponds to "*densha*", not "*noru*" in Japanese. Thereby, only (get;*noru*) is selected as the bilingual word pair, and it is registered to the dictionary for bilingual word pairs. Details of the determination process for the most suitable bilingual word pairs using the similarity values are given in Section 3.4.

Using the acquired ICL rules, the system can decrease the number of candidates of equivalents for SL words. In example 2 of Fig. 10, only "*densha*" and "*noru*" become the candidates of equivalents for "get." In contrast, all independent words (i.e., "*sukottorando*", "*iki*", "*densha*", "*noru*", "*eki*", "*ike*", "*ii*", "*oshie*", and "*itadakeru*") in the TL sentence of bilingual sentence pair 2 become the candidates of equivalents for "get" when ICL rule 2 does not exist. This fact indicates that ICL is effective to solve the sparse data problem. Moreover, ICL rules have the knowledge to cope with the different word order between SL and TL in the local parts of bilingual sentence pairs. For example, in the SL part of ICL rule 2 (to @;@ *ni*), the variable "@" adjoins the right side of "to." In the TL part, the variable "@" adjoins the left side of "*ni*". Therefore, the system can extract bilingual word pairs from parallel corpora with various languages for which the grammatical structures of SL differ from the grammatical structures of TL.

### *3.4. Determination process of bilingual word pairs*

The system determines the most suitable bilingual word pairs according to their similarity values and the similarity values of ICL rules to all extracted bilingual word pairs. This process requires the extracted bilingual

word pairs, ICL rules and a parallel corpus. It is executed every time the bilingual word pairs for SL words are obtained in the process based on SL words and the process based on ICL rules. Details of this process are the following:

P4-(1) The system selects the bilingual word pairs with the highest similarity values among all extracted bilingual word pairs.
P4-(2) The system selects the bilingual word pairs that were extracted using ICL rules with the highest similarity values when several bilingual word pairs have identical similarity values.
P4-(3) The system selects the bilingual word pairs with the TL words that appear in a parallel corpus for the first time when it cannot determine only one bilingual word pair by P4-(1) and P4-(2).

### 3.5. Process based on similarity measure

In the process based on similarity measure, the system extracts bilingual word pairs for SL words from bilingual sentence pairs for which SL words exist using each similarity measure without using ICL only when the similarity values are not greater than the threshold value or when no bilingual word pairs are obtained. That is, this process is needed when sufficient ICL rules cannot be acquired. Also, this process requires SL words and a parallel corpus; it is executed every time SL words are inputted. In this paper, the cosine, Dice coefficient, LLR and Yates' $\chi^2$ are used respectively as similarity measures. Details of this process are the following:

P5-(1) The system selects the bilingual word pairs for which SL words exist from a parallel corpus.
P5-(2) The system obtains candidates of equivalents for SL words by extracting all independent words from TL sentences of the bilingual sentence pairs for which SL words exist.
P5-(3) The system calculates the similarity values between SL words and each extracted independent word. Here, each similarity measure (i.e., cosine, Dice coefficient, LLR or Yates' $\chi^2$) is used with all bilingual sentence pairs in a parallel corpus.
P5-(4) The system chooses the pairs of SL words and independent words, for which the similarity values are highest, as the most suitable bilingual word pairs.
P5-(5) The system selects the bilingual word pairs with the TL words that appear at the first time in a parallel corpus, when several bilingual word pairs for which the similarity values are the same are obtained by P5-(4).

In this process, the system can obtain bilingual word pairs for SL words whenever the independent words exist in TL sentences of bilingual sentence pairs. In that case, all independent words in TL sentences of bilingual sentence pairs for which SL words exist become candidates of equivalents. This fact indicates that it is difficult to solve the sparse data problem, as described in Section 1.1.

In this study, the cosine, Dice coefficient, LLR or Yates' $\chi^2$ is used as a similarity measure in the process based on similarity measure. The cosine, Dice coefficient are measures based on comparison of two vectors. The LLR and Yates' $\chi^2$ are measures based on comparison of two probability values. Measures based on a comparison of two vectors indicate the ratio of two probabilities: the probability that both $W_S$ and $W_T$ occur, and the probability that $W_S$ or $W_T$ occurs along with the other. In this process, $W_S$ and $W_T$ correspond to the SL word and each independent word, respectively. The cosine was defined already as function (1) in Section 1.1. The Dice coefficient is defined as

$$\mathrm{Dice}(W_S, W_T) = \frac{2a}{(a+b)+(a+c)} \tag{2}$$

Definitions of all parameters $(a, b, c)$ in function (2) are identical to those given Table 1. The reliability of word pairs is high when the score is large for both the cosine and Dice coefficient.

Measures based on comparison of two probability values indicate whether the appearance of $W_S$ and $W_T$ is independent or dependent. The LLR is defined as the following.

$$\mathrm{LLR}(W_S, W_T) = a\log\frac{an}{(a+b)(a+c)} + b\log\frac{bn}{(a+b)(b+d)} + c\log\frac{cn}{(c+d)(a+c)} + d\log\frac{dn}{(c+d)(b+d)} \quad (3)$$

The LLR represents the ratio of two probabilities: the probability that $W_S$ depends on $W_T$, and the probability that $W_S$ and $W_T$ are independent of each other. The LLR score becomes 0.0 when $W_S$ and $W_T$ are completely independent. Another measure, the $\chi^2$ statistic, indicates the differences between observed and expected values. The Yates' $\chi^2$ is the measure that the expected value is revised with when the expected frequencies are low. The Yates' $\chi^2$ is defined as the following.

$$\mathrm{Yates}'\chi^2(W_S, W_T) = \frac{n(|ad-bc| - \frac{n}{2})^2}{(a+b)(c+d)(a+c)(b+d)} \quad (4)$$

In functions (3) and (4), the definitions of all parameters $(a, b, c, d, n)$ are the same as those of Table 1.

## 4. Experiments for performance evaluation

### 4.1. Experimental procedure

Five kinds of parallel corpora are used in this study as experimental data. These parallel corpora are for English–Japanese, French–Japanese, German–Japanese, Shanghai-Chinese–Japanese and Ainu[6]–Japanese. They were taken from textbooks (Chikushi, 2001; Emoto & Han, 2004; Harukawa & Snelling, 1998; Nakagawa & Nakamoto, 2004; Oshio, 2004). The number of bilingual sentence pairs was 1794; the average numbers of words in SL sentences and TL sentences were 6.8 and 8.8, respectively. We inputted all 1081 SL words in five parallel corpora into eight systems: the system based on cosine, the system based on cosine in which ICL is used (herein, we call it the system based on cosine + ICL), the system based on Dice coefficient, the system based on Dice coefficient in which ICL is used (herein, we call it the system based on Dice + ICL), the system based on LLR, the system based on LLR in which ICL is used (herein, we call it the system based on LLR + ICL), the system based on Yates' $\chi^2$, and the system based on Yates' $\chi^2$ in which ICL is used (herein, we call it the system based on Yates + ICL). We applied ICL to various similarity measures to confirm the effectiveness of ICL. The initial conditions of all dictionaries are empty. Experiments were repeated for each parallel corpus using each system. In the system using ICL, 0.5 is adopted as its best threshold.[7] Moreover, the systems based only on similarity measure (i.e., the systems based on cosine, Dice coefficient, LLR and Yates' $\chi^2$) correspond to the systems that only process based on similarity measure as described in Section 3.5 is used.

### 4.2. Evaluation standard

We evaluated whether correct bilingual word pairs are obtained or not, and calculated the recall for all 1081 SL words using function (5).

$$\text{Recall } (\%) = \frac{\text{Number correctly extracted bilingual word pairs}}{\text{Number of all correct bilingual word pairs}} \times 100 \quad (5)$$

By function (5), the number of all correct bilingual word pairs in parallel corpora represents the sum of all nouns, verbs, adjectives, adverbs and conjunctions in the parallel corpora: 1081.

[6] Ainu language is spoken by members of the Ainu ethnic group, which originated in northern Japan and Sakhalin. That language is expressed using alphabet characters based on Japanese Katakana because it is a non-character language. Ainu language is independent, but similar to, Japanese and Korean.

[7] This value was obtained through preliminary experiments. Some correct bilingual word pairs are evaluated as erroneous bilingual word pairs when the system using ICL uses a high value as a threshold. In contrast, some erroneous bilingual word pairs are evaluated as correct bilingual word pairs when the system using ICL uses a low value as threshold. Therefore, 0.5, the middle value, became a most suitable threshold.

### 4.3. Experimental results

Table 2 shows experimental results in measures based on comparison of two vectors. In Table 2, all TL in every parallel corpus are Japanese. The recall of the system based on cosine + ICL was more than 5.9% points higher than that of the system based on cosine. Moreover, the recall of the system based on Dice + ICL was more than 7.6% points higher than that of the system based on Dice coefficient. These results indicate that ICL is effective for measures based on comparison of two vectors.

Table 3 shows experimental results for measures based on comparison of two probability values. The recall values of the systems using ICL were more than 4.6% points higher than those of the systems based on LLR and Yates' $\chi^2$. These results indicate that ICL is also effective for measures based on comparison of two probability values.

On the other hand, in our method, the precision values by function (6) are unobtainable because the number of all extracted bilingual word pairs in the dictionary for bilingual word pairs means only the number of bilingual word pairs that have the highest similarity values, not all bilingual word pairs extracted by the system using ICL in Fig. 3. The system chooses only one bilingual word pair ranked the highest among all ranked bilingual word pairs when several candidates of bilingual word pairs are obtained for each SL word. It then registers the chosen bilingual word pair to the dictionary. Therefore, the number of all extracted bilingual word pairs in the dictionary corresponds only to the number of bilingual word pairs ranked the highest. This

Table 2
Results of evaluation experiments in measures based on comparison of two vectors

| SL | Cosine (baseline) (%) | Cosine + ICL (%) | Dice coefficient (%) | Dice + ICL (%) | Number of correct bilingual word pairs |
|---|---|---|---|---|---|
| English | 50.3 | 56.8 | 50.3 | 58.0 | 169 |
| French | 49.6 | 54.6 | 49.6 | 57.1 | 240 |
| German | 54.4 | 61.5 | 54.4 | 62.6 | 195 |
| Shanghai-Chinese | 54.9 | 60.6 | 56.1 | 63.6 | 264 |
| Ainu | 52.6 | 58.2 | 52.1 | 59.2 | 213 |
| Total | 52.5 | 58.4 | 52.6 | 60.2 | 1081 |

Table 3
Results of evaluation experiments in measures based on comparison of two probability values

| SL | LLR (%) | LLR + ICL (%) | Yates' $\chi^2$ (%) | Yates + ICL (%) | Number of correct bilingual word pairs |
|---|---|---|---|---|---|
| English | 52.7 | 57.4 | 53.8 | 58.0 | 169 |
| French | 54.6 | 57.1 | 55.4 | 56.3 | 240 |
| German | 54.4 | 60.5 | 53.3 | 60.0 | 195 |
| Shanghai-Chinese | 57.6 | 62.9 | 57.6 | 63.6 | 264 |
| Ainu | 53.1 | 57.7 | 52.1 | 57.7 | 213 |
| Total | 54.7 | 59.3 | 54.7 | 59.3 | 1081 |

Table 4
Details of recalls in measures based on comparison of two vectors

| SL | Cosine | | Cosine + ICL | | Dice coefficient | | Dice + ICL | |
|---|---|---|---|---|---|---|---|---|
| | 1 (%) | Others (%) | 1 (%) | Others (%) | 1 (%) | Others (%) | 1 (%) | Others (%) |
| English | 35.7 | 78.9 | 43.8 | 82.5 | 34.8 | 80.7 | 44.6 | 84.2 |
| French | 38.6 | 79.7 | 45.5 | 79.7 | 38.6 | 79.7 | 47.2 | 84.4 |
| German | 35.3 | 84.2 | 47.9 | 82.9 | 35.3 | 84.2 | 49.6 | 82.9 |
| Shanghai-Chinese | 40.0 | 79.8 | 48.5 | 80.8 | 41.8 | 79.8 | 50.3 | 85.9 |
| Ainu | 41.3 | 64.4 | 50.5 | 66.3 | 40.4 | 64.4 | 52.3 | 66.3 |
| Total | 38.3 | 76.5 | 47.1 | 77.5 | 38.5 | 76.8 | 48.8 | 79.8 |

Table 5
Examples of bilingual word pairs extracted using ICL

| SL | Correct bilingual word pairs | Erroneous bilingual word pairs | |
|---|---|---|---|
| | | Bilingual word pairs | Equivalents |
| English | (coordinate; *toukatsu suru*) 1.0<br>(post office; *yubin kyoku*) 1.0 | (only; *hidarigawa* [left]) 1.0<br>(interesting; *soto* [outside]) 0.71 | Only<br>Interesting |
| French | (femme; *tsuma* [wife]) 1.0<br>(entrez; *hairu* [enter]) 0.71 | (prendre; *soe* [granish]) 0.58<br>(trois; *kitte* [stamp]) 0.58 | Take<br>Three |
| German | (nämlich; *tsumari* [after all]) 0.71<br>(steht; *tat* [stand]) 0.58 | (südlichen; *noboru* [climb]) 1.0<br>(Neues; *shinbun* [newspaper]) 0.71 | South<br>New event |
| Shanghai-Chinese | (原諒; *yurusu* [permit]) 1.0<br>(名菜; *yumei ryouri* [popular dish]) 1.0 | (正好; *sugoshi* [spend]) 1.0<br>(安静; *benri* [ convenience]) 1.0 | Just<br>Quit |
| Ainu | (huraypa; arau [wash]) 1.0<br>(set; *nedoko* [bed]) 1.0 | (uturano; *iko* [go]) 1.0<br>(koterke; *kedo* [but]) 0.71 | Together<br>Spring |

fact indicates that the number of bilingual word pairs in the dictionary is less than the number of SL words inputted by the user in the system using ICL.

$$\text{Precision}(\%) = \frac{\text{Number of correctly extracted bilingual word pairs}}{\text{Number of all extracted bilingual word pairs}} \times 100 \quad (6)$$

### 4.4. Discussion

We investigated the ratio of correctly extracted bilingual word pairs for which the frequency was 1 to all correct bilingual word pairs for which the frequency was 1. In the systems without ICL, many bilingual word pairs for which the frequency was 1 were extracted as erroneous bilingual word pairs because of data sparseness problems, as described in Section 1.1. Therefore, the extraction of many correct bilingual word pairs for which the frequency is 1 indicates that ICL is effective to solve the sparse data problem. Table 4 shows the details of the ratios in respective frequencies of bilingual word pairs in measures based on comparison of two vectors. The ratio of correctly extracted bilingual word pairs for which the frequency is 1 improved 9.6% points, on average, using ICL. The ratio of correctly extracted bilingual word pairs for which the frequency is greater than 2 to all correct bilingual word pairs for which the frequency is greater than 2 improved 2.0% points on average. In measures based on comparisons of two probability values, the ratio of correctly extracted bilingual word pairs for which the frequency is 1 improved 7.0% points on average using ICL. The ratio of correctly extracted bilingual word pairs for which the frequency is greater than 2 improved 0.7% points on average. Therefore, we confirmed that ICL is effective to solve the sparse data problem because the system using ICL could extract many bilingual word pairs for which the frequency is 1, comparing the system without ICL.

Table 5 shows examples of bilingual word pairs extracted by ICL.[8] In Table 5, "[ ]" indicates equivalents in English. By examples in Table 5, we can confirm that the system using ICL can extract not only bilingual word pairs that the number of words is 1, but also bilingual word pairs that the number of words is greater than 2, e.g., (post office;*yubin kyoku*).

Table 6 shows the recall values by only ICL. This result means the experimental result of the system without process based on similarity measure. In Table 6, the recall is insufficient. In the Ainu–Japanese parallel corpus, the recalls are high among all parallel corpora, and the ratios of correct bilingual word pairs for which the frequency is 1 to all correct bilingual word pairs are low. In contrast, in English–Japanese and French–Japanese parallel corpus, the recalls are low and the ratios of correct bilingual word pairs for which the

[8] In Shanghai-Chinese sentences, the division for each word has been performed already in the textbook (Emoto & Han, 2004) even though Shanghai-Chinese is an agglutinative language.

Table 6
Recall values obtained using only ICL

| SL | Recall (%) | Ratio of correct bilingual word pairs for which frequency is 1 to all correct bilingual word pairs (%) |
|---|---|---|
| English | 33.7 | 66.3 |
| French | 35.0 | 73.3 |
| German | 40.0 | 61.0 |
| Shanghai-Chinese | 43.2 | 62.5 |
| Ainu | 46.9 | 51.2 |
| Total | 40.1 | 63.0 |

frequency is 1 are high. In ICL, the words or word strings for which the frequency is 1 are not used as ICL rules because the common parts, for which the frequency is greater than 2, are used as ICL rules. Therefore, it is difficult to acquire ICL rules when the ratio of correct bilingual word pairs for which the frequency is 1 is high. Furthermore, the system using ICL acquires not only correct ICL rules and bilingual word pairs but also erroneous ICL rules and bilingual word pairs. Fig. 12 gives examples of the acquisition of an erroneous ICL rule and the extraction of erroneous bilingual word pair. In the acquisition of ICL rules of Fig. 12, (@ to;@ *wo*) was acquired as an erroneous ICL rule because of the omission by the set phrase. In bilingual sentence pair 1, "to" corresponds to "*te*" in Japanese. However, in bilingual sentence pair 2, the equivalent of "to" is omitted in Japanese because "is going to" is the set phrase. As the result, the erroneous ICL rule (@ to;@ *wo*) was acquired, and (job;*hanashi*) was extracted as erroneous bilingual word pair from bilingual sentence pair 3 using

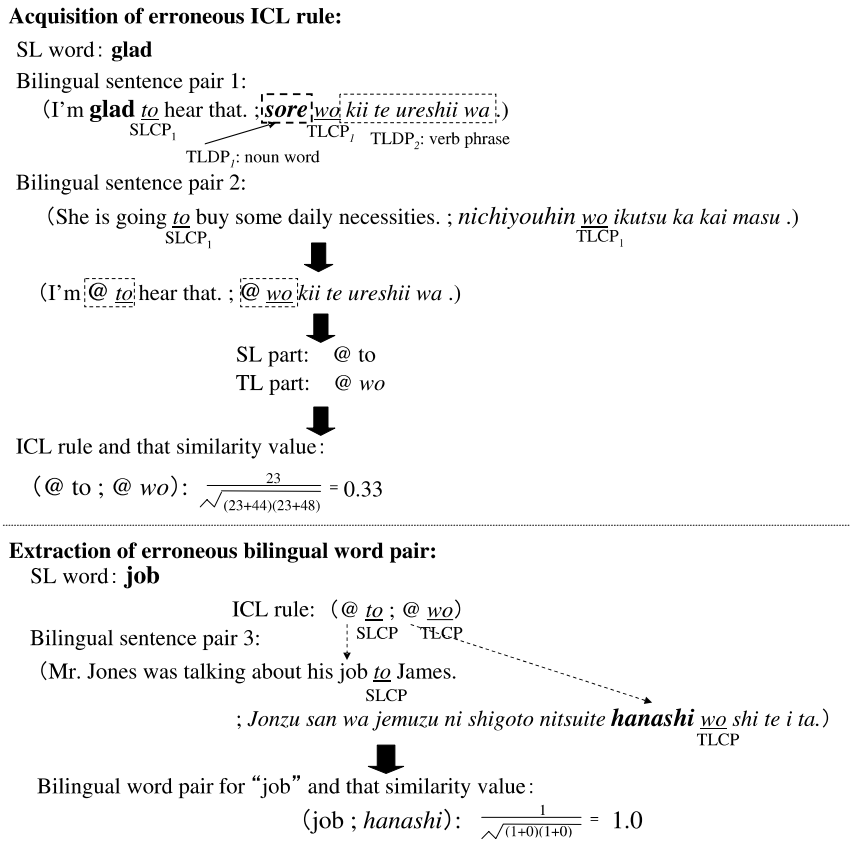


Fig. 12. Examples of acquisition of erroneous ICL rule and extraction of erroneous bilingual word pair in the system using ICL.

Table 7
Results of evaluation experiments in GIZA++

| SL | GIZA++ (%) | GIZA++ + ICL (%) |
|---|---|---|
| English | 47.3 | 54.4 |
| French | 39.6 | 54.2 |
| German | 37.4 | 61.5 |
| Shanghai-Chinese | 62.5 | 60.6 |
| Ainu | 66.6 | 58.2 |
| Total | 51.3 | 57.9 |

(@ to;@ *wo*). In Japanese, "job" corresponds to "*shigoto*", not "*hanashi*". In such a problem, it is effective to use the bilingual word pairs extracted previously. That is, the system determines the equivalent omitted by the set phrase using the extracted bilingual word pairs in the dictionary for bilingual word pairs when it acquires ICL rules.

Moreover, we applied ICL to GIZA++ (herein, we call it the system based on GIZA++ + ICL) by replacing the process based on similarity measure with GIZA++ in Fig. 3. The GIZA++ is the statistical word-alignment model (Och, 2000). Table 7 shows results of evaluation experiments in GIZA++. In the system based on GIZA++, the recalls are very low in English–Japanese, French–Japanese, and German–Japanese parallel corpora. This result indicates that GIZA++ is insufficient when the grammatical structures of SL differ from the grammatical structures of TL, and when the frequencies of many bilingual word pairs are extremely low. On the other hand, in Shanghai-Chinese–Japanese and Ainu–Japanese parallel corpora, the recall values were lower through the use of ICL. In the system based on GIZA++ + ICL, GIZA++ that corresponds to the process based on similarity measure is not executed when the bilingual word pairs with the highest similarity values are obtained by ICL (i.e., the process based on two bilingual sentence pairs, the process based on SL words, the process based on ICL rules; and the determination process of bilingual word pairs). This fact indicates that the recall value of the system based on GIZA++ + ICL is lower when ICL extracts erroneous bilingual word pairs with the highest similarity values and when GIZA++ can extract many correct bilingual word pairs. In Shanghai-Chinese–Japanese and Ainu–Japanese parallel corpora, ICL extracted some erroneous bilingual word pairs with the highest similarity values. Furthermore, GIZA++ extracted many correct bilingual word pairs because the grammatical structures of Shanghai-Chinese and Ainu are similar to the grammatical structure of Japanese. Therefore, the recall values of the system based on GIZA++ + ICL lowered to the recall values of the system based on GIZA++ in Shanghai-Chinese–Japanese and Ainu–Japanese parallel corpora. The grammatical structure of Ainu is SOV, and the grammatical structure of Japanese is also SOV. Although the grammatical structure of Shanghai-Chinese is SVO, the grammatical structure of Shanghai-Chinese is similar to the grammatical structure of Japanese in terms of phrase units (e.g., noun phrase), if not in terms of overall sentences. The recall values improved in English–Japanese, French–Japanese, and German–Japanese parallel corpora through the use of ICL. This fact indicates that ICL is effective in the automatic extraction of bilingual word pairs from the parallel corpora with languages for which the grammatical structures of SL differ from the grammatical structures of TL.

## 5. Conclusion

In this paper, we proposed Inductive Chain Learning (ICL) as a new learning method for extracting bilingual word pairs from parallel corpora with various languages. The system using ICL can automatically extract bilingual word pairs using only parallel corpora without any prior preparation of a bilingual resource (e.g., a bilingual dictionary, a machine translation system). Moreover, the system using ICL can extract bilingual word pairs from parallel corpora with various languages for which grammatical structures of SL differ from the grammatical structures of TL. In addition, the system using ICL can extract not only high-frequency bilingual word pairs, but also low-frequency bilingual word pairs. Evaluation experiments indicated that the system using ICL is very effective to extract bilingual word pairs with various languages and to solve the sparse data problem.

Future studies will solve the problem of word-ambiguity. Moreover, we will apply our method to a multilingual machine translation system and a cross-language information retrieval system.

## Acknowledgement

This work was partially supported by Grants from the High-Tech Research Center of Hokkai-Gakuen University, and an academic research grant of Hokkai-Gakuen University.

## References

Ahrenberg, L., Andersson, M., & Merkel, M. (1998). A simple hybrid aligner for generating lexical correspondences in parallel texts. In *Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics (COLING-ACL'98)* (pp. 29–35).

Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D., & Tyson, M. (1993). FASTUS: A finite-state processor for information extraction from real-world text. In *Proceedings of the 13th international joint conference on artificial intelligence (IJCAI'93)* (pp. 1172–1178).

Bar-Yossef, Z., & Rajagopalan, S. (2002). Template detection via data mining and its application. In *Proceedings of the 11th international world wide web conference (WWW'02)* (pp. 580–591).

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, 19*(2), 263–311.

Chen, A., & Gey, F. C. (2004). Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval, 7*(1–2), 149–182.

Chen, F.S. (1993). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st annual meeting of the association for computational linguistics (ACL'93)* (pp. 9–16).

Chikushi, F. (2001). *Express: French, Hakusui-sha* (in Japanese).

Collier, N., Ono, K., & Hirakawa, H. (1998). An experiment in hybrid dictionary and statistical sentence alignment. In *Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics (COLING-ACL'98)* (pp. 268–274).

Crescenzi, V., Mecca, G., & Merialdo, P. (2001). ROADRUNNER: Towards automatic data extraction from large web sites. In *Proceedings of the 27th international conference on very large data bases* (pp. 109–118).

Dagan, I., Church, K. W., & Gale, W.A. (1993). Robust bilingual word alignment for machine aided translation. In *Proceedings of the workshop on very large corpora: academic and industrial perspectives* (pp. 1–8).

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61–74.

Echizen-ya, H., Araki, K., Momouchi, Y., Tochinai, K. (2002). Study of practical effectiveness for machine translation using Recursive Chain-link-type Learning. In *Proceedings of the 19th international conference on computational linguistics (COLING'02)* (pp. 246–252).

Echizen-ya, H., Araki, K., & Momouchi, Y. (2005a). Automatic extraction of low frequency bilingual word pairs from parallel corpora with various languages. In *Proceedings of the 9th pacific-asia conference on knowledge discovery and data mining (PAKDD'05). Lecture notes in artificial intelligence* (Vol. 3518, pp. 32–37). Springer Publishing.

Echizen-ya, H., Araki, K., & Momouchi, Y. (2005b). Automatic acquisition of adjacent information and its effectiveness in extraction of bilingual word pairs from parallel corpora. In *Proceedings of the 10th international conference on applications of natural language to information systems (NLDB'05). Lecture notes in computer science* (Vol. 3513, pp. 349–352). Springer Publishing.

Emoto, H., Han, G. (2004). *Express: Shanghai, Hakusui-sha* (in Japanese).

Fujii, A., & Ishikawa, T. (2001). Japanese/English cross-language information retrieval: exploration of query translation and transliteration. *Computers and the Humanities, 35*(4), 389–420.

Fung, P., & Church, K. (1994). K-vec: a new approach for alignment parallel texts. In *Proceedings of the 15th international conference on computational linguistics (COLING'94)* (pp. 1096–1102).

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English–Chinese corpus, *Workshop on Very Large Corpora* (pp. 173–183).

Fung, P. (1998). *A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. Lecture notes in artificial intelligence* (Vol. 1529, pp. 1–17). Springer Publishing.

Gale, W. A., & Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics, 19*(1), 75–102.

Güvenir, H. A., & Cicekli, I. (1998). Learning translation templates from examples. *Information Systems, 23*(6), 353–363.

Harukawa, Y., & Snelling, J. (1998). *Express: English, Hakusui-sha* (in Japanese).

Hiemstra, D., de Jong, F., & Kraaij, W. (1997). A domain specific lexicon acquisition tool for cross-language information retrieval. In *Proceedings of the RIAO'97 conference on computer-assisted information searching on internet* (pp. 255–269).

Hirokawa, S., Itoh, E., & Miyahara, T. (2003). Semi-automatic construction of metadata from a series of web documents. In *Proceedings of the 16th Australian conference on artificial intelligence (AI'03). Lecture notes in computer science* (Vol. 2903, pp. 942–953). Springer Publishing.

Hisamitsu, T., & Niwa, Y. (2001). Topic-word selection based on combinatorial probability. In *Proceedings of the 6th natural language processing pacific rim symposium (NLPRS'01)* (pp. 289–296).

Hsu, J. Y., & Yih, W. (1997). Template-based information mining from HTML documents. In *Proceedings of the 14th national conference on artificial intelligence and 9th conference on innovative applications of artificial intelligence (AAAI-IAAI'97)* (pp. 256–262).

Kaji, H., & Aizono, T. (1996). Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *Proceedings of the 16th international conference on computational linguistics (COLING'96)* (pp. 23–28).

Kay, M., & Röscheisen, M. (1993). Text-translation alignment. *Computational Linguistics, 19*(1), 121–142.

Kishida, K., Chen, K., Lee, S., Chen, H., Kando, N., Kuriyama, K., Myaeng, S., & Eguchi, K. (2004). Cross-lingual information retrieval (CLIR) task at the NTCIR workshop 3. *SIGIR Forum, 38*(1), 17–20.

Kumano, A., & Hirakawa, H. (1994). Building an MT dictionary from parallel texts based on linguistic and statistical information. In *Proceedings of the 15th international conference on computational linguistics (COLING'94)* (pp. 76–81).

Kushmerick, N., Weld, D.S., & Doorenbos, R. (1997). Wrapper induction for information extraction. In *Proceedings of the 15th international joint conference on artificial intelligence (IJCAI'97)* (pp. 729–735).

Lee, J., & Bui, T. (2000). A template-based methodology for disaster management information systems. In *Proceedings of the 33rd annual Hawaii international conference on system sciences (HICSS-33)*.

Macklovitch, E., & Hannan, M. L. (1996). Line 'em up: advances in alignment technology and their impact on translation support tools. In *Proceedings of the second conference of the association for machine translation in the Americas (AMTA'96)* (pp. 145–156).

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Matsumoto, Y., Ishimoto, H., & Utsuro, T. (1993). Structural matching of parallel texts. In *Proceedings of the 31st annual meeting of the association for computational linguistics (ACL'93)* (pp. 23–30).

Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., & Asahara, M. (2000). Japanese Morphological Analysis System ChaSen version 2.2.1 manual. Nara Institute of Science and Technology.

McTait, K. (2001). Linguistic knowledge and complexity in an EBMT system based on translation patterns. In *Proceedings of the workshop on EBMT, MT Summit VIII*.

Melamed, I. D. (2001). *Empirical methods for exploiting parallel texts*. MIT Press.

Nakagawa, H., & Nakamoto, M. (2004). *Express: Ainu, Hakusui-sha* (in Japanese).

Nießen, S., & Ney, H. (2004). Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics, 30*(2), 181–204.

Och, F.J. (2000). Giza++: Training of statistical translation models. Available from http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 29*(1), 19–51.

Oshio, T. (2004). *Express: German, Hakusui-sha* (in Japanese).

Pedersen, T., Varma, N. (2003). K-vec++: Approach for finding word correspondences, Available from http://www.d.umn.edu/~tpederse/Code/Readme.K-vec++.v02.txt.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting of the association for computational linguistics (ACL'95)* (pp. 320–322).

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the association for computational linguistics (ACL'99)* (pp. 519–526).

Sadat, F., Yoshikawa, M., & Uemura, S. (2003). Learning bilingual translations from comparable corpora to cross-language information retrieval: hybrid statistic-based and linguistics-based approach. In *Proceedings of the 6th international workshop on information retrieval with Asian languages* (pp. 57–64).

Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics, 22*(1), 1–38.

Tanaka, K., & Iwasaki, H. (1996). Extraction of lexical translation from non-aligned corpora. In *Proceedings of the 16th international conference on computational linguistics (COLING'96)* (pp. 580–585).

Utsuro, T., Hino, K., & Kida, M. (2004). Integrating cross-lingually relevant news articles and monolingual Web documents in bilingual lexicon acquisition. In *Proceedings of the 20th international conference on computational linguistics (COLING'04)* (pp. 1036–1042).

Veronis, J. (2000). *Parallel text processing: alignment and use of translation corpora*. Kluwer Academic Publishers.

Vogel, S., Ney, H., & Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of the 16th international conference on computational linguistics (COLING'96)* (pp. 836–841).

Xu, J., & Weischedel, R. (2003). Cross-lingual retrieval for Hindi. *ACM transactions on Asian language information processing, 2*(2), 164–168.

Yamada, K., & Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th annual meeting of the association for computational linguistics (ACL'01)* (pp. 523–530).

Yamada, Y., Ikeda, D., & Hirokawa, S. (2002). Automatic wrapper generation for multilingual web resources. In *Proceedings of the 5th international conference on discovery science (DS'02)*. *Lecture notes in computer science* (Vol. 2534, pp. 332–339). Springer Publishing.

Zhao, B., Zechner, K., Vogel, S., & Waibel, A. (2003). Efficient optimization for bilingual sentence alignment based on linear regression. In *Proceedings of HLT-NAACL 2003 workshop: building and using parallel texts data driven machine translation and beyond* (pp. 81–87).