Automatic Acquisition of Bilingual Rules for Extraction of Bilingual Word Pairs from Parallel Corpora

Hiroshi Echizen-ya
Dept. of Electronics and Information
Hokkai-Gakuen University
S26-Jo W11-Chome, Chuo-ku
Sapporo, 064-0926 Japan
echi@eli.hokkai-s-u.ac.jp

Kenji Araki
Graduate School of Information Science
and Technology, Hokkaido University
N14-Jo W9-Chome, Kita-ku
Sapporo, 060-0814 Japan
araki@media.eng.hokudai.ac.jp

Yoshio Momouchi
Dept. of Electronics and Information
Hokkai-Gakuen University
S26-Jo W11-Chome, Chuo-ku
Sapporo, 064-0926 Japan
momouchi@eli.hokkai-s-u.ac.jp

Abstract

In this paper, we propose a new learning method to solve the sparse data problem in automatic extraction of bilingual word pairs from parallel corpora with various languages. Our learning method automatically acquires rules, which are effective to solve the sparse data problem, only from parallel corpora without any bilingual resource (e.g., a bilingual dictionary, machine translation systems) beforehand. We call this method **I**nductive **C**hain **L**earning (ICL). The ICL can limit the search scope for the decision of equivalents. Using ICL, the recall in three systems based on similarity measures improved respectively 8.0, 6.1 and 6.0 percentage points. In addition, the recall value of GIZA++ improved 6.6 percentage points using ICL.

1 Introduction

1.1 Sparse data problems in extraction of bilingual word pairs

Many studies of automatic extraction of bilingual word pairs have been reported. Most studies have used similarity measures (Manning and Schütze, 1999; Sadat et al., 2002) because they are language-independent. However, these studies are insufficient because of the sparse data problem. For example, we would like to obtain (book; 本 [*hon*[1]]) as the bilingual word pair from (Your book is on the table.; テーブル/[2]に/あなた/の/本/が/あり/ます. [*teburu ni anata no hon ga ari masu.*]) using the Dice coefficient (Smadja et al., 1996) automatically. The Dice coefficient is defined as

$$Dice(W_S, W_T) = \frac{2a}{(a+b)+(a+c)} \qquad (1)$$

In that equation, 'a' is the number of pieces in which both the **S**ource **L**anguage (SL) word W_S and **T**arget **L**anguage (TL) word W_T were found; 'b' is the number of pieces in which only W_S was found; and 'c' is the number of pieces in which only W_T was found.

In the case of using the Dice coefficient, the system cannot extract only (book; 本 [*hon*]) when the respective frequencies of "book", "本 [*hon*]" and "テーブル [*teburu*]" are 1. That is, the similarity value between "book" and "本 [*hon*]" becomes 1.0($= \frac{2\times1}{1+1}$); the similarity value between "book" and "テーブル [*teburu*]" also be-

[1]Italics means Japanese pronunciation.

[2]'/' in Japanese sentences are inserted after each morpheme because Japanese is an agglutinative language.

comes 1.0($= \frac{2\times1}{1+1}$). This obstacle is common among methods based on similarity measures.

1.2 Basic idea for solution of the sparse data problem

We propose a new learning method to solve this sparse data problem. We call this method **I**nductive **C**hain **L**earning (ICL). For example, in (Your book is on the table.; テーブル/に/あなた/の/本/が/あり/ます. [*teburu ni anata no hon ga ari masu.*]), a system using ICL uses the information that "your" corresponds to "あなた/の [*anata no*]." Moreover, it uses the information that equivalents of words that adjoin the right side of "your" exist on the right side of "あなた/の [*anata no*]" in TL sentences. Using such bilingual rules, the system can extract only (book; 本 [*hon*]). This fact indicates that the system limits the search scope for the decision of equivalents in TL sentences. Consequently, ICL is effective to solve the sparse data problem. In this study, bilingual rules are acquired automatically only from parallel corpora by view of learning (Echizen-ya et al., 2002). The system using ICL extracts bilingual word pairs by applying the acquired bilingual rules to bilingual sentence pairs in parallel corpora. Therefore, the system using ICL causes a chain reaction in the acquisition of bilingual 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, machine translation systems) beforehand. All bilingual rules are acquired automatically solely from the parallel corpora. Moreover, the system using ICL extracts bilingual word pairs using only acquired bilingual rules to solve the sparse data problem.

(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 bilingual rules. The bilingual rules can locate the information to cope with the difference word orders of SL and TL.

(3) The system using ICL can extract bilingual word pairs even when the frequencies of the pairs of the co-occurrence words and the bilingual word pairs are only 1 in a parallel corpus. For example, when the bilingual rule (your @; あなた/の/@[*anata no* @]) exists, the system using ICL can extract (book; 本 [*hon*]) even when the frequency of the pairs of "your" and "book" is only 1. This fact indicates that the system using ICL can extract not only high-frequency bilingual word pairs, but also low-frequency bilingual word pairs.

We applied this ICL to three systems based on the Dice coefficient, Yates' χ^2 (Hisamitsu and Niwa, 1996), and **A**kaike's **I**nformation **C**riterion (AIC) (Akaike, 1974). For evaluation experiments, five kinds of parallel corpora: English – Japanese, French – Japanese, German – Japanese, Shanghai-Chinese – Japanese and Ainu[3] – Japanese parallel corpora were used as evaluation data. Evaluation experiments indicated that, using ICL in the systems based on the Dice coefficient, Yates' χ^2 and AIC, the respective recall values improved 8.0, 6.1 and 6.0 percentage points. In addition, using ICL, the recall of the statistical word-alignment model GIZA++ (Och, 2003) improved 6.6 percentage points. Therefore, we confirmed that ICL is effective to solve the sparse data problem in the extraction of bilingual word pairs from parallel corpora with various languages.

1.3 Related works

Several methods based on the co-occurrence of words have been proposed. (Fung, 1995) proposed a method that specifically examines context heterogeneity, which indicates the number of kinds of words that adjoin SL words. (Rapp, 1999) proposed a method that uses co-occurrence vectors based on the two words that

[3]The Ainu language is spoken by some members of the Ainu ethnic group of northern Japan and Sakhalin. Ainu language is independent from, but similar to, Japanese and Korean.

adjoin SL words on the right side and left side. Moreover, (Fung, 1998; Kaji and Aizono, 1996) proposed methods that uses co-occurrence vectors based on all words that exist in the existing bilingual dictionary, among sentences. (Tanaka and Iwasaki, 1996) presented a translation matrix that provides co-occurring information translated from the source into the target, and obtains bilingual word pairs by determining the best translation matrix. Ultimately, these methods depend on the existing bilingual dictionary. Therefore, it is difficult to extract bilingual word pairs from parallel corpora with various languages when a sufficient bilingual dictionary does not exist. In contrast, the system using ICL automatically can extract bilingual word pairs without an existing bilingual dictionary as a bilingual resource.

Regarding methods for acquisition translation templates, (McTait, 1997; Güvenir and Cicekli, 1998) proposed methods that acquires bilingual templates using common parts and different parts. However, such methods require many similar bilingual sentence pairs to extract sufficient translation templates. Moreover, K-vec (Fung and Church, 1994) is unable to extract low-frequency bilingual word pairs. The algorithm is applicable only to bilingual word pairs that occur with a frequency greater than three.

In addition, statistical word-alignment methods (Brown et al., 1993; Melamed, 2000; Och and Ney, 2003; Nieβen and Ney, 2004) have been proposed, but they are also insufficient. That is, the statistical word-alignment methods cannot extract bilingual word pairs efficiently when the frequencies of many bilingual word pairs are low. (Watanabe and Sumita, 2003) proposed a method by which the decoder uses some translation examples whose source part is similar to the input. However, numerous translation examples are necessary as a bilingual resource. That is, it is difficult to deal with languages for which translation examples are not sufficiently obtainable. In contrast, ICL can extract bilingual rules and bilingual word pairs efficiently, even from a small parallel corpus.

2 Outline

Figure 1 shows an outline of a system using ICL. The ICL corresponds to three processes: a method based on bilingual rules, a method based on two bilingual sentence pairs, and the decision process of bilingual word pairs.

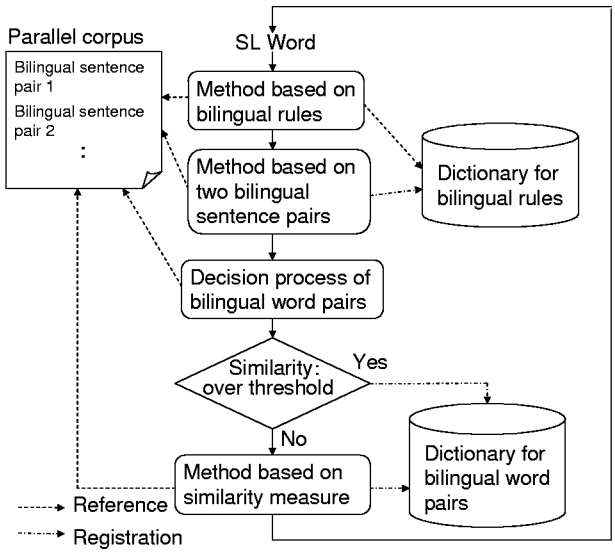

Figure 1: Process flow.

First, the user inputs the SL words of bilingual word pairs. In methods based on bilingual rules, the system extracts bilingual word pairs using the acquired bilingual rules in the dictionary for bilingual rules. In this paper, the bilingual rules are the rules for extracting new bilingual word pairs. In all extracted bilingual word pairs, similarity values between SL words and TL words are assigned using similarity measure. In the method based on two bilingual sentence pairs, the system obtains bilingual word pairs and new bilingual rules using the bilingual sentence pairs that SL words exist and other bilingual sentence pairs. Moreover, in the decision process of bilingual word pairs, the system chooses the most suitable bilingual word pairs using their similarity values when several bilingual word pairs candidates exist. The system compares the similarity values of chosen 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 re-

```
1:  Input: TL sentence of bilingual sentence pair that SL word exists
2:      m = 1
3:      if TLDP_m exists on the left side of TLCP_1 then
4:          If TLDP_m corresponds to 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 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 exists on the right side of TLCP_NTLCP then
24:         if TLDP_m corresponds to 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 words
```

Figure 2: The algorithm of method based on two bilingual sentence pairs.

spective similarity values are greater than the threshold value.

In the method based on similarity measure, the system extracts bilingual word pairs using only one similarity measure (*i.e.*, the Dice coefficient, Yates' χ^2, AIC) from bilingual sentence pairs that SL words exist without ICL. It does so when their similarity values are not greater than the threshold or when no bilingual word pairs are extracted in the ICL process.

3 Process

3.1 Method based on two bilingual sentence pairs

In the method based on two bilingual sentence pairs, the system acquires bilingual rules using the bilingual sentence pairs that SL words exist and other bilingual sentence pairs. The bilingual word pairs for SL words are also extracted. The system obtains bilingual rules using common parts between two bilingual sentence pairs. That is, the word strings for which the frequencies are very low are used as bilingual rules. Using such low-frequency word strings, the bilingual rules are acquired easily only from parallel corpus. In this paper, the respective common parts between SL sentences of two bilingual sentence pairs are called $SLCP_{i=1,...,NSLCP}$; the respective common parts between TL sentences of two bilingual sentence pairs are called $TLCP_{i=1,...,NTLCP}$; the respective different parts between TL sentences of two bilingual sentence pairs are called $TLDP_{m=1,2,3,...}$. In addition, the number of SLCPs is called NSLCP; the number of TLCPs is called NTLCP. The details of the process based on two bilingual sentence pairs are the following:

P1-(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 and TLCPs as the bilingual sentence pairs with SL words. In that case, SLCPs must adjoin SL words in SL sentences.

P1-(2) The system extracts TLDPs that correspond to nouns, verbs, adjectives, adverbs,

or conjunctions from TL sentences of bilingual sentence pairs for which SL words exist. Figure 2 shows the algorithm of this process. In lines 11 and 13 of Fig. 2, ${}_{\mathrm{NTLCP}}\mathrm{C}_2$ indicates $\frac{\mathrm{NTLCP!}}{2!(\mathrm{NTLCP}-2)!}$. That is, it means the number of combinations based on two TLCPs.

P1-(3) The system obtains bilingual word pairs by combining SL words and extracted TLCPs.

P1-(4) The system acquires bilingual rules using the extracted TLDPs. 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 all pairs of each SLCP and variable, and all pairs of each TLCP and variable from bilingual sentence pairs with variables obtained by process (i) of P1-(4).

(iii) The system generates bilingual rules using all combinations of the pairs of SLCPs and variables, and the pairs of TLCPs and variables.

(iv) The system calculates the similarity values between SLCPs and TLCPs in the acquired bilingual rules using the Dice coefficient function (1); it registers the bilingual rules to the dictionary for bilingual rules.

Figure 3 shows an acquisition example of bilingual rules using two English – Japanese bilingual sentence pairs. The system selects bilingual sentence pair 1, for which "house" exists. Furthermore, the system chooses the bilingual sentence pair 2 that have SLCP and TLCPs as the bilingual sentence pairs with SL words by process P1-(1). In Fig. 3, "this" is SLCP in SL sentences of bilingual sentence pairs 1 and 2; it adjoins an SL word "house" in SL sentence of bilingual sentence pair 1. First, the system determines the TLDP that adjoins the left side of TLCP_1 by processes of lines 3 to 8

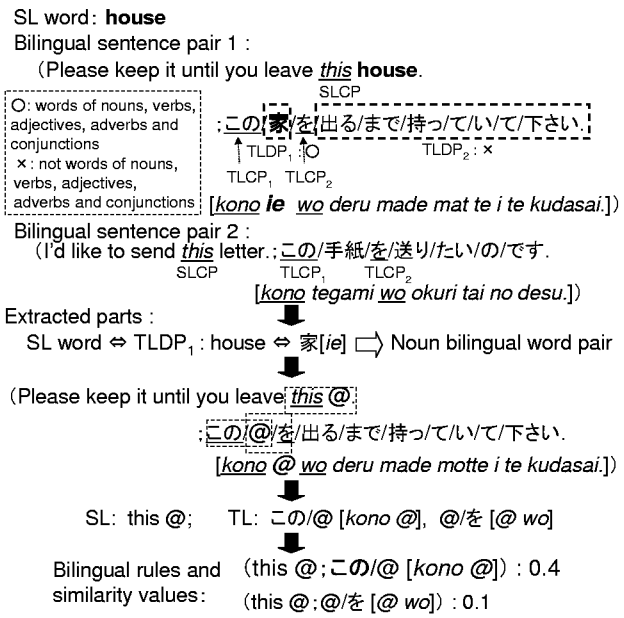

Figure 3: An acquisition example of bilingual rules using two bilingual sentence pairs.

in Fig. 2. However, in TL sentences of bilingual sentence pair 1, the word that adjoins the left side of TLCP_1 ("この [*kono*]") does not exist. Therefore, TLDP is not extracted by this process. The system then determines TLDPs using the parts exist between two TLCPs by the processes of lines 9 to 22 in Fig. 2. In TL sentences of bilingual sentence pair 1, one TLDP exists because the number of combinations based on two TLCPs is 1 by ${}_{\mathrm{NTLCP:2}}\mathrm{C}_2 = \frac{2!}{2!(2-2)!} = 1$. That is, "家 [*ie*]" that exists between TLCP_1 ("この [*kono*]") and TLCP_2 ("を [*wo*]") is determined as TLDP_1. Moreover, the system determines the TLDP that adjoins the right side of $\mathrm{TLCP}_{\mathrm{NTLCP:2}}$ by the processes of lines 23 to 27 in Fig. 2. In TL sentences of bilingual sentence pair 1, "出る/まで/持っ/て/い/て/下さい [*deru made mot te i te kudasai*]" is determined as TLDP_2 because it is the part from the word that adjoins the right side of TLCP_2 ("を [*wo*]") to the word at the end of TL sentence. Among two extracted TLDPs, the TLDP that corresponds to word of noun, verb, adjective, adverb, or conjunction is TLDP_1 ("家 [*ie*]") that is noun word. TLDP_2 ("出る/まで/持っ/て/い/て/下さい [*deru made mot te i te kudasai*]") is

```
1:  Input: SL word
2:      while Selection of bilingual sentence pair that SL word exist, and selection of ICL
            rule that has SLCP and TLCP to the selected bilingual sentence pair
3:          if Variable exists on the right side of TLCP in TL part of ICL rule then
4:              i = 0
5:              while i < NTLCP
6:                  Extraction of TL word (i.e., word of noun, verb, adjective, adverb
                    and conjunction) that adjoins the right side of TLCP_i in TL sentence
7:                  i = i + 1
8:              end
9:          end
10:         if Variable exists on the left side of TLCP in TL part of ICL rule then
11:             i = 0
12:             while i < NTLCP
13:                 Extraction of TL word (i.e., word of noun, verb, adjective, adverb
                    and conjunction) that adjoins the left side of TLCP_i in TL sentence
14:                 i = i + 1
15:             end
16:         end
17:     end
18:     Calculation of similarity value between SL word and each extracted TL word using
        the cosine function (1)
19:     Extraction of bilingual word pair by combining SL word and each TL word
20: Output: Bilingual word pairs
```

Figure 4: The extraction algorithm of bilingual word pairs based on bilingual rules.

verb phrase, not word. Therefore, only (house; 家 [*ie*]) is obtained by combining the SL word ("house") and the extracted TLDP ("家 [*ie*]") by process P1-(3). In addition, the system replaces "house" and "家 [*ie*]" with variable "@" by process (i) of P1-(4). As a result, (this @; この/@[*kono @*]), (this @;@/を [*@ wo*]) are acquired as bilingual rules by process (ii) and (iii) of P1-(4). Similarity values in the acquired bilingual rules (this @; この/@[*kono @*]) and (this @;@/を [*@ wo*]) are calculated using Dice coefficient function (1) by process (iv) of P1-(4). The similarity value of (this @; この/@[*kono @*]) is higher than that of (this @;@/を [*@ wo*]) because (this @; この/@[*kono @*]) is the correct bilingual rule; and (this @;@/を [*@ wo*]) is the erroneous bilingual rule. That is, "this" corresponds to "この [*kono*]", not "を [*wo*]" in Japanese. In this paper, the parts extracted from SL sentences are called SL parts; the parts extracted from TL sentences are called TL parts.

3.2 Method based on bilingual rules

In the method based on bilingual rules, the system extracts bilingual word pairs using the bilingual rules acquired by the method based on two bilingual sentence pairs. The system can limit the search scope for the decision of equivalents in the TL sentences by the use of bilingual rules. Figure 4 gives the extraction algorithm of bilingual word pairs based on bilingual rules.

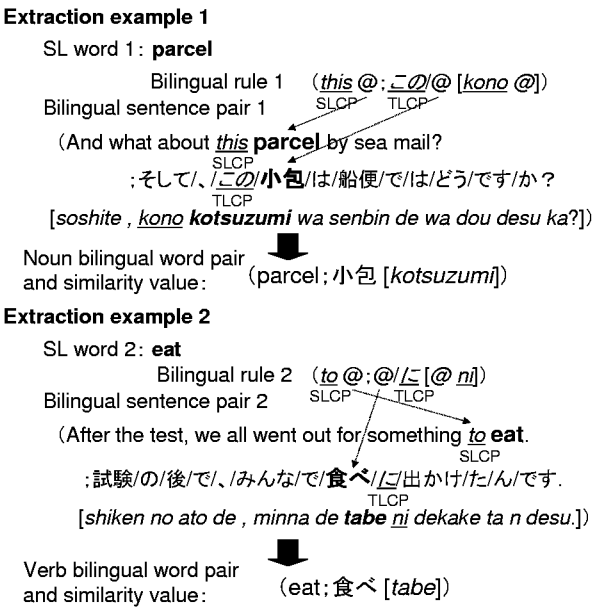

Figure 5: Examples of extraction of bilingual word pairs based on bilingual rules.

Figure 5 shows examples of extraction of bilingual word pairs from English-Japanese bilingual sentence pairs in the method based on bilingual

rules. In example 1 of Fig. 5, (parcel; 小包 [*kotsuzumi*]) is extracted as the noun bilingual word pair using (this @; この/@[*kono @*]) acquired in Fig. 3. First, the system selects bilingual sentence pair 1 that SL word 1 "parcel" exists from a parallel corpus. Moreover, the system selects bilingual rule 1 (this @; この/@[*kono @*]) from the dictionary for bilingual rules because the variable "@" exists on the right side of SLCP ("this") in the SL part of bilingual rule 1, and SL word 1 "parcel" also exists on the right side of SLCP ("this") in the SL sentence of bilingual sentence pair 1. The system then extracts TL words that adjoin the right side of TLCP because the variable "@" exists on the right side of TLCP ("この *kono*") in the TL part of bilingual rule 1. Using bilingual rule 1, noun word "小包 [*kotsuzumi*]", which exists on the right side of TLCP ("この *kono*") is extracted from TL sentence of bilingual sentence pair 1. As a result, the system can obtain (parcel; 小包 [*kotsuzumi*]) as the noun bilingual word pair.

In example 2 of Fig. 5, (eat; 食べ [*tabe*]) is extracted as the verb bilingual word pair using bilingual rule 2 (to @;@/に [*@ ni*]). The system selects bilingual sentence pair 2, in which SL word 2 "eat" exists from a parallel corpus. Moreover, the system selects bilingual rule 2 (to @;@/に [*@ ni*]) from the dictionary for bilingual rules because the variable "@" exists on the right side of SLCP ("to") in the SL part of bilingual rule 2, and SL word 2 "eat" also exists on the right side of SLCP ("to") in SL sentence of bilingual sentence pair 2. The system then extracts TL words that adjoin the left side of TLCP because the variable "@" exists on the left side of TLCP ("に [*ni*]") in the TL part of bilingual rule 2. Using bilingual rule 2, verb word "食べ [*tabe*]", which adjoins the left side of TLCP ("に [*ni*]") is extracted from the TL sentence of bilingual sentence pair 2. The system calculates the similarity value between "eat" and "食べ [*tabe*]" using the Dice coefficient function (1), and registered (eat; 食べ [*tabe*]) into the dictionary of bilingual word pairs. The system determines the most suitable bilingual word pairs according to their similarity values when several bilingual word pairs have been extracted as described in section 3.3.

Using the bilingual rules, the system can decrease the number of candidates of equivalents for SL words. In example 2 of Fig. 5, the system could decrease the number of candidates of equivalents for "eat" using the bilingual rule (to @;@/に [*@ ni*]). All words of nouns, or verbs "試験 [*shiken*]", "後 [*ato*]", "みんな [*minna*]", "食べ [*tabe*]", "出かけ [*dekake*]", and "ん [*n*]" become candidates of equivalents for "eat" when ICL is not used. In contrast, only "食べ [*tabe*]" becomes candidates of equivalents for "eat" using ICL. This fact indicates that ICL is effective to solve the sparse data problem. Moreover, the system can extract bilingual word pairs from parallel corpora of various languages for which the grammatical structure of SL differs from the structure of TL. For example, in the bilingual rule 2 (to @;@/に [*@ ni*]), the variable "@" exists on the right side of "to." In contrast, in the TL part, the variable "@" exists on the left side of "に [*ni*]." Therefore, bilingual rules have the knowledge to cope with the different word order between SL and TL.

3.3 Decision process of bilingual word pair

The system determines the most suitable bilingual word pairs according to their similarity values when several bilingual word pairs have been extracted. The details of this process are the following:

P2-(1) The system selects the bilingual word pairs that have the highest similarity values.

P2-(2) When several bilingual word pairs with identical similarity values exist, the system selects the bilingual word pairs that used bilingual rules with the highest similarity values.

P2-(3) The system selects the bilingual word pairs that appear in a parallel corpus for the first time when it cannot choose only one bilingual word pair by processes P2-(1) and P2-(2).

Table 1: Results of evaluation experiments.

SL	Dice coefficient	Dice +ICL	Yates' χ^2	Yates +ICL	AIC	AIC +ICL	Number of bilingual word pairs
English	49.7%	58.0%	53.8%	59.8%	53.3%	58.6%	169
French	47.9%	56.7%	55.4%	60.4%	55.4%	60.4%	240
German	53.3%	61.0%	53.3%	58.5%	53.8%	59.0%	195
Sh.-Chinese	54.9%	62.9%	57.6%	62.5%	58.3%	62.9%	264
Ainu	54.0%	61.5%	52.1%	62.0%	52.6%	62.4%	213
Total	52.1%	60.1%	54.7%	60.8%	54.9%	60.9%	1,081

3.4 Method based on similarity measure

In the method based on similarity measures, the system extracts bilingual word pairs using only one similarity measure (*i.e.*, the Dice coefficient, Yates' χ^2, AIC) without using ICL when the similarity values are not greater than the threshold value or when no bilingual word pairs are extracted. Moreover, the system chooses the bilingual word pairs that appear in the parallel corpus at the first time when several candidates of bilingual word pairs are obtained.

4 Performance Evaluation

4.1 Experimental Procedure and Evaluation Standard

Five kinds of parallel corpora were used in this paper as experimental data. These parallel corpora are for English – Japanese, French – Japanese, German – Japanese, Shanghai-Chinese – Japanese and Ainu – Japanese. They were taken from textbooks(Harukawa and Snelling, 1998; Chikushi, 2001; Oshio, 2004; Emoto and Han, 2004; Nakagawa and Nakamoto, 2004). The number of bilingual sentence pairs was 1,794; the average numbers of words in SL and TL sentences were 6.8 and 8.8, respectively. We inputted all 1,081 SL words of nouns, verbs, adjectives, adverbs, and conjunctions in five parallel corpora to six systems: a system based on the Dice coefficient; a system based on the Dice coefficient in which AIL is applied (herein, we call it the system based on Dice+ICL); a system based on Yates' χ^2; a system based on Yates' χ^2 in which ICL is applied (herein, the system based on Yates+ICL); a system based on AIC; and a system based on AIC in which ICL is applied (herein, the system based on AIC+ICL). Initially, the dictionary for bilingual word pairs and the bilingual rule dictionary are empty. Moreover, the system uses 0.5 as its best threshold[4]. We repeated the experiments for each parallel corpus using respective systems.

We evaluated whether or not correct bilingual word pairs exist in the dictionary. Moreover, we calculated the recall. The recall is the rate for the number of correct bilingual word pairs to the number of all bilingual word pairs in the parallel corpora (*i.e.*, 1,081).

4.2 Experiments and Discussion

Table 1 shows the results of the experiments. The respective recall values of the systems based on Dice+ICL, Yates+ICL, and AIC+ICL were more than 8.0, 6.1, and 6.0 percentage points higher than those of the systems based on the Dice coefficient, Yates' χ^2, and AIC. These results indicate that ICL is effective for various similarity measures. Particularly, the recall values of the bilingual word pairs for which the frequencies are 1 improved to 11.0, 9.7 and 9.9 percentage points using ICL. In systems without ICL, many bilingual word pairs for which the frequencies are 1 were erroneous bilingual

[4]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.

Table 2: Examples of bilingual word pairs extracted by ICL.

SL	Correct bilingual word pairs	Erroneous bilingual word pairs	
		Bilingual word pairs	Equivalents
English	(cereal; シリアル) 1.0 (boarding house; 下宿) 1.0	(curtains; 新しい [new]) 0.67 (interesting; 外 [outside]) 0.67	curtains interesting
French	(monuments; 記念/建造/物) 1.0 (cherche; 探し [search]) 0.67	(surtout; 関係 [relation]) 1.0 (petit; 所 [place]) 0.67	specially small
German	(nämlich; つまり [after all]) 0.67 (das Foto; 写真 [photograph]) 1.0	(Wege; 橋 [bridge]) 1.0 (Neues; 新聞 [newspaper]) 0.67	lane new event
Sh.-Chinese	(下班；退勤/し [leave office]) 1.0 (大閘蟹; 上海/ガニ [Shanghai crab]) 1.0	(中飯; ご馳走/し [treat]) 1.0 (夜飯; サービス [service]) 0.67	lunch dinner
Ainu	(ekupa; くわえ [take something in one's mouth]) 1.0 (set；寝床 [bed]) 1.0	(apto; 降っ[fall]) 1.0 (tunasno; 起き [get up]) 0.67	rain early

word pairs created by data sparseness problems, as described in section 1.1. Therefore, improvement of the recall values of bilingual word pairs for which the frequencies are 1 indicates that ICL is effective to solve the sparse data problem. On the other hand, the precision values – the rates of the number of correct bilingual word pairs to the number of all extracted bilingual word pairs – are all equal to the recall values. Among all 1,081 SL words, the correct bilingual word pairs or erroneous bilingual word pairs were obtained by the method based on similarity measure when the ICL process extracted no bilingual word pairs. Consequently, the numbers of all bilingual word pairs in the parallel corpora and of all extracted bilingual word pairs became 1,081. That is, the precision values are identical to the recall values. Table 2 shows examples of bilingual word pairs extracted by ICL and their similarity values. Table 2 indicates that 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 over 1.

Furthermore, we applied ICL to GIZA++. Table 3 shows those experimental results. The total recall of GIZA++ +ICL was more than 6.6 percentage points higher than that of GIZA++. Table 3 indicates that ICL is very effective for parallel corpora between languages for which the grammatical structure of SL differs from the grammar structure of TL. Grammatical structures of English, French, and German are SVO, whereas the Japanese grammatical structure is SOV. Using ICL, the recall improved 15.3 percentage points on average in English – Japanese, French – Japanese, and German – Japanese parallel corpora.

Table 3: Experimental results in GIZA++.

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

5 Conclusion

This paper presented **I**nductive **C**hain **L**earning (ICL) as a new learning method to solve the sparse data problem in extraction of bilingual word pairs among various languages. From experimental results, we confirmed that ICL is effective to solve the sparse data problem in extraction of bilingual word pairs from parallel corpora with various languages.

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

6 Acknowledgements

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

Manning, C. D. and Schütze, H. 1999. Foundations of Statistical Natural Language Processing. The MIT Press.

Sadat, F., Déjean, H. and Gaussier, É. 2002. A combination of models for bilingual lexicon extraction from comparable corpora. In *Proceedings of Papilion'02*, pp.16–21.

Smadja, F., McKeown, K. R. and Hatzivassiloglou, V. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, vol.22, no.1, pp.1–38.

Echizen-ya, H., Araki, K. Momouchi, Y., and Tochinai, K. 2002. Study of Practical Effectiveness for Machine Translation Using Recursive Chain-link-type Learning. In *Proceedings of COLING '02*, pp.246–252.

Hisamitsu, T. and Niwa, Y. 2001. Topic-Word Selection Based on Combinatorial Probability. In *NLPRS'01*, pp.289–296.

Akaike, H. 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, AC-19(6), pp.716–723.

Och, F. J. 2003. GIZA++: Training of statistical translation models. Available at http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

Fung, P. 1995. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. Workshop on very large corpora, pp.173–183.

Rapp, R. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of ACL'99*, pp.519–526.

Fung, P. 1998. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora, *LNAI*, Springer Publishing, vol.1529, pp.1–17.

Kaji, H. and Aizono, T. 1996. Extracting Word Correspondences from Bilingual Corpora Based on Word Co-occurrence Information. In *Proc. Coling'96*, pp.23–28.

Tanaka, K. and Iwasaki, H. 1996. Extraction of Lexical Translation from Non-Aligned Corpora. In *Proc. Coling'96*, pp.580–585.

McTait, K. 1997. Linguistic knowledge and complexity in an EBMT system based on translation patterns. In *Proceedings Workshop on EBMT, MT Summit VIII.*

Güvenir, H. A. and Cicekli, I. 1998. Learning translation templates from examples. *Information Systems*, vol. 23, no.6, pp.353–363.

Fung, P. and Church, K. 1994. K-vec: A new approach for alignment parallel texts. In *Proc. Coling'94*, pp.1096–1102.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, vol.19, no.2, pp.263–311.

Melamed, I. D. 2000. Models of translation equivalence among words. *Computational Linguistics*, vol.26, no.2, pp.221–249.

Och, F. J. and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, vol.29, no.1, pp.19–51.

Nieβen, S. and Ney, H. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, vol.30, no.2, pp.181–204.

Watanabe, Y. and Sumita, E. 2003. Example-based decoding for statistical machine translation. In *Proceedings of MT summit IX*, pp.410–417.

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

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

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

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

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