Learning Method for Automatic Acquisition of Translation Knowledge

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

[1] Dept. of Electronics and Information, Hokkai-Gakuen University,
S26-Jo, W11-Chome, Chuo-ku, Sapporo, 064-0926 Japan
{echi, momouchi}@eli.hokkai-s-u.ac.jp
Tel: +81-11-841-1161, Fax: +81-11-551-2951

[2] 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
Tel: +81-11-706-6534, Fax: +81-11-709-6277

Abstract. This paper presents a new learning method for automatic acquisition of translation knowledge from parallel corpora. We apply this learning method to automatic extraction of bilingual word pairs from parallel corpora. In general, similarity measures are used to extract bilingual word pairs from parallel corpora. However, similarity measures are insufficient because of the sparse data problem. The essence of our learning method is this presumption: in local parts of bilingual sentence pairs, the equivalents of words that adjoin the source language words of bilingual word pairs also adjoin the target language words of bilingual word pairs. Such adjacent information is acquired automatically in our method. We applied our method to systems based on various similarity measures, thereby confirming the effectiveness of our method.

1 Introduction

1.1 Problem in Similarity Measures

Bilingual word pairs - pairs of **s**ource **l**anguage (SL) words and **t**arget **l**anguage (TL) words - are extremely important as translation knowledge in the field of machine translation. However, manual extraction by humans of bilingual word pairs of various languages is costly. For that reason, automatic extraction using a system of bilingual word pairs from parallel corpora is effective. Similarity measures [1] are often used to extract bilingual word pairs from parallel corpora because such measures are language independent. However, similarity measures are insufficient because of the sparse data problem. For example, a system based on cosines [1] would seek to extract (letter; 手紙 [*tegami*[1]]) from (I'd like to sent this letter to Japan.; この/手紙/を/日本/に/送り/たい/の/です.[*kono tegami wo nippon ni okuri tai no desu.*]). The cosine is the effective similarity measure. The cosine is defined as

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

[1] Italics means Japanese pronunciation.

R. Khosla et al. (Eds.): KES 2005, LNAI 3682, pp. 1347–1353, 2005.

In function (1), ‘a’ is the number of pieces in which both the SL word W_S and TL word W_T are found, ‘b’ is the number of pieces in which only the W_S is found, and ‘c’ is the number of pieces in which only the W_T is found.

This system based on the cosine cannot clearly extract (letter; 手紙 [*tegami*]) as a correct bilingual word pair when the respective frequencies of “letter”, “手紙 [*tegami*]” and “日本 [*nippon*]” are 1. In that case, the cosine score of “letter” and “手紙 [*tegami*]” becomes $1.0(=\frac{1}{\sqrt{1\times1}})$. The cosine score between “letter” and “日本 [*nippon*]” also becomes $1.0(=\frac{1}{\sqrt{1\times1}})$. Therefore, the system based on the cosine cannot exclusively select (letter; 手紙 [*tegami*]) from among two bilingual word pairs (letter; 手紙 [*tegami*]) and (letter; 日本 [*nippon*]). That is, when several bilingual word pairs have resemblant similarity value candidates, the system based on similarity measures falls into the sparse data problem.

1.2 Motivation

To solve the sparse data problem, we use of this inference: in local parts of bilingual sentence pairs (*e.g.*, phrases, not sentences), the equivalents of words that adjoin the SL words of bilingual word pairs also adjoin the TL words of bilingual word pairs. For example, in (I’d like to send this letter to Japan.; この/手紙/を/日本/に/送り/たい/の/です.[*kono tegami wo nippon ni okuri tai no desu.*]), the system uses the information that “this” which adjoins “letter” corresponds to “この [*kono*].” Moreover, it uses the information that equivalents of words that adjoin the right side of “this” exist on the right side “この [*kono*]” in TL sentences. Consequently, only (letter; 手紙 [*tegami*]) can be extracted as a bilingual word pair. That is, using such adjacent information, the system can limit the search scope for the decision of equivalents in bilingual sentence pairs by extracting only those word pairs that adjoin the adjacent information.

Moreover, the adjacent information is acquired automatically from the perspective of learning [2]. We call this learning method **A**djacent **I**nformation **L**earning (AIL). In this study, we apply AIL to four systems based on the cosine, the Dice coefficient, the **L**og-**L**ikelihood **R**atio (LLR), and Yates’ χ^2 to efficiently extract bilingual word pairs as translation knowledge from five kinds of parallel corpora: English – Japanese, French – Japanese, German – Japanese, Shanghai-Chinese – Japanese, and Ainu[4] – Japanese parallel corpora.

2 Outline

Figure 1 shows an outline of the system using AIL. In Fig. 1, AIL corresponds to three processes: the process based on templates, the process based on two bilingual sentence pairs, and the decision process of bilingual word pairs.

First, the user inputs the SL words of bilingual word pairs. In the process based on templates, the system extracts bilingual word pairs using the templates

[4] Ainu language is spoken by members of the Ainu ethnic group, who mainly reside in northern Japan and Sakhalin. It is independent, but similar to Japanese or Korean.

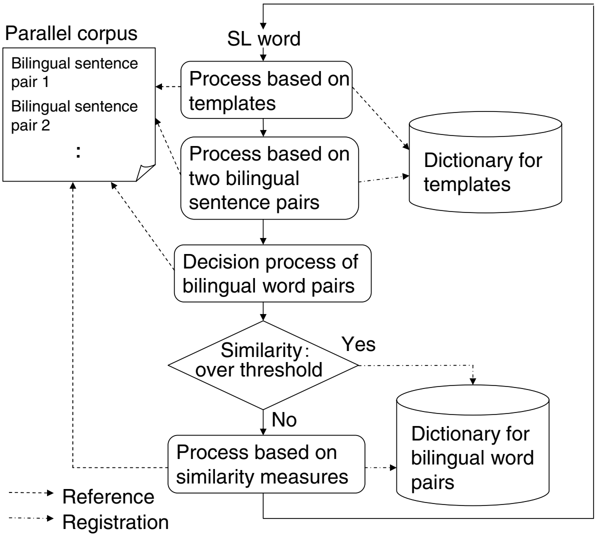
Parallel corpus
Bilingual sentence pair 1
Bilingual sentence pair 2
:
SL word
Process based on templates
Process based on two bilingual sentence pairs
Dictionary for templates
Decision process of bilingual word pairs
Similarity: over threshold
Yes
No
Process based on similarity measures
Dictionary for bilingual word pairs
Reference
Registration

Fig. 1. Process flow.

in the dictionary for templates. In this study, templates are designated as the rules for extracting new bilingual word pairs. Similarity values between SL words and TL words in all extracted bilingual word pairs are assigned. The similarity value is defined by function (2) based on the Dice coefficient.

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

In function (2), f_{ST} is the number of pieces in which both the SL word W_S and the TL word W_T are found, f_S is the number of pieces in which the W_S are found, and f_T is the number of pieces in which the W_T are found. In the process based on two bilingual sentence pairs, the system obtains bilingual word pairs and new templates from two bilingual sentence pairs. Similarity values in all acquired templates are also assigned by function (2). Moreover, during the decision process of bilingual word pairs, the system chooses the most suitable bilingual word pairs using their similarity values when several candidates of bilingual word pairs 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 respective similarity values are greater than the threshold value.

The system extracts bilingual word pairs without AIL in the process based on similarity measures when their similarity values are not greater than the threshold value or when no bilingual word pairs are extracted. Moreover, the extracted bilingual word pairs can be registered into the dictionary efficiently using a morphological analysis system to very minute changes in spellings or words or pronunciation. The system can extract bilingual word pairs even when the scripts of two languages are same because AIL is language independent.

3 Adjacent Information Learning (AIL)

3.1 Process Based on Two Bilingual Sentence Pairs

The system obtains bilingual word pairs and templates using common parts between two bilingual sentence pairs. That is, the bilingual word pairs and the templates can be acquired easily only from a parallel corpus using common parts for which the frequencies are very low, *i.e.* 2. Figure 2 shows examples of extraction of a bilingual word pair and acquisition of a template.

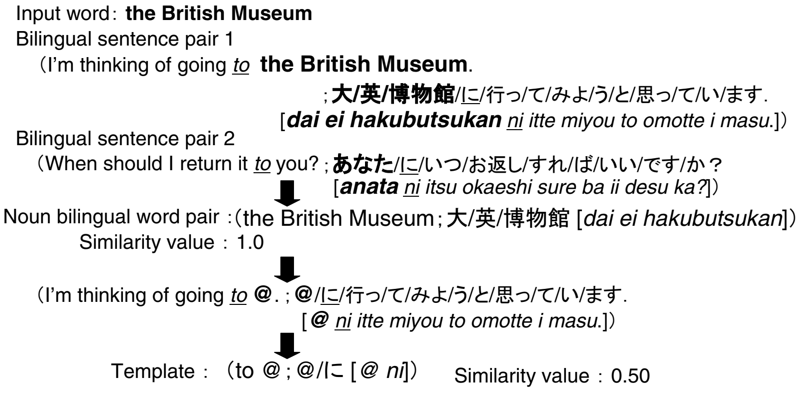
Input word: **the British Museum**
Bilingual sentence pair 1
(I'm thinking of going to **the British Museum**.
; **大/英/博物館**/に/行っ/て/みよ/う/と/思っ/て/い/ます.
[***dai ei hakubutsukan*** *ni itte miyou to omotte i masu.*])
Bilingual sentence pair 2
(When should I return it to you? ; **あなた**/に/いつ/お返し/すれ/ば/いい/です/か？
[***anata*** *ni itsu okaeshi sure ba ii desu ka?*])
Noun bilingual word pair : (the British Museum ; 大/英/博物館 [*dai ei hakubutsukan*])
Similarity value : 1.0
(I'm thinking of going to @. ; @/に/行っ/て/みよ/う/と/思っ/て/い/ます.
[*@ ni itte miyou to omotte i masu.*])
Template : (to @ ; @/に [*@ ni*])
Similarity value : 0.50

Fig. 2. An example of the process based on two bilingual sentence pairs.

First, the system selects bilingual sentence pair 1, for which "the British Museum" exists. Furthermore, the system selects bilingual sentence pair 2 for which "to" that adjoins "the British Museum" exists and for which "に [*ni*]" exists as a common part between two TL sentences. The system extracts "大/英/博物館 [*dai ei hakubutsukan*]", which adjoins the common part "に [*ni*]", from the TL sentence of bilingual sentence pair 1. On the other hand, "行っ/て/みよ/う/と/思っ/て/い/ます [*itte miyou to omotte i masu*]", adjoins "に [*ni*]", is not extracted because it is not the word. Consequently, (the British Museum; 大/英/博物館 [*dai ei hakubutsukan*]), is obtained as a correct noun bilingual word pair. Moreover, the system replaces "the British Museum" and "大/英/博物館 [*dai ei hakubutsukan*]" with the variable "@" in bilingual sentence pair 1, and obtains (to @;@/に [*@ ni*]) as template by combining the common parts "to", "に [*ni*]" and variable "@." Similarity values in all obtained bilingual word pairs and templates are assigned by function (2). On the other hand, when several common parts exist, the system extracts the parts between two common parts from TL sentences. As a result, when several candidates of bilingual word pairs are obtained, the system chooses most suitable bilingual word pairs in the decision process of bilingual word pairs as described in section 3.3.

3.2 Process Based on Templates

The system can extract bilingual word pairs efficiently using templates that have adjacent information. Figure 3 shows an example of extraction of a bilingual word pair using templates. In Fig. 3, (eat; 食べ [*tabe*]) was extracted as the verb bilingual word pair using the template (to @;@/に [@ *ni*]). This template (to @;@/に [@ *ni*]) has information that the equivalents of words that adjoin the right side of "to" exist on the left side "に [*ni*]" in TL sentences. Therefore, only (eat; 食べ [*tabe*]) was easily extracted. The use of such templates is effective to solve the sparse data problem because it can limit the search scope for the decision of equivalents in bilingual sentence pairs.

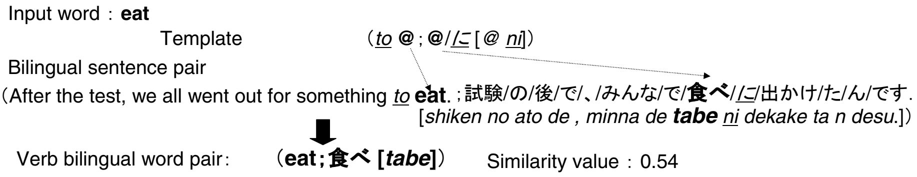
Input word：eat
Template
（to @；@/に [@ ni]）
Bilingual sentence pair
（After the test, we all went out for something to eat.；試験/の/後/で/、/みんな/で/食べ/に/出かけ/た/ん/です.
[shiken no ato de , minna de tabe ni dekake ta n desu.]）
Verb bilingual word pair:
（eat；食べ [tabe]）
Similarity value：0.54

Fig. 3. An example of use of templates.

3.3 Decision Process for Bilingual Word Pairs

The most suitable bilingual word pairs are selected according to their similarity values when several bilingual word pairs have been extracted. That is, the extracted bilingual word pairs are sorted so that the bilingual word pairs that have the highest similarity values are ranked at the top. Moreover, when several bilingual word pairs with equal similarity-value candidates exist, the system selects the bilingual word pairs that appear for the first time in a parallel corpus.

4 Process Based on Similarity Measures

The system extracts bilingual word pairs using only the cosine, the Dice coefficient [3], LLR [4], or Yates' χ^2 [5] without AIL 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 for the first time when several bilingual word-pair candidates are obtained.

5 Experiments for Performance Evaluation

5.1 Experimental Procedure

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 containing conversational sentences. The number of bilingual sentence pairs was 1,794. To confirm AIL's effectiveness, we inputted all 1,081 SL words of nouns, verbs, adjectives, adverbs, and conjunctions into the system based on the cosine, the system based on the cosine in which AIL is applied as described in section 2 (herein, we respectively call it the system based on the cosine+AIL), the system based on the Dice coefficient, the system based on Dice+AIL, the system based on LLR, the system based on LLR+AIL, the system based on Yates' χ^2, and the system based on Yates+AIL. The initial conditions of all dictionaries are empty in those respective systems. We repeated experiments for each parallel corpus using respective systems. The system using AIL uses 0.5[2] as its best threshold value. Moreover, we evaluated whether correct bilingual word pairs are obtained or not, and calculated the extraction rates for all 1,081 SL words.

5.2 Experiments and Discussion

Table 1 shows experimental results. The respective extraction rates of systems using AIL were more than 8.0, 8.0, 6.7 and 6.1 percentage points higher than those of the systems based on the cosine, the Dice coefficient, LLR, and Yates' χ^2. These results indicate that our method is effective for various similarity measures. For example, in Fig. 3, the system without AIL must select "食べ [*tabe*]" as a correct equivalent from many noun and verb candidates "試験 [*shiken*]", "後 [*ato*]", "みんな [*minna*]", "食べ [*tabe*]", and "出かけ [*dekake*]". In contrast, the system using AIL can easily select only "食べ [*tabe*]" using the acquired template (to @;@/に [@ *ni*]).

Table 1. Results of evaluation experiments.

SL	cosine	cosine +AIL	Dice coefficient	Dice +AIL	LLR	LLR +AIL	Yates' χ^2	Yates +AIL
English	52.1%	61.5%	49.7%	58.0%	52.7%	60.4%	53.8%	59.8%
French	50.8%	58.8%	47.9%	56.7%	54.6%	61.3%	55.4%	60.4%
German	52.3%	57.9%	53.3%	61.0%	54.4%	59.5%	53.3%	58.5%
Shanghai-Chinese	56.8%	63.6%	54.9%	62.9%	57.6%	63.3%	57.6%	62.5%
Ainu	52.6%	62.9%	54.0%	61.5%	53.1%	62.0%	52.1%	62.0%
Total	53.1%	**61.1%**	52.1%	**60.1%**	54.7%	**61.4%**	54.7%	**60.8%**

Moreover, we investigated the extraction rates for which the frequencies are 1. In the systems without AIL, many extracted bilingual word pairs for which the frequencies are 1 were erroneous bilingual word pairs because of data sparseness problems, as described in section 1.1. Therefore, improvement of the extraction rates of such bilingual word pairs indicates that AIL is effective to solve the

[2] This value was obtained through preliminary experiments.

sparse data problem. The respective extraction rates of the bilingual word pairs for which the frequencies are 1 improved 11.3, 11.0, 10.9 and 9.7 percentage points using AIL.

Among previous studies, one [6] uses the co-occurrence of words depending on the number of co-occurrence words and their frequency. Such a method is insufficient in terms of efficient extraction of bilingual word pairs. In contrast, the system using AIL requires only a one-word string as the co-occurrence word, *e.g.* only "to" in Fig. 3. Moreover, the system using AIL 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, *e.g.*, "to eat" in bilingual sentence pair of Fig. 3. In a study [7] that acquires low-frequency bilingual terms, bilingual dictionary and MT systems are used for measuring similarity. Therefore, it is difficult to deal with various languages because of the use of large-scale translation knowledge.

6 Conclusion

This paper presented **A**djacent **I**nformation **L**earning (AIL) as a new learning method for solving the sparse data problem in similarity measures. Results showed that AIL is effective for various similarity measures. It is also effective as a solution to the sparse data problem. Future studies will apply this method to a multilingual machine translation system.

References

1. Manning, C. D. and Schütze, H. 1999. Foundations of Statistical Natural Language Processing. MIT Press.
2. Echizen-ya, H., K. Araki, Y. Momouchi, and K. Tochinai. 2002. Study of Practical Effectiveness for Machine Translation Using Recursive Chain-link-type Learning. In *Proceedings of COLING '02*, pp.246–252.
3. Smadja, F., K. R. McKeown and V. Hatzivassiloglou. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, vol.22, no.1, pp.1–38.
4. Dunning, T. E. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, vol.19, no.1, pp.61–74.
5. Hisamitsu, T. and Y. Niwa. 2001. Topic-Word Selection Based on Combinatorial Probability. In *NLPRS'01*, pp.289–296.
6. Kaji, H. and T. Aizono. 1996. Extracting Word Correspondences from Bilingual Corpora Based on Word Co-occurrence Information. In *Proceedings of COLING'96*, pp.23–28.
7. Utsuro, T., K. Hino, and M. Kida. 2004 Integrating Cross-Lingually Relevant News Articles and Monolingual Web Documents in Bilingual Lexicon Acquisition. In *Proceedings of COLING'04*, pp.1036–1042.