Automatic Acquisition of Adjacent Information and Its Effectiveness in Extraction of Bilingual Word Pairs from Parallel Corpora

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
Tel: +81-11-841-1161, Fax: +81-11-551-2951
{echi,momouchi}@eli.hokkai-s-u.ac.jp

[2] Graduate School of Information Science and Technology, Hokkaido University,
N14-Jo, W9-Chome, Kita-ku, Sapporo, 060-0814 Japan
Tel: +81-11-706-6534, Fax: +81-11-709-6277
araki@media.eng.hokudai.ac.jp

Abstract. We propose a learning method for solving the sparse data problem in automatic extraction of bilingual word pairs from parallel corpora. In general, methods based on similarity measures are insufficient because of the sparse data problem. The essence of our method is the use of this inference: in local parts of bilingual sentence pairs (*e.g.*, phrases, not sentences), the equivalents of words that adjoin the source language words of bilingual word pairs also adjoin the target language words of bilingual word pairs. Our learning method automatically acquires such adjacent information. The acquired adjacent information is used to extract bilingual word pairs. As a result, our system can limit the search scope for the decision of equivalents in bilingual sentence pairs by extracting only word pairs that adjoin the acquired adjacent information. We applied our method to two systems based on Yates' χ^2 and AIC. Results of evaluation experiments indicate that the extraction rates respectively improved 6.1 and 6.0 percentage points using our method.

1 Introduction

Manual extraction by humans of bilingual word pairs of various languages is costly. For that reason, automatic extraction of bilingual word pairs from parallel corpora is effective. Many similarity measures [1] are used to extract bilingual word pairs automatically from parallel corpora with various languages because they are language independent. However, they are insufficient. That is, when several bilingual word pairs with close similarity values candidates exist, the system based on similarity measures falls into the sparse data problem. This problem is common among the methods based on similarity measures.

To overcome the sparse data problem, we use the hypothesis that, in local parts of bilingual sentence pairs (*e.g.*, phrases, not sentences), the equivalents of words that adjoin the source language (SL) words of bilingual word pairs also

A. Montoyo et al. (Eds.): NLDB 2005, LNCS 3513, pp. 349–352, 2005.

adjoin the target language (TL) words of bilingual word pairs. Such adjacent information is effective to solve the sparse data problem. That is, 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 for learning [2]. These features allow the application of our method to parallel corpora with various languages. We call this learning method Adjacent Information Learning (AIL). In this paper, we applied AIL to two systems based on Yates' χ^2 [3] and Akaike's Information Criterion (AIC) [4] to extract bilingual word pairs from parallel corpora. Evaluation experiments using parallel corpora with five different languages indicated that the extraction rates respectively improved 6.1 and 6.0 percentage points through the use of AIL. Therefore, we confirmed that AIL is effective to solve the sparse data problem in extraction of bilingual word pairs from parallel corpora.

2 Acquisition of Adjacent Information

The system obtains bilingual word pairs and templates using two bilingual sentence pairs. In this paper, the templates are rules that possess the adjacent information for extracting new bilingual word pairs. The system determines the templates using common parts between two bilingual sentence pairs. Moreover, the system assigns similarity values between SL words and TL words using the Dice coefficient for the obtained bilingual word pairs and templates. Figure 1 shows an acquisition example of adjacent information.

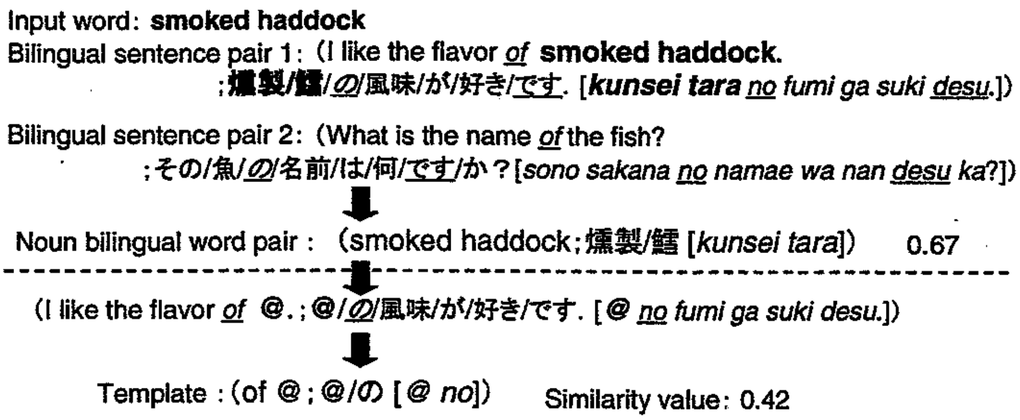
Input word: smoked haddock
Bilingual sentence pair 1: (I like the flavor of smoked haddock.
;燻製/鱈/の/風味/が/好き/です. [kunsei tara no fumi ga suki desu.])
Bilingual sentence pair 2: (What is the name of the fish?
;その/魚/の/名前/は/何/です/か？[sono sakana no namae wa nan desu ka?])
Noun bilingual word pair : (smoked haddock;燻製/鱈 [kunsei tara]) 0.67
(I like the flavor of @.;@/の/風味/が/好き/です. [@ no fumi ga suki desu.])
Template : (of @ ; @/の [@ no]) Similarity value: 0.42

Fig. 1. An example of acquisition of adjacent information.

First, the system selects bilingual sentence pair 1, for which the SL word "smoked haddock" exists. Furthermore, the system selects bilingual sentence pair 2, for which "of" that adjoins "smoked haddock" exists and for which "の [*no*]" and "です [*desu*]" exist as common parts between two TL sentences. The system extracts "燻製/鱈 [*kunsei tara*]", which exists on the left side of the common part "の [*no*]", from the TL sentence of bilingual sentence pair 1. On the other hand,

"風味/が/好き [*fumi ga suki*]", which exists between "の [*no*]" and "です [*desu*]", is also extracted. However, it does not correspond to a noun, verb, adjective, adverb, or conjunction. Consequently, only (smoked haddock; 燻製/鱈 [*kunsei tara*]) is obtained as a correct noun bilingual word pair. Moreover, the system acquires the template (of @;@/の [@ *no*]) by replacing "smoked haddock" and "燻製/鱈 [*kunsei tara*]" with the variable "@" in bilingual sentence pair 1. Similarity values in (smoked haddock; 燻製/鱈 [*kunsei tara*]) and (of @;@/の [@ *no*]) are calculated using the Dice coefficient. The system chooses the most suitable bilingual word pairs and templates using their similarity values when several candidates of bilingual word pairs and templates exist.

The template (of @;@/の [@ *no*]) has the information that the equivalents of words that adjoin the right side of "of" exist on the left side "の [*no*]" in TL sentences. This fact indicates that the system using AIL can limit the search scope for the decision of equivalents in bilingual sentence pairs by extracting ONLY word pairs that adjoin the acquired templates. In contrast, the system without AIL must select correct bilingual word pairs from ALL bilingual word pairs that are nouns, verbs, adjectives, adverbs, and conjunctions in bilingual sentence pairs.

3 Performance Evaluation and Conclusions

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 conversation sentences. The number of bilingual sentence pairs was 1,794. We inputted all 1,081 SL words of nouns, verbs, adjectives, adverbs, and conjunctions into the system based on Yates' χ^2, the system based on Yates' χ^2 in which AIL is applied (herein, we call it the system based on Yates+AIL), the system based on AIC and the system based on AIC+AIL. Initially, the dictionary for bilingual word pairs and the template dictionary are empty. We repeated the experiments for each parallel corpus using respective systems. In addition, we evaluated whether or not correct bilingual word pairs exist in the dictionary and calculated the extraction rate for all SL words.

Table 1 shows experimental results. The respective extraction rates of the systems based on Yates+AIL and AIC+AIL were more than 6.1 and 6.0 percentage points higher than those of the systems based on Yates' χ^2 and AIC. These results indicate that AIL is effective for both Yates' χ^2 and AIC.

Moreover, in the systems based on Yates+AIL and AIC+AIL, the respective extraction rates of the bilingual word pairs for which the frequencies are 1 were more than 9.7 and 9.9 percentage points higher than those of the systems based on Yates' χ^2 and AIC. This fact indicates that AIL is effective to solve the sparse data problem. In some erroneous bilingual word pairs extracted by systems without AIL, their frequencies are 1. The system without AIL extracted such erroneous bilingual word pairs because of the data sparseness problems. Therefore, improvement of the extraction rates of bilingual word pairs for which

Table 1. Results of evaluation experiments.

SL	Yates' χ^2	Yates +AIL	AIC	AIC +AIL	Number of bilingual word pairs
English	53.8%	59.8%	53.3%	58.6%	169
French	55.4%	60.4%	55.4%	60.4%	240
German	53.3%	58.5%	53.8%	59.0%	195
Shanghai-Chinese	57.6%	62.5%	58.3%	62.9%	264
Ainu	52.1%	62.0%	52.6%	62.4%	213
Total	54.7%	**60.8%**	54.9%	**60.9%**	1,081

the frequencies are 1 indicates that AIL is effective to solve the sparse data problem.

Among related works, one study [5] has acquired low-frequency bilingual terms using a bilingual dictionary and MT systems for measuring similarity. It is difficult to deal with various languages because of the use of large-scale translation knowledge. On the other hand, one study [6] used co-occurrence of words depending on the number of co-occurrence words and their frequency. That method is insufficient in terms of efficient extraction of bilingual word pairs. In contrast, AIL requires only a one-word string as the co-occurrence word, *i.e.* only "of." Moreover, 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. Regarding methods [2, 7] for acquisition templates, such methods require similar bilingual sentence pairs to extract effective templates.

Future studies will apply our method to a multilingual machine translation system.

References

1. Manning, C. D. and H. Schütze. 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. Hisamitsu, T. and Y. Niwa. 2001. Topic-Word Selection Based on Combinatorial Probability. In *NLPRS'01*, pp.289–296.
4. Akaike, H. 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723.
5. 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.
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. McTait, K. and A. Trujillo. 1999. A Language-Neutral Sparse-Data Algorithm for Extracting Translation Patterns. In *Proceedings of TMI'99*, pp.98–108.