Automatic Extraction of Low Frequency Bilingual Word Pairs from Parallel Corpora with Various Languages

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, momouch}@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
araki@media.eng.hokudai.ac.jp

Abstract. In this paper, we propose a new learning method for extraction of low-frequency bilingual word pairs from parallel corpora with various languages. It is important to extract low-frequency bilingual word pairs because the frequencies of many bilingual word pairs are very low when large-scale parallel corpora are unobtainable. We use the following inference to extract low frequency bilingual word pairs: the word equivalents that adjoin the source language words of bilingual word pairs also adjoin the target language words of bilingual word pairs in local parts of bilingual sentence pairs. Evaluation experiments indicated that the extraction rate of our system was more than 8.0 percentage points higher than the extraction rate of the system based on the Dice coefficient. Moreover, the extraction rates of bilingual word pairs for which the frequencies are one and two respectively improved 11.0 and 6.6 percentage points using AIL.

1 Introduction

Use of parallel corpora with various languages is effective to build dictionaries of bilingual word pairs because bilingual sentence pairs that are pairs of source language (SL) sentences and target language (TL) sentences include natural equivalents and novel equivalents. Moreover, it is important to extract low-frequency bilingual word pairs because the frequencies of many bilingual word pairs are extremely low when large-scale parallel corpora are unobtainable. Consequently, systems based on similarity measures [1, 2] fall into the sparse data problem because bilingual word pair candidates with close similarity value increase when many low-frequency bilingual word pairs exist.

From the perspective of learning [3], we propose a new method for extraction of low-frequency bilingual word pairs from parallel corpora. We call this new learning method Adjacent Information Learning (AIL). The AIL is based on the inference that the equivalents of the words that are adjacent the SL words

T.B. Ho, D. Cheung, and H. Liu (Eds.): PAKDD 2005, LNAI 3518, pp. 32–37, 2005.

of bilingual word pairs also adjoin the TL words of bilingual word pairs in local parts of bilingual sentence pairs. Our method easily acquires such adjacent information solely from parallel corpora. Moreover, our system can extract not only high-frequency bilingual word pairs, but also low-frequency bilingual word pairs, which typically have the sparse data problem. Thereby, our system can limit the search scope for the decision of equivalents in bilingual sentence pairs.

Evaluation experiments using five kinds of parallel corpora indicated that the extraction rate of our system using AIL was more than 8.0 percentage points higher than the extraction rate of a system based on the Dice coefficient. Moreover, the extraction rate of bilingual word pairs for which the frequencies are one and two respectively improved 11.0 and 6.6 percentage points using AIL. We thereby confirmed that our method is effective to extract low-frequency bilingual word pairs efficiently.

2 Outline

Our system consists of four processes: a method based on templates, a method based on two bilingual sentence pairs, a decision process of bilingual word pairs, and a method based on similarity measures.

First, the user inputs SL words of a bilingual word pair. In the method based on templates, the system extracts bilingual word pairs using the bilingual sentence pairs, the templates, and the SL words. In this paper, templates are defined as rules to extract new bilingual word pairs. Similarity between SL words and TL words is determined in all extracted bilingual word pairs. In the method based on two bilingual sentence pairs, the system obtains bilingual word pairs and new templates using two bilingual sentence pairs and the SL words. Similarity is determined in all templates. Moreover, during the decision process of bilingual word pairs, the system chooses the most suitable bilingual word pairs using their similarity values from among all extracted bilingual word pairs. The system then compares 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 similarity values are greater than the threshold value.

The system extracts bilingual word pairs using the Dice coefficient with bilingual sentence pairs and the SL words in the method based on similarity measures. It does so when their similarity values are not over the threshold or when no bilingual word pairs are extracted.

3 Extraction Process of Bilingual Word Pairs

3.1 Method Based on Two Bilingual Sentence Pairs

In the method based on two bilingual sentence pairs, the system obtains bilingual word pairs and templates using two bilingual sentence pairs. Details of the method based on two bilingual sentence pairs are the following:

(1) The system selects bilingual sentence pairs for which the SL words exist.
(2) The system compares the bilingual sentence pairs selected by process (1) with other bilingual sentence pairs in the parallel corpus. The system selects those bilingual sentence pairs that have the same word strings as those adjoining the SL words, *i.e.*, the common parts, and those that have parts in common with TL sentences.
(3) The system extracts the TL words that correspond to the SL words using the common parts from the bilingual sentence pairs selected through process (1). When the system uses the common parts that exist near words at the beginning of a sentence, it extracts, from the TL sentence, those parts between words at the beginning of a sentence and words that adjoin the left sides of the common parts. When the system uses the common parts that exist near words at the end of a sentence, it extracts, from the TL sentence, those parts between words that adjoin the right sides of common parts and words at the end of a sentence. When the system uses several common parts, it extracts, from the TL sentence, those parts between the two common parts.
(4) The system only selects parts that are nouns, verbs, adjectives, adverbs, or conjunctions.
(5) The system calculates the similarity values between the SL words and the parts selected by process (4) using the Dice coefficient [1].
(6) The system replaces the extracted bilingual word pairs with variables in bilingual sentence pairs.
(7) The system acquires templates by combining common parts and variables.
(8) The system calculates the similarity values between SL words and TL words in the acquired templates using the Dice coefficient; it registers the templates to the template dictionary.

Figure 1 shows an example of acquisition of template: (by @; *@ de*) is acquired as the template. The system replaces the SL word "air mail" and the TL word "*koukubin*" with the variable "@" in bilingual sentence pair by process (6). In this case, "by" and "*de*" are common parts between two bilingual sentence pairs. Consequently, the system obtains (by @; *@ de*) as a template by combining "by @" and "*@ de.*" In this paper, the parts extracted from SL sentences are called SL parts; the parts extracted from TL sentences are called TL parts.

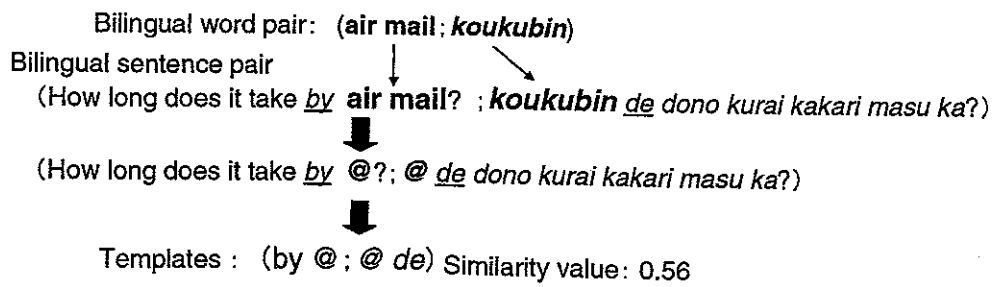

Fig. 1. An example of template acquisition

3.2 Method Based on Templates

In the method based on templates, the system extracts bilingual word pairs using the acquired templates. Details of the extraction process of bilingual word pairs using templates are the following:

(1) The system selects bilingual sentence pairs for which the SL words exist.
(2) The system compares the bilingual sentences selected by process (1) with the templates in the dictionary. Subsequently, the system selects the templates for which SL parts have the same parts as those adjoining the SL words, and for which TL parts have the same parts as those in TL sentences.
(3) The system extracts TL words that correspond to the SL words. The system extracts words that adjoin the left sides of common parts from TL sentences when variables exist on the left sides in TL parts of templates. The system extracts words that adjoin the right sides of common parts from TL sentences when variables exist on the right sides in TL parts of templates.
(4) The system calculates similarity values between the SL words and the parts extracted from TL sentences using the Dice coefficient.

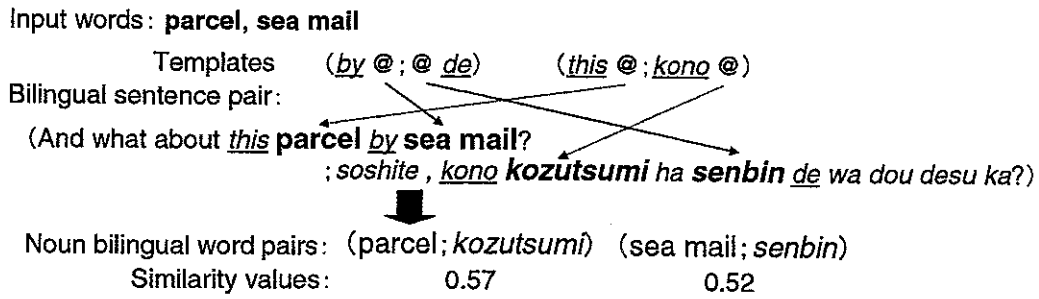

Fig. 2. Examples of extraction of bilingual word pairs

Figure 2 shows examples of extraction of bilingual word pairs from English – Japanese bilingual sentence pairs. In Fig. 2, (parcel;*kozutsumi*) and (sea mail;*senbin*) are extracted respectively as the noun bilingual word pairs using the templates (by @;@ *de*) and (this @;*kono* @). The template (by @;@ *de*) has information that equivalents of words, which adjoin the right side of "by", exist on the left side "*de*" in TL sentences. This fact indicates that the acquired templates have bilingual knowledge that can be used to process the differing word orders of SL and TL. Moreover, our system using AIL can extract bilingual word pairs efficiently without depending on the frequencies of bilingual word pairs using the templates.

3.3 Decision Process of Bilingual Word Pairs and the Method Based on Similarity Measures

In the decision process of bilingual word pairs, the most-suitable bilingual word pairs are selected according to similarity values when several bilingual word pairs

are extracted. The extracted bilingual word pairs are sorted so that bilingual word pairs with the largest similarity values are ranked highest.

Moreover, in the method based on similarity measures, the system extracts bilingual word pairs using only the Dice coefficient without AIL when the similarity values are not greater than a threshold value or when no bilingual word pairs are extracted.

4 Performance Evaluation and Conclusion

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. We inputted all SL words of nouns, verbs, adjectives, adverbs, and conjunctions to our system using AIL and the system based on the Dice coefficient, respectively. The initial conditions of all dictionaries are empty. Moreover, our system using AIL uses 0.5[1] as its best threshold value. We repeated the experiments for each parallel corpus using each system. We evaluated whether correct bilingual word pairs are obtained or not, and calculated the extraction rate for all SL words.

Experimental results indicated that the extraction rate of our system using AIL was more than 8.0 percentage points (from 52.1% to 60.1%) higher than that of the system based on the Dice coefficient. Moreover, in each parallel corpus, the extraction rates improved using AIL. Therefore, our method is effective when using parallel corpora of various languages.

Tables 1 and 2 show extraction rate details in our system using AIL and the system based on the Dice coefficient. In Tables 1 and 2, the extraction rates of the bilingual word pairs for which the frequencies are one and two respectively improved 11.0 and 6.6 percentage points using AIL. This result verified that our system using AIL can extract low-frequency bilingual word pairs efficiently.

In related works, K-vec [4] is applied only to bilingual word pairs for which the frequencies are greater than three. Therefore, it is insufficient in terms of extraction of low-frequency bilingual word pairs. In one study [5] that acquired low-frequency bilingual terms, a bilingual dictionary and MT systems were used for measuring similarity. Therefore, it is difficult to deal with various languages because of the use of large-scale translation knowledge. On the other hand, one study [6] that uses the co-occurrence of words depends on the number of co-occurrence words and their frequency. Therefore, such a method is insufficient in terms of efficient extraction of bilingual word pairs. In contrast, AIL merely requires a one-word string as the co-occurrence word, *e.g.*, only "by" and "this", as shown in Fig. 2. 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 one. In Fig. 2, the respective frequencies of "by sea mail" and "this

[1] This value was obtained through preliminary experiments.

Table 1. Details of extraction rates in our system using AIL

Frequency	English	French	German	Sh.-Chinese	Ainu	Total	Number of bilingual word pairs
1	46.4%	49.4%	51.3%	49.1%	56.9%	**50.4%**	681
2	71.4%	80.0%	71.4%	90.7%	74.4%	**78.6%**	168
others	89.7%	73.5%	79.2%	82.1%	61.5%	75.4%	232
Total	58.0%	56.7%	61.0%	62.9%	61.5%	60.1%	1,081

Table 2. Details of extraction rates in the system based on the Dice coefficient

Frequency	English	French	German	Sh.-Chinese	Ainu	Total	Number of bilingual word pairs
1	35.7%	37.5%	39.5%	40.0%	45.0%	**39.4%**	681
2	64.3%	80.0%	67.9%	74.4%	71.8%	**72.0%**	168
others	89.7%	73.5%	79.2%	83.9%	58.5%	75.0%	232
Total	49.7%	47.9%	53.3%	54.9%	54.0%	52.1%	1,081

parcel", which are pairs formed by the co-occurrence of words and the SL words of bilingual word pairs, are only one. The method [7] that acquires templates requires many similar bilingual sentence pairs to extract effective templates.

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. The MIT Press.
2. 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.
3. 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.
4. Pedersen, T. and N. Varma. 2003. K-vec++: Approach For Finding Word Correspondences. Available at . http://www.d.umn.edu/ tpederse/Code/Readme.K-vec++.v02.txt
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. Tanaka, K and H. Iwasaki 1996. Extraction of Lexical Translation from Non-Aligned Corpora. In *Proceedings of COLING'96*, pp.580–585.
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.