Automatic Building of a Machine Translation Bilingual Dictionary Using Recursive Chain-Link-Type Learning from a Parallel Corpus

Hiroshi Echizen-ya[1], Kenji Araki[2], Yoshio Momouchi[3], and Koji Tochinai[4]

[1] 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,
TEL: +81-11-841-1161(ext.7863), FAX: +81-11-551-2951

[2] Division of Electronics and Information, Hokkaido University, N13-Jo, W8-Chome, Kita-ku Sapporo, 060-8628 Japan
araki@media.eng.hokudai.ac.jp,
TEL: +81-11-706-6534, FAX: +81-11-706-6534

[3] 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,
TEL: +81-11-841-1161(ext.7864), FAX: +81-11-551-2951

[4] Division of Business Administration, Hokkai-Gakuen University, 4-Chome, Asahi-machi, Toyohira-ku Sapporo, 060-8790 Japan
tochinai@econ.hokkai-s-u.ac.jp,
TEL: +81-11-841-1161(ext.2753), FAX: +81-11-824-7729

Abstract. Numerous methods have been developed for generating a machine translation (MT) bilingual dictionary from a parallel text corpus. Such methods extract bilingual collocations from sentence pairs of source and target language sentences. Then those collocations are registered in an MT bilingual dictionary. Bilingual collocations are lexically corresponding pairs of parts extracted from sentence pairs. This paper describes a new method for automatic extraction of bilingual collocations from a parallel text corpus using no linguistic knowledge. We use Recursive Chain-link-type Learning (RCL), which is a learning algorithm, to extract bilingual collocations. Our method offers two main advantages. One benefit is that this RCL system requires no linguistic knowledge. The other advantage is that it can extract many bilingual collocations, even if the frequency of appearance of the bilingual collocations is very low. Experimental results verify that our system extracts bilingual collocations efficiently. The extraction rate of bilingual collocations was 74.9% for all bilingual collocations that corresponded to nouns in the parallel corpus.

1 Introduction

Recent years have brought the ability to obtain much information that is written in various languages using the Internet in real time. However, current machine

R. Monroy et al. (Eds.): MICAI 2004, LNAI 2972, pp. 410–419, 2004.

translation (MT) systems can be used only for a limited number of languages. It is important for MT systems to build bilingual dictionaries. Therefore, many methods have been studied for automatic generation of an MT bilingual dictionary. Such methods are able to produce an MT bilingual dictionary by extracting bilingual collocations from a parallel corpus. Bilingual collocations are lexically corresponding pairs of parts extracted from sentence pairs of source and target language sentences. These studies can be classified into three areas of emphasis. Some use a linguistic-based approach [1]. In a linguistic-based approach, the system requires static, large-scale linguistic knowledge to extract bilingual collocations *e.g.*, a general bilingual dictionary or syntax information. Therefore, it is difficult to apply such static large-scale linguistic knowledge to other various languages easily because developers must acquire linguistic knowledge for other languages.

Other methods for extracting bilingual collocations include statistical approaches [2, 3]. In these statistical approaches, it is difficult to extract bilingual collocations when the frequency of appearance of the bilingual collocations is very low, *e.g.*, only one time. Therefore, the system requires a large bilingual corpus, or many corpora, to extract bilingual collocations. Typically, a statistical approach extracts only bilingual collocations that occur more than three times in the parallel corpus. A third type of method emphasizes the use of learning algorithms that extract bilingual collocations from sentence pairs of source and target language sentences without requiring static linguistic knowledge. We have proposed a method using Inductive Learning with Genetic Algorithms (GA-IL) [4]. As shown in Fig. 1, this method uses a genetic algorithm to generate two sentence pairs automatically with one different part of two source language sentences and with just one different part of two target language sentences. Unfortunately, this method requires similar sentence pairs as the condition of extraction of bilingual collocations. Therefore, these learning algorithm methods require numerous similar sentence pairs to extract many bilingual collocations, even though they require no *ex ante* static linguistic knowledge.

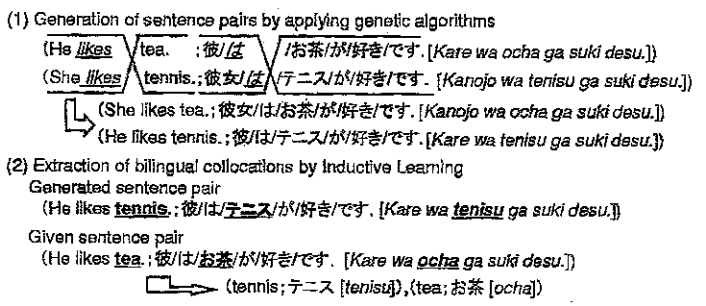

Fig. 1. Example of bilingual collocation extraction using GA-IL

We propose a new method for automatic extraction of bilingual collocations from a parallel corpus to overcome problems of existing approaches. Our method uses the learning algorithm we call Recursive Chain-link-type Learning (RCL) [5] to extract bilingual collocations efficiently using no linguistic knowledge. In this RCL system, various bilingual collocations are extracted efficiently using only character strings of previously-extracted bilingual collocations. This feature engenders many benefits. This RCL system requires no static analytical knowledge, in contrast to a linguistic-based approach. Furthermore, in contrast to a statistical approach, it does not require a high frequency of appearance for bilingual collocations in the parallel corpus. This means that this RCL system can extract bilingual collocations from only a few sentence pairs. Numerous similar sentence pairs are unnecessary, in stark contrast to requirements of a learning-based approach. Evaluation experiment results demonstrate that this RCL system can extract useful bilingual collocations. We achieved a 74.9% extraction rate for bilingual collocations which correspond to nouns. Moreover, the extraction rate of bilingual collocations for which the frequency of appearance was only one in the parallel text corpus was 58.1%.

2 Overview of Our Method

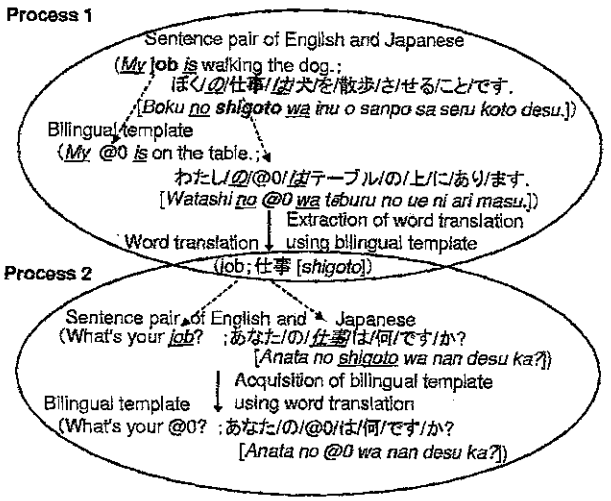

Fig. 2. Diagram of the English-Japanese collocation extraction process

The prominent feature of our method is that it does not require linguistic knowledge. We intend to realize a system based only on learning ability from the view of language acquisition of children. The RCL algorithm imitates part

of that principle in language acquisition because it requires no *ex ante* static linguistic knowledge.

Figure 2 depicts the RCL process, which extracts English-Japanese collocations. This RCL system extracts two types of bilingual collocations. This is an English-Japanese collocation: (job; 仕事 [*shigoto*][5]). It can be registered in an MT bilingual dictionary as a word-level bilingual translation. Hereafter, we call this type of collocation a **word translation**. In contrast, phrases such as (What's your @0?; あなた/の/@0/は/何/です/か?[6] [*Anata no @0 wa nan desu ka?*]) are representative of an English-Japanese collocations that are used as a template for extraction of word translations. This type of collocation is called a **bilingual template**. In this paper, a word translation is a pair of source and target parts; a bilingual template is also a pair of source and target parts. Figure 2 shows a process by which a word translation (job; 仕事 [*shigoto*]) and a new bilingual template (What's your @0?; あなた/の/@0/は/何/です/か? [*Anata no @0 wa nan desu ka?*]) are extracted reciprocally.

In process 1 of Fig. 2, this RCL system extracts (job; 仕事 [*shigoto*]) as a word translation. This (job; 仕事 [*shigoto*]) corresponds to the variables "@0" in the bilingual template (My @0 is on the table.; わたし/の/@0/は/テーブル/の/上/に/あり/ます. [*Watashi no @0 wa tēburu no ue ni ari masu.*]). "My" and "is" adjoin the variable "@0" in the source part of the bilingual template. They are shared parts with the parts in the English sentence "My job is walking the dog." Moreover, "の [*no*]" and "は [*wa*]" adjoin the variable "@0" in the target part of the bilingual template; they are also shared parts with parts in the Japanese sentence "ぼく/の/仕事/は/犬/を/散歩/さ/せる/こと/です. [*Boku no shigoto wa inu o sanpo sa seru koto desu.*]" Therefore, this RCL system extracts the (job; 仕事 [*shigoto*]) by extracting "job" between the right of "My" and the left of "is" in the English sentence, and extracting "仕事 [*shigoto*]" between the right of "の [*no*]" and the left of "は [*wa*]" in the Japanese sentence.

Moreover, this RCL system acquires new bilingual templates using only character strings of the extracted (job; 仕事 [*shigoto*]). In process 2 of Fig. 2, the source part "job" of the word translation (job; 仕事 [*shigoto*]) has the same character strings as the part in the English sentence "What's your job?" In addition, the target part "仕事 [*shigoto*]" of the word translation (job; 仕事 [*shigoto*]) has the same character strings as the part in the Japanese sentence "あなた/の/仕事/は/何/です/か? [*Anata no shigoto wa nan desu ka?*]." Therefore, this RCL system acquires (What's your @0?; あなた/の/@0/は/何/です/か? [*Anata no @0 wa nan desu ka?*]) as the bilingual template by replacing "job" and "仕事 [*shigoto*]" with the variables "@0" for the sentence pair (What's your job?; あなた/の/仕事/は/何/です/か? [*Anata no shigoto wa nan desu ka?*]).

Extracted word translations and bilingual templates are applied for other sentence pairs of English and Japanese to extract new ones. Therefore, word

[5] Italics express pronunciation in Japanese.

[6] "/" in Japanese sentences are inserted after each morpheme because Japanese is an agglutinative language. This process is performed automatically according to this system's learning method [6], without requiring any static linguistic knowledge.

translations and bilingual templates are extracted reciprocally as a linked chain. A characteristic of our method is that both word translations and bilingual templates are extracted efficiently using only character strings of sentence pairs of the source and target language sentences. Thereby, our system can extract bilingual collocations using no linguistic knowledge, even in cases where such collocations appear only a few times in the corpus. Figure 2 shows that (job; 仕事 [*shigoto*]) was extractable even though it appears only one time.

3 Outline

Figure 3 shows an outline of this RCL system's extraction of bilingual collocations from sentence pairs of source and target language sentences. First, a user inputs a sentence pair. In the feedback process, this RCL system evaluates extracted word translations and bilingual templates using the given sentence pairs. The user does not evaluate word translations and bilingual templates directly. In the learning process, word translations and bilingual templates are extracted automatically using two learning algorithms: RCL and GA-IL. In this study, this RCL system extracts English-Japanese collocations.

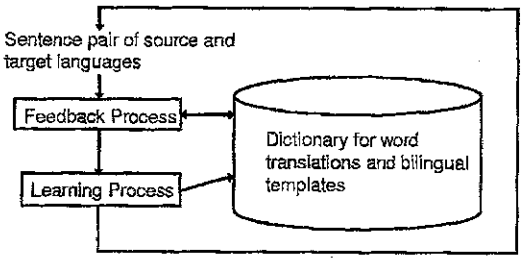

Fig. 3. Process flow

4 Process

4.1 Feedback Process

Our system extracts not only correct bilingual collocations, but also erroneous bilingual collocations. Therefore, this RCL system evaluates bilingual collocations in the feedback process. In this paper, correct bilingual collocation denotes a situation in which source parts and target parts correspond to each other; erroneous bilingual collocations are cases where source parts and target parts do not correspond to one another. In the feedback process, this RCL system first

generates sentence pairs in which source language sentences have the same character strings as source language sentences of the given sentence pairs. It does so by combining bilingual templates with word translations. Consequently, this RCL system can generate sentence pairs in which the English sentences have the same character strings as the English sentences of given sentence pairs.

Subsequently, this RCL system compares Japanese sentences of the generated sentence pairs with Japanese sentences of given sentence pairs. When Japanese sentences of generated sentence pairs have the same character strings as the Japanese sentences of given sentence pairs, word translations and bilingual templates used to generate sentence pairs are determined to be correct. In this case, this RCL system adds one point to the correct frequency of the used word translations and bilingual templates. On the other hand, word translations and bilingual templates used to generate sentence pairs are designated as erroneous when Japanese sentences of generated sentence pairs have different character strings from Japanese sentences of given sentence pairs. In that case, this RCL system adds one point to the error frequency of used word translations and bilingual templates. Using the correct frequency and error frequency, this RCL system calculates the Correct Rate (CR) for the word translations and bilingual templates that were used. Following is a definition of CR. This RCL system evaluates word translations and bilingual templates automatically using CR.

$$\text{CR } (\%) = \frac{\text{Correct frequency}}{\text{Correct frequency} + \text{Error frequency}} \times 100.0 \qquad (1)$$

4.2 Learning Process

Word translations and bilingual templates are extracted reciprocally by this RCL system. We first describe the extraction process of word translations using bilingual templates as in process 1 of Fig. 2. Details of this process are:

(1) This RCL system selects sentence pairs that have the same parts as those parts that adjoin variables in bilingual templates.
(2) This RCL system obtains word translations by extracting parts that adjoin common parts, which are the same parts as those in bilingual templates, from sentence pairs. This means that parts extracted from sentence pairs correspond to variables in bilingual templates. In the extraction process, there are three patterns from the view of the position of variables and their adjoining words in bilingual templates.

Pattern 1: When common parts exist on both the right and left sides of variables in source or target parts of bilingual templates, this RCL system extracts parts between two common parts from source language sentences or target language sentences.

Pattern 2: When common parts exist only on the right side of variables in source parts or target parts of bilingual templates, this RCL system extracts parts from words at the beginning of the sentence to words of

the left sides of common parts in source language sentences or target language sentences.

Pattern 3: When common parts exist only on the left side of variables in source parts or target parts of bilingual templates, this RCL system extracts parts from words of the right sides of common parts to words at the end in source language sentences or target language sentences.

(3) This RCL system yields CR that are identical to those bilingual templates used to extract word translations.

In addition, we describe the acquisition process of bilingual templates using word translations as in process 2 of Fig. 2. Details of this process are:

(1) This RCL system selects word translations in which source parts have identical character strings to those parts in source language sentences of sentence pairs, and in which target parts have the same character strings as parts in the target language sentences of sentence pairs.
(2) This RCL system acquires bilingual templates by replacing common parts, which are identical to word translations, with variables.
(3) This RCL system yields CR that are identical to those word translations used to acquire bilingual templates.

On the other hand, word translations or bilingual templates that are used as starting points in the extraction process of new ones are extracted using GA-IL. The reason for using GA-IL is that our system can extract bilingual collocations using only a learning algorithm with no static linguistic knowledge. In this study, our system uses both RCL and GA-IL.

5 Experiments for Performance Evaluation

5.1 Experimental Procedure

To evaluate this RCL system, 2,856 English and Japanese sentence pairs were used as experimental data. These sentence pairs were taken from five textbooks for first and second grade junior high school students. The total number of characters of the 2,856 sentence pairs is 142,592. The average number of words in English sentences in the 2,856 sentence pairs is 6.0. All sentence pairs are processed by our system based on the outline described in Section 3 and based on the process described in Section 4. The dictionary is initially empty.

5.2 Evaluation Standards

We evaluated all extracted word translations that corresponded to nouns. Extracted word translations are ranked when several different target parts are obtained for the same source parts. In that case, word translations are sorted so that word translations which have the highest CR described in Section 4.1 are ranked at the top. Among ranked word translations, three word-translations, ranked from No. 1 to No. 3, are evaluated by the user as to whether word translations where source parts and target parts correspond to each other are included in those three ranked translations or not.

5.3 Experimental Results

Table 1. Extraction rate of this RCL system

Extraction rate	Detail	
	nouns	compound nouns
74.9% (347)	75.5% (330)	65.4% (17)

There are 463 kinds of nouns and compound nouns in the evaluation data: 437 varieties of nouns and 26 varieties of compound nouns. Table 1 shows the extraction rate of this RCL system in the evaluation data. In Table 1, values in parentheses indicate the number of correct word translations extracted by this RCL system. Moreover, in this paper, a system using only GA-IL is used for comparison to this RCL system. It is difficult to make comparisons among methods [1–3] which extract bilingual collocation because they typically use various static linguistic knowledge. In the system using only GA-IL, the extraction rate for the word translations which corresponded to nouns and compound nouns was 58.7%. Therefore, using RCL, the extraction rate improved from 58.7% to 74.9%. The extraction rate of word translations for which the frequency of appearance is only one time in the parallel text corpus improved from 32.6% to 58.1% through use of RCL. Table 2 shows examples of the extracted correct word translations.

Table 2. Examples of extracted correct word translations

English	Japanese
museum	博物館 [*hakubutsukan*]
machine	機械 [*kikai*]
sumo	すもう [*sumō*]
すもう means a Japanese traditional sport.	
Statue of Liberty	自由/の/女神 [*jiyū no megami*]
Alice in Wonderland	不思議/の/国/の/アリス [*fushigi no kuni no arisu*]
electric guitar	エレキ/ギター [*ereki gitā*]

5.4 Discussion

We confirmed that this RCL system can extract word translations without requiring a high frequency of appearances of word translations. Figure 4 shows the change in extraction rates engendered by this RCL system and the system using only GA-IL for every 100 word translations that correspond to nouns and compound nouns in the 2,856 sentence pairs used as evaluation data. Figure 4

shows 463 word translations that correspond to nouns and compound nouns. The translations are arranged by appearance sequence in 2,856 sentence pairs.

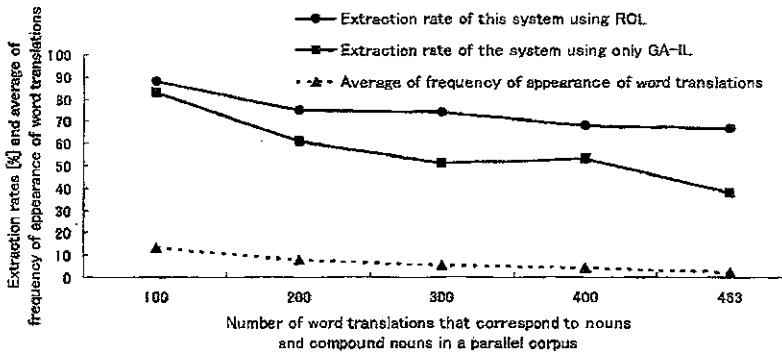

Fig. 4. Change of extraction rates and the average frequency of appearance of extracted word translations in the parallel corpus

In Fig. 4, the dotted line shows average frequencies of appearance of word translations for every 100 word translations that correspond to nouns and compound nouns in evaluation data. The average frequencies of appearance of word translations between Nos. 1 and 100 are high because such word translations appear in many other sentence pairs. The average frequency of appearance of word translations between Nos. 1 and 100 is 13.0. In general, the system extracts word translations easily when their frequency of appearance is high because the probability that the system can extract them is relatively high. Consequently, the extraction rate of word translations between Nos. 1 and 100 is higher than for those in other parts of Fig. 4. On the other hand, the average frequency of appearance of word translations between Nos. 401 and 463 is low because such word translations do not appear in any other sentence pairs. The average frequency of appearance of word translations between Nos. 401 and 463 is 2.4. In general, it is difficult for the system to extract word translations when their frequency of appearance is low because the probability that the system can extract them is relatively low.

Figure 4 depicts the extraction rate of the system using only GA-IL. The rate decreases rapidly as the frequency of appearance of word translations decreases. In contrast, the extraction rate of this RCL system is almost flat except between Nos. 1 and 100. In this RCL system, the decrement of the extraction rate is only nine points between Nos. 101 and 463. In the system using only GA-IL, the decrement of the extraction rate is 23 points between Nos. 101 and 463. These results imply that this RCL system can extract many word translations efficiently without requiring a high frequency of appearance of word translations.

On the other hand, erroneous word translations are also extracted in this RCL system. The precision of extracted word translations was 47.3%. This precision is insufficient. However, in the feedback process described in Section 4.1, this RCL system can evaluate these word translations as erroneous word translations. The rate at which the system could determine erroneous word translations for extracted erroneous word translations was 69.2%. In that case, erroneous word translations mean word translations whose CR is under 50.0%.

6 Conclusion

This paper proposed a new method for automatic extraction of bilingual collocations using **R**ecursive **C**hain-link-type **L**earning (RCL). In this RCL system, various bilingual collocations are extracted efficiently using only character strings of previously-extracted bilingual collocations. Moreover, word translations and bilingual templates are extracted reciprocally, as with a linked chain. Therefore, this RCL system can extract many word translations efficiently from sentence pairs without requiring any static linguistic knowledge or when confronting corpus which contain words with very low frequency of appearance. This study demonstrates that our method is very effective for extracting word translations and thereby building an MT bilingual dictionary.

Future studies will undertake more evaluation experiments using practical data. We also intend to confirm that this RCL system can extract bilingual collocations from sentence pairs using other languages. We infer that this RCL system is a learning algorithm that is independent of a specific language. Moreover, we will apply RCL to other natural language processing systems, *e.g.*, a dialog system, to confirm RCL effectiveness.

References

1. Kumano, A. and H. Hirakawa. 1994. Building an MT Dictionary from Parallel Texts based on Linguistic and Statistical Information. In *Proceedings of Coling '94.*
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. Brown, R. D. 1997. Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation. In *Proceedings of TMI '93.*
4. Echizen-ya, H., K. Araki, Y. Momouchi, and K. Tochinai. 1996. Machine Translation Method Using Inductive Learning with Genetic Algorithms. In *Proceedings of Coling '96.*
5. 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.*
6. Araki, K., Y. Momouchi and K. Tochinai. 1995. Evaluation for Adaptability of Kana-kanji Translation of Non-segmented Japanese Kana Sentences using Inductive Learning. In *Proceedings of Pacling '95.*