Acquisition of Word Translations Using Local Focus-Based Learning in Ainu-Japanese Parallel Corpora

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. This paper describes a new learning method for acquisition of word translations from small parallel corpora. Our proposed method, Local Focus-based Learning (LFL), efficiently acquires word translations and collocation templates by focusing on parts of sentences, not on entire sentences. Collocation templates have collocation information to acquire word translations from each sentence pair. This method is useful even when frequency of appearances of word translations is very low in sentence pairs. The LFL system described in this paper extracts Ainu-Japanese word translations from small Ainu-Japanese parallel corpora. The Ainu language is spoken by the Ainu ethnic group residing in northern Japan and Sakhalin. An evaluation experiment indicated that the recall was 57.4% and the precision was 72.0% to 546 kinds of nouns and verbs in 287 Ainu-Japanese sentence pairs even though the average frequency of appearances of the 546 kinds of nouns and verbs was 1.98.

1 Introduction

In recent years, many studies have addressed methods for building bilingual dictionaries from bilingual corpora [1,2,3,4]. Using bilingual corpora, such methods can obtain natural equivalents. However, these methods require large parallel corpora to acquire many word translations that are corresponding words of source

A. Gelbukh (Ed.): CICLing 2004, LNCS 2945, pp. 300–304, 2004.

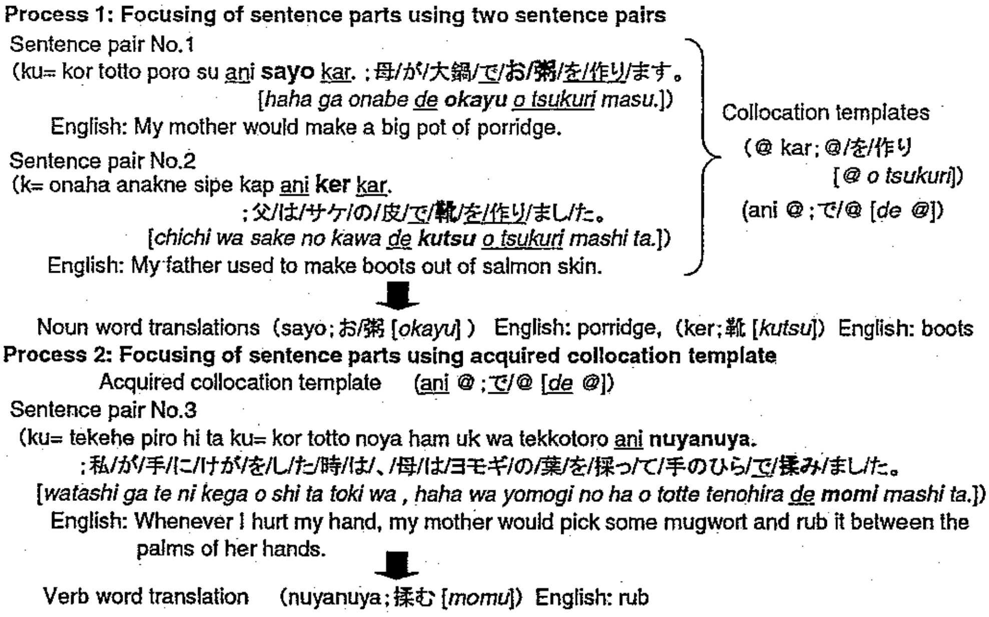
language words and target language words because they cannot acquire many word translations when the frequency of appearances of word translations is low. In an earlier study, we proposed a learning method that acquires bilingual knowledge to solve such problems [5]. A system described in our past work acquires bilingual knowledge using no large parallel corpus. However, the sentences in parallel corpora are very simple. Moreover, many similar sentence pairs are needed in the system based on translation patterns [6]. This fact renders that system inefficient. Sentence pairs are pairs of source language sentences and target language sentences.

This paper proposes a new learning method for acquisition of word translations from small parallel corpora containing many long sentences. We call this method Local Focus-based Learning (LFL). This LFL system efficiently acquires word translations and collocation templates by focusing on parts of sentences, not on the entire sentences. Collocation templates have collocation information to acquire word translations from each sentence pair. This method is useful even when frequency of appearances of word translations is very low in sentence pairs. We used Ainu-Japanese parallel corpora to confirm the effectiveness of our method. In that case, it is difficult to obtain large Ainu-Japanese parallel corpora. However, it is important to acquire word translations from Ainu-Japanese parallel corpora because we can get natural equivalents. Our method is effective to acquire word translations from such small parallel corpora. Ainu language is spoken by the Ainu ethnic group residing in northern Japan and Sakhalin. Although typologically similar in some respects to Japanese, Ainu is a language isolate: it has no known relation to other languages. Evaluation experiment results show that this LFL system acquired many Ainu-Japanese word translations without using large parallel corpora. In 287 Ainu-Japanese sentence pairs, there were 546 kinds of nouns and verbs. We obtained a recall rate of 57.4% with precision of 72.0% even though the average frequency of appearance of the 546 kinds of nouns and verbs is 1.98. From these results, we confirmed that LFL is very effective to acquire word translations.

2 Acquisition of Word Translations Using LFL

This LFL system acquires Ainu-Japanese noun and verb word translations from a few Ainu-Japanese sentence pairs. In Japanese sentences of sentence pairs, part-of-speech information serves as static linguistic knowledge because this LFL system obtains the basic vocabulary item for verbs and each morpheme in Japanese. In addition, a heuristic for position information, based on the similarity of collocation between Ainu and Japanese, is used to prevent the acquisition of erroneous Ainu-Japanese word translations. Japanese is an agglutinative language. Ainu is expressed as a non-agglutinative language even though no formal orthography exists for writing Ainu. Latin-based scripts devised by linguists, as well as the Japanese katakana syllabary, are used variously.

Our LFL is a very simple method. Figure 1 shows an example of acquisition of word translations using LFL. This LFL system focuses on sentence parts by



Process 1: Focusing of sentence parts using two sentence pairs

Sentence pair No.1
(ku= kor totto poro su ani **sayo** kar. ;母/が/大鍋/で/**お/粥**/を/作り/ます。
[*haha ga onabe de* ***okayu*** *o tsukuri masu.*])
English: My mother would make a big pot of porridge.

Sentence pair No.2
(k= onaha anakne sipe kap ani **ker** kar.
;父/は/サケ/の/皮/で/**靴**/を/作り/まし/た。
[*chichi wa sake no kawa de* ***kutsu*** *o tsukuri mashi ta.*])
English: My father used to make boots out of salmon skin.

Collocation templates
(@ kar; @/を/作り
[*@ o tsukuri*])
(ani @ ; で/@ [*de @*])

Noun word translations (sayo; お/粥 [*okayu*]) English: porridge, (ker; 靴 [*kutsu*]) English: boots

Process 2: Focusing of sentence parts using acquired collocation template

Acquired collocation template (ani @ ; で/@ [*de @*])

Sentence pair No.3
(ku= tekehe piro hi ta ku= kor totto noya ham uk wa tekkotoro ani **nuyanuya**.
;私/が/手/に/けが/を/し/た/時/は/、/母/は/ヨモギ/の/葉/を/採っ/て/手のひら/で/**揉み**/まし/た。
[*watashi ga te ni kega o shi ta toki wa , haha wa yomogi no ha o totte tenohira de* ***momi*** *mashi ta.*])
English: Whenever I hurt my hand, my mother would pick some mugwort and rub it between the palms of her hands.

Verb word translation (nuyanuya; 揉む [*momu*]) English: rub

Fig. 1. Acquisition of word translations using LFL

two processes. The first process focuses on sentence parts using two sentence pairs, as shown in process 1 of Fig. 1. In process 1, this LFL system acquires word translations and collocation templates. Collocation templates express collocation information to acquire word translations efficiently from various sentence pairs. This LFL system acquires new word translations through the use of these collocation templates. Details of this process are the following:

(1) This LFL system selects two sentence pairs that have two common parts. In process 1, "ani", "kar" and "で [*de*][1]", "を/作り[2] [*o tsukuri*]" are common parts in sentence pair Nos. 1 and 2.
(2) This LFL system extracts parts between two common parts from Ainu sentences and Japanese sentences; it then obtains word translations by combining them. In process 1, "sayo" and "ker" are extracted from Ainu sentences, and "お/粥 [*okayu*]" and "靴 [*kutsu*]" are extracted from Japanese sentences. As a result, (sayo; お/粥 [*okayu*]) and (ker; 靴 [*kutsu*]) are obtained as noun word translations. The "sayo" and "お/粥 [*okayu*]" mean "porridge" in English, whereas "ker" and "靴 [*kutsu*]" mean "boots" in English.
(3) This LFL system replaces extracted parts with variables "@", and acquires collocation templates by separating variables "@" and their adjoining common parts from sentence pairs. In process 1, (@ kar;@/を [*@ o*]) and (ani @; で/@ [*@ de*]) are acquired as collocation templates.

[1] Italics express pronunciation in Japanese

[2] "/" in Japanese sentences are inserted after each morpheme.

The second process focuses on sentence parts using acquired collocation templates, as shown in process 2 of Fig. 1. Details of this process are the following:

(1) This LFL system selects collocation templates in which Ainu parts, aside from variables, have identical character strings to those parts in Ainu sentences, and in which Japanese parts, aside from variables, have the same character strings as parts in the Japanese sentences. In process 2, (ani @; で/@ [@ *de*]) is selected because "ani" and "で [*de*]" have the same character strings as parts in sentence pair No. 3.

(2) This LFL system extracts words that correspond to variables from Ainu sentences and Japanese sentences. In that case, this LFL system extracts one word from Ainu sentences, and extracts one noun word, one verb word, or one noun phrase, one verb phrase from Japanese sentences. This LFL system obtains word translation by combining them. In process 2, "nuyanuya" and "揉む [*momu*]" are extracted from sentence pair No. 3. Consequently, (nuyanuya; 揉む [*momu*]) is obtained as verb word translation. The "nuyanuya" and "揉む [*momu*]" mean "rub" in English.

This LFL system can acquire various word translations by these processes. In process 1, *n:m* lexical translations are also acquired, and *1:n* lexical translations are acquired in process 2.

3 Experiments and Discussion

Evaluation experiment used 287 Ainu and Japanese sentence pairs as experimental data. These sentence pairs were taken from two books [7,8]. The average number of words in Ainu sentences of all sentence pairs was 11.8. Among these 287 sentence pairs, 546 kinds of nouns and verbs existed. Their average frequency of appearance was 1.98. All sentence pairs were processed by the method described Section 2. In that case, the dictionary was initially empty.

Evaluation experiment showed a 57.4% recall rate with 72.0% precision. The recall rate indicates the rate of acquired correct word translations among the 546 kinds of nouns and verbs. The precision is the rate of acquired correct word translations among the acquired word translations to the 546 kinds of nouns and verbs. In the 546 kinds of nouns and verbs, there were 356 (65.2%) kinds of nouns and verbs for which the frequency of appearance was only. This means that this LFL system can acquire noun and verb word translations which the frequency of appearances is very low.

References

1. 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.
2. Haruno, M., S. Ikehara and T. Yamazaki. 1996. Learning Bilingual Collocations by Word-Level Sorting. In *Proceedings of Coling '96*.

3. Melamed, I.D. 1997. A Word-to-Word Model of Translation Equivalence. In *Proceedings of ACL '97.*
4. Sato, K. and H. Saito 2002. Extracting Word Sequence Correspondences with Support Vector Machines. In *Proceedings of Coling '02.*
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. McTait, K. 2001. Linguistic Knowledge and Complexity in an EBMT System Based on Translation Patterns. In *Proceedings Workshop on EBMT, MT Summit VIII.*
7. Nakamoto, M. and T. Katayama. 1999. The wisdom of the Ainu: UPASKUMA 1. Shin Nippon Kyouiku Tosho.
8. Nakamoto, M. and T. Katayama. 2001. The wisdom of the Ainu: UPASKUMA 2. Shin Nippon Kyouiku Tosho.