PREDICTION METHOD OF WORD FOR TRANSLATION OF UNKNOWN WORD

HISAYUKI SASAOKA  KENJI ARAKI  YOSHIKO MOMOCHI
Department of Electronics and Information Engineering
Faculty of Engineering, Hokkai-Gakuen University
S 26-Jo, W 11-Chome, Chuo-ku, Sapporo, 064 Japan
E-mail:sasa,araki,momouchi@eeli.hokkai-u.ac.jp

KOJI TOCHINAI
Department of Electronics and Information Engineering
Faculty of Engineering, Hokkaido University
N 13-Jo, W 8-Chome, Kita-ku, Sapporo, 060 Japan
E-mail:tochinai@huxd.hokudai.ac.jp

ABSTRACT

The machine translation system can not translate the words which are not registered in its dictionary. The unregistered words are called the unknown words. One of problems on machine translation is that the machine translation system can not translate the unknown words. To resolve this problem, we do research on the automatic extraction method of the units from the pairs of words and their words for translation using inductive learning and the prediction method of words for translation of unknown words using the extracted units. Our methods will improve the precision of translation on machine translation system.

In this paper, we propose the extraction and prediction method. And we describe the result of evaluation experiment with 178 unknown words whose kinds are limited to a derive and a compound word with the system based on our methods. The correct rate of this experiment is 70.8%. And we apply our method to translate unknown English words into Japanese words for translation.

KEY WORDS: machine translation, inductive learning, prediction, extraction, word for translation, unknown word

1 INTRODUCTION

There are many researches of machine translation [1], [2], [3], [4]. Unknown words can not be translated using machine translation system, since they are not registered in its dictionary. It is one of problems on machine translation.

To resolve this problem, some researchers attempt to decrease the appearance of unknown words and some machine translation systems [4] use the dictionary which have quite a lot of vocabulary entries. However the perfect dictionary which register all of words can not be realized. The reason is that many new words are created by word-building capacity of natural language after the dictionary was made and they become unknown words. The another reason is that the work to realize the perfect dictionary is the very heavy labor and there is the limitation of hardware.

The alternative approach is that the field of translation is limited [5]. However the unknown words still appear. The reason is that many new words which are created will appear and become unknown words. Moreover this approach requires preparing different dictionaries whenever the fields of translation are changed. The labor to prepare many different dictionaries is very heavy.

The paper [6] reported that the appearance rate of unknown words in some corpora and described that the effective clue to classify the class of meaning is the information of morpheme. However it merely readjust the effective clue to classify the class of meaning.

This problem on machine translation can not have been resolved. To resolve it, we do research on the prediction method of unknown words. In this paper, the extraction method of units from the pairs of words and their words for translation and the prediction method of words for translation using extracted units are described. And the result of evaluation experiment with our method is described.

2 BASIC IDEA OF OUR METHODS

We consider that the units to predict words for translation are in the pairs of words and words for translation. Figure 1 shows that Word3 for translation is predicted using strings which are extracted from the pair of Word1 and Word1 for translation and the pair of Word2 and Word2 for translation.

The units which are extracted from English or Japanese character strings are called "a Piece of Word(PW)". And the pairs of the English PW and the Japanese PW are called "a Pair of Piece of Word(PPW)".

Inductive learning is defined as the algorithm to acquire rules or knowledges from some examples.
3 OUTLINE OF OUR METHODS

3.1 OVERVIEW

The outline of the system based on our methods is showed in figure 5.

After input of unknown English word, words for translation are predicted using PPWs in the PPW dictionary in the prediction process. And the correct word for translation is given in the case that erroneous prediction result was generated in the proofreading process. Then new PPWs are acquired from the pairs of English words and its Japanese words for translation using inductive learning in the learning process. After the prediction results are judged whether correct or incorrect, the values to rank the order of priority among

Figure 4: Example of PPW

the prediction results are dealt in the feedback process. Therefore the feedback process improve the capability of prediction.

3.2 PREDICTION PROCESS

The words for translation are predicted by PPWs which are in PPW dictionary in our method. Figure 6 shows that the Japanese word for translation is predicted by PPW-1 which is prefix and its word for translation and PPW-2. @1 in PPW-1 is rearranged with PPW2. Therefore the prediction result is generated and it is equal to correct word for translation.

Some results are generated in the case that there are some available combinations of the Japanese PWs. The results should be ranked to inform the most certain results of them. We made some condition based on the heuristics to rank the order among results. The conditions are as follows:

1. The less number of PPW in the result, the more certain it is.
2. The more average of appearance frequency of PPWs in the result, the more certain it is.
3. The more average of correct prediction frequency of PPWs in the result is, the more certain it is.
4. The less average of erroneous prediction frequency of PPWs in the result is, the more certain it is.
5. The more average of correct connection frequency of Japanese PWs in the result is, the more certain it is.
6. The less average of erroneous connection frequency of Japanese PWs in the result is, the more certain it is.

The correct prediction frequency of PPW expresses how many times the PPW has been used in the correct prediction and the erroneous prediction frequency of PPW expresses how many times the PPW has been used in erroneous prediction. The correct connection frequency of Japanese PW expresses how many times the Japanese PW has been used in the correct prediction and the erroneous connection frequency of Japanese PW expresses how many times the Japanese PW has been used in erroneous prediction.

3.3 LEARNING PROCESS
New PWs are acquired from the pairs of English and Japanese strings. However a lot of PWs are in the incorrect correspondence of meaning. Therefore we made one condition to extract English PWs and two conditions to extract PWs.

The condition to extract English PWs is that the length of common character strings is more than two characters.

And two condition to extract PWs are as follow:
1. both English and Japanese string in PAIR-1 is constructed by only one common character string which is extracted from Pair-1 and -2.
2. both Pair-1 and -2 are constructed by one common character string and each different character strings in Pair-1 or -2.

Figure 7 shows the example of PPW extraction applies to condition 1 and figure 8 shows example of PW extraction applies to condition 2.

3.4 FEEDBACK PROCESS
The capability of prediction is improved by the feedback process. The values which are dealt in the feedback process are as follow:
1. the appearance frequency of PPWs in the result
2. in case that the prediction result is correct
   (a) the correct prediction frequency of PPW in the result
   (b) the correct connection frequency of Japanese PWs in the result
3. in case that the prediction result is not correct

4. EVALUATION EXPERIMENT
4.1 INITIAL PW
The pairs of bases and words for translation are defined as PWs. Because they are extracted strings from the pairs of words and their words for translation. The pairs are defined as the pairs of entry words and their Japanese words for translation in the dictionary [7]. The dictionary [7] has 49,336 English entry words. 102,156 pairs of English words and their Japanese words for translation are extracted from it.

The pairs of affixes and words for translation are also defined as PWs. The book [8] was used to collect words which have same affixes. The pairs of affixes and their words for translation were extracted from the collected words. 351 PWs are extracted. They include 69 PWs with same prefixes and each words for translation and 282 PWs with suffixes and each words for translation.

The total number of these PWs is 102,507 and they are defined as the initial PWs.

4.2 DEFINITION OF UNKNOWN WORD
The unknown words are defined as words which are unregistered in the dictionary [7]. 178 unknown words whose kinds are a derivative and a compound word are extracted from 4,425 words in the category of press reportage on Susanne corpus [9]. Susanne corpus includes an approximately 130,000-word and samples each of the four genre groups. Press reportage is one of four genre groups.
4.3 PROCEDURE

Before this experiment, 102,507 initial PPWs were added to PPW dictionary. 178 unknown words were experimented. They were executed with the prediction process, the proofreading process in the case of the erroneous result, the learning process, and the feedback process.

4.4 STANDARD OF EVALUATION

The correct results are classified into two category. They are as follow:

group-1: the result is equal to the correct Japanese word for translation which is registered in the English-Japanese dictionary [10] or the only one part of result is rearranged with the equivalent string.

group-2: the result is partly equal to the correct Japanese word for translation.

The results which can not be classified two groups are erroneous results.

4.5 RESULT AND DISCUSSION

The number of correct prediction results is 126 and the correct rate in all results is 70.8%. 126 correct prediction results include 29 prediction results in the group-1 and 97 prediction results in the group-2. The rate of results in the group-1 is 16.3% and the rate of results in the group-2 is 54.5%. Figure 9 shows the prediction result which are classified to group-1 and figure 10 shows the prediction result which are classified to group-2.

The number of erroneous prediction results is 52 and the rate of erroneous prediction results in all of prediction results is 29.2%. There are some reasons that some erroneous results are generated. One reason is the insufficiency of essential PPWs to predict words for translation. The PPWs are used in our prediction method. If the essential PPWs are not in its PPW dictionary, the prediction process can not generate correct results. To resolve it, we should increase the number of acquired PPWs. We are investigating how to acquire more PPWs. The number of the erroneous prediction results owing to the insufficiency of the essential PPWs is 34 and the rate of the erroneous prediction results owing to the insufficiency of essential PPWs is 19.1%.

Another reason is that there are some compound words whose words for translation can not be made of words for translation of base. For example, there is word "break-up" and one of words for translation "破体 (Kaitai)". The both "break" and "up" can have neither "破 (Kai)" nor "体 (Tai)". Therefore the process can not predict correct result. We consider that the words for translation of such compound words need to be given in advance. The number of prediction results of such unknown words is 18 and their rate is 10.1%.

5 CONCLUSION

This paper proposes the extraction method of units "PPW" and "PPW" using inductive learning and the prediction method using extracted PPWs. And we describe the result of evaluation experiment and the rate of correct prediction results is 70.8%. Therefore we confirm our method is effective to predict Japanese words for translation of unknown English words whose kinds are a derivate and a compound word.

We plan to experiment with a lot of unknown words which are extracted from other categories on Susanne corpus [9].

REFERENCES