Pacific Association for Computational Linguistics

CONSTRAINT ON SOURCES OF UNITS FOR PREDICTION METHOD OF TARGET WORD USING INDUCTIVE LEARNING

Hisayuki Sasaoka[†], Kenji Araki[‡], Yoshio Momouchi[†], Koji Tochinai[‡]

[†]*Faculty of Engineering, Hokkai-Gakuen University,*
Minami 26 Nishi 11, Chuo-ku, Sapporo 064-0926 Japan.
E-mail: { sasa, momouchi } @eli.hokkai-s-u.ac.jp

[‡]*Graduate School of Engineering, Hokkaido University,*
Kita 13 Nishi 8, Kita-ku, Sapporo 060-8628 Japan.
E-mail: { araki, tochinai } @media.eng.hokudai.ac.jp

We have proposed the prediction method of target word in machine translation using inductive learning and confirmed the effectiveness of the method. However, in some cases, our system could not generate target words which fit the context of the field. This paper presents and evaluates a method of improving our proposed method. One of the improvements is the constraint on the selection of source pairs of the units for the prediction. In our method, the units are pairs of character strings in words and target words. And the source pairs for the units are pairs of words and target words. To extract the effective units from source pairs, we consider that the system needs to constrain on source pairs. In the extraction process, our system automatically ranks the source pairs and extracts the units from the source pairs which are ranked above the threshold used the information of common character strings. The threshold is gived by the user. This paper describes the results of evaluation experiments on this method. The number of correct results of the prediction for the target word on this method is larger than the number on our proposed method and we confirm that this method is effective.

Key words: machine translation, inductive learning, prediction, constraint, target word.

1. INTRODUCTION

Some recent researches have reported the method of bilingual alignment for the machine translation using the statistical method. Ahrenberg et al. (Ahrenberg 1998) have aimed the generation of a bilingual lexicon from aligned sentence pairs and reported the evaluation on two fields of text. Their method has based on the word-to-word model (Melamed 1997) and used the probability value of co-occurrences in the sentence pairs. Kitamura and Matsumoto (Kitamura 1997) have presented the automatic extraction method of translation patterns from aligned sentence pairs. Their method (Kitamura 1997) has based on the Dice coefficient and achieved the high rates of precision. These researches have focused the alignment on word level with pairs of sentences.

Some researches have studied statistical translation model. Alshawi et al. (Alshawi 1997) have described that the simple transducer models did not sacrifice accuracy at least for the limited domain application. Brown et al. (Brown 1993) have discussed five statistical models of translation process.

We consider that the system needs to use the basic units within words and target words in the translation process of words including technical terms and name expressions. We have proposed the prediction method of target word (Sasaoka 1997; Sasaoka 1998) using inductive learning (Araki 1995) and have called this method "**P**rediction method for **T**arget words using **I**nductive **L**earning (**PT-IL**)." And we have confirmed the effectiveness of PT-IL on four fields of Susanne corpus. The system has automatically extracted the pairs of the units from the source pairs. We call each unit "a **P**iece of **W**ord (**PW**)" and the pairs of units "a Pair of Pieces of Words (**PPW**)." The system on PT-IL has not yet achieved the high rate of effectiveness. The system should extract more effective PPW's. To extract more effective PPW's, we consider that the system needs to extract PPW's also from source pairs in English-Japanese dictionary. However the system might extract many ineffective PPW's in addition to the effective PPW's. Therefore we propose the constraint method on sources of PPW's and call the method "Prediction method for Target words using Inductive Learning with Constraints on Sources (**PT-ILCS**)". This paper describes the results of evaluation experiment on PT-ILCS.

This paper describes the basic idea and the outline of our experimental system in Section 2 and 3. In Section 4, we evaluate the system on our method and consider the results. And we conclude in Section 5.

2. BASIC IDEA

2.1. Prediction Units

In the extraction of PPW's, our system refers only the information of character strings. Figure 1 shows examples of PPW's. These italic character strings express Japanese phonograms. In the PPW's, the mark "@" means the variable. In the extraction of PPW's, the positions of the variables are equal to the positions of different parts of character strings in the source pairs. The system puts another character string into the position of the variable and then generates the new character strings.

Pairs of Word and Target Word

Word	Target Word
"diamagnetic,	反磁性体" (*han jisei tai*)
"ferromagnetic,	強磁性体" (*kyou jisei tai*)

⇓

Extracted PPW's

PPW 1	"@1 magnetic,	@1 磁性体 " (@1 *jisei tai*)
PPW 2	"dia,	反 " (*han*)
PPW 3	"ferro,	強" (*kyou*)

FIGURE 1. Examples of PPW's extraction

2.2. Information on PT-ILCS

On PT-ILCS, our system selects the source pairs of PPW's using the only information of character strings. Someone would think that the system could constraint on the selection of source pairs using the static knowledge, for example thesaurus. The knowledge needs to be generated and given to the system by researchers. Most of rule-based approaches have used the static information. However, the system using the knowledge would have some serious problems. One of them is the process of irregular cases in the knowledge. Moreover, the knowledge should have high quality and good balance between source and target languages. Therefore, in this research, our system does not use the static knowledge. To resolve these

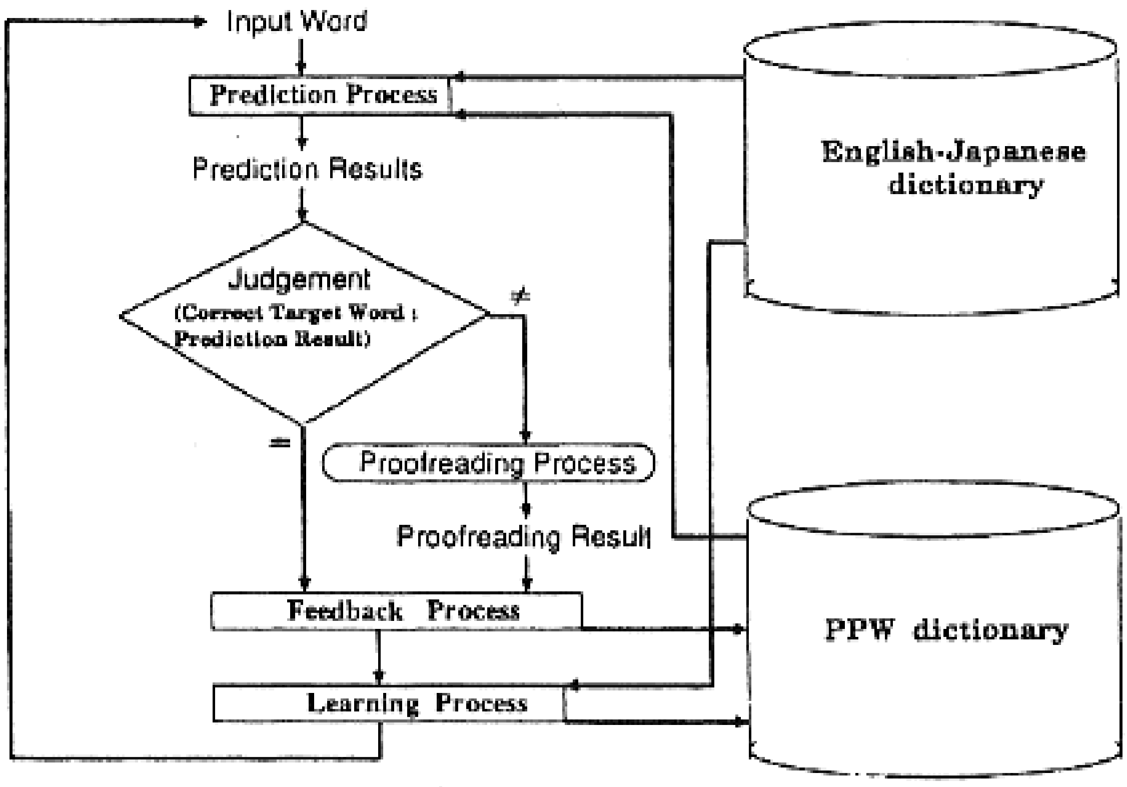

FIGURE 2. Experimental system

serious problems, we will consider the algorithm of the knowledge acquisition for the system in the near future.

2.3. Differences between PT-IL and PT-ILCS

There are two major different points between PT-IL and PT-ILCS. One of them is the constraint on the selection of source pairs of PPW's when the system extracts PPW's from source pairs in English-Japanese dictionary. On PT-IL, the system does not extract PPW's only among the pairs in the dictionary. However one of problems on PT-IL is the insufficient acquisition of effective PPW's for the prediction process. Therefore the system on PT-ILCS extracts PPW's among the pairs in the dictionary. To increase effective PPW's and decrease ineffective PPW's, the system on PT-ILCS limits the source pairs using the information of character strings. The details of this process appear in Section 3.2.

Another point is the definition of the correct target word for the system. On PT-IL, the correct target words are defined as the target words which are registered in English-Japanese dictionary. On PT-ILCS, we define the correct target word as fitting the context of the field. We consider that the definition of a target word increases the number of the effective PPW's and decreases the number of the ineffective PPW's.

3. OUTLINE OF OUR SYSTEM

3.1. Overview

Figure 2 illustrates the outline of our experimental system. Our system uses two dictionaries. One is the PPW dictionary and another is the English-Japanese dictionary. The system executes four processes, which are the prediction process, the judgment process, the learning process and the feedback process. The user executes the proofreading process. Our system translates English into Japanese.

The user inputs an English wors to the experimental system. At first, the system attempts to translate the word using PPW's in PPW dictionary. If the system cannot generate target words, it will attempt acquiring new PPW's from source pairs selected by the system and generate the target word using acquired PPW's. In the case that the system generates some prediction results, it decides the priority order for each prediction result with the numerical value of using PPW's. The referred numerical values are as follows:

A1 The number of appearances in the experimental data
A2 The number of appearances in the prediction results
A3 The number of appearances in the correct prediction results
A4 The number of appearances in the erroneous prediction results

The system values more A1, A2 and A3. On the other hand, it values less A4. On the results of prediction for the target word, the system refers A1, A2, A3 and A4 in sequence and determines the priority order among the results of prediction.

In the next step, the system compares the result of prediction for the target word with the correct target word. The correct target word is the word that appears in the context of the field. On the experiment, the user proofreads the result in the only case that the prediction result does not agree with the correct target word. Thereafter the system does the feedback process. In this process, the system calculates above A1 and A2. In the case that the system can predict the correct result, the system calculates above A3. In another case, the system calculates above A4. At last, in the learning process, it extracts PPW's and adds PPW's to PPW dictionary. We have defined the limitation for the PPW's that are added into PPW dictionary. This limitation is that the character strings of PPW need to construct other experimental data. The reason of the limitation is to increase the number of effective PPW's and to decrease the number of ineffective PPW's.

3.2. Constraint on PT-ILCS

In our method, the constraint on the selection of the source pairs for the prediction units are as follows:

1. Our system ranks the source pairs using the information of character stings.
2. Our system extracts the units from the source pairs that are ranked above the threshold.

In the decision of the priority order for the source pairs, the system ranks the source word with the larger number of character strings above the others. The number is the maximum number of matching the character strings of the word with the character string of the processing word. For examples, between the word 'electrical material' and the word 'material,' the maximum number of matching is 8. In another case, between the word 'electrical material' and the word 'mate', the number is 4. Our system regards the word 'material' as more effective source of PPW's. And it ranks the word 'material' above the word 'mate' in the priority order.

Our system uses the source pairs of PPW's which are ranked higher than the threshold for the priority order among the source pairs of PPW's. The user has defined the threshold in advance.

4. EVALUATION EXPERIMENTS

4.1. Data and Procedure

To confirm the effectiveness of PT-ILCS, we have done the evaluation experiments. The data of these experiments are the names of lectures, the names of research groups and the names of research fields in a faculty of engineering in one university. The number of data is 100. Table 1 indicates the classification according to the numbers of bases in English words among experimental data. Figure 3 shows the examples of experimental data.

The English-Japanese dictionary in our system includes the electrical dictionary "*Gene*" (Kubo 1995). The number of pairs of English and Japanese words in the dictionary "*Gene*" is 102,156. Moreover, we have extracted the pairs of base and the target word from the book (Maeda 1994). The number of these pairs is 351. The total number of pairs in initial English-Japanese dictionary is 102,507.

We have classified the results of the prediction as predictable results or unpredictable results. Among predictable results, we have defined that the correct results are character strings that are equal to the correct target word in the context of the filed and that are ranked within the 10th place in the order of priority among the prediction results. And, in other cases, we defined the results as the erroneous results. We have evaluated the experiment with the rates of recall and the rates of precision. The expressions of the rates are as follows:

$$recall[\%] = \frac{The\ \ number\ \ of\ \ correct\ \ results}{The\ \ number\ \ of\ \ experimental\ \ data} \times 100.0$$

$$precision[\%] = \frac{The\ \ number\ \ of\ \ correct\ \ results}{The\ \ number\ \ of\ \ predictable\ \ results} \times 100.0$$

TABLE 1. A table of classified experimental data according to the number of bases

The numbers of bases	The number of experimental data	[%]
1	0	0.0
2	41	41.0
3	35	35.0
4	13	13.0
More than 5	11	11.0
total	100	100.0

The number of bases:2

Word	Target Word
cryoelectronics:	低温エレクトロニクス (*teionn erekutoronikusu*)
plasma engineering:	プラズマ工学 (*purazuma kougaku*)
opto-electronics:	光エレクトロニクス (*hikari erekutoronikusu*)

The number of bases:3

Word	Target Word
optical-fiber engineering:	光ファイバ工学 (*hikari faiba kougaku*)

The number of bases:4

Word	Target Word
applled electric power laboratory:	電力応用研究室 (*dennryoku ouyou kenkyuusitsu*)

The number of bases:More than 5

Word	Target Word
biological information processing system lab.:	生体情報処理システム研究室 (*seitai jouhou syori sisutemu kennkyuusitsu*)

FIGURE 3. Examples of experimental data

4.2. Results

Table 2 indicates the numbers of the predictable and unpredictable results and Table 3 indicates the numbers of the correct and erroneous results in the experiment on PT-ILCS. And Table 4 shows the recall and precision of the results. On Table 3 and 4, the values within parentheses indicate the number and the rate for the first candidate among the rank of the results of the prediction.

Moreover we have done another experiment on PT-IL. The second experiment has been done on the same procedure as the first experiment. Table 5 indicates the recall and precision of the results in the second experiment. The progresses of rates between two experiments are represented in Figure 4 and 5.

TABLE 2. A table of numbers of the predictable and unpredictable results on PT-ILCS

Data	Predictable results	Unpredictable results
0	0	0
25	8	17
50	22	28
75	34	41
100	53	47

TABLE 3. A table of numbers of the correct and erroneous results on PT-ILCS

Data	Correct results	(First Candidate)	Erroneous results	Predicatable results
0	0	(0)	0	0
25	2	(0)	6	8
50	7	(3)	15	22
75	9	(5)	25	34
100	14	(8)	39	53

TABLE 4. A table of results on PT-ILCS

Data	Recall[%]	(First Candidate)	Precision[%]	(First Candidate)
0	0.0	(0.0)	0.0	(0.0)
25	8.0	(0.0)	25.0	(0.0)
50	14.0	(6.0)	31.8	(13.6)
75	12.0	(6.7)	26.5	(14.7)
100	14.0	(8.0)	26.4	(15.1)

4.3. Consideration

Figure 6 shows one example of correct results. The causes of erroneous results are as follows:

B1 The system used PPW's which have erroneous correspondence between source and target words.

Table 5. A table of results on PT-IL

Data	Recall[%]	Precision[%]
0	0.0	0.0
25	0.0	0.0
50	8.0	33.3
75	6.7	27.8
100	7.0	26.9

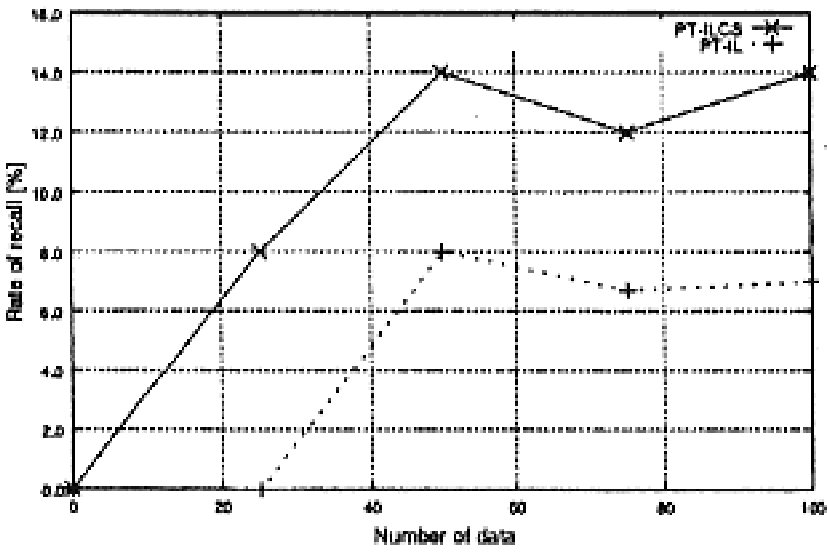

Figure 4. The comparison between the recall on PT-ILCS and the recall on PT-IL

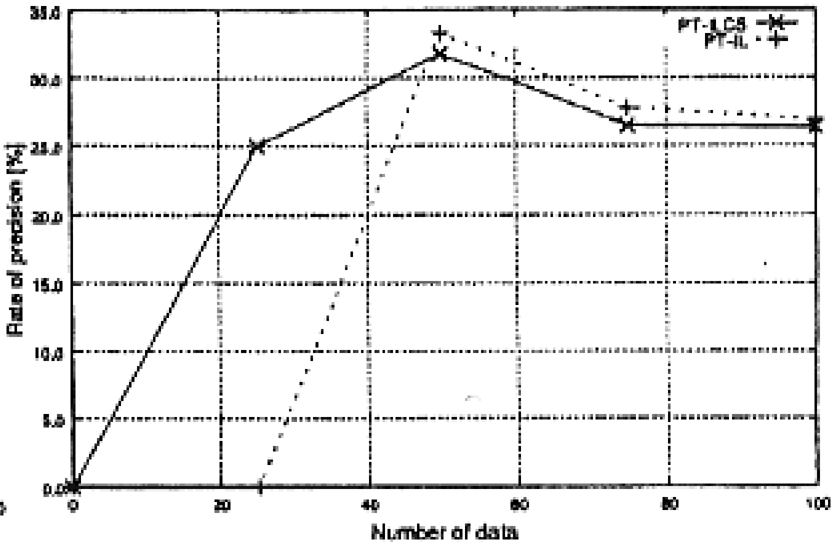

Figure 5. The comparison between the rates of precision on PT-ILCS and on PT-IL

B2 A part of the target word on the erroneous result differed from the part on the correct target word.

B3 The correct results were ranked under the 11th among results of the prediction.

We consider that our system extracted ineffective PPW's at the erroneous position between source and target words. Our system uses the only information of character strings. To decrease the number of erroneous results by the cause B1, the system needs to employ other types of the knowledge. However, the use of static knowledge has some problems. Therefore, we need to consider the algorithm for the knowledge acquisition.

An example of erroneous results by the cause B2 is 'gaseous electronics, 気体エレクトロニクス (*kitai erekutoronikusu*).' The result is constructed by the PPW's 'gaseous, 気体 (*kitai*)' and '@1 electronics, @1 エレクトロニクス (*erekutoronikusu*).' In this example, the correct target word is '気体電子工学 (*kitai densi kougaku*).' The reason of the extracted PPW 'electronics, エレクトロニクス (*erekutoronikusu*)' is that there are the pairs 'power electronics, パワーエレクトロニクス (*pawah erekutoronikusu*)' and 'plasma electronics, プラズマエレクトロニクス (*purazuma erekutoronikusu*)' among experimental data. Therefore, the system corresponded 'electronics' to 'エレクトロニクス (*erekutoronikusu*).' On this experiment, we have regarded these results as erroneous results. However, we regard such results as dependent on target words apprearing in the experimental data. From this, our system can adapt to the field of context.

Experimental Data
Word Target Word
electronic system engineering: 電子システム工学講座
(*densi sisutemu kougaku kouza*)
⇓
Appearing Expeimental Data
Word Target Word
electrical system engineering: 電気システム工学講座
(*denki sisutemu kougaku kouza*)
etc.

PPW Dictionary
English PW Japanese PW
system engineering: 気システム工学講座
(*ki sisutemu kougaku kouza*)
system engineering: システム工学講座
(*sisutemu kouzaku kouza*)
etc.

English-Japanese Dictionary
Word Target Word
electronic book: 電子ブック
(*densi bukku*)
electronic mail: 電子郵便
(*densi yuubin*)
etc.

PPW from English-Japanese Dictionary
English PW Japanese PW
electronic @0: 電 @1
(*den @1*)
electronic @0: 電子 @1
(*densi @1*)
etc.
⇓

Prediction Results
電 システム工学講座
(*den sisutemu kougaku kouza*)
電子 システム工学講座
(*densi sisutemu kouzaku kouza*)
etc.

FIGURE 6. An example of Prediction Process

An example of erroneous results by the cause B3 is 'laser engineering, レーザ工学.' The system ranked this result the 14th place in the priority order among prediction results. We need to improve the decision method for the priority order. And we consider that one method is the system would acquire and employ the rules for connections between PPW's.

5. CONCLUSION

This paper addressed and evaluated the constraint method for source pairs of PPW's on the prediction method of target word using inductive learning. Moreover, we considered the results of evaluation experiment. The number of correct prediction results on PT-ILCS is larger than the number on PT-IL. From the consideration, we confirmed the effectiveness of PT-ILCS. To rise the precision, we have to consider the acquisition method and the application method of the knowledge in the near future.

ACKNOWLEDGMENTS

This research was partially supported by the Grants from High-Tech Research Center in Hokkai-Gakuen University and by the Grants from the Government subsidy for aiding scientific researches (No.098780780 and No.10680367) of the Ministry of Education of Japan.

REFERENCES

AHRENBERG, L., and M. ANDERSSON, and M. MERKEL. 1998. "A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts." In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages29–35.

MELAMED, I. D. 1997. "A Word-to-Word Model of Translation Equivalence." In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages490–497.

KITAMURA, M., and Y. MATSUMOTO. 1997. "Automatic Extraction of Translation Patterns in Parallel Corpora."In Transactions of **IPSJ**, Vol.38, No.4, pages727–736.

ALSHAWI, H., A. BUCHSBAUM and F. XIA. 1997. "A Comparison of Head Transducers and Transfer for a Limited Domain Translation Application", In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages360–365.

BROWN, P., S. PIETRA, V. PIETRA and R. MERCER. 1993. "The Mathmatics of Statisticak Machine Translation: Parameter Estimation", Computational Linguistics, vol.19, pages263–312.

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-II*, pages1–7.

SASAOKA, H., K. ARAKI, Y. MOMOUCHI and K. TOCHINAI. 1997. "Prediction Method of Words for Translation of Unknown Words with Words and Words for Translation using Inductive Learning", In *Proceedings of PACLING-97*, pages282–292.

SASAOKA, H., K. ARAKI, Y. MOMOUCHI and K. TOCHINAI. 1998. "Evaluation of prediction Method of Target Word using Inductive Learning for Unknown Derivative Words and Compound Words." In Transactions of **IEICE**, Vol.J81-D-II, No.9, pages2146–2158.

KUBO, M. 1995. "*Eiwa-Waei Densaku Jiten:* Gene." *Gizyutsu Hyouron Sha.*

MAEDA, K. 1994. "*Tsuyokunaru Eitango.*" *Yuuseidoh.*