Credibility Evaluation of Candidates for Input Prediction Method

Wu Rina[*] and Kenji Araki[†]

[*] Graduate School of Information Science and Technology Hokkaido University
Sapporo 060-0814 Japan
Tel & Fax: +81-11-706-7389
E-mail: wrn@media.eng.hokudai.ac.jp
[†]Graduate School of Information Science and Technology Hokkaido University
Sapporo 060-0814 Japan
Tel & Fax: +81-11-706-6534
E-mail: araki@media.eng.hokudai.ac.jp

Abstract— **In this paper, we introduce a method of credibility evaluation of candidate for input prediction. It is very important how to give the best order of predicted candidates on the input prediction system. In our study, we make use of the Priority Evaluation Function (PEF) and the correlation of some related words to evaluate the prediction candidates and give the best candidate's order if possible. Experimental evaluation of the method has shown that the PEF function is effective.**

I. INTRODUCTION

In recent years, studies of the prediction input system have been done [1,2,3] for several languages. Generally, there are common problem in the input prediction system, that is, a user has to choice the correct one from prediction candidate. In the prediction process, it is difficult that the system decided the correct one from all of the candidates. Usually the system gives some prediction candidates for user's select. Therefore, the user hopes the correct candidate is contained in the upper part of the candidate's order. The ideal candidate's order is that the desired word is on the top of the order of candidates.

As for Chinese PinYin input system, how to decrease the keystrokes number and how to increase input speed is a long-time problems. We proposed the method of input prediction for Chinese PinYin input using the Inductive learning [3] [4]. In our study, the Inductive Learning [5] is defined as the method of knowledge acquirement capability from the inputted Chinese sentences by comparing a pair of the inputted sentences and extracting different parts and common parts recursively. Our system based on Inductive Learning approach can acquire new rules under any situation by its learning capability even if the rule dictionary is empty at the beginning. Furthermore, the system can save a great deal of labor for completing a corpus.

There are two major processes in our proposed method. That is the Inductive Learning process and the Input Prediction Process. For the Input Prediction Process, there are two steps. First step is referring to the prediction dictionary and getting all of prediction candidates. Second step is to give a suitable candidates order by using some algorithm. The first step depends on learning process, because the prediction dictionary is generated in the learning process. The number of rules of prediction dictionary can be increased by improve the capability of the learning process. And the system can present a large number of prediction candidates. However, increasing candidate's number will bring a problem such as the correct candidate moves to low rank of candidate's order. Therefore, it is necessary that the system must have a capability to give a suitable order for candidates. Because of, as an effective input prediction system, it is not desirable that the desired candidate down to the bottom of the candidate's order. Therefore, how to make the predicted candidates into correct order is very important. The best candidate's order is that the desired word is on the top of candidate's order. An effectively prediction process needs the system can give a correct order.

Generally, we use the Priority Evaluation Function (PEF) to evaluate the candidate and decide the order of all candidates. The details of the PEF function are described in section 3.1. In the PEF function, we only use the information of the mutual frequency between reference word and prediction word. Other variables are not related to the current input sentence. However, through the research we find that, the words including in the same sentence except reference word is useful for prediction. We make use of related word to get the best order of prediction candidate. In our study, the related word is defined as two words frequently appear in the same sentence and those words must compose more than two Chinese characters.

In our study, the system generates the prediction dictionary by using the Inductive Learning. At first, the system compares two sentences and divides them into different part and common part. Then, if two sentences matched at least two places at the same time, the system abstracts the common part and put it into the related words dictionary. Therefore, we obtain the related word only using the surface information of the text.

II. RELATED WORD

A. *A Formal view on related word*

Generally, the related word has the broad meaning definition, such as Synonym, Antonym. Related word is widely used in the field of natural language processing [6]. In [6], the related word is used to WWW search. It is very meaningful that uses the related word effectively for natural language processing.

B. Acquisition of related words

In our study, we generate the related word dictionary using the Inductive Learning by comparing two sentences belonging to user corpus. The user corpus contained all of text inputted by the user. In this process, it can get the useful information adapted to the user, because all of inputted sentences are used for learning to generate the related word dictionary.

TABLE I
AN EXAMPLE OF ACQUISITION OF RELATED WORD

Sentence 1	xxx W1 xxxxxx WA xxxxxxxx
Sentence 2	xxxxx W1 xxxxxxx WA xxxxx
Sentence 3	xxx W1 xxxxxx WB xxxxxx
Sentence 4	xxxx W1 xxxxxx WB xxxxx
Sentence 5	xxx W2xxxxxxx WB xxxxx
Sentence 6	xxxxxxx W2 xxxxxxxx WB xxxxx
Sentence 7	xxxxxxx W2 xxxxxxxx WC xxxxx
Sentence 8	xxxxxxx W2 xxxxxxxx WC xxxxx
Matching segment	(W1 WA) (W1 WB) (W2 WB) (W2 WC)
Related word	W1 (WA # f #, WB # f #) W2 (WB # f #, WC # f #)

The example of related word acquisition shows in Table 1. In the Table 1, "x" shows some Chinese character, "W1, W2, WA, WB, WC" shows the common part of two compared sentence. As show in Table 1, compare two sentences, when those sentences matched in two places and all of the common part is long than 2 Chinese characters, we abstract the matched parts as a related word. At the same time, appearance frequency is added to related word. Initial value of appearance frequency is 1.

C. Correlation of the related word

Appearance Frequencies (AF) of related words is used to calculate the correlation of the related words.

In our study, Appear Frequency is the number of times of the related word appeared in the same sentence at the same time. The AF means that, two words appears in same sentence frequently means those words have strong relationship.

III. EVALUATION OF PREDICTION CANDIDATES

We decided to use Priority Evaluation Function and correlation of Related Word to evaluate the prediction candidates and give the best order of candidates if possible.

D. Priority Evaluation Function (PEF)

In our study, each rule of the prediction dictionary has its priority value, which is calculated by the Priority Evaluation Function (PEF), and PEF function defined as below.

$$PEF = \alpha \times A - \beta \times B + \gamma \times F + L. \qquad (1)$$

A: Correct prediction frequency
B: Erroneous prediction frequency
F: The rule appears frequency
L: A number of Chinese characters in the rule
α,β,γ: Coefficients

Function (1) means that the rule in the prediction dictionary will be keep a high degree of priority, when it is in the high correct prediction rate and in the low erroneous prediction rate and also frequently appears in the text.

About the correct prediction frequency A: When the predicted candidate is the correct input word, the value of A for the candidate increase by one.

About the erroneous prediction frequency B: For all of the prediction candidates except correct one, the value of B for those candidates increase by one.

E. Evaluation of prediction candidates

It is clear that, frequently appears in the same sentence at the same time, the two words have some relationship. And also, according to definition of the PEF function, all of the variables contained in the function are related with the rule of prediction dictionary. However, it is very important and very difficult how to unite the information of related words and the PEF function.

In our study, we have combined the PEF function and the correlation of the related word in a line type as defined in function (2).

$$V=PEF+C*AF \qquad (2)$$

V: Credibility of prediction candidate
PEF: Priority Evaluation Function
C: Coefficient
AF: Appear Frequency of related word

IV. EXPERIMENT

F. Goal and data of experiment

We carry out some experiments to certify the efficiency of the method as mentioned below. We obtain the data of experiment from the internet[1].

A. Preliminary experiments

At the first step in this study, we carry out a preliminary experiment to find the most suitable values of α, β, γ, and C.

[1] The data of experiment is published on the web pages as follow: www.zjzw.net, www.hncnlp.com, www.ahetc.gov.cn/.

TABLE 2
Result of preliminary experiment

α	1			**2**			5			10		
β	2	7	12	2	**7**	12	2	7	12	2	7	20
γ	3	3	5	1	**3**	3	3	1	3	3	5	3
P(%)	72.31	72.58	70.74	72.53	**73.11**	73.00	72.64	70.89	70.72	71.21	71.03	70.15

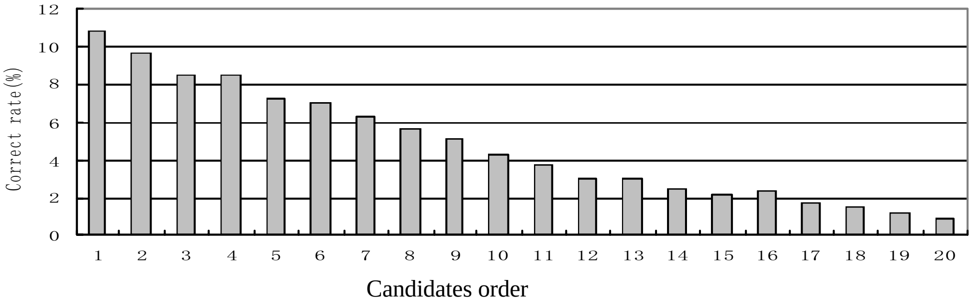

Fig. 1. Distribution in the correct rate.

Like mentioned above, the functions (1) and (2) are used to evaluate the prediction candidate and to give the user a suitable order of candidates. It means that, all of those coefficients are related to the candidate orders. Therefore, the most suitable values of coefficient are the value that gives the prediction candidates the best order.

Pre-experiment 1: Decide the value of α, β and γ

We have used the Greedy method to calculate the correct rate of the prediction candidates that move into front ten inside. In the experiment, the "α" changes the value between "1, 2, 5, 10, 20", the "β" changes the value between "2, 7, 12, 20", and the"γ" changes the value between "1, 3, 5, 9". Then 80 (5*4*4=80) sets of data are used to calculate. We chose 12 sets of better results to show in Table 2. In the Table 2, P is the probability of correct candidate that the position number is smaller than 10. The value of C is established with zero in this experiment.

We decide the values of coefficient α, β and γ according to the preliminary experiment result, that is: $\alpha=5, \beta=7, \gamma=2$, because the P is in the highest value in this case.

Pre-experiment 2: Decide the value of C

After the values of α, β and γ are decided, we use the same method to find the most suitable value for coefficient C. In the experiment, the C changes the value between "1 ~ 10", and at the end of experiment, the C used the value of 3.

B. Effectiveness of PEF function

In the first experiment, we only use the PEF function to evaluate the prediction candidates and calculate the efficiency of PEF function using the Correct Rare (CR). Correct Rate defined as:

$$CR = \frac{np}{NC} \tag{3}$$

CR: Correct rate
np: Number of correct prediction candidate on the Position
NC: Total Number of Correct prediction candidate

According to the function (3), the CR is shown the distribution of correct prediction candidate in the candidates order.

Experimental results show in Figure 1. In the Figure 1, the number of side shaft shows the position of the candidates. In our prediction system, the number of 1 is top of candidate order. The vertical axis of Figure 1 shows the correct rate. For example, the value of correct rate at the position number of 1 is about 10.9%. It means the probability of correct candidate moves to the top of candidate order is 10.9%.

C. Influence of the related word

Then, we use both of the PEF function and the correlation of related word to evaluate the prediction candidates and calculate the influence of the related word. Experimental results show in Table 3. . In the experiment, the number of correct candidate is 337, and between correct candidates, 145

TABLE 3
INFLUENCE OF THE RELATED WORD

Items	Number of correct prediction candidate (only using PEF)	Moved to the high rank (after using the related word)	Moved to the Low rank (after using the related word)	No change (after using the related word)
Sum	145	83	40	22
Rate (%)	43.0	57.2	27.5	15.1

of correct candidates have the position number smaller than 10. In Table 3, the rate in the bottom row shows the changed level when using the correlation of related word. For example, the number of correct candidate moved to the high positions is 83, and it divided by 145 is 57.2. It means that, over half of correct candidates have moved to the high position. Therefore, the order of candidate is influenced by the related words.

G. Results of experiment

The final result of section 4.1.2 shows in Table 4. In the Table 4, the item of candidate's position means that the correct prediction candidate is contained in the range that shows in Table 4's first row. According to the experimental results as show in Table 4, the change of the correct rate in the range of "1-5" is 21.8%, and the change of the correct rate in the range of "11-20" is -23%. It means that, when using the PEF function to evaluate the prediction candidate, about 23% of correct candidates move into front ten inside of the candidate's position range. Therefore, the PEF function is effective.

TABLE 4
FINAL RESULT OF EXPERIMENT

Range of Candidate's positions	1-5	6-10	11-20	Sum
Correct rate At the position's range (%)	45.3	24.7	24.1	94.1
Average correct rate (%)	23.5	23.5	47.1	94.1
Changes (%)	21.8	1.2	-23	_

H. Discussion

According to the result of experiment, the related word is useful in evaluation of the prediction candidates. However, there are some negative influences, that is, some times the position of desired candidate falls down when using the information of related words. It is that, the correlation of the related word not only concerns with the appearance frequency of related words, but concerns with other variable such as distance and word number so on.

V. CONCLUSION AND FUTURE WORKS

During the study of input prediction method for Chinese PinYin input, we noticed that the evaluation of prediction candidates is very important, and give a suitable candidate's order is a difficult problem. We utilize the Priority Evaluation Function and the information of the related word to evaluate the prediction candidate and try to get a best order of candidates. The experimental results have showed that the order of prediction candidates has been changed to more suitable when using the information of related words.

Although the effectiveness of related word is confirmed, much further work is needed. We will focus on calculation of correlation for related words using the distance of two related words and the word numbers so on. And also, we must search a best way to unite the correlation of the related words and the PEF function effectively.

REFERENCES

[1] Liu Changsong, Wu Zhenjun, Qiao Chunlei, Li Yuanxiang: Intelligent Association for Chinese Input Using Statistical Method. *Journal of Chinese Information Processing* Vol.14 No.1.pp32-38.(1999).

[2] Masui, T. Integrating Pen Operations for Composition by Example. *In proceedings of the ACM symposium on User Interface Software and Technology* (UIST'98) (November 1998), ACM Press, pp.211-212.

[3] Wu Rina, K. Araki and K. Tochinai, Assistant Chinese Input Method Using the Association of the Input Words Which Acquired by Inductive Learning, *In proceedings of the Info symposium of Hokkaido* 2002,(April 2002), pp139-140.

[4] Wu Rina, K. Araki and K. Tochinai, A Method for Intelligent Association of Chinese Input Using Inductive Learning, *Proceedings of International Conference on Information Technology & applications Information Technology & Applications ICITA 2002), 234-17, 25-28 November 2002,* BATHURST, AUSTRALIA.

[5] K. Araki and K. Tochinai, Effectiveness of Natural Language Processing Method Using Inductive Learning, *Proceedings of the IASTED International Conference ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING*, pp.295-300, May, 2001, Cancun, Mexico.

[6] M. Harada, S. Shimizu, a Simple Way of Guidance: Making Relevant Keyword From Anonymous User Behavior on WWW Search, NTT Software Labs.