# A Method for Intelligent Association of Chinese Input Using Inductive Learning

**Wu Rina, Kenji Araki, *Member, IEEE*, and Koji Tochinai**

***Abstract*--We introduce an efficient Chinese PinYin input technique of intelligent association using the Inductive Learning. As for the Chinese PinYin input system, a large number of keystrokes and the selection from candidate words limit the speed increasing. Association of input words is one of the most important methods for Chinese character input to decrease the number of keystrokes and to improve the input speed. Our proposed method enables us to promptly input the Chinese words into a computer, that is, not necessary to spell complete character of PinYin only by choosing the desired words from candidates. Moreover, the Inductive Learning approach gives the system based on our method two characteristics. One of them is the dynamically adaptation to current situation, and the other is the capability of acquiring rules even if there are no initial rules. Furthermore, we make use of the relationship between Chinese words or characters in context to generate a new rule. We call it the rule generation process. The average correct association rate is up to 44.3% in the experiment for the performance evaluation. It shows that this method is effective for Chinese input.**

**Keywords: Candidate Rule, HanZi, Intelligent Association, the Inductive Learning, IL-IA, PinYin, Rule Generation**

## I. Introduction

In recent years, the researches of the input system are done since the computer comes into wide use. In Chinese, there are two kinds of characters called PinYin and HanZi. PinYin represents the pronunciation of the HanZi and usually consists of several alphabets. HanZi is Chinese character for expressing Chinese sentences and has several thousands kinds. Therefore, we have to develop an input method to input Chinese text into a computer by keyboard. Generally, two kinds of technique are used for input the Chinese character into a computer, that is: PinYin input method and HanZi character input method. The PinYin input method is not only the most common but also the widest used method of Chinese input and almost occupies 93% of all input methods. The HanZi character input method is mainly used by typewriter due to two reasons as mentioned below. First, the method enables a user to promptly input the Chinese text into a computer, if the user is extremely skilled in keyboard. Second, the user needs to memorize all kinds of the keyboard distribution for radical of HanZi□characters and it is difficult for a normal user. An ordinary procedure of PinYin input system is as follows:

*A. A user inputs a PinYin string corresponding to the pronunciation of a HanZi character using the keys of twenty-six alphabets.*

*B. A system translates the PinYin string into several HanZi characters which is in the same pronunciation different patterns using PinYin-HanZi translator.*

*C. The user selects the desired HanZi character as an input word by hand.*

Since multiple HanZi characters often have the same pronunciation, the user needs to choose the correct one by hand. There are many reasons why it is very slow to input a Chinese text into the computer using PinYin input method. First of all, a PinYin of a Chinese character is generally spelled by several alphabets. Second, the user must choose the correct one from all of HanZi, which is in the same pronunciation. The association of input word is the most commonly used for the input system to decrease the number of the keystrokes as an effective method. As a beginning for consideration with association approach, the word dictionary for association has to be complete. However, it is difficult to complete a dictionary which can satisfy all of users and all kinds of fields, that is, the quality of dictionary is limited by developer's personal knowledge. It is necessary to make a huge dictionary in order to increase the correct association rate. Therefore, a method of intelligent association for Chinese input on a computer based on statistical approach was proposed to solve the previous problems[1]. In this research, it takes much labor to complete the huge corpus for statistics. The performance of such methods is limited by the quality of the used corpus. And also, it is necessary to complete a huge and high quality corpus in order to increase the precision of the system. In addition, it is difficult to make a corpus to adapt to all kinds of field and all kinds of users, since the new words and the new technical terms in every field come out increasingly nowadays.

The "Intelligent ABC Input Method [1]" is a commercial system as an effective technique for Chinese text input on the

Wu Rina is with Graduate School of Engineering, Hokkaido University, Kita 13 Nishi 8, Kita-ku, Sapporo-shi, 060-8628 Japan (fax: +81-11-709-6277, e-mall: wrn@media.eng.hokudai.ac.jp).

K. Araki is with Graduate School of Engineering, Hokkaido University, Kita 13 Nishi 8, Kita-ku, Sapporo-shi, 060-8628 Japan (fax: +81-11-709-6277, e-mall: araki@media.eng.hokudai.ac.jp).

K. Tochinai is with Graduate School of Business Administration, Hokkai-Gakuen University Asahimachi 4-1-40, Toyohira-ku, 062-8605, Sapporo, Japan (fax: +81-11-841-1161, e-mail: tochinai@econ.hokkai-s-u.ac.jp).

ISBN:1-86467-114-9.

[1] It is published in http://www.znabc.com.

computer by PinYin. In this system, a user has to input whole spelling of PinYin which corresponding to the pronunciation of HanZi characters. Then, the system refers the words dictionary to associates inputting words. If there are several candidate words for selection, a user needs to selects the correct words from candidates correctly if the inputting words exist in the candidates. However, in this system, the increasing speed is limited because it is not only necessary to spell all of the PinYin characters but also select the correct one from candidate words. The problem that we have to consider is how to decrease the number of keystrokes to increase input speed of PinYin input system. We consider an effective technique for getting candidate words from previous input. It is called the method of Intelligence Association for Chinese input using Inductive Learning, and the method is described as IL-IA in this paper.

According to past research[2][3][4], the Inductive Learning approach is effective for natural language processing. In our research, we pay attention to the advantage of association technique for Chinese input system, and introduce a new method for inputting words association using the Inductive Learning approach. In our research, the Inductive Learning is defined as the method of knowledge acquirement capability from the inputted Chinese sentences by comparing a pair of the inputted sentences and extracting the common parts and the different 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 initial. Furthermore, the system can save a great deal of labor for completing a corpus.

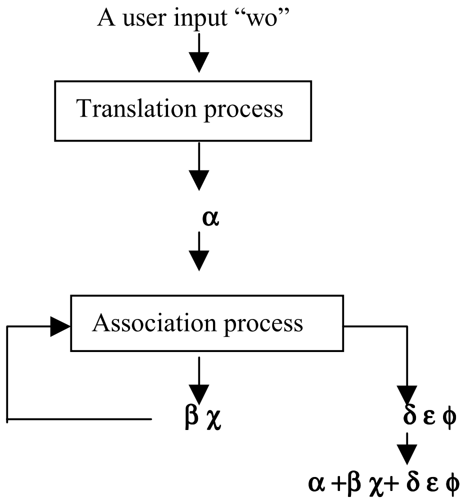

Figure 1: An example of procedure

In this paper, we describe the detail of the system based on our proposed method and the experimental result. Figure 1 shows an example of procedure in our system. In the case of the input HanZi string as follow:

"α β χ δ ε φ. (We are foreign students.) [2]"

The corresponding PinYin of the HanZi character is "wo men shi liu xue sheng." As show in Figure 1, a user inputs the "wo" which is the PinYin of the first HanZi character "α", and thesystem translates it into the HanZi character "α" in the translation process. Then the system associates the next input word "β χ" using the character "α", if the word "β χ" exists in the rule dictionary. At the same time, the system continues to associate the next input word "δ ε φ" using the characters "β χ", if it also exists in the rule dictionary. Finally, the system outputs the string "α β χ δ ε φ" when the association process has succeed. However, if the association process fails, the user needs to input the next character's PinYin "men", and then the system repeats the process as mentioned above.

## II. Outline

Figure 2 shows the outline of our method. With our approach, a user can select the desired words from the candidate words that obtained from the rule dictionary. The rule dictionary is generated in learning process by Inductive□Learning. Because, in the learning□process, all of inputted sentences are used for learning to acquire rules, the system can dynamically adapt to a current situation.

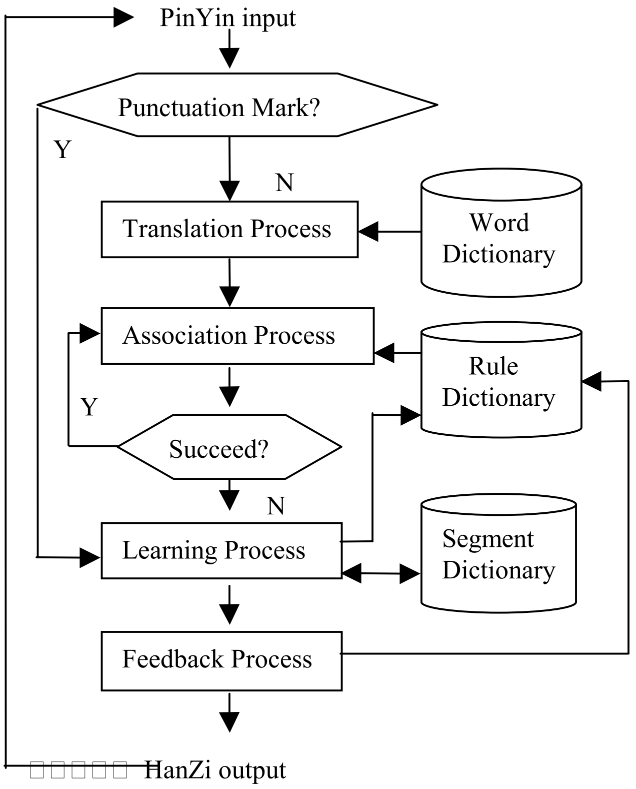

Figure 2: Outline of our method

[2] We use the code "α β χ δ ε φ "instead of the HanZi character in this paper, because the paper can't contain the Chinese HanZi characters.

When a user input the PinYin characters of a word using twenty-six alphabets, it is converted into HanZi characters in the translation process by referring to the word dictionary, which is provided at the beginning. If there are several translation results, the user needs to select correct one from candidates. In the association process, candidate rules are selected automatically from the rule dictionary by the system. And then the system infers a rule which is in the highest priority degree. If the rule is erroneous, the user needs to select the desired rule from candidates, if the desired rule exists in the candidates. When a sentence is inputted, the learning process is performed. In the learning process, a pair of previously inputted sentences is compared and the both of common segments and different segments are extracted in order to generate rule in the learning process. The rule of the rule dictionary has a priority value, which is described later in section 3.

## III. Process

The system based on our proposed method consists of four processes that are translation process, learning process, association process, and feedback process.

### A. *Translation process*

When a PinYin of a HanZi character is inputted, the system translates it into HanZi character by refering to the word dictionary. The word dictionary includes 6,890 basic Chinese HanZi characters and is completed at the beginning. Generally, there are several different HanZi characters in the same pronunciation. Therefore, the user needs to select the correct one from candidates of HanZi character for increasing the correct association rate.

### B. *Learning process*

In the learning process, the system uses the Inductive Learning to acquires rules from a pair of inputted sentences. Table 1 shows an example of extraction from two compared sentences. When a sentence is inputted, the system compares this sentence with all of the inputted sentences which is in the same text, and extracts both of the common parts and the different parts. We call the common part as Common Segment(CS), call the segment in front of CS as Front Segment(FS), and call the segment behind the CS as Behind Segment(BS).

Table1: An example of extraction

| | |
|---|---|
| Sentence1<br>Sentence2 | α β χ δ ε ϕ γ η ι φ κ<br>λ μ ο ε ϕ γ η π θ ρ |
| CS | ε ϕ γ η |
| FS1<br>FS2 | α β χ δ<br>λ μ ο |
| BS1<br>BS2 | ι φ κ<br>π θ ρ |

#### 1) *Acquisition of new segments*

If the number of HanZi characters of CS is less than five, the CS registered into the segment dictionary as a new segment. If the number of HanZi characters of the CS is greater than five, the CS is divided into some new segments using the segments of the segment dictionary. The new segments shows in Table 2 are registered into the segment dictionary and will be used in the rule generation section as mentioned below. The number of HanZi character of the segment in the segment dictionary is more than 1.

Table2: An example of segment acquisition

| | | |
|---|---|---|
| CS | α β χ δ ε ϕ γ | ο π θ |
| Length | > 5 | < 5 |
| S | δ ε | |
| New S | α β χ , ϕ γ | ο π θ |

#### 2) *Acquisition of rules*

The system can acquires rules using FS, CS and BS. The condition and the type of rules show in Table 3. We use the five HanZi characters for the limit number of FS, CS and BS. There have two reasons why use five HanZi characters as a condition. First, a great deal of idioms combined by four HanZi characters in Chinese language is usually used with a particle, which is consists of one HanZi character, and the idiom frequently appears in the Chinese text. Second, a huge number of words in Chinese language generally combined by two or three HanZi characters, and the phrase usually consist of two words.

We call the number of HanZi characters of FS as L (FS), and call the number of HanZi characters of BS as L(BS). The system acquires a new rule by using the procedure as mentioned below.

a) *When the both of L (FS) and L(BS) is less than five, the rule formed in FS+CS+BS is acquired.*

b) *When the both of L (FS) and L(BS) is greater than five, the rule formed in CS is acquired.*

c) *When the L (FS) is greater than five and the L (BS) is less than five, the rule formed in CS+BS is acquired.*

d) *When the L (FS) is less than five and L (BS) is greater than five, the rule formed in FS+CS is acquired.*

All of the acquired rules are registered into rule dictionary and divided into two layers, as shows in Table 3.

Table3: Types and conditions of a rule

| L( FS)<br>L(BS) | 0<L(FS)<=5 | 5<L(FS) |
|---|---|---|
| 0<L(BS)<=5 | FS+CS+BS | CS+BS |
| 5<L(BS) | FS+CS | CS |
| Layer | Second | First |

Each rule of the rule dictionary has its priority value, which is evaluated by the Priority Evaluation Function (PEF), and PEF defined as:

$$PEF = \alpha \times A - \beta \times B + \gamma \times F + L \text{……(1)}$$

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

At the first step in this research, we carry out a tentative experiment to obtain the values of α, β and γ, that is:
α =5,β =7, γ =2.

*3) Rule generation*

When the L (FS) or L (BS) is greater than five, the system can generate new rules by using FS (or BS) and the S (segments of the segment dictionary), as shows in Table 4. At the same time, the FS (or BS) is divided into new segments using S, and the new rules are generated using those new segments. If there are several S that can match to FS (or BS), the system infers the certainty of a new rule using both of the frequency and the number of HanZi characters of the matching segment, and also using the position of matching segment in FS (or BS). The **P**osition of the **M**atching **S**egment is called as PMS and defined as:

$$PMS = \frac{|\,Lfs - Lbs\,|}{Lfs + Lbs} \quad \text{…… (2)}$$

Lfs: The number of HanZi characters of the segment in front of matching segment
Lbs: The number of HanZi characters of the segment behind the matching segment

The procedure of system's inference for the certainty of new rules is as follows:

a) *Whether the value of frequency of the matching segment is the highest or not.*

b) *Whether the value of PMS is the smallest or not.*

c) *Whether the number of HanZi characters of the matching segment is the most or not.*

And also, the generated rules must satisfy the conditions show in Table 3. All of the rules generated in the rule generation section are registered into the rule dictionary and divided into two layers, for speedy selection in association process, which is described in section C.

Table4: An example of rule generation

| FS (or BS) | ξ ψ ζ υ ω ο π θ λ μ φ |
|---|---|
| S | ω ο π θ |
| New rule | ξ ψ ζ υ+ ω ο π θ+ λ μ φ |

Since all of the sentences of current inputted text are used for learning, the system can dynamically adapts to current situation rapidly.

## *C. Association process*

In the association process, the system associates some next input words using the information of previous inputted HanZi characters by referring to the rule dictionary. The rule dictionary is generated in the learning process. And some rules are inferred as candidate rules according to previously inputted HanZi characters. The association procedure is as mentioned below.

1) *The system uses final inputted 1 and 2 HanZi characters refer to the first layer of rule dictionary respectively, and gets some candidate rules formed in CS+BS and CS.*
2) *The system uses final inputted 3, 4, and 5 HanZi characters refer to the second layer of the rule dictionary respectively, and also gets some candidate rules formed in FS+CS+BS and FS+CS.*

In both of two steps, the system can find some candidate rules for selection. At the same time, the system arranges the candidate rules in order of their priority values, and then the system infers a rule which is in the highest priority value as next input words. Moreover, the system outputs all of candidate rules for user's selection. If the inference of the system is correct, the correct association frequency (described as A) of the rule increases by one. If the inference of the system is erroneous, the erroneous association frequency (described as B) of the rule increases by one, and the user has to selects the correct rule from the candidates, if it exists in candidate rules. In addition, the correct association frequency(A) of the rule which is selected by the user increases by one. The system performs all of the works by its inference and learning capability.

## *D. Feedback process*

In the feedback process, the system updates the degree of priorities of the rules in the rule dictionary by changing the values of A, B and F. In the association process, if the system's inference is correct, the value of A for this rule increases by one. And if the system's inference is erroneous, the value of B for this rule increases by one. When the user selects the rule from candidates, the value of A for the rule increases by one. Moreover, in the learning process, when the same rule is acquired, its value of F increases by one. Because the repetition of rule means that the rule has high certainty degree as a Chinese phrase which is used frequently in Chinese text. As mentioned above, the system is upgraded by the repetition of the feedback process.

## IV. Evaluation Experiment

### A. *Goal and data of experiment*

We carry out some experiments to certify the efficiency of the system based on our proposed method as mentioned below.

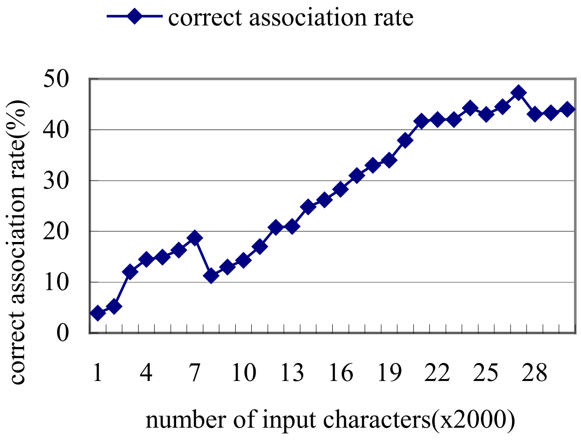

Figure 3: The relation between CAR and the number of inputted character

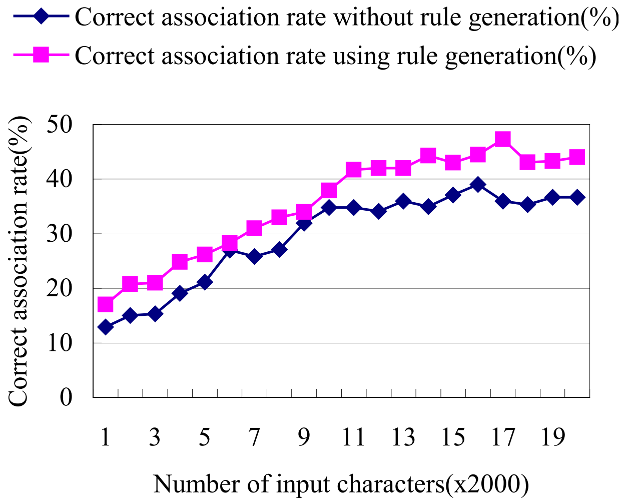

Figure 4: The correct association rate comparing

#### 1) *Association capability*

We calculate the correct association rate of our system for certifying the association capability, using the function of **C**orrect **A**ssociation **R**ate**,** which is described as CAR and defined as:

$$CAR = \frac{NC}{NA} \quad \ldots\ldots.(3)$$

NC: The number of all correctly associated HanZi characters

NA: The number of all inputted HanZi characters

And also we calculate the changes in the correct association rate when the system performs the rule generation application in the learning process.

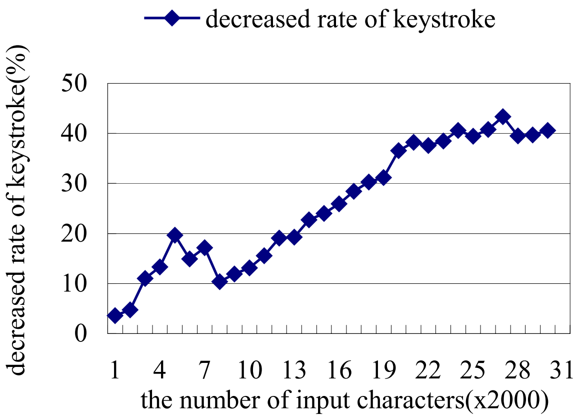

Figure 5: The relation between DKR and the number of inputted character

Table 5: Final result of experiment

| Number of HanZi characters | Number of acquired rules | CAR (%) | DKR (%) |
|---|---|---|---|
| 60,000 | 7,976 | 44.3 | 406 |

#### 2) *Decrease of keystroke*

We calculate the decreasing rate of keystroke of our system for certifying the inputting speed increase, using the function of

**D**ecreasing **R**ate of **K**eystroke, which is described as DKR and defined as:

$$DKR = CAR(1 - \frac{1}{\omega \times \mathrm{AL}}) \ldots\ldots (4)$$

AL: Average HanZi character's number of correctly associated rule

ω: The alphabet character's number of PinYin for each HanZi character

According to the function (4), the DKR is in direct proportion to the CAR. In this experiment, we use the average value of ω□that is ω=3.1.

#### 3) *Adaptation capability*

We calculate the adaptation capability of our system by changing the fields of text frequently.

The rule dictionary is empty at beginning. The data used in this experiments obtained from the Chinese Internet Web

Pages[3]. The number of all HanZi characters in the data is nearly 60,000 and there are 3 kinds of fields.

In the experiment, we limit the candidate rule's number as 20, because too much candidate rules for selection will influence the inputting speed.

### B. *Experimental results*

The results of the experiments show in Figure 3, 4, 5 and Table 5.

As show in Figure 3 and 5, the correct association rate of system has close to 45%, and the decreased rate of keystroke has close to 40% when the number of inputted HanZi characters had achieved 44,000. Figure 4 shows the result of correct association rate when the system performs the rule generation application, and the number of inputted characters is from 20,000 to 60,000. As show in Figure 4, the increase of average correct association rate is about 4%. Table 5 shows the final result of experiment. When the number of all input characters is close to 60,000, the number of rules in rule dictionary is 7,976, and the average correct association rate at final 1,000 characters is 44.3%, and the average decreasing rate of keystroke is 40.6% finally. According to Table 5, the final experimental result can prove the rule acquisition capability by the Inductive Learning, which uses the previous inputted characters under the condition without initial rules. And also can prove the efficiency of the intelligence association capability. According to the result shows in Figure 4, the rule generation capability is effective in increasing the correct association rate of the system.

## V. CONSIDERATIONS

As show in Figure 3, according to the inputted character's number has increased, the value of CAR increased, and then it closed to stable value about 44%, when the inputted character's number has exceeded 40,000. It is that, at the beginning, the rules number in rule dictionary increases, and also the rules number used in correct associate increases, when the number of input character increased. Therefore the correct association rate increases according to the increase of the inputted characters. However, when the number of the inputted characters approaches some values, a great number of useless rules are acquired, and the rule which is in low certainty degree is inferred as rule candidates by the system, because in the experiment, we limits the number of the candidate rules as 20.

## VI. CONCLUSION

In this paper, we introduce a new method called IL-IA, and we evaluated the system using the correct association rate and decreased rate of keystroke. Even if the rule dictionary is empty at the beginning of experiment, it also grows fast by its learning capability. The correct association rate increases 4% when the system added an application of rule generation.

In addition, the more effective rule generation can increase the correct association rate of the system. Therefore we consider that the future work for IL-IA is the increase of the number of effective rules and the decrease of the number of useless rules, by performs the more effective rule generation application.

## ACKNOWLEDGMENT

This work is partially supported by the Grants from the Government subsidy for aiding scientific researches (No.14658097) of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

## 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] K. Araki, Y. Takahashi, Y. Momouchi and Koji. Tochinai, Non-Segmented Kana-Kanji Translation Using Inductive Learning, The Transactions of The Institute of Electronics, Information and Communication Engineering, Vol.J79-D-□(3),1996,pp391-402.

[3] 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.

**[4]** K. Araki and K. Tochinai, Acquisition Words by Inductive Learning and Recognition Words Using Certainty, The Transaction and Communication Engineering, D-□, J75-D-□(7),1992, 1213-1221.

[5] 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.

[6] Masui , T. An Efficient Text Input Method for Pen-based Computers. In proceeding of the ACM conference on Human Factors in Computing Systems (CHI'98)(April1998), Addison-Wesley, pp. 328-335.

[7] 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.

[8] M.Nagao, S.Mori, A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, Proc. COLING94, Kyoto (August 1994), pp611-615.

[9] Z.Wang, K. Araki, and K. Tochinai, Word Segmentation Method Using Inductive Learning for Chinese Text, Proc. IASTED International Conference, Artificial Intelligence and Soft Computing, (July 2000), pp452-458.

[3] We obtain the data of experiments from the Internet Web Page as follow: www.zjzw.net, www.hncnlp.com, and www.ahetc.gov.cn/