International Symposium on Chinese Spoken Language Processing (ISCSLP 2000)
Fragrant Hill Hotel, Beijing
October 13-15, 2000

CHINESE PINYIN INPUT METHOD FOR MOBILE PHONE

Feng ZHANG[i] Zheng CHEN, Mingjing LI, Guozhong DAI

Microsoft Research, China

Intelligent Engineering Lab, Institute of Software, Chinese Academy of Science

ddw@imd.cims.edu.cn zf7615@sina.com

ABSTRACT

Chinese input method is one of the most difficult problems in Chinese Language Processing. And to input Chinese word in mobile phone effectively is an even bigger challenge. In this paper, we propose a new Chinese pinyin input method in mobile phone. This method uses a compact statistical bigram based language model. Also, to meet the special requirements of Chinese pinyin input in mobile phone, we introduce some new features for the search engine and user interface of our system.

1. INTRODUCTION

Mobile phone and wireless communication is undergoing a rapid development in China. And with the introduction of some new technology such as WAP protocol, Internet will be much more tightly integrated with wireless communication. Information retrieval through mobile phone has become one of the hotspots of information technology. Thus an efficient Chinese input method in mobile phone is in great demand.

Chinese input method is one of the key challenges in Chinese Language Processing. There are 406 syllables mapped to more than 7000 common Chinese characters. Furthermore, there are only 8 Keys (2.3…9) can be used to represent the 26 English letters in mobile phone. So the confusion of input is very significant.

There are two main categories of Chinese input method. One is shape-based input method, such as “wu bi zi xing”, the other is pinyin, or pronunciation-based input method, such as “MSPY”, etc. Because of its facility to learn and to use, pinyin is the most popular Chinese input method. Over 97% of the users in China use pinyin for input. This is also true with mobile phone Chinese input methods. But to mobile phone users, the strictly limited key numbers that can be used for input caused many new problems and make it very difficult to produce an effective Chinese pinyin input method in mobile phone.

We studied many Chinese pinyin input methods of mobile phones now used. All have lots of inconvenience, such as, one must select one of the possible pinyin first for choosing the hanzi one wants to type, and one must press a mode switch key to switch mode between typing and selecting. Due to these inconveniences, these products always annoy user very much and have a very slow input speed.

To eliminate these inconveniences and improve input speed, in our method, we rebuilt and compressed our language model originally used in PC pinyin input method. Then we introduced some new features in our viterbi beam search engine [3, 5]. Finally, a new modeless indexing method is introduced to eliminate the requirement of mode switch in our system’s user interface.

The organization of this paper is as follows. In the second section, we briefly describe the architecture of our method. Then in the following three sections, we discuss the new features we introduced to our language model, search engine and user interface. Finally, we give some conclusions.

2. ARCHITECTURE OF SYSTEM

Commonly, a statistical based Chinese pinyin input method can be mainly divided into three parts, N-gram language model, viterbi beam search engine, and user interface. This structure is also applicable to Chinese pinyin input method in mobile phone. The relation and data flow between these three parts are briefly described in the following Figure 1.

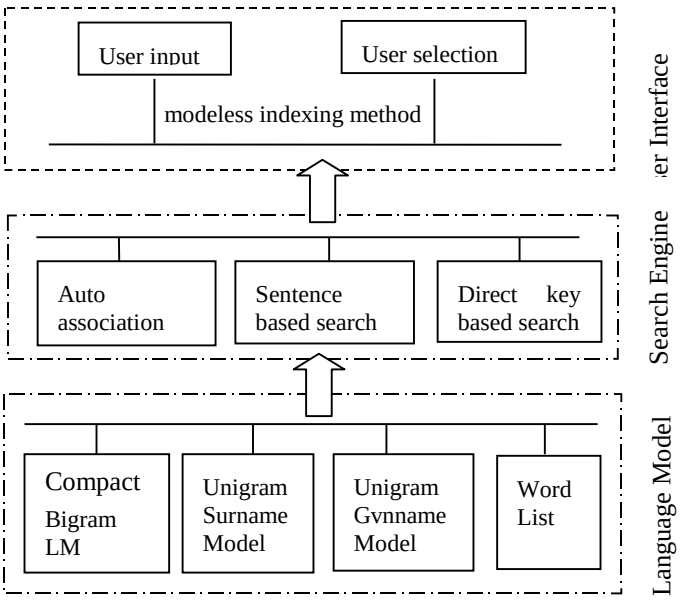

Figure 1. Architecture of mobile phone pinyin input method

In this figure, we describe the main structure of our system. We can also see some new features we introduced especially for mobile phone pinyin input, such as compact bigram language model, direct key based search and modeless indexing method.. We will discuss these features in details in the following sections.

3. COMPACT BIGRAM LANGUAGE MODEL

3.1 Some theory basics

In the conversion of pinyin to Chinese character, for the given pinyin P, the goal is to find the most probable Chinese character H, so as to maximize $\Pr(H \mid P)$. Using Bayes law, we have:

$$\hat{H} = \arg\max_{H} \Pr(H \mid P) = \arg\max_{H} \frac{\Pr(P \mid H)\Pr(H)}{\Pr(P)} \tag{1.1}$$

The problem is divided into two parts, typing model $\Pr(P \mid H)$ and language model $\Pr(H)$.

Conceptually, all H's are enumerated, and the one that gives the largest $\Pr(H, P)$ is selected as the best Chinese character sequence. In practice, some efficient methods, such as viterbi beam search [4], will be used.

In the Chinese language model in equation 1.1, $\Pr(H)$ measures the a priori probability of a Chinese word sequence. Usually, it is determined by a statistical language model (SLM), such as bigram LM.

3.2 Rebuild our model based on Chinese characters

Conventionally, a Chinese bigram language model [5] is based on Chinese words, more than 60000 words are stored. To reduce the model's size, our language model is based on Chinese characters; it has 7432 normally used characters. Although a model based on Chinese characters causes a lower decoding accuracy than a model based on Chinese words, it is enough for a Chinese pinyin input method in mobile phone. We first build a character based word list, and then based on this word list, we rebuilt a bigram language model using the CMU SLM Toolkit from a 20M corpus. Finally, an experiment was conducted to testify the accuracy of a new search engine using our new word list and bigram language model. The result showed when using the 8 keys of mobile phone's keypad as input keys, the decoding accuracy is about 76%, while using the 26 English letter keys i.e. A to Z, the decoding accuracy is about 90%. Compared to the original accuracy of 93% while using word based bigram language model, this result is quite satisfactory. The accuracy of 76% means that if you type the whole pinyin of the sentence you want to input using the keypad of a mobile phone, there will be about one word error in four words. Though this is far from enough, it is somewhat reasonable. The obvious reduction of accuracy is caused by the confusion of mapping keys in mobile phone's keypad to the 26 English letter keys. And this confusion can hardly be avoided in almost all pocket size devices because of their strictly limited key number.

3.3 Compress the model using scalar quantization method

To make the language model more compact, we use scalar quantization method [3] to compress our bigram language model. Before compression, a N-gram language model is stored either in a flat text format, e.g. the ARPA format generated by CMU SLM toolkit, or in the tree-bucket format which stores links between the higher-level language model items and their lower level relatives. The typical size of the two file formats is somewhat too large for us. Because of the limited memory size of mobile phone, we must reduce the size of our language model. We analyzed the distribution of the log probabilities and back-off coefficients in our language model. Then we studied the information loss (measured by perplexity changes) of using different codebook sizes and found that a codebook size of 64 or greater achieves acceptable results. Thus, every floating-point number in the language model which takes 4 bytes of storage can be compressed to 1 byte with little information loss. After compression, our final model size is about 50K. Compared to the originally size of 160K, we get a 68% reduction of size. Our experiment showed that the average keystroke times per word had little change after compression.

3.4 Additional model for Chinese person name input

Inputting Chinese person name and searching a phone number by the name are two of the most frequently used functions to mobile phone users. The composition of Chinese person name is quite different from common Chinese sentence. But all mobile phones now used did not make use of this speciality. This causes great inconvenience while inputting Chinese person name. For example, if you want to input a name beginning with "邸", you will find it really difficult to find this word.

To the speciality of Chinese person name, we developed a set of unigram language models for Chinese people name input. We first build a corpus of Chinese person name collected from "People's Daily", of which the size is about 100K. Then we trained a family name unigram language model and a given name unigram language model. These two unigram models take a little memory storage, and they really increase the performance of inputting Chinese person name in our system extremely.

We also introduced a new feature in our system to improve the efficiency of searching the phone number by person name. Some systems now used can only search a phone number after user inputted the full family name in Chinese. Considering the inefficiency of the Chinese input method they used, this really caused great inconveniences to users. In our system, we built a small database which included the pinyin of all person names in the address book. So when one wants to search a friend's phone number, the only thing he needs to do is to type the keys corresponding to the first letters of his family name and given

name. This small feature surely increases the operability of our system.

4. SEARCH ENGINE

4.1 Select Hanzi directly based on key input

In mobile phone, each key can represent 3-4 English letters. If you want to input one English letter, you should press 1-3 times of corresponding key. For example, “66” means “n”. If you want to enter the pinyin of one Chinese character, you should enter each letter with this method. For example, you need to type “66444” to enter pinyin “ni”. In some products, you only need to type the key that corresponds to the letter one time. For example, you just need to type “64” in the previous example. These systems will display the confusing pinyin to you for choosing, e.g. “mi” and “ni”. This is a great advancement, but still not enough. To this instance, if you want to type “你”, you then have to first press a key to choose “ni” then can you see “你” in candidate line. This is extremely inconvenient for user. Our solution to this problem is search the most possible hanzi in the context directly based on mobile phone key input, in this case, “64”, not English character pinyin, “ni” or “mi”. Now, after typing “64”, user can directly choose “你” or “米” without having to choose “ni” or “mi” first. To achieve this objective, we rebuilt our lexicon, making Chinese words directly related to mobile phone key input, i.e. 2 to 9. Thus, the keystroke times were greatly decreased.

4.2 Auto-association

No user can be as familiar with the mobile phone keyboard as the PC keyboard. While user want to type a Chinese sentence “我们”, of which the pinyin is “women”, he must first find out the corresponding key to every pinyin character he want to type. This is almost the most annoying thing to user while inputting. To solve this problem, we introduced auto-association to our system. That is, while user is inputting a sentence, our system automatically displays possible hanzi before user type full pinyin. The order that these possible hanzi is sorted is based on the context and the characters user typed. For example, user types “640” for “你”, then, in some other systems, he need to type “636” for “们”. In our system, after user type “0”, the candidate line of our system will show the following suggestions “们 的 不 这 …” for user to select. Then, user can directly choose the word he wanted to input. Otherwise, if user want to type “也” instead of “们”, after he types “9”, our system will suggest “也 一 这” as candidates. Compared to other systems, this method provides user with great convenience.

4.3 Automatically select Chinese words based on context

No method now used can make decision for user automatically because of their word-based search engine. In our system, sentence-based input method is used instead of word-based input. System can output the most possible Chinese characters based on the context. User can type the full pinyin of the sentence he want to input; our system will put out the most possible sentence as the result. For example, if you want to input “我们是学生”, you only need to type the corresponding keys, in this case, “96063607440983074364”. Furthermore, although the engine is sentence-based input, user can choose interact with the system or not. And user’s interaction will improve the accuracy of the system’s decision.

5. USER INTERFACE

5.1 Modeless indexing method

In conventional mobile phone input methods, keyboard has two functions: typing and selection. So mode switch is a big burden to the user. To eliminate this inconvenience, we adopted a modeless indexing method to the user interface of our system.

In the past, we use digit 1 – 9 to index the candidates of Chinese character. But the system cannot automatically distinguish the selection key from typing digits. For example, user inputs “646”, after user types “64” system will suggest “1. 们 2. 门 3. 闷 4. 妈 5.民 6.名 7.明 8.么 9.梦”. Then if user go on to type “6”, system will be confused for “6” can be either the selection of “名”, or the input of “646”(“min” or “nin”). So mode switch is needed here.

Our system provides a modeless indexing method that can eliminate the need of mode switch. System automatically adopt the invalid input keys as selection indexes of candidates, that is, once these keys are used as input keys, the typing pinyin will be invalid. In the previous example, system will use key “1, 4, 5, 7, 9” as selection keys because these keys cannot be the input pinyin keys following “64”. So the candidate line will be modified as “1. 们 4. 门 5. 闷 7. 妈 9. 民”. Then if user finds the word he want to input, he can directly press one key for selection instead of press a mode switch key first. This makes the keystroke time greatly reduced and the speed of input significantly improved.

5.2 Prediction of the next key user may type

As we had discussed in previous chapter, one of the most annoying things user faces while inputting Chinese sentence in mobile phone is to find out the corresponding key to every pinyin character he want to type. To relieve this burden of user, we make some prediction of the next key user may type in a sentence. Based on our language model, we calculated the probability of each key that may follow the word user has just typed. Then we show user the key that has the highest probability. So, if the prediction is correct, user can see the key he want to type in the screen and press a certain key to input it instead of find this key in keypad first then press it. This method surely brings some convenience, but to some extent, it also brings user a big burden of watching the hint key in screen. Besides, the accuracy of prediction now is about 60%, which is far from satisfactory. So, how to improve the accuracy of prediction is really a big challenge to us.

6. FUTURE WORK

The small size of the language model now used in our system limits the further improvement of input efficiency. A client-server mode may be a solution to this deficiency. In client side (mobile phone), a small language model is used to help user to input the Chinese characters without time delay. In the server side, some powerful language model can be used to get more exact result. Client side and server side will be synchronized if the bandwidth is enough. And system will automatically learn and adapt to the user.

7. CONCLUSION

An effective Chinese input method in mobile phone is very important for mobile phone users. And such a Chinese input method is also in great demand by many other pocket size devices, whose keypad is almost the same kind of the mobile phone's. In the past, there were not enough research in this area. Almost all currently used Chinese input methods in mobile phone were simply copied from Chinese input method in PC without necessary adaptation. The speciality of the mobile phone 's keypad was not adequately considered in these methods. Therefore the efficiency of Chinese input now used in mobile phone was really poor compared to the efficiency of Chinese input method in PC.

In this paper, we proposed a new method of Chinese pinyin input in mobile phone. Although the main frame of our method is as same as the frame of the popularly implemented Chinese pinyin input method in PC, many inventive features are introduced into our system to meet the special requirements of mobile phone. The efficiency of Chinese input in mobile is greatly improved while using our method. We also conducted some experiments to test the efficiency of our method. Based on our test, the average keystroke times per word of other methods now used is more than 4.2 times, the number of word inputted per minute is less than 9 (we got this approximately statistic by randomly choose ten persons, and let them input fifty sentences in five kinds of mobile phone now used popularly). While using our method, keystroke times per word is 3.3 (this precise statistic was gained through an experiment conducted on 5M test texts), reduced by more than 20%, and word input number per minute is 13, improved by more than 40%. We reached this by improvement in the language model, the search engine, and the user interface used in our method.

8. REFERENCES

[1] Chen Yuan, Chinese Language Processing, Shanghai education publishing company, 1997.12, pp.4.

[2] Zheng Chen & Kai-Fu Lee, A New Statistical Approach to Chinese pinyin Input, submitted to ACL 2000.

[3] Shuo DI, Lei ZHANG, Zheng CHEN, Eric CHANG, Kai-Fu LEE, N-Gram Language Model Compression Using Scalar Quantization and Incremental Coding, ISCSLP 2000

[4] Kai-Fu Lee, Automatic Speech Recognition, Kluwer Academic Publishers, 1989.

[5] Frederick Jelinek, Statistical Methods for Speech Recognition, MIT Press, 1997.Jensen K., Heidorn G., and Richardson S. *Natural Language Processing: The PLNLP Approach*, Kluwer Academic Publishers, 1993.

[i] This paper was performed as the author was a visit student in Microsoft Research China, while leaving Intelligent Engineering Lab of Institute of Software, Chinese Academy of Science.