International Journal of Soft Computing 11 (7): 448-452, 2016
ISSN: 1816-9503

Efficient Phoneme-Based Chinese Text Input Method

Cheng-Huang Tung and Wei-Hong Jhang
Department of Computer Science and Information Engineering, National Pingtung University, 900 Taiwan, No. 51, Mingsheng East Road, Pingtung, Taiwan

Abstract: We previously proposed an Efficient Phoneme-based Chinese Input Method (EPCIM) which yields a conflict-code rate of 13.5%. To reduce further the conflict-code rate in inputting Chinese text, an EPCIM-based method that uses a Chinese dictionary from "Hsin-Ku-Yin" input method is developed in this research. Because of the EPCIM's low conflict-code rate, the large "Hsin-Ku-Yin" Chinese dictionary can be transformed into a very compact EPCIM dictionary which can extend the EPCIM to the Efficient Phoneme-based Chinese Text Input Method (EPCTIM). The EPCTIM uses dynamic programming to resolve ambiguity in the EPCIM code string of a piece of Chinese text and to determine the optimal Chinese word string. Experiments demonstrate that the EPCTIM is very promising because of its low conflict-code rate of 2.67% and its low error rate of 0.32%.

Key words: Efficient phoneme-based Chinese input method, efficient phoneme-based Chinese text input method, conflict-code rate, ambiguous degree, Taiwan

INTRODUCTION

The development of Chinese input methods has become more important because of the increasing popularity of Chinese language worldwide. Chinese characters are non-alphabetic and two-dimensional in structure and thousands of commonly used Chinese characters differ ideographically (Kwong *et al*., 1991). However, the standard Chinese phonetic spelling system can use only 41 phonetic symbols to capture the pronunciation of all Chinese characters. A Chinese character can be pronounced as a sequence of Chinese phonetic symbols and an effective phonetic sequence, defined to be the pronunciation of at least one Chinese character, comprises one to four phonetic symbols (Ministry of Education, 2012). In Taiwan, >10,000 characters are commonly used 5,401 of them are designated as "frequently-used" characters but the number of effective phonetic sequences is only 1,351 (Wellington and Yu, 1988).

Methods of coding chinese characters can generally be divided into the phonetic coding strategies and the radical coding strategies. The former (Wan *et al*., 1982; Ooka and Chien, 1983; Cao and Suen, 1987; Gu, 1994) utilize the phonetic spelling system to encode Chinese characters. They have the well-known advantage of simplicity and directness but suffer from the homophone problem because an effective phonetic sequence may in the worst case be associated with tens of homonyms (Suen and Huang, 1984). The Standard Phonetic Chinese Input Method (SPCIM) utilizes this spelling system to encode Chinese characters. Because of the high conflict-code rate, the SPCIM generally is used with a Chinese word dictionary to form the standard phonetic Chinese text input method (Hsin-Ku-Yin, 2012) which further reduces the conflict-code rate by considering the context.

The radical coding strategy (Kwong *et al*., 1994; Chen and Gong, 1984; Kiang and Cheng, 1982; Shieh, 1987; Chen, 1993; Tsang, 2012) uses fundamental components of Chinese characters as basic radicals in encoding them. To achieve a low conflict-code rate, this strategy typically depends on a one-to-one mapping between most input codes and Chinese characters. Nowadays, the most popular radical coding methods in Taiwan include "Tsang" 2012. To encode all Chinese characters, a radical coding method heuristically defines a few radicals but actually not enough to reconstruct thousands of Chinese characters. It also defines some extra heuristic rules to divide Chinese characters based on the defined radicals. Users must be familiar with not only the defined radicals but also the various rules for separating Chinese characters. This requirement is the major shortcoming and makes radical coding methods hard to learn.

We have already proposed an Efficient Phoneme-based Chinese Input Method (EPCIM) to reduce greatly the conflict-code rate (Jean and Tung, 2000; Tung and Jean, 2009). To eliminate the difficulty of decomposing Chinese characters, the EPCIM defines a large extended radical set including 5,401 frequently-used

Corresponding Author: Cheng-Huang Tung, Department of Computer Science and Information Engineering, National Pingtung, University, 900 Taiwan, No. 51, Mingsheng East Road, Pingtung, Taiwan

Chinese characters, standard radicals and three primitive strokes. Based on the written strokes of a Chinese character, the EPCIM extracts two extended radicals which must include the first and the final strokes, respectively. Since the EPCIM has assigned a phonetic symbol as the feature of an extended radical, it thereby obtains two feature phonetic symbols for the Chinese character. Then, the EPCIM extracts at most two more feature phonetic symbols from the phonetic code of the Chinese character. By combining these feature phonetic symbols, the EPCIM generates the EPCIM phonetic code of the character which comprises at most four phonetic symbols. Based on experimental results, the EPCIM is easy to learn and yields a low conflict-code rate of 13.5%.

In this research to reduce the conflict-code rate of inputting Chinese text, the EPCIM is utilized with a Chinese word dictionary as an Efficient Phoneme-based Chinese Text Input Method (EPCTIM). The number of words in a Chinese dictionary is normally very large-typically >100,000. The Chinese dictionary that is used by Hsin-Ku-Yin (2012), an open-source standard phonetic Chinese text input method, contains >140,000 Chinese words and is adopted in our new method. Because of the low conflict-code rate of the EPCIM, it does not require a dictionary with a very large number of Chinese words. First, the length of words is restricted to four characters and then the phonetic code string of each Chinese word is translated into the corresponding EPCIM phonetic code string to generate the raw EPCIM dictionary. The ambiguous degree of each multi-character word entry in the raw EPCIM dictionary is measured and these unambiguous multi-character words are then deleted to yield the compact EPCIM dictionary which extends the EPCIM to the EPCTIM. When users want to input Chinese text they type the EPCIM phonetic string of the input Chinese text. The EPCTIM translates the EPCIM phonetic code of every Chinese character into a Chinese character set and then uses dynamic programming to determine the optimal word string from the combinations of the character sets. Experimental results demonstrate that the EPCTIM is very promising. The conflict-code rate for inputting 100 studies of 42,088 Chinese characters is improved from 12.28% for the EPCIM to 2.67% for the EPCTIM. Following automatic character selection by the EPCTIM, the total error rate is only 0.32%.

MATERIALS AND METHODS

Review of efficient phoneme-based Chinese input method: This study reviews key points of the Efficient Phoneme-based Chinese Input Method (EPCIM) (Jean and Tung, 2000; Tung and Jean, 2009) using examples.

Extraction of extended radicals: The way in which the EPCIM extracts two extended radicals based on the written strokes of a Chinese character and obtains the feature symbols from the extracted extended radicals is elucidated.

The 41 Chinese phonetic symbols: Pronunciation buh; puh; muh; fuh; duh; tuh; nuh; luh; guh; kuh hu; gee; chi; shi; jr; chr; shr; re; dz; tz sz; ah; oh; uh; ai; a; au; oh; an; un; ang; uea; ea; wu; re; a; ong. Based on the concept that Chinese characters are constructed recursively, the EPCIM defines the set of extended radicals to include:

- About 5401 frequently-used Chinese characters
- Standard radicals, listed in a Chinese dictionary
- Phonetic symbols
- Three primitive strokes used as the extracted extended radicals of Chinese characters that are difficult to decompose

The EPCIM then can extracts two extended radicals based on the written strokes of a Chinese character. Three criteria for extracting extended radicals must be followed. First, the first extracted extended radical must begin with the first stroke and the second must end with the last stroke. Second, except for the three primitive strokes, extracted extended radicals can not intersect any other stoke. Third, the remaining strokes that are not included by the two extracted extended radicals must be as few as possible.

EPCIM Phonetic code extraction: The EPS of a Chinese character has one to four phonetic symbols and the EPCIM adopts at most two phonetic symbols as feature symbols which are combined with the feature symbols for the two extracted extended radicals to form the EPCIM phonetic code of the Chinese character.

Presents some example EPSs for Chinese characters. The EPCIM adopts all symbols of an EPS having only one or two phonetic symbols. When the EPS of a character comprises three phonetic symbols, EPCIM extracts the first two symbols if the final symbol is a tone, or the first and the third symbols otherwise. The remaining tone symbol of a three-symbol EPS is reserved as the backup feature for a later specific use.

Generation of compact EPCIM Chinese dictionary: In this study, the "Hsin-Ku-Yin" Chinese dictionary is firstly translated into the raw EPCIM Chinese dictionary. The ambiguous degree of every multi-character word entry in the raw EPCIM Chinese dictionary is then measured.

Table 1: The number of word entries in the "Hsin-Ku-Yin" dictionary, the raw EPCIM dictionary and the compact EPCIM dictionary

Word length (characters)	Word entries in "Hsin-Ku-Yin" dictionay	Word entries in the raw EPCIM dictionary	Word entries in the compact EPCIM dictionary
1	6,485	6,265	6,265
2	77,933	75,947	19,444
3	27,735	26,753	2,960
4	24,254	22,876	2,785
Total	136,411	131,841	31,454

Following removing the unambiguous multi-character word entries those that remain constitute a compact EPCIM Chinese dictionary.

Raw EPCIM Chinese dictionary: To support the efficient input of Chinese text, some input methods involve large. Chinese dictionaries. For example, input methods "HsinKu-Yin", "Microsoft New Phonetic IME and "Natural Chinese Input" are standard phonetic Chinese input methods and use large Chinese dictionaries to input Chinese text. When users type the phonetic sequences that correspond to a piece of Chinese text these input methods look up the dictionary to resolve ambiguity based on context.

Such a Chinese dictionary normally has hundreds of thousands of Chinese words. Even though the usable Chinese characters are limited to 5401 frequently-used characters and word length is limited to four characters, the number of words in the "Hsin-Ku-Yin" Chinese dictionary, given in Table 1 is still as high as 136,411. The dictionary is so large that may consume extensive resources and reduce system performance. Every word entry in the "Hsin-Ku-Yin" dictionary has three attributes-character code, phonetic code and frequency. Table 2 presents four example word entries.

Compact EPCIM Chinese dictionary: The number of words in the raw EPCIM dictionary may be sufficiently large to influence system performance. A compact EPCIM dictionary is then obtained by deleting unambiguous multi-character word entries in the raw EPCIM dictionary.

An extra feature, ambiguous degree is added to every word entry in the raw EPCIM dictionary. Hence, word Entry E in the raw EPCIM dictionary has four features-character code E.CC, EPCIM phonetic code E.EPC, frequency E. α and ambiguous degree E. β. The concept for the ambiguous degree of a word entry is associated with the number of code conflicts caused by its EPCIM phonetic code. Generally, if a word entry is unambiguous, it means that the transformation from its EPCIM code to its character code is unique even though the word entry is not in the dictionary. This reveals that the unambiguous word entries in the raw EPCIM dictionary can be removed to reduce the size of dictionary. Notably, since the mapping between the Chinese characters, i.e., the single-character words and their EPCIM phonetic codes is the foundation of the EPCIM, every single-character word entry is necessary in the dictionary no matter it is ambiguous no not. Accordingly, the following investigation of ambiguous degree is focused on the multi-character word entries in the raw EPCIM dictionary.

The following proposes the algorithm that generates the compact EPCIM dictionary. First, the ambiguous degree of each multi-character word entry in the raw EPCIM dictionary is measured and the unambiguous one is then deleted. Frequency values of the distinct remaining words are summed. The occurrence probability of each word entry is calculated from its frequency. The remaining word entries form the compact EPCIM dictionary. Notably, the feature of occurrence probability is newly added to each word entry in the compact EPCIM dictionary:

Algorithm 1:

Algorithm 1: Create the compact EPCIM dictionary
Input: the raw EPCIM dictionary
Output: the compact EPCIM dictionary
Begin
//delete the unambiguous multi-character words
for each multi-character word Entry E whose E. CC has at least two characters
E. $\beta = \min \Pi | f(A'_i)|$ if $\Pi | f(A'_i)| > 0$
if E. $\beta A_1 A_2 = 1$; = E.CC delete E from the raw EPCIM dictionary
fsum = 0
for each distinct word entry E
fsum+=
//calculate the occurrence probability for each word entry
for each word entry E
E.P = E.α/ fsum
output the remaining word entries as the compact EPCIM dictionary
End

Table 2 also presents the number of words in the compact EPCIM dictionary. It is 31,454, reduced from 131,841 for the raw EPCIM dictionary. Accordingly, 76.1% (1-31,454/131,841) of word entries in the raw EPCIM dictionary can be deleted.

Efficient phoneme-based Chinese text input method: The compact EPCIM dictionary is used to construct the Efficient Phoneme-based Chinese Text Input Method (EPCTIM). Assume that $D_1 D_2, ..., D_n$ is the EPCIM code

Table 2: Performance evaluations of the EPCIM. the EPCTIM and the Standard Phonetic Chinese Text Input Method (SPCTIM)

Paramiter	EPCIM	EPCTIM	SPCTIM
No. of input characters	42,088	42,088	42,088
No. of ambiguous character sets	5,167	1,123	17,085
conflict-code rate	12.28% (= 5,167/42,088)	2.67% (= 1.123/42.088)	40.59% (= 17.085/42.088)
No. of selection errors	-	136	4.041
Total error rate	-	0.32% (= 136/42.088)	9.60% (= 4.041/42.088)

string for inputting an n-character piece of Chinese text where D_i is the EPCIM code of the i-th Chinese character. After D_i is looked up in the dictionary, the Chinese character set S_i for D_i is obtained.

Assume that function P(i, j) can locate the (j-I+1) character word with the maximal occurrence probability from the combinations of character sets $S_i...S_j$ and then returns the occurrence probability. If no (j-i+1)-character words in the dictionary can be found then p (i, j) = 0. For example, in P (1, 2) can finds only the two-character word in the dictionary from the combinations of characters sets and and thus returns the occurrence probability of word P (1, 3) and P (1, 4) find no three-character and four-character words in the dictionary and so both return the probability zero. Let P_{Si} be the probability product of the optimal word string for character sets $S_1 S_2, ..., S_i$. $P_{so} = 1$ and $P_{si} = 0$, i<0 are defined first. Since the longest word in the dictionary has four characters P_{Si}, i≥1 is defined recursively as follows:

$$P_{S_i} = \max_{j=1...4} (P_{S_{i-j}} P(i-(j-1), i)) \quad (1)$$

Note that P_{Si} can be calculated by dynamic programming and the optimal word string for P_{Si} can be obtained by backtracking the optimal path identified by dynamic programming.

The EPCTIM algorithm which determines the optimal word string for the EPCIM code string of a piece of Chinese text is summarized as follows. The algorithm uses dynamic programming to obtain the path of the optimal word string. If the input is the EPCIM phonetic code of an n-character Chinese text, then an n-stage dynamic programming procedure is performed. In each stage, the derived word and the backward path are recorded. Following backtracking, the optimal word string is obtained in reverse order and then is output as, $W_1W_2.,,,W_m$, m≤n in forward order".

Algorithm 2:

Algorithm 2: Transform the EPCIM code string into the optimal word string

Input: EPCIM phonetic code $D_1D_2...D_n$ of an n-character Chinese text

Output: the optimal word string for P_{Sn}

Begin

get the corresponding character sets $S_1S_2...S_n$ of $D_1D_2...D_n$ by dictionary look-up

P_{so} and $P_{si} = 0$, i<0

for *i*=1 to n

$P_{si} = \max(P_{Si-j} \times P(i-(j-1), i))$

$b = \arg i = 1 \max(P_{Si-j} \times P(i-(j-1), i))$

set the identified b-character word with probability P(i-(b-1), i) as W_{Si} and stage i-b as the backward path of this stage

backtrack from stage n back to stage zero following the backward path to get the optimal word string in reverse order

reassign the optimal word string in forward order as, $W_1W_2...W_m$, m≤n

End

RESULTS AND DISCUSSION

In the experiments, performance of the EPCTIM in inputting the text of 100 Chinese articles is analyzed. These articles contain a total of 42,088 Chinese characters. The EPCIM code string of each article is typed in sequence. The EPCTIM processes the EPCIM code string sentence by sentence. The EPCTIM first translates the EPCIM code string of a sentence into the corresponding Chinese character sets and then uses dynamic programming to identify the optimal word string from the combinations of these character sets. Presents the evaluated performance of the EPCIM, the EPCTIM and the Standard Phonetic Chinese Text Input Method (SPCTIM) when used to input 42,088 Chinese characters in 100 studies. First, the EPCIM reads the EPCIM code of each character. If the corresponding character set contains multiple characters, then the character set is ambiguous and the ambiguity in the set must be resolved by the user manually. In this experiment, the 5,167 ambiguous character sets yield a conflict-code rate of 12.28%(=5,167/42,088). When the EPCTIM is utilized to process the EPCIM code string of the same 100 articles, only 1,123 character sets remain ambiguous after dynamic programming and the conflict-code rate is thus greatly reduced from 12.28-2.67% (= 1,123/42,088). Since the EPCTIM uses dynamic programming to identify the optimal word strings, the EPCTIM can automatically select the characters for the ambiguous character sets according to the optimal word strings. In sum, the total error rate of the EPCTIM for inputting the 42,088 Chinese characters is as low as 0.32%. The SPCTIM combines the SPCIM with the "Hsin-Ku-Yin" Chinese dictionary. Notably, the SPCTIM also uses dynamic programming to resolve ambiguities for the ambiguous character sets.

Using it to process the phonetic code strings of the same 100 studies yields 17,085 ambiguous character sets and a conflict-code rate of 40.59% (=17,085/42,088) which greatly exceeds the 2.67% value achieved using the EPCTIM. With automatic character selection, the SPCTIM then yields a total error rate of 9.60% which is still much >0.32%, achieved using the EPCTIM. Notably, the dictionary used with the SPCTIM has 136,411 word entries whereas the dictionary used with the EPCTIM has only 31,454.

CONCLUSION

An Efficient Phoneme-based Chinese Text Input Method (EPCTIM) which uses a very compact Chinese dictionary to resolve the input ambiguity is constructed. The EPCTIM achieves a low conflict-code rate of 2.67% for inputting the EPCIM code string of Chinese text. Furthermore, the EPCTIM reduces the total error rate of inputting Chinese text to only 0.32%.

RECOMMENDATIONS

Most selection errors of dynamic programming involve a lack of available contextual information concerning the single-character words in the dictionary. However, enlarging the dictionary in advance to resolve the ambiguities among all single-character words is impossible. In the future the authors will develop a search-engine-based ambiguity resolution model. Except identifying only the optimal word string, the proposed method first generates the candidate word strings associated with the ambiguous character sets. Search engines such as Google or Yahoo, allow access to an extra very large corpus and so can be used to determine the best from these candidate word strings. The proposed method uses a search engine to extract the related articles for each candidate word string and then measures the fitness degree of the candidate word string in the articles to locate the optimal word string.

ACKNOWLEDGEMENT

Ted Knoy is appreciated for his editorial assistance.

REFERENCES

Cao, X. and C.Y. Suen, 1987. A new phonetic and ideographic coding technique for Chinese information processing. Comput. Process. Chinese Orient. Lang., 3: 91-106.

Chen, C.K. and R.W. Gong, 1984. Evaluation of Chinese input methods. Comput. Process. Chinese Orient. Lang., 1: 236-247.

Chen, K.J., 1993. A mathematical model for Chinese input. Comput. Proc. Chinese Orient. Lang., 7: 75-84.

Gu, H.Y., 1994. A Chinese-character inputting system using a new type of phonetic keyboard and a compound Markov language model. Proc. ROCLINGC., 1994: 253-262.

Hsin-Ku-Yin, 2012. Open source intelligent Chinese phonetic input method. Chewing Core Team, China. http://chewing.csie.net/index.html

Jean, E.Y. and C.H. Tung, 2000. A phoneme-based Chinese input method with low conflict code rate. Intl. J. Comput. Process. Orient. Lang., 13: 333-349.

Kiang, T.Y. and T.H. Cheng, 1982. A new Chinese indexing system based on separation of character into main and subordinate components. Proc. Intl. Comput. Symp., 1: 131-141.

Kwong, S., H. Wong and Y.S. Yu, 1991. An effective method for storing and retrieval of Chinese characters. Proc. Intl. Conf. Inf. Eng., 1991: 368-376.

Kwong, S., Y.K. Chan and E. Lee, 1994. A postprocessing to reduce conflict code rate for Chinese input methods. Proc. Process. Chinese Orient. Lang., 1994: 82-85.

Ministry Of Education, 2012. Learning program for stroke order of frequently used Chinese characters. Ministry Of Education, Taiwan, http://stroke-order.learningweb.moe.edu.tw/home.do

Ooka, T. and M.H. Chien, 1983. The practical application of the input method converting PINYIN to characters. Proc. Intl. Conf. Text Process. Large Character Set, 1983: 131-136.

Shieh, C.C., 1987. Evaluation report of Chinese input methods and input devices. Institute for Information Industry, Taipei, Taiwan.

Suen, Y. and E.M. Huang, 1984. Computational analysis of the structural compositions of frequently used Chinese characters. Comput. Process. Chinese Orient. Lang., 1: 1-10.

Tsang, J., 2012. Input method. School of Education, The Chinese University of Hong Kong, Hong Kong, China. http://www.fed.cuhk.edu.hk/readwrite/typing/

Tung, C.H. and E.Y. Jean, 2009. A modified phoneme-based Chinese input method for minimizing conflict code rate. Comput. Stand. Interfaces, 31: 292-299.

Wan, S.K., H. Saitou and K.I. Mori, 1982. Experiment on PINYI-HANZI conversion Chinese word processor. Comput. Proc. Chinese Orient. Lang., 1: 213-224.

Wellington, L. and C.P. Yu, 1988. A historical advancement of Chinese language computing. Comp. Proc. Chinese Orient. Lang., 4: 57-81.