DISAMBIGUATION OF PHONETIC CHINESE INPUT BY RELAXATION–BASED WORD IDENTIFICATION

Charng-Kang Fan and Wen-Hsiang Tsai
National Chiao Tung University
Hsinchu, Taiwan 30050
Republic of China

I. INTRODUCTION

Among the various Chinese input methods for computers, the national phonetic input method is the most favored one by casual users in Taiwan. Possible reasons include the following: (1) users need not decompose each character into parts which in most cases are puzzling non–conventional Chinese radicals; (2) everyone learns the national phonetic symbols in primary school; and (3) the number of the national phonetic symbols is only 37. Also, there exists a well–known ordering among the symbols. Hence it is easy for most people to memorize them well, which facilitates the finding of proper symbol keys, in contrast with the difficulty of searching the unnatural radical keys required by radical–based input methods.

Since Chinese characters are monosyllabic, a character can be represented by a syllable which usually consists of one or two vowels and an optional consonant plus a tone marker. The phonetic input method is to key in the syllables and convert them into corresponding characters.

Though the national phonetic input method is convenient for casual users, an inherited drawback does exist. Since only about 1300 distinct syllables are used for more than ten times of Chinese characters, the number of homonyms per syllable is quite large. Hence ambiguities exist in determining the correct character for a given syllable.

II. KNOWN APPROACHES FOR PHONETIC INPUT

According to whether the processing is performed with or without human intervention; whether the processing unit is a character, a word, or a sentence; and whether the contextual relationship of adjacent syllables are used or not, existing methods are briefly surveyed as follows.

(1) Most available Chinese systems [1] require a user to input phonetic symbols, syllable by syllable, and leave the homonym resolution to the user. The user is required to select manually the desired character for an input syllable among a list of homonyms displayed on the computer screen.

(2) Wan, Saiton, and Mori [2] described a method by which users manually inserted a word break in a continuously entered Pinyin string and issued the conversion command to have the computer convert the sequence of syllables between two breaks into characters.

(3) Ho et al. [3] described a method for segmenting a sentence into segments. It uses three special classes (the 'pure word head,' the 'pure word tail,' and the 'pure monosyllabic word') of words as segment markers. The syllables between the markers can then be converted into characters automatically by matching against the stored word dictionary. But emphasis was not put on the use of the method for ambiguity resolution.

(4) Chen et al. [4] proposed an automatic continuous conversion algorithm to convert a string of syllables into characters word by word. The system has a dictionary for word look up, and a file to handle exceptions. Several ad hoc rules (such as "previous word first," "preference for highest usage frequency," etc.) are also employed to resolve the ambiguous cases. It did not use more general contextual information existing in a sentence.

(5) Lin and Tsai [5] proposed an automatic ambiguity resolution method by a relaxation process. It regards the phonetic input method as a task of assigning each individual syllable

to a proper character. Relaxation iterations are applied to reduce the ambiguities in the assignments, using the co–occurrence statistics of syllable pairs to compute the initial probability values and the compatibility coefficients needed in the relaxation process. The contextual information of a sentence is utilized, but only pairwise neighboring character relationships are explored; word relationships are not utilized.

III. PROPOSED APPROACH TO DISAMBIGUATION OF PHONETIC INPUT BY WORD IDENTIFICATION

A. Ideas

A new approach to automatic disambiguation of phonetic input by a relaxation–based word identification process is proposed. This approach is applied to a string of phonetic symbols or syllables which constitute a sentence. The approach is based on the following consideration: (1) the smallest meaningful unit of Chinese is word; (2) there are very few or even no homonym for each multi–syllabic word in contrast to a large number of homonyms for each monosyllabic word; and (3) there exists useful contextual relationship among the words in a sentence, which will be described later. Compared with Lin and Tsai's method [5] which assigns syllables to characters and utilizes the co–occurrence relationships of pairs of characters, the proposed approach assigns syllables to words and utilizes the adjacency relationships of neighboring words. Thus it is a word–oriented approach which employs the more meaningful contextual constraint information among the words in an input sentence.

By regarding the phonetic input text as a string of syllables, the proposed approach basically is a process of word identification which assigns syllables to words. For example, in the sentence "yóu gù wùn tí rcu jièn yì bǐ jiàu hǎu."[1]

(It is better that the suggestions be proposed by the consultant.),

although the numbers of homonyms are 11, 8, 7, 7, 3, 26, 50, 7, 8, and 2 respectively for the syllables according to the Eten Chinese system, the sentence is composed of (or should be segmented into) the words as "yóu" (by), "gù wùn" (the consultant), "tí rcu" (propose), "jièn yì" (the suggestions), "bǐ jiàu" (comparatively), and "hǎu" (good, better), regardless of the number of the enormous combinations of homonyms (approximately 10^9 for this sentence). The essence of the proposed approach is to apply the word identification technique to the syllables of a sentence so that the goal of converting the syllables into characters can be accomplished simultaneously when the segmentation of the sentence into words is completed.

Relaxation is a problem solving paradigm which is used in a lot of class assignment or labelling problems to handle the situations of ambiguities. It iteratively employs the contextual information to modify the previous judgement, which lessens the ambiguities and finally converges to the most likely choice. The application problems to which relaxation has been applied includes noise removal, edge detection, scene labelling, image analysis, handwritten character recognition, etc. [6–15].

B. The Syllable–to–Word Assignment Problem

Let S represent an input sentence and W_j be an arbitrary Chinese word which are composed of n and m syllables, respectively, as follows:

[1] The Suen's phonetic symbols [17] are used to facilitate the pronunciation and reading of the Chinese characters.

$$S = s_1 s_2 \ldots s_n,$$
$$W_j = s_{w_{j1}} s_{w_{j2}} \ldots s_{w_{jm}},$$

where s_i or $s_{w_{jk}}$ denotes a syllable. Let

$$Q = \{s_1, s_2, \ldots, s_n\},$$
$$W = \{\text{all Chinese words}\},$$

then a syllable–to–word assignment A is defined as a mapping from Q to W:

$$A : Q \rightarrow W$$

such that the expression $W_j = A(s_i)$ means to assign s_i, $1 \leq i \leq n$, to the word $W_j = s_{w_{j1}} s_{w_{j2}} \ldots s_{w_{jm}}$, indicating s_i as one of the composing syllables of W_j (i.e., there exists an integer k, $1 \leq k \leq m$, such that $s_{w_{jk}} = s_i$). We will also use $s_i \rightarrow W_j$ or A(i,j) to denote the meaning of $W_j = A(s_i)$. Word identification is thus a consistent syllable–to–word assignment problem such that each syllable in a sentence is correctly assigned to the composing words of the sentence, and the assignment of each syllable is compatible with one another.

C. Initial Probability Values for Relaxation

There usually exist, in a sentence S, multiple words to which a syllable s_i may be assigned. Let AA_i denote the set of all assignments which assigns s_i to such words, i.e.,

$$AA_i = \{ A(i,j) \mid W_j \text{ is in S and } W_j = A(s_i) \}.$$

Let P_{ij} denote the probability estimate that s_i is assigned to W_j, and P_{ij}^0 the initial probability value of P_{ij}. It is proposed to define P_{ij}^0 as

$$P_{ij}^0 = \text{count}(W_j) / \sum_{W_k \epsilon AW_i} \text{count}(W_k)$$

where $\text{count}(W_k)$ is the usage frequency count of word W_k which can be collected in advance, and AW_i is the set of all possible words in a sentence S containing syllable s_i, i.e.,

$$AW_i = \{ W_j \mid A(i,j) \epsilon AA_i \}.$$

D. The Relationships Among Neighboring Words

Following Fan and Tsai [15], the relationship between two neighboring words W_a and W_b in a sentence can be categorized into five classes.

(1) Interleaving, that is,

a. $W_a \neq W_b$ and

b. there exists a W_c such that W_c=suffix(W_a) and W_c=prefix(W_b), or W_c=suffix(W_b) and W_c=prefix(W_a), where the terms prefix and suffix represent the leading and the trailing strings of syllables in a word respectively.

(2) Containment, that is,

a. $W_a \neq W_b$ and

b. W_a=suffix(W_b) or W_a=prefix(W_b) or W_b=suffix(W_a) or W_b=prefix(W_a).

(3) Identity, i.e., $W_a = W_b$.

(4) Adjacency, i.e., suffix(W_a) and prefix(W_b) are adjacent in the sentence (assuming W_a is in front of W_b). This can further be classified into two cases as follows.

a. Loose adjacency such that suffix(W_a)·prefix(W_b) is not a word.

b. Intimate adjacency such that there exists another word W_c which is formed with suffix(W_a) and prefix(W_b).

(5) Irrelevancy, that is, $W_a \neq W_b$ and they are positionally apart from each other in the sentence.

Classes (1) through (4) will also be called relevancy relationships, in contrast to the irrelevancy relationship.

The relationship between two assignments can be defined in terms of the relationship between the involved words. Assignments A(i,j) and A(h,k) are said to be mutually

(1) neutral, if W_j and W_k are irrelevant or if W_j and W_k are loosely adjacent;

(2) supportive, if W_j and W_k are identical;

(3) opposing, if W_j and W_k are of the relationship of containment or interleaving;

(4) quasi–opposing, otherwise (i.e., if W_j and W_k are intimately adjacent).

E. Compatibility Coefficients of Syllable Assignments for Relaxation

Different relationships between assignments result in different effects of the relaxation iterations. Given two assignments A(i,j) and A(h,k), if they are mutually neutral, then neither of them will affect the other. If they are mutually supportive, then both of syllables s_i and s_h are assigned to the same word W_j, thus both supporting the identification of word W_j. If the assignments are mutually opposing, then either the two words W_j and W_k are interleaving, or one of them is contained in the other, either case reflecting a conflict situation. Finally, if the assignments are mutually quasi–opposing, they may or may not oppose each other, depending on the effect of the intermediate words constructible from the syllables in W_j and W_k.

Thus the compatibility coefficients C(ij;hk), indicating the degree of supporting or opposing by A(h,k) to A(i,j), can be derived based on the relationship of A(h,k) with respect to A(i,j). The supportive ones are given positive C(ij;hk) values; the opposing ones, negative values; and the neutral ones, the values of zero. More specifically, the compatibility coefficients are defined in this approach as follows:

C(ij;hk) =

(1) 1, if A(i,j) and A(h,k) are mutually supportive;

(2) 0, if A(i,j) and A(h,k) are mutually neutral;

(3) –0.5, if A(i,j) and A(h,k) are mutually quasi–opposing;

(4) –1, if A(i,j) and A(h,k) are mutually opposing.

F. The Relaxation Iteration

P_{ij}^{0} is the initial estimate of the probability value for Assignment A(i,j). With the supporting or opposing information from its neighboring assignments, the probability estimate P_{ij} is modified in each iteration in the relaxation process to reflect the contextual constraints among the words in the sentence. Let the probability estimate for A(i,j) after the rth iteration be denoted as P_{ij}^{r}, then it is proposed to compute the updated probability estimate P_{ij}^{r+1} after the (r+1)th iteration as follows:

$$P_{ij}^{r+1} = \frac{P_{ij}^{r}\,(1+q_{ij}^{r})}{\sum_{AA_i} P_{ij}^{r}\,(1+q_{ij}^{r})} \qquad (1)$$

with

$$q_{ij}^{r} = \frac{\sum_{s_h \in ENS(i,j)} \left(\sum_{A(h,k) \in ENA(i,j)} C(ij;hk) * P_{hk}^{r} \right)}{ENN(i,j)}\,, \qquad (2)$$

where ENS(i,j) is the set of effective syllables which are relevant to A(i,j), ENA(i,j) is the set of effective neighboring assignments for A(i,j), and ENN(i,j) the cardinality of ENS(i,j).

More specifically,

$$ENS(i,j) = \{ s_h \mid h \neq i, \text{ and there exists an } A(h,k) \epsilon AA_h \text{ such that } A(h,k) \text{ and } A(i,j) \text{ are not mutually neutral } \};$$

$$ENA(i,j) = \{ A(h,k) \mid s_h \epsilon ENS(i,j) \};$$

$$ENN(i,j) = \| ENS(i,j) \|.$$

The inner summation (embraced by the parenthesis) in the numerator of Eq. (2) is the effect of a certain effective neighboring syllable s_h on the probability distribution P_{ij} during the (r+1)th iteration. The outer summation (the whole numerator part) in Eq. (2) is the total effect of the effective neighboring syllables of A(i,j) on P_{ij}. This total effect is normalized by the number of effective neighboring syllables ENN(i,j) of A(i,j) to keep the value of q_{ij}^r within the range [–1,1]. Note that $P_{ij}^{r+1} \geq 0$ and $\sum_{AA_i} P_{ij}^{r+1} = 1$.

The termination condition for the relaxation process must be defined. The most frequently used one is to set an upper limit on the process iteration times or to define a threshold value for the P_{ij} values. Fan and Tsai [15] proposed another condition which is adopted in this study. The condition requires that the P_{ij} value of any desired assignment of each syllable become greater than a pre–defined threshold value and that this P_{ij} value be increased for the last two iterations.

When the termination condition is satisfied, for each syllable within the sentence, the word of the assignment with the highest P_{ij} value is denoted as the word the syllable is assigned to, and then the corresponding characters are determined, too.

IV. EXPERIMENTAL RESULTS

A prototype system is built to test the applicability of the proposed approach. The computer is an IBM PC/AT compatible machine, with 640K RAM and 20M hard disk. A computer dictionary which includes about 4100 word entries is used for the experiment. Each word entry contains its length, phonetic codes, Chinese internal codes (Big 5 codes), and its usage frequency count. The word usage frequency counts recorded in [16] are directly used in the dictionary. The phonetic codes of the characters in a word are used as the index to search for the word in the word dictionary. The dictionary occupies 320K bytes of disk space.

Test data are selected from the articles in the " gwǒ yǔ rì bàu." They are coded in phonetic codes. Each phonetic code represents a syllable and occupies two bytes. The test data include 1318 sentences and 16347 syllables in total.

The processing procedure for each sentence is as follows. First, the dictionary is checked to find all possible words (each word being a string of syllables) corresponding to the syllables of the sentence. Then the syllable–to–word assignment is made. With the usage frequency counts of the words, the initial probability values of the assignments can be computed, and the adjacency relationships of the words can be analyzed to compute the compatibility coefficients between the assignments. Finally, relaxation iterations are performed until the termination condition is met. The result is printed as a segmented sentence with syllables converted into characters and words identified with spaces as markers between them.

The experimental result is summarized as follows. With the termination condition described as in Section III.F, and the threshold value of 0.8, the correct conversion (converting the syllables to Chinese characters) rate is 96.91%; the average processing speed is 0.97 second per syllable; and the average number of iterations per sentence is 17.2. Some examples of the results are shown in Figure 1. The underlined characters are errors.

V. CONCLUSIONS AND SUGGESTIONS

A new approach to Chinese phonetic input is proposed. The input method is regarded as a problem of assigning syllables to words. The advantages of the proposed method are that there are very few or even no homonyms for multi–syllabic words and that the adjacency relationships among words can be utilized for word identification based on the relaxation technique. The feasibility of this approach is proved by its high character conversion rate.

The following are some suggestions for further study.

(1). collection of more word entries in the dictionary to improve the syllable conversion rate. However, the contents of the dictionary essentially will still be limited. Also there exists trade–off between the dictionary volume and the processing speed.

(2). inclusion of the formation rules of compound words and word groups. Such rules may link lots of single syllables to form multi–syllabic strings, and so reduce syllable conversion ambiguity.

REFERENCES

[1]. "Evaluation Report of Chinese Input Methods and Input Devices," Institute for Information Industry, Taipei, Taiwan, R. O. C., June 1987.

[2]. S. K. Wan, H. Saiton, and K. Mori, "Experiment on Pinyin–Hanzi Conversion Chinese Word Processor," Computer Processing of Chinese and Oriental Languages, Vol. 1, No. 4, pp. 213–224, Nov. 1984.

[3]. W. H. Ho, C. C. Hsieh, K. Mei and C. T. Chang, Automatic Recgnition of Chinese Words, National Taiwan Institute of Technology, Taipei, Taiwan, R. O. C., 1983.

[4]. S. I. Chen, C. T. Chang, J. J. Kuo, and M. S. Hsieh, "The Continuous Conversion

Algorithm of Chinese Character's Phonetic Symbols to Chinese Character," Proceedings of 1987 National Computer Symposium, Taipei, Taiwan, R. O. C., Dec. 1987, pp. 437–442.

[5]. M. Y. Lin and W. H. Tsai, "Removing the Ambiguity of phonetic Chinese input by the Relaxation technique," Computer Processing of Chinese and Oriental Languages, Vol. 3, No. 1, pp. 1–24, May 1987.

[6]. A. Rosenfeld and A. C. Kak, Digital Picture Processing, Vol. 2, Academic Press, New York, 2nd ed., 1982.

[7]. A. Rsoenfeld, R. A. Hummel and S. W. Zucker, "Scene Labelling by Relaxation Operations," IEEE Tran. Syst., Man, Cybern., Vol. SMC–6, pp. 420–431, 1976.

[8]. A. Rosenfeld, "Iterative Methods in Image Analysis," Pattern Recognition, Vol. 10, pp. 181–187, 1978.

[9]. W. S. Rutkowsky, "Recognition of Occluded Shapes Using Relaxation," Computer Graphics and Image Processing, Vol. 19, pp. 111–128, 1982.

[10] S. Peleg, "A New Probabilistic Relaxation Scheme," IEEE Tran. PAMI, Vol. PAMI–2, No. 4, pp. 362–369, 1980.

[11]. O. D. Faugeras and K. Price, "Improving Consistency and Reducing Ambiguity in Stochastic Labelling: an Optimization Approach," IEEE Tran. PAMI, Vol. PAMI–3, pp. 412–424, 1981.

[12]. R. A. Hummel and S. W. Zucker, "On the Foundations of Relaxation Labelling Processing," IEEE Tran. PAMI, Vol. PAMI–5, pp. 267–287, 1983.

[13]. M. Y. Lin, and W. H. Tsai, "A New Approach to On–line Chinese Character Recognition by Sentence Contextual Information using the Relaxation Technique," Proceedings of International Conference on Computer Processing of Chinese and Oriental Languages, Toronto, Canada, August 1988.

[14]. S. I. Hanaki and T. Yamazaki, "On–line Recognition of Handprinted Kanji Characters," Pattern Recognition, Vol. 12, pp. 421–429.

[15]. C. K. Fan, and W. H. Tsai, "Automatic Word Identification in Chinese Sentences by the Relaxation Technique," to appear in Computer processing of Chinese and Oriental Languages.

[16]. I. M. Liu, et al., Frequency Counts of Frequently Used Chinese Words, Lucky Book Co., Taipei, Taiwan, R. O. C. 1975.

[17]. C. Y. Suen, Computational Studies of the Most Frequent Chinese Words and Sounds, World Scientific Publishing Co., Singapore, 1986.

```
sentenee 83
由於 愛 面子 的 人 太 多 .
sentence 84
所以 應酬 的 風氣 越 來 越 盛 ,
sentence 85
不 管 甚麼 事 ,
sentence 86
先 吃 一 頓 再 說 ; 光 吃 還 不 夠 ,
sentence 87
還 得 要 有 于 興 節目 ,
sentence 88
於是 各 種 飯店 常 有 變相 營業 的 情形 發生 ,
sentence 89
連帶 地 促使 各 種 行業 蓬勃 發展


sentence 147
在 童年 的 記憶 裡 ,
sentence 148
我 對 外公 的 印像 非常 深刻
sentence 149
他 常常 到 各 地 旅遊 ,
sentence 150
每 到 一個 地方 ,
sentence 151
就 把 當 地 名產 帶 回來 給 我們 這些 孫 兒 吃


sentence 17
每 逢 閱讀 的 課堂 ,
sentence 18
大家 便 沒 開 眼 笑 ,
sentence 19
欣喜 萬分 有 說 不 出 的 快樂 ,
sentence 20
像 是 考 了 第一 名 一樣 。
sentence 21
學校 圖書室 外面 的 走廊 ,
sentence 22
有 一 排 整齊 的 鞋 跪 ,


sentence 162
於是 我 加 快 腳步 ,
sentence 163
向 他 奔 去 ,
```

Figure 1. Some testing results. Underlined ones are errors.