Language, Information and Computation (PACLIC12), 18-20 Feb, 1998, 283-294

An Automatic Chinese Document Revision System Using Bit and Character Mask Approach

June-Jei Kuo
Matsushita Electric Institute of Technology (Taipei) Co., Ltd.
Rm.1002, 10 Fl., No.136, Sec. 3, Jen-Ai Road, Taipei, Taiwan, R.O.C.
kaku@mitt.com.tw

Abstract

The errors in Chinese document are mainly caused in two stages - input and editing. There are homonyms or homophones selection error, ambiguous pronunciation error, word segmentation error, similar shape character error, editing operation error and so on. In order to increase the quality of Chinese text, the conventional Chinese document revision system used the similar characters set and language model with some statistical date. Nevertheless, there are the following problems: (1) The perfect similar character set is difficult to make (2) Due to the copyright problem the large and balanced Chinese corpus is very difficult to be obtained (3) The above editing errors can not be solved simultaneously (4) The average success revision rate is not over 75%.

In this paper we study the Chinese features and phonetic-input-to-character conversion system for Chinese. It is found that the Chinese phonetic information and the related conversion algorithm are much help to detect and revise the input errors in Chinese document. As to the editing errors, a special code structure of Chinese pronunciation which has only one bit difference among similar pronunciations is proposed In addition, the bits and characters mask technology is also proposed respectively. The experimental result of the proposed system show that the average success revision rate of the proposed system is close to 87%.

Keywords: Natural language processing, document revision, bit mask,character mask, phonetic information, dynamic programming

1. Introduction

As the rapid progress of computer technology, the need for document computerization is also rised. So, how to enhance the document quality in the computer becomes an important issue. In addition, the document revision technology will be needed as a preprocessor or postprocessor in many application softwares, such as character recognition, speech recognition, machine translation and so on. Therefore, there are many approaches [1] on automatic document revision. The document revision can be further divided into error detection and error modification. Nevertheless, the document error types is dependent on its processing language. For example, the English spelling checker can not be used to revise Chinese or Japanese document due to the different error types. Thus, the error types of Chinese document will be described below.

The error types of Chinese document can be divided into two parts. The first part is caused by input stage and the other part is caused by editing stage. Moreover, in input stage the different input way, such as phonetic way or radical way, will also have different error types.

In the input part:

(1) Using phonetic way

1. Ambiguous pronunciation error

The phonetic symbols of a Chinese character can be divided into four parts: consonant, intermediate, vowel and tones. The possible ambiguous phonetic symbols in each part will be shown below.

Consonant: 【ㄕ】'sh' and 【ㄙ】's' and 【ㄑ】'ch' and 【ㄒ】'x' and etc.

Intermediate: 【ㄧ】'i' and 【ㄨ】'u and 【ㄩ】'iu'

Vowel: 【ㄙ】'eng' and 【ㄣ】'en', 【ㄢ】'an' and 【ㄤ】'ang' and etc.

Tone: Wrong tone will get wrong character.

For example, if the phonetic symbols of 【興趣】'xing4qiu4' were mistyping as 'xingqi', the output will become 【性器】.

2. Homonyms or homophones selection error

Chinese has some 1300 pronunciations, but there are 13,053 characters in Big5 code system. The usage frequency (learning function) is usually to be used as an important cue, but the homophones and homonyms selection is still unavoidable.

For example, 【同音意義字】or【同音意譯字】are the wrong input of Chinese word 【同音異義字】'homophone'.

3. Candidates selection error

The success conversion rate[4]~[6] using phonetic way is some 92%. The main problem is that there is no effective way to select the overlapping candidates. For example, inputted phonetic string "i3dian4nau3' is converted as "椅墊腦" other than "以電腦".

4. Reference dictionary error

Due to the careless there are some input errors existed in the reference dictionary.

(2) Using radical way

1. Similar shape character error

There are many similar Chinese characters, e.g. "糸" and "糸", "日"and "曰", so it is very easy for a user to input the wrong character using the radical way.

2. Similar radical combination error

There are many Chinese characters which has the similar radical combination. For example, if we input the radicals "smv" rather than "qmv", then the Chinese character "長" will be obtained rather than "表".

In the editing part:

1. Missing character error

   When we use the delete function carelessly, the missing character error will occur. For example, if we delete the character "料" among "一種資料庫" 'a kind of database', the string "一種資庫" will be no meaning.

2. redundant character word error

   When we use the copy function carelessly, the redundant character word will occur. For example, if we copy the character "學" to the Chinese string "去校" twice, the above string will become "去學學校".

3. Wrong character order word error

   When we use both the delete and copy functions carelessly, the wrong order character word error will occur. For example, "國際經濟" 'international Economics' may become "國經際濟" which will be no meaning.

Therefore, the automatic Chinese document revision system which can solve all the above error types is the idea one. In section 2, we survey the conventional Chinese revision system and describe the related problems. The design issues of the new automatic Chinese document revision system and the proposed system architecture will be pointed out in section 3. In section 4, we give an example to further explain our system. The experimental results will be shown and analyzed in section 5. Finally, we will describe the prospects and future development in section 6.

2. The conventional approach and its related problems

The conventional approaches[2][3] for the Chinese document revision use the similar character set and statistic data, e.g. bi-gram or tri-gram. First, they made survey on every Chinese character and collected the related Chinese characters which have the similar meaning or shape or pronunciation or input code into similar character set shown as the figure 1. For example, the Chinese character "力" "power" has the similar shape Chinese characters as "刀" "sword," "刃" "knife'. And, "厲" and "勵" are the its similar pronunciation characters. Nevertheless, the Chinese character "厲" is the similar meaning character to the Chinese character "利". Moreover, Chinese character "力" is the similar pronunciation character to the Chinese character "利". On the other hand, "糸" "hvif' is the similar input code to the character "糸" "vif'.

As to the statistical data, based on a training corpus, the co-occurrence probabilities of different combinations of successive syntactic tags are calculated. First, the corpus is segmented into a number of words. Then, the syntactic tags are assigned to these segmented words. Moreover, if the above two tasks are performed by some automatic procedures, it is necessary to correct the mistakes manually.

人：入S
力：厲p，勵P，刀S，刃S
利：厲M，力P，判S，剃S，剩S，・・・
己：已S，巳S，乙S
干：甘P，乾P，千S
弋：戈S
急：岌p，疾M，
冶：治S
糸：糸I
：
：

(S:similar shape, P:similar pronunciation, M:similar meaning, I:similar input code)

Figure 1 The example of similar character set

Thus, the bigram model[3] is used to calculate the syntactic co-occurrence probability of a number of words. For certain pair of syntactic tags Tx and Ty, P(Tx|Ty) are defined as follows.

$$P(Tx) = \frac{\text{number of Tx used in the training corpus}}{\text{number of all words in the training corpus}}$$

$$P(Tx|Ty) = \frac{\text{number of Tx following Ty in the training corpus}}{\text{number of Ty in the training corpus}}$$

The system architecture and processing flow of the conventional Chinese document revision system is shown as the figure 2. The Chinese document is inputted through the input device, such as scanner, hard disk, keyboard and so on. Thus, the substitution device substitute the characters of input Chinese document into their similar characters by referring the above similar character set. And then, the language model evaluation device[7] contains the evaluation device, such as Markov probabilistic model, and search device, such as dynamic programming. After the finding the optimal string through the lattice of the possible candidates, it compare the characters of optimal character string with the input Chinese string to detect and mark the wrong characters in wrong characters detection device. Finally, it will output the marked Chinese string and optimal string to the storage device. The average success revision rate[2][3] is some 75%.

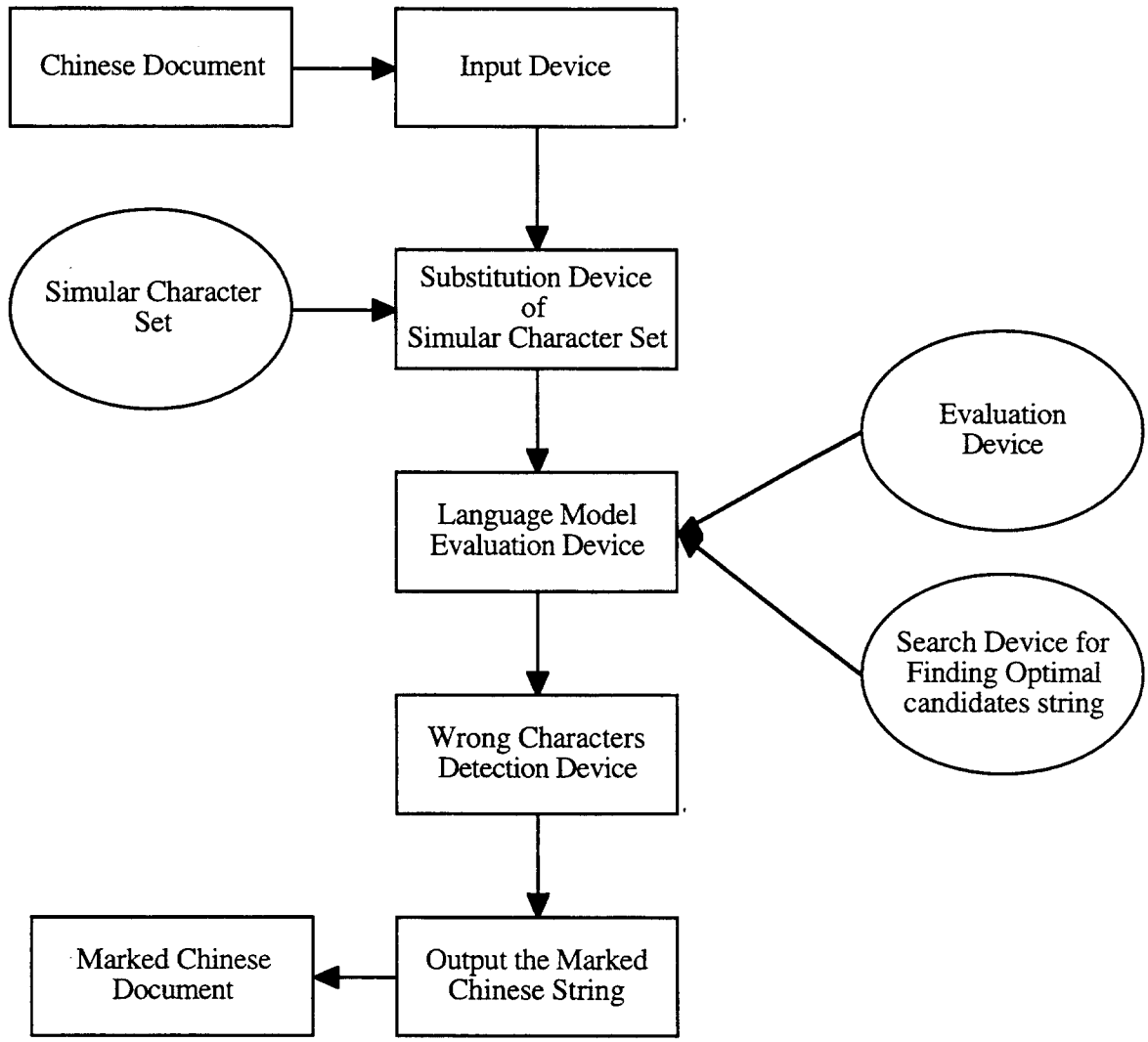


Figure 2 The processing flow of the conventional approach

2.1 the related problems

There are some problems shown as the following.

[1] The revision accuracy rate is largely influenced by the quality of the similar characters set, but how to make the similar characters set effectively and correctly will become a difficult issue.

[2] In order to obtain the character bi-gram or tri-gram(statistical data), a large and balanced Chinese corpus will be necessary, but such a corpus is also difficult to be obtained due to the copyright problem.

[3] the problem of missing or wrong order characters can not be solved effectively in the conventional approaches.

[4] The 75% revision rate is not satisfactory.

3. The proposed automatic Chinese revision system

In order to be able to solve the above mentioned problems of conventional approaches, we study the Chinese feature and try to find some useful information for us to solve both the input and editing errors in Chinese document.

3.1 The survey of Chinese feature

There are 13,053 Chinese characters in BIG5 code, but some 3,000~4,000 characters commonly used in the modern Chinese. Chinese characters are already the basic semantic and syntactic units and can be used independently to express certain meaning in Chinese. Nevertheless, only the Chinese characters are not sufficient for usage. Thus, two or more characters are grouped together to form a word, which is also a complete semantic and syntactic unit in Chinese. Moreover, one character can also be seen as a special one-character word. The number and usage of Chinese word[3] will be shown in the Figure 3.

Word length	Number	Usage
One character	12.1%	64.3%
Two characters	73.6%	34.3%
Three characters	7.6%	0.4%
Four characters	6.4%	0.4%
Five characters or longer	0.2%	0.0%

Figure 3 The number and usage of Chinese word

From the above survey result, it is found that it is uncommon to use more than five successive one-character words in a Chinese document. Thus, the presence of successive one-character words, e.g. three or over three, found in the segmented Chinese document can be used a hint to detect the document errors.

3.2 The survey of phonetic-input-to-character conversion system for Chinese

As to the above wrong selection of homophones or homonyms, it is found that there are many phonetic-input-to-character conversion system for Chinese[4]~[8] shown as the Figure 3 and the average conversion success rate are some 92% which is higher than the average revision rate, e.g. 75%, of Chinese document. So, the phonetic information and the related algorithm will be another important cue to detect and revise the the wrong selection problems of homophones or homonyms.

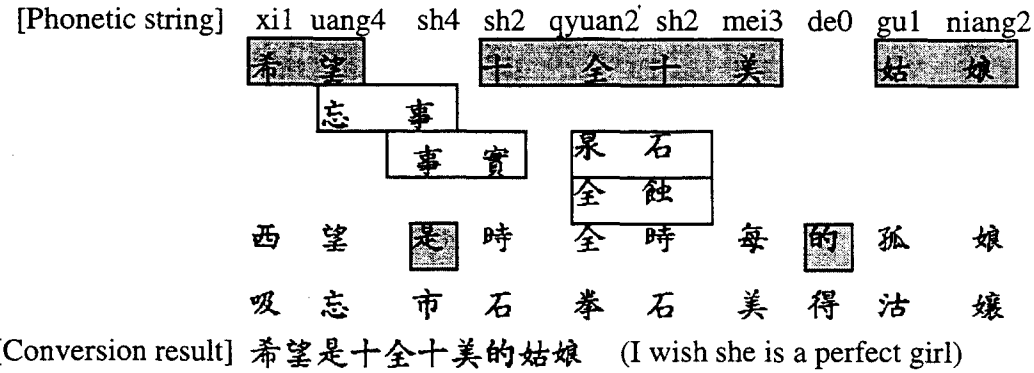


Figure 3 The example of phonetic-input-to-character conversion for Chinese

3.3 The Chinese pronunciation structure and the bit mask technology

There are some 1300 kinds of pronunciation of Chinese characters. Moreover, a pronunciation of a Chinese character can be divided into four parts which are consonant, intermediate, vowel and tone. In addition, there are 20 consonant, 3 intermediates, 13 vowels and five tones respectively. Therefore, the two byte structure[4] , shown as Figure 4, of a Chinese character pronunciation will be used. In that, the similar phonetic symbols are assigned to 1 bit difference and so the similar bits mask technology can be used to solve the ambiguous pronunciation problems effectively. Furthermore, the example of bit mask technology will also be shown as Figure 5. For example, if we mask (or ignore) the last bit of the phonetic symbol "ㄓ" and "ㄕ", both bit representations will be the same.

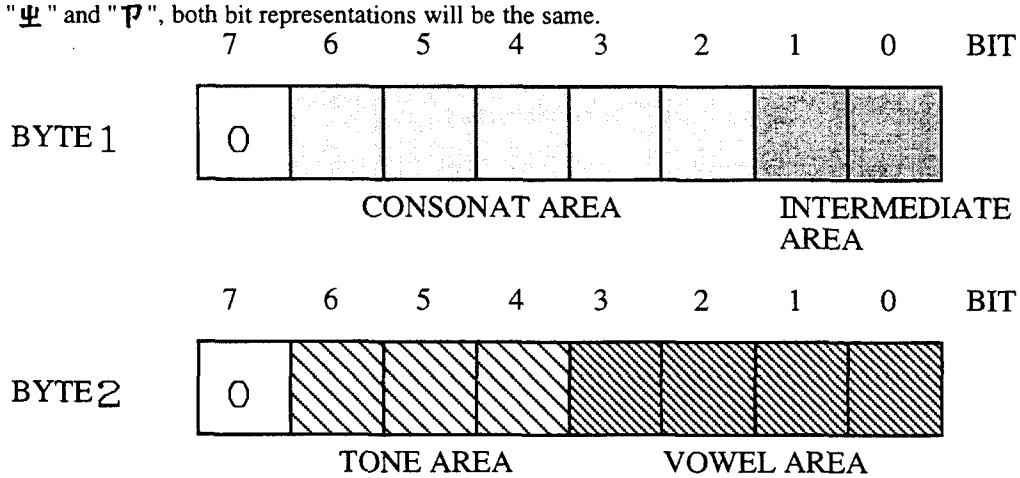


Figure 4 The two byte structure of Chinese character pronunciation

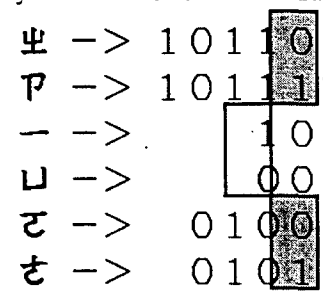


Figure 5 The example of bit mask technology

Therefore, the bit mask technology can help us to find all the similar candidates by referring the dictionary.

3.4 The character mask technology

Because the successive one-character words in Chinese document is uncommon use, so this hint can be used to detect the missing character or redundant character or wrong order words error. The character mask pattern between two or three successive one-character words will be shown below. In this paper, in order not to obtain large number of candidates we only consider two character mask patterns.

The successive two one-character words: **A,B,C**

Two Character mask patterns: ***AB, A*B, AB*, *A, A*, *B, B***

Three character makk patterns: **AB*C,*ABC,A*BC, AB*C, ABC***

(*:ignored character)

For example, if we use the phonetic string of successive one-character words are "z1" and "ku4", then all the possible candidates using the character mask technology will be shown below.

"z1*ku4": "資料庫" 'database' "*z1ku4*": None

"*z1ku4": None "z1*": "知道"'know', "知識"'knowledge',...

"*z1": "樹枝"'branch', ... "ku4*": "褲子"", "酷熱"'hot',...

"*ku4": "倉庫"'warehouse', "內褲"'underwear',...

3.5 The similarity weight with the input text and the word length weight

It is also found that the more similar characters the candidates has comparing with the input text, the more possible the candidates are the optimal candidates. So, the similarity weight will be defined below. For example, the candidate and the corresponding characters in input text are "資料庫" and "資庫" respectively, so its similarity weight will be 2/3.

$$\textbf{Similarity weight} = \frac{\textbf{The number of similar character}}{\textbf{The total character number of the candidate}}$$

As we know, the word length is a very important cue to find the optimal conversion result in the phonetic-input-to-character conversion system for Chinese or Japanese. So, in order to find the optimal path effectively through the candidates, the following word length weight. is also considered in our proposed system. For example , the word length weight of the candidates "資料庫" 'database' is (3-1)*2=4.

Word length weight = (word length-1)*2

3.6 The system architecture and processing flow of the proposed Chinese document revision system

We adopt the the algorithm of the phonetic-input-to-character conversion system for Chinese[8] as our processing processing engine because of its high conversion performance. The related score function of the proposed Chinese document revision system will be described below.

3.6.1 The score function

In the proposed Chinese revision system , we use the backward dynamic programming to find the optimal path through the candidates network. The score function will be shown below. Among them, beside the above character similarity weight and word length weight, the usage weight obtained by long and short term learning function and the semantic similarity weight in the conversion algorithm[8] will also be adopted in order to enhance the performance of the proposed system.

SCORE = USAGE WEIGHT + WORD LENGTH WEIGHT +
CHARACTER SIMILARITY + SEMANTIC SIMILARITY

3.6.2 The configuration and processing flow of proposed Chinese document revision system

The configuration and the processing flow will be shown as Figure 6.

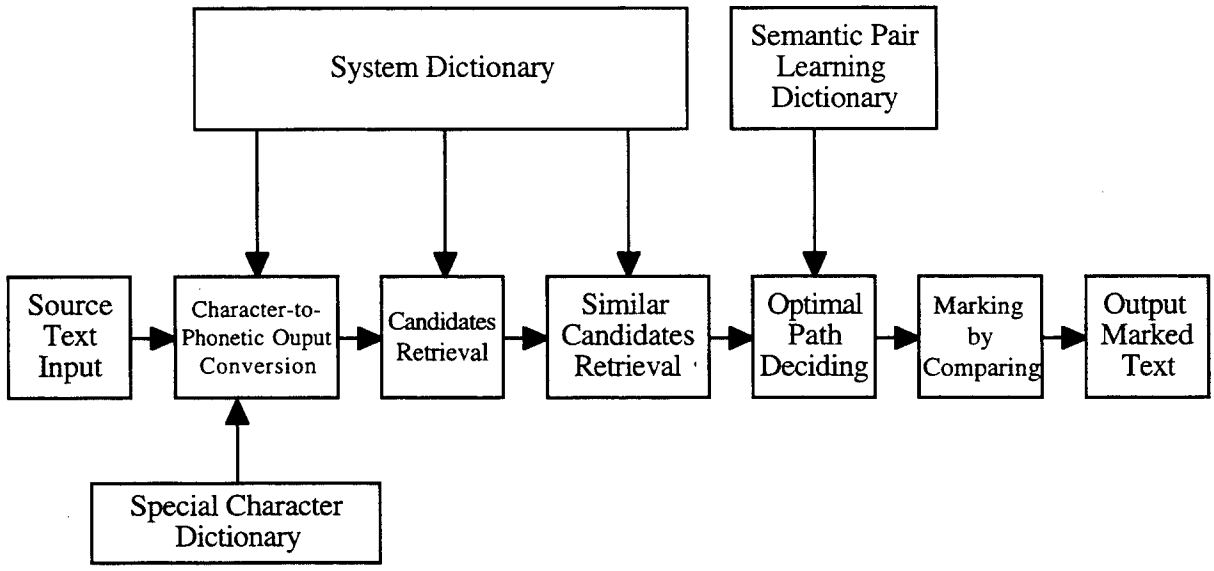


Figure 6 The processing flow of the proposed Chinese revision system

First, the Chinese text will be input from the input device in the source text input, then the character-to-phonetic output conversion software [10] is used to convert the input Chinese text into phonetic string by referring the special character dictionary, e.g. Poinzi dictionary "破音字字典" and system dictionary in the character-to-phonetic conversion module. Thus, candidate retrieval module retrieves all the possible

Chinese candidates by referring the system dictionary. And then, the similar candidates retrieval module uses the bit mask and character mask technology to retrieve all the similar candidates by referring the system dictionary. After that, the optimal path deciding module will use the start and end position of each candidate to link a processing net and use the dynamic programming and the above score function to find the optimal path through the candidates by referring both the system dictionary and the semantic pair learning dictionary[8]. In the marking by comparing module, the optimal path Chinese string will be compared with the input Chinese text and then output the optimal path string and the marked input text to the output device in output marked text module.

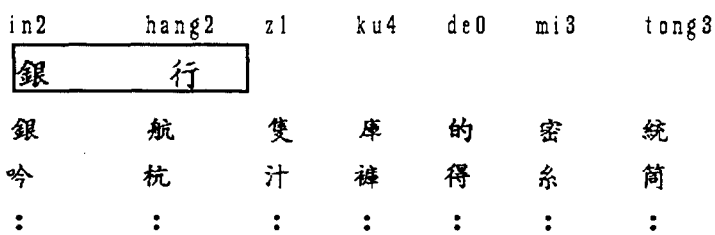
4. Example

In order to further explain our proposed Chinese document revision system, an example will be given in this section below.

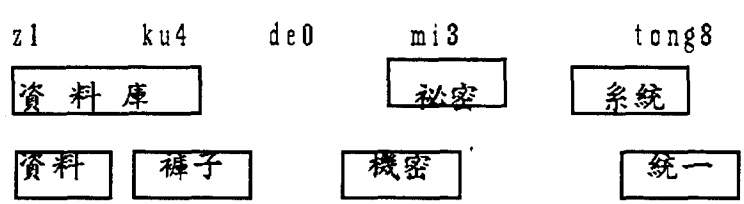
[Input Chinese text] 銀行資庫的系統

[Character-to-phonetic conversion] in2hang2z1ku4de0mi4tong3

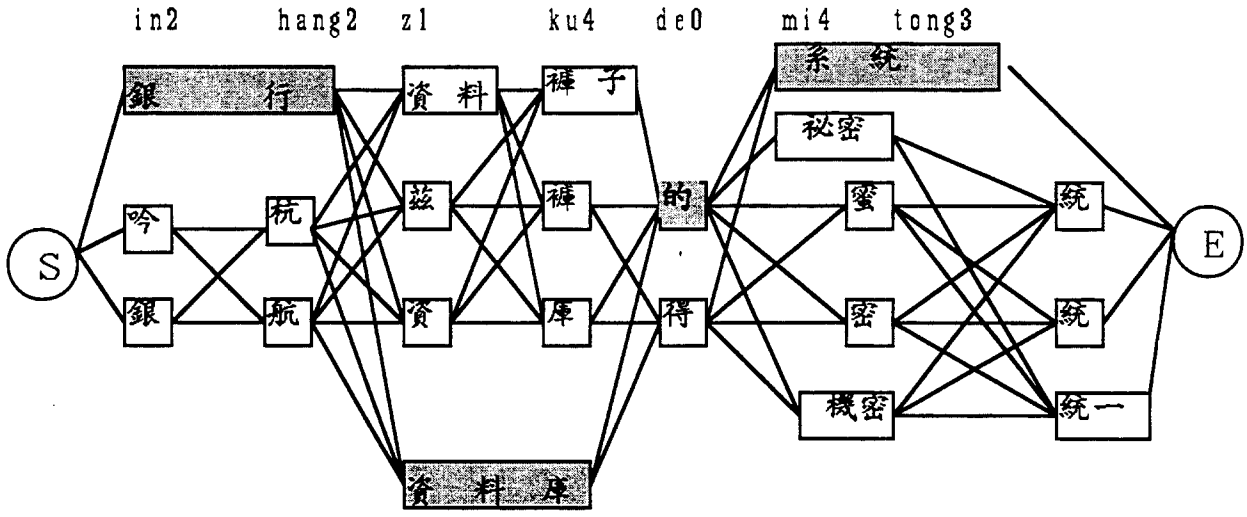
[Candidates retrieval]



[Similar candidates retrieval]



[Net and optimal Path deciding]



[Marked the input text by comparing]

Input Text : 銀行資庫的系統

Optimal String : 銀行資料庫的系統

Marked Output : 銀行資卄料庫的＊系統

(卄 : Missing character , ＊ : Similar radical character)

5. The experimental results and analysis

The 11 articles (some 8,620 characters) selected from the text books of the primary school in Taiwan were used as the test samples. Those articles were modified manually, e.g. deletion, insertion, and so on, and the modification records were also memorised. Although the success conversion rate of Chinese character-to-phonetic conversion[10] is over 99%, but in order to get rid of the influence of character-to-phonetic conversion completely, the phonetic strings of the above articles are checked manually. Moreover, in order to reduce the number of Chinese chartacter candidates we only use those Chinses characters which have been used recently in the learning dictionary and the system were implemented on Power PC and C language. The accuracy of both detection and correction is 86.4% and the average execution speed was some 3 characters/second.

The analysis of errors will be described below.

(1) Because the Chinese names or unknown words are usually successive one-character words, so the proposed revision system can not process them effectively.

(2) There are some errors caused by the typing errors in the system dictionary, e.g. [形影不離] be mistyped as [行影不離] in the system dictionary.

(3) The proposed revision can not process the discourse error, e.g. [他明天死了]' He died tomorrow.'.

(4) The similar candidates are not enough due to few character mask patterns.

6. Concluding remarks

The above proposed Chinese revision system without the necessity of similar Chinese characters set or large marked Chinese corpus was applied to the post processing of Chinese editors and the satisfactory results (both the document quality and editing time) was obtained. Moreover, those dictionaries of our proposed system are the same with the Chinese input conversion system, so the development cost and time can also be decreased. In order to further improve the performance of the proposed Chinese revision system, there are still some future works.

[1] Develop some effective algorithms for processing the unknown words or Chinese names in the Chinese document.

[2] Introduce the heuristic way to decrease the candidates and increase the execution speed.

[3] Introduce some discourse information to solve the semantic contradictions problems.

[4] Introduce the radical information.

[5] Use more Chinese text to evaluate the proposed revision system.

7. References

[1]池原　悟など、”文章校正支援システムにおける自然言語処理”、情報処理、vol.34, No.10,PP1249~1257,1983

[2]施得勝等，“基於統計的中文錯字偵測法“，電腦與通訊，Vol.8, 1992, PP19~26

[3]C.H. Leung and W.K. Kan, "Difficulties in Chinese Typing Error Detection and Ways to the solution", Vol.10, No.1, 1996,PP97~113

[4]J.J. Kuo, J.H. Jou, M.H. Hsieh and F. Maehara, "The development of New Chinese Input Method---Chinese Word-string Input System", Proceeding of International Computer Symposium, PP1470~1479, 1986

[5]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",Proceeding of National Computer Symposium, PP437~442, 1987

[6]M.L. Hsieh, T.T. Lo and C.H. Lin, "A Grammar Approach to Converting Phonetic Symbols into Characters", Proceeding of National Computer Symposium, PP453~461, 1989

[7]R. Sporat,"An Application of Statistical Optimization with Dynamic Programming to phonetic-input-to-character conversion for Chinese",Proceeding of ROCLING III, PP380~390, 1992

[8]J.J Kuo, "Phonetic-input-to-character Conversion System for Chinese Using Syntactic Connection Table and Semantic Distance", Computer Processing of Oriental Language, Vol.10, No.2, 1996, PP195~210

[9]大野　晋、浜西正人、類語国語辞典、角川書店、1986

[10]蔡祈岩，“文句翻譯語音系統中破音字及音韻處理之研究“，成大資訊工程所碩士論文，1995