An Automated Mandarin Document Revision System Using both Phonetic and Radical Approaches

June-Jei Kuo
Panasonic Taiwan Laboratories Co., Ltd.

Abstract

Correcting the errors in an inputted document is an important issue in the pre or post editing process of machine translation. An essential problem of automated document revision system is how to tackle the all kinds of error types, such as literal, hiatus, cacography, homophones and so on, in a document simultaneously and effectively. In this paper we focus on mandarin document and propose a Chinese character phonetic structure that the similar pronunciation characters will have only one bit difference, so the bits and characters mask can be conducted to find the possible candidates for homophone and hiatus errors. Moreover, in order to improve the system performance we also propose a similar character set using each Chinese character radical information that not only the literal or cacography errors can be revised effectively, but also the number of candidate characters can be reduced largely. The dynamic programming is used to decide the optimal path through the candidate lattices and the experimental results show a better performance than those of other methods.

1. Introduction

As the rapid progress of computer technology, the need for document computerization is also risen. So, how to enhance the digital document quality becomes an important issue. In addition, the document revision technology will also be needed as a preprocessor or postprocessor in many application software, such as character recognition, speech recognition, machine translation and so on. Therefore, there are many approaches [1] on automation document revision. The document revision can be further divided into error detection and error modification. Nevertheless, the document error types are related to the 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 mandarin document will be discussed first below.

The error types of mandarin document can be divided into two categories. The first one is the input errors and the other one is the editing errors. Both of them are caused mainly by careless and misuse.

I. The input error types

(1) Ambiguous pronunciation error

The phonetic symbols of a Chinese character can be divided into four parts: consonant, intermediate, vowel and tones. The examples of possible ambiguous phonetic symbol pair in each part will be shown below. ([]: Chuyin phonetic symbols, '': Pinyin phonetic symbols)

Consonant: [ㄕ] 'sh' and [ㄙ] 's'; [ㄑ] 'q' and [ㄒ] 'x'; [ㄓ] 'zh' and [ㄗ] 'z'

Intermediate: [ㄧ] 'i' and [ㄩ] 'iu'

Vowel: [ㄥ] 'eng' and [ㄣ] 'en'; [ㄤ] 'ang' and [ㄢ] 'an'

Tone: wrong tone will lead to get wrong character.

For example, the phonetic symbol string of [興趣] "hobby" is 'xing4qi4. If the phonetic symbol string is mistyped as 'xing4ci4', the Chinese character conversion result will become [性器] "sexual organ".

(2) Similar shape characters error

There are many similar shape Chinese characters, e.g. [系]"department" and [糸]"silk"; [曰]"say" and [日]"day";[長] "long" and [表] "table". Thus, it is very easy to input the wrong character due to the careless and the shape similarity.

(3) Homophones or homonyms selection error

Chinese has some 1345 kinds of pronunciation, but there are at least 13,053 characters in any Chinese character code system, such as Big5. So, one Chinese pronunciation may have over 10 homophones. Thus,

users usually need to select right character or word from the homophones or homonyms whenever they use phonetic-input-to-character conversion system [5]~[7]. Due to the inadvertence it is also very easy for a user to select wrong character or word candidates.

(4) Word segmentation error

As to Chinese phonetic-input-to-character conversion system described above, the average success conversion rate is 92%~96%. The main problem is that there is not effective way to select the suitable one from the overlapping candidates. For example, the inputted phonetic string 'i3 dian4 nau3' will usually be converted to [椅墊腦] 'cushion brain' other than [以電腦] "... by the computer".

(5) Errors in reference dictionaries of character input system

Such errors will cause wrong character input as long as those errors in the reference dictionaries are not correct.

II. The editing error types

(1) Hiatus error (Missing character error)

Whenever we use the delete function of editor software, the missing character error will occur due to careless. For example, if the user delete the character [識] from the character string [一種知識庫] 'a kind of knowledge base' inadvertently, the result character string [一種知庫] will be no meaning.

(2) Literal or cacography error (Misused character error)

For example, the Chinese word [按倍就班] 'step by step' is easy to be inputted as [按步就班].

(3) Redundant characters or wrong character order words error

Whenever we use delete and paste functions carelessly, it is easy to have the redundant or wrong order characters words in the editing text. For example, [國際經濟] 'International Economics' may be changed to be [國經際濟] which is no meaning.

Therefore, the ideal document revision system should be able to solve all the types of error above effectively and simultaneously. In section 2 the conventional mandarin document revision system will be shown and the related problems will be described. Thus, the design issues and the proposed system architecture will be pointed out in section 3. In section 4, an example will be given to further explain the proposed system. In section 5 the experimental results will be given and analyzed. Finally, the prospects and the future development will be described in section 6.

2. The conventional approaches and its related problems

The conventional approaches [2][3] to the mandarin document revision used the similar character set and statistical data to detect and modify the document errors. As to the similar character set, they studied the features of every Chinese character first. And then, they used the clustering technology to divide them into several groups which the characters in each group have the similar meaning or shape (radical input code) or pronunciation with each other, shown as the figure 1. For example, the Chinese character [力] "power" has the similar characters [刀]"knife" and [刃]"edge" in shape. Meanwhile, [厲] and [勵] are its similar characters in pronunciation 'li4'. However, the Chinese character [厲] is also the similar character in both meaning and pronunciation to the Chinese character [利].

As to the statistical data, the co-occurrence probabilities of different combinations of successive syntactic tags in a Chinese tagging corpus are calculated, such as part-of-speech bi-gram or tri-gram. Thus, during the document revision processing the bi-gram model [3] is used to calculate the co-occurrence probability of successive two words by referring their syntactic markers.

人：入 S
力：厲 P，勵 P，刀 S，刃 S
利：厲 P，力 P，判 S，剎 S，踩 S，・・・・・・・
己：已 S，已 S，乙 S，・・・・・・
干：甘 P，乾 P，千 S
弋：戈 S
急：发 P，疾 M
冶：治 S
：

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

Figure 1 Example of similar character set

2.1 The outline of conventional system architecture

The system architecture and processing flow of the conventional mandarin document revision system is

shown as Figure 2. The mandarin document is inputted through the input device, such as scanner, keyboard, hard disk and so on. And then, the substitution device refers the characters of inputted document and the similar character set to find all the possible substitution characters. After that, the language model evaluation device [2][8] use the evaluation device, such as statistical data described above, and the search device to find the optimal character string through candidate lattices. Then, the wrong character detection device compares the characters in the optimal character string with those characters in the inputted character string and marks the wrong characters of the inputted string. Finally, it output both the marked Chinese inputted character string and optimal character string to the storage device. The average success detection and correction rates are some 75% respectively.

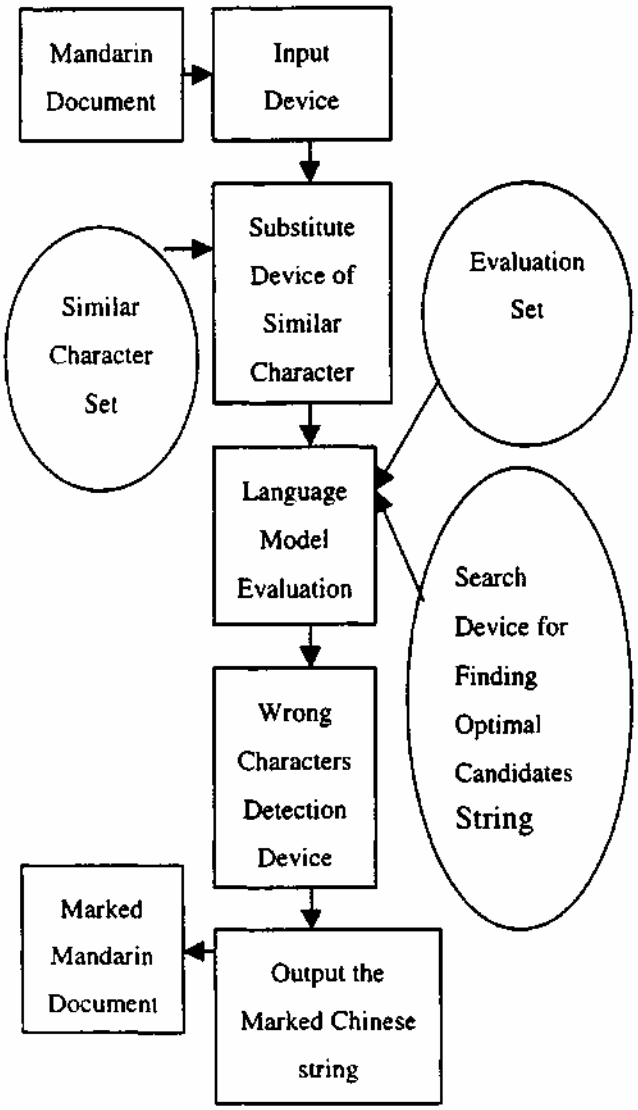


Figure 2 The flow of the conventional approach

2.2 The related problems

There are some problems of above revision system shown as the following.

1. The revision accuracy rate is largely influenced by the quality of similar characters set. However, it is not only very difficult to make the high quality similar characters set, but also there is not effective way to maintain.
2. In order to obtain the bi-gram or tri-gram (statistical data), a large and balanced Chinese corpus will be necessary. Nevertheless, such corpus is very difficult to be obtained due to the budget and copyright problems.
3. The hiatus or wrong character order word problem can not be solved effectively.
4. The 75% revision accuracy rate is not satisfactory.

3. The proposed automated mandarin document revision system

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

3.1 The survey of Chinese feature

There are 13,053 Chinese characters in BIG5 code system, but only 3000 ~ 4000 characters commonly used in the modern Chinese. On the other hand, Chinese characters are already the basic semantic and syntactic units and can be used independently to express certain meaning in Chinese. However, 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. Th number and usage of Chinese words[3] will be shown in Figure 3.

Word length	Number ratio	Usage ratio
One character	12.1%	64.3%
Two character	73.6%	34.3%
Three character	7.6%	0.4%
Four character	6.4%	0.4%
Five character or longer	0.2%	0.6%

Figure 3 The number and usage of Chinese word related to their lengths

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

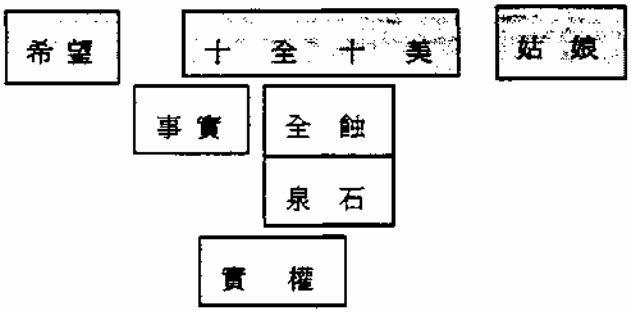
3.2 The survey of Chinese phonetic-input-to-character conversion system

There are many Chinese phonetic-input-to-character conversion system [5]~[9], shown as the Figure 4 and their average conversion success rate is 92%~96%. Moreover, Their conversion accuracy rate is far higher than the 75% accuracy rate of conventional document revision. So, the phonetic information and the related conversion algorithms, e.g. long word preference, can be another important cue to detect and revise the homophones (homonyms) selection errors.

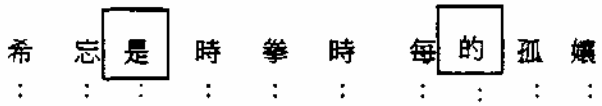
[Phonetic String]

xi1 uang4 sh4 sh2 qyuan2 sh2 mei3 de0 gu1 niang2

[Word Candidates]



[Char. Candidates]



[Conversion Result] 希望是十全十美的姑娘

"I wish she is a perfect girl."

Figure 4 Example of Chinese phonetic-input-to-character conversion

3.3 The survey of Chinese radical-input-to-character conversion system

Beside the phonetic-input-to-character conversion system, the radical-input-to-character conversion will be another important input software, such as Chang-Jie (倉頡) input system. The Chang-Jei input method uses 40 kinds of Chinese radical to form all the Chinese characters. The structure of its reference dictionary will be shown as Figure 5. For example, if users input III (木木木) and press a space key, [森] 'forest' will be inputted. Furthermore, it is evident that the similar shape characters will have similar radical combinations. Thus, if some radicals of a Chinese character are masked, the similar shape character candidates can be obtained. So, the radical information will be helpful to detect and revise the similar shape characters error.

Chinese character	Radicals combination
:	:
辨	YJILJ（火月木中月）
辦	YJKSJ（火月大尸月）
辯	YJYRJ（火月火口月）
辦	YJVFJ（火月女火月）
:	:
森	III（木木木）

Figure 5 The dictionary structure of Chang-Jie input

3.4 The proposed Chinese pronunciation structure and mask technology

In order to solve the ambiguous pronunciation errors, a phonetic symbol structure, which the similar pronunciation symbols will have only one bit difference one another, is proposed. In Chinese, there are some 1345 kinds of pronunciation. As to the phonetic symbols, there are 20 consonants, 3 intermediates, 13 vowels and 5 tones respectively. So, two bytes structure is enough to represent all the Chinese pronunciations and the proposed Chinese pronunciation structure will be shown as Figure 6. (The value of the 7th bit of each byte is set to be 0)

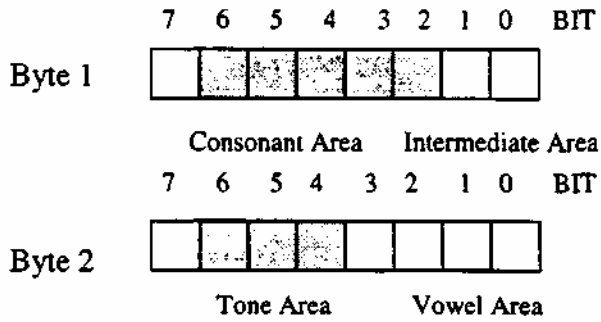


Figure 6 Two bytes structure of Chinese character pronunciation

Meanwhile, the structure example of ambiguous Chinese phonetic symbols will be shown as Figure 7. (*: any)

ㄓ'zh' -> 010110** (byte 1)
ㄗ'z' -> 010111** (byte 1)
ㄧ'yi' -> ******10 (byte 1)
ㄩ'yu' -> ******00 (byte 1)
ㄛ'o' -> ****0100 (byte 2)
ㄜ'e' -> ****0101 (byte 2)

Figure 7 One bit difference example of ambiguous phonetic symbols

3.4.1 The bit mask for ambiguous phonetic symbols

The first bytes of phonetic symbol [ㄓㄨㄥ] and [ㄗㄨㄥ] are '01011001' and '01011101' respectively. On the other hand, the second bytes are the same. So, if the bit 2 of the first byte is masked as '01011*01', it is easy to use this pattern to consult all similar pronunciation candidates from the reference dictionary. Thus, the ambiguous pronunciation errors can be handled easily. For example, as to phonetic symbols [ㄗㄨㄥ ㄒㄧㄣ ㄘㄨㄥ], if the above bit mask technology were used, then the related candidates[中心]'center',[忠心]'loyalty',[衷心]'with my best wish' will be obtained rather than no candidate due to the Chinese ambiguous pronunciation.

3.4.2 The character mask

Because the successive one character words in mandarin document is unusual, so this cue can be used to detect the missing character or redundant character or wrong character order word errors. The character mask pattern between successive one character words will be shown below. In this paper, in order to avoid the large number of candidates we only consider those successive two and three one character words in the processing sentence. Moreover, the proposed one character mask patterns will be shown below.

The successive three one-character words in processing text: A, B, C (*: mask symbol)

Two character word pattern: *A,A*,*B,B*,*C,C*

Three character word pattern: *AB, A*B, AB*, *BC, B*C,BC*

Fourth character word pattern: ABC,A*BC, AB*C,ABC*

A,B,C: the phonetic symbols of successive characters

So, the above mask patterns can be used to get the possible candidates by referring the dictionary. For example, if there is two successive one character words "z1" and "ku4" respectively, then the possible candidates using the mask pattern will be shown as the following.

"*z1":[樹枝]'branch',[果汁]'juice', ...
"z1*":[知道]'know',[資料]'data',....
"*ku4":[倉庫]'warehouse',[內褲]'underwear',.....
"ku4*":[庫存]'storage',[酷熱]'hot',...
"*z1ku4": none "z1*ku4":[資料庫]'database'

3.4.3 The radical mask

In order to tackle the cacography error or similar shape error problem, we can mask few radicals of a Chinese character to retrieve all the similar shape candidates by referring the Chinese radical reference dictionary. For example, the radical combination of character [辨] is [YJIIJ]. Thus, if we use the radical pattern [YJ**J] (*: mask symbol) to retrieve Chinese characters from the radical reference dictionary shown as Figure 5, then the similar shape Chinese characters [辨], [辯], [辮], [瓣] will be obtained.

3.5 The proposed mandarin document revision system

3.5.1 The configuration

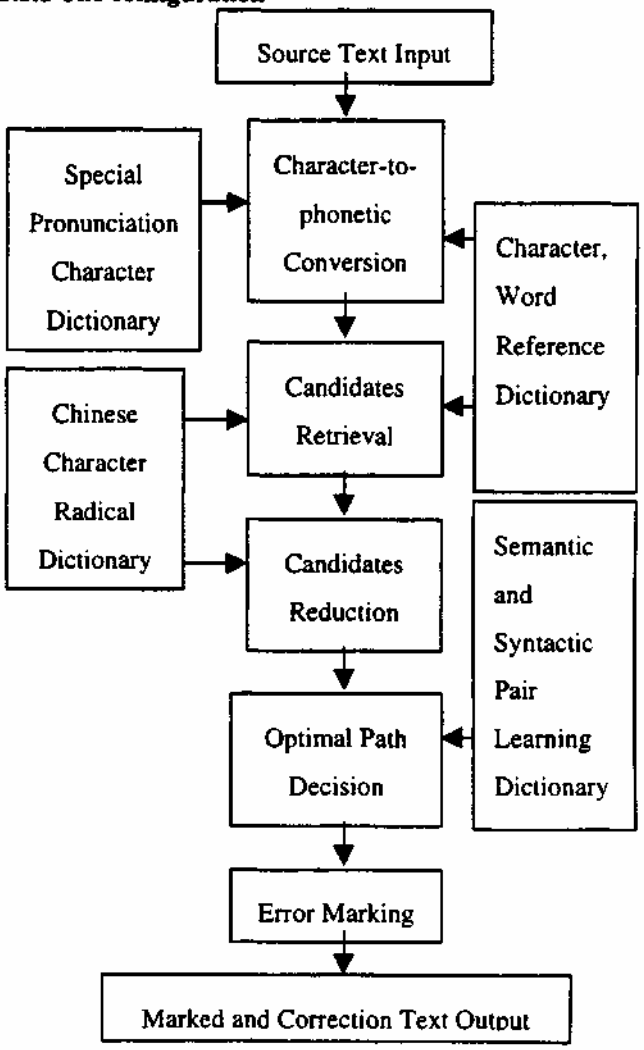


Figure 8 Configuration of the proposed document revision system

The proposed system configuration and processing flow, which introduces the above special Chinese phonetic structure and mask technology will be shown as Figure 8.

The Chinese text will be inputted from the source text input module, then the character-input-to-phonetic conversion module [11] will convert the inputted text into phonetic symbols string by referring the character, word reference dictionary and special pronunciation character dictionary, such as Poinzi dictionary [破音字字典]. Thus, candidate retrieval module will not only use the above phonetic string to retrieve all the possible candidates referring the character, word reference dictionary, but the bit and character mask will also be used to retrieve the similar shape and pronunciation candidates. And then, in order to speed up the performance of the proposed system, candidates reduction module will delete some candidates using radical similarity, which will be described below. After that, the optimal path decision module will use the corresponding start and end position of each candidate in the inputted text to link a directional net. Moreover, the backward dynamic programming will be used to find the optimal path through the candidate lattices by referring the value of the score function shown as Figure 9. In the error mark module the Chinese character string in the optimal path will be compared with the inputted text and mark the error as error detection marks. Finally, both the optimal Chinese string and the marked Chinese string will be outputted.

Max. Score(i) = usage frequency weight(i) + word length weight (i) +text similarity weight(i) + syntactic weight (i,j) +semantic weight(i,j)

i: the processing node , j: the adjacent nodes of node i

Figure 9 Definition of the score function

3.5.2 The usage frequency weight

In order to solve the homophones and homonym selection problem, the usage frequency is an important factor [5][6]. So, we use the character or word usage frequency in the "Modern Mandarin frequency dictionary"[12] as the frequency weight in the score function. For example, the frequency of [科學] 'science' and [地方] 'place' will be 0.0745 and 0.0527 respectively.

3.5.3 The text similarity weight and word length weight

It is also found that the more same characters the candidates have comparing with the inputted text, the more possibility the candidates are the optimal candidates. Thus, the text similarity weight in the score function will be defined below. For example, if the candidate and the corresponding characters in the inputted text are [資料庫]'database' and [資庫] respectively，the text similarity weight of candidate [資料庫] will be 2/3.

$$\text{Text Similarity Weight} = \frac{\text{The number of same characters}}{\text{The total number of the candidate}}$$

On the other hand, the word length is also a very important cue, e.g. the longest word preference algorithm, to decide the optimal conversion result in the phonetic-input-to-character conversion system for mandarin or Japanese. So, the definition of word length weight of a candidate will be shown below. For example, the word length weight of candidate [資料庫] will be (3-1)*2=4.

Word Length Weight = (the character number of a candidate – 1)*2

3.5.4 The syntactic and semantic similarity weight

In order to introduce the syntactic and semantic information, we defined 34 Chinese syntactic markers [9] which are used to mark up our Chinese corpus. Nouns are divided into 10 categories, e.g. A0 (common noun), A1(country name), A2(place name) and so on. Numbers are further divided into 3 categories, e.g. B1(Arabic numerals), B2(number unit) and B3(the characters or words before number). Meanwhile, the semantic code system [10] is also introduced in our Chinese segmented corpus. So, we can obtain the syntactic connection table shown as Figure 10 and the semantic connection table shown as Figure 11 by referring the adjacent semantic and semantic markers in the corpus.

	A0	A1	A2	A3	A4	A5	A6	A7.....	Q0	R0	S0
A1	1	1	1	1	1	1	1	1.....0	1	1	
A2	1	1	1	1	1	1	1	0.....0	1	1	
A3	1	1	1	1	1	1	1	0.....0	1	1	
A4	1	1	1	1	1	1	1	0.....0	1	1	
:	:	:	:	:	:	:	:	:.....:	:	:	
Q0	0	0	0	0	0	0	0	0.....0	0	0	
R0	2	1	1	1	1	1	1	0.....0	0	0	
S0	1	1	1	1	1	1	1	0.....0	0	0	

Q0: ominal, R0: "be" verb, S0: localizer

2: strongly connectable, 1: connectable, 0:unconnectable

Figure 10 Example of the syntactic connection table

Key semantic code	The adjacent semantic codes
061b'animal'	828h'classifier'
135a'strong'	3850'pile'
3950'xonstruction'	2580'strength'
464a'attack'	714'army',5120'children',...
:	:

Figure 11 Example of the semantic connection table

By referring the above syntactic and semantic connection tables, the definition of the syntactic and semantic similarity weights are shown below.

Syntactic similarity weight of node i and node j = st(i,j)*0.5

st(i,j): the value of syntactic connection table by referring the syntactic markers of node i,j

Semantic similarity weight of node i and node j =

1 the four digits of s1 and s2 are the same
0.7 only the left three digits of s1 and s2 are the same
0.4 only the left two digits of s1 and s2 are the same
0.1 only the left one digits of s1 and s2 are the same
0 all different digits

s1: the semantic code of node j

s2: the adjacent semantic codes in the semantic connection table related to the semantic code of node i.

3.5.5 The reduction of candidates using radical similarity

There is still one more problem need to be tackled. As the number of candidates is increasing due to the bit and character mask, the performance of the proposed system will become worse and worse. Thus, it is very important to get rid of the impossible candidates, especially the homophones, as possible as we can.

In the proposed system, the radical combination of each character is used to reduce the impossible candidates. First, we obtained the radical combination of each character in the inputted text by referring the Chinese character radical dictionary. And then, the radical combination of each character candidate will also be extracted. By comparing (or intersecting) the radical combinations of the candidates with the radical combination of their corresponding character in the inputted text, we can delete those candidates which don't have any identical radical with the corresponding character. And, we found that most of the impossible candidates can be reduced successfully. For example, the phonetic symbols of Chinese character [辨] is 'bian4', so its homophones are [辮],[便],[辯],[變] and so on. However, the radical combinations of each character are as the following.

辨 ---→ YJILJ
辮 ---→ YJVFJ
便 ---→ OMLK
辯 ---→ YJYRJ
變 ---→ VFOK

Because the character [便] and [變] have no intersection with the radical combination of character [辨], so the proposed system will delete these two candidates easily. Furthermore, we also use the combination of successive one character word candidates to form a two or more character word by referring the reference dictionary and then those one character word candidates can be deleted, whose example will be shown in section 4.

4. Example

In order to further explain the proposed mandarin document revision system, an example will be given in this section

[Chinese text input] "不遵守者以法究辨 "

[Character-to-phonetic conversion result]
"bu4zuen1shou3zhe3i3fa3jiou4bian4"

[Candidate retrieval with phonetic symbols]

bu4 zuen1 shou3 zhe3 i3 fa3 jiou4 bian4

候選詞：

遵 守

候選字：

部	尊	手	緒	以	法	究	辨
不	樽	首	者	椅	髮	就	便
:	:	:	:	:	:	:	:

[Candidates using bit mask]

bu4	zuen1	shou3	zhe3	i3	fa3	jiou4	bian4
		遮	一	閥	九		扁
		螯	依	乏	久		邊
		這	移	罰	糾		蝙
:	:	:	:	:	:	:	:

[Candidates using character mask]

bu4 zuen1shou3 zhe3 i3fa3 jiou4 bian4

記者 椅子 法律 舅舅 不便
椅墊 法源 就是
: : : :

[Candidate using radical mask]

bu4 zuen1 shou3 zhe3 i3 fa3 jiou4 bian4

守 樂 淺 穴 辦

宏 : : 窮 辮

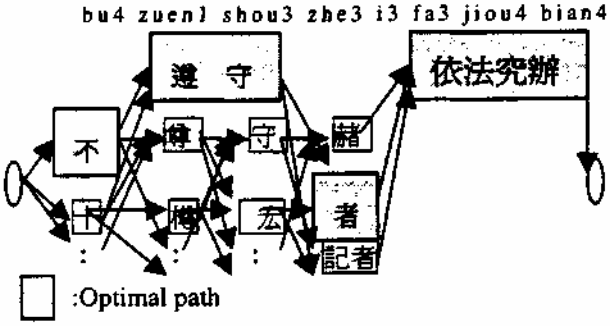
: :

[Word candidate using one character word candidates combination]

bu4 zuen1 shou3 zhe3 i3 fa3 jiou4 bian

依 法 究 辦

[Candidate directional net after reduction]



[Output]

Inputted Text: 不遵守者以法究辨

Optimal string(correction): 不遵守者依法究辦

'Someone who does not obey the law will be prosecuted.'

Marked string(detection): 不遵守者＃以法究＊辨

(＃: ambiguous pronunciation error, ＊: similar shape error)

5. Experimental results and analysis

The 6 articles (some 5,500 characters) selected from the text books of the primary school in Taiwan were used as the test samples. Those articles were inputted all kinds of document error and these errors were also recorded shown as Figure 12. Although the success conversion rate of Chinese character-input-to-phonetic conversion [11] is over 99%, in order to get rid of the influence of this conversion system all the converted phonetic strings are checked manually. As to the reference dictionaries, we used the related reference dictionaries which are used by Chinese phonetic-input-to-character conversion [9] and Chang-Jie radical input rather than special ones. However, in order to evaluate the revision performance we don't consider the character or word shortage factor of the reference dictionaries The system was implemented by C language on PC. The accuracy rate and recall rate shown as Table 1 are some 78.5% and 79.6% respectively, which are better than the conventional system described in section 2. Meanwhile, the average execution time was less 1 character/second.

Table 1 The experimental result

	Correct chars	Total chars	System Correct Chars.	System Wrong Chars.	Precision Rate	Recall Rate
File 1	640	652	516	136	79.1%	80.6%
File 2	960	970	744	226	76.7%	77.5%
File 3	635	644	513	131	79.7%	80.7%
File 4	1175	1186	881	305	74.3%	74.9%
File 5	265	276	230	46	83.4%	86.8%
File 6	1,310	1,322	1,082	240	81.9%	82.6%
Total	4,985	5,050	3,966	1,084	78.5%	79.6%

The analysis of errors will be described below.

(1) The unknown word or those errors related proper names, e.g. company or place name, can not be detected effectively due to lack of morphological and statistical knowledge such as morphological rules or character tri-gram.

(2) Though those redundant character or wrong character order errors in a inputted text can be detected, the redundant characters can not be modified effectively. For example, the revision result to the wrong character order error as [國 經 際 濟] will become [國際經濟*國際經濟*] rather than [國際經濟]'International Economics'.

(3) Most of the detection or modification errors to the similar shape and literal errors are caused by the bad quality of the related reference dictionaries.

(4) The semantic or pragmatic errors in a sentence can not be detected effectively due to lack of related knowledge, such as context, discourse and so on. For example, The wrong time adverb [明天]'tomorrow' in the sentence [他明天死了]'He died tomorrow.' can not be detected.

6. Concluding remarks

The above proposed mandarin document revision system are applied to our Japanese-Chinese machine translation system [4] as the post-processing editors. The performance of post-processing can be improved. For example, the editing time per a operator can be reduced 50%. Moreover, those reference dictionaries in the proposed system are the same with the Chinese input front-end-processors, so the development time and

maintenance cost will also be saved largely. In order to further improve the performance of the proposed system , there are still some future works.

(1) Develop effective algorithms to process the unknown words or character error in proper names.

(2) In order to reduce the word candidates retrieved by bit or character mask patterns, we will continue to survey what the necessary character mask patterns are.

(3) Develop the effective algorithms to solve the redundant or wrong character order errors problem.

(4) Develop the objective evaluation criteria and use more mandarin text to evaluate the system.

(5) Improve the quality of the reference dictionaries.

7. References

[1] 池原 悟ら. (1983). "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

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

[3] C.H. Leung and W.K. Kan. (1996). "Difficulties in Chinese Typing Error Detection and Ways to the solution". In Journal of Computer Processing of Oriental Language, Vol. 10, No.1, pp97~113

[4] J.J. Kuo, J.K. Wu and W.L. Yang. (1991). "The generation of Chinese Text in Japanese-Chinese Machine Translation System". In proceedings of Natural Language Processing Pacific Rim Symposium, PP160~169

[5] J.J. Kuo, J.H. Jou, M.H. Hsieh and F. Maehara. (1986). "The Development of New Chinese Input Method --- Chinese Word-String Input system". In Proceedings of International Computer Symposium, PP1470~1479

[6] S.I. Chen, C.T. Chang, J.J. Kuo and M.S. Hsieh. (1987). "The Continuous Conversion Algorithm of Chinese Character's Phonetic Symbols to Chinese Character ". In Proceedings of National Computer Symposium, pp437~442

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

[8] S. Sporat. (1992). " An Application of Statistical Optimization with Dynamic Programming to Phonetic-input-to-character Conversion for Chinese". In Proceedings of ROCLING III, PP380~390

[9] J.J. Kuo. (1996). "Phonetic-input-to-character Conversion system for Chinese Using Syntactic Connection Table and Semantic distance". In Journal of Computer Processing of Oriental Language. Vol.10, No.2, pp195~210

[10] 類語国語辞典, 角川書店（日本）, 1986

[11] Y.J. Lin, M.S. Yu, S.Y. Hwang and M.J. Ker. (1998). "A Way to Extract Unknown Words Without Dictionary from Chinese Corpus and Its Application". In Proceedings of ROCLING XI, PP217~226

[12] 現代漢語頻率辭典，北京語言學院出版社，1985