Post-processing for Handwritten Chinese Character Recognition

by

Xu Rui-feng

A Thesis Submitted to
the Hong Kong Polytechnic University
in Fulfillment of the Requirements
for the Degree of
Master of Philosophy

Department of Computing
The Hong Kong Polytechnic University

January 13, 2001

Abstract

Some post-processing techniques for improving the performance of Handwritten Chinese Character Recognition (HCCR) system by selecting the most promising candidate characters are presented here.

Aiming to remove mis-recognized and unrecognized characters in the recognition result, three post-processing approaches, namely the one based on contextual linguistics information, the one based on confusing character characteristics produced by a recognizer, and the one based on a hybrid approach, are studied in this thesis and their performance are evaluated and compared.

In the study of the post-processing approach based on contextual linguistics information, the dictionary-based post-processing method is presented. The dictionary-based techniques, including sentence fragments detection and contextual approximate word matching for removing erroneous characters, are studied and its performance is evaluated.

Post-processing Techniques based on statistical language models are then proposed. A Chinese word BI-gram model is established and employed in HCCR post-processing to

identify a most linguistic-promising sentence with the maximum word co-occurrence production by selecting plausible candidate characters. To obtain the description capacity of long-distance restrictions among Chinese sentences, the word BI-Gram model is extended to a distant word BI-Gram model with a maximum distance 3 and prior to post-processing. Their upgrading performances are evaluated and compared.

To recover the unrecognized characters and enhance the theoretical upper improvement limit for the post-processing approach based on contextual linguistics information, the post-processing techniques based on the characteristics of confusing characters produced by recognizer are studied. Analyzing the recognition results for the training samples, the confusing characters for each character category are collected and constructed into a confusing character set. Based on this set, a statistical Noisy-Channel model is used to identify the most promising input character when a candidate sequence is given. This method proves to be effective in removing unrecognized characters. Considering the confusing characters as observed features of character categories, the classification algorithm based on neural networks can be employed to identify the most plausible input as the production of the candidate sequence. All together 3755 character categories in GB2312-80 character-set are clustered into several hundred groups after searching through the transitive closure of the similarity matrix associated with the confusing character set. A group of neural networks for these category groups are established and trained to produce a candidate to match the input character and to adjust the confidence parameter of candidates for a given candidate sequence. A better performance in comparing with the

one based on Noisy-Channel model is achieved.

A three-stage hybrid post-processing system is then built. The post-processing technique based on confusing character characteristics of a recognizer is firstly conducted to append similar-shaped characters into the candidate set. Then the dictionary-based method is employed to append linguistic-prone characters and bind the candidate characters into a word-lattice. Finally the statistical language model is applied to identify a most promising sentence by selecting plausible words from the word-lattice.

On the average, this hybrid post-processing system achieves 6.2% recognition rate improvement for the first candidate when the character recognition rate is 90% for the first candidate and 95% for the top-10 candidates by online HCCR engine. For the offline HCCR engine with the original recognition rate of 81% and 92% for the first and the top-10 candidates, 12% recognition rate improvement for the first candidate is achieved.

Acknowledgements

The completion of this thesis would not have been possible without the help of many people whom I would like to express my heartfelt appreciation.

Firstly, I would like to express my deepest thanks to my supervisor, Prof. Daniel So Yeung. I gratefully acknowledge him who guides my heart to a free and peaceful world that will be of great benefit to all my life. Also I would like to thank him for his kind supervision and continuous support during my M. Phil study at the Hong Kong Polytechnic University. I learned a lot from him in both the technology involved in my M. Phil project and general research methodology.

Another great excellent person whom I would like to express my deep gratitude is Prof. Wenhao Shu, the supervisor of my Master of Science study at Harbin Institute of Technology despite of the fact that I had no chance to finish it. He introduced me into the subject of Chinese character recognition and Chinese information processing.

I would like to express my thanks to Prof. Xiaolong Wang, Prof. Xizhao Wang, and Prof. Eric Tseng who shared with me not only their expertise in computational linguistics and soft computing, but also the enjoyment in life.

Also I would like to express my thanks to Dr. Jiafeng Liu, Dr. Daming Shi, Dr. Fusi Wang, Mr. Yang Lu and Mr. Jianhua Huang, my buddies in Harbin Institute of Technology, who generously made available to me the recognition engine for Online and Offline Handwritten Chinese Characters. Without their help, my research work would surely have suffered.

I would like to thank many of my friends in Hong Kong Polytechnic University, Prof. Dapeng Zhang, Dr. Jiannong Cao, Dr. Jia You, Dr. Jicheng Duan, Dr. Gaoxi Xiao, Mr. Renguo Xiao, Mr. Xiaoqin Zeng, Dr. Yan Wu and Ms. Yue Shu, who constantly encouraged me during the past two years.

In addition to those involved directly with my research work, I would like to thank many of my friends, Ms. Qiu Liu, Ms. Xiaoyu Wang, and Dr. Kang Zhang, who share the pleasure of life with me.

At last, but not the least, I wish to express my deepest appreciation to my parents, my aunt and uncle, my elder brother and his wife, and Ms. Shuqi Jiang, for their endless love and unwavering support. This thesis is dedicated to them.

Contents

List of Tables

List of Figures

Chapter 1

Overview

The purpose of this research is to study the post-processing techniques for improving the recognition rate of Handwritten Chinese Character Recognition (HCCR) system. In this chapter, we will explain the motivation of this research and give an overview of this thesis.

Section 1.1 explains the motivation of using post-processing techniques to enhance the recognition rate of HCCR system. The main problems for HCCR post-processing research are presented in section 1.2 and section 1.3 briefly reviews the reported HCCR post-processing techniques. Section 1.4 gives the objectives of this research and the organization of this thesis is outlined in section 1.5. The contributions of this research are summarized in section 1.6. Finally the publication output of this thesis are listed in section 1.7.

1.1 Motivation

HCCR is undoubtedly a natural input methodology, and it has been an active research area in pattern recognition since 1980's [29, 70, 90, 97]. HCCR is normally considered as a difficult task due to large Chinese character-set, complex structures, many characters of high degree similarity, and wide varieties of writing styles [13, 32]. Depending on the nature of the input characters, HCCR research can be categorized either as Online, such that the input data are expressed in terms of pen action collected from the tablet [12, 39], and Offline which uses the scanned bitmap image as the data source [31, 119].

Generally, the proposed algorithms for recognizing handwritten Chinese characters can be classified into three major categories: statistical method [81], structural method [1, 77], and neural-fuzzy method [120]. With the rapid advancement on efficient feature extraction and pattern classification algorithm, many HCCR systems achieve satisfactory recognition rates. A HCCR engine outputs an ordered set of candidates according to their degrees of similarity to the input sample. In most cases the first candidate is the right choice, but there are still cases in which the correct character is not ranked first position and even sometimes does not appear in the candidate sequence at all. Some reported Online HCCR engines achieve a first candidate accuracy rate of 88-90%, and it increases to 95-96% for the first ten candidates [32, 54]. For Offline HCCR systems, the average recognition rate of 80% for the first candidate and 90% for the top-10 candidates is achieved [82, 96, 105]. Further improvement on the recognition rate is desirable if it is of any practical application. Besides the further improvement on the recognition engine, post-processing techniques which select the most promising

candidates after the recognition of each individual sample for a meaningful sentence, is considered a feasible way. It is motivated by the result of Cognitive Psychology research that the high recognition performance achieved by a human being does not only depend on his ability to recognize each individual character, but also depends on the use of linguistic knowledge and contextual information [3, 40, 86, 92, 106]. Utilizing both the structural information of each individual character and the linguistic knowledge among the sentences in post-processing system, the overall recognition rate of HCCR system can be improved.

In short, our post-processing research aims to improve the overall recognition rate of a HCCR system by identifying the most plausible characters from the candidate set when the recognition engine isn't enhanced.

1.2 Problem Statement

One may observe that there are two kinds of frequently encountered errors in the recognition result, namely the mis-recognized characters (the input character appears in the output candidate sequence but is not ranked first) and the unrecognized characters (the input character dose not appear in the output candidate sequence). The objective of a post-processing technique is to remove these two kinds of errors.

Post-processing is not an all-new topic. Previous works on this topic mainly focus on post-processing for English character recognition. There are a number of attempts to reduce the recognition errors by employing methods either based on compound decision theory [19, 78] or based on letter posteriori probability [67, 74]. Then methods based

on dictionary lookup are designed as the first step to make use of linguistic knowledge in post-processing [17, 99]. The contextual binary N-Gram is also developed to eliminate the least probable words in the candidate set without dictionary lookup [75]. By further taking the factor of confusion between letters into consideration, the post-processing methods based on Viterbi Algorithm is developed [34]. In the recent decades, more computational linguistic knowledge is employed in post-processing research such as word matching, part-of-speech tagging, and syntax analysis [21, 85]. Integrating posteriori, N-Gram and linguistic knowledge together is regarded as a promising way [28].

The development of post-processing methods for Chinese character recognition has gained some progress in the past twenty years. The algorithms employing confusion characters, language model, or N-gram are reported [11, 41, 113, 122]. Attribute to some characteristics of Chinese that differ from the one of English, developing an effective post-processing algorithm for Chinese character recognition is more difficult [45]. Some of the problems to be solved are listed here.

- English is an alphabet based language that consists of only 26 letters, while Chinese, a ideographic language, has a much larger character-set consisting of thousands characters. The hundred-times larger character-set not only leads to a lower recognition rate of individual characters, but also makes the post-processing methods based on English letter posteriori not applicable to Chinese directly. Finding out an effective model for describing the confusing property between Chinese character categories is essential.

- Chinese sentences, unlike those for English or other alphabet-based languages, are written in a continuous string of characters without obvious separators indicating the word boundaries. Prior to further linguistic analysis, an effective word identification and word segmentation algorithm is indispensable.
- The contextual linguistic information is described as a language model. To employ linguistic knowledge in HCCR post-processing, establishing an effective language model, which can be used to exactly describe the Chinese linguistic property with reasonable parameter space and computational cost, is required. Also the research on computational linguistic model is the basic and most important issue of natural language processing research.
- To make full use of two kinds of available information for post-processing, namely, contextual linguistic knowledge and confusing character characteristics of the recognizer, need an effective integration method for a hybrid post-processing system.

1.3 Introduction to Post-processing for Handwritten Chinese Character Recognition

Towards the goal of removing the erroneous characters in the recognition outputs, two kinds of information can be utilized, namely, confusing character characteristics for each individual character category, and contextual linguistic information. Based on the information employed, HCCR post-processing methods can be classified into three

major categories: the approach based on confusing character characteristics of the recognizer which can be used to remove the errors for each individual character category [3, 41, 122], the approach based on contextual linguistic information which can be used to remove the errors occur in a word or a phrase [40, 92, 106], and the hybrid approach which incorporates these two kinds of information together for recognition error removing [113].

The post-processing approach based on the utilization of contextual linguistic information is widely adopted. In this approach, the computational language model is employed to construct a discriminator to detect and remove the errors, and identify the most linguistic-plausible result. It can be further classified as dictionary-based and statistical-based approach. The dictionary-based techniques, including approximate word matching, contextual word matching, and syntax analysis, are employed to analyze the sentence hypotheses and remove the identified errors [4, 15, 49, 86, 110]. This approach works under a top-down strategy. Unlike the dictionary-based approach, modeling Chinese language as a stochastic Markov process, statistical measures for language such as word frequency, character and word co-occurrence frequency, can be employed in a statistical-based approach to evaluate the sentence hypotheses constructed by the candidate characters and to identify the most linguistic plausible sentence hypothesis as the output [9, 18, 79]. This approach works under a bottom-up strategy in which smaller linguistic units are first constructed and evaluated, and then they are used to form larger and complex ones.

The post-processing approach based on confusing character characteristics produced by the recognizer is reported in some works. The similar-shaped character set, the most-error prone character set (MEPC set), and the confusing character set are popularly adopted to describe the similarity among Chinese characters. The input/output statistical characteristics of the recognizer are analyzed and then used in candidate selection. The posteriori statistical models are employed to evaluate the probability of the candidates and the most promising one is selected as the output result [13, 36, 48]. This approach has its advantage in unrecognized character recovering which is very difficult for the post-processing approach based on computational linguistic information, especially for the popular one based on the statistical language model.

Incorporating these two basic approaches, the hybrid approach is expected to achieve a higher improvement performance by making full use of these two kinds of available information [51, 95, 113].

More details of the proposed HCCR post-processing research will be presented in Chapter 2.

1.4 Objective of Thesis

The basic objective of this thesis is to construct a hybrid post-processing system for both Online and Offline HCCR systems for improving the overall performance of the recognition system significantly. Detail tasks to achieve this objective may be summarized as follows:

- Develop a large scale Chinese corpus for building a dictionary, collecting statistics information, analyzing language features, and conducting automatic word segmentation test.
- Design and build a dictionary containing certain statistics on linguistic information such as word frequency, word co-occurrence frequency. Most of the statistical information of each word entry in the dictionary will be obtained from the corpus.
- Design and implement an automatic Chinese word segmentation algorithm for Chinese sentence, the approximate word matching and word segmentation is specially considered.
- Design and implement a dictionary-based post-processing method. The approximate word segmentation and matching are employed to remove the detected errors in the sentence hypothesis.
- Design and implement the post-processing methods based on different statistical language models. Word BI-gram model and long-distance word BI-Gram model will be considered.
- Analyze the results of large-scale recognition experiment for both online and offline HCCR system and construct a confusing character set for the descriptions of confusing properties among Chinese character categories.
- Establish an effective error-recognition model to describe the relevant characteristics and distribution of erroneous characters for the purpose of

recovering unrecognized and mis-recognized characters in the candidate sequences.

- Establish a hybrid post-processing system for HCCR system integrating dictionary-based method, statistical language model and error-recognition model together.
- Investigate the appropriate integration strategy when applying hybrid post-processing techniques to online and off-line HCCR systems.
- Demonstrate the integration of our post-processing system with an online and an offline HCCR system. Reasonable improvement of the first candidate accuracy rate, depending on the corpus characteristics and the original performance of the HCCR system, could be expected.

The outline of this thesis is presented in the following section. It is structured according to the process that leads to achievement of the stated objectives.

1.5 Outline of Thesis

This thesis contains 7 chapters. In Chapter 2, existing HCCR post-processing methods are reviewed. According to the source of information used for removing recognition errors, these methods are categorized into three major approaches. In the review of post-processing approach based on computational linguistic information, after the introduction and discussion of the computational linguistic units that can be employed in Chinese post-processing, the dictionary-based techniques are firstly

presented. The popular statistical-based post-processing techniques are then reviewed and special attention is paid on the employed statistical model and considered linguistic units. Thirdly, the post-processing techniques based on hybrid language model are reviewed. As another major approach, the post-processing techniques based on confusing character characteristics of recognizer are next reviewed. This is followed by a brief review of hybrid post-processing approach.

Some fundamental works on Chinese language modeling and HCCR post-processing are presented in Chapter 3. The development of a large-scale Chinese corpus for building dictionary and retrieving linguistic statistic is presented. Then the establishment of the dictionary recording the word entries and statistical linguistic information is proposed. To evaluate the performance of our post-processing methods, two HCCR engines, one for online and another for offline, are adopted in this research. In the second part of Chapter 3, a brief introduction of these two recognizers, the interface between the recognizer and the post-processing system, and the construction of the testing sample database are given.

Chapter 4 concentrates on our proposed post-processing techniques based on computational linguistic information. As a fundamental component of employing computational linguistic information in language processing, a word segmentation algorithm based on BI-directional Maximum Matching and word BI-Gram model is established. Then, The dictionary-based technique, based on sentence fragment detection and approximate word matching, is studied and its performance is evaluated. Next, the works on the statistical-based techniques is presented. The post-processing

method based on Chinese word BI-Gram model is studied, and then the works based on Distant Chinese word BI-Gram model which is a extension of word BI-Gram model, is presented.

In view of the fact that the post-processing approach based on computational linguistic information sometimes suffers from the problem of unrecognized character, the post-processing approach based on confusing character characteristics is studied in Chapter 5. This approach is designed to recover the unrecognized and mis-recognized characters so as to improve the recognition rate and enhance the theoretical upper improvement limit for the post-processing approach based on computational language model. In this chapter, the method based on linear statistical Noisy-Channel model and the method based on non-linear neural network groups for similar character categories are respectively presented.

The research of hybrid post-processing system that integrates these proposed methods together to improve recognition rate of HCCR system is presented in Chapter 6. The different integration methods of the hybrid post-processing approach for online and offline HCCR system are discussed and their performance improvements are evaluated.

Finally, Chapter 7 concludes our research and directions for further investigations are recommended.

1.6 List of Contributions

The contributions of this thesis are listed as follows:

- A large-scale Chinese text corpus is developed and established. Its memory size is more than 2 GB, with approximately 890 millions of Chinese characters. About 40MB text are selected and manually segmented to collect exact linguistic statistic data and about 500MB data in the text corpus is segmented by our automatic word segmentation program to form a segmentation corpus. These obtained corpora are very important not only for the works of this thesis but also for the future research works related to Chinese language processing.
- A dictionary consists of 101,641 word entries is built. The word frequency information that obtained from the corpus is recorded. Furthermore, the statistics of Chinese character BI-Gram, word BI-Gram and distant word BI-Gram are collected from the corpus and recorded.
- A Chinese word segmentation algorithm is established. The BI-directional Maximum Matching is applied to segmenting a sentence into word sequence. The word BI-Gram model is employed to remove the detected ambiguities during maximum matching.
- A dictionary-based post-processing method is developed. The sentence hypothesis is segmented into word sequence employing an exact word segmentation algorithm, and the located sentence fragments are regarded as potential errors. The dictionary-based approximate word matching is applied to these fragments to remove erroneous characters. This method is proven effective when the original recognition rate is good. In particular we may observe this method has the capacity of recovering some of the unrecognized characters.

- A statistical post-processing method based on Chinese word BI-Gram model is designed. The Viterbi algorithm is employed to identify a sentence hypothesis with the maximum word co-occurrence frequency constructed by the characters in the candidate set. A higher recognition rate improvement compared with the one by employing popular character BI-Gram model and word UNI-Gram model is achieved.
- Extending the word BI-Gram model to long-distance word BI-Gram model by considering the distance parameter, are expected to acquire the description capability of long-distance restriction among Chinese sentence with an acceptable parameter space. The statistical-based post-processing method based on this language model is able to achieve further improvement in comparison with the one based on word BI-gram model.
- In the research of post-processing approach based on the characteristics of the confusing characters, the 3755 character categories in GB2312 character-set are clustered into several hundred groups through searching the transitive closure of the similarity matrix associated with their confusing character set. A group of neural networks for these categories groups are built and trained. Regarding confusing characters as the observed feature of character categories, a classification algorithm based on neural networks is employed to identify the most promising candidate. A good result is obtained.
- A three-stage hybrid post-processing system based on the proposed methods is established. The method based on confusing character characteristics is first

conducted to recover erroneous characters, especially the unrecognized characters. Then the dictionary-based method is used to recover some unrecognized characters by appending linguistic-prone characters to the candidate sequences, and to bind the characters into a word-lattice. Finally, the statistical-based stage is processed to identify a most promising sentence constructed by the words in the word-lattice as the final result.

Benefited from making full use of two kinds of available information, our hybrid post-processing system has achieved a higher performance improvement. For the online HCCR engine with the original recognition rate of 90% for the first candidate and 95% for the top-10 candidates, a 6.2% recognition rate improvement for the first candidate is achieved. As for the offline HCCR engine, our post-processing achieves 12% recognition rate improvement for the first candidate when the original recognition rate for the first and top-10 candidates are 81% and 92% respectively.

The application of our proposed language model is not limited to HCCR post-processing, many other domains such as speech recognition [25, 108], Pinyin-Hanzi Transcription [30, 88], and spelling correction [38] can be benefited by incorporating this model. Furthermore, the idea of hybrid post-processing system could be used to improve the recognition performance of some other oriental ideographic characters such as Japanese and Korean [44, 65, 66] [Nagata 1998] [Lee+ 1996].

Chapter 2

Review of Post-processing for Handwritten Chinese Character Recognition

The related works on HCCR post-processing are reviewed here to provide the necessary background understanding for the work in this thesis. In Chapter 1, existing post-processing methods are categorized into three approaches, first one based on computational linguistic information, the second one based on characteristics of the confusing characters produced by the recognizer and the third one being a hybrid approach. In the following three sections, existing works based on each of these three approaches are reviewed and their characteristics are discussed.

2.1 The Post-processing Approach based on Contextual Linguistic Information

The popular post-processing approach based on contextual linguistic information is inspired by such a phenomenon: some crabbed or blurry handwritten samples are so similar that even a human being cannot distinguish these individual character images exactly, but if these character images are put in the context of a meaningful sentence, they could be easily identified. The principle behind this phenomenon is that both the recognizable features of individual characters and the contextual linguistic information are utilized during the human recognition process. The works on Cognitive Psychology, a study on the human mind and how the mind functions with regard to knowledge acquisition, indicate that the character recognition and sentence understanding of a human being are based on a three-stage word recognition process. Firstly, a 'Character processing' works to identify words from the text and then proceed to recognize individual characters. Then a 'Word Access' processing leads to activate the semantic information, orthographic and other information related to the 'recognized word' in Mental Lexicon [62]. They are then used to acquire the exact meaning and context information for sentence understanding in the 'Word Processing' stage, and during this stage, the errors found in the words can be removed [63]. With the reference of human recognition, the computational linguistic information is employed in post-processing by construct a discriminator to detect and remove the erroneous characters, and identify the most linguistic-plausible result.

To develop a post-processing method based on computational linguistic information, we must handle a basic issue: how to represent, acquire, and apply the linguistic knowledge. According to the representation form of linguistic knowledge and working

strategy in error removing, the post-processing methods based on this approach can be further classified as dictionary-based (or rule-based) approach, statistical-based approach, and the approach based on hybrid language model. Before the introduction of these approaches, the Chinese computational linguistic units that will be employed in language processing are discussed.

2.1.1 The Chinese Computational Linguistic Units

Character

Generally, a Chinese sentence is composed of a number of ordered words, while a word is composed of a number of ordered characters. Character is the minimum language unit. In Aho and Ullman's work [2], a mathematical definition of character is given:

Definition 2.1: An *alphabet* Σ is a finite set of symbols. Each symbol in the alphabet is a character. The alphabet size denoted by $|\Sigma|$ is the number of characters in the alphabet.

In GB2312-80 character-set [27], there are totally 6763 Chinese characters. In order to distinguish from the characters written on paper, in the following parts of this thesis, character category is used to denote the character in the character-set. Character is the minimum language unit that can be employed in Chinese language processing. Considering the character category in the Chinese character-set has a finite number of 6763, which is obviously less than the number of Chinese words, character-based language models are simpler and smaller compared with the word-based one. However,

character-based language model cannot be employed to describe the semantic restrictions among Chinese sentences attribute to the fact that character is not a complete semantic language unit and the meaning of a character must be determined by putting this character into a word.

Word

The definition of word is given as follows [2]:

***Definition 2.2*:** Let alphabet Σ be a finite set of characters. A *dictionary* D is a set of character strings over alphabet Σ. For any element $w \in D$, w is a *word*. The dictionary size denoted by $|D|$ is the number of words in the dictionary.

Attribute to the fact that Chinese sentence is written in a continuous character string without blanks to indicate the boundaries of words, the characters in the sentence are combined into a sequence of words with semantic meaning. However, Chinese language does not have a set of pre-defined rules to bind characters into a word for semantic meaning. They are mainly based on customary use. Therefore, there is no well-defined dictionary available to record all the words in modern Chinese.

However a descriptive definition of 'word' is widely accepted by most linguistic researchers [57, 102].

***Definition 2.3*:** Word is the minimum complete semantic unit that can be put together to form a sentence.

As a result, there are varieties of dictionaries with a wide range of vocabulary sizes from 39,000 to about 104,000 [26, 56, 71, 106, 107, 109, 111]. Some of them are listed in **Table 2.1**.

	Vocabulary Size	Application
National Standard of Frequently Used Word List for Information Processing [26]	39, 016	All language application
WORDDATA from Institute Information Science, Academia Sinica in Taiwan [107]	78, 410	All language application
Lexicon HKU97 [106]	85, 855	Post-processing
Corpus for Modern Chinese Research [109]	47, 006	Corpus-based language analysis
Word Base [71]	about 50, 000	Document Compression and Automatic Correction
PolyU Lexicon [111]	104, 251	Word Segmentation and Post-processing

Table 2.1 Vocabulary Sizes of Some Language Processing Systems

Due to the fact that word is defined as the minimum complete semantic unit, the word-based language model is considered suitable for describing the semantic restrictions among Chinese sentences.

Word-Class

Word-class is not a natural linguistic unit for Chinese languages, but it is rather a computational language unit. A word-class-based language model, in which words with

similar linguistic properties or other specific properties are clustered into smaller number of word-classes, is expected to significantly reduce the parameter space it requires [87].

Since there is no widely accepted definition of word-class, a variety of word clustering strategies has emerged. Lee and Tung [42] present a character-based word-clustering algorithm by clustering the semantic-similar words with the same prefix character into a word class, and 800 word-classes are obtained. In Golden Mandarin III, a mandarin dictation machine system, Lyu [60] presents an algorithm to cluster the words with same part-of-speech property and similar word co-occurrence statistical behavior into a word class. This system has about 2,000 word classes corresponding to nearly 60,000 words. Furthermore, the word classification strategies based on the homogeneity of syntactic and semantic property are also reported [5, 106].

Some reported word clustering strategies and the corresponding number of grouped word class are shown in **Table 2.2**.

In spite of the significant reduction of the parameter space and the problem of data sparseness by replacing a word-based language model to a word-class based language model, the linguistic decryption capacity of a word-class based language model is lower than the conventional word-based one. A major reason is that a large number of words attribute to more than one word-class will unavoidably weaken the performance of a word-class based system.

	Word Clustering Strategy	Number of Grouped Word Class
Chang C. H et al. [5]	Clustering words by Simulated Annealing.	50-1,000
Chang J. S. [7]	Clustering words into unambiguous class, two-way ambiguous class and singleton word class, etc.	4, 238
Chang Y.C et al. [8]	Clustering words with similar statistics of frequency and co-occurrence frequency into a word class	500
Lee and Tung [42]	Clustering the semantic-similar words with the same prefix character into a word class	800
Lyu et al. [60]	Clustering the words with same part-of-speech property and similar word co-occurrence statistical behavior into a word class.	2, 000
Wong P. K. et al. [106]	Clustering words with similar syntactic/semantic into a word class	470

Table 2.2 Some Word-class Clustering Strategy and Corresponding Number of Word-Class

2.1.2 The Dictionary-based Post-processing Approach

In the dictionary-based post-processing approach, the approximate word matching and semantic analysis are utilized to correct the erroneous characters identified in the sentence hypothesis by selecting the first candidate characters in the recognition result to construct a meaningful sentence. This sentence hypothesis is segmented into word sequences and semantically analyzed to locate the ambiguous character strings, called sentence fragments. Then the contextual word matching and linguistic analysis are

utilized to evaluate the similarities between the word entries in the dictionary and the word hypothesis generated by substituting the similar or promising characters for one character in the ambiguous string. The word entry with the largest similarity value is then selected as the corrected result to substitute the ambiguous string [38, 86]. If the ambiguities are still there after approximate word matching process, the syntactic analysis, such as part-of-speech tagging and unification grammars, is conducted to remove these ambiguities [15, 49]. In this way, the erroneous characters can be detected and corrected. This approach adopts a top-down strategy that the errors in smaller language units are corrected by means of the analysis of larger language units [7].

Wong P. K. and Chan C. [105] employed dictionary-based post-processing approach as the baseline language model to improve an off-line hand-written Chinese character recognizer. The algorithm based on maximum word matching and word binding force is utilized to match the bounded words with the entries in the dictionary. The most promising word is identified and used to replace the original recognition result. A 6.8% recognition rate improvement is reported.

One of the advantages of dictionary-based approach is that existing linguistic knowledge can be incorporated into the system directly with small parameter space and low computational cost. Another one is that some of the unrecognized characters can be recovered by contextual approximate word matching, which is very important to enhance the performance of post-processing systems.

However, the correction capability of a dictionary-based post-processing approach suffers from several problems. The first problem is the difficulty to scale up a complete linguistic rule set to resolve the error of ambiguities and ill form. Secondly, the out-of-dictionary problem, that occurs when the matched words in the sentence are not recorded in the dictionary, is difficult to solve. If an out-of-dictionary problem occurs, the dictionary-based technique can do nothing. A Solution is to build a complete Chinese dictionary that records all Chinese words. However, it is infeasible to build such a dictionary because a large number of proper nouns, domain-dependent words, and new words cannot be all recorded. Thirdly, dictionary-based post-processing approach is mainly designed for removing the recognition errors in a multi-character word. The existence of large number of single-character-words in Chinese will tend to weaken the dictionary-based approximate matching. Finally, for a given recognition result with low accuracy, the dictionary-based approach cannot offer a satisfactory result because too many erroneous characters will undermined the word matching process since they fail the strict linguistic rules.

2.1.3 The Statistical-based Post-processing Approach

The statistical-based post-processing approach adopts a bottom-up strategy to that smaller linguistic units are firstly constructed and evaluated, and they are used to form and evaluate the larger and complex units. Finally the sentence hypothesis with the maximum linguistic probability is identified as the output result.

Let characters C_1, C_2, ..., C_n form a sentence, denoted by S, and for each $j=1,...,n$, i_j is the image of C_j. Suppose for each input image i_j, the recognition engine outputs m candidate characters ($C_{jk} : k = 1,...,m$).

A sentence hypothesis $S' = c_{1*}, c_{2*}, ..., c_{n*}$ is constructed by selecting one c_{j*} from $c_{j1}, ..., c_{jm}$, $j=1,...,n$. There are a total of m^n sentence hypotheses. The objective of the statistical-based post-processing module is to find out a sentence hypothesis, $\hat{S}$, that has the maximum linguistic likelihood among all sentence hypotheses constructed by the candidate character for the image of sentence I, i. e.,

$$P(\hat{S} \mid I) = \arg\max_{S'} P(S' \mid I). \tag{2.1}$$

There exist many statistical language models which can be used to evaluate the probabilities of sentence hypotheses and identify the most promising one, $\hat{S}$. They can be classified into five major categories: context-independent model (UNI-Gram model), N-Gram model, long-distance N-Gram model, word-class-based N-Gram model, and Part-of-speech N-Gram model (N-POS-Gram model).

Context-independent Model: UNI-Gram Model

Supposing that the words in the context are independent, the language can be described as a context-independent model. Since it is a special case (N=1) of the N-Gram models (The N-gram model will be discussed in the following section), it is always named UNI-Gram model.

The statistics measure used in UNI-Gram is the probability of linguistic units. It is defined as the ratio of a specified linguistic unit's occurrence times over the total number of linguistic units in the trained text. A typical method to construct a context independent model is to estimate the probability of each word according to its frequency. There are two UNI-Gram language models actually applied to Chinese language processing, namely, Character UNI-Gram model and Word UNI-Gram model.

Character UNI-Gram model is seldom used in Chinese language processing alone. Normally, it is part of in a more complex language model [61, 94].

Word UNI-Gram model is more powerful in comparison with the Character UNI-Gram model. Many reported statistical language processing methods are based on word UNI-Gram model and word frequency is used as the basic statistical measure [55]. Liu [101] presents a method to evaluate the probabilities of different segmentation forms for a sentence as a product of word frequencies. The one with the maximum value of word frequency product is regarded as the most promising one. In Chou's post-processing system based on word UNI-Gram model, word frequency is used to identify the most plausible characters that form a most promising sentence with maximum word frequency combinations [16].

The advantage of the UNI-Gram model is that it requires small parameter space and few training data. As for its disadvantage, it is well known that the relationship between characters or words is ubiquitous in Chinese sentences, and the UNI-Gram model cannot represent such a relationship for it is a context-free language model. Generally

speaking, integrating the UNI-Gram model into a contextual language model will perform better [3].

N-Gram Model

The most widely used statistical language model is the N-gram model. It works under the Markov assumption that the occurrence probability of a language unit depends only on its previous *N-1* language units. In another word, this model uses the previous *N-1* language units as the dependent context of the current one [92, 116]. Character N-gram model and word N-Gram model are the major ones. In what follows we use a word N-Gram to demonstrate how a N-Gram model works.

Supposing a sentence or a word sequence S is composed of the words $w_1, w_2, \cdots, w_k$. The word sequence probability $p(s)$ can be calculated as

$$p(s) = p(w_1, \cdots, w_k) = p(w_1)p(w_2 / w_1) \cdots p(w_k / w_1, \cdots, w_{k-1}) = \prod_{i=1}^{k} p(w_i / w_1, \cdots, w_{i-1}), \quad (2.2)$$

where $p(w_1) = p(w_1 / w_0)$ by convention.

The N-Gram model is to simplify the complication by making the approximation that the i^{th} word of S only depends on the preceding *N-1* words. Therefore,

$$p(s) = \prod_{i=1}^{k} p(w_i / w_1, \cdots, w_{i-1}) \approx \prod_{i=1}^{k} p(w_i / w_{i-n}, \cdots, w_{i-1}). \quad (2.3)$$

Since the number of parameters in N-Gram model grows exponentially with a large value of N, but which limited parameter space and training data, it is infeasible to directly estimate and store the contextual probability for arbitrary N for a N-Gram model. Normally only the cases of $N=2$ and $N=3$ are practically considered, and the N-

Gram model with the value of N being 2 or 3 are usually named BI-Gram model and TRI-Gram model respectively.

BI-Gram model is a widely adopted language model. By using a BI-Gram model, **Equation 2.3** is simplified to,

$$p(s) = \prod_{i=1}^{k} p(w_i / w_1, \cdots, w_{i-1}) \approx \prod_{i=1}^{k} p(w_i / w_{i-1}). \tag{2.4}$$

The basic statistic measure for the BI-Gram model is the co-occurrence frequency of language units. Two computational measures, namely mutual information and t-test value are also frequently used.

Character BI-Gram is a widely used statistical Chinese language model in word segmentation [58, 59, 89, 93], phonemic-to-character conversion [88] and post-processing [48, 100]. The character-based BI-Gram statistical information, which can be derived automatically from the raw corpus, is treated as the main statistical measure in these systems.

In some reported post-processing works, the conventional character BI-Gram model is employed to identify an optimal path with highest character co-occurrence possibility from the candidate set and their performance is good [46, 47]. In order to reduce the storage space, some reported systems based on character BI-Gram do not store all character co-occurrence statistics, but only the statistics of those 'important' characters are recorded. A measure called word-binding force is used to describe the strength of the characters combined to form the word, is introduced by Wong and Chan [104, 106]. It is applied to Chinese word segmentation and post-processing. This measure in fact is based on character mutual information, but only those characters that can form a word

by itself and also can be a part of multi-character word simultaneously are considered. Furthermore, an inter-word-character-based BI-Gram model is proposed by Li [50] and Shyu [83]. In this model, only the co-occurrence frequency of the characters that appear in the first position and the last position of a word, together with the those characters can construct a word by itself, are collected from the corpus and their statistical information are used in the post-processing. Employing this inter-word-character-based BI-Gram model, 60% parameter space can be reduced and a good correction performance is achieved.

Generally speaking, a Chinese character BI-Gram model has a parameter space of at most 6775*6775 combination forms which is acceptable to a PC after filter and compression. The major advantage of a character BI-Gram model is that the parameters of this model can be established from raw corpus directly. It is very important to a practical-oriented statistical language model. Furthermore, character BI-Gram model also shows its advantage in new word detection by analyzing the most frequently occurred character strings, which is an important problem in Chinese language processing [68, 124]. As for its disadvantage, this model can deal with two adjacent characters only, which is not enough for describing the complex relationship of characters in the sentences.

Preliminary experimental results indicate that the word BI-Gram is much more powerful over the character BI-Gram in modeling Chinese language [30, 110]. The advantage of word BI-Gram is due to the fact that word is regarded as the basic semantic unit. That means word BI-Gram model can effectively describes the Chinese

characteristics and support further linguistic analysis. Another advantage is that word BI-Gram model make use of more contextual information compared with character BI-Gram model. As a result, ambiguities occur in the multi-character words can be effectively removed by word BI-Gram model.

In considering the two previous linguistic units as dependent context for the current unit, a TRI-Gram model is expected to yield a better result in comparison with the BI-Gram model. However, its parameter space will increase significantly.

Xia proposes a post-processing method based on character BI-Gram and character TRI-Gram model [9]. In her method, the co-occurrence probabilities between two adjacent characters and among three adjacent ones are collected from the raw Chinese text directly. Regarding the character BI-Gram and TRI-Gram probabilities as the cost of a path, a dynamic programming searching strategy is employed to identify the most plausible sentence from the candidate set. The experimental result is satisfactory.

The number of possible combination of a character TRI-Gram is very huge. Based on GB-2312, a character TRI-Gram has at most 6775*6775*6775 combination forms, that is, hundreds of times to a word BI-Gram model. Therefore, the powerful filter and compression procedure for reducing parameter space is the most important issue when considering character TRI-Gram model.

The N-Gram model with the parameter $N>3$ is normally named higher-order N-Gram model. Theoretically speaking, the N-Gram model will be more refined and a better linguistic description capacity can be obtained along with increasing parameter N. However, for practical system, we must consider that the parameter space for the

language model will grow rapidly when N increases, as well as the increase in errors caused by insufficient training data. Therefore, higher order N-Gram has not been applied to the any Chinese post-processing system. Zhang and Huang [123] make a theoretical analysis and small-scale statistical experiments in order to find the optimal value of N in N-Gram for Chinese language processing. Three factors, namely, approximate expression for Chinese grammatical structure, reconstruction capacity of new words, and the performance for the transcription of Chinese phonemic-to-character conversion, are utilized to evaluate the performance of N-Gram models with different values of N. The optimal N values for these three factors are N=6, N=4, and N=4, respectively. In considering these three factors together, one may conclude that N= *4* is a better value for word-based N-Gram model in Chinese language processing. This conclusion and the three factors are very important to our research work.

Long-Distance N-Gram Model

The conventional N-gram model has a weakness that the required parameter space and training data will greatly increase with an increasing value of N. Therefore, for a practical system, the selected value of N is normally not larger than 3. But that is not enough to describe many long-distance restrictions in the Chinese sentences. Thus, the long-distance N-Gram model is designed to accommodate the long-distance restrictions with slowly increasing parameter space.

Conventional N-Gram model estimates the probability of current linguistic unit according to its previous N-*1* words, while long-distance N-Gram model describes the

dependence of current language unit on its pervious *N-1* distant units [35, 84]. For example, applying a word-based distant BI-Gram model with a distance of 3, the probability of a sentence is given by:

$$P(s) = \prod_{i=1}^{k} P(w_i \mid w_{i-3}, \text{distance} = 3) \cdot P(w_i \mid w_{i-2}, \text{distance} = 2) \cdot P(w_i \mid w_{i-1}, \text{distance} = 1).$$

(2.5)

One may observe that employing the above equation, the required parameter space is only three times over the one for regular BI-Gram model. It is proven an effective method to describe the long-distance restrictions with linearly increased parameter space.

When the considered distance is 1, the long-distance N-Gram is just the conventional N-Gram model. Therefore, the long-distance N-Gram model covers the N-Gram model.

Zeng [121] presents a practical post-processing system based on character long-distance N-Gram model. In that system, a distance-2 character BI-Gram model is employed to estimate the relationship between the candidate character C_i and distant characters C_{i-2}, C_{i+2}, while the conventional BI-directional BI-Gram model is used to estimate $P(C_i \mid C_{i-1})$ and $P(C_i \mid C_{i+1})$. Employing this model, the four adjacent characters are used to determine the plausible candidate character. The experimental results prove this method to be more effective than the conventional BI-Gram in HCCR post-processing.

Furthermore, Yang presents a word-based long-distant BI-Gram model, called N-Window model, to employing in a large vocabulary speech recognition system [115].

The adopted statistical measure is similar to that of Zeng's system but the basic language-processing unit is changed from character to 'frequently used word'. Yang reports that the word-based long-distance BI-Gram model shows a significant advantage in prediction capacity and output recognition accuracy.

Adopting long-distance N-Gram models makes it possible to partly resolve the long-distance restrictions problem, which is a main barrier to acquire a refined description capacity by using N-Gram model. Furthermore, the parameter space for long-distant BI-Gram is obviously less than the one for a TRI-Gram model. Therefore, the long-distance N-Gram is regarded as an effective language model in Chinese language processing.

Word-Class-based N-Gram Model

Word-class-based N-Gram model is another way to acquire the description capacity for long-distance restrictions with reasonable parameter space by clustering similar words into word-class to reduce the number of basic language units [87]. Some word-clustering algorithms are presented in Section 2.1.1.

Substituting word-class for the character or word as the basic linguistic unit, the presented techniques for conventional N-Gram model are easily applied to word-class-based N-Gram model. Take BI-Gram model as an example. Let $c(w)$ denote the class that word w is assigned to, then for a word-class-based BI-Gram:

$$P(s) = \prod_{i=1}^{k} P(c(w_i) \mid c(w_{i-1})) \cdot P(c(w_i) \mid w_i). \tag{2.6}$$

Lee and Tung [40] present an efficient semantically word-clustering algorithm that clustering the words into 800 word-classes. Applying this model to HCCR post-processing, 4% additional recognition accuracy improvement is obtained in comparison with using a character BI-Gram model.

Moreover, Wong and Chan present a word-classification method based on the homogeneity of syntactic and semantic information [106]. 470 word-classes are finally formed, and the BI-Gram statistics between them are collected from corpus. A 10.2% average recognition rate improvement is achieved by employing this model in HCCR post-processing.

In spite of the fact that a word-class-based N-Gram model requires a much smaller parameter space by reducing the number of basic linguistic units, the linguistic description capacity of word-class-based N-Gram model is lower than that of conventional N-Gram model.

Part-of-Speech N-Gram Model (N-Pos-Gram Model)

Similar to the above word-class-based N-Gram model, the widely used Part-of-Speech N-gram model, in short N-POS-Gram model is also designed to reduce the number of basic linguistic unit. In this model, the part-of-speech property of words is treated as the basic linguistic units, and the occurrence probability of a word depends on its part-of-speech and the part-of-speech properties of pervious *N-1* words. For example, considering the POS-BI-Gram model, the probability of a sentence hypothesis is calculated as,

$$P(s) = \prod_{i=1}^{k} P(t(w_i) \mid t(w_{i-1})) \cdot P(t(w_i) \mid w_i), \tag{2.7}$$

where $t(w_i)$ is the part-of-speech property of word w_i, and t is a member of part-of-speech tagging set T.

Chien [14]applies a part-of-speech BI-Gram model to HCCR post-processing. The preliminary experimental results prove that this model can improve the overall output accuracy of a HCCR system. Similar works on employing N-POS-Gram model in Japanese OCR correction [65] and Chinese character recognition [18, 43] are reported.

Due to the fact that a large number of Chinese words have more than one part-of-speech properties, the forecast capacity of this kind of language model is unavoidably weakened. Furthermore, the conditional probability of a word w_i has a part-of-speech t, and $P(t| w_i)$ must be obtained from tagged training corpus. Since it is very difficult to acquire a large-scale tagged corpus, the large-scale parameter training for part-of-speech N-Gram is very difficult. This fact will further affect the performance of N-POS-Gram model.

From the above review of statistical-based post-processing techniques, one may observe that for a statistical language model, if the number of basic linguistic units is small, its forecast accuracy on current words and the capability for correcting the recognition errors would be insignificant, and vice versa.

Generally speaking, the statistical-based post-processing methods are proven effective, and they have demonstrated advantages over a dictionary-based one in several aspects. Firstly, the linguistic knowledge is expressed in terms of the likelihood of

linguistic events, and there is no strict sense of well-formed rules. Secondly, in the statistical-based approach, non-deterministic language behavior can be objectively qualified by objective probabilistic metrics. Therefore, the statistical-based technique can be used to evaluate the probability of several semantic-correct sentence hypotheses and identify the most linguistic likely one. Thirdly, automatic or semi-automatic training of the parameters for statistical-based system is possible via using well-developed optimization techniques. Fourthly, employing the statistical-based post-processing approach, a global optimal result can be obtained that is better than the local optimization result obtained by employing dictionary-based approach. Finally, this approach depends less on particular application domain since the linguistic estimation process is universal for most language application domain. Therefore, this approach is widely adopted in post-processing systems.

The major problem of the statistical-based post-processing approach is the unrecognized character. A basic assumption made by statistical-based post-processing methods is that all input characters must appear in the candidate set in order to ensure a correct sentence to be identified from the candidate set. If it were not the case, the unrecognized character problem will occur, and such errors cannot be corrected by this method alone. Obviously the correction rate of a statistical-based post-processing approach has a theoretical upper limit depending on the recognition accuracy for the first m candidate produced by the recognizer. For example an offline HCCR system with 80% accuracy for the first candidate and 90% accuracy for the first ten candidates [80], 50% of the total errors attribute to unrecognized characters can not be corrected by

employing statistical-based post-processing approach alone. Therefore, incorporating the statistical-based post-processing approach with other approaches that can recover some of the unrecognized characters, a higher correction rate can be expected.

2.1.4 The Post-processing Approach based on Hybrid Language Model

For the post-processing methods in which a statistical word-based statistical language model is employed to identify the most promising recognition result, a word hypotheses binding process, which used to combine the candidate characters into words to construct a word-lattice, is required. Since the dictionary-based approximate word matching has the capacity of recovering some unrecognized characters by appending linguistic-prone character into the candidate set, the method of approximate word matching can be used in conjunction with a statistical-based post-processing system to improve its recognition performance [86].

Sheng and Fan present a post-processing system based on approximate word matching and Markov character BI-Gram model [79]. The recognized candidate characters are processed in two passes. In the first pass, word-level hypotheses are generated using hybrid contextual word matching. Non-dictionary words are recognized using an approximate string-matching algorithm. In the second pass, word-level hypotheses are verified and identified. The Markov character BI-Gram is applied to the established word-lattice and a best result is found. Employing dictionary-based approximate word matching, recognition accuracy is enhanced from 81.25% to 90%.

Combined with the Markov character BI-Gram model, a further 1.5% improvement is achieved.

The reported works on the post-processing techniques indicate that the dictionary-based approach has an advantage in removing erroneous character by employing approximate word matching to correct the error in a recorded multi-character word. On the other hand, the statistical-based approach can select the appropriate single-character word by searching a global optimal result and detect new words that are not recorded in the dictionary base on finding the most likely character string. Combining the dictionary-based top-down error detection and removal, and statistical-based bottom-up optimal result identification, a higher improvement performance can be achieved by utilizing a hybrid language model.

2.2 The Post-processing Approach based on Characteristics of Confusing Characters

Generally speaking, the presented post-processing approach based on computational linguistics information performs well in removing erroneous character. However, the unrecognized character problem puzzles this post-processing approach because the exact evaluation based on linguistics information cannot be correctly constructed for a character string with unrecognized character. The performance of post-processing approach based on computational linguistic information can be further enhanced by taking the characteristics of confusing character into consideration for reducing

unrecognized characters. By analyzing the input/output statistical data of the recognition results, the characteristics of the confusing character can be obtained. These characteristics can be utilized in recovering unrecognized character and in selecting the most plausible candidate based on posteriori statistical method.

Utilizing characteristics of confusing characters in post-processing, an effective way is to append appropriate similar-shaped characters to the given candidate sequence to recover some of unrecognized characters. In many reported works based on this method, the similar character set is established based on the analysis of training recognition result, and it will be used to append promising characters when a real candidate sequence is given. In Chang's work [4], the characters similar to a certain character category in shape are manually collected to construct a similar character set. When an ambiguous candidate character is detected, the similar-shaped characters in the similar character set for this candidate character are appended into the candidate sequence. Then the method based on language model is used to select an appropriate candidate. More practical, Most Error Prone Character set (MEPC set) is automatically established based on statistical analysis of training recognition result, and this set can be employed to recover the erroneous characters [95]. An example of such a post-processing system based on MEPC set is presented by Lee and Lin [41]. In the training phase, all confused characters of a character category came from the first five candidates are collected. Each candidate is assigned a weight in reversed-rank. If the reversed-rank sum of an output candidate is greater than a threshold, the input character is stored in the confusion set for the output candidate as an error-prone character. In the post-

processing phase, once an output candidate falls into the confusion set, its corresponding error-prone characters recorded in the MEPC set will be appended to the candidate sequence for further operation.

Another way to utilize the characteristics of confusing characters in post-processing is to directly identify the most-promising input character for a given candidate sequence based on error pattern and posteriori statistical characteristics of characters. Kaki presents a post-processing method based on Error-Pattern-Correction for Japanese character recognition [36]. The frequently occurred error character strings contain the erroneous characters, and their corresponding correct strings are extracted from the training recognition result and recorded in a database. These two strings form an error-pattern. Once a recorded error-string is detected in the recognition result, the post-processing system will make a correction by substituting a correct-part for an error-part. Error-Pattern-Correction is a simple but effective method because it detects and corrects error only by pattern matching. Furthermore, Chiang and Yu propose a method to distinguish similar Chinese characters by means of the utilization of confusing character pair database [13]. In this database, the frequently erroneous character category pairs are recorded. To distinguish these two characters, the appropriate critical features are selected. In the post-processing stage, once the first and the second candidates are matched with a recorded erroneous character pair in this database, with the HLVQ algorithm, the post-processing engine will recalculate the confidence scores between the input sample and the templates of the two candidate categories based on the sorted critical features. In this way, some mis-recognized characters appear in the first and

second candidates can be corrected. Moreover, to utilize the posteriori character statistical characteristics into post-processing, the input/output statistical characteristics of the recognizer for the training samples are analyzed and recorded. Regarding the output candidate characters and their associated confidence scores as observed features for the input character, when a candidate sequence is given in the post-processing phase, the methods based on posteriori statistical model can be employed to identify the most promising input character. The Noisy-Channel model, developed by Shannon, is widely employed as a posteriori statistical model [10, 37]. If a HCCR procedure is viewed as a transfer process in which input character images are transferred to recognized characters through a recognition engine, some errors may occur due to the existence of Noise. The Noisy-Channel model, a pure posteriori statistical model, is a simple but effective technique in describing and recovering the errors that may occur during such a transfer process. The experimental result of Chang's work [3], which is a typical post-processing system based on confusion matrix and Noisy-channel model, shows that 30% unrecognized characters can be recovered. Machine learning techniques are also employed to identify the input character from the observed candidates. In Sun's work [95], the training recognition results are analyzed to identify and classify the types and distributions of recognition errors. The plausible candidate insertion based on most likely unrecognizable characters set is designed for correcting unrecognized errors, and machine learning technique based on probability conversion map is used to retrieve the most promising input character for a given candidate sequence in order to correct the mis-recognized characters. When this method is applied to English OCR post-

processing, about 46% of total errors are corrected when the original character recognition accuracy is 91%.

Besides these two contributions, the approach based on the characteristics of confusing character can be used together with post-processing methods based on computational linguistic information to avoid unduly dependence on the language model, and to save recognition cost by reducing the searching space.

Analyzing the input/out statistical information of a recognizer, and applying such information to the candidate character identification and interpolation, the post-processing approach based on characteristics of confusing characters has different advantages than the one based on computational linguistic information, namely, context independence stability with respect to a certain recognizer, and better capacity to recover unrecognized characters. However, since this approach never uses contextual linguistics information, its performance for correcting mis-recognized characters is not good.

2.3 The Hybrid Post-processing Approach

Integrating the post-processing approach based on computational linguistics information and the one based on characteristics of confusing characters together, the resulting hybrid post-processing approach is expected to achieve a higher performance improvement by making full use of both the contextual linguistic information and the

characteristics of the recognizer [112]. Two integration strategies for hybrid post-processing approach are reported here.

One popular integration strategy is serial integration. The post-processing technique based on the confusing character characteristics is first conducted to reconstruct the candidate set by appending similar-shaped characters into the candidate set and adjust the confidence scores of the candidates. Then the post-processing technique based on computational linguistic information is applied to the reconstructed candidate set to identify the most linguistic-promising sentence from the candidate set [66]. A hybrid post-processing system for Japanese OCR correction is presented by Nagata [65]. In his system, the method based on Noisy-Channel model is first used to recover unrecognized characters and adjust the confidence scores of the candidates. An approximate word matching algorithm is then employed to retrieve the approximately matched words and the exactly matched words. Based on the part-of-speech TRI-Gram language model, the most plausible word sequence is then selected as the correction candidates from all combinations of exactly matched and approximately matched words. This hybrid post-processing system achieves a 6% recognition rate improvement when the original character recognition accuracy is 90%.

Another integration strategy is parallel integration that combines the computational linguistic information and the characteristics of confusing characters into a discriminator to evaluate the candidate sequence and select the most promising one as the output result. Miao [61] presents a simple hybrid post-processing system that integrates recognition similarity, character frequency, and character TRI-Gram model

together into a N-United-word model for post-processing. The confidence scores of candidate characters, the candidate character frequency W, and f_i, the frequency of a character appear in the i_{th} position of a N-United-Word are integrated to evaluate the possibility of a N-United-word. The word hypothesis with the maximum possibility is selected as the result. This hybrid approach achieves an encouraging performance improvement. Hidden Markov Model (HMM), a popular tool for modeling stochastic sequences with an underlying finite-state structure [72, 73], is another commonly employed parallel integration method [53]. An example work of using HMM for Chinese document recognition is proposed by Li and Ding [50]. In this work, based on the analysis of recognition results for the individual characters, the posterior probabilities of candidates are calculated as a product of the confidence scores of the candidates and characteristics of confusing characters. Hidden Markov Model that combines language model, posterior probabilities of candidates, and Viterbi searching algorithm, are applied to identify an optimal sentence hypothesis. On the average, 6% improvement for the first candidate accuracy is achieved when this method is applied to Chinese recognition systems with an original recognition rate of 90%.

2.4 Summary

In summary, the post-processing approach based on computational linguistic information is suitable for all character recognition systems. The contextual linguistic information can be employed to effectively identify a most promising sentence from the

candidate character set. However its performance sometimes suffers from the problem of unrecognized characters that the evaluation based on language model cannot be well constructed for character stings that contain unrecognized characters. On the other hand, the performance of the post-processing approach based on character characteristics of recognizer highly depends on the character recognition algorithm and the quality of the input handwritten sample. It is a labor-intensive task to prepare a confusion character set for each character recognition system. When the confusion character set for each recognizer is ready, the algorithms based on confusion character set are ritually the same for all character recognition systems. This approach has an obvious advantage of unrecognized character recovering over the approach based on linguistic information, but its overall correction performance is not high since no contextual information is employed. Benefits from making use of contextual linguistic information and characteristics of confusing characters in candidate character recovering and selection, integrating these two post-processing approaches together into a hybrid post-processing approach will lead to a higher correction performance, and therefore this approach will be adopted in the research works of this thesis.

Chapter 3

Fundamental Work for Language Modeling and HCCR Post-processing

In order to establish an effective language model in support of the post-processing approach based on computational linguistic information, the fundamental works for language modeling including corpus constructing and dictionary establishing are reviewed in section 3.1. Since the corpus-based statistical information is proven to be effective in strengthen the language model, a huge text corpus is constructed, and some of the text data in the corpus are processed for further language analysis. Next, the dictionary for recording word entries and relative linguistic statistics is established. In the second half of this chapter, a short introduction for two HCCR systems, one for online and another for offline, that will be used to evaluate the performance of our post-processing system, is given. Then the data interface between the recognizer and the

post-processing system is discussed. Finally, the construction of the testing samples is presented.

3.1 Fundamental Work for Language Modeling

3.1.1 A Huge Text Corpus

Owing to the increasing availability of machine-readable materials along with the rapid development of electronic and Internet publication, a huge corpus can be established and the corpus-based language analysis plays a more important role in natural language processing [33, 91].

A huge text corpus is constructed for extracting statistical linguistic information and supporting the construction of the dictionary. To assure the accuracy of the statistical linguistic information and reasonable distribution of linguistic phenomena, the selection of the text materials must satisfy the following requirements:

1. The typing mistakes in the materials must be small in number,
2. The sentences must be grammatically correct,
3. The text must cover a large variety of domains and writing styles, and
4. The synchronization of the materials should be considered.

Therefore, the official publication materials and newspapers are selected to build a huge text corpus. It is mainly derived from the following sources:

. People Daily China 1994-1999

. A collection of 100 selected Chinese publication in 1994

. China Computer World and China InfoWorld 1990-1998

. Beijing Daily and Beijing Evening Paper 1996-1998

. Chinese Youth Daily 1998

. Great Chinese Cyclopedia

. Home Library series CD including large volume of novels, philosophy works, history books and novels

. Electronic Novels

. Electronic Research Resource

Up to now, the memory size of the text corpus is more than 2.5 GB, with approximately 1, 140 million Chinese characters. Due to the fact that these plain text materials are neither segmented into word sequences nor tagged part-of-speech property, it is regarded as a raw corpus. The character frequency and character co-occurrence frequency are directly extracted from the raw corpus for further application.

A simple corpus management program is implemented to make a rough classification of the materials according to their topic. About 500MB text materials are classified into 8 major domains, namely literature, history, politics, technology, economy, military, entertainment, and spoken language. The distributions of the materials in different topics are shown in **Figure 3.1**.

Chinese sentences, unlike those of English, are written in a continuous string of characters without obvious separators indicating the word boundaries. Prior to further linguistic analysis, a Chinese sentence should be segmented into a sequence of words. During word segmentation, one could frequently encounter ambiguous strings which may lead to different segmentation forms and different sentence meaning. Therefore,

the corpus-based statistical linguistics information will be collected to support an effective word segmentation algorithm. About 40MB samples are pro rata selected from the classified text corpus that covers all the domains and manually segmented. The resulting word sequences are then used to construct a manually segmented text corpus from which the exact word-based statistics are extracted. Those sentences that have segmentation ambiguities are collected and the corresponding correct segmentation forms are recorded. These data are then used to build a 3.1MB ambiguity sub-library for the training of Chinese ambiguity string segmentation algorithm.

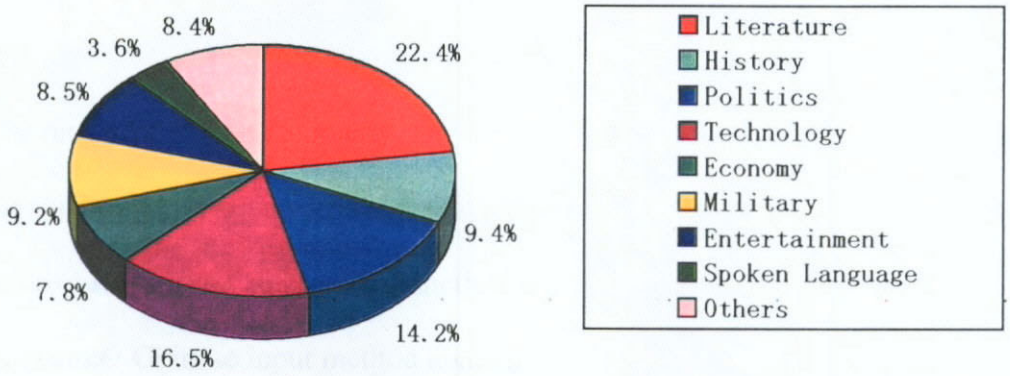


Figure 3.1 The Distribution of Classified Materials in Corpus

Trained by the corpus, the statistical-based automatic word segmentation algorithm based on Chinese word BI-Gram model and word UNI-Gram model (The details of the word segmentation algorithm will be presented in Chapter 4), are employed to segment the raw corpus. The materials in the raw corpus are automatically segmented into word sequences and the detected ambiguous strings are identified by trained segmentation ambiguity analysis program. The automatically segmented data will form an automatically-segmented corpus. This corpus is used to strengthen statistical language analysis and to refine the word-based statistics information.

3.1.2 Dictionary

Dictionary plays an important role in language processing. At first, we establish a basic lexicon to record the words in modern Chinese. Making use of the advantage of the dictionary-based word matching method in removing erroneous characters in multi-character-words while bearing in mind that this method suffers the out-of-dictionary problem, a large lexicon containing most words used in modern Chinese is established. The entries of this lexicon are mainly from the following sources:

. National standard of modern Chinese lexicon for information processing (GB13715) 39, 016 entries

. The modern Chinese dictionary 76, 292 entries

. The dictionary of new words and new phrases 7, 837 entries

. Microsoft Pwin98 Chinese input method lexicon 40, 586 entries

. Richwin97 Chinese input method lexicon 51, 292 entries

. Unregistered words collected from manually segmented corpus 8, 427 entries

. Unregistered words extracted from automatically-segment corpus 2, 022 entries

The method to extract unregistered words from automatically segmented corpus will be presented in **Chapter 4**.

Word entries from these sources are combined and checked, and finally a total of 104, 251 word entries, ranging from one character to eight characters per word with an average word length of 2.77 characters, are listed in our lexicon which is shown in **Figure 3.2**.

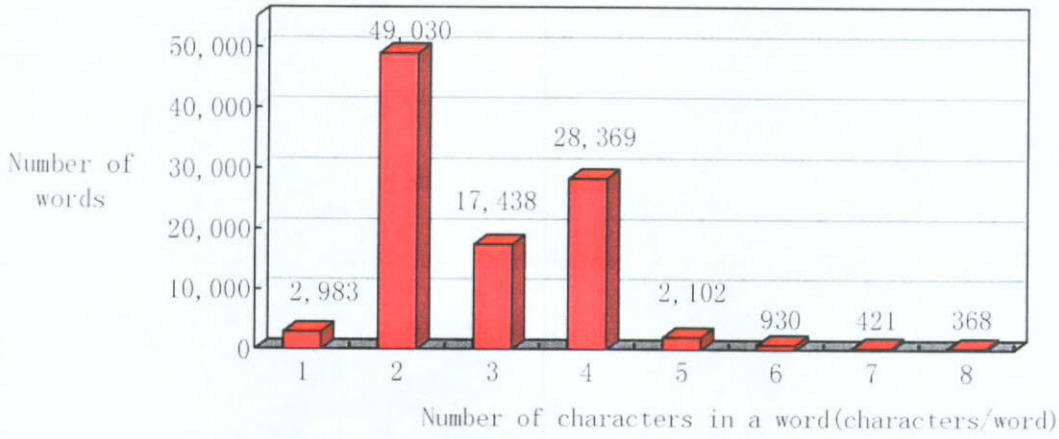


Figure 3.2 Word Length Information of the Lexicon

The occurrence frequencies of words are widely distributed. Based on the statistical experiments for 500MB automatically segmented materials, sorted by word frequency, the trend between the occurred word numbers and the coverage percentage for the number of occurred words is shown in **Figure 3.3**.

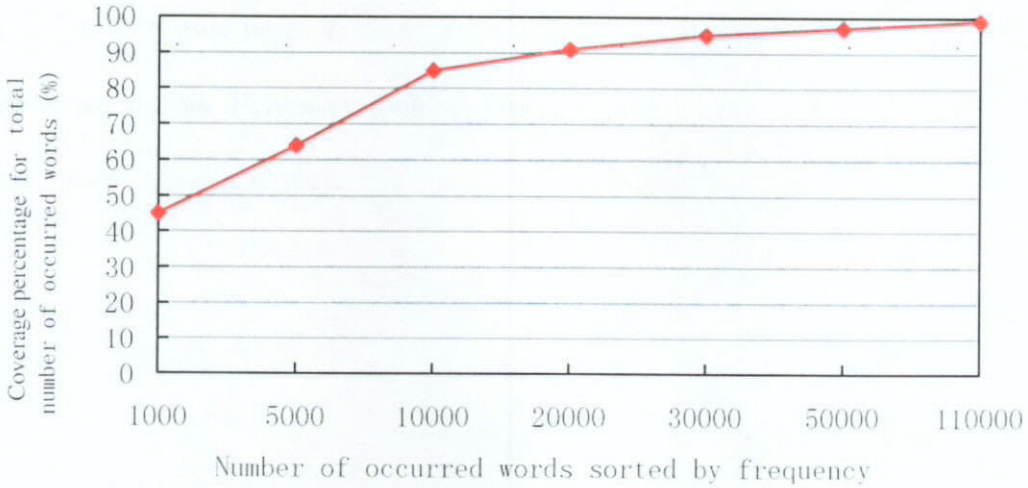


Figure 3.3 Number of Occurred Words Vs. the Coverage Percentage of Total Number of Words

In order to increase the speed of the dictionary lookup, a multi-level indexing table is established. Firstly, all word entries are sorted according to their GB codes and word

length. Each word entry in this lexicon is assigned an unique ordinal word-ID. An index table is established to record the starting and ending word-IDs corresponding to each index character and the total occurrence times of index character. Each word-ID is associated with its corresponding word occurrence times. To support reversed word matching that match the character string with the word entries in an inverted direction, a reversed lexicon is generated from the basic lexicon. Then an index table for reversed words is established. Each record in reversed index table consists of three components, namely the index character, the number of the words associated with the index character, and the word-IDs of these associated words. Then the statistical information of these words are recorded in the dictionary can be retrieved. The structure of the dictionary with a two-level and two-direction indexing table is shown in **Figure 3.4**.

This word-ID indexing scheme offers two advantages. One is that words with variable length can now be represented by word-ID with the same fixed length which will help speed up the dictionary look-up process. The other one is that it will save storage memory space and make the implementation of our proposed word BI-Gram model more efficient.

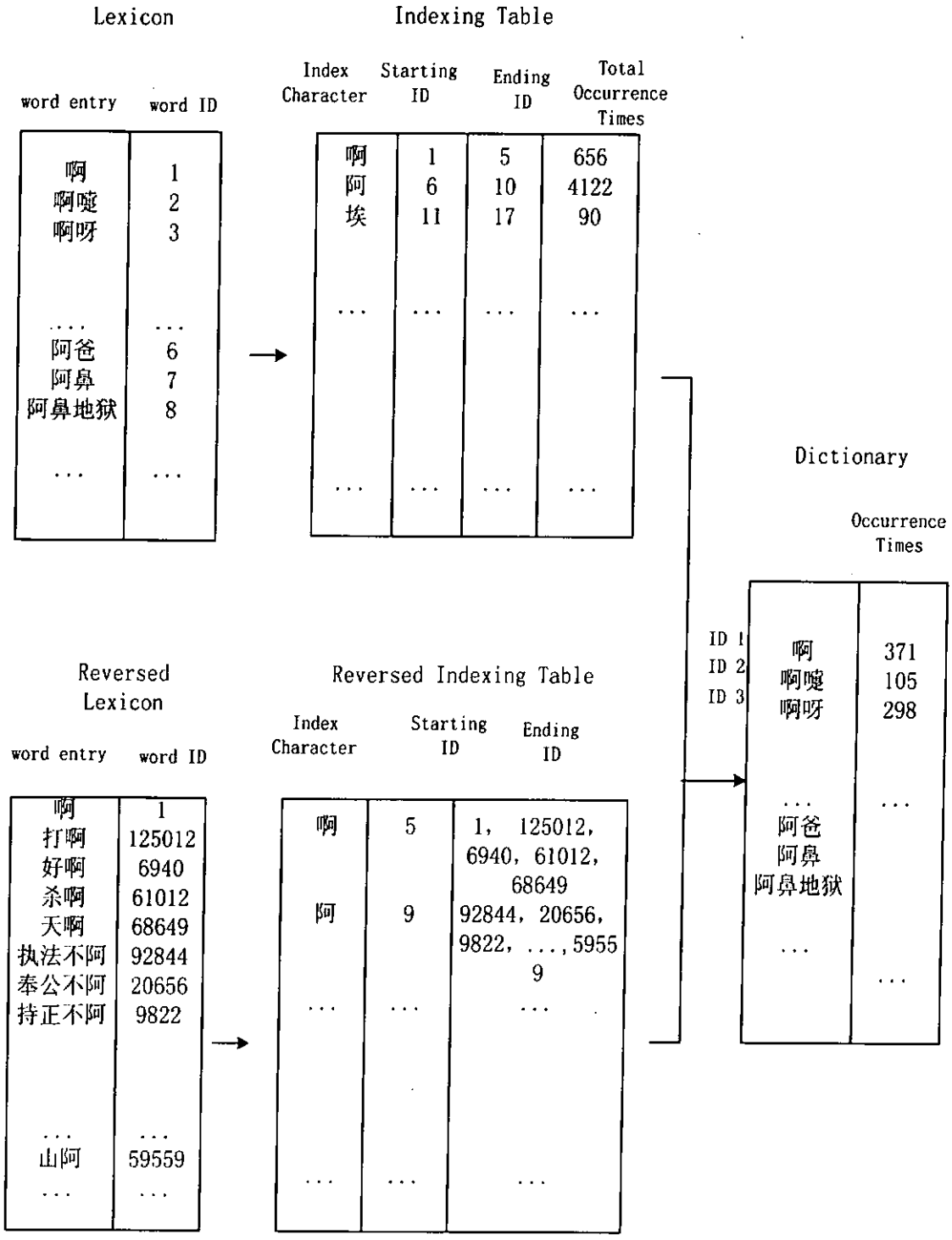


Figure 3. 4 The Structure of Dictionary

3.2 Handwritten Chinese Character Recognizers and Testing Samples

3.2.1 An Online Handwritten Chinese Character Recognizer

An online Chinese character recognizer developed by Jiafeng Liu and Wenhao Shu [54] is used to test our hybrid post-processing system. The input data of this recognizer that collects from the tablet connecting to the computer is expressed in terms of pen-down, pen movement, and pen-up actions. Some collected samples are given in **Appendix I**.

Based on statistical analysis on the distance distributions of interclass and intra-class Chinese character categories, this system uses stroke number, stroke order and endpoint position relationship of strokes as the prime features for recognition, and the stroke vector as an associate feature. For each input handwritten sample I corresponding to character C, a series of operations is performed, i.e., pre-processing, feature extraction, classification and matching. The noise data is firstly removed and the input pen-coordinate data is made smooth. Then the pen-coordinate sequences together with pen-up and pen-down action are analyzed to extract stroke sequences. In order to lessen the intra-class dispersion, line density equalization, which is a non-linear shape normalization algorithm, is utilized to equalize the stroke position. This is followed by a multi-stage classification and matching operation, in which a multi-branch coarse classification is performed to help reducing the matching space. A refined classification is achieved by relaxing the extracted feature sequences in matching against standard templates.

Finally the candidate characters C_1, C_2, ..., C_m make up a candidate sequence with the size of up to 10 candidates. Each candidate character is associated with a confidence value μ indicating the likelihood for the candidate character to be considered as the original one by the recognizer based on previously acquired knowledge. The confidence score is calculated as a function of the Euclidean distance and covariance matrix between the collected character sample I and the standard template of the candidate character. Its detail definition is given in [54]. μ ranges from 0 to 1, and a larger value of μ means more similarity.

For the tested handwritten samples of about 1.3 millions, this recognition engine achieves an average of 90% recognition rate for the first candidate character and 95% for the top-10 candidates.

3.2.2 An Offline Handwritten Chinese Character Recognizer

One may observe that usually recognition rate of an offline handwritten Chinese character recognizer is lower than the one for online case because the number of strokes and stroke orders of the input sample are known in the case of online recognition, but not so for the offline recognition. Therefore, post-processing technique is more important as far as the improvement of the recognition rate of offline HCCR systems is concerned. But it is also more difficult to design because a large number of erroneous characters may undermine the language model for post-processing and more confusing characters are to be identified.

In this study, an offline handwritten Chinese character recognition engine [80, 81] is employed. This writer-independent recognition engine that supports a vocabulary of 3755 GB2312-80 Chinese character categories, uses both the image of a character and its structural information to classify the characters. The scanned character images are firstly normalized to a standard size, and then a noise removal algorithm based on seed filling is employed to remove isolated noise pixels. After thinning the image of strokes, the preliminary classification is carried out on the basis of rapid transformed stroke density features. Then the observed samples will be classified by rules which are acquired by learning from examples. An advanced extension matrix algorithm is employed here to solve the multi-class problem with overlapping area, in which a heuristic search based on the average entropy is used to get approximate solutions of minimal complexity, and a potential function is used to estimate the probability density function of the area of overlap between positive and negative examples. In this way, nonlinear separating hyperplanes between classes may be obtained.

Similar to the online HCCR system, the output of this engine is the candidate character sequence together with their confidence scores μ, ranging from 0 to 1, to describe the similarity between the standard template of the candidate character and the scanned character image. The higher confidence score means more similar, and he details of the calculation for the value of μ is given in [80].

In the previous experiment, this offline HCCR engine reports an 81% average accuracy rate for the first candidate and 90% for the top-10 candidates respectively.

3.2.3 The Interface between Recognizer and Post-processing System

An appropriate interface between the recognizer and the post-processing system is necessary to ensure that enough available information from the recognizer can support the post-processing system. Suggested by Murveit and Ward [64, 103], we can roughly classify the interface between the recognizer and the post-processing system into the following categories:

Top-Best Hypothesis

The top-best hypothesis provides the post-processing system with the first characters selected from each of the candidate sets first candidate character. Some dictionary-based post-processing methods, similar to the one adopted in automatic Chinese correction system, are applied to analyze the top-best hypothesis to identify the erroneous characters and remove the errors by approximate word matching. This kind of interface has the advantage of less data exchange and low computational cost, and more useful information generated by a recognizer can be transferred to the post-processing system such as more good candidates. Furthermore, if the accuracy of the original first candidates is low, the performance of the post-processing system will drop significantly.

N-best Hypotheses

The recognition engine outputs the first N most likely sentences hypotheses constructed by the candidate sets, and the post-processing system further processes them

and chooses the most promising one [65, 76]. The value of N is determined according to the tradeoff between the allowed interaction and the additional work involved. The advantage of this approach is that it reduces the difficulties of the post-processing system to identify the most plausible hypothesis. The disadvantage is that N may increase exponentially with the length of the input sentence in order to keep the correct hypothesis in the N-best hypotheses.

Candidate Character Matrix

In this approach, the output of the recognition engine is a confusing candidate character matrix, in which the candidate characters are sorted according to the similarity between the input sample and the standard template. This interface is adopted in most existing post-processing systems [98]. The advantage of this approach is the high degree of interaction between the two components. Furthermore, benefit to one can make use of the various statistical features of language models and the characteristics of recognition engine to search for the optimal result from the candidate set.

Parallel Integration

In this approach, the constraints of the post-processing procedure are integrated directly in the recognizer to reduce the search space [114]. This is perhaps the most direct and attractive approach, but its implementation is difficult. The advantage of this kind of interface is that it allows considerable interaction between the recognition engine and the language model. Such an interaction is expected to save computational

cost and make the recognition system "smart". The main disadvantage of this approach is that, when complicated natural language models are integrated into the recognition engine, the computational cost increases significantly.

In order to support a higher performance improvement for the recognizer and minimize additional work involved in the integration, the candidate character matrix interface is used in our research.

3.2.4 Testing Samples

There are 200 sets of online handwriting samples written by 200 persons. Each set of samples consists of 6763 samples for the corresponding character categories in the GB2312-80 character set. These samples are collected from writing-pads with a sampling resolution of 150 DPI. Each character sample records the pen-actions data and its writer. To construct the testing samples for evaluating the post-processing system, the text materials containing meaningful sentences with about 100,000 characters are randomly selected in which 80% of all text materials are from the manually segmented corpus and the rest are from automatically segmented corpus. Then character samples are random selected from the character sample library to construct the online handwritten testing samples for the meaningful sentences.

Similar to the construction of the online HCCR testing, about 100,000 offline handwritten samples corresponding to the characters that form meaningful sentence are randomly selected from 200 sets of handwritten samples in which each set of

handwritten sample consists of 3755 character samples with respect to the 3755 character categories in first-level sub-set of GB2312-80 character-set. Since the offline recognition engine chosen can only support a vocabulary of 3755 character categories, only those sentences composed of the characters in first-level sub-set of GB2312-80 character-set are selected during the text material selection process.

Chapter 4

HCCR Post-processing Techniques based on Contextual Language Model

This chapter presents the works on the post-processing techniques based on contextual language model, and their performance are evaluated and then compared. The dictionary-based post-processing approach, which adopts a top-down strategy, is proposed in **Section 4.1**. To segment the Chinese sentences that consist of continuous characters into word sequence, an automatic word segmentation algorithm based on BI-directional Maximum Matching method and Chinese word BI-Gram model is firstly presented. It is an exact matching algorithm. **Section 4.1.2** presents a simple unregistered word identification algorithm to strengthen the word segmentation algorithm by reducing the out-of-dictionary problem. In dictionary-based post-processing system, the word segmentation algorithm is employed to segment the sentence hypothesis generated by the recognizer into word sequence and during this

process, and then the sentence fragments can be located. The dictionary-based approximate word matching is then applied to these sentence fragments to remove the erroneous characters by substituting the linguistic-prone characters for the erroneous ones in the fragments. This approach is evaluated through experimental results which demonstrate the effectiveness of this method in removing the erroneous characters in multi-character words. **Section 4.2** presents our post-processing techniques based on the popular statistical approach. Linguistic statistical information is employed to evaluate the probabilities of the sentence hypotheses constructed by selecting characters from the candidate sets, and the one has the maximum probability is outputted as the result. Our post-processing system implements the BI-Gram model since word is the minimum, complete semantic unit that can form a sentence. In order to further enhance the linguistic description capacity for the long-distance restrictions in Chinese sentences, the word BI-Gram model is extended to a distant word BI-Gram with maximum context distance of 3 words. The experimental results of employing these two statistical language models in erroneous character removal will be analyzed.

4.1 Dictionary-based Top-Down Post-processing Approach

In this approach exact word segmentation program is used to segment the sentence hypothesis produced by the recognizer. Then the sentence fragments which consist of either a minimum of any three characters or two characters in which one of them cannot form a word by itself, are identified. It is followed by applying a dictionary-based approximate to match these sentence fragments, and the erroneous characters are

removed. In **Section 4.1.1**, the word segmentation algorithm based on BI-directional Maximum Word Matching method and word BI-Gram model is presented. To reduce the out-of-dictionary problem that undermines the word segmentation algorithm, Section **4.1.2** presents an unregistered word identification algorithm. The dictionary-based post-processing technique is proposed in Section **4.1.3** and its performance is evaluated.

4.1.1 Word Segmentation Algorithm based on BI-directional Maximum Matching and Word BI-Gram Model

One may observe that Chinese sentences are written in a continuous string of characters without obvious separators indicating the word boundaries. Therefore the sentence should be segmented into word sequences prior to further linguistic analysis. During the segmentation process, some ambiguities may be encountered. For example, the sentence "深入研究生物理论有着重要的现实意义" could be segmented into several different word sequences such as "深入 研究 生物 理论 有 着 重要 的现实 意义" or "深入 研究生 物理 论 有 着重 要的 现实 意义". All of the segmented words in both cases could be found in the dictionary but their meanings are quite different. Since this kind of ambiguity widely exists in Chinese sentences, word segmentation method plays an important role in Chinese language processing.

Usually the ambiguous strings are classified into two major types: overlapping ambiguity and combinatorial ambiguity, depending on the structure of the ambiguous

string. Considering a string '*ABC*'. (Where *A*, *B* and *C* is a character or a word in the lexicon *W*), if both $AB \in W$ and $BC \in W$, the string '*ABC*' is called an overlapping ambiguous string. On the other hand, a combinatorial ambiguity occurs when a string '*AB*' has $A \in W$, $B \in W$ and $AB \in W$. It is reported that 86% of the ambiguities falls into the category of overlapping ambiguity, and therefore our word segmentation algorithm mainly addresses this case.

There are two major approaches proposed in Chinese word segmentation: rule-based approach [117, 118] and statistical approach [6, 104]. In most word segmentation methods, a dictionary is utilized to match the sentences in order to generate the combination hypotheses. Then, the grammatical rules and the structural relationships among the words are used to remove the ambiguities in a rule-based approach. For the statistical approach, statistical information is used to construct discriminate functions for removing the ambiguities. Some hybrid approaches have been proposed in which statistical information is used to identify unknown words after the dictionary-based matching is executed to reduce the segmentation ambiguities [68].

In our system, a statistics-based word segmentation algorithm is designed. Since the Chinese is semantically composed of continuous words, word is regarded as the basic linguistic unit and the word BI-Gram statistical information is employed to resolve the ambiguities. This algorithm consists of two components, namely ambiguity identification and ambiguity removal.

A dictionary-based BI-directional word matching method is utilized to pre-segment a given sentence in order to identify segmentation candidates and to locate overlapping

ambiguous strings [52]. Firstly, a forward maximum matching method is used to pre-segment the sentence and the position of segmentation is recorded. Secondly, the sentence is matched again using backward maximum method. If no difference between these two results can be found, then no ambiguity is detected. Otherwise, the number of words segmented by the two matching methods is compared. The matching method which resulting in fewer words is selected.

The statistical method will be used to remove any remaining ambiguity. At that time potential ambiguities are recorded and the identical parts of segmentation results will be treated as the 'correct-segmented' words. Considering the following sequence

$$X\, A_1 A_2 \ldots A_{i-1}\, A_i \ldots A_{n-1}\, A_n\, Y,$$

where X and Y are the "correct" words and the Chinese characters $A_1 \ldots A_n$ represent the ambiguous string. Suppose the two sequences of possible split spots after pre-segmentation are represented as $pa_1, pa_2 \ldots, pa_k$ *and* $pb_1, pb_2 \ldots, pb_k$. The ambiguous string is segmented into two sequences of words:

$$W_1 W_2 \cdots W_{j-1} W_j W_{j+1} \cdots W_{k-1} W_k W_{k+1} \text{ and } W_1' W_2' \cdots W_{j-1}' W_j' W_{j+1}' \cdots W_{k-1}' W_k' W_{k+1}'.$$

$$\begin{array}{c} X\ W_1\ W_2 \cdots W_{j-1}\ W_j\ W_{j+1} \cdots W_{k-1}\ W_k\ W_{k+1}\ Y \\ pa_1\ pa_2 \cdots pa_{j-1}\ pa_j\ pa_{j+1} \cdots\ pa_{k-1}\ pa_k \end{array}$$

$$\begin{array}{c} X\ W_1'\ W_2' \cdots W_{j-1}'\ W_j'\ W_{j+1}' \cdots W_{k-1}'\ W_k'\ W_{k+1}'\ Y \\ pb_1\ pb_2 \cdots pb_{j-1}\ pb_j\ pb_{j+1} \cdots\ pb_{k-1}\ pb_k \end{array}$$

Let $W_1\ W_2\ W_3\ \ldots W_{k+1}$ be the sequence of words found in the dictionary. The statistical information is utilized to evaluate the linguistic probability between the split

spots. The mutual information of the two words and the *t-test* value of three adjacent words, which are the product of word co-occurrence frequencies, are utilized here.

The word co-occurrence frequency information is collected in advance from training text corpus. The word pairs and their frequencies appear in the training text are collected and then filtered to remove the insignificant and low frequency word pairs. For the remaining word pairs, an effective data compression method is employed to reduce the storage size. In this way, a word BI-Gram statistical information database is obtained, in which the frequently encountered word pairs and their occurrence frequencies are recorded. The details for the construction of this database will be presented in **Section 4.2.1**.

Based on the obtained word frequency and word pair frequency, the mutual information of two words, which is used to measure how strong two events are related, is calculated as follow [20],

$$I(a:b) = \log_2 \frac{P(a,b)}{P(a)P(b)} . \tag{4.1}$$

Here a and b are two independent events with probabilities $P(a)$ and $P(b)$, and $P(a, b)$ denotes the joint probability of a and b.

One may observe that:

$I(a:b) >> 0$ means a and b are highly related.

$I(a:b) \approx 0$ means a and b are nearly independent.

$I(a:b) << 0$ means a and b are very few related.

The mutual information between a and b is a useful measure to express how often the two events co-occur. If $r(a)$ and $r(b)$ denote the occurring times of the word 'a' and

'b' in a corpus containing *N* words and *r(a, b)* is the co-occurring time of the word pair 'ab', then the following equation can be easily derived:

$$p(a) \approx \frac{r(a)}{N}, p(b) \approx \frac{r(b)}{N}, \ p(a,b) \approx \frac{r(a,b)}{N}$$

$$I(a:b) = \log_2 \frac{P(a,b)}{P(a)P(b)} \approx \log_2(N) + \log_2(\frac{f(a,b)}{f(a)f(b)}). \tag{4.2}$$

Suppose *a, b* and *c* are three continuous events. The *t-test* value of *b* relative to *a* and *c* is defined as follows:

$$t_{a,c}(b) = \frac{p(c|b) - p(b|a)}{\sqrt{\sigma^2(p(c|b)) + \sigma^2(p(b|a))}}. \tag{4.3}$$

Here, *p(b|a)* is the conditional probability of the event *b* relative to *a*, likewise *p(c|b)* is the conditional probability of the event *c* relative to *b*. $\sigma^2(p(b|a))$ denotes the variance of *p(b|a)*. Assuming that a word pair *'ab'* occurring in a N-word corpus falls into a binomial distribution with parameters *N* and *p(a/b)*, an approximation could be obtained according to the same method used in calculating the mutual information. The parameters in **Equation 4.3** could be estimated as follow:

$$p(b|a) = \frac{p(a,b)}{p(a)} \approx \frac{r(a,b)}{r(a)}, p(c|b) = \frac{p(b,c)}{p(b)} \approx \frac{r(b,c)}{r(b)},$$

$$\begin{aligned} \sigma^2(p(b|a)) &= \sigma^2(\frac{p(a|b)}{p(a)}) \approx \sigma^2(\frac{r(a|b)}{r(a)}) \\ &= N * \frac{r(a,b)}{N} * (1 - \frac{r(a,b)}{N}) = \frac{r(a,b)}{r^2(a)} \end{aligned} \tag{4.4}$$

The *t-test* information could be treated as the measure describing the binding force of three continuous words. For instance, $t_{a,c}(b)>0$ means the word *'b'* shows a related trend with *'c'*. On the contrary, $t_{a,c}(b)<0$ implies that the word pair *'ab'* has a stronger

association than *'bc'*. Making use of two word pairs BI-gram information, *t-test* is an effective measure to describe the relationship between words.

Based on the mutual information for two adjacent words and the *t-test* value for three adjacent words, the probability between the split spots mentioned above can be calculated by:

$$I(pa_{j-1}) = I(W_{j-1} : W_j),\ \ I(pa_j) = I(W_j : W_{j+1}) \tag{4.5}$$

$$\Delta t(pa_j) = t_{W_{j-1},W_{j+1}}(W_j) - t_{W_j,W_{j+2}}(W_{j+1}) \tag{4.6}$$

For an identified ambiguous string, the statistical information for different split spots suggested by the pre-segmentation program is compared. Once a segmentation form is found to have a larger mutual information value over a threshold, it is considered as being correct. Otherwise, the *t-test* value is used to further evaluate the statistical information for the different split spots, and the segmentation form yielding a larger t-test value difference is considered as the correct one.

The automatic segmentation result by employing our algorithm are compared with the one by employing the popular character BI-Gram model and word UNI-Gram. It is found that a better segmentation accuracy of about 98.7% can be obtained by employing a word BI-Gram model, while 95.6% and 96.4% average segmentation accuracy are obtained by employing character BI-Gram model and word UNI-Gram model respectively. The word segmentation algorithm is not only used to support automatically segmented corpus, which is discussed in Chapter 3, and it is also used in the dictionary-based post-processing system.

4.1.2 Un-registered Words Identification

The analysis for the proposed experimental word segmentation results indicates that the segmentation errors are attributed to two major reasons. One is the out-of-dictionary problem that undermines the ambiguity identification task, and another one is the combination errors which can not be recovered by this method. In view of the fact that there are many unregistered words and new words continue to form, an effective unregistered word identification algorithm is needed. We shall present a simple algorithm based on character co-occurrence frequency and mutual information of the character pairs here.

The continuous character strings that cannot form multi-character words in the segmented result are collected and analyzed. The mutual information between two characters is firstly evaluated. Supposing A and B are two adjacent characters, from **Equation 4.1**, the mutual information for character A and B are calculated. If its value is much larger than 0, this means these two characters have a strong association to form a word. However, some spurious association may be found based on the mutual information of the character pairs. This is because that when the number of occurrences of A or B is small, the mutual information for A and B will be large in spite of the number of occurrences of AB is very small. Therefore, a practical threshold for the number of least occurrence of characters A and B with the value of 10 is selected to filter some un-frequently occurred character pairs. In this way, the two-character-word can be retrieved.

Then this method is extended for finding longer words. Define $P(AX_nB)$ as the probability that A occurs before B with an infix X_n of at most N characters between them. In our lexicon, the longest words have a length of 8 characters, and N is set to 6. Then the mutual information of this string is calculated by,

$$I(AX_nB) = \log_2 \frac{P(AX_nB)}{P(A)P(B)} . \tag{4.7}$$

Once the value of $I(AX_nB)$ is much larger than 0, the string AX_nB is regarded as a word.

Thus, a complete unregistered word identification algorithm is given as follows:

Unregistered Words Identification( )

{

for (n=0; n<=6; n++)

{

Collect the character string AX_nB, that begins at character A and ends at character B with a n-character infix X_n. The occurrences of AX_nB are recorded.

For each recorded character string AX_nB,

If the occurrences of A or B is less than 10, then AX_nB is ignored.

If $MI(AX_nB)$ less than a threshold which is much larger than zero, then AX_nB is ignored.

Else

AX_nB is recorded in a temporary dictionary for n with their occurrences.

Open the temporary dictionary for smaller value of *n-1*, and subtract the occurrences of AX_nB from the occurrences of AX_n or X_nB. Recalculate $MI(AX_n)$ and $MI(X_nB)$ to decide whether it should be recorded.

}

}

Table 4.1 The Flow of Un-registered Word Identification Algorithm

This process identifies the character pairs with the length of 2 to 8 characters as potential words. During the construction of the lexicon, presented in Section 3.1.2, these potential words are checked and the reasonable ones will be recorded in the lexicon. In this way, 10449 new word entries are obtained. The new dictionary will decrease the out-of-dictionary problem and enhance the word segmentation accuracy. A small-scale segmentation experiment for about 40,000 characters shows that the word segmentation accuracy using the original dictionary is 98.4%, and it increase to 99.3% after 10449 new words are appended into the lexicon.

4.1.3 Using Dictionary-based Approximate Word Matching in Post-processing

A dictionary can be used to identify ambiguous strings in the candidate characters produced by the recognizer. The similarity between the ambiguous strings and the word entries in the dictionary is evaluated by using dictionary-based approximate matching and contextual information. The word entry with the largest similarity value is then

selected to replace the ambiguous string. In this way, potential errors are detected and corrections are appended to the candidate set.

The sentence hypotheses that is formed by the first candidate character from each candidate character sequence is firstly segmented into word sequence for further analysis. The BI-directional Maximum Matching method with word BI-Gram model is employed to segment the sentence hypothesis into words and remove the ambiguities. Then, the sentence fragments can be identified. A sentence fragment consists of a minimum of three characters, or two characters which one of them cannot form a word by itself. The characters that cannot form a word by itself are numbered about 2,500 in the GB2312 character-set. The identified sentence fragments are regarded as potential errors, and further analysis are needed.

Firstly, a number of Chinese linguistic rules are identified to help remove some of these sentence fragments. Here are some examples.

1. if a sentence fragment = numeral (一, 二, ..., 十, 百,千, 万, 亿, 兆,...)+ quantifier (such as 个, 种, 吨, 类, 此), then it is ignored.

E.g. The sentence fragment "五十七个" (fifty seven), consists of a numeral "五十七" (fifty seven) and a quantifier "个", and it will be ignored.

2. if a sentence fragment = Chinese surname + less than 3 characters, then it is regarded as a Chinese name and is ignored (The most frequently used Chinese surname is less than 300).

E.g. The sentence fragment "潘仁美" (a Chinese name) is ignored.

3. if a sentence fragment = less than 3 characters + a suffix for a place-name, then it is regarded as a place-name and is ignored. (The most frequently used suffixes for place-name including 省, 市, 州, 乡, 县, 镇, 区, 山, 峰, 江, 河, 湖, 海, and so on.)

E.g. The sentence fragment "澜沧江" (Lan-Cang River) "小笠山" (Xiao-Li Hill) "揭阳市"(Jie-Yang City) are regarded as proper nouns for place-name and they are ignored.

These rules can help to reduce the out-of-dictionary problem. The approximate word matching technique is used to take care of the rest in order to find out the linguistics-prone characters. The approximate matching technique consists of two steps, word hypothesis generation and word hypothesis approximate matching.

Suppose $C = C_i C_{i+1} \cdots C_{i+k}$ is a $k+1$-character fragment starting from the *ith* position in the sentence, and W_{i-1}, W_{i+1} are the two adjacent matched words of this sentence fragment. Then the word hypothesis generation is performed as follows:

1. If in the dictionary, there exists any word W that contains W_{i-1} as its prefix, then W_{i-1} will be combined with its subsequent characters in the sentence fragment C to construct a word hypothesis of the same length as W.

2. For each C_j, $j=i,\ldots, i+k-1$, append the subsequent characters in C such that it is matched with words with C_j as their prefixes. Repeat this until C_{i+k-1} is done.

3. If in the dictionary, there exists any word W that contains W_{i+1} as its suffix, then W_{i+1} will be combined with its preceding characters in C to form a word hypothesis of the same length as W.

The following similarity weighting equation evaluates the similarity *SA* between the word hypothesis *WH* and the word entry *WE*

$$SA(WH,WE) = \alpha \cdot \frac{l_m}{l_e} + \beta \cdot \frac{1}{l_e - l_m} \sum_{s=1}^{l_e - l_m} \frac{tw(C_s)}{t(C_s)} + \gamma \cdot \frac{1}{l_e - l_m} \sum_{s=1}^{l_e - l_m} S(C_h \mid C_e) + \mu \cdot f(WE), \quad (4.8)$$

where, α, β, γ and μ are the weight parameters for the four terms, and $\alpha + \beta + \gamma + \mu = 1$ to ensure *SA(WH,WE)* has a value between 0 and 1;

l_m is the number of matched characters between the *WE* and *WH*, and l_e is the character length of word entry *WE*;

C_h and C_e are the unmatched characters between *WH* and *WE* respectively. Obviously the number of unmatched characters is $l_e - l_m$, in our system, the maximum allowed number of unmatched characters is 1 for the word entries consisting of not more than 4 characters, and 2 for the longer word entries;

$tw(C_s)$ is the number of times that C_s appears in the training corpus as a part of a multi-character word and $t(C_s)$ is the number of times that this character appears in the training corpus;

$S(C_h \mid C_e)$ is the similarity score between C_h and C_e which is defined in **Section 4.2**; and

$f(WE)$ is the frequency of *WE* in the training corpus.

This similarity evaluation equation consists of three terms. The first term gives the similarity between the word hypothesis and the word entry in terms of matched length. The second term describes the corpus-based word construction capacity of the

unmatched characters, and the last term is the average confusing probability of C_{wj} being recognized as C_s.

Calculating the similarity score for each word hypothesis, the ones with a similarity value higher than a given threshold is appended in the word-lattice and the unmatched character is regarded as a linguistics-prone one. If this character doesn't appear in the candidate set, then it is added as a part of an approximately matched word. Otherwise its confidence parameter will be adjusted. Suppose this character C_h appears in the *j-th* position of the *i-th* candidate sequence and C_{i1} is the first candidate character with the confidence score μ_{i1} output by the recognition engine, then

$$\mu_{ij} = \min\{1, \mu_{ij} + SA(WH, WE) \cdot (\mu_{i1} - \mu_{ij})\}. \tag{4.9}$$

Two groups of test samples with the size of about 10,000 are employed to test the performance of our dictionary-based post-processing technique.

For the test samples of the online HCCR system with the original recognition accuracy of 91.78% for the first candidates and 95.2% for the first ten candidates, the dictionary-based post-processing method achieves 41% correction rates for the erroneous characters for the first candidates and yields a 3.37% overall recognition accuracy improvement. Particularly, we observe that 52% of the unrecognized-characters can be recovered. It is an encouraging result. On the other hand, about 0.6% of all of the first candidates are wrongly substituted by the linguistic-prone characters for the original recognition result.

For the test samples of the offline HCCR system with the original recognition accuracy of 81% for the first candidates and 92% for the first ten candidates, the

dictionary-based post-processing method achieves a 1.2% overall recognition rate improvement while about 13% of the unrecognized characters can be recovered.

The smaller improvement for the offline HCCR samples may be caused by the lower recognition rate of the first candidates, which unavoidably weakens the sentence fragments identification and a large number of continuous single characters cannot be effectively evaluated by the dictionary-based post-processing method. Furthermore, the dictionary-based method tends to select words with longer length, and some correctly recognized single-character-word strings are wrongly modified.

4.2 Statistics-based Bottom-Up Post-processing Approach

In this section, two statistics-based post-processing approaches will be discussed.

4.2.1 Statistics-based post-processing approach and N-Gram model

The statistics-based approach, which has a simple mathematical description, can be easily implemented so it is popularly adopted. Statistics-based post-processing approach works employ a bottom-up strategy in which smaller linguistic units are firstly constructed from the candidate set and evaluated, and then they form larger units. Finally, a most promising sentence can be identified from the candidate set. Let $I = i_1, i_2, ..., i_n$ be the character images of $c_1, c_2, ..., c_n$ which form a sentences S. Suppose for each input image, the recognition engine outputs m candidates and thus the n candidate sequences form a $n \times m$ candidate character set. A sentence hypotheses, denoted by S', is formed by selecting one candidate character from each sequence. The statistics-based

post-processing method is used to find a sentence hypothesis, $\hat{S}$, that has the maximum likelihood among a total of m^n sentence hypotheses. This problem can be formulated as:

$$\hat{S} = \arg\max P(S' \mid I). \tag{4.10}$$

By Bayes' rule, the above equation can be rewritten as:

$$\hat{S} = \arg\max \frac{P(I \mid S')P(S')}{P(I)}. \tag{4.11}$$

Considering that $P(I)$ is independent of S', **Equation 4.11** becomes

$$\hat{S} = \arg\max P(I \mid S')P(S'). \tag{4.12}$$

where $P(S')$ is called the language model and $P(I|S')$ is the confidence score of the sentence hypothesis S' which can be computed from a priori likelihood for the candidate character,

$$P(I \mid S') = \prod_{x=1}^{n} P(i_x \mid CC_x), \tag{4.13}$$

where $P(i_x \mid CC_x)$ is the confidence score of each candidate character.

N-gram model is a widely used language model to find out $\hat{S}$ [11, 22, 41, 113, 122]. N-Gram model works under the Markov assumption, that the occurrence probability of each linguistics unit u_i only depends on its previous *N-1* linguistics units $\{u_{i-n+1},\ldots,u_{i-2},u_{i-1}\}$. Suppose a sentence S is composed of the linguistics units $u_1, u_2, \ldots, u_k$, the occurrence probability of the sentence S can be computed as the product of the probabilities of $u_1, u_2, \ldots, u_k$:

$$P(S) = P(\mu_1,\ldots,\mu_k) = P(\mu_1)P(\mu_2 \mid \mu_1)\ldots P(\mu_k \mid \mu_{k-n+1},\ldots,\mu_{k-1}) = \prod_{i=1}^{k} P(\mu_i \mid \mu_{i-n+1},\ldots,\mu_{i-1}). \tag{4.14}$$

One may observe that the memory and computation requirements of N-Gram is too big to process when N is larger than 3. Therefore the BI-Gram and the TRI-Gram model are the most popular N-Gram models. For the BI-Gram model, the **Equation 4.14** is simplified as:

$$P(S) = P(\mu_1)P(\mu_2 \mid \mu_1)P(\mu_3 \mid \mu_2)\ldots P(\mu_k \mid \mu_{k-1}) = \prod_{i=1}^{k} P(\mu_i \mid \mu_{i-1}) \cdot \tag{4.15}$$

In view of the fact that word is the minimum complete semantic unit that can form a sentence, the Chinese word is selected as the basic linguistics unit in our system. Then the sentences S is composed of $w_1, w_2, \ldots, w_i$ and its probability can be computed as:

$$P(S) = P(w_1, \ldots, w_k) = P(w_1)P(w_2 \mid w_1)\ldots P(w_k \mid w_{k-n+1}, \ldots, w_{k-1}) = \prod_{i=1}^{i} P(w_i \mid w_{i-n+1}, \ldots, w_{i-1}) \cdot \tag{4.16}$$

For word-based BI-Gram, (4.15) can be simplified as $P(S) = \prod_{i=1}^{i} P(w_i \mid w_{i-1}) \cdot$ Thus,

$$\hat{S} = \arg\max\ P(I \mid S')P(S') = \arg\max \prod_{j=1}^{i} P(w_i \mid w_{i-1}) \prod_{k=1}^{n} P(w_k \mid i_k), \tag{4.17}$$

and it can be employed to identify the most promising sentence hypothesis based on word BI-Gram.

4.2.2 Post-processing Method based on Chinese Word BI-Gram Model

An important problem to solve when employing Chinese word BI-Gram model in HCCR post-processing is to acquire and record its statistical parameter, i.e. word co-occurrence frequency. As already mentioned in **Section 3.2**, a lexicon is built with a vocabulary of 104, 251 words. Each word entry in the lexicon is assigned an unique ordinal word-ID and a multi-level indexing table is designed for saving storage memory and speed dictionary lookup.

Considering the vocabulary size of words is much larger than the one for characters, which is unlike that of the character BI-Gram model where all possible combination forms can be recorded simultaneously, only the observed co-occurrences of word-pairs will be recorded. The large-scale segmented corpus discussed in **Section 3.1**, is used to extract the statistical information of the word BI-Gram model. The sentences consist of segmented words in the corpus are loaded and the responding word-IDs are generated. A word pointer is set to the first word of the training materials. Once a current word identified by the word pointer, and its following word form a word-pair that has been observed, then the occurrence probability of this word-pair node will increase. Otherwise, a new word-pair node is generated, and its occurrence is set to 1. Then the word pointer is set to the next word, and the same procedure will be repeated until all the sentences in the training corpus are processed.

To speed the searching procedure for verifying whether a word-pair has been encountered, and for sorting procedure on the word-pairs according to the indexing word-IDs, an optimal statistical algorithm is designed. Firstly, we allot an inner-buffer in the memory that can store K word-pairs. During the statistics collection process, the K encountered word-pairs obtained from the training materials are stored in the inner-buffer. Then a quick-sort algorithm is employed to sort the word-pairs according to their word-IDs. Next, the word-pairs and their associated occurrences are combined with the ones recorded in the statistical parameter file that is stored in the outer memory such as hard disk, and re-sorted according to their word-IDs. Once the parameter file is larger than a pre-defined threshold, a new parameter file is created for recording newly

encountered word-pairs. It is aimed to decrease the times needed to merge the inner-buffer and the parameter files. After merging the inner-buffer and the parameter file, the next K word-pairs are loaded from the training materials, and the same procedure will be repeated until all the sentences in the training corpus are processed. Once more than one parameter files are created, the records from these parameter files are combined and sorted.

Figure 4.1 shows the trend of the numbers of word pairs with increasing corpus size. With the increasing training corpus, the number of observed word-pairs increase rapidly in the initial stage, and after more than 100MB corpus being trained, the increase in the number of new observed word-pairs will slow down. When the trained corpus size exceeds more than 160MB, the curve for the number of observed word pairs nearly levels off.

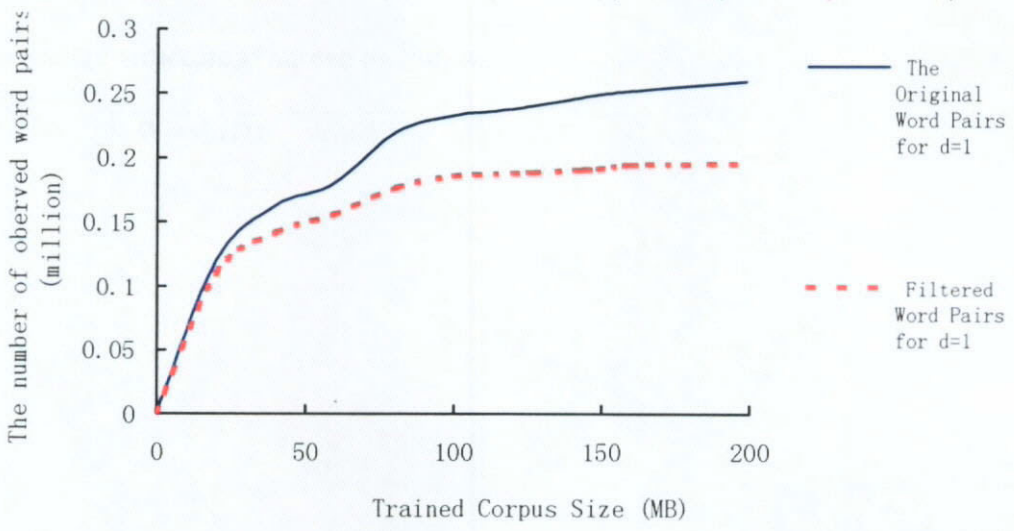


Figure 4.1 The Number of Observed Word-pairs Vs. Trained Corpus Size

Certainly, not all the word pair frequencies need to be recorded in the final word co-occurrence database, due to the existence of a large number of insignificant and burst word pairs. Therefore, a multi-step combination and sparse word pairs removal algorithm is employed. Unlike other methods which use a fixed minimum occurrence times of word pairs as a removal threshold, two factors are considered in our algorithm. One factor is a minimum occurrence times of word-pairs and the other one is a minimum percentage of occurrence times of a word-pair over the occurrence times of the index word. It is motivated by the fact that some low-frequency words show a 'strong' association tendency with another word in spite of infrequent occurrence of the word-pairs. The trend of the number of filtered word-pairs is also shown in **Figure 4.1** with a dotted line.

Since word co-occurrences are represented as a huge sparse matrix, an effective data compression method is needed to reduce the storage space and speed lookup. A basic parameter storage structure is shown as follow:

Word ID1	Word ID2	Occurrence Times

Then word co-occurrence data are re-constructed into a table indexed by word ID1 for the purpose of saving parameter memory and speed lookup. In this way, we need to save the ID1 only once, and the size of parameter memory is obviously reduced. Finally we build a 14MB BI-gram statistical parameter database which is small enough to be processed by a PC.

In order to identify the most plausible output, a multi-stage word transition graph is constructed. Then the problem of searching for the most plausible result becomes an

optimum path searching problem in a multi-stage transition graph. An effective search algorithm, Veterbi searching algorithm [67], is employed to identify the optimum path. Since the Veterbi searching algorithm is well known, only a brief introduction of this algorithm is given here instead of detailed discussion.

The Viterbi searching is essentially a dynamic programming algorithm, consisting of traversing a network of states and maintaining the best possible path score at each state in each frame. It is a time synchronous search algorithm in that it processes all states at time t before moving on to time $t+1$. Viterbi algorithm is applied to limit the search space by pruning out the less likely states.

The time-synchronous nature of the Viterbi search implies that the 2-D space is traversed from start status to ending status. The search is initialized at time $n \times m$ with the path probability at the start state set to 1 and at all other states to 0. In each frame, the computation consists of evaluating all transitions between the previous frame and the current frame, and then evaluating all NULL transitions within the current frame. For non-NULL transitions, the algorithm is summarized by the following expressions:

$$P_j(t) = \max_i (P_i(t-1) \cdot a_{ij} \cdot b_j(t)), i \in \text{ set of predecessor states of } j. \qquad (4.18)$$

where, $P_j(t)$ is the path probability of state j at time t, a_{ij} is the static probability associated with the transition form state i to j, and $b_j(t)$ is the output probability associated with state j. This expression includes NULL transitions that do not consume any input.

Thus, every state has a single best predecessor at each time instant. One can easily determine the best state sequence for the entire search by starting at the final state at the end and following the best predecessor at each step all the way back to the start state.

The complexity of Viterbi decoding is N^2T (assuming that each state can transit to every state at each time step), where N is the total number of states and T is the total duration (number of frames). Since the state space is huge for large vocabulary applications, the beam search heuristic is usually applied to limit the search space by ignoring the less likely states. The combination is often simply referred to as Viterbi beam search and is proved effective in many cases.

We propose a different state node generating method, compared with the regular one that generates nodes together with a searching procedure. In this method we advanced bound the candidate characters are firstly bound into words, and the bounded words are regarded as nodes. Furthermore, the regular method bounds the characters into words from left to right and searches the optimum path without considering the widely overlapping ambiguities, as mentioned in **Section 4.1.1**, for our system, the word hypotheses are generated by bounding the candidate characters from two directions. Each node in the transition graph contains two parameters, namely, word and its confidence score generated by the recognizer. The multi-stage transition graph is then constructed as follows. For each character in the candidate set at the current stage, a node is generated for the matched word which has this character as its prefix. If no matched word is found, a dummy node is generated with its weight equal to zero. Such

a construction starts from the first character and continues until the last character is exhausted. In this way, a multi-stage transition graph is obtained.

Treating the word co-occurrence probability $P(W_i|W_{i-1})$ as the transition probability from one stage to the next stage, the Veterbi dynamic programming search algorithm is employed to find the optimum path in the word transition graph. The word node sequence in the optimum path contains the final output result of the recognition engine.

An experiment to evaluate the performance of our statistics-based post-processing techniques is done. Two popularly employed statistical language models, namely character BI-Gram model and word UNI-Gram model, are implemented and tested. The resulting accuracy improvements, compared with the one by employing the word BI-Gram model, are shown in **Figure 4.2**.

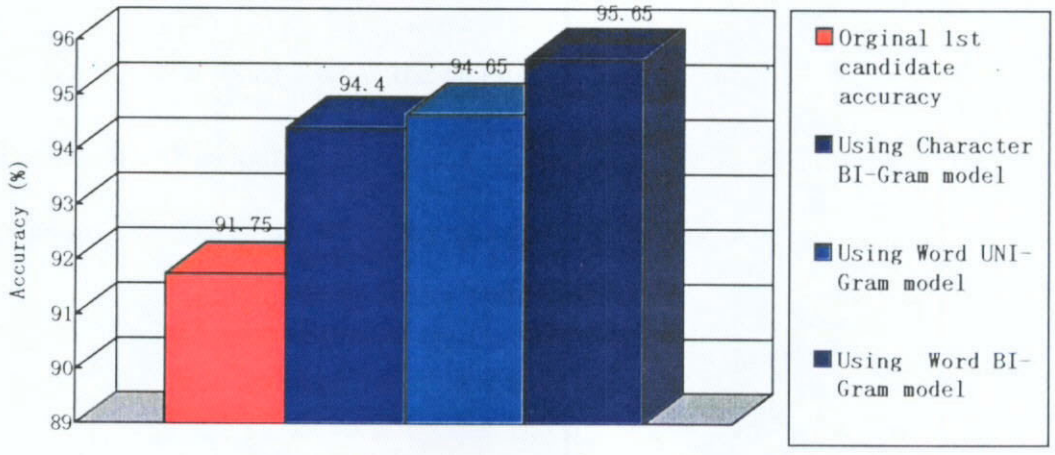


Figure 4.2 The Improved Recognition Rate by Using Different Statistical Language Models Individually

Employing the word BI-Gram model, an average of 3.9% recognition rate improvement over the original system performance is achieved, which is obviously

higher than the 2.65% for character BI-Gram model and 2.9% for the word UNI-Gram model. This result shows that our word BI-Gram model is effective in recovering the confused characters. Furthermore, the result indicates that the word-based language model produces a better performance, compared with the character-based ones.

4.2.3 Post-processing Method based on Distant Chinese Word BI-Gram Model

In the previous section, the word BI-Gram model is proven effective in post-processing and achieves a better improvement of the recognition rate, compared with employing the popular character BI-Gram model and word UNI-Gram model. But a serious problem remains that many long-distance restriction in natural language can not be described by the BI-Gram model. For example, a Chinese phrase "一本有趣的书" (An interesting book) is composed of the words : 一 本 有趣的 书" (an, set, interesting, and book). The word "本" is the quantifier for the noun "书". But the lower order N-Gram such as BI-Gram and TRI-Gram cannot describe this kind of linguistic restriction. Extending the BI-Gram model by considering the distance parameter, the distant BI-Gram model is designed for make a trade-off between the description capability of long distance restriction and an acceptable size of parameter space [84].

Conventional BI-Gram models estimate the probability of the current word according to the previous word, while distant BI-Gram models (with the maximum considering distance d_{max}) describe the probability of current word according to its up to d_{max} pervious words. For example, applying a distant BI-Gram Model with a maximum distance of 2, the probability of a sentence is changed to,

$$P(s)=\prod_{i=1}^{k} P(w_i \mid w_{i-2}, \text{distance}=2)^{\lambda_2} \cdot P(w_i \mid w_{i-1}, \text{distance}=1)^{\lambda_1},\ \lambda_1+\lambda_2=1. \quad (4.19)$$

When $d_{\max}$ is set to 1, the Distance-1 BI-Gram is reduced to the a conventional BI-Gram model.

Distant BI-Gram models makes it possible to partly resolve the long-distance restriction problem, which is a main barrier to acquire refined description capacity by using the BI-Gram model. Furthermore, the parameter space for distant BI-Gram is obviously less than the one for the TRI-Gram. For a distant BI-Gram model with $d_{\max}$, the number of possible combination is only $d_{\max}$ times over the one for a regular BI-Gram, which is obviously less than the one for a TRI-Gram model.

When employing the distant BI-Gram model in HCCR Post-processing, the selection of an appropriate value for $d_{\max}$ is very important. According to Zhang and Huang's work [123] on the study of the optimal value of parameter N in N-Gram Chinese statistical model, three factors are considered and analyzed. The first one is the average number of Chinese words in a phrase which can be treated as an approximate expression for Chinese grammatical structure. For this factor, *N=3* is the best value. The second one is the capability for reconstruction of a new word and *N=4* is a better choice. The last one is the performance of the higher order N-Gram model for Chinese word segmentation in a small-scale experiment and Quad-Gram model achieves the best performance. Considering these three factors, *N=4* for word N-Gram model is regarded as a better choice. With a similar reason, 3 is selected as the maximum distance, i.e., a

distant word BI-Gram model with $d_{\max} = 3$ will be employed in our post-processing system. Thus, the probability of a sentence is calculated as

$$P(s) = \prod_{i=1}^{k} P(w_i \mid w_{i-3}, \text{distance} = 3)^{\lambda_3} \cdot P(w_i \mid w_{i-2}, \text{distance} = 2)^{\lambda_2} \cdot P(w_i \mid w_{i-1}, \text{distance} = 1)^{\lambda_1},$$

$$\lambda_1, \lambda_2 \lambda_3 > 0, \textit{and } \lambda_1 + \lambda_2 + \lambda_3 = 1, \tag{4.20}$$

where $\lambda_1, \lambda_2, \textit{and } \lambda_3$ are the weights parameters with respect to each distance value.

The corpus for training our previous word BI-Gram model is used for training the distant word BI-Gram model with $d_{\max}$ =3. Thus, the training corpus are scanned for three passes, and each time the statistical word-pairs data corresponding to a increasing considering distance are recorded. For a scanning pass with the distance parameter *d*, all observed word-pairs with distance *d* are recorded. The parameter storage and sorting method is the same as the one for word BI-Gram statistics collection. With the increasing training corpus size, the number of observed word-pairs increases rapidly in the initial stage, and after more than 80-100MB, the increase in the number of newly observed word-pairs will slow down. When the trained corpus size exceeds 160MB, the curve for the number of observed word pairs begins to level off. The comparison of the number of observed word-pairs curves with respect to different distances *d* are shown in **Figure 4.3**.

From **Figure 4.3** one may observe that the number of observed word-pairs increases with *d*. Since the word co-occurrences are represented as a huge sparse matrix, a filter program is conducted to remove the many insignificant and burst word-pairs and hence the storage space is reduced. The trend of the number of filtered word-pairs is

also shown in **Figure 4.3**. One observes that for the filtered cases, the number of observed word-pairs is smallest for d=3. This is contrary to the trend for the un-filtered ones. That means together with the increased distance, the restrictions between two distant words will be weakened. From the statistical results and some linguistics research results, the weight parameters $\lambda_1, \lambda_2, and\ \lambda_3$ are experientially set to 0.60, 0.25and 0.15 respectively, and

$$P(S) = \prod_{i=1}^{n} P(w_i \mid w_{i-3}, \text{distance} = 3)^{0.15} \cdot P(w_i \mid w_{i-2}, \text{distance} = 2)^{0.25} \cdot P(w_i \mid w_{i-1}, \text{distance} = 1)^{0.60} \ . \quad (4.21)$$

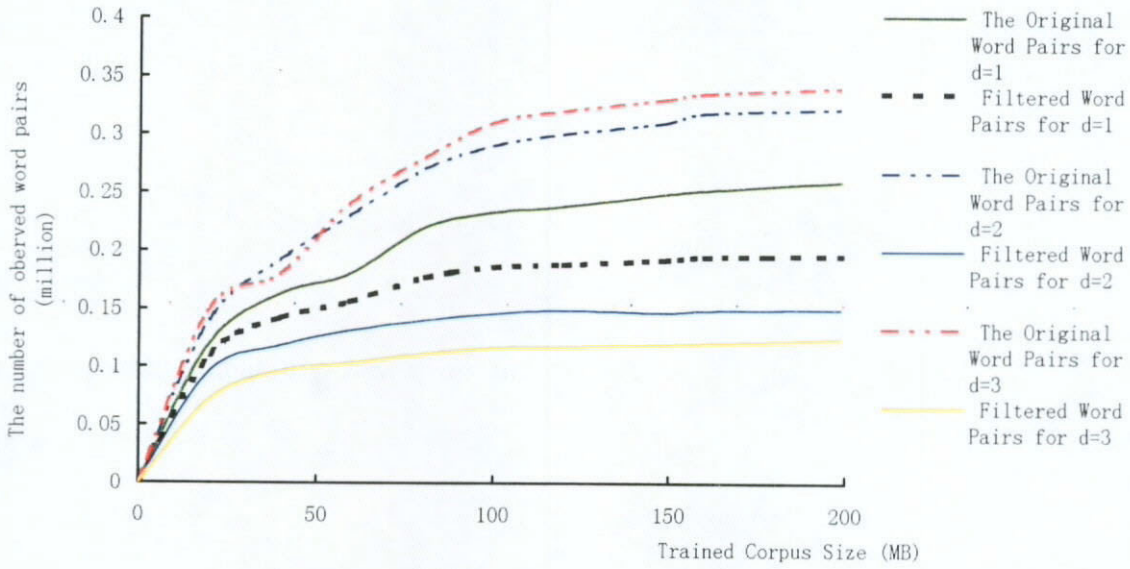


Figure 4.3. The Number of Observed Word-pairs Vs. Proceeded Corpus Size for Different *d*

The processed results are further compressed and indexed to reduce the redundancy and save storage space. Finally, three BI-Gram statistical parameter databases with respect to *d=1,2,3* are obtained, the storage sizes are 14MB, 11MB, and 9MB respectively. That means we can approximately describe the long distance restriction

with d_{max} =3 with only about 2.4 times parameter storage space which is much less than the theoretical parameter space requirement for TRI-Gram.

To employ distant Chinese word BI-Gram model in HCCR post-processing, the candidate characters will be bounded in advance from two directions and a word-lattice is constructed. Then the search algorithm based on the distant word BI-Gram model will be employed to identify a path with the maximum likelihood as the output result. The distant word BI-Gram is used to evaluate the linguistic probability of the current unit according to the distant units, and the delayed Viterbi algorithm which evaluates the path based on the product of several past units, is employed to search for the optimum path. For each node in the word-lattice, the word entry is associated with its frequency and recognition confidence scores. If an isolated character which cannot form a word by itself is observed, then a dummy node is generated and its parameter is set to zero.

Treating the probability based on the distant word BI-Gram model, as the transitional probability from one state to the next state, the Veterbi dynamic programming search algorithm is employed to search for the most promising result consisting of words in the optimum path.

The experiment for evaluating the performance of our distant Chinese word BI-Gram model is done for the offline HCCR result with a known lower accuracy, and it is compared with the ones for character BI-Gram model, word UNI-Gram model, and word BI-Gram model. The acquired recognition rate improvements for the first candidate characters are shown in **Figure 4.4**.

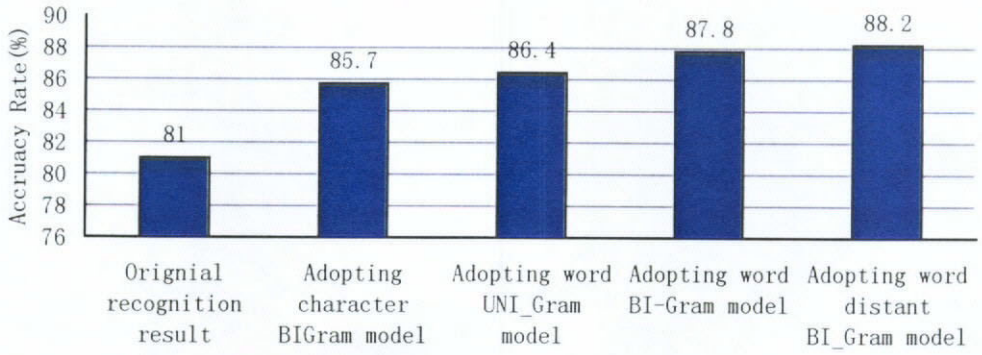


Figure 4. 4 The Improved Recognition Rate by Employing Different Statistical Language Models

Employing the distant word BI-Gram model, a 7.2% recognition rate improvement is achieved, which is obviously higher than the 4.7% improvement by employing character BI-Gram model and 5.4% improvement by employing word UNI-Gram, but it is only slightly better than the 6.8% improvement by employing word BI-Gram model. This result indicates that selecting word as the basic linguistics unit, a word-based language model can achieve a better performance compared with the character-based ones. Analyzing the unsuccessfully cases by employing distant word BI-Gram model, we notice that many of the errors are due to unrecognized characters so that a correct word hypothesis cannot be formed. Therefore, the statistics-based post-processing approach should be integrated with other approach which can recover some of the unrecognized characters, and a higher performance can be expected.

Chapter 5

HCCR Post-processing Techniques based on Characteristics of Confusing Characters

This chapter presents our works on the post-processing techniques based on the characteristics of the confusing characters. Making use of the character posteriori probability, this approach is designed for recovering unrecognized characters and adjusting the confidence scores of the candidates. The motivation of this post-processing approach is presented in **Section 5.1**. The confusing character set which is for recording the confusing characters and their confusing probabilities for each character category, is defined in **Section 5.2** and its properties are discussed. **Section 5.3** describes the character recognition as a signal transmission process through a noisy channel, and a linear statistics-based Noisy-Channel model is employed to identify the most plausible character for a given candidate set and adjust the confidence score for each candidate. Considering the confusing character sequence and their confusing

probability as the features of a character category, neural networks for similar character category groups can be used to classify the given candidate sequence and identify the most promising one as the output result. This part of the work is presented in **Section 5.4**. Its performance in correcting recognition error is evaluated and compared with the one by employing the Noisy-Channel model.

5.1 Motivation of The Post-processing Approach based on Characteristics of Confusing Characters

In **Chapter 4**, the post-processing technique based on computational linguistic information was presented and it has been proven effective in removing erroneous characters. However, this method is puzzled by the fact that exact evaluation of a linguistics unit cannot be correctly carried out due to the presence of unrecognized characters.

We will use the following correction rate to evaluate the performance of a variety of post-processing techniques:

$$CorrectionRate = \frac{A_1^* - A_1}{100\% - A_1}, \tag{5.1}$$

where A_1 is the accuracy rate for the first candidate output by the recognizer, and A_1^* is the improved first candidate accuracy rate after the post-processing is done. A major assumption made by popular post-processing methods based on statistical language model is that all input characters must appear in the candidate set in order to ensure a correct sentence output. If it were not the case, the mis-recognized character problem

will occur, and errors cannot be corrected by this method alone. Obviously the correction rate of a statistical-based post-processing approach is related to A_m, namely the accuracy rate for the first m candidates (for most reported recognition systems, m is ranging from 3 to15). Thus, employing the statistical-based post-processing approach individually has an upper theoretical correction rate limit,

$$CorrectionRate_{\max} = \frac{A_m - A_1}{100\% - A_1} . \tag{5.2}$$

For the online HCCR engine adopted in our experiments, its A_1 is 90% and A_m is 96%. The upper theoretical limit for the correction rate by using a statistical-based post-processing technique is then (96%-90%) / (100%-90%) = 60%. As for the offline HCCR engine with average A_1 equal to 80% and A_m equal to 90%, the upper theoretical limit for the correction rate is 50%. Obviously, many errors cannot be removed by the statistical-based post-processing approach. The purpose of adding a post-processing stage based on the characteristics of the confusing characters produced by the recognizer is to enhance A_m by appending some possible mis-recognized characters to the candidate set, so that the upper correction rate limit is raised.

Furthermore, for the post-processing methods based on computational linguistic information, the output results are determined by the particular language model being used. It leads to an unduly dependence on the language model. For this reason, some frequently used characters are often selected while those infrequently used characters are usually ignored. This may produce erroneous result. Hence we propose to incorporate the post-processing techniques based the characteristics of confusing

characters into the computational linguistic information based method to minimize such a problem.

5.2 Establishment of Confusing Character Set

One of the key factors for the success of post-processing methods based on characteristics of confusing characters is the collection and analysis of the statistical characteristics of the confusing characters. In view of the fact that candidates for a character category are normally encountered in a set of similar-shaped characters rather than from the whole character-set, the confusing character set, which is for recording the confusing characters and their confusing probabilities for each character category, is established from the recognition results based on the training samples in order to save computation cost and reduce the influence of burst recognition errors.

A total of over 1.35 millions online handwritten samples, with 200 samples for each character category in GB2312 character-set written by 200 persons, are collected as the training samples. The candidate sequence and their confidence scores, output by our HCCR recognizer, are recorded. Analyzing the recognition results of training samples, the confusing-prone characters for each character category are collected for constructing the confusing character set for an online HCCR system. In the same way, the recognition results for over 0.75 millions of offline handwritten samples are collected to construct the confusing character set for an offline HCCR system.

For a character category A_i, the recognition results of its 200 training samples $\{ I(R) = IA_{i1}, IA_{i2},..., IA_{i200} \}$ form the matrix $\{C = (C_{ij}), i = 1,...,200, j = 1,...,20\}$ with up to the first 20 candidates being considered, and the confidence score for each C_{ij} is μ_{ij}.

For each encountered candidate character category $B=C_{ij}$ in this matrix, we use the following equation to calculate the similarity score between B and the sample character category A by comparing the confidence scores of category A and category B in the recognition results of training samples of category A,

$$S(B,A) = \min\{1, \frac{\sum_{(i,j)\in F} \mu_{ij}}{\sum_{(k,l)\in G} \mu_{kl}}\}, F = \{(i,j) \mid C_{(i,j)} = B, i = 1,...,200, j = 1,...,20\},$$
$$G = \{(k,l) \mid C_{kl} = A, k = 1,...,200, l = 1,...,20\} \quad (5.3)$$

The value of the similarity score *S(B,A)* ranges from 0 to 1, and the larger *S(B, A)* means more similarity between character category A and B. If *S(B,A)* is bigger than a threshold r, B is treated as a confusing-prone character of A and it is appended into the confusing character set of category A. An element in the confusing character set, *(B, S(B,A), T)*, records the confusing character category B, its corresponding similarity score *S(B,A)*, and the encountered times T. Repeat this procedure until all the confusing-prone characters for each character category are recorded.

One may notice that the value of the threshold r influences the size of the confusing character set and the coverage of the encountered times of the recorded confusing characters over the number of all encountered candidates. The relationship between the value of threshold r, the average size of confusing character sets, and the average

candidate character's coverage percentage for training samples are respectively shown in **Figure 5.1** and **Figure 5.2**.

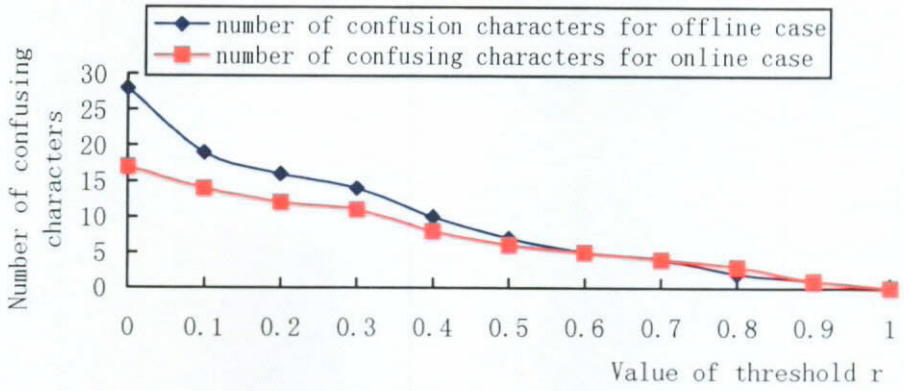


Figure 5.1. Threshold *r* vs. Average Size of Confusing Character Sets

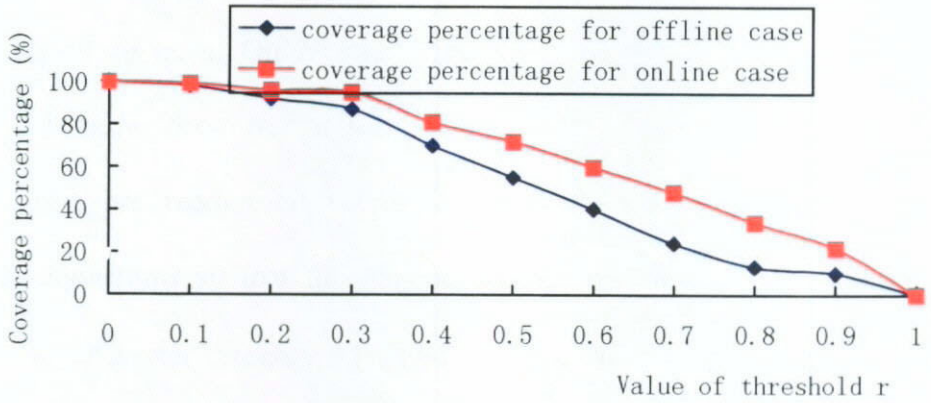


Figure 5.2 Threshold *r* vs. Candidate Character Coverage Percentage

Take these two factors into account, $r = 0.5$ is selected as the default threshold for the online case, and the corresponding average number of confusing characters is 6, while the average candidates coverage is 72%. As for the offline case, $r = 0.4$ is selected as the default threshold, and the corresponding average number of confusing characters and candidates coverage are 10 and 71% respectively.

Some example confusing character sets for the online case are given below.

人 人 0.995 199 八 0.976 197 入 0.972 196 卜 0.933 195 义 0.560 195 火 0.442 182

揩 揩 0.932 199 楷 0.737 189 措 0.585 129 擂 0.549 134 摺 0.508 132 谐 0.437 108

木 木 0.988 198 术 0.952 194 火 0.932 190 不 0.910 193 歹 0.863 191 太 0.852 186

少 0.815 192 仆 0.419 116

One may observe that the confusing character sets have two properties, namely, the recognizer dependent and nonsymmetrical [122].

Recognizer-dependent. For a character category *A*, not only its corresponding confusing characters, but their total numbers are depending on the recognizer. Normally a recognizer with higher recognition rate will have a small confusing character set and vice versa. On the other hand, the confusing characters for a character category are different for different recognizers. This is quite obvious because the character images are recognized based on various feature extraction and pattern classification algorithms so that the obtained recognition results are different. For example for the character category "干" and "千", a stroke-order independent offline recognizer will treat them as a confusing character pairs due to their similar shapes, but not for a stroke-order dependent online recognizer because the orientation of their top stokes are directly opposite..

Nonsymmetrical. This means if character category *B* appears in the confusing character set of *A*, *A* does not always belong to the confusing character set of *B*. For example, in our online system, character category "木" is sometimes recognized as

"不" when the noise filter module wrongly removes the top part of stroke "丨" as a noisy data, but "不" is seldomly recognized as "木".

Regarding the confusing characters and their corresponding confusing probabilities as the features of a character category for a recognizer, classification algorithms are used to evaluate the given candidate sequence and identify the most promising character.

5.3 Noisy-Channel Model in Post-processing based on Characteristics of Confusing Characters

HCCR procedure could be viewed as a transmission process in which input character images are transferred to recognized characters through a recognition engine. During this process, some errors may occur due to noise in the channel. In [3], a Noisy-Channel model was shown to be an effective technique in describing and recovering the errors that may occur during such a transfer process.

In the post-processing, confusing character sets can be used to recover some of the unrecognized characters by appending similar-shaped characters. Suppose B is a candidate character in the candidate sequence for an input sample I with the confidence score $\mu(B,I)$. We search for the confusing 3-tuples $(B,A,\ C(B|A))$ indexed by B, satisfying

$$\mu(B,I) > 0.7 * C(B \mid A) \ , \quad \text{where 0.7 is a experimental parameter.} \tag{5.4}$$

If A does not appear in this candidate sequence, then A is appended as a possible unrecognized character. The confidence score of this appended candidate is assigned the value $0.5*C(B|A)$.

Suppose $C_1, C_2, ..., C_n$ are candidate characters for the input sample of character category C. According to Bayes' Rule,

$$\begin{aligned} & P(C^* \mid C_1 C_2 \cdots C_n) \\ &= \underset{i=1,\ldots,n}{Max} P(C_i \mid C_1 C_2 \cdots C_n) \\ &= \underset{i=1,\ldots,n}{Max} \frac{P(C_1 C_2 \cdots C_n \mid C_i)}{P(C_1 C_2 \cdots C_n)} \\ &= \underset{i=1,\ldots,n}{Max} \frac{\prod_{j=1}^{n} P(C_j \mid C_i)}{P(C_1 C_2 \cdots C_n)} \end{aligned} \quad . \tag{5.5}$$

If only the first ten candidates are considered, then $P(C^* \mid C_1 C_2 ... C_{10})$ is simplified to:

$$P(C^* \mid C_1 C_2 ... C_{10}) = \frac{\underset{i=1,\ldots,10}{Max} \prod_{j=1}^{10} P(C_j \mid C_i)}{P(C_1 ... C_{10})}, \tag{5.6}$$

where $P(C_j \mid C_i)$ is the confusing parameter $C(C_j \mid C_i)$. The character with the highest probability, namely, C^*, is regarded as the possible input character and then the similarity distance of each candidate character is adjusted as follows:

$$\mu_{new}(C_i \mid I) = C(C_i \mid C^*) \cdot \mu_{old}(C_i \mid I). \tag{5.7}$$

After the similarity parameters are adjusted, the candidate characters are re-sorted in an ascending order according to their similarity parameter values.

An experiment of employing Noisy-Channel and confusing character set in the post-processing system for online HCCR system shows an average of 1.5% and 2.0% improvement for the first candidate accuracy rate and the first ten candidates accuracy rate being achieved when the original accuracies are 91.4% and 96.2% respectively. That means about 17.4% mis-recognized characters are corrected by employing Noisy-

Channel model, while 52% of all the unrecognized characters are recovered. It is an encouraging result.

The result for the offline case by employing the Noisy-Channel model and confusing character shows that, for the testing sample with 82.3% and 90.1% accuracy for the first candidate and the first ten candidates, the accuracy is raised to 82.9% and 92.4% respectively. The corresponding correction rate of 7.8% and 23.3% is obviously lower than the online one.

The nature of the character recognition is a complex, non-linear process. But in this method it is modeled as a pure statistical, linear one. Hence the improvement on the recognition rate is not impressive. Some non-linear techniques should be considered instead.

5.4 Neural Networks for Similar Character Category Groups in Post-processing based on Characteristics of Confusing Characters

Unlike the post-processing method being discussed which is based on widely used Noisy-Channel model, a post-processing technique use neural networks for similar character category groups to identify the most promising character for a given candidate sequence.

Regarding the confusing characters and their confusing probabilities as the features of a character category, for a given candidate character sequence, we treat the post-processing procedure based on characteristics of confusing characters as a classification problem. In the reported research works, artificial neural networks (or named neural

networks in short) is proven effective in small-scale non-linear classification, but its capability will be obviously weakened when the scale of classification space is enlarged. For our post-processing system, a neural network for classifying thousands of character categories is difficult to be established, maintained, and trained. Therefore, we cluster these thousands of character categories into several hundreds of similar character categories groups to ensure the neural networks can perform well. For each similar category group, we establish and train a neural network for classifying the categories in this group. When a candidate sequence is given, the candidate characters and their corresponding confidence scores are treated as the observed features. The associated neural networks for this sequence are then activated, and the network output result for the input candidate sequence is calculated and synthesized. The network outputs are sorted according to adjusted confidence scores and the result will be used to replace the original recognition result. Since the neural networks can effectively describe the nonlinear recognition process, and perform well for small-scale classification problem, this method is expected to achieve a higher improvement performance. Here, we use the offline HCCR post-processing system with lower original recognition rate as the example system.

5.4.1 Similar Character Category Clustering

To ensure the neural networks perform well, the a total of 3755 character categories in Level 1 subset of GB2312-80 character-set are clustered into several hundreds of similar character category groups to reduce the classification space for each the neural

network. The category clustering method based on similarity matrix is an effective technique which performs by means of searching the transitive closure of similarity matrix [23, 24]. The description of this algorithm is briefly given below.

Let *CC* denote the entire *N* categories. A similarity measure *R* defined on *CC* refers to a map from $CC \times CC$ to [0,1] satisfying the following properties:

(1) (Reflexive) $R(u,u)=1$, and

(2) (Symmetric) $R(u,v)=R(v,u).\ u,v \in CC.$ (5.8)

Specifically, if $CC=\{e_1,e_2,...,e_n\}$, then a matrix $S=(s_{ij})_{N\times N}$ can be defined by $s_{ij}=R(e_u,e_v)$. Customarily this matrix *S* is called a similarity matrix. The similarity matrix is reflexive ($s_{ij}\geq 0, s_{ii}=1$) and symmetric ($s_{ij}=s_{ji}.\ i,j\in CC$), but does not necessarily satisfy the fuzzy transitive condition $s_{ij}\geq \vee_k(s_{ik}\wedge s_{ij})$, where $\wedge$ and $\vee$ stand for max and min operation respectively. As we know, the transitive condition is indispensable for category clustering. Thus, the similarity matrix should be transformed into its equivalence matrix that simultaneity satisfying reflexive, symmetric and transitive condition. The minimum equivalence matrix of this similarity matrix is defined as its transitive closure, denoted by $TC(S)=(t_{ij})_{N\times N}$. Usually *TC(S)* can be obtained by searching for such an integer *k* so that S^k satisfies the equivalence condition of the similarity matrix *S*.

The searching for the transitive closure is shown in **Figure 5.3**.

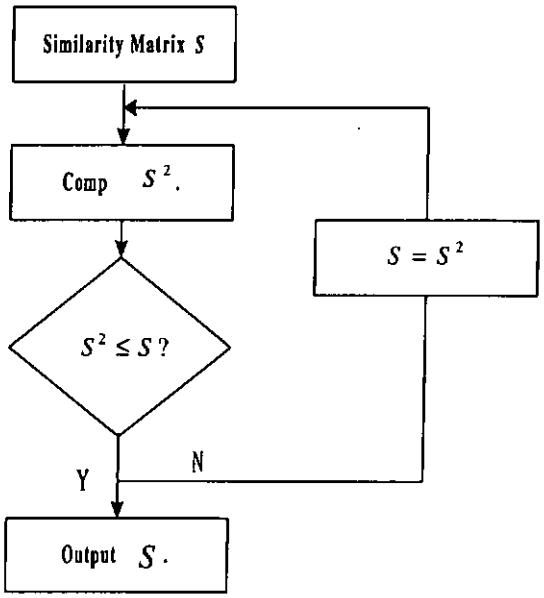


Figure 5.3 The Flow of Searching For the Transitive Closure of Similarity Matrix

In this search procedure, the multiplication for computing S^2 is defined as

$$s_{ij}^2 = \vee_k (s_{ik} \wedge s_{kj}), \; i, j = 0, 1, \ldots, n. \tag{5.9}$$

Once the transitive closure *TC(S)* is obtained, the *N* categories $\{e_1, e_2, \ldots, e_n\}$, can be clustered into several groups in terms of the criterion that e_i and e_j belong to the same cluster if and only if t_{ij} is equal to or bigger than a given threshold α.

Employing this technique to cluster the similar character categories into groups, the first problem is the generation of the similarity matrix *S*. Let $CC = \{e_1, e_2, \ldots, e_{3755}\}$, denote a total of *N=3755* character categories in GB2312-80 character-set, and each character category is identified by a 3755-dimensional feature vector $F_j (j = 1, \ldots, 3755)$. The value of each feature item is based on the similarity score of character category pairs. $F_j = 0$ if and only if the character category pair doesn't appear in either of their corresponding confusing character sets. For the non-zero feature item, the similarity scores of the

character category pairs are normalized and averaged to satisfy the reflexive and symmetric property. Suppose e_i is the description vector for the character category C_i and F_{ij} is the feature item for C_i and character category C_j. Character category C_j appears in the confusing character set of category C_i with similarity $S(C_j,C_i)$ as well as $S(C_i,C_i)$ is the similarity parameters of recognizing C_i itself by the offline HCCR engine. $S(C_i,C_j)$ and $S(C_j,C_j)$ are corresponding parameter which be can found in the confusing character set of category C_j. Then, F_{ij} is calculated as

$$F_{ij} = \begin{cases} 0, & \text{wh en } C_i \textit{ and } C_j \text{ is not a confusing character pair} \\ \frac{1}{2}\left(\frac{S(C_j,C_i)}{S(C_i,C_i)} + \frac{S(C_i,C_j)}{S(C_j,C_j)}\right) & \text{Otherwise} \end{cases} \tag{5.10}$$

The obtained sparse feature matrix S is reflexive and symmetric and is regarded as the similarity matrix of *CC*. It is easy to check whether the similarity matrix satisfies the transitive condition. If not, then computing S^2, S^4, and so on until the transitive closure *TC(S)* is obtained. In this way, the entire 3755 character categories can be clustered into several hundreds of groups according to a given threshold α. One may observe that the number of groups depends on the value of α. All the categories will constitute one group if α is set to 0 whereas each category will form a group if α is selected as 1. When α changes from 0 to 1, the number of groups will increase from 0 to *N*. Therefore the selection of the value α is very important which will influence the clustering performance. To find out an appropriate value of α so that the category clustering is "reasonable", one should pay attention to three evaluation indexes of the clustering, namely intra-similarity, inter-similarity, and average group size.

Intra-Similarity. For a given clustered group CL with r categories, the intra-similarity of CL is defined as

$$SM_{CL} = \frac{2}{r(r-1)} \sum_{m,n \in CL(m<q)} R_{mn}, \tag{5.11}$$

which describe the similarity among the members in a group.

Suppose S is clustered into m_α groups $\{CL_1, CL_2, ..., CL_{m_\alpha}\}$ when the threshold is set to α, the intra-similarity of S is defined as the average intra-similarity of all the m_α groups,

$$SM_{S-Intra}(\alpha) = \frac{1}{m} \sum_{j=1}^{m} SM_{CL_j}. \tag{5.12}$$

The value of SM_{CL} and $SM_{S-Intra}$ are in [0,1], and a large value of $SM_{S-Intra}(\alpha)$ means the members in each group are close to each, i.e., a better performance of clustering. One may observe that with an increasing α, the number of categories in a group will decrease and $SM_{S-Intra}(\alpha)$ will increase.

Inter-similarity. For any pair of clustered groups CL_1 and CL_2, we define the inter-similarity as

$$SM_{CL_1,CL_2} = \frac{1}{r_1 \cdot r_2} \sum_{m \in CL_1, n \in CL_2} R_{mn}, \tag{5.13}$$

which describes the similarity between these two groups, where r_1 and r_2 are numbers of category in CL_1 and CL_2 respectively. Suppose S is clustered into m_α groups $\{CL_1, CL_2, ..., CL_{m_\alpha}\}$ when the threshold is set to α, the inter-similarity of S is defined as the average inter-similarity among all m_α groups,

$$SM_{S-Inter}(\alpha) = \frac{1}{m(m-1)} \sum_{1 \le i < j \le m} SM_{CL_i, CL_j} . \tag{5.14}$$

The value of SM_{CL_i,CL_j} and $SM_{S-Inter}$ are in [0,1], and a small value of $SM_{S-Inter}(\alpha)$ means the better performance of clustering. One may observe that with an increasing α, $SM_{S-Inter}(\alpha)$ will also increase.

Group Size. One may notice that the requirement on the value α for a better clustering performance based on intra-similarity and inter-similarity respectively are just the opposite. Therefore, a trade-off for these two evaluation indexes for a "reasonable" clustering is needed. Simultaneously, to ensure the neural networks for similar character category groups perform well, the average size of similar character group should be taken into account. Suppose S is clustered into m_α groups $\{CL_{1_\alpha}, CL_{2_\alpha}, ..., CL_{m_\alpha}\}$ with the size of each group being $\{n_{1_\alpha}, n_{2_\alpha}, ..., n_{m_\alpha}\}$ when the threshold is set to α. The average size of category groups for S is defined as,

$$\bar{n}_\alpha = \frac{1}{m_\alpha} \sum_{i=1}^{m} n_{i_\alpha} . \tag{5.15}$$

Considering the classification capacity of neural networks and the amount of training samples, needed an average size of category groups between 3-15 is acceptable. A series of clustering experiments for different threshold α's have been conducted to find out an appropriate α ensuring a "reasonable" trade-off among the intra-similarity, the inter-similarity and average group size. Finally, the value of α is selected as 0.47 and a total of totally 421 similar character category groups, consisting of 3 to 8

character categories with on the average of 6.1 characters in each group, are obtained. These similar character category groups account for 2568 among the total 3755 categories and the remaining 1187 categories from odd-clusters in which each cluster has only one character category.

5.4.2 Neural Networks for Confusing Character Classification

For each similar character categories group, a corresponding fully connected feed-forward neural network (BP neural network) is established, which is shown in **Figure 5.4**, for classifying these similar character categories.

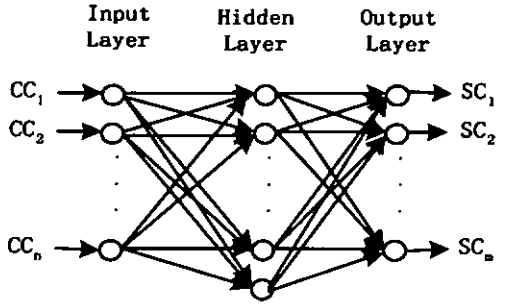


Figure 5.4 A Neural Network for Classifying Similar Character Categories

The input nodes are the confusing characters $\{ CC_1, CC_2, \ldots, CC_n \}$ of the character categories $\{ SC_1, SC_2, \ldots, SC_n \}$ in a similar character category group, and $\{ SC_1, SC_2, \ldots, SC_n \}$ are treated as the output nodes of the neural networks. The number of output nodes is determined by the size of similar character category group, ranging from 3 to 8. As for the number of input nodes, it ranges from 16 to 38. The number of hidden nodes depends on the number of output nodes and is ranges from 16 to 30.

For each neuron in this network, it outputs an activation value as a function of its net input through an activation function. Here, an unipolar sigmoid function is selected as the activation function,

$$f(x) = \frac{1}{1+e^{-x}}. \tag{5.16}$$

The back-propagation learning algorithm, which is one of the most popular and effective algorithms, is selected for training this fully connected feed-forward network. The recognition results of the image samples with respect to $SC_1, SC_2, \ldots, SC_n$ are treated as the training samples of the network. Suppose a recognition candidate sequence for input image $I(R) = ISC_i$ is $C_{i1}, C_{i2}, \ldots\ldots, C_{i.20}$. The input nodes CC_k which appear in this candidate sequence are assigned the input value of the confidence score μ_i. For the output nodes, the expected output for SC_i is set to 1 and for the other output nodes, they are set to 0. The input signals propagate forward through the network until the outputs of all the output nodes have been obtained. Compute the error value for the output layer and propagate the errors backward to update the weights. This is a learning step. After a learning step is finished, the next training pattern is submitted and the learning step is repeated until the entire *200*n* recognized candidate sequences are exhausted. If the total error is not satisfied, a new learning cycle will be initiated. When the total error is less than a pre-given threshold, the network training is terminated. These trained neural networks will be used to classify the confusing characters for a given candidate sequence.

5.4.3 Employing Neural Networks for Confusing Character Classification in Post-processing

For a given candidate character sequence as observed features, the neural networks for similar character category groups can be used to identify the most promising character as a classification result, which its flow is shown in **Figure 5.5**. A recognition candidate sequence $\{C_1, C_2, \ldots, C_i\}$ is firstly transferred to a neural network selector. Such a selector will analyze the candidate sequence and activate those "relative" neural networks for similar character categories groups. The "relative" neural networks refer to those having at least 3 matched character categories between the input nodes and the candidate sequence. For each activated neural network, a process outlined below is performed.

Step 1. Assign the input signals for the n nodes of the input layer. For each input node, once CC_j $(j=1,\ldots,n)$ matches a candidate character C_p, then its input signal is assigned as the confidence score of C_p, namely μ_{C_p}. Otherwise, the input signal is assigned zero.

Step 2. Compute the output signals for the output layer. If the output signal has a value larger than a threshold, the character categories for this output node together with its output will be recorded and transferred to the following synthesizer. In this step, some character categories corresponding to high values of output nodes, which have not already appeared in the candidate sequence, will be appended in the output of the neural networks. That means some mis-recognized characters could be recovered.

The outputs of these activated neural networks are transferred to a synthesizer, which will compute the average output results of those character categories appearing in the result of more than one activated neural networks. The obtained characters are sorted according to their neural network outputs as the final output.

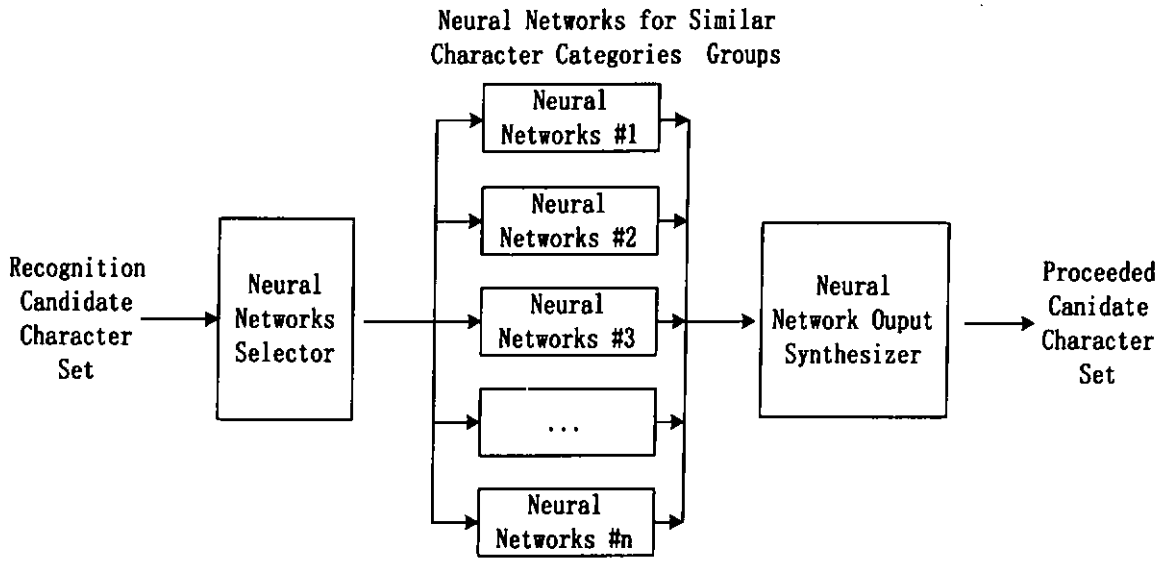


Figure 5.5 The Schematic Diagram of Employing Neural Networks for Similar Character Category Groups in Post-processing

An experiment is conducted to evaluate the performance for improving recognition accuracy by employing Noisy-Channel model and neural networks for confusing character categories. A group of testing samples consists of about 20,000 offline handwritten samples is established. The recognition accuracy rate produced by the recognizer is 81% and 92% for the first candidate and the first ten candidates respectively. The improved accuracy rate after post-processing is shown in **Figure 5.6**.

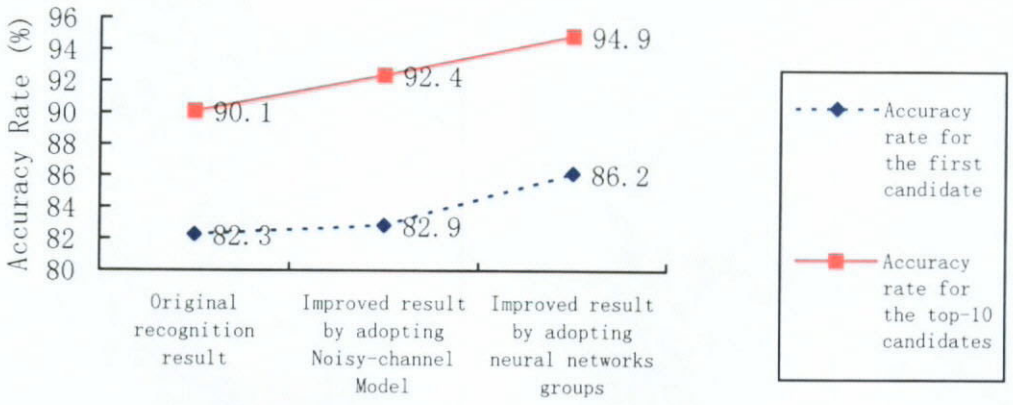


Figure 5.6 The Improved Accuracy for the Post-processing System based on Characteristics of Confusing Characters

The preliminary experiments show an average of 3.9% and 4.8% accuracy rate improvement for the first candidate and the first ten candidates respectively, which is higher than the corresponding 0.6% and 2.3% accuracy rate improvement achieved by employing the Noisy-Channel model. One may observe that after the post-processing based on confusing character set and neural networks groups, an average of 48% of the unrecognized characters are recovered. Furthermore, one may also observe that 72% of the 'expected' characters are moved onwards in the candidate sequence while 9% of those characters are wrongly moved backwards. It leads to a 3.9% overall accuracy rate improvement for the first candidate, i.e., about 50% correction rate for the mis-recognized characters can be achieved.

For the experimental result by using the presented post-processing techniques based on computational linguistic information and characteristics of confusing characters, one may observe that the dictionary-based post-processing technique has the advantage of

recovering some of the unrecognized characters by appending linguistic-prone character and removing erroneous characters in the multi-character words. The statistical-based post-processing approach is effective in removing mis-recognized characters and identify the most linguistic-plausible sentence, and post-processing technique based on characteristics of confusing characters has the advantage of being context-free, stability with regard to the recognizer, and capability to recover unrecognized characters. The hybrid post-processing system, which integrates all these techniques together for a expected higher improvement performance, will be presented in **Chapter 6**.

Chapter 6

Hybrid HCCR Post-processing System

In the previous two chapters, the post-processing approach based on computational linguistics information and characteristics of confusing characters were discussed respectively. In the first part of this chapter, **Section 6.1** briefly summarizes the characteristics of these two approaches and compares their performance on removing erroneous characters. These techniques have their own advantage in removing different type of recognition errors, and the hybrid post-processing approach, which integrates these techniques together, is motivated to achieve a higher improvement performance. In **Section 6.2**, a three-stage hybrid post-processing system is implemented for improving the performance of an online HCCR system. In the first stage of this system, the confusing character set and post-processing techniques based on characteristics of confusing characters are used to append possible unrecognized similar-shaped characters into the candidate character set and the confidence scores of the candidate characters are recomputed. Secondly, the dictionary-based post-processing technique is

conducted to identify the fragments in the candidate sentence and then append contextual linguistics-prone characters into the candidate set by means of approximate word matching. Finally, in the third stage, the techniques based on statistical language model are employed to identify a most promising sentence from the sentence hypotheses constructed from the candidate characters. Its performance is evaluated and discussed. In view of the lower original recognition accuracy of the offline handwritten Chinese characters will weaken the dictionary-based sentence fragments identification and linguistics-prone characters accession, the proposed three-stage post-processing method is modified. In the second stage of post-processing method, the dictionary-based approximate word matching is no longer identifying sentence fragments and removing erroneous characters, but appending linguistic-prone characters and binding candidate characters into word-lattice by means of approximate word matching. Its working flows and performance evaluation is given in **Section 6.3**.

6.1 Characteristics of Presented Post-processing Techniques and Motivation of Hybrid Post-processing Approach

In **Chapter 4** and **Chapter 5**, the dictionary-based post-processing approach, the statistical-based post-processing approach and the post-processing approach based on characteristics of confusing characters were presented. These approaches are proven effective in removing erroneous characters in the recognition result. The experimental results of employing these approaches in online and offline HCCR post-processing system are analyzed and their characteristics are shown.

The dictionary-based post-processing approach, which adopts a top-down strategy, performs sentence analysis to identify the sentence fragments and remove the erroneous characters. This approach could effectively remove the errors encountered in the multi-characters words, but could not remove those erroneous characters when they are single-character words. Employing approximate word matching to generate word hypotheses and evaluate these word hypotheses for identifying a most linguistic-prone one as the corrected result, this approach can remove both the mis-recognized errors and unrecognized errors in the original recognition result. Even some errors due to the use of wrong characters by the writer in the original recognition result can be removed. For example, "闻名遐迩" is a Chinese phrase, but many people write it in a wrong character formation as "闻名遐尔". This erroneous occurs in the Microsoft Chinese Input Method. Employing the proposed dictionary-based approximate word matching technique could replace the erroneous character formation by the correct one, thus some of the errors caused by wrong use of characters may be corrected. The dictionary-based approach performs well when the accuracy of original recognition result is high, but if the original accuracy is low, the performance of the dictionary-based post-processing approach will be obviously weakened due to the fact that a large number of erroneous characters may undermines the sentence fragments identification and word hypotheses generation. Furthermore, the dictionary-based approach cannot ensure a global optimal result will be identified.

With a bottom-up strategy, the statistical-based post-processing approach is designed for identifying a global optimal result with maximum linguistic likelihood that

is evaluated by the statistical information among the linguistic units. This approach is proven effective in removing mis-recognized characters. Evaluating both the probability of a character as a single-character word and the probability of a character as part of a multi-character word, employing this approach can remove the erroneous character encountered in multi-character words and single-character words. Simple statistical-based mathematical description of linguistic phenomena and corpus-based parameter training ensures this approach performs well in removing recognition error. If it is assumed that the correct characters can be selected from the candidate set and then form a linguistic plausible sentence, the statistical-based approach has a major weakness that employing this approach cannot remove the unrecognized character errors. Therefore, its correction performance is also influenced by the accuracy of the original recognition result, especially by the accuracy of the first n-characters. But this is not as obvious as the case for the dictionary-based approach.

The post-processing approach based on the characteristics of confusing characters makes use of other available information besides contextual linguistic information. Erroneous characters are collected to construct the confusing character set for each character category. Such a confusing character set can be employed to evaluate a given candidate sequence and identify the most promising one. This approach can be employed to remove the errors for each individual recognition sequence, thus its correction performance is the same for characters in the single-character words or in the multi-character words. Furthermore, the correction performance of this approach does not depend on the accuracy of the original recognition result. The importance of this

approach cannot be ignored especially in practical applications. For example, some writers always write a particular character in their own special form and order of strokes that could lead to recognition errors. These errors can be recovered after analyzing the experimental recognition result to construct a confusing character sequence and then employing this confusing character sequence in post-processing.

A brief comparison of these post-processing approaches is shown in **Table 6.1**.

	Dictionary-based Approach	Statistical-based Approach	Approach based on Characteristics of Confusing Characters
Mis-recognized characters removing	Good	Better	Good
Unrecognized character removing	Good	No	Better
Removing errors in single-character word	No	Better	Good
Removing errors in multi-characters word	Good	Good	Good
Influenced by the original recognition accuracy	Strong	Obvious	Inconspicuous
Utilization of contextual linguistic information	Good	Better	No

Table 6.1 A Comparison for the Characteristics of Post-processing Approaches

From the above comparison, one may observe that the statistical-based post-processing approach has a better overall performance over the other two approaches in erroneous characters removal because the utilization of contextual information ensures a global optimal result to be identified, but its capacity is often weakened by the unrecognized characters. The experiment has shown that both the dictionary-based approach and the approach based on characteristics of confusing characters can be employed to recover some of the unrecognized characters. Employing different information, i.e., linguistics information and confusing characters, the recovered

unrecognized characters for these two approaches are not the same. Naturally, the reduced unrecognized character errors will enhance the correction performance of the statistical-based approach. Therefore, constructing a hybrid post-processing approach is motivated to make use of the advantages of these methods for a higher recognition accuracy improvement. Its major strategy is to employ dictionary-based post-processing approach and the approach based on characteristics of confusing characters to recover some of the unrecognized characters and remove some mis-recognized characters. Then the statistical-based post-processing approach is applied to the modified candidate character set for identifying the most linguistic promising optimal result. In view of the fact that correction by the approach based on characteristics of confusing characters is independent of the encountered context, this approach is used firstly. Then the dictionary-based approach is used to further recover the unrecognized characters and remove mis-recognized characters. Finally, the statistical-based approach is invoked to identify the optimal result. However, in any practical application, since there is an obvious difference on the recognition accuracy achieved by online and offline HCCR system, different integration methods may be considered. The following two sections present these two cases separately.

6.2 Hybrid Post-processing System for Online HCCR System

A serial three-stage hybrid post-processing system is established for improving an online HCCR system. The schematic diagram for a complete recognition system with post-processing is shown in **Figure 6.1**.

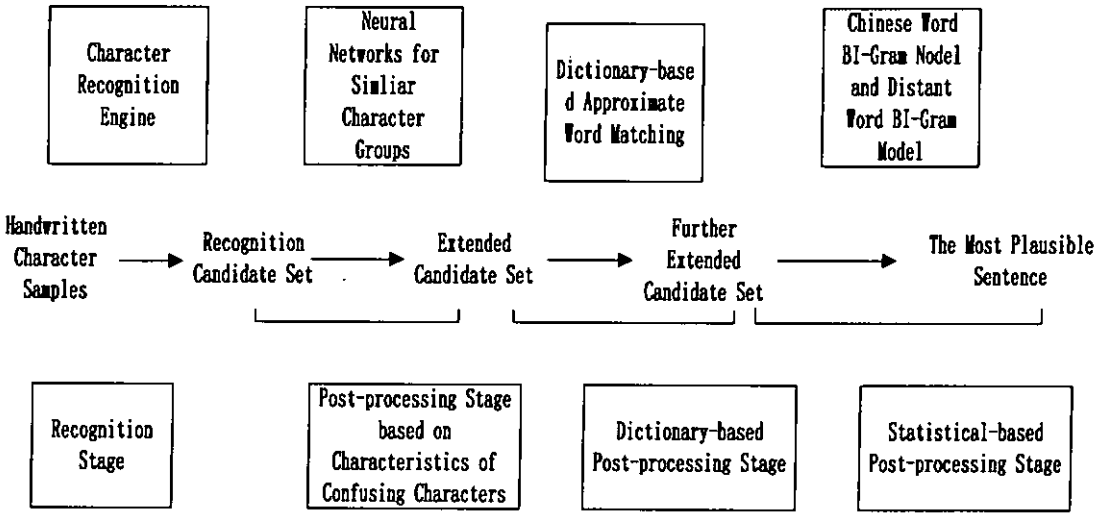


Figure 6.1 The Schematic Diagram of the Hybrid Post-processing System for Online HCCR System

In the first stage, the confusing character set and the techniques based on characteristics of confusing characters are utilized to extend the candidate character set by appending the similar-shaped error-prone characters and adjust the confidence scores of candidates. The method based on neural networks for similar character category groups that was described in **Chapter 5** is employed here. A given candidate character sequence $\{ C_1, C_2, \ldots, C_i \}$ is firstly transferred to a neural network selector. This neural network selector will activate the "relative" neural networks that have at least 3 matched characters between the input nodes of the neural networks and the candidate character sequence. For each activated neural network, the original candidate characters and their associated confidence scores are assigned to the input nodes, and the output result for the output layer is computed. Then output of these activated neural networks are

transferred to a synthesizer for the purpose of averaging the output result of those character categories which appear in the result of more than one activated neural networks. The obtained candidate characters are combined and then sorted according to their neural network outputs as the modified candidate set. During this stage, some unrecognized characters are recovered and the confidence scores of candidate characters are adjusted.

Secondly, the dictionary-based post-processing stage is conducted to further extend the candidate set. The sentence hypothesis consists of the first candidate characters is segmented into word sequence using the word segmentation algorithm based on word BI-Gram which were presented in **Chapter 4**. After the word segmentation, the continuous individual character strings which consist of either a minimum of any three characters or a minimum of two characters in which one of them cannot form a word by itself, are regarded as sentence fragments. Next, the BI-directional approximate word matching is applied to these fragments in order to find out linguistic-prone characters.

For the approximately matched word entries, the confidence scores of the unmatched characters will be adjusted according to the similarity between the approximately matched word and word entries in the dictionary if the approximately matched character encountered in the candidate set is not ranked first. Otherwise, the approximately matched characters will be appended into the candidate set. During this stage, the candidate set is further extended by appending linguistic-prone characters. Meanwhile, the bounded words will form a word-lattice which will be used in the following word-based statistical post-processing approach.

Finally, in the statistical-based post-processing stage, the Chinese word BI-Gram model and the distant Chinese word BI-Gram model are respectively employed to identify the most plausible sentence constructed by the characters from the extended candidate set. Their performances are evaluated and compared with some popular post-processing techniques.

The large online handwritten Chinese test samples presented in Chapter 3 with 91.8% recognition accuracy for the first candidate and 95.2% for the first ten candidates, are used to test our hybrid post-processing system. The recognition results for the handwritten samples are transferred to the post-processing system. After the first stage, the one based on characteristics of confusing characters, improve the accuracy rates for the first candidate and the first ten candidates to 94.9% and 97.4% respectively. That means about 46% unrecognized characters are recovered. Meanwhile, 70% of the 'expected' characters are moved onwards in the candidate sequence while 7% of those characters are wrongly moved backwards. It leads to a 3.1% accuracy improvement for the first candidate, i.e., about 37% correction rate for the mis-recognized characters is achieved.

The dictionary-based post-processing stage is worked next. The approximate word matching is applied to the detected sentence fragments, and the linguistic-prone characters are identified. After this stage, a further 1.2% improvement for the first ten candidates is achieved. The accuracy of the first ten candidates for the twice-extended candidates set is 98.6%, which means 71% of the unrecognized characters are recovered.

As for the accuracy of first candidate, it rises to 95.4%. Such an enhanced candidates set helps the statistical-based post-processing achieve a good performance improvement.

In the last stage, the post-processing technique based on word BI-Gram model and distant word BI-Gram model, together with the popular character BI-Gram model and word UNI-Gram model are respectively employed. Their recognition accuracy performance improvements are shown in **Figure 6.2**.

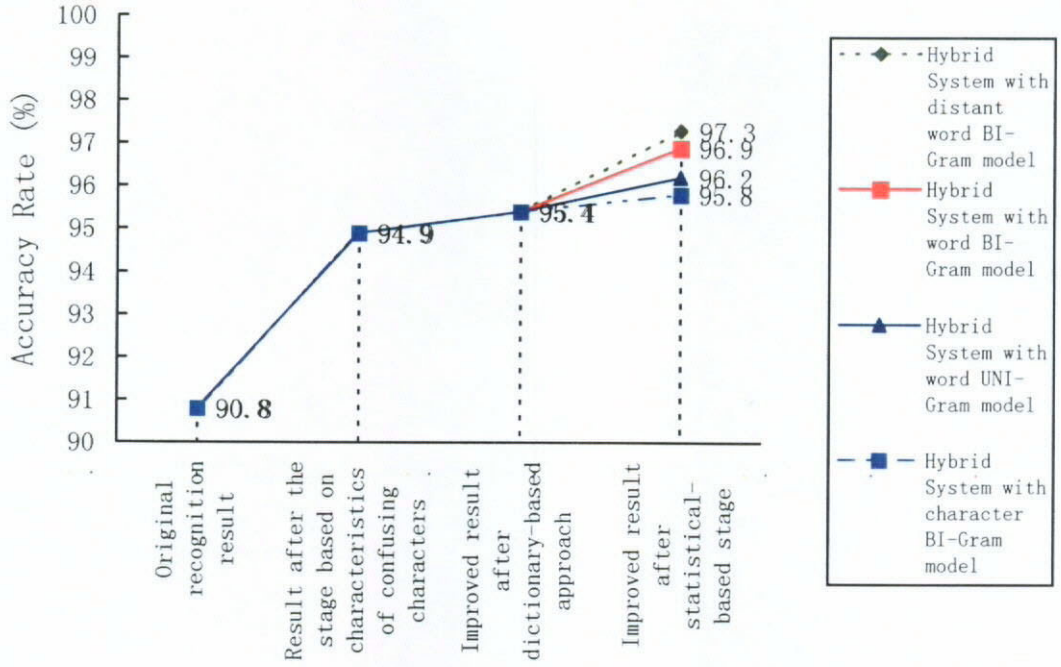


Figure 6.2 The Improved Accuracy of the First Candidate for the Hybrid Post-processing Systems with Different Statistical Language Models

Using the method based on distant word BI-Gram model, the 97.3% accuracy rate for the first candidate is achieved, which is higher than the 96.9% accuracy achieved by employing word BI-Gram model, 95.8% accuracy achieved by employing character BI-Gram model, or 96.2% accuracy achieved by employing word UNI-Gram model. One may observe that the final 96.9% accuracy is already higher than the original first ten

candidates. This result proves that our hybrid post-processing strategy can effectively raise the theoretical enhancement limit for the statistical-based post-processing technique, both for our proposed language models and the popularly adopted ones.

6.3 Hybrid Post-processing System for Offline HCCR System

The proposed hybrid post-processing system for online HCCR system is proven effective, but when we apply this system to the offline HCCR system, the correction performance especially for the dictionary-based stage is lower because the obvious lower recognition rate for the first candidate will unavoidably weaken the sentence fragments identification. Therefore, the second stage of the proposed hybrid post-processing is modified to fit for offline HCCR system. In the new three-stage processing system, the dictionary-based post-processing will not detect sentence fragments but the approximate word matching is directly applied to the candidate set in order to recover the unrecognized characters, adjust the confidence scores of candidates and bind candidate characters into word-lattice.

Suppose $C = C_i C_{i+1} \cdots C_{i+k}$ is a $k+1$-character string starting from the *i-th* position in the sentence, and W_{i-1} W_{i+1} are the adjacent matched words of this string. Then the word hypothesis is generated as follows:

1. If in the dictionary there exists a word entry W that matched C, then C is bound as a word hypothesis.

2. If in the dictionary, there exists any word W that contains W_{i-1} as its prefix, then W_{i-1} will be bound with its subsequent characters in C to construct a new word hypothesis with the same length as W.

3. For each C_j, $j=i,\ldots, i+k-1$, append the subsequent characters in C to form a word hypothesis that matches a word entry with C_j as its prefixes. Repeat this until C_{i+k-1} is done.

4. If in the dictionary, there exists any word W that contains W_{i+1} as its suffix, then W_{i+1} will be bound with its preceding characters in C to form a word hypothesis of the same length as W.

Then **Equation 4.7** is employed here to evaluate the similarity between the word entries in the dictionary and the approximately matched character string. The ones with the similarity value larger than a threshold are regarded as linguistic-prone words and the unmatched character is regarded as linguistic-prone character. Once this character is not found in the candidate set, it is appended into the candidate set as a part of the approximately matched word. Otherwise, its confidence score will be adjusted by **Equation 4.8**.

Generally speaking, the hybrid post-processing system for offline HCCR system differs from the one for online HCCR system in the dictionary-stage. Unlike the post-processing system for online HCCR system, the sentence hypothesis is analyzed to detect sentence fragments which are then presented to the approximate word matching. For the offline HCCR system, the approximate word matching is applied to the entire

candidate character set to generate linguistic-prone words without sentence fragment identification. This method leads to an increased computational cost but more linguistics-prone words are evaluated and more linguistic-prone characters are discovered. Furthermore, during the linguistic-prone word's evaluation, the candidate characters are bound in a word-lattice, which is employed in the statistical-based post-processing stage.

The offline handwritten test samples are used to evaluate the performance of this three-stage post-processing system. In the first experiment, the performance of dictionary-based post-processing stage is evaluated and compared with the one by employing sentence fragment detection and approximate word matching. The results are given in **Figure 6.3**.

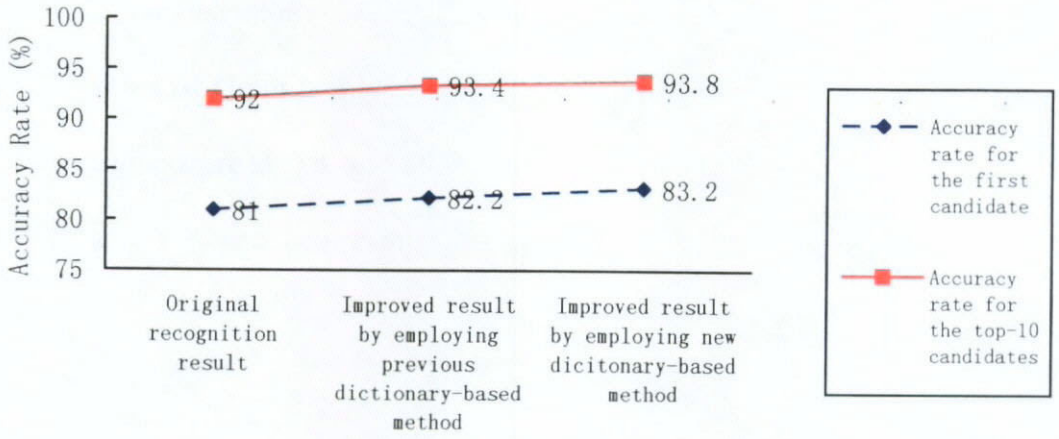


Figure 6.3 The Improved Accuracy for the Dictionary-based Post-processing Methods

From the observed experimental results, one notices that for the test samples with 81% accuracy rate for the first candidate and 92% for the first ten candidates, a 2.2% and 1.8% accuracy improvement for the first candidate and the first ten candidates are respectively achieved by employing our current method while the previous dictionary-

based method achieves the improvement of 1.2% and 1.4% respectively. The lower original recognition result obviously affects the accuracy of sentence fragment identification which is important to our previous dictionary-based matching method.

The second experiment is conducted to evaluate the performance of a complete hybrid post-processing system for offline HCCR. Different dictionary-based post-processing methods are employed. Word BI-Gram model and distant word BI-Gram model are respectively employed in the third post-processing stage. Their performance improvements are evaluated and shown in **Figure 6.4**. By employing the new three-stage hybrid post-processing method with the statistical-based component based on distant word BI-gram, the accuracy for the first candidate enhances from 81% to 92.1% when the original accuracy of the first ten candidates is 92%, as well as the improved accuracy by the previous three-stage hybrid post-processing system is 91.3%. Respect to employing word BI-Gram in the statistical-based component, the improved accuracy for the first candidate are 90.3% and 88.9% respectively for the new three-stage hybrid post-processing system and the previous one. The experimental results have shown that the new dictionary-based post-processing stage is effective in enhancing the recognition accuracy for the first ten candidates to ensure that the statistical-based post-processing method obtains a higher recognition accuracy improvement. Meanwhile, the statistical-based post-processing method based on word BI-Gram model and distant word BI-Gram model is effective in the optimal result identification when the original recognition accuracy is lower.

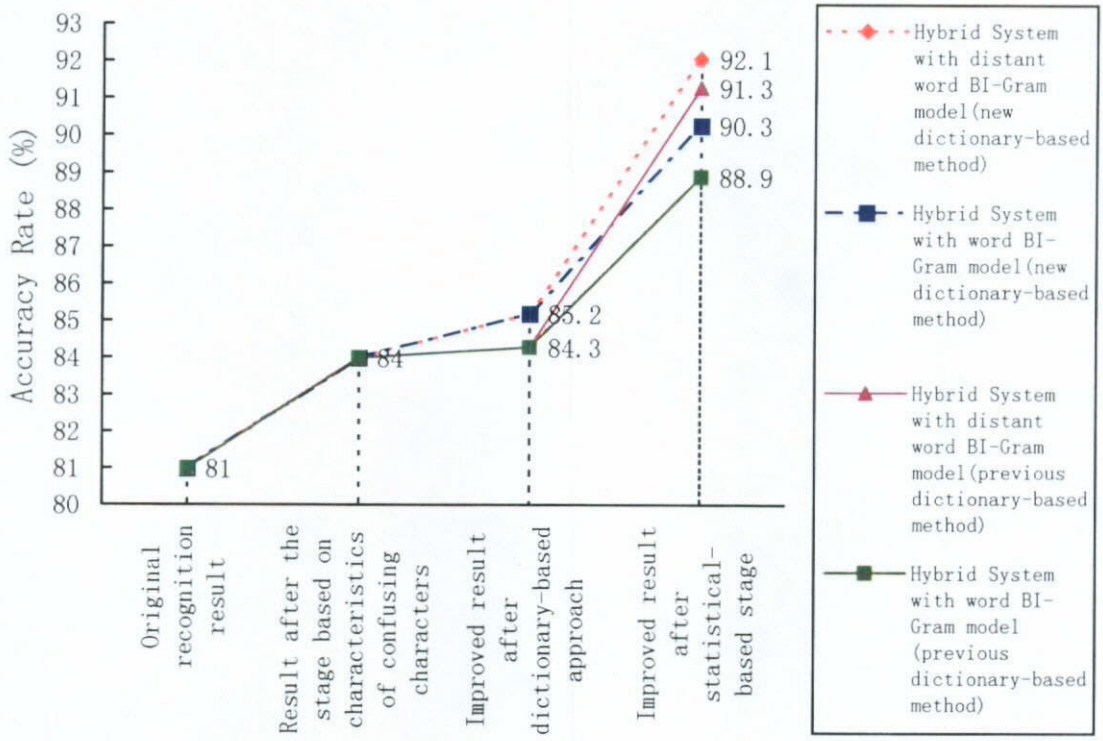


Figure 6.4 The Improved Accuracy of the First Candidate for the Hybrid Post-processing Systems for Offline HCCR

In this Chapter, the advantage and weakness of our proposed post-processing method are analyzed. Integrating the approach based on characteristics of confusing characters, dictionary-based approach and statistical-based approach together, our hybrid post-processing has shown good performance in enhancing the performance of the HCCR system. In view of the different character recognition accuracy for the online and offline systems, different integration methods are employed and the result is good.

Chapter 7

Conclusion and Future Research

This chapter presents the conclusion of this thesis and the direction of future research. Section 7.1 presents the main components and the major achievements of this thesis. In Section 7.2, the directions of our further research on the post-processing techniques for Handwritten Chinese Character Recognition system are proposed.

7.1 Summary and Major Achievements of This Thesis

The major objective of this thesis is to develop a post-processing system for effectively improving the accuracy of both online and offline HCCR system. Recognition errors are categorized into mis-recognized characters and unrecognized characters. Post-processing techniques are then developed to remove these two kinds of errors.

The popular post-processing approach based on computational linguistic information is first investigated. The dictionary-based technique segments the sentence hypothesis, which consists of each first candidate character for the handwritten samples, into word sequence and then attempts to locate the sentence fragments. The dictionary-based approximate word matching is applied to these sentence fragments in order to generate word hypotheses and evaluate their linguistic probabilities. The linguistic-prone word hypotheses are used to substitute the corresponding character string in the sentence fragments. In this way, some unrecognized characters and mis-recognized characters are removed. The experimental result has shown that this approach performs well when the original character recognition accuracy is high. Especially, this approach could effectively remove recognition errors encountered in multi-character words.

Another popular post-processing approach based on computational linguistic information is statistical-based. In this approach, the candidate characters in the candidate set are bound into words and a word-lattice is formed, then the statistical language model is employed to identify an optimal word combination in the word-lattice to form a sentence with the maximum linguistic likelihood as the output result. Considering the characteristics of Chinese language and the requirement of the computational cost, a Chinese word BI-Gram model is employed here as the statistical-based language model for evaluating the probability of sentence hypotheses that consist of words in the word-lattice. This method is proven effective to improve the recognition accuracy. Based on this work, the Chinese word BI-Gram model is extended to a distant Chinese word BI-Gram with a maximum distance of 3 in order to

obtain the linguistic description capacities for the long-distance semantic restrictions among Chinese sentences with an acceptable enlarged parameter space and computational cost. A higher recognition improvement performance is achieved by employing this statistical language model in the post-processing system. The statistical-based post-processing approach is proven effective in removing the mis-recognized characters and identifying an optimal sentence hypothesis. It could remove the errors encountered in both single-character words and multi-character words. However, the unrecognized characters cannot be removed by individually employing this post-processing approach, thus this approach has a theoretical upper limit of recognition accuracy improvement.

The post-processing approach based on the characteristics of confusing characters is then studied. This approach is developed to make use of characteristics of confusing characters produced by the recognizer, besides the computational linguistic information, to remove recognition errors. The popular post-processing method based on linear statistical-based Noisy-channel model is implemented, and this method has advantages different from the proposed methods based on computational linguistic information. In view of the recognition process being a complex nonlinear one, the post-processing method based on neural networks groups for similar character categories are developed. Regarding the obtained candidate sequence and its confidence score of a handwritten sample as its observed feature, neural networks are employed to classify the most promising character. To ensure the neural networks have good classification performance, the character categories in GB2312-80 character-set are clustered into

several hundreds of similar-shaped category groups. Then a group of neural networks for these category groups are built and trained. These neural networks are employed to identify the most-promising one from the candidate sequence of a character supported by the recognizer. This approach could remove both the mis-recognized and unrecognized characters. Due to the fact that no contextual linguistic information is utilized in this approach, a global optimal solution for the recognition result could not be obtained.

After analyzing the advantages and weaknesses of these post-processing methods, a three-stage hybrid post-processing system is developed, which integrates these methods together, to obtain a higher recognition accuracy improvement performance. In the first stage of this system, the method based on characteristics of confusing characters is conducted to remove erroneous characters and recover some of the unrecognized characters through appending similar-shaped characters into the candidate set. Secondly the dictionary-based method is processed to recover some of the unrecognized characters by appending linguistic-prone characters, and bind the candidate characters into words to construct a word-lattice. In the last statistical-based stage, the proposed distant word BI-Gram model is employed to identify a sentence having the maximum linguistic likelihood probability among the words in the word-lattice. Benefiting from the utilization of two kinds of available information, i.e. contextual linguistic information and characteristics of confusing characters produced by the recognizer, and the utilization of advantages of each post-processing method on removing mis-recognized characters and unrecognized characters, our hybrid post-processing

approach has shown a better recognition accuracy performance improvement for the HCCR systems.

The achievements in this thesis are briefly listed as follows:

Corpus

As a fundamental work for employing Chinese linguistic information in post-processing and other Chinese language processing research, a large-scale Chinese text corpus with approximately 890 millions of Chinese characters are built. 40MB manually word-segmented text corpus is built for extracting linguistic statistics. Those sentences which have segmentation ambiguities are collected to build a 3.1MB ambiguity sub-library for supporting Chinese word segmentation research. Strengthened by the word statistical information and ambiguity sub-library, the automatic word segmentation system is employed to segment 500 MB text files into word sequences.

Dictionary

A dictionary with 101,644 word entries is built. Each word entry in this dictionary consists of the word, the corresponding unique word-ID sorted by the GB code, and its frequency in the corpus. To speed up word lookup, an index table is established for this dictionary.

Word-based Statistics Database

To support statistical-based word segmentation and HCCR post-processing, the word-based statistics database is constructed. The segmented text corpus is utilized to extract the word-based statistics by filtering and compressing the word co-occurrence statistics from the more than 100MB automatically segmented corpus, a 14MB word BI-Gram statistics database is built. In a similar way, the statistics database for distant word BI-Gram model are built. For words with distance of 2 and 3, the obtained statistics databases have the size of 11MB and 9MB respectively.

Word Segmentation

A BI-directional maximum word segmentation algorithm, strengthened by ambiguity removal component based on word BI-Gram model, is developed. For a test sample of about 40MB text files, this word segmentation algorithm achieves 98.7% accuracy. An un-registered word identification algorithm is developed to extract new words encountered in the training text.

Dictionary-based Post-processing Method

A dictionary-based post-processing method is developed. It is based on sentence fragment detection and approximate word matching for erroneous character removal. For a test sample for the online HCCR system with original recognition accuracy of 91.8% for the first candidate and 95.2% for the first ten candidates, this method achieves a 3.37% recognition accuracy improvement. For a test sample of offline

HCCR system with original recognition accuracy of 81% and 92% for the first and the first ten candidates respectively, a 1.2% recognition accuracy improvement is achieved.

Statistical-based Post-processing Method

Two statistical language models, namely word BI-Gram and distant word BI-Gram model, are established and employed in our post-processing system to identify a most promising sentence constructed by the characters in the candidate set. For a test sample for online character recognition system with accuracy of 91.78% for the first candidate and 95.2% for the first ten candidates, an average of 3.9% recognition rate improvement is obtained by employing word BI-Gram model in a post-processing system. Employing distant word BI-Gram model, a 7.2% recognition accuracy improvement is achieved for the testing offline character sample with 81% recognition accuracy for the first candidate and 92% for the first ten candidates.

Confusing Character Set

The large-scale recognition experimental results are used to establish the confusing character set in which similar characters for each character category are recorded. For the online HCCR system, the average number of confusing characters for each character category is 6. For the offline HCCR system, this number increases to 10.

Similar Character Category Group

Through searching for the transitive closure of the similarity matrix associate with the confusing character set, the 3755 character categories in GB2312-80 Level-1 character-set are clustered into 421 similar character category groups, that consist of 3 to 8 character categories with an average of 6.1 characters in a group.

Post-processing Method based on Characteristics of Confusing Characters

A Post-processing method based on characteristics of confusing characters is developed to enhance the recognition accuracy of the recognizer. The neural networks for the similar character category groups are employed in the post-processing system to identify the most promising one for a given candidate sequence. For about 20,000 offline character samples, the accuracy for the first candidate increases from 82.3% to 86.2% by employing this post-processing system. As for the accuracy of the first ten candidates, it increases from 90.1% to 94.9%.

Hybrid Post-processing System

Integrating the dictionary-based method, the statistical-based method and the method based on characteristics of confusing characters into a three-stage hybrid post-processing system, a higher recognition accuracy improvement is achieved. For the online character sample of 90.8% and 95.4% accuracy for the first candidate and the first ten candidates, the improved recognition accuracy is 97.3%. When the hybrid post-processing system is employed to improve the offline HCCR system with the first

candidate's accuracy of 81% and first ten candidates' accuracy of 92%, the overall recognition accuracy for the first candidate is raised to 92.1%.

	Recognition Accuracy Improvement for the Online HCCR System (%)		Recognition Accuracy Improvement for the Offline HCCR System (%)	
	First Candidate	First Ten Candidates	First Candidate	First Ten Candidates
Dictionary-based Post-processing Method	3.4	2.4	2.2	1.8
Statistical-based Post-processing Method (Word BI-Gram Model)	3.9	0	2.8	0
Statistical-based Post-processing Method (Distant Word BI-Gram Model)	4.1	0	7.2	0
Post-processing Method based on Characteristics of Confusing Characters (Noisy-Channel Model)	1.5	2.0	2.2	2.5
Post-processing Method based on Characteristics of Confusing Characters (Neural Networks Group)	3.1	2.4	3.9	4.8
Hybrid Three-stage Post-processing System (with Word BI-Gram Model)	5.1	3.6	10.3	6.2
Hybrid Three-stage Post-processing System (with Distant Word BI-Gram Model)	5.5	3.6	11.1	6.2

Table 7.1 Recognition Accuracy Improvement Performances of Post-processing Methods

In **Table 7.1**, the recognition accuracy improvement performance of proposed post-processing methods for online and offline systems are compared.

7.2 Future Research

The performance of our proposed hybrid post-processing system for HCCR recognition system is satisfactory. To further enhance the performance of recognition accuracy improvement, some directions of further research on the post-processing technique for HCCR system are suggested.

Cache-based Language model for Global/Local Linguistic Statistics

The proposed post-processing methods utilizing statistical linguistic information are all based on global linguistic statistic such that the word frequency and word co-occurrence frequency are the ones encountered in the corpus. However this global linguistic statistics sometimes could be misleading because some words and word pairs encountered frequently in the current processing context may have low occurrence frequency in the corpus. A cache-based language model [35] is expected to collect local linguistic statistics under the current processing context in order to strengthen the global linguistics statistic for a better linguistic description capacity.

The Rule-based/Statistical-based Hybrid Language Model

There are mainly two types of language model that could be employed in language processing. One type is based on dictionary and linguistic rules. There are many

difficulties to use this type of language model to process large-scale real texts since these may exist many un-registered linguistic rules. Another type is based on linguistic statistics, in which the popular N-Gram model is used. This statistical-based language model is affected by three major problems: long-distance restriction, recursive nature, and partial language understanding. In order to take the advantage of these types of language model, the new hybrid language model which makes use of grammatical rules or semantic rules to improve the n-gram technique, is expected to provide a better linguistic description capacity.

Employing Language Model in Character Image Segmentation for Real Recognition Process

The aim of this research is to study how post-processing techniques could be used to identify the most promising candidate characters which are output by a character recognizer. However the recognition errors caused by incorrect segmentation of the character images are ignored. In a real recognition system, this problem does widely exist, especially for the offline handwritten character recognition systems. The language model and the post-processing techniques should be integrated into character image segmentation method to enhance the recognition accuracy of the real recognition process.

Integrating Post-processing Techniques into Recognition System

In our proposed work of this thesis, the post-processing system is designed for processing the candidate character sets produced by the recognition system. Integrating post-processing technique into a recognition engine is expected to reduce the searching space and speed up the pattern classification for the recognition of the character samples.

We believe that these further research efforts can lead to a higher recognition accuracy improvement for large-scale real document processing.

List of Publications

1. Xu R. F., and Yeung D. S. "Experiments on the use of corpus-based word BI-Gram in Chinese word segmentation". In *Proceedings of IEEE International Conference on System, Man and Cybernetics 1998*, Vol. V, pp.4222-4227, San Diego, (1998)

2. Xu R. F., Yeung D. S., and Wang X. L. "A hybrid post-processing approach for handwritten Chinese character recognition". In *Proceedings of the International Conference on Machine Translation and Computer Language Information Processing*, Vol.1, pp.152-158, Beijing, (1999)

3. Xu R. F., Yeung D. S., and Shu W. H. "Using confusing character, dictionary matching and word BI-Gram language model for improving handwritten Chinese character recognition". In *Proceedings of the International Conference on Artificial Intelligence*, Vol. III, pp.1271-1277, Las Vegas, (2000)

4. Xu R. F., Yeung D. S., Shu W. H. and Liu J. F. "A hybrid post-processing system for online handwritten Chinese character recognition", submitted to *International Journal of Pattern Recognition and Artificial Intelligence*, (2001)

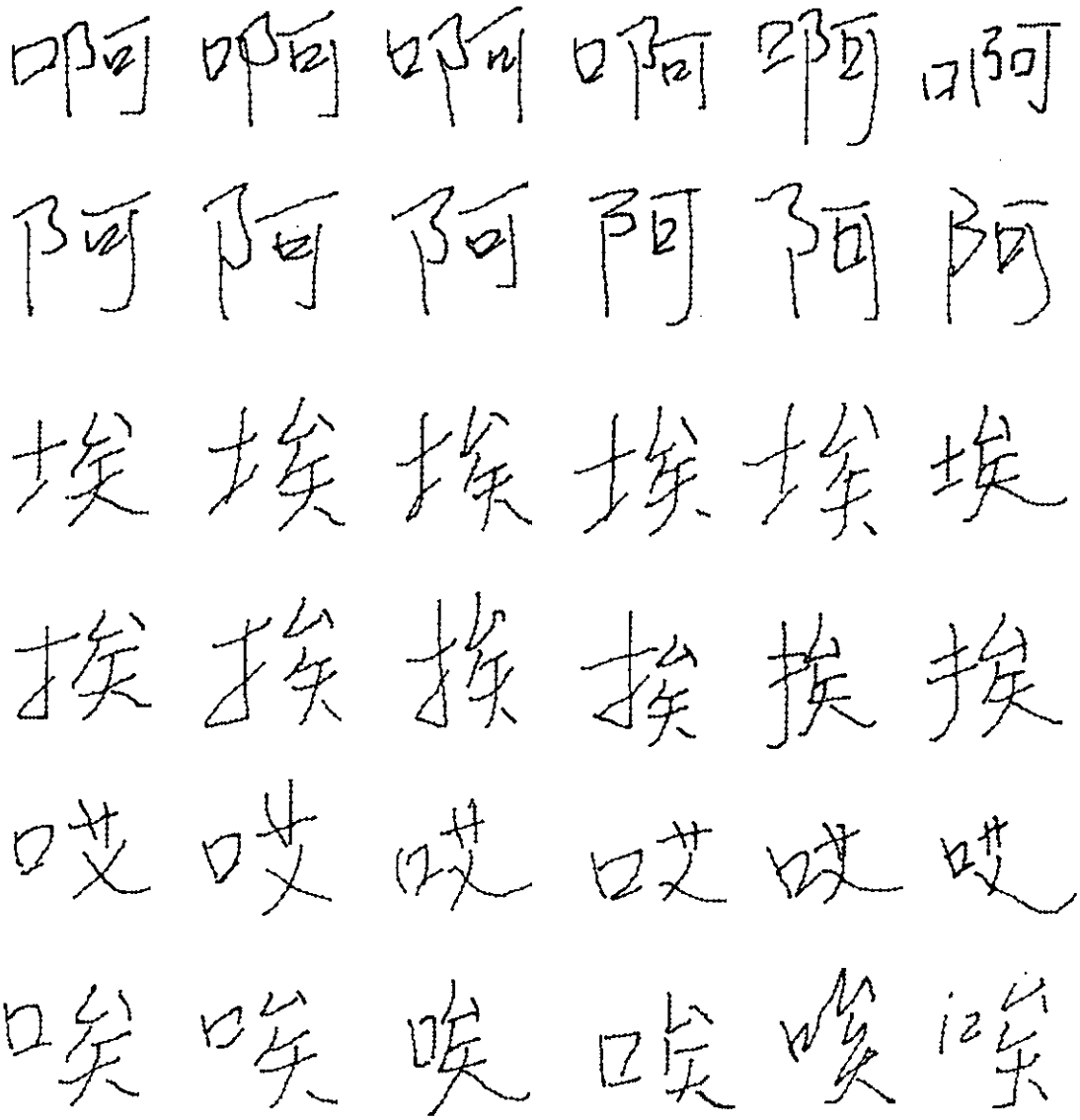
5. Xu R. F., Wang X.Z., and Yeung D. S. "Using Neural Network Classifier in Post-processing System for Handwritten Chinese Character Recognition", to appear in

IEEE International Conference on System, Man and Cybernetics 2001, (2001)

6. Xu R. F., Yeung D. S., and Shi D. M. "An improved hybrid post-processing system for offline handwritten Chinese character recognition using neural network classifier and language model", submitted to *Pattern Recognition* (2001)

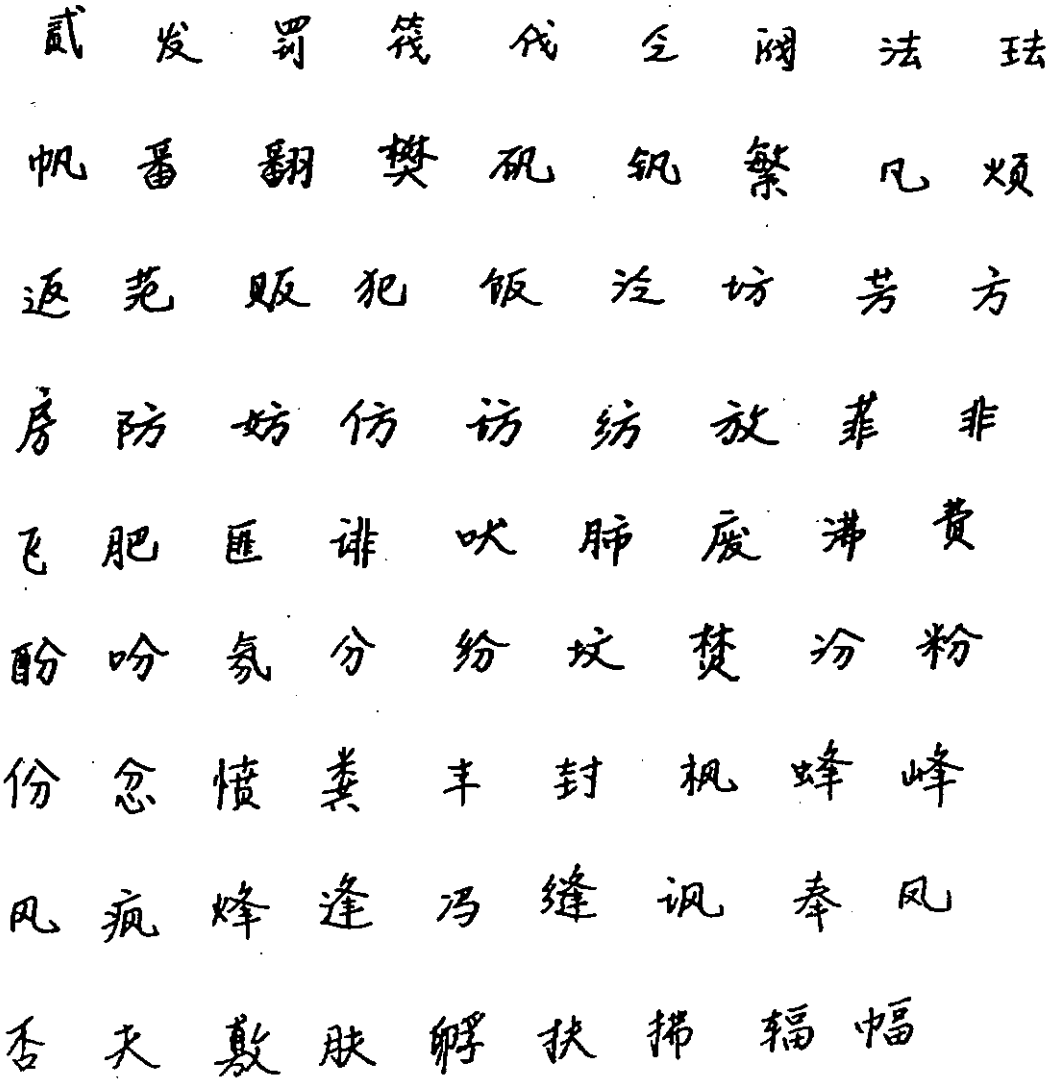
Appendix A

Online Handwritten Chinese Character Samples



Appendix B

Offline Handwritten Chinese Character Samples



Appendix C

An Example of HCCR Post-processing

Handwritten Chinese Character Samples

Character Recognition Result (Candidate Character Set)

如 1.00 何 0.93 把 1.00 握 1.00 世 1.00 界 1.00 经 1.00 济 1.00 岁 1.00 展 1.00 大 1.00 超 0.96 势 0.90

杠 1.00 忡 0.90 挝 1.00 据 0.98 毋 1.00 卑 0.98 组 1.00 法 1.00 发 0.98 屉 0.63 丈 1.00 题 0.20 热 0.87

加 0.98 俘 0.63 挫 0.99 掳 0.96 母 0.98 养 0.12 绎 0.99 洗 0.99 收 0.98 铱 0.57 下 1.00 销 0.20 拐 0.79

朽 0.83 皖 0.63 拙 0.97 掺 0.50 尘 0.83 易 0.12 绕 0.98 沛 0.98 忭 0.90 居 0.50 犬 0.99 弱 0.17 担 0.74

她 0.79 例 0.57 拈 0.95 撞 0.50 共 0.79 早 0.12 纾 0.87 沦 0.98 冬 0.87 屋 0.50 寸 0.99 匙 0.17 拇 0.74

帅 0.74 快 0.57 拔 0.95 摆 0.50 述 0.69 客 0.12 轻 0.87 沉 0.83 忙 0.83 屡 0.43 上 0.98 耗 0.12 抱 0.74

册 0.69 仲 0.43 抠 0.90 墟 0.43 芒 0.57 建 0.12 络 0.57 泡 0.83 枚 0.83 屏 0.43 于 0.98 抿 0.50

凡 0.20 件 0.43 挞 0.74 措 0.43 迈 0.57 果 0.12 终 0.57 沧 0.63 产 0.74 干 0.95 换 0.50

丸 0.12 待 0.43 担 0.74 摘 0.36 君 0.12 结 0.57 浓 0.63 孚 0.74 士 0.95 梨 0.12

杰 0.12 佰 0.43 拐 0.74 兵 0.63 工 0.95

Post-processing Stage based on Characteristics of Confusing Characters for Appending Similar-Shaped Characters and Adjust the Confidence Scores of Candidates

如 1.00 何 0.95 把 1.00 握 1.00 世 1.00 界 1.00 经 1.00 济 1.00 岁 1.00 展 1.00 大 1.00 超 0.96 势 0.93
加 1.00 忡 0.82 挝 0.98 据 0.99 毋 0.95 卑 0.98 组 1.00 法 1.00 发 0.99 屉 0.70 丈 1.00 趋 0.47 热 0.88
杠 0.98 仲 0.75 挫 0.98 掳 0.86 母 0.90 易 0.25 绎 0.90 洗 1.00 收 0.85 居 0.63 下 1.00 题 0.30 拐 0.70
朽 0.80 例 0.57 拙 0.97 摆 0.62 尘 0.85 果 0.20 绕 0.88 沛 0.95 忭 0.77 屋 0.52 犬 0.99 销 0.22 担 0.66
她 0.79 快 0.57 拈 0.97 撞 0.50 共 0.80 早 0.12 纾 0.88 沦 0.95 冬 0.67 屡 0.50 寸 0.99 弱 0.15 拇 0.66
帅 0.74 件 0.43 拔 0.90 掺 0.47 述 0.69 客 0.12 轻 0.78 沉 0.83 枚 0.65 屏 0.44 上 0.98 匙 0.12 抱 0.64
册 0.53 俘 0.43 抠 0.80 措 0.40 芒 0.57 建 0.12 络 0.62 泡 0.80 忙 0.53 铱 0.33 于 0.98 耗 0.12 抿 0.50
凡 0.12 佰 0.43 挞 0.62 摘 0.33 迈 0.42 养 0.12 结 0.60 沧 0.43 产 0.44 干 0.94 换 0.46
丸 0.12 待 0.43 担 0.50 墟 0.26 君 0.12 终 0.43 浓 0.23 孚 0.41 士 0.93 梨 0.12
杰 0.12 皖 0.22 拐 0.44 兵 0.33 工 0.90

Dictionary-based Sentence Segmentation and Fragments Detection

如何 把握 世界 经济 岁 展 大 超 势

The underlined character string "岁 展 大 超 势" is regarded as the sentence fragment.

Then the dictionary-based approximate word matching is applied to the sentence fragment to append linguistic-prone characters and adjust the confidence scores of candidates.

如 1.00 何 0.95 把 1.00 握 1.00 世 1.00 界 1.00 经 1.00 济 1.00 岁 1.00 展 1.00 大 1.00 超 0.96 势 0.95
加 1.00 忡 0.82 挝 0.98 据 0.99 毋 0.95 卑 0.98 组 1.00 法 1.00 发 1.00 屉 0.70 丈 1.00 趋 0.82 热 0.90
杠 0.98 仲 0.75 挫 0.98 掳 0.86 母 0.90 易 0.25 绎 0.90 洗 1.00 收 0.87 居 0.63 下 1.00 耗 0.32 拐 0.70
朽 0.80 例 0.57 拙 0.97 摆 0.62 尘 0.85 果 0.20 绕 0.88 沛 0.95 忭 0.77 屋 0.62 犬 0.99 题 0.30 担 0.66
她 0.79 快 0.57 拈 0.97 撞 0.50 共 0.80 早 0.12 纾 0.88 沦 0.95 冬 0.67 屡 0.50 寸 0.99 销 0.22 拇 0.66
帅 0.74 件 0.43 拔 0.90 掺 0.47 述 0.69 客 0.12 轻 0.78 沉 0.83 枚 0.65 屏 0.44 上 0.98 弱 0.15 抱 0.64
册 0.53 俘 0.43 抠 0.80 措 0.40 芒 0.57 建 0.12 络 0.62 泡 0.80 忙 0.53 铱 0.33 于 0.98 匙 0.12 抿 0.50
凡 0.12 佰 0.43 挞 0.62 摘 0.33 迈 0.42 养 0.12 结 0.60 沧 0.43 产 0.44 干 0.94 换 0.46
丸 0.12 待 0.43 担 0.50 墟 0.26 君 0.12 终 0.43 浓 0.23 孚 0.41 士 0.93 梨 0.12
杰 0.12 皖 0.22 拐 0.44 兵 0.33 工 0.90

Dictionary-based Word Lattice Construction

如 1.00 何 0.95 把 1.00 握 1.00 世 1.00 界 1.00 经 1.00 济 1.00 岁 1.00 展 1.00 大 1.00 超 0.96 势 0.95
加 1.00 忡 0.82 挝 0.98 据 0.99 毋 0.95 卑 0.98 组 1.00 法 1.00 发 1.00 屉 0.70 丈 1.00 趋 0.82 热 0.90
杠 0.98 仲 0.75 挫 0.98 掳 0.86 母 0.90 易 0.25 绎 0.90 洗 1.00 收 0.87 居 0.63 下 1.00 耗 0.32 拐 0.70
朽 0.80 例 0.57 拙 0.97 摆 0.62 尘 0.85 果 0.20 绕 0.88 沛 0.95 忭 0.77 屋 0.62 犬 0.99 题 0.30 担 0.66
她 0.79 快 0.57 拈 0.97 撞 0.50 共 0.80 早 0.12 纾 0.88 沦 0.95 冬 0.67 屡 0.50 寸 0.99 销 0.22 拇 0.66
帅 0.74 件 0.43 拔 0.90 掺 0.47 述 0.69 客 0.12 轻 0.78 沉 0.83 枚 0.65 屏 0.44 上 0.98 弱 0.15 抱 0.64
册 0.53 俘 0.43 抠 0.80 措 0.40 芒 0.57 建 0.12 络 0.62 泡 0.80 忙 0.53 铱 0.33 于 0.98 匙 0.12 抿 0.50
凡 0.12 佰 0.43 挞 0.62 摘 0.33 迈 0.42 养 0.12 结 0.60 沧 0.43 产 0.44 干 0.94 换 0.46
丸 0.12 待 0.43 担 0.50 墟 0.26 君 0.12 终 0.43 浓 0.23 孚 0.41 士 0.93 梨 0.12
杰 0.12 皖 0.22 拐 0.44 兵 0.33 工 0.90

如何　把握　世界　易经　经济　发展　居上　干耗　趋势
加快　尘界　收屋　屋上　趋热
共建
芒果

Statistical-based Post-processing Stage for Optimal Sentence Hypothesis Identification

The sentence hypothesis

如何 把握 世界 经济 发展 大 趋势

is selected as the final output result.

During this post-processing procedure, both the mis-recognized character and unrecognized character encountered in the original recognition result are recovered.

Bibliography

[1] Agui T. and Nagahashi N. "A description method of Hand-printed Chinese character". *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 1, pp. 20-24 (1979)

[2] Aho A. V. and Ullman J. D. *The theory of parsing, translation, and compiling*, Volume: Parsing, Prentice-Hall (1972)

[3] Chang C. H. "Automatic evaluation of language processing systems using the reversible Noisy Channel model". *Communications of COLIPS*, Vol. 7, No. 1 (1997)

[4] Chang C. H. "A pilot study on automatic Chinese spelling error correction". *Communications of COLIPS*, Vol. 4, No. 2, pp.143-149 (1994)

[5] Chang C. H. and Chen C. D. "Some issues on applying SA-class BI-Gram language models". *In Proceedings of ROCLING VII*, pp.171-186 (1994)

[6] Chang C. H. and Chen C. D. "A study on corpus-based classification of Chinese words". *Communications of COLIPS*, Vol. 5, No. 1 (1995)

[7] Chang J. S. and Lin Y. J. "An estimation of the entropy of Chinese- a new approach to constructing class-based N-Gram Models". R *In Proceedings of ROCLING VII*, pp.149-167 (1994)

[8] Chang Y. C. et al. "Methodology implementation and application of word-class based language model in Mandarin speech recognition". *In Proceedings of ROCLING VII*, pp.17-31 (1994)

[9] Chang X. G. and Xia Y. "A local search approach for statistics-based text recognition post-processing". *In Proceedings of 5th National Conference on Chinese Character Recognition*, pp. 231-238 (1995)

[10] Chang J. S. and Chen S. D. "The Post-processing of Optical Character Recognition based on statistical Noisy Channel and language model". *In Proceedings of PACLIC*, pp. 127-131 (1995)

[11] Chen W. T. and et al. "Lexicon-driven handwritten word recognition using optimal linear combinations of order statistics". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 1, pp.77-82 (1999)

[12] Chen, K. J., et al., "A system for On-line recognition of Chinese characters". *Pattern Recognition and Artificial Intelligence*, Vol. 2, No. 1, pp. 139-148, (1988)

[13] Chiang C. C. and Yu S. S. "A method for improving the machine recognition of confusing Chinese characters". *In Proceedings of IEEE International Conference on Pattern Recognition*, pp. 79-83 (1996)

[14] Chien C. H. "A Markov Language Model in Handwritten Chinese Text Recognition Application". *Master thesis*, Institute of computer Science and Information Engineering, NCTU, Taiwan (1991)

[15] Chien L. F., Chen K. J., and Lee L. S. "A Best-first language processing model integrating the Unification Grammar and Markov language model for speech

recognition applications". *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No. 2, pp. 221-240 (1993)

[16] Chou B. H. and Chang J. S. "The language models in Optical Chinese Character Recognition". *In Proceedings of ROCLING V*, Tai Wan, pp. 259-286, (1992)

[17] Doster W. "Contextual post-processing system for cooperation with a multiple-choice character recognition system". *IEEE Transactions on Computers*, Vol. C-26, No. 11, pp. 1090-1101 (1977)

[18] Du L., Wu J., and Sun Y. F. "Statistic Model based Chinese character recognition post-processing". *In Proceedings of 6th National Conference on Chinese Character Recognition*, pp. 175-180 (1997)

[19] Duda R. O. and Hart P. E. "Experiments in the recognition of Handprinted text: context analysis". *In Proceedings of AFIPS 1968*, Vol. 33, Washington, pp.1139-1149 (1968)

[20] Fano R., *Transmission of information*, MIT press, Cambridge MA (1961)

[21] Favata J. T. "General word recognition using approximate segment-string matching". *In Proceedings of 4th IEEE International Conference on Document Analysis and Recognition*, Vol. 1, pp. 92-96 (1997)

[22] Fong L., and Cheung H. K. "N-Gram estimates in probabilistic models for Pinyin to Hanzi transcription." *In Proceedings of IEEE International Conference on Intelligent Processing Systems*, Vol. 3, pp.1798-1803 (1997)

[23] Fu G., "Optimization methods for Fuzzy Clustering". *Fuzzy Sets and System*, Vol. 93, pp. 301-309 (1998)

[24] Fu G., "An algorithm for computing the transitive closure of a Fuzzy Similarity Matrix". *Fuzzy Sets and System*, Vol. 51, pp. 189-194 (1992)

[25] Fu W. K., Lee C. H., and Clubb O. L. "A survey on Chinese speech recognition". Communications of COLIPS, Vol.6, No.1, pp.1-17 (1996)

[26] GB-13715, National Standard of People's Republic of China - Word Segmentation Standard of Modern Chinese for Information Processing (1992)

[27] GB2312-80, National Standard of People's Republic of China - Modern Chinese character-set for information processing (1980)

[28] Golding A. R. and Schabes Y. "Combining TriGram-based and feature-based methods for context-sensitive spelling correction". *In Proceedings of 34th ACL Annual* (1996)

[29] Govindan V. K. "Character recognition - A review". *Pattern Recognition*, Vol. 23, no.7, pp. 671-683 (1990)

[30] Gu H. Y., Tseng C. Y., and L. S. Lee, "Markov modeling of Mandarin Chinese for decoding the phonetic sequence into Chinese characters". *Computer Speech and Language*, Vol. 5, No. 4, pp. 363-377 (1991)

[31] Gu X. F., et al. "Hand-printed Chinese character recognition system (HCCR)". *Communication of COLIPS*, Vol. 5, No. 2 (1997)

[32] Hildebrandt T. H., and Liu W. T. "Optical recognition of Handwritten Chinese Characters: Advances since 1980". *Pattern Recognition*, Vol. 26, No. 2, pp. 205-225 (1993)

[33] Hsu Y. L., Chang J. S., and Su K. Y. "Computational tools and resources for linguistic studies". *Computational Linguistics and Chinese Language Processing*, Vol. 2, No. 1, pp.1-40 (1997)

[34] Hull J. J. and Srihari N. "Experiments in text recognition with binary N-gram and Viterbi algorithms". *IEEE. Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-4, No. 5, pp. 520-530 (1982)

[35] Iyer R. M. and Ostendorf M. "Modeling long distance dependence in language: topic mixtures versus dynamic cache models". *IEEE Transactions on Speech and Audio Processing*, Vol.7 No.1, pp.30-39 (1999)

[36] Kaki S., Sumita F., and Iida H. "A method for correcting errors in speech recognition using the statistical features of character co-occurrence". *In Proceeding of COLING-ACL98*, pp.653-657 (1998)

[37] Kernighan M. et al. "A Spelling correction program base on a Noisy Channel model". *In Proceedings of COLING90*, Vol. 2, pp. 205-210 (1990)

[38] [Kukich K.] Kukich K. "Techniques for automatically correcting words in text". *ACM Computing Surveys*, Vol. 24, No. 4, pp.377-439 (1992)

[39] Kwong S., Man K. F., and Lai L. "On-line Chinese character recognition system". *Communication of COLIPS*, Vol. 5, No. 1, pp. 19-27 (1995)

[40] Lee H. J. and Tung C. H. "A language model based on semantically clustered words in a Chinese character recognition system". *In Proceedings of IEEE 3rd International Conference on Document Analysis and Recognition*, pp. 450-453 (1995)

[41] Lee H. J. and Lin Y. C., "Using confusing characters to improve character

recognition rate". *In Proceedings of IEEE International Conference on System, Man and Cybernetics*, California, USA, pp. 4195-4200 (1998)

[42] Lee H. J., and Tung C. H. "A language model based on semantically clustered words in a Chinese Character Recognition system". *In Proceedings of IEEE 3rd International Conference on Document Analysis and Recognition*, pp. 450-453 (1995)

[43] Lee H. S. et al. "A Markov language model in Chinese text recognition". *In Proceedings of IEEE 2nd International Conference on Document Analysis and Recognition*, pp. 72-75 (1993)

[44] Lee G., Lee J. H., and Yoo J. H. "Multi-level post-processing for Korean character recognition using morphological analysis and linguistic evaluation". *E-Print Archie comp-lg, no.9604011* (1996)

[45] Lee. L. S., Chien. L. F., and et al. "An efficient natural language processing system specially designed for Chinese language". *Computational Linguistics*, Vol. 17, No. 4, pp.347-374 (1991)

[46] Lee L. S., Tseng C. Y., et al., "Golden Mandarin (I) – A real-time Mandarin speech dictation machine for Chinese language with very large vocabulary". *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No.2 (1993)

[47] Lee L. S., et al., "Golden Mandarin (II) – an Improved single-chip real-time Mandarin dictation machine for Chinese language with very large vocabulary". *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 1993*, Vol. 2, pp. II503-506 (1993)

[48] Lee Y. S. and Chen H. H. "Analysis of error count distributions for improving the

post-processing performance of OCCR". *Communications of COLIPS*, Vol. 6, No. 2, pp. 81-86 (1996)

[49] Leung C. H., and Kan W. K. "Difficulties in Chinese typing error detection and ways to the solution". *Computer Processing of Oriental Languages*, Vol. 10, No. 1 (1996)

[50] Li G. H., Xia Y., et al, "A Chinese character recognition post-processing method based on character BI-Gram". *In Proceedings of 6th National Conference on Chinese Character Recognition*, Beijing, China, pp. 181-186 (1997)

[51] Li Y. X., Ding X. Q., and Liu C. S. "Post-processing study of Chinese document recognition based on HMM". *Chinese Journal of Chinese Information Processing*, Vol. 13, No. 4, pp. 29-33 (1999)

[52] Liang N. Y. and Zhen Y. B. "A Chinese word segmentation model and a Chinese word segmentation system PC-CWSS". *Communications of COLIPS*, Vol. 1, pp. 51-55 (1991)

[53] Liu C. L., Dai R. W., and Liu Y. J. "Hidden Markov Model and its application in character recognition". *In Proceedings of 5th National Conference on Chinese Character Recognition*, pp. 163-172 (1996)

[54] Liu J. F. "Study and implementation large-set practical On-line Handwritten Chinese Character Recognition system". *Ph.D. Thesis*, Harbin Institute of Technology, China (1996)

[55] Liu T., Wang K. Z., et al, "The maximum probability segmentation algorithm of ambiguous character strings". *In Proceedings of the 4th national joint conference on*

computational linguistics, pp.182-187, Beijing (1997)

[56] Liu Y., Tan Q., and Shen X. K., *The Word Segmentation Standard of Modern Chinese and Automatic Segmentation Method for Information Processing*, The Tsinghua University Press (1992)

[57] Liu Y., and Liang N. Y. *The Frequency Dictionary of Frequently Used Modern Chinese Words*, The Chinese Aerospace Press (1990)

[58] Lua K. T. "Experiments on the use of BI-Gram mutual information in Chinese natural language processing". *In Proceedings of International Conference on Oriental Language*, Hawaii, pp. 306- 313 (1995)

[59] Luk W. P. "Chinese word segmentation based on Maximum-Matching and Bigram techniques". *In Proceedings of ROCLING VII*, Tai Wan, pp. 273-282 (1994)

[60] Lyu et al. "Golden Mandarin III – A user-adaptive prosodic segment based Mandarin dication Machine for Chinese language with very large vocabulary". *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing 1995*, pp. 57-60 (1995)

[61] Miao L. F., Zhang S., and Zhou C., "A study post-processing approach to Chinese Character Recognition based on N-United-Word". *Chinese Journal of Chinese Information Processing*, Vol. 8, No. 2, pp. 39-46 (1994)

[62] Miller G. A. "The background to modern Cognitive Psychology". In J. Miller, *State of Mind*, New York: Pantheon, (1983)

[63] Morton J. "Word Recognition", In J. Morton, and J. C. Marshau (eds.), *Psycholinguistic Series 2*, MIT Press, (1979)

[64] Murveit H. et al. "Integrating natural language constraints into HMM-based speech recognition," *In Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing,* Albuquerque, NM, pp.573-576 (1990)

[65] Nagata M., "Context-based spelling correction for Japanese OCR". *In Proceedings of COLING 96*, vol. 2, pp. 806-811 (1996)

[66] Nagata M., "Japanese OCR error correction using character shape similarity and statistical language model". *In Proceedings of COLING-ACL 98*, vol. 2, pp. 922-928 (1998)

[67] Neuhoff D. L. "The Viterbi algorithm as an aid in text recognition". *IEEE Transactions on Information Theory*, pp. 222-226 (1975)

[68] Nie J. Y., Hannan M. L. and Jin W. Y. "Combining dictionary, rules and statistical information in segmentation of Chinese". *Computer processing of Chinese and Oriental language*, Vol. 9, No. 2, pp. 125-143 (1995)

[70] Plamondon R. and Srihari S.N. "On-Line and Off-Line Handwriting Recognition: A comprehensive survey". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63-84 (2000)

[71] Qin X. L., and Huang X. X. "The technique and realization of building the large-scale words and phrases base". *In Proceedings of 1998 International Conference on Chinese Information Processing*, pp.188-193 (1998)

[72] Rabiner L. R. and Huang B. H. "An introduction to Hidden Markov Models". *IEEE ASSP Magazine*, Jan. 1986, pp.4-16 (1986)

[73] Rabiner L. R. "An tutorial on Hidden Markov Models and selected applications in speech recognition". *Proceedings of IEEE*, Vol.77, No.2, pp.2257-286 (1989)

[74] Raviv J. "Decision making in Markov chains applied to the problem of Pattern Recognition". *IEEE Transactions on Information Theory*, Vol. IT-3, pp. 536-551 (1967)

[75] Riseman M. and Hanson R. "A contextual post-processing system for error correction using binary N-Grams". *IEEE Transactions on Computer*, Vol. C-23, No.5, pp.480-493, (1974)

[76] Schwartz R. et al., "The optimal N-best algorithm: An efficient procedure for finding the Top N sentence hypotheses". *In Proceedings of Drapa Speech and Natural Language* (1989)

[77] Sekita I., Toraichi K., et al. "Feature extraction of Hand-printed Japanese characters by spine function for relaxation matching". *Pattern Recognition*, Vol. 21, no. 1, pp. 1-7 (1988)

[78] Shahidul-Hussain A. B., "Compound sequential probability ratio test for the classification of statistically dependent patterns". *IEEE Transactions on Computers*, Vol. C-23, pp. 398-410 (1974)

[79] Sheng L. D., and Fan J. S. "Post-processing research for Handwritten Chinese Character Recognition." *In Proceedings of 6th National Conference on Chinese Character Recognition*, pp. 200-211 (1997)

[80] Shi D. M., "Study and implement of Off-line Handwritten Chinese Character Recognition". *Ph.D. Thesis*, Harbin Institute of Technology, China (1997)

[81] Shi D. M., Damper R. I. and Shu W. H. "Chinese character recognition using

Genetic Algorithms and extension matrix algorithms". *Communications of COLIPS*, Vol. 9, No. 2, pp. 137-154 (2000)

[82] Shi D. M., Gunn S. R., et al. "Recognition rule acquisition by an advanced extension matrix algorithm". *International Journal of Engineering Intelligent Systems*, Vol. 8, No. 2, pp. 97-101 (2000)

[83] Shyu K. H., Tsay M. K. and Chen C. S., "An OCR based translation system between Simplified and Complex Chinese characters". *Computer Processing of Chinese and Oriental Languages*, Vol. 9, No. 1, pp. 59-68 (1995)

[84] Simons M. et al. "Distant BI-Gram language modeling using maximum entropy". *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp.787-790 (1997)

[85] Sinha R. M. K. and Prasada B. "Visual text recognition through contextual processing". *Pattern Recognition*, Vol. 21, No. 5, pp.463-479 (1988)

[86] [Sinha R. M. K. 1993] Sinha R. M. K., Prasada B., et al. "Hybrid contextual text recognition with string matching". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 9, pp.915-925 (1993)

[87] Song R., Qiu C. J., et al. "BI-Orderly-Neighborship and its application to Chinese word-segmentation and proof reading". *In Proceedings of International Conference on Chinese Computing 96*, pp. 428-433, Singapore (1996)

[88] Sproat R. "An application of statistical optimization with dynamic programming to phonemic-input-to-character conversion for Chinese". *Proceedings of ROCLING III*, Taiwan, pp. 377-390 (1990)

[89] Sproat R. and Shih C. "A statistical method for finding word boundaries in Chinese text". *Computer Processing of Chinese and Oriental Languages,* Vol. 4, No. 4, pp. 336-349 (1990)

[90] Stallings W. "Approach to Chinese character recognition". *Pattern Recognition,* Vol. 8, no. 2, pp. 87-98 (1976)

[91] Su K. Y., Chiang T. H. and Chang J. S. "An overview of corpus-based statistics-oriented (CBSO) techniques for Natural Language Processing". *Computational Linguistics and Chinese Language Processing,* Vol. 1, No. 1, pp. 101-157 (1996)

[92] Suen C.Y. "N-Gram statistics for Natural Language Understanding and text processing". *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. PAMI-1, No. 2, pp. 164-172 (1979)

[93] Sun M. S., Huang C. N., et al, "Using character BI-gram for ambiguity resolution in Chinese word segmentation". *Computer research and development,* Vol. 34, No. 5, pp. 332-339 (1997)

[94] Sun S. W. "A contextual post-processing for Optical Chinese Character Recognition". *IEEE International Symposium on Circuits and Systems,* pp. 2641-2644, (1991)

[95] Sun W., Liu L. M., et al. "Intelligent OCR processing". *Journal of The American Society for Information Science,* Vol. 43, No. 6, pp. 422-431 (1992)

[96] Tang Y. Y., et al. "Offline recognition of Chinese handwriting by multi-feature and multilevel classification," *IEEE Transactions on Pattern Analysis and Machine*

Intelligence, Vol. 20, No. 5, pp. 556- 561 (1998)

[97] Tappent C. C., Suen C. Y., et al. "The state of the art in Online Handwriting Recognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 8, pp. 787-808 (1990)

[98] Tomita M. "An efficient word lattice parsing algorithm for continuous speech recognition". *In Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 1569~1572 (1986)

[99] Toussaint G. T. "The use of context in Pattern Recognition". *In Proceedings of Pattern Recognition Image Processing*, pp.1-10 (1977)

[100] Tung C. H. and Lee H. J. "2-stage character recognition by detection and correction of erroneously-identified characters". *In Proceedings Of the 2nd IEEE International Conference on Document Analysis and Recognition*, pp. 834-837 (1993)

[101] Wang X. L. "The fewest segmentation problem and its solution". *Science in China*, Vol. 1989, pp. 1030-1032 (1989)

[102] Wang H., et al. *The Frequency Dictionary of Modern Chinese Words*, The Beijing Linguistic University Press (1986)

[103] Ward K. et al. "On the need for a theory of integration of knowledge sources for spoken language understanding". *In Proceedings of AAAI Workshop on Integration of Natural Language and Speech Processing*, Seattle, WA, pp.23-30 (1994)

[104] Wong P. K. and Chan C. "Chinese word segmentation based on maximum matching and word binding force". *In Proceedings of COLING 96*, Japan, pp.200-203 (1996)

[105] Wong P. K., and Chan C. "Off-Line Handwritten Chinese Character Recognition as a compound Bayes decision problem". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 9, pp. 1016-1023 (1998)

[106] Wong P. K. and Chan C. "Post-processing statistical langauge models for a Handwritten Chinese Character Recognizer". *IEEE Transactions on System, Man, and Cybernetics*, Vol. 29, No. 2, pp. 286-291 (1999)

[107] WORDDATA, Chinese Knowledge Information Processing Group, Technical Report, No. 93-05, Institute of Information Science, Academic Sinica, Taiwan, (1993)

[108] Wu J., Wang Z. Y., and Ren Y. S. "Stochastic language models for Chinese speech recognition based on Chinese spelling". *In Proceedings of IEEE ISSIPNN 94*, pp.674-677 (1994)

[109] Xing H. B. "Statistical results and analysis of basic components in Chinese characters and words". *In Proceedings of 1998 International Conference on Chinese Information Processing*, pp.56-62, (1998)

[110] Xu D. X., Zhu P. J., and Huang T. Y. "Using high-level linguistic knowledge for Chinese speech recognition". *In Proceedings of 9th IEEE International Conference on Pattern Recognition*, Vol. 1, pp. 197-200 (1988)

[111] Xu R. F. and Yeung D. S. "Experiments on the use of corpus-based word BI-gram in Chinese word segmentation," *In Proceedings of IEEE International Conference on System, Man, and Cybernetics*, California, Vol. 5, pp. 4222-4227 (1998)

[112] Xu R. F., Yeung D.S., and Wang X.L. "A hybrid post-processing approach for handwritten Chinese character recognition". *In Proceedings of the International*

Conference on Machine Translation and Computer Language Information Processing, Vol. 1, pp. 152-158 (1999)

[113] Xu R. F., Yeung D. S., and Shu W. H. "Using confusing character, dictionary matching and word BI-Gram language model for improving Handwritten Chinese Character Recognition". *In Proceedings of the International Conference on Artificial Intelligence 2000*, Las Vegas, USA, Vol. III, pp. 1271-1277, June 2000 (2000)

[114] Yang J. and Wang Y. Q. "Automatic recognition of Handwritten Chinese text based on linguistics knowledge". *Journal of Computer Research and Development*, Vol. 35, No. 7, pp. 668-672 (1998)

[115] Yang L. Y. "Language models in large vocabulary speech recognition system". *Master Thesis*, Department of Computer Science, Tsinghua University (1991)

[116] Yannakoudakis E. J., Tsomokos I., and Hutton P. J. "N-Grams and their implication to Natural Language to natural language under-standing". *Pattern Recognition*, Vol. 23, pp. 509-528 (1990)

[117] Yao T. S., Zhang G. P., et al. "A rule-based Chinese automatic segmentation system". *Chinese Journal of Chinese information processing*, Vol. 4, No. 1, pp. 37-42 (1990)

[118] Yeh C. L. and Lee H. S. "Rule-based word identification for Mandarin Chinese sentences – A unification approach". *Computer processing of Chinese and Oriental Languages*, Vol. 4, No. 2, pp. 97-118 (1991)

[119] Yeung D. S., and Fong H. S. "Handwritten Chinese Character Recognition by rule-rmbedded Neocognitron". *Neural Computing & Applications*, Vol. 2, pp. 216-226

(1994)

[120] Yeung D. S., and Fong H. S. "A fuzzy substroke extractor for Handwritten Chinese Character". *Pattern Recognition*, vol. 29 no. 12, pp. 1963-1980, (1996)

[121] Zeng X. N. "A TRI-Gram model based application in post-processing for Handwritten Chinese Character Recognition". *In Proceedings of 6th National Conference on Chinese Character Recognition*, China, pp. 195-199, 1997

[122] Zhang D. X., Ma S. P., et al. "Handwritten similar Chinese characters recognition based on combining statistics with Neural Networks method". *Chinese Journal of Chinese Information Processing*, Vol. 13, No. 3, pp.33-39 (2000)

[123] Zhang S. W. and Huang T. Y. "A study of the value of parameter N in N-Gram statistical model in Chinese language". *Chinese Journal of Chinese Information Processing*, Vol. 12, No.1, pp.35-41 (1999)

[124] Zhou G. D. and Lua K. T. "Detection of unknown Chinese words using a hybrid approach". *Computer Processing of Oriental Languages*, Vol. 11, No. 1, pp. 63-75 (1997)