The Postprocessing of Optical Character Recognition based on Statistical Noisy channel and language model

Jason J. S. Chang and Shun-Der Chen
Department of computer Science, National Tsing Hua University
e-mail : jschang@cs.nthu.edu.tw
dr798306@cs.nthu.edu.tw

Abstract

The techniques of image processing have been used in optical character recognition (OCR) for a long time. The recognition method evolved from early "pattern recognition" to "feature extraction" recently. The recognition rate is raised from 70% to 90%. But the character by character recognition technique has its limitation. Using language models to assist the OCR system in improving recognition rate is the topic of many recent researches.

Recently, the related research on Chinese nature language processing has improved rapidly. These improvement include the Chinese word segmentation, syntax analysis, semantic analysis, collocation analysis, statistical language models.

In this paper, we will propose a new techniques for Chinese OCR postprocessing and postediting. We combine noisy channel model and the technique of natural language processing to implement an OCR postprocessing system. From the result of experiments, we found noisy channel model very effective for postprocessing. Under the approach, it is possible to recover the correct character, even when it is not in the candidate list produced by the OCR system.

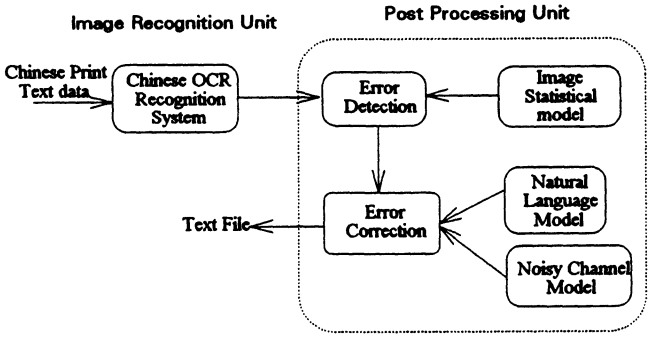
1. Introduction

Recently, the research of human-machine interface is becoming more and more important. To reduce the workload of user entering text and data into a computer system, there are many human-machine interfaces developed, such as optical character recognition (OCR), speech recognition, etc. However, some errors are usually introduced in the recognition process. Therefore, means for finding and correcting such errors in indispensable.

Research in character recognition for Chinese faces more difficulties than for other languages. Firstly, Chinese has a huge character set. Secondly, Chinese characters have more complex structure than alphabetic characters and there are a large number of similar character groups. Therefore, we need some kind of contextual information to detect and correct errors. Language models can be used to provide such contextual information to enhance the correction rate and speed of a man-machine interface system. For example, Tsuyoshi [14] use character frequency and the morphological analysis to improve the handwritten Japanese OCR system.

Recently the techniques of Chinese OCR have advanced greatly, the recognition rate has reached 95% and 90% for printed and handwritten text respectively. In view of the limitation of the techniques of OCR, we hope to use natural language models to detect the 5 to 10% errors and suggested their possible correction in a postediting environment. With an effective postprocessing technique and user-friendly posteditor, the usability of OCR technology can be greatly enhanced.

2. The framework of the System

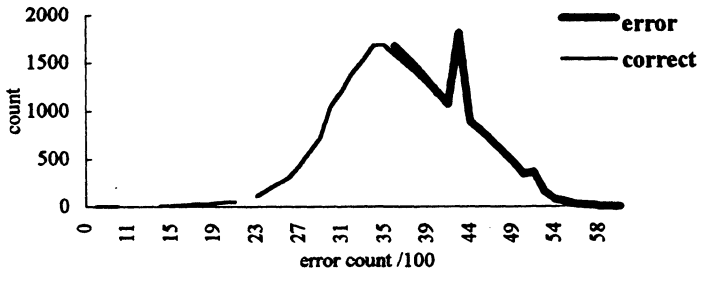


In this paper, we have adopted a multi-stage postprocessing approach. In the detection stage, the statistical model is used to re-evaluate the confidence of the error counts provided by the image model. The purpose of this stage is to identify the places where the first candidate is wrong. In correction stage, we combine the noisy channel and language model to suggest possible corrections.

3. The Error Detection Model

To evaluate the confidence of the error counts, we analyse the output produced by the recognition system (UMAX Standard Chinese OCR System). First, we scan the most frequent 5401 Chinese characters. And we analyzed the relationship between the number of the correct candidates with a certain error count (see Figure 1). Using the distribution of the relation, we can come up with a threshold for detecting point when the first candidate is in error. To further improve the precision of detection, a dictionary is used to reconfirm these detection points.

Figure 1. number of correct candidate vs. error count



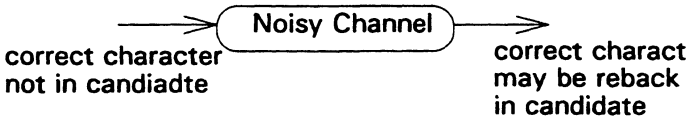
4. The Error Correction Model

4.1 Language Model

In the first experiment of the correction stage, we used the word segmentation model [18] locally to correct the detected points. This was very effective in correcting errors found in long words consisting of 2 or more characters. As for single-character word, the character bigram model was used. We trained the character bigram information from a 300,000 Chinese character balance corpus. From the result of experiment, the correction stage eliminated 2 thirds of the error produced by OCR.

4.2 Noisy Channel Model

Our previous work[20] shows that the upmost difficultly in postprocessing is found in the case where the correct character is missing from the candidate list. Therefore, we exploit noisy channel model to introduce some characters into the candidate list to resolve the problem. In our experiment, that can be done successfully for 40% the case.



5. Conclusion

This paper proposes a new approach to the OCR postprocessing. It combines the information provided by the image processing unit and contextual information to detect the recognition errors. In the past, the OCR system can not solve the problem of the missing correct character in the candidate list. In out experiment, we find it that the noisy channel model is very effective in solving this problem.

In the future, we will enhance the noisy channel model and introduce the long distance contextual information to improve our system. It seems that this model is applicable to other man-machine interface applications. We are currently studying the possibility of apply the model to speech recognition.

Table 1: The experiment result of error detection

Title	百年與十分鐘				戴勳章逛街的人			
	total errors find in first candidate 45				total errors find in first candidate 73			
	alarms	true positive	Precision	Recall	alarms	true positive	Precision	Recall
UMAX	90	19	21.1%	42.2%	122	31	25.4%	42.5%
Threshhold	221	45	20.4%	100%	272	72	26.5%	98.6%
Dictionary	94	43	46%	95.6%	163	71	43.6%	97.3%

Table 2 : The experiment result of error correction

百年與十分鐘						
total errors find in first candidate 45						
	alarms	detection (true positive)	alarm precision (uncorrect error)	correction (false negative)	correction (true positive)	correct rate
UMAX	90	19	54%	35	10	22%
Word Segment	68	19	79%	24	21	47%
Noisy Channel	68	14	74%	19	26	58%
Character Bigram	68	9	69%	13	32	71%

戴勳章逛街的人						
total errors find in first candidate 73						
	alarms	detection (true positive)	alarm precision (uncorrect errors)	correction (false negative)	correction (true positive)	correct rate
UMAX	122	31	57%	54	19	26%
Word Segment	109	36	86%	42	31	43%
Noisy Channel	109	31	86%	36	37	51%
Character Bigram	109	22	85%	26	47	64%

* alarm precision (uncorrect errors) = detection (true positive) / correction (false negative)

Reference

[1] Andrew R. Golding, A Bayesian hybrid method for context-sensitive spelling correction. In Proceeding of the third workshop on Very Large Corpora, 30 June, 1995.

[2] Chao-Huang Chang, Word Class Discovery for Postprocessing Chinese Handwriting Recognition. In Proceeding of COLING-94, Page 1221-1225, 1994.

[3] D. G. Elliman, A Review of Segmentation and Contextual Analysis Techniques for Text Recognition. Pattern Recognition, Vol 23, No. 3/4, pp. 337-346, 1990.

[4] Frank Smadja, Retrieving Collocations from Text: Xtract, 1993 Association for Computation Linguistics.

[5] H. Takahashi, etc., A Spelling Correction Method and its Application to an OCR System. Pattern Recognition, Vol 23, No. 3/4, pp. 363-377, 1990.

[6] K. T. Lua and K. W. Gan, Recognizing Chinese Characters Through Interactive Activation and Competition. Pattern Recognition, Vol 23, No. 12, pp. 1311-1321, 1990.

[7] Mark D. Kernighan, Kenneth W. Church, and William A. Gale, A Spelling Correction Program Based on a Noisy Channel Model. In Proceeding of COLING-90, vol 2, Page 205-210, 1990.

[8] Masaki YAMASHINA & Fumihiko OBASHI, Collocation Analysis in Japanese Text Input.

[9] Jyn-Sheng Chang and Yuh-Juh Lin, An estimation of the entroy of Chinese - A new approach to building class-based n-gram models. Rocling VII, 1994.

[10] Rose & L.J. Evett, Text Recognition using Collocations and Domain Codes. In Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, pp.65-73, Columbus, Ohlo, 1993.
[11] Tao Hong, Integration of Visual Inter-word Constraints and Linguistic Knowledge in Degraded Text Recognition. In Proceeding of 32nd Annual Meeting of Association for Computational Linguistics, pp. 328-330, Las Cruces, New Mexico, 27-30 June, 1994.
[12] Tao Hong and Jonathan J. Hull, Degraded Text Recognition Using Word Collocation and Visual Inter-word Constraints.
[13] Tetsuo Araki et al., An Evaluation of a Method to Detect and Correct Erroneous Characters in Japanese input through an OCR using Markov Models.
[14] Tsuyoshi Kitani, An OCR Post-Processing method for Handwritten Japanese Documents.
[15] YASUHITO TANAKA & SHO YOSHIDA, The Acquisition of Knowledge Data for Natural Languages ~ Word-to-Word Relationships Obtained by Analyzing Data Found in the Asahi Newspaper ~
[16] 張照煌(Chao-Huang Chang)，中文錯別字自動訂正方法初探（初稿）
[17] 周並弘、張俊盛，語言模式在中文文字辨識上的應用，Rocling，1992.
[18] 張俊盛、陳志達、陳舜德，限制式滿足及機率最佳化的中文斷詞方法. Rocling IV, 1991.
[19] 王榮宗、王駿發，語言模式在中文語音辨識上的應用. Rocling VII, 1994.
[20] 文字自動辨認系統語言分析模式, 交通部電信研究所期末報告，1994

Output Example:

[Example 1]

掉在地上必然會發出金屬一樣響脆的聲音
0000000000000000000XX0000XX00000000
掉在地上必然會發出金石一樣響脆的聲音
####################????##????########
############################??########
掉在地上必然會發出金石一樣響脆的聲音
####################?d?d##?d??########
掉在地上必然會發出金屬一樣響脆的聲音
####################?dmd##?d??########

[Example 2]

熨燙得非常整齊
XX000000000000
使燙得非常整齊
??############
??############
使燙得非常整齊
??############
熨燙得非常整齊
Nd############

the output of UMAX Chinese OCR system

10 使 04872 愛 04973 度 05178 便 05210 良 05243 庆 05252 瘦 05281 要 05320 更 05329 夜 05333
10 燙 03713 吏 04832 更 04944 愛 05152 史 05169 曳 05274 受 05288 变 05320 菱 05425 斐 05433
10 得 03538 符 04202 待 04295 存 04529 侍 04750 行 04751 捍 04787 捍 04794 持 04826 特 04917
10 非 02347 井 03497 菲 03713 并 03990 許 04104 升 04175 拌 04181 弄 04207 卉 04245 咋 04303
10 常 03206 帝 04445 帶 04627 索 04631 希 04715 兮 04741 肯 04894 申 04930 哼 04977 牛 05000
10 整 03501 塾 04834 斐 05165 奎 05238 鯊 05294 坚 05315 贊 05379 簧 05422 莹 05428 幣 05436
10 齊 03293 齋 04313 卉 04633 并 04738 萍 04765 猝 04786 捭 04803 奔 04848 帝 04853 芥 04859