Research on OCR Post-processing Applications for Handwritten Recognition Based on Analysis of Scientific Materials

Zhijuan Hu[1], Jie Lin[2], and Lu Wu[3]

[1] Room 3022, 17th Building, NO.4800, Caoan Road, Jiading District, Shanghai, China
huzhj2008@yahoo.cn
[2] Room 506,YunChou Building, NO.1239, SiPing Road, Yangpu District, Shanghai, China
jielinfd@163.com
[3] Room 918, NO.95, West Beijing Road, Jingan District
wulu@fangdi.com.cn

Abstract. This paper studied the application of OCR post-processing techniques in Real estate transactions registration and proposed a dictionary-based post-processing method. This paper introduced briefly the design of database and post-processing program of this system. Average accuracy rate was enhanced largely compared to that of pretreatment when the system conducted models test. The experimental results showed that this model was very practical, and can significantly improve the recognition accuracy rate, which verified the validity of the approach.

Keywords: OCR, Post-processing, Handwritten recognition.

1 Introduction

In recent years, OCR study for Chinese character recognition has made great progress. At the same time, many OCR software systems have been commercialized successfully to market and achieved a great deal of economic and social benefits. However, for Chinese text recognition, during to its complicated structure and considerable variability in written, character recognition rate is subject to certain restrictions. In particular, for offline handwritten Chinese characters, OCR recognition rate is not good enough, some good OCR software generally keep accuracy rates of about 50-60%. At present, a considerable part of the key information in the registrations of real estate transactions needs to be identified and extracted from the offline handwritten, it is of great significance to improve the recognition rate of the text by verifying and correcting through OCR post-processing techniques which based on character recognition and integrated professional fields knowledge, probability statistics and statistical language model.

2 OCR Post-processing Techniques

OCR[1] post-processing method is by analyzing the syntax, semantics and phrase of OCR recognition pretreatment results in which every single character may be

D. Jin and S. Lin (Eds.): CSISE 2011, AISC 104, pp. 131–135, 2011.
springerlink.com

represented by several candidates through using certain language and knowledge models, in order to rectify some incidental mistakes and finally to realize the purpose of improve recognition accuracy rate. A common goal of OCR post-processing systems [2] is to ensure that the words or sentences generated by corrections to the OCR output are correct in the sense that they belong to the language of the task and the specialty. The key issue of post-processing is how to use context information to determine a plausible word from multiple candidate word as a final result to be outputted.

Most widely used OCR post-processing method [3] are those based on probability and Statistical models, based on dictionary and mixed method. In terms of specific implementation technology, probability and statistical-based models can be classified into two categories: Hidden Markov Method and n-gram methods (especially Bi-gram and Tri-gram). A hidden Markov model (HMM) [4] is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved state. An n-gram [5] model models sequences, notably natural languages, using the statistical properties of n-grams. N-gram models are widely used in statistical natural language processing. Bigrams or digrams are groups of two written letters, two syllables, or two words, and are very commonly used as the basis for simple statistical analysis of text. They are used in one of the most successful language models for speech recognition [6]. They are a special case of N-gram. Trigrams are a special case of the N-gram, where N is 3. They are often used in natural language processing [7] for doing statistical analysis of texts.

The dictionary-based post-processing method is constituted by two main steps: the first one is to choose candidate characters; secondly, calculate the similarity between the candidate and the terms in the dictionary, and output the most similar terms as plausible results. This paper adopted a dictionary-based post-processing method to recognize and distinguish key information in original property files which were scanned in daily registration in every district of Shanghai Real Estate Trading Center. By saving recognition results, the database can be enriched. For clerks they can also compare and identify the content while entering data manually, hence, improving the accuracy further.

3 OCR Post-processing System

The post-processing system, which is a combination of post-processing program and DBMS, is responsible for processing pre-results. The system will offer post-processing results to the user after retrieving the dictionary database by post-processing program. DBMS is maintained by DBA is a less important part in this whole system. Since in this system we adopted dictionary based method, the design of dictionary is also an important component of the system. We will introduce the two important parts: Dictionary database design and post-processing program in the following.

4 Dictionary Database Design

A key problem of dictionary-based post-processing method is how to build Chinese terms database. The database should meet some requirements. First, it should

performance perfect term storage and maintenance functions; Secondly, it must reflect the frequency of words being used; Thirdly, it should help to improve the search speed as much as possible, only by this way, it will facilitate post-processing program. For different industries, the content of the professional entries in the dictionary is changing greatly.

This paper conducted a study in Real estate industry. Accordingly every term in the dictionary contain name of the term, its type and also its used frequency. We also group the dictionary by type so that when executing similarity calculation, the program will search terms of a specific type which is known already. In this way, recognition accuracy rate can be enhanced by avoiding the confusion of different types which is not the same as the given one.

5 OCR Post-processing Model

1. Define a String: $S=S_1+S_2+\ldots+S_{m;}$

Where S_i (i=1…m;m represents the length of the string) is the characters to be identified. For each S_i there are several candidate characters which are denoted as X_{ij}. This definition can be described as following:

S_1	S_2	S_3	…	S_m
(λ_{11},X_{11})	(λ_{21},X_{21})	(λ_{31},X_{31})	…	(λ_{m1},X_{m1})
(λ_{12},X_{12})	(λ_{22},X_{22})	(λ_{32},X_{32})	…	(λ_{m2},X_{m2})
⋮	⋮	⋮	…	⋮
(λ_{1n},X_{1n})	(λ_{2n},X_{2n})	(λ_{3n},X_{3n})	…	(λ_{mn},X_{mn})

Where: X_{ij} is candidate word j for S_i; λ_{ij} is the similarity degree of candidate word j for S_i.

2. Set s string $W= X_{11}+ X_{21} +X_{31} \ldots+X_{m1;}$

Where we get the first candidate for every S_i,that is those have the most similarity, and form a string. Then we search in the dictionary to find whether there a same string, if there is one, output it as the most possible results.

3.if there is none term which is the same as string W, then search all the terms in the database for those contain at least one of the first candidates, and compose entries $Record=\{T_1,T_2,T_3\ldots T_L\}$ (where L is the number of eligible terms). Then introduce a variable TValue.

$$\text{Set TValue}=\sum_{i=1}^{m}\sum_{j=1}^{n}\lambda_{ij}$$

Where n is the length of every term;When the character in the term is the same as the candidate word, λ_{ij} is the similarity degree of candidate word j for S_i; When the character in the term is different from the candidate word, se t $\lambda_{ij}=0$.

4. For every term T_k(k=1…L), calculate corresponding the value of TValue, order by value and then output several front T_k, which has relatively bigger TValue value, as temporary results.

The prerequisite of above way to select record is that at least one of the first candidate word is the correct final results. If all of X_{11}, X_{21}, X_{31} …,X_{m1} are wrong,

two situation will happen as following:(1)There are temporary results, however, they are wrong. When this happens, the corresponding TValue of terms are relatively low. So when the biggest TValue is very low, there is a need for reselect terms. (2)There is no term at all, the searching results are empty.

We adopted the same way to handle those two instances. Firstly, search in the database for all terms that contain at least one of the candidate words, not the first ones, but all candidates; secondly, calculate corresponding TValue for every selected terms; thirdly, output those terms with the highest values.

The biggest disadvantage of this mean is large computation; therefore, it would be adopted on when those two situations happen.

6 Experiment and Analysis

In this paper, we recognized four common templates, each contain about 100 proofs. Therefore, this experiment is universality in some degree.

1. Experiment result:

Table 1.

templates	Total number	Pretreatment accuracy rate	The first temporary result accuracy rate	The second temporary result accuracy rate	The third temporary result accuracy rate
Template 1	102	35.33%	65.55%	68.02%	70.57%
Template 2	90	31.43%	63.06%	69.76%	70.03%
Template 3	95	39.26%	67.57%	71.79%	73.34%
Template 4	105	27.75%	58.92%	62.34%	65.89%
Average		33.44%	63.78%	67.98%	69.96%

2.Analysis

(1) This approach has very high accuracy rate, which could enhance accuracy rate by nearly 30 percent. Two main reasons contribute to this success. One is that we give the most similar terms instead of the exact ones; the other one is the correct result is in the database without any exception. In this way, we limit search range, therefore, we improve accuracy rate greatly.

(2) A higher pretreatment accuracy rate will lead to a higher post-processing rate. This is because post-processing program use pretreatment results as query conditions to search in the database. Consequently, only high pretreatment accuracy rate can improve query results.

7 Conclusion

In conclusion, this paper adopted word-based language model to realize post-processing program. Experiment results show that this program is effective in enhancing recognition accuracy rate. However, since this approach is an application in specific industry, some limitations still exist, and further study is needed. At the same time, post-processing resulted is closely related with pretreatment results, there is a lot work to be done in increasing pretreatment results.

Acknowledgement. The authors would like to thank the fund: Application of OCR techniques in Real estate transactions registration (FG2009019).

References

1. http://en.wikipedia.org/wiki/Optical_character_recognition
2. Perez-Cortes, J.C., Amengual, J.-C., Arlandis, J., Llobet, R.: Stochastic error-correcting parsing for OCR post-processing. In: Proceedings of the International Conference on Pattern Recognition, pp. 4405–4408 (2000)
3. Long, C., Zhuang, L., Zhu, X.-y.: A Post-processing Approach for Handwritten Chinese Address Recognition. Journal Of Chinese Information Processing 20(6), 69–74 (2006)
4. Wang, X., Yang, Y., Xie, B.: HMM-based off-line handwritten Chinese characters recognition using Krawtchouk moments. In: 6th World Congress on Intelligent Control and Automation, WCICA 2006, pp. 10068–10072 (2006)
5. http://en.wikipedia.org/wiki/N-gram#n-gram_models
6. Collins, M.: A new statistical parser based on bigram lexical dependencies. In: Proceedings of the 34th Annual Meeting of the Association of Computational Linguistics, Santa Cruz, CA, pp. 184–191 (1996)
7. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)