An OCR Post-processing Approach Based on Multi-knowledge

Li Zhuang and Xiaoyan Zhu

Department of Computer Science and Technology, Tsinghua University
State Key Laboratory of Intelligent Technology and Systems
Beijing, 100084, P.R. China
zhuangli98@mails.tsinghua.edu.cn, zxy-dcs@tsinghua.edu.cn

Abstract. This paper proposes an OCR post-processing approach based on multi-knowledge, which integrates language knowledge and candidate distance information given by the OCR engine. In this approach, statistical language model and semantic lexicon are combined, and candidate distance information is used to reduce the size of the search space. The experimental results show that this approach is very effective. After post-processing, the recognition accuracy rate on the test set increases from 58.45% to 83.73%, which means 60.84% error reduction.

1 Introduction

In most OCR systems, independent character recognition engine is often used to recognize each segmented part of an image, where only shape and structure of the character are considered. In order to improve the recognition accuracy rate, it is necessary in post-processing to use language knowledge, which introduces context information, to correct the image recognition results. Post-processing approaches based on language knowledge include using a lexicon [1][2] or some syntax and semantic rules [3] to correct the spelling of words, and using some statistical language models (SLM) [4][5] to select out the best sequence from the candidate characters given by the OCR engine. Because of the complexity of language, all kinds of language knowledge sometimes are used together to obtain better performance [6].

An OCR engine outputs not only candidate characters, but also candidate distance information of each candidate character, which is also important in OCR post-processing. Currently, candidate distance is usually transformed to reliability of the corresponding candidate character to be utilized. Generally speaking, the bigger the reliability of a candidate character, the smaller the corresponding candidate distance. In early period, the reliability was calculated by using some empirical formulas [7][8]. Afterwards, a statistical approach was proposed [9], which calculates the reliability according to the distribution of candidate characters and correct characters with different candidate distances. It reflects some statistical characteristics, and its complexity is low, therefore it achieves good results in some applications. However, the use of candidate distance is still limited in OCR post-processing.

R. Khosla et al. (Eds.): KES 2005, LNAI 3681, pp. 346–352, 2005.

This paper proposes an OCR post-processing approach based on multiknowledge, which integrates language knowledge and candidate distance information. In this approach, statistical language model and semantic lexicon are combined, and candidate distance information is used to reduce the size of search space. The experimental results show that this approach is very effective. After post-processing, the recognition accuracy rate on the test set increases from 58.45% to 83.73%, which means 60.84% error reduction.

This paper is organized as follows. In Section 2, the multi-knowledge based approach that integrates language knowledge and candidate distance information is introduced. The experimental results are given in Section 3. In Section 4, some conclusions and future works are given.

2 The Multi-knowledge Based Approach

In the proposed approach, first the candidate distance information is used to construct the search space, where only the candidates whose distances satisfy a specific condition are added into the space; and then, statistical language model and semantic lexicon are used on the search space, with Viterbi algorithm to find the best sequence as the result as usual. In the following, statistical language model, semantic lexicon and use of candidate distance information will be introduced respectively.

2.1 Statistical Language Model

The most widely used statistical language model in OCR post-processing is n-gram model, which supposes that the occurrence probability of a word w_i is only related with the previous $n-1$ words, i.e.

$$P(w_i|w_1^{i-1}) = P(w_i|w_{i-n+1}^{i-1}) \tag{1}$$

There are some other models based on n-gram [10], one of which is distance-m n-gram model [11]. This model is the same as n-gram except using the history information with an interval of $m-1$ to the word w_i, i.e.

$$P(w_i|w_1^{i-1}) = P(w_i|w_{i-m-n+2}^{i-m}) \tag{2}$$

Distance-m n-gram model can utilize long-distance history information. However, it is reported that its precision decreases quickly along with the increase of the value of m [11]. Therefore, it is often combined with conventional n-gram model to maintain the performance as well as to use more contextual information.

In our experiment, six different models were used, including bigram, trigram, bigram+trigram, bigram+distance-2 bigram, trigram+distance-2 trigram and bigram+trigram+distance-2 bigram+distance-2 trigram. The last four models are combined via linear interpolation. For example, trigram+distance-2 trigram model can be described as

$$P(w_i|w_{i-3}w_{i-2}w_{i-1}) = \lambda_1 P_1(w_i|w_{i-2}w_{i-1}) + \lambda_2 P_2(w_i|w_{i-3}w_{i-2}) \tag{3}$$

where $P_1(w_i|w_{i-2}w_{i-1})$ and $P_2(w_i|w_{i-3}w_{i-2})$ correspond to the trigram model and the distance-2 trigram model respectively, and λ_1, λ_2 are the weight parameters which satisfy $\lambda_1 + \lambda_2 = 1$.

2.2 Semantic Lexicon

Many applications in natural language processing require extensive common language knowledge. Recently, there are many attempts to overcome the lack of available large knowledge bases by using semantic lexicon (SL). The most famous semantic lexicon is WordNet [12], which was developed by Princeton University. In WordNet, entries are sets of synonyms, called synsets, each of which represents a concept. Between the concepts there are semantic relations. For example, the most common relation between nouns is the ISA (is-a) relation that organizes the noun synsets into hierarchies.

Similar semantic lexicons for other languages have been developing by many countries, such as EuroWordNet for Italian, Spanish etc [13], KoreaNet for Korean [14], and HowNet [15] for Chinese. In our experiment, a sub-lexicon of HowNet was used, which describes relations between nouns.

2.3 Use of Candidate Distance Information

The analysis of distribution of candidate distance shows that it is almost impossible for a candidate to be the correct character when its candidate distance is too much bigger than the distance of the corresponding first candidate. On the other hand, the first candidate distance reflects the reliability of the first candidate character. The bigger the first candidate distance, the more possible that the correct character is not the first candidate, and the distance difference between the correct character and the first candidate may be more.

Based on the above analysis, an approach to reduce the size of search space by using the candidate distance information can be proposed. For the candidates of the same character, the first candidate distance d_1 is used to obtain the threshold $th = f(d_1)$, where $f(d)$ is called threshold function. If and only if another candidate distance d_i satisfies the condition $d_i - d_1 < th$, the corresponding candidate is added into the search space. Here the threshold function can be selected differently, but it should be an increment function so that when the first candidate distance increases, the number of candidate characters added into the search space may increase, which accords with the analysis proposed previously.

Generally speaking, the size of the reduced space is much smaller than that of the original space that contains all the candidates given by the OCR engine. Therefore it can speed up the search process significantly.

Figure 1 is an example of Chinese text, where each line contains six candidates for the same character and the number following every candidate is the corresponding candidate distance. When the threshold function is selected as $f(d) = 0.1d$, only underlined candidates are in the reduced space, while all the candidates are in the original space. The difference between them is obvious.

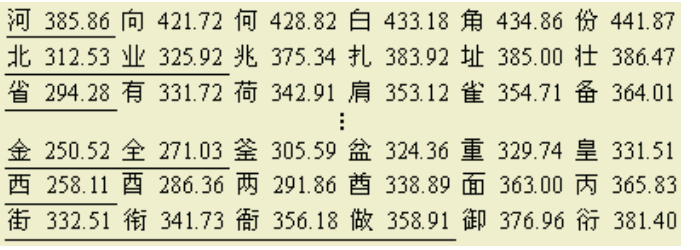

Fig. 1. Comparison of the reduced space and the original space.

3 Experiment

3.1 Data

We put the multi-knowledge based approach into Chinese address OCR post-processing. A training corpus containing 196,009 Chinese addresses is used to train the character-based n-gram and distance-2 n-gram models. Another corpus containing 20,000 Chinese addresses is used as a held-out data to optimize the weight parameters of linear interpolation language models. The test set is a corpus containing 15,000 handwritten Chinese addresses. After OCR, ten candidates for each character and the corresponding candidate distances are given. The evaluation criterion is the recognition accuracy rate of whole address, and the baseline on the test set is the recognition accuracy rate when results consist of the first candidates, which is 58.45% (8768 correct addresses).

3.2 Results

The results are shown in Table 1, where B, T, DB and DT denote bigram, trigram, distance-2 bigram and distance-2 trigram respectively, and the threshold function was selected as $f(d) = 0.1d$.

Table 1. The results with different n-gram types.

N-gram type	SLM with original space	SLM with reduce space	SLM-SL with original space	SLM-SL with reduce space
B	54.87%	70.49%	61.33%	79.64%
T	68.17%	**73.92%**	75.91%	83.13%
B+T	67.73%	73.63%	75.80%	82.97%
B+DB	55.07%	70.37%	62.41%	80.51%
T+DT	68.23%	73.23%	77.29%	83.71%
B+T+DB+DT	**68.23%**	73.25%	**77.29%**	**83.73%**

The advantages of the multi-knowledge based approach are shown from two sides. Firstly, the comparison between column 2 and column 4, or column 3 and column 5 in Table 1 shows that for every SLM, the recognition accuracy

rate with semantic lexicon used is higher than that without semantic lexicon used. The reason is that the semantic lexicon introduces more effective language knowledge. Especially, APO (a-part-of) relation between nouns is mainly used here. In Chinese, there is such a relation between place-names in an address (a latter place-name is a part of a former place-name in geographical meaning). The relation can help to correct the search result very effectively, so that the recognition accuracy rate increases. Secondly, the comparison between column 2 and column 3, or column 4 and column 5 in Table 1 shows that in every instance, the recognition accuracy rate on the reduced space is higher than that on the original space. The fact indicates that the search space can influence the performance of a language model significantly. Because there is only one or no correct candidate for each character, if all the candidates are used, there must be lots of wrong candidates in the search space. The effect of them is like much noise in speech recognition, and these wrong candidates will probably lead the search along a wrong direction. Among the language models we used, bigram model and bigram+distance-2 bigram model are the most imprecise, so they are influenced badly. When using SLM only, the results of them are even worse than the baseline. However, when using candidate distance to reduce the size of search space, many wrong candidates are deleted, thus the "noise" in recognition decreases, and the recognition accuracy rate is improved clearly.

In addition, use of the reduced space can speed up the post-processing significantly. For example, with trigram model, when using the original space, 6 hours is needed for the test set on a computer with 3.2GHz CPU, i.e., 1.44 seconds for per address in average. While using the reduced space, only 3 minutes is needed on the same computer, i.e., 0.012 seconds for per address in average. The speed is improved by 120 times or so.

From the above results, it can be seen that the multi-knowledge based approach (column 5 in Table 1) integrates the advantages of using both semantic lexicon and reduced search space, so that it is very effective. On the test set, the best recognition accuracy rate increases from 58.45% to 83.73%, which means 60.84% error reduction.

3.3 Discussion

When using candidate distance to reduce the size of search space, it can be noticed obviously that different threshold function can make different search space and then influence the result of post-processing. Here we used positive proportion function $f(d) = ad$ as the threshold function, and observed the results on the same test set with changing the proportion coefficient a. The results when using trigram and semantic lexicon are shown in Figure 2.

From Figure 2, it can be seen clearly that along with the increase of the proportion coefficient a, the recognition accuracy rate first increases, and then decreases. When $a = 0.1$, the best recognition accuracy rate 83.13% is achieved. The reason is that along with the increase of the proportion coefficient, more and more candidates are added into the search space. At the beginning, the candidates added can bring correct characters, so that the chance of selecting

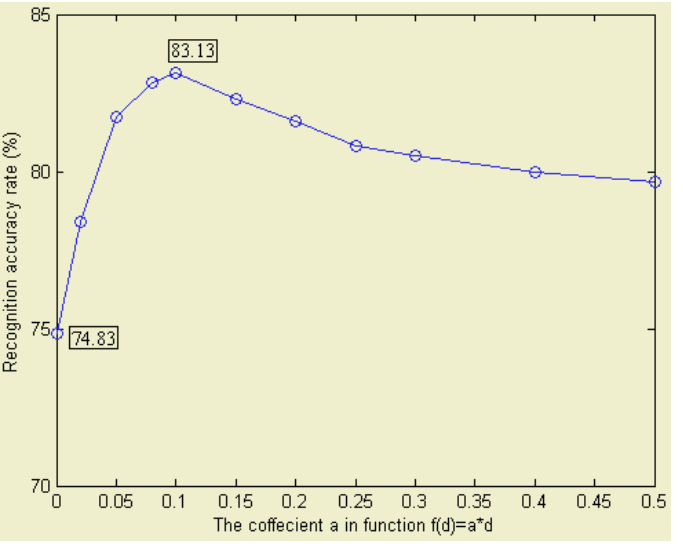

Fig. 2. Recognition accuracy rate to different proportion coefficient a.

out the correct result increases, so does the recognition accuracy rate. However, after adding an amount of candidates, the "noise" in the search space increases quickly, which results in the decrease of the recognition accuracy rate.

Besides the influence of different threshold function, it also can be noticed that the precision of the reduced space depends on the original image recognition result. If the precision of the original result is poor, most correct characters are in the back of the candidate list, and the distance difference between correct character and the first candidate may be much. Thus many correct candidates will not be in the reduced space, and the post-processing result may be very poor. Therefore, this approach cannot be used for a bad OCR engine. Fortunately, most applied OCR engines are good enough to satisfy the requirement of this approach, which provides good condition for using it widely.

4 Conclusion and Future Work

This paper proposes an OCR post-processing approach based on multi-knowledge, in which statistical language model and semantic lexicon are combined, and candidate distance information is used to reduce the size of search space. In this approach, more semantic information and a smaller and more precise search space are used, which makes it very effective. In an application of Chinese address OCR post-processing, the recognition accuracy rate on the test set increases from 58.45% to 83.73%, which means 60.84% error reduction.

There are two key points in this approach: the language model and the threshold function. In this paper, only the simplest forms of them are discussed. In the future, more research about them will be carried out, including using some more

precise language models, and looking for a proper function according to some analysis of original recognition results. In addition, we will test the approach in processing some general contents.

Acknowledgements

The authors are thankful to Fujitsu Laboratories Ltd. for providing the experimental data and partly supporting the work. Moreover, this work is supported by the Natural Science Foundation of China (Grants No. 60272019 and 60321002).

References

1. Wimmer Z., Dorizzi B., "Dictionary preselection in a neuro-Markovian word recognition system", Proceedings of the 5th International Conference on Document Analysis and Recognition, Bangalore, India, 1999: pp.539-542.
2. Procter S., Illingworth J., "Mokhtarian F. Cursive handwriting recognition using hidden Markov models and a lexicon-driven level building algorithm", Proceedings of the IEE Vision, Image and Signal Processing 2000, 147(4): pp.332-339.
3. Marti U., Bunke H., "A full English sentence database for off-line handwriting recognition", Proceedings of the 5th International Conference on Document Analysis and Recognition, Bangalore, India, 1999: pp.705-708.
4. Brakensiek A., Willett D., Rigoll G., "Unlimited vocabulary script recognition using character n-grams", Proceedings of the 22nd DAGM Symposium, Tagungsband Springer-Verlag, Kiel, Germany, 2000: pp.436-443.
5. Zhuang L., Bao T., Zhu X.Y., "A Chinese OCR spelling check approach based on statistical language models", Proceedings of the IEEE International Conference on System, Man and Cybernetics, Hague, Netherlands, 2004: pp.4727-4732.
6. Golding A.R., Schabes Y., "Combining trigram-based and feature-based methods for context-sensitive spelling correction", Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, CA, 1996: pp.71-78.
7. Lee H.J., Tung C.H., Chang Chien C.H., "A Markov language model in Chinese text recognition", Proceedings of the 2nd International Conference on Document Analysis and Recognition, Tsukuba, Japan, 1993: pp.72-75.
8. Tung C.H., Lee H.J., "Increasing character recognition accuracy by detection and correction of erroneously identified characters", Pattern Recognition, 1994, 27(9): pp.1259-1266.
9. Li Y.X., Zhu X.Y., "Post-processing for handwritten Chinese address recognition", Proceedings of the International Conference on Intelligent Information Technology, Beijing, China, 2002.
10. Goodman J.T., "A bit of progress in language modeling", Computer Speech and Language, 2001, 15(4): pp.403-434.
11. Rosenfeld R., "A maximum entropy approach to adaptive statistical language modeling", Computer Speech and Language, 1996, 10(3): pp.187-228.
12. http://wordnet.princeton.edu/
13. http://www.dcs.shef.ac.uk/research/groups/nlp/funded/eurowordnet.html
14. Moon Y., "Design and implementation of WordNet for Korean nouns", Journal of the Korea Information Science Society, 1996, 2(4): pp.437-445.
15. http://www.keenage.com/