2011 International Conference on Document Analysis and Recognition

Trie-Lexicon-Driven Recognition for On-line Handwritten Japanese Disease Names using a Time-synchronous Method

Bilan Zhu and Masaki Nakagawa
Department of Computer and Information Science,
Tokyo University of Agriculture and Technology,
Tokyo 184-8588, Japan
{zhubilan, nakagawa}@cc.tuat.ac.jp

Abstract— **This paper describes a lexicon-driven approach to on-line handwritten Japanese disease name recognition using a time-synchronous method. A trie lexicon is constructed from a disease name database containing 21,713 disease name phrases. It expands the search space using a time-synchronous method and applies the beam search strategy to search into a segmentation candidate lattice constructed based on primitive segments. This method restricts the character categories for recognizing each character candidate pattern from the trie lexicon of disease names and preceding paths during path search in the segmentation candidate lattice, and selects an optimal disease name from the disease name database as recognition result. The experimental results demonstrate the effectiveness of our proposed method, which improves character recognition rate from 94.56% to 99.97% compared with a general-purpose Japanese text recognizer and speeds up recognition time as 4.3 times faster as the general recognizer.**

Keywords- On-line recognition; disease name recognition; character recognition

I. INTRODUCTION

Recently, electronic clinical record systems are being introduced into hospitals and clinics. Keyboard-based systems are the mainstream. They are expanding quickly. However, they are hard for elder doctors and require doctor's attention even for young doctors so that they often disrupt the communication between patients and doctors. Therefore, pen-based systems are also being developed. The electronic clinical record systems that recognize and process the electronic clinical records inputted from digital pens are demanded so that we need to design and construct a disease names recognizer for the clinical record systems.

Applying a general-purpose on-line handwritten Japanese text recognizer [1] to recognize disease names would cause misrecognitions that are not disease names registered in the database. Actually disease name phrases are specific words, and domain specific methods are more effective. Address recognition is most popular. There can be many distinct recognizers for varieties of specific domains but we need to construct a specific recognizer for each domain to realize high recognition performance. If we could construct a generic method to construct domain specific recognizers, it would help construct them.

We can look each disease name phrase as a word and apply English word recognition methods to recognize it. However, English word recognition methods are apt to consider a model for each word and compare a word pattern with all the words to recognize the word pattern with the result that large processing costs are consumed. Gunter et al. [2] and Liwicki et al. [3] use a HMM based English word recognition method where each character is modeled as a HMM, word HMMs are constructed by concatenating character HMMs during recognition and an input word pattern is compared with models of all the words. Shetty et al [4] proposes a segmentation based English word recognition method using conditional random fields that also needs to compare an input word pattern with all the words.

The most similar technique to disease name recognition is address recognition, in which the postal addresses are also domain specific. Fu et al. [5] apply a general-purpose string recognition method to Chinese address recognition, based on the fact that Chinese addresses are not so restricted.

The expansion methods of the search space for string recognition can be divided into two classes: character-synchronous methods and time-synchronous methods [6]. The former expands the search space with the same depth of characters while the latter expands the search space with the same segment. Liu et al. [7] present a trie-based Japanese address recognition system which expands the search space using a character-synchronous method with the beam search strategy to select the most plausible paths. Koga et al. [8] use a similar trie-lexicon-driven method for Japanese addresses.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.228

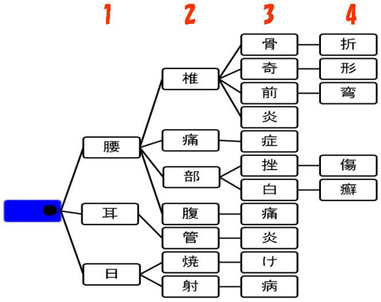


Figure 1. A portion of the trie lexicon of disease names.

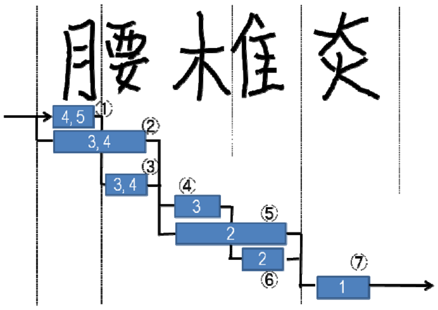


Figure 2. Segmentation candidate lattice.

We are inspired with the Japanese postal address recognition method [7, 8] and extend it to expand the search space in the segmentation candidate lattice using a time-synchronous method for on-line handwritten Japanese disease name recognition. We restrict the character categories of recognizing each character candidate pattern from the trie lexicon of disease names and preceding paths, as well as the length of disease names. Our method proposed in this paper is generic to construct domain specific recognizers and it helps to construct a different domain specific recognizer by applying different database. Experimental results on the collected disease name database demonstrate the superiority of our proposed method.

The rest of this paper is organized as follows: Section 2 describes our proposed on-line handwritten Japanese disease name recognition method. Section 3 presents some investigations and statistical information for the disease name database. Section 4 presents the experimental results and Section 5 draws our conclusion.

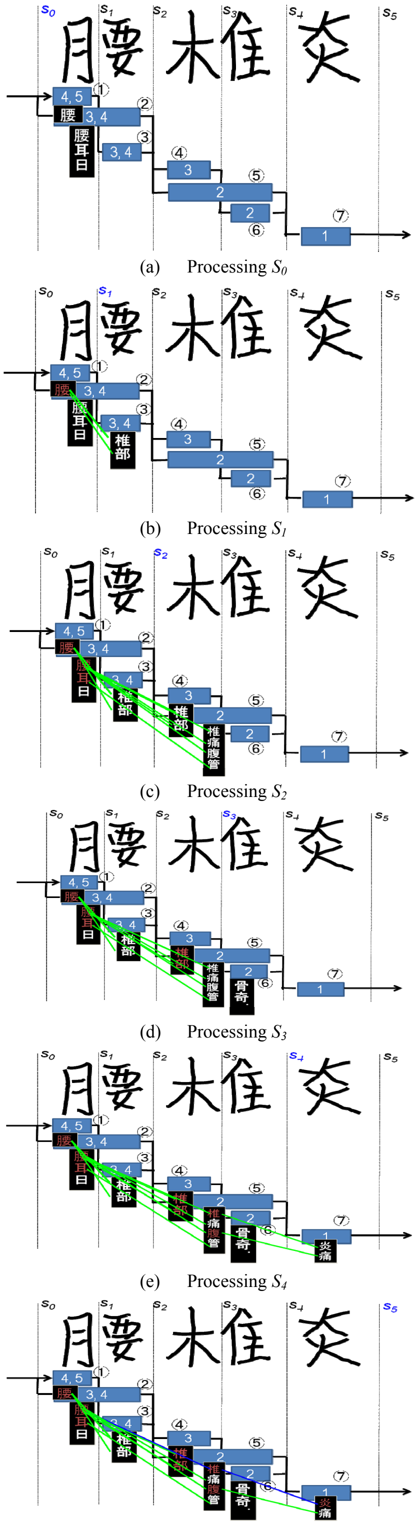
II. RECOGNITION

We construct a trie lexicon from disease name database which contains 21,713 disease name phrases as shown in Fig. 1.

Then we process each on-line handwritten disease name pattern as follows:

(1) **Segment candidate lattice construction.**

An on-line disease name pattern is over-segmented into primitive segments according to the features such as spatial information between adjacent strokes. Then one or more consecutive primitive segments form a candidate character pattern. The combination of all candidate patterns is



(a) Processing S_0

(b) Processing S_1

(c) Processing S_2

(d) Processing S_3

(e) Processing S_4

(f) Processing S_5

Figure 3. Search and recognition.

represented by a segmentation candidate lattice, where each node denotes a candidate character pattern and each arc denotes a segmentation point. Fig. 2 shows a segmentation candidate lattice that has seven nodes denoted by ①, ②, …, ⑦.

(2) **Setting possible length to the terminal for each node.**

Possible length to the terminal for each node is calculated as shown in Fig. 2 where the numbers shown in each node box are the possible length for the node. We can restrict the searched candidate paths by the possible length with the result of improved recognition accuracy and recognition speed.

(3) Search and recognition.

We apply the beam search strategy to search segmentation candidate lattice. The search processes are executed in order of the segmentation point. When searching this method restricts the character categories of recognizing each character candidate pattern from the trie lexicon of disease names and preceding paths, and the paths are evaluated for the likelihood of candidate segmentation and its string class according to the path evaluation criterion proposed in [1] that combines the scores of character recognition and geometric features (character pattern sizes, inner gaps, single-character positions, pair-character positions, candidate segmentation points) with the weighting parameters estimated by the genetic algorithm. This method selects an optimal disease name from the disease name database as recognition result.

Denote **X** = $x_1 \ldots x_n$ as successive candidate character patterns of one path, and every candidate character pattern x_i is assigned candidate class c_i. Then $f(\mathbf{X},\mathbf{C})$ is the score of the path $(\mathbf{X},\mathbf{C})$ where $\mathbf{C} = c_1 \ldots c_n$. The path evaluation criterion is expressed as:

$$f(\mathbf{X},\mathbf{C}) = \sum_{i=1}^{n} \left\{ \begin{array}{l} \sum_{h=1}^{5} [\lambda_{h1} + \lambda_{h2}(k_i - 1)] \log P_h \\ \lambda_{61} \log P(g_{j_i} \mid SP) + \lambda_{62} \sum_{j=j_i+1}^{j_i+k_i-1} \log P(g_j \mid NSP) \end{array} \right\} + n\lambda \quad (1)$$

where P_h, h=1,…,5, stand for the probabilities of $P(b_i|C_i)$, $P(q_i|C_i)$, $P(x_i|C_i)$, $P(p^u{}_i|C_i)$ and $P(p^b{}_i|C_{i-1}C_i)$, respectively. b_i, q_i, $p^u{}_i$ and $p^b{}_i$ are the feature vectors for character pattern sizes, inner gaps, single-character positions and pair-character positions, respectively. SP is the segmentation point and NSP is the non-segmentation point. g_i is the between-segment gap feature vector. k_i is the number of primitive segments contained in candidate character pattern x_i. λ_{h1}, λ_{h2} (h=1~6) and λ are the weighting parameters.

We show an example of searching in Fig. 3 and use it to describe our processes. We make the search in order from segmentation point S_0 to S_5.

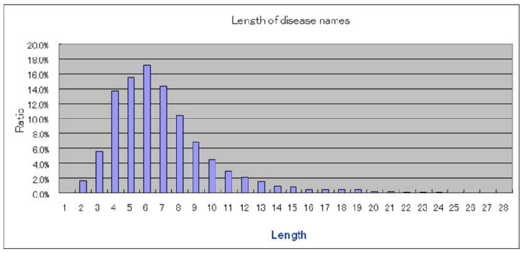


Figure 4. Lengths of disease names.

Firstly, we process the segmentation point S_0 as shown in Fig. 3 (a). The segmentation point S_0 is the start point and there is no node before it while there are two nodes (① and ②) after it. We also search the trie lexicon as shown in Fig. 1 from its start trie nodes. The start trie nodes are [腰], [耳] and [日] that can be set as character categories of ① and ②. For node ①, the possible lengths to the terminal are 4 and 5, and the trie nodes (character categories of recognizing) [耳] and [日] are erased because their lengths to the terminal are 3 and do not satisfy the possible lengths of node ①, and the category [腰] remains. We set its recognition score by a character recognizer. The node ② is set three character categories [腰], [耳] and [日] similarly.

Secondly, we process the segmentation point S_1 as shown in Fig. 3 (b). For the segmentation point S_1 there is a node ① before it and there is a node ③ after it. For all the nodes before each segmentation point, we evaluate all paths according to the path evaluation criterion proposed in [1] and sort them, and then only select several top paths and erase other paths. The number of the selected top paths is called as the beam band. In the instance of Fig. 3 the beam band is set as two and for S_1 there is only a path ending at [腰]. Then from the remaining preceding path we set the character categories of the node ③ after S_1. Form the remaining preceding path and the trie lexicon, character categories can be [椎], [痛], [部] and [腹], and [痛] and [腹] are erased from the possible lengths of ③ and [椎], [部] remains.

Then, we apply the same method to process the segmentation point S_2, S_3 and S_4 in order as shown in Fig. 3 (c), (d), (e), respectively.

Lastly, we process the last segmentation point S_5 as shown in Fig. 3 (f). For all the nodes before the last segmentation point, we evaluate all paths according to the path evaluation criterion proposed in [1] and sort them, and then select the optimal path as the recognition result.

III. Investigations and Statistical Information

We show some investigations and statistical information for the disease name database in the section as follows:

(1) Number of disease names.

The number of disease name phrases is 21,713.

(2) Lengths of disease names.

Fig. 4 shows the lengths of disease names (lenghs of characters). The average length is 6.9.

(3) Numbers of divergences of the trie nodes.

The number of the start trie nodes is 1,345. Fig. 5 shows divergences of the trie nodes (the number of alternatives at the n-th position from the start). We can see that the numbers of the character categories is greatly restricted by the trie lexicon, and restricting character categories of recognizing each node from the trie lexicon can decrease largely the amount of recognitions from thousands of character categories and the misrecognition from similar characters.

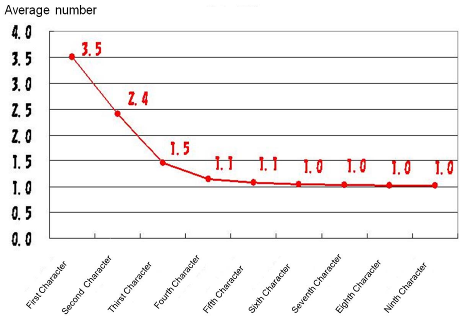


Figure 5. Numbers of divergences of the trie nodes.

IV. EXPERIMENTS

We trained our character recognizer and geometric scoring functions using a Japanese on-line handwriting database Nakayosi [9]. We also trained the SVM classifier for the candidate segmentation point probability and the weighting parameters of path evaluation score using the database Kondate. The details for training are described in [1].

After training, we use 1,112 handwritten disease name samples that contain 3,803 characters to evaluate the proposed disease name recognition method. The experiments were implemented on a Genuine Intel(R) CPU U1400 1.20 GHz with 1.49 GB.

We compare the performance of our method proposed in this paper with that of a general-purpose Japanese text recognizer [1]. For fair comparison, the two methods use the same classifiers for character recognition and geometric context. For the general-purpose Japanese text recognizer the path evaluation use language context score from tri-gram instead of using the trie lexicon, and the tri-gram table is prepared from the year 1993 volume of the ASAHI newspaper and the year 2002 volume of the NIKKEI newspaper. Table 1 shows the recognition results. The recognition time is the time for recognizing all disease names.

Table 1. Recognition results

	Recognition rate	Recognition time
Disease name recognizer	99.97%	9m37s
General-purpose recognizer	94.56%	40m33s

From the results, we can see that the disease name recognizer improved character recognition rate largely from 94.56% to 99.97% and is 4.3 times faster compared with the general-purpose Japanese text recognizer.

For disease name recognizer there is a single misrecognition as shown in Fig. 6 that is because the correct category is not in the dictionaries of the character recognizer and geometric scoring functions.

Fig. 7 shows some examples recognized correctly. We can see that disease name [うっ血肝] is correctly recognized by the disease name recognizer while it is incorrectly recognized by the general-purpose recognizer. That is because the disease name recognizer decreases largely misrecognition from similar characters by restricting the character categories of recognizing each character pattern. Moreover, even for the disease name patterns not written correctly such as [18 常染色体異常] as shown in Fig. 7, the disease name recognizer can also correctly recognize them.

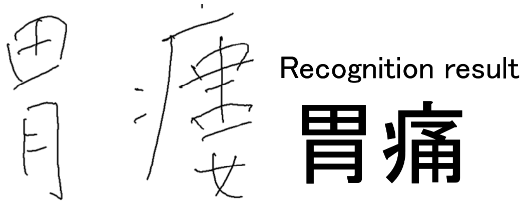


Figure 6. Misrecognition.

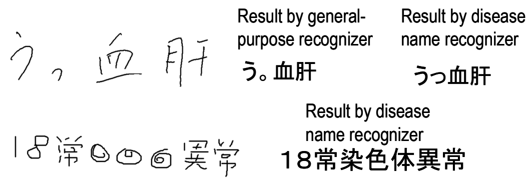

Result by general-purpose recognizer
う。血肝
Result by disease name recognizer
うっ血肝
Result by disease name recognizer
18常染色体異常

Figure 7. Correctly recognized results.

V. CONCLUSION

This paper presented a lexicon-driven approach to on-line handwritten Japanese disease name recognition, using the beam search strategy for path search in the segmentation candidate lattice. By restricting the character categories with the trie-lexicon, recognition errors from similar characters are reduced dramatically compared with the general-purpose text recognizer. The recognition speed is also improved remarkably. Even for the disease name patterns not written correctly, the disease name recognizer can also correctly recognize them.

ACKNOWLEDGMENT

This work is being partially supported by the R&D fund for "development of pen & paper based user interaction" under Japan Science and Technology Agency.

REFERENCES

[1] B. Zhu, X.-D. Zhou, C.-L. Liu and M. Nakagawa, "A Robust Model for On-line Handwritten the Japanese Text Recognition," International Journal on Document Analysis and Recognition (IJDAR), Vol. 13, No. 2, pp.121-131, 2010.

[2] S. Gunter and H. Bunke, "HMM-based handwritten word recognition: on the optimization of the number of states, training Iterations and Gaussian components," Pattern Recognition, 37, pp. 2069-2079, 2004.

[3] M. Liwicki and H. Bunke, "HMM-based On-line Recognition of Handwritten Whiteboard Notes," Proc. 10th Int'l Workshop on Forntiers in Handwriting Recognition (IWFHR), pp. 595-599, 2006.

[4] S. Shetty, H. Srinivasan, S. Srihari, "Handwritten word recognition using conditional random fields," Proc. 9th ICDAR, pp. 1098-1102, 2007.

[5] Q. Fu, X. Q. Ding, T. Liu, Y. Jiang, and Z. Ren, "A novel

segmentation and recognition algorithm for Chinese handwritten address character strings," 18th Int. Conf. Pattern Recognition, pp. 974-977, 2006.

[6] M. Cheriet, N. Kharma, C.-lin Liu and C. Y. Suen, Character Recognition Systems: A Guide for Students and Practioners, John Wiley & Sons, Inc., Hoboken, New Jersey, 2007.

[7] C.-L. Liu, M. Koga, and H. Fujisawa, "Lexicon-Driven Segmentation and Recognition of Handwritten Character Strings for Japanese Address Reading," IEEE Trans. Pattern Analysis and Machine Intelligence, 24(11), pp.1425-1437, 2002.

[8] M. Koga, R. Mine, H. Sako, and H. Fujisawa, "A lexicon driven approach for printed address phrase recognition using a trie dictionary," (in Japanese) IEICE Trans. Information and Systems, J86-D (2), pp.1297-1307, 2003.

[9] M. Nakagawa, K. Matsumoto, "Collection of on-line handwritten Japanese character pattern databases and their analysis," International Journal on Document Analysis and Recognition (IJDAR), 7(1), pp.69-81, 2004.