Recognition of Japanese historical text lines by an attention-based encoder-decoder and text line generation

Anh Duc Le[1], Daichi Mochihashi[2], Katsuya Masuda[3], Hideki Mima[3], Nam Tuan Ly[4]

1. The Center for Open Data in the Humanities, Japan, anhp@ism.ac.jp
2. The Institute of Statistical Mathematics, Tokyo, Japan, daichi@ism.ac.jp
3. Center for Research and Development of Higher Education, University of Tokyo, Japan, masuda@he.u-tokyo.ac.jp, mima@t-adm.t.u-tokyo.ac.jp
4. Tokyo University of Agriculture and Technology, Japan, namlytuan@gmail.com

ABSTRACT

Inspired by the recent successes of attention based encoder-decoder (AED) approach on image captioning, machine translation, we present an AED model as an end-to-end recognition system for recognizing Japanese historical documents. The recognition system has two main modules: a dense convolution neural network for extracting features, and a Long Shor Term Memory (LSTM) decoder integrating with attention model for generating target text. We can train the model end-to-end. The model requires only input text line images and corresponding output characters. Therefore, we don't need annotations for characters and save a lot of time for making annotations. We also present a method to generate artificial text lines to solve the imbalance problem of the current annotated database. The results of experiments on the annotated and artificial databases demonstrate the effectiveness of the text line generation. Our recognition system achieved Character Error Rate of 23.76% and 22.52% by training with and without artificial text lines, respectively. Moreover, our recognition system outperforms the CNN-LSTM system, which achieved the state-of-art results in other document recognition tasks.

Applied computing → Document management and text processing → Document capture → Document analysis

Keywords

Japanese historical documents; text line recognition; text line generation; attention model; encoder-decoder approach

ACM Reference format:

Anh Duc Le, Daichi Mochihashi, Katsuya Masuda, Hideki Mima, Nam Tuan Ly. 2019. Recognition of Japanese historical text lines by an attention-based encoder-decoder and text line generation. In Proceedings of *The 5th International Workshop on Historical Document Imaging and Processing, Sydney, NSW, Australia, September 20-21, 2019 (HIP 2019), 5 pages.*

https://doi.org/10.1145/3352631.3352641

HIP '19, September 20–21, 2019, Sydney, NSW, Australia
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7668-6/19/09..$15.00
https://doi.org/10.1145/3352631.3352641

1. INTRODUCTION

Since historical documents are an invaluable resource for historians in exploring social aspects, lifestyles, even weather in the previous era, many countries have been preserved their historical documents. The popular method is to construct digital libraries to preserve historical documents and made them available to the public. This helps researchers from domestic and abroad to access historical documents. The traditional method is scanning books to images. Then, experts read them and provide transcriptions. This approach is labor-intensive, time-consuming, and requiring many experts. So that, it is not feasible to provide transcription for a large number of historical documents in libraries. Document analysis and recognition can speed up the digitalization process.

The goal of this research is to research and develop a recognition system with high accuracy for historical documents. The project will inherit the previous researches in recognition of historical documents and latest deep learning such as attention model, encoder-decoder approach to build an end-to-end recognition system.

For historical documents, Kim et al. [1] developed a recognition system for handwritten Hanja historical documents in Korea. In China, a huge project was led by Digital Heritage Publishing Ltd. [2] to digitize more than 4.7 million pages of Siku Quanshu (the largest collection of books on Chinese history was written during the Qianlong period (1711–1799)). Li et al. [3] adopted Style Transfer Mapping method to improve historical Chinese character recognition. Font characters are employed for training to improve the recognition system. Phan et al. [4] developed a document layout analysis and recognition for Nom script (a character system for Vietnamese from the end of the tenth century to the beginning of the twentieth century). They also employed font characters and handwritten characters to improve the performance of the recognition system. In Japan, there are several projects for recognizing historical documents. Horiuchi et al. [5] employed modular neural networks to recognize Kuzusi characters. Nguyen et al. [6] also presented a recognition system to recognize multiple sentences of Kuzusi characters based on CNN and BiLSTM. Masuda et al. [7] proposed a method to use massive text data to reduce errors of character recognition. However, all of them are developed for specific historical documents and do not open to the research community. In this research, we aim to build an open source recognition system for Japanese historical documents. The recognition system should be easy to train and reuse by other researchers from document analysis and digital humanities fields.

The document recognition system has two main steps: page layout analysis and text recognition. Layout analysis analyzes documents to paragraphs and text lines. Then, text recognition recognizes each text line and provide character codes. Commercial Optical

Character Recognition has good performance on modern printed and handwritten document. However, they are still insufficient for historical documents. The challenges are from complex and various layouts, degraded and damaged documents, large of vocabulary, out of vocabulary characters, and so on. Moreover, the critical problem is the lack of annotated documents for training a recognition system. In this paper, we focus on text recognition step. To overcome these challenges, we employ the attention-based encoder-decoder model (AED) to develop an end-to-end recognition system. AED recognition system does not require the character segmentation process like traditional text recognition methods. It has been applied for recognizing handwritten mathematical expressions [8], [9], handwritten Vietnamese [10], and scene text [11]. The system is easy to train because it requires only images and corresponding transcriptions without character annotation. It also speedups the annotation process.

In this paper, we investigate the AED model for recognizing Japanese historical documents. We do initial experiments and report the results on the annotated text lines. Then, we proposed a method to generate artificial text lines to solve the imbalance problem of the current annotated database. We show the effectiveness of the AED recognition system and artificial text lines in the experiment section.

The rest of this paper is organized as follows. The AED recognition system for historical documents is presented in Section II. The historical text lines dataset and text line generation method are presented in Section III. Experimental results are present in Section IV, and conclusions are drawn in Section V.

2. OVERVIEW OF THE ATTENTION BASED ENCODER-DECODER RECOGNITION SYSTEM

The structure of the AED model is shown in Figure 1. This is similar to the end-to-end recognition system for handwritten mathematical expression [8, 9]. It has two main modules: a dense convolution neural network for feature extraction from a historical text line image, and an LSTM decoder with an attention model for generating the target text. They are described in the following sections.

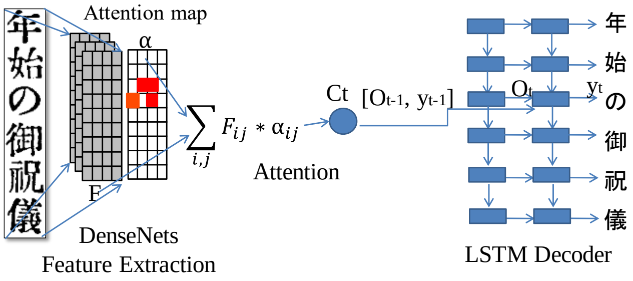

Figure 1. Structure of the attention based encoder-decoder model for recognizing historical documents.

2.1 DenseNet feature extraction

The related works [9, 13] have verified that DenseNet outperforms the VGG and ResNet by proposing direct connections from any preceding layers to succeeding layers. The i^{th} layer receives the feature-maps of all preceding layers, $x_0, \ldots, x_{i-1}$, as input:

$$\boldsymbol{x_i} = \boldsymbol{H_i}([\boldsymbol{x_0}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_{i-1}}]) \qquad (1)$$

where H_i refers to the convolutional function of the i^{th} layer and $[x_0, x_1, \ldots, x_{i-1}]$ refers to the concatenation of the output of all preceding layers. Dense connections help the network reuse and learn features of cross layers. However, dense connections require more memory because of the growth of connections through depth. To limit connections and keep the same input size, DenseNet is divided into densely connected blocks as Figure 2. In each dense block, we add a blottleneck layers (1x1 convolution layer) before the 3x3 convolution layer to reduce the computational complexity. The dense blocks are connected by transition layers which contain convolutional and average pooling layers. For compression, the transition layer reduces a half of feature maps. The detailed implementation is described as follows. We employ a convolutional layer with 48 feature maps and a max pooling layer to process input image. Then, We employ three dense blocks of growth rate (output feature map of each convolutional layer) $k = 24$ and the depth (number of convolutional layers in each dense block) $D = 16$ to extract features. The size of the output features is *H*x*W*x*C*.

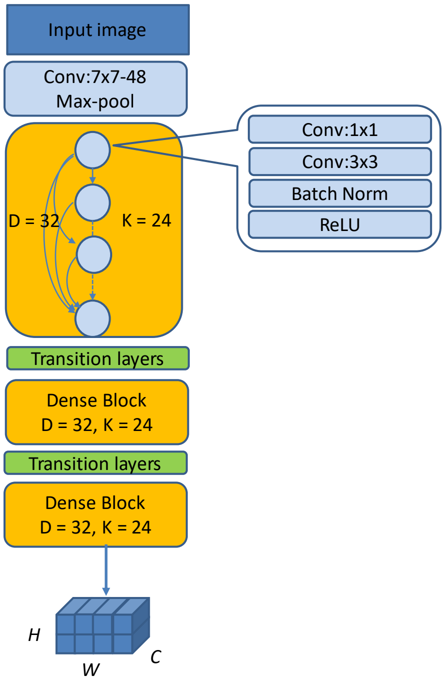

Figure 2. The architecture of DenseNet for features extraction.

2.2 Attention-based LSTM decoder

The LSTM decoder generates one character at a time. At each time step *t*, the decoder predicts symbol y_t based on the current output O_t as the following equations:

$$\boldsymbol{p}(\boldsymbol{y_t}|\, \boldsymbol{y_1}, \ldots, \boldsymbol{y_{t-1}}, \boldsymbol{F}) = \textbf{softmax}(\boldsymbol{W_{out}} * \boldsymbol{O_t}) \qquad (2)$$

O_t is calculated from the embedding vector of the previous decoded symbol E_{yt-1}, the currently hidden state of the decoder h_t, and the current context vector c_t.

$$\boldsymbol{O_t} = (\boldsymbol{E_{y_{t-1}}} + \boldsymbol{W_h} * \boldsymbol{h_t} + \boldsymbol{W_c} * \boldsymbol{c_t}) \qquad (3)$$

$$\boldsymbol{h_t} = \boldsymbol{LSTM}(\boldsymbol{h_{t-1}}, \boldsymbol{c_t}, \boldsymbol{E_{y_{t-1}}}) \qquad (4)$$

c_t is computed by a weighted sum of the input features *F* and the attention probability *α* produced by an attention model.

$$c_t = \sum\nolimits_{u,v} \alpha_{t(u,v)} * F_{u,v} \qquad (5)$$

$$\alpha_{t(u,v)} = \frac{exp\ (e_{t(u,v)})}{\sum_{i,j} \mathbf{exp}(e_{t(i,j)})} \qquad (6)$$

$e_{t(u,v)}$ denotes the energy of $F_{u,v}$ at time step t conditioned on the previous hidden state of the decoder h_{t-1} and coverage vector $Cov_{t(u,v)}$, and the feature $F_{u,v}$. The coverage vector is initialized as a zero vector, and we compute it based on the summation of all past attention probabilities α . The coverage vector contains information about the parts of input images that the recognition system recognized. Therefore, it help the decoder avoid repeated attention.

$$e_{t(u,v)} = v_{att}^T \tanh\left(\mathbf{W}_h * \mathbf{h}_{t-1} + \mathbf{W}_F * F_{u,v} + \mathbf{W}_{cov} * Cov_{t(u,v)}\right) \qquad (7)$$

$$Cov_{t(u,v)=} \sum\nolimits_{l=1}^{t-1} \alpha_{l(u,v)} \qquad (8)$$

We initialize the hidden of the decoder by zero vector. The character generation process is repeated until the decoder produces an end-of-line character.

2.3 Training

We employ cross-entropy as the objective function to maximize the probability p of predicted symbols as follows:

$$f = -\sum\nolimits_{i=1}^{|D|} \quad \sum\nolimits_{t=1}^{|L_i|} \mathbf{log}\, p(g_{i,t}|y_1, \ldots, y_{t-1}, F_i) \qquad (9)$$

where D is the text line training set, L_i is a text line, $g_{i,t}$ is a ground truth of the text line, and F_i is the output of DenseNet.

We use mini-batch stochastic gradient descent to learn the parameters. The initial learning rate is set to 0.1 and batch size is set to 8. The training process is stopped when the character error rate on the validation set does not improve after 15 epochs. The parameters for DenseNet is set as described in Section II. Based on the experiments of the related works on recognition of handwritten mathematical expression [8, 9], we set the size of the LSTM decoder to 512.

3. TEXT LINES GENERATION

3.1 Training dataset

We employ annotated documents from the related work [12] for this research. The dataset contains 922 pages from historical magazines in Japan from 1870 to 1945. We randomly select 80% of pages for training, 10% of pages for validation and the rest for testing. For each page, we extract text lines and their ground truth for experiments. The numbers of lines for training, validation, testing are shown in Table 1. The number of categories is 5,398 which contains many character categories that do not use in current Japanese character system. Figure 3 shows some example of text lines from the database. The database contains both vertical and horizontal text lines.

Table 1. Statistics of the Japanese historical documents dataset.

	# pages	# text lines	# characters
Training	736	30,812	603,355
validation	93	4,048	78,077
Testing	93	3,910	76,679

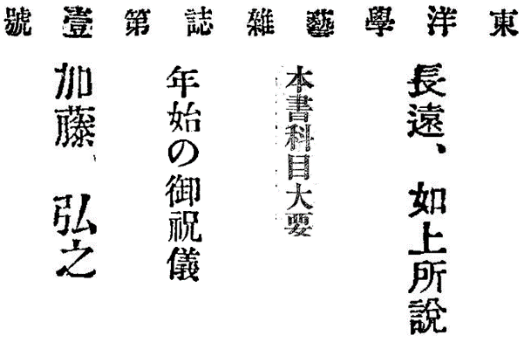
東洋學藝雜誌第壹號
長遠、如上所說
本書科目大要
年始の御祝儀
加藤、弘之

Figure 3. Text line examples in the Japanese historical document database.

The critical problem of the current database is an imbalance between character categories. We observe that there are 320 categories that have more than 400 characters (large samples set) and 2697 categories that have less than 10 characters (a few samples set) in the dataset. Since the current dataset is small and imbalanced, it is hard to train a deep learning model with such imbalanced data.

3.2 Text lines generation

For this purpose, we also propose a method to generate more training text lines. First, we prepare a Japanese historical text corpus (*TC*) and a set of isolated Japanese historical characters (*C*). Then, for each text line on *TC*, we take corresponding characters on *C* and concatenate them into a text line image. The detail of the pattern generation method is described below.

Japanese historical text corpus: Taiyo (The Sun) is a Japanese magazine published from 1895-1928. It is an important primary source for Japanese colonial studies, known for its literary criticism, literature, and translations of western authors. We extracted 65,020 lines from the text corpus.

Japanese historical characters set: We extracted isolated characters from training documents. For character categories having a small number of patterns, we add patterns generated from fonts. We also add Gaussian noise and employ distortion models such as scaling, rotation, and shear to make font image more natural. Figure 4 shows the process of generating font images.

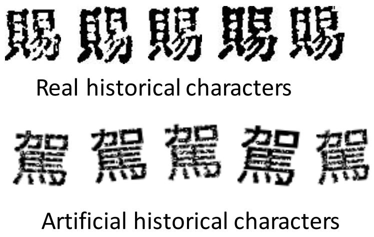

Figure 4. Samples of real and artificial historical characters.

Pattern generation process

1. Get a sentence *S* from the corpus.
2. For each character of the sentence *S*, an image of this character is randomly chosen from *C.*

3. Selected isolated characters are concatenated in horizontal or vertical directions. Since the majority of training data is vertical text lines, we randomly select 5% of sentences on the corpus to generate horizontal text lines and the remaining to generate vertical text lines.

Figure 5 shows some real and artificial text lines. Artificial text lines are pretty similar to real text lines.

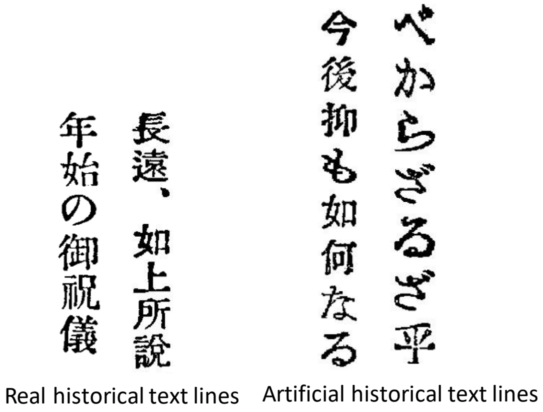
長遠、如上所説
年始の御祝儀
べからざるざ乎
今後抑も如何なる
Real historical text lines
Artificial historical text lines

Figure 5. Samples of real and artificial historical text lines.

4. EVALUATION

4.1 Databases

We employ real historical text lines extracted pre-modern Japanese historical documents and artificial historical text lines for evaluating the AED recognition system. The detail of the training data was described in the previous section. First, we compare the performance of the AED recognition system and the CNN-LSTM system trained by only real historical text lines, and real historical and artificial historical text lines. Then, we visualize the attention probabilities to show the recognition process.

4.2 Evaluation metric

In order to measure the performance of our system on handwriting recognition, we use the Character Error Rate (CER) metric which is generally employed for evaluating handwriting recognition systems. The detail of this metric is shown in the following equation:

$$\boldsymbol{CER} = \boldsymbol{100} * \frac{\sum_{i=1}^{n} ED(S_i, R_i)}{\sum_{i=1}^{n} |S_i|} \qquad (10)$$

where S_i is the *i-th* string belonging the set of target strings (ground truth text) S. R_i is the corresponding output string of the S_i string. $|S_i|$ is the number of words in S_i. *ED* is the edit distance function which computes the Levenshtein distance between two strings S_i and R_i.

4.3 Results

Table 2 shows the CER of our system and CNN-LSTM on the validation and testing sets. We employ CNN-LSTM system which achieved the state-of-art results in other document recognition tasks such as handwriting recognition [14] and scene text recognition [15]. We reused the setting provided in the related work for recognizing handwritten Japanese [14]. Without artificial text lines, we achieved 23.76% of CER on testing sets. With artificial text lines, we achieved 22.52% of CER on the testing sets. The CER of our system outperforms LSTM-CNN system.

Table 2. The results of AED and CNN-LSTM recognition systems on the testing set.

Method	**Training data**	**Testing(%)**
CNN-LSTM	Without artificial text lines	25.99
	With artificial text lines	23.00
AED	Without artificial text lines	23.76
	With artificial text lines	22.52

Figure 6 shows the recognition process of the AED recognition system. The system generates characters until it reaches the <end> symbol. At each time step, the decoder focuses on a part of the input image (red part in the image) to generate the corresponding character. We observed that the attention model provide very precise attention positions.

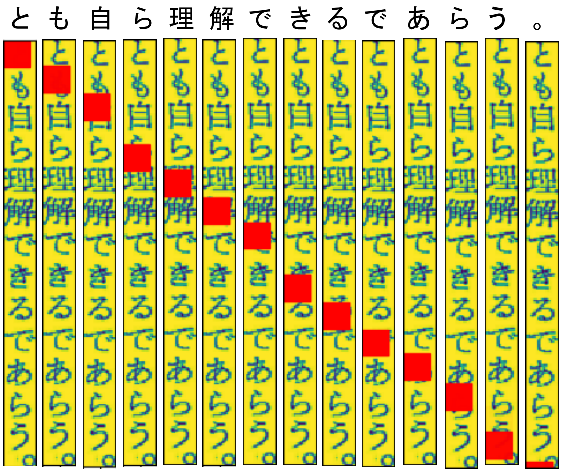

Figure 6. Visualization of the recognition process of the attention based encoder-decoder model.

Figure 7 shows an example of the recognition result. The historical document text line is in the left, and the recognition result is in the right. The miss-recognition is shown in red.

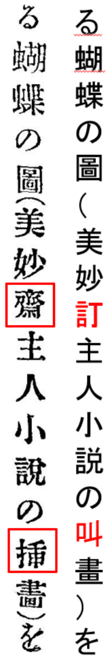
る蝴蝶の圖（美妙訂主人小説の叫畫）を
る蝴蝶の圖(美妙齋主人小説の插畫)を

Figure 6. An example of the recognition result.

5. CONCLUSION

In this paper, we have presented the attention based encoder-decoder model for recognizing Japanese historical documents. The model is trained end-to-end with input text line images and target characters. We also propose a text line generation method to solve the imbalance problem of the annotated database. We achieved 23.76% and 22.53 of CER on the testing set by training the recognition system without and with artificial text lines, respectively. We also visualized attention probabilities and observed the attention model provide very precise attention positions.

6. REFERENCES

[1] Kim, M.S., Jang, M.D., Choi, H.I., Rhee, T.H., Kim, J.H., Kwag, H.K.: Digitalizing scheme of handwritten Hanja historical documents. In: Proceedings of the 1st International Workshop on Document Image Analysis for Libraries, USA, pp. 321–327, Jan. 2004

[2] http://www.itventuresltd.com/eng/dhp.htm

[3] Bohan Li, Liangrui Peng, Jingning Ji, Historical Chinese Character Recognition method based on Style Transfer Mapping, 2014 11th IAPR International Workshop on Document Analysis Systems.

[4] Truyen Van Phan, Kha Cong Nguyen, and Masaki Nakagawa. A Nom historical document recognition system for digital archiving, International Journal on Document Analysis and Recognition (IJDAR), Vol. 18, pp. 1-16, (Dec. 2015)

[5] Tadashi Horiuchi, Satoru Kato, A Study on Japanese Historical Character Recognition Using Modular Neural Networks, 2009 Fourth International Conference on Innovative Computing, Information and Control.

[6] Hung Tuan Nguyen, Nam Tuan Ly, Cong Kha Nguyen, Cuong Tuan Nguyen, Masaki Nakagawa: Attempts to recognize anomalously deformed Kana in Japanese historical documents, Proc. of the 2017 Workshop on Historical Document and Processing, pp. 31-36, Kyoto, Japan (11.2017).

[7] Katsuya MASUDA, 大域的情報を用いた OCR 文字誤り訂正, 言語処理学会年次大会発表論文集 2015

[8] Anh Duc Le, Masaki Nakagawa, Training an End-to-End System for Handwritten Mathematical Expression Recognition by Generated Patterns, ICDAR 2017, pp. 1056-1061.

[9] Jianshu Zhang, Jun Du, and Lirong Dai, Multi-Scale Attention with Dense Encoder for Handwritten Mathematical Expression Recognition, ICPR 2018.

[10] Anh Duc Le, Hung Tuan Nguyen, Masaki Nakagawa, Recognizing Unconstrained Vietnamese Handwriting By Attention Based Encoder Decoder Model, ACOMP 2018

[11] Suman K. Ghosh, Ernest Valveny, Andrew D. Bagdanov, Visual attention models for scene text recognition, arXiv:1709.02054

[12] 永野雄大, 幡谷龍一郎, 持橋大地, 増田勝也, CNN を用いた近代文献画像からのテキスト領域抽出, PRMU 2018

[13] Huang, Gao and Liu, Zhuang and van der Maaten, Laurens and Weinberger, Kilian Q., Densely connected convolutional networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[14] Nam Tuan Ly, Cuong Tuan Nguyen, Masaki Nakagawa, Training an End-to-End Model for Offline Handwritten Japanese Text Recognition by Generated Synthetic Patterns, ICHFR 2018, pp. 74-79.

[15] Baoguang Shi, Xiang Bai and Cong Yao, An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition, IEEE Trans. Pattern Anal. Mach. Intell., Vol.39, pp. 2298--2304, 2017.