KuroNet Kuzushiji Recognition and Its Impact on Historical Big Data Research

*Asanobu Kitamoto[1,2], Tarin Clanuwat[1,2]

1. ROIS-DS Center for Open Data in the Humanities, 2. National Institute of Informatics

1. Introduction

Japan has carefully preserved historical materials such as rare books, documents, and records for over a thousand years, which makes it a rare country in the world that has this large amount of historical documents in the scale of hundred million items. However, most Japanese today can't read those materials anymore as they were written in "Kuzushiji". To solve this problem, we started research on Kuzushiji optical character recognition (OCR) using machine learning. This research will have a major impact on historical big data research, which aims to study Japanese history, culture and also natural phenomena such as past disasters based on the interpretation of historical materials.

2. KuroNet Kuzushiji Recognition

First, we released a large-scale Kuzushiji dataset (http://codh.rois.ac.jp/char-shape/). As of November 2019, this dataset includes 1,086,326 character images of 4,328 character types cropped from 44 classical books. Kuzushiji recognition community became vibrant as several research groups started Kuzushiji recognition by machine learning (deep learning).

Our group is also doing research on Kuzushiji recognition algorithm called KuroNet using object detection and recognition algorithms [1]. We had an idea that object detection and recognition technique could be used by regarding a Kuzushiji as an object. This idea was more successful than we expected and led to a birth of KuroNet, which is a robust Kuzushiji recognizer even in a complex layout.

In order to expand Kuzushiji recognition research to the world, we hosted Kaggle Kuzushiji recognition competition on the world biggest data science platform Kaggle. In this competition, machine learning researchers and engineers from around the world tackled the problem of Kuzushiji recognition, and most of the winning algorithms were also based on object detection and recognition algorithms [2]. The top five algorithms in the competition have already been open sourced, and we are currently working on incorporating these results into KuroNet.

KuroNet will be released as open source software in the future, but we decided to release first as a web service so that anyone can interact with. We made the Kuzushiji recognition web service available from IIIF Curation Viewer.

(1) KuroNet Kuzushiji recognition service for multiple characters: it can recognize multiple characters within a specified region, and show the result on IIIF Curation Viewer.

(2) KogumaNet Kuzushiji recognition service for a single character: it can recognize one character within a specified region and show the ranking of candidate characters. This service is based on TensorFlow.js that runs on a client-side browser so that it does not require any server-side process.

3. Impact on Historical Big Data research

KuroNet Kuzushiji recognition serves as a starting point for data structuring workflow in historical big data research. If KuroNet can generate plain text from digital images with Kuzushiji, it can be used as draft to aid human transcription, or the basic data for constructing semi-structured data by adding tags. To realize this plan, there remains several research challenges in addition to improving the accuracy of KuroNet.

First, since KuroNet gives recognition results as an unordered set of Unicode and coordinates of a character, serializing these characters into string sequence is not an easy task because some document has complex layout that is hard to order even for human. Second, building an ecosystem to further expand the dataset for machine learning is a critical issue in order to expand the application of KuroNet to handwritten historical documents. We believe that KuroNet can handle handwritten historical documents if we have substantial dataset. How to design an ecosystem for creating a high-quality and large-scale dataset is another challenge of KuroNet.

References

[1] Tarin CLANUWAT et al. "KuroNet: Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning", 15th International Conference on Document Analysis and Recognition (ICDAR2019), arXiv:1910.09433, 2019.

[2] Asanobu KITAMOTO et al. "Progress and Results of Kaggle Machine Learning Competition for Kuzushiji Recognition", Proceedings of IPSJ SIG Computers and the Humanities Symposium 2019, pp. 223-230, 2019.

Keywords: Kuzushiji, Historical Big Data, Machine Learning, Historical documents, KuroNet, Character Recognition

