ISSN 2436-5408 (Online)
ISSN 1346-2601 (Print)

Recognizing Kuzushiji in Japanese Historical Documents
——International ARC Seminar Review

Biligsaikhan Batjargal (Senior researcher, Kinugasa Research Organization, Ritsumeikan University)
E-mail biligee@fc.ritsumei.ac.jp

1. Introduction

This paper provides a review about "miwo", an AI Kuzushiji recognition application for Japanese historical documents, which was presented by Tarin Clanuwat at the 95th International ARC Seminar held on December 8, 2021 [1]. This review is organized as follows: in Section 2, the outline of Clanuwat's presentation will be explained. Some related work will be introduced in Section 3 by the reviewer. A summary will be given in Section 4.

2. Outline of the presentation

After the brief introduction, Clanuwat mentioned how important is to read old cursive writing (i.e., Kuzushiji) in a classical Japanese illustrated book for understanding the historical events. She took a use case by referencing to『Enkouan Zuikan Zue』[2] written by Kouriki Tanenobu (1756~1831), which is a pictorial diary about what happened around Nagoya during the Meiwa 4th year to Anei 7th year (1767-1778). Clanuwat argued that if we read cursive writing in『Enkouan Zuikan Zue』properly, it could be evidence that Red Aurora might have been seen at night in Nagoya on 28 July, 1778. Further, in order to emphasize the importance of having a computerized Kuzushiji recognition system, she mentioned:

> Japan had been using Kuzushiji for over a thousand years. There are over a billion Kuzushiji documents preserved in Japan. People who can read Kuzushiji fluently is just 0.01% of the whole Japanese population. This is why we need help from machines [3].

Therefore, it is crucial to have an artificial intelligence (AI) system to recognize Kuzushiji.

2-1. Challenges in Kuzushiji recognition

The problems of recognizing Kuzushiji using AI were defined as follows:

i) Kuzushiji is written in a cursive style, and in many cases, characters are connected or overlap each other.

ii) Many characters of classical Hiragana or Hentaigana were written in various different ways in Kuzushiji, while these had been standardized to be written as a single way in modern Japanese. Thus, many rules in Kuzushiji don't exist in modern Japanese anymore.

iii) There is a high level of ambiguity and similarity between characters. The same character in Kuzushiji can be transcribed into different characters in modern Japanese. Moreover, a few characters in Kuzushiji look very similar and it is hard to recognize without considering the character in context.

iv) The layout of Kuzushiji characters is quite hard and there are various layouts, which do not follow a single simple rule or text sequence.

v) A large fraction of the characters, e.g., many Kanji characters with very specific meanings may have only appeared once or twice.

2-2. "miwo", an AI Kuzushiji recognition application

At the 95th International ARC Seminar, Clanuwat introduced a mobile application "miwo", which is capable of transcribing Kuzushiji to modern Japanese. "miwo" was released on 31 August, 2021. As of December 8, 2021, "miwo" has over 33,000 downloads and transcribed 200,000 images in three months [4]. It 1) recognizes Kuzushiji automatically, 2) displays bounding boxes around the recognized characters, 3) allows to edit the recognition results, and 4) outputs the results as text when a user uploads an image that contains Japanese cursive text.

2–3. Underlying technologies behind "miwo"

"miwo" uses a Kuzushiji recognition model KuroNet [5),6)], which recognizes an entire page of text as an image. The KuroNet is an end-to-end model that uses the residual U-Net architecture [7)] with additional regulations. It was trained using character locations instead of character sequences. The KuroNet captures both long-range and local dependencies without pre-processing. The KuroNet performs character segmentation and recognition at the same time by utilizing object detection algorithms. By using the KuroNet model, "miwo" is able to process an entire scanned image of a page in approximately 1.2 seconds with the average F1 score of 0.902.

2–4. Utilized dataset

The Kuzushiji recognition model (the KuroNet) of "miwo" was trained using pre-modern Japanese books' Kuzushiji dataset (i.e., The Kuzushiji dataset), which was updated and released in November 2019 [8)]. The Kuzushiji dataset was created by the National Institute of Japanese Literature (NIJL) and is curated by the ROIS-DS Center for Open Data in the Humanities (CODH). The Kuzushiji dataset contains 1,086,326 characters extracted from 6,151 pages of the 44 classical Japanese books published in the dataset of pre-modern Japanese text. There are 4,328 distinct characters (character types) in the Kuzushiji dataset [9)]. The dataset of pre-modern Japanese text contains not only Kuzushiji data but also scanned images, bibliographic data as well as some full-text and tagged proper nouns of unlabeled 3,126 classical Japanese books owned by the NIJL and other related organizations [10)].

2–5. Lessons learned and feedbacks from "miwo"

Clanuwat shared her thoughts in developing a real-world application as follows: Although there are several decent models, the highest accuracy model may not be the best for production, since a trade-off between recognition time, computing power and accuracy is crucial. High quality data prepared by domain experts is very important. "miwo" needs a lot of improvements, because the KuroNet achieved 95% accuracy on the Kuzushiji dataset only, which is a small portion of the Kuzushiji documents in the real-world. Moreover, the KuroNet needs more trials in vast data of different periods such as handwritten documents and letters, as the accuracy drops notably in recognizing documents other than printed books of the Edo period (1603-1868). Regarding user expectations, either having too high expectations such as AI being 100% correct, or having too low expectations such as AI being unable to recognize anything is not good. If we rely on AI entirely, no matter how advanced the AI models are, in the end if we humans can't comprehend the AI outputs, it is still not trustworthy. Clanuwat also mentioned that they received many feedbacks from users, including a suggestion to use "miwo" in a classical Japanese literature class.

3. Related work

In recent years, many efforts have been devoted to Kuzushiji research using computers. Many approaches are proposed and some of them are already in use. Prior to deployment of AI in Kuzushiji research, Historiographical Institute of the University of Tokyo released Kuzushiji Digital Database [11)], and Nara National Research Institute for Cultural Properties (a.k.a. NABUNKEN) also released Mokkanko - Wooden Tablet Database [12)]. Both databases are now merged to the Multi-database Search System for Historical Chinese Characters [13)], and searchable along with NIJL's Kuzushiji dataset. Moreover, NABUNKEN and Historiographical Institute of the University of Tokyo jointly developed "MOJIZO" [14)], an image similarity search service of Kuzushiji or wooden tablets (mokkan). Kengo Terazawa and Toshio Kawashima developed an online spotting system named "Document and image search system" [15)] to retrieve the similar instances of text region from scanned images for a given query image, and applied it to pre-modern documents [16)] that might have written in Kuzushiji. Moreover, "KuLA", a Kuzushiji learning application has been developed for assisting users to read historical materials of pre-Edo period, written in Kuzushiji [17)]. In order to transcribe Kuzushiji, KuLA is linked to the "Minna de Honkoku" [18)], which is capable of annotating the historical materials through crowdsourcing.

In 2019, Toppan Printing Co., Ltd. publicized a Kuzushiji recognition service and Application Programming Interface (API) named

Fuminoha [19], that has been developed since 2015 as collaborative research with the NIJL by utilizing funding supports of the "Project to Build an International Collaborative Research Network for Pre-modern Japanese Texts (NIJL-NW project)" [20),21)]. The Kuzushiji recognition models KuroNet and KogumaNet [22)] along with API were also launched online by the CODH in 2019. In overall, many machine learning models were developed for recognizing Kuzushiji [23)]. These Kuzushiji recognition models are widely used in many applications. For instance, "miwo" utilizes KuroNet, and "Minna de Honkoku" uses Fuminoha and KogumaNet selectively to recognize Kuzushiji automatically by AI [24)]. "Kuzushiji decoding support and transcription guidance system" [25)] of the Art Research Center (ARC), Ritsumeikan University also utilizes Fuminoha for obtaining AI suggestions to transcribe obfuscated characters [26)]. ARC's "Kuzushiji decoding support and transcription guidance system" also utilizes the results of the similar characters of "Document and image search system" along with Fuminoha. Using ARC's system, even beginners can transcribe Japanese historical documents with the supports of experts as well as getting hints from AI consecutively. It allows to accumulate 1) unreadable character images for asking experts' supports, and 2) correctly annotated characters for utilizing further improvements of AI models. It is an effective learning system that can be used in university classes or learning courses. Online self-study is possible, though it would be effective for group learning under the guidance of experts and teachers. ARC's system is already in an educational practice at the Ritsumeikan University [27)].

4. Summary

This review discussed a mobile application "miwo" as an example of recognizing Kuzushiji in Japanese historical documents. Nowadays, there are several real-word web or mobile applications to recognize Kuzushiji. Obviously, AI-driven approaches are becoming dominant. However, at this stage, human experts' assistance is still essential as the AI results and predictions are not perfect yet. Thus, having an interactive learning system, that is capable of receiving experts' assistance and guidance is viable to avoid the 100% reliance on AI.

[Notes]

1) Tarin Clanuwat. "miwo" AI Kuzushiji Recognition Application for Japanese Historical Document, The 95th International ARC Seminar, December 8, 2021. https://www.arc.ritsumei.ac.jp/e/news/pc/009275.html (accessed: 2022-01-06)
2) National Diet Library Digital Collections. 『猿猴庵随観図絵』高力種信, https://dl.ndl.go.jp/info:ndljp/pid/2537160 (accessed: 2022-01-06)
3) Same as 1).
4) Same as 1).
5) Alex Lamb, Tarin Clanuwat, and Asanobu Kitamoto. KuroNet: Regularized Residual U-Nets for End-to-End Kuzushiji Character Recognition. *Springer Nature, Computer Science Special Issue on Document Analysis and Recognition*, 2020, vol.1:177, pp. 1-15, https://doi.org/10.1007/s42979-020-00186-z
6) Center for Open Data in the Humanities. KuroNet くずし字認識サービス(AI OCR), 2019, http://codh.rois.ac.jp/kuronet/ (accessed: 2022-01-06)
7) Tran Minh Quan, David G. C. Hildebrand, and Won-Ki Jeong. FusionNet: A deep fully residual convolutional neural network for image segmentation in connectomics, *CoRR*, 2016, http://arxiv.org/abs/1612.05360
8) Center for Open Data in the Humanities. 日本古典籍くずし字データセット, 2019, http://codh.rois.ac.jp/char-shape/ (accessed: 2022-01-06)
9) Center for Open Data in the Humanities. 日本古典籍くずし字データセット 書名一覧, 2019, http://codh.rois.ac.jp/char-shape/book/ (accessed: 2022-01-06)
10) Center for Open Data in the Humanities. 日本古典籍データセット, 2019, http://codh.rois.ac.jp/pmjt/ (accessed: 2022-01-06)
11) Historiographical Institute of the University of Tokyo.「電子くずし字字典データベース」, https://wwwap.hi.u-tokyo.ac.jp/ships/shipscontroller (accessed: 2022-01-06)
12) Nara National Research Institute for Cultural Properties.「木簡庫」(Wooden Tablet Database), https://mokkanko.nabunken.go.jp/en/ (accessed: 2022-01-06)
13) Nara National Research Institute for Cultural Properties. Multi-database Search System for Historical Chinese Characters, https://mojiportal.nabunken.go.jp/en/ (accessed: 2022-01-06)

14) Nara National Research Institute for Cultural Properties and Historiographical Institute of the University of Tokyo. “MOJIZO”: Image matching search for mokkan or cursive characters, https://mojizo.nabunken.go.jp/ (accessed: 2022-01-06)
15) Kengo Terazawa and Toshio Kawashima. Word Spotting Online, *The Computers and the Humanities Symposium JinMonCon 2011*, Information Processing Society of Japan, pp. 329-334
16) Future University Hakodate.「文書画像検索システム」, http://records.c.fun.ac.jp/ (accessed: 2022-01-06)
17) Yuta Hashimoto, Yoichi Iikura, Yukio Hisada, SungKook Kang, Tomoyo Arisawa, and Daniel Kobayashi-Better. The Kuzushiji Project: Developing a Mobile Learning Application for Reading Early Modern Japanese *Texts, Digital Humanities Quarterly*, Volume 11 Number 1, 2017.
18) Yuta Hashimoto, くずし字の学習支援と市民参加翻刻. 2nd CODH Seminar - Old Japanese Character Challenge - Future of Machine Recognition and Human Transcription, 10 February, 2017, https://doi.org/10.20676/00000007
19) Toppan Printing Co., Ltd. 古文書解読とくずし字資料の利活用サービス「ふみのは」, 2019, https://www.toppan.co.jp/biz/fuminoha/ (accessed: 2022-01-06)
20) Yamamoto Sumiko and Osawa Tomejiro. Labor saving for reprinting Japanese rare classical books: The development of the new method for OCR technology including kana and kanji characters in cursive style, *Journal of Information Processing and Management.* 2016, vol. 58, no. 11, pp. 819-827, doi: http://doi.org/10.1241/johokanri.58.819
21) National Institute of Japanese Literature. Collaborative Research to Convert into Text All Images Gathered by the “Project to Build an International Collaborative Research Network for Pre-modern Japanese Texts”, くずし字 OCR 技術に関する実証試験結果, https://www.nijl.ac.jp/pages/cijproject/images/kuzushi-ji_ocr.pdf (accessed: 2022-01-06)
22) Center for Open Data in the Humanities. KogumaNet くずし字認識サービス(一文字), 2019, http://codh.rois.ac.jp/char-shape/app/single-mobilenet/ (accessed: 2022-01-06)
23) Center for Open Data in the Humanities. Kaggle Competition: Kuzushiji Recognition, 2019, http://codh.rois.ac.jp/competition/kaggle/ (accessed: 2022-01-06)
24) Rekihaku, National Museum of Japanese History / Earthquake Research institute, The University of Tokyo / Research Group for Historical Earthquakes, Kyoto University. Minna de Honkoku. https://honkoku.org/index_en.html (accessed: 2022-01-06)
25) Shikoku News.「崩し字」、AIで解読します／立命館大が支援システム, 2019, http://www.shikoku-np.co.jp/national/science_environmental/20190513000675 (accessed: 2022-01-06)
26) SankeiBiz.「崩し字」をAIが画像解読　立命館大教授ら全国初の高精度システム, 2019, https://www.sankeibiz.jp/business/news/190515/bsj1905150500001-n1.htm (accessed: 2022-01-06)
27) 山路 正憲, 岡田 崇, 岡 敏生, 秋元 良仁, 大澤 留次郎, 赤間 亮. 古典籍デジタルアーカイブの活用を促進するディープラーニング型くずし字翻刻支援システムと指導システム, Annual Convention, Japan Art Documentation Society, 2019, http://www.jads.org/news/2019/20190608-09.html (accessed: 2022-01-06)