Press Release

大学共同利用機関法人 情報・システム研究機構 Research Organization of Information and Systems

大学共同利用機関法人 情報・システム研究機構 データサイエンス共同利用基盤施設 Joint Support-Center for Data Science Research (ROIS-DS)

大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

大学共同利用機関法人 人間文化研究機構 国文学研究資料館 National Institute of Japanese Literature

July 9, 2019

Please note that the embargo will lift as follows:
TV, radio, and internet: Wednesday, July 10, 2019, 01:00GMT
Newspaper: Wednesday, July 10, 2019, morning edition, GMT

Worldwide Competition to Develop AI for Historical Japanese Character (Kuzushiji) Recognition

- The Competition, hosted on Kaggle, will start in July-

Japan has classical books, documents, and other historical records that date back over a thousand years. However, while these documents may number in the hundreds of millions, most modern Japanese speakers cannot read kuzushiji, or cursive Japanese script. Despite ongoing efforts to digitize and release historical documents, reading the vast quantity of kuzushiji remains a serious issue in research on Japanese culture and in explaining historical disasters and other natural phenomena from the past.

To explore ways of using artificial intelligence (AI) to help resolve societal issues, a global competition entitled “Kuzushiji Character Recognition: Opening the Door to A Thousand Years of Japanese Literate Culture” is being held on the world’s largest machine learning platform “Kaggle” from July to October 2019. The competition will not only advance the development of innovative kuzushiji recognition methods but also heighten global interest in Japanese culture through the kuzushiji dataset.

The competition is hosted by the Research Organization of Information and Systems' Center for Open Data in the Humanities (CODH, Director: Asanobu KITAMOTO) at the Joint Support-Center for Data Science Research, along with the organization's National Institute of Informatics (NII, Director General: Masaru KITSUREGAWA) and National Institutes of Japanese Literature (NIJL, Director General: Robert CAMPBELL) at the National Institute for the Humanities.

Japan has classical books, documents, and other historical records dating back over a thousand years. However, while these documents may number in the hundreds of millions, most modern Japanese speakers cannot read kuzushiji, or cursive Japanese script. Despite ongoing efforts to digitize and release historical documents, reading the vast quantity of kuzushiji remains a serious issue in research on Japanese culture and in explaining historical disasters and other natural phenomena from the past.

It is estimated that only a few thousand individuals across Japan are able to properly read kuzushiji[1], and there are limitations in how many of the existing documents these individuals can transcribe[2]. Therefore, research is being conducted in two directions to assist in resolving this issue. The first is in the development of a participatory transcription system [3]. Citizens, together with experts, participate in the system not only for transcribing kuzushiji, but also for learning skills to read it better, thereby increasing the number of people who can read kuzushiji. The second is the use of computers for automatic recognition. A number of groups have tried automatic transcription using optical character recognition (OCR) for machines to read characters. However, kuzushiji OCR for practical use remains challenging for machine learning systems, due to vast amount of character types, difficulty in cleanly separating characters, diversity in layout, and variations in writing style between books.

On the other hand, rapid advancement of AI using deep learning (machine learning) in computer vision technology opens up new possibilities to develop innovative methods for kuzushiji OCR. To collect ideas in an open manner to improve the performance of kuzushiji OCR, CODH, NII, and the NIJL are hosting a competition entitled "Kuzushiji Recognition Challenge: Opening the Door to a Thousand Years of Japanese Literate Culture [4]" from July to October, 2019, using the global machine learning competition platform Kaggle[5] (Figure 1).

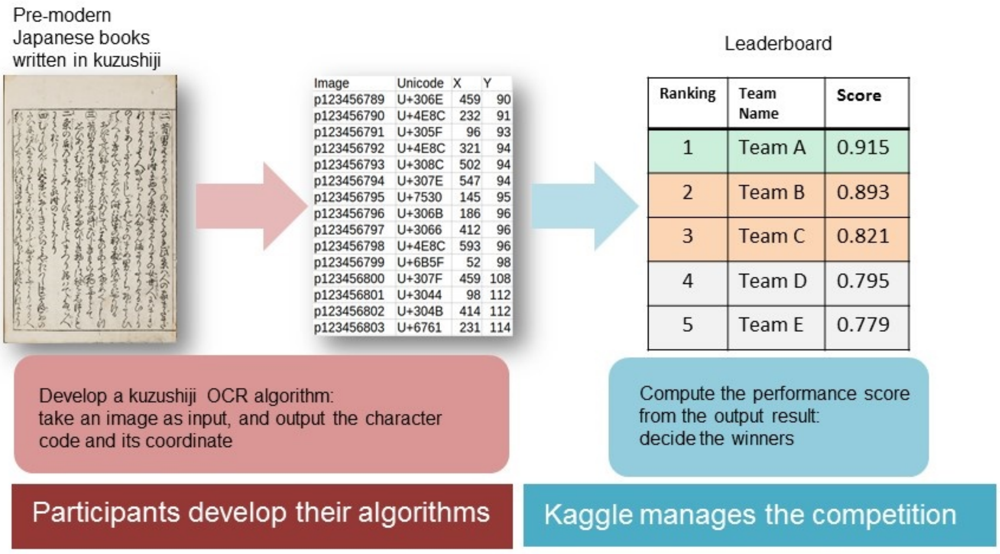
Pre-modern
Japanese books
written in kuzushiji
Leaderboard
Develop a kuzushiji OCR algorithm:
take an image as input, and output the character code and its coordinate
Compute the performance score from the output result:
decide the winners
Participants develop their algorithms
Kaggle manages the competition

Image	Unicode	X	Y
p123456789	U+306E	459	90
p123456790	U+4E8C	232	91
p123456791	U+305F	96	93
p123456792	U+4E8C	321	94
p123456793	U+308C	502	94
p123456794	U+307E	547	94
p123456795	U+7530	145	95
p123456796	U+306B	186	96
p123456797	U+3066	412	96
p123456798	U+4E8C	593	96
p123456799	U+6B5F	52	98
p123456800	U+307F	459	108
p123456801	U+3044	98	112
p123456802	U+304B	414	112
p123456803	U+6761	231	114

Ranking	Team Name	Score
1	Team A	0.915
2	Team B	0.893
3	Team C	0.821
4	Team D	0.795
5	Team E	0.779

Figure 1 Overview of the competition

The competition provides a modified version of Kuzushiji Dataset, an open dataset prepared by NIJL and curated by CODH. During the three-month competition, participants will develop kuzushiji OCR algorithms to recognize kuzushiji from raw images. The algorithms selected as winners will be shared as open source after the competition.

Similar competitions have been held in Japan in the past, but on a smaller scale hosted by a domestic academic society [6], but to extend the competition to a global scale, this competition will be hosted on internationally recognized Kaggle platform, wherein more than three million AI researchers and engineers worldwide participate. This is the third

Japanese organization, after Recruit and Mercari, to host a Kaggle competition, and the first Japanese organization to host a research competition. Moreover, in the history of Kaggle, this is the first competition for a humanities-related data. A person in Kaggle commented, "Even with our 337 competitions to date, this competition covers an area that has not yet been explored on Kaggle. Based on the rise of computer vision techniques, the competition could be very exciting for our community to tackle."

Developing an innovative method to recognize kuzushiji through this competition could lead to research and development to support new technologies to understand Japanese culture in the past, such as AI-based transcription, full text search on AI-based character recognition. If AI can help some of the tasks that have been done by experts, they can focus more on high-level interpretation of historical documents. Hence this kind of new technologies will be essential in data-driven research [7] for Japanese culture; namely, digitizing historical documents for open access, transcribing historical documents both by machines and by citizen, analyzing data of the past world through collaboration of researchers in the natural sciences and humanities, and returning research results back into the wider society.

Detailed information on this competition can be found on Kaggle website on the launch day in the mid of July, 2019. Participants can submit algorithms over a period of three months until October deadline, after which the organizers will work with Kaggle in selecting the top five entries. These winners will be awarded at the "Japanese Culture and AI" symposium to be held in Tokyo on November 11.

Reference

Asanobu KITAMOTO, Tarin CLANUWAT, Tomo MIYAZAKI, Kazuaki YAMAMOTO, "Analysis of Character Data: Potential and Impact of Kuzushiji Recognition by Machine Learning", Journal of IEICE (Institute of Electronics, Information, and Communication Engineers), Vol. 102, No. 6, pp. 563-568, June 2019 (in Japanese)
http://doi.org/10.20676/00000349

Media contact:

PR office, Headquarter, Research Organization of Information and Systems
TEL: +81-(0)3-6402-6214 E-mail：koho@rois.ac.jp

Publicity Team, National Institute for Informatics, Research Organization of Information and Systems
TEL: +81(0)3-4212-2164; E-mail: media@nii.ac.jp

Center for Collaborative Research on Pre-modern Texts, National Institute of Japanese Literature
TEL:+81-50-5533-2988; E-mail : cijinfo@nijl.ac.jp

[1] From "Wahon no Susume" Mitsutoshi Nakano, Iwanami Shinsho, (2011). Also refer to the press release from National Institute of Informatics entitled "Edo Jidai no Moji no Jikei Dataset wo Kokubunken tono Kyoudou de Kouchiku Kikai to Ningen no Gakushuu no tame no Open Data toshite Koukai" [Dataset of Pre-modern Character Shapes Published as Open Data for Human and Machine Learning in Cooperation with the National Institute of Japanese Literature] (https://www.nii.ac.jp/userimg/press_20161117.pdf).

[2] Transcription of kuzushiji is a process whereby humans read kuzushiji and input the corresponding contemporary Japanese character.

[3] "Minna de Honkoku" (https://honkoku.org/) is a participatory (crowdsourced) transcription system mainly developed by Yuta Hashimoto, Assistant Professor of the National Museum of Japanese History, in cooperation with the Kyoto University Paleoseismology Study Group and Earthquake Research Institute at the University of Tokyo. This group also collaborate with CODH in a joint research project.

[4] For further details, see following websites.
The competition website (https://www.kaggle.com/c/kuzushiji-recognition)
#The competition page will be released on the launch day
CODH website (http://codh.rois.ac.jp/competition/kaggle/)

[5] Headquartered in the US (and owned by Google), Kaggle (https://www.kaggle.com/) is the world's largest platform for machine learning competitions. Competitions at Kaggle are held in an open research and development process wherein (1) corporations or researchers submit challenges they wish to solve and related data; (2) AI researchers and engineers worldwide submit algorithms to solve the challenges; (3) the submitted algorithms are ranked by performance, and winners are determined; and (4) winners share the outcome of the competition with the host and obtain the prize.

[6] The 23rd PRMU Algorithm Contest, 2019 Kuzushiji Recognition Challenge (https://sites.google.com/view/alcon2019) runs from May 31 to August 31, 2019, and is hosted by the IEICE Society of Pattern Recognition and Media Understanding (supported by the CODH). While it uses a similar dataset with the Kaggle competition's, the difficulty of the challenge is different.

[7] "Data-driven research" is a research method for acquiring new knowledge from

evidences obtained from the collection and analysis of (big) data, often by large-scale data processing such as machine learning (AI).