Human-assisted OCR of Japanese Books with Different kinds of Microtasks

Kosetsu Ikeda[1], Ryota Hayashi[2], Kiyonori Nagasaki[3], Atsuyuki Morishima[2]
[1]Chiba University, Japan
[2]University of Tsukuba, Japan
[3]International Institute for Digital Humanities, Japan

Abstract
Human-assisted OCR is a common approach for transcribing books and has been used for many digital library projects. This paper reports our project for transcribing the book collections of National Diet Library in this approach. Our project is unique in two ways. First, we try to extend the human-assisted OCR approach by distributing microtasks in many ways other than just showing tasks in the specific Web page on PC screens. Second, we deal with Japanese books which have thousands of characters, some of which look similar to each other. This paper shows that we can expect high-quality results even if we transcribe Japanese texts with microtasks and the number of preformed microtasks to be stable if we distribute microtasks to equipment with witch worker perform microtasks in their daily lives.
Keywords: Digital transcription; Crowdsourcing; Microtask
DOI: Citation info is to be added.

Acknowledgements: The authors are grateful to the contributors to Crowd4U, whose names are partially listed at http://crowd4u.org. This research was partially supported by the Grant-in-Aid for Scientific Research (#25240012) from MEXT, Japan and by CREST, JST.
Contact: lumely@chiba-u.jp

1 Introduction

Recently, digital transcription with crowdsourcing attracts much attention, and there are a large number of crowdsourced digital transcription projects in the library and digital archive domains (Australian Newspaper Digitisation Program, n.d.; National Archives Transcription Pilot Project, n.d.; Terras, 2012; Causer, Tonra, & Wallace, 2012; Ishihara, Itoko, Sato, Tzadok, & Takagi, 2012). A common approach in digital transcription projects is the human-assisted optical character recognition (human-assisted OCR) approach, where workers correct the results of OCR software or directly transcribe characters that the OCR fails to transcribe (reCAPTCHA, n.d.).

This paper reports our project for transcribing the book collections of National Diet Library (NDL) (National Diet Library Digital Collections, n.d.) in this approach. Our project is unique in two ways. First, we try to extend the human-assisted OCR approach by distributing microtasks in many ways other than just showing tasks on PC screens. Most of existing projects ask crowd workers to perform such tasks on the specific Web page. In contrast, we try to extend the approach by distributing tasks in many other ways. For example, we distribute tasks to the task-on-the-floor (shortly TOF) system, where people walking perform tasks while walking on the floor, and to the smartphone lockscreen application, where people who want to unlock their smartphones perform tasks.

Second, we deal with Japanese books which have thousands of characters, some of which look similar to each other. This raises an interesting question because people performing microtasks may not pay a great attention to the task compared to the case where they perform tasks on the specific Web page.

The main contributions are as follows. First, we explain our ongoing project whose approach is novel. To the best of our knowledge, our project is the first to try to distribute tasks in many ways other than PC screens in the digital library domains. Second, we show our preliminary results that suggest that the microtask approach is effective to some extent although the Japanese language contains thousands of characters some of which are similar to each other.

In our project, we only extract text data from digital images and do not consider the layout of book pages.

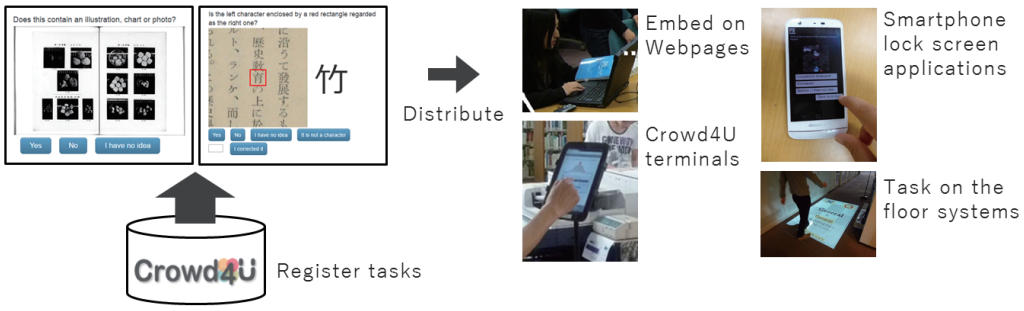

Figure 1: Overview of Crowd4U

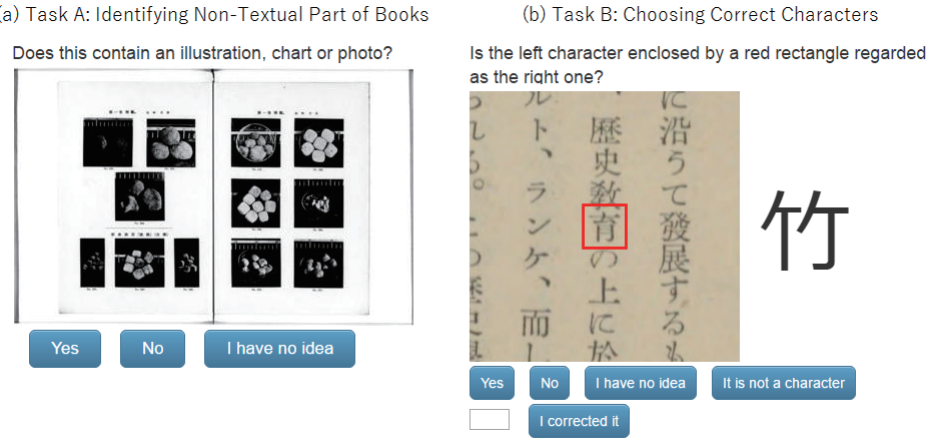

Figure 2: Microtasks used in the project

2 The Hondigi Project and Crowd4U

Hondigi is an academia-based project to try to transcribe books in the National Diet Library digital collections, whose copyrights are expired. As a subproject of Hondigi, we have been exploring the effective usages of microtasks for transcribing books.

In this project, we use Crowd4U (Morishima, 2013), a nonprofit open microvolunteering and crowdsourcing platform for academic and public purposes. Figure 1 shows the overview of Crowd4U. A prominent feature of Crowd4U is the ability to distribute the registered tasks in many ways. Currently, tasks can be distributed to (1) any Web pages with embedded javascripts, (2) Crowd4U terminals that are located six universities in the world, (3) lockscreen applications that can be downloaded from Google Play, and (4) the TOF system that are located in four universities in Japan.

3 Microtasks Design

This section explains two tasks currently used in the project.

Task A: Identifying Non-Textual Part of Books.

The purpose of this task is to identify non-textual part of book contents. The identified part will be published as a list of figures.

Figure 2 (a) shows an example of this task. This task asks workers whether the shown image contains an illustration, chart, or photo. Each worker chooses one of “Yes ” “No” or “I have no idea.” We apply majority voting for obtaining the final decision.

Task B: Choosing Correct Characters

This task asks workers whether the result of OCR is correct or not. Figure 2 (b) shows an example. The OCR outputs a ranked list of character candidates with a confidence value for each character. The task

first shows the candidate with the highest confidence. The target character is surrounded by a red box, and each worker chooses one of "Yes," "No," "I don't know" or "It it not a character," or directly enter the shown character to a text field. We used two versions of Task B. The first version do not allow workers to directly enter the characters. As we will show in the next section, this caused a problem and we added the text field to the task in the second version.

We ask more than one worker to perform the task and changes the shown candidate in the following way.

- If three workers answer "yes," the final result is the character.
- If two workers answer "no," we change the shown character to the next candidate.
- If three workers answer "It is not a character," the final result is null.
- If a worker enters a character, we increase the number of "no" and we use the entered character as the next candidate.

4 Preliminary Results

We registered the tasks to Crowd4U and volunteer workers performed the tasks with PC screens, lockscreen applications, and the TOF systems.

(1) Number of Performed Tasks

We started to distributed the tasks on Feb. 2, 2015. We had 153,201 results as of Apr. 10, 2016. Figure 3 shows how the monthly number of performed tasks in the period. The two solid lines represent the total numbers for Task A and B, respectively. The dotted lines are breakdowns of the number for Task A.

Some observations are as follows.

First, the number of tasks dramatically increases if we distribute tasks to the lockscreen applications. The numbers on the line with diamonds are 1.36 times larger than those with triangles on average. This is because we distributed only Task A to the lockscreen applications.

Similarly, there is a huge impact of the TOF system on the number. The numbers on the solid line with squares are twice larger than those on the dotted line with diamonds on average. The only difference between them is that the former includes the number of tasks performed with the TOF systems.

We observe that the number of tasks are affected by several factors. First, workers tend to perform many tasks when the task are announced for the first time, but eventually they are getting tired and only a few active users continue to perform tasks. Therefore, it is effective to use distribution frameworks that do not depend on the motivation of workers, such as the TOF systems.

We placed the systems on university campuses and most of workers are university students. Therefore, the seasonal factor exists. For example, the number of tasks performed on the TOF systems becomes small on Summer and Spring vacations of Japanese universities.

(2) Quality of the Task B Results

For this task, we used OCR results generated by the ABBYY FineReader (ABBYY FineReader, n.d.) as a trial. We started to distributed the tasks on Nov. 12, 2015. We obtained 20,644 results and 3,979 characters as of Apr. 10, 2016. We selected the 285 out of the 3,979 characters in the first two paragraphs of the book "Iwanami Kōza Nihon Rekishi" (The Historical Science Society of Japan (Ed.), 1935), and we manually compared the characters with the direct OCR results (i.e., the top-ranked characters of the OCR outputs).

Table 1 shows the comparison result. The F_1 score of the extracted characters is 0.13 points higher than that of the direct OCR results. The direct OCR results contained two misidentified characters (i.e., the parts of books that are not characters but identified as characters) but the crowd successfully removed them. On the other hand, there are five characters that the workers could not output the correct characters. This is because the workers answered "No" to all candidates in the list of characters the OCR output. The problem happened only with the first version of Task B and we found that no such problems happened with the revised version of Task B.

Table 2 explains the classification of incorrectly recognized characters. The first and second categories represent minor errors. The first category "Form is different" means that the meaning of character

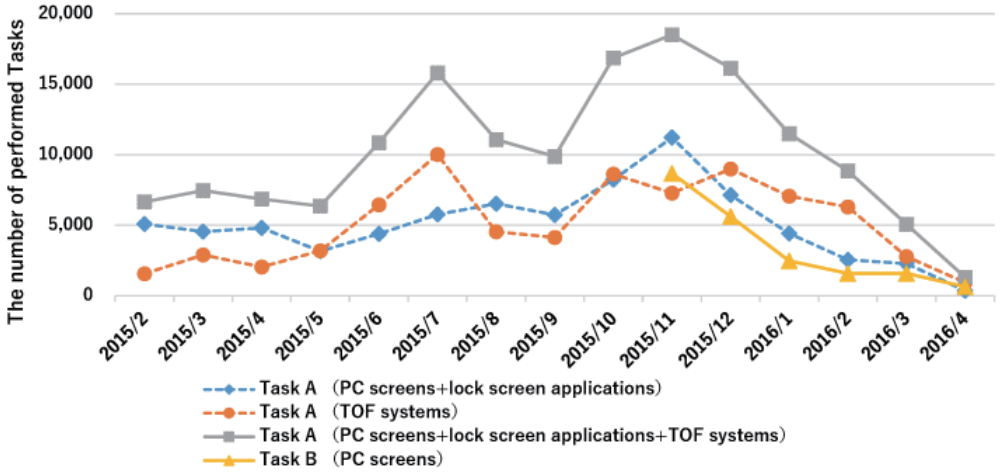

Figure 3: Monthly number of performed tasks

is correct but the form is different. Some Japanese characters have different forms with the same meaning. The second category "Size is different" means that the shape is correct, but the size is different. The third to fifth categories are serious errors. As we explained, the OCR misidentified two characters and the workers could not give answers to five characters.

Workers succeeded in removing completely different answers although they are performing tasks in ways where they cannot necessarily concentrate on the tasks. And workers failed to correct different size characters because the difference of sizes of characters in our task is hardly distinguishable.

Table 1: Digital transcription results (for 285 characters)

	Number of characters	Precision	Recall	F_1	TP	FP	FN
OCR results	287	0.75	0.76	0.76	213	69	0
Task results	280	0.90	0.88	0.89	253	27	5

Table 2: Classification of errors

	OCR results	Task results
Form is different	5	5
Size is different	14	22
Completely different	48	0
Not a character	2	0
Couldn't give an answer	0	5

5 Discussions

A weakpoint of the volunteer-based crowdsourcing is that the number of performed tasks is not stable. The number tend to be the largest when the introduction of new tasks are announced and gradually becomes small. We found that distributing tasks to equipment placed in their everyday lives, such as lockscreen applications and the TOF system, is effective to keep the number large. Since the forms of microtasks that can be distributed in such ways are limited to simple ones, effectively mixing tasks with different granularities is desired.

OCR software is not good at dealing with text containing figures and tables. In particular, some Japanese books contains texts that are overlapped by official seals. We found that crowdsourcing is effective to deal with such cases.

Some Japanese characters can appear in text in different sizes and they have to be distinguished. We found that the naive task design cannot deal with such cases. Using a square to surround the candidate character in the task may help workers grasp its size.

So far, we used fixed-sized microtasks showing only one character to each worker. We noticed that the size is often too small since OCR results are generally good. We will consider dynamically changing the granularity of tasks depending on the quality of OCR results.

6 Summary

This paper reported our project for transcribing the book collections of NDL. In the project, we try to extend the human-assisted OCR approach by distributing microtasks in many ways other than just showing tasks on PC screens and apply the approach to Japanese books which have thousands of characters. This paper explained the overview and the initial findings in our first attempt. Our experience so far suggests that (1) we can expect high-quality results even if we transcribe Japanese texts with microtasks. (2) If we distribute microtasks to equipment with which worker perform microtasks in their daily lives, the number of performed microtasks becomes stable, and (3) The task design is important for the efficiency of the transcription. Future studies include introducing tasks to correct the direction of the character. And to dynamically change the task granularity for improving efficiency.

References

ABBYY FineReader. (n.d.). Retrieved from http://finereader.add-soft.jp/

Australian Newspaper Digitisation Program. (n.d.). Retrieved from http://www.nla.gov.au/content/newspaper-digitisation-program

Causer, T., Tonra, J., & Wallace, V. (2012). Transcription maximized; expense minimized? crowdsourcing and editing the collected works of jeremy bentham. *Literary and Linguistic Computing*, *27*(2), 119-137.

Ishihara, T., Itoko, T., Sato, D., Tzadok, A., & Takagi, H. (2012). Transforming japanese archives into accessible digital books. In *Proc. jcdl'12* (pp. 91–100). New York, NY, USA: ACM.

Morishima, A. (2013). Cylog/crowd4u: A case study of a computing platform for cybernetic dataspaces. In P. Michelucci (Ed.), *Handbook of human computation* (pp. 561–572). New York, NY: Springer New York.

National Archives Transcription Pilot Project. (n.d.). Retrieved from http://www.archives.gov/citizen-archivist/

National diet library digital collections. (n.d.). Retrieved from http://dl.ndl.go.jp/

reCAPTCHA. (n.d.). Retrieved from https://www.google.com/recaptcha/

Terras, M. (2012). Present, not voting: Digital humanities in the panopticon. In *Understanding digital humanities* (pp. 172–190). London: Palgrave Macmillan UK.

The Historical Science Society of Japan (Ed.). (1935). Iwanami kōza nihon rekishi. In (Vol. 1: Sōsetsu and Kodai (Introduction and Ancient)). Iwanami Shoten, Publishers. Retrieved from http://kindai.ndl.go.jp/info:ndljp/pid/1263939