Vol.2016-CH-110 No.10
2016/5/14

翻デジにおけるマイクロタスク活用の試み

池田 光雪[1,a)]　林 亮太[1,b)]　永崎 研宣[2,c)]　森嶋 厚行[,d)]

概要：近年，刊行された資料のデジタル化，及び公開が盛んに行われており．例えば国立国会図書館デジタルコレクションには我が国において明治期以降に刊行された図書・雑誌のデジタル化資料が多数収録されている．しかし，国立国会図書館デジタルコレクションの資料も含め多くの資料は未だ画像形式でしか提供されておらず，解析や検索による利用が困難である．一方，デジタル化された資料を用い翻刻を行うシステムはこれまでに多数提案されてきたが，いずれも人手による入力や修正に非常にコストがかかるという問題があった．「翻デジ」は，そのような背景の下，クラウドソーシングによって国立国会図書館デジタルコレクションのデジタル翻刻を実際に行うプロジェクトである．本稿では，様々な形態でのクラウドソーシングによるマイクロタスク型クラウドソーシングを用いたデジタル翻刻を行う手法を提案し，これまでの「翻デジ」の経験に基づいた議論を行う．我々の知る限り，本稿は，マイクロタスク型クラウドソーシングを利用したデジタル翻刻の実プロジェクトとその経験に基づく知見を示した初めての論文である．

キーワード：デジタル翻刻，クラウドソーシング，マイクロタスク

1. 翻デジとクラウドソーシング

近年，テクストのデジタル化，いわゆるデジタル翻刻が盛んに行われている．例えば，国立国会図書館デジタルコレクション [14] では著作権が切れた入手が困難な資料等を公開しているが，現在，その多くは画像のみの提供となっている．そこで，画像でしか提供されていない資料をデジタル翻刻することで，全文検索が可能になる，様々な解析が容易になるといった様々な恩恵を得ることが出来ると考えられる．

デジタル翻刻には様々な手法があるが，まず OCR で機械的に読み取りをして，その結果を人手で確認，修正するというプロセスが採用されていることが多い．しかし，日本語は欧米に比べ非常に多彩な文字を持つが故に，日本語資料のデジタル翻刻においては OCR 結果の修正に莫大なコストが必要である．

一方，情報工学分野ではクラウドソーシングに注目が集まっている．クラウドソーシングとは群衆（crowd）と外部委託（outsourcing）を組み合わせた用語であり，不特定多数の人間に対し何らかの作業を依託することを意味する．このクラウドソーシングを用いることで，負荷を分散させたデジタル翻刻が可能になると考えられる．

翻デジとは，クラウドソーシングを用いた日本語資料の翻刻プラットフォームである．永崎は日本デジタル・ヒューマニティーズ学会の SIG-TranscribeJP における活動として 2014 年に国立国会図書館デジタルコレクションの資料をターゲットとした翻デジ 2014 を立ち上げ，デジタル翻刻を進める傍ら資料の典拠性や文字の同定などを検討してきた [16]．さらに 2015 年には，国デコ翻デジ@ JADH×Crowd4U としてマイクロタスク型クラウドソーシングプラットフォームである Crowd4U [3] を用いて図書館領域の問題の解決を試みる L-Crowd [4] と連携し，1 回 1 回はごく短時間で作業が可能なマイクロタスクによりデジタル翻刻を行ってきた．

本稿では，国デコ翻デジ@ JADH×Crowd4U で行われたタスク結果の報告や，マイクロタスク型クラウドソーシングを用いたデジタル翻刻の今後の展望を論ずる．

なお，本稿ではデジタル翻刻を紙媒体上のテクストをテクストデータにすることを指す用語として扱い，テクストのレイアウトについては考慮しないこととする．

我々の知る限り，本稿は，実際にマイクロタスク型クラ

[1] 筑波大学図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba
[2] 一般財団法人人文情報学研究所
International Institute for Digital Humanities
[3] 筑波大学 図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba
[a)] mitsu@klis.tsukuba.ac.jp
[b)] ryota.hayashi.2014b@mlab.info
[c)] nagasaki@dhii.jp
[d)] mori@slis.tsukuba.ac.jp

ウドソーシングを用いてデジタル翻刻を行ったプロジェクトの知見に関する初めての報告である．利用者を限定しているが複数人によるデジタル翻刻システムとして，共同翻刻用ソフトウェアである SMART-GS [10] や SAT 大蔵経データベースにおける Web コラボレーションシステム [15] がある．

また，クラウドソーシングを活用した図書館等におけるデジタル翻刻の事例は特に海外を中心として数多く存在し．例えば Australian Newspaper Digitisation Program [1] ではオーストラリアの新聞を，National Archives Transcription Pilot Project [6] は米国国立公文書館が所蔵する資料をそれぞれクラウドソーシングを用いて電子化している．また，University College London が主導するクラウドソーシングプロジェクト，Transcribe Bentham では，日本を含む世界中からボランティアが参加し，人文社会科学研究の基礎資料として通用するデータを構築するに至っている [2], [5]．日本においては，著作権の消滅した作品を主な対象とした青空文庫 [11] の取り組みが有名である．しかし，これらはあくまでも対象資料を限定した専用システムを利用した「取り組み」であり，クラウドソーシングの可能性に着目して様々な試みを展開する，日本語を扱うデジタル翻刻プロジェクトは我々の知る限り存在しない．

石原らのデジタル翻刻システム EBIS に関する論文 [8] では，クラウドソーシングでも利用可能なように設計したと記載されている．しかし，クラウドソーシングには不可欠な不特定多数の入力を制御する機能や，実際にクラウドソーシングを行ったという記述はない．

一方，L-Crowd プロジェクトでは，最初の試みとして，国立国会図書館の書誌データを対象として，マイクロタスク型クラウドソーシングを用いて誤同定された資料を発見する取り組みを行った [9]．しかし，我々の知る限りマイクロタスク型クラウドソーシングを用いて日本語資料のデジタル翻刻を行っている先行事例は存在しない．そこで，本研究ではマイクロタスク型クラウドソーシングを用いた効率の良いデジタル翻刻について検討する．

2. マイクロタスク型クラウドソーシングプラットフォーム Crowd4U

本節では，マイクロタスク型クラウドソーシングプラットフォームである Crowd4U の説明を行う．Crowd4U とは，大学等の研究者が協力して構築・運用する非営利・オープン・汎用のプラットフォームであり，マイクロタスクという 1 回あたり 10 秒程度の，ごく短時間で作業できる作業が多数登録されている．以降，作業の単位を単にタスク，タスクを行う人をワーカと表記する．

図 1 に Crowd4U の概要を示す．まず，Crowd4U ではタスクを宣言的に定義し，タスクプールに保存する．タスクは 4 つの選択肢の中から 1 つ，あるいは複数を選択するような簡単なものから，何らかの文字を入力させるといったことまで自由な設計が出来る．さらに，あるワーカに入力させた内容を別のワーカに確認させるなど，段階的なタスク設計も可能である．タスクプール内のタスクは各端末・システムから呼び出され，実行される．

タスクは各端末・システムに自由に配信できるが，操作する端末やシステムによって向き不向きがあるため，それぞれに応じたビューの設定やタスク設計を行うことが望ましい．例えば，最大 4 つの選択肢の中から 1 つを選ぶようなタスクであればスマートフォン上でも行うことが容易であり，ロックアプリとして開発を行っている．ここで，ロックアプリとはスマートフォンをロック状態から復帰させるときに実行させるアプリケーションである．単純なアプリケーションが行われるかはユーザの意欲に直結するが，ロックアプリでは日常的にタスクが行われることが期待される．

さらに，最大 3 つの選択肢の中から 1 つを選ぶようなごく簡単なタスクであれば床にタスクを投影し，その上を通過する人の動きを Kinect センサーを用いて計測することで，床を歩くだけでタスクを解かせるといったこともできる [12]．これは別途床システムとして開発しているが，本稿ではその詳細な説明は省略する．

なお後述する翻デジマイクロタスクも含め，Crowd4U に登録されたタスクは誰でも，いつでも，どこでも行うことができる．タスクを行うにあたりユーザ登録は不要だが，ユーザ登録をすることで何時間タスクを行ったかという証明書の発行や，タスク処理数のランキングに参加することができるようになる．

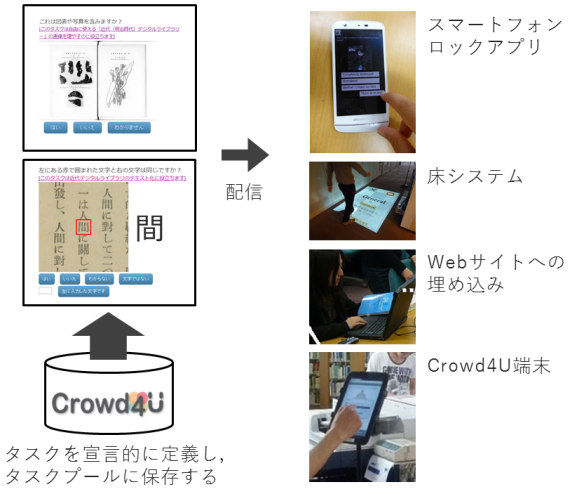

図 1　Crowd4U の概要

3. 翻デジマイクロタスク

この節では，現在行っている翻デジに関する 2 種類のタスクである図表包含判定タスクと OCR 結果校正タスクの

説明を行う．このうち，OCR 結果校正タスクは運用に伴い得られた知見を元に改良したため，改良前と改良後の両方を説明する．

3.1 図表包含判定タスク

このタスクでは，国立国会図書館デジタルコレクション収録の資料に対し見開きページ内に図表や写真が含まれるかを判定する．2015 年 2 月 2 日から実施しており，2016 年 4 月 10 日現在までに累計 76,306 タスクが，さらに，2015 年 2 日 3 日から 2016 年 4 月 10 日現在まで，2 章で述べた床システムにより 76,895 タスクが別途行われており，累計 153,201 タスクが行われている．

このタスクでは現在，著作権を気にすることなく使える画像を抜き出すことを目的としているが，OCR においてノイズとなりうる図表を除去することに資するなど，OCR 結果の改善に役立てることも検討している．

本タスクの例を図 2 に示す．このタスクでは見開き 2 ページを表示し，図表や写真が含まれるかをワーカに尋ね，ワーカは「はい」「いいえ」「わかりません」の 3 つの選択肢の中から回答する．そして，同じタスクを複数のワーカに課し，図表や写真が含まれるかを多数決で判定する．例えば，3 人のワーカがそれぞれ「はい」「はい」「いいえ」と回答した場合，その見開きページには図表が含まれるものと判定する．なお，「わかりません」という回答は無視される．

図 2 図表包含判定タスクの例

3.2 OCR 結果校正タスク（文字の入力なし）

このタスクでは，岩波講座日本歴史．第 1 [13] の OCR 結果 1 文字 1 文字に対し，その正誤を判定する．2015 年 11 月 12 日から 12 月 15 日まで，累計 9,996 タスクが行われている．

タスクの例を図 3 に示す．まず，OCR では認識した 1 文字 1 文字に対し，候補とその確からしさ（信頼度）を複数出力する．このタスクでは認識した文字を赤枠で囲み，信頼度が最も高い文字を並べて表示し，OCR の認識結果が正しいかをワーカに尋ねる．ワーカは「はい」「いいえ」「わからない」「文字ではない」の 4 つの選択肢から 1 つを選ぶ．同じタスクを複数のワーカに課し，累計 3 人のワーカが「はい」と回答した場合はその文字の OCR 認識が正しいとみなす．累計 2 人のワーカが「いいえ」と回答した場合は次に信頼度が高い候補で処理を繰り返し，累計 3 人のワーカが「文字ではない」と回答した場合は認識した箇所が文字ではないと記録する．例えば，1 人目のワーカが「はい」，2 人目のワーカが「いいえ」，3 人目のワーカが「はい」と回答したとき，4 人目のワーカが「はい」と答えた場合はその文字は正しいとして次の文字の判定に移る．「いいえ」と答えた場合はその文字は誤りとして，次に信頼度が高い候補で判定を続ける．なお，候補が「いいえ」により除外された場合は，本タスクではその箇所について文字を出力しない．

しかし，本タスクでは OCR 結果の候補内に正答の文字が全く含まれないことがあったことと，文字ではない箇所を文字として大量に認識してしまったデータをそのまま投入してしまったことが原因となり実行されたタスク数に対し思うように校正が進まなかったため，目視で文字ではない部分が含まれるページを落とし，3.3 で示す，文字の入力も可能にしたタスクに切り替えた．

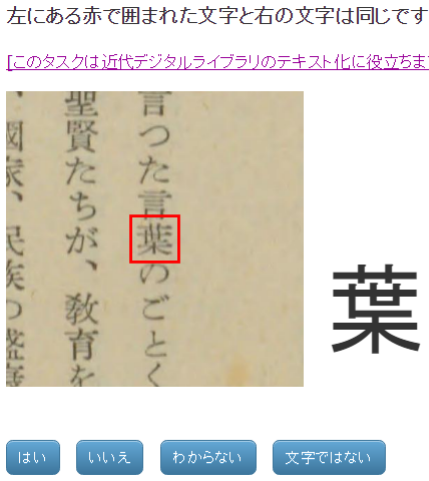

図 3 OCR 結果校正タスク（文字の入力あり）の例

3.3 OCR 結果校正タスク（文字の入力あり）

このタスクでは 3.2 で述べたタスク同様，岩波講座日本歴史．第 1 [13] の OCR 結果 1 文字 1 文字に対しその正誤を判定する．2015 年 12 月 8 日から実施しており，2016 年 4 月 10 日現在までに累計 10,630 タスクが行われている．

本タスクの例を図 4 に示す．3.2 のタスクとの最大の違いは，ワーカが「はい」「いいえ」「わからない」「文字ではない」の 4 つの選択肢に加え，文字を入力できるようになったことである．ワーカが「はい」，「いいえ」，「わから

ない」，「文字ではない」と回答した場合の処理は 3.2 のタスクと同様に行うが，ワーカが文字を入力した場合は「いいえ」としてカウントされ，さらにその文字を次に信頼度が高い候補として設定する．例えば，OCR が認識したある候補に対し 1 人目のワーカが「はい」，2 人目のワーカが「いいえ」と回答し，3 人目のワーカが文字を入力した場合は，その候補は誤っていたと処理され，3 人目のワーカが入力した文字を候補としてまた他のワーカに判定を行わせる．

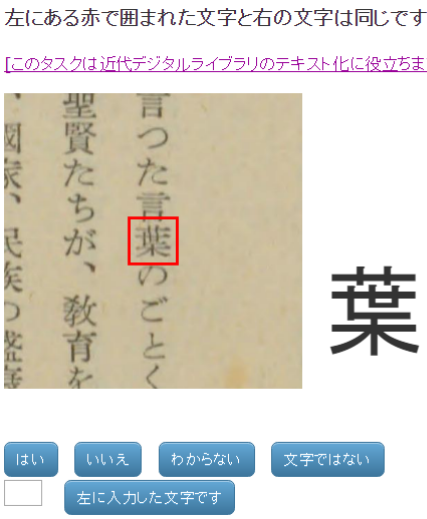

図 4　OCR 結果校正タスク（文字の入力なし）の例

4.　タスク結果の分析

この節では，3 節で述べたタスクの実行結果を述べ，その分析を行う．まず，タスクの実行回数の推移を述べる．次に，3.2 と 3.3 で示したタスクにおいて，校正が正しく行われたかを検証する．

4.1　タスクの実行回数

3 節で示した 3 つのタスクの実行回数が月ごとにどのように変化しているのかを調べた．その結果を図 5 に示す．ただし，3.2 と 3.3 のタスクは OCR 結果校正タスクとして 1 つにまとめている．

まず，黄色の破線で表した OCR 結果校正タスクと青色の点線で表した図表包含判定タスクでは月によって行われたタスク数に最大数千の差が出ているが，これは後者のみスマートフォンのロックアプリにも配信が行われていることが原因だと考えられる．また，黄色の破線で表した OCR 結果校正タスクと灰色の実戦で示した図表包含判定タスクでは 1 万以上の差があるが，これはスマートフォンのロックアプリに加え，床システムの有無による差であると考えられる．

全体として，2015 年 11 月以降に行われたタスク数は単調減少の傾向にあるが，これはリリースしたタイミングでは注目が集まるが，継続して定期的にタスクをこなすワーカはそう多くはないということが原因として推測される．従って，安定した数のタスクを処理するためには床システムやスマートフォンのロックアプリのような日常的に行われるシステムにも配信をすることが重要だと考えられる．ただし，床システムにおいても 2015 年 8 月，2015 年 9 月，2016 年 3 月は他の月に比べタスク実行数が大きく減少している．これは，床システムは大学の学内に設置していることが要因となり，長期休暇中の期間に一時的にタスク数が減ったのだと考えられる．

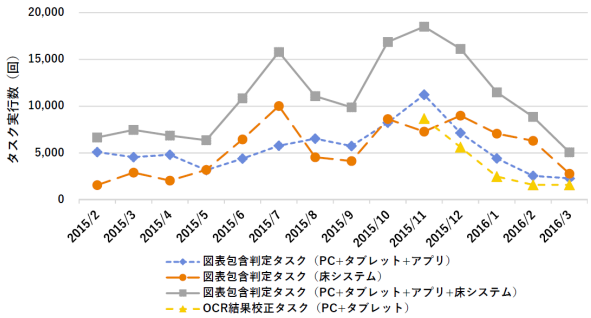

図 5　タスク実行回数の月別推移

4.2　OCR 校正結果の分析

3.2 と 3.3 で示したタスクでは，累計 20,644 タスクが実行され，3,979 文字が校正結果として得られた．

このうちデジタル翻刻対象の先頭 285 文字に対し，OCR をかけるだけで得られた文字列とそれをタスク結果として得られた文字列を目視で元の画像と比較し，正しくデジタル翻刻が出来ているかどうかを検証した．結果を表 1 に示す．

まず，単に OCR を掛けた場合の F 値が 76.22 なのに対し，タスクの校正結果は 89.56 になり，13.34 の向上がみられた．また OCR の誤り 69 文字の中には 2 文字，本来は文字ではない箇所を文字としていた誤認識があったが，タスクによりそれらを除去することができた．一方，タスク結果がデジタル翻刻対象に対し 5 文字減っているが，これは当該文字に対する全候補に対し，ワーカが正しくないと回答したことが原因である．

次に，OCR 認識が誤りだった文字とその正答の対応を表 2 に，タスクによる校正結果が誤りだった文字とその正答の対応を表 3 に，さらにそれらの誤りの類型をまとめたものを表 4 に示す．ここで，表 4 における『「あ」と「ぁ」等の同形の誤り』は表 2, 3 における平仮名の大文字小文字，及び「〇」と「。」のみを対象としており，「<」と「く」のような誤りは全く違う文字の誤りとしている．

まず，OCR 認識では正しかった文字がタスク結果においては誤った文字となってしまったケースは 1 件も存在しなかった．次に，旧字体・異体字・新字体間の誤りについて

はシステムからワーカに校正の指針を提示しなかったことが原因だと考えられる．そもそも，旧字体・異体字・新字体といった字体の微細な違いが存在する場合の校正は永崎が [16] で述べているように，文章の内容についてのテクスト解析を行うのであればなるべく現代の扱いやすい漢字に置き換えるべきであり，文献学的研究や字体史研究等を行うのであれば微細な違いも可能な限り分けるべきという 2 つの立場がある．前者の立場であれば一定のルールを設け現代の文字に置き換える，後者の立場であれば多漢字フォントを用いるか外字を使うということが考えられるが，どちらの立場に立つかについては今後検討が必要である．

タスク結果における誤りのうち，最も多い誤りは「あ」と「ぁ」,「〇」と「。」のような同形の文字間での誤りであり，明かな誤りや文字ではない箇所を文字とした誤認識はタスクにより全て除去できていた．同形だが大きさが異なり違う文字とされるものを誤る問題については，大文字小文字が区別しやすいフォントを用いる，候補を枠で囲む，ガイド線を付与するなどによりワーカが判別しやすくすることで改善が可能であると考えられる．

最後に，OCR 認識では何らかの文字であると認識していたが，タスク結果においては文字と認識されず，元のデジタル翻刻対象に対し 5 文字が欠落してしまった．これは，3.2 で示したタスク設計が OCR 認識の候補の中から正解を選ぶ，すなわち OCR 認識の候補の中に正解が必ず存在するということを暗に前提としていたことが原因である．この問題は 3.3 のようにワーカに文字を入力させることで解決が図られていると考えられる．

表 1 デジタル翻刻の結果

	文字数	適合率	再現率	F 値
デジタル翻刻対象	285	-	-	-
OCR 結果	287	75.96	76.49	76.22
True positive	213	-	-	-
False positive	69	-	-	-
False negative	0	-	-	-
タスク校正結果	280	90.36	88.77	89.56
True positive	253	-	-	-
False positive	27	-	-	-
False negative	5	-	-	-

5. ディスカッション・今後の展開

4 節で述べたように，タスクの実行がタブレットや PC からに限られてしまうと安定してタスクが行われなくなる．従って，タスクの一部をスマートフォンのロックアプリや床システムのような日常的に行うことが可能なシステムで実行が可能なように設計することで，より短時間でデジタル翻刻が可能になると考えられる．この際，文字を入力させるといった比較的複雑なタスクは少ない数で済むよう，粒度の異なる複数のタスクを組み合わせることができればより効率的に問題が解決できるようになることが考えられる．

また，OCR においては図表がノイズになる．さらに，国立国会図書館デジタルコレクションに収録されている資料の中には本文の上から蔵書印が押されているものや，書き

表 2 OCR 認識が誤りだった文字とその正答

OCR 認識	正答	回数	備考
0	つ	4	
0	の	12	
0	り	1	
0	。	3	
5	ぃ	1	
9	つ	1	
A	人	3	
L	し	2	
N	二	1	
T	で	1	
ぁ	あ	9	
ぅ	う	1	
っ	つ	1	
ゃ	や	2	
ゎ	わ	1	
社	社	2	
進	進	1	
精	精	1	
問	間	1	
兌	見	1	
達	達	1	
宵	育	1	
有	育	1	
尊	尋	1	
.		1	文字ではない箇所を誤判定
’	、	1	
,	、	1	
\	ゝ	1	1 文字繰り返し文字
-	〳	1	2 文字繰り返し文字
-		1	文字ではない箇所を誤判定
>	、	5	
<	く	4	
〇	り	1	

表 3 タスク結果が誤りだった文字とその正答

校正結果	正答	回数
ぁ	あ	9
ぃ	い	1
ぅ	う	1
っ	つ	6
ゃ	や	2
ゎ	わ	1
社	社	2
精	精	1
達	達	1
進	進	1
〇	。	2

表 4 OCR 認識とタスク結果の誤り内訳

	OCR 認識	タスク結果
旧字体・異体字・新字体間の誤り	5	5
「あ」と「ぁ」等の同形の誤り	14	22
「A」と「人」等の全く違う文字の誤り	48	0
文字ではない箇所を文字と誤認識	2	0
認識すべき文字の欠落	0	5

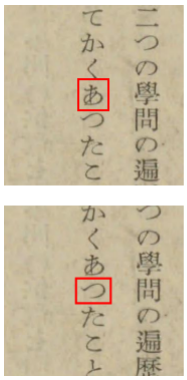

図 6 校正の誤りの例

込みされているものもある．従来はこれらのノイズは目視で除去していたが，この部分もクラウドソーシングにより除去できるようになることが望ましい．

さらに，現行の OCR 校正のタスク設計では実行された総タスク数に対し校正された文字が概ね 5 タスクで 1 文字でありそう多いとは言えず，さらなる効率化の余地がある．例えば，1 文字単位ではなく文字列単位で確認を行い，誤りが含まれると判定された文字列に対してのみ文字単位で OCR 校正を行うことで，大幅にタスク数が削減できることが考えられる．ここで，文字列の望ましい長さは OCR の適合率によって変わるが，文字列長を動的に変化させることで OCR の適合率に関わらずデジタル翻刻が効率良く行えるようになることが見込まれる．

これらを踏まえ，本文箇所判定タスク，OCR 修正箇所判定タスク，文字校正タスクという粒度の異なる 3 段階のマイクロタスクを用いることでより効率良くデジタル翻刻を行うことを目指す．

本文箇所判定タスク

一般に，OCR ではレイアウト解析，ブロック解析，行解析，文字解析を段階的に行う．このうち，3.3 で示したタスクでは文字解析の結果だけを利用していたが，他の解析の結果も活用することでさらなら効率化が望まれる．

このタスクでは，OCR で認識したブロック領域を示し，その領域に本文が含まれるかどうかを「はい」「いいえ」「わからない」の 3 つの選択肢の中からワーカに選択させるタスクである．これにより，本文ではない箇所を後のタスク対象から外すことができるため，総タスク数の削減が見込まれる．

また，このタスクは 3 つの選択肢から 1 つを選ぶという非常にシンプルなタスクであるため，スマートフォンのロックアプリや床システムといった日常に紐付いたタスクとして多くの数が処理されると考えられる．

OCR 修正箇所判定タスク

このタスクでは，まず本文箇所判定タスクで本文が含まれていると判定された領域に含まれる文字のうち，数文字を結合する．次に，それぞれの文字で最も信頼度の高い文字も同様に結合して併記し，それらが完全に一致しているかをワーカに「はい」「いいえ」「わからない」の 3 つの選択肢の中から 1 つを選ばせる．複数人に「はい」と判定された場合は当該文字列に含まれる文字は全て正しいとし，「いいえ」と判定された場合は誤りを含む文字を含むとして次のタスクに処理を回す．ただし，文字列の長さが一定以上の場合は文字列を 2 分割し，このタスクを繰り返すことで総タスク数の削減を図る．

ここで，1 度に判定する文字列の長さは可変であり，初期値として 2 を割り当てるが判定結果に応じて動的に変更する．直感的には，「はい」，すなわち誤りが含まれないと判定されるほど一度に判定する文字列を長くし，「いいえ」，すなわち誤りを含むと判定されるほど一度に判定する文字列を短くする．

行やページを跨ぐ場合に対応するため，事前に OCR が認識した文字単位で画像を切り出すか，IIIF [7] のような動的に画像を処理できる仕組みを用いる必要がある．

文字ではなく文字列で一度に複数の文字を判定することは，総タスク数を削減できる可能性がある他，文脈が利用できるようになることが利点として挙げられる．例えば，4.2 で述べたように現行のタスクでは大文字と小文字の区別が付きにくいという問題があったが，文脈を利用することで判断ミスが軽減されると考えられる．

文字校正タスク

このタスクでは，OCR 修正箇所判定タスクで誤りが含まれると判定された文字列を文字に分割しなおして，3.3 で説明したタスクと同様に，OCR の対象と認識した文字の中で最も信頼度が高い者を並べ，ワーカにそれら 2 つは同一かどうかを尋ねる．ワーカは「はい」「いいえ」「わからない」「文字ではない」の 4 つの選択肢の中から 1 つか，あるいは正しい文字を入力させる．

ただし，3.3 で示したタスクではワーカが文字を入力した場合，「いいえ」としてカウントし，入力された文字を次に信頼度が高い候補として処理を進めていたが，さらに入力された文字に対し「はい」と回答されたとして処理を進める．

6. まとめ

本稿では，マイクロタスク型クラウドソーシングを用いたデジタル翻刻の取り組みの紹介と，これまで行われたタスク結果の分析，これからの展望を述べた．本研究により，1. マイクロタスク型クラウドソーシングを用いた場合でも OCR に比べ比較的高品質なデジタル翻刻が可能になること，2. 日常的に行うことが可能なようにタスクを設計することで安定的にタスクが行えること，3. 翻刻の効率と実行

されるタスク数はトレードオフの関係にあり，そのデザインが重要なことの 3 点が知見として得られた．

今後の課題としては文字の向きが異なる場合にも対応したタスクの設計や，端末やシステムによって再現率・適合率が変わるかといった比較・検証が挙げられる．

謝辞 本研究の一部は科研費 (#25240012) による．

参考文献

[1] Australian Newspaper Digitisation Program(オンライン), 入手先 ⟨http://www.nla.gov.au/content/newspaper-digitisation-program⟩ (参照 2016-04-11).
[2] Causer, T., Tonra, J. and Wallace, V.: Transcription maximized; expense minimized? Crowdsourcing and editing The Collected Works of Jeremy Bentham*, *Literary and Linguistic Computing*, Oxford University Press, Vol. 27, No. 2, pp. 119–137 (2012).
[3] Crowd4U(オンライン), 入手先 ⟨https://crowd4u.org/ja/⟩ (参照 2016-04-01).
[4] L-Crowd(オンライン), 入手先 ⟨https://crowd4u.org/ja/projects/lcrowd⟩ (参照 2016-04-01).
[5] Terras, M.: Present, not voting: Digital Humanities in the Panopticon: closing plenary speech, Digital Humanities 2010, *Literary and Linguistic Computing*, Oxford University Press, Vol. 26, No. 3, pp. 257–269 (2011).
[6] National Archives Transcription Pilot Project(オンライン), 入手先 ⟨http://www.archives.gov/citizen-archivist/⟩ (参照 2016-04-11).
[7] International Image Interoperability Framework(オンライン), 入手先 ⟨http://iiif.io/⟩ (参照 2016-04-13).
[8] Ishihara, T., Itoko, T., Sato, D., Tzadok, A. and Takagi, H.: Transforming Japanese Archives into Accessible Digital Books, *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM, pp. 91–100 (2012).
[9] Morishima, A., Tomita, S., Kawashima, T., Harada, T., Uda, N., Sato, S. and Abematsu, Y.: A Crowdsourcing Approach for Finding Misidentifications of Bibliographic Records, *Proceedings of the iConference 2014*, iSchools, pp.166–191 (2014).
[10] 相原健郎, 林晋: 画像化主義に基づく文献資料研究用ツール SMART-GS とその発展, 情報処理学会研究報告 *(デジタルドキュメント)*, 2011-DD-79, pp.1–5 (2011).
[11] 青空文庫 (オンライン), 入手先 ⟨http://www.aozora.gr.jp/⟩ (参照 2016-04-11).
[12] 太田千尋, 森嶋厚行, 中村聡史, 寺田努, 北川博之: 歩行中のマイクロタスク処理におけるデータ品質向上手法とその評価, 第 *8* 回データ工学と情報マネジメントに関するフォーラム *(DEIM2016)*, D6-3, 7p (2016)
[13] 国史研究会編: 岩波講座日本歴史. 第 1 (総説・古代), p.40, 岩波書店 (1935). 入手先 ⟨http://kindai.ndl.go.jp/info:ndljp/pid/1263939⟩
[14] 国立国会図書館：国立国会図書館デジタルコレクション (オンライン), 入手先 ⟨http://dl.ndl.go.jp/⟩ (参照 2016-04-01).
[15] 永崎研宣, 鈴木隆泰, 下田正弘: 大正新脩大蔵經テキストデータベース構築のためのコラボレーションシステムの開発, 情報処理学会研究報告 *(人文科学とコンピュータ)*, 2006-CH-070, pp. 33–40 (2006).
[16] 永崎研宣: 日本語クラウドソーシング翻刻に向けて (<特集 >デジタル時代の日本語), 情報の科学と技術, 社団法人情報科学技術協会, Vol. 64, No. 11, pp. 475–480 (2014).