DEIM Forum 2017 P7-5

日本語資料のマイクロタスク型Human-assisted OCRの試み

池田　光雪† 　林　亮太†† 　永崎　研宣††† 　森嶋　厚行††††

†† 千葉大学アカデミック・リンク・センター　〒 263–8522 千葉県稲毛区弥生町 1–33
†† 筑波大学図書館情報メディア研究科　〒 305–8550 茨城県つくば市春日 1–2
††† 一般財団法人人文情報学研究所　〒 113–0033 東京都文京区本郷 5–26–4–11F
†††† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †lumely@chiba-u.jp, ††ryota.hayashi.2014b@mlab.info, †††nagasaki@dhii.jp,
††††mori@slis.tsukuba.ac.jp

あらまし　近年，著作権保護期間が満了しているといった利用が容易な資料に対し，画像でのみ提供されている資料をクラウドソーシングによりデジタル翻刻を行うなどの取り組みを通じて，資料のさらなる活用に有用なデータを生成するという動きが広まっている．それらにより，従来の専門家のみによるアプローチでは事実上困難であったデジタルアーカイブの拡充が今後益々進むことが考えられる．本稿では，国デコ翻デジ@ JADH×Crowd4U として取り組んできたマイクロタスク型クラウドソーシングによるデジタル翻刻の試みを報告すると共に，本アプローチが有効であったことを示す．さらに，今後の展開についても論ずる．

キーワード　クラウドソーシング, マイクロタスク, デジタル翻刻

1. はじめに

近年，クラウドソーシングによるデジタルアーカイブの拡充が盛んに行われている．特に，著作権保護期間が満了しているような利用が容易な資料に対し，画像でのみ提供されている資料のデジタル翻刻，いわゆる文字起こしをすることでさらなる活用に有用なデータを生成するという動きが広まっている．たとえば，2017 年 1 月には京都大学古地震研究会が歴史災害のオンライン翻刻プラットフォームである「みんなで翻刻」[13] をオープンした．みんなで翻刻はくずし字で筆記された古文書・古記録を対象としており，解読にはある程度の専門知識が求められるにも関わらず，公開から 38 日で約 95 万字が入力され，翻刻対象としている 3,193 枚の画像中，1,325 枚の翻刻が完了するなど，爆発的な注目を集めている．これらの背景には，従来，専門家が単独で行う翻刻作業には莫大なコストが必要であったという一方で，資料に対する関心が高く，翻刻作業に携わりたい非専門家が多く存在しており，両者をクラウドソーシングにより結びつけることができたということがあると考えられる．

このように，クラウドソーシングによるデジタル翻刻は大きな可能性を秘めているにも関わらず，我々の知る限りこれらに着目した研究は現在ほとんど研究は行われていない．一方我々は，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 [4] を用いて，図書館領域の問題の解決を試みる L-Crowd [5] 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国立国会図書館デジタルコレクション [15] に所蔵された資料を対象としたデジタル翻刻を行ってきた [10]．本稿では，国デコ翻デジ@ JADH×Crowd4U で行われているタスクから得られた知見を踏まえ，OCR によって得られたデータをマイクロタスク型クラウドソーシングによって校正することでデジタル翻刻するアプローチを Human-assisted OCR として定義し，新しい Human-assisted OCR のタスク設計を提案する．

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．

本稿の構成は次の通りである．2 節では関連研究について述べる．3 節では本提案手法を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 について述べる．4 節ではこれまで実施してきたマイクロタスク型クラウドソーシングによるデジタル翻刻のタスク設計を述べ，5 節ではそれを改善した新しいタスク設計を提案する．6 節では本稿のまとめ及び今後の課題について述べる．

2. 関連研究

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 [9] 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 [16] がある．

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海外を中心として数多く存在し，Australian Newspaper Digitisation Program [2] ではオーストラリアの新聞を，National Archives Transcription Pilot Project [7] 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 [3], [6]．また，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ているか，非常に限られた人数で 1 資料の翻刻を行う「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非常に限られている．

日本点字図書館と国立国会図書館は視覚障害者をはじめとする視覚による表現の認識に障害のある人々の読書環境向上のためアクセシブルな電子書籍製作実験プロジェクトを実施しており，オンラインコミュニティサイトであるみんなでデイジー [18] を運用している．みんなでデイジーでは参加にあたりユーザ登録及び認証が必要だが，OCR の校正を一文字単位や文字列単位で行うことで翻刻を行うことができる．一方，我々が提案するデジタル翻刻ではユーザ登録や認証を必要としないプラットフォームである Crowd4U を利用しており，不特定多数が翻刻に参加することが可能である．さらに，スマートフォンのロックを解除する際にタスクを表示するアプリや，床にタスクを投影しその上を歩くだけでタスクを行うことができるシステムにタスクを配信することができ，日常動作とタスクを紐付けることで安定して翻刻作業を行うことができるという特徴を持つ．

3. 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．タスクは自由な設計ができ，4 つの選択肢の中から 1 つを選ばせる簡単なものから，画像を表示し，その画像に対応した何らかの文字を入力させるような複雑なことも可能である．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できる．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，ごく短時間で最大 3 つの選択肢の中から 1 つを選ぶような非常に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]．

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．

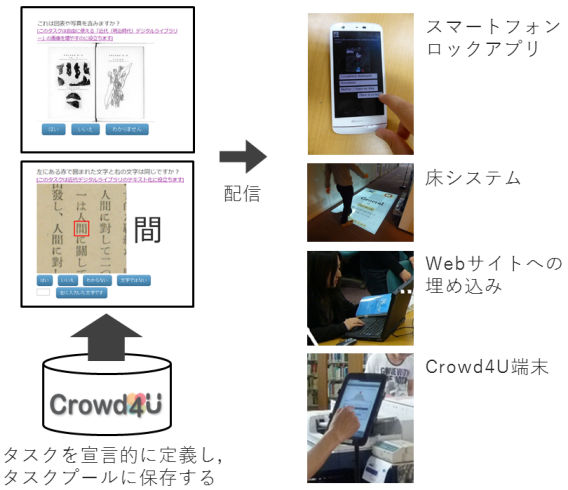

図 1　Crowd4U の概要

4. 翻デジマイクロタスク

本節では，国デコ翻デジ@ JADH×Crowd4U として実施してきたデジタル翻刻タスクの説明を行う．

このタスクでは，OCR ソフト ABBYY FineReader [1] を試用して得られた岩波講座日本歴史．第 1 [14] 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 2016 年 12 月 21 日まで，累計 40,743 タスクが行われている．

本タスクの例を図 2 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 つを選ぶか，直接正しい文字を入力する．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，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．また，全ての候補が「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．

本タスクでデジタル翻刻の対象とした文書に対し，先頭の 2 文，計 285 文字における OCR のみによる処理結果と，タスク結果の適合率，再現率，F 値の比較を表 1 に示す．

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，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，本節で示したタスクの初期バージョンにおいてはワーカに正しい文字の入力を許しておらず，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．本タスク結果で適合率が低くなった多くの原因は，デジタル翻刻対象においては旧字体や異体字だった文字がタスク結果においては新字体にしたことが理由として挙げられる．しかし，永崎が [17] 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，どちらの立場に立ってデジタル翻刻を行うかは，デジタル翻刻を行うことによってどのような活用を行うのかによって考える必要がある．

さらに，今回の評価において，字体や字形デザインの違いは可能な限り別のものとして扱い厳密に判定したが，それらが異なっていても同じ文字であると見なす包摂規準というものが定められている．この基準に従うことで OCR 結果，タスク結果共に適合率は上がると考えられる．

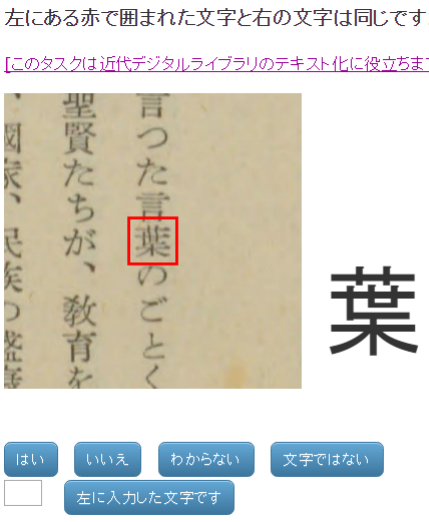

図 2　OCR 結果校正タスクの例

5.　提案タスクデザイン

前節で述べたデジタル翻刻のタスクは，実行された総タスク数に対し校正された文字は概ね 5 タスクで 1 文字と，多いとは言えず，さらなる効率化の余地がある．

表 1　285 文字に対するデジタル翻刻結果の比較

	文字数	適合率	再現率	F 値	TP	FP	FN
OCR 結果	287	0.75	0.76	0.76	213	69	0
タスク結果	280	0.90	0.88	0.89	253	27	5

本節では，3 種類のタスクを組み合わせることにより適合率を下げることなく，少ないタスク数でより多くのデジタル翻刻を行うことのできるタスク設計を提案する．

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 [8] 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．たとえば，4. 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，4. 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，4. 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. まとめと今後の課題

本稿では，これまで行ってきた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を述べ，それを改善した新しいタスク設計を提案した．

今後の課題として，タスクの粒度に関する検討が挙げられる．高い再現率を得ることができるが，1 文字の翻刻に必要な平均タスク数が大きすぎる問題がある．従って，5. で示すように 1 回のタスクで判定を行う文字数を 2 文字以上とすることや，OCR が苦手とする同形でサイズが異なる「〇」「。」などの文字を除き，OCR の信頼度が一定以上の文字はタスクの対象から除外するなどの工夫を行うことで翻刻結果の質をできる限り下げることなく，必要タスク数を大きく削減する必要がある．

謝　　辞

本研究の一部は，JST CREST および JSPS 科研費 (#25240012) の支援による．

文　　献

[1] ABBYYFineReader(オンライン)，http://finereader.abbysoft.jp/
[2] Australian Newspaper Digitisation Program(オンライン)，http://www.nla.gov.au/content/newspaper-digitisation-program
[3] 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).
[4] Crowd4U(オンライン)，https://crowd4u.org/ja/
[5] L-Crowd(オンライン)，https://crowd4u.org/ja/projects/lcrowd
[6] 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).
[7] National Archives Transcription Pilot Project(オンライン)，http://www.archives.gov/citizen-archivist/
[8] International Image Interoperability Framework(オンライン), http://iiif.io/
[9] 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).
[10] 池田光雪, 林亮太, 永崎研宣, 森嶋厚行: 翻デジにおけるマイクロタスク活用の試み, 人文科学とコンピュータ研究会第 *110* 回発表会, pp.1–7 (2016).
[11] 青空文庫 (オンライン)，http://www.aozora.gr.jp/ 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] 京都大学古地震研究会, “みんなで翻刻”, http://honkoku.org/, accessed 2017-01-16.
[14] 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). http://kindai.ndl.go.jp/info:ndljp/pid/1263939
[15] 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)，http://dl.ndl.go.jp/
[16] 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).
[17] 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).
[18] 日本点字図書館, “みんなでデイジー”, http://mindeji.lab.ndl.go.jp, accessed 2017-01-16.