様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ５ 月 ２５ 日現在

機関番号：６２５０１
研究種目：若手研究
研究期間：2018～2019
課題番号：１８Ｋ１８３３８

研究課題名（和文）クラウドソーシングと機械学習を統合した歴史資料翻刻システムの開発

研究課題名（英文）Integration of Crowdsourcing and Machine Learning for Large-scale Transcription of Pre-modern Historical Manuscripts

研究代表者
橋本　雄太（Hashimoto, Yuta）

国立歴史民俗博物館・大学共同利用機関等の部局等・助教

研究者番号：１０８０２７１２

交付決定額（研究期間全体）：（直接経費）　　1,600,000円

研究成果の概要（和文）：本研究は、文字認識技術とクラウドソーシングによる人海戦術を統合することで、日本語の歴史文献資料を効率的にテキスト化する手法を開発することであった。当初は本研究で文字認識技術の開発に取り組む予定であったが、「くずし字」の自動認識技術が急速に発展したことを踏まえて方針を転換し、文字認識研究者と協業して翻刻プラットフォームの開発にあたった。
その成果として、AI文字認識に対応した翻刻プラットフォーム「みんなで翻刻」を2019年7月に公開した。2020年4月までに「みんなで翻刻」では、約800人の参加者により250万字以上の歴史資料が翻刻されるという成果を挙げている。

研究成果の学術的意義や社会的意義
AI認識に対応した「みんなで翻刻」は、300日の短期間で250万字ものテキスト化を成し遂げた。本研究成果の直接的な意義のひとつは、AIの支援を通じて市民による翻刻作業の効率化が実際に可能であることを実際に示したことにある。
より大きな観点での成果は、技術の適切な組み合わせによって、①人文学研究者、②市民、③AI技術（およびその研究者）の三者が互恵的な関係を築くことが可能であると示したことにある。AI技術の発展が人文学研究と市民参加型研究の将来にもたらす影響について、これまで様々な議論がなされてきたが、本研究の成果は重要な参考事例のひとつになるはずである。

研究成果の概要（英文）：This research program aimed to develop a efficient method for transcribing pre-modern Japanese documents written with cursive characters (kuzushiji). Although the program initially planned to develop an OCR technology for kuzushiji on its own, this field has rapidly advanced over the past few years against the author's expectation.
This led the author to collaborate with the AI researchers who study the OCR technology for kuzushiji, rather than to compete with them. Through this collaboration, the author launched in 2019 a new version of "Minna de Honkoku", a crowdsourced transcription platform that supports automatic recognition of kuzushiji. Since its launch, 2.5 million characters have been transcribed on this platform by more than 800 participants.

研究分野：人文情報学

キーワード：クラウドソーシング　翻刻　IIIF　くずし字　歴史資料　文字認識

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

わが国には世界でも類を見ない点数の歴史文献資料が残されており、すでにデジタルアーカイブで公開される資料だけでも 10 万点を超える。しかしながら、江戸期以前の文献資料は手書きの「くずし字」で書かれており、既存 OCR 技術の適用が困難なことから、これまでにテキスト化（翻刻）された資料は 1 万点にも満たない。全文テキストの不在は、資料に書かれた内容の把握と情報検索に大きな困難をもたらしている。

一方で、近年は大規模な翻刻を実現するための技術開発が進み、歴史文献資料の全文翻刻の実現が現実味を帯びつつある。その第一の技術は、クラウドソーシングによる翻刻である。現代では使用されない「くずし字」のクラウドソーシング翻刻はこれまで困難とされてきたが、申請者が 2017 年 1 月に公開した『みんなで翻刻』では、公開 9 ヶ月間で 3300 名の参加者により 260 点（画像 4500 枚、文字数 310 万字）の文献資料が翻刻されるという成果を得た。

第二の技術は、機械学習、特に深層学習を利用した自動文字認識である。限定的な条件下ではあるものの、すでに畳込みニューラルネットワーク（CNN, Convolutional Neural Network）によって手書きのくずし字が高精度で認識できることが先行研究により示されている（早坂ほか 2016）。

一方で、上記の 2 つの技術にはそれぞれに難点がある。クラウドソーシング翻刻は確かに効果的であるものの、現時点では数百点規模の翻刻しか実現できておらず、スケーラブルでない。また機械学習を利用した自動認識は、手書きで書かれた古文書の多様な形態に対応することが難しく、柔軟性に欠ける。しかしながら、上記の 2 アプローチを組み合わせ、人間のもつ柔軟な認知能力と、機械の高度なパターン認識能力との協調を実現することで、歴史資料の高効率な翻刻が実現する可能性がある。

２．研究の目的

上述の背景と課題意識のもと、本研究の目的は、クラウドソーシング翻刻と機械による自動認識の手法を統合し、高効率な歴史資料のテキスト化手法を開発することであった。具体的な研究課題として設定したのは次の 3 点の実現である：

① 言語モデルを援用し文脈をとらえる「くずし字」認識の実現
② 人間の作業者と機械の自動認識との協調を実現する翻刻支援システムの開発
③ 歴史災害資料を対象とした翻刻支援システムの評価実験の実施

ただし、以下に述べるように、研究計画の立案時点から「くずし字」認識の研究分野には大きな変化が生じた。これを踏まえて上記の目標を変更し、AI 研究者との協業のもと②③の実現を優先的な目標とした。

３．研究の方法

研究計画の立案当時は、研究課題のひとつとして「くずし字」画像認識の研究を実施する予定であった。しかしながら、研究計画の申請時点から研究開始までの期間に、この分野で大きな変化があった。国文学研究資料館が公開した「くずし字データセット」を利用したくずし字認識の研究が急速に発展したのである。特に人文学オープンデータ共同利用センター（CODH）のタリン・カラーヌワット氏が開発した KuroNet と、凸版印刷株式会社が開発したくずし字 AI 認識システムは、歴史資料翻刻の実用に十分に耐えるレベルであった。また、2019 年にはくずし字認識の Kaggle コンペティションが開催されるなど、AI 研究からのこの分野への新規参入も拡がりつつある。

そこで本研究で独自に文字認識技術を開発する方針を転換し、AI 分野の研究者と協業することで当初の研究目的の遂行にあたった。具体的には、CODH のカラーヌワット氏と凸版印刷株式会社と研究協力し、彼らの開発した文字認識モデルを組み込んだクラウドソーシング翻刻プラットフォームの開発にあたった。

４．研究成果

その成果として、AI 文字認識に対応した新バージョンの翻刻プラットフォーム「みんなで翻刻」（https://honkoku.org/）を 2019 年 7 月に公開した。「みんなで翻刻」はもともと 2017 年に東京大学地震研究所が所蔵する災害資料のクラウドソーシング翻刻プラットフォームとして公開されたものであるが、新バージョンはデジタルアーカイブにおける画像共有の国際標準 IIIF（International Image Interoperability Framework）に対応し、IIIF 形式で画像を公開する国内外の任意のデジタルアーカイブから資料画像を取り込むことができる。2020 年 5 月時点では、ユネスコ世界記憶遺産に登録された東寺百合文書や、近世の疫病関連資料など、計 8 件の翻刻プロジェクトが「みんなで翻刻」上で公開されている。

また、新バージョンの「みんなで翻刻」には、2 種類のくずし字認識エンジンを搭載している。ひとつは CODH のカラーヌワット氏が開発した koguma-net である。Koguma-net は Tensorflow.js を用いて構築されており、Web ブラウザ上で動作する。もうひとつは凸版印刷株式会社が開発したくずし字認識システムである。こちらは凸版印刷の Web サーバー上に配備した認識エンジンを API 経由で利用している。2020 年 5 月時点では、いずれの認識エンジンも 1 文字ごとの認識にしか対応していないが、近日中に複数文字の同時認識や文字位置の推定にも対応する予定で

ある。

これらの認識エンジンを用いて、クラウドソーシング翻刻の参加者が文字の識別に困難を覚えた際、AI の判断を仰ぐことができるようなユーザーインターフェイスを「みんなで翻刻」上に実装した。図 1 に示すように、参加者が翻刻対象の資料画像の一部分を指定すると、AI の文字認識結果をスコア付きで表示する。CODH 製のエンジンと凸版印刷製のエンジンは切り替え可能であり、両者の認識結果を比較しながら最終的に作業者が文字の読み方を決定することができる。

それでは AI による文字認識は、実際にクラウドソーシング翻刻の効率化に寄与しただろうか。システムログによると、2019 年 7 月から 2020 年 5 月までの間に、AI による文字認識は 25,480 回使用された。同じ期間に記録された資料の編集回数は 10,246 回であるから、1 回の編集につき平均 2.5 回文字認識 AI が使用されたことになる。この数字からは、AI による文字認識が翻刻作業の支援機能として頻繁に利用されていたことが分かる。AI による支援が作業者に対して与えた具体的な効果（AI の正答率や利用傾向など）については、今後オンラインアンケートとログ分析による調査を予定している。

新バージョンの「みんなで翻刻」では、2019 年 7 月の公開から 10 ヶ月の間に 820 人のユニークユーザーが翻刻作業に参加し、すでに 594 点の歴史資料が翻刻されている。画像単位では、現在公開されている 23,731 枚中 3,906 枚の翻刻が完了した。入力文字数は 250 万文字に達するなど、急速なペースで翻刻作業が進んだ。今後詳しい分析が必要であるが、多数の市民による翻刻作業への参加を後押ししたのは、AI による支援の貢献が大きいものと思われる。

また、CODH との協業の延長上で、「みんなで翻刻」で入力された翻刻文を AI の訓練に利用するための研究を開始している。ただし「みんなで翻刻」上のテキストには資料画像中の座標情報が付与されていない。そこで、カラーヌワット氏により KuroNet を利用して文字の画像中の位置を推定する作業が進められている。この試みが実現すれば、旧バージョンも併せて 800 万文字を超える翻刻テキストを、くずし字認識 AI の訓練データとして利用することが可能になる。国文学研究資料館によって公開されているデータセットのサイズは 100 万文字程度であるから、かねてからの課題であったデータセットの整備が一挙に進む可能性がある。

AI 技術の発展が人文学研究と市民参加型研究の将来にもたらす影響について、これまで様々な議論がなされてきた。本研究のより大きな観点での成果は、図 3 に示すように、技術の適切な組み合わせによって①人文学研究者や学芸員などの専門家、②市民、③AI 技術（およびその研究者）の三者が互恵的な関係を築くことが可能であると示したことにある。AI 技術の発展が人文学研究と市民参加型研究にもたらす影響を論じる上でも、本研究の成果は重要な参考事例のひとつになるはずである。

図 1 文字認識結果の表示ダイアログ

図 2 KuroNet による文字位置の推定結果

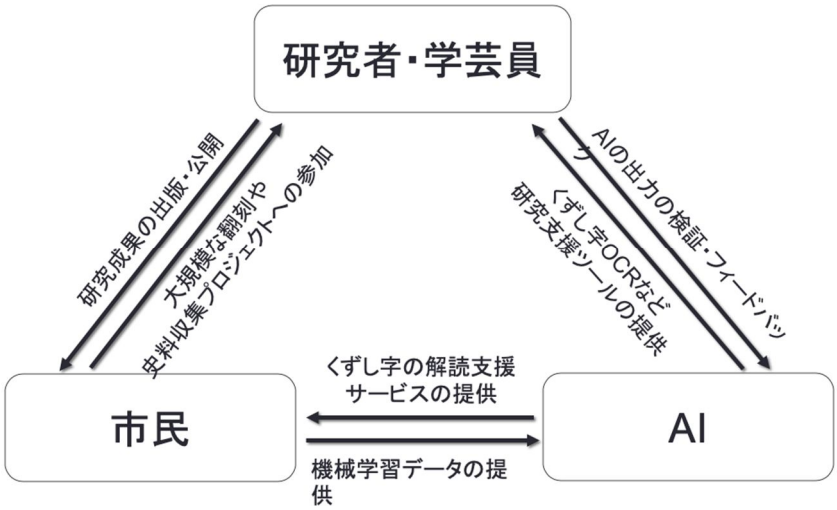

図 3 専門家・市民・AI の互恵的関係

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 橋本雄太	４．巻 848
２．論文標題 『みんなで翻刻』プロジェクト	５．発行年 2019年
３．雑誌名 日本歴史	６．最初と最後の頁 68-73
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計6件（うち招待講演　3件／うち国際学会　3件）

１．発表者名 Yuta Hashimoto, et al.
２．発表標題 Minna De Honkoku: Learning-Driven Crowdsourced Transcription Of Pre-Modern Japanese Earthquake Records
３．学会等名 Digital Humanities 2018（国際学会）
４．発表年 2018年

１．発表者名 橋本雄太, 宮川真弥
２．発表標題 日本語文献史料の構造化記述のための軽量マークアップ言語の開発
３．学会等名 人文科学とコンピューターシンポジウム2018
４．発表年 2018年

１．発表者名 Yuta Hashimoto
２．発表標題 Digital Humanities Research in National Museum of Japanese History
３．学会等名 The International Conference for Museums of Language & Writing 2019（招待講演）（国際学会）
４．発表年 2019年

１．発表者名
Yuta Hashimoto

２．発表標題
Honkoku2: Towards a Large-scale Transcription of Pre-modern Japanese Manuscripts

３．学会等名
The 9th Conference of Japanese Association for Digital Humanities (JADH2019)（国際学会）

４．発表年
2019年

１．発表者名
橋本雄太

２．発表標題
市民参加とAI―「みんなで翻刻」開発者の立場から

３．学会等名
日本文化とAIシンポジウム2019（招待講演）

４．発表年
2019年

１．発表者名
橋本雄太

２．発表標題
デジタルアーカイブの研究事例―みんなで翻刻を中心に―

３．学会等名
第20回図書館総合展 国立国会図書館フォーラム（招待講演）

４．発表年
2018年

〔図書〕　計1件

１．著者名	４．発行年
今村文彦　監修／鈴木親彦　責任編集	2019年
２．出版社	５．総ページ数
勉誠出版	208
３．書名	
災害記録を未来に活かす（デジタルアーカイブ・ベーシックス 2）	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考