研 究 終 了 報 告 書

「資料調査のためのオンデバイスくずし字認識」

研究期間： 2020年12月～2021年8月
研 究 者： カラーヌワット・タリン

1． 研究のねらい

日本は大量の歴史的資料がよく保存されており、その規模は 1 億点とも 10 億点ともいわれ、これらの資料を読み解ければ、これまで知られていなかった多くの事実が見えてくると考えられる。ところが、そこに立ちはだかるのが「くずし字」の問題である。くずし字とは日本で 1000 年以上も使われてきたが、明治 33 年の小学校令以降は学校教育でくずし字を教えなくなったため、ほとんどの現代日本人はくずし字が読めなくなっている。大量に残された歴史的資料に比べて読める人が少ないという問題が日本の歴史的資料の保存と活用を阻む一つの原因となっている。活字になった歴史的資料は研究者が重要性の高く自分が興味を持つ資料を優先して翻刻するため、活字化されている資料の数やジャンルは限定されている。一方、くずし字の読めない他分野の研究者や一般の人が興味のあるくずし字資料を調べようとしても、自力で調査ができないため、研究も進まない。日本文化と資料の情報を守るには多くの人からの協力が必要であり、一般の人でも資料を利用できるようになることが非常に大事である。この問題を解決するには、誰でも使えるくずし字認識ソフトウェアが必要であり、研究者の研究活動もより効率的になると考えられる。

しかし、研究代表者が開発したくずし字認識モデル KuroNet は現在一般公開された最も精度の高いモデルであるが、この KuroNet サービスは IIIF で公開されていない資料には対応しておらず、手持ちの資料をすぐに調査できないという問題がある。KuroNet モデルをオフライン認識で使おうとしても、高性能の GPU を用意し、パソコンを設定し、Ubuntu、PyTorch、CUDA などの各種ソフトウェアをインストールして Python スクリプトを実行しなければならい。しかし、くずし字認識の大半のユーザである一般の人、高齢者、文系研究者、図書館、博物館の学芸員にはハードルが高すぎる。そこで、資料調査の現場で誰でもくずし字認識サービスを使えるようにすることを目標に、オンデバイスくずし字モバイル認識アプリを開発する。

2． 研究成果

（1）概要

2020年度は研究の環境を整える準備期間であり、データ準備を行った。オンデバイスくずし字認識アプリは認識可能の文字種類の数が限られているため、資料に出現する頻度の高い文字を選択し、くずし字データセットのデータフォーマット、データクリーニングの作業が完了した。さらに、当初の研究計画になかった辞書機能とテキスト出力機能を追加した。2021年4月～6月にくずし字認識アプリを慶應義塾ミュージアム・コモンズ（Kemco）の特別展覧会「交景：クロス・スケープ」、特別プログラム「文字形－AI が開く くずし字の風景」[1]でアプリの体験

[1] 慶應義塾ミュージアム・コモンズ「交景：クロス・スケープ」、特別プログラム「文字形－AI が開く くずし

版を展示するため、アプリの初期バージョンの開発を完了した。

2021年度にアプリのオンデバイス文字認識モデルを開発した。Kemco の来館者の意見や感想を参考にし、アプリの最終テスティングを行っている。

（2）詳細

ACT-X で開発したくずし字認識アプリは「みを」と名付けた。「みを」は『源氏物語』第一四巻「みをつくし」にちなんだ名前で、「みを（船の水路）を示すために立ててある杭」の意であり、「みをつくし」が人々の水先案内となるように、「みを」アプリがくずし字資料を読むための道案内となることを目指している。

アプリの主要部分のインターフェースは以下の通りである。

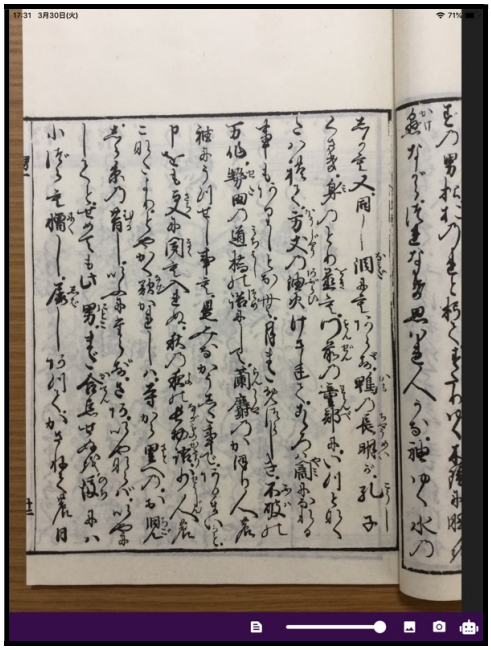

図 1 資料の写真を撮影する。

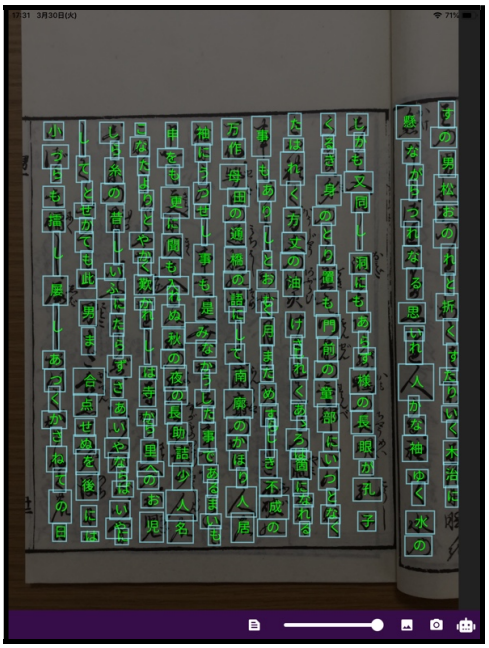

図 2 認識結果を表示する。

最初のくずし字データの準備期間ではくずし字データセットにある旧字体文字を新字体に全部変換した。さらに、くずし字認識アプリのオンデバイスで認識可能な文字の種類を制限するため、資料に出現する頻度の高い文字を選択し、くずし字データセットのデータフォーマット、データクリーニングの作業が完了した。そして、当初の計画書になかった新しく追加した機能二つある。一つ目は辞書機能である。ひらがなの辞書機能は国立国語研究所の「変体仮名フォント」[2]を利用し、変体仮名のバリエーションを表示する。そして、漢字辞書機能は現代漢字辞書ではあるが、Kanji API Dev[3]を利用している。

字の風景」のプレスリリース https://www.keio.ac.jp/ja/press-releases/files/2021/4/19/210419-2.pdf

[2] https://cid.ninjal.ac.jp/kana/font

[3] https://kanjiapi.dev/

ところが、Kemco の来館者や文学研究者の感想と意見から見ると、くずし字アプリを授業で使用したいという声が多かったので、人間のくずし字を学習するための機能も必要と考えた。追加した現代日本語の漢字辞書は多少利用者に役に立つが、各文字の AI モデルの推定結果を確率順で表示したほうが解読に役に立つのではないかと考え、漢字辞書の代わりにオンデバイス画像認識（MobileNetV2）で文字認識を行う機能を追加した。推定文字の上5位まで表示するようにした。

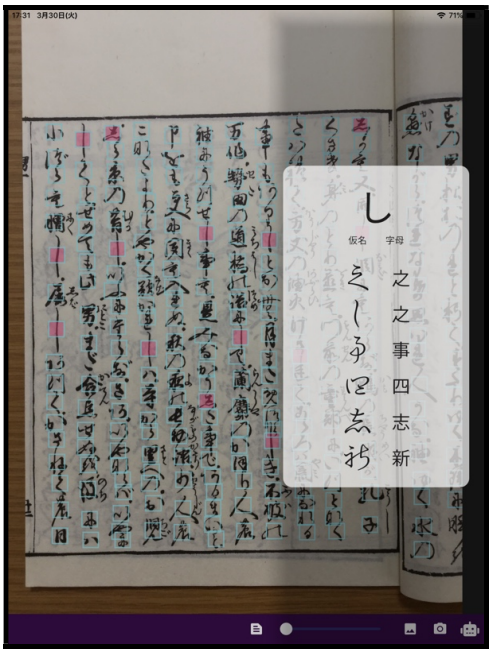

図 3 辞書機能。「し」のバリエーションを表示する。

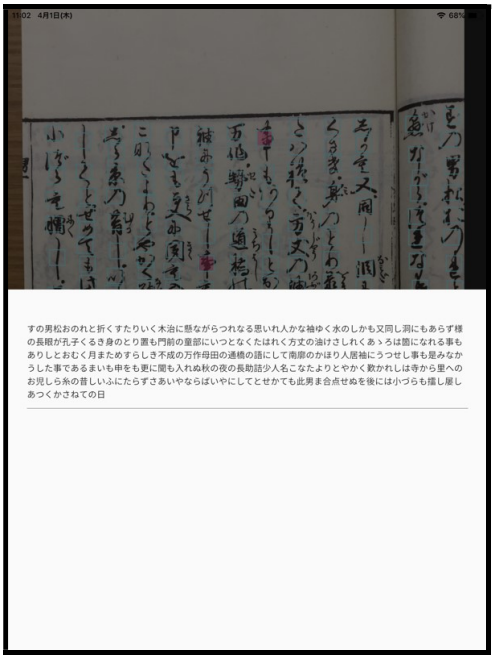

図 4 テキスト出力機能。

二つ目の新機能はテキスト出力機能である。しかし、くずし字資料は複雑なレイアウトが多いため、テキスト出力は簡単に解決できる問題ではない。研究代表者が試した手法は Adaptive Rule-Based モデルと Deep Autoregressive Sequence モデルで、論文を執筆中である。

Adaptive Rule-Based は文字の大きさと距離を計算し、テキストシークエンスを推定するものであり、機械学習を利用していない。一方、Deep Autoregressive Sequence モデル（Deep-AR）はディープラーニングモデルで、Autoregressive モデルで文字のシークエンスを推定する手法である。結果的に、シークエンス認識の精度は Deep-AR のほうが Adaptive Rule-Based より若干上だが、計算するスピードとGPUリソース、そしてモデルの推定が間違った場合、どの文字をミスするのかはほぼランダムで、なぜミスしたのかを説明できないため、アプリに採用したのは Adaptive Rule-Based モデルにした。

最後に当初の研究計画にはなかったのだが、来館者の意見と感想を収集したところ、もう一つ必要な機能は認識結果を簡単に修正できるものである。AI、特にくずし字認識は100%正しく認識することができないので、間違いがあれば、どう修正するかは大事な課題である。人間からのフィードバックは機械にとっても重要な情報であり、AI は高性能であっても人間の修正が必要というメッセージも AI 利用者に伝えたいので、認識結果の修正と人間のフィードバッ

クを収集できる機能を追加する予定である。ACT-X の期間終了まで一般公開を目指している。

3. 今後の展開

研究代表者がこのくずし字認識スマホアプリの研究から感じたのはAIモデル、アルゴリズムの研究が大事だが、ユーザーがその AI をどう利用するのも重要である。ユーザーがアプリを利用しているところを見ると、結果を表示するだけでは使いやすい、読みやすい、勉強になると言えない。今後の展開では人間と AI がくずし字を学習するために、インターフェースにどう工夫するのが課題で、ヒューマンコンピュータインタラクションの研究も進めたい。

4. 自己評価

この「みを」くずし字認識モバイルアプリの公開で、誰でも手軽にくずし字資料の内容を確認できるようになり、日本文化への興味を高めることができる。研究代表者がアプリのデモ動画を SNS に投稿し、動画の再生は150万回以上、4万回に共有され、コメントも3000件以上もらい、大反響を呼んだ。研究代表者は個人的な理由で9ヶ月しか研究期間がなかったのだが、できるだけ早くアプリを公開しようと考え、スケジュール管理は逆にうまくできなかったため、公開するアプリは当初の研究計画から多少違うものになった。ところが、慶應義塾大学の展覧会との共同プロジェクトで、ユーザーの反応、意見、感想を直接に聞くことができ、大変勉強になった。

研究代表者がこの研究を通して、もう一つの感じたことは AI の利用者が AI に対する理解が大事である。くずし字が機械で読めるようになったら、研究者がいらないという意見や、くずし字を勉強する必要ないというのもいくつかあった。AI には100%の精度を出すのがありえないのであり、研究者の仕事もくずし字を読むだけではないというメッセージを一般の人々にもっと伝えたいと考えている。

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：0件

（2）特許出願

無し。

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1. くずし字認識アプリ「みを」を開発した。2021年4月から6月まで、慶應義塾ミュージアム・コモンズの特別展覧会「交景：クロス・スケープ」、特別プログラム「文字形－AI が開くくずし字の風景」でアプリの体験版を展示した。
2. 招待公演リスト
    1. 国立歴史民族博物館、「みんなで翻刻サミット」、「AI とみんなで翻刻」（2021年2

月15日）

2. JST Sakura Science Club, “Japanese Culture and AI”（2021年3月）
3. Nissan Institute of Japanese Studies, Oxford School of Global and Area Studies, University of Oxford, UK. “Deciphering pre-modern Japanese manuscripts: kuzushiji recognition systems and AI”（2021年4月30日）
4. 慶應義塾ミュージアム・コモンズの特別展覧会の慶應義塾中等部ワークショップ（2021年5月11日）
5. Lenovo Japan 社内公演「くずし字とAI」（2021年5月14日）
6. Google Asia Pacific (APAC) Spotlight on Women in Research APAC “Opening the door to a thousand years of Japanese culture with AI”（2021年5月19日）
7. The Alan Turing Institute, UK（TBA、2021年7月30日）