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　４ 年　 ６ 月 ２３ 日現在

機関番号：８２６５７
研究種目：若手研究
研究期間：2019～2021
課題番号：１９Ｋ１３０８５

研究課題名（和文）ディープラーニングによるEnd-to-End日本古典籍くずし字認識の研究

研究課題名（英文）End-to-end Pre-modern Japanese Kuzushiji Recognition with Deep Learning

研究代表者
Ｃｌａｎｕｗａｔ　Ｔａｒｉｎ（Clanuwat, Tarin）

大学共同利用機関法人情報・システム研究機構（機構本部施設等）・データサイエンス共同利用基盤施設・特任助教

研究者番号：１０８３５１７７

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：本研究は機械学習によるくずし字認識の研究であり、研究代表者はくずし字認識モデルKuroNetを開発し、認識精度が９０％に達した。KuroNetの特徴は典型的な文字認識の処理順を逆転させ、難易度が高いレイアウト解析を後に回して文字認識を先に行うという方法である。そして、2019年に国際的AIコンペKaggle Kuzushiji Recognitionを開催した。このコンペの問題設定やデータ準備の作業に対してKuroNetは不可欠だった。さらに、IIIF KuroNetくずし字認識サービスが公開された。最後にKuroNetを利用しAIくずし字認識スマホアプリ「みを」を開発し無料アプリで公開した。

研究成果の学術的意義や社会的意義
日本ではくずし字が千年以上も前から使われており、数億点規模の資料が保管されているが、現代日本人はそれらの資料を利用できない問題がある。この問題は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。本研究で開発した認識モデルやさまざまなサービスはこの問題を解決するためである。そして、誰でも簡単にくずし字認識モデルを利用することを実現した。国内外の博物館や図書館が公開した画像はIIIF KuroNetくずし字認識サービスで、ユーザの手持ちの資料は「みを」アプリでくずし字認識を行える。最後にくずし字データセットやKaggleコンペを通して、海外の研究者にも日本文化への関心が高めたともいえよう。

研究成果の概要（英文）：This research is a study of machine-learning-based Kuzushiji character recognition. We developed a Kuzushiji character recognition model called KuroNet with 90% accuracy on the test data. The KuroNet features a method that reverses the typical processing order for character recognition, by perform character recognition first then perform layout analysis. This method works well with Kuzushiji document which the layout can be very hard. In 2019, we hosted the international AI competition Kaggle Kuzushiji Recognition. KuroNet was indispensable for the problem setup and data preparation work for this competition. Furthermore, We also released the IIIF KuroNet Kuzushiji Recognition service. Finally, KuroNet was used to develop the AI Kuzushiji Recognition smartphone application "miwo" and released as a free app for both Android and iOS.

研究分野：日本文学

キーワード：くずし字　機械学習　文字認識　古典籍　日本文学

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

１．研究開始当初の背景

日本は世界的に見ても大量の歴史的資料がよく保存されている国である。保管されている資料の点数は明確ではないが、『国書総目録』によると古代から 1867 年の時点までても、約 170 万点の古典籍が登録されており、現在まで未登録や新発見を含め、数百万点を超える膨大の資料が残されていると考えられる。さらに、手紙、個人日記、記録など、書籍になっていないものを含めると、日本各地に眠っている歴史的資料の規模は数億点ともいわれる。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ことだろう。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。

日本ではくずし字が 1000 年以上も前から使われており、江戸期以前の資料のほとんどはくずし字で書かれてきた。活版印刷への移行が進んだ明治時代にもくずし字は残ったが、それが消えるきっかけとなったのが明治 33 年（1900 年）の小学校令である。それまでは、一つの音に対応するひらがなは複数存在したが、小学校令を契機にそうした変体仮名は整理され、現代の一音一字に統一されることになり、さらに、一般の教育課程だと、漢字の楷書しか教えられなくなり、その後、学校教育でもくずし字の姿が消えてしまったのである。現代の高校教育では古文の授業があるとしても、現代日本人はたった 150 年前に書かれた資料を解読することができない。今やくずし字がきちんと読める人は数千人程度（＝人口の 0.01%程度）ともいわれ、大量に残された歴史的資料に比べて読める人があまりに少ないというアンバランスな状況が、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。

このような事情ゆえ、日本文化と資料の情報を守り、専門ではない人でもくずし字資料を利用ができ、さらに研究者のより有効的な研究活動を可能にするため、機械学習を用いたくずし字認識の研究が重要な課題となる。そのため、本研究は人工知能の一種であるディープラーニングによる日本古典籍くずし字認識の研究、つまり、くずし字で書かれている資料を機械で文字認識を行い、現代日本語のテキストに変換する技術を研究することである。

２．研究の目的

歴史的資料を保管するのは資料を撮影し、デジタル画像化されている方法が一般的である。しかし、大量のデジタル画像があっても、その資料にはどのような内容があるのか、画像内の情報を得ることができない。画像内のテキストからテキストを出力するには文字認識の技術が必要となるのは間違いないが、機械によるくずし字認識の研究は決して新しい研究課題ではないのに、従来の文字認識アルゴリズムは、認識精度が 50%以下にとどまるなど、実用性の高いレベルには至らなかったのはなぜなのだろうか。それは一般的な文字認識の手法はくずし字に適していないのである。この問題を解決するため、研究代表者は新しいの手法を提案し、この手法でのくずし字認識システムを開発し、３年間以内、くずし字認識サービスを一般公開するのが目的である。本研究の目的は以下のとおりである。

1. 機械によるくずし字認識を研究し、モデルの認識精度を 80%～90%に向上させる。
2. 国内外のくずし字認識の研究を広めて、機械学習によるくずし字認識コンテストを開催する。
3. 本研究の成果でくずし字認識サービスを一般公開する。

３．研究の方法

研究代表者の研究方法は一般的な文字認識と異なる物体検出の手法を提案する主な理由はくずし字認識のレイアウト解析が難しいからである。典型的な文字認識の手法は画像データをま

ずレイアウト解析し、文字分割を行い、最後に文字認識をする。この手法では、画像を入力した後の最初の処理がレイアウト解析となり、最初のレイアウト解析で失敗してしまうと、その後の文字認識も失敗してしまい、全体として OCR の精度が向上しない。つまり最も難しいレイアウト解析を最初に行うという処理順になっていることが、くずし字 OCR の精度が向上しない最も大きな原因ではないかと考える。

これに対して研究代表者は、難易度が高いレイアウト解析を後に回して文字認識を先に行うという、処理順を逆転させたくずし字認識モデルの研究を進めた。そして文字認識の部分には、画像中のどこに何があるかを画像中から直接探し出す物体検出（object detection）技術を適用することで、レイアウト解析をしなくても文字認識ができるようにした。この手法で最初のバージョンのくずし字認識「KuroNet」を開発した。

研究開始当初の KuroNet は U-Net というアルゴリズムを活用した。厳密にいうと、U-Net は物体検出の手法ではなく、Semantic Segmentation であるが、方向性は同じである。しかし、当時 2018 年では KuroNet のアルゴリズムは単純でいくつかの問題を抱えていた。まずモデルに入力する古典籍 1 ページの画像サイズは 512x512 pixels にとどまり、認識できる文字種も最大で 409 字種に限られていた。画像サイズに 512 pixels を選択した理由は、1 ページの画像がこのサイズであれば、人間の目でまだ文字が読めたからである。しかしこの選択はよくなかった。また、GPU メモリーの制限の問題もあり、多くの文字を認識することができなかった。

2019 年 9 月に発表した KuroNet は、さまざまな問題を解決するためにモデルを大幅に改善したバージョンである。まず U-Net より安定した Residual U-Net（特に FusionNet という Residual U-Net のバリエーション）を採用した。次に GPU メモリーの問題を解決するために、Teacher Forcing というテクニックを取り入れ、画素ごとに文字があるかを判断し、文字のある確率が高い画素だけ文字認識を行う手法とした。さらに精度を向上させるために、Mixup Regularization も採用した。Mixup Regularization は学習画像の Opacity を 70%、無関係に選んだランダム画像の Opacity を 30%にして、学習画像の上にランダム画像をノイズとして重ねて学習する方法である。このような Data Augmentation を取り入れることで、KuroNet の精度は 10%ほど高くなっただけでなく、GPU メモリーを節約するテクニックにより、入力画像のサイズも 976x976 pixels にまで拡大することができた。そして、2020 年 2 月に発表した改善版の KuroNet では、学習の際に画像をランダムに切り取る Random Cropping を採用した結果、認識精度はさらに向上しただけでなく、さまざまな文字サイズへの対応も改善し、1 ページの認識時間は 2 秒程度で、テストデータの平均精度は 85%～90%に達した。この KuroNet の開発が成功したことで、本研究のさまざまな成果に繋がった。

４．研究成果

本研究のタイムラインは以下のとおりである。

- 2018 年 8 月：くずし字認識モデルの研究開始。当時３つの文字しか認識できなかった。
- 2018 年 12 月：KuroNet という名前をつけ、認識可能文字数は 409 文字に向上した。
- 2019 年 2 月：国際的くずし字認識コンペ Kaggle Kuzushiji Recognition を開催決定。
- 2019 年 7 月～9 月：Kaggle コンペ開催した。
- 2019 年 11 月:「日本文化と AI」シンポジウムを開催、IIIF Curation Viewer 上の KuroNet くずし字認識サービスを公開した。
- 2020 年：KuroNet、Kaggle 優勝者モデルを API 化し、テキスト出力 API を開発した。
- 2021 年 8 月：「みを」くずし字認識スマホアプリを AppStore と GooglePlay で無料アプ

リとして公開した。

研究成果１：KuroNet

現在、KuroNet の精度は 85%～90%で、Kaggle の優勝者モデルの精度（95%）より低いが、KuroNet の最も大きな成果は認識精度ではない。まず、KuroNet が精度 80%の超えた初めてのくずし字認識モデルであり、一般的文字認識と異なるが、この手法だとくずし字認識がうまくいくかもしれないということを証明してくれた。次に、Kaggle コンペが KuroNet をベースにして設計したものともいえよう。それは KuroNet がベースラインモデルとして使えたことが大きかった。コンペの現実な問題設定や、データ分割などのデータ準備の作業に対して、ベースラインモデルは不可欠だった。さらに、参加者のモデルの評価方法も KuroNet と同じ F1Score を用いた。

研究成果２：Kaggle Kuzushiji Recognition コンペ

#	△pub	Team Name	Notebook	Team Members	Score
1	–	tascj			0.950
2	–	Konstantin Lopuhin			0.950
3	–	Kenji			0.944
4	▲1	YoudaoOCR			0.942
5	▼1	See--			0.940
6	–	abc			0.939
7	–	K_mat			0.934
8	–	t-hanya			0.920
9	–	Ollie, Nanashi, and Tom			0.910
10	–	Zenkei_R&D			0.903

研究代表者がプロジェクトリーダーとして、世界最大の機械学習コンペプラットフォームである Kaggle 上で、2019 年 7 月から 10 月にかけて「くずし字認識」コンペを開催した。最終的には、293 Teams、338 Competitors、2,652 Entries という規模となった。また、ディスカッションのトピックは 63、参加者が実験し公開されたコード（パブリックなノートブック）は 40 である。

優勝者のスコアは 0.95079 であった。2 位のスコアは 0.950688 であり、1 位と 2 位の差は小数点以下 4 桁の差となった。ベースラインモデルである KuroNet のスコアが 0.902 相当であることを考えると、コンペによって自分たちでは試せない優れた解法を集めることができた。

参加者の国籍は厳密にはわからないが、トップ 10 となったチームを分析すると、日本を拠点とするチームが少なくとも 4 つ、その他、中国、ロシア、ドイツなど、世界規模で上位入賞者が生まれた。そしてくずし字が読めることは、コンペの上位入賞には必須でないことも判明した。

研究成果３：IIIF KuroNet くずし字認識サービス

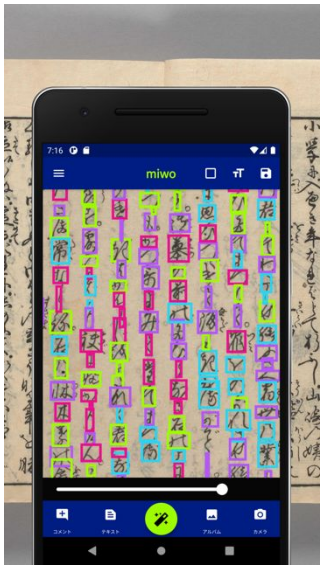
2019 年 11 月に公開した KuroNet くずし字認識サービスは、国内外の多くの美術館、図書館が

利用する画像公開方式 IIIF に基づき、世界中で公開されている資料にプログラミングなしでくずし字認識を適用できるサービスである。なぜ KuroNet サービスを IIIF 上で公開したのかは以下の理由がある。

まず、IIIF で公開されているくずし字資料は全世界に数多く存在しており、これらの資料を認識できるサービスを立ち上げることが最優先の課題だった。例えば、国文学研究資料館が新日本古典籍総合データベースで公開している資料を翻刻し、検索できるようにすれば、今後の人文学研究に大きな貢献になることが期待できる。2020 年、新型コロナウイルス感染症（COVID-19）の影響は多くの研究者が受けており、現場で資料調査を行うことが難しくなっている。そのため、オンラインで資料調査できることの重要性はますます大きくなってきている。

次に、KuroNet も Kaggle のくずし字認識モデルも、撮影環境や画質など、画像の状態に認識精度は大きく影響されるという問題がある。図書館や美術館から IIIF で一般に公開されている画像は、同じ角度、同じ大きさで、できるだけ画質が高くなるようにきれいに撮影されたのがほとんどである。こうした画像は、くずし字認識モデルにとってベストの状態となっている。このような理由から、KuroNet くずし字認識サービスは、当面の間は IIIF 画像のみを対象としたサービスとして運用する計画である。

研究成果４：AI くずし字認識スマホアプリ「みを」



AI くずし字認識スマホアプリ「みを」は研究代表者の JST、ACT-X「AI 活用で挑む学問の革新と創成」に採択されたプロジェクトではあるが、アプリのベースモデルは KuroNet と Kaggle 優勝者のモデルである。KuroNet くずし字認識サービスは IIIF 上のサービスで、博物館や図書館が IIIF を用いて多くの資料画像をくずし字認識を行うことができるようにした。とはいえ、このサービスは IIIF で公開されていない資料に対応していないため、手持ちの資料をすぐに調査できないといった問題があった。そこで、くずし字認識システムに入力する画像データをスマホのカメラで撮影し、すぐにくずし字認識を行えるアプリがあると良いのではないかと考えた。そこで誰でも活用できるサービスとして開発したのが AI くずし字認識アプリ「みを」である。2022 年 6 月現在で、アプリダウンロード回数が 5 万回以上、ユーザが認識した画像は約 60 万まで登った。大学のくずし字の授業でも「みを」アプリを教材として使われている。

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　6件／うち国際共著　2件／うちオープンアクセス　3件）

１．著者名 北本 朝展 ， カラーヌワット タリン ， ボーバー・イリザー ミケル	４．巻 35
２．論文標題 Kaggle くずし字認識 ―世界規模の人文系コンペ開催への挑戦―	５．発行年 2020年
３．雑誌名 人工知能学会誌	６．最初と最後の頁 366-376
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 有
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 ―

１．著者名 Alex Lamb ， Tarin Clanuwat ， Asanobu Kitamoto	４．巻 1
２．論文標題 KuroNet: Regularized Residual U-Nets for End-to-End Kuzushiji Character Recognition	５．発行年 2020年
３．雑誌名 SN Computer Science	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s42979-020-00186-z	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 カラーヌワット タリン ， 北本朝展	４．巻 -
２．論文標題 くずし字認識の進化とサービス化の展開	５．発行年 2020年
３．雑誌名 人文科学とコンピュータシンポジウム じんもんこん2020論文集	６．最初と最後の頁 3-10
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 有
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 ―

１．著者名 北本 朝展 ， カラーヌワット タリン ， 宮崎 智 ， 山本 和明	４．巻 102
２．論文標題 文字データの分析――機械学習によるくずし字認識の可能性とそのインパクト――	５．発行年 2019年
３．雑誌名 電子情報通信学会誌	６．最初と最後の頁 563-568
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 オープンアクセスとしている（また、その予定である）	国際共著 ―

1．著者名 北本 朝展 ， カラーヌワット タリン ， Alex LAMB ， Mikel BOBER-IRIZAR	4．巻 -
2．論文標題 くずし字認識のためのKaggle機械学習コンペティションの経過と成果	5．発行年 2019年
3．雑誌名 人文科学とコンピュータシンポジウム じんもんこん2019論文集	6．最初と最後の頁 223-230
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 有
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 －

1．著者名 Tarin Clanuwat, Alex Lamb, Asanobu Kitamoto	4．巻 2019
2．論文標題 KuroNet: Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning	5．発行年 2019年
3．雑誌名 The International Conference on Document Analysis and Recognition (ICDAR) Proceeding	6．最初と最後の頁 1-8
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 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 カラーヌワット・タリン，北本朝展	4．巻 2021
2．論文標題 資料調査のための AI くずし字認識スマホアプリ「みを」	5．発行年 2021年
3．雑誌名 人文科学とコンピュータシンポジウム じんもんこん2021論文集	6．最初と最後の頁 -
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 有
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9件（うち招待講演　6件／うち国際学会　3件）

1．発表者名 カラーヌワット・タリン
2．発表標題 くずし字認識の進化とサービス化の展開
3．学会等名 人文科学とコンピュータシンポジウム じんもんこん2020
4．発表年 2020年

１．発表者名
Tarin Clanuwat

２．発表標題
Kuzushiji and Premodern Japanese Studies: Learning Resources and Artificial Intelligence Initiatives

３．学会等名
Centre for Japanese Research, the University of British Columbia, Canada（招待講演）（国際学会）

４．発表年
2020年

１．発表者名
カラーヌワット・タリン

２．発表標題
AIとみんなで翻刻

３．学会等名
みんなで翻刻サミット（招待講演）

４．発表年
2020年

１．発表者名
Tarin Clanuwat

２．発表標題
Japanese Culture and AI

３．学会等名
JST Sakura Science Club, Japan Science and Technology Agency（招待講演）（国際学会）

４．発表年
2021年

１．発表者名
カラーヌワット・タリン

２．発表標題
世界中のアイデアを集めるくずし字コンペの開催

３．学会等名
日本文化とAIシンポジウム

４．発表年
2019年

１．発表者名
カラーヌワット・タリン

２．発表標題
くずし字ｘAIオンラインで世界に開く日本古典籍

３．学会等名
DMC 研究センターシンポジウム、第９回 大学教育のミライ：オープンエデュケーションのその先へ（招待講演）

４．発表年
2019年

１．発表者名
カラーヌワット・タリン

２．発表標題
Kuzushiji and AI : A Case Study of Multidisciplinary Research

３．学会等名
次世代日本研究者協働研究ワークショップ（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
カラーヌワット・タリン

２．発表標題
ＡＩによるくずし字認識、古典文学と情報学の世界的なコラボレーション

３．学会等名
総研大文化フォーラム2019（招待講演）

４．発表年
2019年

１．発表者名
カラーヌワット・タリン

２．発表標題
資料調査のための AI くずし字認識スマホアプリ「みを」

３．学会等名
人文科学とコンピュータシンポジウム じんもんこん2021

４．発表年
2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

KuroNetくずし字認識サービス（AI OCR） http://codh.rois.ac.jp/kuronet/ KuroNetくずし字認識サービス http://codh.rois.ac.jp/kuronet/ AIくずし字認識（一文字） http://codh.rois.ac.jp/char-shape/app/single-mobilenet/

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

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関