# データ駆動型人文学研究の発展とAIによるくずし字認識

オープンデータの大規模化とAI（人工知能）の高性能化により、人文学の分野においてもデータ駆動型研究が広がっています。データ駆動型研究とは、大規模データの収集と分析を通して新しい知識を得る方法論であり、AIによる「くずし字」認識もその一つです。本稿では、AIによるくずし字認識を通したAIの活用方法やOCRの原理、AI活用の可能性等についてご紹介いただきます。

**北本朝展**
情報・システム研究機構
データサイエンス共同利用基盤施設
人文学オープンデータ共同利用センター長
国立情報学研究所
コンテンツ科学研究系准教授

東京大学大学院工学系研究科電子工学専攻修了。博士（工学）。学術情報センター研究開発部助手、国立情報学研究所コンテンツ科学研究系准教授、JSTさきがけ研究者などを経て、平成29年より現職。

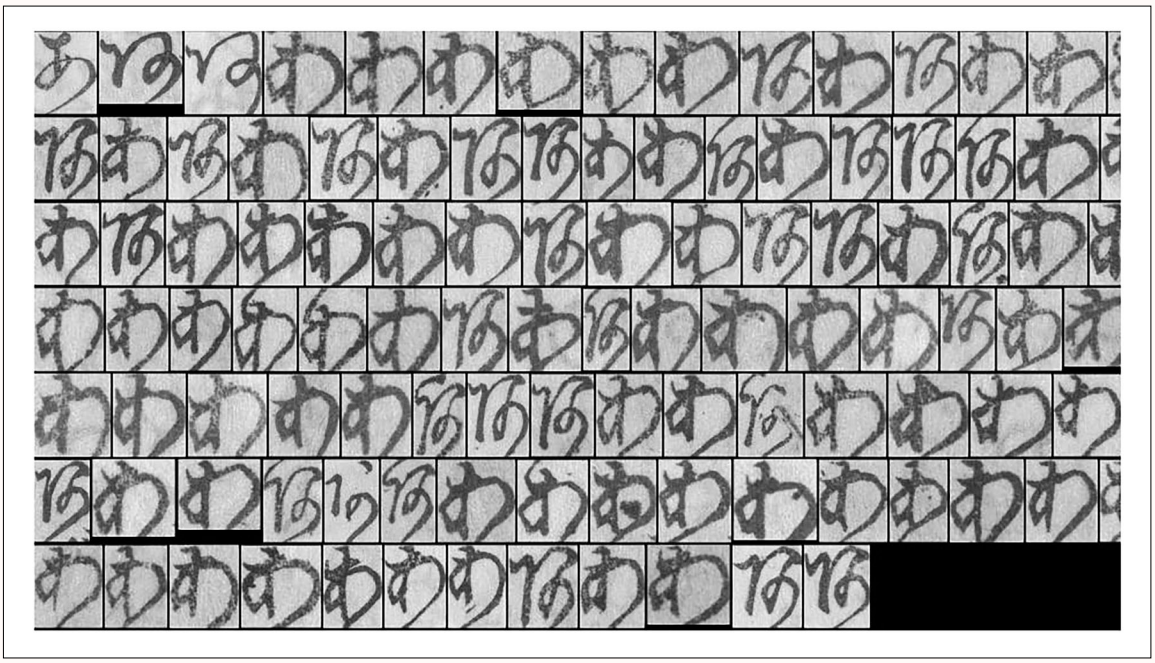
## はじめに

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す。その規模は1億点とも10億点ともいわれ、日本各地のミュージアムや宗教施設、旧家などに今も多くの歴史的資料が眠っています。各地の歴史を伝えるこうした資料を読み解くことができれば、これまで知られていなかった多くの事実が明らかになるでしょう。

ところが、そこに立ちはだかるのが「くずし字」の問題です。くずし字とは日本で1000年以上も使われてきた手書き文字のことで、江戸期以前の資料のほとんどがくずし字で書かれています。明治時代に入ってもくずし字は引き続き使われたものの、明治33年には小学校令により、それまで多数存在したひらがな（変体仮名）を現代のひらがな（一音一字）に統一することになり、それ以降は学校教育でくずし字を教えなくなりました。その結果として、ほとんどの現代日本人はくずし字で書かれた歴史的資料を読むことができなくなりました。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）ともいわれます。大量に残された歴史的資料に比べ、あまりに読める人が少ないというアンバランスな状況で、歴史的資料を人間が一つひとつ読み解いていくにも限界があります。どうすればよいのでしょうか。

そこに登場したのがAI（人工知能）です。AIの中でも特に機械学習と呼ばれる、機械に正解データを与えて学習させる方法が急速な発展を遂げ、様々な分野への活用が進みつつあります。そうした流れにのって、AIにくずし字を読ませてみてはどうでしょうか。数年前ならこれは大変に難しいことだったかもしれませんが、機械学習手法の一つであるディープラーニング（深層学習）がここ数年で急速に発展し、状況は激変しました。さらに国文学研究資料館とROIS-DS人文学オープンデータ共同利用センターが協力して公開するオープンデータ「くずし字データセット」[1)]（図－1）の登場により、学習に用いるデータセットも飛躍的に大規模化しました。機械学習手法の高性能化とオープンデータの大規模化という2つの流れが合流することで、AIがくずし字を読める時代がようや

図ー１　くずし字データセットの例



すべて平仮名の「あ」。変体仮名の存在により複数の字形が混在している。

くやってきた、というのが筆者の実感です。

では、AIがくずし字を読めると何が嬉しいのでしょうか。歴史的資料の中で、読まれているものはごく一部であり、大半はくずし字が読めず内容もわからないため、活用が進んでいません。もしAIがくずし字を読んで現代日本人にも内容がつかめるようになれば、興味を持つ人が増える可能性は大いにあります。さらに歴史的資料から地域の歴史に根差す文化的価値を発見できれば、それが長期的にはツーリズムなどの経済的価値につながっていくかもしれません。このようにAIくずし字認識は、地域の価値を見直すツールにもなりうるのです。

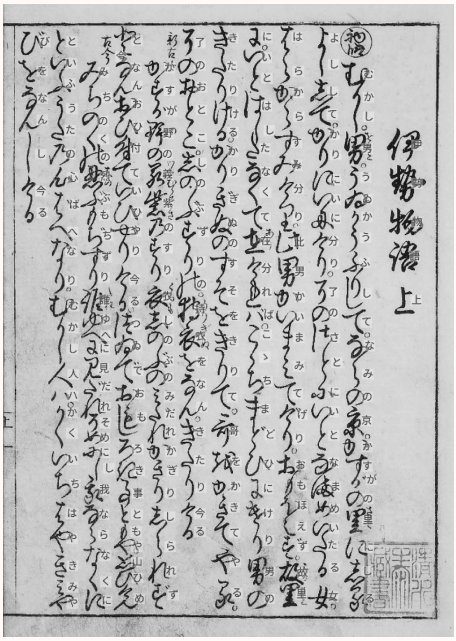
## AIによるくずし字認識

AIとは何でしょうか。現在のブームの中心にある機械学習は、前例をもとに判断を下すことに非常に長けた機械を作り出す技術です。前例は必ずしもまったく同一でなくてもよく、前例とは多少違うケースでもうまく補ってくれます。前例主義は「深く考えなくても」そこそこよい判断を下せる点が魅力的で、より多くの経験を積めば、よりよい判断が下せるようになることも期待できます。こうして判断と正解のペアを機械にたくさん教えこむことで、判断の正解率を上げていこうというのが機械学習の基本的なアイデアです。くずし字認識もこうした前例主義がうまく働く問題の一つです。文字はコミュニケーションの道具ですから、相手に伝わるように前例にしたがって同じ文字を書くことが望ましいからです。「この文字はどこかで見たことがある」と機械が気づくことができれば、あとは前例にしたがって正解できます。

筆者ら[2)]は、くずし字を対象としたOCR（Optical Character Recognition）の核心部分となる、KuroNetという文字認識システムを開発しています。画像を与えると、そこに含まれる文字をすべて抽出する作業を、１秒程度で完了できるシステムです。このシステムでは、ディープラーニングの一種である「U-Net」という手法を使っていますが、くずし字のように続けて書かれる文字（連綿体）の認識を苦にしないという点に特徴があります。人間にとってはくずし字をどこで切るかは難しい問題ですが、機械の場合は連綿体を苦にしないというのは意外な結果でした。

ここでOCRの原理についても簡単に説明しておきたいと思います。最近はRPA（Robotic

**図－２　KuroNet による古典籍のくずし字認識の結果**




『国書総目録』には古代から江戸期（1867 年まで）の古典籍 170 万点が登録されている。全国の古典籍の数は 300 万点ほど、古文書を含めると 1 億点以上と推定されている。これらはくずし字が読めないと解読できない。

**どんな研究？**

- ディープラーニングによるくずし字 OCR システムを研究。
- 一般的な文字認識アルゴリズムと違い、中間ステップも自動的に学習する End-to-End アルゴリズムを開発。

**何がわかる？**

- くずし字を自動的に現代日本語文字に変換できる。
- 古典籍全文検索機能など、さまざまな機能を開発し、人文学研究を効率化する。

**一般的な文字認識の手法**

画像 → レイアウト解析 → 文字分割 → 文字認識

**End-to-end 文字認識の手法**

画像 → 文字認識

◀くずし字の右横に表示されている文字が認識結果（認識誤りもそのまま表示）。「日本古典籍データセット」（国文学研究資料館所蔵）を利用。

Process Automation）などのブームにより、機械による文字認識（OCR）が注目を集めています。印刷文書や本などをスキャンしてデジタル化した画像を機械に与えると、そこに書いてある文字を自動的に読み取って文章にしてくれるというのが、OCRの基本的な機能です。さらにRPA向けのOCRでは、ビジネスにおいて価値が高い手書き文字の認識や構造化データ（表や帳票）の抽出など、より高度な機能を備えることもあります。こうしたOCRを実現するのに必要となるのは、以下の３つの機能です。第一に、文字がどこにどのように書かれているかを分析するレイアウト解析です。第二に、そこに実際にどんな文字が書かれているかを分析する文字認識です。そして第三に、辞書を照合するなどの方法で認識した文字の誤りを修正する言語処理です。この３つの機能をこの順番で連続実行するのがOCRの典型的な手法です。

しかし、くずし字の場合はレイアウト解析が非常に難しいという問題があります。江戸時代の本は木版印刷のため、どんな複雑なレイアウトでも自由自在、古文書の場合も、原稿用紙のように行がきちんと揃っているわけではありません。このような難題を解決するため、KuroNetでは文字を直接認識するという戦略を用いました。つまりレイアウト解析を省略し、どこに文字があるかを直接取り出すのです。このアイデアが功を奏し、特に江戸時代の木版印刷の本などの場合、条件がよければ９割以上の文字がAIで読めるようになりました（図－２）。この結果を使えば、くずし字が読めない現代日本人でもなんとか内容を推測できるため、歴史的資料の活用への道が開けてくるかもしれません。

## くずし字認識から歴史的資料の活用へ

では、AIがくずし字を読むことができれば、くずし字を読めない現代日本人でも歴史的資料が自由自在に活用できるようになるのでしょうか。実は残念ながらそうではありません。越えなければならない壁は、その後にいくつも控えています。

まず「翻刻」と「翻訳」という言葉を区別する必要があります。くずし字認識はあくまで「翻刻」、つまり、くずし字を現代の文字に変換することを目標とします。その結果として得られるのはいわゆる古文・漢文です。皆さんは古文・漢文がすらすら読めるでしょうか。くずし字を読める人よりは多いかもしれませんが、現代日本人にとって依然として高いハードルである点

は変わりません。

おそらく多くの人は「翻訳」、つまり古文を現代文に「現代語訳」するところまでをAIにやってほしいと考えるでしょう。このような機械翻訳もAIの中では研究が盛んな分野で、英語と日本語のようにまったく異なる言語間でも精度が向上しつつあります。そして、古文と現代文は同じ日本語ですから、異なる言語間の翻訳よりも高い精度が達成できる可能性もあります。しかし「翻刻」と「翻訳」は、同じAIとはいっても技術体系は大きく異なり、研究者も別々です。また「翻刻」用のくずし字データセットが大規模なオープンデータとして公開されているのに比べ、「翻訳」用の古文・現代文対訳データセットには大規模なオープンデータがありません。現代語訳の多くは書籍の形式で出版されているため、著作権の保護対象となるからです。オープンデータとして使える古文・現代文対訳データセットをどうやって作るか。それが「翻訳」の研究を進める上での大きな課題です。

さて、もし翻刻と翻訳を連続して実行できるAIが実現すれば、現代日本人は歴史的資料を自由自在に活用できるのでしょうか。そこに立ちはだかるのが「読解」や「史料批判」と呼ばれる問題です。歴史的資料には、単純な（書き）間違いだけではなく、視点の偏りや意図的な誤りなどが含まれていることがあるため、文字になっているからといってその内容を無条件に信用してはなりません。とはいえ、これは歴史的資料に特有の問題ではなく、現代のフェイクニュースなどとも共通する問題です。ニュースやソーシャルメディアの文章の意図を読み解くリテラシーと同様に、歴史的資料を正しく読み解くリテラシーも必要となるのです。ところが現代ニュースと違い、歴史的資料の背景となる社会に関する知識を現代日本人が身に付けるのは簡単ではありません。AIがいくら発展しても、すべてを自動化できるわけではなく、AIの出力結果を内容の面からチェックできる専門家は今後も不可欠なのです。

## おわりに

日本の各地には多くの歴史的資料が保存されていますが、それらが価値を認められることなく廃棄される危険性が日々高まっています。そうした資料を未来に受け継ぐためには、それらを地域の価値ある資産として復活させるアイデアが必要であり、そこにはAIを活用する余地があるように思います。もしAIくずし字認識が初心者レベルを超えれば、初心者はAIと上級者の両方から教えてもらえるようになります。それによって歴史的資料の翻刻が進めば、その結果をAIにフィードバックすることでAIの性能をさらに向上させることができます。そして歴史的資料の内容が見えてくると、より多くの人々が興味を持つようになり、そこに価値を認める人が増えてきます。このように、郷土の魅力を歴史的資料から発掘する「ふるさと翻刻」活動にAIくずし字認識技術を組み込むことで、歴史的資料のリバイバル（復興・再生）を実現することが筆者らの長期的な目標です。

このような目標を実現するには、時代や地域を越えた汎用AIくずし字認識に向けて、より大規模なデータセットを構築することが重要な課題です。そしてそうしたデータセットの構築はそれ単体で行う事業というよりも、各種の活動を巻き込む中で成長するものだと考えています。筆者らもそのような活動を支える情報基盤として、くずし字認識システムのオープンソース公開やAPI提供を今後は進めていく計画です。またこうした研究の進展については、令和元年11月11日に一橋講堂で開催する「日本文化とAIシンポジウム2019」（参加無料・要事前申込、詳しくはHP[3]）でもご紹介しますので、こちらにもぜひお越しください。

1）http://codh.rois.ac.jp/char-shape/
2）カラーヌワット・タリン（ROIS-DS人文学オープンデータ共同利用センター・国立情報学研究所）、Alex Lamb（モントリオール大学MILA）との共同研究による。
3）http://codh.rois.ac.jp/symposium/japanese-culture-ai-2019/