

2019. 9. 9

畑 啓之

江戸時代のくずし字は英語よりもはるかに難しい 人工知能による解読期待

江戸時代に書かれたくずし字は、文字をくずすことによりその筆記速度が3倍となったそうである。従って、一般の手紙文の多くはこのくずし字が使用されている。文字が読めない人が多かったというが、なるほど、この文字ならばそのことも理解できるというものである。下に示したWebの引用には「今はもう使われなくなった『くずし字』と呼ばれる文字で書かれた紙の書物や文書が数多く残されていますが、その文献を判読できる日本の人文学教授は10%にも満たないというのが現状です」と驚くべき内容が記されている。専門家でも読めない文字なのか？ この文字並びをすらすらと解読できた先人にひたすらの敬意！

今私が一番読みたい文章は江戸時代の以下のものである。しばらく眺めてはいるが、これまたなかなかのものである。近いうちに「翻刻（活字体に直すこと）」してやるぞと意気込みだけは高い。翻刻の後には意味理解が待っている。



**SCIENCE**

## 史料のデジタル化で研究革新

新旧史料の差分を検出

寛政武鑑  
旧 1789年 新 1791年

画像認識

伊勢物語 (1756年)

AI技術でくずし字を活字に起こしてデジタルデータとして使いやすくする

くずし字の史料を翻刻

古い史料(1789年)と新しい史料(1791年)を画像認識技術で比較、変更の前後を色分けして表示する(上)

(写真は国文学研究資料館所蔵・人文学オープンデータ共同利用センター提供)

# AIが変える人文学研究

歴史学や文学など人文学の分野で研究手法の変革が起きている。文字や絵など様々な情報をデジタル化し、ビッグデータや人工知能(AI)などの最新技術を活用する動きが広がってきた。統計的な解析手法で緻密な分析ができるようになり、経験豊富な専門家に迫ろうとしている。定説を覆したり思わぬ発見が生まれやすくなり期待が寄せられている。

文学や歴史学といった人文学にデジタル技術を使う試みは「デジタル・ヒューマニティーズ(人文学)」という新しい分野として世界で活発になっている。日本でも情報・システム研究機構が17年に「人文学オープンデータ共同利用センター」を新設した。同センターの研究グループはさっそく、画像認識技術を駆使して江戸時代の木版印刷史料を、時代を追って解析する方法を開発した。

対象は、当時の定期刊行物の「武鑑」だ。大名や幕府役人の人事や経済活動に関するデータが記載されている。このデータの改訂箇所を時系列で調べれば、当時の武家社会の社会・経済活動を詳細にたどれるのではないかと。改訂前と改訂後のページを専用の機器で読み込み、削除や追加などの変更部分を自動的に表示できるようにした。

北本朝展・同センター長らはこうした史料の「差分研究」を提案している。「昔前ま

では技術的に難しかった史料のデジタル化やデータ解析が可能になった。現在広く行われているビッグデータ解析と同じ手法を使える(北本センター長と説明する)。

同グループは、日本の歴史文献の「くずし字」を、自動的に活字に置き換える「翻刻」技術も開発した。現在のAIの主流技術である機械学習を使い、江戸時代の古籍の場合で90%近い正確さで翻刻できるようになった。

活字で表記されれば、くずし字が苦手の研究者も多くの文献にあたれる。翻刻した史料はテキストデータとして処理できるため、大量の文献の中からキーワードやその頻度を調べるなど、文学作品の分析に使われ始めている手法も可能になる。

独自のデータベースを作り歴史研究を進めようとするグループは他にも現れている。「データ駆動型」の手法は自然科学で定着してきたが、人文学分野ではまだ試行錯誤の段階だ。検証がこれから活発になるだろう。

(編集委員 吉川和輝)

「古文書 AI 解読」をキーワードに Google 検索した、上位 3 記事の抜粋

古文書・浮世絵の崩し字を AI が解読 産経新聞 2019.5.25

古文書や浮世絵などに書かれた「崩し字」を、人工知能（AI）が画像から解読するシステムを立命館大文学部の赤間亮教授らのチームが、凸版印刷（東京）と共同開発したと 13 日、発表した。

歴史を『読み解く』:AI で日本の古文書の膨大な文章をより多くの人々へ

BY ISHA SALIAN · JUNE 10, 2019

<https://blogs.nvidia.co.jp/2019/06/10/japanese-texts-ai/>

今はもう使われなくなった「くずし字」と呼ばれる文字で書かれた紙の書物や文書が数多く残されていますが、その文献を判読できる日本の人文学教授は 10% にも満たないというのが現状です。

くずし字には何千種類もの文字があり、データセット内でほとんど出現しない字も多いため、ディープラーニング モデルによる認識は困難です。それでもなお、同チームの文書認識モデル KuroNet は以前のモデルをしのぐ平均 85% の精度を誇ります。

この最新バージョンのニューラルネットワークは、2,000 種類以上の文字を認識可能です。300 種類未満の文字からなる比較的簡単な文書なら、精度は 95% 程度にまで跳ね上がると、カラーヌワット氏は説明します。「データセットの中でもっとも厄介な文書の 1 つは辞書です。珍しい言葉や一般的でない言葉が多く収録されていますから。」

AI 活用した古文書解読プロジェクト「みんなで翻刻」がリニューアル 2019 年 7 月 24 日

<https://ict-enews.net/2019/07/24honkoku/>

国立歴史民俗博物館・京都大学古地震研究会・東京大学地震研究所のメンバーを中心に開発を進める古文書史料の市民参加型翻刻プラットフォーム「みんなで翻刻（ほんこく）」が、22 日にリニューアル公開した。

「翻刻」とは、くずし字で書かれている古文書を、活字化して現代人が読めるようにする作業のこと。

2017 年 1 月に公開された「みんなで翻刻」では、約 5000 人の市民が参加。600 万文字以上の「くずし字」で書かれた史料が解読された。

最新の AI を使った「くずし字」の自動認識機能も搭載。凸版印刷、および人文学オープンデータ共同利用センター (ROIS-CODH) が開発した「くずし字認識システム」が利用でき、「くずし字」に慣れない初心者でも AI の支援で翻刻作業に参加できる。

翻刻されたテキストは、オープンデータとして公開、歴史研究や科学研究に役立てられる。