古文書解読に向けたくずし字AI-OCR技術開発

企業リポート

岡　敏生*，福井 尚子**

A Kuzushiji AI-OCR Technology Towards Deciphering Japanese Ancient Documents

Key Words：Classical Documents, AI-OCR, Cursive Characters

1．はじめに

一説によると日本国内には古文書（こもんじょ）が数十億点残存すると言われている。それらはわれわれに数百年前の社会や文化、当時の災害などについて貴重な情報を与えてくれる。その多くは各地に散在しているが、それらを有効活用できれば大きな観光資源になりうる。

しかし古文書の多くは現代の日本人には解読が困難なくずし字で書かれており、解読できる人は0.1%もいない（図1はくずし字で書かれた古文書）。そのため古文書の解読は遅々として進まず、本来の価値と比べて十分に活用されているとは言いがたい状況にある。

また古文書の内容が分からず価値を判断できないため、破棄、劣化、紛失に至るという事象も発生しているとの指摘もある。そうなるとそれら文化的遺産は不可逆的な形で失われてしまい、二度と取り戻すことができない。

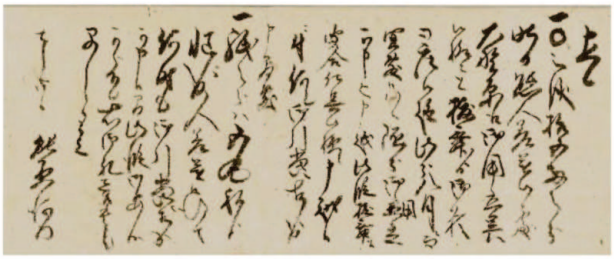

図1 くずし字で書かれた古文書の例

くずし字AI-OCR (Optical Character Recognition) の意義は、多くの日本人に読めない古文書の解読を支援し古文書の本来の価値を取り戻すことにある。

TOPPANにおけるくずし字研究の経緯

TOPPANにおけるくずし字AI-OCR技術開発は2015年に始まった[6]。当初は公立はこだて未来大学にて研究開発された「文書画像検索システム」と凸版印刷（株）（現：TOPPANグループ）で研究中だったディープラーニングによる識別器を併用する形で開発された。本取り組みは当社にとって、開発中のディープラーニング技術を実課題に適用する初めての試みであったが、くずし字とディープラーニングという文脈で見ても最初（期）の取り組みと思われる。その後、ディープラーニング技術の発展にしたがって、継続的な改良をつづけてきている。

2．関連する取り組み

本節では、デジタル人文学領域においてくずし字に関連する代表的な取り組みを紹介する。

まず情報データベースに関する取り組みとして、国文学研究資料館が中心となり日本語の歴史的典籍約30万点をデジタル化しWeb公開する事業（歴史的典籍NW事業[3]）が挙げられる。

くずし字学習という観点では、大阪大学文学研究科を中心に開発されたくずし字学習支援アプリ

* Toshio OKA

東京大学大学院　新領域創成科学研究科基盤情報学修了（2006年）
現在、TOPPANデジタル（株）技術戦略センター情報技術研究室　主任研究員
博士（科学）
E-mail：toshio.oka@toppan.co.jp

** Naoko FUKUI

2015年に凸版印刷（株）（現TOPPAN（株））に入社、くずし字OCR研究開発の立ち上げに携わる。現在はくずし字関連事業のサービス開発やGUIデザインを担当。
E-mail：naoko.fukui@toppan.co.jp

図2　くずし字認識の処理フロー

「KuLA」[2]や立命館大学での取り組み[1]がある。

また国立歴史民俗博物館、東京大学地震研究所、京都大学古地震研究会のメンバーが中心に立ち上げた「みんなで翻刻」プロジェクト[10]では、多数の人々が協同的に史料の翻刻（古い時代の文字をテキスト化すること）を行っている。

くずし字の解読を支援するツールとしては本稿で取り上げる古文書カメラ®[5]以外にも、人文学オープンデータ共同利用センターで開発された「みを」[9]がある。また国立国会図書館の次世代デジタルライブラリー[4]では、古典籍資料についてAI-OCR技術により全文テキストが生成されており検索が可能になっている。これら3つのツールは主に近世の文書を対象にしているが、他にも奈良文化財研究所・東京大学史料編纂所が公開している「木簡・くずし字解読システム」[11]がある。

3．くずし字AI-OCR

3.1　技術課題

まずくずし字を認識する上で特徴的な課題として、連綿体（いわゆる続け字）で記載されている文字を認識しなくてはならないことが挙げられる。

加えて古典的な史料の多様性も大きな課題となる。具体的には、媒体として古典籍（歴史的価値の高い書籍、特に江戸期以前）と古文書では字形のバラつきが大きく異なる。版本（主に木版印刷）や写本（筆写）から構成される古典籍は字形が安定しており、相対的に認識しやすい。対して古文書や古記録は一般的に字形のバラつきが大きい。

また書体も多様性を構成する要素になる。たとえば公式文書で使われるお家流と勘亭流（かんていりゅう）では全く字形が異なる。古文書/古典籍を文字認識する場合、現代人にも読める楷書も認識対象に含まれることになる。さらに変体仮名によるバラつきもある。同じ平仮名の「あ」でも、元の字母が「安」と「阿」では異なる字形となる。

日本語OCR全般に言えるが、文字種が欧米の言語と比べて多く、必要な学習データ数がそれに比例して多くなってしまう。さらに出現頻度はZipfの法則に近い形に従うため、低い出現頻度の文字種の認識が難しくなる。

図3　くずし字認識コンペティションサイト

3.2　TOPPANくずし字AI-OCRの特徴

われわれは3.1節に挙げた技術課題を解決するために、アルゴリズムまで含めた形でチューニングを行っているが、加えて下記の2点が大きな特徴といえる。

文字位置の推定機能

多くの日本人にくずし字が難読であることを踏まえると、認識結果について文字ごとにどこに書かれているかが分かることが利便性の観点で重要な要素となる。

OCRの典型的なアプローチとして、行画像から行テキストシーケンスに変換する手法があるが、その場合、仮にアテンション機構を使ったとしても正確に文字単位の位置を同定することは難しい。

当社が展開する「ふみのは®」サービス[8]が大学等での教育目的でも利用されていることも踏まえ、当社エンジンには文字位置推定の仕組みが搭載されている。

データ収集体制の構築

機械学習技術を構築する上で、学習データセットは極めて重要な要素となる。当社ではデータ収集体制を構築し、継続的に収集している。データ収集がOCR精度の改善につながり、それがデータ収集を効率化するという好循環サイクルが出来上がっている。

図 4　古文書解読とくずし字資料の利活用サービス「ふみのは ®」

3.3 コンペティションによる技術導入

昨今のディープラーニング技術の急速な進展を踏まえると、社内開発だけでは技術進化を十分に取り込めない。そこで機械学習コンペティションによる技術導入も図った（図 3）[7]。SIGNATE 社の協力を得て、行領域抽出タスクと行内文字認識タスクの 2 部門でコンペティションを開催し、延べ 1,097 名の優秀な技術者に参加いただいた（投稿者数は 133 名）。入賞者による工夫はすでに当社の OCR エンジンの改善に役立てられている。

4．ふみのは ® サービス

TOPPAN では 2011 年から OCR に関する事業に取り組んでいるが、くずし字に関して「ふみのは」という名称でサービス展開を行っている（図 4）。具体的には大学の授業や市民参加型ワークショップなどを対象としたふみのは ® ゼミのほか、古文書を解読するサービスや古文書解読用のスマートフォンアプリ（古文書カメラ ®）の提供を行っている。

ふみのは ® サービス活用事例

すでに多数の利用実績があるが、ここでは典型的な事例を紹介する。岐阜県垂井町は、郷土史家の収集した膨大な史資料の寄託依頼を受けた。当該事例では寄託に向けた調査・整理業務に「ふみのは ® ゼミ」を導入いただいた結果、整理段階で概要を把握することが可能となり、調査期間を計画より 1 年短縮することに成功した。

5．おわりに

本稿では古文書の有効活用に向けて TOPPAN グループで開発しているくずし字 AI-OCR 技術とその関連サービスを紹介した。古文書には難読文字が多く、効果的な活用にはさらなる機能改良や精度改善が不可欠である。TOPPAN グループは文化貢献のため、ひきつづき技術開発を進める計画である。

参考文献

1) 赤間亮、岡敏生：AI 技術を応用したくずし字翻刻学習・指導システム，画像ラボ（2020）
2) くずし字学習支援アプリ KuLA：https://kula.honkoku.org/
3) 国文学研究資料館：日本語の歴史的典籍の国際共同研究ネットワーク構築計画，(2013-)
4) 国立国会図書館次世代デジタルライブラリー，https://lab.ndl.go.jp/dl/
5) 古文書カメラ：https://camera.fuminoha.jp/
6) 凸版印刷株式会社：江戸期以前のくずし字を高精度でテキストデータ化する新方式 OCR 技術を開発，https://www.holdings.toppan.com/ja/news/2015/07/newsrelease150703_2.html, (2015)

7) 凸版印刷株式会社 くずし字認識チャレンジ：
https://signate.jp/competitions/580
https://signate.jp/competitions/581
8) ふみのは：
https://www.toppan.com/ja/joho/fuminoha/
9) みを (miwo)：http://codh.rois.ac.jp/miwo/
10) みんなで翻刻：https://honkoku.org/
11) 木簡・くずし字解読システム：
https://aimojizo.nabunken.go.jp/