ディープラーニングによる変体仮名の翻刻およびWWWアプリケーション開発の試み

早坂 太一・大野 亙・加藤 弓枝（国立高等専門学校機構 豊田工業高等専門学校）
山本 和明（人間文化研究機構 国文学研究資料館 古典籍共同研究事業センター）

国文学研究資料館古典籍共同研究事業センターにより構築が進められている「日本語の歴史的典籍データベース」は，これを有効活用することで，異分野を融合させた研究の展開も期待されるが，いかに資料が集積されたとしても，多くの研究者にとっては，書かれている文字が「くずし字」であることが障壁となる．本研究は，世界的に注目されている人工知能技術である，ディープラーニングを用いたくずし字の自動翻刻システムの構築を目的とする．オープンデータとして公開されているいくつかの歴史的典籍内の変体仮名に対して，人工知能による認識の精度を算出するとともに，学習したモデルをWWWアプリケーションとして実装した．

Recognition of *Hentaigana* by Deep Learning and Trial Production of WWW Application

Taichi Hayasaka / Wataru Ohno / Yumie Kato
(National Institute of Technology, Toyota College)
Kazuaki Yamamoto (Center for Collaborative Research on Pre-modern Books, National Institutes for the Humanities, National Institute of Japanese Literature)

Effective utilization of "*Pre-modern Japanese book database*" constructed by the project supervised by Center for Collaborative Research on Pre-Modern Texts, NIJL will push forward the development of the interfiled study. It may become obstruction for the researchers with a little knowledge of classical literature, however, because historical Japanese texts have been written by *Kuzushiji* (*Hentaigana* and cursive script). In this article we report an attempt of recognizing *Hentaigana* by deep learning, which is the artificial intelligence technology regarded throughout the world. Using the convolutional neural networks, we obtained a rate of correct distinction of *Hentaigana* in several pre-modern texts in open database. Furthermore, we developed the WWW software application to recognize *Hentaigana*.

1．まえがき

近年，くずし字に関する研究が注目される契機となったのは，国文学研究資料館により平成26年度より開始された「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」[1]である．この計画では，研究基盤整備として，約30万点の歴史的典籍を画像データ化し，既存の書誌情報データと統合させた「日本語の歴史的典籍データベース」の構築を行うことになっている．あらゆる分野の書籍が含まれる，膨大な画像データを有効活用できれば，例えば，津波や噴火などの天変地異の歴史を教訓とした防災研究のように，人文科学のみならず，自然科学系分野を融合させた研究の展開も期待される．しかしながら，いかに資料が集積されたとしても，多くの研究者にとっては，それらに書かれている文字が「くずし字」であることが障壁となる．

上記のような経緯で，くずし字翻刻に関する研究は以前にも増して待ち望まれるようになった．殊に，コンピュータ技術を利用したアプローチは，最も先行研究が多い分野である．そのような状況下において，著者ら[2]はネオコグニトロンを用い，変体仮名を対象とした翻刻に挑戦した．翻刻精度は65%あまりと悪くはないが，今後，仮名文字だけでなく漢字も含めた大量のデータを対象とする場合には，計算時間および精度といった観点から，従来のモデルでは手に余る状況が予想される．

本研究では，従来の方法よりも格段に優れた性能を示すことから，様々な分野で導入が進められつつあるディープラーニング(deep learning) [3]

を利用した，くずし字翻刻のための人工知能を構築し，それを用いて「いかなる場面や人々でも，くずし字翻刻を行うことができる」ソフトウェアを開発することを目的としている．ディープラーニングは，大量のデータを扱うことが可能であるという特徴を有するため，歴史的典籍に含まれるあらゆるくずし字の翻刻に対しても，極めて有用な方法であることが予想される．

2．人工知能によるくずし字翻刻の有用性

現行のくずし字翻刻に関する研究は，複数の区分に跨がる研究もあるが，主に以下の三系統に分けられる．

1) 学習者のくずし字解読能力・効率を高める方法に関する研究
2) コンピュータ技術によるくずし字自動翻刻に関する研究
3) 変体仮名の文字コード標準化に関する研究

特に，2) に関する研究は，最も先行研究の蓄積があり(例えば[4])，進捗度の大きい分野であると考えられる．特に，2015 年 7 月に報道された，凸版印刷株式会社よる「くずし字を高精度でテキストデータ化する OCR 技術の開発」[5]は記憶に新しい．同社は 2013 年より古文書をデータ化する「高精度全文テキスト化サービス」を提供してきたが，この技術を公立はこだて未来大学が開発した文書画像検索システム[6]と組み合わせることで，くずし字で記されている古典籍のＯＣＲ技術を開発したものである．このシステムについては，国文学研究資料館の協力の下で動作検証が行われている．テキストデータ化済みの文献を，ＯＣＲ処理に用いるくずし字データベースとして使用することで，くずし字で記された文献を 80%以上の精度でテキストデータ化することが可能であることが発表されると，報道機関によって驚きをもって伝えられた．

また，ＯＣＲ技術を用いないテキストデータ化に関する研究に関しては，中京大学が挑戦を始めたことが報道されたことも記憶に新しい[7]．この研究は，解読が難しいとされる明治時代から戦中までに書かれた文書の解読システムの構築を目指したものである．特に，台湾に保管されている台湾総督府時代の行政文書を解読しながらシステムをつくるという．これらの文書を読み取れるようになれば，江戸時代から現代まで幅広い文書が解読できるほか，中国語の識別も可能になるという．また，走り書きのカルテや中国の古文書を解読するなどの活用法も想定されている．

同じく 2) に分類される本研究において用いるディープラーニング[3]は，ヒト脳内における多数の神経細胞による情報のやりとりを，数式によりモデル化したニューラルネットワークが基になっている．ディープラーニングにより翻刻を行うモデルを構築するには，GPGPU (general-purpose computing on graphics processing units) をはじめとする最新の計算機技術を必要とするが，一度モデルを構築しさえすれば，ニューラルネットワークと同様に，翻刻に要する時間はごく僅かである．また，学習に用いる文字画像を多数用意する必要はあるが，学習後のモデルには，それぞれの典籍やそれらが書かれた時代で異なる可能性のあるくずし字の特徴が反映されているため，OCR 技術のように，翻刻の際に膨大なデータベースを用意する必要はない．つまり，人工知能技術の導入によって，一般的に普及している携帯情報端末でも動作する，小規模なアプリケーション・ソフトウェアとして，「いつでも／どこでも／誰でも自動翻刻」を実現することが可能になると考えられる．

3．ディープラーニングによる学習

3.1. データセット

本研究では，変体仮名画像を，それぞれ平仮名「あ」「い」…「ゑ」「を」「ん」の 48 クラスに分類する学習を行った．ここで，濁点が画像中に含まれていても，分類上は考慮しない．

変体仮名を一文字ずつ，64×63 ピクセルの大きさにリサイズし，ネガ・ポジを反転した JPEG 形式のグレイスケール画像として用意し，学習用，学習途中のテスト用，および学習後のテスト用にそれぞれ分類した．学習に用いるデータとして，『五體字類』[8]の 1,473 文字，『和翰名苑』仮名字体データベース[9]の 3,265 文字，および正保 4 年(1647)に出版された『古今和歌集』[13]から 3,140 文字，計 7,878 文字の変体仮名画像を用意した．学習途中のテストに用いるデータは，慶長年間頃に出版された『平治物語』巻一[10]から最初の 150 字の変体仮名画像を，学習後のテストに用いるデータは，承応 3 年(1654)に出版された『源氏物語』桐壺[11]から 10,026 字の変体仮名画像を，それぞれ切り出して利用した．

なお，後述するネットワークモデルへは，62×62 ピクセルに切り取られたものが入力される．

3.2. ネットワークモデル

本研究で用いた convolutional neural network (CNN)と呼ばれる典型的なネットワーク構造を図 1 に示す．入力層から出力層へ向けて，畳み込み層(convolutional layer)とプーリング層(pooling layer)，および ReLU(rectified linear unit)

がセットで並び，これが複数層重なっている．その後，全結合層(fully connected layer)と呼ばれる，隣接層間のユニットが全て結合された層が配置される．最後に，softmax関数により，それぞれの平仮名に分類される確率が出力される．本研究では，三つの畳み込み層と二つの全結合層によるCNNモデルを，繰り返し回数40,000回で事前学習させた後，そのネットワークを初期解として，四つの畳み込み層と二つの全結合層によるCNNモデルを，繰り返し回数60,000回で学習させたものを採用した．

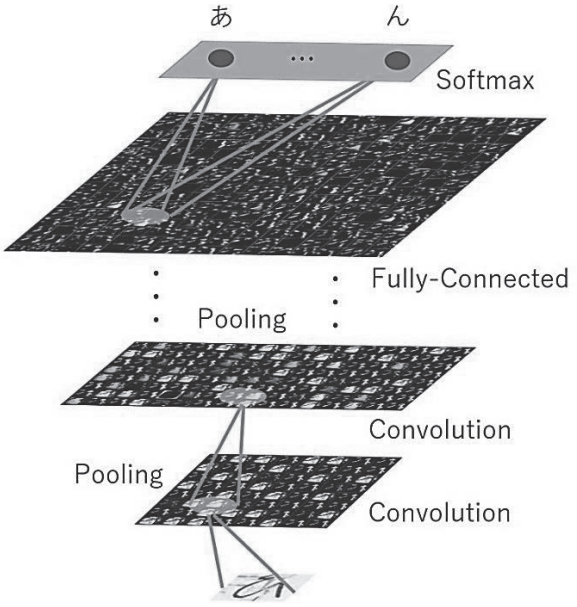

図1. ディープラーニング(CNN)における典型的なネットワークの構造
Figure 1. Typical structure of convolutional neural networks in deep learning.

3.3. 数値解析環境

本研究における一連の数値計算は，代表的なディープラーニング用ライブラリであるCaffe[12]を用いて行われた．GPGPUによる計算の高速化を図るために，計算機環境として，OSはUbuntu 14.04，CPUはIntel Core i5，GPUはnVidia GeForce GTX 750を搭載したパーソナルコンピュータHP EliteDesk800 G1 TWRを利用した．

4．変体仮名の分類

学習したCNNモデルに，学習中のテストデータ『平治物語』巻一150字の変体仮名を入力し，分類させた結果を表1に示す．表1において，「第一候補」とは，正解となる平仮名の分類確率が最上位であったもの，「10%以上」とは，分類確率は最上位ではないが，その値が10%以上であったものの割合を，それぞれ示している．分類確率10%以上までを考慮すると，85%を超える精度が得られている．

表1. CNNモデルによる『平治物語』巻一 における変体仮名の認識結果
Table 1. Recognition results of *Hentaigana* in *Heiji Monogatari* by CNN.

	第一候補	10%以上
分類結果	75.3%	10.0%

次に，学習後のテストデータとして『源氏物語』桐壺10,026字の変体仮名を入力し，分類させた結果を表2に示す．表2より，分類確率10%以上までを考慮すると，85%に近い精度が得られていることがわかる．

表2. CNNモデルによる『源氏物語』桐壺 における変体仮名の認識結果
Table 2. Recognition results of *Hentaigana* in *Genji Monogatari* by CNN.

	第一候補	10%以上
分類結果	74.3%	9.9%

『源氏物語』桐壺 10,026 字のデータの中で，「第一候補」および「10%以上」を合わせて90%以上認識できた平仮名を表3に示す．表3にある仮名数は24と，全体の半数である．また，「濁点割合」とは濁点を含む文字の割合であり，その値よりも誤認識率の方が低いことから，濁点が変体仮名認識に与える影響が限定的であることが示唆される．

表3において，文字数が300以上の変体仮名の画像例を図2に示す．形状特徴が明らかに異なる複数の字母を持つ変体仮名に対しても，高い認識率が得られていることがわかる．

表3に示す好ましい結果とは逆に，「第一候補」および「10%以上」を合わせて70%未満の認識率であった平仮名を表4に示す．表4にある仮名については，アスペクト比の影響から，他の仮名と混同されている傾向がある．例えば，図3に示すように，「え」を「し」と（52.8%），「り」を「か」と（49.3%），「よ」を「に」と（24.1%）分類していることが多い．ただし，「す」については，濁点が含まれていることによる影響であることが否めない．

表 3. 『源氏物語』桐壺 における
認識率の高い変体仮名
Table 3. *Hentaigana* in *Genji Monogatari* with higher recognition rates.

	文字数	第一候補	10%以上	濁点割合
む	43	97.7%	2.3%	—
ゑ	9	100.0%	0.0%	—
の	410	99.0%	0.5%	—
を	186	98.4%	1.1%	—
れ	149	99.3%	0.0%	—
や	99	98.0%	1.0%	—
め	95	91.6%	7.4%	—
ん	92	93.5%	5.4%	—
ひ	182	97.8%	1.1%	15.9%
い	229	96.9%	1.7%	—
ふ	126	96.0%	2.4%	3.2%
ぬ	45	75.6%	22.2%	—
あ	155	89.0%	7.7%	—
ろ	64	87.5%	7.8%	—
お	241	86.7%	8.3%	—
た	360	90.0%	5.0%	12.2%
へ	185	88.6%	4.9%	25.9%
け	175	89.1%	4.0%	28.0%
し	543	88.0%	5.0%	8.1%
わ	56	75.0%	17.9%	—
き	324	84.6%	7.1%	13.9%
て	324	86.4%	5.2%	16.0%
ま	293	84.0%	7.5%	—
ち	105	83.8%	7.6%	13.3%

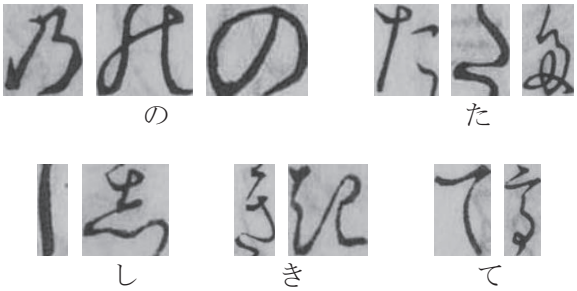

図 2. 『源氏物語』桐壺 における
認識率の高い変体仮名の例
Figure 2. Examples of *Hentaigana* in *Genji Monogatari* with higher recognition rates.

表 4. 『源氏物語』桐壺 における
認識率の低い変体仮名
Table 4. *Hentaigana* in *Genji Monogatari* with lower recognition rates.

	文字数	第一候補	10%以上	濁点割合
ゐ	16	50.0%	18.8%	—
す	182	53.3%	14.8%	41.2%
ね	34	47.1%	17.6%	—
る	248	48.0%	16.1%	—
ら	207	43.0%	19.3%	—
せ	132	42.4%	12.1%	3.0%
よ	79	27.8%	26.6%	—
り	353	38.5%	15.6%	—
そ	103	20.4%	25.2%	20.3%
え	89	29.2%	3.4%	—

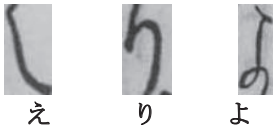

図 3. 『源氏物語』桐壺 における
認識率の低い変体仮名の例
Figure 3. Examples of *Hentaigana* in *Genji Monogatari* with lower recognition rates.

これらの結果については，学習に用いたデータの「典型」が，テストデータのそれに合致していたのではないか，ということは否定できない．あらゆる時代の歴史的典籍に対し，さらなる精度の向上を目指すためには，今後，WWW 上のオープンデータを利用して，学習およびテストに用いるデータ数を充実させることが必要であると考えられる．

5．WWWアプリケーションの実現

古典籍の画像データを読み込み，マウスで選択された 1 文字分の変体仮名を翻刻する WWW アプリケーションを試作した(http://vpac.toyota-ct.ac.jp/hayasaka/kuzushiji/)．ブラウザ画面の例を図 4 に示す．

読み込まれた画像に対し，openCV 2.4 を利用して，グレイスケール変換，ネガ・ポジ反転，コントラスト調整，さらにリサイズを施し，Caffe によって学習された CNN モデルに入力することで，平仮名ごとの分類確率が出力され，グラフとして表示される．プログラミング言語は java script および python2.7 を，API として jQuery

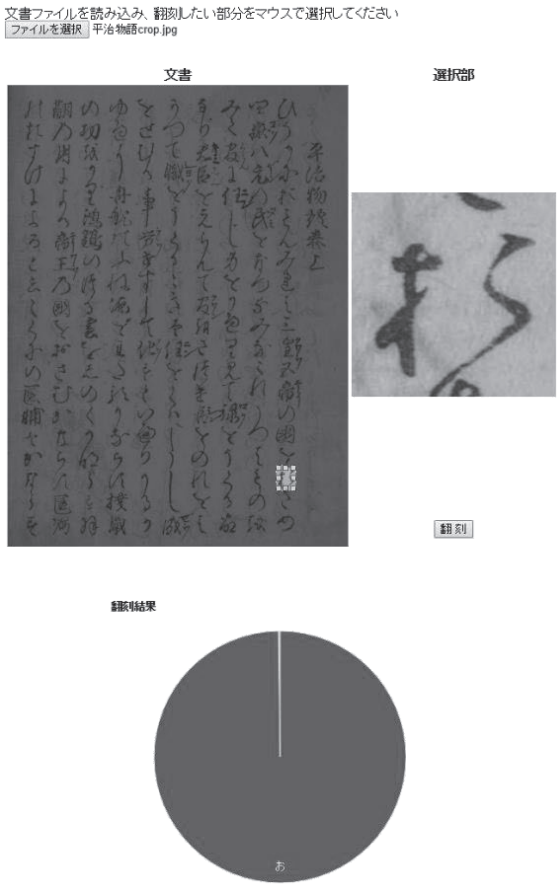

図 4. 開発したWWWアプリケーションによる変体仮名翻刻の例
Figure 4. Example of machine reprinting of *Hentaigana* in our developed WWW application.

(ImageSelect プラグインを含む) および Google Chart を使用した.

WWW サーバのハードウェアとして, Apple Mac Mini を用い, GPU ではなく, CPU による演算を行わせた. 表示については, クライアント側の計算機環境に依存するが, サーバ側で 1 文字あたりの分類にかかる時間は約 0.4 秒であった. 高性能なハードウェアや GPGPU を利用しなくとも, 十分な演算速度による翻刻が実現できることが伺える.

6. むすび

本研究では, 日本語の歴史的典籍の自動翻刻を目的として, ディープラーニングにより, 変体仮名を対象とした文字認識を行わせ, さらに試作ではあるが, それを WWW アプリケーションとして実現した. 結果として, 精度は決して高くなかったが, 有効な学習データを揃えることで, それぞれの文字の本質的な特徴を獲得できることから, こうしたアプローチがくずし字の認識にも有効であることが示唆された. 学習データを増加させることによって, 認識率の向上に繋がることが期待される.

今後は, 認識率の向上を目指すことはもちろんであるが, くずし字を一般の人々でも扱いやすくするべく, このアプローチをアプリケーション・ソフトウェアとして実装することが, 課題として挙げられる. 例えば, 茶席などで床の間の掛け軸をタブレットやスマートフォンで撮影すると, 書かれているくずし字や言葉を知ることができるアプリなどが考えられる. また, くずし字を撮影してデータ化し, その情報から生成されたキャラクタ同士で対戦するようなコンピュータゲームが開発できれば, 児童や生徒らがくずし字に親しむきっかけを与えることができると考えられる. どのようなアプリケーション・ソフトウェアであれば, くずし字に対して, より興味を持たせることが可能かを検討し, 仕様を策定していきたい.

近い将来, 人工知能技術の発展により, 一方的な情報伝達や単純作業を伴う労働が駆逐されるという懸念がある. ゆえに, 本研究の成果が, 翻刻作業に人間を必要としなくなるという指摘が想定される. 然り, 人工知能技術開発の究極の目標は, 人手を介さない知的作業の実現にあるとも言えるが, 例えば, 機械翻訳技術が急速に発展している現在でも「翻訳」という職業はなくならないように, 歴史的典籍が持つ「古人の心」を伝えるためには, やはり文学研究者の力が必要となる.

本研究の成果は, 海外を含む様々な地域および分野の研究者が, 日本に膨大に残る歴史的典籍を判読することを支援する「夢の技術」へと進展していくと考えられる. このことは, 日本の歴史的典籍の海外における利用価値を高めることにも繋がる. また, 研究者のみならず, 一般の人々でも, 本研究の成果を利用して, 歴史的典籍に記された知識の遺産を有効活用することが期待される. このように, 持続可能な社会を実現するためにも, 本研究が果たす役割は少なくないと考えられる.

謝辞

この研究は, 本研究は JSPS 科研費 JP16K02433 の助成, および平成 28 年度内藤科学技術振興財団研究助成を受けたものです.

参考文献

[1] 国文学研究資料館:歴史的典籍に関する大型プロジェクト, <https://www.nijl.ac.jp/pages/cijproject/>（参照 2015-10-14）

[2] 早坂太一，大野亙，加藤弓枝 : ネオコグニトロンによる日本語の歴史的典籍におけるくずし字の認識, 豊田工業高等専門学校研究紀要, No.48, pp.5-12（2015）
[3] 岡谷貴之 : 深層学習，講談社（2015）
[4] 和泉勇治，加藤寧，根元義章，山田奨治，柴山守，川口洋 : ニューラルネットワークを用いた古文書個別文字認識に関する一検討, 情報処理学会研究報告，1999-CH-045（2000）
[5] 凸版印刷株式会社 : ニュースリリース，<http://www.toppan.co.jp/news/2015/07/newsrelnews150703_2.html>（参照 2015-10-14）
[6] 公立はこだて未来大学 : 文書画像検索システム, <http://records.c.fun.ac.jp/>（参照 2016-9-6）
[7] 中日新聞 : 崩し字の壁 崩せ 自動解読システム 中京大挑戦，<http://edu.chunichi.co.jp/?action_kanren_detail=true&action=education&no=6016>（参照 2015-8-10）
[8] 法書会編 : 五體字類，<http://www.let.osaka-u.ac.jp/~okajima/PDF/5tai/>（参照 2015-11-12）
[9] 岡田一祐 :『和翰名苑』仮名字体データベース, <https://kana.aa-ken.jp/wakan/>（参照 2016-8-16）
[10] 国立国会図書館 : 国立国会図書館デジタルコレクション 平治物語，<http://dl.ndl.go.jp/info:ndljp/pid/2544708>（参照 2016-1-14）
[11] 国立情報学研究所 : 国文研古典籍データセット（第 0.1 版）, 源氏物語, <http://jcbsv.nii.ac.jp/oa/NIJL0-1/items/NIJL0001.zip>（参照 2016-7-25）
[12] Berkeley Vision and Learning Center: Caffe, <http://caffe.berkeleyvision.org/>（参照 2015-11-11）
[13] 国立情報学研究所 : 国文研古典籍データセット（第 0.1 版）, 二十一代集, <http://jcbsv.nii.ac.jp/oa/NIJL0-1/items/NIJL0002.zip>（参照 2016-7-25）