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　５ 年　 ５ 月 １６ 日現在

機関番号：１２３０１
研究種目：挑戦的研究（萌芽）
研究期間：2018～2022
課題番号：１８Ｋ１９８００

研究課題名（和文）計算機による古文書の自動的な活字化

研究課題名（英文）Text Recognition of Historical Japanese Documents

研究代表者
長井　歩（Nagai, Ayumu）

群馬大学・大学院理工学府・助教

研究者番号：７０３７５５６７

交付決定額（研究期間全体）：（直接経費）　　1,800,000円

研究成果の概要（和文）：行単位に分割された崩し字画像を入力に，認識したテキストを出力する深層学習のシステムを開発した．3文字の文字列認識が課題のPRMUアルコンで41チーム中2位，ページ単位の認識が課題のKaggleのコンテストで293チーム中6位となった．
さらに，肉筆の崩し字認識率向上を目的に，肉筆の崩し字のページ画像と対応するページ単位の翻刻テキストを入力に，くずし字の行画像とその翻刻テキストのペアを出力するシステムを開発した．これは肉筆のくずし字だけから成る一定以上の規模のデータとしては初めてである．この崩し字データを従来の公開データと共に別途学習に使うと，従来に比べ4.5%程度正解率が向上した．

研究成果の学術的意義や社会的意義
本研究の意義は，崩し字で書かれた版本や写本を計算機で自動的に活字化（翻刻）することである．江戸時代の古文書はその99%以上が翻刻されておらず，最後に残された最大の文字文化である．しかし多くの現代人にはそれを容易には読みこなせない問題がある．古文書を読むためには専門的な知識と訓練を要し，現状では圧倒的に人手が足りていない．この問題を解消すべく，計算機による自動的な古文書の翻刻に貢献した．現在では版本に対しては95%前後の正解率を叩き出すが、版本ではない肉筆の古文書の崩し字には、読みにくい文字がまだまだ沢山ある。これらの難易度の高い崩し字の認識も視野に見据え、正解率を高める１つの方法を提案した。

研究成果の概要（英文）：We have developed a deep neural network that outputs transcription from input images of lines of historical Japanese cursive. We placed 2nd out of 41 teams in the PRMU competition, where recognition of 3-character strings was the task, and 6th out of 293 teams in the Kaggle contest, where recognition of page units was the task.
Furthermore, with the aim of improving the recognition rate of autograph historical Japanese cursive, we developed a system that takes as input a page image of cursive and the corresponding ground-truth text for each page, and outputs a pair of a line image of cursive and its ground-truth text. This is the first time that data of a certain size consists only of autograph historical Japanese cursive. Using this data together with the conventional public data for training improved the accuracy rate by about 4.5% compared to the conventional data alone.

研究分野：人工知能

キーワード：くずし字認識　文字認識　深層学習　翻刻

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

「日本を今一度せんたくいたし申候」との坂本龍馬の手紙は，日本人にとってあまりにも有名である．しかし，活字ではなく龍馬の筆跡で読んだことがある人は殆どいないだろう．現代の日本人にとって，150 年余り前の生き生きとした坂本龍馬の手紙を読むよりも，英和辞典を片手に 400 年以上前のシェイクスピアの初版本を読む方が断然楽である．何故こんなことになってしまったのだろうか．その最大の理由は，現代人には崩し字を読めないことであろう．
実は日本は古文書大国である．200 年ほど前の古文書を神保町で誰でも手軽に買えるような国は，世界広しと言えどなかなかないと聞く．それなのに，崩し字が読めないから読まないのは，あまりにも悲しいことではないだろうか．150 年前の日本人は 1000 年前の日本人の文字が読めたのにもかかわらず，である．これは日本人にとって大きな損失である．事実，ロバート・キャンベル氏は，「日本語自体がとんでもないビッグデータを蓄積しているのに，くずし字が読めないためにアクセスできないのは本当にもったいない」と表現している．また，歴史家の磯田道史氏は「昨今の歴史小説は出版数は多いが，面白いものが少な」く，その理由は「生の古文書が読めない」ために「誰かがすでに書いて活字になった本をもとに想像をふくらませて歴史を書」いているのが原因だと述べている．限られた活字化された古文書の中で堂々巡りするのではなく，翻刻されていない膨大な古文書をもっと身近なものにする必要がある．
計算機で活字や手書き文字を認識する研究は 20 年以上前から行われているが，その成果だけで古文書を活字化できるとは到底思えない．主な理由は，崩し方の多様性と，上下の文字同士の連結性の２つである．現代人が書く手書き文字は活字から大きくかけ離れることはないのに対し，崩し字には多様な崩し方がある．さらに崩し字は上下の文字が連結していることが多い．文字認識の従来研究は 1 字 1 字が切り離されていることを前提とすることが多く，そもそも前提が違う．1 文字ずつ分離できたとしても，その文字単独ではとても読めないレベルにまで崩されていることも多い．これらの理由から，1 文字ごとに文字を認識する従来の研究では本課題に太刀打ちできない．近年，人工知能が学術界のみならず，産業界や世間一般からも脚光を浴びている．その火付け役となったのは，画像の被写体を推定する物体認識という分野における推定精度の目覚しい向上である．成功の背景には深層学習という，ニューラルネットを大規模化しつつ，学習方法を洗練させた技術の存在がある．文字認識においても深層学習の成果を取り入れるのが目的達成の近道と考えられる．

２．研究の目的

崩し字で書かれた版本や写本を計算機で自動的に活字化（翻刻）することである．江戸時代の古文書はその 99%以上が翻刻されていない．龍馬の手紙は例外的存在なのである．江戸時代以前の古文書は数が少なく，発見されているものは殆ど活字化されているので，江戸時代の古文書は最後に残された最大の文字文化である．しかし多くの現代人にはそれを容易には読みこなせない問題点がある．古文書を読むためには専門的な知識と訓練を要し，現状では圧倒的に人手が足りていない．この問題を解消すべく，計算機によって自動的に古文書を活字化する研究を切り拓くことを目指す．

３．研究の方法

ニューラルネットを大規模化させつつ，学習方法を洗練させた技術である深層学習の成果を積極的に取り入れた．特に文字列認識の分野で成果をあげている CRNN という手法を核に，比較的単純な工夫から込み入った方法まで様々なアイデアを導入することによって崩し字認識の向上を目指した研究を行った．
まずは出現頻度の少ない文字の学習効率を上げるために，そのような文字を含む文字列のデータを人工的に加工して増やしたり，言語モデルを導入したりである．
次に，まず複数のモデルを学習し，それらの推定結果を統合することによってモデルの改善を繰り返し行うことによって崩し字認識の精度を上げた．
最後に，古文書の画像と正解テキストがページ単位でしか対応していないデータから，崩し字の１行の画像とその正解テキストのペアを自動的に収集することによって，難易度の高い新たな訓練データを作り，それを訓練に利用することによって崩し字認識の精度を上げた．

４．研究成果

(1) 既に行単位に分割された崩し字の文字列の画像を入力として，テキストの文字列を出力する深層学習のシステムを開発した．このシステムは，申請者が知る限り最高の認識性能を達成した．近年，崩し字認識の研究事例が増えてきつつあるが，その殆どは１文字単位に分割された崩し字の画像を入力としている．しかし崩し字は前後の文字が連結していることが多く，１文字単位への分割自体が非常に難易度の高い問題である．申請者はこの問題を避けるべく，１行の崩し字の文字列を入力の単位とした．その他にも様々な工夫を導入した．出現頻度の少ない文字の学習効率を上げるために,そのような文字のデータを人工的に加工して増やしたり(データ拡大)，画像データとは別に膨大なテキストの情報を用いて認識精度を上げたり（言語モデルの導入）などである．また，古文書特有の 2 文字が合体した字（合字）を，合字としてではなく 2 文字として学習させた場合の認識精度や，誤認識を起こしやすい文字の組み合わせについても解明した．

(2) 行単位に分割された崩し字の文字列画像を入力として，テキストの文字列を出力する深層学習のシステムを改良したことと，入力画像がページ単位の場合にも対応できるようにしたことである．前者については，１年目に行単位の画像を入力としたシステムを開発したが，乱数の種を変えた上でそのシステムで複数回実行すると，出力として複数のテキスト候補を得られる．それらの候補の情報と言語モデルとを組み合わせて最も自然なテキストを出力し，そのテキストを正解文字列と仮定して次の学習に活用することによって学習精度を向上するという一連の学習サイクルを繰り返すことによって正解率を最大 10%程度向上させた．後者については，ページ単位での認識，言い換えると複数行に渡る崩し字を認識できるようなシステムを開発した．行単位での認識は申請者の知る限り最高級の認識性能に既に達しているので，入力であるページ画像を複数の行に分割する部分を開発し，行分割された画像を行認識のシステムに引き渡すことによってページ単位での認識を実現した．この研究は，3 文字の文字列認識を課題とした PRMU アルコンで 41 チーム中 2 位，ページ単位の認識を課題とした Kaggle の崩し字認識コンテストで 293 チーム中 6 位になるという成果に結びついた．

(3) プライベートな文書に出現する肉筆の崩し字の認識精度を上げるための研究である．近年公開されている大量のくずし字の画像データを使って学習すると、我々のシステムも含め 95%前後の正解率を叩き出すことができる。その結果、崩し字認識の基礎研究は一段落付いたようにも見える。しかしそれは違う。公開されているくずし字のデータは手書き文字には違いないが、殆どが庶民向の版本なので読みやすいくずし字である。具体的には、漢字より平仮名を好んで使っていたり、変体仮名の種類が少なかったり、くずしの程度が軽かったりする。教育レベルが低い一般大衆にとっても読みやすいように、清書専門の職人が工夫して書いた文字である。それに対し、版本ではない肉筆の文書の崩し字の中には、版本と違い読みにくい文字がたくさんある。それらの多くはプライベートな文書であって、一般大衆向けに書かれたものではないからである。版本のくずし字認識はエキスパートなレベルにまで向上したかもしれないが、肉筆の崩し字認識はその限りではない。肉筆の崩し字認識には今まで以上の工夫が必要である．そこで肉筆のくずし字のページ画像と対応するページ単位の翻刻テキストを入力として与えると，くずし字の行画像とその翻刻テキストのペアを出力するような学習システムを開発した．これは肉筆のくずし字だけから成る一定以上の規模のデータとしては初めてのものである．更に，このくずし字データを従来の公開データと共に別途学習に使うと，従来に比べ 4.5%程度正解率が向上した．

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Ayumu Nagai	４．巻 13108
２．論文標題 Generation of a Large-Scale Line Image Dataset with Ground Truth Texts from Page-Level Autograph Documents	５．発行年 2021年
３．雑誌名 Lecture Notes in Computer Science	６．最初と最後の頁 354-366
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-92185-9	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3件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Ayumu Nagai
２．発表標題 Generation of a Large-Scale Line Image Dataset with Ground Truth Texts from Page-Level Autograph Documents
３．学会等名 The 28th International Conference on Neural Information Processing (ICONIP 2021)（国際学会）
４．発表年 2021年

１．発表者名 Ayumu Nagai
２．発表標題 On the Improvement of Recognizing Single-line Strings of Japanese Historical Cursive
３．学会等名 The 15th International Conference on Document Analysis and Recognition (ICDAR 2019)（国際学会）
４．発表年 2019年

１．発表者名 Ayumu Nagai
２．発表標題 Recognizing Japanese Historical Cursive with Pseudo-Labeling-aided CRNN as an Application of Semi-Supervised Learning to Sequence Labeling
３．学会等名 The 17th International Conference on Frontiers in Handwriting Recognition (ICFHR 2020)（国際学会）
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

くずし字の肉筆データ（ground truth付き） https://gadget.inf.gunma-u.ac.jp/dl/autograph.tar.gz

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