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　２ 年　 ６ 月 ３０ 日現在

機関番号：１４６０２
研究種目：基盤研究(B)（一般）
研究期間：2017～2019
課題番号：１７Ｈ０１８２９

研究課題名（和文）近代書籍からの知の抽出

研究課題名（英文）Extracting Knowledge from Japanese Early-Modern Printed Books

研究代表者
城　和貴（Joe, Kazuki）

奈良女子大学・生活環境科学系・教授

研究者番号：９０２８３９２８

交付決定額（研究期間全体）：（直接経費）　　12,300,000円

研究成果の概要（和文）：本研究では四つの成果を得ることができた。まず、2017年度にそれまでの認識手法を統合し、学習データが少ないものの、実用化に近い2,678種の近代書籍文字を9割以上の認識率を記録した。次に学習データを増やすために深層学習を利用して未知の近代書籍文字種を自動生成する手法について2018年度に発表した。さらに2019年度には既存認識手法を一新し、深層学習を利用することで2017年度と同等の性能を示し、さらに転移学習を行うことで、9割程度だった認識率を一気に98%まで引き上げることができた。また、実用化には必須のレイアウト解析にも深層学習を利用できることを示した。

研究成果の学術的意義や社会的意義
近年個人所有のHDD等記憶メディアが劇的に大容量化し、インターネットを介して自由にデータのアクセスが可能になったことから、紙媒体でしか記録が残されていなかった近代書籍等のアーカイブ化が急速に行われている。しかしながら画像でのアーカイブ化では全文検索が不可能であり、現在のような規格が規定されていなかった頃の活版印刷に対応した自動テキスト化技術の確立は急務の課題である。本研究はその技術の確立を目指したもので、現時点で実用化に極めて近いレベルまで研究が進展している。

研究成果の概要（英文）：Four results were obtained in this study. First, we integrated the previous recognition methods in 2017, and although there is little training data, we are close to a practical application. The recognition rate of 2,678 Japanese early-modern printed characters was recorded at more than 90%. Next, to increase the training data, we used deep learning to automatically generate unknown early-modern printed character types to be presented in 2018. In addition, in 2019, the existing recognition methods was revamped, and by using deep learning, to get the same as in 2017. In addition, by performing transfer learning, the recognition rate has been increased from around 90 % to 98%. We also showed that deep learning can be used for layout analysis, which is essential for practical applications.

研究分野：人工知能

キーワード：自動テキスト化　深層学習　CNN　レイアウト解析　言語翻訳

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

研究代表者は平成 19 年 12 月より国立国会図書館関西館に非常勤調査員として月 4 回勤務しているが，近代デジタルライブラリのテキスト化の可能性について同館電子図書館課のメンバーと議論を行った．同館所蔵の当時 143,000 冊（現在は約 35 万冊）に及ぶ書籍画像を，青空文庫のように人手でテキスト化を行うのは予算的に不可能であるため，国内の大手 IT 企業に随意契約で近代書籍の自動テキスト化について調査を行わせたところ，既存の OCR 技術では全く役にたたないとの結果であった．一般に文字認識は比較的古くから研究されている分野であり、当時既にオフラインの印字文字認識ならびに手書き文字認識は基礎的な研究が終了しており、商用化がなされていた。研究代表者は近代書籍画像に対する文字認識は、印字認識、手書き文字認識に続く第三の文字認識、すなわち活版印刷文字認識という新しい研究分野となりうる可能性を感じた。そこで予備実験として小規模のデータセットを用意し、手書き文字認識の古典的な手法を適用してみた。データセットとしては、既にテキスト化されている青空文庫を利用することで，比較的容易に 256 種の漢字を選び出し，近代デジタルライブラリ―の画像から手作業での文字切り出しを行った．その結果、出版元が同じであっても時代によって認識できない程全く違うフォントが使われていることも判明した．以上のことを踏まえて，近代デジタルライブラリより，2,678 種類の漢字を異なる出版社・時代から各 6 セット以上作成し，認識実験を行ったところ，学習用データが少なくとも 5 セット確保できる漢字については 90%程度の認識率を得ることを確認した．

２．研究の目的

本研究課題での目的は，研究代表者が過去 12 年に渡って手掛けた結果，ほぼ実用化の目途がついた近代書籍用自動テキスト化技術を実際のドメインに適用できるように拡張し，近代書籍の自動テキスト化の基礎技術を完成させることである．そのためにドメインごとに活版印刷活字データを充実させる方法と，新たに深層学習を利用した未知文字種の自動生成を行う．また，文字認識では文字切り出しという前処理を行う必要があるが、文字認識率がある程度高性能になってくると、切り出しの精度が問題となる。本研究課題では、帝国議会議事録を対象とした複合的なレイアウト解析手法を開発する。さらに、近代書籍のテキスト化が自由に行えるようになったとしても，我国で太平洋戦争前に一般的に使われていた文語体を我々特に若い世代の人が読み理解するのは困難であるため，近代書籍からテキスト化された文語体文章を，深層学習を利用して現代口語に自動変換する手法を提案し、予備実験でその手法を評価する．

３．研究の方法

自動テキスト化の基礎技術を完成させる目標に対して、１）ドメインごとに文字種を収集する方法と、２）未知の文字種を自動生成する手法のうち、１）に関しては近代期の日本産婆学会誌を借り受けることができて、約 2 万ページの画像アーカイブを作成し、その出現文字種を調査することでドメインごとの文字種収集が有効であることを示した。２）は深層学習ニューラルネットを使って、明朝体等現代フォントを入力して、特定の近代書籍の当該文字を出力するように学習させることで、未知の文字種についてもそれと同じ現代フォントを入力すれば当該近代書籍文字画像を自動生成できることを示した。

レイアウト解析に関しては、深層学習を利用した画像の汎用クラスタリング手法として知られているセマンティックセグメンテーションを利用して、帝国議会議事録のレイアウト解析を効率的に行う手法を開発した。

近代文語体と現代口語体の相互自動翻訳に関しては、一般的な言語のニューロ翻訳機をサーベイし、sec2sec や Convsec2sec 等を使って予備実験を行った。本サブテーマでは学習データを集めるのに苦労したが、森鴎外の単一の著書に絞って学習を行わせ、自動翻訳が可能であることを示した。

４．研究成果

下記研究業績のうち、1-5 は査読付き国際会議論文で、6-7 は国内研究会での口頭発表である。

業績 8 では、ヒストグラムを使った古典的なレイアウト解析手法とセマンティックセグメンテーションを用いた手法を帝国議会議事録の特定の一ページに適用して比較し、セマンティックセグメンテーションを利用した方が良い性能を出すことを報告し、業績 1 で評価も含めた報告を行った。

業績 6 では近代期の日本産婆学会誌から約 2 万ページを選んで出現文字種の分析を行い、文字種の出現頻度が小説等他ドメインのそれとはまったく違うことを示し、効率的な文字種収集には多様なドメインを対象にすべきことを報告した。

業績 4 では CNN 型のニューラルネットで、画像変換を行うように学習を行わせた。入力画像は現代の特定フォントで、覚えるべき画像は特定の近代出版者の書籍から取ってきた同じ種類の文字である。ある程度の学習セットで学習した後、近代書籍には乗っていない文字種の現代フ

ォントを入力してやると、未知の文字種の画像が生成することを示した。業績 2 では、より効率よく文字種の自動生成を行うためのニューラルネットの構成法について報告した。
業績 7 では、森鴎外の舞姫の原文と翻訳文から学習セットを作り、それを Convsec2sec を用いて学習させることで、ある程度の自動翻訳が可能であることを示した。しかしながら、入手可能な学習セットは入手困難であることが判明した。
業績 5 では、本研究課題開始以前に収集した 2,678 種類の近代書籍文字種 6 セットを使って、3 種類の特徴抽出器と 2 種類の識別器によるアンサンブル学習を行うことで、90%以上の認識が可能であることを示した。このことは、学習セットを増やせば実用的な認識が可能であることを示している。ところが、業績 3 で、現在の JIS 第一水準 20 種類のフォントと上記の近代書籍文字 5 セットを CNN に学習させることで、98%の認識が可能であることを示した。これは CNN を利用する場合、その構造的な性質上、現代フォントが近代書籍文字の転移学習を可能としたことを示している。

1. Sayaka Iida, **他**: Layout Analysis using Semantic Segmentation for Imperial Meeting Minutes (2019).
2. Yuki Takemoto, **他**: Structure of Neural Network Automatically Generating Fonts for Early-Modern Japanese Printed Books (2019).
3. Suzuka Yasunami, **他**: Applying CNNs to Early-Modern Printed Japanese Character Recognition (2019).
4. Yuki Takemoto, **他**: Automatic Font Generation for Early-Modern Japanese Printed Books (2018).
5. Kaori Fujimoto, **他**: Early-Modern Printed Character Recognition using Ensemble Learning (2017).
6. **藤田未希，他：近代書籍における低出現頻度文字種の獲得** (2019-12-4).
7. **林 英里香，他：近代文語体と現代口語体の自動翻訳への試み**(2018-12-10).
8. **飯田 紗也香, 他：帝国議会会議録におけるレイアウト解析手法の比較** (2018-09-18).

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5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 Yuki Takemoto, Yu Ishikawa, Masami Takata, Kazuki Joe	４．巻 On-site Edition
２．論文標題 Automatic Font Generation for Early-Modern Japanese Printed Books	５．発行年 2018年
３．雑誌名 The 2018 International Conference on Parallel and Distributed Processing Techniques and Applications	６．最初と最後の頁 326-332
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 －

１．著者名 Yuki Takemoto;Yu Ishikawa;Masami Takata;Kazuki Joe	４．巻 1
２．論文標題 Structure of Neural Network Automatically Generating Fonts for Early-Modern Japanese Printed Books	５．発行年 2019年
３．雑誌名 The 2019 International Conference on Parallel and Distributed Processing Techniques and Applications	６．最初と最後の頁 182-188
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 －

１．著者名 Sayaka Iida, Yuki Takemoto, Yu Ishikawa, Masami Takata, Kazuki Joe	４．巻 1
２．論文標題 Layout Analysis using Semantic Segmentation for Imperial Meeting Minutes	５．発行年 2019年
３．雑誌名 The 2019 International Conference on Parallel and Distributed Processing Techniques and Applications	６．最初と最後の頁 135-141
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 －

１．著者名 Suzuka Yasunami, Norie Koiso, Yuki Takemoto, Yu Ishikawa, Masami Takata, Kazuki Joe	４．巻 1
２．論文標題 Applying CNNs to Early-Modern Printed Japanese Character Recognition	５．発行年 2019年
３．雑誌名 The 2019 International Conference on Parallel and Distributed Processing Techniques and Applications	６．最初と最後の頁 189-195
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 －

１．著者名 JOE Kazuki;Kaori Fujimoto;Yu Ishikawa;Masami Takata	４．巻 1
２．論文標題 Early-Modern Printed Character Recognition using Ensemble Learning	５．発行年 2017年
３．雑誌名 The 2017 International Conference on Parallel and Distributed Processing Techniques and Applications, Final Edition	６．最初と最後の頁 288-294
掲載論文のDOI（デジタルオブジェクト識別子） なし	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2件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 飯田 紗也香, 竹本 有紀, 石川 由羽, 高田 雅美, 城 和貴
２．発表標題 帝国議会会議録におけるレイアウト解析手法の比較
３．学会等名 情報処理学会数理モデル化と問題解決研究会
４．発表年 2018年

１．発表者名 林 英里香, 竹本 有紀, 石川 由羽, 高田 雅美, 城 和貴
２．発表標題 近代文語体と現代口語体の自動翻訳への試み
３．学会等名 情報処理学会数理モデル化と問題解決研究会
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

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
研究分担者	高田　雅美 (Takata Masami) (20397574)	奈良女子大学・生活環境科学系・講師 (14602)	

６．研究組織（つづき）

	氏名 （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	石川　由羽 (Ishikawa Yu) (20814370)	滋賀大学・データサイエンス教育研究センター・助教 (14201)	