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　 ６ 月 ２５ 日現在

機関番号：１４６０２
研究種目：基盤研究(B)（一般）
研究期間：2020～2022
課題番号：２０Ｈ０４４８３

研究課題名（和文）近代書籍からの知の再構築

研究課題名（英文）Reconstructing Knowledge from Early-Modern Books

研究代表者
城　和貴（Jo, 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）　　13,600,000円

研究成果の概要（和文）：近代書籍文字認識ではレイアウト解析において新聞等に見られる多段多見出し出版物に適した手法を提案し有効性を確認した。認識部分では学習データをクローリングで取り出す手法を実装し、人間が手作業で行うより数百倍早く収集できる環境を構築した。また、GANを利用して、特定の近代書籍出版者のデータにない文字種を人工的に作り出す手法を確立した。さらに認識エンジンとして、それまでのCNNから深層距離学習に変更することで、99%以上の認識率を確認し、近代書籍文字認識研究の完成を得た。近代文語体から現代口語体への機械翻訳では、学習データ対を6万文整備し、Transformerで十分な精度の翻訳が可能なことを示した。

研究成果の学術的意義や社会的意義
本研究成果は画像としてアーカイブ化された近代書籍のテキスト化を自動的に行えることを示したもので、テキスト化された近代文語体の文章を現代口語体に自動翻訳することで、近代書籍の知を再構成して利用することが可能となる。現在スタンフォード大学フーバー研究所でアーカイブ化が進められている邦字新聞（明治以降の日本人移民が現地で出版した日本語の新聞の総称）に本研究成果が利用される予定である。また、本研究の知見は令和6年度に公開される国会図書館のNDLOCR2で一部利用されており、NDLOCR2では近代書籍に対応した初めてのOCRとなる。

研究成果の概要（英文）：In early-modern printed character recognition, we proposed a method suitable for multi-column, multi-heading publications such as newspapers in layout analysis and confirmed its effectiveness. In the recognition part, we implemented a method to retrieve training data by crawling, and built an environment that can collect training data hundreds of times faster than human workers can do it manually. We also established a method to artificially create character types not found in the data of specific early-modern book publishers using GAN. Furthermore, by changing the recognition engine from CNN to deep metric learning, we confirmed a recognition rate of over 99%, thereby completing our research on early-modern printed character recognition. For neural translation from early-modern literary style to present colloquial style, we prepared 60,000 training data pairs and showed that the Transformer is capable of translating with sufficient accuracy.

研究分野：情報工学

キーワード：デジタルアーカイブ　文字認識　レイアウト解析　ディープラーニング　ニューロ翻訳

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景
本研究グループは本科研費（2020 年度～2022 年度）による研究を始めるまでに、近代書籍文字認識手法と近代文語体現代口語体相互自動翻訳手法、特定の近代書籍に特化したレイアウト解析手法、近代文語体と現代口語体に関する機械翻訳に関する基礎研究を行ってきた。当時近代書籍に対応した OCR は皆無であったため、本研究グループの研究成果は多くの注目を浴びていた。

２．研究の目的
本科研費による研究では、1）低出現頻度文字クローラを利用した近代書籍文字認識、2）GAN を用いた近代書籍文字の自動生成、３）複数のレイアウト解析技術をハイブリッドに融合した近代書籍用レイアウト解析、４）近代文語体と現代口語体の相互翻訳のためのデータセット作成とその適用、５）近代書籍文字認識エンジンの改善、の 5 サブテーマを目的とした。

３．研究の方法
サブテーマ１）では、2019 年度までに使えるようになっていた近代文字認識エンジンを利用して、国立国会図書館の近代書籍デジタルライブラリー内の近代書籍画像を自動的にクローリングし、予め指定していた文字種の画像を発掘する。
サブテーマ２）では、複数種類の GAN を利用して近代書籍文字画像を生成する。すなわち、限られたセットの特定の出版者の近代書籍文字画像を学習データとし、現在の明朝体等の文字画像を教師データとして GAN を学習させることで、学習データに存在しない文字種であって、その文字種を生成することが可能となる。
サブテーマ３）では、新聞等のように多段多見出しの書籍に対応した近代書籍用レイアウト解析を実現するために、既存の技術を複合利用することで文字抽出を行える手法を開発する。
サブテーマ４）では、スタンフォード大学フーバー研究所で整備を行っている邦字新聞デジタルアーカイブの画像データから、近代文語体のテキストと、その現代語訳を対として人力で作成する。なお、邦字新聞とは明治以降に世界中に散らばっていった邦人移民が現地で出版した新聞の総称のことである。
サブテーマ５）では、2019 年度に本研究グループが文字認識エンジンとして使っていた CNN よりも性能が良いと思われる深層距離学習を近代書籍文字認識に適用することである。

４．研究成果
サブテーマ１）では、2020 年夏に国際会議で研究成果を発表した。その後、さらに改良することにより、人手で低出現頻度文字を発掘し収集するより、200 倍以上高速に発掘収集が行えるようになった。
・Nanami Fujisaki, Yu Ishikawa, Masami Takata, Kazuki Joe: Crawling Low Appearance Frequency Characters Images for Early-Modern Japanese Printed Character Recognition, The 2020 International Conference on Parallel and Distributed Processing Techniques and Applications, pp.683-695 (2020).

サブテーマ２）では、StarGAN、fgNN、CycleGAN を利用して、JIS 第 2 水準までの文字種に相当する近代書籍風文字セット生成できることを示した。これらの研究成果は 2021 年度情報処理学会研究会、2022 年度情報処理学会論文誌で発表を行った。
・竹本有紀，石川由羽，高田雅美，城和貴：特定の近代書籍出版者における低出現頻度文字種の獲得方法，情報処理学会論文誌数理モデル化と応用，Vol.15(3)，pp71-89 (2022).
・角張 凜、飯田 紗也香、城 和貴: CycleGAN を用いた近代書籍風文字の生成とそのデータ拡張への応用，情報処理学会数理モデル化と問題解決研究会, 2021-MPS-136(15),1-6 (2021-12-13).
・倉田 帆風、藤﨑 菜々美、飯田 紗也香、高田 雅美、城 和貴: 近代書籍文字認識に有効なデータ拡張の一手法，情報処理学会数理モデル化と問題解決研究会，2021-MPS-136(10),1-6 (2021-12-13).

サブテーマ３）では、CRAFT と既存手法である解像度ピラミッドを複合して使うことにより、多段多見出しの近代書籍である帝国議会会議録の文字抽出がほぼ完ぺきにできることを示した。また、2022 年度に国会図書館が発表した近代書籍に対応した OCR である NDLOCR よりも遥かに性能が良いことを示した。この研究成果は 2022 年度内に情報処理学会論文誌に採択され、2023 年度に出版される予定である。
・飯田紗也香，竹本有紀，石川由羽，高田雅美，城和貴：多段組多サイズ見出しで構成される近代書籍のレイアウト解析，情報処理学会論文誌数理モデル化と応用，出版予定 (2023)

・飯田 紗也香 , 竹本 有紀 , 石川 由羽 , 髙田 雅美 , 城 和貴: 近代書籍における文字切り出し手法の検討, 情報処理学会数理モデル化と問題解決研究会, 2020-MPS-132(4),1-6 (2021-2-22)

サブテーマ４)では､邦字新聞デジタルアーカイブの画像データから､近代文語体のテキストと、その現代語訳を対として整備した。これは本学文学部学部生 15 人年を使って人手で整備し、6 万対の学習データを得ることとなった。この研究成果は 2022 年夏に国際会議で発表を行った。
・Honoka Nishikawa, Yuki Takemoto, Sayaka Iida, Yu Ishikawa, Masami Takata, Kaoru Ueda, Kazuki Joe: Translating Early-modern Written Style into Current Colloquial Style in Hoji Shinbun, The 2022 International Conference on Parallel and Distributed Processing Techniques and Applications, in press, (2022).
・藤井 千香子、竹本 有紀、石川 由羽、高田 雅美、城 和貴: 教師なし学習を用いた近代文語体と現代口語体の相互翻訳の検討, 情報処理学会数理モデル化と問題解決研究会, 2021-MPS-136(9),1-6 (2021-12-13).

サブテーマ５）では、それまで CNN を使って 96%程度の認識率だったものが深層距離学習を使うことで、一気に 98%に改善できることを示した。この研究成果は 2022 年夏に国際会議で発表を行った。
・Norie Koiso, Yuki Takemoto, Sayaka Iida, Yu Ishikawa, Masami Takata, Kazuki Joe: Application of Deep Metric Learning to Early-modern Japanese Printed Character Recognition, The 2022 International Conference on Parallel and Distributed Processing Techniques and Applications, in press, (2022).

５．主な発表論文等

〔雑誌論文〕　計6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 飯田紗也香，竹本有紀，石川由羽，髙田雅美，城和貴	４．巻 -
２．論文標題 多段組多サイズ見出しで構成される近代書籍のレイアウト解析	５．発行年 2023年
３．雑誌名 情報処理学会論文誌数理モデル化と応用	６．最初と最後の頁 -
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 1.Norie Koiso, Yuki Takemoto, Sayaka Iida, Yu Ishikawa, Masami Takata, Kazuki Joe	４．巻 -
２．論文標題 Application of Deep Metric Learning to Early-modern Japanese Printed Character Recognition	５．発行年 2023年
３．雑誌名 Proceedings of The 2022 International Conference on Parallel and Distributed Processing Techniques and Applications	６．最初と最後の頁 -
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 2.Honoka Nishikawa, Yuki Takemoto, Sayaka Iida, Yu Ishikawa, Masami Takata, Kaoru Ueda, Kazuki Joe	４．巻 -
２．論文標題 Translating Early-modern Written Style into Current Colloquial Style in Hoji Shinbun	５．発行年 2023年
３．雑誌名 Proceedings of The 2022 International Conference on Parallel and Distributed Processing Techniques and Applications	６．最初と最後の頁 -
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 竹本有紀，石川由羽，高田雅美，城和貴	４．巻 -
２．論文標題 特定の近代書籍出版者における低出現頻度文字種の獲得方法	５．発行年 2022年
３．雑誌名 情報処理学会論文誌数理モデル化と応用	６．最初と最後の頁 -
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 Nanami Fujisaki, Yu Ishikawa, Masami Takata, Kazuki Joe	4．巻 -
2．論文標題 Crawling Low Appearance Frequency Characters Images for Early-Modern Japanese Printed Character Recognition	5．発行年 2021年
3．雑誌名 Proceeding of 2020 PDPTA (in press)	6．最初と最後の頁 -
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 飯田 紗也香 , 竹本 有紀 , 石川 由羽 , 髙田 雅美 , 城 和貴	4．巻 2020-MPS-132(4)
2．論文標題 近代書籍における文字切り出し手法の検討	5．発行年 2021年
3．雑誌名 情報処理学科数理モデル化と問題解決研究会報告	6．最初と最後の頁 1-6
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 無
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5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
2．発表標題 近代書籍のためのCRAFTを用いたレイアウト解析手法
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 2022年

1．発表者名 藤井 千香子
2．発表標題 教師なし学習を用いた近代文語体と現代口語体の相互翻訳の検討
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 2021年

１．発表者名
倉田 帆風

２．発表標題
近代書籍文字認識に有効なデータ拡張の一手法

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
2021年

１．発表者名
角張 凜

２．発表標題
CycleGANを用いた近代書籍風文字の生成とそのデータ拡張への応用

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
2021年

１．発表者名
飯田 紗也香

２．発表標題
近代書籍における文字切り出し手法の検討

３．学会等名
情報処理学科数理モデル化と問題解決研究会報告

４．発表年
2021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 （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	石川　由羽 (ISHIKAWA YU) (20814370)	滋賀大学・データサイエンス学系・助教 (14201)	

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