「招待講演」
近代書籍からの知の抽出と再構成

城　和貴

奈良女子大学　〒630-8506 奈良市北魚屋西町

E-mail:　joe@ics.nara-wu.ac.jp

あらまし　　本講演では筆者が 2008 年度より行ってきた近代書籍のテキスト化と得られた近代書籍テキストの現代口語体への機械翻訳の手法について報告する．近代書籍とは，明治・大正・昭和初期に刊行された書籍で活版印刷による出版であるため，既存の印字文字認識手法や手書き文字認識手法をそのまま使ったのではテキスト化できないことが知られている．本講演では近代書籍文字認識の問題点とその解決法，さらにテキスト化された近代文語体を現代人が読むことのできる現代口語体に機械翻訳する手法について紹介する．

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，機械翻訳

Extracting and Reconstructing Knowledge from Early-Modern Books

Kazuki Joe

Nara Women's University, Kitauoyanishimachi, Nara city, 6308506, Japan

E-mail:　joe@ics.nara-wu.ac.jp

Abstract In this presentation, the author will report on the textualization of early-modern books and the method of machine translation of the early-modern book texts into current colloquialisms, which has been conducted by the author since 2008. Early-modern books are books published in the Meiji, Taisho, and early Showa periods and printed in letterpress, so it is known that existing printed or handwritten character recognition methods cannot be used for the text conversion. In this lecture, I will introduce the problems and solutions of early-modern book character recognition, as well as the method of machine translation of the converted early-modern literary texts into current colloquial texts that can be read by people of today.

Keyword Digital archiving, character recognition, layout analysis, machine translation

1. はじめに

我国における文字認識研究は 1967 年に東芝 TR-2 が世界初の手書き数字認識装置として発表されて以来，オフライン印字文字認識，オンライン手書き文字認識，オフライン手書き文字認識と続き，2000 年頃にはオフライン手書き文字認識研究までもが成熟したと思われていた．著者は 2008 年に国立国会図書館関西館（以下 NDL）で非常勤調査員を兼務することになったが，当時 NDL では近代デジタルライブラリの整備にリソースを多く使っていた．その際，利用の拡大が続いていた Google Scholar の全文検索が近代書籍に対してほぼ適用できないことに気がついた．そこで商用の OCR を使って近代書籍のテキスト化を試みたところ，ほぼ読めないという結論に至った．この事実は当時独自に随契で調査していた NDL も把握しており，著者は既に 20 世紀中に成熟したと思われていた印字文字，手書き文字認識に続いて，近代書籍文字認識研究が手つかずの分野として存在することを認識したのである．以下，本稿では 2 章で近代書籍文字認識の問題点と解決手法，3 章で解決手法を用いた邦字新聞 OCR の開発，4 章で近代文語体と現代口語体の自動翻訳に関して説明する．

2.近代書籍文字認識

2.1.問題点

印字・手書き文字認識は通常特徴抽出器をどのように構築するか，得られた特徴ベクトルをどのように識別するかの 2 点が基本的な問題点であった．20 世紀におけるこれらの研究には多種多様なものが考案され利用されていた．近代書籍文字に関しては，印字文字ではあるがフォントの規格等は存在しない活版印刷であったため，出版者ごとに独自のフォントを持つと考えられた．その出版者の数は NDL で確認できる限りでも 2 万者を超えている．つまり，2 万通りのフォントセットに対する文字認識ということになるが，近代書籍に

は現在の JIS 第 2 水準に分類される漢字も多いと予想されたため，近代書籍文字認識には少なくとも数百出版者の JIS 第 1 ならびに第 2 水準までの学習データが必要であると想定された．そこで筆者らは小規模なデータセットに既存手書き文字認識手法を適用した予備実験を行った結果[1]，ある程度の性能が出たため，近代デジタルライブラリから 262 種類 10 出版者のデータを手作業で集め，同様の手法で認識実験を行った結果，92.7%の認識率であった[2]．この認識率を上げるにはデータを集めれば良いと考え，少なくとも JIS 第 1 水準相当の文字種を 100 出版者程度集めようと試みた．ところが，数百種類程度までは順調に集まるが，1,000 種類くらいになるとそもそも見つからず，実験計画を根本から見直すことになった．この原因は漢字の出現頻度はジップの法則[1]に従う事実にあった．データセットの充実はすぐには解決できない問題であったため，取り敢えず人海戦術で集めた 2,678 種類×6 出版者のデータに対して，三種類の特徴抽出法と二種類の識別器を使ってアンサンブル学習をすることで 90%程度の認識率は確認できた[3]．ここで問題点として明確になったのは，JIS 第 2 水準までの約 6,000 種類の近代書籍文字フォントをどうやって集めるかであった．学習データ以外の問題では，近代書籍に多数見受けられるルビの除去があった．現在の出版物ではルビは規格化されているため容易に除去できるが，活版印刷でのルビは複雑を極め除去が困難であると思われた．この問題に関しては進化計算を利用することで早期に問題は解決した[4]．

2.2.解決手法

多数の出版者による JIS 第 1 および第 2 水準の近代書籍文字を収集する手法に関しては，実際の出版物から取得する方法と，そのような文字種フォントを予測して作り出す 2 種類の方法が考えられた．

前者に関しては[5]で提案はしたものの，実装するのに困難を極めたが，[6]で基本的なシステムの実装は行うことができた．このシステムでは，取り敢えず現状集まっているデータで近代書籍文字認識エンジンを作成し，近代デジタルライブラリの書籍データを自動的に読み込み，明らかに認識できない文字を表示して人間に文字種を指定してもらい，新たな学習データとして認識エンジンの性能を上げていく，というものであった．しかしながら，ハードウェア性能が十分ではないことから，実用的なデータ収集は困難であった．この手法は[7]で実用化に至り，近代書籍をクローリングして連続して自動的に読み込み認識エンジンで認識できない文字種の画像を次々に記録することで，近代書籍 1 ページあたり 1 分未満で低出現頻度文字種を探索することが可能となり，人間が手作業で行うより数千倍効率の良い収集システムが完成した．

後者に関しては，特定の近代書籍出版者の出版した書籍から学習用のデータを取り出し，明朝体等の現在のフォントを当該出版者の字体に自動変換させる手法を提案[8]したが，変換手法に進化計算を使っただけでは十分な変換は得られなかった．[9]ではディープラーニングを用いた変換用のニューラルネットを利用し，後に GAN を用いた手法と合わせて JIS 第 2 水準までの文字種なら任意の近代書籍出版者の文字種を作り出せることを示した．

2.3. より良い解決手法

本研究では文字認識手法に当初より古典的な特徴抽出と識別器を利用していたが，ディープラーニングの代表的なものとして知られる CNN を利用してみたところ，より良い性能を出すことが判明した[11]．さらに[11]では，学習データに少数の近代書籍文字と現在利用されている数十種類のフォントを使ったところ，近代書籍文字認識の認識率が大幅に向上した．[12]では同じ条件で学習手法に CNN ではなく深層距離学習を利用することで，認識率は実用上問題ないところまで達したことを報告した．

3.邦字新聞 OCR

3.1. 邦字新聞

邦字新聞とは，明治から昭和初期までの時代にアメリカ大陸，アジアにおいて日本人移民やその第二，第三世代により刊行された海外日系新聞である．現在，スタンフォード大学フーバー研究所により収集された邦字新聞画像データが，邦字新聞デジタル・コレクション[13]として公開されている．邦字新聞は様々な出版者により刊行され，当時の暮らしや政治を知るための極めて重要な資料であり，全文検索機能が強く求められているが，フーバー研究所で利用していた商用 OCR（漢字対応版）では全文検索はおろか文字認識さえもできていない．著者はフーバー研究所より直接協力を求められ，近代書籍文字認識研究の成果を利用することにした．

3.2. レイアウト解析

一般に書籍等の画像に文字認識を適用する場合，レイアウト解析を適用して文字を切り出すことが必須となる．通常の書籍であれば大きな問題なく行切り出し文字切り出しが行われるが，新聞に代表される多段組多サイズ文字種の書籍ではレイアウト解析の難易度が桁違いに上がる．近代デジタルライブラリにも数十年に渡る帝国議会議事録という多段組多サイズ文字種の記録書籍があり，これに対応したレイアウト解析の研

[1] 出現頻度が k 番目に大きい要素が全体に占める割合は 1/k に比例する現象

究を行ってきた．[14]ではセマンティックセグメンテーションを利用した手法を検討したが性能が十分ではなく，結局 CRAFT に解像度ピラミッドを組み合わせた手法で単体 PC で多段組多サイズフォントの書籍に対するレイアウト解析が十分な性能を出せることを示した[15].

3.3. 行単位のルビ除去

邦字新聞の本文内に使われる漢字のほとんどには，ルビが振られている．邦字新聞において漢字にルビが振られる理由の 1 つに，当時の識字率の低さが挙げられる．ルビの除去は文字認識研究における重要な位置づけを持っていたし，著者の研究においても早いうちに着手していた[4]．しかしながら，現在の OCR は下記に述べるように文字ごとの認識を行うのではなく，行ごとに行うのが主流となっている．

現在の OCR の文字認識手法には 1 行の文字列画像を系列データとして扱うシーケンス認識が利用されている．入力される行領域文字画像に対し，CNN によって文字列特徴量を得る．そして，双方向 LSTM により文字列画像列に対する時間ごとの文字種確率列を得る．最後に，最も確率の高い文字種から得られる予測文字列に対して CTC 関数による最適化が行われ，テキストが出力される．

ルビの除去は行ごとに行われ，不十分な切り取りや過剰な切り取りがあったとしても，文字列ごとの一括認識であればあまり問題にはならない．従って，[4]のように文字ごとの正確なルビ除去も必ずしも必要とはならない．しかしながら，邦字新聞においてはルビのある行とない行が同一記事の中で混在することが知られている．ルビのない行に対してルビ除去を行うと文字列ごとの認識に悪影響を与えてしまい，テキスト化が困難となる．そこで邦字新聞用の OCR を開発するために，フーリエ記述子を用いたルビ除去手法を開発した[16].

3.4. 邦字新聞 OCR

我々は NDLOCR[17]をベースに，3.2 の解像度ピラミッドと CRAFT を利用したレイアウト解析とフーリエ記述子を利用した行ごとのルビ除去，さらに同一記事が分散配置される問題を Okapi-BM25 を利用した手法で解決した実装を施し，邦字新聞 OCR を完成させ，スタンフォード大学フーバー研究所で設置運用を実現した[18].

4.近代文語体と現代口語体

近代書籍から手作業で文字画像切り出しを行っていたころ，作業をしてくれていた学生たちの不満は，何が書いてあるかよくわからない，であった．近代書籍の文体は江戸時代以前の古文程ではないが，確かに現代人に詠み辛いものである．つまり，近代書籍のテキスト化に成功しても，そのままでは現代人に利活用してもらいにくいということである．

4.1. 機械翻訳機の利用

近代文語体の文章を現代口語体に自動翻訳するのに機械翻訳技術を利用することは自然な流れであるが，その学習データをどのように確保するかという問題がある．異言語間の機械翻訳には数十万から数百万対の文章データが必要であると言われている．異言語も英語とドイツ語のように似ているものの場合と，英語と日本語のように全く似ていないものでは，必要となる翻訳対データの量がまったく異なる．近代文語体と現代口語体は，ある意味，非常に似通った言語であるため，機械翻訳を適用するのに数万程度の翻訳対データがあればかなりの翻訳が可能になると思われた．ところが，そのような翻訳対データをテキストとして入手することは難しかったので，明治時代の著名な作家の原著テキストと，その翻訳テキストを与えることで機械翻訳を試みた[19].

学習データとして森鴎外の「即興詩人」を選び，現代語訳は神西清のものを利用したが，現代語訳は意訳であるため，質の良い翻訳データ対とは言い難かった．機械翻訳は Seq2Seq と ConvSeq2Seq の二種類を使い，学習データは 1,880 の翻訳対で行ったが，あまり良好な結果は得られなかった．

また，翻訳対データがなくとも，任意の言語コーパスがあれば自動翻訳可能という研究を発見し，それを実装した[20]．その結果，精度は良くないものの，十分な量のコーパスがあれば，かなりなところまで機械翻訳可能なことが判明した．さらに，コーパスは同じ分野のコーパスであれば，翻訳精度は向上することも確認できた．

4.2. 邦字新聞からの学習データ作成

近代文語体と現代口語体の自動翻訳研究に着手したころ，前述スタンフォード大学フーバー研究所より邦字新聞デジタルアーカイブに関する協力要請を受けたわけだが，この邦字新聞を利用して翻訳データ対を作成することにした．当初は 1 名の学生が自力で翻訳対を数千作成し，[19]と同様の手法で実験を行った結果，データ数は不十分であるものの，データの拡張を行うことで近代文語体と現代口語体の自動翻訳が可能であることが見えてきた[21]．このデータセットを作成するために，本学文学部の学部生 10 名程度に翻訳バイトを行ってもらい，5 万翻訳対のデータを整備し，ほぼ満足のいく成果を得ることができた[22].

上記自動翻訳において，近代文語体の熟語で現代口語体に存在しないものは翻訳が難しい．翻訳対データを数十倍に増やせば対応可能かもしれないが，それよ

りは翻訳の前処理後処理で対応させた方が早いであろう．そのために必要な近代文語体コーパスは探した限りにおいては見つけることができなかった.そのため，近代文語体の誤字検出の仕組みだけは行った[23].

なお，本機械翻訳の学習で利用した翻訳データ対の作成は，コロナ禍であったからこそ可能であった[i]ことを申し添えておく.

5.おわりに

明治，大正，昭和初期に刊行された近代書籍は活版印刷が使われており，現代の印字文字認識では認識が難しい．さらに近代書籍の保存状態が劣悪であれば認識は更に悪くなる．20 世紀中に成熟した手書き文字・印字文字認識に対し，第 3 の文字認識として近代書籍文字認識が研究分野として発見された．筆者はこの分野を 15 年に渡って研究し,実用化に道筋をつけることができた．更に自動テキスト化された近代書籍の近代文語体は現代人には解読しづらいことから，現代口語体に機械翻訳させる道筋を示した.

これにより近い将来現代語で近代書籍の全文検索や，逆に近代文語体で現代の書籍等の全文検索も可能となるであろう．さらに，異なる言語間の自動翻訳も実用化レベルに達していることを考えれば，任意の言語間の翻訳（水平方向の翻訳）と，特定の言語の過去の言語体系での翻訳（垂直方向の翻訳）が自由自在に可能となる未来が見えてくる.

文　献

[1] Chisato Ishikawa, Naomi Ashida, Yurie Enomoto, Masami Takata, Tsukasa Kimesawa, Kazuki Joe: Recognition of Multi-Fonts Character in Early-Modern Printed Books, The 2009 International Conference on Parallel and Distributed Processing Techniques and Applications(PDPTA), Vol.2, pp.728-734 (2009).

[2] Manami Fukuo, Yurie Enomoto, Naoko Yoshii, Masami Takata, Tsukasa Kimesawa, Kazuki Joe: Evaluation of the SVM Based Multi-Fonts Kanji Character Recognition Method for Early-Modern Japanese Printed Books, PDPTA2011, Vol.2, pp.727-732 (2011).

[3] Kazumi Kosaka, Kaori Fujimoto, Yu Ishikawa, Masami Takata, Kazuki Joe: Comparison of Feature Extraction Methods for Early-Modern Japanese Printed Character Recognition, PDPTA2016, Final Edition, pp.408-414 (2016).

[4] Taeka Awazu, Manami Fukuo, Masami Takata, Kazuki Joe: A Multi-Fonts Kanji Character Recognition Method for Early-Modern Japanese Printed Books with Ruby Characters, 3rd International Conference on Pattern Recognition Applications and Methods, pp. 637-645 (2014).

[5] 黒田佳世，榎本友里枝，高田雅美，城和貴:近代デジタルライブラリーテキスト化支援のためのポータルサイトの設計，情報処理学会数理モデル化と問題解決研究会, MPS-81-35 (2010).

[6] Kazumi Kosaka, Taeka Awazu, Yu Ishikawa, Masami Takata, Kazuki Joe: An Effective and Interactive Training Data Collection Method for Early-Modern Japanese Printed Character Recognition, PDPTA2015, Vol.1, pp. 276-282 (2015).

[7] Nanami Fujisaki, Yu Ishikawa, Masami Takata, Kazuki Joe: Crawling Low Appearance Frequency Characters Images for Early-Modern Japanese Printed Character Recognition, PDPTA2020, pp.683-695 (2020).

[8] 岩田彩，上坂和美，粟津妙華，石川由羽，高田雅美，城和貴: 近代書籍用 OCR のための学習用特定フォントセットの自動生成手法，情報処理学会数理モデル化と問題解決研究会, 2015-MPS-105(10), 1-6 (2015).

[9] Yuki Takemoto, Yu Ishikawa, Masami Takata, Kazuki Joe: Automatic Font Generation for Early-Modern Japanese Printed Books, PDPTA2018, On-site Edition, pp.326-332 (2018).

[10] 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).

[11] Suzuka Yasunami, Norie Koiso, Yuki Takemoto, Yu Ishikawa, Masami Takata, Kazuki Joe: Applying CNNs to Early-Modern Printed Japanese Character Recognition, PDPTA2019, pp.189-195 (2019).

[12] Norie Koiso, Yuki Takemoto, Sayaka Iida, Yu Ishikawa, Masami Takata, Kazuki Joe: Application of Deep Metric Learning to Early-modern Japanese Printed Character Recognition, PDPTA2022, in press, (2022).

[13] 邦字新聞デジタル・コレクション: https://hojishinbun.hoover.org/?l=ja （参照 2023-11-11）

[14] Sayaka Iida, Yuki Takemoto, Yu Ishikawa, Masami Takata, Kazuki Joe: Layout Analysis using Semantic Segmentation for Imperial Meeting Minutes, PDPTA2019, pp.135-141 (2019).

[15] 飯田紗也香，竹本有紀，石川由羽，髙田雅美，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，Vol.16(2), pp-67-79 (2023).

[16] 熊谷もも，邦字新聞に対応した OCR システムの開発，奈良女子大学修士論文，(2024).

[17] NDLOCR: https://github.com/ndl-lab/ndlocr_cli（参照 2023-11-11）

[18] 熊谷 もも，邦字新聞 OCR の概要と設置，スタンフォード大学上田研究室セミナー，スタンフォード大学フーバー研究所，(2023.10).

[19] 林 英里香，竹本 有紀，石川 由羽，高田 雅美，城 和貴：近代文語体と現代口語体の自動翻訳への試み，情報処理学会数理モデル化と問題解決研究会，2018-MPS-121(18),1-6 (2018).

[20] 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).

[21] 稲見 郁乃，竹本 有紀，石川 由羽，高田 雅美，上田 薫，城 和貴: 邦字新聞における近代文語体と現代口語体の自動翻訳の検討，情報処理学会数理モデル化と問題解決研究会，2020-MPS-131(12),1-6 (2020).

[22] 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, PDPTA2022, in press, (2022).

[23] 福元 春奈, 竹本 有紀, 石川 由羽, 高田 雅美, 城 和貴: 近代書籍文字認識に対応した誤字検出, 情報処理学会数理モデル化と問題解決研究会, 2022-MPS-141(21),1-6 (2022).

[24] Kazuki Joe: Digital extraction of knowledge from early-modern books, Impact, Vol. 2021, Num. 3, pp. 89-91, Science Impact Ltd , (2021).

[i] コロナ禍で飲食店等のバイトができなくなった学生が多数生まれた.