日本語 BERT モデルによる近代文の誤り訂正

謝素春　松本章代
東北学院大学大学院 人間情報学研究科
s2195101@g.tohoku-gakuin.ac.jp

概要

近代の資料は重要な価値がある．現在文書の電子化には光学文字認識（OCR）がよく使われているが，既存の OCR モデルの識別では近代文書を正確に獲得することが困難なため，識別エラーを訂正する必要がある．現在，言語モデルを OCR の誤り訂正に運用する手法が用いられているが，公開されている日本語言語モデルは主に現代文データで学習したものであり，近代文に対しては性能をうまく発揮できない可能性が高い．そこで本研究は，近代文のデータを収集し，近代文言語モデルを構築する．また，近代文の誤り訂正データセットを構築し，モデルの誤り訂正性能を検証した．さらに比較実験を通じて，近代文と現代文モデルの間の転移性を検証した．

1 はじめに

近代，すなわち明治時代（1868）から昭和初期（1945）までの歴史は日本に対して大きな変化を与えた．近世から現代までのかけ橋として，当時の教育，経済，政治，および文化は現在の日本にも影響している．その当時の紙資料は今でも数多く保管され，重要な価値を持っている．古文書であれば，従来は人手によって手書きの文章を解読し，現代文に書き換えて出版していたが，現在では，紙の資料を画像化し，人工知能に基づく手法を用いて，自動的に文字列に変換する方法が用いられている．しかし残念ながら，現在古文書のテキスト化は手書き字やくずし文字で保存された江戸時代の古典籍を対象とした研究がメインとになっており[1, 2, 3]，活字体である近代の出版物に対しての研究はほとんど見当たらない．そこで，近代の文書をテキスト化する研究をおこないたいと考えた．

画像からテキストを変換するには光学文字認識 (Optical Character Recognition; OCR)を利用するのが一般的だが，文書の保存状況やフォントの違いと異

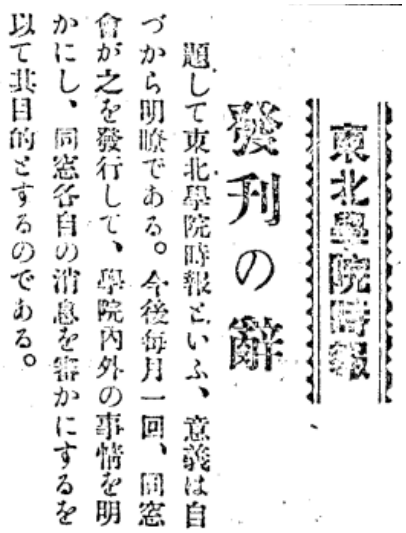
東北學院時報
發刊の辭
題して東北學院時報といふ、意義は自づから明瞭である。今後毎月一回、同窓會が之を發行して、學院内外の事情を明かにし、同窓各自の消息を審かにするを以て其目的とするのである。

図 1 近代文書の一例 1916 年の東北学院時報

体字の使用[4]，および利用した OCR などによって，識別の精度が大きく変わる一方，現代文のテキスト化と比べ，近代文のテキスト化精度はまだ高くない．正確な文章を得るためには，OCR の識別エラーに対して，校正を行う必要がある．

近代文書のテキストにおいて，従来は主に人手より文字入力および校正が行われているが，近年では OCR を用いての自動テキスト化と人手校正を組み合わせる手法を用いられた．しかし，文字が不鮮明のため脱字が発生する場合や，OCR の学習データ不足のため識別できない文字がある場合が散見される，その一例を図 1 に示す．それに対して，増田[4]は言語情報および字形情報を同時に考慮した OCR 訂正手法を提案した．しかし，近代文書における誤り検出の困難さにより，テキスト全体の訂正率にはあまり寄与しなかった．また近代文コーパスで学習した Kindai-OCR[i] [5]，NDLOCR[ii] を開発したが，近代資料で実際に検証した結果，識別精度はまだ改善する余地があることが示された．

近年 OCR の誤り訂正タスクにおいて，BERT[iii] モデルを用いる手法を見られる．BERT モデル[6]は，

[i] https://github.com/ducanh841988/Kindai-OCR
[ii] https://github.com/ndl-lab/ndlocr_cli
[iii] https://github.com/google-research/bert

Google 社が提出した事前学習済み言語モデルであり，ほかの言語モデルと比べ，文脈を考慮した単語の分散表現を獲得する特徴がある．BERT を用いた手法は分類タスク，質問応答などを含む様々なタスクにおいて優れた結果を示した[6].

誤り検出と誤り訂正のタスクにおいても，BERT が有効であることが確認されている．Kaneko ら[7]は，文法誤り訂正(GEC)タスクにおいて BERT を組み込むことで，優れた結果を得た．Yamakoshi ら[8]は BERT をベースにした分類器を使用し，日本語の法律用語の誤りの検出と訂正タスクにおいて，他のモデルを上回る性能が出せることを示した．近代文の識別誤りにおいても，事前学習済み BERT モデルが有効と予想される．

上記のタスクでは代表的な事前学習済みの汎用型日本語 BERT モデルである東北大 BERT[iv] と京大 BERT[v] [9]を利用した．ほかに NICT[vi] や Laboro 社[vii]によりいくつかの日本語 BERT モデルも公開されている．それらのモデルは主に日本語 Wikipedia をコーパスとして学習し，ほかに経済新聞やインターネット上のオープンデータから学習しているモデルもあるが，いずれも現代文を用いて訓練したモデルであり安易に近代文に適応しても良い結果は得られないと予想される．

近代文用の BERT としては青空文庫 (2020 年時点)で事前学習したモデル[viii] がある．ただし，青空文庫の多数は，新字新仮名で書かれたものである．その中に同じ文章を新字新仮名と旧字旧仮名２つのバージョンが同時に存在する場合もある．青空文庫 BERT はそれらのデータを現代文として学習に利用しているため，近代文に対する性能は高くないと予想される．

そこで，本研究は新たな近代文データセットを収集し，近代文用の BERT モデルを構築する．また近代文の誤り訂正データセットで微調整学習(fine-tuning)する．最後に，モデルの誤り訂正性能を検証する．さらに，近代文モデルと既存の現代文モデルをそれぞれ近代文と現代文の誤り訂正データセットを用いて比較実験を行う．

[iv] https://github.com/cl-tohoku/bert-japanese
[v] https://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese
[vi] https://alaginrc.nict.go.jp/nict-bert/
[vii] https://laboro.ai/activity/column/engineer/laboro-bert/
[viii] https://github.com/akirakubo/bert-japanese-aozora
[ix] https://www.ninjal.ac.jp/
[x] https://www.aozora.gr.jp/

2 近代文 BERT モデルの構築

現代文と比べて近代文は独自の仮名や漢字遣いが存在し，文法も異なる部分が多いため，近代日本語のテキスト化において誤字訂正は事前学習済みの近代 BERT モデルを使用することで性能が向上できると予想されている．しかし，現時点性能を確認された近代文用の BERT モデルが存在していない．そこで本研究では, 近代文用の BERT モデルを構築する．また，事前学習済みのモデルを用い，近代文の誤り訂正データセットで微調整学習し，近代文モデルの性能検証を行う．それに加え，現代文と近代文の誤り訂正データセットにおいて，現代文 BERT と近代文 BERT の比較実験をおこない，近代文用 BERT の有効性を調査する．

2.1 近代文コーパス

近代の紙資料は数多く存在するが，テキスト化されたものは少ないうえ，公開されているデータはさらに少なく，データの入手が困難である．本研究の事前学習用データは，2022 年 7 月 30 日時点の青空文庫のデータおよび国立国語研究所[ix]より公開された近代語コーパスより獲得した．

青空文庫[x] は著作権が切れた日本語文章が数多く収録されており，2022 年時点は 1000 名以上の作家の作品が収録されている．その中に明治期から昭和初期の文章が多数存在し，その一部は入力時に新字新仮名の文章に変換されている．旧字旧仮名の文字に統一するため，本研究は新字新仮名と旧字旧仮名の文章を分けて収集し処理を行った．

近代語コーパス[xi] は『太陽コーパス』，『近代女性雑誌コーパス』，『明六雑誌コーパス』，『国民之友コーパス』[xii] の 4 つのコーパスにより構成されている．明治前期から昭和戦前期まで（1945 年）の近代資料より選ばれた代表的な近代資料である[10]. これらのデータは文章入力後，人手による校正も行われており，近代日本語のモデル構築に適切なデータといえる．

近代語コーパスも青空文庫と同様に，文章の間は

[xi] https://clrd.ninjal.ac.jp/cmj/
[xii] 『近代女性雑誌コーパス』,『明六雑誌コーパス』,『国民之友コーパス』はクリエイティブ・コモンズ　表示 - 非営利 - 改変禁止 3.0 非移植 (CC BY-NC-ND 3.0) で公開されている。

1 つの空白行で区切り，文章の内容は 1 文 1 行に分割して処理する．正確に分割できない文に対しては人手で句読点を追加して文を分割した．上記 2 つのコーパスから作成した近代文データセットは合計 627MB，約 508 万文である．

2.2 モデルの事前学習

BERT Base モデル[6]を用いて近代文 BERT を構築した. 分かち書きありの設定とし, 近代文語 Unidic[xiii] [11]による MeCab[xiv] を使用した．事前学習の実装は東北大 BERT を利用し，100 万ステップの学習を実施した．

3 近代文識別率の事前調査

既存の OCR が近代文に対する識別率を調査するため, Tesseract OCR[xv] , EasyOCR[xvi] , PaddleOCR[xvii] , Kindai-OCR [5]，および Cloud Vision OCR[xviii]の 5 種類の OCR モデルを比較した. その結果, 識別率が比較的に高い Cloud Vision OCR を識別モデルとして採用した．

近代文書の識別率を客観的に評価するため，近代文章の画像で実験を行った．近代の出版物は様々あるが，東北学院時報[xix]および『文明論の概略』の一部を実験対象とした．東北学院時報は，1916 年に創刊され，2022 年現在まで 100 年以上発行を続けている．創刊号から現存するものはすべて画像として PDF 化され，ネットで公開されている．そのうち，1916 年の第 1 号から第 3 号の内容より 59 枚のキャプチャ画像を作成し,テキスト化の実験対象とする．また，『文明論の概略』は，1931 年に岩波書店から出版された．第 1 章より 61 枚のキャプチャ画像[xx]を作成し，同じくテキスト化の実験対象とする．実験対象から 120 文のキャプチャ画像および対応テキストを含むデータを人手より作成した(『文明論の概略』のテキストはブログ[xxi] 上のデータに基づいて作成した)．キャプチャ画像の選定基準としては，主に縦書きの 1 文から 2 文単位でキャプチャ画像を作成し，複雑なレイアウトを含む内容は今回選択していない．Cloud Vision OCR で検証した結果，近代文に対する平均識別率は 89%だった．さらに識別エラーに対して分析し，近代文の識別エラーは主に誤字，脱字であることを確認した．

4 評価実験

4.1 近代文誤り訂正データセット

事前調査により近代文の識別エラーは誤字と脱字が主たる要因となることがわかった．本研究ではまず誤字を中心に検証する．また，脱字は未来のタスクとする．近代文に対する誤り訂正能力を検証するには, 近代文の誤り訂正データセットが必要である．本研究では，近代文データに基づき，類似文字の書き換えにより近代文誤り訂正データセットを作成した．作成方法は下記の通りである．

1. 類似文字作成: フォントより文字を画像化し, 画像のバイナリデータを取得する. その後，文字画像をペアごとに類似度を計算し，類似度が高い top-6 を類似文字とする．フォントに含まれていない文字は，異体字リスト[xxii] より類似文字を生成する．
2. 近代語コーパスを文ごとに改行して正規化する．その後，長さ 10 以下また 200 以上の文を取り除く[12].
3. 文ごとに，近代文語 Unidic を用いた MeCab を利用して分かち書きする. その後, ランダムに単語を選択し，その中の 1 文字を該当する類似文字に変更する．
4. 変更された文字以外のテキストと組み合わせることで，正解文と誤字を含む文のペアを作成する．

4.2 近代文モデルの検証

近代文 BERT モデルの性能を検証するために，構築した近代文誤り訂正データセットで微調整学習したモデルを，評価データでの正解率を評価基準として性能を評価する．188832 ペアの近代文エラーデータセットから，学習データ，検証データおよび評価データをそれぞれ 165087, 11846, 11899 で設定した．

[xiii] https://csd.ninjal.ac.jp/lrc/index.php?UniDic
[xiv] https://taku910.github.io/mecab/
[xv] https://github.com/tesseract-ocr/tesseract
[xvi] https://github.com/JaidedAI/EasyOCR
[xvii] https://github.com/PaddlePaddle/PaddleOCR
[xviii] Cloud Vision OCR: https://tinyurl.com/2nc8vp3g
[xix] https://jihou.tohoku-gakuin.jp/
[xx] https://dl.ndl.go.jp/pid/1278790/1/1
[xxi] https://web.flet.keio.ac.jp/~ueda/bunmei-1.html
[xxii] 異体字リスト: https://tinyurl.com/2mdo9ymr

表 1 近代文 BERT の誤り訂正の正解率

Model	Corpus	Epoch	Acc
近代文 BERT	近代	5	0.39
近代文 BERT	近代	10	0.54
近代文 BERT	近代	15	0.56

表 2 近代文 BERT と東北大 BERT の比較実験

Model	Corpus	Train	Test	Epoch	Acc
近代	近代	165087	11899	15	0.56
東北大	近代	85673	4143	15	0.20
近代	現代	104199	2252	10	0.10
東北大	現代	143027	2252	5	0.76

学習時に学習率は 1e-5 で設定し，最大シーケンス長は 60，batch size は 32 で，それぞれ 5, 10, 15 エポックで学習し，評価データでの性能を検証する．評価時は，エラー箇所を指定せず，文中にあるすべてのトークン位置に対して語彙数での分類を実行する，尤度が最も高い予測（分類）結果を当該位置で選択されたトークンとする．提供された正解文と完全に一致する場合のみを正解とする．

4.3 比較実験設定

比較実験では，近代文データセットより学習したモデルと現代文データセットより学習したモデルに対し，現代文と近代文の誤り訂正データセットでの微調整学習により比較実験し，性能評価を行う．現代文用のモデルは，東北大の BERT base モデルとする．評価用のデータセットに関しては，近代文に対する性能検証は近代文誤り訂正データセットを用いるが，現代文に対する性能評価は現代文の誤り訂正データセットである京大の日本語 Wikipedia 入力誤りデータセット[xxiii] [12]を用いる．京大の誤りデータセットには複数のエラータイプのデータが含まれるが，一貫性を保持するため，誤字のエラータイプのみを用いる．

5 結果と考察

近代文の誤り訂正データセットでの検証結果を表 1 に示す．微調整学習の Epoch 数を増やすと，評価データでのエラー訂正の正解率が向上した，Epoch 数が 15 の時に正解率が最も高く，0.56 になった．それにより，事前学習した近代文 BERT モデルは誤り訂正のタスクにおいて有効であり，誤り訂正データセットでの微調整学習も有効であることがわかった．

比較実験の結果を表 2 に示す．近代文誤り訂正タスクにおいて，近代文 BERT は現代文 BERT より正解率が約 0.36 向上した．このことから，近代文 BERT は有意義であることがわかる．また，近代文 BERT および現代文 BERT は，それぞれ逆のデータセットによるタスクにおいて正解率が低くなった．このことから，現代文と近代文はある程度の共通点はあるが，近代あるいは現代文一方のデータのみで学習したモデルはもう 1 つの言語種類のタスクでは適切に機能しない可能性が示唆された．

また，近代文 BERT が近代文タスクで達成した正解率は現代文 BERT が現代文タスクで達成した正解率より低い．これは事前学習時の学習データ量の違いが要因と考えられる．東北大 BERT は 4GB, 約 9.3 億の単語量で学習したのに対して，近代文 BERT は 627MB で 1.4 億のデータで訓練した．事前学習量の差は 6.4 倍であるため，正解率の差は事前学習のデータ量による違いと考えられる．

6 まとめ

本研究は青空文庫および近代文コーパスを用い，近代文用の BERT 事前学習モデルを構築し，その性能を検証した．類似文字変換により生成された近代文の誤り訂正タスクによりその性能を検証した．その結果，近代文 BERT モデルは，現代文用の BERT モデルより高い性能を示した．また，比較実験より近代文または現代文のみで学習したモデルは，もう 1 つの言語のタスクには対応できないことを実験的に示した．

また，微調整学習した近代文用の BERT モデルを OCR の識別エラーで評価したが，性能は十分に高くなかった．今後の課題として，近代文の事前学習のデータ量を増やし，近代文の誤り訂正タスクおよび近代文の識別誤り訂正の実運用における性能を改善できるかを検証していく．

[xxiii] 日本語 Wikipedia 入力誤りデータセット: https://tinyurl.com/2gkurns6

謝辞

本研究は Google の TPU Research Cloud (TRC) program の TPU 支援を受けたものです.

参考文献

[1] **Dilbag Singh, C.V. Aravinda, Manjit Kaur, Meng Lin, Jyothi Shetty, Vikram Raju Reddicherla, and Heung-No Lee.** Dknet: Deep kuzushiji characters recognition network. In IEEE Access, Vol. 10, pp. 75872-75883, 2022.

[2] **Lamb, Alex, Tarin Clanuwat, and Asanobu Kitamoto.** KuroNet: Regularized residual U-Nets for end-to-end Kuzushiji character recognition. In SN Computer Science**,** Vol. 1, No. 177, pp. 1-15, 2020.

[3] **Anh Duc Le, Tarin Clanuwat, and Asanobu Kitamoto.** A Human-Inspired Recognition System for Pre-Modern Japanese Historical Documents. In IEEE Access, Vol. 7, pp. 84163-84169, 2019.

[4] **増田勝也**. 言語情報と字形情報を用いた近代書籍に対する OCR 誤り訂正. 人文科学とコンピュータ研究会 2016 論文集，pp. 57-62, 2016.

[5] **Anh Duc Le, Daichi Mochihashi, Katsuya Masuda, Hideki Mima, and Nam Tuan Ly**. Recognition of Japanese historical text lines by an attention-based encoder-decoder and text line generation. In Proceedings of the 5th International Workshop on Historical Document Imaging and Processing, pp. 37-41, 2019.

[6] **Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova**. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.1 (Long and Short Papers), pp. 4171-4186, 2019.

[7] **Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui.** Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4248-4254, 2020.

[8] **Takahiro Yamakoshi, Takahiro Komamizu, Yasuhiro Ogawa, and Katsuhiko Toyama.** Japanese Mistakable Legal Term Correction using Infrequency-aware BERT Classifier. In IEEE Big Data, pp.4342-4351, 2019.

[9] **柴田知秀，河原大輔，黒橋禎夫**. BERT による日本語構文解析の精度向上．言語処理学会 第 25 回年次大会，pp. 205-208, 2019.

[10] **近藤明日子，田中牧郎**.『明六雑誌コーパス』の仕様.『近代語コーパス設計のための文献言語研究成果報告書』, pp. 118-143, 2012.

[11] **小木曽智信，小町守，松本裕治**. 歴史的日本語資料を対象とした形態素解析. 自然言語処理, Vol. 20, pp.727-748, 2013.

[12] **田中佑，村脇有吾，河原大輔，黒橋禎夫**. 日本語 Wikipedia の編集履歴に基づく入力誤りデータセットと訂正システムの改良. 自然言語処理, Vol 28, pp. 995-1033, 2021.