LLM を利用した古典籍における文字認識誤り訂正

西尾響 [1] 古宮嘉那子 [1] 竹内綾乃 [2] 小木曽智信 [3]
[1] 東京農工大学 [2] 国文学研究資料館 [3] 国立国語研究所
s241236w@st.go.tuat.ac.jp, kkomiya@go.tuat.ac.jp,
takeuchi.ayano@nijl.ac.jp, togiso@ninjal.ac.jp

概要

本研究では，古典籍の光学的文字認識（Optical Character Recognition：OCR）結果に含まれる文字認識誤りの訂正を目的とし，誤り箇所の検出と検出された誤り箇所の修正から成る 2 段階モデルで古典籍の誤り訂正を行う. また、古典籍コーパスによる LLM の追加学習を行い性能の変化を検証する. 実験では 2 段階モデルの学習・評価を国文学研究資料館から提供を受けた OCR 結果と人手による文字起こしデータをタスクの学習に用い、LLM の追加学習には小学館コーパスを用いた. モデルは GPT-4o mini を使用し学習は全てファインチューニングにより行った. その結果，2 段階モデルは誤り修正だけを行う 1 段階モデルと比べ性能が向上し、誤り箇所の検出の効果を確認できた. しかし、古典籍による LLM の追加学習を行うと、誤り修正の性能が低下することが分かった.

1 はじめに

歴史的資料のデジタルアーカイブにおいて光学的文字認識（Optical Character Recognition：OCR）は不可欠だが，手書き文字のばらつきやくずし字・変体仮名等により誤りが発生し，検索・校訂・翻刻など後段の作業コストを大きく増加させる.

一方，現代語における文字認識誤り訂正の手法として誤り箇所推定と誤り訂正を組み合わせた 2 段階モデルが一定の効果を示している [1]. 2 段階モデルではまず誤り検出モデルを学習し、これを用いて誤り個所の推定を行う. 次に、入力テキストに誤り箇所を示すタグを付与することで訂正が必要な箇所を明確化し，推定された誤り箇所のみを誤り訂正モデルで訂正する．このアプローチにより，文脈を考慮しつつ不要な訂正を抑えた効率的な文字認識誤り訂正が可能となる. そこで, 本研究では 2 段階モデルを古典籍における文字認識誤り訂正の手法として用いる.

既存の LLM の日本語事前学習は現代日本語が中心であると考えられる．そこで本研究では 2 段階モデルによるアプローチに加え古典籍コーパスを用いた次単語予測タスクによる追加学習を行うことで古典籍に対する言語理解を促進し、更なる性能の向上を試みる.

2 関連研究

文字認識誤り訂正に関する研究は多い. 竹内ら [2] は日本語の文字エラー訂正の候補を文字トライグラムから作成して品詞 ngram と統計言語モデルを利用してそこから絞る方式を提唱した. Nagata ら [3] は統計言語モデルを利用して OCR の日本語のエラー訂正を行っている. Sakamoto ら [4] は漢字の DL 距離を編集距離として利用して漢字認識のエラーを訂正している. Nguyen ら [5] は機械学習データを利用しないベトナム語の OCR のエラーを訂正する手法を提唱した。山登り法を用いた文字編集距離を利用したエラー訂正手法である. 謝ら [6] は BERT を古典籍の文字認識エラー訂正に利用している. Clanuwat ら [9] は，くずし字 OCR の出力を入力とし，LLM で後処理する OCR text refiner を提案した．また，古典籍テキストで次トークン予測により基盤モデルを追加学習した上で refiner を学習している.

2 段階モデルに関する研究では、Schaefer ら [7] と Nguyen ら [8] が深層学習を用いた誤り検出と誤り訂正の 2 段階モデルを提案している. Schaefer ら [7] はまず Bi-LSTM を用いてエラー箇所を推定し、その後 LSTM による機械翻訳でエラー訂正を行う方式を提案している。鈴木ら [1] は日本語の現代語における文字認識誤り訂正の手法として LLM を用いた誤り検出と誤り訂正の 2 段階モデルを提案している. 本研究では, 現代日本語で事前学習された BERT や T5 で歴史的な日本語の語義曖昧性解消 [10] や翻訳 [11] を行った論文があることから、GPT モデルを用

いて古典籍の OCR の誤り訂正を行うことも可能であると想定し、本研究を行った.

3　LLM を利用した 2 段階モデル

本研究では誤り検出，誤り訂正からなる 2 段階モデルを利用する.[1)] 以下にそれぞれ構成について述べる.

3.1　誤り検出

第 1 段階では，OCR の結果から誤り箇所を推定する．GPT-4o mini をファインチューニングすることにより誤り検出モデルを作成し、これを利用した. OCR 結果のテキストを入力とし、誤り部分をエラータグで囲んだ OCR 結果のテキストを出力するよう学習した.

3.2　誤り訂正

第 2 段階では，指定された誤り箇所のみを訂正する．GPT-4o mini をファインチューニングすることにより誤り訂正モデルを作成し、これを利用した. 誤り訂正モデルの入力は誤りを含む書き起こしであり、出力は、正解の書きおこしである. この際、入力となる誤りを含んだテキストの誤り部分をエラータグで指定した. なお、推論には、第一段階の推定結果を用いて、誤り個所を指定する. これに対し、学習には、第一段階の推定結果を用いず、正解データから、実際の誤り個所を指定した. 誤り訂正モデルの入力形式は，誤り検出モデルにより誤りと判定されたトークンを error タグで囲んだタグ付きの文章と，タグなしの文章を<sep>トークンで連結した形式である. 出力は，誤り訂正後のテキストとする.

3.3　古典籍のコーパスによる追加学習

既存の LLM の日本語事前学習は現代日本語が中心であると考えられる．そこで、2 段階モデルの学習を行う前に古典籍コーパスを用いた次単語予測タスクによる学習をファインチューニングで行い，古典籍に対する言語理解を促進する.

1)　2 段階訂正モデルの提案については，現在ジャーナル論文として投稿中である.

4　データ

4.1　古典籍データ

本研究では、日本語の古典籍データを用いた実験を行った. 古典籍データとは、国書データベースで公開されている国文研所蔵資料の画像データを、NDL 古典籍 OCR vers.3 により OCR 処理したテキスト化データである. また、もともとの画像データから人手で書き起こししたデータを正解の書き起こしとして利用した. どちらも、国文学研究資料館が処理し、共同研究において提供されたデータである.

提供を受けたのは、源語秘訣、薬師通夜物語、東海道中詩などの計 126 作品である. 人手の書き起こしは二社が行っており、ある会社（A 社）が書き起こしたものが 82 作品、別の会社（B 社）が書き起こしたものが 54 作品ある. このうちの 10 作品には、ふたつの会社による書き起こしデータが存在した. 我々はこれらの人手書き起こしデータのうち、人手で対応箇所の確認が取れなかった 4 作品を除く 122 作品を対象に実験を行った. 二社の書き起こしがある作品については A 社の書き起こしを採用した.

また、追加学習に用いるコーパスには、国語研究所より提供を受けた小学館新編日本古典文学全集より抽出した古文単言語コーパス約 13 万文を用いた.

4.2　誤り例

表 1 に古典籍データ OCR 結果と正解データの例を示す. 上の例は「滑稽雌黄」の一節であり、下の例は「大井河行幸和歌考証」の一節である. 古典籍は歴史的な日本語で書かれており、現代語よりも漢字を多く用いている. また、使用されている漢字も現代では使用されないような難解なものを含む.

OCR	正解（人手転記）
滑稽雌黄	
消稽唯黄巻之一終同	滑稽雌黄巻之一終
大井河行幸和歌考証	
叟酒舎大為河行幸安歌考証	瓊廼舎大い河行幸安歌考證

表 1　作品『滑稽雌黄』および『大井河行幸和歌考証』から抽出した OCR 誤り例（黄色は誤り箇所とその対応箇所）

5　実験

本実験では，3 章で述べた手法について，4 章で述べたデータを用い有効性を検証する．モデル

は全工程において GPT-4o mini を利用した. また、GPT-4o-mini には、gpt-4o-mini-2024-07-18 を利用した.

5.1 誤り検出

学習には，文字認識結果と人手の書き起こしデータを比較して作成したデータセットを使用した．文字認識結果の誤り部分をエラータグで囲み、これを教師値としてファインチューニングを行った. 入力は OCR 結果であり、出力は入力のエラー部分をエラータグで囲んだものである. データセットは、(学習：テスト)=(4:1) の比率に設定して五分割交差検定を行った. 評価には，Accuracy，Precision，Recall，F1 スコアを使用した. トークンごとの誤り箇所検出性能を評価するため，各指標はサブトークンベースで計算を行った．これには RoBERTa のトークナイザーを利用してトークナイズを行った. 文字ベースではなく、サブトークンベースで評価を行ったのは、後段の処理がトークンベースの入力の LLM による誤り訂正であるためである. また、図 1 にファインチューニングに用いたプロンプトを示す.

次の文章は OCR の出力結果です。文字認識が誤っている部分にのみ<error>タグ</error>を付けてください.

注意:
・誤りのある語句だけを<error>タグ</error>で囲むこと。
・それ以外の補足説明、前置き、警告などは一切出力しないこと。
・タグ付きの文章だけを返してください。

図 1 誤り検出モデルのファインチューニングに用いたプロンプト

5.2 誤り訂正

学習には，誤り検出モデルの学習時に使用したデータを基に，誤り箇所を示す error タグを付与したデータを使用した．データセットは、誤り検出と同様のデータの分割方法で，問題が被らないように 5 分割し (学習：テスト)=(4:1) の比率に設定して五分割交差検定を行った. 具体的には，文字認識結果の中で誤っているトークンを error タグで囲み，そのタグ付き text とタグなしの元の text を<sep>トークンで結合したものを入力として与えた．出力としては，訂正後の正解データを使用した．一方，評価およびテスト時には，誤り検出モデルの予測結果に基づき，誤り箇所にエラータグを付与したデータを使用した．このデータも学習時と同様に，タグ付き text とタグなし text を<sep>トークンで結合した形式で入力した．評価には，BLEU スコア [12]，文字認識率（CRR: Character Recognition Rate），および単語認識率（WRR: Word Recognition Rate）を使用した．BLEU スコアの算出には，evaluate ライブラリの sacrebleu を用いた[2)]．文字認識率は文字誤り率 (CER: Character Error Rate) をもとに 1 – CER として計算し，単語認識率は，単語誤り率 (WER: Word Error Rate) を基に 1 – WER として求めた．これらの指標を用いることで，出力結果が正解データとどの程度一致しているかを定量的に評価した.

また、図 2 にファインチューニングに用いたプロンプトを示す.

次の文章は手書き文字の OCR 結果です。<error>タグで囲まれた箇所に OCR による誤りがあります。
<error>タグで囲まれた部分のみを訂正し，本来書かれていた通りの文字列に戻してください。
入力は，<error>タグ付きの文と，タグなしの文が<sep>で連結された形式です。

図 2 誤り訂正モデルのファインチューニングに用いたプロンプト

5.3 古典籍コーパスによる追加学習

学習には、小学館コーパスの古典籍テキストを Mecab[3)] で分かち書きしたデータを使用した. データは 1025 トークンごとに分割し、前 1024 トークンを入力、最後のトークンを出力として与えた.

6 結果

6.1 誤り検出の結果

誤り箇所推定では，Accuracy，Precision，Recall，F1 スコアの 4 つの指標で評価した．表 2 から結果が 95 を超える高い Accuracies, Precisions, Recalls, F1 であり、非常に高性能にエラー箇所を特定できることが分かった. 誤り検出モデルの性能が高ければ、高性能にエラー箇所を指定して次の誤り修正が行えるため、誤り検出モデルの性能は重要である.

	Accuracy	Precision	Recall	F1
GPT-4o-mini	95.90	95.95	95.82	95.87

表 2 古典籍テキストにおける誤り検出モデルの性能

2) https://huggingface.co/spaces/evaluate-metric/sacrebleu
3) https://taku910.github.io/mecab/

6.2 誤り訂正の結果

表 3 に古典籍データにおける誤り訂正性能（BLEU/CRR/WRR）の比較を示す. また、付録の表 7, 表 8 に 2 段階モデルでの誤り訂正により 誤り訂正のみを行った 1 段階モデルよりも BLEU が改善した例を示す. 誤り訂正のみを行った場合より誤り検出と誤り訂正を行った 2 段階モデルの方が、BLEU、CRR、WRR のすべての指標において性能が上回った. このことにより、2 段階モデルの古典籍に対する有効性が示された. なお、訂正前と比較して、2 段階モデルは、BLEU では 4.46 、CRR では 3.82 、WRR は 5.75 の上昇であった。

誤り検出	誤り訂正	BLEU	CRR	WRR
–	GPT-4o-mini	75.94	84.21	81.17
GPT-4o-mini	GPT-4o-mini	**78.53**	**86.82**	**84.00**
訂正前		74.07	83.00	78.25

表 3 古典籍データにおける誤り訂正性能（BLEU/CRR/WRR）の比較

6.3 2 段階モデル＋ 追加学習モデルの結果

小学館コーパスを用いて次トークン予測タスクによる追加学習を行ったモデルの誤り検出性能と誤り訂正性能について，それぞれ 6.1 節と 6.2 節と同一条件で評価した．表 4 と表 5 に、それぞれ 小学館コーパスによる追加学習後の古典籍データにおける誤り検出性能と誤り訂正性能（BLEU/CRR/WRR）の比較を示す. その結果，追加学習を行ったモデルでは，誤り検出性能は追加学習を行っていない場合と同程度であるが、誤り訂正性能（BLEU/CRR/WRR）が追加学習を行っていない場合より低下する傾向が確認された．このことから，古典籍コーパスによる追加学習が必ずしも誤り訂正に有利に働くとは限らないことが示唆される．

	Accuracy	Precision	Recall	F1
GPT-4o mini	95.90	95.92	95.90	95.91

表 4 小学館コーパスによる追加学習後の古典籍データにおける誤り検出性能

Setting	BLEU	CRR	WRR
2 段階 + 単語予測モデル	56.783	63.52	61.22

表 5 小学館コーパスによる追加学習後の古典籍データにおける誤り訂正性能（BLEU/CRR/WRR）の比較

7 考察

小学館コーパスによる追加学習を行った結果、誤り訂正において性能が低下した. この一因として表 6 に示した例のように出力が短くなる出力が多くなったことが挙げられる. これは追加学習における次単語予測学習時の, 一単語での出力が影響したと考えられる. この現象は追加学習がファインチューニングによって行われたことにより起こった可能性があり、追加学習をフルパラメータで行った場合は起こらない可能性がある.

項目	内容（長い箇所は省略）
input	<error>吉田求者事宜</error>といふべきなり。故に国天下のあるじたらん人は。常にその<error>関</error>意あり度ことなり〇…（中略）…されば<error>文上之写</error><sep>吉田求者事宜といふべきなり。故に…（中略）…されば文上之写
input_tagless	吉田求者事宜といふべきなり。故に国天下のあるじたらん人は。常にその関意あり度ことなり〇…（中略）…されば文上之写
prediction	文上之写
reference	といふべきなり。故に国天下のあるじたらん人は。常にその関意あり度ことなり〇…（中略）…かくのごとくまるで四五十年豊作つゞきて…（中略）…されば
BLEU	0.0
input_BLEU	93.073

表 6 小学館コーパスによる追加学習により性能が低下した例

8 おわりに

本研究では，LLM を用いた古典籍 OCR 誤り訂正の 2 段階モデル（誤り検出＋誤り訂正）を，古典籍データに適用した結果，誤り訂正のみのモデルより性能が向上することを確認した．また、小学館コーパスにより LLM の追加学習を行った場合の性能の低下を確認した. 今後は追加学習をフルパラメータでの学習が可能なオープンモデルに対して行うことで性能の向上を試みる．

謝辞

本研究は、科研費 JP23H03511、JP24H00738、および国文学研究資料館 令和 7 年度国文研プロジェクト型共同研究 「大規模言語モデルを用いた OCR 読み取り結果のエラー訂正と現代語への翻訳」、国立国語共同研究プロジェクト 「開かれた共同構築環境による通時コーパスの拡張」、公益財団法人 三菱財団 人文科学研究助成「自然言語処理を利用した古文解析」の助成を受けたものです。

参考文献

[1] 鈴木里菜，臼井久生，尾崎太亮，Nguyen Tuan Hung，古宮嘉那子，石岡恒憲，中川正樹．RoBERTa と T5 を用いた 2 段階モデルによる国語答案の文字認識誤り訂正．言語処理学会第 31 回年次大会 発表論文集，pp. 1051–1055，2025．

[2] 竹内孔一，松本裕治．統計的言語モデルを用いた OCR 誤り訂正システムの構築．情報処理学会論文誌，Vol. 40，No. 6，pp. 2679–2689，1999．

[3] Masaki Nagata. Japanese OCR Error Correction using Character Shape Similarity and Statistical Language Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pp. 922–928, 1998.

[4] 阪本浩太郎，阿部川明優，佐竹真樹，岸川至白，阪本エリーザ，石下円香，渋木英潔，森辰則．契約書 OCR の単語誤り訂正における漢字の偏旁冠脚を考慮した木編集距離の検討．言語処理学会第 26 回年次大会 発表論文集（Web），論文番号 P1-35，2020．

[5] Quoc-Dung Nguyen, Nguyet-Minh Phan, Pavel Krömer, Duc-Anh Le. An Efficient Unsupervised Approach for OCR Error Correction of Vietnamese OCR Text. *IEEE Access*, Vol. 11, pp. 58406–58421, 2023.

[6] 謝素春，松本章代．日本語 BERT モデルによる近代文の誤り訂正．言語処理学会第 29 回年次大会 発表論文集，論文番号 C7-4，2023．

[7] Robin Schaefer, Clemens Neudecker. A Two-Step Approach for Automatic OCR Post-Correction. In *Proceedings of LaTeCH-CLfL 2020*, pp. 52–57, 2020.

[8] Thi-Tuyet-Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickaël Coustaty, Antoine Doucet. Neural Machine Translation with BERT for Post-OCR Error Detection and Correction. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2020)*, pp. 333–336, 2020.

[9] Tarin Clanuwat, Tianyu Zhao, Yuki Imajuku, Asanobu Kitamoto. Beyond OCR: Enhancing Classical Japanese Transcription with Large Language Models. 人文科学とコンピュータシンポジウム（CH）2024, pp. 75–82, 2024.

[10] Soma Asada, Kanako Komiya, Masayuki Asahara. All-Words Word Sense Disambiguation for Historical Japanese. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation (PACLIC 37)*, pp. 201–209, Hong Kong, China, 2023. Association for Computational Linguistics.

[11] Hisao Usui, Kanako Komiya. Translation from Historical to Contemporary Japanese Using Japanese T5. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pp. 27–35, Tokyo, Japan, 2023. Association for Computational Linguistics.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

付録

表 7 は 2 段階モデルの誤り訂正モデルでの入出力の例を示しており、表 8 は表 7 と同じ文章を用いて 1 段階モデルの誤り訂正モデルでの入出力の例を示したものである. なお、表 7 の入力は誤り検出モデルで error タグを付与した文と，タグなしの文が<sep>トークンで連結された形式である.

項目	内容
input	<error>〇日コ三日一日の□□□□〇一一一三・一一〇三・四一〇〇〇〇の十一しく十二六〇・〇一〇一〇〇日の</error> 竹取翁物語<error>紺</error>四<sep> 〇日コ三日一日の□□□□〇一一一三・一一〇三・四一〇〇〇〇の十一しく十二六〇・〇一〇一〇〇日の竹取翁物語紺四
input_tagless	〇日コ三日一日の□□□□〇一一一三・一一〇三・四一〇〇〇〇の十一しく十二六〇・〇一〇一〇〇日の竹取翁物語紺四
prediction	竹取翁物語註四
reference	竹取翁物語解四
BLEU	54.752
input_BLEU	5.034

表 7　2 段階モデルでの誤り訂正の例

項目	内容
input	〇日コ三日一日の□□□□〇一一一三・一一〇三・四一〇〇〇〇の十一しく十二六〇・〇一〇一〇〇日の竹取翁物語紺四
prediction	のべ
reference	竹取翁物語解四
BLEU	0.0
input_BLEU	5.034

表 8　1 段階モデルでの誤り訂正の例