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

鈴木里菜 [1]　臼井久生 [1]　尾崎太亮 [1]　Nguyen Tuan Hung[1]　古宮嘉那子 [1]
石岡恒憲 [2]　中川正樹 [1]
[1] 東京農工大学大学院　[2] 大学入試センター 研究開発部
{s248289x, h-usui, hiroaki-ozaki}@st.go.tuat.ac.jp　{fx7296, kkomiya}@go.tuat.ac.jp
tunenori@rd.dnc.ac.jp　nakagawa@cc.tuat.ac.jp

概要

本研究では，手書き答案の文字認識誤り訂正を目的として，RoBERTa による誤り箇所推定と T5 による誤り訂正を組み合わせた 2 段階モデルを提案する．既存研究では，T5 を用いたモデルで答案全体を対象に訂正を行っていたため，必要のない箇所まで訂正される問題があった．これに対し，本手法では，RoBERTa を用いて誤り箇所を推定し，その結果に基づき T5 で該当箇所のみを訂正する．中学生 185 名による国語ドリルの記述式問題の手書き答案をデータとして実験を行った結果，提案手法は T5 単独モデルに比べ訂正精度が向上し，不要な訂正を抑える効果が確認された．

1　はじめに

手書き答案の文字認識は，自動採点システムの実現に欠かせない技術である．しかし，手書き文字認識には，筆跡や文脈の多様性，認識ソフトウェアの性能限界による誤りが避けられない課題として存在する．そのため，文字認識結果の誤り訂正は，正確な自動採点やそれに基づくフィードバックのために不可欠なプロセスとなっている．

これまで，文字認識誤り訂正の手法として，N-gram やルールベースの訂正手法，さらには BERT[1] を用いた文脈理解に基づくモデルが提案されてきた．さらに，生成モデルである T5[2] を用いる手法も試みられ，一定の効果を示している [3, 4]．過去研究 [3, 4] では，文字認識結果を入力に訂正結果を T5 に出力させるが，その場合，誤り箇所を特定せずに全体を対象に訂正を行うため，必要のない箇所まで訂正してしまうという課題があった．

本研究では，この課題に対処するため，誤り箇所を正確に特定し，その箇所だけを訂正する仕組みとして，誤り箇所推定と誤り訂正を組み合わせた 2 段階モデルを提案する．まず，RoBERTa[5] を用いて誤り箇所を推定し，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する．その後，T5 を用いて，推定された誤り箇所のみを訂正する．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．

2　関連研究

文字認識誤り訂正に関する研究では，竹内ら [6] や Nagata ら [7] が統計言語モデルや文字 N-gram を用いたエラー訂正手法を提案している．また，Sakamoto ら [8] や Nguyen ら [9] は，編集距離や山登り法を活用した OCR エラー訂正に取り組んだ．

大規模言語モデルを活用した手法としては，謝ら [10] が BERT を日本語古文の文字認識誤り訂正に応用しており，文脈情報を活用するアプローチを提案している．また，中村ら [11] や藤武ら [12] は，T5 を用いて音声認識や OCR 精度向上の課題に対応している．いずれも，T5 がエラー訂正において有効であることを示している．

推定＋書き換えの 2 段階モデルにおいては，Schaefer ら [13] が Bi-LSTM を用いてエラー箇所を推定し，その結果をエンコーダデコーダ型モデルに入力して訂正する手法を提案している．また，Nguyen ら [14] は，BERT でエラー箇所を推定した後，文字ベースのニューラル機械翻訳を適用するアプローチを提案している．Nguyen らの研究では，エラー箇所を含むトークンとその前後 2 トークンを文字レベルで表現した入力シーケンスを使用して訂正を行い，文全体を対象とするのではなく局所的な訂正を行う構造を採用している．

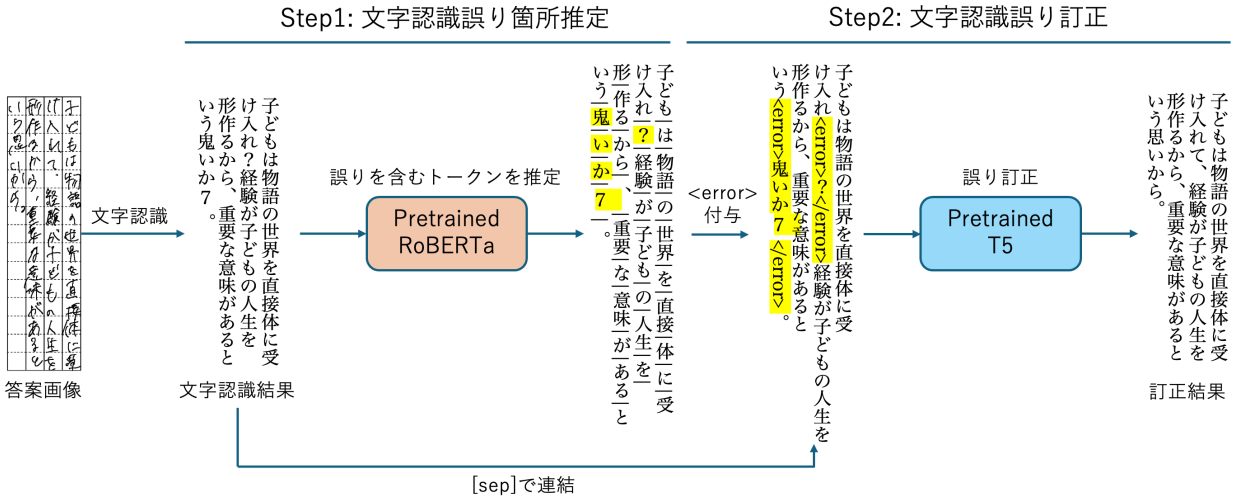

図 1　RoBERTa と T5 を用いた文字認識誤り訂正 2 段階モデルの概要

本研究では，RoBERTa による誤り箇所推定と T5 による柔軟な書き換えを組み合わせた 2 段階モデルを提案する．提案手法では，文全体を入力とし，誤り箇所を error タグで明示的に指定することで，文脈を広範囲に活用可能な構造を実現している．このように，大規模言語モデルの能力を活かしながら，日本語の文字種の多様性に対応し，高精度な文字認識誤り訂正を目指す．

3　RoBERTa と T5 を用いた文字認識誤り訂正 2 段階モデル

本研究では，文字認識誤り訂正を目的として，RoBERTa と T5 を組み合わせた 2 段階モデルを構築した．この提案手法の概要を図 1 に示す．

誤り箇所推定　第 1 段階では，日本語で事前学習した RoBERTa を用いて文字認識結果から誤り箇所を推定した．具体的には，入力された文字列を RoBERTa を用いてトークン単位で分類し，各トークンが誤りを含むか否かを系列ラベリングタスクとして判定した．これにより，誤りの有無を示す二値のラベルを出力する．

誤り訂正　第 2 段階では，日本語で事前学習された T5 を用いて，推定された誤り箇所のみを訂正した．ここでの入力形式は，誤り推定モデルにより誤りと判定されたトークンに error タグを付与したタグ付きの文と，タグなしの原文を [SEP] トークンで連結した形式として，出力は，誤り訂正後の文字列とする．

4　データ

答案データは，T5 を用いた既存研究 [3, 4] と同様に，受験研究社出版の「10 分間復習ドリル 国語読解」の回答を使用した．3 冊のドリルを使用し，それぞれに対して中学生 65 名，57 名，66 名が回答している．また，選定した問題は記述式問題 25 問で，全ての答案は 60 字以内で記述されている．

答案の画像データは，アイラボ株式会社の縦書き文字認識システムを用いてテキスト化した．この文字認識システムは，各答案に対し最大 5 件の候補 (5-best) を出力する．これにより，文字認識結果として合計 6,656 件のデータが得られた．この自動認識された文字認識データを訂正対象として使用し，訂正後の正解データには，答案画像を人手で文字起こししたデータを用いた．

5　実験

本実験では，3 章で述べた手法について，国語ドリルの答案をデータとして用い有効性を検証する．誤り箇所推定と誤り訂正の両タスクに対して評価を行った．

5.1　文字認識誤り箇所推定

実験設定　RoBERTa モデルとして，日本語で事前学習されたモデルである rinna/japanese-roberta-base[1] を使用した．学習には，文字認識結果と人手の文字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．具体的には，文字認識結果の各トークンに対し，人手の文字起こしと一致しない場合には「1」，一致する場合には「0」のラベルを付与し，これを fine-tuning に用いた．

データセットは問題が被らないように 5 分割し，(学習：評価：テスト)=(3:1:1) の比率に設定して 5 分

1)　https://huggingface.co/rinna/japanese-roberta-base

割交差検証を行った．学習時の学習率は 3e-5 に設定し，バッチサイズは 8 を使用した．学習は 3 エポック実行し，重み減衰率には 0.01 を採用した．これらのハイパーパラメータは，事前実験でのパフォーマンスを基に設定した．また，トークン化には AutoTokenizer を使用した．

評価指標　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，各指標はラベルごとに計算を行った．

5.2　文字認識誤り訂正

実験設定　T5 モデルとして，日本語で事前学習されたモデルである retrieva-jp/t5-base-medium[2] を使用した．学習には，誤り箇所推定の学習時に使用した誤り箇所ラベルデータを基に，誤り箇所を示す error タグを付与したデータを使用した．具体的には，文字認識結果の中で誤り箇所ラベルが付与されたトークンを error タグで囲み，そのタグ付き文字列とタグなしの元の文字列を [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，誤り箇所推定モデルの予測結果に基づき，誤り箇所に error タグを付与したデータを使用した．このデータも同様に，タグ付き文字列とタグなし文字列を [sep] トークンで結合した形式で T5 モデルに入力した．

データセットは誤り箇所推定と同様に，問題が被らないように 5 分割し，(学習：評価：テスト)=(3:1:1) の比率に設定した 5 分割交差検証を行った．学習時のパラメータは，学習率を 1e-4，バッチサイズを 16，エポック数を 50 と設定した．これらの値は事前実験での性能を基に決定した．また，生成時のパラメータとしては，繰り返しの制御のため repetition penalty を 10.0 に設定した．

評価指標　誤り訂正の評価には，BLEU スコア [15]，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[3] を用いた．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．

2) https://huggingface.co/retrieva-jp/t5-base-medium
3) https://huggingface.co/spaces/evaluate-metric/sacrebleu

6　実験結果

本章では，5 章で実施した 2 段階モデルの性能評価結果について，誤り箇所推定と誤り訂正の各タスクごとの結果を示す．

6.1　誤り箇所推定の結果

表 1　RoBERTa による誤り箇所推定の評価結果

	Accuracy	Precision	Recall	F1 Score
スコア	0.9148	0.9140	0.9140	0.9140

RoBERTa を用いた誤り箇所推定では，Accuracy，Precision，Recall，F1 スコアの 4 つの指標で評価した．表 1 に示すように，Accuracy は 0.9148 で最も高く，F1 スコアも 0.9140 と高い値を示した．Precision と Recall はそれぞれ 0.9140 でバランスの取れた性能を示した．このことから，誤り箇所推定モデルは，文脈情報のみから誤り箇所を十分推定できることが確認できた．

6.2　誤り訂正の結果

表 2　T5 による誤り訂正の評価結果

	訂正前			訂正後		
手法	**BLEU**	**CRR**	**WRR**	**BLEU**	**CRR**	**WRR**
T5	48.56	74.68	51.18	52.70	67.91	59.73
RoBERTa+T5				**54.11**	**69.73**	**62.91**

表 2 は，T5 単独モデルを用いた既存手法および提案手法である RoBERTa+T5 の 2 段階モデルを用いた誤り訂正の評価結果を示している．T5 単独モデルの評価結果は既存研究 [4] から引用したものである．

RoBERTa+T5 の 2 段階モデルでは，BLEU スコアが訂正前の 48.56 から 54.11 に向上し，T5 単独モデルの 52.70 を上回った．また，CRR は訂正前の 74.68 から 69.73 に低下したものの，T5 単独モデルの 67.91 よりは上回る結果となった．さらに，WRR は訂正前の 51.18 から 62.91 に向上し，T5 単独モデルの 59.73 を上回った．

OCR システムは一般的に CRR の最大化を目指して設計されているため，訂正プロセスで文脈に基づく修正が行われると CRR が低下することは想定される．一方で，BLEU スコアおよび WRR は向上しており，自動採点における文字認識誤り訂正では，これらの指標が特に重要とされる．これは，文字単

表 3　既存手法と提案手法による誤り訂正例

手法	モデルの入力	訂正後	正しい文字起こし
例 1			
T5	人が畑でつくる野菜はみんな色も形もかわいらしく・大事にしてぁ**げたい**	人が畑でつくる野菜はみんな**色が**かわいらしく、大事にして**あげたい**	人が畑でつくる野菜はみんな色も形もかわいらしく、大事にして**あげたい**
RoBERTa+T5	人が畑でつくる野菜はみんな色も形もかわいらしく<error>・</error >大事にして<error>ぁ**げたい**</error>[sep] 人が畑でつくる野菜はみんな色も形もかわいらしく・大事にしてぁげたい	人が畑でつくる野菜はみんな色も形もかわいらしく、大事にして**あげたい**	人が畑でつくる野菜はみんな色も形もかわいらしく、大事にして**あげたい**
例 2			
T5	**三**すごく好きで、**越えよう**としてもっ湧いて出てきてしまうから。	**三**すごく好きで、湧いて出てきてしまうから。	すごく好きで、**仰えよう**としても湧いて出てきてしまうから。
RoBERTa+T5	**三**すごく好きで、**越えよう**としても<error>っ</error>湧いて出てきてしまうから。[sep] 三すごく好きで、越えようとしてもっ湧いて出てきてしまうから。	**三**すごく好きで、**越えよう**としても、湧いて出てきてしまうから。	すごく好きで、**仰えよう**としても湧いて出てきてしまうから。

位の正確性を示す CRR よりも，文全体の整合性や単語レベルの正確性が採点結果に直接影響を及ぼすためである．

これらの結果から，提案手法である 2 段階モデルは，既存手法に比べて優れた訂正精度を達成し，文字認識誤り訂正における有効性が示された．

7　考察

提案手法である RoBERTa+T5 の 2 段階モデルでは，BLEU スコアおよび WRR が訂正前と比較して向上した一方で，CRR は低下した．この要因として，主に訂正プロセスにおける文脈変更が影響を及ぼした可能性がある．特に，T5 による訂正が過剰修正となり，本来正しい文字列を誤りとして訂正してしまうケースが考えられる．誤り箇所推定モデルによる推定ミスも一因として挙げられる．

表 3 に示した例 1 では，提案手法が「かわいらしく・大事にしてぁげたい」を正しい表現「かわいらしく、大事にしてあげたい」に訂正した一方，既存手法は「色がかわいらしく」と不適切な訂正を行った．提案手法は error タグで誤り箇所を明示し，生成モデルに訂正箇所を限定的に指示することで，文脈を考慮した訂正が可能となった．それに対して，既存手法は文全体を対象とするため，不要な訂正が生じやすいことがわかる．一方で，誤り箇所推定のミスが訂正結果に影響を与える例も見られた．表 3 の例 2 では，RoBERTa による誤り箇所推定で「三すごく」や「越えよう」の部分が誤りと認識されず，訂正後もそのまま残ってしまった．誤り箇所が漏れることで，生成モデルがその箇所を訂正対象として認識できず，結果的に訂正精度が低下するということが，提案手法の課題として挙げられる．

今後の精度向上に向けた取り組みとして，データ拡張や prefix の追加が有効であると考えられる．既存研究 [3, 4] では，データ拡張による学習データの多様性向上が誤り訂正モデルの性能向上に寄与することが示されている．また，入力時に prefix として訂正に有効な情報を付与する手法は，モデルの文脈理解を助け，誤り箇所推定および訂正の精度を向上させる可能性がある．例えば，文長や誤りの特徴など，訂正に有効な情報を prefix として追加することで，さらなる性能向上が見込まれる．これらの手法を提案手法に取り入れることで，さらなる性能向上が期待される．

8　おわりに

本研究では，手書き答案の文字認識誤り訂正を目的として，RoBERTa による誤り箇所推定と T5 による誤り訂正を組み合わせた 2 段階モデルを提案した．第 1 段階では，日本語で事前学習された RoBERTa を用いて，文字認識結果から誤り箇所を特定するモデルを構築した．第 2 段階では，生成モデルである T5 を用い，誤り箇所を文脈に基づいて訂正する手法を考案した．提案手法の有効性を検証するため，中学生の国語ドリル答案を用いて実験を行った．実験の結果，提案手法は T5 単独のモデルに比べて性能が向上し，誤り訂正の精度を高めることができた．また，不必要な訂正を抑制することで，より正確な訂正結果が得られることを示した．

今後の課題としては，誤り箇所推定モデルと生成モデルの精度向上，異なるデータセットへの適用可能性の検討，データ拡張手法を導入することで，学習データの多様性を高める取り組みなどが挙げられる．

謝辞

本研究は JSPS 科研費 JP22K12145，JP23K28201 JP24H00738 の助成を受けたものである．答案収集は，本学における人を対象とする研究に関する倫理審査委員会の承認を得て実施した (No.230402-0411)．また，本研究において答案データを提供くださったワコム株式会社，および文字認識データを提供くださったアイラボ株式会社に深く感謝申し上げる．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. **NAACL-HLT2019**, p. 4171–4186, 2019.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **J. Mach. Learn. Res.**, Vol. 21, No. 140, pp. 1–67, 2020.

[3] 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. T5 を用いた日本語記述式答案の文字認識誤り訂正. 言語処理学会 第 30 回年次大会 発表論文集, pp. 1148–1153, 2024.

[4] Rina Suzuki, Hisao Usui, Hiroaki Ozaki, Hung Tuan Nguyen, Kanako Komiya, Tsunenori Ishioka, and Masaki Nakagawa. Error Correction of Japanese Character-Recognition in Answers to Writing-Type Questions Using T5. **Document Analysis Systems**, pp. 229–243, 2024.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. **arXiv:1810.04805 [cs.CL]**, 2019.

[6] 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.

[7] Masaaki Nagata. Japanese OCR error correction using character shape similarity and statistical language model. In **COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics**, 1998.

[8] 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. **The Association for Natural Language Processing**, pp. 137–140, 2020.

[9] Q.D. Nguyen, N.M. Phan, P. Krömer, and D.A. Le. An efficient unsupervised approach for OCR error correction of vietnamese OCR text. **IEEE Access 11**, pp. 58406–58421, 2023.

[10] 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, pp. 1616–1620, 2023.

[11] 中村朝陽, 李聖民, 田村鴻希, 吉永直樹. 前後の発話を文脈として考慮するニューラル音声認識誤り訂正. 情報処理学会, pp. 1–7, 2022.

[12] 藤武将人. 証憑を用いた日本語 OCR 誤り訂正ベンチマークの構築. 言語処理学会 第 30 回年次大会 発表論文集, pp. 2373–2377, 2024.

[13] Robin Schaefer and Clemens Neudecker. A Two-step Approach for Automatic OCR Post-Correction. **Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage**, pp. 52–57, 2020.

[14] Thi Tuyet Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickael Coustaty, and Antoine Douset. Neural machine translation with bert for post-ocr error detection and correction. **Proceedings of the ACM/IEEE Joint Conference on Digital Libraries**, 2020.

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