T5 を用いた日本語記述式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　[2] 大学入試センター 研究開発部
{s208649t，h-usui，hiroaki-ozaki}@st.go.tuat.ac.jp　fx7297@go.tuat.ac.jp
kkomiya@go.tuat.ac.jp　tunenori@rd.dnc.ac.jp　nakagawa@cc.tuat.ac.jp

概要

本研究では，事前学習済みの日本語 T5 モデルを用いて，手書き答案の文字認識結果の誤り訂正を行う．文字認識の誤り訂正には，これまで N-gram を用いた手法や BERT を用いた手法が提案されてきた．本論文では，自動で手書き答案を読み取った文字認識結果に対して人手で訂正を行い，これらのペアを用いて T5 を fine-tuning することで，文字認識誤りの訂正モデルを作成する．今回用いた答案データは，中学生約 50 名によって回答された，国語ドリルの記述式問題の日本語手書き答案である．さらに，誤認識のパターンをより多く学習させるため，学習データのデータ拡張を行った．実験の結果，T5 を fine-tuning することで，文字認識結果をより実際の答案に近い文章に訂正できることを示した．また，データ拡張の結果，訂正モデルの BLEU 値は向上した．

1　はじめに

教育業界では，2020 年に学習指導要領が改訂され，長い文章や資料を読み解き，自分の考えを記述させる問題が増加傾向にある．しかし，記述式問題は，採点者間の採点基準の差や，採点者の労力など，様々な問題がある．そのため，採点業務の効率化や教育サービスのレベル向上へ向けた取り組みとして，採点の自動化が進んでいる．

自動採点は，大きく 2 つのステップに分けられ，1 つ目は手書き答案を画像モデルを用いて文字認識する処理，2 つ目は文字認識結果に基づいて答案を採点する処理である．1 つ目の処理である手書き答案の文字認識の精度は，採点システムの性能に直結する重要な要素である．しかし，人間の手書き文字を正確に文字認識することは困難であり，正しく採点するためには，採点の前に文字認識結果に対して校正を行う必要がある．特に，理解可能な文として文字認識ができなければ，採点は難しい．言語モデルによる文字認識誤り訂正は，単語を答えさせる問題の場合，本来誤っていた回答を正答に訂正してしまう可能性があるため単純な適用は難しい．しかし記述式の答案では，このような可能性はかなり低いと予想できる．

そこで本研究では，日本語の記述式答案に対し，意味を考慮した文字認識結果の訂正手法として，事前学習済みの日本語 T5 モデルを用いた手法について，その効果を検証する．T5 モデル [1] は，Google 社が発表した大規模言語モデルであり，多くの自然言語処理ベンチマークで最高値を記録した．同じく Google によって開発された BERT[2] がエンコーダモデルであるのに対し，T5 はエンコーダデコーダモデルであるため，文字認識誤り訂正を変換モデルとして実現することが可能である．本研究では，この T5 を，手書き答案を読み取った文字認識結果とそれを人手で訂正したデータのペアを用いて fine-tuning することで，文字認識結果の訂正タスクに適用させる．

2　関連研究

日本語の文字認識誤りの訂正手法は多く提案されている．竹内ら [3] は，文字 trigram による文字誤り候補の生成と品詞 N-gram による候補の選択を組み合わせた手法を提案している．また，漢字を対象とした手法として，阪本ら [4] は漢字の部首認識誤りを訂正するための編集距離として漢字 DL 距離を提案し，そこから算出される類似度を用いた訂正手法を提案した．大規模言語モデルを用いた手法としては，謝ら [5] によって提案された日本語 BERT モデルを用いた手法がある．謝らは，近代語辞書を用いて事前学習した BERT モデルを，近代文誤り訂正データセットで fine-tuning することで，近代文の文

字認識誤り訂正を行った．

T5 を用いた誤り訂正の例としては，中村ら [6] による音声認識誤り訂正や，相馬ら [7] によるプログラムコードの誤り訂正，Shahgir ら [8] によるバングラ語の文法誤り訂正などがある．どの手法も，誤りのあるデータと誤りのないデータを学習データとして T5 を fine-tuning し，誤り訂正モデルを構築している．どちらも訂正への効果があることが示されているため，記述式答案の文字認識誤り訂正においても T5 モデルが有効であると予想される．

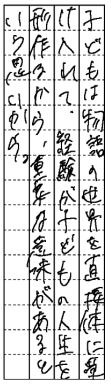
3　T5 を用いた文字認識誤り訂正

本研究では，事前学習済み日本語 T5 モデルである retrieva-jp/t5-base-medium[1] を用いて，文字認識誤り訂正を行った．T5 はテキストを入力としてテキストを出力するエンコーダデコーダモデルであり，本研究では，文字認識誤りを含むテキストを入力して，誤りを修正したテキストを出力する，End2End の誤り訂正モデルを作成することを目的とする．T5 モデルを誤り訂正タスクに適用させるため，記述式答案の文字認識データを入力，答案画像を人手により文字起こししたテキストデータを出力としたデータセットで fine-tuning を行う．

4　データ

答案データとして，中学生約 50 名によって回答された，受験研究社出版「10 分間復習ドリル　国語読解」の答案画像データから，記述式答案 25 問を抜粋し，使用した．答案はすべて 60 字以内で回答されている．これらの答案画像データをアイラボ株式会社の手書き文字認識エンジンを用いてテキスト化したものをシステムの入力とする．システムの出力として利用する正解には，答案画像データを見ながら答案を人手でテキスト化したものを用いた．図 1 に答案画像とその文字認識結果，人手の文字起こしデータの例を示す．

手書き文字認識エンジンは 2 種類使用しており，縦書き文字認識モデルによる答案の認識結果の 5-best と，単文字認識モデルによる各文字の認識結果の 5-best がある．これらを文字認識誤りの例として，文字認識誤り訂正データセットを作成した．



子どもは物語の世界を直接体に受け入れて、経験が子どもの人生を形作るから、重要な意味があるという思いから。

子どもは物語の世界を直接体に受け入れ？経験が子どもの人生を形作るから、重要な意味があるという鬼いか？。

図 1　答案画像 (左)，認識結果 (中)，人手の文字起こし (右)

5　実験

作成したデータセットを用いて T5 モデルの fine-tuning を行い，文字認識誤り訂正モデルを作成した．用いたモデルは retrieva-jp/t5-base-medium であり，Tokenizer は T5Tokenizer を用いて retrieva-jp/t5-base-medium の Tokenizer を指定した．また，誤り訂正を評価するための指標として，人手による文字起こしデータを参照訳として利用とした BLEU 値 [9] を算出した．BLEU 値の算出には，evaluate ライブラリの sacrebleu[2] を使用した．BLEU 値は 0〜100 までの実数値で表され，生成した文章が参照訳に近いほど値が高くなる．文字認識結果の BLEU 値と，誤り訂正モデルによる訂正結果の BLEU 値を比較し，誤り訂正モデルによる訂正結果の BLEU 値が文字認識結果の BLEU 値を上回った場合，訂正モデルは有効であるとする．

実験は，縦書き文字認識モデルによる文字認識結果のみを用いた実験と，拡張データを学習データとして加えた実験の 2 つを行った．

5.1　実験 1: 縦書き認識モデルの文字認識結果を用いた実験

この実験では，縦書き文字認識モデルによる文字認識結果の 5-best を入力例とした文字認識誤り訂正データセットを用いて T5 を fine-tuning することで，文字認識誤り訂正モデルを作成する．データセットの件数は，6,656 件であった．BLEU 値がより高くなるパラメータを探索するため，あらかじめ指定したハイパーパラメータの組み合わせで grid search を行った．またこの際，精度検証のため 5 分割交差検証を行った．データ数の比率は (学習：検証：テスト)=(3:1:1) とした．

1）https://huggingface.co/retrieva-jp/t5-base-medium

2）https://huggingface.co/spaces/evaluate-metric/sacrebleu

交差検証を行うためのデータセットの分割は，(1) 問題別で 5 分割する方法，(2) 答案別で 5 分割する方法の 2 種類で行った．データ分割のフローを図 2 に示す．

(1) 問題別に分割する方法では，ドリルの問題ごとに学習，検証，テストデータを分割する．こうすることで，学習に利用されなかった問題の回答の文字誤り訂正についての能力を測る．本実験における主な目的は，この能力の測定である．

(2) 答案別に分割する方法では，ドリルの問題を問わず，学習，検証，テストデータを分割する．結果として同じ問題を別の生徒が回答した答案が，学習，検証，テストデータに共通して存在する可能性がある．学習データが十分存在すれば，(2) 答案別の分割の実験の設定程度に性能が上昇すると期待できるため，参考として実験を行った．

実験の際，fine-tuning の学習率は [0.00001, 0.0001, 0.001] の 3 通り，エポック数は [5, 10, 15] の 3 通り，さらに生成の際の繰り返しに対して与えるペナルティである repetition penalty は [5.0, 10.0] の 2 通りで，これらを組み合わせた 18 通りの設定で実験を行った．バッチサイズは 16 とした．

5.2 実験 2: 拡張データを用いた実験

この実験では，単文字認識モデルの文字認識結果を用いて縦書き文字認識モデルの文字認識結果を変換して擬似文字認識データを作成し，これを実験 1 で使用したデータに加えて T5 を fine-tuning することで，文字認識誤り訂正モデルを作成する．作成したモデルの BLEU 値を実験 1 の結果と比較し，データ拡張が有効であるかどうかを検証する．

拡張データの作成

4 章で示したデータのうち，縦書き文字認識モデルによる文字認識結果の best データ，単文字認識モデルによる各文字の認識結果の 5-best データとその尤度を用いて，拡張データを作成する．

1. 単文字認識モデルによる認識結果における各文字の尤度を softmax 関数を用いて確率値に変換する．
2. 縦書き認識モデルによる best データの各文字に対して，1 で求めた確率値の 1/100[3)] の確率で，ランダムに単文字認識の 5-best のうちいずれかの文字と置き換える．
3. 1，2 を 5 回行い，各答案に対して変換データを 5 つ作成する．

上記の手法で作成した拡張データを加えた結果，文字認識結果と人手の文字起こしデータのペアは 11,271 件となり，実験 1 で用いたデータの約 2 倍となった．この拡張データを加えたデータセットを，実験 1 で行った問題別のデータ分割方法で 5 分割しこれを学習データとして，5 分割交差検証を行った．この時，検証データとテストデータは，実験 1 の問題別に分割したデータと同じものを使用した．fine-tuning のパラメータは，実験 1 と同条件で grid search を行った．

6 実験結果

実験の結果，最も高かった文字認識誤り訂正後の BLEU 値を表 1，表 2 に示す．表 1，表 2 の訂正前の BLEU 値は，入力文である文字認識結果の BLEU 値である．表 1 は，実験 1，実験 2 で問題別で 5 分割交差検証を行った結果であり，表 2 は実験 1 で答案別で 5 分割交差検証を行った結果である．すべてのパラメータにおける結果は，付録の表 4〜表 6 に示す．また，各実験で最も BLEU 値が高かったパラメータ設定を表 7 に示す．

表 1 問題別のデータ分割の実験における BLEU 値

	訂正前	訂正後
実験 1 (データ拡張なし)	48.56	52.70
実験 2 (データ拡張あり)		56.26

表 2 答案別のデータ分割の実験における BLEU 値

	訂正前	訂正後
実験 1 (データ拡張なし)	48.70	53.01

表 1 より，問題別のデータ分割の実験において，訂正前の BLEU 値より，実験 1 の文字認識誤り訂正後の BLEU 値の方がよいことが分かる．これより，T5 による文字認識の訂正は有効であるといえる．さらに，表 1 の実験 1 と実験 2 の結果を比較すると，実験 1 の縦書き認識モデルの文字認識結果を用いた実験の BLEU 値を，実験 2 の拡張データを用いた実験が上回っているため，拡張データが有効であることが分かる．

表 2 より，答案別のデータ分割の実験において，訂正前の BLEU 値より，実験 1 の文字認識誤り訂正後の BLEU 値の方がよいことが分かる．さらに，表

3) 尤度から求めた確率値のまま単文字認識の 5-best の文字と置き換えた場合，実際の文字認識結果よりも誤りが多発してしまうため，1/100 をかけている．この値は予備実験により決定した．

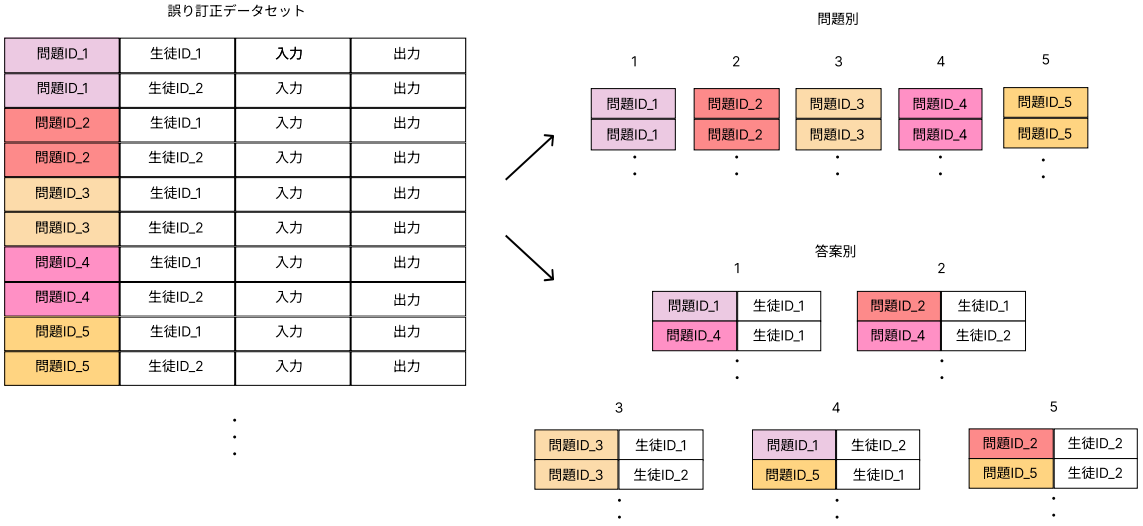


図 2　問題別と答案別のデータ分割方法

1 の実験 1 の結果と比較すると，問題別のデータ分割より答案別の方が BLEU 値が高いことが分かる．このことから，問題別の実験設定では，学習データの量が不足していることが示唆される．しかし，実験 2 の拡張データを用いた実験の方が，同じ問題の別の生徒による回答を含む，答案別の実験設定の結果より BLEU 値が上回っているため，データ拡張の有効性が分かる．

7　考察

実験 2 のデータ拡張後の訂正モデルによる文字認識誤り訂正の例を表 3 に示す．

表 3　文字認識誤り訂正モデルの生成例

入力	読いだ方がいいだろうと u う本モたのうわずに買うこと
生成結果	読んでおいた方がいいだろうと思う本をためらわずに買うこと
正しい文字起こし	読んだ方がいいだろうと いう本もためらわずに買うこと

表 3 の文字認識誤り訂正の例を見ると，正しい文字起こしが「読んだほうがいいだろうという本」であるのに対し，文字認識結果の訂正後の文章は「読んでおいた方がいいだろうと思う本」となっており，理解しやすい日本語を生成していることが分かる．この生成結果には T5 のテキスト生成モデルとしての特徴が表れており，この特徴によって本来誤っていた回答を正答に訂正してしまう懸念はある一方，文意を著しく変更しているわけではないため，採点結果には大きく影響しないのではないかと考えられる．

また，本研究では誤り訂正の評価指標として BLEU 値を用いたが，BLEU 値は翻訳の性能を評価するのに有効な手法ではあるものの，細かい意味的な違いを評価することはできない．そのため，採点システムの前処理としての誤り訂正の指標として，どのような指標が適切であるかを考えていく必要がある．また，本手法は，採点システムに用いることを目的として文字認識誤り訂正を行った．そのため，今後，実際に本手法の訂正結果を採点システムの入力として利用し，評価していく予定である．

8　おわりに

本研究では，自動で手書き答案を読み取った文字認識結果に対して人手で訂正を行い，これらのペアを用いて T5 を fine-tuning することで，文字認識誤りの訂正モデルを作成した．用いた答案データは，中学生約 50 名によって回答された，国語ドリルの記述式問題の日本語手書き答案である．実験の結果，T5 を fine-tuning することで，文字認識結果をより実際の答案に近い文章に訂正できることを示した．また，データ拡張の結果，訂正モデルの精度は向上したため，本研究で行ったデータ拡張手法は有効であったといえる．さらに，今後の課題として，実験のパラメータ設定や評価方法の改善，さらに本手法を使った認識結果による採点を行い，その効果を確認することなどが挙げられる．

謝辞

本研究は，科研費 23H03511 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.220707-04111）

参考文献

[1] Colin Raffel Katherine Lee Sharan Narang Michael Matena Yanqi Zhou Wei Li Peter J Liu et al. Noam Shazeer, Adam Roberts. 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.

[2] 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.

[3] 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.

[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, pp. 137–140, 2020.

[5] 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.

[6] 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, 2022.

[7] 相馬菜生, 梶浦照乃, 高橋舞衣, 倉光君郎. 大規模言語モデルへの追加事前学習による誤り訂正モデルのコードへの適用. **DEIM Forum 2023**, pp. 1b–5–4, 2023.

[8] H.A.Z. Sameen Shahgir , Khondker Salman Sayeed. Bangla grammatical error detection using t5 transformer model. **arXiv:2303.10612 [cs.CL]**, 2023.

[9] Kishore Papineni Salim Roukos, Todd Ward and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. **ACL**, pp. 311–318, 2002.

表 4　実験 1 の各パラメータにおける BLEU 値　問題別

repetition penalty		5.0			10.0		
エポック数		5	10	15	5	10	15
学習率	0.00001	10.97	11.17	2.29	13.38	10.94	7.12
	0.0001	11.75	6.07	**52.70**	27.60	10.92	37.78
	0.001	0.52	23.50	11.28	1.05	7.30	1.43

表 5　実験 1 の各パラメータにおける BLEU 値　答案別

repetition penalty		5.0			10.0		
エポック数		5	10	15	5	10	15
学習率	0.00001	13.73	9.21	2.91	17.50	11.94	3.09
	0.0001	23.74	22.95	**53.01**	0.00	20.53	42.94
	0.001	2.72	6.08	2.43	0.49	1.72	10.05

表 6　実験 2 の各パラメータにおける BLEU 値

repetition penalty		5.0			10.0		
エポック数		5	10	15	5	10	15
学習率	0.00001	22.29	5.95	3.36	22.17	2.81	4.00
	0.0001	7.56	18.37	33.85	12.60	21.04	**56.26**
	0.001	3.41	10.41	16.60	3.71	0.50	25.02

表 7　各実験で最も検証データの BLEU 値が高かったパラメータ

データ分割	実験	エポック数	学習率	repetition penalty
問題別	実験 1	15	0.0001	5.0
答案別	実験 1	15	0.0001	5.0
問題別	実験 2	15	0.0001	10.0