日本語小論文自動採点システムに関する画像データの活用

森本仁　竹内孔一
岡山大学大学院　pebe6zn6@s.okayama-u.ac.jp, takeuc-k@okayama-u.ac.jp

概要

本研究では，テキストの答案データと同時に画像の答案データを入力として利用した場合の自動採点システムの効果について議論する．テキスト特徴量と画像特徴量を直接分類層に入力するモデルとLLaVA モデルの二つのモデルを作成し，BERT のみを利用したモデルとの比較実験を行った．その結果，文法能力について評価を行っている文書力については画像を取り込むことで一部の課題において評価精度が向上した．また画像データによるモデルの性能向上の原因として答案の長さの影響が示唆される実験結果を得た．

1　はじめに

小論文試験は論理的な思考力や自分の考えを表現する力を測定する方法として様々な試験で取り入れられている．その一方で複数の採点者間による細かな採点基準の違いや，採点を長時間行う上での疲労の蓄積，心境の変化などの要因から人間が小論文を一律な基準で採点することは難しい．加えて，人的労力や時間的なコストがかかることから小論文試験を実施するためのハードルが存在する．このような背景から現在多数の小論文自動採点システムが提案されている．

事前学習済み言語モデルである BERT[1] を用いたモデルとしては英語小論文データセットである ASAP[1)] を利用して Yongjie らの R2BERT モデルを採用した自動採点システムが高い精度を出したことが報告されている [2]．模範解答との共通形態素を利用したモデル [3] や文法特徴を取り入れたモデル [4, 5] のような小論文の特徴に即したモデルも構築されている．また，生成系言語モデルである GPT-4[6] を用いた自動採点システムも構築されており [7, 8]，採点根拠をフィードバックさせるなど，教育分野への貢献についても模索されている [9]．

これらの自動採点システムでは，モデルに入力するデータとして電子化されたデータを使用している．ただし，試験の場合，一般には紙の解答用紙に手書きで解答する．そこで，まず近年発展してきている Vision & Language モデルである LLaVA を利用して言語生成モデルで採点させたが高い精度が得られなかった．そこで本研究では，画像の答案データに採点に有効な情報があるかを明らかにするために，テキストの答案と共に解答者が記述した答案を画像データと同時に利用した場合の自動採点モデルへの効果について実験的に明らかにする．

2　実験

2.1　データセット

本研究では GSK2021-B 日本語小論文データ[2)] を利用する．この小論文データは 2016 年から 2017 年にかけて大学生と大学院生を対象に行われた講義の受講者の解答データである．この解答データは理解力，論理性，妥当性，文書力の 4 つの項目についてそれぞれ 1 から 5 点で評価されている．本実験では，4 つの項目の内，理解力と文書力について実験を行う．理解力は課題を理解し，設問に即した内容が書けているかを問う項目であり，主に指定された内容を解答が含んでいるかを評価している．文書力は文法的に正しく書けているかを問う項目であり，文章の長さや誤字数を基に解答を評価している．

本研究では小論文データの課題の内「グローバリゼーションの光と影」,「自然科学の構成と科学教育」の 2 つのテーマについて実験を行う．それぞれのテーマには 3 つの設問が用意されている．以降，各講義テーマをそれぞれ g，s で表し，g の設問 1 は g1 と表す．各課題の小論文データの件数と最大文字数は表 1 に示す．また，提供されている画像データは複数の課題が 1 つの画像にまとめられているため，図 1 に示すように各課題で切り取ってモデルに入力する．

今回の実験では，各課題のデータを学習データ，

1）https://paperswithcode.com/dataset/asap

2）https://www.gsk.or.jp/catalog/gsk2021-b/

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1　各課題の小論文データ数及び最大文字数

課題	小論文データ数	最大文字数
g1	328	300
g2	327	250
g3	327	300
s1	327	100
s2	325	400
s3	327	800

図 1　使用する画像データ例

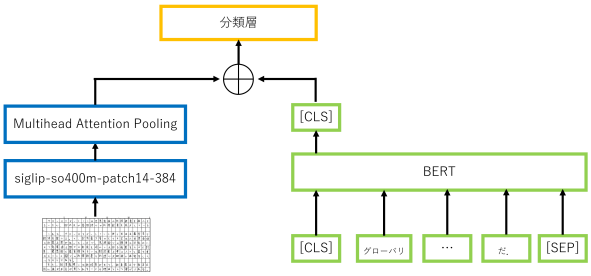

図 2　BERT+画像モデルの概要

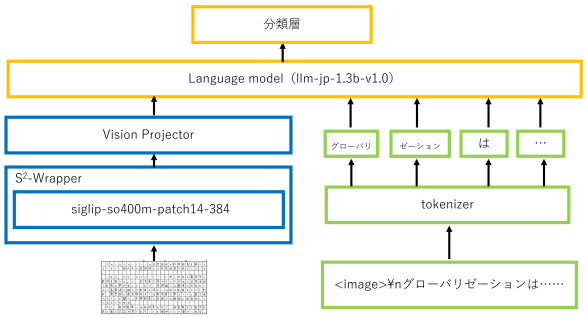

図 3　LLaVA モデルの概要

開発データ，テストデータにそれぞれ 60：20：20 の割合で分割し，5-fold 交差検証を行って評価する．評価値には 2 次の重み付きカッパ係数 (QWK) を用いる．

2.2　モデル

本研究では BERT によるテキストの特徴量と画像特徴量を組み合わせたモデルと LLaVA[10] により画像特徴量を LLM のモデルに入力したモデルを作成する．両モデルは 5 クラス分類を実行し，最終的に 1 点から 5 点のいずれかを出力する．また画像エンコーダとして本研究では siglip-so400m-patch14-384 をどちらのモデルでも利用している．

2.2.1　BERT+画像モデル

BERT+画像モデルでは BERT をテキストデータの特徴量を抽出するために使用し，siglip-so400m-patch14-384 を画像特徴量を抽出するために使用する．図 2 にその概要を示す．

まず，BERT のトークナイザーを使用してトークン化したテキストデータを BERT に入力する．BERT の最終層から CLS トークンを取り出し，このトークンの特徴ベクトルをテキストデータから得られたテキスト特徴ベクトルとして利用する．次に，画像エンコーダを利用し，画像から埋め込みベクトルを獲得する．最終層の埋め込みベクトルを Multihead Attention Pooling したベクトルを画像特徴ベクトルとして利用する．このテキスト特徴ベクトルと画像特徴ベクトルを結合して線形分類層に渡し，5 クラスのクラス分類を行う．

本研究では画像特徴量を追加した場合に採点にどのような結果をもたらすのか実験的に明らかにするため，BERT と siglip-so400m-patch14-384 は凍結し，線形分類層のみを学習させる．

2.2.2　LLaVA モデル

LLaVA モデルでは LLM に画像特徴量を含んだ特徴量を入力し，出力として 5 クラス分類を行う．今回は LLM として llm-jp-1.3b-v1.0[3] を使用する．図 3 にその概要を示す．

始めに<image>トークンを含んだプロンプトを読み込む．<image>トークンは取り込んだ画像の特徴ベクトルに置き換えられ，そのほかのテキストはトークナイズされ，最後にそれらを組み合わせて LLM に入力される．

画像は S^2-Wrapper[11] を用いた画像エンコーダによって取り込み，その最終層の出力を Vision Projector を用いて LLM に入力可能な形に変形する．S^2-Wrapper は図 4 に示すように，サイズの大きい画像を分割して画像エンコーダに入力することでより高画質な画像の入力を可能とする手法である．Vision Projector は 2 層の線形層で構成されており，画像エンコーダの出力を LLM の入力に変換してい

3）　https://huggingface.co/llm-jp/llm-jp-1.3b-v1.0

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

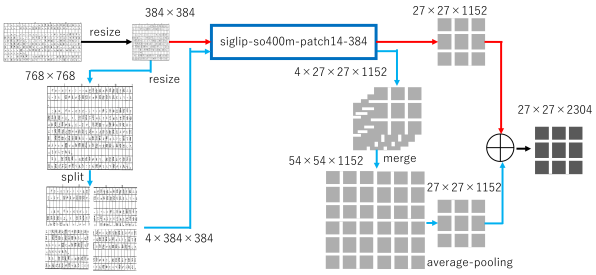

図 4　S^2-Wrapper の概要

表 2　モデルごとの理解力の採点結果

	BERT	BERT+画像	LLaVA
g1	**0.3033**	0.1377	-0.0260
g2	**0.7162**	0.6536	0.0559
g3	**0.2771**	0.2659	0.0309
s1	**0.5438**	0.4536	0.2106
s2	**0.5119**	0.4034	0.0167
s3	0.2348	**0.2912**	0.1860

表 3　モデルごとの文書力の採点結果

	BERT	BERT+画像	LLaVA
g1	0.5744	**0.6339**	0.4320
g2	0.6264	**0.6387**	0.3082
g3	0.6597	**0.6629**	0.5790
s1	**0.3530**	0.2410	-0.0221
s2	**0.6951**	0.6908	0.3517
s3	**0.6912**	0.6869	0.1362

表 4　文字数を追加したモデルの採点結果

	BERT+len(理解力)	BERT+len(文書力)
g1	0.3014	0.6310
g2	0.6711	0.6380
g3	0.3192	0.6399
s1	0.5026	0.2296
s2	0.4456	0.6980
s3	0.3440	0.7102

る．Vision Projector を通した特徴ベクトルをトークナイズされたテキストデータと結合する．

結合した特徴量を LLM に入力する．LLM の最終トークンの埋め込みベクトルを線形分類層が受け取り，5 クラスのクラス分類を行う．

本研究では 2.2.1 章と同様に画像特徴量を追加した場合，採点にどのような結果をもたらすかを実験的に明らかにするため，LLM と siglip-so400m-patch14-384 は凍結し，線形分類層と Vision Projector のみを学習させる．

3　実験結果と考察

理解力，文書力に対する自動採点モデルの QWK をそれぞれ表 2，表 3 に示す．今回は画像を取り込んでいない場合との比較モデルとして BERT の CLS トークンのみを利用した BERT モデルをベースラインモデルとして用意した．

理解力については課題 s3 を除き，BERT モデルの QWK が 3 つのモデルの中で高く，課題 s3 のみ BERT+画像モデルが BERT モデルを上回った．文書力では BERT+画像モデルがテーマ g に関しては QWK が 3 つのモデルの中で最も高く，テーマ s では BERT モデルが高い値を得た．理解力と文書力について比較すると，理解力では一部を除き画像特徴量が採点に悪影響を及ぼしているが，文書力ではテーマ s に関しても QWK の差が BERT モデルと BERT+画像モデルの間で小さく，採点に少ないながらも貢献していることが分かる．LLaVA モデルについても BERT を利用したモデルと比べて精度自体は大きく劣るが，理解力に比べ文書力では大きく精度が向上しており，文書力の採点で画像ベクトルが貢献したことが分かる．

この要因として文書力は文字数に関する評価を行っており，画像に含まれる余白の割合が採点に影響したと考えられる．そこで，文字数の長さを特徴量として BERT モデルに入力したモデルを BERT+len モデルとして追加実験を行った結果を表 4 に示す．理解力では課題 g3 と s3 の QWK が向上し，文書力では課題 g3 と s1 以外の精度が向上した．文書力において BERT+len モデルと BERT+画像モデルの QWK のスコアの傾向は近く，画像データから文字列の特徴量を取り込んでいることが分かる．また理解力において BERT+len モデルで課題 g3 と s3 の QWK が向上した点について，これらの課題は自由度が高く解答者間の解答の違いが大きいという共通点がある．そのため BERT 単体では類似度の比較は難しかったと考えられる．理解力は内容について評価する評価項目であるが，文字数が少ない解答については内容自体が少ないため必然的に低めの評価が付けられる．そういった文字数と評価との相関関係が自由度の高い 2 つの課題において評価の基準となり，QWK が向上したのだと考えられる．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

4 おわりに

本研究では，自動採点システムに画像の解答データを入力した場合の効果について検証した．その結果，文法能力について評価する文書力については精度の向上に寄与することを示した．追加実験として文字数の長さを特徴量として追加した実験を行った結果，画像データから文字数の特徴を得ているということが明らかとなった．また，自由度の高い問題については内容を評価した理解力についての評価でも文字数の特徴量が精度向上に貢献することが明らかとなった．本研究では学習層を最小限にして実験を行ったため，BERT や画像エンコーダ，LLM の学習を行い，画像入力に適したモデルを作成するのが今後の課題である．また今回の画像データの入力では文字数の要素が大きいという部分が判明したが，文字の形状や誤字など画像から得られるデータは他にも考えられるため，そういった要素との関係性を明らかにすることでより精度の高いモデルの構築を検討したい．

謝辞

議論に参加してくださいました竹内研究室の諸氏に心より感謝致します．

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. **CoRR**, Vol. abs/1810.04805, , 2018.

[2] Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. On the Use of Bert for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 3416–3425, Seattle, United States, July 2022. Association for Computational Linguistics.

[3] 竹内孔一, 大野雅幸, 泉仁宏太, 田口雅弘, 稲田佳彦, 飯塚誠也, 阿保達彦, 上田均ほか. 研究利用可能な小論文データに基づく参照文書を利用した小論文採点手法の開発. 情報処理学会論文誌, Vol. 62, No. 9, pp. 1586–1604, 2021.

[4] 土肥康輔, 須藤克仁, 中村哲. エッセイ自動採点における文法特徴と学習者レベルの関係. 言語処理学会第 29 回年次大会 発表論文集, pp. 211–216, 2023.

[5] 土肥康輔, 須藤克仁, 中村哲. 文法項目の多様性と誤り情報を利用したエッセイ自動採点. 言語処理学会第 30 回年次大会 発表論文集, pp. 1160–1164, 2024.

[6] GPT-4 Technical Report, 2024.

[7] Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. Rating Short L2 Essays on the CEFR Scale with GPT-4. In **Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)**, pp. 576–584, 2023.

[8] SeongYeub Chu, JongWoo Kim, Bryan Wong, and Mun-Yong Yi. Rationale Behind Essay Scores: Enhancing S-LLM's Multi-Trait Essay Scoring with Rationale Generated by LLMs, 2024.

[9] 中本さや香, 嶋田和孝, 岡本芳明, 中河内孝. 日本語小論文に対する採点およびフィードバックの生成. 言語処理学会 第 30 回年次大会 発表論文集, pp. 1142–1147, 2024.

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.

[11] Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When Do We Not Need Larger Vision Models? **CoRR**, Vol. abs/2403.13043, , 2024.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).