ルーブリックに基づいたタグを付与した日本語小論文データの構築と自動採点への効果

成岡 智也[1]　竹内 孔一[2]
[1] 岡山大学工学部　[2] 岡山大学大学院環境生命自然科学学域
pak13nxr@s.okayama-u.ac.jp
takeuc-k@okayama-u.ac.jp

概要

本研究では，小論文にルーブリックに基づいたアノテーションを行い，そのデータを用いて小論文自動採点を行う方法を提案する．ルーブリックの評価部分に対応した小論文の文書内構造の情報を，小論文にタグをつけるという形でアノテーションを行った．そうして得られるタグを小論文と同時にトークン化し，採点モデルにスコアを予測させ，小論文のみでの自動採点と比較することで，提案手法の有効性を明らかにした．

1　はじめに

大量の小論文を人間が平等に採点することは困難であり，またそれに伴う労力も大きい．近年，このような問題を解決するために，大規模言語モデルを用いた様々な小論文自動採点手法の研究が行われている．[1] しかし，小論文自動採点分野では，更なる精度の向上や，学習データを大量に必要にしてしまう点など様々な課題が見られる．

そこで，小論文を自動で採点するにあたって，追加情報を与えることを考える．小論文は，基本的に各問題に対して採点基準 (ルーブリック) を基準に採点が行われる．小論文の内容に関して得点を決める際，その要素が小論文の中にどの程度含まれるかによって得点が決まる．先行研究では，小論文の構造を扱った研究がある [1][2][3][4] が，ルーブリックで評価される部分を日本語の小論文に対して付与したデータを作成し，小論文自動採点を適用した研究は我々の知る限り見当たらない．このような背景から，本論文ではルーブリックに基づいたアノテーションを行ったデータを作成するとともに，そのアノテーション結果を自動採点に利用し，小論文のみを利用した自動採点手法と比較し，有効性を明らかにする．

2　関連研究

大規模言語モデルによる小論文自動採点に小論文アノテーションを用いた例として，Ku らの研究が挙げられる [2]．この研究では，中国語小論文において，各小論文に対しての採点根拠を小論文外部にアノテーションし，それらを学習させることによって，モデルが未知の小論文に対してより良い採点根拠を出力させることを行っている．また，エンコーダに入力するトークンとして小論文以外のものを用いた例として，Chu らの研究が挙げられる [5]．この研究では，英語小論文において，GPT3.5 turbo と Llama3.1-8B-Instruct を用いて各小論文に対応する「根拠」を生成し，それぞれを共通のエンコーダに入力して自動採点を行うことによって，スコア予測において全体的な QWK の向上がみられている．本研究では，日本語小論文に対して内部にアノテーションを行い，その構造を追加で用いて自動採点を行っている．

3　ルーブリック評価部分のアノテーションデータ構築

本研究で使用している小論文データは，言語資源協会から公開されている日本語小論文データ [6] である．この小論文データは，2016 年及び 2018 年に開講された講義の受講者の回答データから作成されている．今回利用した小論文データは「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」，「東アジア経済の現状」，「批判的思考とエセ科学」の計 4 つのテーマであり，それぞれの講義テーマごとに 3 つの設問が用意されている．各設問ごとに 300 件前後の小論文がある．以降，各講義テーマは「グローバリゼーションの光と影」を global,

表 1　文書構造分析タグの種類

課題名	タグ
global_q1	光，影，格差縮小，格差拡大
global_q2	光，影，具体例，企業名
global_q3	光，影，具体例，見解
science_q1	実証性の説明，再現性の説明， 客観性の説明
science_q2	自然相手，持続役割，根拠，客観確保， 共通役割
science_q3	導入，科学リテラシー，展開，今後
easia_q1	相互依存，協力・協業の実態，具体例
easia_q2	概略，脱する方法
easia_q3	日本，韓国，中国，協調と対立
criticize_q1	論理的・合理的思考の説明， 目標志向的思考の説明， 内省的・熟慮的思考の説明
criticize_q2	導入，相関あり，因果なし， 妥当でない理由，カラーテレビ， 平均余命，まとめ
criticize_q3	実例，方法，証拠，説明，要因

「自然科学の構成と科学教育」を science，「東アジア経済の現状」を easia，「批判的思考とエセ科学」を criticize と表現する．加えて各設問を q1，q2，q3 と表現し，これらを組み合わせて「グローバリゼーションの光と影」の設問 1 を global_q1 という風に表現する．この各設問に対する各解答に，ルーブリックの内容をもとにしたタグを付けるアノテーションを行った．以降，このタグを「文書構造分析タグ」または単に「タグ」と呼ぶ．各設問に対応する文書構造分析タグの種類は，表 1 の通りとなっている．

global の設問および science の設問には，「光?」のようにタグに?がついているものも存在する．これは，同名のタグと比較して確信度が低い場合に用いられている．本研究では?がついているタグは一律に使用していない．

また，例えば設問 global_q2 に関して，データ中のタグの種類は表に記載のある通り 4 種類であるが，うち「具体例」タグと「企業名」タグは理解力のスコアに関係がない．このような場合，一部タグは使用していない場合がある．

実際のアノテーションされた小論文は，図 1 のような形式になっている．

図 1 は，設問 global_q1 の小論文の一部である．例えば，1 行目に着目すると，「グローバリゼーションは世界全体の所得格差を縮小する」の部分に「光」「格差縮小」タグが，「国内及び各国間での所得格差を拡大している」の部分に「影」「格差拡大」タグがアノテーションされていることがわかる．

4　実験

本章では，3 章で述べたデータセットを用いて小論文の採点を行い，小論文採点における文書構造分析タグの有効性について確かめる．以下，各節で実験設定，実験結果，考察を述べる．また，本実験では，小論文採点モデルとして BERT (Bidirectional Encoder Representations from Transformers) [7] を使用する．

4.1　実験設定

実験では，BERT を用い，各小論文の 1 点から 5 点で評価されている「理解力」のスコアを 5 クラス分類によって予測する．本実験では，HuggingFace の BERT[1)] を利用する．評価指標には Accuracy と QWK (Quadratic Weighted Kappa) を用いる．それぞれ，以下の式で示す．

$$P(a,b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{if } a \neq b \end{cases} \tag{1}$$

$$\text{Accuracy} = \frac{\sum_{l=1}^{n} P(A_l, B_l)}{n} \tag{2}$$

$$\text{QWK} = 1 - \frac{\sum_{i,j} O_{ij} W_{ij}}{\sum_{i,j} E_{ij} W_{ij}} \tag{3}$$

Accuracy は，式 (1)，(2) によって求められる．n は各設問における小論文の総数であり，A_l，B_l は l 番目の小論文に対してそれぞれ正解のスコア，予測したスコアである．Accuracy は 0 から 1 の値を取り，1 に近いほど一致率が高い．また，QWK は式 (3) によって求められる．i，j はそれぞれ正解のスコア，予測したスコアである．W_{ij} は正解のスコアと予測したスコアが離れているほど大きくなる値であり，観測された一致度 O_{ij} と期待される一致度 E_{ij} の双方に W_{ij} を重み付けする．QWK は，予測したスコアと正解のスコアが離れていればいるほど大きなペナルティを与える，小論文自動採点モデルで一般的に用いられる評価指標である．QWK は-1 から 1 の値を取り，1 に近いほど一致率が高い．

また，アノテーションをしてある小論文について，以下の図 2 ような方法を用いてトークナイズする．

これは，A というタグが 2 箇所に，B というタグが 1 箇所についている小論文の例である．各タグに

1）　https://huggingface.co/tohoku-nlp/bert-base-japanese-v2

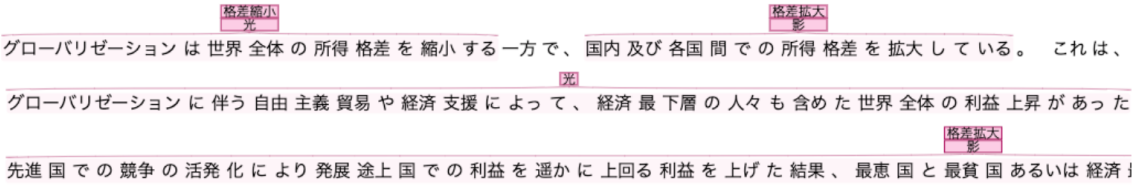

図 1　アノテーションされた小論文データ

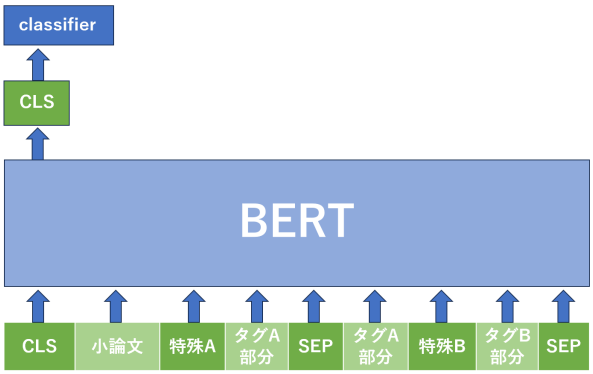

図 2　BERT による採点モデルと入力トークン例

対応した特殊トークンをタグの切り替わる位置に挿入している．また，同じタグの文は，SEP トークンで分割している．これを入力した際に出力される CLS トークンを取得し，最終的な予測値を得る．

各小論文データは 6:2:2 の割合で分割し，それぞれを学習データ，検証データ，テストデータとして用いる．学習に関して，バッチサイズは 16 で行う．各 epoch が終了するごとにその時点のモデルで検証データに対して Accuracy を評価し，その Loss 値がそれ以前の epoch 終了時のものの中で得られた最小の Loss 値より大きかった回数が 5 となった時点で学習を終了し，その時点で最も検証データに対しての Loss 値が少なかったモデルを使用してテストデータを評価する．

4.2　実験結果

以降，前節で提案したモデルを用いて自動採点を行った結果を示す．表 2 に，文書構造分析タグを用いずに自動採点を行った際の Accuracy と QWK，及び前節で提案した方法で文書構造分析タグを用いて自動採点を行った際の Accuracy と QWK を示している．

文書構造分析タグを用いずに行った採点での Accuracy の平均値は 0.5842，QWK の平均値は 0.4772 であった．対して，文書構造分析タグを用いた採点での Accuracy の平均値は 0.6567，QWK の平均値は 0.6241 であった．設問 global_q2，設問 global_q3，設問 science_q2，設問 easia_q2 においては，タグを用いて実験を行った場合より，タグを用いなかった場合のほうが高い結果が得られる例が見られた．そのほかの設問に関しては，Accuracy，QWK ともにタグを用いたほうが高い結果が得られた．

表 2　実験結果

	BERT のみ		BERT+タグ	
課題名	Acc.	QWK	Acc.	QWK
global_q1	0.5821	0.4548	**0.6866**	**0.5973**
global_q2	**0.8030**	**0.8322**	0.7879	0.7648
global_q3	**0.5522**	0.2165	0.5373	**0.4071**
science_q1	0.6176	0.5201	**0.7500**	**0.7807**
science_q2	0.5075	**0.6555**	**0.5522**	0.6211
science_q3	0.6418	0.0381	**0.6716**	**0.4827**
easia_q1	0.6000	0.6284	**0.7000**	**0.6927**
easia_q2	**0.4667**	0.4164	0.4167	**0.5197**
easia_q3	0.7333	0.5272	**0.7677**	**0.6277**
criticize_q1	0.5833	0.7017	**0.6833**	**0.7750**
criticize_q2	0.4833	0.4568	**0.6333**	**0.6662**
criticize_q3	0.4407	0.2788	**0.6949**	**0.5581**
Average	0.5842	0.4772	**0.6567**	**0.6241**

4.3　考察

全体の考察　4.2 節で述べた通り，設問全体での Accuracy と QWK の値の平均値は，ともにタグを用いたモデルのほうが高い．この結果から，文書構造分析タグを用いた小論文自動採点を行うことは，採点結果の Accuracy と QWK を向上させるにあたってよい影響を与えていると考えることができる．特に QWK に関しては，設問 science_q1，設問 science_q3，設問 criticize_q3 などかなりの向上が見られるものもあったため，特によい影響を与えていると考えることができる．代表して，設問 science_q1 の混同行列を図 3，図 4 に示す．

図 3 がタグをつけていない場合の混同行列，図 4 がタグをつけている場合の混同行列である．これを見ると，5 点の小論文を 4 点と誤って評価する例や，3 点の小論文を 5 点と誤って評価する例が少なくなったことなどがわかる．これによって Accuracy，QWK が上がったと考えられる．一方，5 点，4 点

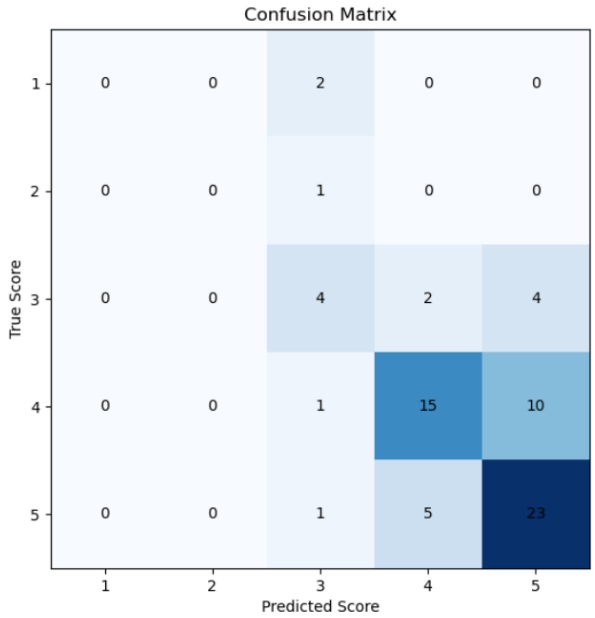

図 3　science_q1　タグ無し

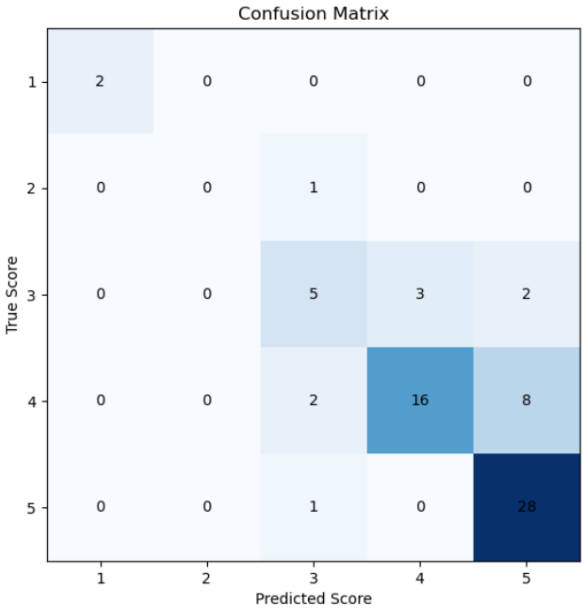

図 4　science_q1　タグ有り

の小論文と比較してその他の点数の母数が少なく，データセットに母数が少ないとタグがあっても予測が難しいといった側面は見られるため，こういった点は課題として挙げられる．

性能が低下したものの考察　設問 global_q2，設問 global_q3，設問 science_q2，設問 easia_q2 では，タグを用いることで Accuracy と QWK のいずれか，もしくは双方において低くなってしまう結果が得られた．しかし，設問 science_q2 には大きな差があるわけではないことから，性能が落ちたのではなくタグがあまり採点に影響しなかったと考えるべきである．設問 global_q3，設問 easia_q2 に関しては，QWK が上がってはいるものの Accuracy は低下しているため，性能が向上した，低下したと一概に評価することは難しい．設問 global_q2 のみ Accuracy，QWK の双方において低下が見られる．これに関して，設問 global_q2 が他の設問と比較して，Accuracy および QWK が高いことに起因しているのではないかと考えられる．このことから，タグを用いての小論文採点は小論文単体でスコアを評価することが簡単である設問には不適切である可能性が考えられる．これは，スコアを推定するための入力として小論文のみで十分であるにも関わらず，過剰な情報をタグによって与えてしまった結果，スコア予測に悪影響を与えてしまったということが想定される．

5　まとめ

本研究では，BERT を用いた小論文自動採点において，小論文のアノテーションを行うことで得られる文書構造分析タグを用いる手法を提案し，その性能評価を行った．実験により，タグを用いることで，Accuracy と QWK の双方の向上が期待できることを明らかにした．しかし，一部の小論文課題に対してはスコア予測の性能の低下も見られたため，文書構造分析タグの利用方法についての課題があることも確認した．今後は，今回性能が低下した小論文課題に対しても有効に働くようなタグの利用方法の検討や，大規模言語モデルを用いたアノテーション及びタグ付けの自動化に取り組んでいく予定である．

6　データ

本研究で構築したデータは整理した後，公開する予定である[2)]．

2）　https://www.cl.cs.okayama-u.ac.jp/?page_id=2 で案内する予定である．

謝辞

本研究は JSPS 科研費 JP22K00530 の助成を受けたものです.

参考文献

[1] 石井雄隆, 舟山弘晃, 松林優一郎, 乾健太郎. 国語記述式問題自動採点システムの開発と評価. 日本教育工学会研究報告集, 2024.

[2] Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Plaes: Prompt-generalized and level-aware learning framework for cross-prompt automated essay scoring. In **The International Conference on Computational Linguistics**, 2024.

[3] 加藤嘉浩. 論理構造グラフを用いた自動採点モデル. 言語処理学会　第 30 回年次大会　発表論文集, 2024.

[4] Yuning Ding, Marie Bexte, and Andrea Horbach. Score it all together: A multi-task learning study on automatic scoring of argumentative essays. In **Association for Computational Linguistics**, 2023.

[5] SeongYeub Chu, JongWoo Kim, Bryan Wong, and MunYong Yi. Rationale behind essay scores: Enhancing s-llm's multi-trait essay scoring with rationale generated by llms. In **arXiv:2410.14202**, 2024.

[6] 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. 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, 2021.

[7] 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, Volume 1(Long and Short Papers)**, 2019.