E-007

BERT による国語記述式答案の自動採点における採点アノテーションの重要性
The Importance of Scoring Annotation in Automated Short Answer Scoring by BERT

高野 峻矢† 市川 治†
Shunya Takano† Osamu Ichikawa†

1. はじめに

自然言語処理技術を用いた自動採点システムによって，採点結果の即時の返却が可能であること，答案を送信できる電子端末さえあれば場所を選ばない利用が可能であること等の利点がある．そのようなインタラクティブな学習の実現のために理化学研究所革新知能統合研究センター自然言語理解チームは採点基準を活用した自動採点技術の開発を進めている[1]．その一例として，水本らは LSTM を含めた深層学習のモデルを拡張し，教師ありアテンション機構を使用して Segmentation Cue（採点用のアノテーション）を推定することで，国語記述式答案を対象として採点項目別に点数を出力する自動採点モデルを提案した[2]．水元らが提案したモデルは 1 つの問題につき 1000 答案程度の訓練データがあれば人間と同等程度の採点精度を得ることができる．しかし 1000 答案もの訓練データを人手によって用意するのは，実際の教育現場では現実的ではない．

本報告で提案する自動採点モデルでは，自然言語処理において高い性能を示すことが知られている BERT[3] を用いて，アノテーションを直接推定する，それに基づいて，同じく BERT によって得られる埋め込み情報を統合し，それらを勾配ブースティングの機械学習フレームワークである LightGBM[4]へ入力し，採点項目別に点数を推定する．本手法により 1 つの問題につき 100 答案程度の訓練データで人間と同等程度の採点精度を得ることができた．

2. 理研記述問題採点データセット

本報告では文献[2]で用いられている，理化学研究所が作成した理研記述問題採点データセット[5]を用いる．このデータセットは代々木ゼミナールで 2014 年から 2015 年に行われた高校生対象模擬試験の国語記述問題について，各受験者の答案に採点基準に基づくアノテーションと採点を付与したものである．アノテーションとは，図 1 のように採点基準ごとに，採点の根拠となる単語に「1」，根拠とならない単語に「0」のアノテーションがそれぞれ付与されている．また答案が採点基準を満

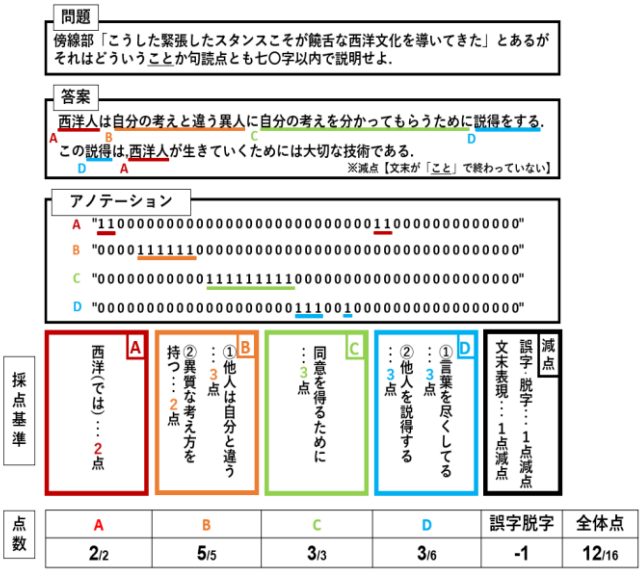

図 1　理研記述問題採点データセットの例

たすかどうかで項目点が付与され，それらの合計として全体点が計算されている．図 1 の例では採点基準として A，B，C，D の 4 項目があり，それぞれについてアノテーションが付与されている．そして，それらに基づいて項目ごとの採点がなされる．さらに，項目点と減点とを合計して全体点が計算される．

3. 提案モデル

3.1　項目点・全体点予測モデル

図 2 に提案モデルを示す．まず答案をトークンと呼ばれる文字列の単位に分割し，そのトークンを ID に変換して事前学習済みの BERT モデルに通す．その後トークンは文脈を考慮した 768 次元の埋め込みベクトルに変換される．さらに採点基準ごとに，「1」のアノテーションが付与された，採点の根拠となる単語トークンの埋め込みベクトルを寄せ集める．寄せ集めたベクトルに対して，次元ごとに最大値を取ることにより，新たに 768 次元のベクトルを生成する．その後，そのベクトルを特徴量として LightGBM モデルを学習し，それぞれの項目点を予測する．最後に全ての項目点を合計し全体点を計算する．ただし，アノテーションが全て「0」の場合， 寄せ集めるベクトルが存在しないため，LightGBM モデルを経由せず，その項目点を 0 点とする．

† 滋賀大学 データサイエンス学部
Faculty of Data Science, Shiga University

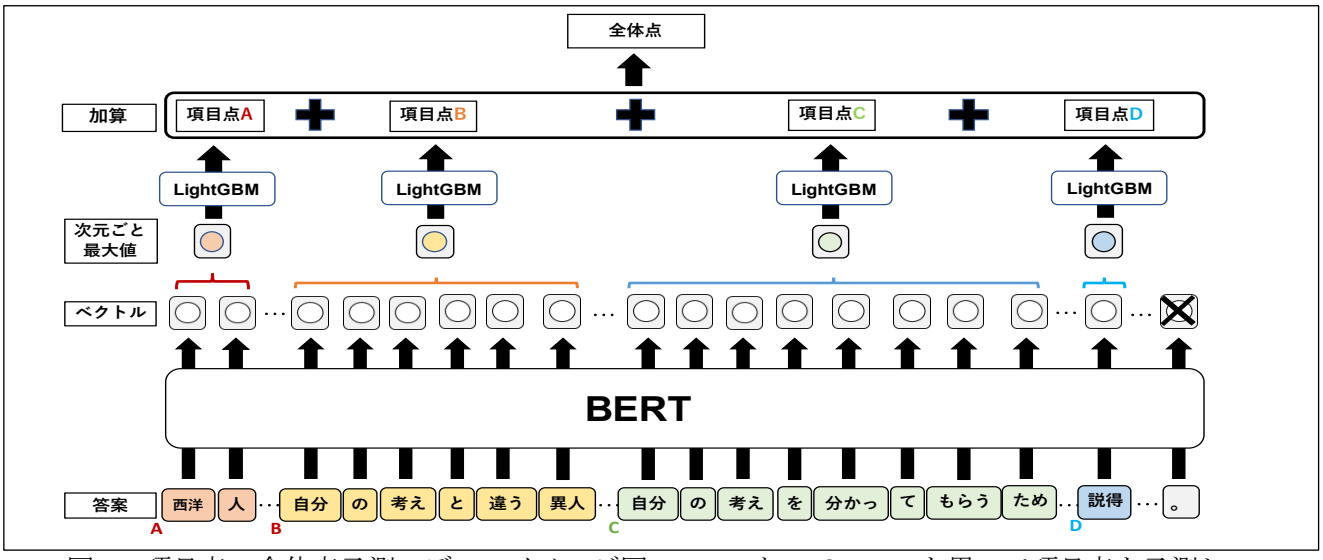

図 2　項目点・全体点予測モデルのイメージ図．BERT と LightGBM を用いて項目点を予測し，それらの合計によって全体点を計算する．

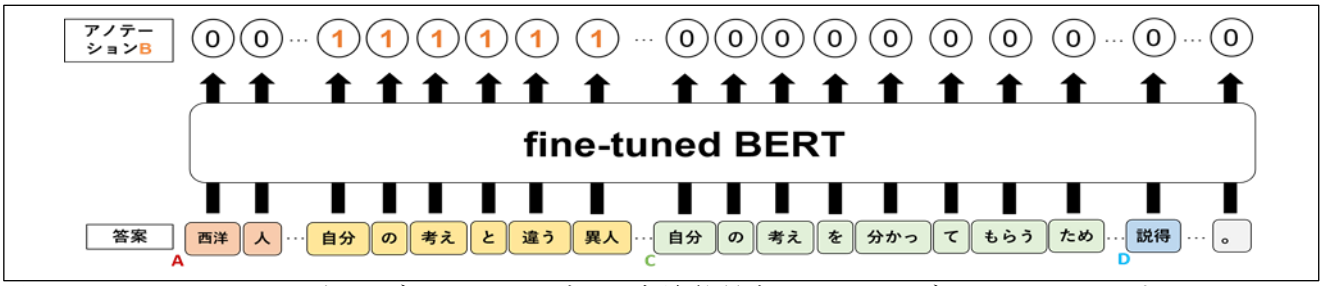

図 3　アノテーション予測モデルのイメージ図．事前学習済み BERT モデルを fine-tuning することで，アノテーションを「1」，「0」のラベルで予測する．

3.2　アノテーション予測モデル

3.1 項に示したモデルではアノテーションが利用可能であることが前提となっている．しかし，本来自動採点の対象となる答案にはアノテーションは付与されていないはずである．よって採点対象の答案からアノテーションを予測するモデルも必要となる．本報告ではこのモデルを BERT の事前学習済みモデルを fine-tuning[*1] することで作成する．図 3 に提案モデルを示す．この図では，採点基準 B を例にとり，B を表現する単語について，「1」のアノテーションが出力される．

4. 評価実験

3.1 項と 3.2 項で示したモデルの性能を評価するために以下の実験を行った．なお事前学習済みの BERT モデルとして東北大学乾・鈴木研究室が公開している日本語 BERT モデル[6]を使用する．

[*1]fine-tuning とは既存の学習済モデルの重みデータの一部再学習を行うこと．

表 1　データセットの統計量

問題	Q1	Q2	Q3	Q4	Q5	Q6
配点	16	12	12	15	15	14
平均値	6.8	4.0	5.3	5.5	4.6	5.5
標準偏差	3.5	1.8	2.1	2.7	2.6	3.1
採点基準数	4	4	4	3	3	3
字数制限	70	50	60	70	70	60

4.1　使用するデータセット

データセットは 2 節に示した理研記述問題採点データセットを使用する．全 9 問題中 6 題を使用し，答案は各問題 2100 件ある．これは，従来研究[2]と同じ設定である．なお採点基準は複数の加点項目に加え，誤字・脱字，文末表現などを対象とした減点項目から構成されているが，本実験では，加点項目のみの合計を解答の得点とした．表 1 にデータの統計量を示す．また 3.1 項や 3.2 項で示したモデルの性能評価のために，答案の 100 件を訓練データ，残りの 2000 件をテストデータとしたデータの組み合わせを，セット間で訓練データが

表 2　予備実験の結果 (QWK)

	問題					
QWK	Q1	Q2	Q3	Q4	Q5	Q6
ケース 1 (100 訓練)	0.96	0.94	0.91	0.87	0.94	0.94
ケース 2 (300 訓練)	0.98	0.96	0.94	0.91	0.95	0.95
人間の採点精度	0.96	0.94	0.76	0.84	0.82	0.90

表 3　予備実験の結果 (MSE)

	問題					
MSE	Q1	Q2	Q3	Q4	Q5	Q6
ケース 1 (100 訓練)	0.83	0.35	0.63	1.65	0.86	1.20
ケース 2 (300 訓練)	0.57	0.22	0.46	1.20	0.61	0.89

被らないように 5 セット用意した．同じように答案の 300 件を訓練データ，残りの 1800 件をテストデータとしたデータの組み合わせを，訓練データが被らないように 5 セット用意した．前者をケース 1 (100 答案訓練)，後者をケース 2 (300 答案訓練) と呼ぶ．

4.2　(予備実験) 正解のアノテーションが与えられた場合の自動採点性能

3.1 項で示したモデルの性能について，性能の上限を調べるため，データセットに付与されている正解のアノテーションを利用して，項目点・全体点を予測した．評価指標は Quadratic Weighted Kappa (QWK)[*2] と Mean Squared Error (MSE) を使用し，5 セットの訓練データとテストデータの組で計算した評価指標の平均値を表 2，表 3 にそれぞれ示す．また文献[2]で検証された人間の採点精度 (QWK) も表 2 に合わせて示す．

4.3　(実験 1) アノテーション予測

3.2 項で示したモデルの性能評価を行う．図 3 は簡単のため，採点基準 B に関するアノテーションのみを予測するイメージ図にしているが，実際には各問題の全ての採点基準についてアノテーションを予測するため，合計で 21 個の BERT モデルを用意する必要がある．また BERT モデルは epoch 数を 3，バッチサイズを 16，最適化アルゴリズムを Adam，損失関数を交差エントロピー関数と設定し，事前学習済み BERT モデルを fine-tuning することで作成した．評価指標は正解率，適合率，再現率，F 値を使用する．

[*2] Quadratic Weighted Kappa (QWK) とはマルチクラス分類用の評価指標で，クラス間に順序関係があるような場合に使用される．0 から 1 までの値を取り，値が大きいほど，予測の当てはまりが良いことを表す．本報告では予測した全体点に対して端数処理を行い整数に変換し，その整数の点数をクラスとして扱う．また文献[2]において，人間の採点精度 (QWK) をどのように計算したのかは，詳しい記述はないので，本報告と同じように計算したと仮定して論じる．

表 4　実験 1 の結果 (正解率)

正解率 (Accuracy)		問題					
		Q1	Q2	Q3	Q4	Q5	Q6
採点基準	A	0.998	0.982	0.942	0.997	0.951	0.947
	B	0.974	0.998	0.925	0.898	0.988	0.955
	C	0.987	0.981	0.975	0.965	0.987	0.974
	D	0.985	0.995	0.980	—	—	—

表 5　実験 1 の結果 (適合率)

適合率 (Precision)		問題					
		Q1	Q2	Q3	Q4	Q5	Q6
採点基準	A	0.981	0.952	0.908	0.942	0.886	0.842
	B	0.935	0.876	0.886	0.880	0.872	0.911
	C	0.966	0.993	0.741	0.903	0.952	0.817
	D	0.929	0.814	0.587	—	—	—

表 6　実験 1 の結果 (再現率)

再現率 (Recall)		問題					
		Q1	Q2	Q3	Q4	Q5	Q6
採点基準	A	0.970	0.945	0.874	0.953	0.854	0.878
	B	0.928	0.996	0.847	0.867	0.922	0.903
	C	0.935	0.969	0.874	0.911	0.971	0.884
	D	0.934	0.921	0.682	—	—	—

表 7　実験 1 の結果 (F 値)

F 値 (F-measure)		問題					
		Q1	Q2	Q3	Q4	Q5	Q6
採点基準	A	0.977	0.946	0.891	0.949	0.870	0.860
	B	0.931	0.939	0.868	0.872	0.897	0.907
	C	0.950	0.981	0.802	0.907	0.961	0.850
	D	0.931	0.864	0.625	—	—	—

ケース 2 (300 答案訓練)について，5 セットの訓練データとテストデータの組で計算した評価指標の平均値を表 4，表 5，表 6，表 7 にそれぞれ示す．

4.4　(実験 2) 自動で推定されたアノテーションを使用した自動採点

3.1 項で示したモデルと 3.2 項で示したモデルを両方用いた場合の性能評価を行う．3.2 項で示したモデルによってアノテーションを予測し，そのアノテーションを用いて項目点・全体点を予測する．評価はケース 1 (100 答案訓練) とケース 2 (300 答案訓練)で行った．評価指標は QWK と MSE を使用し，5 セットの訓練データとテストデータの組で計算した評価指標の平均値を表 8，表 9 にそれぞれ示す．また文献[2]で検証された人間の採点精度(QWK)も表 8 に合わせて示す．

4.5　実験結果のまとめと考察

表 2 より正解のアノテーションが与えられた場合の自動採点の性能について，いずれの問題においても，ケース 1，ケース 2 の両方の QWK が人間の採点精度における QWK よりも高くなった．

表 8　実験 2 の結果 (QWK)

	問題					
QWK	Q1	Q2	Q3	Q4	Q5	Q6
データセット 1	0.96	0.94	0.91	0.87	0.94	0.94
データセット 2	0.98	0.96	0.94	0.91	0.95	0.95
人間の採点精度	0.96	0.94	0.76	0.84	0.82	0.90

表 9　実験 2 の結果 (MSE)

	問題					
MSE	Q1	Q2	Q3	Q4	Q5	Q6
ケース 1 (100 訓練)	0.83	0.35	0.63	1.65	0.85	1.20
ケース 2 (300 訓練)	0.56	0.21	0.46	1.20	0.61	0.89

さらにケース 2 (300 答案訓練) の場合，いずれの問題においても QWK が 0.90 以上と高い値となった．また表 4，表 5，表 6，表 7 よりアノテーションの予測精度について，各評価指標が半数以上の BERT モデルにおいて 0.90 以上と高い値となった．その一方で問題 Q3 の採点基準 D における BERT モデルの適合率，再現率，F 値が，他のモデルよりも特に低い値になっている．これは問題 Q3 の採点基準 D における採点の根拠が，その他の採点基準と異なり，「悔しい」や「悩む」といった人間の心情にあるため，BERT をもってしても人間の 心情を表す表現を，その意味を含んだ数値ベクトルに変換することが困難であったのではないかと推察している．そして表 2，表 3，表 8，表 9 より自動で推定されたアノテーションを使用した場合の自動採点の性能について，正解のアノテーションが与えられた場合とほぼ同等の QWK，MSE となった．自動で推定されたアノテーションは人手のそれと完全には一致していないが，アノテーションされた単語の埋め込みベクトルを寄せ集める際に，次元ごとの最大値をとるという操作を行っているため，実質的にほぼ同じ出力が得られたと解釈される．よって正解のアノテーションが与えられていない新規の採点課題であっても，100 答案あるいは 300 答案といった少量の答案に人手でアノテーションを付与して訓練データを作成すれば，3.2 項で示したモデルを用いて残りの答案のアノテーションを予測することができ，そして，正解のアノテーションが与えられた場合と同等の性能で自動採点が可能であるといえる．

5. おわりに

本報告では BERT と LightGBM を用いて，理研記述問題採点データセット中の 1 つの問題につき，100 答案程度の訓練データで，人間と同等程度の採点精度得ることができるということを明らかにした．さらに 300 答案程度の訓練データで，いずれの問題においても評価指標 QWK が 0.90 以上と高い値になることも明らかにした．今後，採点基準について加点項目しか扱っていないため，減点項目を含めるなどの改良を行っていきたい．

謝辞

本研究は，国立情報学研究所の IDR データセット提供サービスにより国立研究開発法人理化学研究所から提供を受けた「理研記述問題採点データセット」を利用した．科研費 (17H06107 および 19K02999) の助成を受けた．

参考文献

[1]乾健太郎 (2021). 「記述式答案の採点・評価を支援する言語処理技術」(全国的な学力調査の CBT 化検討ワーキンググループ第 7 回) . https://www.mext.go.jp/content/20210302-mxt_chousa02-000013129-3.pdf (参照 2021.6.10.)

[2] Tomoya Mizumoto, Hiroki Ouchi, Yoriko Isobe, Paul Reisert, Ryo Nagata, Satoshi Sekine and Kentaro Inui. “Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring”. In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 14)*. pp. 316-325, August 2019.

[3] 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*. pp.4171-4186, June 2019.

[4] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye and Tie-Yan Liu. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. In: *Advances in Neural Information Processing Systems*. pp.3149-3157, December 2017.

[5] 理化学研究所 (2020) ： 理研記述問題採点データセット．国立情報学研究所情報学研究データレポジトリ．データセット：https://doi.org/10.32130/rdata.3.1.

[6] 東北大学乾研究室.「Pretrained Japanese BERT models」. https://github.com/cl-tohoku/bert-japanese (参照 2021.6.10.)