E-005

連続的なスコアを推論する自動採点モデル構築のための基礎検討

Basic study for developing an automatic scoring model to infer continuous scores

春日 優虎† 浦野 昌一†
Yuto Kasuga Shoichi Urano

1. はじめに

大学入学共通テストにおいて記述式問題の導入が検討されるなど，学習者の思考力・表現力を計測するための試験改革が随所で進められている．しかしながら，大学入学共通テストにおいては，採点コストの肥大化と，それに伴う外部委託への情報漏洩の観点から議論が重ねられており，現在もなお導入が見送られているという現状がある．特に採点コストの肥大化という観点においては，いかにして採点の正確性・公平性を担保しつつ採点に要するコストを削減するかが重要である．

筆者らはこれまでに採点コスト削減のための手法として，BERT の分類モデルを用いた記述式問題の自動採点に関する研究を行ってきた．しかしながら現状では，「不正解」，「部分正解」，「正解」の 3 値分類タスクとして採点モデルを構築しており，部分正解に分類された答案については目視での検査が必要となっていた．分類モデルとしての精度は高かったものの，部分正解ラベルへ分類される答案が多く，その度に目視による採点を行う必要があるため，採点コスト削減の観点からは有用な結果が得られずにいた[1]．

そこで本稿では，3 値分類から拡張して，0 点から満点に至るまでの全ての点数に答案を振り分けることのできる採点モデルを実現するための基礎検討を行なう．

2. BERT を用いた記述式自動採点モデル

BERT[2]とは、Google が開発した大規模言語モデル(LLM)の一種であり、Transformer[3]の Encoder モデルの一種である。BERT は Transformer が持つ Attention 機構と呼ばれる仕組みにより文章データから文意を反映させたベクトル(分散表現)を得ることができる。本研究では BERT モデルの構築には Haggingface の Transformers を使用し，事前学習モデルは東北大学乾研究室の BERT 日本語モデルを使用している．以下の手順により分散表現を取得する．

中間点(0 点と満点の中間値)と満点，0 点のラベルを持つ学習データを用いて BERT 分類モデルをファインチューニングし，3 値分類器を作成する．続いて，上記 3 ラベルを含むすべての得点のデータを学習済みモデルに入力し，BERT から出力される文頭トークン[CLS]のみを取り出す．

得られた分散表現を 3 次元に次元圧縮し，可視化した結果を図 1 に示した．図 1 から，0 点から中間点にかけて，また中間点から満点にかけて直線関係が読み取れたため，本稿ではこの特性を活かして，連続的な得点を推論するモデルを構築していく．

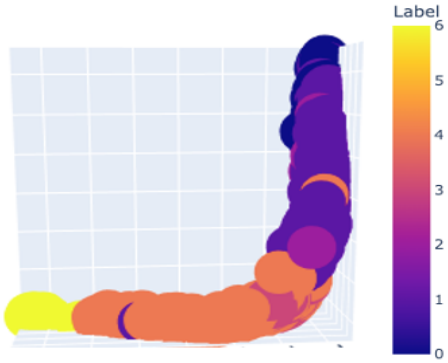

図 1　3 次元空間上における連続性

3. 提案手法

3.1 分散表現の抽出

N件の学習データ$\boldsymbol{y_i}$ $(i = 1, 2, 3, \cdots, N)$により学習を行なった BERT モデルへ学習データを再入力することで，分散表現を出力し，それを PCA によりK次元に圧縮して得られる分散表現を$\boldsymbol{y'}_i$とする．

$$\boldsymbol{y'}_i = \mathrm{PCA}\left(\mathrm{BERT\_MODEL}(\boldsymbol{y_i})\right) \tag{1}$$

3.2 各クラスの中心座標の定義

学習データ$\boldsymbol{y_i}$に対応する教師ラベルをl_iとする．l_iはそれぞれ 0, 1, 2 の値を持っており，それぞれ 0 点，中間点，満点を表すとする．l_iおよび$\boldsymbol{y'}_i$を用いて，(2)式～(3)式のように 0 点，中間点，満点クラスの中心座標を求める．

$$\boldsymbol{center}_{\boldsymbol{zero}} = \frac{\sum_{i\in\{i|l_i=0\}} \boldsymbol{y'}_i}{\sum_{i\in\{i|l_i=0\}} 1} \tag{2}$$

$$\boldsymbol{center}_{\boldsymbol{middle}} = \frac{\sum_{i\in\{i|l_i=1\}} \boldsymbol{y'}_i}{\sum_{i\in\{i|l_i=1\}} 1} \tag{3}$$

$$\boldsymbol{center}_{\boldsymbol{full}} = \frac{\sum_{i\in\{i|l_i=2\}} \boldsymbol{y'}_i}{\sum_{i\in\{i|l_i=2\}} 1} \tag{4}$$

また，2 つの距離$\boldsymbol{d_{low}}$およびd_{high}を(5)式～(6)式のように定める．

$$\boldsymbol{d_{low}} = \left\|\boldsymbol{center}_{\boldsymbol{zero}} - \boldsymbol{center}_{\boldsymbol{middle}}\right\| \tag{5}$$

$$d_{high} = \left\|\boldsymbol{center}_{\boldsymbol{full}} - \boldsymbol{center}_{\boldsymbol{middle}}\right\| \tag{6}$$

3.3 座標間距離の算出

M件のテストデータ$\boldsymbol{x_j}(j = 1,2,3,\cdots,M)$についても学習データと同様の手順により分散表現$\boldsymbol{x'}_j$を得る．

$$\boldsymbol{x'}_j = \mathrm{PCA}\left(\mathrm{BERT\_MODEL}(\boldsymbol{x_j})\right) \tag{7}$$

† 明治大学大学院 先端数理科学研究科 ネットワークデザイン専攻
Meiji University Graduate School of Advanced Mathematical Sciences Network Design Program

テストデータの分散表現$\boldsymbol{x'_j}$と 3 クラスの中心座標$\boldsymbol{center_{zero}}$, $\boldsymbol{center_{middle}}$, $\boldsymbol{center_{full}}$の間のユークリッド距離をそれぞれ計算することで，$\boldsymbol{x'_j}$を(8)式で示す 3 次元ベクトルに変換する．

$$\boldsymbol{a_j} = [a_{j,zero} \quad a_{j,middle} \quad a_{j,full}] \tag{8}$$

例えば$a_{j,zero}$は以下のように計算される．

$$a_{j,zero} = \sqrt{\sum_{k=1}^{K} \left(x_{j,k} - center_{zero,k}\right)^2} \tag{9}$$

3.4 未定義得点の補完

0 点，中間点，満点の間の得点を含め，テストデータ$\boldsymbol{x_j}$に対する予測スコアz_jを(10)式で定義する．ここで，満点をS点とする．

$$z_j = \begin{cases} \mathrm{round}\left(\frac{S}{2.0} \times \frac{a_{j,zero}}{d_{low}}\right) & \text{if } j \in \{j | \max(\boldsymbol{a_j}) = a_{j,full}\} \\ \mathrm{round}\left(S - \frac{S}{2.0} \times \frac{a_{j,full}}{d_{high}}\right) & \text{if } j \in \{j | \max(\boldsymbol{a_j}) = a_{j,zero}\} \\ -1 & others \end{cases} \tag{10}$$

なお，$z_j = -1$となった場合は「採点不可」とし，目視確認を行うものとする．

4. シミュレーション

4.1 使用データセット

本稿では，理化学研究所が提供する記述問題データセット「Y14_1-2_1_3」を使用する[4]．本データセットには 4 つの採点項目(A～D)が設けられており，それぞれに対する得点が与えられているが，本稿では採点項目 D のみを扱うこととする．採点項目 D は 6 点満点の採点項目であり，受験者の答案に応じて 0 点から 6 点までの 7 段階で得点がアノテーションされている．

4.2 シミュレーション条件

本シミュレーションでは，2100 件のデータセットから，0 点，中間点(3 点)，満点(6 点)がつけられたデータを不均衡になりすぎないよう 1083 件抽出し，これを学習データ 920 件，テストデータ 163 件に分割した(ここで得られたテストデータをテストデータ A とする)．また，0 点，3 点，6 点以外の得点が付与されたデータセット 607 件をテストデータ B とする．

本稿では，以下の 2 つのシミュレーションを行う．

4.2.1 シミュレーション A

学習済み 3 値分類モデルにテストデータ A を入力し，得られた分散表現を PCA により 60 次元に圧縮する．これらの分散表現を 3.3 項に示すような距離を成分とするベクトルに変換し，最も値が小さいクラスを予測クラスとする．このシミュレーションにより，3 値分類モデルとしての精度と，距離を特徴量とすることの有用性を検証する．

4.2.2 シミュレーション B

学習済み 3 値分類モデルにテストデータ B を入力し，得られた分散表現を PCA により 60 次元に圧縮する．これらの分散表現を用いて 3.3 項および 3.4 項に示すような手法を用いることで，学習データに含まれていない 1 点，2 点，4 点，5 点への分類を行う．シミュレーション B により，本稿の目的である連続的なスコアの推論精度を計測する．

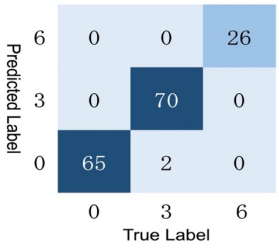

図 2　混同行列 A

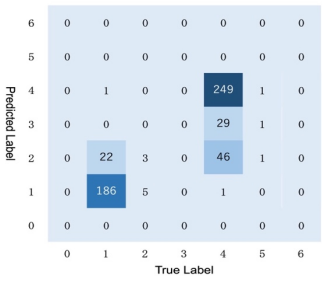

図 3　混同行列 B

4.3 シミュレーション結果・考察

シミュレーション A では，各クラスに属する学習データの中心座標との距離を測定し，最も距離が小さいクラスを予測ラベルとする分類手法により 0 点，3 点，6 点のテストデータに対する自動採点を行なった．結果として，図 2 に示すように非常に高精度な分類を行うことができた．このことから，ラベル間の距離に関する情報を採点モデルに用いることの有用性は示されたといえる．

シミュレーション B では，3 値分類器としてファインチューニングしたモデルに対し，学習に使用しないラベルを正解ラベルとして持つテストデータを入力して得られた分散表現が，図 1 に示すように 3 次元空間上で連続性を持っていたことに着想を得，学習データの中心座標間および学習データとの間の距離を用いて得点の予測を行なった．結果として，図 3 から読み取れるように，1 点および 4 点の多くの答案を正しく採点することができた．また誤った採点が行われた答案についても，多くは正解ラベルと近しい得点がつけられていることが読み取れる．なお，シミュレーション B において「採点不可」がつけられた答案数は 607 件中 62 件となり，約 90%を自動採点することができた．

5. おわりに

本稿では，3 値分類モデルから出力される分散表現間の距離情報を用いることで，連続的な得点を予測する採点手法を提案した．テストデータ B のラベルの偏りから，精度は高くないものの，混同行列からは一定の妥当性は認められた．しかし，自動化できる答案の件数を先行研究と比較して飛躍的に増加させることができた一方で，誤分類の件数が増加されてしまったため，今後は自動化できる件数を維持しつつ採点の精度をより向上させていくことを目指す．

参考文献

[1] 春日優虎,浦野昌一：「Back Translation による小規模データセット拡張手法を用いた記述式問題自動採点モデルの構築」,人工知能学会全国大会論文集 4Xin2-17 (2024)

[2] Jacob Devlin, Ming-WeiChang, Kenton Lee, Kristina Toutanova, “*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*”, in Proc. NAACL-HLT 2019, pp. 4171–4186 (2019)

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, “*Attention is All you Need”*, arXiv, 1706.03762 (2017)

[4] 理化学研究所, "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 (2020)