記述式答案採点モデルの採点基準に対する整合性の検証

浅妻佑弥 [1,2]　舟山弘晃 [1,2]　松林優一郎 [1,2]　水本智也 [2]　乾健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所
{asazuma.yuya.r7,h.funa}@dc.tohoku.ac.jp
{y.m, kentaro.inui}@tohoku.ac.jp　tomoya.mizumoto@a.riken.jp

概要

記述式答案自動採点タスクにおいて，採点基準と合致する採点モデルを構築することは重要な要件であるが，訓練済みのモデルに対して採点基準との整合性を効率的に検証する手段は確立されていない．有力な手法として，モデルの内部動作を可視化できる特徴量帰属法 (Feature Attribution) が存在するが，解答毎に検証を行う必要があるため，多数のデータを使用した上で整合性を証明することは困難だった．本研究では，クラスタリングアルゴリズムを特徴量帰属法で求めた説明系列に適用することで，少ない労力で採点モデルと採点基準間の整合性検証を可能にした．

1　はじめに

記述式答案自動採点 (Short Answer Scoring) とは，ある問題に対して解答された数十文字程度の答案を自動で採点するタスクである [1]．機械学習モデルで解ける形式に落とし込む場合，採点結果を目的変数に設定し，答案の文章を説明変数とする回帰問題として扱い，問題文や採点基準（ルーブリック）は補助的な特徴量として扱うのが一般的である [2]．

しかしながら，本来の採点過程では採点基準との合致が求められる．ゆえに，採点基準に沿った答案箇所が主な得点源となり，モデルが利用する特徴量となることが求められるが，現状の多くの取り組みでは答案の文章が主たる入力であり，学習の過程で使用する特徴量が決定されるため，基準に沿った採点をモデルが行うことを保証することはできない．たとえ，補助的な特徴量として採点基準を導入しても，機械学習モデルが利用するかは学習によって決定されるため，正確な保証になるわけではない．

この不安定性の問題に対応する手段として，特徴量帰属法 (Feature Attribution) の活用が考えられる．主に説明可能な AI (XAI) 分野において扱われ，出力に対する入力特徴量の寄与度をモデルを使用して計算する手法群の総称である．モデルが使用した特徴量を根拠箇所としてヒートマップの形式で可視化することが可能であり，モデルの動作や判断の手がかりを得るための手法として定評がある [3]．

しかしながら，特徴量帰属法は答案毎にベクトルの形式で寄与度を計算するため，自然言語で記述された採点基準とは形式が異なり，単純な自動化は難しい．また，可能な限り多くの答案で検証できることが望ましいが，訓練に使用する答案の数は最低でも数百・数千に及ぶため，全ての答案に対する検証を行うことは困難である．

本研究では，特徴量帰属法を使用した採点基準との整合性検証が抱える実務上の問題点を解消するために，スペクトラルクラスタリング [4] を使用したクラスタ分析による整合性検証ツールの開発を行った．そして，考案した手続きの有効性を示すため，採点基準を含むデータセット上で動作確認を行い，少ない労力で訓練済みの機械学習モデルと採点基準間の整合性検証が可能であることを確認した．

2　関連研究

記述式答案採点モデルの解釈を目的とした既存研究は限られている．Mizumoto ら [5]，Zeng ら [6] は注意機構 [7] の重みをモデルの解釈として提供しているが，モデルの一部の特徴量でしかないため，採点基準の検証手段としての使用には適さない．

一方で，Tasuku ら [8] は特徴量帰属法で生成した説明系列と採点基準を元としたアノテーションの重複率を計測することで，モデルが採点基準に沿った動作を行うか検証している．しかしながら，検証したい全てのサンプルに対して採点基準を元としたアノテーションを実施する必要があり，新規のデータセットに適応するための障壁が高い．そのため，少ない労力で採点基準との整合性を検証できる手法を本研究で取り扱う．

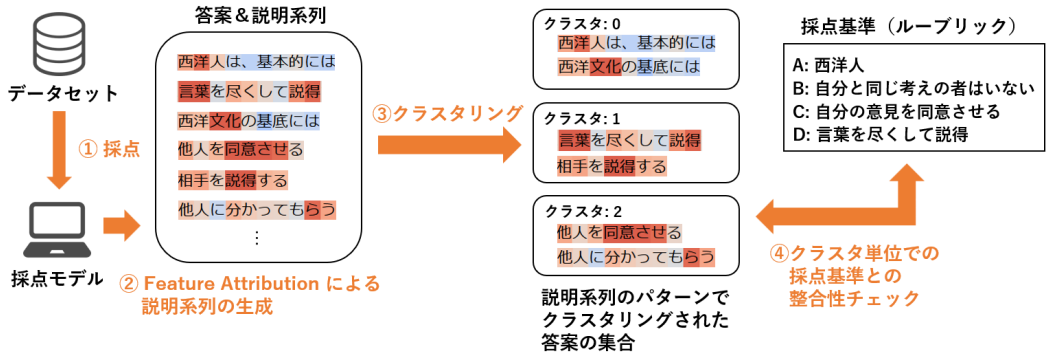

図 1　本研究の概略図．訓練に使用した答案に対して特徴量帰属法による説明系列を生成し，クラスタリングによって類型の答案を集約する．各々のクラスタに対して採点基準と比較することで整合性のチェックを行う．

3　手法

3.1　特徴量帰属法

特徴量帰属法 (Feature Attribution)[1)] はモデルの出力に対する入力特徴量の寄与度を計算するための手法である．入力サンプルの集合を $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$，機械学習モデルを $f(\boldsymbol{x})$ とする．あるサンプル $\boldsymbol{x} = (x_1, x_2, ..., x_m)$ に対するモデルの出力得点が $\boldsymbol{y}$ であるとき，出力得点に対する入力系列の重要度を説明系列 $\boldsymbol{e} = (e_1, e_2, ..., e_n)$ として計算する．各 e_i の値は得点に対する x_i の寄与度となることが望まれる．

本研究では，寄与度の計算に Integrated Gradients [9] を使用する．勾配計算を利用する解釈手法 [10] の中でも有力な手法であり，モデルと入出力を利用して以下の式で計算する．

$$e_i = (x_i - x_i') \int_0^1 \nabla f(\boldsymbol{x}' + \alpha(\boldsymbol{x} - \boldsymbol{x}'))\mathrm{d}\alpha \qquad (1)$$

ここで，$\boldsymbol{x_i'}$ はベースラインと呼ばれる比較の基準となる入力サンプルであり，本研究では零テンソルを使用する．摂動ベースの手法 [11][12] と比較して計算量が少なく，微分可能であれば大規模なモデルでも実行可能であるため，本研究で採用する．

3.2　SpRAy

SpRAy[13] は 特徴量帰属法 による解釈の分析を行う手法である．論文内において，PASCAL VOC 2007 に対して学習された分類モデルが，馬カテゴリの分類に画像の透かし表記を利用することを明らかにしている．本研究では，SpRAy を自然言語処理領域に拡張し，記述式答案自動採点タスクで利用する．

1)　本論文の執筆時には合意ある日本語名称が存在しない．理解の為に，論文内では特徴量帰属法の名称を使用する．

SpRAy の主な特徴として，スペクトラルクラスタリングを使用したクラスタ分析を行う．ラプラシアン行列の固有値分解によるグラフの連結成分分解の問題を解くことでクラスタリングを行う手法であり，説明系列の集合を k 個のクラスタに分類するとき，アルゴリズムは以下の 4 ステップで実行される．

1. 任意の二つの説明系列 $\boldsymbol{e_a}$ と $\boldsymbol{e_b}$ の関係性から親和性行列 A を構築する．
2. A をグラフ構造に帰着し，対称正規化ラプラシアン行列 L_{sym} を求める．
3. L_{sym} の固有値分解を行う．固有値の昇順に k 個の固有値ベクトルを選択して行列 E を作る．
4. E に対して素朴なクラスタリング手法を実施する．本研究では k-means 法 [14] を使用する．

ここで，固有値を昇順に整列した系列の前後の差分 $\lambda_i^{gap} = \lambda_{i+1} - \lambda_i$ は固有値ギャップと呼称され，最適なクラスタの数を推定できるヒューリスティックな指標として扱うことができる．[15]

4　実験

本章で，記述式答案自動採点データセットを使用した検証実験を行い，少ない労力でモデルと採点基準間の整合性検証が可能であることを示す．

4.1　データセット

本研究では，理研記述問題採点データセット [16][5][17] を使用して実験を行った．このデータセットは，入力とする答案文章，採点者による得点，採点の根拠を示すアノテーションから構成される．データセット内には複数の問題文に対するデータが含まれているが，本実験では問題 Y14_1-2_1_3 を使用する．得点は複数の採点項目によって構成さ

表 1 問題 Y14_1-2_1_3 の採点項目．4 つの採点項目から構成され，合算した点数が最終的な得点となる．

項目	配点	満点となる文章スパンの例
A	2	”西洋”，”西洋人”
B	5	”自分と同じ考えの者はいない”
C	3	”自分の意見を同意させるため”
D	6	”言葉を尽くして説得し”

れる．表 1 に，採点項目の詳細を示す．

4.2 機械学習モデル

実験のために，水本ら [5] の研究を基にした機械学習モデルを構築する．エンコーダ層に BERT[18] を導入し，デコーダ層に注意機構 [7] を使用する．学習済みの BERT モデルには文字単位の日本語 BERT モデル[2)]を使用する．後の実験のために，採点項目毎の部分点をベクトル形式で予測する個別採点モデルと，採点結果を合算した点数をスカラー形式で予測する全体採点モデルを構築する，得点を予測する回帰問題としてモデルを訓練した．モデルの性能評価は付録 A に記載する．

4.3 クラスタ分析の前処理

モデルの学習後，使用した訓練サンプルに対して Integrated Gradients による説明系列を生成する．ただし，埋め込み層に対する勾配の計算は不可能であるため，埋め込み後の入力に対して寄与度を計算する．クラスタリングにあたり説明系列間の類似性から親和性行列を構築する必要があるが，入力が可変長系列であるため，全ての入力に対して総和を取った説明系列間でコサイン類似度を計算する．親和性行列からラプラシアン行列への変換は，5 つの近傍の頂点を接続行列として扱い，対称正規化ラプラシアン行列を構築する．また，予測スコアが最大の点数の二割以下であるサンプルについては除外する．白紙の答案に対して採点理由を求めることが不可能であるように，得点が著しく低い答案に対する説明を求めることが困難なためである．

4.4 全体採点モデル

まず，全体採点モデルに対してクラスタ分析を行う．ラプラシアン行列に対して求めた固有値の特性を図 2 に示す．λ_1・λ_2 の固有値ギャップが大きく，λ_5 以降は緩やかに増加することが確認できる．この

2） https://github.com/cl-tohoku/bert-japanese

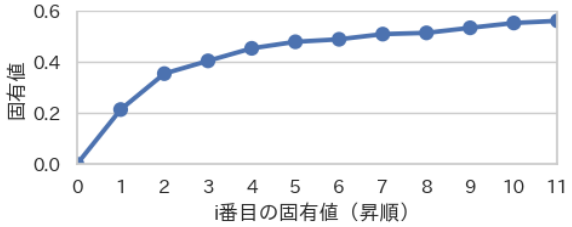

図 2 全体採点モデルにおける固有値の推移．

結果より，クラスタ数を $k = 5$ に設定してクラスタ分析を行った．ヒートマップで色付けしたクラスタ毎の文章例を図 3 に示す．なお，ヒートマップの濃度は，トークン毎に説明系列の総和を取り標準化を施した値を使用した．概ね，各クラスタに含まれる顕著な特徴は以下の結果になった．

- クラスタ 0, 4: 西洋 (項目 A)，対決 (項目 B)
- クラスタ 1, 2: 同意 (項目 C)，説得 (項目 D)
- クラスタ 3: 類似性の少ない答案の集合

3 番以外のクラスタで，表 1 に示す採点基準を中心に集約される傾向が確認できた．クラスタ 3 のみ，類似性が少ない答案と説明系列が集約されている．ゆえに，3 番以外のクラスタに属する訓練サンプルに関しては基準との整合性あると結論付けられる．クラスタ 3 に所属するサンプルのみ，個別に検証することになるが，今回の実験では問題のある事例は発見できなかった．

以上の手順によって，少ない労力で採点基準との整合性を確認できた．なお，採点基準に反するような説明系列を発見することはなかった．

4.5 個別採点モデル

4.4 節と同様の手順で，個別採点モデルに対するクラスタ分析を行った．分析自体は全ての採点項目に対して実行したが，本文には採点項目 B における結果を記載する．ヒートマップで色付けしたクラスタ毎の文章例を図 4，ラプラシアン行列に対して計算した固有値の特性を図 5 に示す．λ_5 以降は緩やかに増加することが確認できるため，クラスタ数は $k = 5$ に設定した．図 4 において，各クラスタに含まれる顕著な特徴は以下の結果になった．

- クラスタ 0: 西洋文化の基底 (負)，対決，異人
- クラスタ 1:「対決」, 答案が類似する．
- クラスタ 2: 西洋 (負)，異人，人間
- クラスタ 3: 西洋 (負)，他人は自分とは異なる
- クラスタ 4:「対決」,「異人」

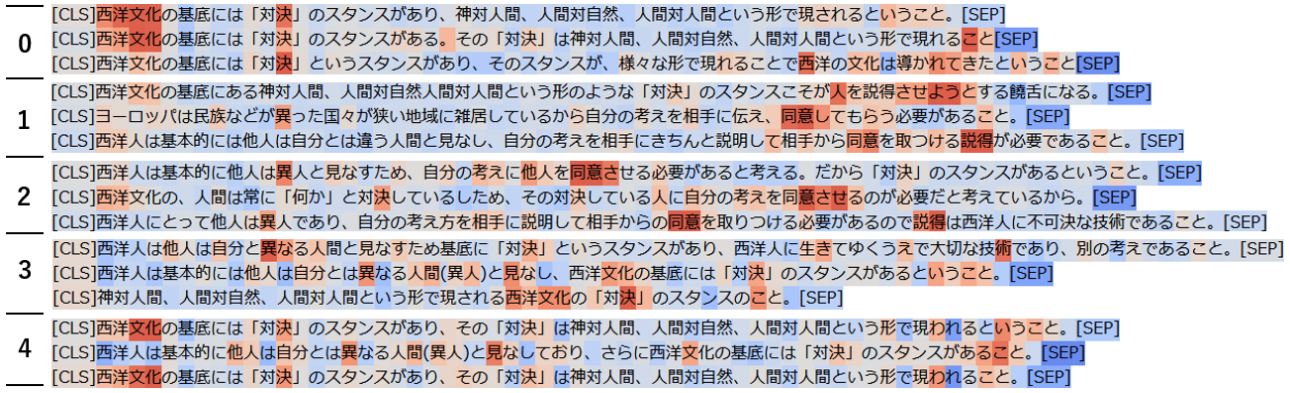

図 3　全体採点モデルで生成した説明系列を対象としてクラスタ分析を行った結果．各クラスタに含まれる文章と説明系列を三件ずつ例示する．文頭の番号は各クラスタの ID を示す．

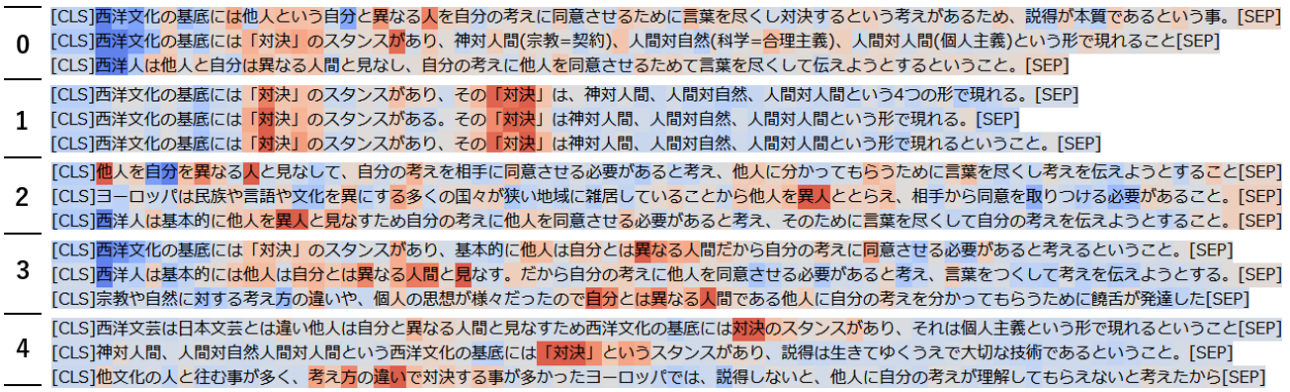

図 4　個別採点モデル・採点項目 B で生成した説明系列を対象としてクラスタ分析を行った結果．各クラスタに含まれる文章と説明系列を三件ずつ例示する．文頭の番号は各クラスタの ID を示す．

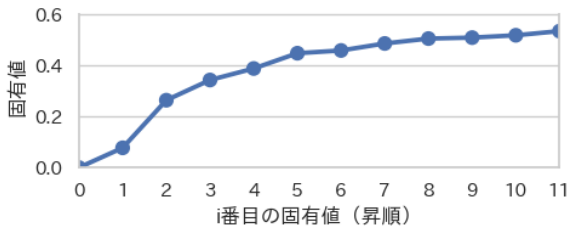

図 5　個別採点モデル・項目 B における固有値の推移．

全体採点モデルの時と同様に，一貫した説明のパターンをクラスタ内で確認できる．採点項目 B に該当する表現に高い寄与度が割り振られているパターンが多いが，一方で採点項目 A であるはずの”西洋文化”に対して強い負の寄与度が確認できる．ゆえに，採点項目 B に含まれない基準”西洋文化”を使用して答案を採点している可能性が浮上した．

5　議論

あくまで固有値ギャップはヒューリスティックな指標であるため，依然として最適なクラスタ数を設定することは困難である．大きなクラスタ数を指定すれば詳細な分割が行われるが，検証の労力が増大しクラスタリングの意義を損ねてしまう．一方で，小さすぎるクラスタ数を指定すれば，一つのクラスタに複数のパターンが含まれることになる．現に，4.4 節における実験では，一つのクラスタ内に複数の採点項目のパターンが観測された．そもそも，固有値ギャップ特性によるクラスタ数の推定に対して批判的な意見も存在する [19]．

また，本論文の実験結果は人間視点の定性的な評価に留まっているため，今後の研究ではクラスタ間の距離・階層関係を利用した定量的な評価の実施を計画している．

6　おわりに

本研究では，記述式答案自動採点タスクにおける採点基準検証の省労力化を目的として，特徴量帰属法とクラスタリングを使用した検証手法を開発した．有効性を示すために，採点基準を含むデータセット上で動作確認を行い．訓練済みの機械学習モデルに対して整合性検証が可能であることを確認した．また，実際に採点基準から外れた動作の疑いがある事例を発見することができた．

謝辞

本研究は JSPS 科研費 22H00524，JST 次世代研究者挑戦的研究プログラム JPMJSP2114 の助成を受けたものである．また，研究を進めるにあたり，頻繁に議論に参加していただいた Tohoku NLP グループの皆様へ感謝いたします．

参考文献

[1] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. Fast and easy short answer grading with high accuracy. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1070–1075, San Diego, California, June 2016. Association for Computational Linguistics.

[2] Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. Investigating neural architectures for short answer scoring. In **Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 159–168, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[3] Ričards Marcinkevičs and Julia E Vogt. Interpretability and explainability: A machine learning zoo mini-tour. December 2020.

[4] Andrew Y Ng, C S Division, U C Berkeley, Michael I Jordan, C S Div, Dept Of, Stat U C Berkeley, and Yair Weiss. On spectral clustering: Analysis and an algorithm. https://proceedings.neurips.cc/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf. Accessed: 2023-1-11.

[5] 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. pp. 316–325, August 2019.

[6] Zijie Zeng, Xinyu Li, Dragan Gasevic, and Guanliang Chen. Do deep neural nets display human-like attention in short answer scoring? In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 191–205, Seattle, United States, July 2022. Association for Computational Linguistics.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. September 2014.

[8] Tasuku Sato, Hiroaki Funayama, Kazuaki Hanawa, and Kentaro Inui. Plausibility and faithfulness of feature Attribution-Based explanations in automated short answer scoring. In **Artificial Intelligence in Education**, pp. 231–242. Springer International Publishing, 2022.

[9] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. March 2017.

[10] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. December 2013.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. February 2016.

[12] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. May 2017.

[13] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. **Nat. Commun.**, Vol. 10, No. 1, p. 1096, March 2019.

[14] J MacQueen. Some methods for classification and analysis of multivariate observations. In **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics**, Vol. 5.1, pp. 281–298. University of California Press, January 1967.

[15] Ulrike von Luxburg. A tutorial on spectral clustering. November 2007.

[16] 理化学研究所. 理研記述問題採点データセット. July 2020.

[17] Hiroaki Funayama, Tasuku Sato, Yuichiroh Matsubayashi, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. Balancing cost and quality: An exploration of Human-in-the-Loop frameworks for automated short answer scoring. In **Artificial Intelligence in Education**, pp. 465–476. Springer International Publishing, 2022.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. pp. 4171–4186, June 2019.

[19] Song Wang, Karl Rohe, Pengsheng Ji, and Jiashun Jin. DON'T MIND THE (EIGEN) GAP. https://www.stat.cmu.edu/~jiashun/Research/Selected/SCC-disc3.pdf. Accessed: 2023-1-13.

A　採点モデルの性能評価

採点モデルを訓練した後に，テストセットを使用した性能の評価を行った．RMSE と QWK の二つの指標で測定した結果を記載する．

A.1　全体採点モデル

全体採点モデル，テストセットに対する性能評価の結果を表 2 に示す．

A.2　個別採点モデル

個別採点モデル，テストセットに対する性能評価の結果を表 3 に示す．採点項目ごとに性能の測定を行った．

表 2　全体採点モデル・テストセットにおけるモデルの性能．問題 Y14_1-2_1_3 を使用．

採点項目	RMSE	QWK
ALL	0.068	0.948

表 3　個別採点モデル・テストセットにおける採点項目毎の性能．問題 Y14_1-2_1_3 を使用．

採点項目	RMSE	QWK
A	0.060	0.999
B	0.088	0.915
C	0.054	0.994
D	0.047	0.990