自動採点モデルにおける Attention 層に着目した採点根拠の分析

宮田創太 [1]　横井健 [1]
[1] 東京都立産業技術高等専門学校
public.miya0169@gmail.com　takeru@metro-cit.ac.jp

概要

本研究は、採点根拠の指標として利用できる Attention 層に着目し、自動採点モデルの採点方針を探り、解答文の特徴を分析することを目指す。提案手法として、解答文のクラスタリング・Attention 層の出力ベクトルのクラスタリングを通して、採点者・自動採点モデルの採点方針を可視化した。その後、クラスタリング結果と採点結果をもとに、採点者・自動採点モデルから見た解答文の特徴分析を行った。具体的には、頻出単語の統計、Attention の重みと SHAP 値の統計、スピアマンの順位相関係数を用いて解答文の特徴分析を行った。

その結果、解答文の特徴として、高得点の解答文は問題の話題に関連した単語を多く含み、低得点の解答文は抽象的な単語を多く含むことが分かった。また、採点者・自動採点モデルが具体性、語彙、単語のスペルに注目して採点をしていることが、単語で明らかにすることができた。

1　はじめに

教育現場での英語や国語の筆記試験において、記述式問題という出題形式がある。記述式問題は、プロンプト (問題、主題、トピック) に対して、1、2 文程度の短文や、長い小論文で解答する形式である。通常、この形式の問題の解答文を採点するときは、採点基準がいくつか存在する。記述式問題を採点するときの問題点として挙げられるのが、一問一答や選択肢問題と比べると採点する項目が多く、時間的コストが高い点である。そこで、自然言語処理を用いた自動採点技術の研究が行われている。

自動採点の理想は、日常教育に使用でき、解答者が納得する自動採点である。この理想を目指すための課題は、採点根拠の説明が必要なことである。具体的には、与えられた解答文に対し、自動採点モデルがなぜこの点数を付けたのか明らかにする必要がある、ということである [1]。採点者が個人で採点方針を持つように、自動採点モデルも何らかの採点方針を持っている。課題の解決策として、自動採点モデルが採点項目ごとに点数を予測すること、アノテーションを用いること、フィードバック文の生成が先行研究で挙げられている [2][3]。自動採点モデルは、通常ブラックボックス化されており、モデル内でどんな予測を行っているか分からない。そのため、採点根拠を説明するためには、まずはモデル内の予測過程を何らかの形で示し、自動採点モデルの採点方針を探る必要がある。

本研究では、採点根拠の指標として利用できる Attention 層に着目し、自動採点モデルの採点方針を探り、解答文の特徴を分析することを目的とする。

2　提案手法

まず、解答文のクラスタリングと、Attention 層の出力ベクトルのクラスタリングを行う。解答文のクラスタリングでは、解答文そのものを対象としてクラスタリングを行い、結果を出力する。結果から、採点者が解答文をどのように分類しているか観察する。Attention 層の出力ベクトルのクラスタリングも同様に行う。Attention 層の出力ベクトルをクラスタリングしてそのまま出力すると、どの解答文を入力として Attention 層が出力したベクトルか分かりにくい。そのため、あらかじめ Attention 層の出力ベクトルと採点済みの解答文を対応付けしながら、クラスタリング結果を出力するようにする。Attention 層の出力ベクトルは、単語ごとの重要度を含む解答文全体の情報を持つベクトルであり、クラスタリングによって、自動採点モデルの Attention 層が解答文をどのように分類しているか観察する。

最後に、それぞれのクラスタリング結果と、採点結果をもとに、解答文の特徴を分析する。クラスタ

ごとに解答文から、頻出する単語やフレーズを集計しランキングをする。解答文を対象としたクラスタリングでは、解答文の中からそのまま集計する。Attention 層の出力ベクトルを対象としたクラスタリングでは、SHAP[4] を利用して、各単語の予測への影響度を数値化し、解答文中で Attention の重みや SHAP 値がしきい値を超える部分の中から集計する。そして、クラスタどうしで比較し、頻出単語やフレーズ、順位相関係数を分析する。

3 実験

解答文のベクトル化には Word2Vec を使用し、自動採点モデルは Attention + LSTM のモデルを使用した。データセットには Automated Student's Assessment Prize（ASAP）データセット [5] を使用した。ASAP データセットは、小論文問題に対するアメリカの高校生の解答を集めたデータセットである。その中でも Prompt1 の解答文に対し実験を行った。おおまかな問題内容は「地元の新聞宛に手紙を書き、コンピュータが人々に与える影響についてあなたの意見を述べ、読者を説得せよ」である。主な採点基準は以下のとおりである。

- 具体的な説明ともに理由を提示しているか
- 文章構成がしっかりとしているか
- 流暢で、読みやすい言語を使っているか
- 読者を意識した記述がされているか

解答文（生データ）1 個あたりの平均単語数は 350、解答文（生データ）1 個あたりの最大単語数は 785、前処理後の解答文 1 個あたりの最大単語数は 383 である。また、解答数は 1,783、Score（点数）の範囲は 12〜2 である。このデータを 1:4 でテスト用データと学習用データに分けて使用した。クラスタリングについては以下のとおりに行った。

- 解答文のクラスタリング
  前処理を行った解答文に対し、K-means 法を用いてクラスタリングを行った。その後、t-SNE[6] を用いてクラスタリング結果を 2 次元空間上で表せるようにした。本実験では TF-IDF を用いて出した語彙の特徴量+単語数を示す特徴量を持つ解答文を 2 次元にした。
- Attention 層の出力ベクトルのクラスタリング
  前処理を行った解答文に対し、自動採点を行い点数を予測した後、Attention 層の出力ベクトルと解答文を対応付けし、K-means 法を用いてクラスタリングを行った。その後、t-SNE[6] を用いてクラスタリング結果を 2 次元空間上で表せるようにした。本実験では Attention 層の出力ベクトルを 2 次元にした。

クラスタ数 k は、シルエットスコアやエルボー法を行った結果を参考にして $k = 4$ に決定した。

各クラスタリング結果について、頻出単語の統計をまとめ、クラスタ同士で比較し、共通して出現している単語や、特出している単語を調べた。Attention 層の出力ベクトルのクラスタリングについては、以下の 2 つを追加でまとめ、予測に対する単語・フレーズの影響度を調べた。

- 0 を上回る、もしくは 0 を下回る SHAP 値を持った出現回数 5 回以上の単語・フレーズ Top20・Bottom20
- Attention の重み平均の高い出現回数 5 回以上の単語・フレーズ Top20

また、採点者の点数と解答文に含まれる単語数、予測した点数と解答文に含まれる単語数でスピアマンの順位相関係数をそれぞれ算出し、点数と単語数の相関関係について調べた。

4 結果と考察

解答文のクラスタリングをクラスタ数 $k = 4$ で行った結果を図 1 に、点数分布を図 2 に示す。Attention 層の出力ベクトルのクラスタリングをクラスタ数 $k = 4$ で行った結果を図 3 に、予測点数の分布を図 4 に示す。図 1〜図 4 より、どちらのクラスタリング結果も 4 つのクラスタに分かれており、またクラスタごとに点数分布が異なることが分かる。この結果から、今回のクラスタリングは採点に関連性があると考えられる。

解答文のクラスタリングにおける頻出単語をまとめた結果、time、friends、family という単語が共通してよく使われていた。最も高い点数分布の Cluster1 では、他に online や society、information といった問題の話題に沿った単語がよく使われていた。この理由は、コンピュータを使う時間が家族や友人と過ごす時間とトレードオフかどうかという話題や、コンピュータの有効性や将来性に触れているからであると考えられる。問題にもその話題について触れられている部分がある。

Attention 層の出力ベクトルのクラスタリングにおける、SHAP の分析について、Cluster0（高い点数

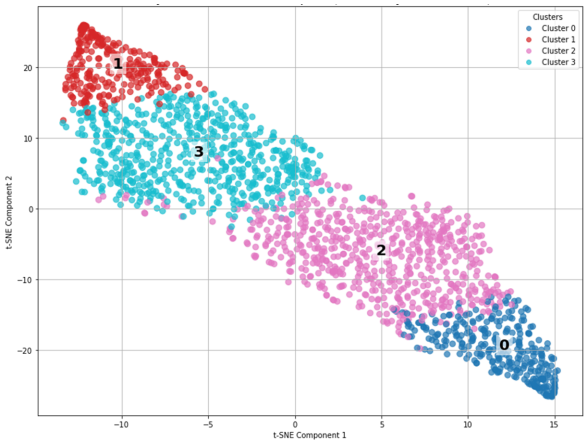

図 1　解答文のクラスタリング クラスタ数 $k = 4$

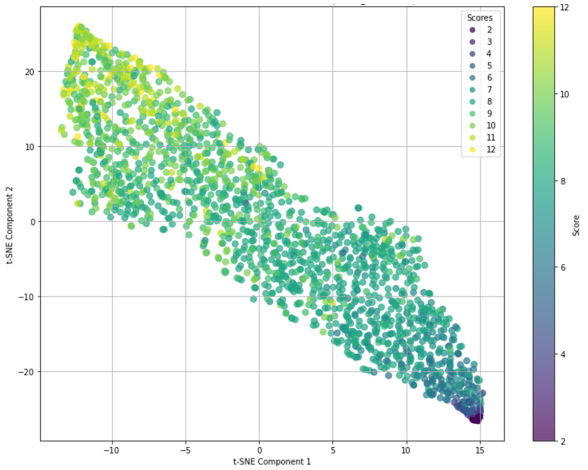

図 2　解答文のクラスタリング 点数分布

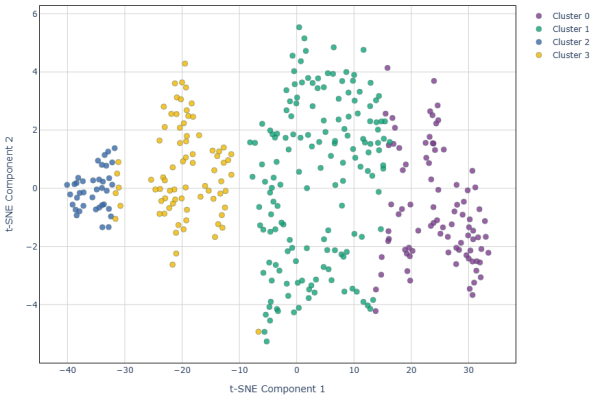

図 3　Attention 層の出力ベクトルのクラスタリング クラスタ数 $k = 4$

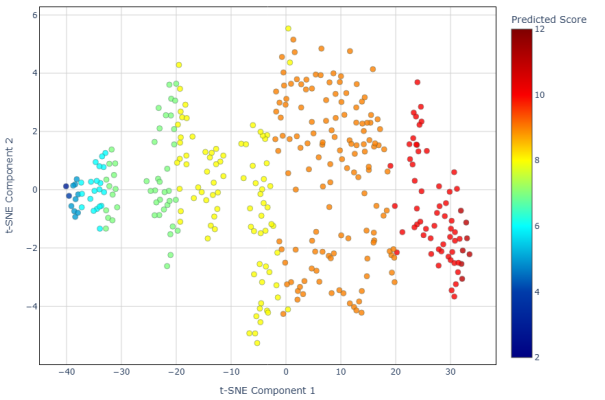

図 4　Attention 層の出力ベクトルのクラスタリング 予測点数分布

分布）の結果を図 5 に、Cluster2（低い点数分布）の結果を図 6 に示す。縦軸：単語名、横軸：SHAP 値の平均で示されている。青い棒グラフが Top20、赤い棒グラフが Bottom20 の結果である。図 5 から、advances、valuable、technology、science などの問題の話題に関連した、具体的な説明を持ちやすい単語がプラスに働いている一方、fun や stuff などの口語的な単語や、alot や soceity などのスペルミスをしている単語がマイナスに働いていることが分かる。図 6 から、fun や stuff などの口語的な単語や、bad や good などの具体的な説明を持ちにくい単語がマイナスに働いており、その SHAP 値の絶対値は図 5 と比べて大きいことが分かる。以上の結果から、自動採点モデルにとって、高得点の解答文は具体的な説明をもった単語・フレーズが多く述べられており、口語表現やスペルミスで減点されていると考えられる。一方で、低得点の解答文は抽象的な説明をもった単語・フレーズが多く述べられており、口語表現やスペルミスに限らず、読者に伝わりにくい文章で大きく減点されていると考えられる。

Attention 層の出力ベクトルのクラスタリングにおける、Attention の重み平均の分析について、Cluster0（高い点数分布）の結果を図 7 に、Cluster2（低い点数分布）の結果を図 8 に示す。Attention の重み平均の棒グラフと、該当する単語の SHAP 値平均の折れ線グラフを重ねて表示している。縦軸：単語名、棒グラフ横軸（グラフ下部）：Attention の重み平均、折れ線グラフ横軸（グラフ上部）：SHAP 値平均で示されている。図 7 から、calories、obese、reports などの問題の話題（コンピュータを使う時間が運動する時間とトレードオフかどうか）に関連した、専門的な単語が重く見られており、SHAP 値はプラスの単語が多いことが分かる。この結果から、自動採点モデルにとって、これらの単語が Attention 層で重く見られており、採点時に重要であると考えられる。図 8 から、nothing や something などの曖昧な表現の単語が重く見られており、SHAP 値はマイナスの単語が多いことが分かる。他のクラスタと比べる

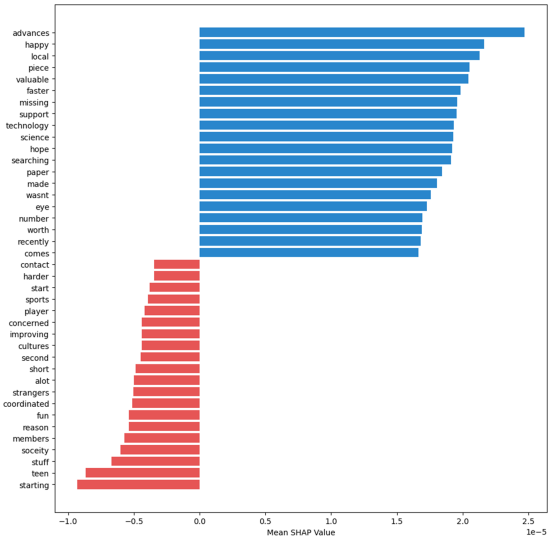

図 5　Attention の出力ベクトルのクラスタリングにおける Cluster0（高い点数分布）の SHAP 分析

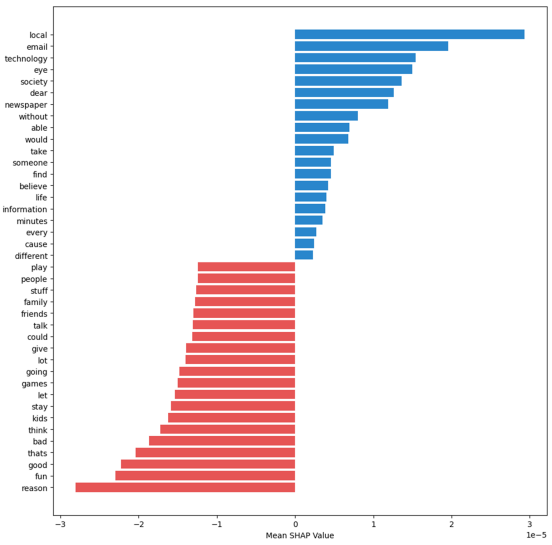

図 6　Attention の出力ベクトルのクラスタリングにおける Cluster2（低い点数分布）の SHAP 分析

と、Attention の重み平均 Top20 が 0.4 低いことが分かる。この結果から、低得点の解答文は、具体的な説明を持った単語が少なく、解答文に特徴がないために、Attention 層も重要だと感じる部分が観測できなかったと考えられる。また、他のクラスタにおいて、knowlege、reson などスペルミスも重く見られ、SHAP 値がマイナスになっていることから、採点時における減点対象だと考えられる。

スピアマンの順位相関係数について、採点者の点数と単語数では 0.81($p < 0.05$)、自動採点モデルが予測した点数と単語数では 0.95($p < 0.05$) となり、どちらの場合も正の相関があることが分かる。このことから、今回の ASAP[5] の問題においては、単

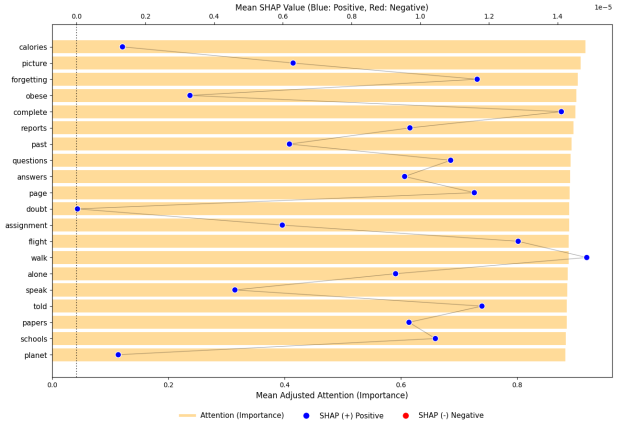

図 7　Attention の出力ベクトルのクラスタリングにおける Cluster0（高い点数分布）の Attention の重み分析

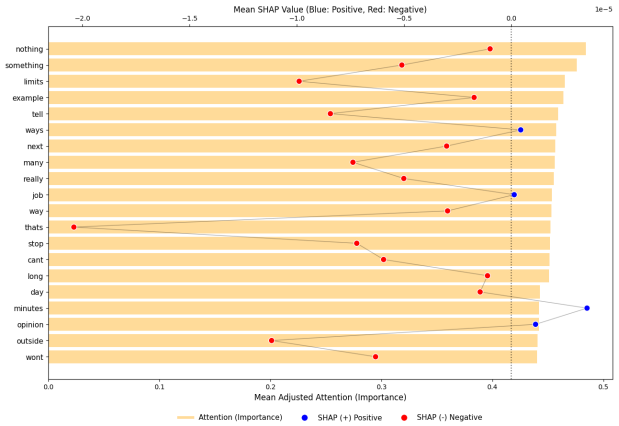

図 8　Attention の出力ベクトルのクラスタリングにおける Cluster2（低い点数分布）の Attention の重み分析

語数・文章量が多いほど、特徴量が多く、語彙力を持った文章と判断され、高得点が予測されやすいと考えられる。

5　まとめ

本研究は、採点根拠の指標として利用できる Attention 層に着目し、自動採点モデルの採点方針を探り、解答文の特徴を分析することを目指した。その結果、自動採点モデルが点数予測をする際の傾向を分類し、点数分布に沿っていることを確認することができた。採点者・自動採点モデルが具体性、語彙、単語のスペルに注目して採点をしていることが、単語やフレーズで明らかにすることができた。また、解答文の特徴として、話題に沿っているかという内容の是非だけでなく、文章の長さも点数分布に影響していることが分かった。

今後の展望として、さらなる重要単語・フレーズの整理、部分点など他の特徴量と組み合わせた場合のクラスタリングや分析などが考えられる。

参考文献

[1] Vivekanandan Kumar and David Boulanger. Explainable automated essay scoring:deep learning really has pedagogical value. **Frontiers in education**, Vol. 5, No. 572367, 2020.

[2] 佐藤汰亮, 舟山弘晃, 塙一晃, 浅妻佑弥, 乾健太郎. 根拠箇所に基づく自動採点結果の説明. 言語処理学会 第28回年次大会, pp. 459–464, 2022.

[3] Seong Yeub Chu, Jong Woo Kim, Bryan Wong, and Mun Yong Yi. Rationale behind essay scores: Enhancing s-llm's multi-trait essay scoring with rationale generated by llms. In **Findings of the Association for Computational Linguistics:NAACL 2025**, pp. 5796–5814, 2025.

[4] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In **Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)**, pp. 4768–4777, 2017.

[5] Kaggle. The hewlett foundation: Automated essay scoring, 2012. https://www.kaggle.com/c/asap-aes/data (参照 2025-11-27).

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. **Journal of Machine Learning Research**, Vol. 9, No. 11, pp. 2579–2605, 2008.