解釈可能性の高い自動採点モデルを用いた小論文採点支援システムの構築

水野友暉　竹内孔一
岡山大学大学院
psqs3ox3@s.okayama-u.ac.jp
takeuc-k@okayama-u.ac.jp

概要

本研究では，日本語の小論文採点を支援するシステムの開発を目的とし，モデルの解釈可能性向上を図るための手法を提案する．提案手法には，Attention，Masking Token，および Sparse Autoencoder の 3 つを提案する．特に Sparse Autoencoder を用いて特徴量ごとに概念を特定し，モデルの内部構造の解釈を可能にすることで，採点結果の透明性と信頼性を向上させることを目指す．実験結果より，Sparse Autoencoder は概念抽出の精度が高く，解釈可能性の向上に寄与することが確認された．一方で，各特徴量と得点の関連性を明確にする課題が残されており，今後の研究ではこれを解決することで，小論文採点支援システムのさらなる改良を図る．

1　はじめに

現在，2025 年度において，日本語の小論文採点を支援するシステムは，小規模な採点に対応した試験的なシステムが Web 上で公開されている [1][2]．一方で，大規模な採点に対応したシステムは未だ普及していない．そこで，本研究では小論文採点支援システムのインターフェースを Web アプリケーションとして構築する．LLM による自動採点機能を実装するにあたり，自動採点の信頼度を採点画面に表示することで，モデルの解釈可能性の向上を図り，ユーザーエクスペリエンスを向上させることを目指す．本研究では，モデルの解釈可能性向上を図るための手法として，Attention，Masking Token，および Sparse Autoencoder の 3 つを提案する．特に，Sparse Autoencoder を用いて特徴量ごとに概念を特定し，モデルの内部構造の解釈を可能にすることで，採点結果の透明性と信頼性を向上させることを目指す．実験結果より，Sparse Autoencoder は概念抽出の精度が高く，解釈可能性の向上に寄与することが確認された．一方で，各特徴量と得点の関連性を明確にする課題が残されており，今後の研究ではこれを解決することで，小論文採点支援システムのさらなる改良を図る．

2　関連研究

高橋らによる研究（2024）[3] では，得点予測の精度向上に加えて予測確信度の推定も可能な深層学習を用いた論述回答自動採点モデルを提案している．従来のモデルが分類器として設計されていたのに対し，本研究では回帰モデルを採用し，得点予測精度を向上させるとともに，マルチタスク学習の枠組みを活用することで分類と回帰のハイブリッド型モデルを開発している．実験結果により，提案モデルは得点予測および確信度推定の双方において，従来手法と同等以上の性能を示すことが確認された．

また，Anthropic（2023）[4] の研究では，言語モデルに対して，スパースオートエンコーダを用いて重ね合わせを展開し，特徴量を可視化する手法を提案している．この手法により，各ニューロンがどのような特徴を捉えているかを明確にすることが可能となり，モデルの解釈可能性の向上に大きく貢献した．

そして，Hoagy ら（2023）[5] の研究では，ニューラルネットワークの内部表現の解釈可能性を阻害する要因である多義性の解決を目指して，スパースオートエンコーダを用いた新しい手法を提案している．この手法では，言語モデルの活性化ベクトルをスパースに復元可能な特徴の集合として再構築し，解釈可能性やモノセマンティック性が向上した特徴を抽出することに成功している．

3 提案手法

3.1 Attention

BERT は Transformer をベースにしたモデルであり，その仕組みとして「Self-Attention」が用いられている．これは，入力文中の各トークンが他のトークンとどの程度関連があるかを表しており，これを用いることでモデルがタスク実行時にどのトークンに注目しているかを可視化することができる．また，トークンの中でも，文頭にある [CLS] トークンは，全てのトークンの情報を集約したものであり，文全体の意味を表しているため，BERT 最終層の [CLS] トークンから他のトークンへの Attention を可視化することで，採点時にモデルがどのトークンに注目しているか分析する．

3.2 Masking Token

BERT のスペシャルトークンである [MASK] トークンは，入力文の一部を MASK することができる．また，モデルの推論時にも入力トークンを [MASK] トークンに置き換えることで，入力文の一部をモデルに見せないようにすることができる．これを利用して，推論時に入力文の一部のトークンを MASK し，各点数予測のロジットがどのように変化するかを分析することで，モデルが採点時に重視しているトークンを可視化することを試みる．入力テキストの 1 トークンを MASK した際，「5 点」に対する予測値がどのように変化したかを示している．MASK をすることで「5 点」に対する予測値が小さくなった場合，そのトークンの重要性は「5 点」と判断するために必要なトークンである，すなわち重要性が高いトークンであると考えることができる．

3.3 Sparse Autoencoder

3.3.1 アーキテクチャ

図 1 に本実験で用いる Sparse Autoencoder のアーキテクチャを示す．BERT 最終層の [CLS] トークンの出力を入力とし，中間層のユニットの数を入力の次元数よりも大きくすることで，ニューロンの多義性による重ね合わせの解消を図る．

3.3.2 学習手順

以下の手順でモデルの学習を行う．

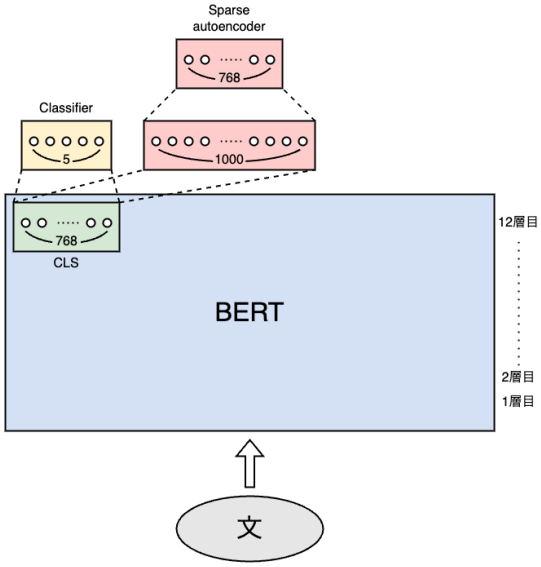

図 1 Sparse Autoencoder

表 1 各変数の説明

変数	説明
n	学習に用いるサンプル数
y_i	i 番目のサンプルに対する正解ラベル
$\hat{y_i}$	i 番目のサンプルに対する予測ラベル
λ	正則化項の強さを制御するハイパーパラメータ
W_i	Sparse Autoencoder の中間層の特徴量のベクトル

1. BERT 最終層と Classifier を式 (1) で表される誤差が小さくなるように，BERT 最終層及び Classifier を学習
2. Sparse Autoencoder を式 (2) で表される誤差が小さくなるように，Sparse Autoencoder を学習

$$Loss = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{1}$$

$$Loss = -\sum_{i=1}^{n} y_i \log(\hat{y_i}) + \lambda \sum_{i=1}^{n} |W_i| \tag{2}$$

式 (1) に示した二乗誤差（MSE）を最小化することで，BERT 最終層と Classifier が入力に対してより正確な予測を行うように学習を進める．

一方，式 (2) で示したクロスエントロピーとスパース性を考慮した誤差を最小化することで，Sparse Autoencoder が必要な特徴を抽出しつつ不要な重みを小さく抑えるように学習する．

上記の手順において，式 (1) および式 (2) で用いられる変数を表 1 に示す．

また，λ の値には先行研究 [5] と同様に 0.00086 を

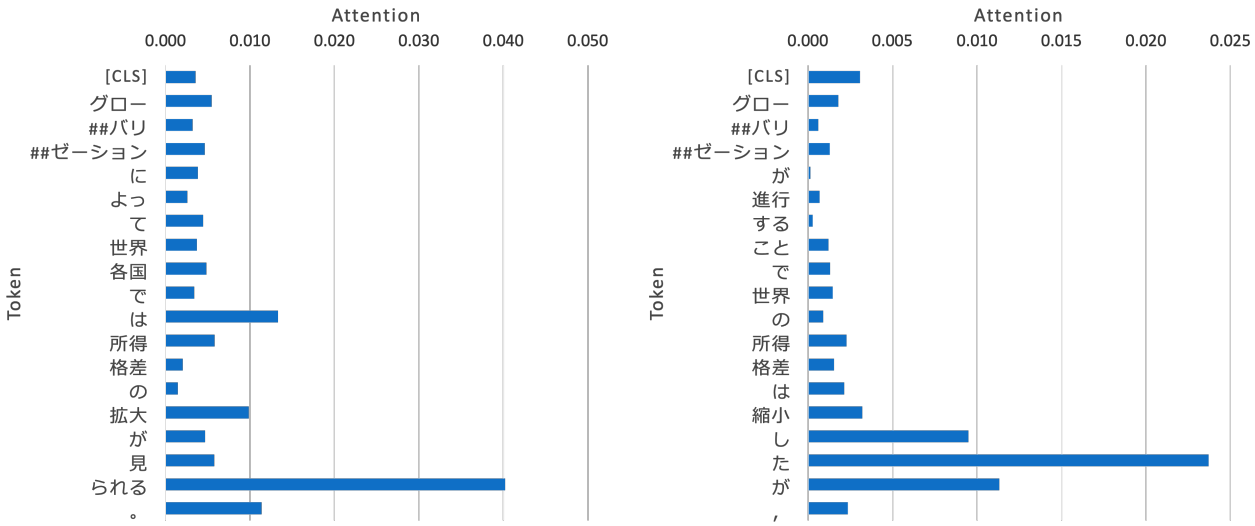

図 2　BERT 最終層の [CLS] トークンからの対象トークンへの Attention，左図：ラベル「2 点」の答案に対する Attention，右図：ラベル「4 点」の答案に対する Attention

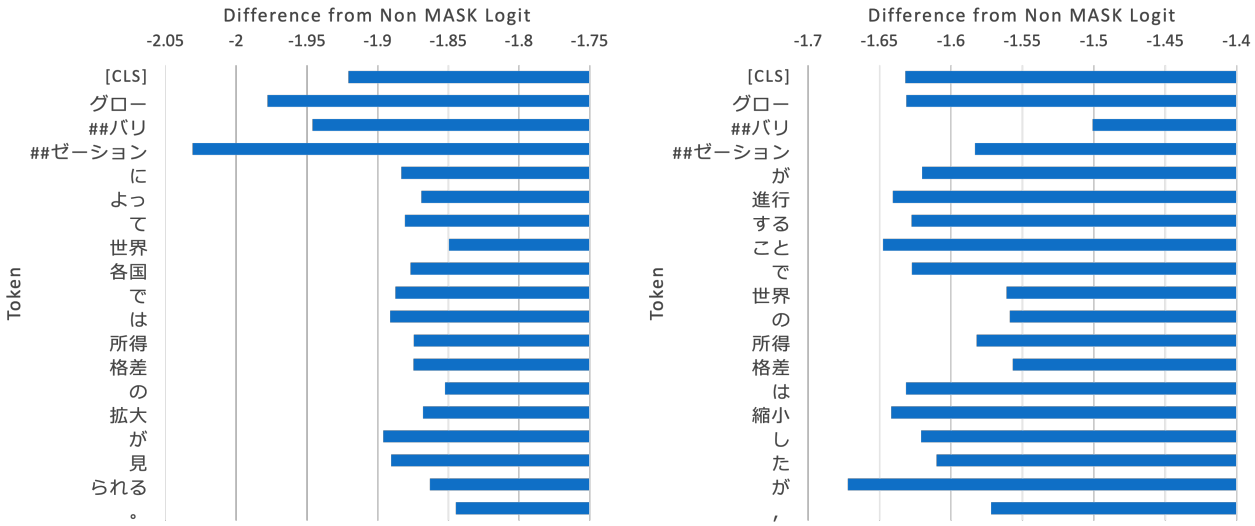

図 3　入力テキストを 1 トークンを MASK した際のラベル「5 点」に対する予測値の変化

用いる．

3.3.3　特徴抽出

3.3.2 章で学習したモデルも用いて，Sparse Autoencoder の中間層の各特徴量が持つ概念を特定する．具体的には，推論時に学習データから抽出した形態素を入力して，各特徴量が活性化した形態素を分析する．これにより，モデルがどのような語彙や特徴を捉えているかを可視化し，特徴量ごとに対応する概念（例：特定の品詞や文脈上の役割など）を把握することができる．分析結果をもとに，特徴量と概念の対応関係を評価し，モデルの解釈可能性の向上を狙う．

4　実験

4.1　実験設定

モデルには，東北大学が公開している事前学習済み日本語 BERT モデル[1)]ををを使用する．また，データセットには，327 件の小論文データ（訓練データ：300 件，テストデータ：27 件）[2)]を用いる．

表 2　訓練データの得点内訳

得点	1 点	2 点	3 点	4 点	5 点
件数	0	88	154	54	4

1）　https://huggingface.co/tohoku-nlp/bert-base-japanese-v2
2）　https://www.gsk.or.jp/catalog/gsk2021-b/

4.2 実験結果

4.2.1 Attention

図 2 に結果を示す．「所得格差の拡大」や「グローバリゼーションが進行」といった設問と関連性が高い文には反応しなかった．一方，文末や文の切れ目となるトークンに Attention が集中していることが確認された．また，他のラベル（「1 点」，「3 点」，「5 点」）に対しても同様に Attention が文末や文の切れ目にかかる傾向が見られた．

4.2.2 Masking Token

図 3 に結果を示す．「グローバリゼーション」などの設問のトピックとなる語彙が重要性が高くなる結果も散見されたが，ほとんどの回答では，図 3 の右図に示すように，トークン間の差異が非常に小さくなる傾向が見られた．また，他のラベル（「1 点」，「2 点」，「3 点」，「4 点」）に対する予測値の変化に関しても同様に，トークン間の差異が小さくなる傾向が見られた．

4.2.3 Sparse Autoencoder

表 3 3 つの特徴量が活性化した上位 20 個の形態素

特徴量 140	特徴量 366	特徴量 613
値段	社国	大半
負け	道	従業
現	上がら	全て
用い	拠点	収め
有効	生みださ	多く
予盾	費	すべて
原因	共有	ほとんど
場合	着手	一部
税金	雇わ	自社
為	道路	参加
考え	扱	場所
理由	支店	設置
状態	燃	運用
感じ	輸出	全く
結果	多国	これ
支払う	又	前述
困難	(	集まっ
賃金	本社	なかっ

表 3 に結果を示す．ノイズが含まれるものの，特徴量ごとに概念が存在されていることが確認された．例えば，特徴量 140 は「値段」や「税金」といった『経済』に関する概念，また，特徴量 366 は「道」や「拠点」といった『インフラ』に関する概念，そして，特徴量 140 は「大半」や「全て」といった『数量・割合』に関する概念を捉えていると分かる．

5 考察

4.2.1 章と 4.2.2 章の結果から，Attention と Masking Token では，設問に関連する重要性が高い単語・文章の抽出は困難であると考えられる．一方，4.2.3 章の結果から，Sparse Autoencoder は，特徴量ごとに概念を抽出することが可能であることが示された．これにより，Sparse Autoencoder は，モデルの解釈可能性を向上させることができるため，採点支援システムにおいて有用であると考えられる．しかし，現段階において，各特徴量と得点間の関連性が不明瞭であるため，今後は特徴量と得点の対応関係を明確にすることが課題であると言える．また，本実験では BERT を用いて検証を行なっているため，GPT モデルなど他のモデルを用いて検証を行うことで異なる特徴表現を抽出し，それらの表現を比較・分析することで，モデル間の違いや特徴抽出の汎化性能についてさらなる洞察を得ることが期待される．

6 おわりに

本稿では，小論文採点支援システムにおいて，UX の改善を図るために，モデルの解釈可能性を向上させる以下のの手法を提案した．

- Attention
- Masking Token
- Sparse Autoencoder

実験結果から，Attention と Masking Token は，重要性が高い単語・文章の抽出は困難であると考えられる．また，Sparse Autoencoder は，特徴量ごとに概念を抽出することが可能であることが示された．しかし，各特徴量と得点間の関連性が不明瞭であるため，今後は特徴量と得点の対応関係を明確にすることが課題であると考えられる．今後の方針としては，これらの提案手法を定量的に評価し手法の有効性を検証したのち，最終的に採点支援システムに組み込むことを目指す．

謝辞

議論に参加して下さいました竹内研究室の諸氏に心より感謝致します.

参考文献

[1] 石岡恒憲, 亀田雅之. コンピュータによる小論文の自動採点システム jess の試作. 計算機統計学, Vol. 16, No. 1, pp. 3–19, 2003.

[2] 石岡恒憲. Jess®: 日本語小論文 評価採点システム, 2002. Accessed on 2025/1/7.

[3] 高橋祐斗, 宇都雅輝. 確信度と得点の予測精度を両立する論述回答自動採点モデル. 言語処理学会第 30 回年次大会, pp. 1136–1141, 2024.

[4] Anthropic. Towards monosemanticity: Decomposing language models with dictionary learning, 2023. Accessed on 2025/1/7.

[5] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023.

[6] 齊藤隆浩, 古宮嘉那子, 石岡恒憲, 中川正樹. Jglue データを用いた模範解答との差異に基づく汎用採点モデルの構築. 言語処理学会第 30 回年次大会, pp. 1182–1186, 2024.

[7] 中本さや香, 嶋田和孝, 岡本芳明, 中河内孝. 日本語小論文に対する採点およびフィードバックの生成. 言語処理学会第 30 回年次大会, pp. 1142–1147, 2024.

[8] Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8669–8687, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[9] Aditya Patkar, Suraj Chandrashekhar, and Ram Mohan Rao Kadiyala. AdityaPatkar at WASSA 2023 empathy, emotion, and personality shared task: RoBERTa-based emotion classification of essays, improving performance on imbalanced data. In Jeremy Barnes, Orphée De Clercq, and Roman Klinger, editors, **Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis**, pp. 531–535, Toronto, Canada, July 2023. Association for Computational Linguistics.

[10] 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, 2022.

[11] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics.