自由記述アンケートを用いた成績根拠文の自動生成

EP21114 山本 樹　　指導教授：山下 隆義

1. はじめに

教育現場では，受講する各学生の講義内容に対する理解度は異なることが多い．そのため，教員は学生一人一人の学習状況を把握することが求められる．しかし，学生の人数が多く，限られた時間の中で，教員が学生一人一人の学習状況を把握していくことは，教員にとって負担である．そこで，受講者のアンケートから成績を予測する研究が注目されている．従来手法 [1] により AI モデルを用いた成績予測が可能となったが，成績に対する根拠がないため，教員にフィードバックできない．そこで，本研究では，講義後アンケートと成績から，成績に対する判断根拠文を LLM により自動生成することを目的とする．

2. 従来研究

小池らは，毎週の生徒の理解度が，最終成績に表れると仮定して，毎週のアンケート文から最終成績を予測する手法を提案した [1]．これにより，毎週のアンケート文から最終成績を予測することが可能となった．一方で，本手法は予測した成績に対する根拠を示すことができない．そのため，教員に成績に対する根拠をフィードバックできず，依然として教員の負担が高いという問題がある．

3. アンケートデータセットの構築

九州大学のデジタル信号処理の講義で実施したアンケートの回答文と，回答者の成績が与えられたデータセットを用いる．アンケートは 2020 年から 2022 年の間に，377 名の学生に実施された．アンケートの質問内容を表 1 に示す．

表 1：アンケート内容

質問	質問内容
Q1	今日の内容を自分なりの言葉で説明してください．
Q2	今日の内容で分かったことを書いてください．
Q3	今日の内容で分からなかったことを書いてください．
Q4	質問があれば書いてください．
Q5	今日の講義の反省・感想を書いてください．

本データセットには成績に対する判断根拠文が含まれないため，LLM の追加学習が不可能である．そこで，GPT-4 を用いて，成績に対する判断根拠文を生成する．図 1 に判断根拠文生成時のプロンプトを示す．プロンプトにはアンケートの質問文，回答文及び最終成績を用いる．

```
アンケートの回答：
可視化はデータを見やすくし，直接的に理解できるようにする手法です．

これが成績 D の人の以下の質問に対する回答文です．
質問：今日の内容を自分なりの言葉で説明してみてください．
成績は上から A，B，C，D，F の 5 段階で，F は不合格です．

上の文章で成績が D である理由を説明してください．
```

図 1：判断根拠文生成のプロンプト

4. LLM による成績根拠文生成モデル

学生の成績に対する根拠を提示するモデルを実現するために，大規模言語モデル（LLM）を使用するアプローチを考える．具体的には，日本語での文章生成に特化した LLM である，Llama-3-ELYZA-JP-8B を Low-Rank Adaptation (LoRA) [2] によって追加学習し，根拠生成に特化させる機構を提案する．

LoRA とは，モデルが持つ巨大なパラメータ行列 W の勾配更新量 ΔW を，低ランク行列で近似することで，学習時における計算効率の向上を図る手法である．LoRA によるチューニング時の線形変換過程を式 (1) に示す．

$$W' = W + BA \tag{1}$$

ここで，W' はチューニング後の重み，A, B は LoRA が導入する低ランク行列，W は事前学習済みモデルの重み行列である．LoRA によるチューニングでは，モデルが持つ重み W を凍結し，分割された重み BA を更新することで，より効率的な学習を可能とする．

5. 評価実験

本実験では，提案手法によって生成された判断根拠文に対して，定量評価とアンケート評価により評価を行い，有効性を検証する．

5.1 実験条件

生成された判断根拠文の有効性をアンケートにより評価する．評価実験には，定量評価とアンケート評価の 2 つを行う．定量評価として MoverScore[3] を利用する．MoverScore とは，文章間の変換コストを指標として示す評価手法である．値は 0 から 1 の範囲で，1 に近いほど文章間の意味的な差異が小さいことを表す．アンケート評価として LoRA によるチューニング前後の判断根拠文に対して「どちらが成績をより説明するか」という質問を実施する．アンケートの回答者は 30 人である．また，GPT-4 によって生成した根拠文を正解データ，LoRA によるチューニング前後のモデルで生成した根拠文を比較データとして利用する．

5.2 実験結果

定量評価として，LoRA によるチューニング前後での MoverScore の比較を表 2 に示す．

表 2：チューニング前後の MoverScore の比較

チューニング前	チューニング後
0.85	0.87

表 2 より，MOverScore の平均値はチューニング後のモデルの方が高いことが確認できる．しかし，この MoverScore の平均値の差異が有効か不明であるため，アンケートを用いて，有効か評価する．図 2 に成績 D の学生に対する正解データ及びチューニング前後の生成文を示す．

正解データ　：可視化の定義が述べられているが，具体的な手法や応用例が不足しているため．

チューニング前　：成績が D である理由は，可視化をデータの可読性向上に役立つ手法と述べているが，具体的な手法や応用例が不足しているため．

チューニング後　：この回答者は，質問の要求である「今日の授業の内容を自分なりの言葉で説明してみてください」という要求を満たしていない．具体的な内容が書かれていないため，質問者が求める「自分なりの言葉」での説明が不可能である．従って，成績は D である．

図 2：判断根拠文の生成結果

図 2 より，チューニング後の根拠文は正解データと主観的に比較すると，同等の根拠文が生成可能であることが確認できる．次に，アンケート結果を表 3 に示す．

表 3：30 名に対するアンケート結果

チューニング後が良いと回答した人数	割合 [%]
21/30	70

表 3 より，7 割の回答者がチューニング後の判断根拠文の方が良いと答えており，LoRA によるチューニングを行うことで，成績に対する判断根拠文を適切に表現可能であることを確認した．

6. おわりに

本研究では，講義後のアンケートに判断根拠文を追加したデータセットを構築し，成績に対する判断根拠を生成可能な言語モデルを LoRA によるチューニングにより実現した． LoRA によるチューニングを行った場合，より適切な成績に対する判断根拠文が生成可能であることを確認した．今後の展望として，別のチューニングを行った時の成績に対する判断根拠文の生成を検討する．

参考文献

[1] M. Koike, *et al.* “Enhancing the Accuracy of Predicting Students Grades in Open-Ended Questions through Adjustments to Attention Weights”, EDM, 2024.

[2] E. Hu, *et al.* “Low-Rank Adaptation of Language Models”, ICLR, 2022.

[3] Z. Wei, *et al.* “MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance”, ICLR, 2022.