5W-07

自動採点における文生成を用いた学習データの削減の検討

宮田 創太 †
東京都立産業技術高等専門学校 †

横井 健 ‡
東京都立産業技術高等専門学校 ‡

1 はじめに

記述式問題を採点する上での問題点として、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ことが挙げられる。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 [1][2]。自動採点技術において取り組むべき課題として、学習データを減らし、学年・クラス単位での自動採点を可能にすることが挙げられている。本研究では、ChatGPT を使い、オリジナルの解答文から文を生成し、学習データにオリジナルの解答文と生成文を使用することで、実際に必要な学習データを減らして自動採点モデルに学習させ、その採点精度を調べる。

2 提案手法

2.1 文の生成

まず、生成するもとになるオリジナルの解答文を選出する。doc2vec を用いた K-means 法によるクラスタリングを行い、オリジナルの解答文を仕分ける。その後、各クラスタから数個ずつ解答文を選出する。そして、ChatGPT を用いて、選出したオリジナルの解答文から文を数十個生成する。この生成した文を学習データとして用いる。

2.2 モデルの学習方法

2.2.1 半教師あり学習

まず、点数のラベルがついたオリジナルの解答文をモデルに学習させる。次に、ラベルのついていない生成文をモデルに入力し、予測結果 (点数) を出力させる。その後、予測結果と生成文をモデルに学習させる。これを繰り返し、自動採点モデルを構築する。

2.2.2 教師あり学習

生成文に、生成元であるオリジナルの解答文と同じ点数をそのままラベルとして付ける。点数のラベルがついたオリジナルの解答文と生成文をモデルに学習させる。

3 実験

3.1 実験方法

まず、2.1 節で示した手法を行った。オリジナルの解答文の選出については、(解答の最高点 +1) 個のクラスタで解答文を仕分け、各クラスタから 2 個ずつ解答文を選出し、選出した文から 50 個ずつ文を生成した。次に、2.2 節で示した 2 種類の手法と、オリジナル解答文のみでの教師あり学習で、自動採点モデルにそれぞれ学習を行った。そして、テストデータ (ラベルのついていないオリジナルの解答文) を入力として、自動採点モデルに点数を予測させた。これを 5 回繰り返し、各回で重み付きカッパ係数 (QWK) を測定し、その平均値を出力した。

データセットには「理研記述問題採点データセット [3]」を用いた。解答の最高点が 14 点である短答記述式問題 1 つに対し予測を行った。解答件数は 2,100 件である。この解答文か

Training data reduction in automated short answer scoring with sentence generation

† Souta Miyata, Tokyo Metropolitan College of Industrial Technology

‡ Takeru Yokoi, Tokyo Metropolitan College of Industrial Technology

表 1　半教師あり学習で自動採点を行った結果

学習データ個数 (オリジナル)	テストデータ個数	1 回当たりに学習した生成文の個数	QWK の平均
400	500	500	0.0001
800	500	500	0.0692
1,200	500	500	0.2385
1,600	500	500	0.4274

ら 1,500 個の文を生成し、すべて学習データに使用した。

自動採点モデルには先行研究で提案された LSTM による自動採点モデル [4] を参考に、日本語の短答記述式問題で自動採点ができるようなモデルを実装、使用した。予測の際、入力は解答文、出力は予測した点数である。

3.2　実験結果・考察

実験結果をまとめたものを表 1、表 2、表 3 に示す。

半教師あり学習については、オリジナル解答文の学習データ個数が増えると、QWK の平均が大きくなっているが、0.5 にも満たさないため、学習数が足りないと考えられる。

教師あり学習についても、オリジナル解答文の学習データ個数が増えると、QWK の平均が大きくなった。学習データ個数が 1,600 個の時は、オリジナルの解答文のみでの教師あり学習と比べると、QWK が約 0.27 も大きくなった。しかし、学習データ個数が 2,000 個になると QWK の平均が減少した。これはモデルが過学習を起こしたためと考えられる。

以上を踏まえると、生成文を用いて教師あり学習を行うことは採点精度の向上に繋がると考えられる。しかし、有用性のある採点精度には満たないため、自動採点モデルの手法を改善すれば、さらに精度が良くなると考えられる。

4　まとめ

文生成を用いて、学習データに生成文を追加することで、実際に必要な学習データを減らして自動採点モデルに学習させ、その採点精度を調べた。その結果、自動採点モデルの採点精度が向上することが分かった。

表 2　教師あり学習で自動採点を行った結果

学習データ個数 (オリジナル)	テストデータ個数	QWK の平均
10	100	0.4047
100	100	0.4499
400	100	0.5172
800	100	0.5323
1,200	100	0.5718
1,600	100	0.7247
2,000	100	0.6697

表 3　教師あり学習 (オリジナルの解答文のみ) で自動採点を行った結果

学習データ個数 (オリジナル)	テストデータ個数	QWK の平均
1,600	100	0.4537

謝辞

本研究では、国立情報学研究所の IDR データセット提供サービスにより国立研究開発法人理化学研究所から提供を受けた「理研記述問題採点データセット」を利用した。

参考文献

[1] T. Mizumoto, et al. “Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring, ” BEA 14, pp. 316-325, 2019.

[2] H. Funayama, et al. “Reducing the Cost: Cross-Prompt Pre-Finetuning for Short Answer Scoring, ” In Artificial Intelligence in Education. AIED2023.

[3] 理化学研究所（2022）: 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. データセット: https://doi.org/10.32130/rdata.3.1

[4] D. Alikaniotis, et al. “Automatic Text Scoring Using Neural Networks, ” ACL, pp.715 – 725, 2016.