E-010

事前学習済みモデルを利用した日本語小論文採点手法の構築
System of Japanese Essay Scoring Using Pre-trained Models

藩 宇偉[†]　　竹内 孔一[†]

Pan Yuwei　Takeuchi Koichi

1. はじめに

近年の AI の発達に伴い、文書の自動採点に関する研究が活発化している。英語では、2012 年に kaggle が Automated Student Assssment Prize コンペティションを開催し、ASAP AEG データセット*を公開した。日本語では、少量ではあるが、日本語母語話者が記述した採点済み小論文データが公開されている[1]。本研究では小論文データ[1]を対象に実験する。事前学習モデルは様々な分野で大きな発展を遂げている。そこで、本研究では近年利用されている BERT[2]と GPT3[3]を使って、既に採点済みの日本語小論文を対象に学習を行い、推定結果について分析する。

2. 自動採点モデル

本研究は各小論文データに対して 1〜5 のスコアをつける多クラス分類タスクとする。

BERT は、トランスフォーマーのエンコーダを使ってテキストを記号に変換して計算するモデルである。各小論文を一つシーケンスとして入力し、512 個のトークンを取り出し、その後多クラス分類を行う。

GPT3 は、トランスフォーマーのデコードを使った、巨大な自己回帰の事前学習済みモデルである。GPT3 の最大のバージョン（davinci）は、1750 億のパラメータを持ち、Fine-Tuning をしなくても、入力に応じて望ましい結果を出力することができる。学習速度の向上とコスト削減のため、本手法では、OpenAI 社から公開されている 3 番目に大きなモデルを選択した。受け取れるトークンは最大で 2048 個、モデルのパラメータは 13 億個となっている。

3. 実験

本章では、使用した実験データ、実験設定、評価尺度、実験結果について説明する。

3.1 実験データ

実験データは、日本語小論文データのうち、講義の内容として「グローバリゼーション」(以降「グローバル」)、「自然科学」の 2 つの講義を選んだ。各テーマには三つの課題があり、各課題に対して答案の文字数は 100 から 800 までである。本実験のデータセットを以下の表 1 に示して、採点の基準は理解力と文章力の二つ点数だけである。

3.2 実験設定

学習における、バッチサイズは両実験とも 8 と設定する。データセットの分布に偏りがあり、サンプル数が少ないため、バリデーションセットは設定しないこととした。トレーニングセットとテストセットは 87%と 13%に設定した。

BERT には HuggingFace の MeCab 形態素解析器を利用して日本語に特化した BERT ベースモデルを使った。学習率は 10^{-5} とし、最適化手法を Adam は利用した。

GPT3 については、OpenAI が公開されている GPT3 の babbage モデルを使用した。公式 Fine-tuning API を使用しているため、評価指標 QWK は各エポック後に一度だけ取得できなかったため、公式デフォルトパラメータを用いて、四つのエポックの学習後直接に最終モデルを取得した。推定の場合、Temperature は 0 に設定する必要があり、そうでない場合、2 点や 3 点といった異なる出力が観測された。一つのスコアを生成するために Maximum length は 1 に設定する必要がある。データセットの準備として、BERT と同じランダムシードを用いてトレーニングセットとテストセットを分割した。その後、OpenAI の Fine-Tuning 公式ガイドを参考に、図 1 のようにトレーニングセットを GPT3 が読み込める JOSN の形式に変換した。

表 1　採点するデータセット

テーマ	課題	答案数	長さ	採点基準
グローバル	1	328	300	理解力と文章力
	2	327	250	
	3	327	300	
自然科学	1	327	100	理解力と文章力
	2	325	400	
	3	327	800	

{"prompt":"グローバリゼーションによって、１９８１年は、１日１ドル未満で生活する人が３０%だったが、２１世紀に入ると、その割合が２０%になったように、世界的にみると、所得格差は減少している。しかし、各国でみると、賃金を安くして働くことができる外国人がたくさん入って来たりしたため、格差は大きくなっている。この所得格差の拡大、縮小が起こった理由として、各国では、外国人に職を取られ、貧しい人が大勢出てしまったからであり、世界では、人口の多い東アジアの人達が、グローバリゼーションによって活躍する人が増加したからである。 ->","completion":" 2"}
……

図 1　GPT3 読み込むデータセット形式

3.3 評価尺度

スコアの 5 点文章を 1 点として判定する場合と、スコアの 3 点文章を 1 点として判別する場合では、前者の方に対して、この判別に対するペナルティを大きくする必要がある。そこで、本実験では、2 次重み付きカッパ係数（Quadratic Weighted Kappa、以下は QWK）を使う。この

[†]岡山大学自然科学研究科　Graduate School of Natural Science and Technology, Okayama University.

*https://www. kaggle.com/c/asap-aes/data

QWK 係数が 1 に近いほど、システムの性能が高いことを意味する。

3.4 実験結果

各実験で異なるトレーニングセットとテストセットを選び、BERT と GPT3 に対して、それぞれ 2 回に実験し、平均の QWK を計算する。実験結果を表 2 に示す。

表 2　各課題 2 回実験した QWK

テーマ	課題	理解力		文章力	
		BERT	GPT3	BERT	GPT3
グローバル	1	.366	.098	.568	.440
	2	.869	.677	.594	.541
	3	.213	.240	.221	.529
自然科学	1	.797	.761	.000	.000
	2	.632	.467	.749	.722
	3	.270	.199	.785	.782
平均		**.524**	**.407**	**.486**	**.502**

3.5 結果分析と考察

QWK の平均に比べると 2 つのモデルの性能はほぼ同じレベル（一部大きな差を除く）であり、同じ課題において、GPT3 の QWK 値が低い場合には、BERT の QWK 値も同様に低い。また、エポックは 4 回で止まったことと GPT3 モデルの収束が予測できなかったため、GPT3 は本実験での性能は本来もつ性能を引き出していない可能性がある。

また、表 2 から同じモデルのもとでは、理解力、文章力の評価ともに、異なる課題における QWK に大きな差があることがわかる。この差の原因は 2 つ考えられる。1 つ目は、小論文の点数分布の問題である。例えば、自然科学の小論文の文章力の点数の分布を見ると、図 2 のような分布になっている。課題 1 の分布が非常に偏っているため、採点が難しくなっていることがわかる。

2 つ目は小論文の長さの問題で、表 1 では自然科学の小論文の長さは 100、400、800 となっている。理解力の自動採点については、文章の長さが長くなるにつれて QWK が低くなっていく。一方で、文章力はこの傾向に当てはまらない。これは評価軸の性質の違いが現れていると考えられる。すなわち、理解力は内容に応じたスコアが付与されているが文書力は文字数や誤字など形式的な側面で文書を評価しているため、この特徴の違いが傾向の違いに現れたと考えられる。

4. おわりに

本研究では、事前学習済みモデル BERT と GPT3 を利用し、小論文に対する理解力と文書力について QWK に基づく採点性能を比較した。この実験では、GPT3 は、ほとんどの場合、BERT より低い結果になっているが、これはデータセットが少ないことと収束の判断ができない点で性能を発揮していない可能性がある。QWK の値から 0.6 を越えた課題もありその部分はどちらのモデルも有効であることがわかる（QWK>0.6）。複数の採点基準に基づく採点では一つのモデルを使うのは最適ではなく、それぞれの項目の採点にカスタマイズしたモデルが必要となる可能性も考えられる。日本語では平尾ら[4]は外国人が書いた小論文に対して自動採点を行い、BERT は全体的に性能が高い値を示した。英語では、Mayfield ら[5]が ASAP AEG データセットに基づいて複数のモデルを評価していたが、BERT など事前学習済みモデルは必ずしも最適ではないことを実験的に示した。一方で、近年多様な事前学習済みモデルが提案されており、異なる事前学習済みモデルでの評価をさらに進める必要があると考えられる。

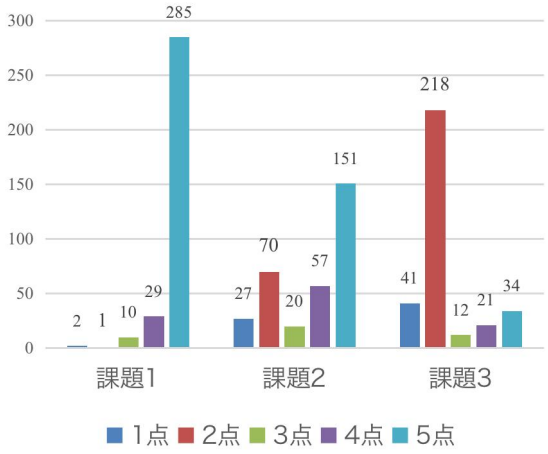

図 2　テーマ「自然科学」の三つの課題それぞれの小論文の文章力の得点分布(学習セット＋テストセット)

謝辞

本研究の一部は科学研究費 22K00530 の助成を受けた。

参考文献

[1]竹内 孔一, 大野 雅幸, 泉仁 宏太, 田口 雅弘, 稲田 佳彦, 飯塚 誠也, 阿保 達彦, 上田 均. 研究利用可能な小論文データに基づく参照文書を利用した小論文採点手法の開発. 情報処理学会論文誌 62(9) 1586-1604 2021 年 9 月.

[2]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[3]Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[4]平尾礼央, 新井美桜, 嶋中宏希, 勝又智, 小町守. 複数項目の採点を行う日本語学習者の作文自動評価システム。言語処理学会第 26 回年次大会発表論文集, 2020.3.

[5]Mayfield E, Black A W. Should you fine-tune BERT for automated essay scoring? Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications. 2020: 151-162.