BERT を利用した日本語小論文採点支援システムの検討

江島知優
岡山大学工学部
chihiro.100522@s.okayama-u.ac.jp

堀江遼河
岡山大学大学院自然科学研究科
ptmn0udx@s.okayama-u.ac.jp

竹内孔一
岡山大学大学院自然科学研究科
takeuc-k@okayama-u.ac.jp

1 はじめに

記述式問題に対する自動採点は近年の大学入試改革等で需要が高まっているのに伴い，様々な手法が提案されている．英語の文章や外国人が書いた日本語の文章を対象とした研究としては佐藤 [1] のトランズダクティブ学習や平尾ら [2] の作文自動評価システムがある．佐藤は記述式答案の自動採点は学習の過程で評価データの情報を明示的に利用するトランズダクティブ学習を用いることが適切であると主張し，英語で書かれた記述式答案のデータセットである ASAP-SAS を用いてクラスタリングをしたモデルとクラスタリングをしていないモデルおよびランダムのモデルとの比較を行い，クラスタリングをしたモデルの精度が最も高いことを示した．平尾らは外国人の日本語学習者が書いた作文を集めた GoodWriting データセットを利用して全体・内容・構成・言語の 4 観点で BERT (Bidirectional Encoder Representations from Transformers) [3] を用いた手法と従来の素性を用いた手法との比較を行い，BERT を用いた手法が構成以外の 3 観点で精度が高いことを示した．

日本人が書いた日本語文章を対象とした研究としては船山 [4] の確信度推計手法による研究，内田ら [5] の項目反応理論に基づく能力推定値を活用した手法などが挙げられる．船山は代々木ゼミナールの国語長文読解問題データセットを用いて事後確率を用いた確信度推定手法と TrustScore を用いた確信度推定手法を検証し，TrustScore を用いた手法のほうが精度が高く，また、学習に利用するデータ数を減らした場合でも TrustScore を用いた手法の精度が維持されることを示した．内田らは項目反応理論（IRT）を用いて客観式問題から求められる受験者の能力推定値を組み込んだ新たな深層学習自動採点モデルを提案し t 検定の結果提案手法が従来手法より有意水準 5%で高い精度を示したことを確認した．他にも，清野ら [6] による Neural Attention モデルや，大野ら [7] による IDF の評価手法がある．

このように記述式問題の自動採点に対しては多くの研究が行われているがその多くは 100 文字以下のいわゆる短答式記述問題と呼ばれるものを対象としている．本研究では，100 文字を超える小論文をターゲットとして BERT を用いたモデルを作成し，性能の分析を行った．

2 小論文自動採点システム

本研究では 1 点から 5 点の各整数値の値をクラスとしたクラス分類問題に取り組んだ．図 1 に作成した BERT モデルの構造を示す．BERT モデルでは，小論文を符号化し冒頭に [CLS] トークンを，末尾に [SEP] トークンを挿入したものが入力として与えられる．符号化された本文の長さがモデルが読み取れる系列長を超えた場合，本文の最大系列長を超えた部分は削除される．図中の E は入力の埋め込み表現, を，C は [CLS] トークンの隠れベクトルを，Ti は i 番目のトークンの隠れベクトルをそれぞれ表している．今回のモデルでは BERT の最終層の出力は各クラスに対するベクトルであり，最大のものを推定した点数としている．

3 評価実験

本章では評価実験に使用した小論文やその設問内容，実験結果を説明する．

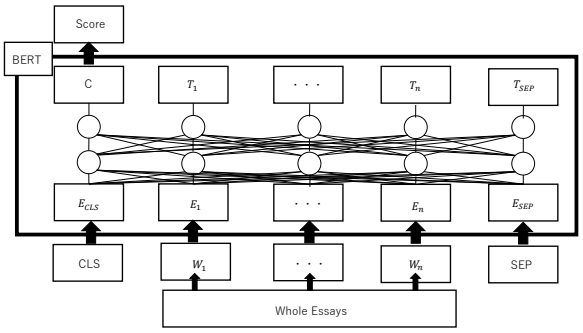

図 1　BERT を利用した小論文自動採点システム

3.1　小論文データ

小論文データとして講義の受講者の解答データを用いている．講義は「グローバリゼーションの光と影」,「自然科学の構成と科学教育」,「東アジア経済の現状」,「批判的思考とエセ科学」の 4 種類である．これらの答案は人手で 1 点から 5 点までの 5 段階で採点されており，その点数を正解データとする．

3.2　小論文の設問内容

本節では今回の実験に利用した小論文の設問内容を示す．

講義名：「グローバリゼーションの光と影」
問 1　「グローバリゼーションは、世界、または各国の所得格差をどのように変化させましたか。また、なぜ所得格差拡大、または縮小の現象が現れたと考えますか。300 字以内で答えなさい。」
問 2　「多国籍企業は、グローバリゼーションの進展の中でどのような役割を果たしましたか。多国籍企業の具体例をあげて、250 字以内で答えなさい。」
問 3　「文化のグローバリゼーションは、私たちの生活にどのような影響を与えましたか。また、あなたはそれをどのように評価しますか。具体例をあげて、300 字以内で答えなさい。」

講義名：「自然科学の構成と科学教育」
問 1　「「科学的」とはどのような条件をみたす必要があるのか 100 字以内で答えよ。」
問 2　「講義で解説した自然科学の二つの側面を参考に、自然科学が果たす役割について 400 字以内で論ぜよ。」
問 3　「「Scientific and Technological Literacy for All」の狙いを考慮し、これからの科学教育はどうあるべきか 500 字以上 800 字以内で論ぜよ。」

講義名：「東アジア経済の現状」
問 1　「日中韓の相互依存の強さを、データを示して簡潔に述べなさい。また、相互依存を示す経済協力・協業の具体例をあげ、合わせて 300 字以内で答えなさい。」
問 2　「「中所得国の罠」の概略を説明し、どうしたらそれを乗り越えることができるか 250 字以内で説明しなさい。」
問 3　「日中韓には少子化や環境問題など 3 国に共通する経済問題がある一方、それぞれの国に特有の課題も多くあります。それぞれの国が抱えている特徴的な経済問題をあげ、東アジアにおける協調と対立の構造を 300 字以内で説明しなさい。」

講義名：「批判的思考とエセ科学」
問 1　「「批判的思考」の定義に関連して、「批判的思考」に関する研究で共通に見出される「批判的思考」の 3 つの観点を述べなさい。100 文字。」
問 2　「講義で紹介した右のグラフを根拠に「長生きするためにはカラーテレビを多く所有すれば良い」と主張することが妥当ではない理由を 400 字以内で述べなさい。ただし、このようなグラフが形成される理由の説明を加えること。」
問 3　「各自で「ニセ科学」の可能性があると思う実例を挙げ、その実例が「ニセ科学」であることを証明するためには、どのような方法で、どのような証拠を得て、どのように説明する必要があるのかを論じなさい。また、その実例がニセ科学でも信じてしまいやすい要因は何かについても考察し、説明しなさい。ただし、講義で扱った事例以外のものを挙げること。500 字以上 800 字以内。」

3.3　実験設定

BERT の言語モデルは Mecab を利用した日本語訓練済みの HuggingFaceBERT[1)] を使用した．このモデルの語彙サイズは 32,000 である．今回の実験では最大シーケンス長とバッチサイズの組み合わせを 2 つ用いた．1 つは最大シーケンス長が 512，バッチサイズが 16 であり，もう 1 つは最大シーケンス長が 400，バッチサイズが 32 である．また，小論文データの数は「グローバリゼーションの光と影」が 328,「自然科学の構成と科学教育」が 327,「東アジ

1）https://github.com/huggingface/transformers

ア経済の現状」および「批判的思考とエセ科学」が290であり，そのうち50をテストデータ，残りを訓練データとして利用している．訓練データのうち50個を*Validation_data*としている．

3.4 評価尺度

本節では今回作成したシステムの精度を測る方法を示す．人手で採点した結果と採点支援システムの出力した結果を入力とし，*Accuracy*および重み付きカッパ係数（QWK: Quadratic Weighted Kappa）での評価を行う．*Accuracy*の計算式を式(1)(2)に示す．

$$eq(a,b) = \begin{cases} 1 & (a = b) \\ 0 & (a \neq b) \end{cases} \tag{1}$$

$$Accuracy = \frac{\sum_{i=1}^{n} eq(A_i, B_i)}{n} \tag{2}$$

nは小論文のデータ数でありA_i,B_iはi番目の点数データである．*Accuracy*が1に近いほど一致率が高いと言える．

3.5 実験結果

小論文の各設問に対して*Validation_Accuracy* (*Val_Acc*)と*Test_Accuracy* (*Test_Acc*)およびQWKをそれぞれ算出した．表1には最大シーケンス長512，バッチサイズ16で実験した結果を，表2には最大シーケンス長400，バッチサイズ32で実験を行った結果を示している．表中では各講義名について「グローバリゼーションの光と影」を「グローバル」,「自然科学の構成と科学教育」を「自然科学」,「東アジア経済の現状」を「東アジア」,「批判的思考とエセ科学」を「批判的思考」とそれぞれ表現している

表1 バッチサイズ16でのシステムの評価

		Val_Acc	*Test_Acc*	*QWK*
グローバル	設問1	0.60	0.48	0.000
	設問2	0.68	0.58	0.000
	設問3	0.66	0.76	0.000
自然科学	設問1	0.60	0.38	0.000
	設問2	0.36	0.24	0.000
	設問3	0.52	0.60	0.000
東アジア	設問1	0.68	0.32	0.132
	設問2	0.36	0.26	0.000
	設問3	0.46	0.44	0.000
批判的思考	設問1	0.48	0.52	0.589
	設問2	0.38	0.36	0.000
	設問3	0.44	0.52	0.000

表2 バッチサイズ32でのシステムの評価

		Val_Acc	*Test_Acc*	*QWK*
グローバル	設問1	0.60	0.48	0.000
	設問2	0.84	0.74	0.628
	設問3	0.66	0.76	0.000
自然科学	設問1	0.60	0.38	0.000
	設問2	0.52	0.44	-0.013
	設問3	0.52	0.60	0.000
東アジア	設問1	0.66	0.34	0.119
	設問2	0.58	0.32	0.322
	設問3	0.68	0.64	0.426
批判的思考	設問1	0.44	0.52	0.606
	設問2	0.50	0.38	0.355
	設問3	0.44	0.52	0.000

3.6 分析

*Accuracy*の結果からBERTは従来研究されていた短答式記述問題に加え，500字以上800字以下といった長文記述問題に対しても5クラス分類の期待値である0.2より高い精度で推定できている．しかし，QWKの値を確認すると特にバッチサイズ16の場合に0.000といった値が多く見られる．これはBERTの推定値がすべて同じ値になったことを示す．BERTの問題点として1点などの極端な点数を推定することが難しく，平均値や最頻値に近い値を予測する傾向が強いことが挙げられ，この問題点によりバッチサイズ16のQWKが低くなってしまったと考えられる．バッチサイズを32に変更することで「東アジア経済の現状」や「批判的思考とエセ科学」のQWKの値には改善がみられる一方で，「自然科学の構成と科学教育」等のQWK値は低いままである．このようにバッチサイズの変更によって推定の精度は大きく変動するため今後も検討が必要である．

4 おわりに

本研究では，HuggingFaceBERTを用いた小論文自動採点システムを作成し*Accuracy*およびQWKによる性能評価を行った．平均値に近い値を予測してしまうといったBERTの特性上，推測値がすべて同じ値になってしまうという問題点も見られたがバッチサイズを大きくする等の工夫によって精度を改善することが可能であり，BERTが小論文自動採点に有効であることを示した．

謝辞

本研究の遂行にあたって岡山大学運営費交付金機能強化経費「小論文、エッセイ等による入学試験での学力の三要素を評価するための採点評価支援システムの開発導入」の助成を受けた.

参考文献

[1] 佐藤俊. 評価データのクラスタリングを用いた記述式答案自動採点のためのトランズダクティブ学習. 言語処理学会第 26 回年次大会発表論文集, 2020.

[2] 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, pp. 1181–1184, 2020.

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

[4] 舟山弘晃. 記述式答案自動採点のための確信度推定手法の検討. 言語処理学会第 26 回年次大会発表論文集, 2020.

[5] 内田優斗, 宇都雅輝. 項目反応理論に基づく能力推定値を活用した短答記述式問題自動採点手法. 言語処理学会第 26 回年次大会発表論文集, 2020.

[6] 清野光雄竹内孔一. ニューラルネットワークを利用した日本語小論文の自動採点の検討. FIT2019 講演論文集, 2019.

[7] 大野雅幸, 泉仁宏太, 竹内孔一, 小畑友也, 田口雅弘, 稲田佳彦, 飯塚誠也, 阿保達彦, 上田均. 小論文自動採点データ構築と理解力および妥当性評価手法の構築. 言語処理学会第 24 回年次大会, pp. 368–371, 2018.