機械学習を利用した日本語小論文採点手法の比較

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
岡山大学学術研究院自然科学学域
takeuc-k@okayama-u.ac.jp

1 概要

本論文では日本語小論文採点システムの構築について記述する．採点手法として BERT や BOW を利用し小論文の文書ベクトルの獲得を行い，SVR や XGBoost などの分類器で学習することで人手に近い採点を得ることを目標とする．精度を QWK (Quadratic Weighted Kappa) を用いて評価することで作成したモデルを比較した．実験の結果から BERT と BOW を組み合わせた方がより高性能であった．また課題として書くべき内容が明確な設問に対して SVR の精度が高く，自由記述の設問は XGBoost の精度が高い傾向があることが実験で示された．

2 はじめに

近年，大学入試改革において，資質や能力の育成が重点となっている．初等中等教育を通じて論理的な思考力や表現力の育成は重要視されており，大学入学共通テストでは思考力，判断力，表現力を測ることを目的として役割を果たしている．これらを評価する手段として，2021 年の大学入学共通テストでは当初国語と数学において記述式問題が採用されると発表されていた．しかしながら，大学入学共通テストの記述式問題を導入することが見送られることが発表された．採点者間における基準の差異が発生したり，質の高い採点者の確保が可能かどうかといった点が課題となっている．また，採点者間だけでなく，同一採点者であっても大量の答案をミスなく採点することは難しく，採点者の心理状態や疲労などにより公正，公平な採点の確保が課題となっている．したがって，人手による大量の記述式問題の採点は人員的かつ公平性の観点から実施難度が高く，導入が難しい．これらの問題点から，記述式問題を一貫して採点することができるシステムの構築が必要とされている．

3 関連研究

関連研究として，竹谷ら [1] の小論文とキーフレーズを比較して採点を行う手法や清野ら [2] による Neural Attention を組み込んだモデルや竹内ら [3] の IDF で評価する手法がある．

本手法では BERT を用いたモデルを作成し，得られる分散表現を用いて評価を行う．

4 小論文採点システム

本提案システムは，採点結果として 1 点から 5 点を出力する．図 1 にシステムの全体像を示す．

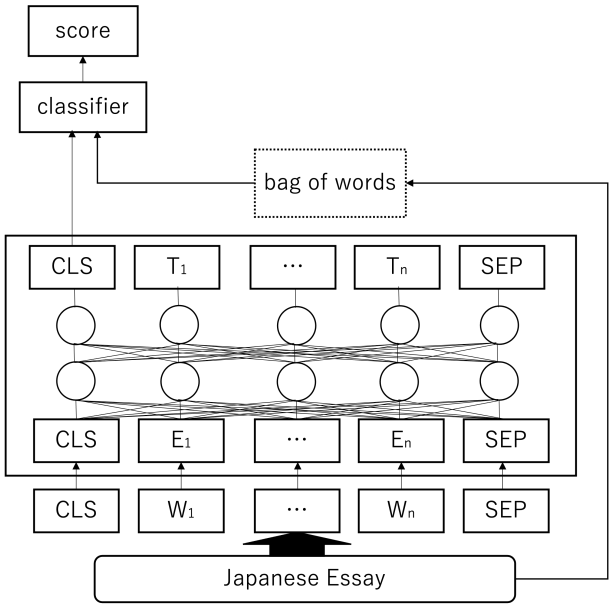

図 1 小論文採点システムの全体図

4.1 文書ベクトル獲得モジュール

BERT (Bidirectional Encoder Representations from Transformers)[4] は 2018 年に Google が開発した事前学習済み自然言語処理タスクを行うモデルである．前後の単語から該当単語の分散表現ベクトルを獲得するため，文脈を考慮したベクトルの得られる．BERT は入力文の先頭に CLS があり，CLS は入

力文全体に対する分散表現ベクトルが埋め込まれている．本研究では一つの小論文から得られる文書をCLS を使用することで分散表現ベクトルに変換し，実験を行う．また，追加実験として bag of words とBERT を組み合わせたベクトルを使用したモデルの作成を行う．bag of words は文章に対して形態素解析を行い，出現した形態素の頻度をベクトルとしたものである．bag of words の語彙数は 32,000 として実験を行う．BERT と bag of words のベクトルは結合し，一つの大きなベクトルとして獲得させる．本実験で利用する BERT は HuugingFce の BERT[1] を利用する．トークナイザーとして MeCab の Wordpiece 版を利用する．

4.2 分類器モジュール

本実験ではサポートベクターマシン (SVM) とXGBoost をそれぞれ分類器として利用し学習をさせる．本問題を回帰問題としてサポートベクターマシンは回帰モデルでの実装を行う．点数データはすべて整数で表現されているため，サポートベクターマシンでも整数値で出力する．回帰モデルでの出力は小数値のため少数第一位で四捨五入を行い，1 点から 5 点の整数値に丸め込みを行う．XGBoost はブースティングで決定木の構築を繰り返して，最適な結果を出力する手法である．入力されたベクトルと正解データから最適な木構造を構築し，学習を行う．その際に構築される木の深さは 6 とした．

5 評価実験

BERT および BOW を利用した場合のベクトルと分類器による採点性能の評価を行うために評価実験を行う．評価実験では，採点済み答案データを学習データとしテストデータに分けて，学習データで分類器を学習し，テストデータで評価する．以下では各詳細について述べる．

5.1 小論文データ

本研究で使用する小論文データは講義に参加したそれぞれ約 300 人の学生の答案である．表 1 に実験で利用する小論文課題と答案数う，ならびにテストデータ，学習データに利用する答案数について示す．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 講義であり，設問は 3 つずつある．表では「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」と省略している．記述が一定の文字数に満たないものや空白であるものについては除外しているため，同講義内で回答者数は一致していない．

本研究で利用している小論文データは言語資源協会から公開されている[2]．

表 1 各講義の設問ごとの小論文データと実験に使用するデータ数

講義名	設問	全データ	テスト	学習
グローバル	設問 1	328	66	262
	設問 2	327	66	261
	設問 3	327	66	261
自然科学	設問 1	327	66	261
	設問 2	325	65	260
	設問 3	327	66	261
東アジア	設問 1	290	58	232
	設問 2	288	58	230
	設問 3	288	58	230
批判的思考	設問 1	290	58	232
	設問 2	290	58	232
	設問 3	290	58	232

5.2 設問内容

本節では講義の設問内容を示す．

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

1） https://github.com/cl-tohoku/bert-japanese

2） https://www.gsk.or.jp/catalog/gsk2021-b

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。」

5.3　評価尺度

本節では提案システムの精度を測る方法について示す．システムの評価は人手で採点された点数を正解データとして，システムの出力した点数データと比較を行う．本研究では重み付きカッパ値 QWK を元に評価を行う．QWK はシステムの採点結果と人手の採点結果との評価ズレの程度を測定する．2 つの評価のズレを 2 乗し重み付けしたものが計算される．結果が 1 に近いほどシステムの評価は高いと言える．$num(a)$ はある採点者が a と採点した回数であり，$ob(a,b)$ は 2 人の採点者が a,b と採点した回数である．

$$byChance(a,b) = \frac{num(a) \times num(b)}{n} \tag{1}$$

$$QWK = 1 - \frac{\sum_{a,b=1}^{5} ob(a,b) \times |a-b|^2}{\sum_{a,b=1}^{5} byChance(a,b) \times |a-b|^2} \tag{2}$$

作成したシステム間に有意差が認められるかの判定を T 検定にて行う．対応する同一の標本である小論文データ n を用いて，システムで採点されたデータを x_i,y_i とし，その差を d とする．s_d はデータの差 d の標準偏差である．検定統計量 T は以下で表される．

$$T = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \tag{3}$$

$$\bar{d} = \frac{\sum d}{n} = \frac{\sum (x_i - y_i)}{n} \tag{4}$$

本実験では有意水準を 0.05 として両実験に有意に差があるかを観測する．

5.4　実験結果

小論文の各設問に対して出力された結果とその平均を表 2 に示す．また，T 検定を適用した P 値について表 3 に記載する．

表 2　BERT+分類器モデルの QWK

		BERT+SVR	BERT+XGBoost
グローバル	設問 1	0.175	0.047
	設問 2	0.233	-0.093
	設問 3	0.217	0.277
自然科学	設問 1	0.148	0.116
	設問 2	0.427	0.336
	設問 3	0.345	0.164
東アジア	設問 1	0.156	-0.349
	設問 2	0.109	0.042
	設問 3	0.122	0.208
批判的思考	設問 1	0.552	0.256
	設問 2	0.184	0.219
	設問 3	0.401	0.470
平均		0.256	0.141

表 3　BERT+分類器モデルの P 値

モデル 1	モデル 2	P 値
BERT+SVR	BERT+XGBoost	0.053

表 3 では P 値が 0.05 より高いため両者に有意差は認められていない．しかし，表 2 で SVR での実装

は，XGBoost での実装に比べて各設問で平均値に近い出力をしていることがわかる．そのため，安定した平均数値を出す SVR での実装が良いことが考えられる．

追加で行なった BERT と BOW の組み合わせについての結果とその平均を表 4 に示す．また，T 検定を適用した P 値について BERT+SVR の結果を含めて表 5 に記載する．

表 4　BERT+BOW+分類器モデルの QWK

		BERT+ BOW+SVR	BERT+ BOW+XGBoost
グローバル	設問 1	0.420	0.436
	設問 2	0.446	0.743
	設問 3	0.429	0.166
自然科学	設問 1	0.756	0.695
	設問 2	0.607	0.412
	設問 3	0.360	0.210
東アジア	設問 1	0.596	0.596
	設問 2	0.625	0.534
	設問 3	0.321	0.443
批判的思考	設問 1	0.813	0.836
	設問 2	0.621	0.280
	設問 3	0.440	0.455
平均		0.535	0.484

表 5　BERT+BOW+分類器モデルの P 値

モデル 1	モデル 2	P 値
BERT+BOW+SVR	BERT+BOW+XGBoost	0.095
BERT+SVR	BERT+BOW+SVR	0.000
BERT+SVR	BERT+BOW+XGBoost	0.005

表 5 では BERT+BOW+SVR と BERT+BOW+XGBoost の P 値が 0.05 より高いため両者に有意差は認められていない．しかしながら，BERT だけでの実装に比べ BOW を組み込んだモデルは有意差があり，平均値から BOW を組み込むことで精度が上がることがわかる．表 4 においてもいくつかの設問で差はあるが，全体的に大きな差は出ていない．設問に対する回答内容が絞られるような自由記述の要素が少ない設問 (1 および 2(ただしグローバルを除く)) の場合には SVR が高い傾向が見られた．また，自由記述に近い設問 (グローバルの設問 2 や設問 3) ではおおむね XGBoost が高い値を示している．

5.5　考察

BERT 単体では分散表現ベクトルから点数との関連性を見いだすことが難しいため，SVR や XGBoost では精度が低く，両者に大きな差は出ていないと考えられる．一方 BOW を埋め込むことで単語自体の出現頻度と分散表現ベクトルの両者の特徴量を利用することが可能になるため，精度の向上が見られたと考えられる．BERT+BOW において XGBoost が自由記述の設問において精度が高く見られたのは決定木で独自のルールが構築されたためだと考える。

6　おわりに

本論文では日本語小論文の採点システムを構築し，BERT や BOW を使った際と SVR や XGBoost を使った際でのそれぞれのシステムを QWK のを用いて評価した．BERT+分類器の実装では SVR が平均値に近い安定した値を出力することを示した．また，文書ベクトルとして BERT だけでなく BOW を取り込むことで QWK が向上することを実験的に示した．モデルの比較として，SVR と XGBoost では設問の特性によって精度の差が観測された．概ね，小論文の記述すべき内容がはっきりしている場合は SVR が高く，自由記述の幅が大き設問に対しては XGBoost が高い傾向を示した．

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」の支援を受けた．

参考文献

[1] 竹谷謙吾, 高井浩平, 森康久仁, 須鎗弘樹. 日本語記述式問題の自動採点システムの提案. In *IEICE Conferences Archives*. The Institute of Electronics, Information and Communication Engineers, 2018.

[2] 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. In *IEICE Conferences Archives*. The Institute of Electronics, Information and Communication Engineers, 2019.

[3] 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, Vol. 62, pp. 1586–1604, 2021.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.