模範答案のみを利用した日本語小論文採点支援システム

江島知優
岡山大学大学院自然科学研究科
chihiro.100522@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

概要

記述式問題の自動採点では近年，BERT (Bidirectional Encoder Representations from Transformers) 等を用いた機械学習によるモデルが提案されている．本研究では BERT の tokenizer を利用して小論文の bag of words を作成し，模範解答で訓練されたニューラルネットワークを通すことで得られたベクトルを用いて自動採点を行うシステムを作成した．このシステムに対して 100 字から 800 字の日本語小論文データを用意し，インドメインで学習を行った手法と One-shot learning を応用した手法について評価実験により精度を求める．

1 はじめに

一般に小論文の採点はルーブリックや模範解答などを利用して人手で行われている．しかしたくさんの答案を人手で採点するとなると採点者の負担が非常に大きくなるため自動採点システムの作成が必要である．記述式問題の自動採点では項目反応理論に基づく能力推定値を活用した手法 [1] や IDF の評価手法 [2] などさまざまな手法が提案されている．近年では BERT [3] を用いた手法 [4] や Neural Attention モデルを利用した手法 [5] も提案されている．実験に用いられているデータセットに注目してみても 100 文字以下の短答式記述問題と呼ばれるものを対象として確信度推定を行う手法 [6] や外国人の日本語学習者の文章を対象としたトランズダクティブ学習手法の研究 [7] など多様な研究が行われている．日本人が書いた日本語小論文では答案ごとの文章量や文章力に大きな差がつかないため，文章中に出現する単語に注目した自動採点システムを作成することが考えられる．本論文では竹内らの参照文書を利用した小論文採点手法 [8] でも利用されている小論文データから BERT の tokenizer を利用して bag of words を作成し，模範解答で訓練されたニューラルネットワークとサポートベクター回帰 (SVR：Support Vector Regression) を組み合わせたモデルを提案し 100 文字から 800 文字の長い小論文に対して性能の分析をするための評価実験を行った．

また，実際に小論文自動採点を行う際に同課題同設問の採点済み小論文を多数集めることは難しいため，One-shot learning の考え方を応用して自動採点を行うことにも取り組んだ．One-shot learning は少数のデータで学習する手法であり Matching Network[9] や Siamese Neural Network[10] などの方法がある．データの分布を推定する手法 [11] なども提案されている．本研究では 1 つの模範解答と同課題他設問の答案をもとに One-shot learning を実施し，精度評価を行った．

学習データとテストデータのみでの採点手法との違いとして，1 つの模範答案を利用するということが本研究の最大の特徴である．

2 小論文自動採点システム

本研究では 1 点から 5 点の小論文に対して自動採点システムを作成した．図 1 に作成した自動採点システムの概要を示す．各小論文に対して BERT の tokenizer を利用して 32000 次元の bag of words ベクトルを作成する．bag of words ベクトルでは小論文中に出現した単語はその回数を値とし，出現しない単語の値は 0 とする．次にニューラルネットワークの教師データとして 5 点の模範答案 1 つとネガティブサンプリングを行うために 0 点とした同課題他設問の答案 k 各 1 つの計 3 答案で同様に bag of words を作成する．この教師データで訓練されたニューラルネットワークに bag of words ベクトル 32000 次元を入力として与え 1epoch 学習を行うことで中間層 768 次元のベクトルを得る．これを SVR への入力として最終的な推定点数を求める．この際の推定点数には丸め処理を行い，1 点から 5 点の整数値としている．また，データの可視化のために UMAP を用いて次元削減した結果を 2.1 節で述べる．

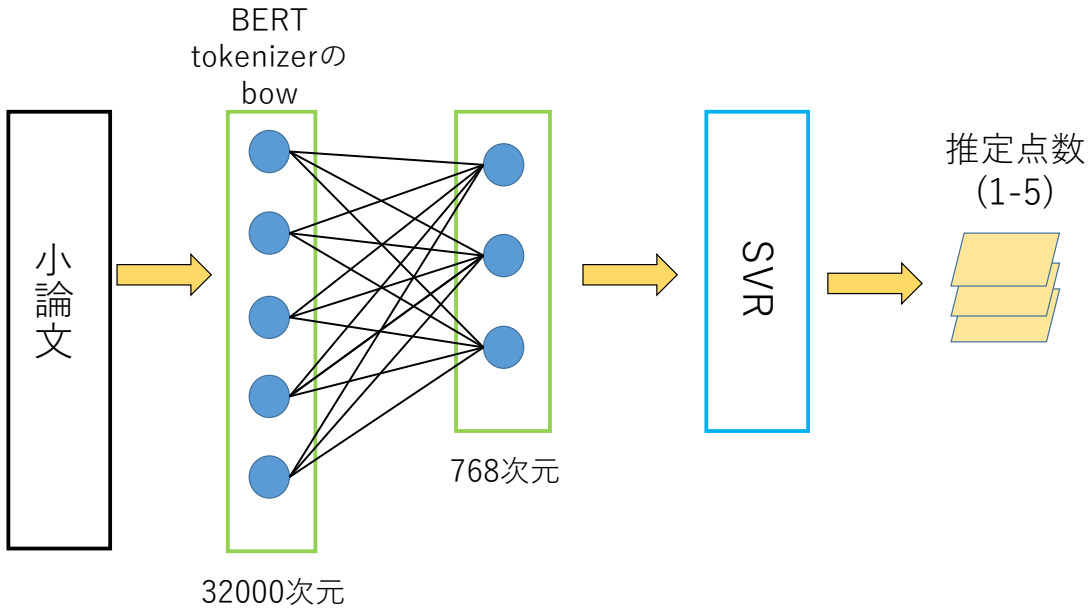

図 1　BERT の tokenizer を利用した自動採点システム

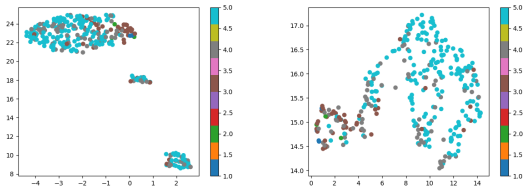

図 2　「自然科学の構成と科学教育」設問 1 で模範答案での学習なし (左) と学習あり (右) の UMAP

2.1　UMAP

本研究では実験の前段階として UMAP[12] による可視化を行った．UMAP は次元削減のための多様体学習技術であり，データの可視化や学習の前処理に応用されている．図 2 に「自然科学の構成と科学教育」設問 1 で bag of words 32000 次元ベクトルと，模範答案および同課題他設問の答案を教師データとしてニューラルネットワークで学習後の 768 次元ベクトルそれぞれの UMAP を示す．2 つの UMAP を比較すると学習後のものでは低い点数の小論文データが図の左下部分に集中し，各点数ごとの分離が容易になっているため精度の向上が期待できる．

3　評価実験

本章では評価実験に使用した小論文，実験設定，評価尺度，実験結果を説明し結果の分析を行う．

3.1　小論文データ

小論文データとして 2016 年から 2017 年にかけて実施した講義の受講者の解答データを用いる．「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 つの講義からなり，それぞれの講義に対して各 3 つの設問が用意されている．これらの答案は理解力・論理性・妥当性・文章力の 4 観点について人手で 1 点から 5 点までの 5 段階で採点されている．今回はこの 4 観点のうち理解力に対して実験を行った．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，「東アジア経済の現状」および「批判的思考とエセ科学」が 290 である．

3.2　実験設定

bag of words を作成するために Mecab を利用した日本語訓練済みの HuggingFaceBERT[1)] を BERT の言語モデルとして使用した．このモデルの語彙サイズは 32000 であるので 32000 次元の bag of words が作

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

成できる．今回は大きく 2 つの実験を行った．1 つ目はインドメインで 2 つの手法を比較する実験である．2 章で述べた小論文自動採点システムについての精度評価を行う．比較対象として模範答案で学習させる前の 32000 次元の bag of words でも同様の実験を行った．この実験では採点対象の設問の小論文データを 8:2 に分割し，8 割を SVR の学習データ，2 割をテストデータとして利用している．

2 つ目の実験は One-shot learning の手法を小論文自動採点に応用するというものである．実験 1 でも使用した模範答案と同課題他設問の答案に対してニューラルネットワークの中間層 768 次元のベクトルを求め SVR を学習させる．この実験ではテストデータとして採点対象の設問の全小論文データを利用している．

3.3 評価尺度

本節では今回作成したシステムの評価尺度を説明する．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，Correlation，重み付きカッパ係数（QWK: Quadratic Weighted Kappa）および RMSE での評価を行う．Accuracy の計算式を式 (1)(2) に示す．

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

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

小論文のデータ数を n，i 番目の点数データを A_i,B_i とする．Accuracy は人手の採点点数とシステムの推定点数の一致率を表し 1 に近いほど一致率が高いと言える．

Correlation は 2 つの結果の相関係数を算出する指標である．人手の採点結果とシステムの推定点数の一方で高く点数をつけたものはもう一方でも高くつけているか，低いものは低く点数付けしているかどうかを測る．式 (3)(4)(5)(6) に示す．

$$Cov(CS, VS) = \sum_{i=1}^{E} (CS_i - \overline{CS})(VS_i - \overline{VS}) \tag{3}$$

$$\sigma_{cs} = \sum_{i=1}^{E} (CS_i - \overline{CS}) \tag{4}$$

$$\sigma_{vs} = \sum_{i=1}^{E} (VS_i - \overline{VS}) \tag{5}$$

$$Correlation = \frac{Cov(CS, VS)}{\sigma_{cs}\sigma_{vs}} \tag{6}$$

人手で採点したデータの集合を CS，システムで採点したデータの集合を VS とする．CS と VS の共分散は式 (3) により求められ，式 (4) で CS の式 (5) で VS の標準偏差をそれぞれ算出する．これらより相関係数は (6) によって表すことができる．ここで E は採点する小論文のデータ数である．Correlation は 2 つの採点結果の関係性の強さを表し，1 に近いほど関係性が強い．

次に QWK について説明する．QWK は 2 つの採点結果の点数の差の 2 乗を重み付けし，ズレが大きいほど大きなペナルティを与える指標である．式 (7)(8) に示す．

$$byChance(a,b) = \frac{num(a) \times num(b)}{n} \tag{7}$$

$$QWK = 1 - \frac{\sum_{a,b=1}^{5} ob(a,b) \times |a-b|^2}{\sum_{a,b=1}^{5} byChance(a,b) \times |a-b|^2} \tag{8}$$

1 人の採点者が a と採点した回数を $num(a)$ とし，2 人の採点者が a,b と採点した回数を $ob(a,b)$ で表す．QWK は 1 に近いほど実際の点数とシステムの推定点数との相関が強いといえる．

最後に RMSE について式 (9) に示す．これは 2 つの採点結果の点数の平均の差を表す評価指標である．RMSE は 0 に近いほど平均の差が小さいことを表す．

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} |A_i - B_i|^2}{n}} \tag{9}$$

3.4 実験結果

小論文の各設問に対して Accuracy，Correlation，QWK，RMSE をそれぞれ算出した．表 1 では小論文を 32000 次元の bag of words に変換したものを SVR への入力とした結果と，その 32000 次元ベクトルを採点対象の模範答案と同課題他設問の答案を用いて学習して得た 768 次元のベクトルを SVR への入力とした結果を示している．また表 2 には模範答案 1 つと同課題他設問の解答文書 2 つで SVR を訓練し，One-shot learning を行った際の結果を示す．

3.5 分析

インドメインでの結果より bag of words 32000 次元のベクトルを採点対象の模範答案と同課題他設問

表 1　インドメインで 8:2 学習を行った際のシステムの評価

			学習なし 32000 次元ベクトルでの自動採点				模範答案学習後の自動採点			
		文字数	Accuracy	Corr	QWK	RMSE	Accuracy	Corr	QWK	RMSE
グローバル	設問 1	300 字	0.500	0.347	0.342	0.769	0.561	0.403	0.402	0.759
	設問 2	250 字	0.636	0.686	0.680	0.728	0.591	0.742	0.736	0.674
	設問 3	300 字	0.652	0.431	0.423	0.696	0.561	0.419	0.416	0.728
自然科学	設問 1	100 字	0.530	0.638	0.622	0.985	0.682	0.776	0.773	0.674
	設問 2	400 字	0.523	0.654	0.650	0.754	0.585	0.694	0.693	0.713
	設問 3	500-800 字	0.485	0.317	0.312	0.929	0.348	0.372	0.357	0.913
東アジア	設問 1	300 字	0.500	0.633	0.615	0.743	0.466	0.599	0.582	0.799
	設問 2	250 字	0.586	0.766	0.752	0.682	0.534	0.732	0.729	0.754
	設問 3	300 字	0.569	0.396	0.391	0.766	0.534	0.414	0.396	0.754
批判的思考	設問 1	100 字	0.500	0.737	0.696	1.106	0.707	0.912	0.906	0.541
	設問 2	400 字	0.500	0.633	0.631	0.809	0.414	0.581	0.577	0.910
	設問 3	500-800 字	0.397	0.122	0.114	0.871	0.431	0.228	0.221	0.910
平均			0.532	0.530	0.519	0.820	0.535	0.573	0.566	0.698

表 2　One-shot learning を行った際のシステムの評価

		Accuracy	Corr	QWK	RMSE
グローバル	設問 1	0.430	0.156	0.114	0.967
	設問 2	0.480	0.168	0.130	0.967
	設問 3	0.046	0.358	0.079	1.609
自然科学	設問 1	0.254	0.679	0.412	0.955
	設問 2	0.151	0.407	0.174	1.498
	設問 3	0.159	0.256	0.108	1.692
東アジア	設問 1	0.272	0.181	0.088	1.338
	設問 2	0.264	0.280	0.164	1.205
	設問 3	0.462	0.238	0.180	0.901
批判的思考	設問 1	0.521	0.653	0.416	0.990
	設問 2	0.445	0.492	0.471	0.855
	設問 3	0.317	0.380	0.318	1.091
平均		0.317	0.354	0.221	1.172

の答案を用いて学習した手法が bag of words の学習なしで SVR への入力とした場合よりも平均 QWK が 0.037 向上するなど 4 つの評価指標すべてで精度向上が確認できた．特に短い文章では単語が得点に与える影響が大きく，QWK が向上したと考えられる．UMAP に関してもこの結果より 2.1 節でも述べたように UMAP を参照することでシステムの精度向上が期待できるといえる．

また，One-shot learning を利用した手法では，「自然科学の構成と科学教育」の設問 1 が QWK0.412，「批判的思考とエセ科学」の設問 1 が QWK0.416 というように 100 文字以内の小論文に対して精度が高くなっており，短い文章に対して本手法が有効であることを示した．One-shot learning でもインドメインでの結果と同様に短い文章では高得点の文章に同一の単語が登場しやすく精度が高くなったと考えることができる．その一方で制限文字数が長い設問については具体例を挙げて議論させるものも多く，精度が悪くなってしまった．この問題に対しては attention を利用した機構を新たに設ける等の検討をしていきたい．

4　おわりに

本研究では BERT の tokenizer の bag of words を用いた小論文自動採点システムを作成し Accuracy，Correlation，QWK および RMSE といった評価指標を用いて性能評価を行った．bag of words 32000 次元のベクトルを学習なしで SVR への入力とした場合と比較すると，採点対象の模範答案と同課題他設問の答案を用いて学習した場合において 12 の設問の平均 QWK が向上する等，本手法が小論文自動採点に有効であることを示した．また，同様に模範答案と同課題他設問の答案を用いての One-shot learning にも取り組んだが制限文字数が長い設問において精度が悪くなるといった問題が見られたため今後の検討課題としたい．

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] 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.

[2] 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.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.
[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, pp. 1181–1184, 2020.
[5] 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.
[6] 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.
[7] 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.
[8] 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ジャーナル 62 巻 9 号, pp. 1586–1604, 2021.
[9] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In **NeurIPS**, 2016.
[10] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In **ICML Deep Learning workshop**, 2015.
[11] Shuo Yang, Lu Liu, and Min Xu. Free lunch for few-shot learning: Distribution calibration. In **ICLR**, 2021.
[12] Leland McInnes, John Healy, and James Melville. Umap:uniform manifold approximation and projection for dimension reduction. In **Journal of Open Source Software 3**, 2018.