アノテータのバイアスを考慮した記述・論述式自動採点手法

岡野将士
電気通信大学大学院
okano@ai.lab.uec.ac.jp

宇都雅輝
電気通信大学大学院
uto@ai.lab.uec.ac.jp

1 はじめに

近年，新しい時代に必要となる能力を評価する手法として記述・論述式試験が注目されている．しかし，大規模試験に記述・論述式試験を導入する場合，時間的・金銭的コストの高さや採点の公平性の担保の難しさといった課題が存在する．自動採点手法はこれらの解決策の一つとして注目されている．

自動採点を実現する方法として，事前に定義された特徴量を用いる手法 [1, 2, 3, 4] や深層学習モデルを用いた手法 [5, 6, 7, 8, 9] が存在する．これらの手法の多くは，教師あり機械学習を用いてモデル学習を行う．しかしながら，教師あり機械学習では，モデル学習に利用するデータセットの質がモデルの構築精度に影響を与えるという欠点が知られている．教師あり機械学習を用いた自動採点モデルでは，モデル学習に利用する採点済み答案データセット中の得点はバイアスのない正確な得点であると仮定する．しかし，大規模試験では多数のアノテータが分担して採点を行うことが一般的であり，そのような場合，個々の答案に対する得点がアノテータの特性（甘さ/厳しさなど）に強く依存してしまう [10]．このようなアノテータの特性の影響を受けたデータを利用した場合，学習されるモデルもその影響を受け，予測性能が低下することが報告されている [11]．

他方で，教育・心理測定の分野において，アノテータの特性の影響を考慮して真の得点を推定できる手法が多数提案されている．具体的には，数理モデルを用いたテスト理論の一つである項目反応モデルに，アノテータの特性を表すパラメータを加えたモデルとして提案されている [10, 12, 13, 14, 15, 16]

そこで本研究では，アノテータの特性を考慮した項目反応モデルを教師あり機械学習を用いた自動採点モデルに組み込んだ，アノテータのバイアスに頑健な新たな自動採点手法を提案する．具体的には，アノテータが与える得点データから項目反応モデルを用いて各答案の真の得点を推定し，これを目的変数として自動採点モデルを学習する．この手法は様々な自動採点モデルで利用できるが，本研究では特徴量ベース自動採点モデルとして EASE[4]，深層学習自動採点モデルとして，LSTM に基づくモデル [7] と BERT を用いたモデル [9] への組み込みを行う．提案手法を利用することで，アノテータのバイアスに頑健なモデル学習と得点予測が期待できる．

2 データ

本研究では，ある記述・論述式問題に対する J 人の受験者 $\mathcal{J} = \{1, \ldots, J\}$ の答案集合 $\boldsymbol{A}$ と，各答案を R 人のアノテータ $\mathcal{R} = \{1, \ldots, R\}$ で分担して採点した得点集合 $\boldsymbol{U}$ で構成されるデータを想定する．

答案集合 $\boldsymbol{A}$ は，受験者 $j \in \mathcal{J}$ の答案 e_j の集合であり，得点集合 $\boldsymbol{U}$ は答案 e_j に対してアノテータ $r \in \mathcal{R}$ が K 段階 $\mathcal{K} = \{1, \ldots, K\}$ で与えた得点 U_{jr} の集合として，$\boldsymbol{U} = \{U_{jr} \in \mathcal{K} \cup \{-1\} | j \in \mathcal{J}, r \in \mathcal{R}\}$ と定義される．ここで，$U_{jr} = -1$ は欠測データを表す．欠測データは答案 e_j にアノテータ r が割り当てられていない場合に生じる．実際の採点場面ではアノテータの負担軽減のために，個々の答案に数名のアノテータを割り当てて採点が行われるため，このような欠測が生じる．

3 自動採点モデル

3.1 特徴量ベース自動採点モデル

事前に定義した特徴量を用いた自動採点手法は自動採点を実現する手法として古くから研究されており，実際の試験現場でも用いられている [1, 2, 3]．それらの中でも特に，EASE（Enhanced AI Scoring Engine）[4] は自動採点のコンペティションで 3 位に入賞したモデルであり，数多くの研究で比較対象として用いられている．このモデルでは，使用語彙や品詞情報，単語数などに基づく特徴量を用いて得点の推定を行っている．具体的には，これらの特徴量を答案から抽出し，それらを回帰モデルに入力する

ことで自動採点を実現している．また，回帰モデルとしてベイジアンリッジ回帰（BLRR）とサポートベクター回帰（SVR）が使用されることが多い [3].

3.2 深層学習自動採点モデル

深層学習自動採点モデルは対象答案の単語系列を深層学習モデルに入力することで，人手で設計した特徴量を利用することなく，採点を行う手法であり，近年多くの手法が提案されている [5, 6, 7, 8, 9]. LSTM に基づくモデル [7] はそれら最先端研究のベースライン手法として知られている．このモデルでは，答案の単語系列を入力し，5 つの層（Lookup Table Layer・Convolution Layer・Recurrent Layer・Pooling Layer・Linear Layer with Sigmoid Activation）を通して得点を予測する．LSTM は 3 層目の Recurrent Layer で用いられ，得点予測に有効な特徴量を文脈を考慮して抽出する．また，5 層目の Linear Layer with Sigmoid Activation では Pooling 層の出力ベクトル $\boldsymbol{M}_j$ から得点を表すスカラー値を求める．具体的には，シグモイド関数 σ を用いて，$\hat{U}_j = \sigma(\boldsymbol{W}\boldsymbol{M}_j + b)$ で計算する．ここで，$\boldsymbol{W}$ と b は重みとバイアスを表すパラメータである．この際，$\hat{U}_j$ は $(0, 1)$ の値を取るため，一次変換を行い，実際の得点尺度に合わせる．

また，様々なタスクで最高精度を達成している BERT[17] を用いたモデル [9] も提案されている．このモデルでは，答案の単語系列を入力し，多層の双方向 Transformer（BERT）を通すことで，中間表現 $\boldsymbol{M}_j$ を生成する．この中間表現 $\boldsymbol{M}_j$ を LSTM に基づくモデルと同様の Linear Layer with Sigmoid Activation に通すことで得点を計算する．

3.3 従来手法の問題点

上述した自動採点モデルの多くは，教師あり機械学習を用いてモデル学習を行う．この際，採点済みの答案データセットを教師データとして活用する．具体的には，次式で定義される平均二乗誤差（mean squared error：MSE）を損失関数として，誤差逆伝搬法で学習することが一般的である．

$$MSE(\boldsymbol{U}, \hat{\boldsymbol{U}}) = \frac{1}{J}\sum_{j=1}^{J}(U_j - \hat{U}_j) \tag{1}$$

ここで，U_j は e_j の得点を，$\hat{U}_j$ は e_j の予測得点を表す．しかし，学習データ中の得点データはアノテータの特性に強く依存する．この場合，自動採点モデルにも，アノテータの特性の影響が反映され，予測精度が低下してしまうことが指摘されている．本研究では，この問題を解決するために，項目反応理論を用いる．

4 項目反応モデル

項目反応理論 (Item Response Theory：IRT) は，コンピュータ・テスティングの普及とともに近年様々な分野で実用化が進められている数理モデルを用いたテスト理論の一つである．本研究では，最先端研究である宇都・植野のモデル [13, 18] を利用する．

宇都・植野のモデル [13, 18] では，受験者 $j \in \mathcal{J}$ のある答案に対し，アノテータ $r \in \mathcal{R}$ が得点 $k \in \mathcal{K}$ を与える確率は次式で定義される．

$$P_{jrk} = \frac{\exp\sum_{m=1}^{k}[\alpha_r(\theta_j - \beta_r - d_{rm})]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}[\alpha_r(\theta_j - \beta_r - d_{rm})]} \tag{2}$$

ここで，α_r，β_r，d_{rk} はそれぞれアノテータ r の一貫性，厳しさ，得点 k に対するアノテータ r の厳しさを表し（ただし，パラメータの識別性のために，$d_{r1}=0$，$\sum_{k=2}^{K} d_{rk}=0$ を仮定），θ_j は受験者 j の真の能力を表す潜在変数である．この θ_j は採点対象となる答案の数が受験者ごとに一つであることから，その受験者の答案 e_j の真の得点（以降では IRT 得点と呼ぶ）を表す潜在変数とみなせる．本研究のアイディアは，このモデルによって推定される IRT 得点 θ_j を用いて自動採点モデルを学習することにある．

5 提案手法

本研究では，アノテータの特性を考慮した項目反応モデルを自動採点モデルに組み込むことで，学習データに含まれるアノテータのバイアスに頑健な自動採点手法を提案する．提案手法は，IRT による得点補正と，自動採点モデルの学習の二段階で構成される．モデルの学習段階と得点の予測段階について，手順を説明する．

1. **モデル学習**

   1) 得点データ $\boldsymbol{U}$ から，IRT モデルを用いて θ_j を推定する．この θ_j を，答案 e_j に対する得点とする．2) 得点 θ_j を予測するように，自動採点モデルを学習する．具体的には自動採点モデルの損失関数を次の MSE で定義し，誤差逆伝播法によりパラメータを学習する．

$$MSE(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{1}{J}\sum_{j=1}^{J}(\theta_j - \hat{\theta}_j)^2 \tag{3}$$

   ここで $\hat{\theta}_j$ は，自動採点モデルの予測値を表す．

表 1　検証モデルの設定

	Convolution Layer	Recurrent Layer	Pooling Layer
CNN-LSTM(MoT)	あり	LSTM	Mean over Time
CNN-LSTM(Last)	あり	LSTM	Last pooling
LSTM(MoT)	なし	LSTM	Mean over Time
LSTM(Last)	なし	LSTM	Last pooling
2L-LSTM(MoT)	なし	2-Layer	Mean over Time
2L-LSTM(Last)	なし	2-Layer	Last pooling
Bidirectional LSTM	なし	Bidirectional	Last pooling

2. **得点予測**

得点予測は，前節で学習されたモデルを用いて θ_j を予測することで行う．ただし，IRT では一般に θ_j の分布として標準正規分布を仮定するため，θ_j は $[-\infty, \infty]$ の範囲の値をとり，元の得点とは尺度が変わってしまう．元の得点尺度に合わせるために，本実験では次のように IRT モデルに基づく期待得点 $\hat{U}_j$ を求める．

$$\hat{U}_j = \frac{1}{R}\sum_{r=1}^{R}\sum_{k=1}^{K} k \cdot P_{jrk} \tag{4}$$

なお，提案手法の本来の利用方法でないが，提案手法では個々のアノテータが与える得点も予測することができる．具体的には，アノテータ r が $e_{j'}$ に与える得点は次式で予測できる．

$$\hat{U}_{j'r} = \sum_{k=1}^{K} k \cdot P_{j'rk} \tag{5}$$

6　実データによる評価実験

6.1　実データ

本実験では，実データとして自動採点モデルのベンチマークデータとして広く利用されている Automated Student Assessment Prize（ASAP）を使用する．ASAP は 2012 年にヒューレット財団がスポンサーとなって開催されたコンペティションのデータであり，8 つの異なるトピックに対する答案データと得点データで構成されている．ただし，ASAP のデータにはアノテータの情報が含まれていないため，提案手法を直接は適用できない．そのため，新たにアノテータを雇用して ASAP の答案データを再度採点し，本実験で用いる得点データを収集した．具体的には，先行研究で予測精度が最も高かったトピック 5 の 1805 個の答案に対して，Amazon Mechanical Turk で募集した英語ネイティブ 38 名のアノテータを 1 つの答案あたり 3〜5 名割り当てて，ASAP と同様に 5 段階の採点を行った．ASAP の得点データとの相関は，平均で 0.675 であった．

6.2　得点予測の頑健性の評価

本節では，提案手法を利用することで，アノテータのバイアスに頑健な自動採点モデルを学習できるかを評価する．本実験では，個々の答案を採点するアノテータを変化させても，安定した性能の自動採点モデルを学習できるかによってこれを評価する．

具体的には，項目反応モデルの研究における実験手順 [19, 20] を参考に，以下の手順で評価実験を行った．1) 得点データから IRT モデルのアノテータに関するパラメータを推定した．2) 各答案に与えられた複数のアノテータの得点からランダムに 1 つの得点を選択することで，各答案に単一の得点が与えられたデータセットを作成した．同様の手続きで 10 パターンの異なるデータセットを作成した．これらのデータセットを $\{\boldsymbol{U}'_1, \ldots, \boldsymbol{U}'_{10}\}$ とする．3) n 番目の得点データセット $\boldsymbol{U}'_n$ から各答案に対する IRT 得点を推定した．推定時には，手順 1 で推定したアノテータに関するパラメータを所与とした．4) 得られた IRT 得点と答案文のデータセットを用いて，5 分割交差検証法で各答案の予測得点を求めた．5) 手順 4 を $n = \{1, \ldots, 10\}$ について行ったあと，n 番目のデータセットから求めた予測得点と n' 番目の得点データセットから推定した予測得点とのカッパ係数，重み付きカッパ係数（Linear Weighted Kappa：LWK），2 次重み付きカッパ係数（Quadratic Weighted Kappa：QWK），平均絶対誤差（Mean Absolute Error：MAE），平均平方二乗誤差（Root Mean Square Error：RMSE），相関係数を $n \in \{1, \ldots, 10\}$，$n' \in \{1, \ldots, 10\}$ の全ての組み合わせについて求め，それらの平均を算出した．

比較のために，IRT を利用しない既存の自動採点手法についても同様の実験を行った．具体的には，手順 2 で作成したデータセット $\{\boldsymbol{U}'_1, \ldots, \boldsymbol{U}'_{10}\}$ を用いて，手順 4，5 と同様に 5 分割交差検証法で得点を予測し，予測された得点同士の一致性指標を求めた．

また，提案手法・従来手法ともに，深層学習自動採点モデルとしては LSTM と BERT を用いたモデル，特徴量ベース手法としては EASE について上記の実験を行った．なお，LSTM 自動採点モデルについては，各層の有無や構成について複数の方式が提案されている．そのため，表 1 に記載した複数の構成のモデルについて検証を行った．また，EASE に

表 2　予測の頑健性の評価結果

	カッパ係数			LWK			QWK			MAE			RMSE			相関係数		
	提案	既存	p 値	提案	既存	p 値	提案	既存	p 値	提案	既存	p 値	提案	既存	p 値	提案	既存	p 値
CNN+LSTM(MoT)	**.749**	.624	**	**.778**	.727	**	.818	**.830**	**	**.139**	.215	**	**.191**	.301	**	**.937**	.931	*
CNN+LSTM(Last)	**.639**	.459	**	**.701**	.551	**	**.678**	.663	.098	**.160**	.302	**	**.212**	.400	**	**.829**	.783	**
LSTM(MoT)	**.831**	.697	**	**.845**	.779	**	**.881**	.863	**	**.102**	.175	**	**.142**	.237	**	**.965**	.958	**
LSTM(Last)	**.612**	.371	**	**.624**	.514	**	**.682**	.670	.121	**.229**	.397	**	**.300**	.518	**	**.804**	.775	**
2L-LSTM(MoT)	**.828**	.661	**	**.842**	.752	**	**.879**	.846	**	**.107**	.197	**	**.147**	.268	**	**.963**	.946	**
2L-LSTM(Last)	**.665**	.420	**	**.679**	.561	**	**.728**	.711	*	**.207**	.359	**	**.272**	.470	**	**.848**	.820	**
Bidirectional LSTM	**.608**	.386	**	**.624**	.508	**	**.701**	.649	**	**.216**	.362	**	**.282**	.470	**	**.816**	.772	**
BERT	**.790**	.629	**	**.808**	.743	**	**.876**	.851	**	**.121**	.233	**	**.159**	.311	**	**.960**	.935	**
EASE(BLRR)	**.879**	.851	**	**.888**	.881	**	**.919**	.916	.15	**.065**	.094	**	**.085**	.307	**	**.984**	.917	**

表 3　予測得点の評価結果

	カッパ係数			LWK			QWK			MAE			RMSE			相関係数		
	提案	既存	p 値	提案	既存	p 値	提案	既存	p 値	提案	既存	p 値	提案	既存	p 値	提案	既存	p 値
CNN+LSTM(MoT)	**.266**	.208	**	**.434**	.395	**	**.601**	.579	**	**.633**	.688	**	**.791**	.859	**	**.734**	.664	**
CNN+LSTM(Last)	**.124**	.062	**	**.241**	.178	**	**.367**	.310	**	**.799**	.862	**	**.988**	1.057	**	**.509**	.405	**
LSTM(MoT)	**.280**	.252	**	**.449**	.438	*	.616	**.618**	.739	**.613**	.664	**	**.768**	.829	**	**.757**	.691	**
LSTM(Last)	**.198**	.154	**	**.344**	.314	.085	**.496**	.477	.340	**.713**	.785	**	**.890**	.976	**	**.632**	.541	**
2L-LSTM(MoT)	**.283**	.234	**	**.452**	.421	**	**.621**	.605	.061	**.612**	.672	**	**.765**	.836	**	**.760**	.685	**
2L-LSTM(Last)	**.221**	.174	**	**.375**	.344	*	**.533**	.516	.268	**.682**	.753	**	**.854**	.937	**	**.671**	.584	**
bidirectional	**.167**	.093	**	**.312**	.238	**	**.465**	.395	**	**.740**	.825	**	**.920**	1.016	**	**.600**	.481	**
BERT	**.311**	.285	**	**.477**	.474	.490	.642	**.649**	.111	**.597**	.656	**	**.750**	.821	**	**.773**	.702	**
EASE(BLRR)	**.252**	.235	**	**.412**	.405	.115	.574	**.575**	.846	**.630**	.638	*	**.788**	.894	**	**.745**	.627	**

* は p 値が 0.05 未満，** は p 値が 0.01 未満を表す．

ついても用いる回帰モデルは複数考えられるが，先行研究で精度が高い BLRR のモデルを用いる．さらに，提案手法と既存手法で性能に有意な差があるかを確認するために，各指標の平均値について，提案手法と既存手法で t 検定を行った．

実験結果を表 2 に示す．表中では提案手法と既存手法で性能が高い方を太字で示している．表 2 から，ほぼ全ての条件において，提案手法が有意に高い性能を示していることが確認できる．このことから，IRT 得点を目的変数として自動採点モデルを学習する提案手法により，アノテータのバイアスに頑健な自動採点を実現できたことがわかる．

6.3　提案手法の得点予測精度評価

本節では，各アノテータの得点 U_{jr} の予測精度を評価するために，U_{jr} の予測得点と実際の得点との一致度を，前節と同様の一致性指標を用いて 5 分割交差検証法で評価した．具体的な実験手順は次の通りである．提案モデルでは，手順 1，2，3 は前節と同様に行い，手順 4 において式（4）で期待得点を求める代わりに各アノテータの得点の予測値を式（5）で求め，予測された得点と実際の得点との一致性指標を求めた．既存モデルでは，前節の手順 2 で作成したデータセット $\{\boldsymbol{U}'_1, \ldots, \boldsymbol{U}'_{10}\}$ を用いて 5 分割交差検証法で得点予測を行い，予測された得点と実際の得点との一致性指標を求めた．

結果を表 3 に示す．表 3 では，提案手法と既存手法で性能が高い方を太字で示している．提案手法と既存手法の性能を比較すると，ほぼ全ての場合で提案手法が高い性能を示している．これは，IRT によって補正された得点は文章の質を素点そのものよりも正確に反映しているため，提案手法では文章と得点の関係がより適切に学習できたことが要因と考えられる．このことから，提案手法は予測得点の頑健性向上に加え，アノテータが与えた得点の予測にも有効であることが確認できた．

7　まとめ

近年，自動採点技術に注目が集まっているが，教師あり機械学習を用いたモデルでは学習データセットによるモデルの不安定さが指摘されてきた．本研究では，IRT を用いて各答案の真の得点を推定し，それを自動採点モデルに学習させることで，この問題を解決する手法を提案し，実データ実験から有効性を示した．今後は，このモデルを end-to-end にすることで，IRT 得点の推定にテキストの情報も活用し，さらなる性能改善を目指したい．

参考文献

[1] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning and Assessment*, Vol. 4, No. 3, pp. 1–30, February 2006.

[2] Hongbo Chen and Ben He. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1741–1752. Association for Computational Linguistics, October 2013.

[3] Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 431–439. Association for Computational Linguistics, 2015.

[4] Tauber James, Paruchuri Vik, Huang Diana, Jarvis John, Aune Nate, and Kern John. Enhanced ai scoring engine, (2021-1 閲覧）. https://github.com/edx/ease.

[5] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 715–725. Association for Computational Linguistics, August 2016.

[6] Jiawei Liu, Yang Xu, and Lingzhe Zhao. Automated essay scoring based on two-stage learning. CoRR,arXiv, December 2019.

[7] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1891. Association for Computational Linguistics, November 2016.

[8] Fei Dong and Yue Zhang. Automatic features for essay scoring – an empirical study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1072–1077. Association for Computational Linguistics, November 2016.

[9] Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. Language models and automated essay scoring. arXiv, September 2019.

[10] 宇都雅輝, 植野真臣. パフォーマンス評価のための項目反応モデルの比較と展望. 日本テスト学会誌, Vol. 12, No. 1, pp. 56–75, May 2016.

[11] Evelin Amorim, Marcia Cançado, and Adriano Veloso. Automated essay scoring in the presence of biased ratings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, Vol. 1, pp. 229–237. Association for Computational Linguistics, June 2018.

[12] Masaki Uto and Maomi Ueno. Empirical comparison of item response theory models with rater's parameters. *Heliyon, Elsevier*, Vol. 4, No. 5, pp. 1–32, May 2018.

[13] Masaki Uto and Maomi Ueno. A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika, Springer*, Vol. 47, No. 2, pp. 469–496, 2020.

[14] Richard J. Patz, Brian W. Junker, Matthew S. Johnson, and Louis T. Mariano. The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, Vol. 27, No. 4, pp. 341–384, January 2002.

[15] Richard J. Patz and Brian W. Junker. Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, Vol. 24, No. 4, pp. 342–366, December 1999.

[16] J.M. Linacre. *Many-faceted Rasch Measurement*. MESA Press, January 1989.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186. Association for Computational Linguistics, 2019.

[18] 宇都雅輝, 植野真臣. ピアアセスメントにおける異質評価者に頑健な項目反応理論. 電子情報通信学会論文誌. D, 情報・システム, Vol. 101, No. 1, pp. 211–224, January 2018.

[19] 宇都雅輝. 論述式試験における評点データと文章情報を活用した項目反応トピックモデル. 電子情報通信学会論文誌 D, Vol. 102, No. 8, pp. 553–566, August 2019.

[20] Masaki Uto. Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. In *Proceedings of the Conference on Artificial Intelligence in Education*, pp. 494–506. Springer, June 2019.