アノテータ特性を考慮した項目反応モデルを組み込んだ深層学習自動採点手法

岡野将士 [1]　宇都雅輝 [1]
[1] 電気通信大学大学院
{okano,uto}@ai.lab.uec.ac.jp

概要

近年注目されている深層学習を用いた小論文自動採点モデルを利用するためには，採点済みの小論文データセットを用いてモデル学習を行う必要がある．一般にモデル学習はアノテータが与えた得点を真値と仮定して行うが，小論文の採点結果はアノテータの特性に依存することが知られており，そのようなバイアスデータを用いてモデルを学習すると自動採点モデルの性能が低下してしまう．この問題を解決するために本研究では，アノテータの特性を考慮した項目反応理論を組み込んだ深層学習自動採点手法を提案する．

1　はじめに

近年の急速な社会変化に伴い，学校教育では論理的思考力などの育成が求められ，そのような能力を評価する手法の一つとして小論文試験が注目されている．しかし，小論文試験を大規模な試験で実施する場合，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 [1, 2, 3]．自動採点手法はこれらの問題の解決策の一つとして注目されている [4]．

自動採点を実現する手法として，近年では，深層学習を用いた手法が多数提案され，高精度を実現している [5, 6, 7, 8]．深層学習を用いた自動採点モデルを利用するためには，採点済み小論文のデータセットを用いてモデルを学習する必要がある．その際，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．しかしながら，大規模試験では多数のアノテータが分担を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はアノテータの特性（甘さ/厳しさなど）に強く依存してしまう [9]．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，性能が低下することが報告されている [10, 11, 12, 13]．

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取り除くことができる手法が提案されている．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 IRT（Item Respose Theory：項目反応理論）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 [9, 14, 15, 16, 17, 18, 19, 20]．

岡野・宇都 [21, 22, 23] はこのような IRT モデルを深層学習自動採点モデルと組み合わせて用いるアプローチを提案している．具体的には，アノテータの特性を考慮した IRT モデルを用いて訓練データ中の得点からアノテータのバイアスの影響を取り除いた得点（以降では IRT 得点と呼ぶ）を推定し，この得点を元に自動採点モデルの学習を行う手法である．この方法によって，アノテータのバイアスに頑健なモデル学習が可能になることが示されている．しかし，この手法では IRT 得点の推定にアノテータが与えた観測得点のみを用いており，答案文の情報は使用していない．一方で，答案文の内容自体も IRT 得点推定の有益な情報となりうるため，答案文の内容も加味するようにモデルを拡張することでこのアプローチの性能を更に向上できると予測される．

そこで本研究では，IRT モデルと深層学習自動採点モデルを二段階で適用するのではなく，end-to-end で学習できるように拡張する．具体的には，深層学習自動採点モデルの出力層に IRT モデルを組み込んで end-to-end で学習する．この手法では，IRT 得点の推定にアノテータが与えた観測得点だけでなく答案の文章も活用できるため，従来の IRT モデルと比べて IRT 得点の推定精度が改善し，このアプローチの全体的な性能改善につながる．本論文では，実データ実験により提案手法の有効性を示す．

2 データ

本研究では，深層学習自動採点モデルの学習データとして，ある小論文問題に対する J 人の受検者 $\mathcal{J}=\{1,\ldots,J\}$ 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，受検者 $j\in\mathcal{J}$ 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，$\boldsymbol{U}=\{U_{jr}\in\mathcal{K}\cup\{-1\}|j\in\mathcal{J},r\in\mathcal{R}\}$ と定義する．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 深層学習自動採点モデル

深層学習自動採点モデルは小論文答案の単語系列を深層学習モデルに入力することで得点を推定する手法であり，近年多くの手法が提案されている [5, 6, 7, 8]．本研究では Long short-term memory（LSTM）に基づくモデル [24] と Bidirectional Encoder Representations from Transformers（BERT）に基づくモデル [25, 26] を使用する．

LSTM に基づくモデル [24] は深層学習自動採点モデルの基礎モデル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．

BERT に基づくモデル [25, 26] では，答案の単語系列を入力し，中間表現を生成する．この中間表現を 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．

これらの深層学習自動採点モデルは，一般に大量の採点済み答案データを訓練データとして用いてモデル学習を行う．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．また，各答案に複数のアノテータが割り当てられている場合，U_j にはアノテータが与えた観測得点の平均などを用いる．しかし，このような観測得点データはアノテータの特性に強く依存することが知られている [9]．そのようなバイアスデータをモデル学習に使用すると，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 [10, 11]．本研究では，この問題を解決するために IRT を用いる．

4 項目反応理論

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，小論文採点の文脈で利用できる宇都・植野のモデル [15, 16] を利用する．

宇都・植野のモデル [15, 16] では，受検者 $j\in\mathcal{J}$ の答案に対し，アノテータ $r\in\mathcal{R}$ が得点 $k\in\mathcal{K}$ を与える確率が次式で定義される．

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

ここで，α_r はアノテータ r の採点の一貫性，β_{rm} は得点 m に対するアノテータ r の厳しさを表す．ただし，パラメータの識別性のために，$\beta_{r1}=0$ を仮定する．また，θ_j は受検者 j の答案 e_j に対応した潜在得点であり，アノテータバイアスの影響を取り除いた得点とみなせる．本研究ではこの潜在得点を IRT 得点と呼ぶ．

5 IRT 得点を用いた自動採点手法

岡野・宇都 [21, 22, 23] は上記の IRT モデルと自動採点モデルを組み合わせて用いる手法を提案した．この手法では観測得点データ $\boldsymbol{U}$ から IRT 得点 θ_j 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．以下で詳細を説明する．

モデル学習は，IRT モデルによる得点推定と自動採点モデルの学習の二段階で行われる．具体的な手順は以下の通りである．1）アノテータが与える得点データ $\boldsymbol{U}$ から，アノテータの特性の影響を取り除いた各答案 e_j の IRT 得点 θ_j を式（2）の IRT モデルを用いて推定する．2）手順 1 で求めた IRT 得点 θ_j を予測するように自動採点モデルを学習する．具体的には損失関数を $MSE(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})=\frac{1}{J}\sum_{j=1}^{J}(\theta_j-\hat{\theta}_j)^2$ と定義し，誤差逆伝播法によりパラメータを学習す

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

学習されたモデルを用いて新たな答案 $e_{j'}$ の得点を予測する手順は以下の通りである．1）答案 $e_{j'}$ の IRT 得点 $\theta_{j'}$ を自動採点モデルを用いて予測する．2）IRT 得点 $\theta_{j'}$ とアノテータの特性パラメータを用いて，IRT モデルに基づく期待得点を以下の式で求め，この値を予測得点とする．

$$\hat{U}_{j'} = \frac{1}{R}\sum_{r=1}^{R}\sum_{k=1}^{K} k \cdot P_{j'rk} \tag{3}$$

この手法を用いることで，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可能になった．しかし，モデル学習の手順 1 で行われる IRT 得点の推定にはアノテータが与える観測得点のみを用いており，答案文は情報として使用しない．一方で，答案文の内容自体も IRT 得点推定の有益な情報となりうるため，答案文の内容も加味するようにモデルを拡張することでこのアプローチの性能を更に向上できると予測される．

6 提案手法

以上を踏まえ，本研究では，深層学習自動採点モデルの出力層にアノテータの特性を考慮した IRT モデルを組み込み，end-to-end で学習できるように拡張した手法を提案する．具体的には，深層学習自動採点モデルの最終層にあたる Linear Layer with Sigmoid Activation を活性化関数を利用しない Linear Layer に変更し，その出力を式（2）における IRT 得点 θ_j とみなす手法を提案する．提案手法による LSTM と BERT に基づくモデルの概念図を図 1 に示す．提案手法に基づくモデルの学習は end-to-end で行い，深層学習モデルのパラメータと同時に IRT のパラメータも推定する．具体的には，以下の損失関数に基づき誤差逆伝播法でパラメータ学習を行う．

$$MSE(\boldsymbol{U},\hat{\boldsymbol{U}}) = \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{n_j}\sum_{r=1}^{R}(U_{jr}-\hat{U}_{jr})^2 I_{jr}\right] \tag{4}$$

ここで，I_{jr} は $U_{jr} = -1$ のときに 0，それ以外のときに 1 を返す関数であり，変数 n_j は $n_j = \sum_{r=1}^{R} I_{jr}$ で定義される．また，学習されたモデルを用いて新たな答案 $e_{j'}$ の得点を予測する手順は 5 章で説明した従来手法と同様である．

提案手法では，IRT 得点の推定にアノテータが与えた観測得点だけでなく答案の文章も活用できるため，従来の IRT モデルと比べて IRT 得点の推定精度が改善し，本アプローチの性能が改善することが期待できる．

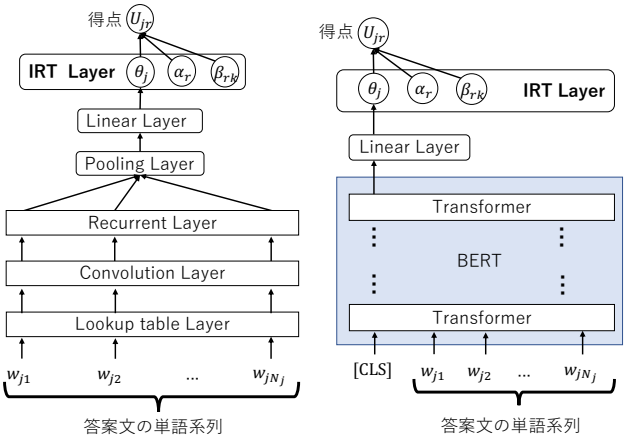

図 1 提案手法の概念図

7 実データ実験

7.1 実データ

本実験では，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）[27] を使用する．ASAP は 8 つの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 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．元データとの相関は 0.675 であった．

本研究では，異なるアノテータが採点したデータを元にモデル学習を行ったとしても安定した得点を予測できるモデルの実現を目指している．そのような評価を行うために本実験では，各答案に与えられた複数の観測得点からランダムに一つの得点データを残すことでアノテータの割り当てを変えた得点データセットを複数パターン用意する．ただし，訓練データ中の全ての答案に単一の得点のみが与えられている場合，アノテータの特性パラメータを同一尺度上で推定するために必要な等化が保証されない．そこで，訓練データ中の半分の答案は元の複数アノテータによるデータをそのまま使用し，残りの半分の答案にはランダムに選択した 1 名のアノテー

表 1　IRT 得点の推定精度評価結果

	提案手法			既存
	CNN-LSTM	LSTM	BERT	IRT
RMSE	0.219	0.215	0.188	0.487
相関	0.920	0.920	0.970	0.738

タの得点を残すようにデータを作成する．以上の手順で ASAP データセットから新たにアノテータの割り当てが異なる 10 個のデータセットを作成し，これらを $\{\boldsymbol{U}'_1, \ldots, \boldsymbol{U}'_{10}\}$ とした．

7.2　IRT 得点の推定精度評価

本節では，提案手法を用いることで，各答案に対する IRT 得点 θ の推定精度が向上するかを評価する．IRT 得点の推定精度は，アノテータが変わっても安定した得点が推定されている場合に高いと解釈できる [21, 22, 23]．そのため，アノテータの割り当て方が異なるデータセット $\boldsymbol{U}'_n$ を用いて各答案の IRT 得点 $\boldsymbol{\theta_n}$ を推定する操作を繰り返し行い，推定された値を比較することによって IRT 得点の推定精度を評価する．具体的には，次の手順に基づいて評価実験を行った．1）n 番目のデータセット $\boldsymbol{U}'_n$ の 4/5 を訓練データとして，提案手法によりモデルの学習を行った．2）手順 1 で得られたモデルを所与として，データセット $\boldsymbol{U}'_n$ における残り 1/5 のテストデータに対して IRT 得点を推定した．3）以上を学習データとテストデータの切り分けを変えて 5 回繰り返すことで，全ての受検者の IRT 得点 $\boldsymbol{\theta_n}$ を求めた．以上の操作を $n = \{1, \ldots, 10\}$ について行ったあと，n 番目のデータセットから求めた $\boldsymbol{\theta_n}$ と n' 番目の得点データセットから推定した $\boldsymbol{\theta_{n'}}$ との平均平方二乗誤差（Root Mean Square Error：RMSE），相関係数（Correlation）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．

以上の実験を複数の構成のモデルで行った．具体的には，LSTM に基づくモデルにおける Convolution layer の有無が異なるモデル（「CNN-LSTM」と「LSTM」）と BERT に基づくモデル（「BERT」）を用いた．また，比較のために，同様の実験を，答案の文章情報を利用しない式（2）の IRT モデル（「既存 IRT」）でも実施した．

実験結果を表 1 に示す．全ての条件において提案手法のモデルが高い性能を示しており，IRT 得点の推定精度の改善に有効であることが確認できた．

表 2　得点予測の頑健性評価結果

LWK	提案手法	IRT 得点を用いた手法	観測得点を用いた手法
CNN-LSTM	**0.682**	0.614	0.535
LSTM	**0.768**	0.694	0.678
BERT	**0.706**	0.688	0.694

相関	提案手法	IRT 得点を用いた手法	観測得点を用いた手法
CNN-LSTM	**0.899**	0.862	0.776
LSTM	**0.945**	0.926	0.903
BERT	**0.931**	0.923	0.920

7.3　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，評価者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評価者を変化させても，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，手順 1〜3 は前節と同様に行い，手順 3 で得られた IRT 得点 $\boldsymbol{\theta_n}$ を所与として，式（3）を用いて予測得点 $\hat{\boldsymbol{U}}_{\boldsymbol{n}}$ を求めた．この手順を，前節と同様に $n = \{1, \ldots, 10\}$ について行い，$\hat{\boldsymbol{U}}_{\boldsymbol{n}}$ と $\hat{\boldsymbol{U}}_{\boldsymbol{n'}}$ との重み付きカッパ係数（Linear Weighted Kappa：LWK），相関係数を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，5 章で説明した IRT 得点を用いてモデル学習を行う手法（「IRT 得点を用いた手法」）と観測得点の平均値を用いてモデル学習を行う手法（「観測得点を用いた手法」）を用いて同様の実験を行った．

実験結果を表 2 に示す．性能が最も高い結果を太字で示している．全ての条件において，提案手法のモデルが高い性能を示していることから，提案手法により自動採点モデルのアノテータバイアスに対する頑健性の向上が確認できた．

8　まとめ

本研究では，深層学習自動採点モデルとアノテータの特性を考慮した IRT モデルを統合し，end-to-end でモデル学習を行う手法を提案した．また，実データ実験により，IRT 得点の推定精度と自動採点モデルの得点予測精度が改善されることを示した．

今後は，様々なデータに適用し提案手法の有効性を評価する．また，より高精度の自動採点モデルに組み込むことでさらなる精度改善を目指す．

謝辞

本研究は JSPS 科研費 JP19H05663，JP20K20817，JP21H00898 の助成を受けたものです．

参考文献

[1] 河原宣央. 国語科の評価問題における記述式問題の採点過程に関する研究　採点基準と採点答案の分析を通して. Technical report, 広島県立教育センター, 2017.

[2] 野澤雄樹. 記述式項目の使用に関する教育測定学的考察. 教育心理学年報, Vol. 58, pp. 131–148, 2019.

[3] 荒井清佳, 石岡恒憲. 小論文課題の複数人による採点の基礎的な分析: 採点者による得点の違いについて. 大学入試研究ジャーナル, No. 26, pp. 53–58, 2016.

[4] 石岡恒憲. 記述式答案、AI 採点「ほぼ人間並み」12 万件で精度検証. 日本経済新聞. 日経電子版. https://www.nikkei.com/article/DGXZQOCD02D8Y0S1A201C2000000/, December 2021. (Accessed on 1/9/2022).

[5] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. In **Proceedings of the International Joint Conference on Artificial Intelligence**, pp. 6300–6308. International Joint Conferences on Artificial Intelligence Organization, July 2019.

[6] Masaki Uto. A review of deep-neural automated essay scoring models. **Behaviormetrika**, 2021.

[7] Paraskevas Lagakis and Stavros Demetriadis. Automated essay scoring: A review of the field. In **Proceedings of the International Conference on Computer, Information and Telecommunication Systems**, pp. 1–6, 2021.

[8] Dadi Ramesh and Suresh Kumar Sanampudi. An automated essay scoring systems: a systematic literature review. **Artificial Intelligence Review**, pp. 1 – 33, 2021.

[9] 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.

[10] 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.

[11] Stefanie A. Wind, Edward W. Wolfe, George Engelhard, Peter W. Foltz, and Mark Rosenstein. The influence of rater effects in training sets on the psychometric quality of automated scoring for writing assessments. **International Journal of Testing**, Vol. 18, pp. 27 – 49, 2018.

[12] Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. O2u-net: A simple noisy label detection approach for deep neural networks. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 3325–3333, 2019.

[13] Shikun Li, Shiming Ge, Yingying Hua, Chunhui Zhang, Hao Wen, Tengfei Liu, and Weiqiang Wang. Coupled-view deep classifier learning from multiple noisy annotators. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, pp. 4667–4674, 2020.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

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

[20] 宇佐美慧. 採点者側と受験者側のバイアス要因の影響を同時に評価する多値型項目反応モデル. 教育心理学研究, Vol. 58, No. 2, pp. 163–175, 2010.

[21] Masaki Uto and Masashi Okano. Robust neural automated essay scoring using item response theory. In **Artificial Intelligence in Education**, pp. 549–561. Springer International Publishing, 2020.

[22] 岡野将士, 宇都雅輝. 評価者バイアスの影響を考慮した深層学習自動採点手法. 電子情報通信学会論文誌 D, Vol. 104, No. 8, pp. 650–662, 2021.

[23] 岡野将士, 宇都雅輝. 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. 言語処理学会第 27 回年次大会, pp. 900–904. 言語処理学会, March 2021.

[24] 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.

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

[26] Masaki Uto, Yikuan Xie, and Maomi Ueno. Neural automated essay scoring incorporating handcrafted features. In **Proceedings of the International Conference on Computational Linguistics**, pp. 6077–6088. International Committee on Computational Linguistics, 2020.

[27] Automated Student Assessment Prize. https://www.kaggle.com/c/asap-aes/data. (Accessed on 1/11/2022).