様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 ２０ 日現在

機関番号：１２６１２
研究種目：挑戦的研究（萌芽）
研究期間：2020～2022
課題番号：２０Ｋ２０８１７

研究課題名（和文）深層学習と項目反応理論を融合した評価者バイアスに頑健な小論文自動採点手法

研究課題名（英文）Robust automated essay scoring method integrating deep neural networks and item response theory

研究代表者

宇都　雅輝（Uto, Masaki）

電気通信大学・大学院情報理工学研究科・准教授

研究者番号：１０７３２５７１

交付決定額（研究期間全体）：（直接経費）　　4,800,000円

研究成果の概要（和文）：近年，深層学習を用いた小論文自動採点モデルが高精度を達成している．そのような自動採点モデルの訓練に利用される大量の得点付き小論文データセットは，一般に多数の評価者が分担して小論文を採点することで構築される．しかし，そのような場合，各小論文に与えられる得点が各評価者の甘さ/厳しさなどの特性に依存してしまう．そのように評価者バイアスの影響を受けたデータをモデルの訓練に利用すると自動採点の性能が低下する．この問題を解決するために，本研究では，研究代表者が長年研究してきた「評価者バイアスの影響を考慮して真スコアを推定できる数理モデル」を活用することで，評価者バイアスに頑健な自動採点手法を開発した．

研究成果の学術的意義や社会的意義
自動採点技術の性能は年々更新されてきているものの，その性能改善は微小であり，大幅な性能改善には手法の抜本的な見直しが必要であると考えられる．本研究で指摘している既存手法の問題点は，既存のすべての自動採点手法に当てはまる根本的な問題でありながら，既存研究で見落とされてきた観点である．本研究は，この問題に対して，理論的かつシンプルな解決策を提案するものであり，自動採点の性能を大幅に改善できる可能性を有するとともに，将来的に自動採点手法の基礎フレームワークとなりうる学術的にインパクトの大きい研究であると考える．

研究成果の概要（英文）：In automated essay scoring (AES), scores are automatically assigned to essays as an alternative to grading by humans. Conventional AES models generally require training on a large dataset of graded essays. However, assigned grades in such a training dataset are known to be biased owing to effects of rater characteristics when grading is conducted by assigning a few raters in a rater set to each essay. Performance of AES models drops when such biased data are used for model training. Researchers in the fields of educational and psychological measurement have recently proposed item response theory (IRT) models that can estimate essay scores while considering effects of rater biases. This study therefore proposed a new method that trains AES models using IRT-based scores for dealing with rater bias within training data.

研究分野：教育工学

キーワード：小論文自動採点　項目反応理論　深層学習　評価者バイアス　信頼性　テスト理論

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年，論理的思考力や創造力などの高次な能力を測定する手法の一つとして小論文テストのニーズが急速に高まっている．特にわが国では，大学入試への記述式問題の導入や英語 4 技能試験の普及などの背景を受け，小論文テストのニーズは今後ますます増加すると予測できる．他方で，小論文テストには，採点コストの高さや評価者バイアスの影響による採点の信頼性低下などの問題があり，大規模試験やハイステークス試験での利用が困難であることが指摘されてきた．小論文自動採点手法は，これらの問題を解決する方法の一つとして，実用化が強く求められている技術である．

これまでに提案されてきた自動採点手法は，図 1 のように，1）答案文からの特徴量抽出と，2）その特徴量ベクトルを入力とする線形・非線形回帰による評点予測，の 2 つのフェーズで構成される．ここで，特徴量抽出の方法には大きく二つのアプローチが存在する．一つは，事前に設計した特徴量を用いる方法であり，ETS の e-rater や大学入試センターの JESS などで古くから用いられてきた．もう一つの方法は，深層学習モデルやトピックモデルなどの機械学習モデルを利用して，評点予測に有効な特徴量をデータから自動的に獲得する方法である．この手法は深層学習技術の発展とともに 2016 年以降急速に普及しているものであり，人工知能や自然言語処理，教育工学のトップカンファレンスである AAAI，ACL，EMNLP，NAACL，AIED などで毎年新たなモデルが提案され，精度が更新され続けている．

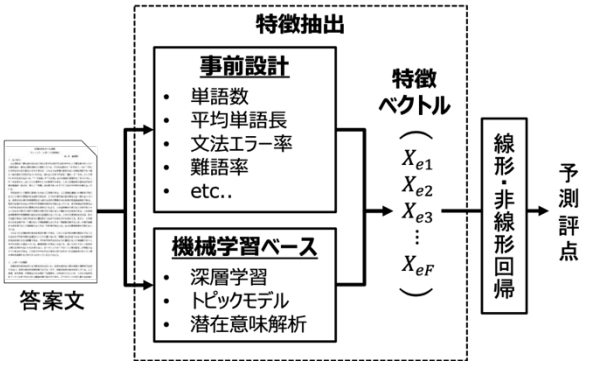

図 1. 一般的な自動採点手法の枠組み

これらの従来の自動採点モデルでは，モデルパラメータ（機械学習モデルのパラメータや特徴量ベクトルと評点の関係を表す線形・非線形回帰モデルのパラメータ）を，採点済み答案のデータセットを用いて事前に学習する必要がある．採点済み答案は数百から数千単位で必要となるため，答案の採点作業は複数の評価者で分担して行うことが一般的である．しかし，この場合，各答案に与えられる評点が，評価者の特性差（甘さ/厳しさや評価基準の差異など）によるバイアスを受けてしまうことが知られている．このようなバイアス・データを利用した場合，学習される自動採点モデルにも評価者バイアスの影響が反映されてしまい，モデルの性能が著しく低下することが近年指摘されている．従来研究ではデータセット中の評点をバイアスのない真の値とみなすことで，この問題を無視あるいは軽視してきた．しかし，小論文の採点において十分に評価者バイアスを取り除くには，長時間の評価者トレーニングや明確な採点基準の整備，採点中の作業管理などの高コストな運用が必要であり，一般にはバイアスが存在しないと仮定することは現実的ではない．この問題は，全ての自動採点モデルに共通する本質的なものであるとともに，今後の性能向上のボトルネックとなりうる重要な問題である．

一方，研究代表者は，小論文テストのように評価者による採点作業を伴うパフォーマンステストにおいて，評価者バイアスの影響を考慮して真のスコアを推定できる数理モデルを開発してきた[1-7]．この数理モデルは，情報処理技術者試験や一部の公務員試験，英検や SPI などで利用されている近代的テスト理論の一つである項目反応理論に基づくモデル（項目反応モデル）である．応募者が開発してきた項目反応モデルは，評価者のバイアスを取り除いたスコアリング手法として世界最高精度を更新し続けており，医療系大学間共用試験や英検，リクルートキャリア社，ベネッセ教育総合研究所などの様々なパフォーマンステストで実用化されている．

２．研究の目的

本研究の目的は，研究代表者が開発してきた評価者特性を考慮できる項目反応モデルを自動採点モデルに統合することで，評価者バイアスに頑健な自動採点手法を開発することである．

３．研究の方法

本研究では，評価者バイアスを考慮した項目反応モデルと，最先端の深層学習自動採点モデルを統合した新たな自動採点技術を開発する．具体的には，図 2 の上段のように，項目反応モデルによって推定されるスコアを学習する自動採点モデルを開発する．従来手法では，図 2 の下段のように評価者の評点を直接学習するため，自動採点モデルに評価者のバイアスが反映されてしまうことが問題であった．これに対し，提案手法では，観測評点から項目反応モデルによって評価者バイアスを考慮した真のスコアを推定し，そのスコアを自動採点モデルに学習させるため，

理論的には評価者バイアスの影響を受けることなく自動採点モデルを学習できる．

本研究は次のスケジュールで実施した．まず，令和 2 年度には，先端的な深層学習自動採点モデルの実装を行うとともに，学習用データセットの収集を行った．小論文自動採点の研究では，Automated Student Assessment Prize (ASAP) と呼ばれるベンチマークデータが一般に利用されるが，このデータセットでは評価者の情報が公開されていない．そこで，ASAP 内の答案文を数十名の評価者に分担して採点させることで，本研究で利用できるデータセットを収集した．令和 3 年度には，項目反応モデルと自動採点モデルを統合した提案技術を開発し，実データ実験による提案技術の有効性評価を行った．令和 4 年度には，研究成果を整理し，国際会議および論文誌への投稿を行った．

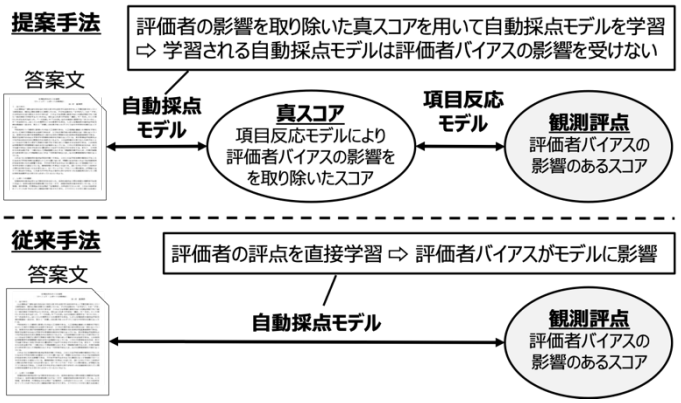

図 2. 提案手法のアイディア

４．研究成果

上述の通り，本研究では，評価者特性を考慮した項目反応モデルを深層学習自動採点モデルに組み込んだ，評価者バイアスに頑健な新たな自動採点手法を提案した．具体的には，評価者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深層学習自動採点モデルを学習する方法である．この手法は様々な深層学習自動採点モデルで利用できるが，本研究では伝統的な深層学習自動採点モデルである LSTM (Long short-term memory)に基づくモデルと，現在幅広く利用されているモデルの一つである BERT(Bidirectional Encoder Representations from Transformers) を用いたモデルへの組み込みを行った．提案手法は，これまで等閑視されてきた学習データ中の評価者バイアスの問題に対処した初めての手法である．本研究では，実データを用いた実験を行い，提案手法により，様々な自動採点モデルに対して，評価者バイアスに頑健なモデル学習と得点予測を実現できることを示した．

本研究の成果は，教育分野における人工知能活用に関する権威ある国際会議 International Conference on Artificial Intelligence in Education; AIED (2020) に採択され，Best paper runner-up award を受賞した．また，査読付き論文が，電子情報通信学会論文誌 D (2021) と IEEE Transactions on Learning Technologies (2022) に掲載された．また，国内の学会発表では，人工知能学会研究会で若手奨励賞を，日本テスト学会で大会発表賞をそれぞれ受賞した．

さらに，提案技術の発展として，複数観点で自動採点できるように提案技術を拡張した技術も開発した．その研究成果は，電子情報通信学会論文誌 D (2023) と言語処理分野の主要国際会議の一つである International Conference on Computational Linguistics; COLING (2022)に採択され，人工知能学会研究会で若手奨励賞を，電子情報通信学会教育工学研究会で研究奨励賞を，それぞれ受賞した．さらに，関連する成果は，教育システム情報学会では論文賞を受賞した．

引用文献

(1) Masaki Uto, Maomi Ueno (2016) Item Response Theory for Peer Assessment. IEEE Transactions on Learning Technologies, IEEE Computer Society, Vol.9, No.2,
(2) Masaki Uto, Maomi Ueno (2018) Empirical Comparison of Item Response Theory Models with Rater's Parameters. Heliyon, Elsevier, Vol.4, No 5.
(3) Masaki Uto, Duc-Thien Nguyen, Maomi Ueno (2019) Group optimization to maximize peer assessment accuracy using item response theory and integer programming, IEEE Transactions on Learning Technologies.
(4) Masaki Uto (2019) Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. International Conference on Artificial Intelligence in Education (AIED), pp. 494-506
(5) Masaki Uto, Maomi Ueno (2020) A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. Behaviormetrika, Springer, Vol. 47, Issue. 2, pp. 469-496.
(6) Masaki Uto (2021) A multidimensional generalized many-facet Rasch model for rubric-based performance assessment. Behaviormetrika, Springer, Vol.48, Issue 2, pp.425-457.
(7) Masaki Uto (2022) A Bayesian Many-Facet Rasch Model with Markov Modeling for Rater Severity Drift. Behavior Research Methods, Springer.

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 柴田 拓海、宇都 雅輝	４．巻 J106-D
２．論文標題 多次元項目反応理論と深層学習を用いた複数観点同時自動採点手法	５．発行年 2023年
３．雑誌名 電子情報通信学会論文誌D 情報・システム	６．最初と最後の頁 47～56
掲載論文のDOI（デジタルオブジェクト識別子） 10.14923/transinfj.2022JDP7007	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Uto Masaki、Okano Masashi	４．巻 14
２．論文標題 Learning Automated Essay Scoring Models Using Item-Response-Theory-Based Scores to Decrease Effects of Rater Biases	５．発行年 2021年
３．雑誌名 IEEE Transactions on Learning Technologies	６．最初と最後の頁 763～776
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TLT.2022.3145352	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Uto Masaki	４．巻 48
２．論文標題 A review of deep-neural automated essay scoring models	５．発行年 2021年
３．雑誌名 Behaviormetrika	６．最初と最後の頁 459～484
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s41237-021-00142-y	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 青見 樹、堤 瑛美子、宇都 雅輝、植野 真臣	４．巻 J104-D
２．論文標題 項目反応理論による小論文自動採点機のモデル平均	５．発行年 2021年
３．雑誌名 電子情報通信学会論文誌D 情報・システム	６．最初と最後の頁 784～795
掲載論文のDOI（デジタルオブジェクト識別子） 10.14923/transinfj.2021JDP7002	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 岡野 将士、宇都 雅輝	4．巻 J104-D
2．論文標題 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	5．発行年 2021年
3．雑誌名 電子情報通信学会論文誌D 情報・システム	6．最初と最後の頁 650～662
掲載論文のDOI（デジタルオブジェクト識別子） 10.14923/transinfj.2021JDP7010	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 内田 優斗、宇都 雅輝	4．巻 38
2．論文標題 受験者の能力を考慮した深層学習ベース短答記述式問題自動採点手法	5．発行年 2021年
3．雑誌名 教育システム情報学会誌	6．最初と最後の頁 218～228
掲載論文のDOI（デジタルオブジェクト識別子） 10.14926/jsise.38.218	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計14件（うち招待講演　2件／うち国際学会　4件）

1．発表者名 Itsuki Aomi, Emiko Tsutsumi, Masaki Uto, Maomi Ueno
2．発表標題 Integration of Automated Essay Scoring Models using Item Response Theory
3．学会等名 International Conference on Artificial Intelligence in Education（国際学会）
4．発表年 2021年

1．発表者名 柴田拓海，宇都雅輝
2．発表標題 多次元項目反応理論と深層学習に基づく複数観点同時自動採点手法
3．学会等名 言語処理学会第28回年次大会
4．発表年 2022年

１．発表者名
岡野将士，宇都雅輝

２．発表標題
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

３．学会等名
言語処理学会第28回年次大会

４．発表年
2022年

１．発表者名
柴田拓海，宇都雅輝

２．発表標題
深層学習と多次元項目反応理論を用いた複数観点同時自動採点手法の開発

３．学会等名
電子情報通信学会 教育工学研究会

４．発表年
2021年

１．発表者名
柴田拓海，宇都雅輝

２．発表標題
多次元項目反応理論と深層学習を用いた複数観点同時自動採点手法

３．学会等名
人工知能学会 第93回 先進的学習科学と工学研究会

４．発表年
2021年

１．発表者名
岡野将士，宇都雅輝

２．発表標題
深層学習自動採点技術を組み込んだ一般化多相ラッシュモデル

３．学会等名
日本テスト学会 第19回大会

４．発表年
2021年

１．発表者名
青見樹，堤瑛美子，宇都雅輝，植野真臣

２．発表標題
項目反応理論を用いた自動採点モデルの統合手法

３．学会等名
第35回人工知能学会全国大会

４．発表年
2021年

１．発表者名
Masaki Uto, Yikuan Xie, Maomi Ueno

２．発表標題
Neural Automated Essay Scoring Incorporating Handcrafted Features

３．学会等名
International Conference on Computational Linguistics（国際学会）

４．発表年
2020年

１．発表者名
Masaki Uto, Masashi Okano

２．発表標題
Robust neural automated essay scoring using item response theory

３．学会等名
International Conference on Artificial Intelligence in Education（国際学会）

４．発表年
2020年

１．発表者名
Masaki Uto, Yuto Uchida

２．発表標題
Automated short-answer grading using deep neural networks and item response theory

３．学会等名
International Conference on Artificial Intelligence in Education（国際学会）

４．発表年
2020年

１．発表者名
岡野将士・宇都雅輝

２．発表標題
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

３．学会等名
言語処理学会第27回年次大会

４．発表年
2021年

１．発表者名
岡野将士・宇都雅輝

２．発表標題
項目反応理論と深層学習を用いた評価者バイアスに頑健な小論文自動採点手法

３．学会等名
行動計量学会第48回大会

４．発表年
2020年

１．発表者名
Masaki Uto

２．発表標題
Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability

３．学会等名
第19回情報科学技術フォーラム FIT2020，情報処理学会（トップコンファレンスセッション）（招待講演）

４．発表年
2020年

１．発表者名
宇都雅輝

２．発表標題
パフォーマンス型試験の課題とその解決に向けた人工知能研究の現在

３．学会等名
SCATE-21研究会（招待講演）

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関