論　文

項目反応理論による小論文自動採点機のモデル平均

青見　樹[†a)]　堤　瑛美子[†b)]　宇都　雅輝[†c)]　植野　真臣[†d)]

Automated Essay Scoring Model Averaging by Item Response Theory

Itsuki AOMI[†a)], Emiko TSUTSUMI[†b)], Masaki UTO[†c)], and Maomi UENO[†d)]

あらまし　小論文自動採点は，人間評価者に代わって自動採点モデルが小論文の採点を行う自然言語処理におけるタスクの一つである．近年では多くの自動採点モデルが提案されており，それぞれに異なった特徴を有している．本研究では，評価者特性を考慮した項目反応理論を用いて自動採点モデルのモデル平均を行う新たな手法を提案する．具体的には自動採点モデルを一人の評価者とみなして評価者特性を考慮した項目反応モデルを適用することで，それぞれの自動採点モデルの特徴を考慮した統合を行う．実験を通して，提案手法が単体の自動採点モデルや，単純な予測スコアの平均化手法と比べて予測精度を向上させることを示す．さらに，提案手法が統合した自動採点モデルの特徴を捉え，安定したスコアの予測を行うことができることを示す．

キーワード　項目反応理論，パフォーマンス評価，自動採点，モデル平均

1. まえがき

近年，膨大な小論文の採点コストを削減するために小論文自動採点に関する研究が注目されている．小論文自動採点とは，人間評価者に代わって自動採点モデルが小論文の採点を行うタスクであり，主に自然言語処理や教育工学の分野で研究が行われている．従来の自動採点モデルは，特徴量ベースモデルと深層学習ベースモデルの主に二つに大別できる [1], [2].

特徴量ベースモデルは，小論文の文書から単語数や誤字の数といった特徴量を抽出し，主に回帰によって小論文のスコアを予測するモデルである．代表的なモデルとしては， TOEFL (Test of English as a Foreign Language) や GRE (Graduate Record Examination) で導入されている e-rater [3] が挙げられる．このモデルの他にも，多様な特徴量ベースモデルが提案されている [4]〜[7]．特徴量ベースモデルでは，特徴量の重要度などを解析することができ，モデルの解釈性が高いという利点を有している．しかし，高い解釈性と高い予測精度を得るためには，教育の専門家による背景知識や経験によって適切な特徴量を選択する必要がある．

他方で，深層学習手法を用いて単語の系列を直接入力として，スコアの予測を行うモデルが提案されている [8], [9]．Taghipour and Ng が提案した LSTM (long short-term Memory) をベースとしたモデル [9] をはじめとして，多くの深層学習手法を用いたモデルの研究がなされている [10]〜[14]．深層学習ベースモデルは，人手では設計が難しい潜在的な特徴量を学習することが出来るため，特徴量ベースモデルでは学習が難しい小論文の採点を行うことが期待される．

自動採点モデルは性質の多様化が進み，それぞれのモデルは異なる利点を有している．本研究の主なアイデアは，多様な自動採点モデルが予測したスコアを平均化することで，スコアの予測精度の向上を目指すというものである．しかし，自動採点モデルの特性が多様であるがゆえに，単純にスコアを平均化するだけでは精度の向上が妨げられる恐れがある．

この問題に対する解決策として本研究では，項目反応理論 (Item response theory: IRT) [15] を利用する．IRT は数理モデルを用いたテスト理論である．IRT の拡張モデルとして，評価の一貫性や厳しさといった人間評価者の特性を考慮してスコアを推定できるモデルが多数提案されており [16]〜[20]，高精度なスコアの

† 電気通信大学大学院情報理工学研究科，調布市
Graduate School of Informatics and Engineering, The University of Electro-Communications, 1–5–1 Chofugaoka, Chofu-shi, 182–8585 Japan
a) E-mail: aomi@ai.lab.uec.ac.jp
b) E-mail: tsutsumi@ai.lab.uec.ac.jp
c) E-mail: uto@ai.lab.uec.ac.jp
d) E-mail: ueno@ai.is.uec.ac.jp
DOI:10.14923/transj.???????????

推定が実現されている [21], [22]．IRT の詳細は 3 章で述べる．本研究では，自動採点モデルを人間評価者とみなすことで IRT モデルを適用し，小論文のスコアの予測精度の向上を図る．提案手法は，各自動採点モデルの特性を考慮しつつ各モデルの予測スコアを統合することができるため，単一の自動採点モデルや単純なスコアの平均化手法と比べてより正確な予測スコアを得ることが期待できる．本論文では，提案手法のスコアの予測精度が，単一の自動採点モデルと単純なスコアの平均と比べて向上することを実データによる実験を通して示す．

2. 小論文自動採点モデル

本節では，これまでに提案された自動採点モデルを，特徴量ベースモデルと深層学習ベースモデルの二つに大別して紹介する．

2.1 特徴量ベースモデル

特徴量ベースモデルは，専門家などが選択したいくつかの特徴量を用いて，小論文のスコアを予測するモデルである．代表的なモデルとしては TOEFL 等で採用されている e-rater [3] が挙げられる．このモデルは，主に文法の誤用，平均単語長，文長，語彙の困難度といった特徴量を用いて重回帰によってスコアの予測を行う．さらに近年では，多彩な特徴量ベースモデルの提案がされている．Phandi et al. は，ベイジアンリッジ回帰を用いてある課題で学習したモデルを別の課題でスコアを予測する手法を提案した [4]．このモデルは， Domain adaptation と呼ばれる元領域で学習した知識を目標領域で適応するタスクにおいて，一般的に用いられる EasyAdapt [23] を応用して学習を行う．また，Beigman klebanov et al. は，単語の話題性に着目し，これらを特徴量として応用した自動採点モデルを提案した [5]．一方，Nguyen and Litman は，自然言語処理のタスクの一つである論証マイニング [24] の知見を自動採点モデルに導入した [6] ．具体的には，論証マイニングで一般的に用いられる要素分類 (Classifying Argument Components) や 関係分類 (Identifying Argumentative Relation) に関する特徴量を用いてスコアを予測する．さらに，Cozma et al. は，HISK (histogram intersection string kernel) と呼ばれる文字列カーネル [25] と BOSWE (bug-of-super-word-embedding) [26] を組み合わせた特徴量を用いて予測を行うモデルを提案している [7] .

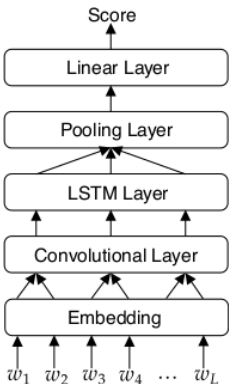


図 1: Taghipour and Ng の LSTM ベースモデル
Fig. 1 LSTM-based model by Taghipour and Ng.

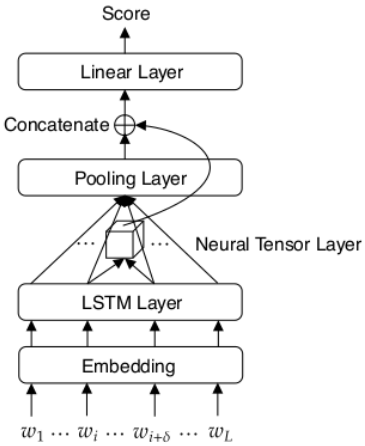


図 2: Tay et al. の SkipFlow モデル
Fig. 2 SkipFlow by Tay et al.

2.2 深層学習ベースモデル

深層学習ベースモデルは，深層学習手法を用いて単語の系列を直接入力として，小論文のスコアを予測するモデルである．Taghipour and Ng により提案された LSTM を用いたモデル (図 1) [8] が，精度の指標である二次の重み付きカッパ係数 (quadratic weighted Kappa: QWK) において従来の特徴量ベースモデルを上回る精度が報告されて以降，数多くのモデルが提案されてきた．例えば，Alikaniotis et al. は，スペルミスなどの情報を用いて各単語が小論文のスコアにどのように影響を与えるかを word-embedding の学習に反映させ，LSTM ベースのモデルで拡張させた [9] ．また，Tay et al. は，Taghipour and Ng のモデルに SKIPFLOW と呼ばれる離れた単語間の特徴を考慮して学習を行う

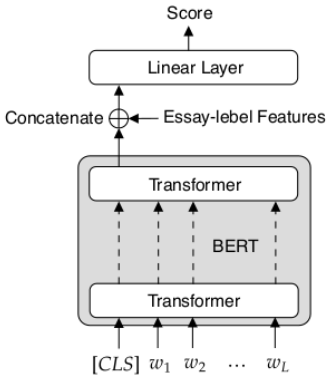


図 3: Uto et al. の BERT ベースハイブリッドモデル
Fig. 3　BERT-based hybrid model by Uto et al.

機構を追加し，長文の小論文に対して離れた単語間の意味関係を考慮できるモデルを提案した (図 2) [12]. Wang et al. は， REINFORCE アルゴリズム [27] による深層強化学習の枠組みを自動採点モデルの学習に導入し， 回帰ベースの予測だけでなく分類ベースの予測の可能性を提示した [13] . Yue et al. は，多様な課題に適応するために，半教師あり学習のフレームワークを提案し，学習した課題とは別の課題で予測を行う際の QWK を向上させた [14] .

さらに， LSTM の代替として Transformer [28] の機構を用いたモデルが提案されている．例えば， Mayfield and Black は，事前学習された BERT (Bidirectional Encoder Representation from Transformers) [29] を fine-tune する自動採点モデルを提案した [30] .

2.3　ハイブリッドモデル

特徴量ベースモデルと深層学習ベースモデルを組み合わせたハイブリッドモデルの研究も行われている．例えば，Dasgupta et al. は一般的な LSTM ベースのモデルの出力と，人手で設計した特徴量を入力とするモデルの出力を結合したモデルを提案している [31]．また， Uto et al. は従来の深層学習ベースモデルの出力に，事前に作製した特徴量ベクトルを結合して学習を行うというフレームワークを提案している [32]．具体的には， LSTM や BERT を始めとした様々な深層学習手法をベースに複数の特徴量ベクトルを結合したモデルを提案している (図 3).

2.4　自動採点モデルの統合

このように自動採点モデルは性質の多様化が進み，モデルごとに異なる特徴と利点を有している．つまり，これらの自動採点モデルが予測したスコアを平均化することで，スコアの予測精度を向上させることが期待できる．しかし，自動採点モデルの特性が多様であるがゆえに，単純にスコアを平均化するだけでは特定のモデルの影響を受けるため，精度の向上が妨げられる恐れがある．本研究では，各自動採点モデルの特徴を考慮して予測スコアの統合を行うために，受験者の能力を適切に測定できる IRT を用いることを提案する．

3.　項目反応理論

IRT [15] は，e ラーニングや e テスティングの基盤技術として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．IRT では，観測されたテストにおける受験者の反応から，テスト項目の特性と受験者の能力を同時に推定する．これらのモデルを利用する利点として，以下が挙げられる．

（1） テスト項目の特性を考慮しつつ，受験者の能力が推定できる．

（2） 異なるテスト項目に対する受験者の反応を，同一尺度で評価できる．

（3） 欠損値が含まれている場合でも，容易に推定できる．

従来の IRT モデルでは，課題における受験者のスコアで構成される受験者 × 課題の二相データにおける定式化がなされてきた．しかし，本論文で扱うような複数の評価者が受験者の小論文を採点する小論文試験におけるデータは，一般には受験者 × 課題 × 評価者の三相データである．このようなデータに対応するために，近年では評価者特性を考慮したモデルが多数提案されている [16]〜[20].

評価者特性を考慮した最も一般的なモデルとして，多相ラッシュモデル (MFRM: many-facet Rasch model) [16] が知られている．X_{ijr} を評価者 $r \in \mathcal{R} = \{1,\ldots,R\}$ が受験者 $j \in \mathcal{J} = \{1,\ldots,J\}$ に課題 $i \in \mathcal{I} = \{1,\ldots,I\}$ の小論文に与えるカテゴリカルスコア $k \in \mathcal{K} = \{1,\ldots,K\}$ とする．MFRM では，$X_{ijr} = k$ となる確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} \left[\theta_j - \beta_i - \beta_r - d_m\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\theta_j - \beta_i - \beta_r - d_m\right]}. \quad (1)$$

ここで，θ_j は受験者 j の潜在的な能力，β_i は課題 i の困難度，β_r は評価者 r の厳しさ，d_m はスコア $k-1$ から k に遷移する困難度を表すパラメータである．モデルの識別性のために，$\beta_1 = 0, d_1 = 0, \sum_{k=2}^{K} d_k = 0$ を

仮定する．

MFRM では，全ての課題について識別力が一定であることと，全ての評価者が同等の一貫性を持つことが仮定されるが，現実ではこれらの仮定が成り立つことは少ない．そこで，これらの制約を緩和したモデルとして課題識別力の差異と評価者一貫性の差異を考慮できるモデルが提案されている [19], [20], [33]．本研究では，その中で IRT モデルである Uto and Ueno が提案した generalized MFRM (g-MFRM) [20] を導入する．このモデルでは，$X_{ijr} = k$ となる確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} \left[\alpha_i \alpha_r (\theta_j - \beta_i - \beta_r - d_{rm})\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\alpha_i \alpha_r (\theta_j - \beta_i - \beta_r - d_{rm})\right]}. \tag{2}$$

ここで，α_i は課題 i の識別力，α_r は評価者 r の一貫性，d_{rk} は評価者 r のスコア k に対する厳しさを表すステップパラメータである．モデルの識別性のために，$\sum_{i=1}^{I} \log \alpha_i = 0, \sum_{i=1}^{I} \beta_i = 0, d_{r1} = 0, \sum_{k=2}^{K} d_{rk} = 0$ を仮定する．

小論文自動採点における研究では，それぞれの小論文の課題についてモデルの学習を行うことが一般的である．これに倣うと， IRT モデルでは課題数 $I = 1$ として学習を行うため，モデルの識別性の仮定より α_i と β_i を無視できる．このとき，式 (1) は，

$$P_{jrk} = \frac{\exp \sum_{m=1}^{k} \left[\theta_j - \beta_r - d_m\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\theta_j - \beta_r - d_m\right]}, \tag{3}$$

となり，また，式 (2) は，

$$P_{jrk} = \frac{\exp \sum_{m=1}^{k} \left[\alpha_r (\theta_j - \beta_r - d_{rm})\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\alpha_r (\theta_j - \beta_r - d_{rm})\right]}, \tag{4}$$

となる．一般に， IRT モデルのパラメータはデータに含まれる観測されたスコアを用いて，EM (expectation-maximization) アルゴリズムや， MCMC (Marcov chain Monte Carlo) 法によって推定される．

一般に最尤法で推定される IRT モデルにおける能力推定の予測誤差は，フィッシャー情報量の逆数に漸近的に一致することが知られている [15]．そのため，IRT では, 能力測定精度を表す指標としてフィッシャー情報量が一般に利用される．式 (3), (4) で示される MFRM や g-MFRM のフィッシャー情報量 $\mathcal{I}(\theta_j)$ は次式で定義される．

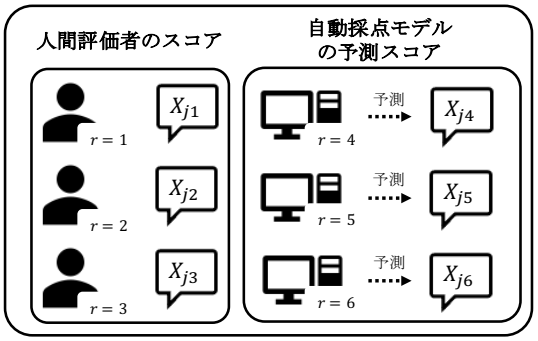


図 4: 提案手法への入力データ (3 人の人間評価者と 3 つの自動採点モデルの場合)

Fig. 4 The Input of the proposed method when three human raters ($r = 1, 2, 3$) and three essay scoring models ($r = 4, 5, 6$) exists.

$$\mathcal{I}(\theta_j) = \sum_{r=1}^{R} \left[\sum_{k=1}^{K} k^2 P_{jrk} - \left(\sum_{k=1}^{K} k P_{jrk} \right)^2 \right]. \tag{5}$$

これらのモデルは，単にスコアを合計したり平均値を行う採点モデルと比べてより高精度に受験者の能力を推定できることが知られている．本研究では，自動採点モデルを人間評価者とみなすことで IRT モデルを適用する．それぞれの自動採点モデルが予測したスコアを用いて IRT モデルのパラメータを推定することで，自動採点モデルの評価特性を考慮したスコアを予測することができる．なお， IRT モデルを自動採点モデルに組み込むことを提案しているものもあるが，本研究のように複数の自動採点モデルを統合するための手法ではない [21], [22]．次節では，提案手法の詳細について述べる．

4. 提 案 手 法

本節では，本研究で提案する複数の自動採点モデルを統合する手法について述べる．本研究では，自動採点モデルを IRT モデルにおける評価者の一人としてみなすことで， IRT モデルのパラメータを推定し，小論文のスコアの予測に用いる．具体的には，各小論文における人間評価者のスコアと学習済みの自動採点モデルの予測スコアを用いて MFRM，g-MFRM のパラメータを推定する．例えば，図 4 のように 3 人の人間評価者のスコア X_{j1}, X_{j2}, X_{j3} と 3 つの自動採点モデルの予測スコア X_{j4}, X_{j5}, X_{j6} がある場合には，$\boldsymbol{X} = \{X_{j1}, X_{j2}, \cdots, X_{j6}\}, (j = 1, \cdots, J)$ を MFRM，g-

MFRM への入力データとして，パラメータを推定する．新たな小論文のスコア予測は次の手順で行う．まず，個別の自動採点モデルで対象小論文の予測スコアを求める．次に，得られた予測スコアを用い，評価者パラメータの推定値を所与として MFRM，g-MFRM に基づく受験者の潜在的な能力 $\hat{\theta}_j$ を推定する．最後に，評価者パラメータと能力値を所与として期待スコア $\hat{X}_j$ を次式で求め，提案手法の予測スコアとする．

$$\hat{X}_j = \frac{1}{|\mathcal{R}_{\text{human}}|} \sum_{r \in \mathcal{R}_{\text{human}}} \sum_{k=1}^{K} k \cdot P_{jrk} \tag{6}$$

ここで，$\mathcal{R}_{\text{human}}$ は人間評価者の集合を示す．期待スコアの算出は，受験者の潜在的な能力 $\hat{\theta}_j$ を元の人間評価者のスコアの尺度に合わせるために行う．

5. 評価実験

5.1 データセット

本研究では，評価実験に用いるデータセットとして ASAP (Automated Student Assessment Prize) データセット[1]を用いる．このデータセットは，過去に Kaggle のプラットフォームによって開催されたデータコンペティションで用いられ，現在では数多くの小論文自動採点の研究に用いられている [4], [7], [8], [12]〜[14], [32], [34]．表 1 に示すように，ASAP データセットは八つの異なる課題で構成されている．それぞれの課題について，英語を母語とする米国の学生が記述した小論文と，小論文に対する人間評価者のスコアが付与されており，各課題ごとに受験者は異なる．表 1 に，各課題における小論文数（受験者数）と平均単語数，スコアのレンジを示した．ここで本研究では，一般的な小論文自動採点の研究に従い，自動採点モデルを課題ごとに学習して評価を行った．また，ASAP データセットでは，各小論文につき人間評価者によって付与された一つの基準となるスコアが対応付けられているため，提案手法における人間評価者数について $|\mathcal{R}_{human}| = 1$ とした．

表 1: ASAP データセットの基礎統計
Table 1 Statistics of the ASAP dataset.

課題番号	小論文数 (受験者数)	平均単語数	スコアレンジ
1	1,783	350	2-12
2	1,800	350	1-6
3	1,726	150	0-3
4	1,772	150	0-3
5	1,805	150	0-4
6	1,800	150	0-4
7	1,569	250	0-30
8	723	650	0-60

5.2 実験設定

本実験では，5 分割交差検証によって小論文のスコアの予測精度で評価を行った．また，それぞれの分割の割合について，先行研究 [8] と同様に，データセットの 60% を トレーニングデータ， 20% を検証データ，20% をテストデータとした．評価指標は，自動採点モデルの研究において広く採用され， ASAP コンペティションでの標準的な指標として利用された，2 次重み付きカッパ係数 (QWK) を用いた．QWK の定義式と説明を付録に示す．また，一般的な精度指標として平均平方二乗誤差 (RMSE) も用いた．

次に，スコアの統合を行う自動採点モデルを以下に示す．

• EASE (SVR), EASE (BLRR). Phandi et al. [4] で用いられた EASE (Enhanced AI Scoring Engine)[2] は，ASAP コンペティションで入賞した特徴量抽出ツールである．EASE では次のような特徴量を用いる．

- 文字数や単語数といった長さに関する特徴量
- POS (Part of speech) タグに関連する特徴量
- 課題ごとの特徴を表す特徴量
- Bag of words による特徴量

本研究では， SVR (support vector regression) と BLRR (Bayesian linear ridge regeression) の二つの回帰モデルを用いた．また，先行研究 [4] と同様に scikit-learn [35] を用いて実装を行った．

• XGBoost. 本研究では，EASE に含まれない特徴量として，先行研究 [34], [36] で用いられた構文木をベースとする特徴量を用いたモデルを採用した．

構文木をベースとする特徴量としては次のような特徴量を用いた．

- 小論文に含まれる節の数に関する特徴量
- 節に含まれる単語数に関する特徴量
- 構文木の深さに関する特徴量

構文木の構成には，CoreNLP [37] を用いた．また，先行研究 [36] と同様に回帰モデルとして XGBoost [38] を用いた．

• LSTMMoT. 深層学習ベースモデルとして， LSTM

（注1）：https://www.kaggle.com/c/asap-aes/

（注2）：https://github.com/edx/ease/

表 2: 実験で利用した自動採点モデルとその平均化手法の一覧
Table 2 The base AES models and AVG methods

	モデルの略称	モデルの特徴
BASE	EASE (SVR), EASE (BLRR)	文字数や単語数，POS タグ，課題ごとの特徴，Bag of words などの特徴量を用いる特徴量ベースモデル
	XGBoost	構文木をベースに求められる，小論文に含まれる節の数，節に含まれる単語数，構文木の深さに関する特徴量などを用いる特徴量ベースモデル
	LSTMMoT	LSTM を利用した代表的な深層学習ベースモデル
	SkipFlow	LSTMMoT をもとに，離れた単語間の意味関係を捉えやすいように拡張された深層学習ベースモデル
	BERT+F	BERT を利用した深層学習ベースモデルに，人手で設計した特徴量ベクトルを統合したハイブリッドモデル
AVG	MEAN	BASE モデルが予測したスコアの平均値を算出する手法
	VOTING	BASE モデルが予測したスコアから多数決 (hard-voting) でスコアを決定する手法
	STACKING	BASE モデルが予測したスコアを説明変数，人間評価者の得点を目的変数とする線形回帰モデルを用いてスコアを予測する手法
	Proposal (MFRM), Proposal(g-MFRM)	BASE モデルを IRT モデルにおける評価者の一人とみなしてスコアを予測する提案手法

ベースのモデルとして最も一般的なモデルである Taghipour and Ng のモデル [8] を採用した．なお，図 1 に示した convolution layer はオプションの層であり，本実験では用いない．また，本研究ではこのモデルの実装に PyTorch[3] を用いた．

• SkipFlow. 本研究ではさらに深層学習ベースモデルとして， LSTM ベースのモデルに SKIPFLOW と呼ばれる機構を導入した SkipFlow モデル [12] を採用した．このモデルは，図 2 に示す LSTM layer の出力のペア $(h_i, h_{i+\delta})$ を Neural Tensor Layer [39] への入力として用いる．本実験ではこの幅 δ を 20 とした．また，モデルの実装には PyTorch を用いた．

• BERT+F. 本研究では，ハイブリッドモデルとして Uto et al. [32] で提案された事前学習済みの BERT に特徴量を加えて fine-tune するモデルを採用した．本研究では，事前学習済みの BERT として， uncased BERT-base を使用し，実装には PyTorch を用いた．

小論文の字句解析には NLTK tokenizer[4] を用いた．また，他の詳細なハイパーパラメータの設定は元の研究の設定に準じた値を使用した．

さらに，単純なモデル平均手法 (以下，AVG 法) と提案手法を以下の手順で用いる．

• MEAN. BASE モデルの予測したスコアを算術平均する．

• VOTING. BASE モデルの予測したスコアから多数決 (hard-voting) でスコアを決定する．

• STACKING. BASE モデルの予測したスコアを説明変数，人間評価者の得点を目的変数とする線形回帰モデルを用いてスコアを予測する．個別の BASE モデルはトレーニングデータを用いて学習し，STACKING のための線形回帰モデルは検証データを利用して学習した．

• 提案手法 (MFRM，g-MFRM). 人間評価者のスコアと BASE モデルの予測したスコアから MFRM，g-MFRM のパラメータを推定する．個別の BASE モデルはトレーニングデータを用いて学習し，MFRM，g-MFRM は検証データを利用して学習した．推定した評価者パラメータを所与として，テストデータにおける BASE モデルの予測スコアから受験者の能力 $\hat{\theta}_j$ を推定し，予測スコア $\hat{X}_j$ を求める．なお，ここではテストデータ中の小論文に対する BASE モデルの予測スコアのみを利用しており，人間評価者が与えたスコアデータは使用していないことに注意してほしい．

以降，提案手法をそれぞれ Proposal (MFRM), Proposal (g-MFRM) と呼ぶ．先行研究 [20] に従い， IRT モデルのパラメータの推定には Stan [40] を利用した No-U-Turn sampler [41] による ハミルトニアンモンテカルロ法を用いた．パラメータの事前分布や MCMC 法の詳細な設定も先行研究 [20] に従った．本実験で使用するモデルの一覧を表 2 にまとめた．これらの BASE モデル，AVG 法，提案手法を用いてスコアの予測精度を比較する．

5.3 実 験 結 果

表 3 に，各 BASE モデルと各 AVG 法の QWK と

(注3)：https://pytorch.org/
(注4)：http://www.nltk.org/

表 3: 各 BASE モデル， AVG 法の QWK と RMSE
Table 3　QWK score of the BASE models and the AVG methods.

			課題番号								
		自動採点モデル	1	2	3	4	5	6	7	8	平均値
QWK	BASE	EASE (SVR)	0.558	0.533	0.564	0.571	0.659	0.749	0.545	0.350	0.566
		EASE (BLRR)	0.804	0.603	0.656	0.717	0.784	0.761	0.730	0.675	0.716
		XGBoost	0.814	0.640	0.593	0.660	0.763	0.657	0.692	0.676	0.687
		LSTMMoT	0.777	0.619	0.651	0.730	0.770	0.760	0.750	0.460	0.690
		SkipFlow	0.798	0.652	0.657	0.729	0.783	0.778	0.751	0.614	0.720
		BERT+F	0.827	0.637	0.672	0.620	0.780	0.673	0.720	0.681	0.701
	AVG	MEAN	0.820	0.667	0.673	0.730	**0.805**	0.774	0.768	0.678	0.739
		VOTING	0.833	0.660	**0.675**	0.731	0.794	0.770	0.745	0.666	0.734
		STACKING	0.831	0.664	0.649	0.739	0.786	0.784	0.770	0.700	0.740
		Proposal (MFRM)	0.821	0.626	0.663	0.685	0.777	0.728	0.768	0.674	0.718
		Proposal (g-MFRM)	**0.838**	**0.686**	0.668	**0.743**	0.796	**0.785**	**0.793**	**0.717**	**0.753**
RMSE	BASE	EASE(SVR)	1.874	0.941	0.748	0.893	0.729	0.801	5.286	9.790	2.633
		EASE(BLRR)	0.892	0.628	0.620	0.634	0.598	0.620	3.019	4.077	1.386
		XGBoost	0.897	0.626	0.699	0.741	0.649	0.752	3.311	4.283	1.495
		LSTMMoT	0.952	0.636	0.636	0.637	0.638	0.617	2.968	4.845	1.491
		SkipFlow	0.954	0.621	0.662	0.683	0.646	0.631	3.031	4.686	1.489
		Bert+F	0.849	0.614	0.619	0.735	0.610	0.718	3.084	4.101	1.416
	AVG	MEAN	0.903	0.605	0.627	0.639	**0.589**	0.615	2.876	4.263	1.390
		VOTING	0.840	0.595	**0.615**	0.627	0.595	0.615	2.986	4.459	1.417
		STACKING	0.838	0.587	0.630	**0.622**	0.601	**0.600**	2.803	**3.994**	1.334
		Proposal (MFRM)	0.837	0.593	0.616	0.642	0.597	0.635	2.783	4.096	1.350
		Proposal (g-MFRM)	**0.824**	**0.576**	0.617	0.623	0.594	0.603	**2.739**	3.999	**1.322**

表 4: AVG 法での比較
Table 4　Comparison between AVG methods.

		MEAN	VOTING	STACKING	Proposal (MFRM)	Proposal (g-MFRM)
QWK	平均	0.739	0.734	0.740	0.718	**0.753**
	p 値	0.039	0.039	0.036	0.007	-
RMSE	平均	1.390	1.417	1.334	1.350	**1.322**
	p 値	0.076	0.159	0.153	0.040	-

RMSE を示した．提案手法である Proposal (g-MFRM) は課題番号 3 の BERT+F を除いて，他の全ての BASE モデルの QWK を上回った．さらに，平均 QWK では全ての比較手法に対して高い値となった．表 3 から，単純な平均化手法である MEAN，VOTING，STACKING では，ほぼ全ての課題番号において BASE モデルと比べて精度が向上した．単純な平均化手法と Proposal (g-MFRM) を比較すると，課題番号 3，5 を除いて Proposal (g-MFRM) の QWK が単純な平均化手法と比べて高くなった．QWK が向上した理由として，提案手法ではそれぞれの BASE モデルの特性を考慮しつつスコアを推定できることが挙げられる．Proposal (g-MFRM) の精度が高い課題では，BASE モデル間の QWK の差が大きい傾向にある．例えば，課題番号 1,7 では EASE (SVR)，課題番号 6 では XGBoost と BERT+F，課題番号 8 では EASE (SVR) と LSTMMoT が他の BASE モデルと比べて QWK が低下している．このような場合に単純な平均化手法では自動採点モデルの特徴を考慮できないために QWK が低下しやすいが，Proposal (g-MFRM) ではこれを考慮できるため，高い予測精度を維持する結果となった．

さらに，AVG 法の Proposal (g-MFRM) と他の AVG 法に対して対応のある t 検定を行った．この検定を行い，検定の多重性を考慮して hommel 法により補正した p

値を表 4 に示す．この結果から， Proposal (g-MFRM) は 有意水準 5% において他の単純な平均化手法と比べて QWK の 有意な差が認められた．

一方, RMSE においては, 他の平均化手法と比べて有意差は認められなかったものの, Proposal (g-MFRM) が最も低い値を示した．また，QWK と同様に，Proposal (g-MFRM) の性能が高い課題では BASE モデル間の RMSE の差が大きい傾向があることも確認できた．

表 3 から，IRT モデル間で比較を行うと， Proposal (MFRM) は Proposal (g-MFRM) と比べて QWK が劣ることがわかった．他の単純な平均化手法と比べても Proposal (MFRM) の精度は下回っていた．この結果から， MFRM のようなシンプルな IRT モデルでは自動採点モデルの特徴を考慮できておらず，提案手法に g-MFRM を導入することの有効性が示唆された．

5.4 受験者の能力における分析

本節では， g-MFRM において推定された受験者の能力 $\hat{\theta}$ の値でデータを分け，各受験者の能力層における各自動採点モデルの性能分析を行う．

表 5 は， g-MFRM において推定された $\hat{\theta}$ について，低い能力の受験者 ($\hat{\theta} \leq -0.5$)，中程度の能力の受験者 ($-0.5 < \hat{\theta} \leq 0.5$)，高い能力の受験者 ($0.5 < \hat{\theta}$) の三つにデータを分割し，それぞれの自動採点モデルの QWK を示したものである．太字は AVG 法の間で最も QWK が大きいものを示す．表 5 より, 各 BASE モデルはデータセットや能力によって大きく QWK が異なり，それぞれ自動採点モデルには特徴があることがわかる．例えば，低い能力の受験者においては，EASE (SVR), XGBoost, LSTMMoT, SkipFlow が他の BASE モデルと比べて平均 QWK が高く，高い能力の受験者においては，EASE (BLRR), BERT+F が他の BASE モデルと比べて 平均 QWK が高くなった．これらを統合する提案手法の Proposal (g-MFRM) は，表 4 に示したように単一の自動採点モデルと他の AVG 法と比べて QWK が向上した．さらに表 5 で各 $\hat{\theta}$ の範囲ごとにみると，特に低い能力の受験者において， Proposal (g-MFRM) は課題番号 3 を除いて他の単純な平均化手法の QWK を上回る結果となった．BASE モデルのそれぞれの特徴を考慮できるため，他の単純な平均化手法と比べて安定して精度が向上している．

ここで，図 5 は，課題番号 1,2 における g-MFRM のフィッシャー情報量 $\mathcal{I}(\theta_j)$ を示したグラフである．表 5 より，課題番号 1 は低い能力の受験者について Proposal (g-MFRM) の精度が大きく向上している．このとき図 5a を見ると，低い能力の受験者の θ の範囲でフィッシャー情報量も相対的に大きな値を示した．また，課題番号 2 においても，中程度の能力の受験者について Proposal (g-MFRM) の精度が向上し，中程度の能力の受験者の θ_j の範囲ではフィッシャー情報量の値が大きくなったことが図 5b からわかる．フィッシャー情報量が大きいときに g-MFRM の θ_j の推定値の標準誤差が小さくなるため，受験者の能力を正確に捉えている範囲では，提案手法である Propsal (g-MFRM) の精度向上に寄与していることがわかる．

さらに，図 6 は，課題番号 3 における g-MFRM のフィッシャー情報量 $\mathcal{I}(\theta_j)$ を示したグラフである．表 5 より， Proposal (g-MFRM) は他の課題とは対照的に低い能力の受験者の QWK が著しく劣化していた．このとき，低い能力の受験者のフィッシャー情報量の値は，相対的に値が小さい．フィッシャー情報量が大きいときには Proposal (g-MFRM) の精度向上に寄与していたが，逆にフィッシャー情報量が小さい場合には精度の劣化がみられた．なお，他の課題においても上記と同様に解釈可能であった．

以上の結果から，実際の小論文の採点のような実用的なシチュエーションにおいて，Proposal (g-MFRM) では，フィッシャー情報量を確認することで自動採点モデルの評価ができることも期待される．

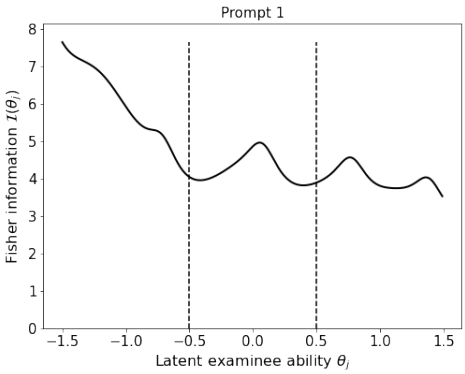
6. むすび

本研究では， IRT を用いた新しい自動採点モデルの平均化の手法を提案した．まず，それぞれの自動採点モデルは受験者の能力に応じて予測されるスコアが異なるため，単純に平均化されたスコアでは予測精度が低下してしまう恐れがあることを述べた．この問題を解決するために，本研究では自動採点モデルの特性を考慮することができる IRT モデルを用いてスコアを予測するアイデアを提示した．それぞれの自動採点モデルを一人の評価者とみなすことで，IRT モデルに適用した．実データを用いた実験の結果，提案手法は単一の自動採点モデルと比べてスコアの予測精度が向上した．さらに，複数の自動採点モデルの予測スコアを単純に平均化する手法と比べても，予測精度が向上し，有意な差が認められた．また, IRT モデルにおけるフィッシャー情報量が大きい際に，予測精度が向上していることを示し，自動採点モデルの評価の一つの指標としての可能性を提示した．今後の研究では，様々なデータセットを用いて提案手法の性能を評価する必

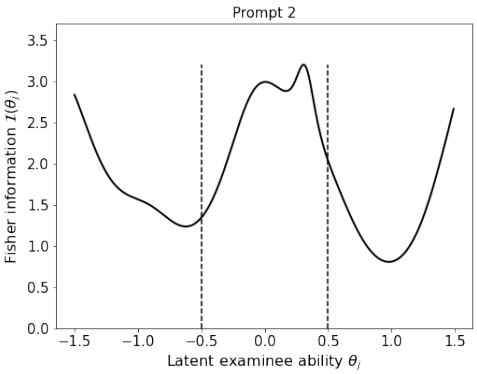
表 5: 各 $\hat{\theta}$ の範囲における自動採点モデルの QWK
Table 5 QWK score of essay scoring models in each range of $\hat{\theta}$.

		課題番号								
	自動採点モデル	1	2	3	4	5	6	7	8	平均値
低い能力の受験者 ($\hat{\theta} \leq -0.5$)										
BASE	EASE (SVM)	0.540	0.487	0.138	0.049	0.382	0.460	0.341	0.326	0.340
	EASE (BLRR)	0.533	0.314	0.125	0.144	0.439	0.438	0.295	0.346	0.329
	XGBoost	0.770	0.528	0.060	0.051	0.317	0.403	0.408	0.586	0.390
	LSTMMoT	0.745	0.570	0.039	0.331	0.395	0.540	0.452	0.116	0.399
	SkipFlow	0.682	0.497	0.048	0.259	0.341	0.574	0.455	0.327	0.398
	BERT+F	0.661	0.358	0.056	0.080	0.359	0.354	0.322	0.355	0.318
AVG	MEAN	0.752	0.521	**0.075**	0.153	0.421	0.451	0.462	0.511	0.418
	VOTING	0.748	0.531	0.000	0.235	0.416	0.486	0.479	0.516	0.426
	STACKING	0.755	0.491	0.037	0.256	0.403	0.542	0.448	0.406	0.417
	Proposal (g-MFRM)	**0.792**	**0.549**	−0.002	**0.292**	**0.425**	**0.551**	**0.522**	**0.524**	**0.457**
中程度の能力の受験者 ($-0.5 < \hat{\theta} \leq 0.5$)										
		課題番号								
	自動採点モデル	1	2	3	4	5	6	7	8	平均値
BASE	EASE (SVM)	0.109	0.047	0.234	0.188	0.234	0.161	0.196	0.108	0.160
	EASE (BLRR)	0.362	0.120	0.059	0.314	0.309	0.334	0.335	0.366	0.275
	XGBoost	0.307	0.069	0.107	0.135	0.290	0.096	0.169	0.321	0.187
	LSTMMoT	0.306	0.086	0.179	0.276	0.174	0.277	0.354	0.305	0.245
	SkipFlow	0.276	0.116	0.058	0.202	0.231	0.245	0.297	0.277	0.213
	BERT+F	0.331	0.232	0.137	0.132	0.338	0.110	0.238	0.366	0.236
AVG	MEAN	0.310	0.172	0.040	0.343	**0.384**	0.265	0.326	0.335	0.272
	VOTING	0.365	0.136	0.081	0.329	0.345	**0.301**	0.297	0.340	0.274
	STACKING	**0.366**	0.171	**0.092**	**0.400**	0.323	0.284	**0.383**	**0.407**	**0.303**
	Proposal (g-MFRM)	0.341	**0.248**	0.071	0.351	0.367	0.177	0.339	0.396	0.286
高い能力の受験者 ($0.5 < \hat{\theta}$)										
		課題番号								
	自動採点モデル	1	2	3	4	5	6	7	8	平均値
BASE	EASE (SVM)	0.129	0.161	0.046	0.191	0.186	0.039	0.117	−0.140	0.091
	EASE (BLRR)	0.425	0.279	0.245	0.399	0.395	0.318	0.382	0.319	0.345
	XGBoost	0.374	0.258	0.087	0.282	0.304	0.130	0.303	0.247	0.248
	LSTMMoT	0.272	0.235	0.208	0.329	0.323	0.256	0.278	0.282	0.273
	SkipFlow	0.253	0.228	0.002	0.304	0.377	0.168	0.191	0.004	0.191
	BERT+F	0.424	0.269	0.256	0.242	0.376	0.168	0.392	0.285	0.302
AVG	MEAN	0.353	**0.280**	0.213	**0.434**	**0.441**	0.309	0.358	0.095	0.310
	VOTING	**0.418**	0.215	**0.290**	0.378	0.395	0.325	0.351	0.197	0.321
	STACKING	0.392	0.209	0.266	0.371	0.385	0.320	**0.360**	0.323	0.328
	Proposal (g-MFRM)	0.407	0.202	0.262	0.358	0.403	**0.344**	0.335	**0.357**	**0.334**

要がある．特性の異なる課題においても，提案手法が有効であることを示したい．また，様々な自動採点モデルを追加することで精度向上が期待できるため，より特徴的な自動採点モデルを組み込むことを検討する．さらに，近年では深層学習手法を用いた IRT の研究も盛んであり，より高い精度で受験者の能力を推定できることが知られている [42], [43]．このようなモデルを導入し，提案手法の精度向上に努めたい．



(a) 課題番号 1 のフィッシャー情報量 $\mathcal{I}(\theta_j)$



(b) 課題番号 2 のフィッシャー情報量 $\mathcal{I}(\theta_j)$

図 5: Proposal (g-MFRM) の QWK が向上するときのフィッシャー情報量 $\mathcal{I}(\theta_j)$
Fig. 5 The Fisher information $\mathcal{I}(\theta_j)$ when Proposal (g-MFRM) performance is good.

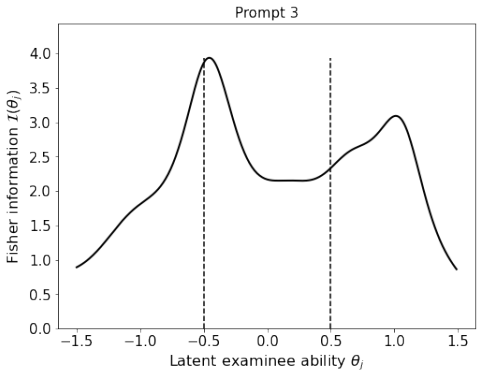


図 6: 課題番号 3 のフィッシャー情報量 $\mathcal{I}(\theta_j)$
Fig. 6 The test information $\mathcal{I}(\theta_j)$ in prompt 3.

謝辞 本研究は JSPS 科研費 19H05663, 19K21751 の助成を受けたものです.

文　　献

[1] Z. Ke and V. Ng, “Automated essay scoring: A survey of the state of the art,” Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pp.6300–6308, July 2019.

[2] M.A. Hussein, H.A. Hassan, and M. Nassef, “Automated language essay scoring systems: a literature review,” PeerJ Computer Science, vol.5, p.e208, 2019.

[3] Y. Attali and J. Burstein, “Automated essay scoring with e-rater® v.2,” The Journal of Technology, Learning and Assessment, vol.4, no.3, 2006.

[4] P. Phandi, K.M.A. Chai, and H.T. Ng, “Flexible domain adaptation for automated essay scoring using correlated linear regression,” Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp.431–439, 2015.

[5] B. Beigman Klebanov, M. Flor, and B. Gyawali, “Topicality-based indices for essay scoring,” Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp.63–72, 2016.

[6] H. Nguyen and D. Litman, “Argument mining for improving the automated scoring of persuasive essays,” Thirty-Second AAAI Conference on Artificial Intelligence, pp.5892–5899, 2018.

[7] M. Cozma, A. Butnaru, and R.T. Ionescu, “Automated essay scoring with string kernels and word embeddings,” Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp.503–509, 2018.

[8] K. Taghipour and H.T. Ng, “A neural approach to automated essay scoring,” Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp.1882–1891, 2016.

[9] D. Alikaniotis, H. Yannakoudakis, and M. Rei, “Automatic text scoring using neural networks,” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.715–725, 2016.

[10] T. Dasgupta, A. Naskar, L. Dey, and R. Saha, “Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring,” Proceedings of the Fifth Workshop on Natural Language Processing Techniques for Educational Applications, pp.93–102, 2018.

[11] Y. Farag, H. Yannakoudakis, and T. Briscoe, “Neural automated essay scoring and coherence modeling for adversarially crafted input,” Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp.263–271, 2018.

[12] Y. Tay, M. Phan, A.T. Luu, and S.C. Hui, “SkipFlow: Incorporating neural coherence features for end-to-end automatic text scoring,” Thirty-Second AAAI Conference on Artificial Intelligence, pp.5948–5955, 2018.

[13] Y. Wang, Z. Wei, Y. Zhou, and X. Huang, “Automatic essay scoring incorporating rating schema via reinforcement learning,” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp.791–797, 2018.

[14] Y. Cao, H. Jin, X. Wan, and Z. Yu, "Domain-adaptive neural automated essay scoring," Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.1011–1020, 2020.

[15] F.M. Lord, Applications of item response theory to practical testing problems, Routledge, Abingdon-on-Thames, 1980.

[16] J.M. Linacre, Many-facet Rasch measurement, MESA Press, Chicago, 1989.

[17] C.M. Myford and E.W. Wolfe, "Detecting and measuring rater effects using many-facet rasch measurement: part I," Journal of Applied Measurement, vol.4, no.4, pp.386–422, 2003.

[18] T. Eckes, Introduction to Many-Facet Rasch Measurement, Peter Lang, Bern, 2015.

[19] M. Uto and M. Ueno, "Item response theory without restriction of equal interval scale for rater's score," Artificial Intelligence in Education, pp.363–368, 2018.

[20] M. Uto and M. Ueno, "A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo," Behaviormetrika, vol.47, pp.469–496, 2020.

[21] M. Uto, "Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability," Artificial Intelligence in Education, pp.494–506, 2019.

[22] M. Uto and M. Okano, "Robust neural automated essay scoring using item response theory," Artificial Intelligence in Education, pp.549–561, 2020.

[23] H. Daumé III, "Frustratingly easy domain adaptation," Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp.256–263, Association for Computational Linguistics, Prague, Czech Republic, June 2007.

[24] A. Peldszus and M. Stede, "From argument diagrams to argumentation mining in texts: A survey," International Journal of Cognitive Informatics and Natural Intelligence, vol.7, no.1, pp.1–31, 2013.

[25] R.T. Ionescu, M. Popescu, and A. Cahill, "Can characters reveal your native language? a language-independent approach to native language identification," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.1363–1373, Association for Computational Linguistics, Doha, Qatar, Oct. 2014.

[26] A. Butnaru and R.T. Ionescu, "From image to text classification: A novel approach based on clustering word embeddings," Proceedings of the 21st International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES), pp.1784–1793, 2017.

[27] R.J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," Machine Learning, vol.8, no.3, pp.229–256, 1992.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," Proceedings of the 31st International Conference on Neural Information Processing Systems, pp.6000–6010, 2017.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp.4171–4186, 2019.

[30] E. Mayfield and A.W. Black, "Should you fine-tune BERT for automated essay scoring?," Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp.151–162, 2020.

[31] T. Dasgupta, A. Naskar, L. Dey, and R. Saha, "Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring," Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pp.93–102, 2018.

[32] M. Uto, Y. Xie, and M. Ueno, "Neural automated essay scoring incorporating handcrafted features," Proceedings of the 28th International Conference on Computational Linguistics, pp.6077–6088, International Committee on Computational Linguistics, Dec. 2020.

[33] M. Uto and M. Ueno, "Item response theory for peer assessment," IEEE Transactions on Learning Technologies, vol.9, no.2, pp.157–170, 2016.

[34] C. Jin, B. He, K. Hui, and L. Sun, "TDNN: A two-stage deep neural network for prompt-independent automated essay scoring," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.1088–1097, 2018.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and ÉdouardDuchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol.12, no.85, pp.2825–2830, 2011. http://jmlr.org/papers/v12/pedregosa11a.html

[36] J. Liu, Y. Xu, and Y. Zhu, "Automated Essay Scoring based on Two-Stage Learning," arXiv e-prints, vol.arXiv:1901.07744, Jan. 2019.

[37] C.D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S.J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," Association for Computational Linguistics (ACL) System Demonstrations, pp.55–60, 2014. http://www.aclweb.org/anthology/P/P14/P14-5010

[38] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.785–794, 2016.

[39] R. Socher, D. Chen, C.D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," Advances in Neural Information Processing Systems 26, pp.926–934 , 2013. http://papers.nips.cc/paper/5028-reasoning-with-neural-tensor-networks-for-knowledge-base-completion.pdf

[40] B. Carpenter, A. Gelman, M.D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," Journal of Statistical Software, vol.76, no.1, 2017. https://doi.org/10.18637

[41] M.D. Hoffman and A. Gelman, "The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo," Journal of Machine Learning Research, vol.15, no.47, pp.1593–1623, 2014. http://jmlr.org/papers/v15/hoffman14a.html

[42] C.K. Yeung, “Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory,” Proceeding of the 12th International Conference on Educational Data Mining (EDM), pp.683–686, 2019.
[43] 木下涼，植野真臣，“深層学習によるテスト理論: Item Deep Response Theory，” 電子情報通信学会論文誌 D，vol.J103-D, no.4，pp.314–329，2020.

付　　録

ここでは，本研究で評価指標として利用した 2 次重み付きカッパ係数 (QWK) について説明する．QWK は，2 つの離散値ベクトル間の一致度を表すカッパ係数の一つであり，データが順序尺度の場合に利用される．

ここで，サイズ N の 2 つの離散値ベクトルを $\mathbf{X} = \{X_1, ..., X_N\}$，$\mathbf{Y} = \{Y_1, ..., Y_N\}$ とする．$X_n \in \mathbf{X}$，$Y_n \in \mathbf{Y}$ は各ベクトルの n 番目のデータを表し，個々のデータは 1 から K の K 段階から値をとるものとする．このとき，2 つのベクトル $\mathbf{X}$ と $\mathbf{Y}$ の QWK は次式で定義される．

$$1 - \frac{\sum_{x=1}^{K}\sum_{y=1}^{K}(x-y)^2 P^o_{xy}}{\sum_{x=1}^{K}\sum_{y=1}^{K}(x-y)^2 P^e_{xy}} \tag{A·1}$$

ここで，P^o_{xy} は実際の一致確率，P^e_{xy} は偶然の一致確率と呼ばれ，$I(X_n = x, Y_n = y)$ を $X_n = x$ かつ $Y_n = y$ のときに 1，それ以外のとき 0 を返す関数，$n_{xy} = \sum_{n=1}^{N} I(X_n = x, Y_n = y)$ とするとき，それぞれ以下で計算される．

$$P^o_{xy} = \frac{n_{xy}}{N} \tag{A·2}$$

$$P^e_{xy} = \frac{\sum_{z=1}^{K} n_{xz}}{N}\frac{\sum_{z=1}^{K} n_{zy}}{N} \tag{A·3}$$

（xxxx 年 xx 月 xx 日受付）

青見　　樹

2019 年電気通信大学情報理工学部卒．同年，同大学大学院情報理工学研究科情報・ネットワーク工学専攻博士前期課程入学，現在に至る．ベイジアンネットワーク，e テスティング，自然言語処理などの研究に従事．

堤　瑛美子

2020 年電気通信大学大学院情報理工学研究科情報・ネットワーク工学専攻博士前期課程修了．同年，同大学大学院情報理工学研究科博士後期課程入学，現在に至る．アダプティブラーニング，e テスティング，e ラーニングなどの研究に従事．

宇都　雅輝　（正員）

2013 年電気通信大学大学院情報システム学研究科博士後期課程了．博士（工学）．長岡技術科学大学特任助教を経て，2015 年に電気通信大学助教に着任，現在に至る．e テスティング，e ラーニング，人工知能，ベイズ統計，自然言語処理などの研究に従事．

植野　真臣　（正員）

1992 年神戸大学大学院教育学研究科修了，1994 年東京工業大学大学院総合理工学研究科修了．博士（工学）．東京工業大学，千葉大学，長岡技術科学大学を経て 2006 年より電気通信大学助教授，2013 年より教授，現在に至る．人工知能，e テスティング，e ラーニング，ベイズ統計，ベイジアンネットワークなどの研究に従事．

Abstract Automated Essay Scoring (AES) is the task of automatic grading essays instead of using human raters. Many AES models offering different benefits have been proposed in the past few decades. This study proposes a new AES model averaging framework using item response theory. The proposed framework can improve scoring accuracy because it averages prediction scores from various AES models while considering characteristics of each model for evaluation of examinee ability.

Key words item response theory, performance assessment, automated essay scoring, model averaging