令和二年度 修士論文

項目反応理論による
小論文自動採点機のモデル平均

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 情報数理工学プログラム

学籍番号 1931002

青見 樹

主任指導教員 植野 真臣 教授

指導教員 川野 秀一 准教授

目次

表目次

図目次

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 (Ke and Ng 2019, Hussein et al. 2019).

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 (Attali and Burstein 2006) 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 (Phandi et al. 2015, Beigman Klebanov et al. 2016, Nguyen and Litman 2018, Cozma et al. 2018)．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 (Taghipour and Ng 2016, Alikaniotis et al. 2016)．Alikaniotis et al. (2016) が提案した LSTM (long short-term Memory) をベースとしたモデルをはじめとして，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 (Dasgupta et al. 2018, Farag et al. 2018, Tay et al. 2018, Wang et al. 2018, Cao et al. 2020)．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) (Lord 1980) 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 (Linacre 1989, Myford and Wolfe 2003, Eckes 2015, Uto and Ueno 2018, Uto and Ueno 2020)，高精度なスコアの推定が実現されている (Uto 2019, Uto and Okano 2020)．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 (Attali and Burstein 2006) 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. (2015) 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．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 (Daumé III 2007) を応用して学習を行う．また，Beigman Klebanov et al. (2016) 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．一方，Nguyen and Litman (2018) は，自然言語処理のタスクの一つである論証マイニング (Peldszus and Stede 2013) の知見を自動採点モデルに導入した. 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. (2018) は，HISK (histogram intersection string kernel) と呼ばれる文字列カーネル (Ionescu et al. 2014) と BOSWE (bug-of-super-word-embedding) (Butnaru and Ionescu 2017) 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．

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．

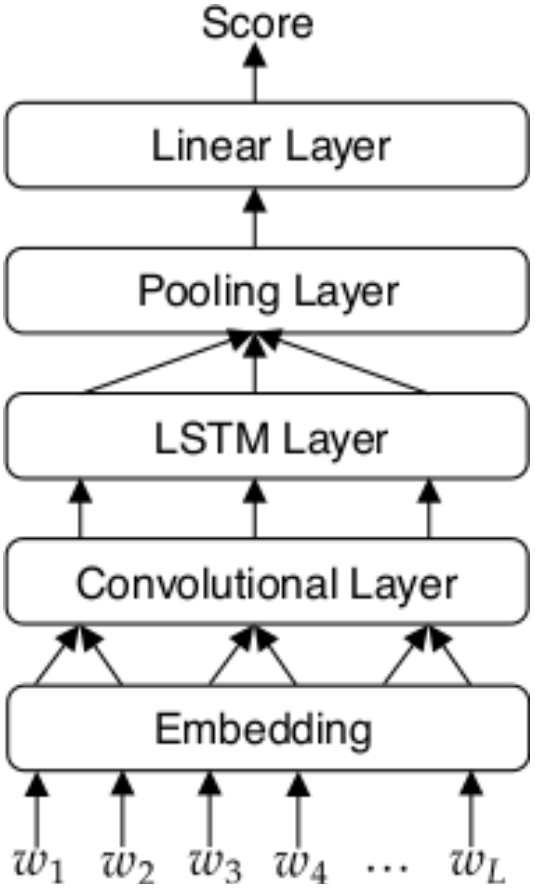

図 1: Taghipour and Ng (2016) の LSTM ベースモデル

Taghipour and Ng (2016) により提案された LSTM を用いたモデル (図 1) 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. (2016) 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．

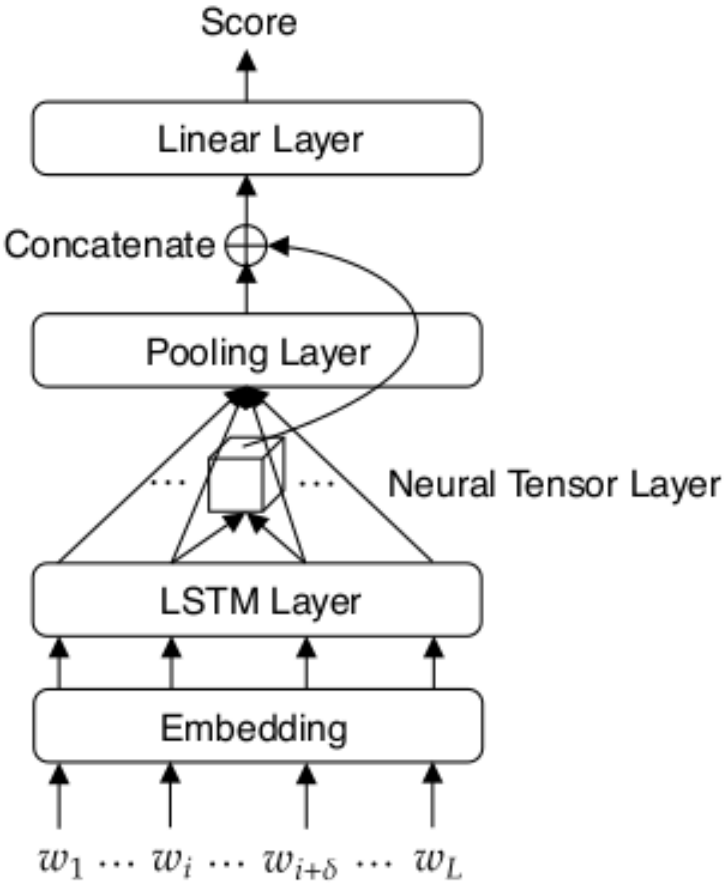

図 2: Tay et al. (2018) の SkipFlow モデル

また，Tay et al. (2018) は，Taghipour and Ng (2016) 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)．Wang et al. (2018) は，REINFORCE アルゴリズム (Williams 1992) 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．Cao et al. (2020) 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.

さらに，LSTM の代替として Transformer (Vaswani et al. 2017) の機構を用いたモデルが提案されている．例えば，Mayfield and Black (2020) は，事前学習された BERT (Bidirectional Encoder Representation from Transformers) (Devlin et al. 2019) を fine-tune する自動採点モデルを提案した.

2.3 ハイブリッドモデル

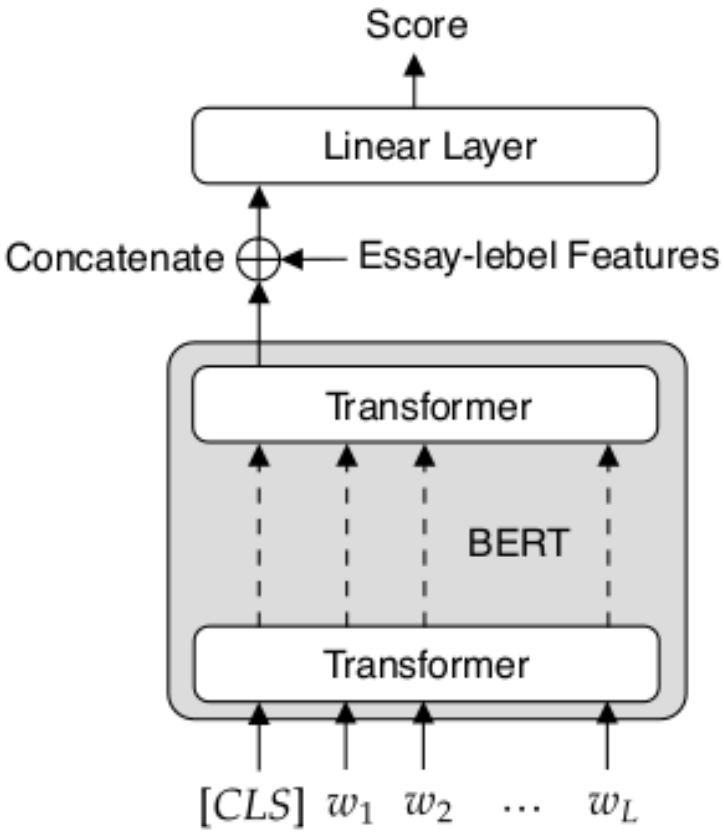

図 3: Uto et al. (2020) の BERT ベースハイブリッドモデル

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. (2018) 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．また，Uto et al. (2020) 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．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 (Lord 1980) 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，テスト項目と受験者の能力を潜在変数モデルとして定式化する．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 (Linacre 1989, Eckes 2015, Myford and Wolfe 2003, Uto and Ueno 2018, Uto and Ueno 2020).

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) (Linacre 1989) が知られている．X_{ijr} を評価者 $r \in \mathcal{R} = \{1, \dots, R\}$ が受験者 $j \in \mathcal{J} = \{1, \dots, J\}$ に課題 $i \in \mathcal{I} = \{1, \dots, I\}$ の小論文に与えるカテゴリカルスコア $k \in \mathcal{K} = \{1, \dots, K\}$ とする．MFRM では，$X_{ijr} = k$ となる確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} \left[\theta_j - \beta_i - \beta_r - d_m\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\theta_j - \beta_i - \beta_r - d_m\right]}. \tag{3.1}$$

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 (Uto and Ueno 2016, Uto and Ueno 2018, Uto and Ueno 2020)．本研究では，その中で最先端の IRT モデルである Uto and Ueno が提案した generalized MFRM (g-MFRM) (Uto and Ueno 2020) 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．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} \left[\alpha_i \alpha_r (\theta_j - \beta_i - \beta_r - d_m) \right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\alpha_i \alpha_r (\theta_j - \beta_i - \beta_r - d_m) \right]}. \tag{3.2}$$

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 (3.1) は，

$$P_{jrk} = \frac{\exp \sum_{m=1}^{k} \left[\theta_j - \beta_r - d_m \right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\theta_j - \beta_r - d_m \right]}, \tag{3.3}$$

となり，また，式 (3.2) は，

$$P_{jrk} = \frac{\exp \sum_{m=1}^{k} \left[\alpha_r(\theta_j - \beta_r - d_m)\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\alpha_r(\theta_j - \beta_r - d_m)\right]}, \tag{3.4}$$

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．

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 (Lord 1980)．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 (3.3), (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．

$$\mathcal{I}(\theta_j) = \sum_{r=1}^{R} \left[\sum_{k=1}^{K} k^2 P_{jrk} - \left(\sum_{k=1}^{K} k P_{jrk} \right)^2 \right]. \tag{3.5}$$

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 (Uto 2019, Uto and Okano 2020)．次節では，提案手法の詳細について述べる．

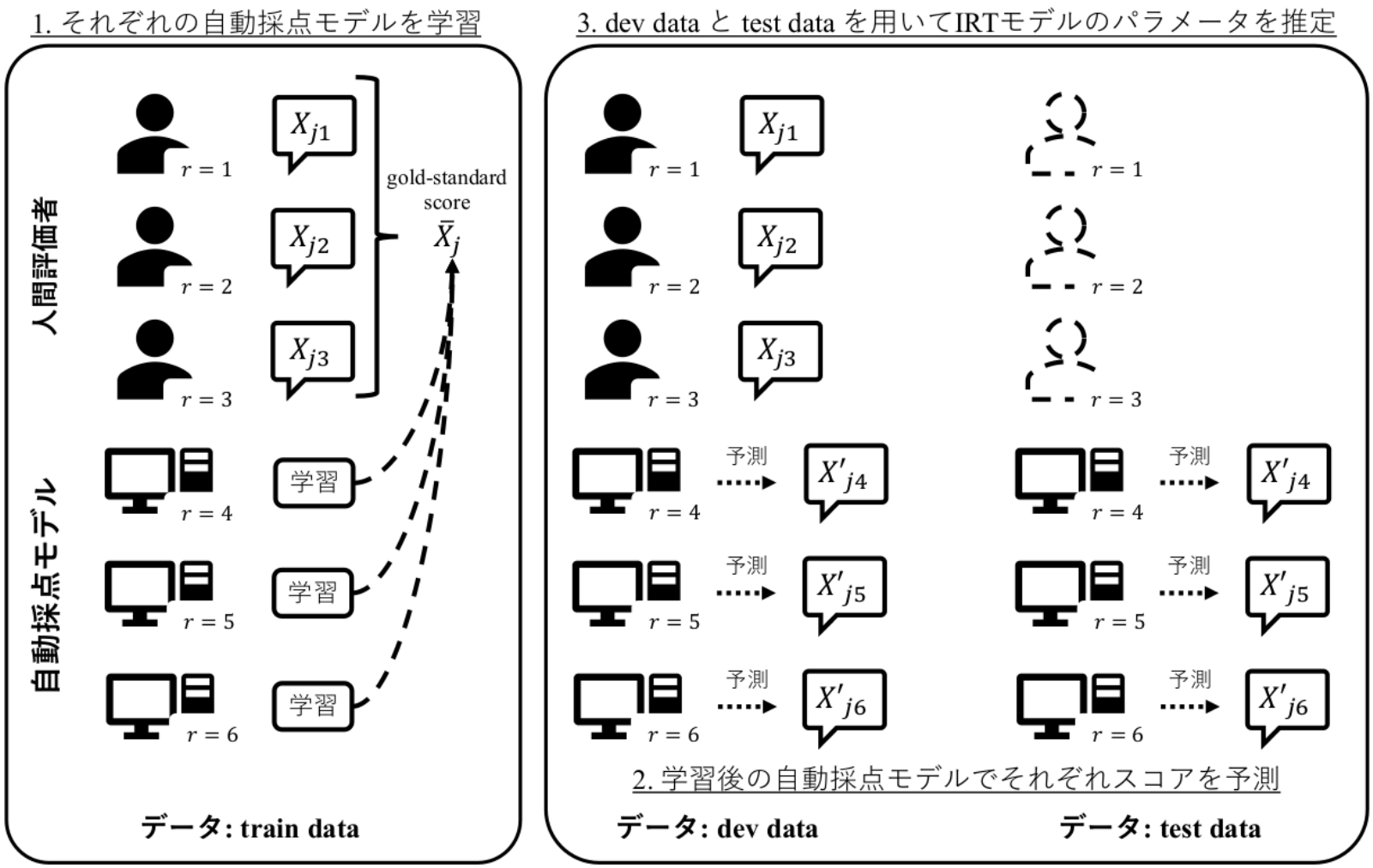

図 4: 提案手法の概略図 (3 人の人間評価者 $(r = 1, 2, 3)$ と 3 つの自動採点モデル $(r = 4, 5, 6)$ の場合)

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．提案手法の概略図を図 4 に示す．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．ここで，提案手法の学習のためにあらかじめ，学習データの一部を検証データ (dev と表記する) とする．この dev データを除いたものを train と表記する．また，ここではテストデータを test と表記する．提案手法の具体的な手順は次の通りである．

1. train データを用いて，自動採点モデルをそれぞれの方法で学習する．このとき一般的に小論文自動採点における研究では，複数の人間評価者のスコアを平均化したものや，合計したものを教師信号として用いる．
2. dev データ，test データについて，1 で学習した自動採点モデルでそれぞれスコアを予測する．
3. dev データ中の人間評価者のスコアと自動採点モデルの予測スコア，test データにおける自動採点モデルの予測スコアを用いて，IRT モデルのパラメータを推定する．
4. 3 で推定された受験者の潜在的な能力 $\hat{\theta}_j$ を含むパラメータを用いて，test データの小論文の期待スコア $\hat{X}_j$ を次のように計算する．

$$\hat{X}_j = \frac{1}{|\mathcal{R}_{\text{human}}|} \sum_{r \in \mathcal{R}_{\text{human}}} \sum_{k=1}^{K} k \cdot P_{jrk} \tag{4.1}$$

ここで，$\mathcal{R}_{\text{human}}$ は人間評価者の集合を示す．この手順は，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．

提案手法では，各自動採点モデルの評価特性を考慮しながら，様々な自動採点モデルの予測スコアを統合できる．これにより，単純なスコアの平均化手法や単一の自動採点モデルと比べ，精度の高い予測スコアを得ることが期待できる．

5　評価実験

5.1　データセット

表 1: ASAP データセットの基礎統計

課題番号	小論文数	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

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 (Phandi et al. 2015, Cozma et al. 2018, Taghipour and Ng 2016, Jin et al. 2018, Tay et al. 2018, Wang et al. 2018, Cao et al. 2020, Uto et al. 2020)．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

[*1] https://www.kaggle.com/c/asap-aes/

題ごとに受験者は異なる.

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.

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 (Taghipour and Ng 2016) と同様に，データセットの 60% を train データ，20% を dev データ，20% を test 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 QWK を用いた.

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. (2015) 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

[*2] https://github.com/edx/ease/

研究 (Phandi et al. 2015) と同様に scikit-learn (Pedregosa et al. 2011) 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 (Jin et al. 2018, Liu et al. 2019) 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 (Manning et al. 2014) を用いた. また, 先行研究 (Liu et al. 2019) と同様に回帰モデルとして XGBoost (Chen and Guestrin 2016) を用いた.
- LSTMMoT. 深層学習ベースモデルとして, LSTM 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 (2016) のモデルを採用した. 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 (Tay et al. 2018) 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 (Socher et al. 2013) 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. (2020) 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,

[*3] https://pytorch.org/

uncased BERT-base を使用し，実装には PyTorch を用いた．

本研究では，小論文の字句解析に 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．

本研究では上に示した自動採点モデルを統合した提案手法を，それぞれの自動採点モデル単体 (以下，BASE モデル) と，次の単純なモデル平均手法 (以下，AVG 法) と比較する．

- 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．

さらに，提案手法では MFRM と g-MFRM の二つの IRT モデルを用いる．以降，提案手法にこれらの IRT モデルを用いた手法を Proposal (MFRM), Proposal (g-MFRM) と呼ぶ．先行研究 (Uto and Ueno 2020) に従い，IRT モデルのパラメータの推定には Stan (Carpenter et al. 2017) を利用した No-U-Turn sampler (Hoffman and Gelman 2014) 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 (Uto and Ueno 2020) に従った．

[*4] http://www.nltk.org/

表 2: 各 BASE モデル，AVG 法の QWK

		課題番号								
	自動採点モデル	1	2	3	4	5	6	7	8	平均値
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
	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**

5.3 実験結果

表 3: AVG 法での比較

	MEAN	VOTING	Proposal (MFRM)	Proposal (g-MFRM)
平均 QWK	0.739	0.734	0.718	**0.753**
p 値	0.039	0.039	0.007	-

表 2 に，各 BASE モデルと各 AVG 法の QWK 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 2 から，単純な平均化手法である MEAN と VOTING も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．精度が改善した理由として，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 モデル間の精度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 モデルと比べて精度が低下している．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精度が低下するが，Proposal (g-MFMR) 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 3 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．

表 2 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．

表 4: 各 $\hat{\theta}$ の範囲における自動採点モデルの QWK

低い能力の受験者 ($\hat{\theta} \leq -0.5$)										
		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)	0.540	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
	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
	Proposal (g-MFMR)	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
	Proposal (g-MFRM)	0.407	0.202	0.262	0.358	0.403	**0.344**	0.335	**0.357**	**0.334**

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 4 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 4 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 3 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 4 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．

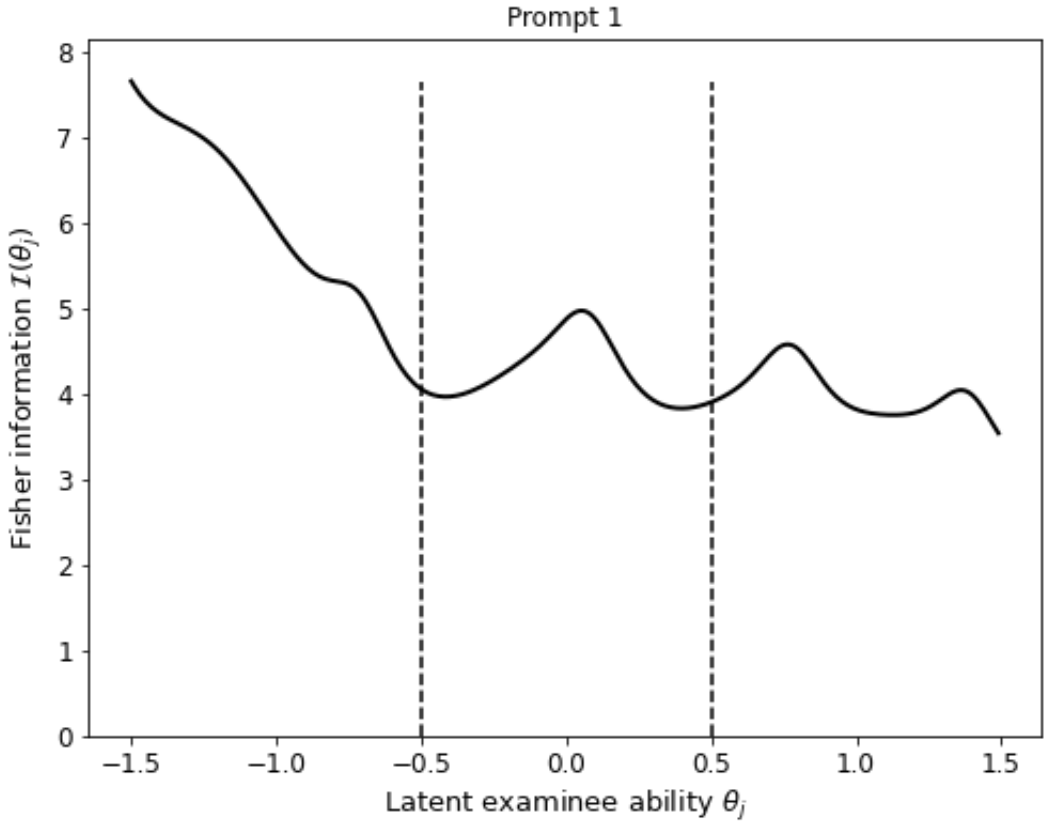

図 5: 課題番号 1 のフィッシャー情報量 $\mathcal{I}(\theta_j)$

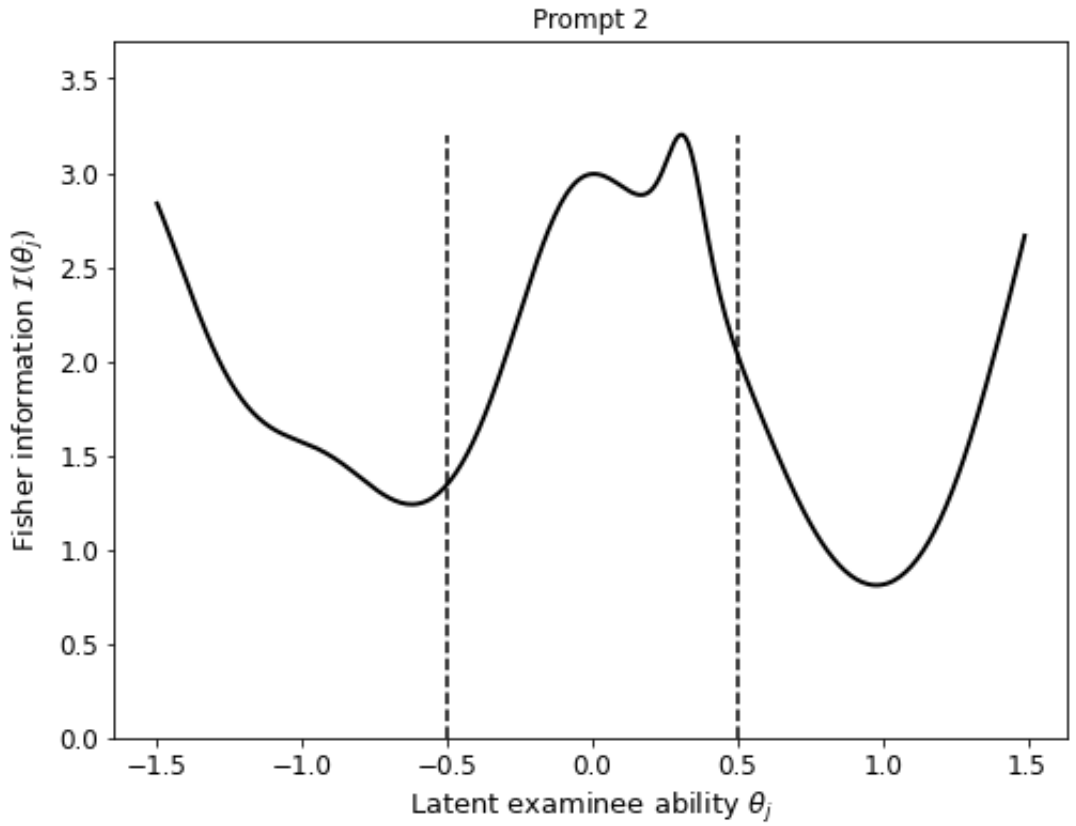

図 6: 課題番号 2 のフィッシャー情報量 $\mathcal{I}(\theta_j)$

ここで，図 5, 6 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 4 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 5 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 6 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．

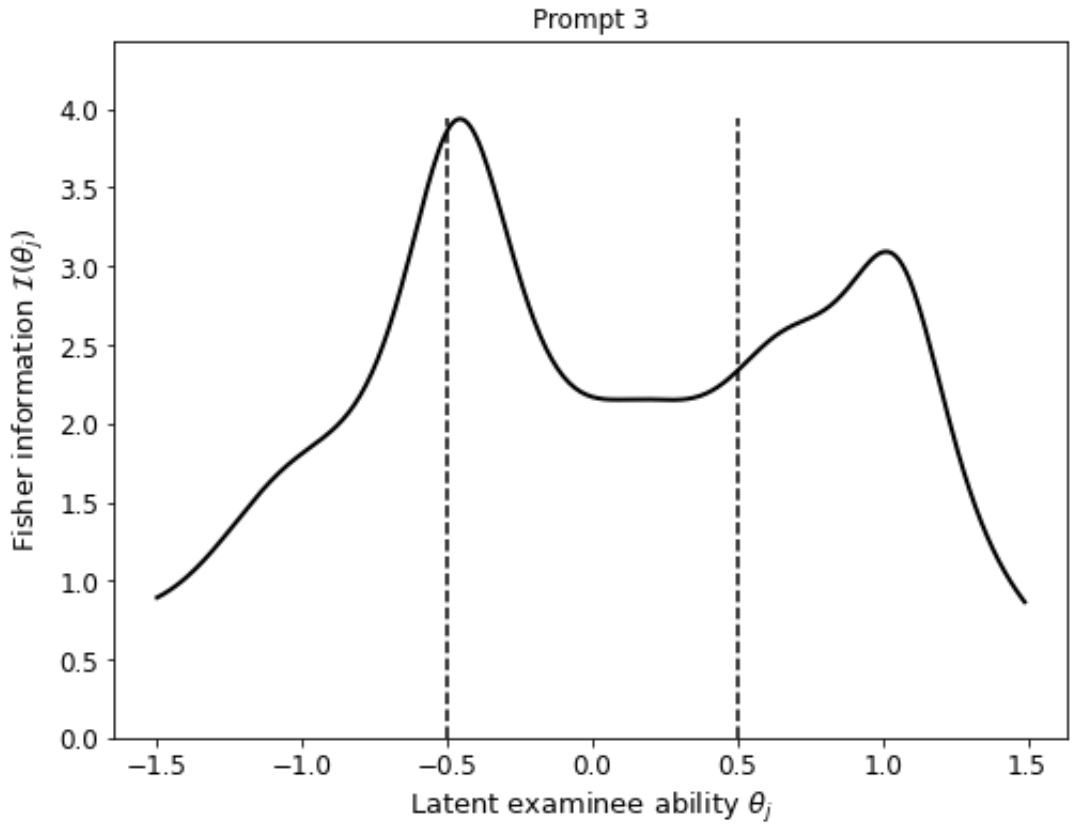

図 7: 課題番号 3 のフィッシャー情報量 $\mathcal{I}(\theta_j)$

さらに，図 7 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 4 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．

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 (Yeung 2019, 木下涼 and 植野真臣 2020)．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．

謝辞

本論文を作成するにあたり，指導教員の植野真臣教授，宇都雅輝准教授から，丁寧かつ熱心なご指導を賜りました．ここに感謝の意を表します．また，ゼミや日常の議論を通じて多くの示唆や知識を頂いた川野秀一准教授，西山悠准教授，研究室の先輩・同期・後輩に感謝いたします．

参考文献

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

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-raterÂ® v.2. *The Journal of Technology, Learning and Assessment*, **4** (3).

Beigman Klebanov, B., Flor, M., and Gyawali, B. (2016). Topicality-based indices for essay scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 63–72.

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

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

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, **76**(1).

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

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

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

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

Daumé III, H.June . (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 256–263. Association for Computational Linguistics, Prague, Czech Republic.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, 4171–4186.

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement*. Peter Lang, Bern.

Farag, Y., Yannakoudakis, H., and Briscoe, T. (2018). Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the 2018 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 263–271.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, **15**(47), 1593–1623.

Hussein, M. A., Hassan, H. A., and Nassef, M. (2019). Automated language essay scoring systems: a literature review. *PeerJ Computer Science*, **5**, e208.

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

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

Ke, Z. and Ng, V.7 . (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 6300–6308.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press, Chicago.

Liu, J., Xu, Y., and Zhu, Y.Jan. . (2019). Automated Essay Scoring based on Two-Stage Learning. *arXiv e-prints*, arXiv:1901.07744: arXiv:1901.07744.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge, Abingdon-on-Thames.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J.,

and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, 55–60.

Mayfield, E. and Black, A. W. (2020). Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 151–162.

Myford, C. M. and Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet rasch measurement: part I. *Journal of Applied Measurement*, **4**(4), 386–422.

Nguyen, H. and Litman, D. (2018). Argument mining for improving the automated scoring of persuasive essays. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 5892–5899.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay,É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**(85), 2825–2830.

Peldszus, A. and Stede, M. (2013). From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence*, **7**(1), 1–31.

Phandi, P., Chai, K. M. A., and Ng, H. T. (2015). Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 431–439.

Socher, R., Chen, D., Manning, C. D., and Ng, A. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems 26*, 926–934. (2013).

Taghipour, K. and Ng, H. T. (2016). A neural approach to automated

essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1882–1891.

Tay, Y., Phan, M., Luu, A. T., and Hui, S. C. (2018). SkipFlow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 5948–5955.

Uto, M. (2019). Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. In *Artificial Intelligence in Education*, 494–506.

Uto, M. and Okano, M. (2020). Robust neural automated essay scoring using item response theory. In *Artificial Intelligence in Education*, 549–561.

Uto, M. and Ueno, M. (2016). Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, **9**(2), 157–170.

Uto, M. and Ueno, M. (2018). Item response theory without restriction of equal interval scale for rater's score. In *Artificial Intelligence in Education*, 363–368.

Uto, M. and Ueno, M. (2020). A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika*, **47**, 469–496.

Uto, M., Xie, Y., and Ueno, M.Dec. . (2020). Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6077–6088. International Committee on Computational Linguistics.

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

Wang, Y., Wei, Z., Zhou, Y., and Huang, X. (2018). Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 791–797.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, **8**(3), 229–256.

Yeung, C. K. (2019). Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. In *Proceeding of the 12th International Conference on Educational Data Mining (EDM)*, 683–686.

木下涼 and 植野真臣. (2020). 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*, **J103-D**(4), 314–329.