アノテータのバイアスに頑健な小論文自動採点手法

岡野 将士　　宇都 雅輝
電気通信大学
{okano, uto}@ai.lab.uec.ac.jp

1 はじめに

近年，新しい時代に必要となる資質･能力として論理的思考力や批判的思考力が求められ，これら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，大規模な小論文試験には，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問題が存在する．自動採点技術（Automated essay scoring）はこれら問題の解決策の一つとして古くから注目されており，現在も多くの研究がなされている．

自動採点手法としては，事前に定義された特徴量（Handcrafted features）を用いる手法が古くから利用されている．例えば，この手法の代表例である e-rater[3] は，TOEFL や Graduate Management Admissions Test（GMAT）の作文試験で実際に用いられている．e-rater は，12 個の特徴量を説明変数，得点を目的変数とする重回帰モデルを用いて予測を行う．この際，モデルの重みパラメータは経験的に決定される．この手法は汎用的かつ低コストで利用できるが，対象とする小論文問題に固有の特徴量を組み込みにくいという問題が残る [1].

この問題を解決する手法として，深層学習モデルに基づく自動採点手法が近年多数提案されている（e.g.,[1, 4]）．この手法では，個々の小論文問題に対する採点済み答案のデータセットを用いて深層学習モデルを学習することで，予測に有効な特徴量を対象問題ごとに獲得できる．この手法は，データ収集のコストは大きいが，個別の問題に固有の特徴量も予測に利用でき，高精度な自動採点が実現されている [4].

深層学習自動採点手法のような教師あり機械学習モデルに基づく自動採点手法では，データセット中の個々の答案文に与えられた得点は真の値と仮定する．しかし，大規模な小論文試験では，多数の評価者（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評価者の特性（甘さ/厳しさなど）に強く依存することが知られている [5]．このような評価者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 [2].

他方で，近年，教育・心理測定の分野において，このような評価者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，小論文試験などで実際に活用されている [6]．具体的には，数理モデルを用いたテスト理論の一つとして様々な客観式テストで利用されてきた項目反応モデルに，評価者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．

そこで，本研究では，この項目反応モデルで得られる得点を予測するように自動採点機を学習する手法を提案する. このアプローチは様々な深層学習自動採点モデルで利用できるが，本研究では現在最も一般的に利用されている LSTM（Long short-term memory）に基づくモデル [4] を利用する．提案手法を利用することで，個々の答案を採点する評価者の特性に依存せず，自動採点モデルを学習できる．本論文では，実データ実験により提案モデルの有効性を示す.

2 データ

本研究では，J 人の受験者 $\mathcal{J}=\{1,\cdots,J\}$ に小論文問題を与え，その答案を R 人の評価者（アノテータ）集団 $\mathcal{R}=\{1,\cdots,R\}$ が K 段階得点 $\mathcal{K}=\{1,\cdots,K\}$ で採点する場合を考える．なお，実際には採点負担を軽減するために，個々の答案に R 人のうちの数名を割り当てて採点が行われる．したがって，受験者 $j\in\mathcal{J}$ の答案を e_j で表し，答案 e_j に対する評価者 r の得点を U_{jr} とすると，得点データは $\boldsymbol{U}=\{U_{jr}\in\mathcal{K}\cup\{-1\}|j\in\mathcal{J},r\in\mathcal{R}\}$ と定義できる．ここで，$U_{jr}=-1$ は欠測データを表す.

また，語彙集合を $\mathcal{G}=\{1,\cdots,G\}$ とすると，各答案 e_j は単語の系列として，$\boldsymbol{e}_j=\{\boldsymbol{w}_{jn}|n=\{1,\cdots,N_j\}\}$ と定義できる．ここで，$\boldsymbol{w}_{jn}$ とは答案 e_j 内の n 番目の単語を表す one-hot ベクトル，N_j は答案 e_j 内の単語数である．

本研究ではこのデータから深層学習自動採点モデルを学習する．

3 LSTMを用いた自動採点モデル

本研究では，深層学習を用いた自動採点モデルの基礎モデルとして知られている LSTM を用いた自動採点モデル [4] を用いる．このモデルでは単語系列を入力とし，4 つの層を通して得点を予測する．以下でモデルの各層についての説明を行う．

1 層目の Lookup Table Layer では各単語を埋め込みベクトル表現 (word embeddings) に変換し，2 層目の Recurrent Layer で埋め込みベクトルの系列を LSTM に入力する．次に 3 層目の Mean over Time Layer では，Recurrent Layer の出力列 $\mathcal{H} = (\boldsymbol{h}_{j1}, \boldsymbol{h}_{j2}, \cdots, \boldsymbol{h}_{j(N_j)})$ に対する平均ベクトル $\boldsymbol{M}_j = \frac{1}{N_j}\sum_{n=1}^{N_j} \boldsymbol{h}_{jn}$ を計算する．最後に Linear Layer with Sigmoid Activation Layer で Mean over Time Layer の出力ベクトルを $\hat{y}_j = \sigma(\boldsymbol{W}\boldsymbol{M}_j + b)$ でスカラーの値に対応させ，その値を推定得点 $\hat{y}_j$ とする．ここで，$\boldsymbol{W}$ と b はそれぞれ線形モデルの重みとバイアスである．このとき，$\hat{y}_j$ は 0 から 1 の値を取る．得点尺度がこれと異なる場合には，$\hat{y}_j$ を一次変換し，実際の得点尺度に合わせる．

モデルの学習で用いる損失関数には，平均二乗誤差（mean squared error：MSE）を用いる．

ここで，教師ラベルとなる真の得点については，評価者が 1 名の場合にはその評価者の得点を，評価者が複数の場合には得点の合計点や平均点を一般に利用する．しかし，1 章でも述べたように，答案ごとに評価者が異なる場合，得られる得点が評価者の特性（甘さ/厳しさなど）に強く依存することが知られている [5]．したがって，これらの得点を真の値としてモデルを学習すると，学習したモデルが評価者特性の影響を受けてしまい，予測精度が低下する [2]．これは教師あり機械学習モデルに基づくすべての自動採点モデルに当てはまる課題である．

他方で，このような評価者の影響を取り除いて各答案の得点を推定できる手法として，項目反応理論に基づくモデルが提案されている．本研究では，このモデルを利用して，評価者バイアスに頑健な自動採点モデルを学習することを目標とする．

4 項目反応理論

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，IRT モデルに評価者（アノテータ）の特性を表すパラメータを加えた宇都・植野のモデル [6] を利用する．

宇都・植野のモデル [6] は，代表的な多値型 IRT である一般化部分採点モデルに評価者特性パラメータを付与した拡張モデルであり，受験者 j のテスト問題 i に対する答案に対し，評価者 r が得点 k を与える確率を次式で定義する．

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

θ_j は受験者 j の真の能力を表す潜在変数であり，α_i，β_i はそれぞれテスト問題 i の識別力，困難度を表す．また，α_r，β_r，d_{rk} はそれぞれ評価者 r の一貫性，厳しさ，得点 k に対する評価者 r の厳しさを表す．ただし，パラメータの識別性のために，$\prod_{i=1}^{I}\alpha_i = 1$，$\sum_{i=1}^{I}\beta_i = 0$，$d_{r1} = 0$，$\sum_{k=2}^{K} d_{rk} = 0$ を仮定する [6]．

2 章で述べたように，本研究ではテスト問題が一つの場合を想定している．この設定では，採点対象となる答案の数は受験者ごとに一つのみとなるため，θ_j は受験者 j の能力を表すとともに，その受験者の答案の補正得点（評価者特性の影響を取り除いた得点）を表す潜在変数とみなせる．

本研究の主なアイディアは，このモデルによって評価者特性の影響を考慮して推定される得点 θ_j を予測するように LSTM 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，評価者の特性を考慮した IRT モデルと LSTM を用いた自動採点モデルを統合することで，評価者バイアスに頑健な自動採点手法を提案する．

提案手法の概念図を図 1 に示す．図 1 のように提案手法は，IRT による得点補正のフェーズと，LSTM モデルの学習フェーズの二段階で構成される．具体的には次の通りである．1) 評価者が与える評点データ $\boldsymbol{U}$ 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) 手順 1 で求めた得点 θ_j を予測するように，LSTM を用いた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．

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

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

得点予測は，前節で学習されたモデルを用いて θ_j を

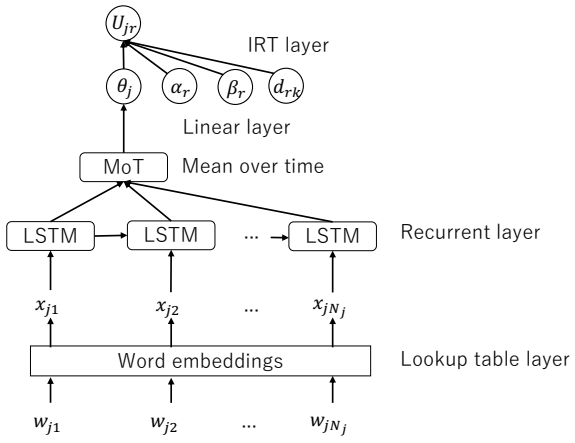

図 1: 提案モデルの概念図

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．

$$\hat{U}_j = \sum_{r=1}^{R} \frac{1}{R} \sum_{k=1}^{K} k \cdot P_{jrk} \tag{3}$$

6 評価実験

ここでは，実データ実験を通して評価者バイアスを取り除いた自動採点の有効性を評価する．

6.1 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）を実データとして使用する．ASAP は，8 つの異なるトピックに対するエッセイ答案データとそれに対する得点データで構成されている．ただし，答案は複数の評価者で採点されていると記載されているが，ASAP のデータには評価者を識別できる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．そこで本研究では，新たに評価者を採用して ASAP の答案データを再度採点させることで本実験に利用できるデータを収集した．ここでは，先行研究で予測精度が最も高かったトピック 5 の答案データを利用した．具体的にはトピック 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 37 名の評価者を 1 つの答案あたり 3〜5 名割り当てて採点を行った．採点基準は ASAP で公開されているものを使用し，5 段階で評価を行った．得られた得点データは ASAP の得点データとの相関が平均で 0.656 であった．

表 1: IRT の有効性評価結果

	IRT	素点	p 値	t 値
RMSE	0.526	0.914	$p < 0.001$	129.9
QWK	0.454	0.409	$p < 0.001$	16.48

6.2 項目反応理論の有効性評価

4 章で述べたように評価者特性を考慮した IRT モデルを用いることで評価者に依存しない安定した得点を推定することができる．本節では，予備実験としてこの点を評価する．

評価者に依存しない安定した得点を推定できるのであれば，異なる評価者が採点したデータを元に学習したとしても得点の差異が安定して小さくなると期待される．この考え方に基づき，ここでは次の手順で IRT の有効性を評価した．1）実データを用いてマルコフ連鎖モンテカルロ法によりモデルパラメータを推定した．2）各答案に与えられた複数の評価者による得点の中からランダムに 1 つ選択して得られる得点データセットを 10 パタン作成した．これらデータセットを $\{\boldsymbol{U}'_1, \cdots, \boldsymbol{U}'_{10}\}$ とする．3）手順 1 で推定した項目・評価者パラメータを所与として，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 θ を推定し，式（3）に基づいて期待得点を求めた．この操作を全てのデータセットに対して行った．4）n 番目のデータセットから求めた期待得点と n' 番目の得点データセットから推定した期待得点との二乗平均平方根誤差 (Root Mean Squared Error:RMSE) と 2 次の重み付きカッパ係数 (Quadratic Weighted Kappa:QWK) を $n \in \{1, \cdots, 10\}$，$n' \in \{1, \cdots, 10\}$ の全ての組み合わせについて求め，RMSE と QWK の平均を算出した．

本実験では，比較のために，IRT を利用しない場合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 $\{\boldsymbol{U}'_1, \cdots, \boldsymbol{U}'_{10}\}$ を用いて，手順 4 と同様に RMSE と QWK の平均を算出した．この実験では，RMSE が低いほど，また QWK が大きいほど，評価者によらない安定的な得点予測ができたことを意味する．また，IRT を利用する場合としない場合で，RMSE と QWK の平均値に有意な差があるかを確認するために，t 検定を行った．

実験結果を表 1 に示す．表 1 から，IRT モデルを利用した場合の方が，RMSE が有意に小さく，QWK が有意に大きくなったことが確認できる．この結果は，

IRT が評価者の特性差を考慮して真の得点を精度よく推定できたことを意味するものであり，先行研究 [6] と一致した結果となっている．

6.3 提案手法の性能評価

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，以下の手順に従って評価実験を行った．1）6.2 の手順 1，2 と同様に，ランダムに選んだ評価者 1 名分の各得点データ $\boldsymbol{U}'_n$ から，IRT 得点 θ を推定した．2）得られた θ と答案文のデータセットを 5 分割し，4/5 を学習データとして提案モデルを学習したのち，1/5 のテストデータに対して得点予測を行った．予測は，IRT 得点の予測値 θ を所与として式（3）で期待得点を求めることで行った．これを全ての分割パタンで行うことで，全てのデータに対して期待得点を求めた．3）手順 2 を $n = \{1, \cdots, 10\}$ について行ったあと，6.2 の手順 4 と同様に，n 番目のデータセットから求めた期待得点と n' 番目の得点データセットから推定した期待得点との RMSE と QWK を $n \in \{1, \cdots, 10\}$，$n' \in \{1, \cdots, 10\}$ の全ての組み合わせについて求め，それらの RMSE と QWK の平均を算出した．

比較のために，IRT を利用しない従来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，各得点データセット $\boldsymbol{U}'_n$ を用いて手順 2 と同様に 5 分割交差検証法で全答案の予測得点を推定したあと，手順 3 と同様にデータセットごとに得られた予測得点間の RMSE と QWK の平均を算出した．また，提案手法と従来手法で RMSE と QWK の平均値に有意な差があるかを確認するために，t 検定を行った．

なお，本実験では，深層学習モデル学習に Python-Keras で実装したプログラムを利用し，ハイパーパラメータやデータの前処理は先行研究 [4] に合わせた．

実験結果を表 2 に示す．表 2 から，IRT を利用する提案手法の方が RMSE が有意に小さく，QWK が有意に大きくなったことが確認できる．このことは，IRT で推定した能力値 θ を目的変数として自動採点モデルを学習することで，評価者に頑健な自動採点を実現できることを示している．

表 2: 提案手法の有効性評価結果

	提案手法	従来手法	p 値	t 値
RMSE	0.160	0.260	$p < 0.001$	30.00
QWK	0.785	0.678	$p < 0.001$	20.35

7 おわりに

近年注目が集まっている深層学習を用いた自動採点手法では，各答案に対する得点にアノテータのバイアスの影響があると想定される場合，学習結果もその影響を受けてしまい，得点予測の性能が低下するという問題がある．本研究では，IRT を用いてアノテータのバイアスを考慮した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た．また，実データ実験により，提案手法を用いることでどのような評価者のデータを元に学習したとしても推定得点が安定する自動採点モデルを作れることを示せた．

今後の課題として，提案アプローチを他の自動採点モデルへ適用することを検討したい．また，様々な実データに対して提案手法を適用し，有効性を評価したい．さらに，2 段階で行なっていた推定を end-to-end にすることも検討したい．

参考文献

[1] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 715–725. Association for Computational Linguistics, 2016.

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

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

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

[5] Masaki Uto and Maomi Ueno. Empirical comparison of item response theory models with rater's parameters. *Heliyon*, Vol. 4, No. 5, p. e00622, 2018.

[6] Masaki Uto and Maomi Ueno. Item response theory without restriction of equal interval scale for rater's score. In *Artificial Intelligence in Education*, pp. 363–368. Springer International Publishing, 2018.