修 士 論 文 の 和 文 要 旨

研究科・専攻	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		
氏　　名	岡野　将士	学籍番号	2031035
論 文 題 目	評価者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		

要　　旨

近年，論理的思考力や表現力などを評価する手段の一つとして，小論文試験のニーズが高まっている．しかし，大規模試験に小論文試験を導入する場合，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点で課題が存在する．これらの課題を解決する手法の一つとして，自動採点技術が注目されている．

近年の自動採点手法の多くは，深層学習技術を用いたモデルとして定式化されている．それらのモデルを自動採点に活用するためには，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．一般にモデル学習の際には，評価者が与えた得点を真値と仮定する．しかし，大規模試験では多数の評価者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は評価者の特性（甘さ/厳しさなど）に依存することが知られている．こ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．

他方で，そのような評価者特性の影響を取り除いて得点を推定する方法として，数理モデルを用いたテスト理論の一つである項目反応モデルに，評価者の特性を表すパラメータを加えたモデルが提案されている．岡野・宇都（2021）は，このような項目反応モデルを用いて訓練データ中の得点データから評価者バイアスの影響を取り除いた得点を推定し，それを用いて自動採点モデルを学習する手法を提案しており，それにより自動採点の精度が向上することを示している．しかし，従来手法では，項目反応モデルに基づく得点の推定に評価者が与えた得点データのみを用いており，答案文の情報は使用していない．一方で，評価者特性を取り除いた得点を推定する際には答案文の内容自体も有益な情報となりうる．

そこで本研究では，項目反応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したモデルを提案する．提案モデルでは，評価者バイアスを取り除いた得点の推定を，答案文の内容も加味して高精度に行うことができ，このアプローチのさらなる性能向上が期待できる．本研究では，実データ実験を通じて，提案手法の有効性を示す．

令和３年度　修士論文

評価者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

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
学籍番号 2031035

岡野　将士

主任指導教員：宇都 雅輝　准教授
指導教員：川野 秀一　准教授

2022年1月28日

目次

図目次

表目次

第1章

はじめに

第4次産業革命と呼ばれるような近年の急速な社会変化に伴い，学校教育では従来の知識・技能の習得とともに思考力・判断力・表現力などの育成が重視されるようになった[1]．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，入学試験や資格試験などの大規模試験に小論文試験を導入する場合，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点で課題が存在する[2, 3, 4, 5]．自動採点手法（Automated essay scoring：AES）はこれらの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[6, 7, 8, 9]．

自動採点を実現する手法としては，事前に定義された特徴量（Handcrafted features）を用いる手法が古くから研究されている（e.g.,[10, 11, 12, 13, 14, 15]）．この手法では，答案文から採点に用いる特徴量を抽出し，機械学習モデルに入力することで得点を出力する．例えば，e-rater[10]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自動採点を行う．このような手法は，特徴量設計が一度完了すれば様々な小論文試験に容易に適用できるという利点を有する．一方で，特徴量の設計が精度に大きく影響を与えることから，高精度を達成するためには対象とするデータセットの性質に合わせた特徴量のチューニングや再設計が必要であることが指摘されてきた[16, 17]．

この問題を解決する手法の一つとして，特徴量の設計を必要としない自動採点手法が近年多数提案されている（e.g.,[16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28]）．具体的には，答案文の単語系列を深層学習モデルに直接入力することで得点を出力する手法である（以降，この手法を用いた自動採点モデルを深層学習自動採点モデルと呼ぶ）．深層学習自動採点モデルを利用するためには，対象とする小論文問題ごとに採点済み小論文のデータセットを収集し，モデル学習を行う必要がある．このため，データ収集のコストは大きいものの，個別のデータセットに固有の特徴量を自動で抽出でき，高精度な自動採点が実現できる[22, 24]．

上述の通り，これらのモデルの学習には，事前採点済みの答案文を集めた訓練データが必要である．モデル学習時には，訓練データ中の各答案に評価者が与えた得点（以降では観測得点と呼ぶ）はバイアスのない正確な値であると仮定する．しかしながら，大規模な小論文試験では，評価者の採点負担軽減のために，多数の評価者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，観測得点は評価者の特性（甘さ/厳しさなど）に強く依存してしまうことが知られている[29]．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[30, 31, 32, 33]．

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得点を推

定できる手法が多数提案されている．具体的には，数理モデルを用いたテスト理論の一つとして様々な客観式テストで利用されてきた項目反応（Item Response Theory：IRT）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（e.g.,[29, 34, 35, 36, 37, 38, 39, 40, 41]）．これらのモデルでは，評価者の与えた得点の集合から最尤推定やベイズ推定によって評価者特性の影響を取り除いた得点を推定する．これらのモデルは，小論文試験を含む様々な試験に適用され，評価者バイアスを取り除いた高精度な得点推定を実現できることが示されてきた．

岡野・宇都[42, 43, 44]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，評価者の特性を考慮したIRTモデルを用いて訓練データ中の観測得点から評価者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．この手法によって，評価者のバイアスに頑健なモデル学習と得点予測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得点の推定に評価者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出力層にIRTモデルを組み込んでend-to-endで学習する手法である．この手法では，IRT得点の推定に評価者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．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[18]と，最先端モデルの一つであるBERT（Bidirectional Encoder Representations from Transformers）を用いたモデル[25]を用いる．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．

なお，近年，青見ら[45, 46]は複数の自動採点モデルの得点をIRTモデルを用いて統合するモデルを提案している．しかし，このモデルでは，採点精度の改善を主目的としており，本研究のように訓練データ中の評価者バイアスを考慮することはできない．

第2章

データ

本研究では，深層学習自動採点モデルの訓練データとして，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，それらの答案を$R$人の評価者$\mathcal{R} = \{1, \ldots, R\}$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$j \in \mathcal{J}$の小論文答案e_jの集合であり，個々の答案e_jは単語の系列として次式で定義できる．

$$e_j = \{\boldsymbol{w}_{jn} | n = \{1, \ldots, N_j\}\} \tag{2.1}$$

ここで，N_jはe_j内の単語数を表し，$\boldsymbol{w}_{jn}$は答案e_j内のn番目の単語を表すG次元のone-hotベクトルである（Gは答案集合$\boldsymbol{A}$に出現する語彙の数を表す）．

また，得点集合$\boldsymbol{U}$は，答案e_jに対して評価者$r \in \mathcal{R}$がK段階$\mathcal{K} = \{1, \ldots, K\}$で与えた観測得点$U_{jr}$の集合として次式で定義できる．

$$\boldsymbol{U} = \{U_{jr} \in \mathcal{K} \cup \{-1\} | j \in \mathcal{J}, r \in \mathcal{R}\} \tag{2.2}$$

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に評価者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評価者の負担軽減のために，個々の答案に数名の評価者を割り当てて採点が行われるため，一般にこのような欠測が生じる．

第3章

自動採点モデル

自動採点を実現する手法として，小論文答案から抽出した特徴量を入力として得点を推定する「特徴量ベースのアプローチ」と，小論文答案の単語系列を入力として得点を推定する「深層学習ベースのアプローチ」が知られている．本章では，これら２つのアプローチに基づく自動採点モデルをそれぞれ紹介する．

3.1 特徴量ベースのアプローチ

特徴量を用いて自動採点を実現するアプローチは古くから研究されており，実際の試験現場でも運用されている．この手法は採点に用いる特徴量を事前に設定し，それらの特徴を形態素分析ツールなどを用いて答案文から抽出する．そして，それらを機械学習モデルへの入力として，得点を推定するアプローチである．ここでは，TOEFLなどの試験で実際に用いられているe-rater[47, 10]と，小論文自動採点に関する国際的なコンペティションで入賞したフリーソフトウェアのEASE[48]を紹介する．

3.1.1 e-rater

e-rater[47, 10]は試験作成・評価を実施する組織としては世界最大規模となるEducational Testing Service（ETS）によって開発された自動採点モデルである．e-raterは実際にTOEFLなどのいくつかの試験で評価者を補助する目的で用いられている．また，一部の機能を用いたCriterionと呼ばれるサービスがオンラインで提供されており，多くの教育機関がe-raterのシステムを用いることができる[49, 50].

e-rater version.2では以下の12種類の特徴量を入力とし, 6点満点の段階得点を推定する[10, 51, 52].

Grammer,Usage,Machanics, and Style Measure

- Grammer：冠詞や語順，語形変化など，文法の誤りの割合に関する特徴量
- Usage：主語・動詞の関係や時制など，文章の構成の誤りの割合に関する特徴量
- Machanics：スペル，句読点，文頭の大文字など，文章の体裁面の誤りの割合に関する特徴量
- Style：単語の使用頻度や極端な文章長など，文章の自然さに関する特徴量

Organization and Development

- Overall organization：談話（Discourse）のパラグラフ数に基づく特徴量

- Development：単語数による，各々のパラグラフでの議論の展開の深さに関する特徴量

Lexical Complexity

- Vocabulary：単語頻度指標[53]に基づく語彙の難易度に関する特徴量
- Word length：答案に用いられる単語の平均的な長さに関する特徴量
- Repetitive word use：同一単語の繰り返し使用頻度に関する特徴量

Prompt-Specific Vocabulary Usage

- $\max.\cos.$：各得点カテゴリに対するサンプルの答案文とのコサイン類似度に基づく特徴量
- $\cos.w/6$：サンプルとして用意された最高得点の答案文とのコサイン類似度に基づく特徴量

Essay Length

- Word count：答案に用いた単語の個数による，答案文の長さに基づく特徴量

e-raterでは，これらの特徴量を答案文から算出し，経験則によって定めた特徴量ごとの重みに基づいて，重回帰モデルにより得点を推定する．

3.1.2　EASE

EASE（Enhanced AI Scoring Engine）[48]はヒューレット財団がスポンサーとなり開催された自動採点のコンペティションであるAutomated Student Assessment Prize[54]で入賞した自動採点モデルである．このモデルはフリーソフトウェアとして誰でも使えるように公開されているため，多くの研究で比較モデルとして用いられている．

Phandi et al. [12]によると，EASEは以下の特徴量を用いて答案文を評価する．

Length：　文字・単語・句読点数による，答案文の長さに基づく特徴量
Part of speech (POS)：　品詞n-gramを用いた文法誤りの数・割合に基づく特徴量
Prompt：　課題特有の単語の総数や文全体に対する使用割合に基づく特徴量
Bag of words：　高得点の小論文に用いられる文章表現や適切な語句の出現回数に基づく特徴量

EASEでは，これらの特徴量を答案から抽出し，回帰モデルに入力することで得点を推定する．また，回帰モデルとしてはベイジアンリッジ回帰（Bayesian Linear Ridge Regression：BLRR）とサポートベクター回帰（Support Vector Regression：SVR）が使用されている．

3.2　深層学習ベースのアプローチ

近年発表されている多くの自動採点モデルには深層学習技術が用いられている[6]．自動採点に深層学習を用いることで，個別のデータセットに固有の特徴量を自動で抽出することができ，その結果，高精度な自動採点が実現されている．ここでは，深層学習技術を用いた自動採点モデルとして初期に登場したLSTMを用いたモデル[18]と様々な自然言語処理分野のタスクで最高精度を達成しているBERTを用いたモデル[25, 26, 27, 28, 55]を紹介する．

3.2.1　LSTMを用いた自動採点モデル

LSTMを用いた自動採点モデル[18]はKaveh et.al. によって2016年に発表され，その後の様々な深層学習技術を用いた自動採点モデルのベースラインとなったモデルである．このモデルの概念図を図 3.1に示す．

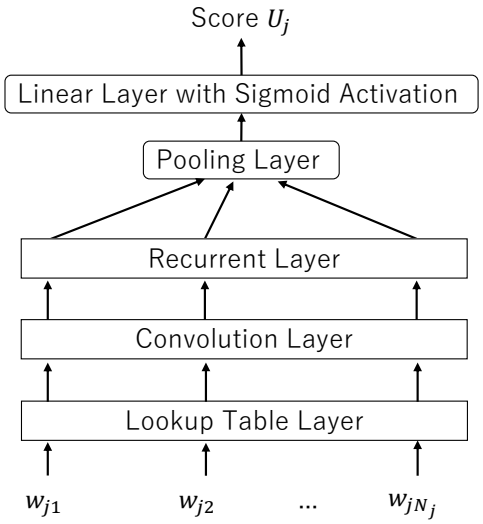


図3.1: LSTMを用いた自動採点モデルの概念図

このモデルは答案の単語系列を入力とし，以下の5つの層を通して得点を予測する．

Lookup Table Layer:

この層では，答案中の各単語を単語の意味を表す埋め込み表現(word embeddings)ベクトルに変換する．具体的には，one-hot表現で表したG次元の単語ベクトル$\boldsymbol{w}_{jn}$と$V \times G$次元の埋め込み行列(word embeddings matrix)$\boldsymbol{E}$との内積$\boldsymbol{x}_{jn} = \boldsymbol{E}\boldsymbol{w}_{jn}$を計算することで，答案文中の各単語を$V$次元の埋め込み表現ベクトル$\boldsymbol{x}_{jn}$に変換する．

Convolution Layer:

この層では，単語系列の局所的な特徴を抽出するために，畳み込みニューラルネットワーク（Convolutional Neural Network：CNN）を用いてn-gramレベルの特徴量を抽出する．具体的には, Sをウィンドウ幅とすると, s番目のウィンドウを表す局所的な単語系列$\{\boldsymbol{x}_{js} \dots \boldsymbol{x}_{j(s+S)}\}$について，それを連結（Concatenatation）したベクトル$\bar{\boldsymbol{x}}_{js}$を線形変換$\boldsymbol{y}_{js} = \boldsymbol{C}\bar{\boldsymbol{x}}_{js} + \boldsymbol{b}^v$する操作を全ての$s \in \{1, \dots, N_j - S\}$について繰り返す．ここで$\boldsymbol{C}$と$\boldsymbol{b}^v$はパラメータであり，全てのウィンドウで同じ値をとる．また，この層からの出力系列の長さがN_jに維持されるように，出力系列にはゼロパディングが適用される．なお，この層は，LSTMに基づく最近の自動採点モデルでは省略されることもある．

Recurrent Layer:

この層では, 時系列データを処理する深層学習モデルであるRNN（Recurrent Neural Network）を用いて，小論文の得点予測に有効な特徴量を入力系列の時系列的な関係を考慮して抽出す

る．RNNとしてはLSTMが一般に用いられる．LSTMを構成するブロックの概念図を図3.2に示す．LSTMでは，n番目の入力$\boldsymbol{y}_{jn}$に対して以下の式を計算することで，特徴量ベクトル$\boldsymbol{h}_{jn}$を得る．

$$\boldsymbol{a}_{jn} = \sigma(\boldsymbol{W}^a \boldsymbol{y}_{jn} + \boldsymbol{Z}^i \boldsymbol{h}_{j(n-1)} + \boldsymbol{b}^a) \tag{3.1}$$
$$\boldsymbol{f}_{jn} = \sigma(\boldsymbol{W}^f \boldsymbol{y}_{jn} + \boldsymbol{Z}^f \boldsymbol{h}_{j(n-1)} + \boldsymbol{b}^f) \tag{3.2}$$
$$\tilde{\boldsymbol{m}}_{jn} = \tanh(\boldsymbol{W}^m \boldsymbol{y}_{jn} + \boldsymbol{Z}^c \boldsymbol{h}_{j(n-1)} + \boldsymbol{b}^m) \tag{3.3}$$
$$\boldsymbol{m}_{jn} = \boldsymbol{a}_{jn} \circ \tilde{\boldsymbol{m}}_{jn} + \boldsymbol{f}_{jn} \circ \boldsymbol{m}_{j(n-1)} \tag{3.4}$$
$$\boldsymbol{o}_{jn} = \sigma(\boldsymbol{W}^o \boldsymbol{y}_{jn} + \boldsymbol{Z}^o \boldsymbol{h}_{j(n-1)} + \boldsymbol{b}^o) \tag{3.5}$$
$$\boldsymbol{h}_{jn} = \boldsymbol{o}_{jn} \circ \tanh(\boldsymbol{m}_{jn}) \tag{3.6}$$

ここで$\boldsymbol{h}_{jn}$はn番目の入力に対するLSTMブロックの出力ベクトルであり，$\boldsymbol{W}^a$，$\boldsymbol{W}^f$，$\boldsymbol{W}^m$，$\boldsymbol{W}^o$，$\boldsymbol{Z}^a$，$\boldsymbol{Z}^f$，$\boldsymbol{Z}^m$，$\boldsymbol{Z}^o$は重み行列，$\boldsymbol{b}^a$，$\boldsymbol{b}^f$，$\boldsymbol{b}^m$，$\boldsymbol{b}^o$はバイアスを表すベクトル，$\boldsymbol{a}_{jn}$，$\boldsymbol{f}_{jn}$，$\boldsymbol{o}_{jn}$はそれぞれn番目の単語に対する入力ゲート・忘却ゲート・出力ゲートを表す．また，$\tilde{\boldsymbol{m}}_{jn}$と$\boldsymbol{m}_{jn}$はメモリセルを表し，入力系列の長期的な依存関係を保持する．$\circ$はアダマール積，σはシグモイド関数を表す．

なお，Recurrent Layerには単方向LSTMが利用されることが多いが，Bidirectional LSTMや多層LSTMなどが使われる場合もある．これらを用いたモデルの概念図は付録Aに示した．

Pooling Layer:

この層では，Recurrent Layerの出力系列を固定長の単一のベクトルに変換する．具体的には，Recurrent Layerの出力$\mathcal{H} = (\boldsymbol{h}_{j1}, \boldsymbol{h}_{j2}, \ldots, \boldsymbol{h}_{jN_j})$の時間方向の平均値$\boldsymbol{M}_j$を次式を用いて計算する．

$$\boldsymbol{M}_j = \frac{1}{N_j} \sum_{n=1}^{N_j} \boldsymbol{h}_{jn} \tag{3.7}$$

このプーリング法は一般にMean over Time（MoT）と呼ばれる．なお，他の典型的なプーリング法としては，最後の入力$\boldsymbol{h}_{jN_j}$のみを出力するLast poolingもしばしば利用される．

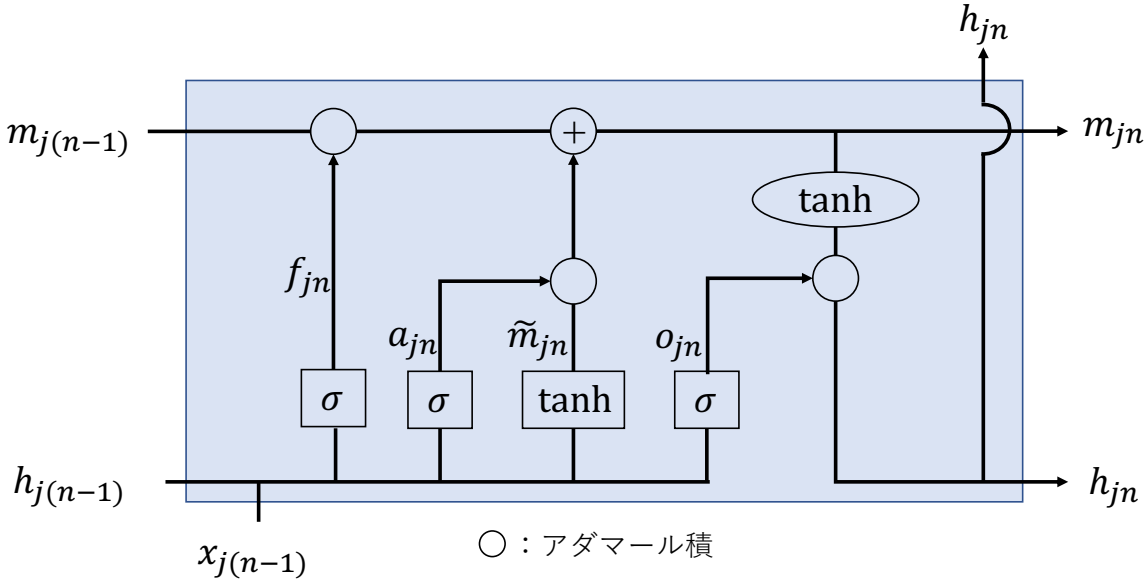


図3.2: LSTMブロックの概念図

Linear Layer with Sigmoid Activation:

この層では，Pooling Layerの出力ベクトル$\boldsymbol{M}_j$から得点に対応するスカラー値を次式で求める．

$$\hat{U}_j = \sigma(\boldsymbol{W}\boldsymbol{M}_j + b) \tag{3.8}$$

ここで，$\boldsymbol{W}$とbはそれぞれ重みとバイアスを表すパラメータである．なお，$\hat{U}_j$は0から1の値を取るため，得点尺度がこれと異なる場合には，$\hat{U}_j$を一次変換し実際の得点尺度に合わせる．例えば，1〜KのK段階得点の場合，$K\hat{U}_j + 1$と変換する．

3.2.2　BERTを用いた自動採点モデル

BERT[56]は2018年にGoogleが発表し，当時様々なタスクで最高精度を達成した自然言語処理モデルである．小論文や短答記述式問題の自動採点タスクにおいても高精度を達成している[25, 26, 27, 28, 55]．このモデルの概念図を図3.3に示す．

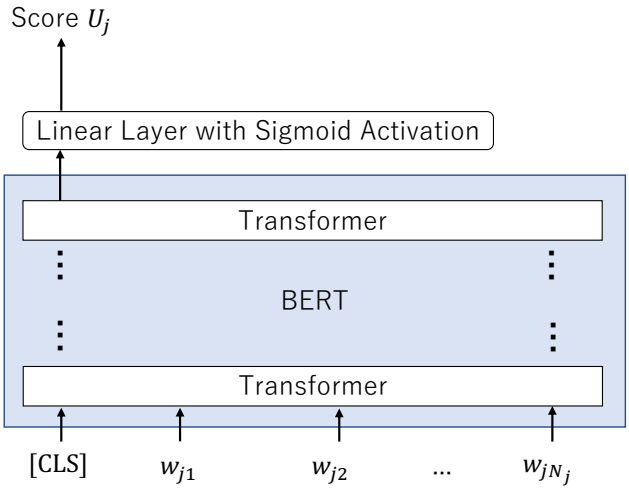


図3.3: BERTを用いた自動採点モデルの概念図

BERTでは，以下の2段階でモデルの学習を行う．

1. **事前学習（pre-training）**
   大量の教師無し文書データから以下のタスクを行うことで，汎用的な言語モデルを学習する．
   Masked Language Model:
   入力テキストの一部の単語を隠し，その単語を予測するタスク
   Next Sentence Prediction:
   2つの文章に対し，それらが隣接したものかを予測するタスク
2. **ファインチューニング（fine-tuning）**
   対象のタスクにモデルを適用させるため，事前学習で得られたモデルを所与として，教師ありデータセットを用いてモデルの再学習を行う．

なお，BERTを自動採点に用いる場合には，図3.3に示すように，全ての答案文の先頭に[CLS]とい

う特殊なトークンを挿入しておく必要がある．このトークンに対応する出力が個々の答案の特徴を表すベクトル表現となる．したがって，自動採点タスクにおいてBERTを用いる際には，このベクトルを3.2.1節で紹介したLinear Layer with Sigmoid Activationに通すことで得点を計算する．

3.3　自動採点モデルの学習と問題点

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を訓練データとして用いてモデル学習（BERTの場合，ファインチューニング）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．

$$MSE(\boldsymbol{U}, \hat{\boldsymbol{U}}) = \frac{1}{J}\sum_{j=1}^{J}(U_j - \hat{U}_j)^2 \tag{3.9}$$

ここで，U_jは小論文答案e_jに対する訓練データ中の観測得点を，$\hat{U}_j$は小論文答案e_jに対して深層学習自動採点モデルが推定した予測得点を表す．また，各答案に複数の評価者が割り当てられている場合，U_jには複数の評価者が与えた観測得点の平均などを用いる．しかし，1章でも説明したように，評価者によって与えられる観測得点データは評価者の特性に強く依存することが知られている[29]．そのように評価者特性の影響を受けた得点データをモデル学習に使用すると，学習された自動採点モデルにも評価者の特性の影響が反映され，予測精度が低下してしまう[30, 31, 32, 33]．

他方で，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として，評価者の特性を表すパラメータを加えた項目反応理論が提案されている．次章では，この項目反応理論について説明する．

第4章

項目反応理論

項目反応理論（Item Response Theory：IRT）は，近年さまざま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．項目反応理論では，テスト問題に対する受検者の正答確率を，テスト問題の特性を表す項目パラメータと，受検者の能力を表す能力パラメータの関数で表す．項目反応理論を利用することで，テスト問題の特性を考慮した受検者の能力推定が可能となる．

一般的なIRTモデルではテスト問題への正誤を表す2値データを扱う．次節では，まず，2値データを扱う最も単純なIRTモデルとしてラッシュモデルを紹介する．

4.1 ラッシュモデル

ラッシュモデル[57]は，現在も広く利用されているIRTモデルの一つである．このモデルでは，受検者jがテスト問題iに正答する確率を次式で定義する．

$$P_{ij} = \frac{\exp(\theta_j - \beta_i)}{1 + \exp(\theta_j - \beta_i)} \tag{4.1}$$

ここで，β_iはテスト問題iの困難度を表すパラメータであり，θ_jは受検者jの能力を表す潜在変数となる．パラメータの解釈を説明するために，困難度の値が異なる4つの問題に対する反応曲線を図4.1に示す．

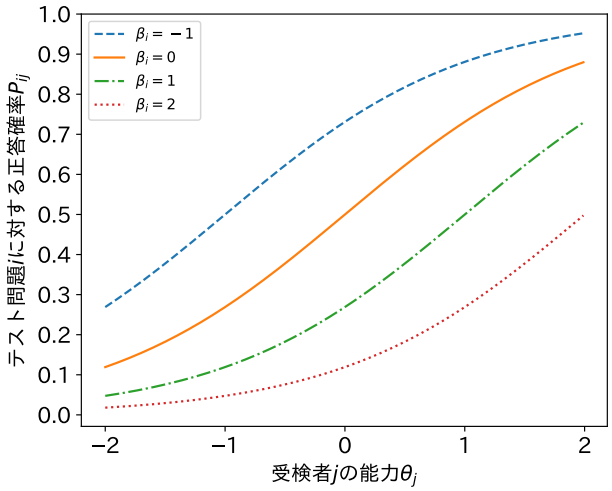


図4.1: ラッシュモデルの反応曲線

図4.1の横軸は受検者の能力θ_j，縦軸はテスト問題に対して正答する確率を示している．これらの反応曲線は，$\theta_j = \beta_i$の際に正答確率が0.5となるようにθ_j軸方向に平行移動している．そのため，困難度β_iが大きいほど，能力が高い受検者でないと正答しにくいことが表現されている．

このモデルはデータがモデルに合致する場合，少数のデータからでも安定した能力推定を行うことができる．一方で，反応曲線の傾きの大きさが一様なため，複雑な項目特性は表現することができず，データへの適合が悪いことも多い．このような課題を解決するため，より多様な問題特性を考慮したモデルが複数提案されている．そのようなモデルの代表例として，次節では，2パラメータ・ロジスティックモデルを紹介する．

4.2　2パラメータ・ロジスティックモデル

2パラメタ・ロジスティックモデル（two-parameter logistic model：2PLM）は国内外で最も多く使われているIRTモデルである．このモデルでは，受検者jがテスト問題iに正答する確率を次式で定義する．

$$P_{ij} = \frac{\exp(\alpha_i(\theta_j - \beta_i))}{1 + \exp(\alpha_i(\theta_j - \beta_i))} \tag{4.2}$$

ここで，α_iはテスト問題iの識別力を表すパラメータである．

パラメータの解釈を説明するために，識別力や困難度の値が異なる4つの問題に対する反応曲線を図4.2に示す．

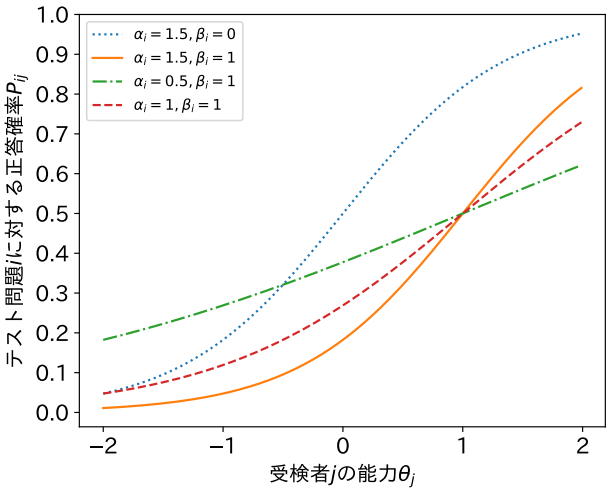


図4.2: 2PLMの反応曲線

困難度β_iは先ほどのラッシュモデルと同様にテスト問題iにおいて正答確率が0.5となる能力値θの値を表す．また，識別力α_iは能力値$\theta_j = \beta_i$付近の能力をどの程度の精度で識別できるかを表し，大きい値であるほど敏感に正答者と誤答者を識別できる．

上述したラッシュモデルや2パラメータ・ロジスティックモデルなどは，採点結果が2値（正答・誤答など）の場合にのみ適用することができるため，本研究で扱うような多段階の採点結果で構成されているデータには適用できない．本研究で扱うような多値データを扱う際には，多値型IRTモデルを用いることになる．次節では代表的な多値型IRTモデルである一般化部分採点モデルを紹介する．

4.3　一般化部分採点モデル

一般化部分採点モデル（Generalized Partial Credit Model：GPCM）[58]はMurakiによって提案された多値型IRTモデルである．このモデルでは，テスト問題iにおいて受検者jが得点kを得る確率を次式で定義する．

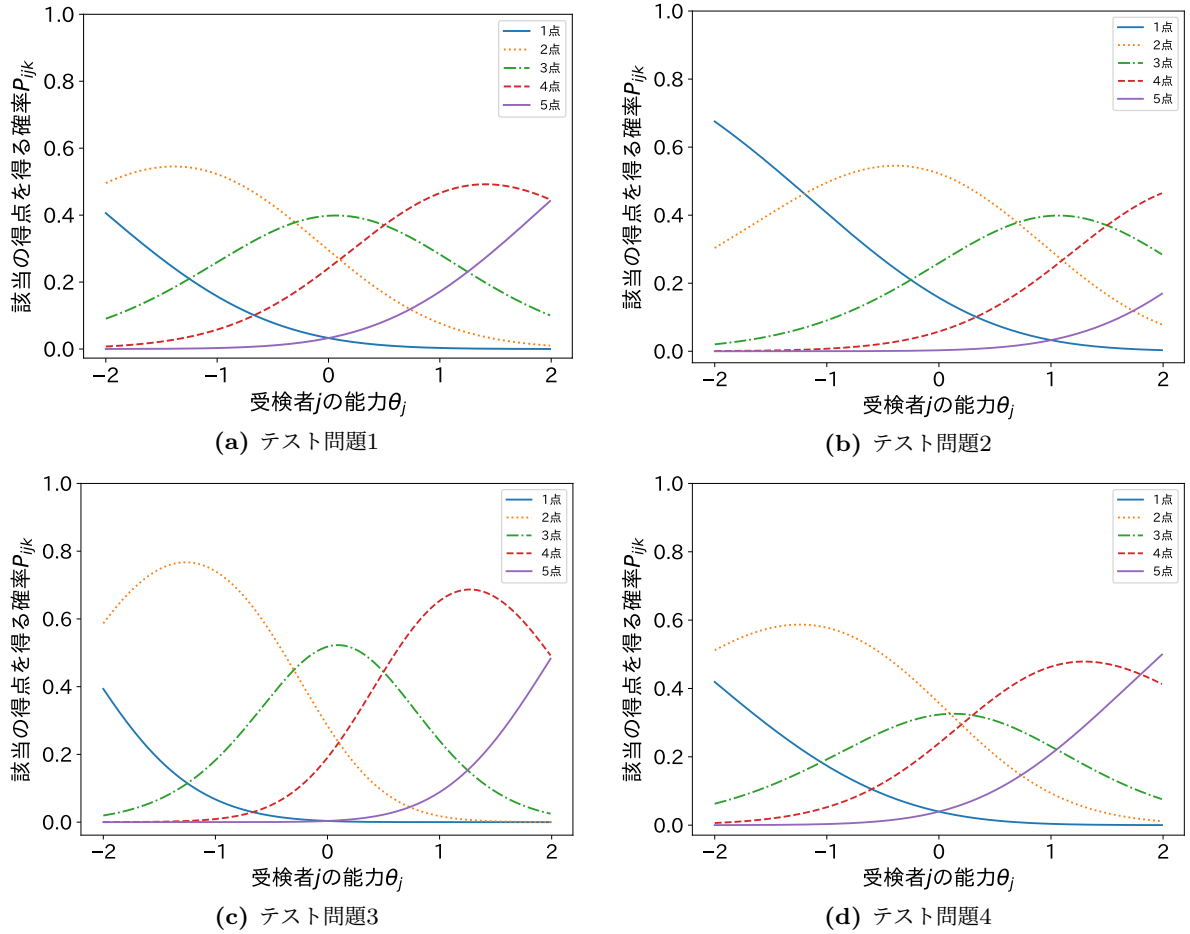
$$P_{ijk} = \frac{\exp \sum_{m=1}^{k} [\alpha_i (\theta_j - \beta_i - d_{im})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_i (\theta_j - \beta_i - d_{im})]} \tag{4.3}$$

ここで，d_{ik}はテスト問題iにおいて得点kを得る困難度を表すパラメータである．ただし，モデルの識別性のために，$d_{i1} = 0$，$\sum_{k=2}^{K} d_{ik} = 0 : \forall i$と制約される．

パラメータの解釈を説明するために，カテゴリ数$K = 5$とし表4.1のパラメータを所与とした4つの問題に対する反応曲線を図4.3に示す．

表4.1: 図4.3で使用したパラメータ

	α_i	β_i	d_{i1}	d_{i2}	d_{i3}	d_{i4}	d_{i5}
テスト問題1	1	0	0	−2.2	−0.3	0.5	2.0
テスト問題2	1	1	0	−2.2	−0.3	0.5	2.0
テスト問題3	2	0	0	−2.2	−0.3	0.5	2.0
テスト問題4	1	0	0	−2.2	0.1	0.3	1.8



(a) テスト問題1　**(b)** テスト問題2

(c) テスト問題3　**(d)** テスト問題4

図4.3: GPCMの反応曲線

これらの図から，受検者の能力が高いほど，低い得点カテゴリへの反応確率が低く，高い得点カテゴリへの反応確率が高くなることが分かる．また，図4.3aと図4.3bを比較すると，困難度パラメータβ_iの値が高いほど反応曲線が右に移動することが読み取れる．これは，困難度β_iの値が大きいほど，能力の低い受検者が高い得点を得にくくなることを表している．図4.3aと図4.3cを比較すると，識別力α_iの値が高いほど能力値θ_jの変化に伴う各得点カテゴリへの反応確率の変化が大きいことが確認できる．これは，識別力α_iの値が大きいほど，能力の微小な違いを精度良く識別できることを表現している．また，得点に対する困難度を表すパラメータd_{ik}については図4.3aと図4.3dの比較から，隣接する得点カテゴリーの値との差が大きくなるほど，その得点を得る確率が高くなる能力値の範囲が大きくなることが確認できる．具体的には，テスト問題1に比べテスト問題4は$d_{i3} - d_{i2}$を大きく，$d_{i4} - d_{i3}$を小さくしたことで，得点2を得る確率が高くなる能力値の範囲を広く，得点3を得る確率が高くなる範囲を狭くなっている．

このモデルを用いることで，多段階の採点データに対してもIRTモデルを適用することができる．他方で，小論文試験などのパフォーマンス評価において，得られる得点が評価者の特性に依存する問題が知られている．このような問題を解決するアプローチの一つとして，評価者特性パラメータを付与したIRTモデルが多数提案されてきた．次節では，代表的なモデルとして多相ラッシュモデル（many-facet Rasch model）を紹介する．

4.4　多相ラッシュモデル

多相ラッシュモデル（many-facet Rasch model）[40]はラッシュモデルに対し課題と受検者以外の要因を表すパラメータを付与したモデルである．ラッシュモデルに評価者の特性を表すパラメータを付与した多相ラッシュモデルでは，評価者rがテスト問題iに対する受検者jの答案を正解とする確率を次式で定義する．

$$P_{ijr} = \frac{\exp(\theta_j - \beta_i - \beta_r)}{1 + \exp(\theta_j - \beta_i - \beta_r)} \tag{4.4}$$

ここで，β_rは評価者rの評価の厳しさを表す．図4.1，図4.2，図4.3のβ_iと同様に，β_rの値が高いほど能力の低い受検者の答案に対して正解を与えにくいことを表している．そのため，β_rの値が高いほど厳しい評価者であると解釈できる．

また，多相ラッシュモデルは多値型IRTモデルとしての定義もされている．その場合，受検者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．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\theta_j - \beta_i - \beta_r - d_m]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\theta_j - \beta_i - \beta_r - d_m]} \tag{4.5}$$

ここで，d_kは得点$k-1$からkに遷移する困難度を表すパラメータである．ただし，パラメータの識別性のために，$\beta_1 = 0$，$d_1 = 0$を仮定する．

このモデルを用いることで，評価者特性を考慮して受検者の能力推定を行うことが可能になった．しかし，このモデルではテスト問題の識別力パラメータα_iが存在しないため，全てのテスト問題で識別力が一定であると仮定している．また，評価者間の識別力に差がないことも仮定している．しかし，現実には，評価者による識別力の違いが生じることが想定される．具体的には，時間変化などにより評価の一貫性が保たれない場合などがある．さらに，各得点カテゴリに対する評価者の評価基準も一定であることを仮定しているが，各得点に対する基準の解釈は評価者ごとに異なることが多い．

以上のように，多相ラッシュモデルは一般には満たされないことが多い強い仮定をモデルに課しているため，それらの仮定を緩和したモデルが近年多数提案されている．

4.5　一般化多相ラッシュモデル

多相ラッシュモデルの拡張モデルの一つである一般化多相ラッシュモデル（Generalized many-facet Rasch model）[37]では，受検者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．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\alpha_r \alpha_i (\theta_j - \beta_i - \beta_{rm})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_r \alpha_i (\theta_j - \beta_i - \beta_{rm})]} \tag{4.6}$$

ここで，α_rは評価者rの一貫性，β_{rk}は得点kに対する評価者rの厳しさを表すパラメータである．ただし，パラメータの識別性のために，$\prod_{i=1}^{I} \alpha_i = 1$，$\sum_{i=1}^{I} \beta_i = 0$，$\beta_{r1} = 0$を仮定する．

このモデルを用いることで，複数評価者によって段階得点で採点されたデータセットからテスト問題や評価者の識別力の違いも踏まえて受検者の能力推定を行うことができる．

以降で説明する項目反応理論を用いた深層学習自動採点モデルでは，このモデルを基礎モデルとして利用している．

第5章

項目反応理論を用いて推定したIRT得点を用いた深層学習自動採点手法

前節で紹介した評価者特性を考慮したIRTモデルと既存の深層学習自動採点モデルを組み合わせて用いることで，評価者バイアスに頑健な自動採点モデルを実現する手法が岡野・宇都[42, 43, 44]によって提案されている．この手法では，得点データ$\boldsymbol{U}$からIRTに基づく得点θ_j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．この手法は3章で紹介したLSTMを用いたモデル[18]とBERTを用いたモデル[25]を採用して設計されているが，様々な深層学習自動採点モデルで利用できるアプローチである．以下で詳細を説明する．

5.1 モデル学習

モデル学習は，IRT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．具体的な手順は次の通りである．

1. 評価者が与える得点データ$\boldsymbol{U}$に前節で紹介した一般化多相ラッシュモデルを適用し，評価者特性の影響を取り除いた各答案e_jの得点θ_jを推定する．ただし，自動採点モデルの学習に利用される一般的なデータでは，問題ごとに受検者集団や評価者集団が異なるため，自動採点モデルの学習は一般に問題ごとに独立に行われる．一般化多相ラッシュモデルを問題ごとに独立に適用する場合，式(4.6)は識別性の制約から次式で表せる．

$$P_{jrk} = \frac{\exp \sum_{m=1}^{k} [\alpha_r(\theta_j - \beta_{rm})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_r(\theta_j - \beta_{rm})]} \tag{5.1}$$

   この場合，採点対象となる答案の数が受検者ごとに一つになるため，θ_jは受検者jの答案の真の得点を表す潜在変数とみなせる．以降では，この潜在得点をIRT得点と呼ぶ．この手順1では，このIRT得点を観測得点$\boldsymbol{U}$から推定している．
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次の平均二乗誤差で定義し，誤差逆伝播法によりパラメータを学習する．

$$MSE(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{1}{J} \sum_{j=1}^{J} (\theta_j - \hat{\theta}_j)^2 \tag{5.2}$$

   ここで$\hat{\theta}_j$は自動採点モデルが予測した答案e_jのIRT得点を表す．なお，既存の自動採点モデル

では出力層にシグモイド関数を用いているため，学習に使用するθ_jの値を$[0,1]$の範囲に変換する必要がある．IRTでは，θは標準正規分布に従うと仮定するため，θの値の99.7%は$[-3,3]$に含まれる．そこで，この研究では，モデル学習を行う前に，$[-3,3]$の範囲を$[0,1]$に線形変換した上で，上記の方法で学習を行っている．なお，変換前の得点が-3以下の場合には0に，3以上の場合には1に変換している．

このようにモデルを学習することで，この手法では，個々の答案を採点する評価者の特性に依存しない自動採点モデルを学習できると期待できる．

5.2　得点予測

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し，得られた値を$[-3,3]$の尺度に線形変換する．
2. $\theta_{j'}$と評価者パラメータを用いて，IRTモデルに基づく期待得点を次式で求める．

$$\hat{U}_{j'} = \frac{1}{R}\sum_{r=1}^{R}\sum_{k=1}^{K} k \cdot P_{j'rk} \tag{5.3}$$

この期待得点は，観測得点$\boldsymbol{U}$と同一の得点尺度となるとともに，個別の評価者の特性に依存しない得点であるため，この値をこの手法による予測得点とする．

なお，本手法の本来の利用方法でないが，個々の評価者が与える得点も予測することができる．具体的には，評価者rが$e_{j'}$に与える得点は次式で予測できる．

$$\hat{U}_{j'r} = \sum_{k=1}^{K} k \cdot P_{j'rk} \tag{5.4}$$

5.3　本手法の課題

この手法を用いることで，評価者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得点の推定には評価者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．

第6章

提案手法

前節で述べた課題を解決するために，本研究では，深層学習自動採点モデルの出力層に評価者特性を考慮した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．これにより，IRT得点の推定に評価者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，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．提案手法のアプローチは様々な深層学習自動採点モデルに適用できるが，本研究では第3.2節で紹介した「LSTMを用いた自動採点モデル」と「BERTを用いた自動採点モデル」を基礎モデルとして利用する．

6.1 提案手法の構成

提案手法の概念図を図6.1と図6.2に示す．提案手法では，従来の深層学習自動採点モデルの最終層の後にIRTモデルに対応する層（以降ではIRT Layerと呼ぶ）を追加する．具体的には，まず，従来の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（5.1）におけるIRT得点θ_jに対応づける．さらに，IRT Layerには，各評価者の特性パラメータα_r，β_{rk}を，学習可能なパラメータとして設定する．IRT Layerは，それらの評価者パラメータとLinear Layerから出力されたIRT得点θ_jを使用して，与

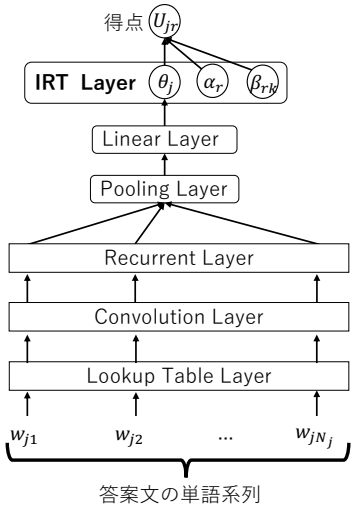


図6.1: LSTMを用いた提案手法の概念図

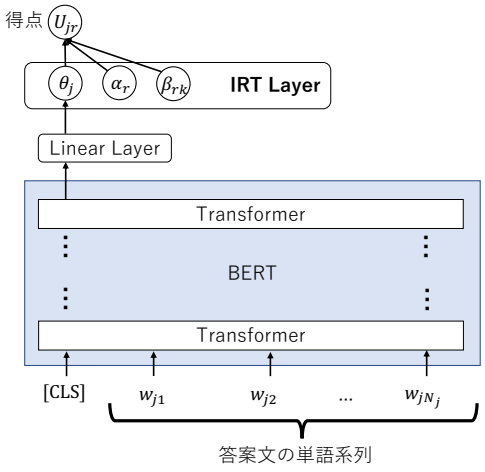


図6.2: BERTを用いた提案手法の概念図

えられた答案に対して評価者rが得点kを与える確率P_{jrk}を式（5.1）で計算する．予測得点には，得られたP_{jrk}から式（5.3）で求めた期待得点を利用する．なお，IRTモデルではθ_jが正規分布に従うと仮定するため，提案手法ではθ_jの平均が0になるように正則化を適用している．

以上のように構成することで，IRT得点θ_jの推定に評価者の与えた観測得点だけでなく答案の文章情報も同時に活用できるようになる．

6.2　提案手法に基づくモデル学習

答案e_jの単語系列と各評価者rによって与えられた得点U_{jr} の集合$\boldsymbol{U}$を入力として，誤差逆伝播法によりパラメータ学習を行う．この際の損失関数は次式で定義された平均二乗誤差を用いる．

$$MSE(\boldsymbol{U}, \hat{\boldsymbol{U}}) = \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{n_j}\sum_{r=1}^{R}(U_{jr} - \hat{U}_{jr})^2 I_{jr}\right] \tag{6.1}$$

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．これは実際に評価者が割り振られているデータのみを使って平均二乗誤差を計算していることを表す．

6.3　提案手法に基づく得点予測

新たな答案$e_{j'}$の得点予測は，その答案に対して評価者rが得点kを与える確率$P_{j'rk}$を学習されたモデルから計算し，IRTモデルに基づく期待得点を次式で求めることで行う．

$$\hat{U}_{j'} = \frac{1}{R}\sum_{r=1}^{R}\sum_{k=1}^{K} k \cdot P_{j'rk} \tag{6.2}$$

なお，5章の従来手法と同様に，提案手法でも個々の評価者が与える得点を予測できる．具体的には，評価者rが$e_{j'}$に与える得点は次式で予測できる．

$$\hat{U}_{j'r} = \sum_{k=1}^{K} k \cdot P_{j'rk} \tag{6.3}$$

第7章

評価実験

7.1 実データ

本実験では，実データとしてAutomated Student Assessment Prize（ASAP）[54]を使用する．ASAPは2012年にヒューレット財団がスポンサーとなって開催されたコンペティションのデータであり，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[6]．ASAPは，8つの異なるトピックに対する小論文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．データ数は，合計12978で，トピックごとの平均は1622.25である．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，新たに評価者を雇用して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38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．ASAPデータセット中の得点データとの相関は，平均で0.675であった．

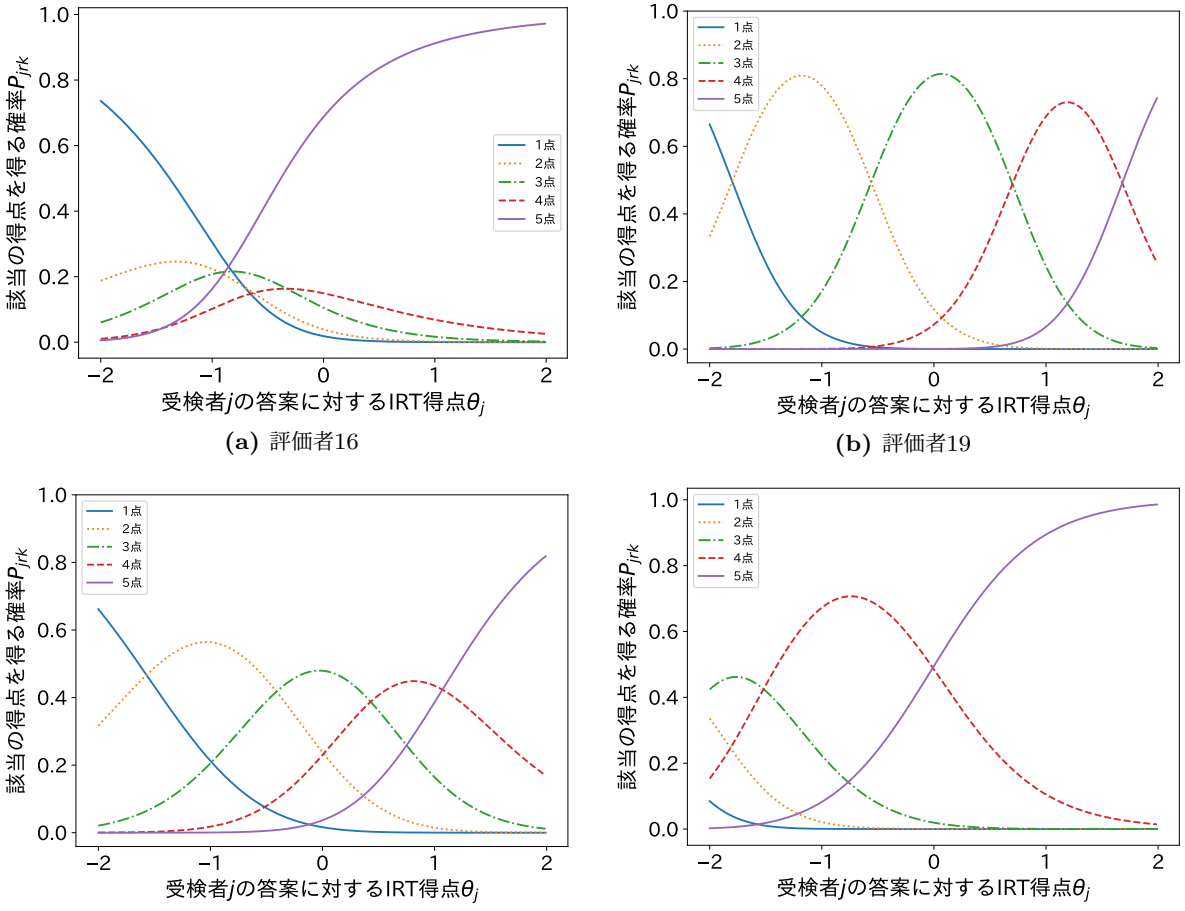
7.2 評価者特性の分析

得られた得点データの基礎統計量として，評価者ごとの観測得点の平均値と分散，各得点カテゴリの出現頻度を表7.1の「基礎統計量」列に示す．これらの値から，評価者ごとに得点の与え方の傾向に差異があることが読み取れる．例えば，評価者16は最高得点と最低得点を相対的に多く使用する傾向があり，評価者19は反対に得点が中心化する傾向がある．また，評価者27は使用得点の偏りが相対的に小さく，幅広く得点を用いる傾向にある．他にも，評価者31は高得点に使用が偏る傾向があり，評価が甘いと解釈できる．このように同じ評価基準を用いて採点を行っても，観測得点に各々の評価者の特性が反映されていることがわかる．

さらに，本データに式（5.1）のIRTモデルを適用して得られた評価者パラメータの推定値を表7.1の「評価者パラメータ」列に示す．これらのパラメータ値も評価者ごとに差異があることが確認できる．また，先ほど具体例として挙げた特徴的な評価者16，19，27，31の反応曲線を図7.1に示す．この反応曲線からも，先ほど言及した特徴を読み取ることができる．

また，評価者間の特性差の有無を客観的に確認するために，IRTで広く利用される適合度統計量で

ある標準化Outfitと標準化Infitを用いた検証を行った．具体的には，評価者間で特性差がないと仮定したIRTモデル（具体的には，式(5.1)のIRTモデルにおいて評価者間で特性値が等しいと制約したモデル）を想定し，そのモデルに基づく標準化Outfitと標準化Infitを評価者ごとに求めた．標準化Outfitと標準化Infitの計算方法については付録Bに示す．標準化Outfit/Infitが$[-2, 2]$の範囲外となった場合，評価者特性を一定と仮定したモデルがその評価者の観測得点データに5%水準で有意に適合しなかったと解釈される[59, 60, 61]．これは，その評価者に固有の特性を考慮しないとデータに適合しないことを意味しており，その評価者は他の評価者と比べて有意に異なる特性を有すると解釈できる．表7.1に各評価者の標準化Outfit/Infitを示した．表7.1の太字は，値が$[-2, 2]$の範囲外であることを意味する．表7.1から，多くの評価者において標準化Outfit/Infitが$[-2, 2]$の範囲外となっていることがわかる．以上から，本実験においては有意に特性の異なる多様な評価者が存在したことが確認できる．このことは，評価者バイアスを考慮することの必要性を示唆している．



(a) 評価者16　**(b)** 評価者19

(c) 評価者27　**(d)** 評価者31

図7.1: 一般化多相ラッシュモデルの反応曲線

表7.1: 評価者ごとの得点データの基礎統計量と評価者パラメータ，および標準化Outfit/Infit

評価者	基礎統計量							評価者パラメータ					標準化	標準化
	平均	標準偏差	各得点の出現頻度					α_r	β_{r2}	β_{r3}	β_{r4}	β_{r5}	Outfit	Infit
			1点	2点	3点	4点	5点							
1	3.318	1.128	8	45	53	55	34	2.95	-1.83	-0.51	0.25	1.13	**-3.84**	**-3.51**
2	3.292	1.147	17	29	57	64	28	1.39	-1.29	-0.83	0.19	1.48	-1.47	-1.22
3	3.528	1.111	10	20	68	51	46	2.68	-1.46	-1.11	0.13	0.79	**-5.27**	**-5.14**
4	3.431	1.145	11	30	59	54	41	1.80	-1.90	-1.02	0.09	0.84	**-5.54**	**-5.51**
5	3.569	0.847	7	6	70	93	19	0.82	-1.18	-3.22	-0.42	2.36	**-3.14**	**-3.68**
6	4.308	0.938	3	11	14	62	105	3.20	-2.35	-1.40	-1.08	0.12	**-4.64**	**-4.76**
7	3.995	1.088	4	22	26	62	81	2.00	-2.56	-1.23	-0.90	0.17	**-5.53**	**-5.58**
8	4.369	0.965	4	8	20	43	120	2.09	-2.27	-1.95	-1.23	-0.52	-1.58	-0.85
9	3.272	1.009	6	39	69	58	23	3.48	-2.14	-0.78	0.31	1.27	**-5.26**	**-5.28**
10	4.492	0.747	0	4	18	51	122	2.63	-2.70	-2.25	-1.38	-0.41	-0.95	-0.32
11	3.395	1.147	11	31	64	48	41	2.31	-1.76	-0.93	0.19	0.83	**-4.41**	**-4.67**
12	4.219	0.941	3	7	31	58	97	1.83	-2.02	-1.96	-0.84	-0.01	**-3.73**	**-3.42**
13	3.185	1.001	6	42	80	44	23	2.06	-2.14	-0.74	0.65	1.37	**-3.30**	**-3.38**
14	3.344	1.067	12	32	49	81	21	3.09	-1.74	-0.89	-0.19	1.34	**-7.19**	**-7.32**
15	3.867	1.153	6	26	31	58	75	2.08	-2.07	-0.92	-0.54	0.37	**-4.50**	**-4.55**
16	3.944	1.458	24	16	21	20	114	1.06	-0.71	-0.93	-0.33	-1.44	1.97	**3.18**
17	3.556	0.97	4	23	62	74	33	2.07	-2.38	-1.32	-0.16	1.09	**-7.80**	**-8.24**
18	3.195	1.169	17	39	56	55	28	3.26	-1.72	-0.63	0.30	1.25	**-5.20**	**-5.16**
19	3.026	0.995	11	47	77	46	14	3.45	-1.80	-0.56	0.70	1.68	**-2.25**	**-2.25**
20	3.738	1.081	3	26	50	56	60	3.57	-2.24	-1.01	-0.12	0.67	**-7.11**	**-7.57**
21	3.779	1.158	6	27	40	53	69	3.37	-1.91	-0.85	-0.20	0.49	**-5.29**	**-5.40**
22	3.651	1.195	10	30	35	63	57	5.16	-1.56	-0.69	-0.17	0.68	**-6.26**	**-6.28**
23	3.872	1.032	5	15	43	69	63	3.22	-2.32	-1.51	-0.49	0.56	**-7.69**	**-8.38**
24	3.551	0.865	1	22	65	84	24	1.74	-2.66	-1.41	-0.21	1.50	**-8.29**	**-8.77**
25	4.292	0.946	2	9	28	47	109	2.07	-2.33	-1.88	-1.01	-0.37	-1.22	-0.81
26	3.544	1.212	14	28	40	64	49	3.55	-1.71	-0.86	-0.19	0.71	**-6.25**	**-6.52**
27	3.148	1.209	17	47	54	44	33	1.71	-1.57	-0.41	0.43	1.07	-0.84	-0.40
28	3.738	1.017	5	14	61	62	53	2.48	-1.96	-1.59	-0.24	0.65	**-7.36**	**-7.94**
29	3.277	1.23	21	31	50	59	34	1.34	-1.26	-0.78	0.03	1.15	-1.41	-1.02
30	3.349	0.918	4	29	76	67	19	1.75	-2.56	-1.19	0.26	1.66	**-6.01**	**-6.38**
31	4.451	0.71	0	4	13	69	109	2.12	-2.65	-2.11	-1.52	-0.01	**-3.34**	**-3.04**
32	3.446	1.008	3	33	66	60	33	2.45	-2.40	-0.88	0.20	1.14	**-6.29**	**-6.40**
33	3.579	1.113	8	27	50	64	46	2.09	-1.86	-0.94	-0.08	0.88	**-5.68**	**-5.93**
34	3.256	1.463	1	35	31	26	51	0.97	-0.72	-0.21	-0.54	0.60	1.40	**2.34**
35	3.733	1.186	5	32	44	43	71	2.21	-2.46	-0.98	-0.15	0.30	**-4.57**	**-4.65**
36	4.005	0.969	4	12	31	80	68	2.31	-2.14	-1.28	-0.59	0.59	**-6.65**	**-6.97**
37	3.077	1.219	18	50	60	33	34	1.43	-1.56	-0.38	0.75	0.81	0.62	0.89
38	4.287	0.751	0	2	29	75	89	2.03	-2.07	-2.54	-1.06	0.24	**-4.83**	**-4.87**

7.3　実験に使用するデータの準備

本研究では，異なる評価者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評価者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，評価者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[62, 63, 64]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評価者による観測得点をそのまま使用し，残りの半分の答案についてのみランダムに選択した1名の評価者の観測得点を残すようにデータセットを作成した．この手順でASAPデータセットから新たに評価者の割り当てが異なる10個の得点データセットを作成し，これらを$\{\boldsymbol{U}'_1, \ldots, \boldsymbol{U}'_{10}\}$とした．

7.4　実験で用いる深層学習自動採点モデル

提案手法は様々な自動採点モデルに適用することが可能である．そこで，本実験についても複数の深層学習自動採点モデルを用いて提案手法に基づく自動採点モデルを作成し，評価を行った．

まず，LSTMに基づくモデルについてはConvolution Layer（CNN）の有無やRecurrent LayerとPooling Layerの構成の違いにより様々なバリエーションが考えられる．表7.2に本実験で作成したLSTMに基づく自動採点モデルを示す．なお，ハイパーパラメータは先行研究と同様に設定した．具体的には，埋め込みベクトルの次元数Gは50，Convolution Layerのウィンドウサイズ数Sは3，出力ベクトルの次元数は50，Recurrent LayerのLSTMの出力ベクトルの次元数は50，Recurrent LayerのDropout率についてはLookup table LayerやConvolution Layerからの入力$\boldsymbol{x_{j(n-1)}}$に対しては0.5，直前のLSTMブロックからの入力$\boldsymbol{h_{j(n-1)}}$に対しては0.1とし，エポック数は30とした．

また，BERTに基づくモデルでは，事前学習されたBASE版BERT[56]を使用した．なお，提案手法では評価者パラメータの推定を自動採点モデルのパラメータ推定と同時に行う必要があるが，BERTではファインチューニング時のパラメータの学習率が小さく設定されているため，事前におおよその値を与えた方が効率的である．そのため，本実験では訓練データからStanによって推定したIRTの評価者パラメータを初期値として与えた．ファインチューニングのエポック数は3とした．

表7.2: 実験に用いたLSTMに基づくモデルのバリエーション

	Convolution Layer	Recurrent Layer	Pooling Layer
LSTM(MoT)	–	LSTM	Mean over Time
CNN-LSTM(MoT)	✓	LSTM	Mean over Time
2layer LSTM(MoT)	–	2layer LSTM	Mean over Time
LSTM(Last)	–	LSTM	Last pooling
CNN-LSTM(Last)	✓	LSTM	Last pooling
2layer LSTM(Last)	–	2layer LSTM	Last pooling
Bidiretional LSTM	–	Bidirectional LSTM	Last pooling

7.5　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，評価者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[42, 43, 44]．そのため，評価者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. データセット$\boldsymbol{U}'_n$の4/5を訓練データとして，モデルの学習を行った．
2. データセットの残り1/5をテストデータとして，IRT得点を推定した．
3. 以上を訓練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$\boldsymbol{\theta_{n'}}$との平均絶対誤差（Mean Absolute Error：MAE），平均平方二乗誤差（Root Mean Square Error：RMSE），相関係数（Correlation：Cor），決定係数（Coefficient of determination：R^2）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．また，これらの評価指標の詳細は付録Cに示す．

これらを前節で示した複数の深層学習自動採点モデルを利用した提案手法で行った．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（5.1）のIRTモデルでも実施した．具体的には，訓練データからIRTモデルを用いて評価者パラメータを推定し，テストデータに対しては評価者パラメータを所与として期待事後確率（Expected a posterior：EAP）推定でIRT得点を推定した．

実験結果を表7.3に示す．表7.3から，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がわかる．このことから，提案手法がIRT得点推定の精度改善に有効であったことが確認できた．

表7.3: IRT得点の推定精度評価実験の結果

		MAE	RMSE	Cor	R^2
提案手法	LSTM(MoT)	0.13	0.18	0.95	0.89
	CNN+LSTM(MoT)	0.16	0.22	0.92	0.84
	2L-LSTM(MoT)	0.14	0.20	0.94	0.88
	LSTM(Last)	0.15	0.22	0.92	0.83
	CNN+LSTM(Last)	0.19	0.26	0.85	0.69
	2L-LSTM(Last)	0.17	0.24	0.90	0.79
	bidirectional(Last)	0.18	0.25	0.89	0.76
	BERT	0.14	0.19	0.97	0.94
既存IRTモデル		0.35	0.49	0.74	0.47

7.6　自動採点モデルの頑健性評価

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，以下の手順に基づいて評価実験を行った．

提案手法の実験手順

1. データセット$\boldsymbol{U}'_n$の4/5を訓練データとして，モデルの学習を行った．
2. データセットの残り1/5の小論文答案をテストデータとして，式（6.2）で予測得点を求めた．
3. 以上を訓練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，全ての受検者の予測得点$\hat{\boldsymbol{U}}_{\boldsymbol{n}}$を求めた．

以上の手順を，前節と同様に$n = \{1, \ldots, 10\}$について行い，n番目のデータセットから求めた$\hat{\boldsymbol{U}}_{\boldsymbol{n}}$と$n'$番目の得点データセットから推定した$\hat{\boldsymbol{U}}_{\boldsymbol{n'}}$とのMAE，RMSE，Cor，$R^2$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．また，自動採点タスクの研究において，一致度合いを評価する指標として用いられる，カッパ係数（Cohen's Kappa：Kappa），重み付きカッパ係数（Linear Weighted Kappa：LWK），2次の重み付きカッパ係数（Quadratic Weighted Kappa：QWK），正解率（Accuracy）についても求めた．ただし，これらの指標は予測得点が離散値になる必要があるため，予測得点を四捨五入して得られた整数値の得点を用いて，これらの指標を計算した．

また，比較のために，5章で説明したIRT得点を用いて自動採点モデルの学習を行う手法（IRT得点を用いた手法）と自動採点モデルの一般的な使用法である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）でも同様の実験を行った．なお，観測得点を用いた手法では，訓練データの平均点を利用して自動採点モデルの学習を行った．また，本実験では実験結果に手法間で有意な差があるか確認するために，多重比較検定も行った．

実験結果を表7.4に示す．表では，提案手法とIRT得点を用いた自動採点手法，観測得点を用いた自動採点手法の中でモデルの性能が最も高い結果を太字で示している．多重比較検定のp値は各評価指標の下に括弧書きで示した．括弧書きの中の左の値はIRT得点を用いた手法とのp値を表し，右の値は観測得点を用いた手法とのp値を表す．表7.4から，ほぼ全ての条件においてIRTモデルを用いた手法（提案手法とIRT得点を用いた手法）が単純な自動採点手法よりも高い性能を示したことが分かる．このことから，評価者特性を考慮したIRTモデルを用いることで，評価者バイアスに頑健な自動採点モデルが実現できたことが分かる．

また，IRTを用いた手法同士を比較すると，相関や一致度に対応する指標ではほとんどの場合で提案手法が高い精度を示している．一方で，MAEやRMSEなどの誤差指標では，IRT得点を使用した従来手法の方が，わずかに高い性能を示している．これらの違いを確認するために，各手法で学習したモデルの予測得点の分布を比較する．ここでは，先ほどの実験で性能の高かったLSTM(MoT)に基づく提案手法とIRT得点を用いた手法について，予測得点$\hat{\boldsymbol{U}}_{\boldsymbol{1}}$，$\hat{\boldsymbol{U}}_{\boldsymbol{2}}$の分布を求め，それぞれの結果を図7.2，図7.3に示す．これらの得点分布を比較すると，IRT得点を用いた手法では最低点付近（1点から1.5点）や最高点付近（4.8点から5点）などの極端な得点が出力されにくい傾向があることが分かる．これはIRT得点を用いた手法で活性化関数として用いているシグモイド関数の特性により，極端

に高い得点や低い得点を予測得点として出しにくくなっていることが原因であると考えられる．これにより，予測得点の範囲が縮小し，MAEやRMSEなどの誤差が小さくなったと考えられる．本実験では，訓練データを変えた時の予測得点同士を比較しているため，予測得点の分布が縮小するだけで，MAEやRMSEなどの誤差指標は小さくなってしまう．したがって，本実験では，相関係数や一致度の指標を重視すべきである．上述の通り，それらの指標においては，多くの場合で提案手法が高い性能を示していたことが分かる．

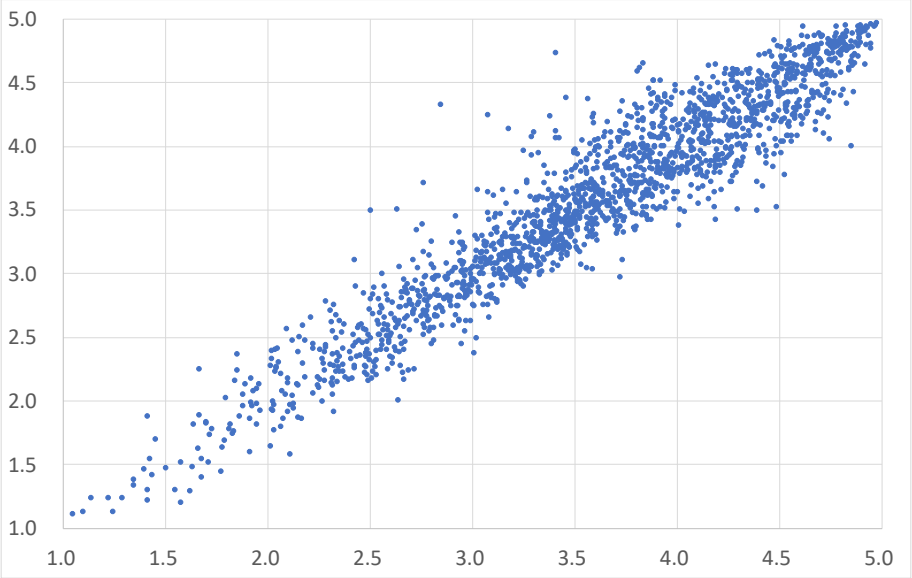


図7.2: 提案手法の予測得点分布

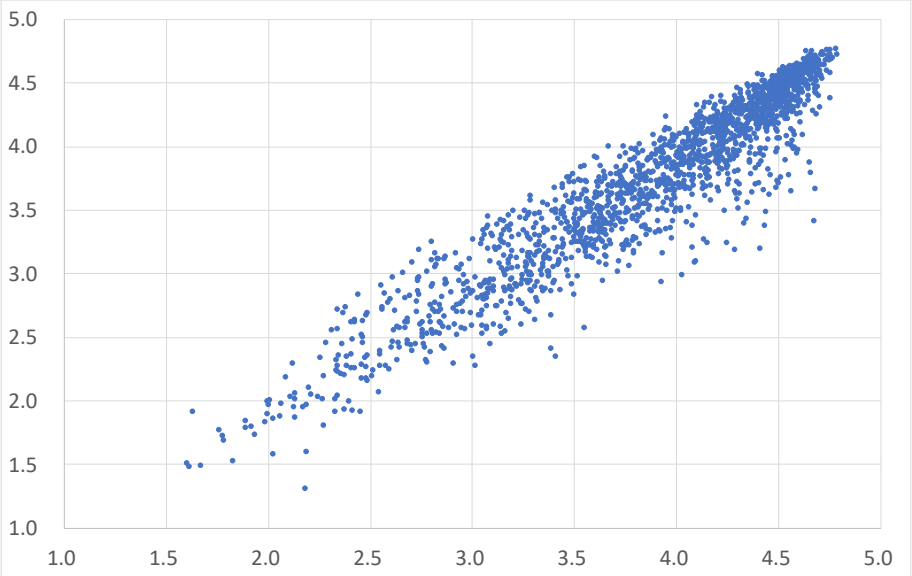


図7.3: IRT得点を用いた手法の予測得点分布

† 縦軸は$\hat{U}_1$，横軸は$\hat{U}_2$を用いて学習した際の予測得点を表す．

表7.4: 頑健性評価実験の結果

(a) MAE

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	0.21 (0.79,0.00)	**0.20** (—,0.00)	0.30 (—,—)
CNN+LSTM(MoT)	0.29 (0.03,0.00)	**0.28** (—,0.00)	0.49 (—,—)
2L-LSTM(MoT)	0.23 (0.00,0.00)	**0.20** (—,0.00)	0.32 (—,—)
LSTM(Last)	**0.24** (0.80,0.00)	**0.24** (—,0.00)	0.41 (—,—)
CNN+LSTM (Last)	0.37 (0.00,0.00)	**0.30** (—,0.00)	0.56 (—,—)
2L-LSTM(Last)	0.25 (0.00,0.00)	**0.23** (—,0.00)	0.38 (—,—)
bidirectional(Last)	0.28 (0.06,0.00)	**0.27** (—,0.00)	0.46 (—,—)
BERT	0.25 (0.00,0.22)	**0.21** (—,0.00)	0.27 (—,—)

(b) RMSE

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.27** (1.00,0.00)	**0.27** (—,0.00)	0.41 (—,—)
CNN+LSTM(MoT)	0.39 (0.04,0.00)	**0.37** (—,0.00)	0.66 (—,—)
2L-LSTM(MoT)	0.29 (0.00,0.00)	**0.26** (—,0.00)	0.43 (—,—)
LSTM(Last)	**0.31** (1.00,0.00)	**0.31** (—,0.00)	0.55 (—,—)
CNN+LSTM(Last)	0.48 (0.00,0.00)	**0.39** (—,0.00)	0.75 (—,—)
2L-LSTM(Last)	0.32 (0.02,0.00)	**0.30** (—,0.00)	0.49 (—,—)
bidirectional(Last)	0.36 (0.17,0.00)	**0.34** (—,0.00)	0.63 (—,—)
BERT	0.32 (0.00,0.15)	**0.26** (—,0.00)	0.34 (—,—)

(c) Cor

	提案モデル	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.94** (0.00,0.00)	0.93 (—,0.00)	0.90 (—,—)
CNN+LSTM(MoT)	**0.90** (0.00,0.00)	0.86 (—,0.00)	0.78 (—,—)
2L-LSTM(MoT)	**0.94** (0.00,0.00)	0.93 (—,0.00)	0.89 (—,—)
LSTM(Last)	**0.92** (1.00,0.00)	0.91 (—,0.00)	0.82 (—,—)
CNN+LSTM(Last)	0.82 (0.00,0.00)	**0.85** (—,0.00)	0.71 (—,—)
2L-LSTM(Last)	**0.91** (1.00,0.00)	0.90 (—,0.00)	0.84 (—,—)
bidirectional(Last)	**0.89** (1.00,0.00)	**0.89** (—,0.00)	0.78 (—,—)
BERT	**0.93** (0.10,0.22)	0.92 (—,1.00)	0.92 (—,—)

(d) R^2

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.88** (0.00,0.00)	0.83 (—,0.00)	0.78 (—,—)
CNN+LSTM(MoT)	**0.78** (0.00,0.00)	0.70 (—,0.00)	0.50 (—,—)
2L-LSTM(MoT)	**0.87** (0.00,0.00)	0.83 (—,0.00)	0.75 (—,—)
LSTM(Last)	**0.82** (0.59,0.00)	0.80 (—,0.00)	0.61 (—,—)
CNN+LSTM(Last)	0.62 (0.26,0.00)	**0.65** (—,0.00)	0.35 (—,—)
2L-LSTM(Last)	**0.80** (0.25,0.00)	0.77 (—,0.00)	0.61 (—,—)
bidirectional(Last)	**0.76** (1.00,0.00)	0.75 (—,0.00)	0.49 (—,—)
BERT	**0.84** (0.58,0.61)	0.83 (—,1.00)	0.82 (—,—)

(e) Kappa

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.68** (0.00,0.00)	0.62 (—,0.00)	0.54 (—,—)
CNN+LSTM(MoT)	**0.56** (0.00,0.00)	0.53 (—,0.00)	0.36 (—,—)
2L-LSTM(MoT)	**0.65** (0.23,0.00)	0.64 (—,0.00)	0.51 (—,—)
LSTM(Last)	**0.60** (0.05,0.00)	0.56 (—,0.00)	0.41 (—,—)
CNN+LSTM(Last)	0.47 (0.25,0.00)	**0.49** (—,0.00)	0.30 (—,—)
2L-LSTM(Last)	**0.59** (0.22,0.00)	0.57 (—,0.00)	0.39 (—,—)
bidirectional(Last)	**0.55** (0.00,0.00)	0.50 (—,0.00)	0.35 (—,—)
BERT	0.59 (0.97,0.39)	**0.60** (—,0.02)	0.57 (—,—)

(f) LWK

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.77** (0.00,0.00)	0.69 (—,0.33)	0.68 (—,—)
CNN+LSTM(MoT)	**0.68** (0.00,0.00)	0.61 (—,0.00)	0.54 (—,—)
2L-LSTM(MoT)	**0.75** (0.00,0.00)	0.71 (—,0.00)	0.66 (—,—)
LSTM(Last)	**0.69** (0.01,0.00)	0.65 (—,0.00)	0.58 (—,—)
CNN+LSTM(Last)	**0.59** (1.00,0.00)	0.58 (—,0.00)	0.47 (—,—)
2L-LSTM(Last)	**0.68** (0.00,0.00)	0.64 (—,0.00)	0.55 (—,—)
bidirectional(Last)	**0.66** (0.00,0.00)	0.61 (—,0.00)	0.52 (—,—)
BERT	**0.71** (0.00,0.74)	0.67 (—,0.04)	0.69 (—,—)

(g) QWK

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.86** (0.00,0.00)	0.78 (—,0.00)	0.81 (—,—)
CNN+LSTM(MoT)	**0.81** (0.00,0.00)	0.72 (—,0.24)	0.70 (—,—)
2L-LSTM(MoT)	**0.85** (0.00,0.00)	0.79 (—,1.00)	0.80 (—,—)
LSTM(Last)	**0.80** (0.00,0.00)	0.76 (—,0.30)	0.73 (—,—)
CNN+LSTM(Last)	**0.72** (0.02,0.00)	0.69 (—,0.00)	0.63 (—,—)
2L-LSTM(Last)	**0.78** (0.00,0.00)	0.73 (—,1.00)	0.72 (—,—)
bidirectional(Last)	**0.77** (0.00,0.00)	0.73 (—,0.00)	0.68 (—,—)
BERT	**0.82** (0.00,1.00)	0.76 (—,0.00)	**0.82** (—,—)

(h) Accuracy

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.79** (0.25,0.00)	0.77 (—,0.00)	0.67 (—,—)
CNN+LSTM(MoT)	0.70 (0.00,0.00)	**0.72** (—,0.00)	0.53 (—,—)
2L-LSTM(MoT)	0.76 (0.00,0.00)	**0.79** (—,0.00)	0.66 (—,—)
LSTM(Last)	**0.74** (0.64,0.00)	0.73 (—,0.00)	0.57 (—,—)
CNN+LSTM(Last)	0.65 (0.00,0.00)	**0.69** (—,0.00)	0.48 (—,—)
2L-LSTM(Last)	0.73 (0.00,0.00)	**0.76** (—,0.00)	0.58 (—,—)
bidirectional(Last)	**0.71** (0.23,0.00)	0.69 (—,0.00)	0.53 (—,—)
BERT	0.72 (0.00,0.16)	**0.78** (—,0.00)	0.70 (—,—)

† 多重比較検定の結果を括弧書きで記載した．括弧書きの中の左の値はIRT得点を用いた手法とのp値を表し，右の値は観測得点を用いた手法とのp値を表す．

7.7　観測得点の予測精度評価

本節では，提案手法を用いることで，各評価者が与えた観測得点U_{jr}の予測精度が改善するかを評価する．具体的には，観測得点U_{jr}と自動採点モデルによって推定された予測得点$\hat{U}_{jr}$との一致度を，前節と同様の指標を用いて評価した．

本実験では，モデル学習についてはどの手法についても7.6節と同様に行い，得点予測については，期待得点ではなく各評価者が与えた得点を推定するようにした．具体的には，提案手法については手順2において式（6.2）で期待得点を求める代わりに各評価者の得点を式（6.3）で求め，実際の観測得点との一致度を評価した．IRT得点を用いた手法についても提案手法と同様の変更を行った．観測得点を用いた手法では，評価者ごとの予測得点は推定できないため，自動採点モデルの予測得点と各評価者が実際に与えた観測得点との一致度を評価した．

実験結果を表7.5に示す．表では，提案手法とIRT得点を用いた手法，観測得点を用いた手法の中で性能が最も高い手法の結果を太字で示している．多重比較検定のp値は各評価指標の下に括弧書きで示した．括弧書きの中の左の値はIRT得点を用いた手法とのp値を表し，右の値は観測得点を用いた手法とのp値を表す．表7.5から，ほぼ全ての条件において，自動採点モデルのみを用いる観測得点を用いた手法と比べて，IRT得点を用いた手法や提案手法が高い性能を示したことが分かる．これは，IRTによって補正された得点は文章の質を素点そのものよりも正確に反映しているため，提案手法では文章と得点の関係がより適切に学習できたことが要因と考えられる．このことから，IRTを用いることで，自動採点モデルの予測得点の頑健性向上に加え，評価者得点の予測にも有効であることが確認できた．

また，提案手法とIRT得点を用いた手法を比較すると，どの深層学習自動採点モデルを用いるかによって最高精度の手法が異なっている．具体的には，BERTモデルとPooling LayerにMoTを採用したLSTMモデルを用いた際には，全ての場合で提案手法が高い性能を示している．一方で，Pooling LayerにLast poolingを用いたLSTMモデルの場合には，IRT得点を用いた手法がやや高い性能を示す傾向がある．BERTモデルとMoTを用いたLSTMモデルは7.5節においてIRT得点θの推定精度が比較的高かったものである．このことから，IRT得点θの推定精度が高いモデルにおいては提案手法による評価者得点の推定精度が高くなることが示唆される．

表7.5: 観測得点の予測精度評価実験の結果

(a) MAE

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.56** (0.25,0.01)	0.57 (—,0.01)	0.68 (—,—)
CNN+LSTM(MoT)	**0.59** (0.39,0.01)	0.60 (—,0.01)	0.73 (—,—)
2L-LSTM(MoT)	**0.55** (0.01,0.01)	0.57 (—,0.01)	0.69 (—,—)
LSTM(Last)	0.59 (0.15,0.01)	**0.58** (—,0.01)	0.72 (—,—)
CNN+LSTM(Last)	0.68 (0.01,0.01)	**0.62** (—,0.01)	0.78 (—,—)
2L-LSTM(Last)	**0.60** (1.00,0.01)	**0.60** (—,0.01)	0.71 (—,—)
bidirectional(Last)	0.62 (0.01,0.01)	**0.59** (—,0.01)	0.74 (—,—)
BERT	**0.55** (0.01,0.01)	0.58 (—,0.01)	0.66 (—,—)

(b) RMSE

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.73** (0.01,0.01)	0.75 (—,0.01)	0.87 (—,—)
CNN+LSTM(MoT)	**0.77** (0.70,0.01)	0.78 (—,0.01)	0.94 (—,—)
2L-LSTM(MoT)	**0.72** (0.01,0.01)	0.75 (—,0.01)	0.87 (—,—)
LSTM(Last)	0.77 (0.39,0.01)	**0.76** (—,0.01)	0.90 (—,—)
CNN+LSTM(Last)	0.87 (0.01,0.01)	**0.80** (—,0.01)	0.99 (—,—)
2L-LSTM(Last)	**0.77** (0.48,0.01)	0.78 (—,0.01)	0.89 (—,—)
bidirectional(Last)	0.79 (0.02,0.01)	**0.77** (—,0.01)	0.93 (—,—)
BERT	**0.72** (0.01,0.01)	0.75 (—,0.01)	0.83 (—,—)

(c) Cor

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.78** (0.08,0.01)	0.77 (—,0.01)	0.68 (—,—)
CNN+LSTM(MoT)	**0.75** (0.58,0.01)	0.74 (—,0.01)	0.63 (—,—)
2L-LSTM(MoT)	**0.78** (0.11,0.01)	0.77 (—,0.01)	0.67 (—,—)
LSTM(Last)	0.75 (0.25,0.01)	**0.76** (—,0.01)	0.64 (—,—)
CNN+LSTM(Last)	0.66 (0.01,0.01)	**0.73** (—,0.01)	0.58 (—,—)
2L-LSTM(Last)	**0.75** (1.00,0.01)	**0.75** (—,0.01)	0.65 (—,—)
bidirectional(Last)	0.73 (0.01,0.01)	**0.76** (—,0.01)	0.62 (—,—)
BERT	**0.78** (0.01,0.01)	0.77 (—,0.01)	0.70 (—,—)

(d) R^2

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.60** (0.01,0.01)	0.58 (—,0.01)	0.42 (—,—)
CNN+LSTM(MoT)	**0.56** (0.70,0.01)	0.54 (—,0.01)	0.33 (—,—)
2L-LSTM(MoT)	**0.60** (0.01,0.01)	0.58 (—,0.01)	0.43 (—,—)
LSTM(Last)	0.56 (0.39,0.01)	**0.57** (—,0.01)	0.38 (—,—)
CNN+LSTM(Last)	0.42 (0.01,0.01)	**0.52** (—,0.01)	0.26 (—,—)
2L-LSTM(Last)	**0.55** (0.48,0.01)	0.54 (—,0.01)	0.41 (—,—)
bidirectional (Last)	0.52 (0.02,0.01)	**0.55** (—,0.01)	0.34 (—,—)
BERT	**0.61** (0.01,0.01)	0.57 (—,0.01)	0.48 (—,—)

(e) Kappa

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.38** (0.15,0.01)	0.37 (—,0.01)	0.27 (—,—)
CNN+LSTM(MoT)	**0.36** (0.06,0.01)	0.33 (—,0.01)	0.23 (—,—)
2L-LSTM(MoT)	**0.40** (0.01,0.01)	0.37 (—,0.01)	0.26 (—,—)
LSTM(Last)	0.35 (0.39,0.01)	**0.36** (—,0.01)	0.25 (—,—)
CNN+LSTM(Last)	0.27 (0.03,0.01)	**0.31** (—,0.01)	0.20 (—,—)
2L-LSTM(Last)	**0.34** (1.00,0.01)	**0.34** (—,0.01)	0.23 (—,—)
bidirectional(Last)	0.32 (0.01,0.01)	**0.35** (—,0.01)	0.23 (—,—)
BERT	**0.40** (0.01,0.01)	0.36 (—,0.01)	0.27 (—,—)

(f) LWK

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.56** (0.01,0.01)	0.54 (—,0.01)	0.45 (—,—)
CNN+LSTM(MoT)	**0.54** (0.02,0.01)	0.51 (—,0.01)	0.42 (—,—)
2L-LSTM(MoT)	**0.57** (0.01,0.01)	0.54 (—,0.01)	0.45 (—,—)
LSTM(Last)	0.52 (0.58,0.01)	**0.53** (—,0.01)	0.43 (—,—)
CNN+LSTM(Last)	0.45 (0.03,0.01)	**0.49** (—,0.01)	0.38 (—,—)
2L-LSTM(Last)	**0.52** (1.00,0.01)	**0.51** (—,0.01)	0.42 (—,—)
bidirectional(Last)	0.50 (0.01,0.01)	**0.53** (—,0.01)	0.41 (—,—)
BERT	**0.58** (0.01,0.01)	0.53 (—,0.01)	0.46 (—,—)

(g) QWK

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.73** (0.01,0.01)	0.70 (—,0.01)	0.62 (—,—)
CNN+LSTM(MoT)	**0.71** (0.01,0.01)	0.67 (—,0.01)	0.59 (—,—)
2L-LSTM(MoT)	**0.73** (0.01,0.01)	0.70 (—,0.01)	0.62 (—,—)
LSTM(Last)	**0.69** (0.58,0.01)	**0.69** (—,0.01)	0.60 (—,—)
CNN+LSTM(Last)	0.61 (0.01,0.01)	**0.66** (—,0.01)	0.55 (—,—)
2L-LSTM(Last)	**0.68** (0.58,0.01)	0.67 (—,0.01)	0.59 (—,—)
bidirectional(Last)	0.67 (0.03,0.01)	**0.69** (—,0.01)	0.58 (—,—)
BERT	**0.74** (0.01,0.01)	0.69 (—,0.01)	0.64 (—,—)

(h) Accuracy

	提案手法	IRT得点を用いた手法	観測得点を用いた手法
LSTM(MoT)	**0.54** (1.00,0.01)	**0.54** (—,0.01)	0.46 (—,—)
CNN+LSTM(MoT)	**0.52** (0.39,0.01)	0.51 (—,0.01)	0.43 (—,—)
2L-LSTM(MoT)	**0.55** (0.02,0.01)	0.54 (—,0.01)	0.46 (—,—)
LSTM(Last)	0.52 (0.39,0.01)	**0.53** (—,0.01)	0.44 (—,—)
CNN+LSTM(Last)	0.46 (0.03,0.01)	**0.49** (—,0.01)	0.40 (—,—)
2L-LSTM(Last)	0.51 (0.97,0.01)	**0.52** (—,0.01)	0.44 (—,—)
bidirectional(Last)	0.50 (0.03,0.01)	**0.53** (—,0.01)	0.43 (—,—)
BERT	**0.56** (0.01,0.01)	0.53 (—,0.02)	0.46 (—,—)

† 多重比較検定の結果を括弧書きで記載した．括弧書きの中の左の値はIRT得点を用いた手法とのp値を表し，右の値は観測得点を用いた手法とのp値を表す．

第8章

まとめ

深層学習を用いた自動採点手法では，訓練データセット中の各答案に対する得点が評価者特性に依存する場合，そのようなデータを元に学習したモデルにも評価者特性の影響が反映され，得点予測の性能が低下するという問題がある．この問題を解決するために，IRTを用いて評価者バイアスの影響を取り除いたIRT得点を推定し，それを自動採点モデルに学習させる手法が提案されている．本研究では，このアプローチを改善し，自動採点モデルにIRTを表す層を加えた新たな自動採点手法を提案した．また，実データ実験から，提案手法では評価者バイアスに頑健な得点予測を実現でき，得点予測の精度自体も改善できることが確認できた．

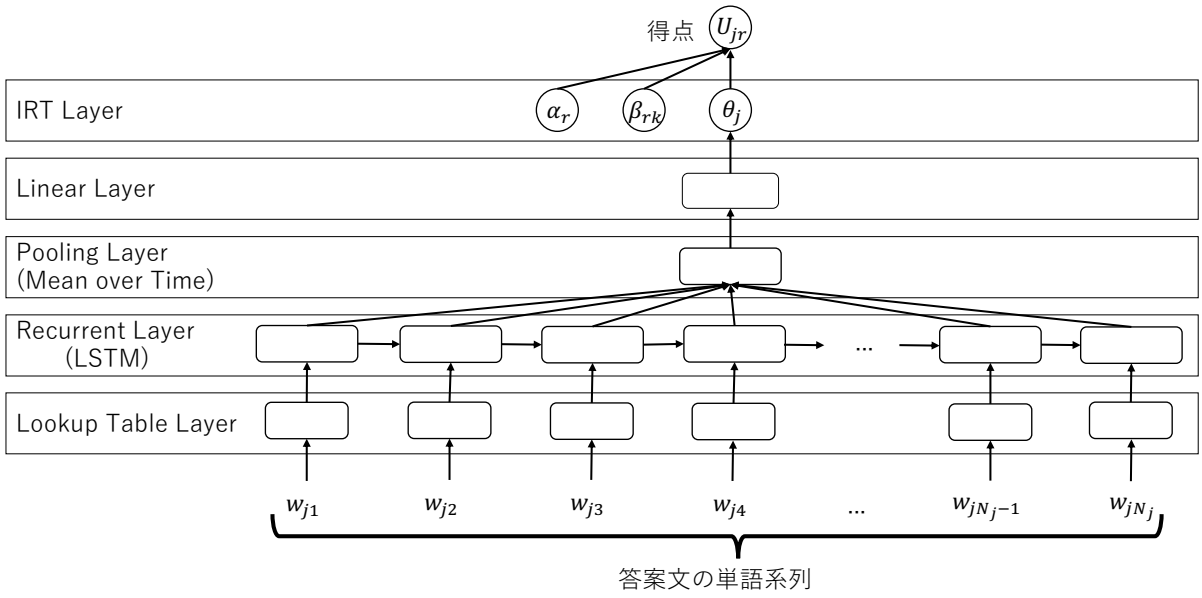
なお，本研究では，多数の受検者の答案を多数の評価者で分担して評価するような状況を想定した．一方で学校現場では，学期末試験に代表されるように，多数の受検者が共通の試験を受験するが，その評価はクラスごとに講義担当者1人で行うという場面がある．このような場合には，IRTを適用しても評価者特性を考慮した得点は理論上推定できないため[62, 63, 64]，提案手法を利用しても評価者バイアスを取り除いた自動採点は実現できない．一方で，一部の答案を複数のクラス担当者で採点するなど，採点デザインを工夫すれば，IRTを適切に適用できるため，提案手法も有効に機能すると考えられる．今後は，このような状況も含めた多様なデータセットへの手法適用や様々な自動採点モデルへの組み込みを通して，提案手法の有効性を確認していきたい．

付録A

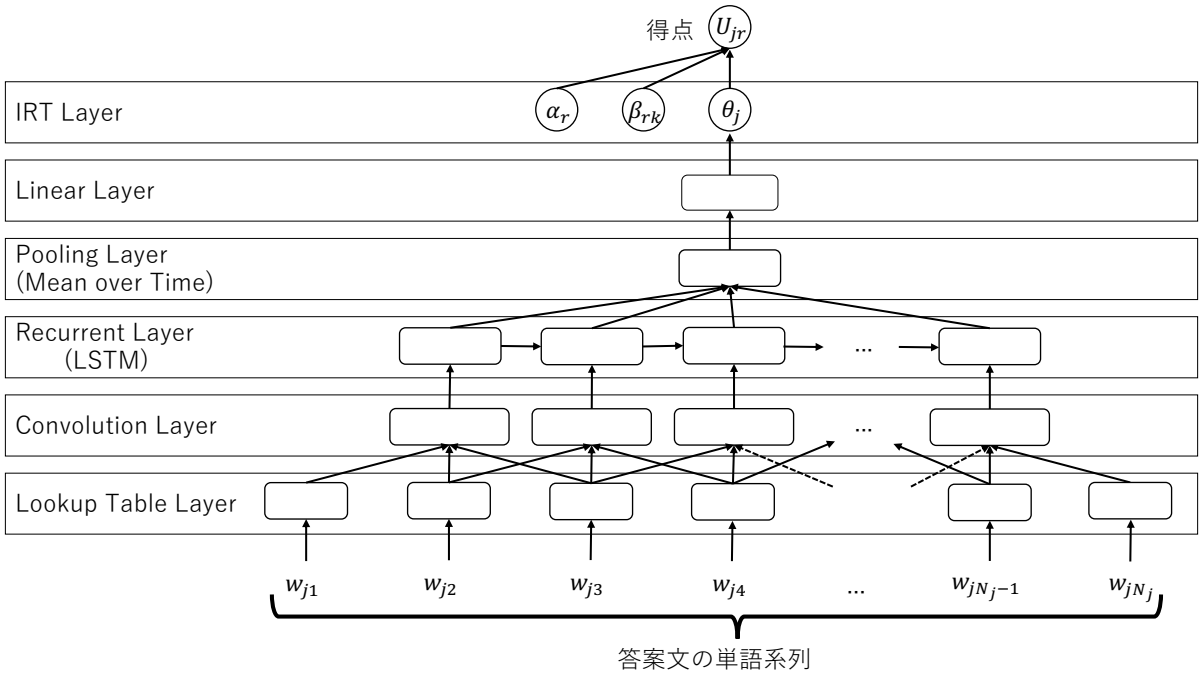
各AESモデルに対する提案手法の概念図

本研究で提案した手法は様々な自動採点モデルに適用することが可能である．そのため，実データ実験では，7.4節で示したように，様々な設定の自動採点モデルを用いて提案手法を検証した．本節では，提案手法を様々なLSTMベースの自動採点モデルに適用した際の概念図と，それらの自動採点モデルの構成に対する補足を述べる．まず，各自動採点モデルを用いた際の概念図を図A.1から図A.7に示す．

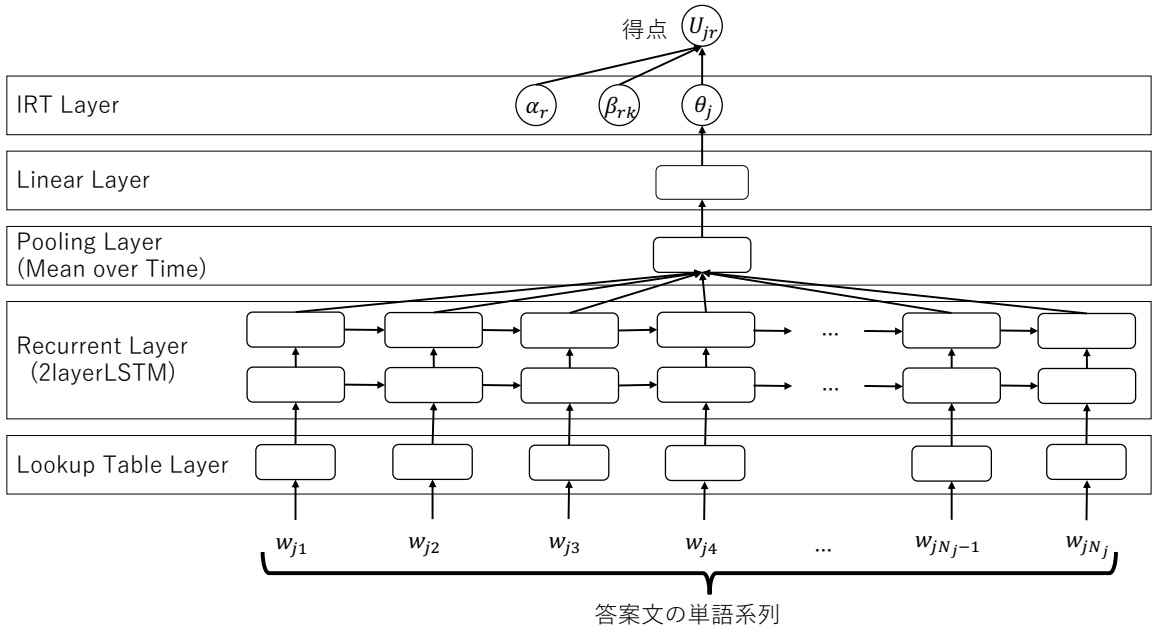
LSTMに基づく自動採点モデルのRecurrent Layerには図A.1や図A.4のように単方向のLSTMを用いることが多いが，他にも様々な構成法が知られている．例えば，図A.3や図A.6のように単方向のLSTMを2層で適用する手法（2layer LSTM）がある．この場合，LSTMブロックの各出力を次のLSTM Layerの各ブロックの入力とする．また，図A.7 のように，双方向のLSTMを用意し，これらのLSTM層の最終ブロックの出力を結合して用いる手法（Bidirectional LSTM）も存在する．この手法では，得点推定のために前後の文脈を用いることができるため，単語の予測などのタスクにおいては有効とされている．また，Pooling LayerについてはMean over Timeを採用した際は3.2で示したように，LSTMブロックの各出力の平均ベクトルを計算し，Last poolingを採用した際は，最後のLSTMブロックの出力のみを用いる．



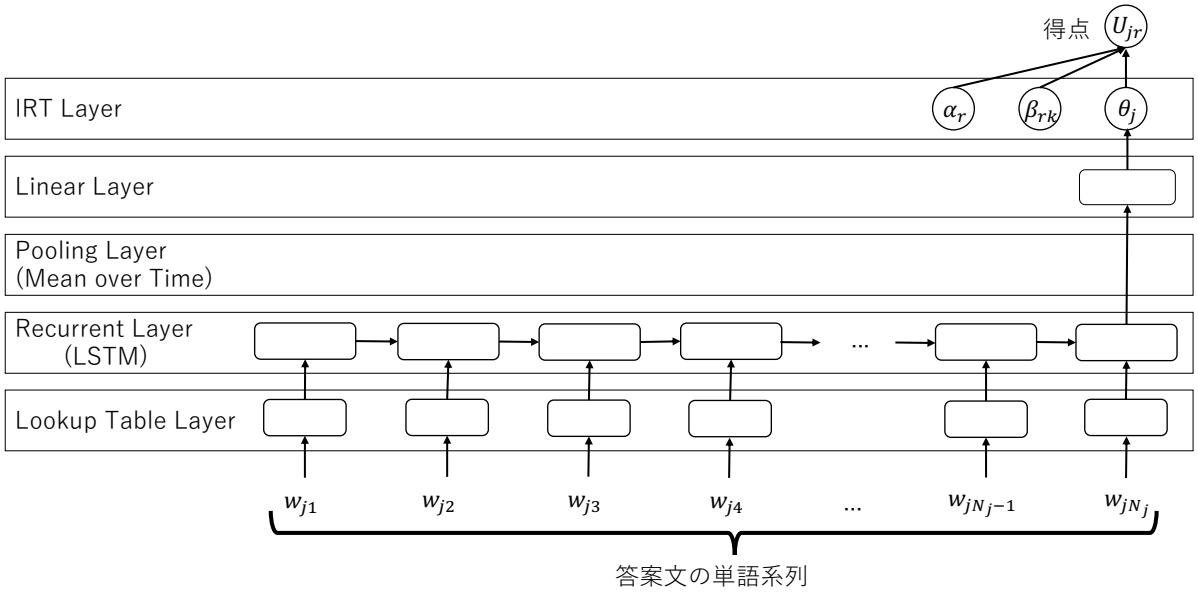
図A.1: 自動採点モデルにLSTM(MoT)を用いた際の提案手法の概念図



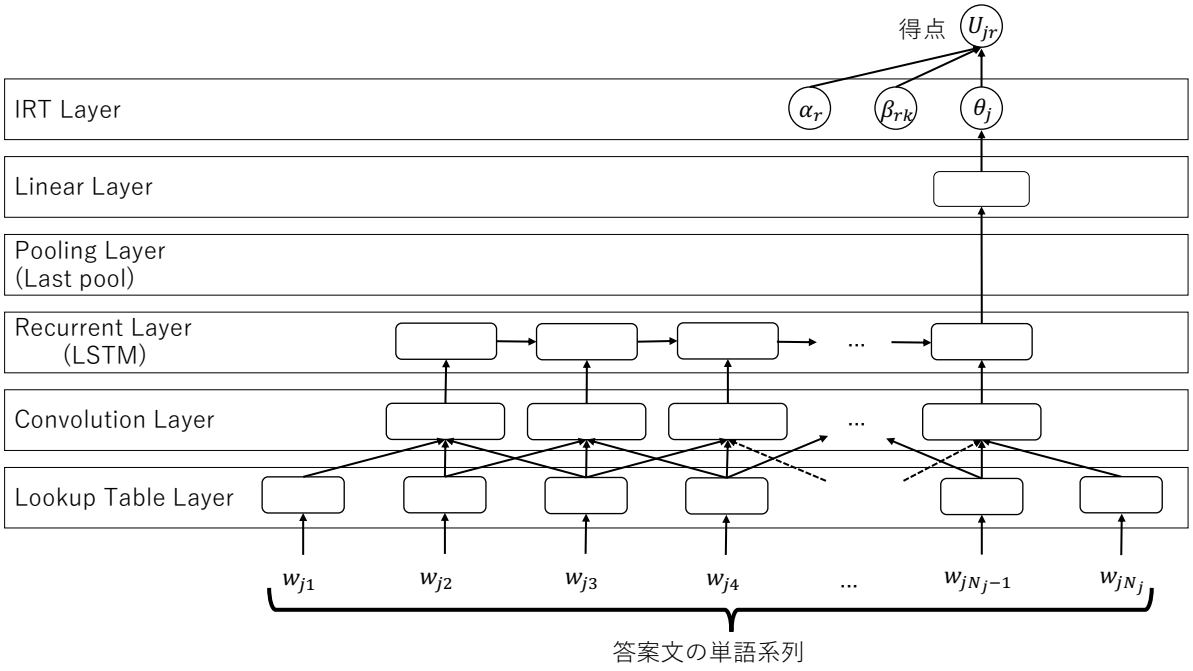
図A.2: 自動採点モデルにCNN-LSTM(MoT)を用いた際の提案手法の概念図



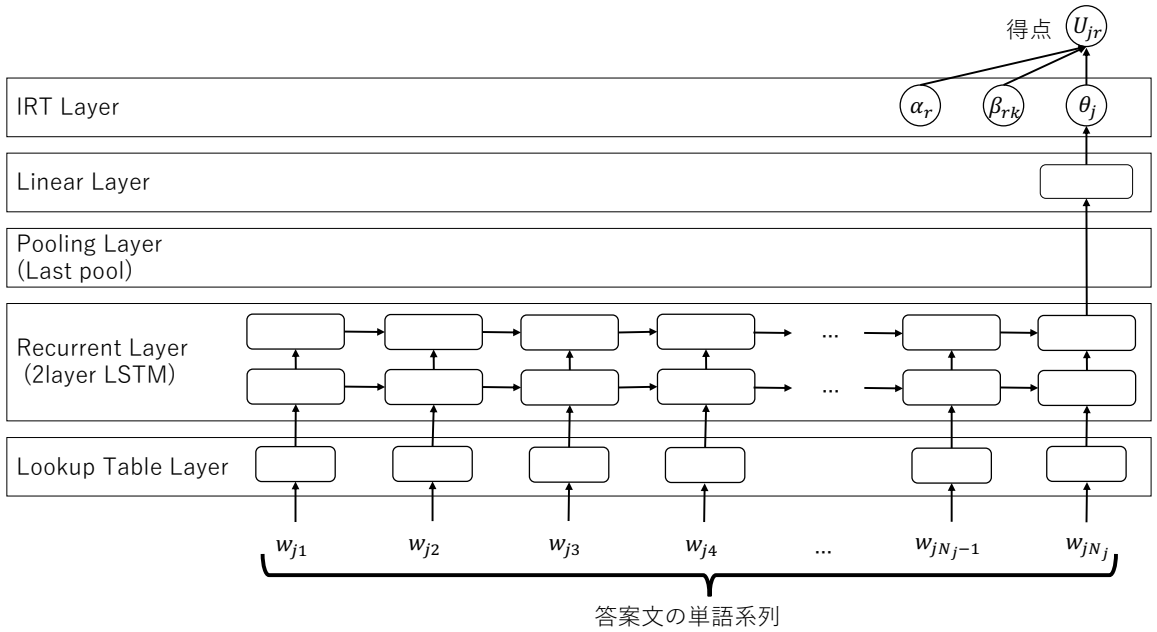
図A.3: 自動採点モデルに2layer LSTM(MoT)を用いた際の提案手法の概念図



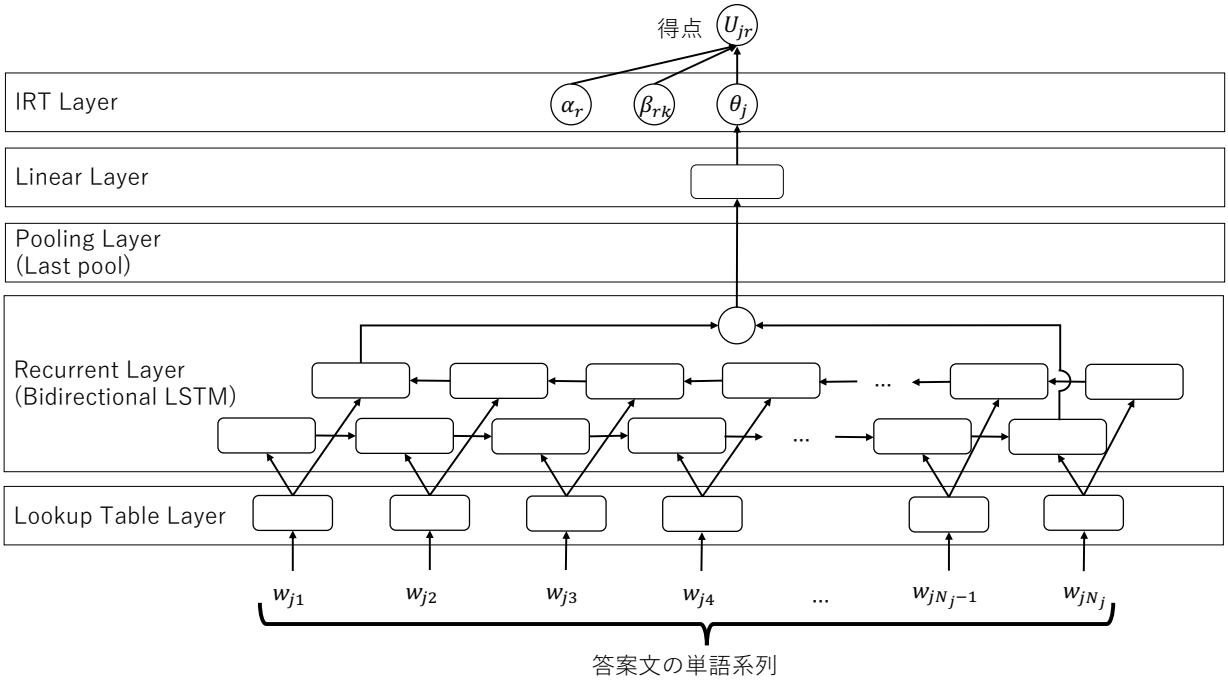
図A.4: 自動採点モデルにLSTM(Last)を用いた際の提案手法の概念図



図A.5: 自動採点モデルにCNN-LSTM(Last)を用いた際の提案手法の概念図



図A.6: 自動採点モデルに2layer LSTM(Last)を用いた際の提案手法の概念図



図A.7: 自動採点モデルにBidirectional LSTMを用いた際の提案手法の概念図

付録B

標準化Outfit/Infit

ここでは，7.2節で導入した評価者ごとの標準化Outfitと標準化Infitについて説明する．7.2節で説明したように，本研究では，評価者間で特性が等しいと仮定したIRTモデルを用いて，これらの指標を求めた．ここで，評価者特性が等しいと仮定したIRTモデルは，式(5.1)において全ての評価者$r \in \mathcal{R}$に対して$\alpha_r = 1$, $\beta_r = 0$, $d_{rm} = d_m$と制約したモデルであり，次式と一致する．

$$P_{jk} = \frac{\exp \sum_{m=1}^{k} (\theta_j - d_m)}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} (\theta_j - d_m)} \tag{B.1}$$

このIRTモデルを用いて，評価者rのOutfitは

$$\frac{\sum_{j=1}^{N} \zeta_{jr} \delta_{jr}^2}{\sum_{j=1}^{N} \zeta_{jr}}, \tag{B.2}$$

評価者rのInfitは

$$\frac{\sum_{j=1}^{N} \rho_j \zeta_{jr} \delta_{jr}^2}{\sum_{j=1}^{N} \rho_j \zeta_{jr}} \tag{B.3}$$

で，それぞれ定義できる．ここで，ζ_{jr}はU_{jr}が欠測データのときに0，そうでないときに1を取るダミー変数であり，δ_{jr}とω_jはそれぞれ次式で定義される．

$$\delta_{jr} = \frac{U_{jr} - \sum_{k=1}^{K} k P_{jk}}{\sqrt{\rho_j}}, \tag{B.4}$$

$$\rho_j = \sum_{k=1}^{K} (k - \sum_{k'=1}^{K} k' P_{jk'})^2 P_{jk} \tag{B.5}$$

標準化Outfitと標準化Infitは，上記で求められたOutfit値とInfit値にWilson-Hilferty変換[65, 66]を適用することで得られる[59, 60, 61].

付録C

評価指標について

本節では, 実データ実験で利用したカッパ係数, 重み付きカッパ係数, 2次重み付きカッパ係数, 平均絶対誤差, 平均平方二乗誤差, 相関係数を定義する．具体的には, 2つの得点集合$\mathcal{X} = \{X_1, \ldots, X_N\}$, $\mathcal{Y} = \{Y_1, \ldots, Y_N\}$を想定して，これらの指標を定義する．ここで，$N$はデータ数，$X_n \in \mathcal{X}, Y_n \in \mathcal{Y}$は各集合の$n$番目の得点を表し，個々の得点は1から$K$の$K$段階で与えられるとする．なお，定義の一般性を担保するため，本節で使用する記号はこれまでの章や節と意味が異なることに注意されたい．

C.1　カッパ係数

カッパ係数は以下の式で計算される．

$$1 - \frac{1 - P^o}{1 - P^e} \tag{C.1}$$

ここで，P^oは実際の一致確率，P^eは偶然の一致確率と呼ばれ，それぞれ以下で計算される．

$$P^o = \sum_{z=1}^{K} \frac{n_{zz}}{N} \tag{C.2}$$

$$P^e = \sum_{z=1}^{K} \frac{\sum_{y=1}^{K} n_{zy}}{N} \frac{\sum_{x=1}^{K} n_{xz}}{N} \tag{C.3}$$

ここで，$I(X_n = x, Y_n = y)$を$X_n = x$かつ$Y_n = y$のときに1，それ以外のとき0を返す関数とすると，$n_{xy} = \sum_{n=1}^{N} I(X_n = x, Y_n = y)$で定義される．

C.2　重み付きカッパ係数

重み付きカッパ係数は以下の式で計算される．

$$1 - \frac{\sum_{x=1}^{K} \sum_{y=1}^{K} w_{xy} P^o_{xy}}{\sum_{x=1}^{K} \sum_{y=1}^{K} w_{xy} P^e_{xy}} \tag{C.4}$$

ここで，P^o_{xy}とP^e_{xy}はそれぞれ次式で計算される．

$$P^o_{xy} = \frac{n_{xy}}{N} \tag{C.5}$$

$$P^e_{xy} = \frac{\sum_{z=1}^{K} n_{xz}}{N} \frac{\sum_{z=1}^{K} n_{zy}}{N} \tag{C.6}$$

またw_{xy}は次式で計算される．

$$w_{xy} = |x - y| \tag{C.7}$$

C.3　2次の重み付きカッパ係数

2次の重み付きカッパ係数は，式 (C.4)におけるw_{xy}を次式に変更することで求められる．

$$w_{xy} = (x - y)^2 \tag{C.8}$$

C.4　平均絶対誤差

平均絶対誤差は得点の差の絶対値の平均であり，次式で計算される．

$$\frac{1}{N}\sum_{n=1}^{N}|X_n - Y_n| \tag{C.9}$$

C.5　平均平方二乗誤差

平均平方二乗誤差は次式で計算される．

$$\sqrt{\frac{1}{N}\sum_{n=1}^{N}(X_n - Y_n)^2} \tag{C.10}$$

C.6　相関係数

相関係数は次式で計算される．

$$\frac{\sum_{n=1}^{N}(X_n - \mu^{\mathcal{X}})(Y_n - \mu^{\mathcal{Y}})}{\sqrt{\sum_{n=1}^{N}(X_n - \mu^{\mathcal{X}})^2 \sum_{n=1}^{N}(Y_n - \mu^{\mathcal{Y}})^2}} \tag{C.11}$$

ここで，$\mu^{\mathcal{X}}$，$\mu^{\mathcal{Y}}$は得点集合$\mathcal{X}$，$\mathcal{Y}$の平均値を表す．

研究成果

本研究に関する成果は以下の通りである．

査読付き論文誌

- 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.J104-D, No.8，2021.
- Masaki Uto, Masashi Okano, “Learning Automated Essay Scoring Models Using Item Response Theory-Based Scores to Decrease Effects of Rater Biases,”IEEE Transactions on Learning Technologies，2022.

国際会議

- Masaki Uto, Masashi Okano, “Robust neural automated essay scoring using item response theory,”International Conference on Artificial Intelligence in Education (AIED)，2020.（**Best paper runner-up award 受賞**）

研究会・ワークショップ

- 岡野将士，宇都雅輝，“評価者バイアスを考慮した小論文自動採点手法”，情報処理学会 第241回自然言語処理研究会，2019.
- 岡野将士，宇都雅輝，“アノテータの採点バイアスに頑健な小論文自動採点手法”，言語処理学会 第26回年次大会，2020.
- 岡野将士，宇都雅輝，“評価者バイアスに頑健な小論文自動採点手法”，人工知能学会 第88回先進的学習科学と工学研究会，2020.（**若手奨励賞受賞**）
- 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，2021.
- 岡野将士，宇都雅輝，“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”，日本テスト学会 第19回大会，2021.（**大会発表賞受賞**）
- 岡野将士，宇都雅輝，“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”，言語処理学会 第28回年次大会，2022.

謝辞

本研究を進めるに当たり，指導教官の宇都雅輝准教授からは丁寧かつ熱心なご指導を賜りました．厚く感謝を申し上げます．また様々な助言を頂きました川野秀一准教授，研究に関して様々な指導や議論をしていただいた先輩方，研究室の皆様に深く感謝申し上げます．

参考文献

[1] 中央教育審議会，“幼稚園、小学校、中学校、高等学校及び特別支援学校の学習指導要領等の改善及び必要な方策等について,” https://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/__icsFiles/afieldfile/2017/01/10/1380902_0.pdf, Dec. 2016. (Accessed on 10/04/2021).

[2] 池田央，テストの科学：試験にかかわるすべての人に，オンデマンド版，教育測定検定所，2007.

[3] 河原宜央，“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.

[4] 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, March 2019.

[5] 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，国立大学入学者選抜研究連絡協議会，Feb. 2016.

[6] Z. Ke and V. Ng, “Automated essay scoring: A survey of the state of the art,” Proceedings of the International Joint Conference on Artificial Intelligence, pp.6300–6308, International Joint Conferences on Artificial Intelligence Organization, July 2019.

[7] M. Uto, “A review of deep-neural automated essay scoring models,” Behaviormetrika, vol.48, pp.1–26, July 2021.

[8] P. Lagakis and S. Demetriadis, “Automated essay scoring: A review of the field,” Proceedings of the International Conference on Computer, Information and Telecommunication Systems, pp.1–6, Nov. 2021.

[9] D. Ramesh and S.K. Sanampudi, “An automated essay scoring systems: a systematic literature review,” Artificial Intelligence Review, pp.1–33, Sept. 2021.

[10] Y. Attali and J. Burstein, “Automated essay scoring with e-rater® v.2,” Journal of Technology, Learning and Assessment, vol.4, no.3, pp.1–30, Feb. 2006.

[11] H. Chen and B. He, “Automated essay scoring by maximizing human-machine agreement,” Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.1741–1752, Association for Computational Linguistics, Oct. 2013.

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

[13] M. Dascalu, W. Westera, S. Ruseti, S. Trausan-Matu, and H. Kurvers, “ReaderBench

Learns Dutch: Building a Comprehensive Automated Essay Scoring System for Dutch Language," Proceedings of the Conference on Artificial Intelligence in Education, vol.10331, pp.52–63, Springer International Publishing, June 2017.

[14] P. Hastings, S. Hughes, and M.A. Britt, "Active learning for improving machine learning of student explanatory essays," Proceedings of the Conference on Artificial Intelligence in Education, vol.10947, pp.140–153, Springer International Publishing, June 2018.

[15] L. Yao, Shelby J.Haberman, and M. Zhang, "Prediction of writing true scores in automated scoring of essays by best linear predictors and penalized best linear predictors," ETS Research Report Series, vol.2019, no.1, pp.1–27, April 2019.

[16] D. Alikaniotis, H. Yannakoudakis, and M. Rei, "Automatic text scoring using neural networks," Proceedings of the Annual Meeting of the Association for Computational Linguistics, vol.1, pp.715–725, Association for Computational Linguistics, Aug. 2016.

[17] J. Liu, Y. Xu, and L. Zhao, "Automated essay scoring based on two-stage learning," CoRR,arXiv, Dec. 2019.

[18] K. Taghipour and H.T. Ng, "A neural approach to automated essay scoring," Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.1882–1891, Association for Computational Linguistics, Nov. 2016.

[19] F. Dong and Y. Zhang, "Automatic features for essay scoring – an empirical study," Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.1072–1077, Association for Computational Linguistics, Nov. 2016.

[20] Y. Tay, M.C. Phan, L.A. Tuan, and S.C. Hui, "Skipflow: incorporating neural coherence features for end-to-end automatic text scoring," Proceedings of the Conference on Association for the Advancement of Artificial Intelligence, pp.5948–5955, AAAI Press, Nov. 2018.

[21] F. Dong, Y. Zhang, and J. Yang, "Attention-based recurrent convolutional neural network for automatic essay scoring," Proceedings of the Conference on Computational Natural Language Learning, pp.153–162, Association for Computational Linguistics, Aug. 2017.

[22] Y. Farag, H. Yannakoudakis, and T. Briscoe, "Neural automated essay scoring and coherence modeling for adversarially crafted input," Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol.1, pp.263–271, Association for Computational Linguistics, June 2018.

[23] 水本智也，磯部順子，関根聡，乾健太郎，"採点項目に基づく国語記述式答案の自動採点," 言語処理学会第24回年次大会 発表論文集，pp.552–555，言語処理学会，March 2018.

[24] 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 Annual Meeting of the Association for Computational Linguistics, vol.1, pp.1088–1097, Association for Computational Linguistics, July 2018.

[25] P.U. Rodriguez, A. Jafari, and C.M. Ormerod, "Language models and automated essay scoring," arXiv, Sept. 2019.

[26] T. Liu, W. Ding, Z. Wang, J. Tang, G.Y. Huang, and Z. Liu, "Automatic short answer

grading via multiway attention networks," Proceedings of the Conference on Artificial Intelligence in Education, vol.11626, pp.169–173, Springer International Publishing, June 2019.

[27] C. Sung, T.I. Dhamecha, and N. Mukhi, "Improving short answer grading using transformer-based pre-training," Proceedings of the Conference on Artificial Intelligence in Education, vol.11625, pp.469–481, Springer International Publishing, June 2019.

[28] J. Lun, J. Zhu, Y. Tang, and M. Yang, "Multiple data augmentation strategies for improving performance on automatic short answer scoring," Proceedings of the Conference on Association for the Advancement of Artificial Intelligence, vol.34, pp.13446–13453, AAAI Press, April 2020.

[29] 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.

[30] E. Amorim, M. Cançado, and A. Veloso, "Automated essay scoring in the presence of biased ratings," 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.

[31] S.A. Wind, E.W. Wolfe, G.E. Jr., P. Foltz, and M. 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, no.1, pp.27–49, Nov. 2018.

[32] J. Huang, L. Qu, R. Jia, and B. Zhao, "O2u-net: A simple noisy label detection approach for deep neural networks," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp.3325–3333, Oct. 2019.

[33] S. Li, S. Ge, Y. Hua, C. Zhang, H. Wen, T. Liu, and W. Wang, "Coupled-view deep classifier learning from multiple noisy annotators," Proceedings of the AAAI Conference on Artificial Intelligence, vol.34, pp.4667–4674, April 2020.

[34] M. Uto and M. Ueno, "Empirical comparison of item response theory models with rater's parameters," Heliyon, Elsevier, vol.4, no.5, pp.1–32, May 2018.

[35] 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，Jan. 2018.

[36] M. Uto and M. Ueno, "Item response theory without restriction of equal interval scale for rater's score," Proceedings of the Conference on Artificial Intelligence in Education, vol.10948, pp.363–368, Springer International Publishing, June 2018.

[37] M. Uto and M. 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.

[38] R.J. Patz, B.W. Junker, M.S. Johnson, and L.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, Jan. 2002.

[39] R.J. Patz and B.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, Dec. 1999.

[40] J.M. Linacre, Many-faceted Rasch Measurement, MESA Press, Jan. 1989.

[41] 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，March 2010.

[42] M. Uto and M. Okano, “Robust neural automated essay scoring using item response theory,” Proceedings of the Conference on Artificial Intelligence in Education, pp.549–561, Springer International Publishing, July 2020.

[43] 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，Aug. 2021.

[44] M. Uto and M. Okano, “Learning automated essay scoring models using item response theory-based scores to decrease effects of rater biases”. IEEE Transactions on Learning Technologies，2022.

[45] I. Aomi, E. Tsutsumi, M. Uto, and M. Ueno, “Integration of automated essay scoring models using item response theory,” Proceedings of the Conference on Artificial Intelligence in Education, pp.54–59, Springer International Publishing, June 2021.

[46] 青見樹，堤瑛美子，宇都雅輝，植野真臣，“項目反応理論による小論文自動採点機のモデル平均,” 電子情報通信学会論文誌. D, 情報・システム，vol.104，no.11，pp.784–795，Nov. 2021.

[47] J. Burstein, K. Kukich, S. Wolff, C. Lu, and M. Chodorow, “Computer analysis of essays,” The annual meeting of the National Council of Measurement in Education, pp.1–13, 1998.

[48] P. Vik, H. Diana, J. John, T. James, A. Nate, and K. John, “Ease,” `https://github.com/edx/ease`. (Accessed on 12/20/2021).

[49] Educational Testing Service, “Criterion,” `https://etsjapan.jp/criterion/index.html`. (Accessed on 12/20/2021).

[50] R. Long, “A review of ETS’s Criterion online writing program for student compositions,” Language Teaching, vol.37, pp.11–18, May 2013.

[51] J. Burstein, M. Chodorow, and C. Leacock, “Automated essay evaluation: The criterion online writing service,” AI Magazine, vol.25, no.3, p.27, Sep. 2004.

[52] 石岡恒憲，“小論文およびエッセイの自動評価採点における研究動向,” 人工知能学会誌，vol.23，no.1，pp.17–24，Jan. 2008.

[53] H.M. Breland, R.J. Jones, L. Jenkins, M. Paynter, J. Pollack, and Y.F. Fong, “The college board vocabulary study,” ETS Research Report Series, vol.1994, no.1, pp.i–51, June 1994.

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

[55] M. Uto, Y. Xie, and M. Ueno, “Neural automated essay scoring incorporating hand-crafted features,” Proceedings of the International Conference on Computational Linguistics, pp.6077–6088, International Committee on Computational Linguistics, Dec. 2020.

[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, pp.4171–4186, Association for Computational Linguistics, June 2019.

[57] G. Rasch, Probabilistic models for some intelligence and attainment tests., ERIC, 1993.

[58] E. Muraki, “A generalized partial credit model,” pp.153–164, Springer New York, 1997.

[59] W. Benjamin and M. Geofferey, “Rating scale analysis. Rasch measurement.,” Journal of the American Statistical Association, vol.78, pp.94–105, June 1983.

[60] J. Linacre, “What do Infit and Outfit, mean-square and standardized mean?,” Rasch Measurement Transactions, vol.16, no.2, p.878, Oct. 2002.

[61] T. Eckes, Introduction to Many-Facet Rasch Measurement, Peter Lang Publishing Incorporated, Jan. 2015.

[62] 宇都雅輝, “評価者特性パラメータを付与した項目反応モデルに基づくパフォーマンステストの等化精度,” 電子情報通信学会論文誌. D, 情報・システム, vol.101, no.6, pp.895–905, June 2018.

[63] M. İlhan, “A comparison of the results of many-facet Rasch analyses based on crossed and judge pair designs,” Journal of Educational Sciences: Theory & Practice, vol.16, pp.579–601, April 2016.

[64] M. Uto, “Accuracy of performance-test linking based on a many-facet rasch model,” Behavior Research Methods, vol.53, pp.1440–1454, Nov. 2020.

[65] E.B. Wilson and M.M. Hilferty, “The distribution of chi-square,” Proceedings of the National Academy of Sciences of the United States of America, vol.17, no.12, pp.684–688, Dec. 1931.

[66] M. Schulz, “The standardization of mean-squares,” Rasch Measurement Transactions, vol.16, no.2, p.879, Oct. 2002.