第19回日本統計学会春季集会
企画セッション「項目反応理論の世界」

# 項目反応理論と深層学習を統合した記述式回答自動採点技術

電気通信大学 大学院情報理工学研究科 准教授
宇都 雅輝

2025年3月8日

# 目次

1. 項目反応理論の概要
2. 記述式回答自動採点技術の概要
3. 項目反応理論と記述式回答自動の融合技術
   - 評価者バイアスに頑健な自動採点手法
   - 自動採点機のアンサンブル手法
   - 複数の評価観点で得点を予測する自動採点手法
4. 質疑応答

# 目次

1. **項目反応理論の概要**
2. 記述式回答自動採点技術の概要
3. 項目反応理論と記述式回答自動の融合技術
   - 評価者バイアスに頑健な自動採点手法
   - 自動採点機のアンサンブル手法
   - 複数の評価観点で得点を予測する自動採点手法
4. 質疑応答

# 項目反応理論（Item Response Theory: IRT）

学力テストやアンケートなどの分析に広く利用される心理・教育測定のための統計数理手法

テスト項目への受検者の反応データから，IRTモデルと呼ばれる数理モデルを用いて受検者の能力スコアと項目の特性値（難易度や識別力など）を分離して推定

| | 項目1 | 項目2 | 項目3 | 項目4 | 能力スコア |
|---|---|---|---|---|---|
| 受検者1 | × | | × | ○ | 0.2 |
| 受検者2 | | × | × | ○ | -0.8 |
| 受検者3 | ○ | ○ | | × | 1.1 |
| 項目難易度 | 1.1 | 0.1 | 0.2 | -0.2 | |
| 項目識別力 | 1.0 | 0.7 | 1.4 | 0.3 | |

# 2パラメータロジスティックモデル

正誤反応データを扱う代表的なIRTモデルの一つ

受検者 $j$ がテスト項目 $i$ に正答する確率を
項目の特性を表すパラメータ と
受検者の能力を表すパラメータ の関数として定式化

正誤反応データ

項目 $i$ の特性パラメータ
- 識別力
- 困難度

$$p(x_{ij}=1)=\frac{\exp(\alpha_i(\theta_j-b_i))}{1+\exp(\alpha_i(\theta_j-b_i))}$$

受検者$j$の能力スコア

パラメータは最尤推定やベイズ推定で算出

[対数尤度]

$$\log L=\sum_j\sum_i \log p(x_{ij})$$

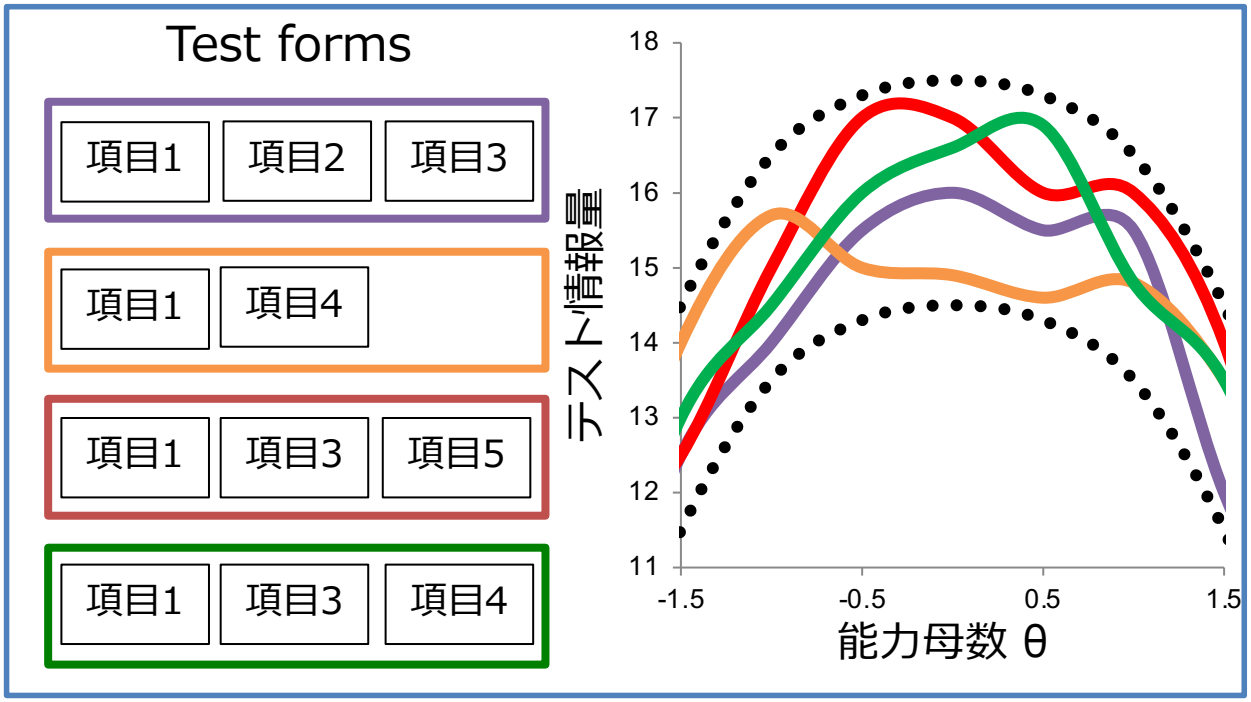
# IRTの利点

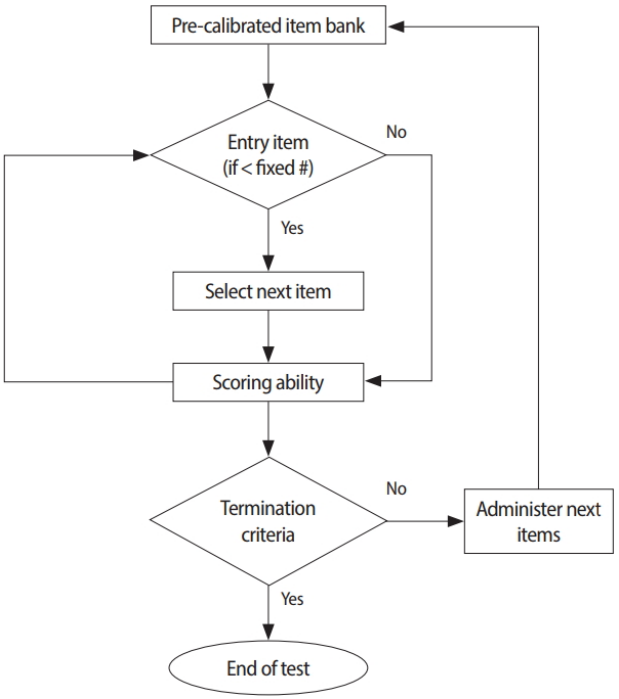
- 個々のテスト項目の特性を考慮した信頼性の高い能力推定が可能である
- 異なる項目群が出題された受検者の能力スコアを同一尺度上で比較できる
  ※ ただし，適切に比較可能にするためのデータ構造には条件もある
- 反応データの生成プロセスを受検者の能力と項目特性値に基づいて説明できるため，解釈性が高い
- 欠測値の扱いが容易である

# IRTを用いた高度なテスト運用

TOEFLやITパスポート，医学共用試験などで広く実用化

1. 所望の難易度・予測誤差となる**テストの自動構成**
2. 能力に合った項目を適応的に出題することで、短時間かつ少数項目で高精度な能力評価を実現する**適応型テスト**



# 多様な拡張モデル

**一般化部分採点モデル/段階反応モデル：**
多値型の得点データを扱うIRTモデル

**多次元項目反応モデル：**
受検者の能力を多次元的に推定できるIRTモデル

**多相ラッシュモデル：**
項目特性だけでなく，評価者などの別の要因の影響も同時に考慮できるIRTモデル

その他にもデータやテストの特徴に応じた多様なIRTモデルが知られており，広く活用されている

# （参考）多相ラッシュモデルの拡張

## 一般化多相ラッシュモデル

- Masaki Uto, Maomi Ueno (2020) A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. Behaviormetrika.
- Masaki Uto, Jun Tsuruta, Kouji Araki, Maomi Ueno (2024) Item response theory model highlighting rating scale of a rubric and rater-rubric interaction in objective structured clinical examination. PLOS ONE.

## 多次元型多相ラッシュモデル

- Masaki Uto (2021) A multidimensional generalized many-facet Rasch model for rubric-based performance assessment. Behaviormetrika.

## 時系列型ベイジアン多相ラッシュモデル

- Masaki Uto (2023) A Bayesian Many-Facet Rasch Model with Markov Modeling for Rater Severity Drift. Behavior Research Methods.

# 本発表の概要

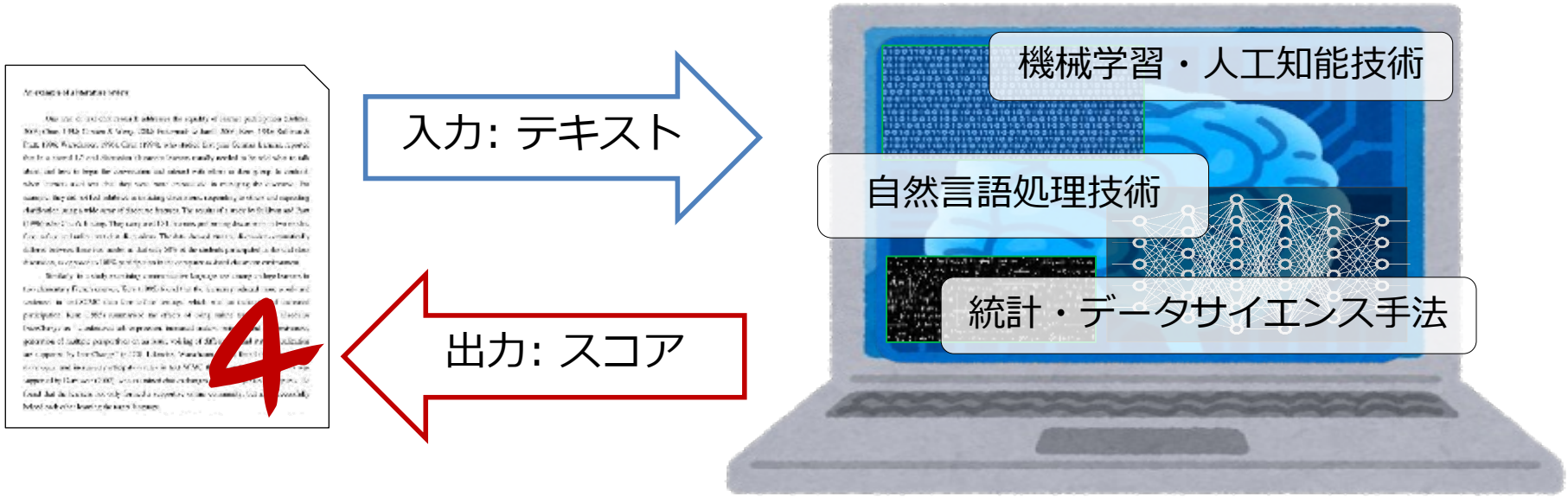
**深層学習に基づく記述式回答の自動採点技術とIRTを融合した先端技術群**を紹介

# 目次

1. 項目反応理論の概要
2. **記述式回答自動採点技術の概要**
3. 項目反応理論と記述式回答自動の融合技術
   - 評価者バイアスに頑健な自動採点手法
   - 自動採点機のアンサンブル手法
   - 複数の評価観点で得点を予測する自動採点手法
4. 質疑応答

# 記述式回答自動採点技術

人工知能技術を活用して人間評価者の代わりに小論文や短答記述式回答，レポートなどを採点する技術

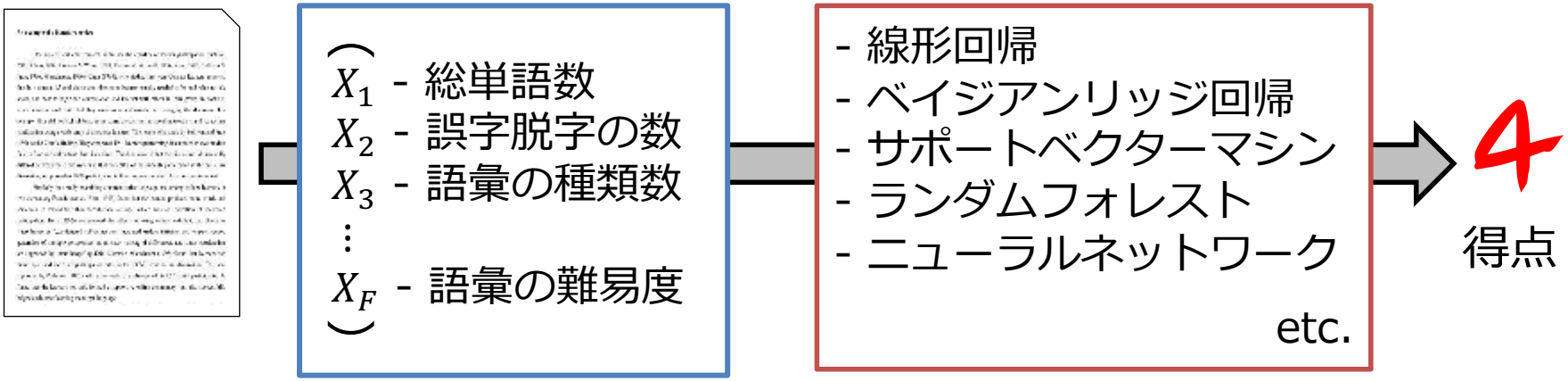



**<u>代表的な二つのアプローチ</u>**

1. 特徴量ベースのアプローチ
2. 深層学習ベースのアプローチ

# 特徴量ベースのアプローチ

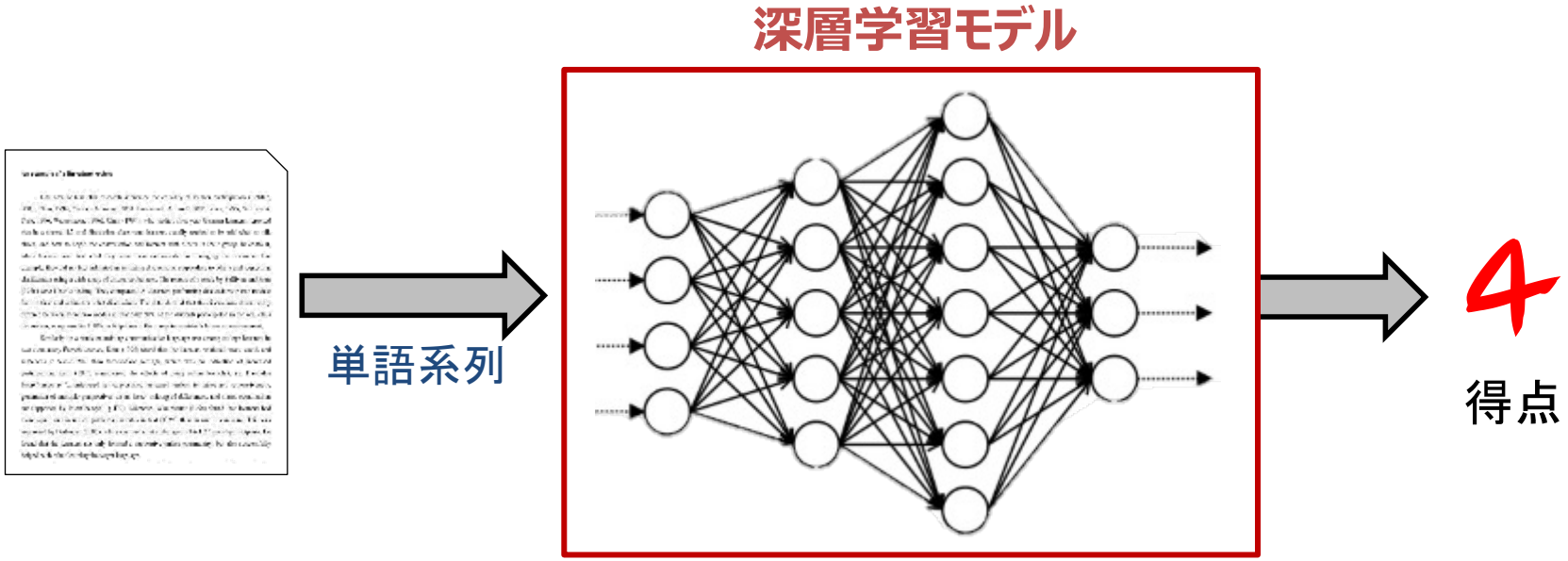
専門家が事前に設計した特徴量を利用
特徴量と得点の関係を機械学習モデルで学習




**代表的な特徴量ベースのシステム**
**e-rater**：ETSが開発し，TOEFLやGREなどで実用化
**JESS**：大学入試センターが開発した日本語対象のシステム
**EASE**：ヒューレット財団開催の自動採点コンペティション（Automated Student Assessment Prize）で上位入賞したシステム

# 深層学習ベースのアプローチ

深層学習を用いて文章の単語系列から直接得点を予測
⇨ 人手での特徴量設計が不要

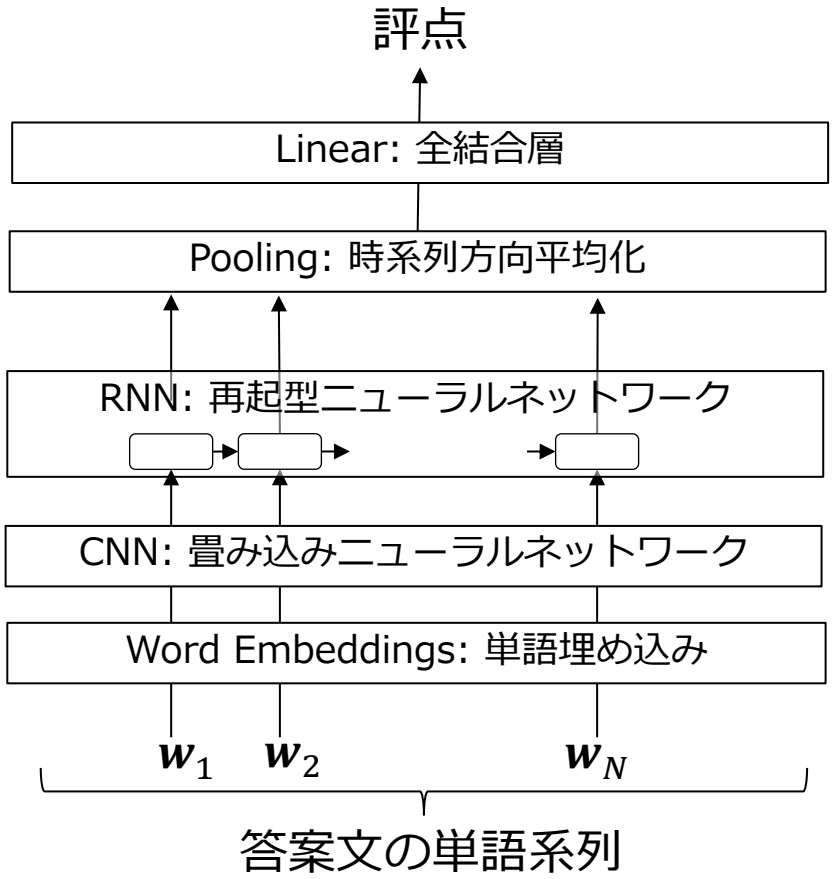



- 2016年頃に初期モデルが提案されたアプローチ
- 現在も人工知能・言語処理のトップカンファレンスで研究が続いており，高性能化が進行中

# 深層学習自動採点モデルの例

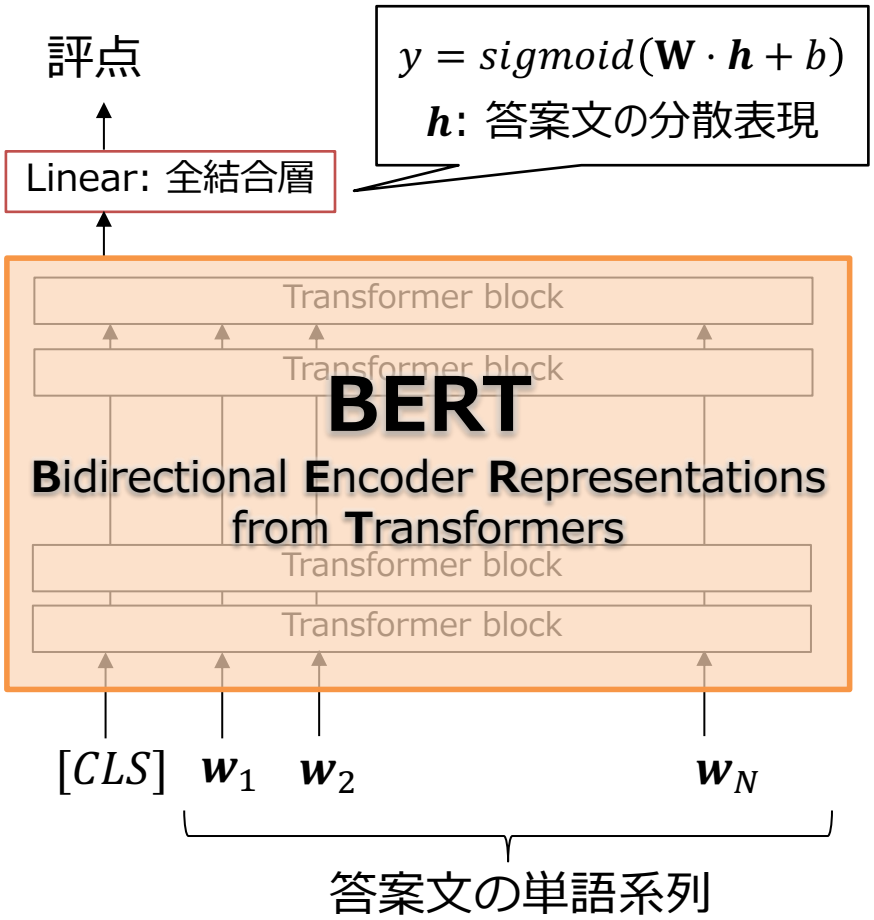
## RNNベースモデル

Kaveh & Hwee (2016) *A neural approach to automated essay scoring*. EMNLP.



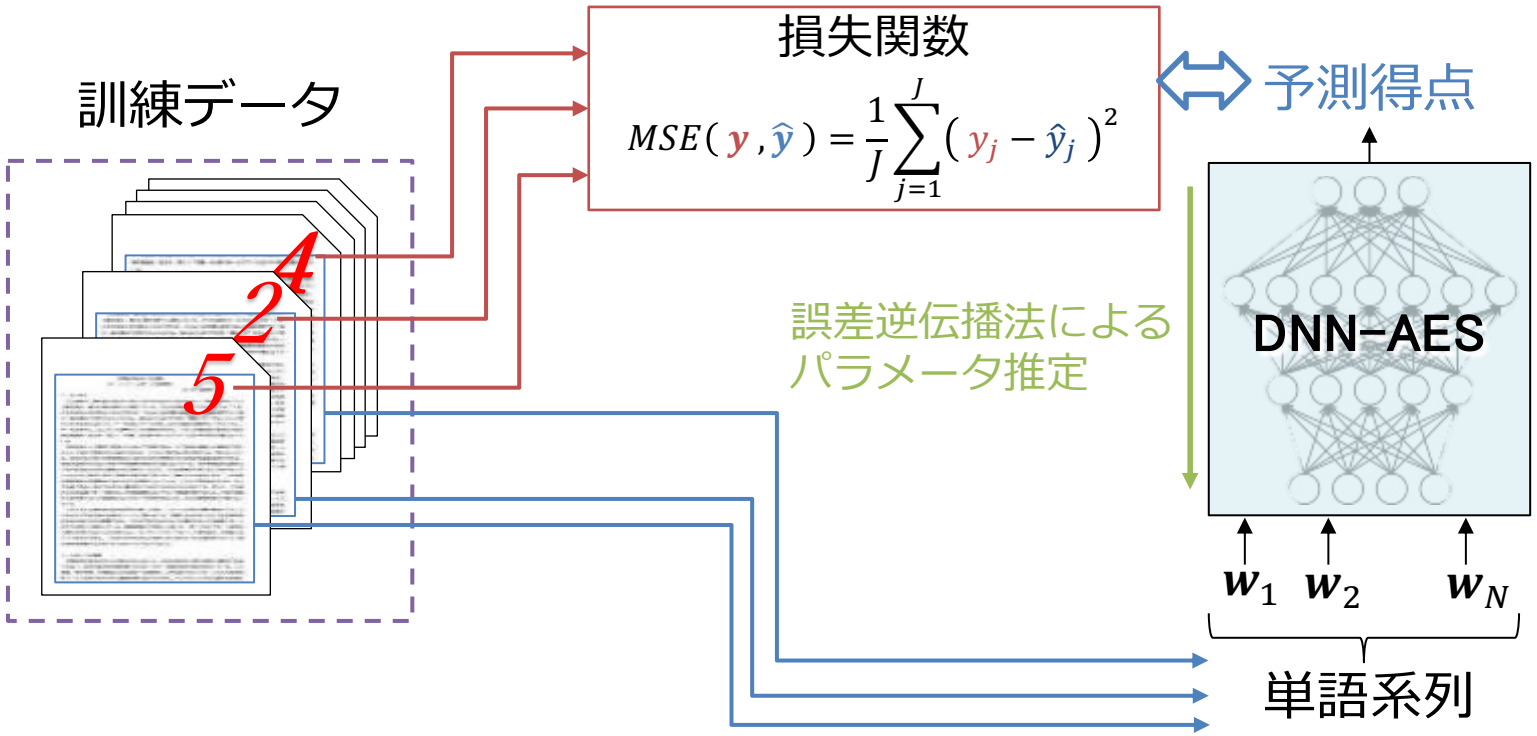

## BERTベースモデル

Devlin et al. (2018) *BERT: Pre-training of deep bidirectional Transformers for Language Understanding*. arXiv.



# 自動採点モデルの構築

大量の採点済みの答案データを用いてモデルパラメータを学習することで，自動採点モデルを構築




※ 最近はLLMに基づく手法も提案されているが，ある程度の訓練データがある場合には，従来のパラダイムの方が一般に高精度

# 目次

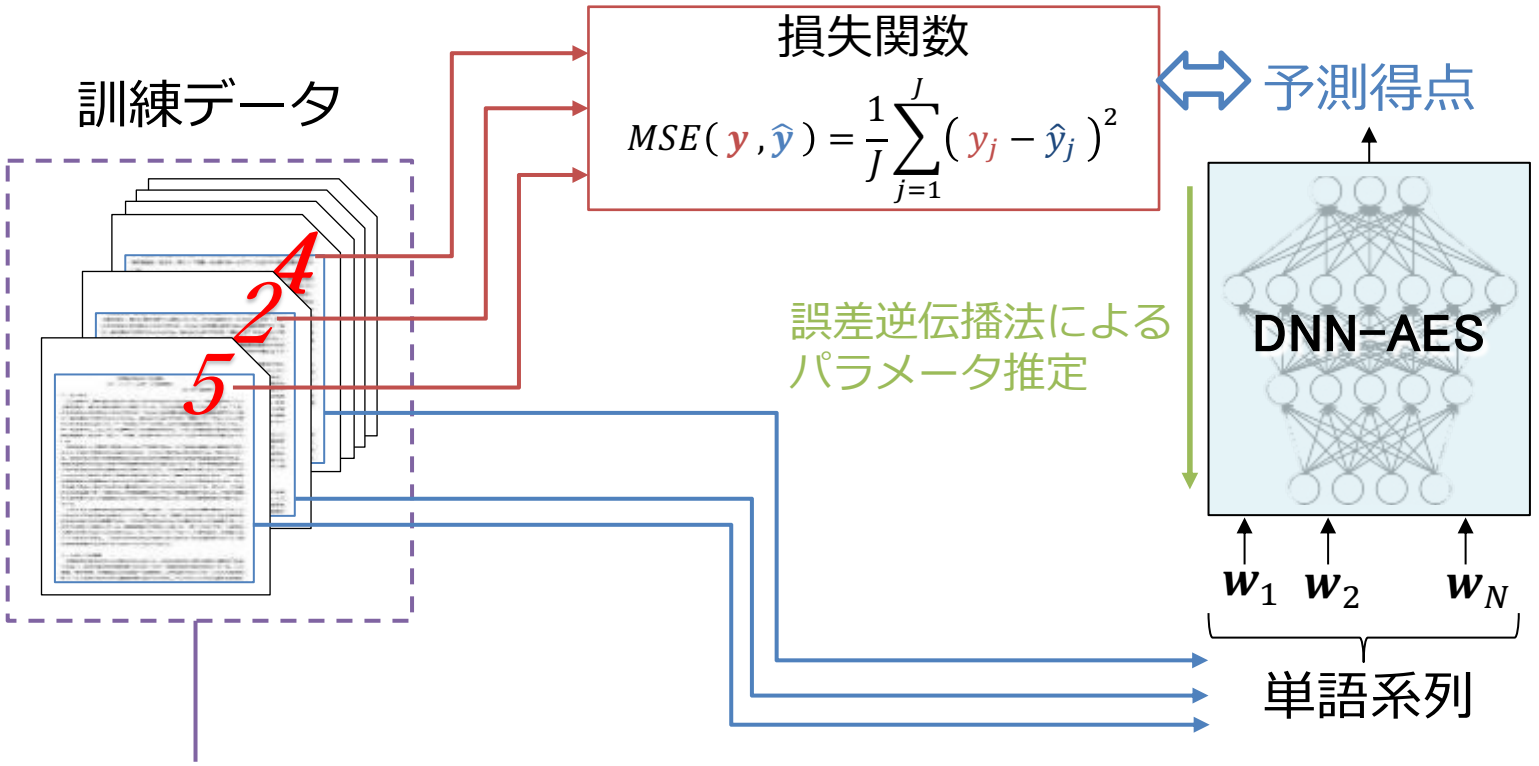
1. 項目反応理論の概要
2. 記述式回答自動採点技術の概要
3. **項目反応理論と記述式回答自動の融合技術**
   - 評価者バイアスに頑健な自動採点手法
   - 自動採点機のアンサンブル手法
   - 複数の評価観点で得点を予測する自動採点手法
4. 質疑応答

# 目次

1. 項目反応理論の概要
2. 記述式回答自動採点技術の概要
3. **項目反応理論と記述式回答自動の融合技術**
   - **評価者バイアスに頑健な自動採点手法**
   - 自動採点機のアンサンブル手法
   - 複数の評価観点で得点を予測する自動採点手法
4. 質疑応答

# 自動採点モデルの学習（再掲）

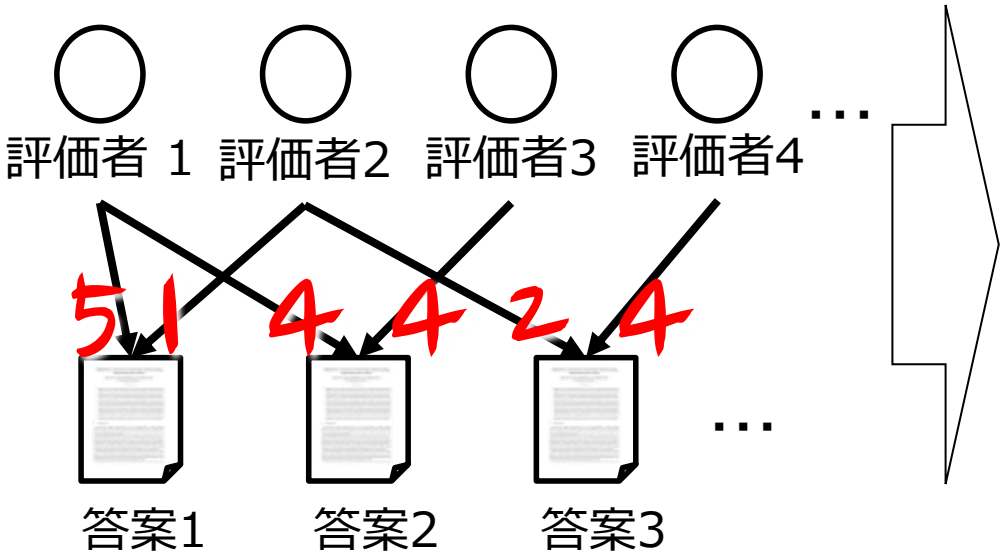
大量の採点済みの答案データを用いてモデルパラメータを学習することで，自動採点モデルを構築




**訓練データ中の各答案文への得点は正しいと仮定**

# 評価者（アノテータ）バイアスの影響

訓練データは、事前に収集した大量の答案を複数の評価者で分担して採点することで一般に作成される




| | 評価者 | | | | 平均 |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| 答案 1 | 5 | 1 | - | - | 3 |
| 答案 2 | 4 | - | 4 | - | 4 |
| 答案 3 | - | 2 | - | 4 | 3 |

自動採点モデルは
平均点に基づいて学習

平均点は評価者の特性（甘さ/辛さなど）に強く依存

訓練データ中の評価者バイアスの影響が自動採点モデルにも反映されてしまい自動採点の性能が低下

# 一般化多相ラッシュモデル

Uto & Ueno (2020) A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. Behaviormetrika, Springer.

評価者の特性を考慮したIRTモデルの一つ

受検者 $j$ の回答に評価者 $r$ が評点 $k$ を与える確率

評価者 $r$ の特性パラメータ

一貫性　厳しさ　得点別の厳しさ

観測評点を表す変数

$$P(x_{jr}=k)=\frac{exp\sum_{m=1}^{k}(\alpha_r(\theta_j-\beta_r-d_{rm}))}{\sum_{l=1}^{K}exp\sum_{m=1}^{l}(\alpha_r(\theta_j-\beta_r-d_{rm}))}$$

受検者 $j$ の能力

能力値と評価者特性は観測評点データの集合から推定

**評価者バイアスの影響を取り除いた受検者の能力スコアの推定が可能**

※ 元々のモデルは評価者特性に加えて項目の特性も考慮できる3相型モデルとして提案されているが，ここでの用途に合わせて2相型で説明している

# 評価者バイアスに頑健な自動採点モデル

評価者特性を考慮したIRTモデル（一般化多相ラッシュモデル）を組み込んだ自動採点手法

**能力スコアの推定**

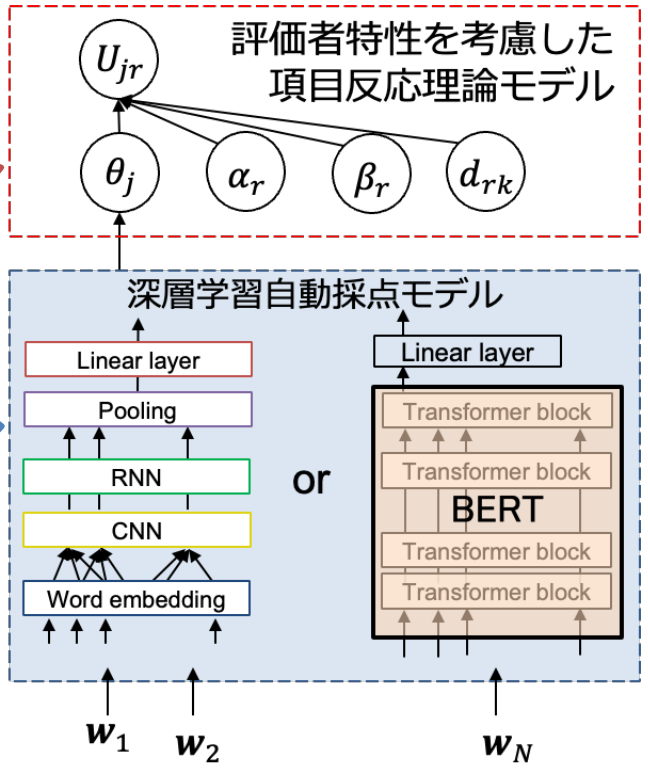
観測得点のデータから、評価者バイアスの影響を取り除いた能力スコア $\theta_j$ (答案の潜在得点)を推定

| | 評価者 | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 答案 1 | 5 | 1 | - | - |
| 答案 2 | 4 | - | 4 | - |
| 答案 3 | - | 2 | - | 4 |

**自動採点モデルの学習**

得られた能力スコア $\theta$を目的変数として、自動採点モデルを学習

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



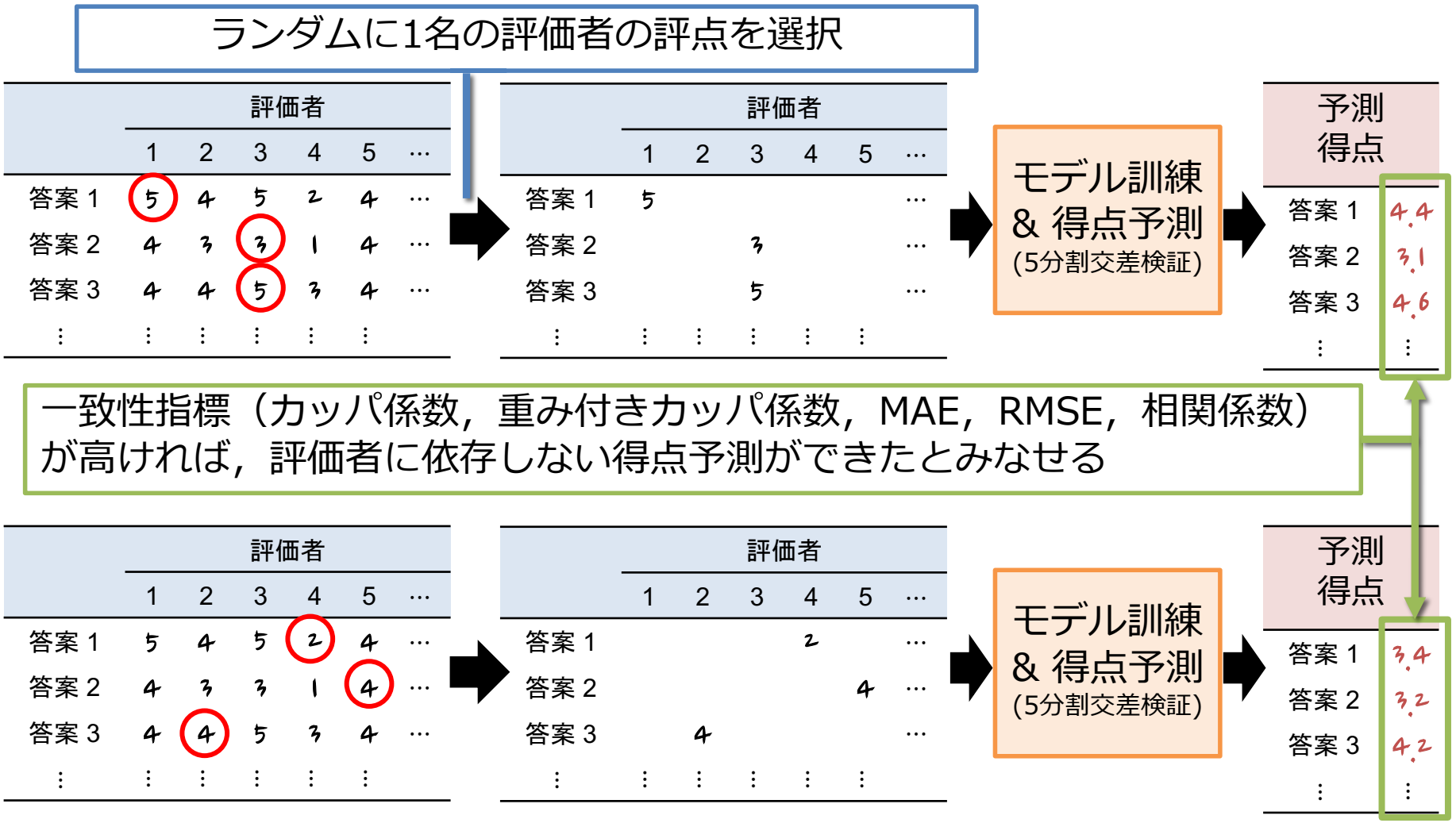


Uto, Okano (2022) Learning Automated Essay Scoring Models Using Item Response Theory-Based Scores to Decrease Effects of Rater Biases. IEEE Transactions on Learning Technologies.

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

# 評価実験（説明割愛）

個々の答案を採点する評価者が変わっても，安定した得点予測を行うことができるかを評価



# 実験結果（説明割愛）

IRTを利用しない従来手法と比較
様々な構成の深層学習自動採点モデルで検証

| | カッパ係数 | | | 重み付きカッパ | | | RMSE | | | 相関係数 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 提案 | 従来 | P値 | 提案 | 従来 | P値 | 提案 | 従来 | P値 | 提案 | 従来 | P値 |
| LSTM | **0.749** | 0.624 | <.01 | **0.778** | 0.727 | <.01 | **0.191** | 0.301 | <.01 | **0.937** | 0.931 | <.05 |
| LSTM w/o CNN | **0.831** | 0.697 | <.01 | **0.845** | 0.779 | <.01 | **0.142** | 0.237 | <.01 | **0.965** | 0.958 | <.01 |
| 2層LSTM | **0.828** | 0.661 | <.01 | **0.842** | 0.752 | <.01 | **0.147** | 0.268 | <.01 | **0.963** | 0.946 | <.01 |
| 双方向LSTM | **0.608** | 0.386 | <.01 | **0.624** | 0.508 | <.01 | **0.282** | 0.470 | <.01 | **0.816** | 0.772 | <.01 |
| BERT | **0.790** | 0.629 | <.01 | **0.808** | 0.743 | <.01 | **0.159** | 0.311 | <.01 | **0.960** | 0.935 | <.01 |

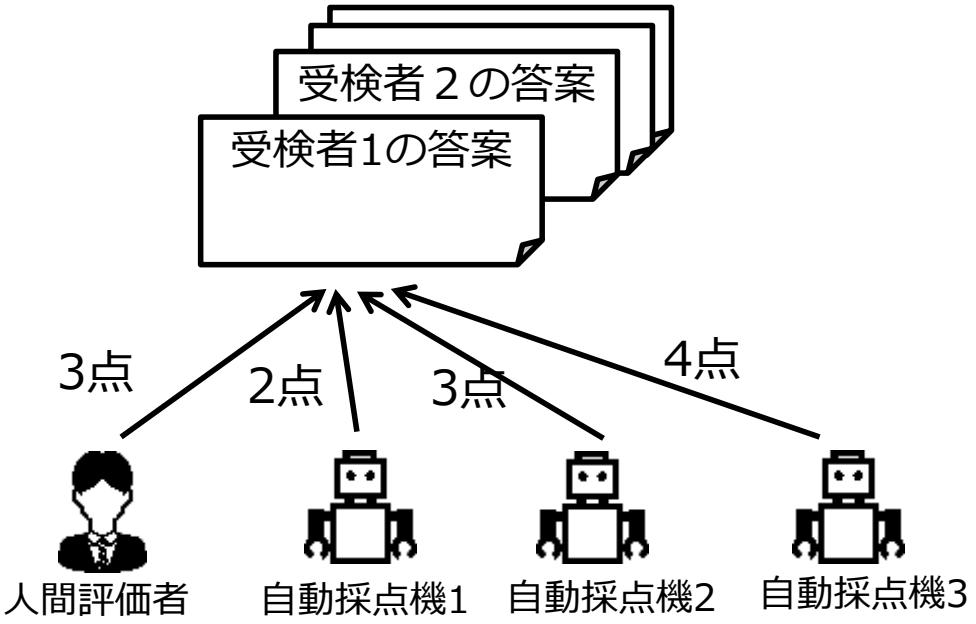
- 全ての条件で提案手法が高い性能
- 様々な自動採点モデルに容易に組み込んで性能向上が可能

# 目次

1. 項目反応理論の概要
2. 記述式回答自動採点技術の概要
3. **項目反応理論と記述式回答自動の融合技術**
   - 評価者バイアスに頑健な自動採点手法
   - **自動採点機のアンサンブル手法**
   - 複数の評価観点で得点を予測する自動採点手法
4. 質疑応答

# 自動採点モデルのアンサンブル学習手法

特徴の異なる様々な自動採点モデルを統合（アンサンブル）することで自動採点の性能改善を狙った研究



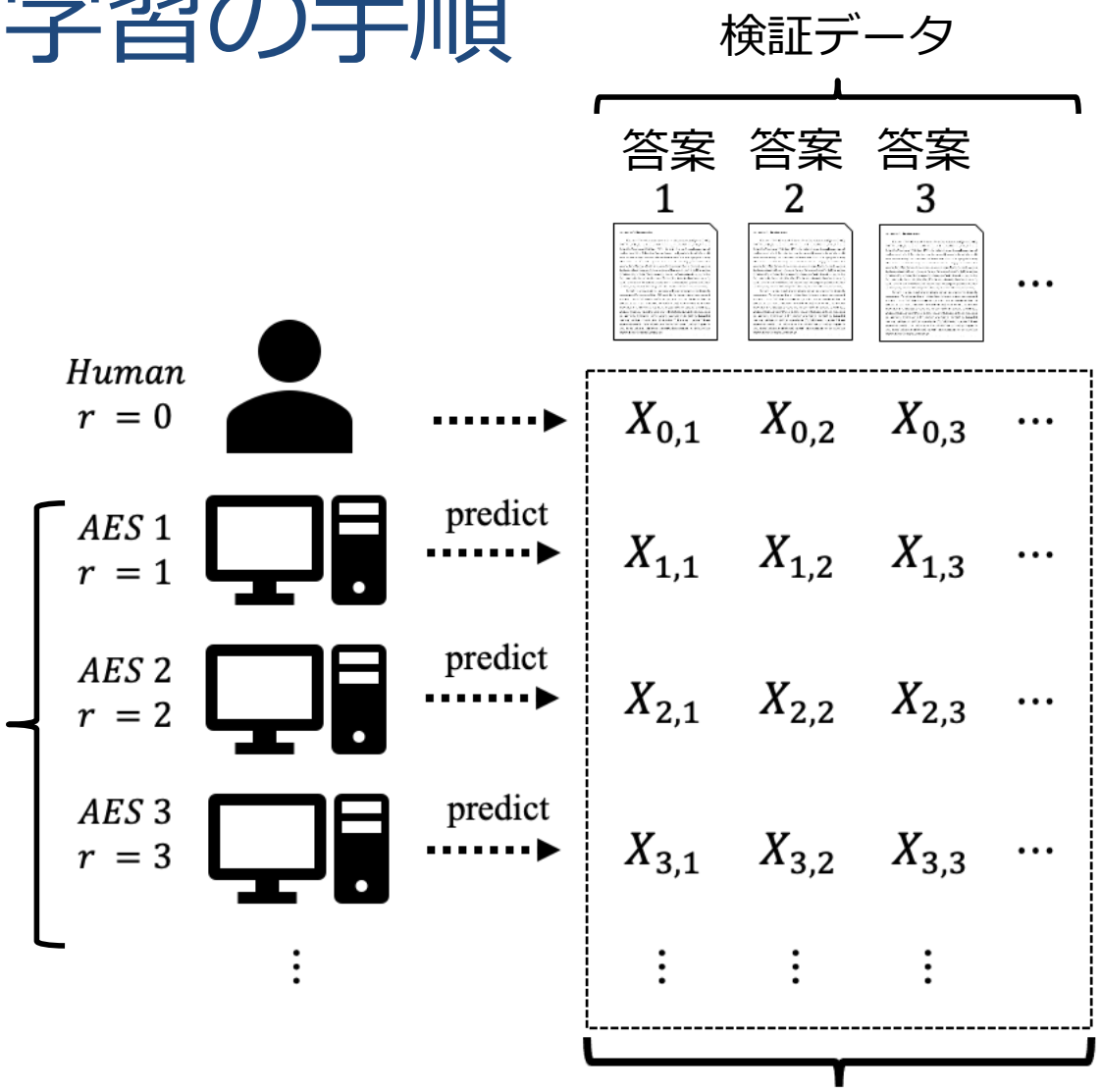

**提案手法の特徴**

- 評価者特性を考慮した項目反応理論を用いて各自動採点モデルの特性を考慮して統合
- 最終的には、統合したスコアを人間の評価尺度に変換して予測得点を出力する

Uto, Aomi, Tsutsumi, Ueno (2023) Integration of Prediction Scores from Various Automated Essay Scoring Models Using Item Response Theory. IEEE Transactions on Learning Technologies.

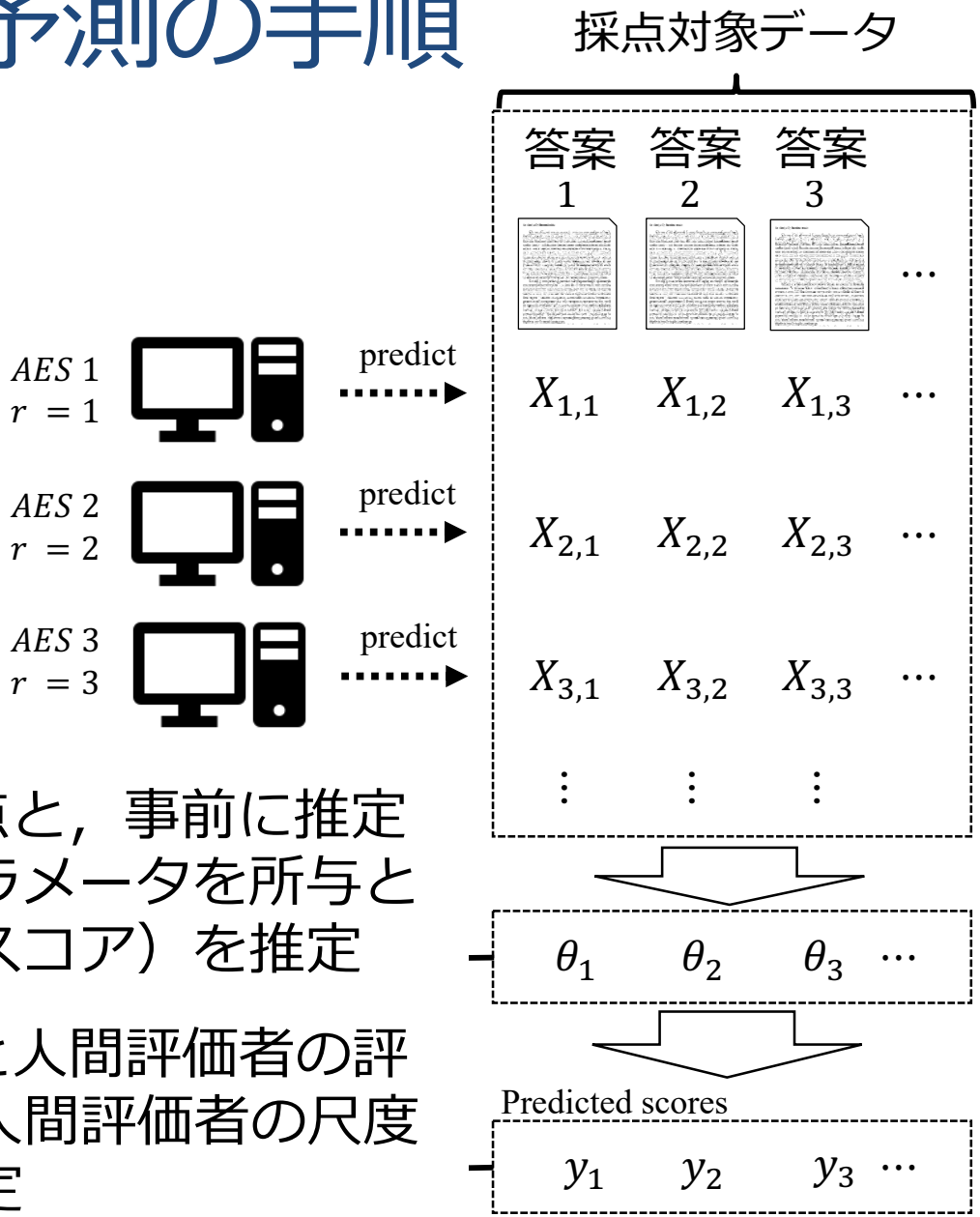
# 学習の手順

**Step 1:**
訓練データ（採点済み答案データセット）を用いて複数の自動採点機を構築




**Step 2:**
検証データ中の記述回答に対する各自動採点機の予測データを用いて評価者IRTで評価者パラメータを推定

# 得点予測の手順




**Step 1:**各自動採点の予測得点と，事前に推定された自動採点機の評価者パラメータを所与として能力スコア（答案の潜在スコア）を推定

**Step 2:**得られた能力スコアと人間評価者の評価者パラメータに基づいて，人間評価者の尺度に沿った段階得点スコアを推定

$$y_j = \sum_{k=1}^{K} k \cdot P\left(X_{j,r^*} = k | \theta_j\right)$$ $r^*$:人間評価者

# 性能評価

ベンチマークデータセット（ASAPデータセット）を利用した5分割交差検証による評価

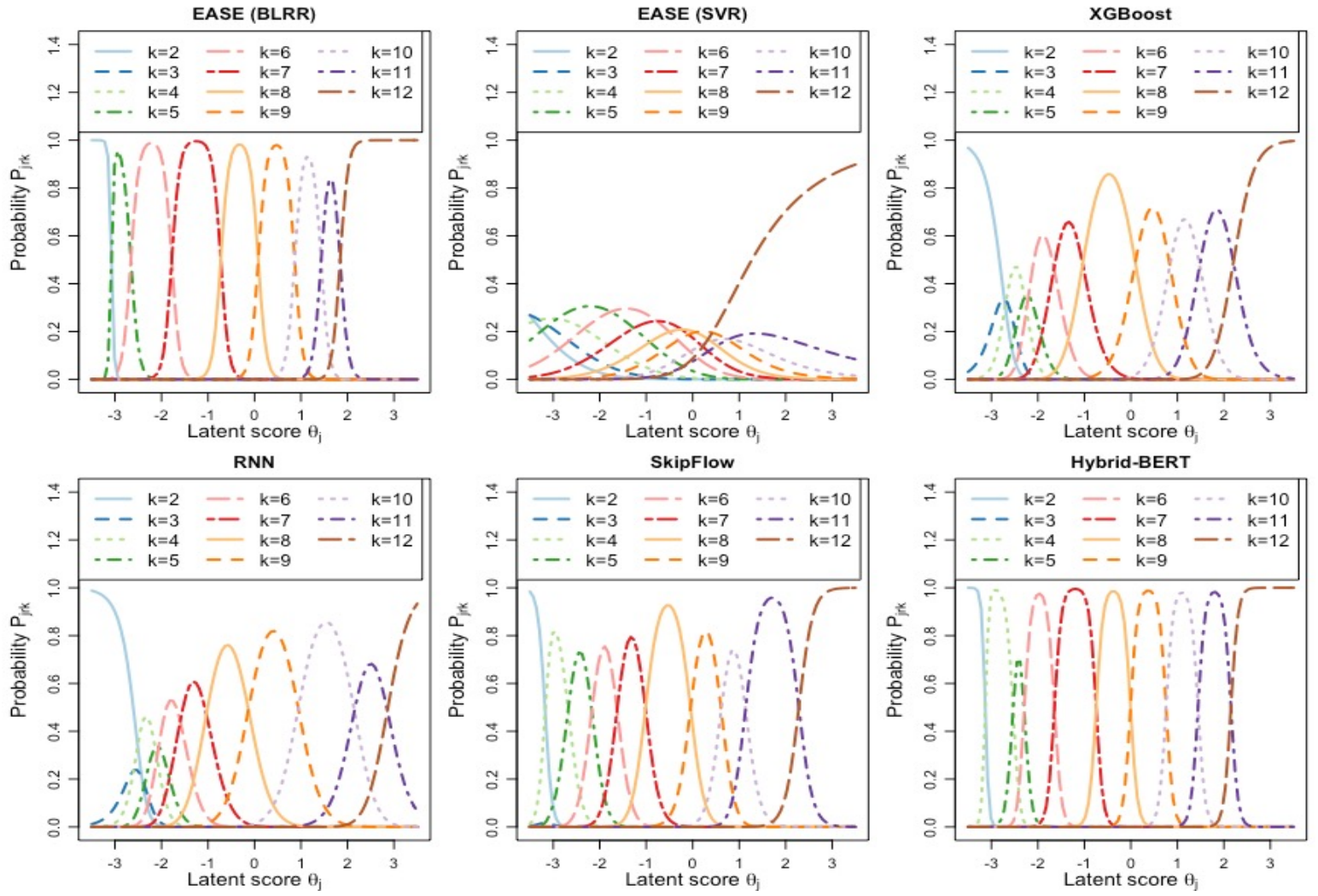
評価指標：重み付きカッパ係数

| | | Prompt | | | | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | |
| Individual models | EASE (BLRR) | 0.8038 | 0.6029 | 0.6555 | 0.7171 | 0.7845 | 0.7612 | 0.7300 | 0.6656 | 0.7151 |
| | EASE (SVR) | 0.5578 | 0.5328 | 0.5644 | 0.5711 | 0.7397 | 0.6902 | 0.5451 | 0.3757 | 0.5721 |
| | XGBoost | 0.8138 | 0.6397 | 0.5929 | 0.6596 | 0.7627 | 0.6573 | 0.6921 | 0.6704 | 0.6861 |
| | RNN | 0.7769 | 0.6185 | 0.6511 | 0.7299 | 0.7542 | 0.7661 | 0.7496 | 0.5074 | 0.6942 |
| | SkipFlow | 0.7984 | 0.6516 | 0.6568 | 0.7294 | 0.7841 | 0.7820 | 0.7512 | 0.6138 | 0.7209 |
| | Hybrid–BERT | 0.8271 | 0.6372 | 0.6716 | 0.6204 | 0.7803 | 0.6728 | 0.7202 | 0.6723 | 0.7003 |
| Conventional integration methods | MEAN | 0.8210 | 0.6771 | 0.6644 | 0.7185 | 0.7959 | 0.7725 | 0.7674 | 0.6722 | 0.7361 |
| | VOTE | 0.8343 | 0.6620 | 0.6749 | 0.7287 | 0.7937 | 0.7710 | 0.7484 | 0.6700 | 0.7354 |
| | STACKING (Linear) | 0.8313 | 0.6644 | 0.6492 | **0.7386** | 0.7861 | 0.7839 | 0.7701 | 0.6922 | 0.7395 |
| | STACKING (Ridge) | 0.8316 | 0.6630 | 0.6477 | **0.7386** | 0.7867 | 0.7835 | 0.7703 | 0.6925 | 0.7392 |
| | STACKING (SVR) | 0.8221 | 0.6230 | 0.6561 | 0.7235 | 0.7804 | 0.7704 | 0.7714 | 0.5810 | 0.7160 |
| | STACKING (Boosting) | 0.8270 | 0.6599 | 0.6366 | 0.7367 | 0.7878 | 0.7838 | 0.7568 | 0.6439 | 0.7291 |
| Proposed method | GMFRM | **0.8365** | **0.6785** | 0.6695 | 0.7375 | **0.7972** | **0.7850** | 0.7893 | 0.7095 | **0.7562** |
| | Consistency-fixed GMFRM | 0.8351 | 0.6657 | **0.6755** | 0.7223 | 0.7851 | 0.7608 | **0.7979** | 0.6902 | 0.7416 |
| | Severity-fixed GMFRM | 0.8313 | 0.6673 | 0.6645 | 0.7380 | 0.7968 | 0.7734 | 0.7875 | **0.7099** | 0.7461 |
| | Threshold-fixed GMFRM | 0.8309 | 0.6690 | 0.6505 | 0.7117 | 0.7905 | 0.7598 | 0.7716 | 0.6944 | 0.7348 |
| | MFRM | 0.7944 | 0.6089 | 0.6630 | 0.6868 | 0.7769 | 0.7284 | 0.7710 | 0.6669 | 0.7120 |

The bold values indicate the maximum QWK values for each prompt. The underlined values represent second largest values.

**単一の自動採点モデルを利用する場合や従来のスタッキング手法と比べて高精度な自動採点を達成**

# 自動採点機ごとの特性解釈も可能




図：自動採点機ごとの反応曲線

# 目次

1. 項目反応理論の概要
2. 記述式回答自動採点技術の概要
3. **項目反応理論と記述式回答自動の融合技術**
   - 評価者バイアスに頑健な自動採点手法
   - 自動採点機のアンサンブル手法
   - **複数の評価観点で得点を予測する自動採点手法**
4. 質疑応答

# 複数の評価観点に基づく自動採点手法

- これまでの自動採点では、単一の総合得点だけを予測する手法が主流
- 現実の人間による採点場面では，ルーブリック（評価基準表）に基づいて複数観点で採点される場合があり、自動採点も観点別で行いたいニーズがしばしば生じる

| | 問題解決力 | | 論理的思考力 | | |
|---|---|---|---|---|---|
| | **評価観点1**<br>(問題設定) | **評価観点2**<br>(結論の導出) | **評価観点3**<br>(根拠の提示) | **評価観点4**<br>(対立意見の検討) | **評価観点5**<br>(全体構成) |
| **3** | 与えられたテーマから問題を設定し，論ずる意義も含め，その問題を取り上げた理由や背景について述べている． | 設定した問題に対し，展開してきた自分の主張を関連づけながら，結論を導いている．結論は一般論にとどまらず，独自性を有している． | 自分の主張の根拠が述べられており，かつ根拠の真実性を立証する信頼できる複数の事実・データが示されている． | 自分の主張と対立するいくつかの意見を取り上げ，それらすべてに対して論駁(問題点の指摘)を行っている． | 問題の設定から結論にいたる論理的な組み立て，記述の順序，パラグラフの接続が整っている．概要は本文の内容を的確に要約している． |
| **2** | 与えられたテーマから問題を設定し，その問題を取り上げた理由や背景について述べている． | 設定した問題に対し，展開してきた自分の主張を関連づけながら，結論を導いている． | 自分の主張の根拠が述べられており，かつ根拠の真実性を立証する信頼できる事実・データが少なくとも一つ示されている． | 自分の主張と対立する少なくとも一つの意見を取り上げ，それに対して論駁(問題点の指摘)を行っている． | 問題の設定から結論にいたる論理的な組み立て，記述の順序，パラグラフの接続がおおむね整っている． |
| **1** | 与えられたテーマから問題を設定しているが，その問題を取り上げた理由や背景の内容が不十分である． | 結論は述べられているが，展開してきた自分の主張との関連づけが不十分である． | 自分の主張の根拠は述べられているが，根拠の真実性を立証する信頼できる事実・データが明らかにされていない． | 自分の主張と対立する意見を取り上げているが，それに対して論駁（問題点の指摘）がなされていない． | 問題の設定から結論にいたるアウトラインはたどれるが，記述の順序やパラグラフの接続に難点のある箇所が散見される． |
| **0** | 1未満の水準 | 1未満の水準 | 1未満の水準 | 1未満の水準 | 1未満の水準 |

＊ 松下ほか（2013）レポート評価における ルーブリックの開発とその信頼性の検討. 大学教育学会誌. をもとに作成

# 複数の評価観点に基づく自動採点手法

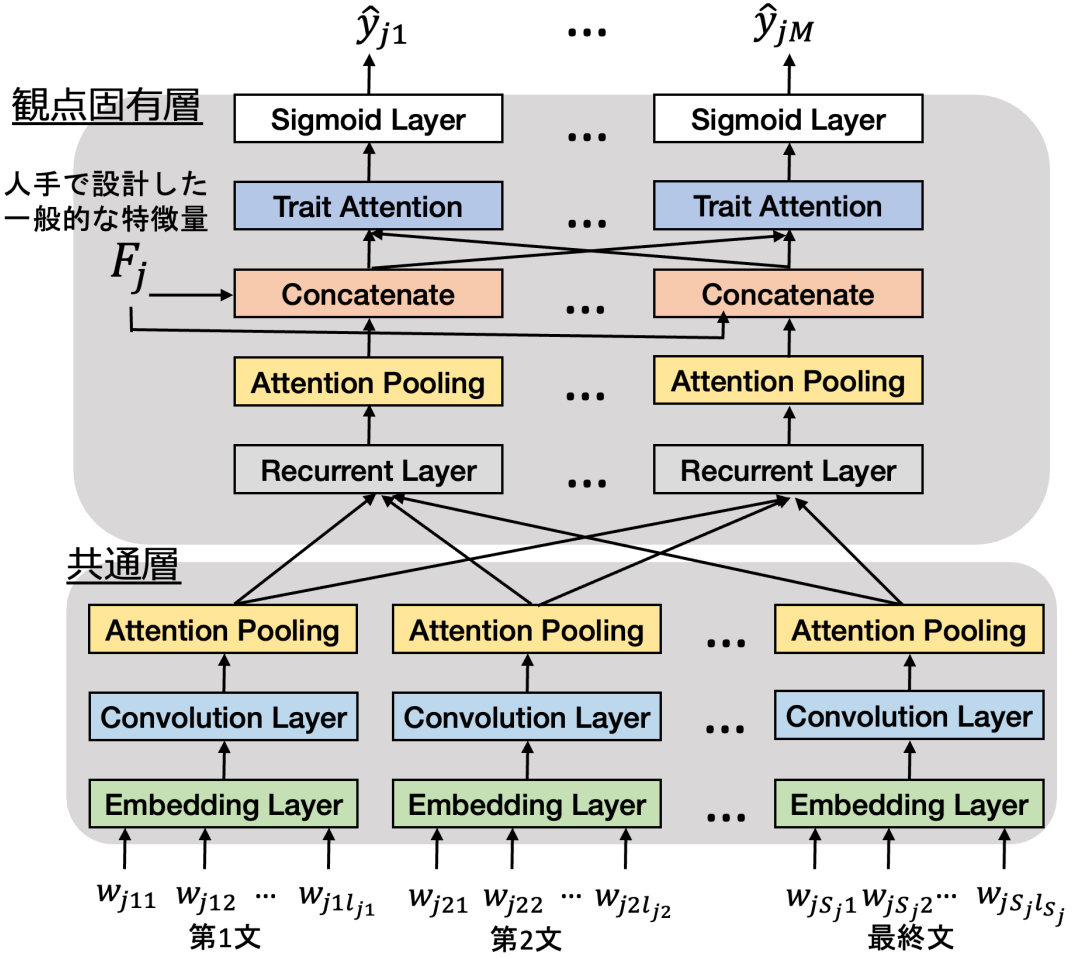
Ridley et al. (2021) Automated cross- prompt scoring of essay traits. AAAI.

総合得点に加えて評価観点別の得点も予測できる手法

右は2021年当時の最高精度を達成した手法

現在もベースラインに使用される有効な手法だが，複雑な構造であり予測の解釈性が低い

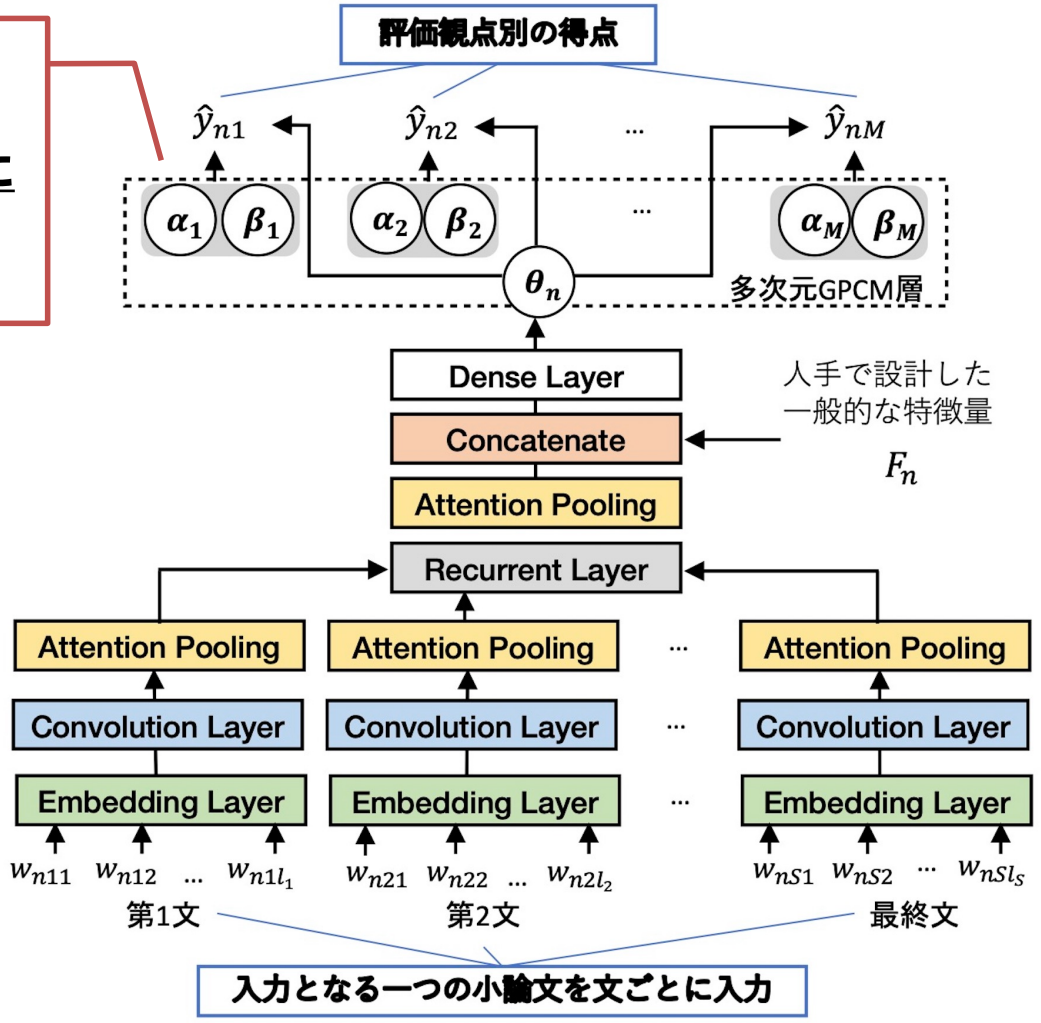
試験の改善や受検者への説明責任のために予測根拠の解釈性の向上は重要



# 複数の評価観点に基づく自動採点手法

Shibata & Uto (2022) Analytic Automated Essay Scoring based on Deep Neural Networks Integrating Multidimensional Item Response Theory, International Conference on Computational Linguistics (COLING).

多次元IRTを出力層に組み込むことで**予測根拠の解釈性を向上した手法を提案**



# 多次元一般化部分採点モデル

- 多値型得点に対応した多次元型項目反応モデル
- 受検者$j$が評価観点$m$において得点$k$を得る確率を以下で定義

**多次元識別力**
評価観点$m$が受検者の$d$次元目の能力を識別する力

**困難度**
評価観点$m$において得点$u$を得る難しさ

$$P_{jmk} = \frac{\exp(k\sum_d \alpha_{dm}\boldsymbol{\theta}_{dj} + \sum_{u=1}^{k}\beta_{mu})}{\sum_{v=1}^{K_m}\exp(v\sum_d \alpha_{dm}\boldsymbol{\theta}_{dj} + \sum_{u=1}^{v}\beta_{mu})}$$

**受検者 $j$ の$d$次元目の能力**

**特徴：**

- 評価観点の背後に複数の能力次元（因子）を仮定し，各評価観点が各能力次元をどの程度測定しているかを推定できる
- 多次元能力尺度上で受検者の能力を推定できる

# 多次元IRTを利用した理由

複数の評価観点が少数の（ただし1次元とは限らない）能力次元が仮定できる可能性あるため

| | 問題解決力 | | 論理的思考力 | | |
|---|---|---|---|---|---|
| | **評価観点1**<br>(問題設定) | **評価観点2**<br>(結論の導出) | **評価観点3**<br>(根拠の提示) | **評価観点4**<br>(対立意見の検討) | **評価観点5**<br>(全体構成) |
| **3** | 与えられたテーマから問題を設定し，論ずる意義も含め，その問題を取り上げた理由や背景について述べている． | 設定した問題に対し，展開してきた自分の主張を関連づけながら，結論を導いている．結論は一般論にとどまらず，独自性を有している． | 自分の主張の根拠が述べられており，かつ根拠の真実性を立証する信頼できる複数の事実・データが示されている． | 自分の主張と対立するいくつかの意見を取り上げ，それらすべてに対して論駁(問題点の指摘)を行っている． | 問題の設定から結論にいたる論理的な組み立て，記述の順序，パラグラフの接続が整っている．概要は本文の内容を的確に要約している． |
| **2** | 与えられたテーマから問題を設定し，その問題を取り上げた理由や背景について述べている． | 設定した問題に対し，展開してきた自分の主張を関連づけながら，結論を導いている． | 自分の主張の根拠が述べられており，かつ根拠の真実性を立証する信頼できる事実・データが少なくとも一つ示されている． | 自分の主張と対立する少なくとも一つの意見を取り上げ，それに対して論駁(問題点の指摘)を行っている． | 問題の設定から結論にいたる論理的な組み立て，記述の順序，パラグラフの接続がおおむね整っている． |
| **1** | 与えられたテーマから問題を設定しているが，その問題を取り上げた理由や背景の内容が不十分である． | 結論は述べられているが，展開してきた自分の主張との関連づけが不十分である． | 自分の主張の根拠は述べられているが，根拠の真実性を立証する信頼できる事実・データが明らかにされていない． | 自分の主張と対立する意見を取り上げているが，それに対して論駁（問題点の指摘）がなされていない． | 問題の設定から結論にいたるアウトラインはたどれるが，記述の順序やパラグラフの接続に難点のある箇所が散見される． |
| **0** | 1未満の水準 | 1未満の水準 | 1未満の水準 | 1未満の水準 | 1未満の水準 |

＊ 松下ほか（2013）レポート評価における ルーブリックの開発とその信頼性の検討．大学教育学会誌．をもとに作成

# 複数の評価観点に基づく自動採点手法

Shibata & Uto (2022) Analytic Automated Essay Scoring based on Deep Neural Networks Integrating Multidimensional Item Response Theory, International Conference on Computational Linguistics (COLING).

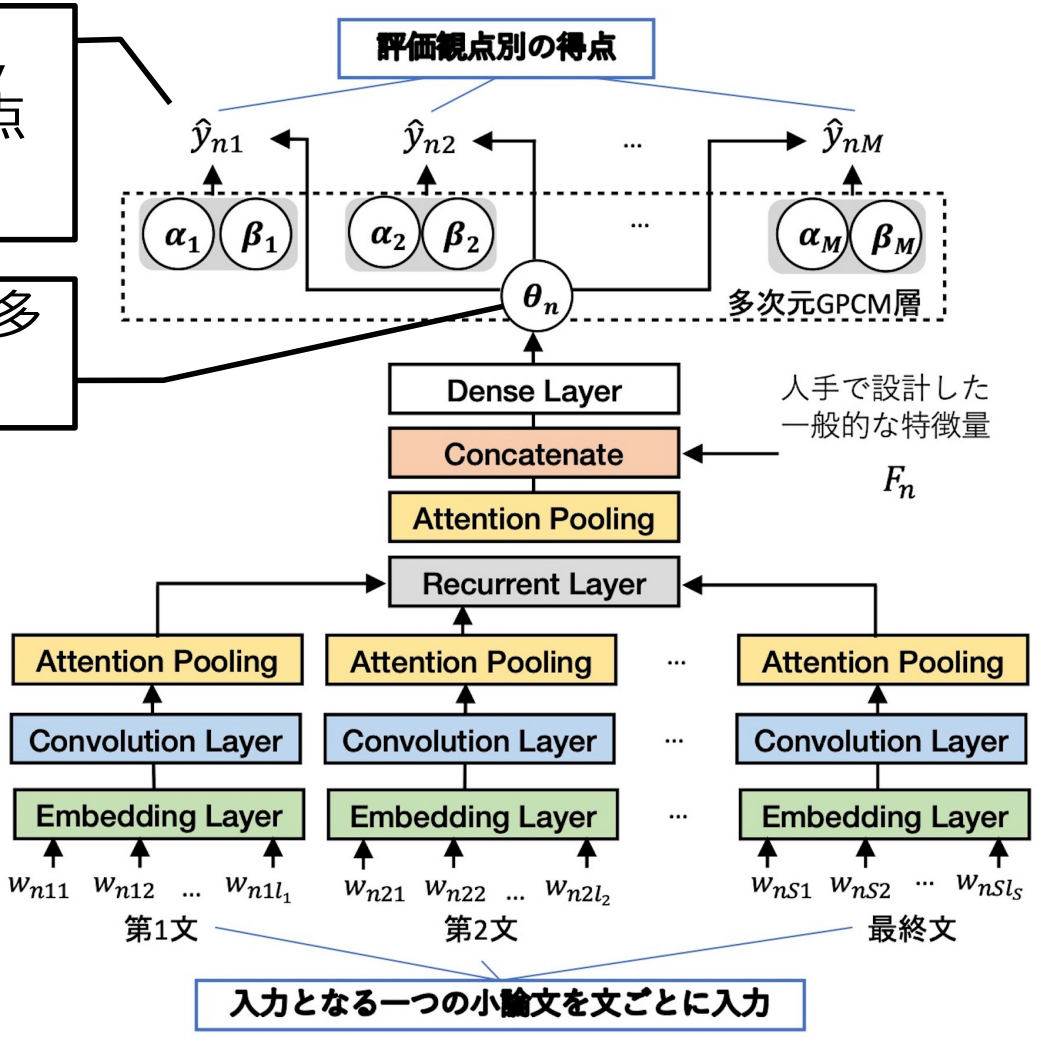
その多次元能力値を所与として，多次元IRTに基づいて各評価観点の得点を予測

答案文の潜在スコアに対応する多次元能力値を予測

モデル訓練は出力層も含めてEnd-to-endで実施（損失関数は多クラス交差エントロピー）

$$\mathcal{L} = -\frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\sum_{k=1}^{K_m} y_{jmk}\log(P_{jmk}) + C\sum_{j=1}^{J}\|\boldsymbol{\theta}_j\|^2$$

→ 推定された各評価観点の特性と各答案の多次元スコアが，背後の能力尺度と合わせて解釈



# 提案手法による解釈性

個別の評価観点の特性を定量的に解釈できる

多次元IRTと同様に，背後にある能力尺度を解釈できる

| 評価観点 | 識別力 | | 困難度 | | | | |
|---|---|---|---|---|---|---|---|
| | $\alpha_{21}$ | $\alpha_{22}$ | $\beta_{22}$ | $\beta_{23}$ | $\beta_{24}$ | $\beta_{25}$ | $\beta_{26}$ |
| 全体得点 | **3.24** | 0.20 | -6.56 | -4.29 | 0.09 | 4.64 | 6.12 |
| 内容 | **2.27** | 2.13 | -6.21 | -2.02 | 0.81 | 2.97 | 6.28 |
| 構成 | **2.48** | 2.00 | -5.72 | -1.58 | 1.52 | 3.30 | 7.26 |
| 語彙 | 1.75 | **2.86** | -6.05 | -2.06 | 1.18 | 3.82 | 6.88 |
| 流暢性 | 1.15 | **3.09** | -6.39 | -3.27 | 0.32 | 3.63 | 6.75 |
| 体裁 | 1.00 | **2.90** | -5.57 | -2.09 | 0.90 | 3.75 | 7.04 |

**内容面**に関連

**文章表現**に関連

# 予測精度の評価実験

## ASAP++：小論文の観点別自動採点のベンチマークデータ

| 課題番号 | 小論文数 | 評価観点 | 全体得点の範囲 | 観点別得点の範囲 |
|---|---|---|---|---|
| 1 | 1783 | 全体得点, 内容, 構成, 語彙, 流暢性, 体裁 | 2-12 | 1-6 |
| 2 | 1800 | 全体得点,内容, 構成, 語彙, 流暢性, 体裁 | 1-6 | 1-6 |
| 3 | 1726 | 全体得点,内容, 問題との整合性, 流暢性, 物語性 | 0-3 | 0-3 |
| 4 | 1772 | 全体得点,内容, 問題との整合性, 流暢性, 物語性 | 0-3 | 0-3 |
| 5 | 1805 | 全体得点,内容, 問題との整合性, 流暢性, 物語性 | 0-4 | 0-4 |
| 6 | 1800 | 全体得点,内容, 問題との整合性, 流暢性, 物語性 | 0-4 | 0-4 |
| 7 | 1569 | 全体得点,内容, 構成, 体裁, 文章表現 | 0-30 | 0-6 |
| 8 | 723 | 全体得点,内容, 構成, 語彙, 流暢性, 体裁, 意見 | 0-60 | 2-12 |

- 5分割交差検証
- 評価指標：二次重み付きカッパ係数（QWK）

| モデル | 課題番号 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Avg. |
| AAAI2021 | **0.685** | **0.655** | **0.660** | 0.720 | **0.706** | **0.750** | 0.694 | 0.568 | **0.680** |
| 提案-1dim | 0.656 | 0.617 | 0.620 | 0.713 | 0.686 | 0.731 | 0.638 | 0.549 | 0.652 |
| 提案-2dim | 0.666 | 0.631 | 0.637 | **0.722** | 0.699 | 0.732 | **0.704** | **0.576** | 0.671 |
| 提案-3dim | 0.679 | 0.633 | 0.642 | 0.704 | 0.698 | 0.734 | 0.696 | 0.553 | 0.667 |

有意差なし

有意には性能を落とすことなく予測できた

# ご清聴ありがとうございました

本日紹介した以外にも，評価者特性を考慮したIRTや記述式回答自動採点技術，問題自動生成などについて，様々な研究を行っています．興味がありましたら下記をご参照ください．

研究室ウェブサイト

基盤Sシンポ2023年12月動画