科研費基盤研究（S）シンポジウム 「eテスティング最前線」

項目反応理論を用いた小論文自動採点

電気通信大学 大学院情報理工学研究科 准教授
宇都 雅輝

2021年1月29日

小論文試験・記述式試験の活用と問題

表現力や思考力などの高次の能力を測定する手法の一つとして注目

[問題点1] 採点結果が評価者の特性に依存

⇨ 評価者の特性を考慮して能力を推定できる
項目反応理論

[問題点2] 採点コストが膨大

⇨ 自動採点技術
項目反応理論を用いた新たな手法を紹介

小論文試験・記述式試験の活用と問題

表現力や思考力などの高次の能力を測定する手法の一つとして注目

[問題点1] 採点結果が評価者の特性に依存

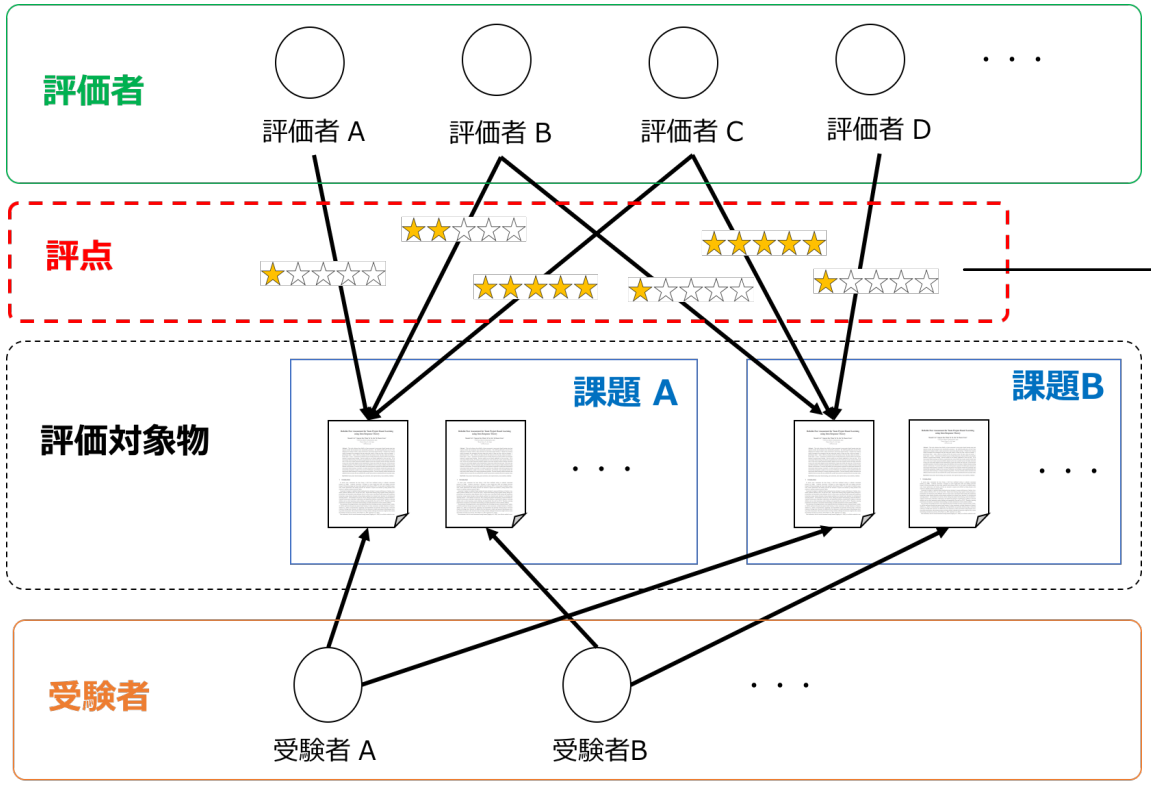
⇨ 評価者の特性を考慮して能力を推定できる
項目反応理論

[問題点2] 採点コストが膨大

⇨ 自動採点技術
項目反応理論を用いた新たな手法を紹介

小論文試験のデザイン

受検者に複数の小論文課題を出題し，その回答を複数の評価者で分担して採点



一般に各受験者の能力は観測評点の平均値として算出

しかし，個々の評点は**評価者や課題の特性に依存**

単純な平均点では高精度な能力評価は困難

評価者と課題の特性を考慮した項目反応モデル

- Patz & Junker (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. Journal of Educational and Behavioral Statistics.
- Patz, Junker & Johnson (1999) “The hierarchical rater model for rated test items and its application to large-scale educational assessment data,” *Journal of Educational and Behavioral Statistics*.
- DeCarlo, Kim & Johnson (2011) “A hierarchical rater model for constructed responses, with a signal detection rater model,” Journal of Educational Measurement.
- Ueno & Okamoto (2008). Item response theory for peer assessment. In Proc. IEEE international conference on advanced learning technologies
- Uto & Ueno (2016). *Item response theory for peer assessment*. IEEE Transactions on Learning Technologies, IEEE Computer Society.
- Uto & Ueno (2020) A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. Behaviormetrika, Springer.

⇨ **評価者と課題の特性を最も柔軟に表現できる最先端モデルのひとつ**

評価者と課題の特性を考慮した項目反応モデル

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

課題 i への受検者 j の回答に評価者 r が評点 k を与える確率

観測評点を表す変数

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

評価者 r の特性パラメータ
- 一貫性 (α_r)
- 厳しさ (β_r)
- 得点別の厳しさ (d_{rm})

受検者 j の能力 (θ_j)

課題 i の特性パラメータ
- 識別力 (α_i)
- 困難度 (β_i)

能力値とパラメータは
観測評点のデータから推定

	課題1			課題2		
	評価者1	評価者2	評価者3	評価者1	評価者2	評価者3
受験者1	3	2	1	3	1	1
受験者2	5	4	2	3	3	1
受験者3	5	5	3	4	4	2
受験者4	5	3	2	3	2	1
受験者5	5	4	2	4	2	1

⇨ **評価者と課題の特性を考慮した高精度な能力評価が可能**

モデルの性能評価

データ

30人の被験者に4つの小論文課題を行わせ，
被験者同士で相互に5段階評価させたデータ

従来モデル・素点平均との性能比較

1. 情報量基準に基づくモデル適合度の比較
2. 能力推定精度の比較

[評価方法] 評点データの一部をランダムに欠測させたデータを10パタン生成し，それらのデータから推定されたスコア間の相関に基づいて評価

性能評価結果

	情報量規準		能力推定精度	
	WAIC	周辺尤度		
提案モデル	**11384.58**	**11200.32**	**0.752**	
提案モデル with fixed d_{rk}	11492.09	11380.25	0.710	*
多相ラッシュモデル (1989)	11401.92	11242.64	0.705	*
Uto & Ueno (2016)	11471.67	11350.67	0.713	*
素点平均	-	-	0.672	*

* は提案モデルと比べて1%で有意差ありを表す

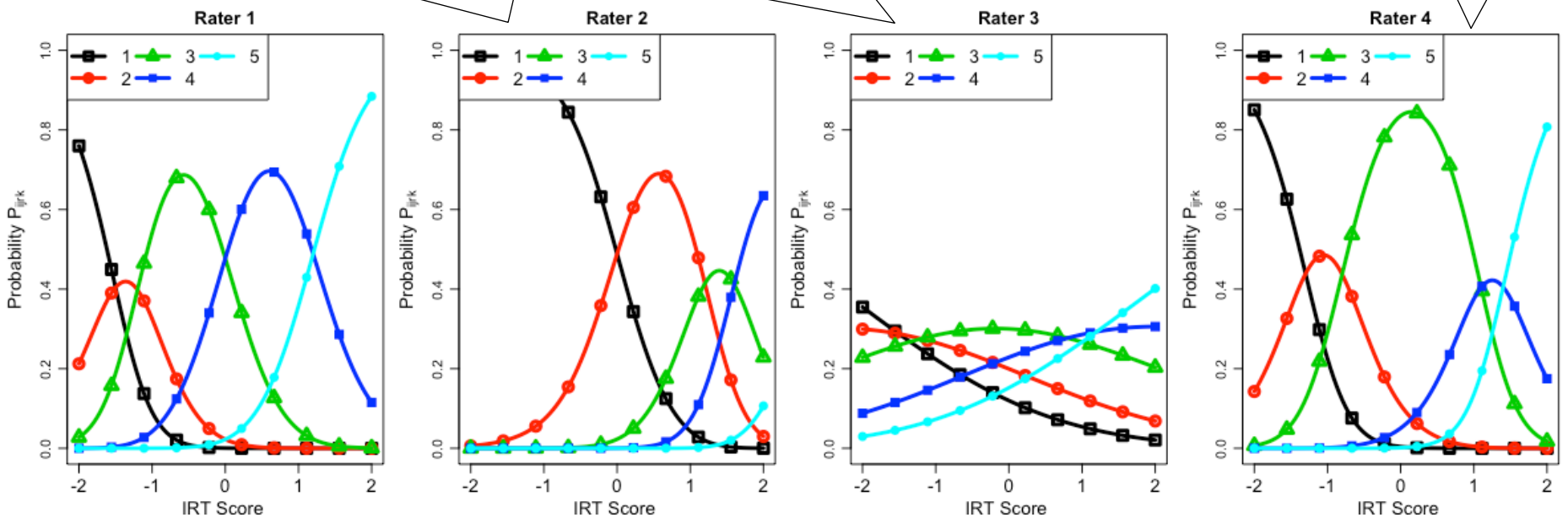
モデル適合度・能力推定精度ともに
本モデルが最高性能を達成

評価者特性の分析

低得点の使用率が高い
⇨ **厳しい評価者**

同レベルの受検者への評点がばらつく⇨ **一貫性が低い評価者**

得点が中心化



	一貫性	厳しさ	得点カテゴリへの厳しさ			
評価者1	1.5	0.0	-1.5	-1.2	0.0	1.2
評価者2	1.5	1.5	-1.5	-0.3	0.1	1.2
評価者3	0.2	0.0	-1.5	-1.2	0.8	1.2
評価者4	1.5	0.0	-1.3	-0.8	1.1	1.4

評価者へのフィードバックや
評価者研修などに活用可能

医療系大学間共用試験の実技試験 OSCE での実用

医療系大学間共用試験の実技試験OSCEにおいて実用化・実証実験が進行中

医学系OSCE（医療面接）例

歯学系OSCE（基本的臨床技能）例

*公益社団法人医療系大学間共用試験実施評価機構（2019）臨床実習開始前の「共用試験」第17版

- 宇都（2018, 2019, 2020）OSCEにおけるIRT利用について．公益社団法人医療系大学間共用試験実施評価機構 試験信頼性向上部会講演会.
- 宇都・森本・野上・内田・吉田・片桐・葛西・川上・江藤・齋藤・仁田（2020）OSCEにおける項目反応理論の適用．第52回医学教育学会全国大会.

ルーブリック評価への適用

ルーブリック評価のための技術拡張

	問題解決力 / 論理的思考力				
	評価項目1 (問題設定)	**評価項目2** (結論の導出)	**評価項目3** (根拠の提示)	**評価項目4** (対立意見の検討)	**評価項目5** (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

ルーブリック（採点基準表）を利用した評価の特徴

1. 評価者と課題に加え評価項目の特性にも評点が依存
2. 背後に複数次元の能力尺度が想定される場合がある

ルーブリック評価のための４相項目反応モデル

評価者と課題に加えて，評価項目の特性も考慮して受検者の能力を推定できるモデル

受検者 j の課題 i への回答に対し，評価者 r が評価項目 c に基づいて評点 k を与える確率を次式で定義

観測得点 (4相)

評価項目 c の特性パラメータ

識別力　困難度　得点カテゴリに対する基準

$$p(x_{ijrc} = k) = \frac{exp\sum_{m=1}^{k}(\alpha_i\alpha_r\alpha_c(\theta_j - \beta_i - \beta_r - \beta_c - \tau_r d_{cm}))}{\sum_{l=1}^{K} exp\sum_{m=1}^{l}(\alpha_i\alpha_r\alpha_c(\theta_j - \beta_i - \beta_r - \beta_c - \tau_r d_{cm}))}$$

宇都雅輝・植野真臣（2020）ルーブリック評価における項目反応理論．電子情報通信学会論文誌D. Vol.J103, No.05. pp.459-470.

ルーブリック評価のための多次元項目反応モデル

受検者の能力を多次元尺度で測定できるモデル

受検者 j の回答に対して，評価者 r が評価項目 c に基づいて評点 k を与える確率を次式で定義

$$P_{cjrk} = P^*_{cjrk} - P^*_{cjrk+1}$$

$$\begin{cases} P^*_{cjrk} = \left\{1 + \exp\left(-\alpha_r \left(\sum_{l=1}^{L} \alpha_{cl}\,\theta_{jl} - \beta_{ck} - \varepsilon_r\right)\right)\right\}^{-1} \\ P^*_{cjr0} = 1 \\ P^*_{cjrK} = 0 \end{cases}$$

能力の次元数

評価者 rの厳しさ

評価者 r の一貫性

評価項目cにおいて評点kを得る困難度

評価項目c の l次元目の識別力

受検者jのl次元目の能力

八木嵩大・宇都雅輝 (2019) パフォーマンス評価における多次元項目反応モデル. 電子情報通信学会論文誌D. Vol.J102, No. 10, pp.708-720.

ルーブリック評価のためのモデルの利点

1. 評価者や課題の特性に加えて評価項目の特性も考慮できるため，より高精度な能力測定が可能

2. ルーブリックや個々の評価項目の特性を定量的に分析可能

3. ルーブリック自体の評価・改善に応用可能

評価者割り当て最適化への応用

評価者割り当ての最適化

評価場面では，個々の回答に数名の評価者を割り当てて採点を実施する場合がある

この場合，評価者の割り当て方に評価精度が依存

項目反応理論と整数計画法を用いて
評価者割り当てを最適化する手法を開発

- Uto, Nguyen& Ueno (2020) Group optimization to maximize peer assessment accuracy using item response theory and integer programming, IEEE Transactions on Learning Technologies.
- Nguyen Duc Thien・宇都雅輝・植野真臣 (2018) ピアアセスメントにおける項目反応理論を用いたグループ構成最適化. 電子情報通信学会論文誌D.

評価者割り当て最適化の考え方

各受検者に対して**フィッシャー情報量**を最大化するように評価者を割り当て

評価者 r が受検者 j をどの程度の精度で評価できるかを表す指標

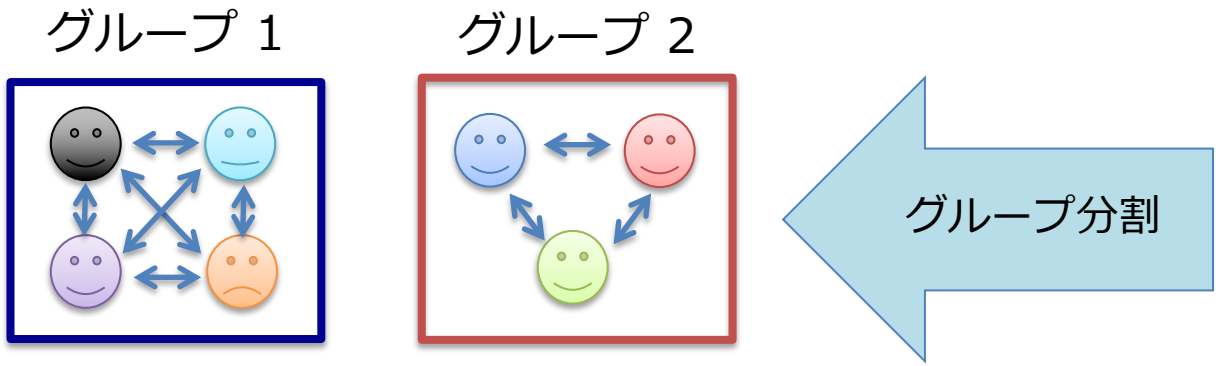
$$I_{ir}(\theta_j) = \alpha_i^2 \alpha_r^2 \sum_k \frac{\left(p^*_{ijrk}(1-p^*_{ijrk}) - p^*_{ijrk+1}(1-p^*_{ijrk})\right)^2}{p_{ijrk}}$$

現実の制約（評価者一人あたりの採点回数や受検者あたりの評価者数など）を満たす，最適割り当てを探索するために整数計画問題として定式化

相互評価のグループ構成問題として提案

MOOCsのような大規模学習環境では，レポートなどの評価を学習者同士の相互評価で行うことが多い

評価負担の軽減のために小グループに分割し，グループ内のメンバ同士で評価を実施



評価者割り当ての一つと解釈できる

グループ構成最適化手法

各学習者に与えられるフィッシャー情報量の下限を最大化するMin-Max整数計画問題として定式化

$$maximize: \quad y$$

$$subject\ to: \quad \sum_{\substack{r\in\mathcal{J} \\ r\neq j}} \sum_{g\in\mathcal{G}} I_{ir}(\theta_j)\, x_{igjr} \geq y$$

学習者 j に対する情報量

学習者に与えられる情報量の下限 y に対する制約式

$$\sum_{g\in\mathcal{G}} x_{igjj} = 1$$

各学習者が単一のグループに属することを制約式

$$n_l \leq \sum_{j} x_{igjj} \leq n_u$$

各グループの構成人数に関する制約式

$$x_{igjr} = x_{igrj}$$

情報量を最大化する最適な評価者割り当てを行うことで，
効率的に能力測定の精度を向上できたことを実験で確認

ここまでのまとめ

評価者バイアスの影響を取り除いて受検者の能力を推定できる項目反応モデルとその応用・実用例を紹介

人間評価者の主観採点を伴う様々な評価場面で利用可能

[関連論文]

- Uto (2020) Accuracy of performance-test linking based on a many-facet Rasch model. Behavior Research Methods, Springer.
- Uto &Ueno (2020) A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. Behaviormetrika, Springer.
- Uto, Nguyen & Ueno (2020) Group optimization to maximize peer assessment accuracy using item response theory and integer programming, IEEE Transactions on Learning Technologies.
- Uto (2019) Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. International Conference on Artificial Intelligence in Education (AIED).
- 宇都（2020）テスト理論と人工知能に基づくパフォーマンス評価の新技術．教育システム情報学会論文誌.
- 宇都・植野（2020）ルーブリック評価における項目反応理論. 電子情報通信学会論文誌D.
- 八木・宇都 (2019) パフォーマンス評価における多次元項目反応モデル. 電子情報通信学会論文誌D.
- 宇都(2019) 論述式試験における評点データと文章情報を活用した項目反応トピックモデル. 電子情報通信学会論文誌D.

小論文試験・記述式試験の活用と問題

表現力や思考力などの高次の能力を測定する手法の一つとして注目

[問題点1] 採点結果が評価者の特性に依存

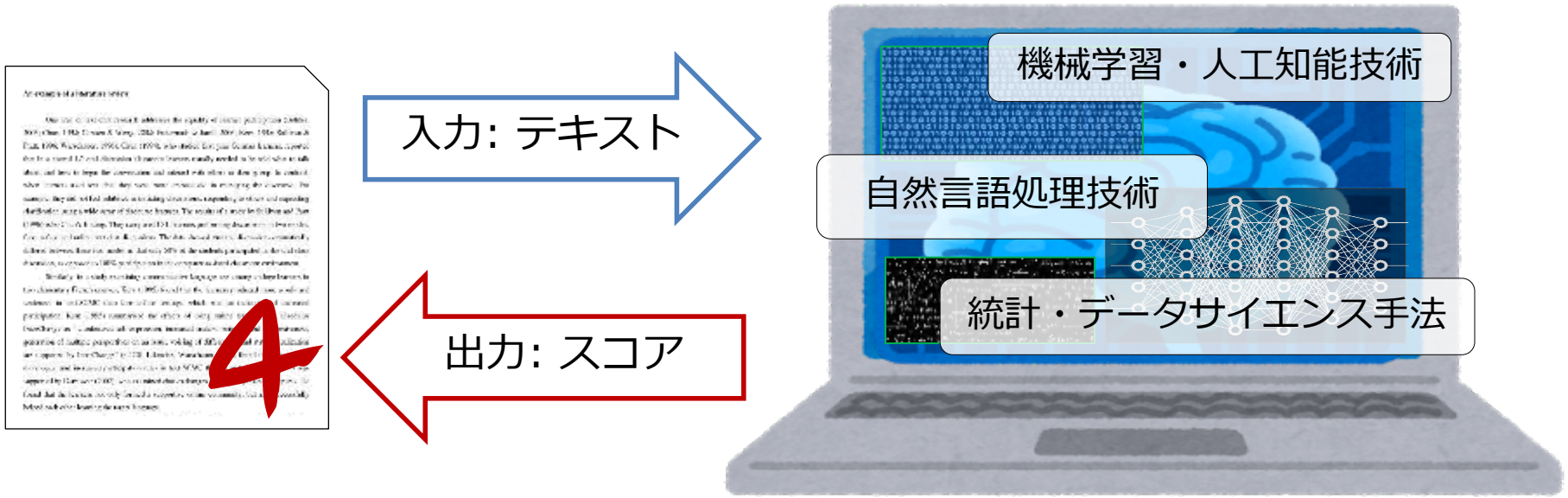
⇨ 評価者の特性を考慮して能力を推定できる
項目反応理論

[問題点2] 採点コストが膨大

⇨ 自動採点技術
項目反応理論を用いた新たな手法を紹介

小論文自動採点技術

人工知能技術を活用して人間評価者の代わりに採点



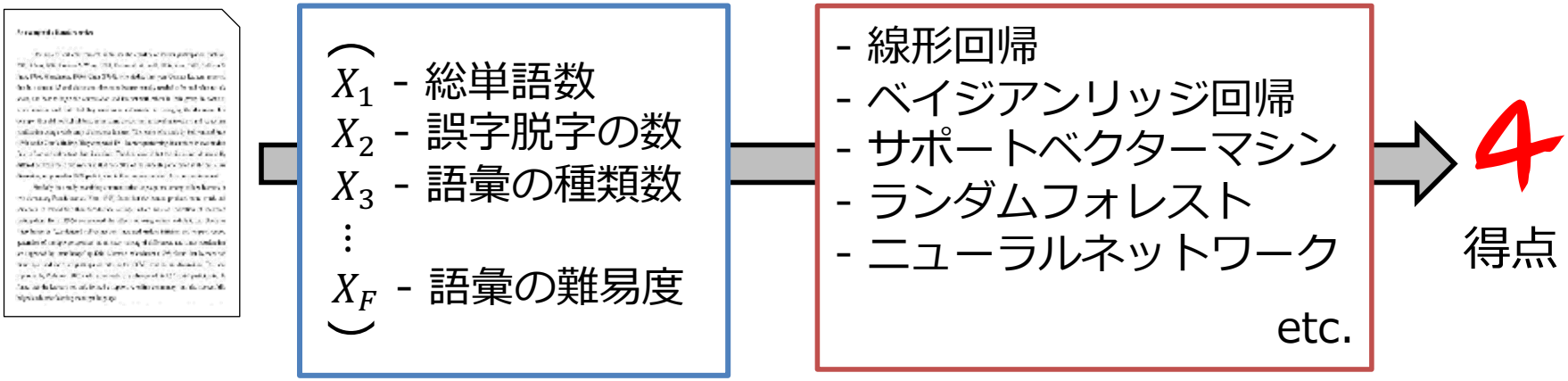
代表的な二つのアプローチ

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

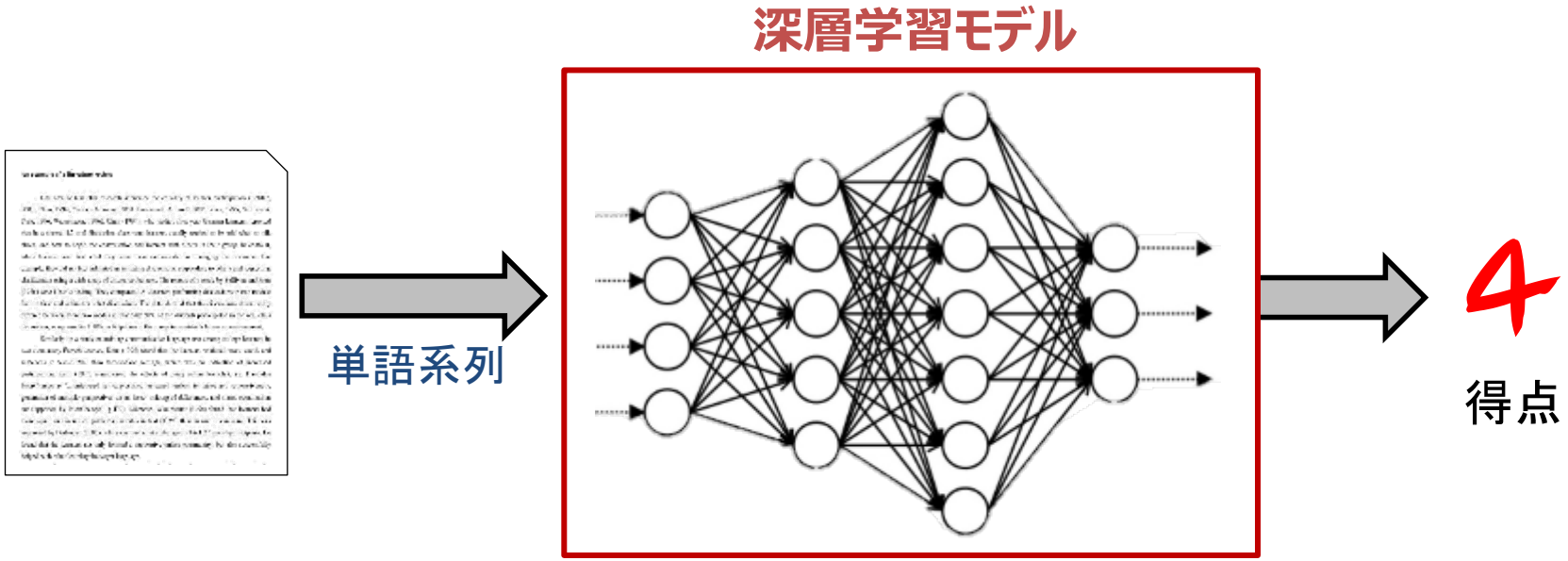
JESS：大学入試センターが開発した日本語対象のシステム

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

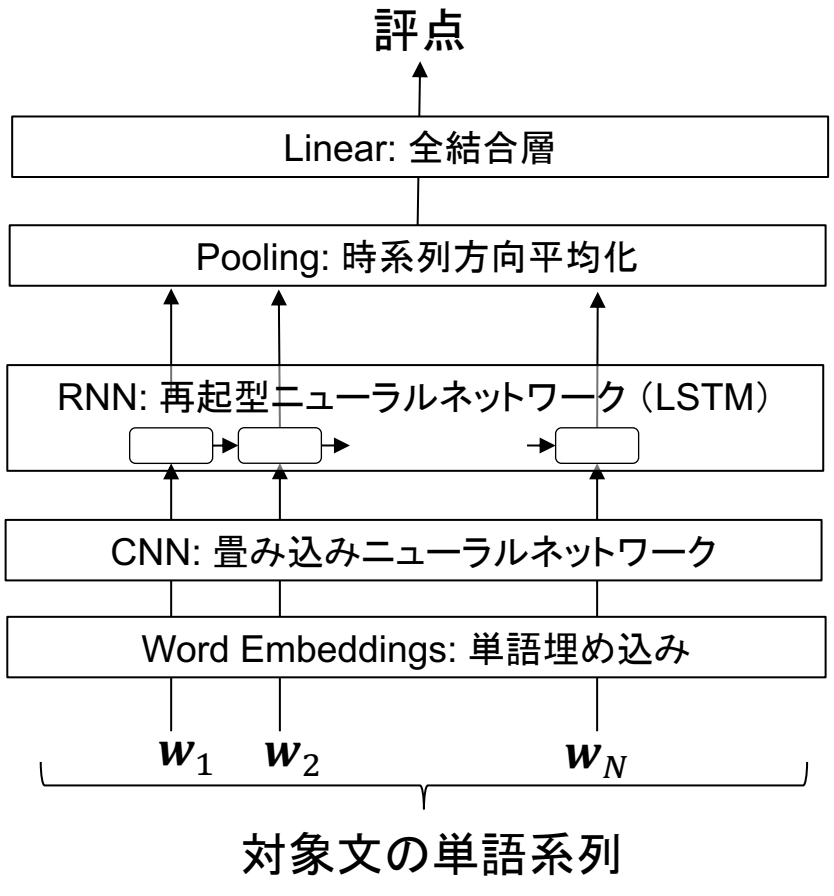


- 2016年頃に初期モデルが提案された新しいアプローチ
- 現在も人工知能・言語処理のトップカンファレンスで研究が続いており，高性能化が進行中

深層学習自動採点モデルの例

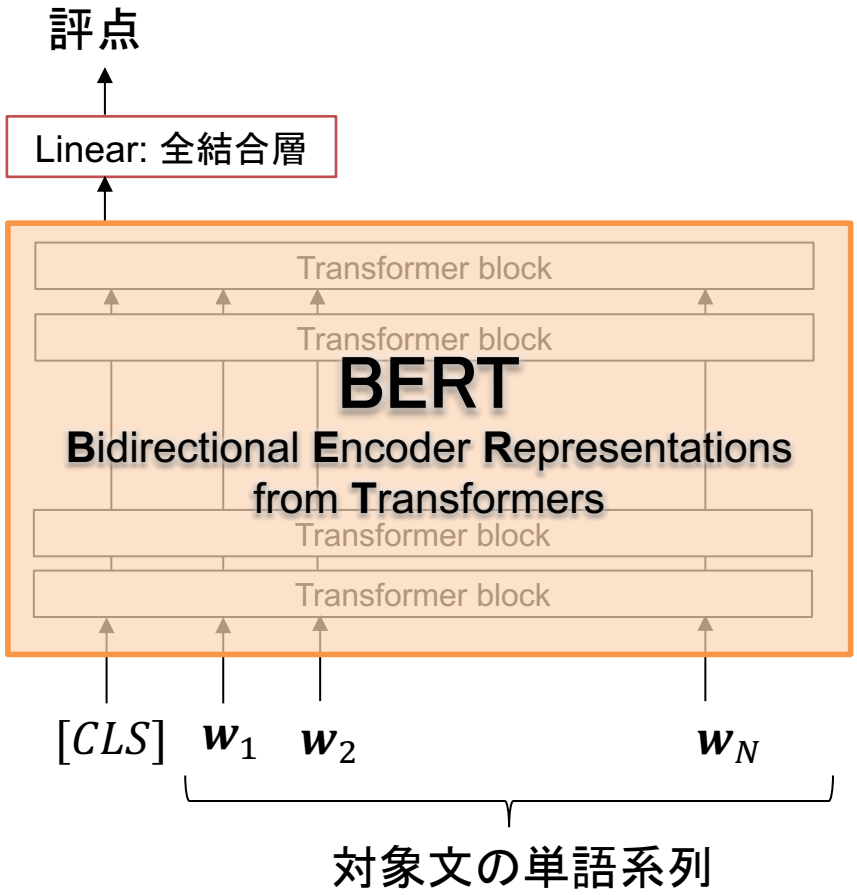
RNNベースモデル

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



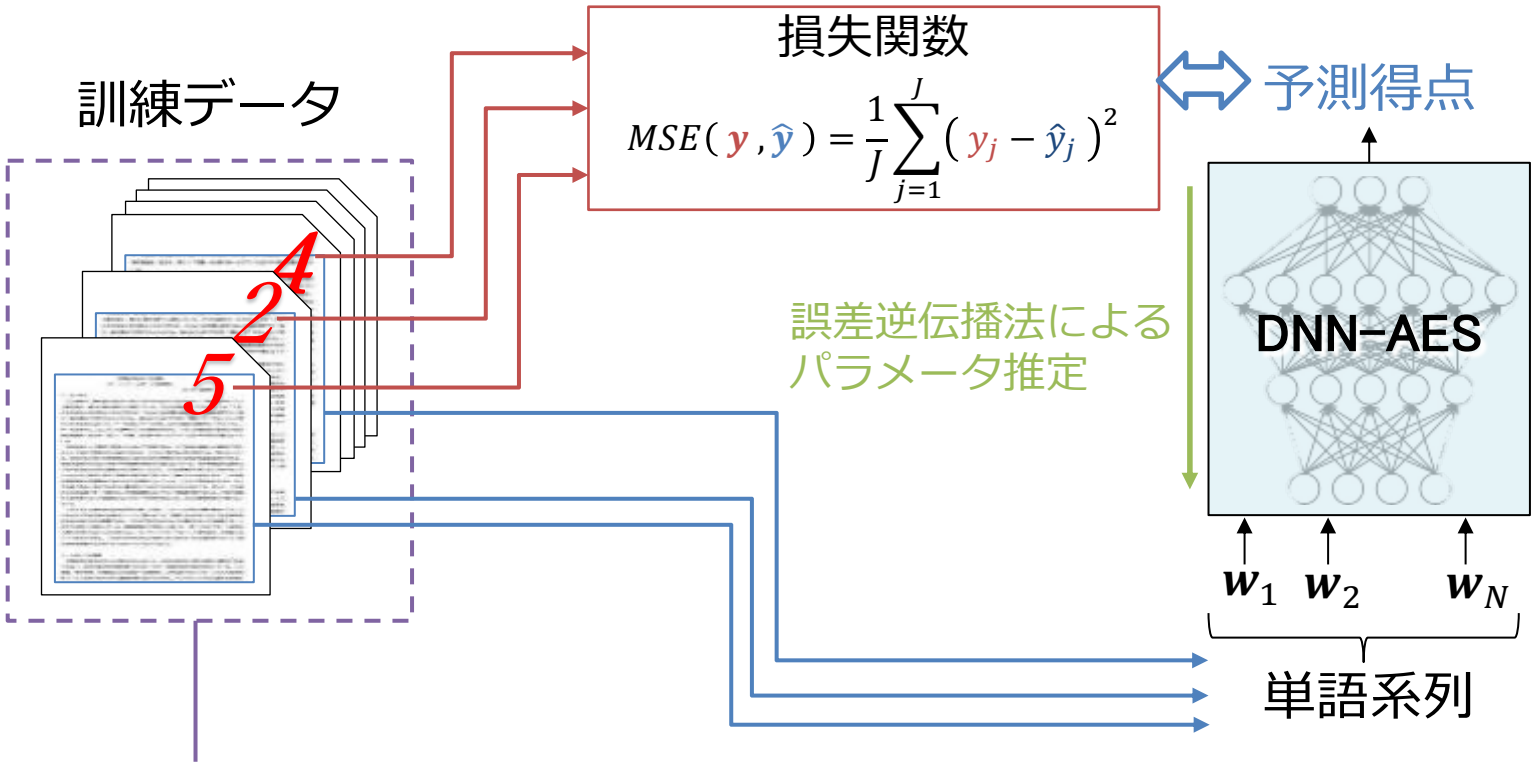
BERTベースモデル

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



自動採点モデルの学習

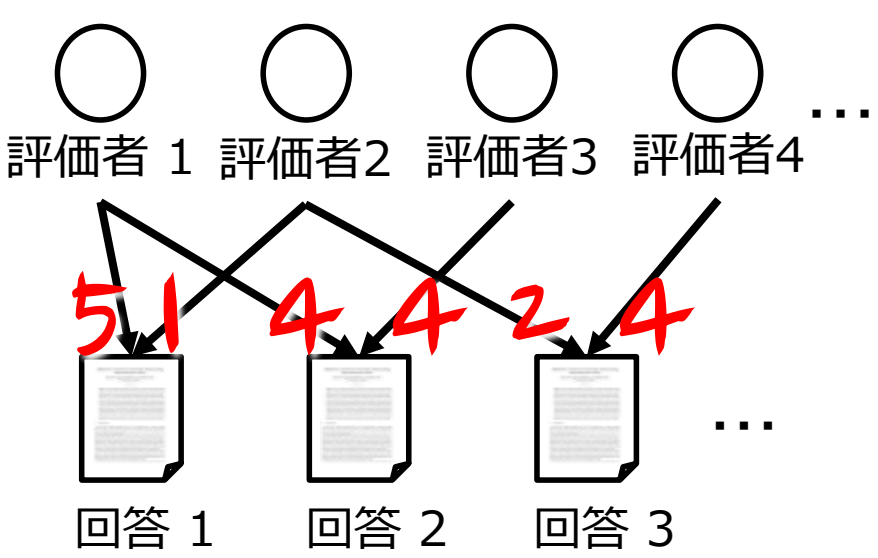
モデルを利用するためには，大量の採点済みの答案データを用いたモデルパラメータの学習が必要



訓練データ中の各答案への得点は正しいと仮定

評価者バイアスの影響

訓練データ作成の際，大量の答案の採点作業は複数の評価者で分担して行われることが多い



	評価者				平均
	1	2	3	4	
回答 1	5	1	-	-	3
回答 2	4	-	4	-	4
回答 3	-	2	-	4	3

自動採点モデルは
平均点に基づいて学習

平均点や合計点は評価者の特性に強く依存

⇨ 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，予測性能が低下

項目反応理論を用いた頑健な自動採点モデル

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>

観測評点データ

	評価者			
	1	2	3	4
回答 1	5	1	-	-
回答 2	4	-	4	-
回答 3	-	2	-	4

項目反応モデル

$$P(U_{jr}=k) = \frac{exp\sum_{m=1}^{k}[\alpha_r(\theta_j - b_r - d_{rm})]}{\sum_{l=1}^{K} exp\sum_{m=}^{l}[\alpha_r(\theta_j - b_r - d_{rm})]}$$

IRTスコア θ	
回答 1	θ_1
回答 2	θ_2
回答 3	θ_3

このスコアを利用して
自動採点モデルを学習

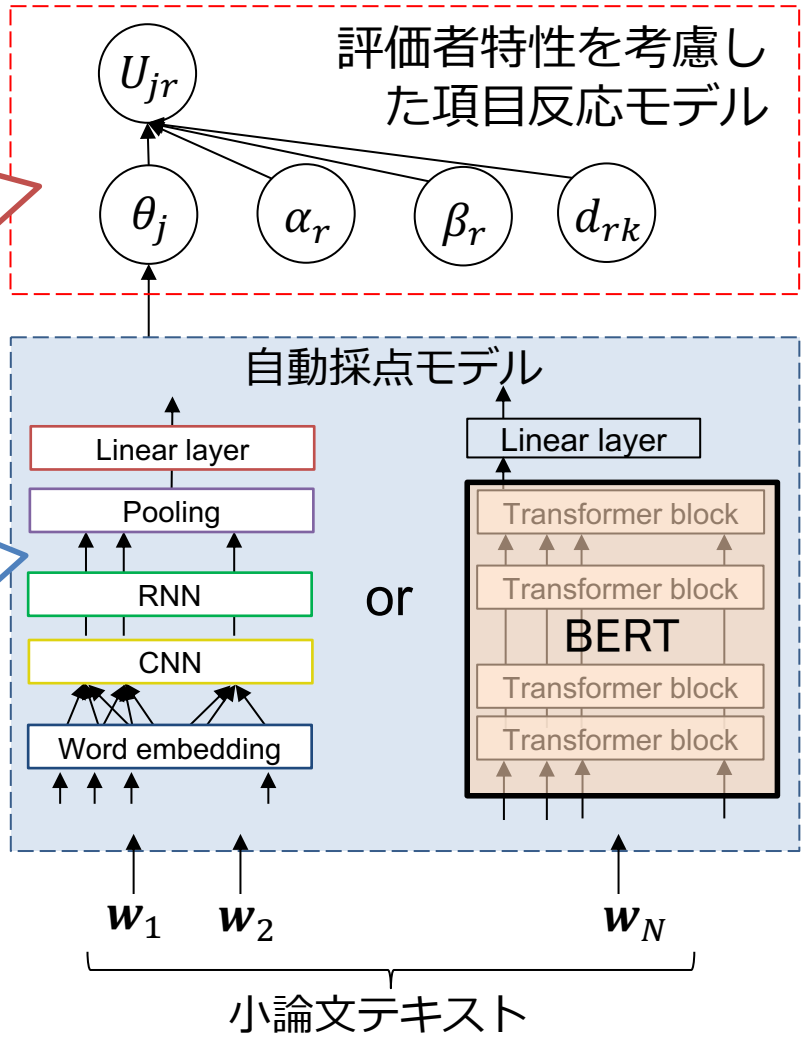
提案手法：モデル学習

手順1: IRTスコアの推定

観測評点のデータから，評価者バイアスの影響を取り除いたスコアθ_j を推定

手順2: 自動採点モデルの学習

得られたスコアθを目的変数として，自動採点モデルを学習

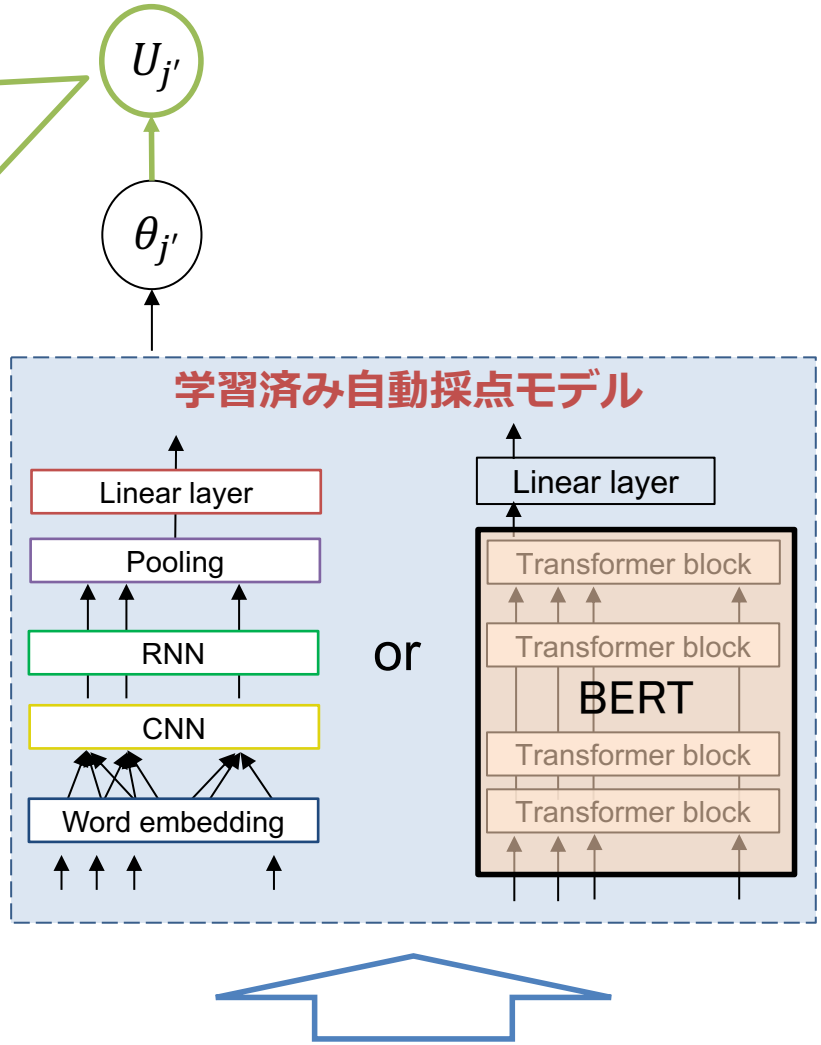


提案手法：得点予測

手順2：元の得点尺度に合わせるために，$\theta_{j'}$を所与として次式で期待得点を計算

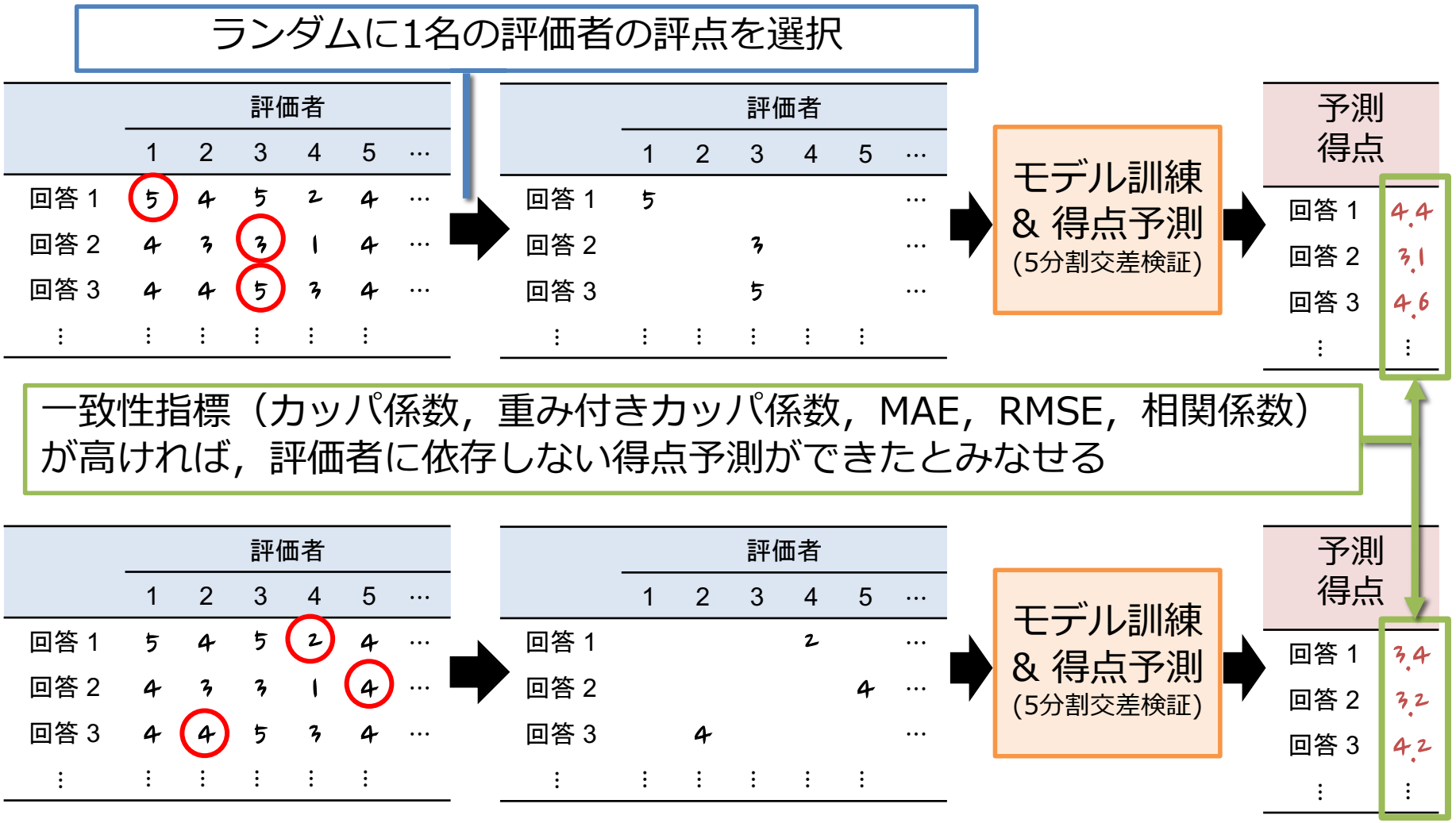
$$U_{j'} = \frac{1}{R}\sum_{r=1}^{R}\sum_{k=1}^{K} k \cdot P\left(U_{j'r} = k|\theta_{j'}\right)$$

手順1: 小論文テキストを学習済み自動採点モデルに入力しIRTスコア$\theta_{j'}$を予測



評価実験

個々の小論文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

項目反応理論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

その他の自動採点関係の研究

特徴量を組み込んだ深層学習自動採点モデル

Uto, Xie & Ueno (2020) Neural Automated Essay Scoring Incorporating Handcrafted Features. Proceedings of the 28th International Conference on Computational Linguistics (COLING).

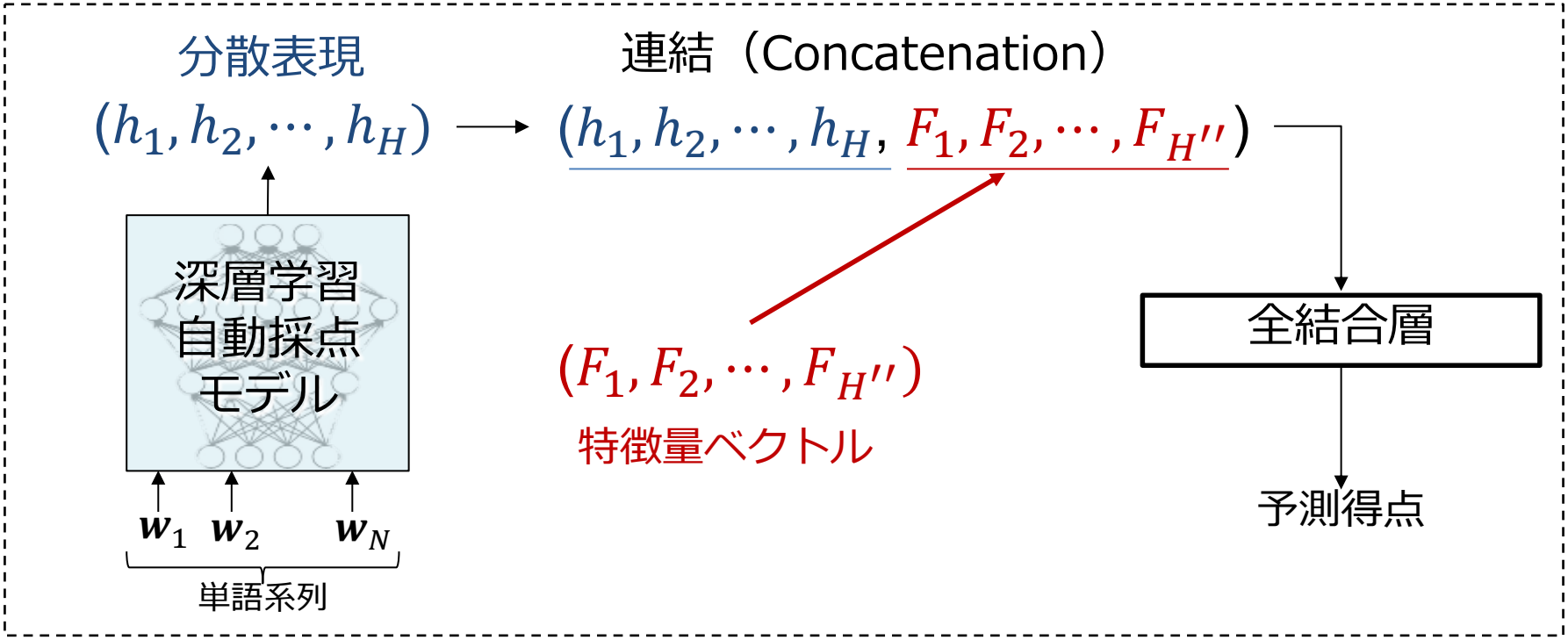
⇨ **小論文自動採点タスクにおいて最高精度を達成**

項目反応理論を用いた短答記述式問題自動採点手法

Uto& Uchida (2020) Automated short-answer grading using deep neural networks and item response theory. International Conference on Artificial Intelligence in Education (AIED).

特徴量を組み込んだ深層学習自動採点モデル

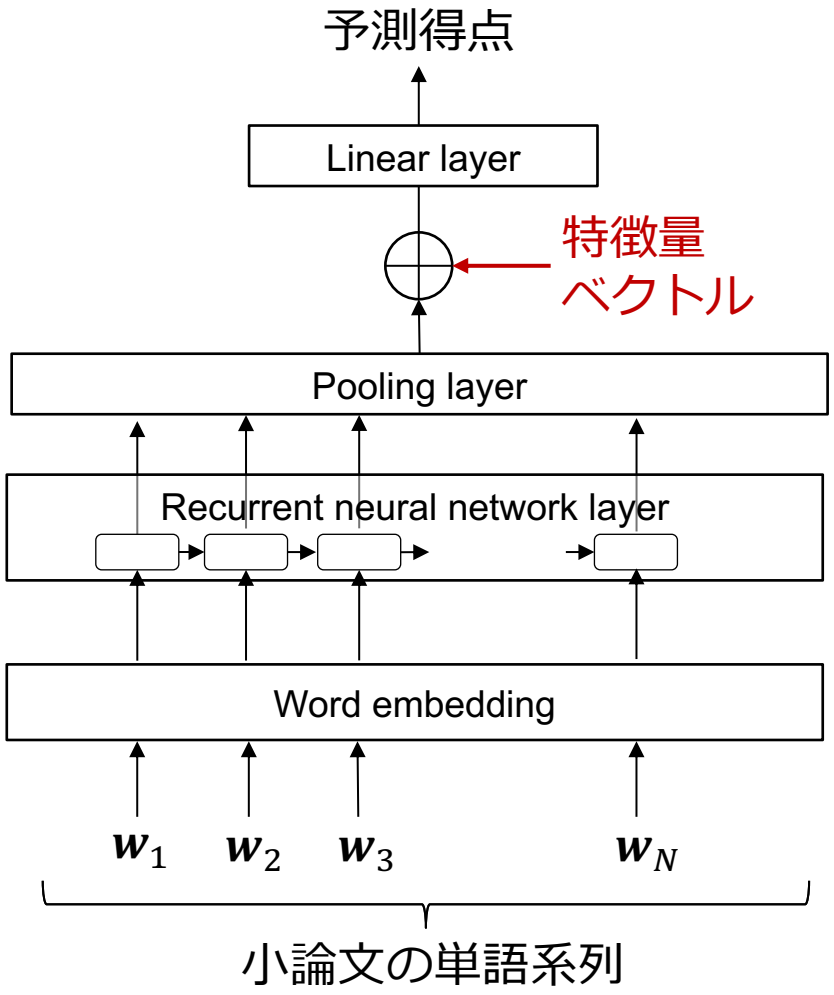
深層学習自動採点モデルに人手で設計した特徴量を統合する方法を提案



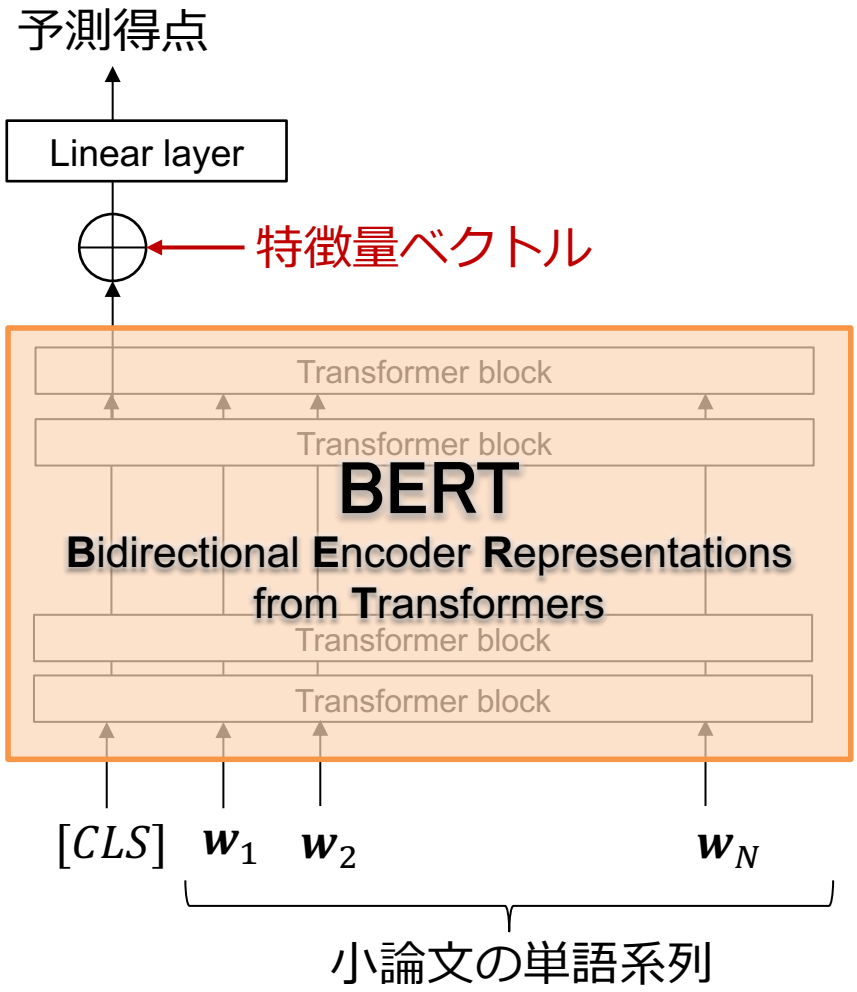
Uto, Xie & Ueno (2020) Neural Automated Essay Scoring Incorporating Handcrafted Features. Proceedings of the 28th International Conference on Computational Linguistics (COLING).

提案モデルの構成

RNNベースモデルへの組み込み



BERTベースモデルへの組み込み



精度評価

ベンチマークデータセット（ASAP）で予測性能を評価

5分割交差検証

評価指標：二次重み付きカッパ係数

ASAPデータセットの基礎情報

Prompt	# of essays	Score range	Average essay length
1	1783	2–12	350 words
2	1800	1–6	350 words
3	1726	0–3	150 words
4	1770	0–3	150 words
5	1805	0–4	150 words
6	1800	0–4	150 words
7	1568	0–30	250 words
8	721	0–60	650 words

	Prompt									
	1	2	3	4	5	6	7	8	Avg.	*p*-value
LSTM	0.373	0.407	0.516	0.773	0.753	0.767	0.635	0.174	0.550	0.018
+ Essay-level features	0.801	0.621	0.602	0.778	0.771	0.777	0.761	0.645	0.720	
LSTM with MoT	0.717	0.522	0.616	0.775	0.796	0.783	0.749	0.562	0.690	0.015
+ Essay-level features	0.821	0.649	0.617	0.790	0.787	0.807	0.794	0.694	0.745	
2-layer LSTM	0.435	0.414	0.530	0.791	0.698	0.768	0.639	0.163	0.555	0.017
+ Essay-level features	0.778	0.620	0.592	0.779	0.779	0.769	0.762	0.643	0.715	
Bidirectional LSTM	0.484	0.419	0.500	0.777	0.738	0.721	0.625	0.218	0.560	0.014
+ Essay-level features	0.779	0.597	0.582	0.778	0.762	0.765	0.756	0.661	0.710	
BERT	0.829	0.391	0.762	0.886	0.876	0.584	0.818	0.540	0.711	0.021
+ Essay-level features	**0.852**	0.651	**0.804**	**0.888**	**0.885**	**0.817**	**0.864**	0.645	**0.801**	
Conventional hybrid	0.729	0.635	0.631	0.787	0.802	0.793	0.773	0.693	0.730	0.073
+ Essay-level features	0.823	**0.674**	0.601	0.795	0.790	0.811	0.806	**0.714**	0.752	
Logistic regression	0.822	0.648	0.666	0.704	0.783	0.672	0.724	0.600	0.702	-

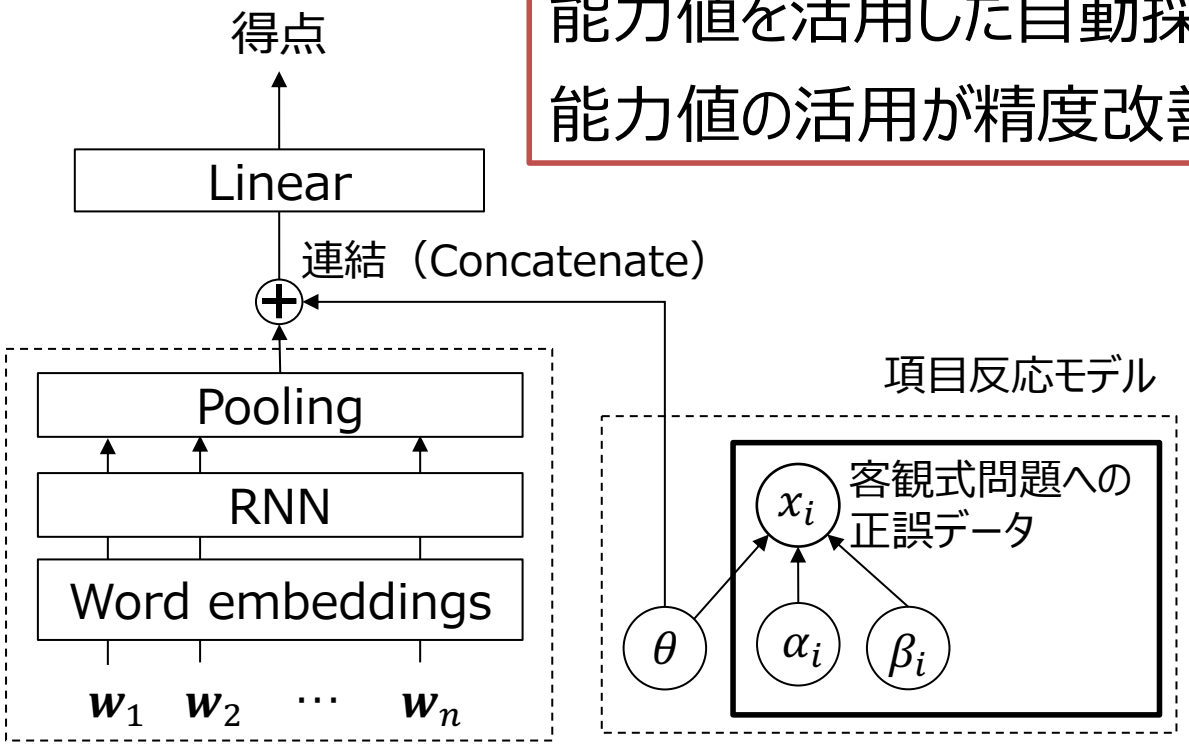
小論文自動採点タスクにおいて最高精度を達成

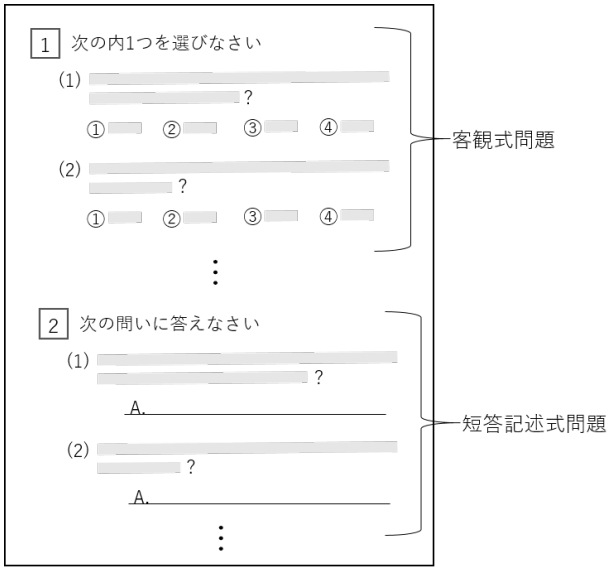
短答記述式問題の自動採点モデル

短答記述式問題が客観式問題を含むテストの一部として出題される場合，それらの問題が測定する能力は部分的に共通すると仮定

客観式テストの正誤情報から推定される受検者の能力値を活用した自動採点モデルを開発

能力値の活用が精度改善に寄与することを確認





Uto & Uchida (2020) Automated short-answer grading using deep neural networks and item response theory. International Conference on Artificial Intelligence in Education (AIED).

まとめ

[問題点1] 採点結果が評価者の特性に依存

- 評価者特性を考慮した項目反応理論とその応用例を紹介した
- 人間の主観採点を伴う様々な評価場面で信頼性改善に有効な技術

[問題点2] 採点コストが膨大

- 項目反応理論を活用した自動採点技術を紹介した
- 自動採点は実用化に向けた更なる技術発展が必要（少数訓練データからの学習など）
- 学習支援などへの応用・発展も望まれる（採点根拠の提示や自動フィードバックなど）

本科研費に関する現在までの研究成果（2019年度以降の査読付き論文）

1. Uto (2020) Accuracy of performance-test linking based on a many-facet Rasch model. Behavior Research Methods, Springer.
2. Uto & Ueno (2020) A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. Behaviormetrika, Springer, Vol. 47, Issue. 2, pp. 469-496.
3. Uto, Nguyen & Ueno (2020) Group optimization to maximize peer assessment accuracy using item response theory and integer programming, IEEE Transactions on Learning Technologies, IEEE Computer Society, Vol.13, No.1, pp.91-106.
4. Uto, Xie & Ueno (2020) Neural Automated Essay Scoring Incorporating Handcrafted Features. Proceedings of the 28th International Conference on Computational Linguistics (COLING), pp.6077-6088.
5. Uto & Okano (2020) Robust neural automated essay scoring using item response theory. International Conference on Artificial Intelligence in Education (AIED), pp.549-561.
6. Uto & Uchida (2020) Automated short-answer grading using deep neural networks and item response theory. International Conference on Artificial Intelligence in Education (AIED), pp.334-339.
7. Uto (2019) Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. International Conference on Artificial Intelligence in Education (AIED), pp. 494-506.
8. 内田・宇都(2021) 受験者の能力を考慮した深層学習ベース短答記述式問題自動採点手法. 教育システム情報学会論文誌.
9. 宇都・植野（2020）ルーブリック評価における項目反応理論. 電子情報通信学会論文誌D. Vol.J103, No.05. pp. 459-470.
10. 八木・宇都 (2019) パフォーマンス評価における多次元項目反応モデル. 電子情報通信学会論文誌D. Vol.J102, No. 10, pp.708-720.
11. 宇都 (2019) 論述式試験における評点データと文章情報を活用した項目反応トピックモデル. 電子情報通信学会論文誌D. Vol.J102, No.8, pp.553-566.

本科研費に関する現在までの研究成果（2019年度以降の受賞歴）

1. **Best paper runner-up award, International Conference on Artificial Intelligence in Education (AIED)** (2020)
   Masaki Uto, Masashi Okano: Robust neural automated essay scoring using item response theory.
2. **人工知能学会 先進的学習科学と工学研究会 若手奨励賞**（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」
3. **日本行動計量学会奨励賞 肥田野直・水野欽司賞**（2019）
4. **NLP 若手の会第14回シンポジウム 萌芽研究賞**（2019）
   内田優斗・宇都雅輝「受験者の解答履歴データを組み込んだ短答式問題自動採点手法」
5. **人工知能学会 平成30年度 研究会優秀賞**（2019）
   宇都雅輝「レイティングデータとテキスト情報を用いて受験者の能力を推定する項目反応トピックモデルの提案」
6. **人工知能学会 先進的学習科学と工学研究会 若手奨励賞**（2019）
   八木嵩大・宇都雅輝「パフォーマンス評価における多次元段階反応モデルの提案と評価」

おわりに

シンポジウムのウェブサイト上に「報告論文集」として関連論文の一部を公開しています.

http://www.ai.lab.uec.ac.jp/kakenhi/

私の研究に関しては，一連の研究の概要を『パフォーマンス評価のための項目反応理論とその小論文自動採点への応用』(pp.121-126)でまとめております.

個別の研究の詳細はそれ以降の各論文を参照ください.