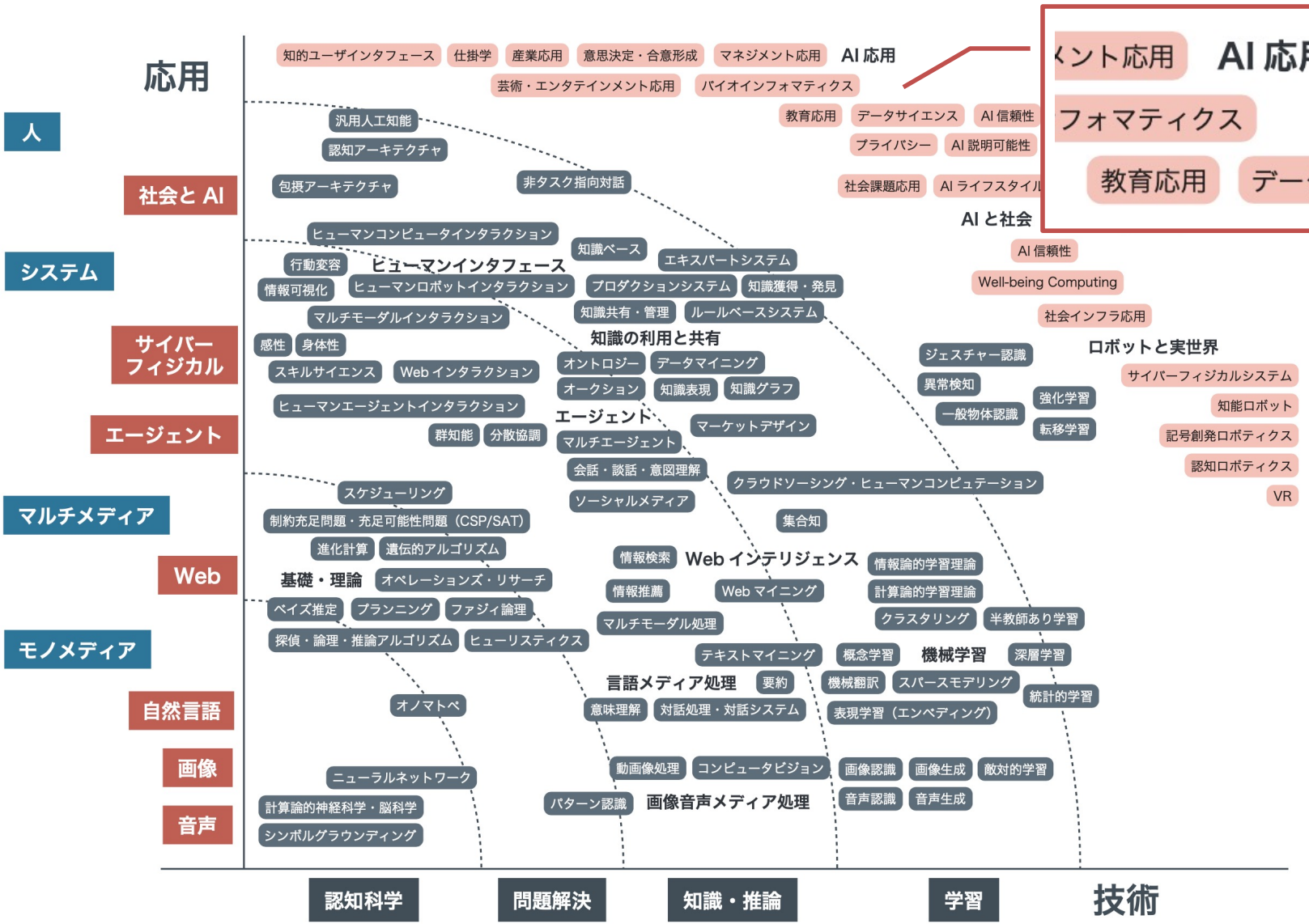
日本テスト学会第19回大会
実行委員会企画録画講演「テスティングの現状と将来展望」

テストとAI

電気通信大学 大学院情報理工学研究科 准教授
宇都 雅輝

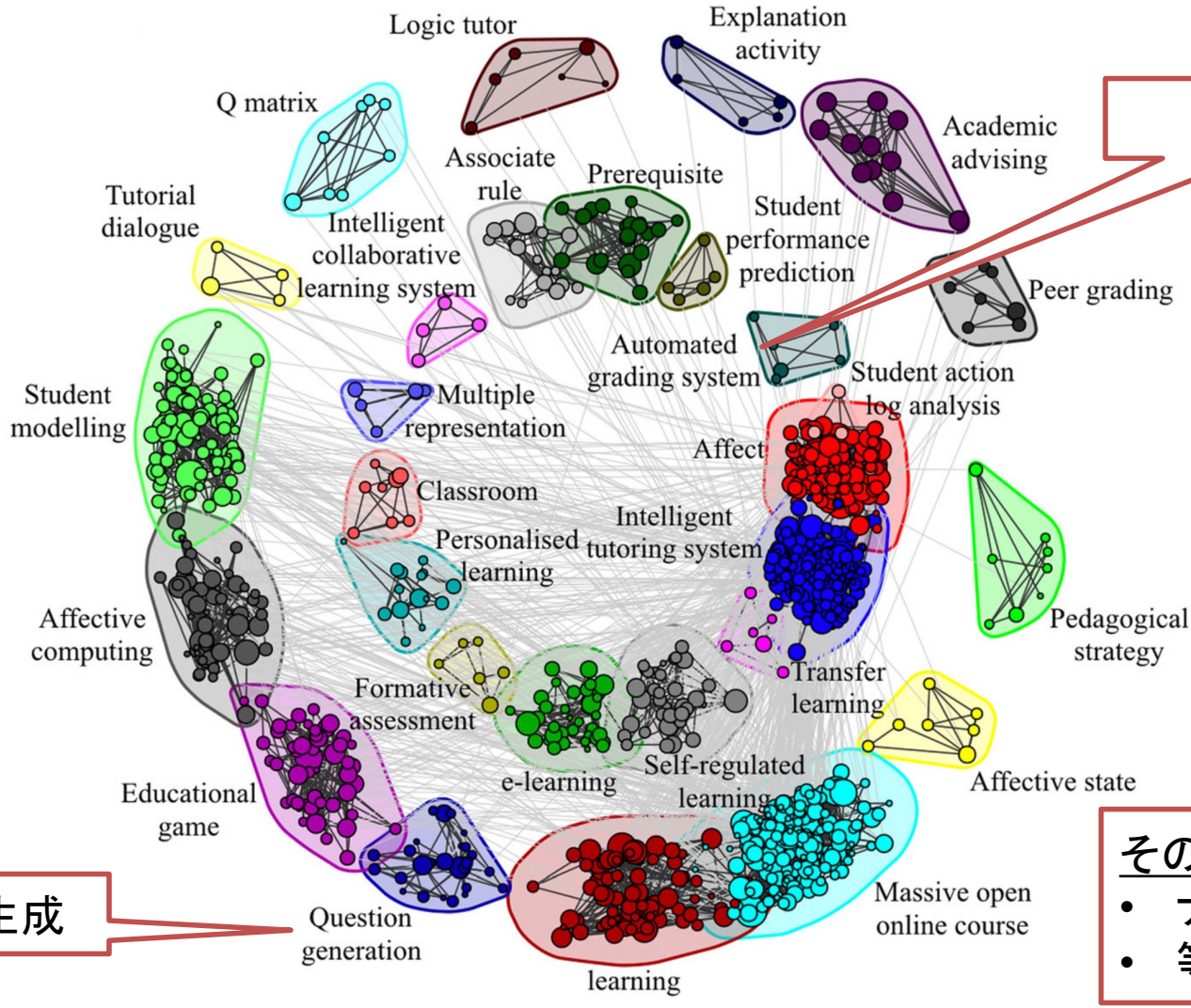
2021年 9月17日～26日

AI技術の発展と応用の拡大



AIマップβ 2.0（人工知能学会）https://www.ai-gakkai.or.jp/resource/aimap/

教育・テスティング分野におけるAI応用



自動採点

問題生成

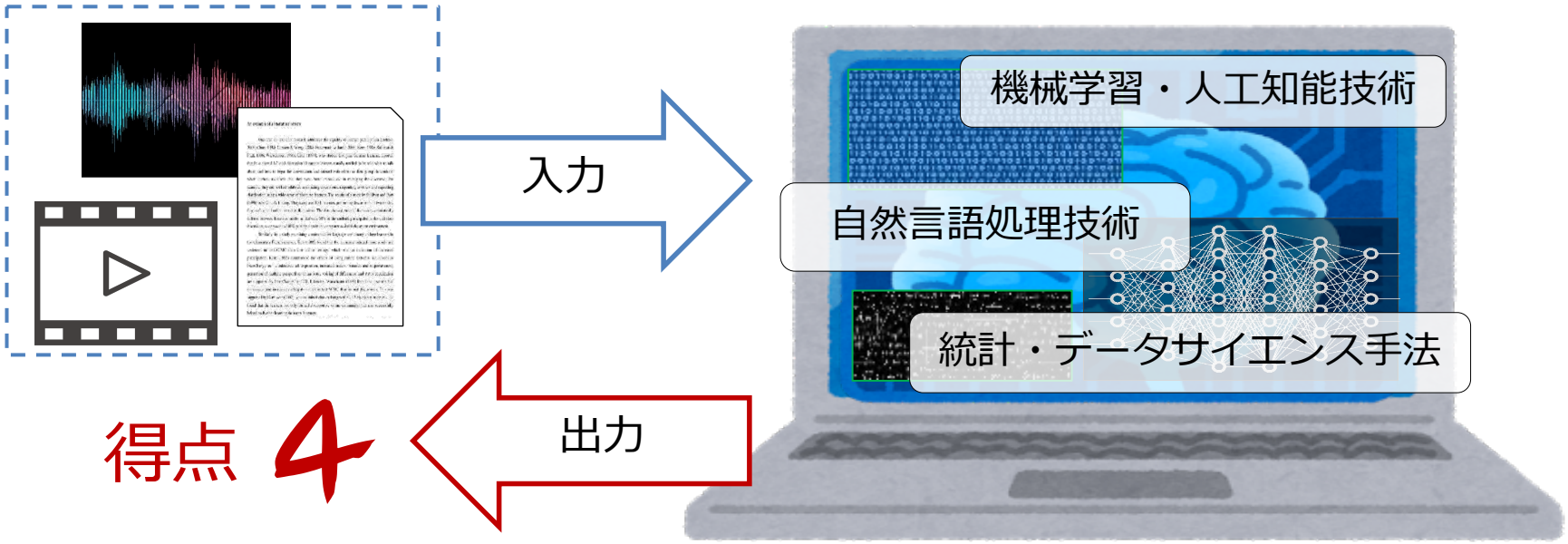
その他の応用タスク
- ナレッジトレーシング
- 等質テスト自動構成

Shihui & Nancy (2021) Mapping Artificial Intelligence in Education Research: a Network-based Keyword Analysis. International Journal of Artificial Intelligence in Education 31 (1).

自動採点技術

受検者のパフォーマンスを採点する人工知能技術

小論文・短答記述回答・スピーキング・プレゼン・面接 …



パフォーマンス評価の普及に伴い近年ニーズが増加

パフォーマンス評価

受験者に現実的な課題を与えて，その成果物やプロセスを評価者が採点する形式の評価

客観式評価では測定しにくい，実践的・現実的な能力測定を期待

[問題点]

1. **評価の信頼性に関わる問題**

   人間の評価者による主観的な採点を伴うため
   評価者の特性（甘さ/厳しさなど）に得点が依存

2. **評価にかかるコストの問題**

   大規模な試験では，採点にかかるコストが大きい

本講演の内容

パフォーマンス評価の二つの問題を解決する研究を紹介

1. 評価の信頼性に関わる問題

⇨ **評価者バイアスの影響を取り除いて能力を推定できる項目反応理論**

2. 評価にかかるコストの問題

⇨ **人工知能技術（深層学習）を利用した自動採点技術**

＃記述・論述式試験の自動採点を中心に紹介

本講演の内容

パフォーマンス評価の二つの問題を解決する研究を紹介

1. 評価の信頼性に関わる問題

⇨ **評価者バイアスの影響を取り除いて能力を推定できる項目反応理論**

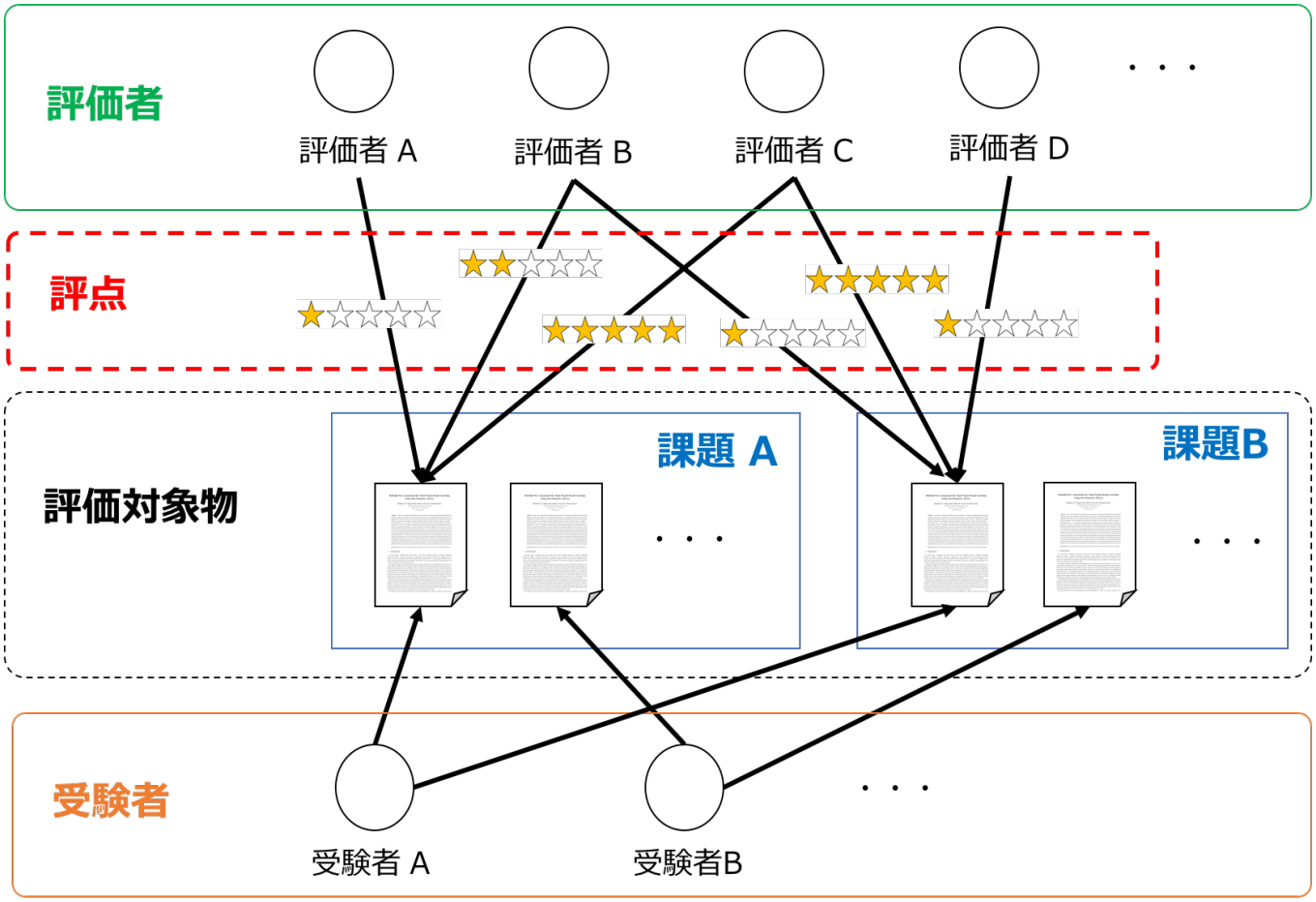
2. 評価にかかるコストの問題

⇨ **人工知能技術（深層学習）を利用した自動採点技術**

＃記述・論述式試験の自動採点を中心に紹介

一般的なパフォーマンス評価のデザイン

- 受験者は複数のパフォーマンス課題を実施
- 個々のパフォーマンスは複数の評価者が採点



パフォーマンス評価で得られるデータ

全ての評価者が全ての受験者を採点する場合

	課題1			課題2			平均点
	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	1.83
受験者2	5	4	2	3	3	1	3.00
受験者3	5	5	3	4	4	2	3.83
受験者4	5	3	2	3	2	1	2.67
受験者5	5	4	2	4	2	1	3.00
課題別平均点	3.40			2.33			
評価者別平均点				4.00	3.00	1.60	

- 課題ごとに平均点が異なる　⇨ 課題の難易度が異なる
- 評価者ごとに平均点が異なる ⇨ 評価の厳しさが異なる

全受験者が全課題に回答し全評価者が採点する状況であれば，「平均点」を利用しても公平といえる

しかし実際には受験者ごとに課題や評価者が異なることが一般的

受験者ごとに評価者や課題が異なる場合

受験者が異なる課題に回答した場合

	課題1			課題2			平均点（順位）	真得点（順位）
	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	1	1	1.67 (5)	1.83 (5)
受験者2				3	3	1	2.33 (4)	3.00 (2)
受験者3				4	4	2	3.33 (2)	3.83 (1)
受験者4	5	3	2				3.33 (2)	2.67 (4)
受験者5	5	4	2				3.67 (1)	3.00 (2)

⇨ 課題困難度が異なるため，どの課題を行なったかでスコアや順位が変動

受験者ごとに評価者が異なる場合

	課題1			課題2			平均点（順位）	真得点（順位）
	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			3			3.00 (3)	1.83 (5)
受験者2		4			3		3.50 (2)	3.00 (2)
受験者3			3			2	2.50 (5)	3.83 (1)
受験者4	5			3			4.00 (1)	2.67 (4)
受験者5		4			2		3.00 (3)	3.00 (2)

⇨ 評価者の厳しさが異なるため，誰が採点したかでスコアや順位が変動

代表的な評価者・課題バイアス

評価者特性によるバイアス要因

- **厳しさ･甘さ：**全体として低い（高い）得点を与える傾向の程度
- **一貫性：**測定対象能力を評点に反映する度合い
- **尺度範囲の制限：**特定の評点に評価が集中する傾向の程度

  例）5段階中3点に評価が集中（中心化傾向）

  例）5段階中1点と5点に評価が集中（極端化傾向）

課題特性によるバイアス要因

- **困難度**：得られる評点が全体として低くなる傾向の度合い
- **識別力**：能力が得点に反映される度合い

評価者と課題の特性を考慮した項目反応モデル

- Linacre (1989) “Many-faceted Rasch Measurement” MESA Press.
- 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.
- Uto & Ueno (2016). *Item response theory for peer assessment*. IEEE Transactions on Learning Technologies.
- Shin, Rabe-Hesketh & Wilson (2019) Trifactor models for multiple-ratings data. Multivariate Behavioral Research
- Uto & Ueno (2020) A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. Behaviormetrika.

  ⇨ **評価者と課題の特性を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

一般化部分採点モデルを多相化した拡張モデル

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**

パラメータの識別性

このモデルは無制約ではパラメータ値が一意に定まらない

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

識別性のための制約として以下を仮定

- $\theta_j \sim N(0, 1)$
- $\prod_i \alpha_i = 1$
- $\sum_i \beta_i = 0$
- $d_{r1} = 0 \; ; \; \forall r$
- $\sum_m d_{rm} = 0 \; ; \; \forall r$

パラメータ推定手法

マルコフ連鎖モンテカルロ法（MCMC）によるベイズ推定（Expected A Posteriori法を採用）

ハミルトニアンMCMCアルゴリズムの一種であるNo-U-turn samplerアルゴリズムを利用

実装にはStanを利用（＊コードは以下で公開しています）

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

完全データに適用した場合の例

	課題1			課題2			平均点	能力値 θ
	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	1.83 (5)	-1.18 (5)
受験者2	5	4	2	3	3	1	3.00 (2)	0.15 (2)
受験者3	5	5	3	4	4	2	3.83 (1)	1.19 (1)
受験者4	5	3	2	3	2	1	2.67 (4)	-0.13 (4)
受験者5	5	4	2	4	2	1	3.00 (2)	0.09 (3)
課題別平均点	3.40			2.33				
課題困難度	-0.52			0.89				
評価者別平均点				4.00	3.00	1.60		
評価者の厳しさ				-1.39	-0.21	1.59		

#評価者の厳しさと課題困難度以外の特性値は割愛

- 難しい課題ほど課題困難度パラメータ値が大きい
- 厳しい評価者ほど評価者の厳しさパラメータ値が大きい
- 能力値 θ は平均点と相関
  ⇨ この場合ではモデルの利用の意義は薄い

受験者ごとに評価者1名だった場合

	課題1			課題2			平均点	能力値 θ
	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			3			3.00 (3)	-1.22 (5)
受験者2		4			3		3.50 (2)	0.19 (2)
受験者3			3			2	2.50 (5)	1.10 (1)
受験者4	5			3			4.00 (1)	-0.34 (4)
受験者5		4			2		3.00 (3)	0.11 (3)

能力値 θ と平均点では全く傾向が異なる

※1. この例では能力スコアは課題と評価者特性値を所与として推定
※2. 課題・評価者の特性値が未知の場合，上記のデータでは
Linkageできないので注意（詳しくは下記文献等を参照）

Masaki Uto (2021) Accuracy of performance-test linking based on a many-facet Rasch model. Behavior Research Methods, Vol. 53, No. 4, pp. 1440-1454.

項目反応理論と素点平均の比較

平均点（順位）	
全データ	評価者1名
1.83 (5)	3.00 (3)
3.00 (2)	3.50 (2)
3.83 (1)	2.50 (5)
2.67 (4)	4.00 (1)
3.00 (2)	3.00 (3)

能力値 θ（順位）	
全データ	評価者1名
-1.18 (5)	-1.22 (5)
0.15 (2)	0.19 (2)
1.19 (1)	1.10 (1)
-0.13 (4)	-0.34 (4)
0.09 (3)	0.11 (3)

誰が採点したかでスコアも順序関係も大きく変動する

順序の変動はなく，数値の差異も非常に小さい

受験者ごとに課題が異なる場合も同様の傾向になる

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

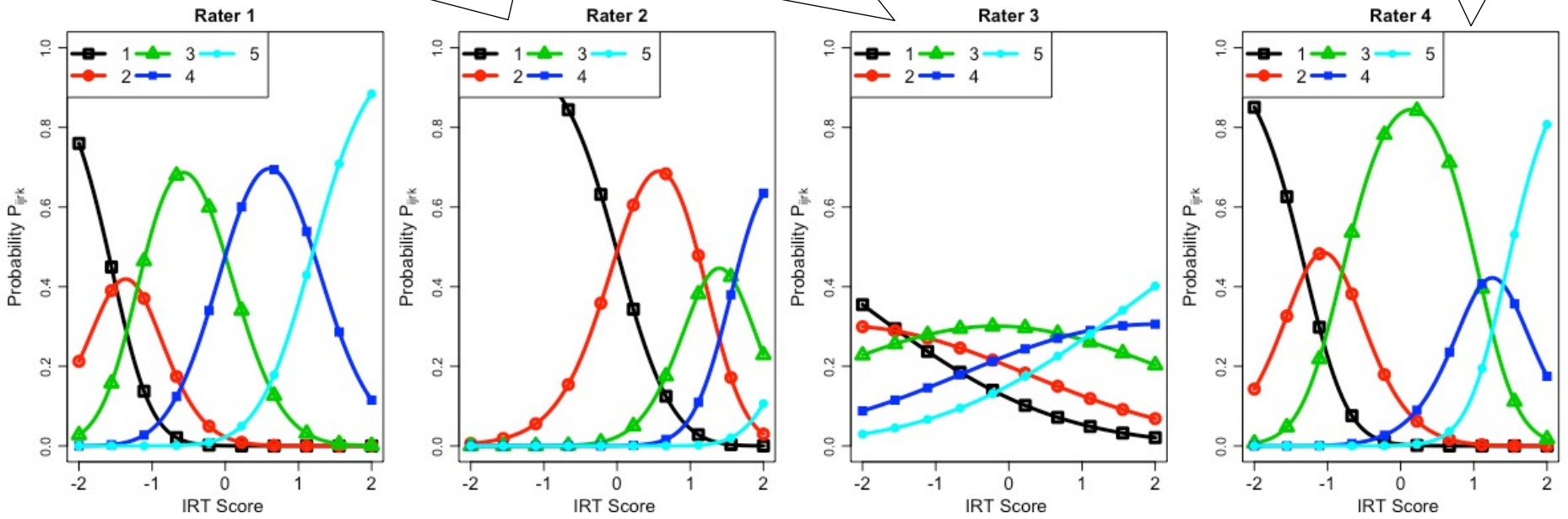
モデル適合度・能力推定精度ともに
本モデルが最高性能を達成

評価者特性の分析にも利用可能

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

ルーブリック評価への拡張

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

$$P\left(x_{ijr}=k\right)=\frac{exp\sum_{m=1}^{k}\left(\alpha_r\left(\sum_{l=1}^{L}\alpha_{cl}\ \theta_{jl}-\beta_c-\beta_r-d_{cm}\right)\right)}{\sum_{l=1}^{K}exp\sum_{m=1}^{l}\left(\alpha_r\left(\sum_{l=1}^{L}\alpha_{cl}\ \theta_{jl}-\beta_c-\beta_r-d_{cm}\right)\right)}$$

- 能力の次元数 L
- 評価項目cにおいて評点mを得る困難度 d_{cm}
- 評価項目c の l次元目の識別力 α_{cl}
- 受検者jのl次元目の能力 θ_{jl}
- 評価項目c の困難度 β_c

Masaki Uto (2021) A multidimensional generalized many-facet Rasch model for rubric-based performance assessment. Behaviormetrika, 48 (2) pp.425-457.

４相型多次元モデル

４相データから，評価者・課題・評価項目の特性を同時に考慮しつつ，多次元尺度上で能力を推定できる新しいモデルを一般セッションで発表しています．

一般セッション5（発表番号2）
『ルーブリック評価のための多次元4相型項目反応モデルの提案』新田 森，宇都 雅輝

本講演の内容

パフォーマンス評価の二つの問題を解決する研究を紹介

1. 評価の信頼性に関わる問題

⇨ **評価者バイアスの影響を取り除いて能力を推定できる項目反応理論**

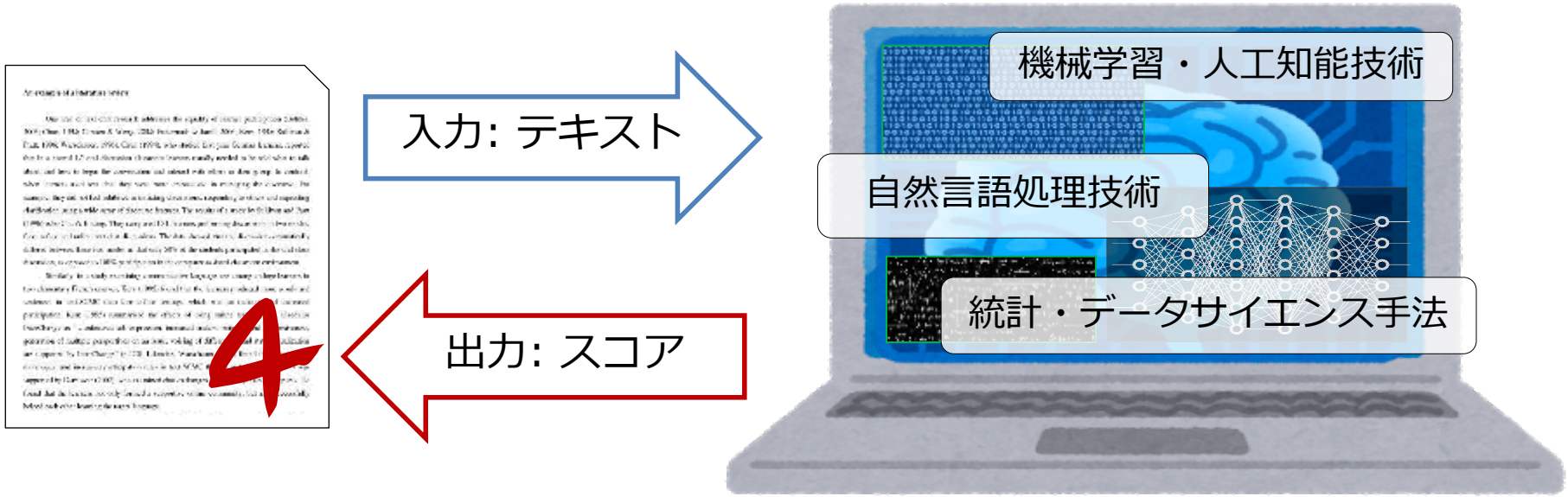
2. 評価にかかるコストの問題

⇨ **人工知能技術（深層学習）を利用した自動採点技術**

＃記述・論述式試験の自動採点を中心に紹介

記述・論述式自動採点技術

回答文に対する得点付けを自動化する技術



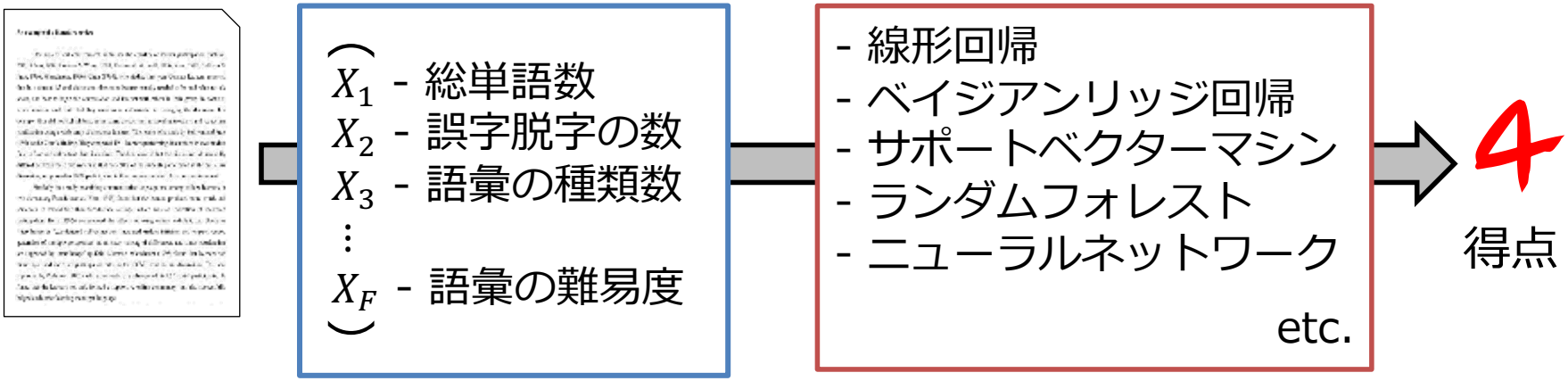
代表的な二つのアプローチ

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

特徴量ベースのアプローチの特徴

Pros.

- 特徴量設計が一度完了すれば比較的容易に利用可能
- どの特徴量がどのように得点に寄与するのかが分析できるため，採点根拠の解釈性が高い

Cons.

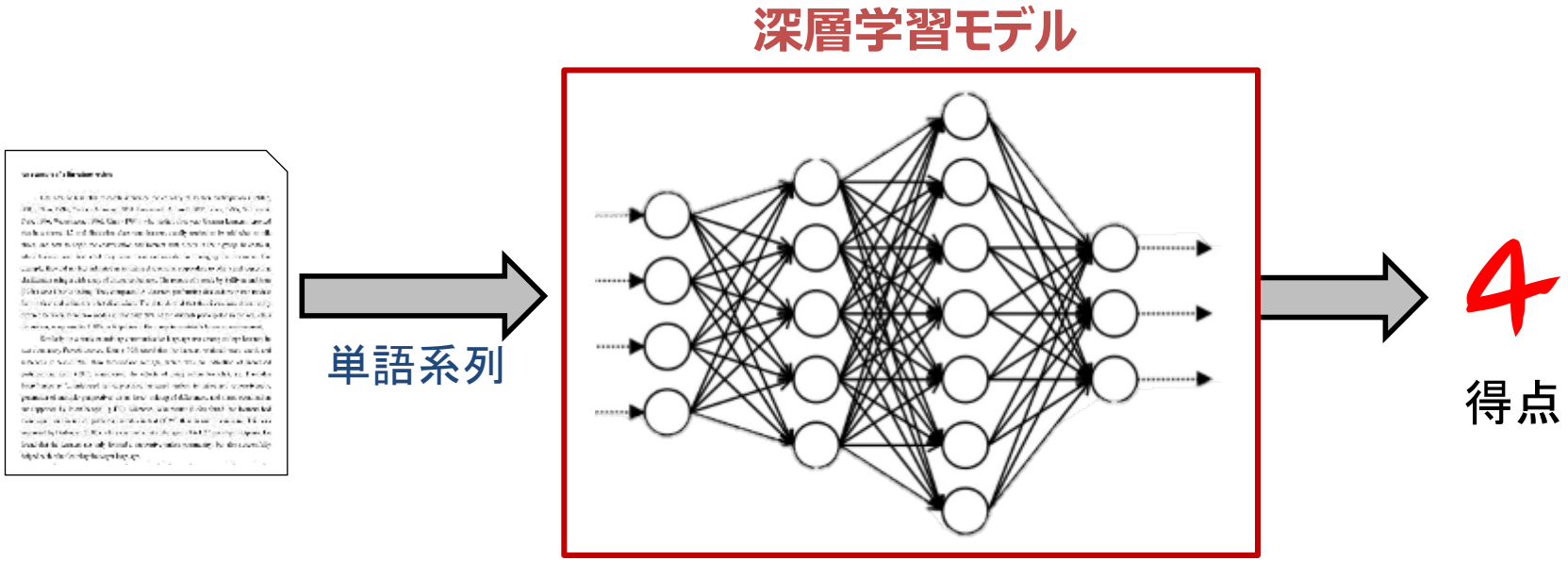
- 高精度の達成には特徴量の綿密なチューニングが必要
- 実装が比較的難しい特徴量が存在（構文解析や潜在意味解析，論理マイニングなど言語処理分野・人工知能分野の広い技術が必要）

⇨ 特徴量設計を省いて自動採点する手法が提案

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

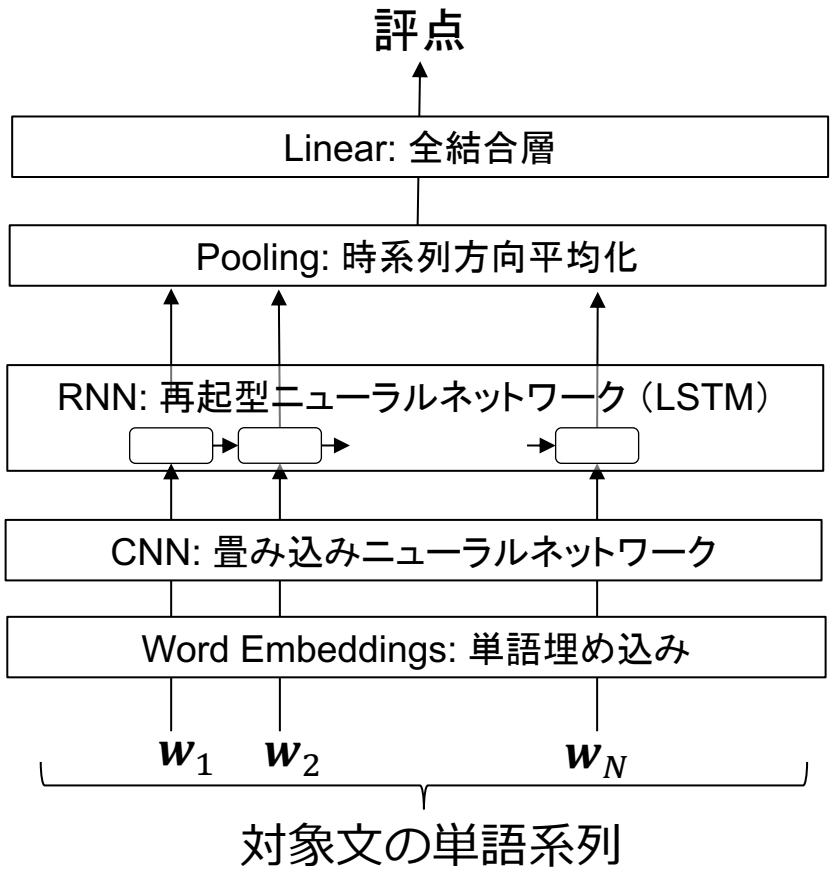


- 2016年頃に初期モデルが提案された新しいアプローチ
- 現在も人工知能・言語処理のトップカンファレンスで研究が続いており，高性能化が進行中

深層学習自動採点モデルの例

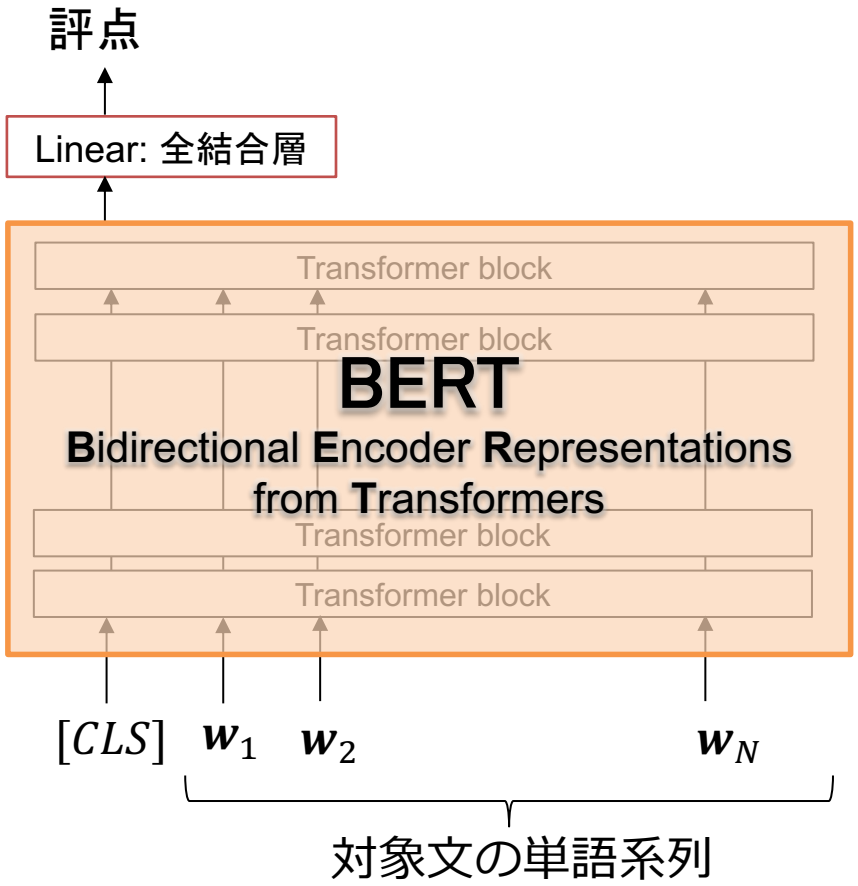
RNNベースモデル

Taghipour & Ng (2016) A neural approach to automated essay scoring. EMNLP.



BERTベースモデル

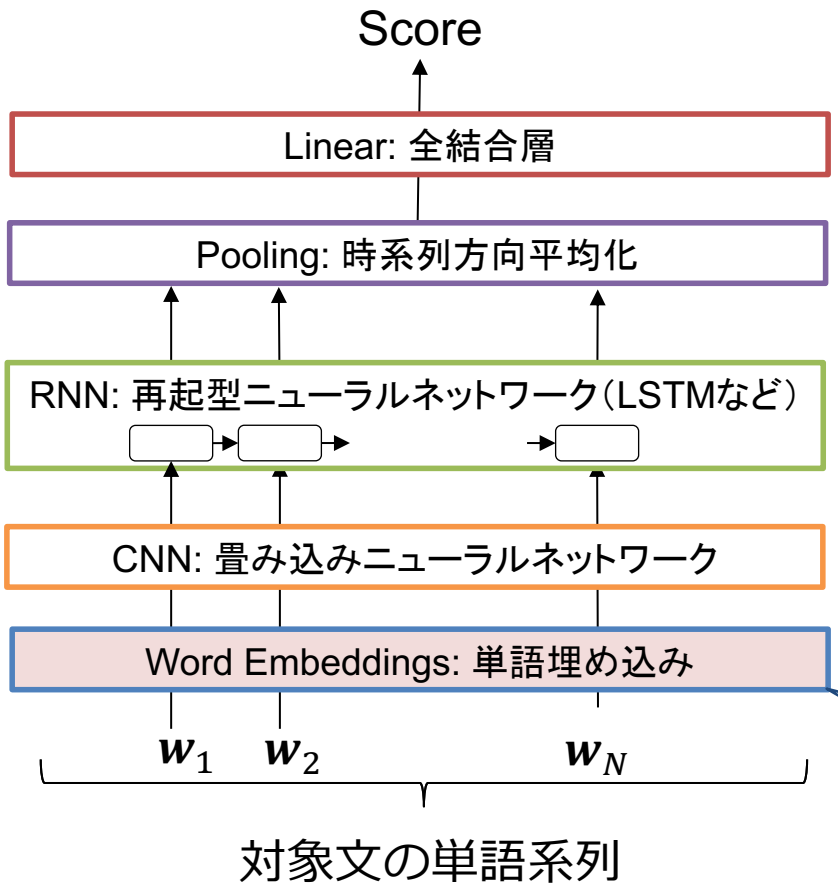
Devlin et al. (2018) *BERT: Pre-training of deep bidirectional Transformers for Language Understanding*. arXiv.



深層学習自動採点モデルの代表例

RNNベースモデル

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

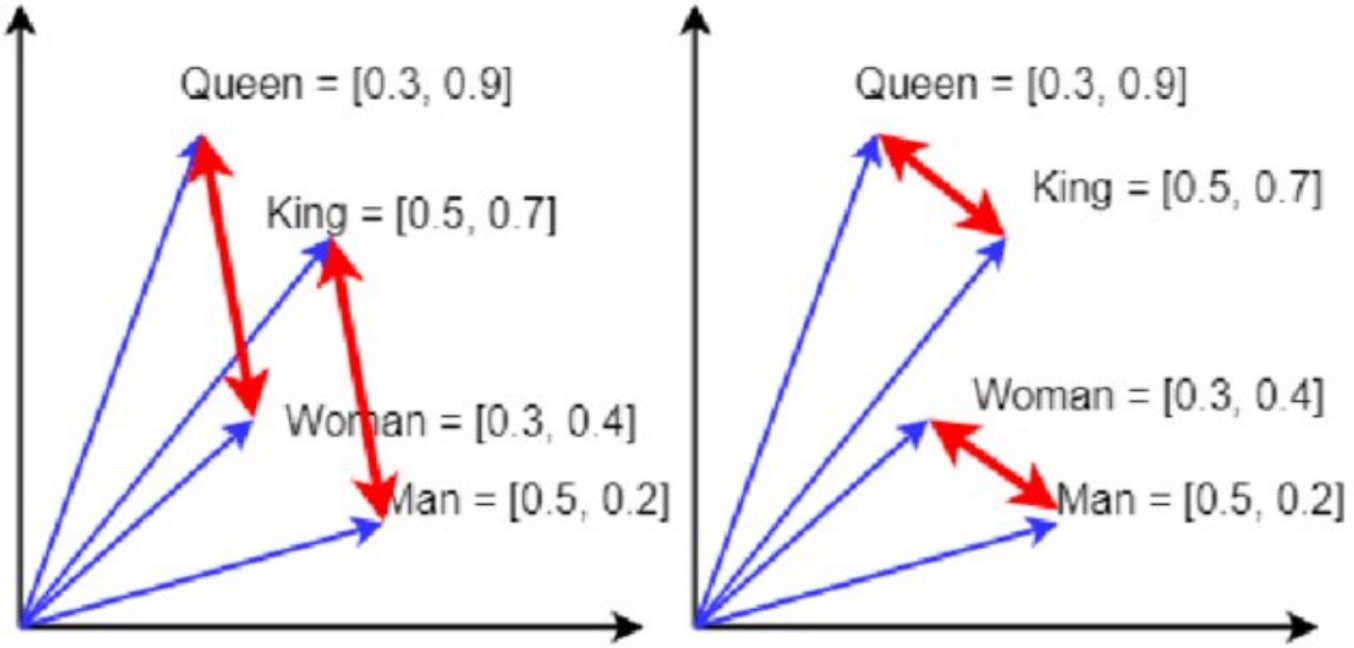


各単語を分散表現に変換

単語分散表現

類似した概念の単語同士が近いベクトル値を取るような低次元（数十～数百次元）の実数ベクトル表現

2次元の分散表現の例：

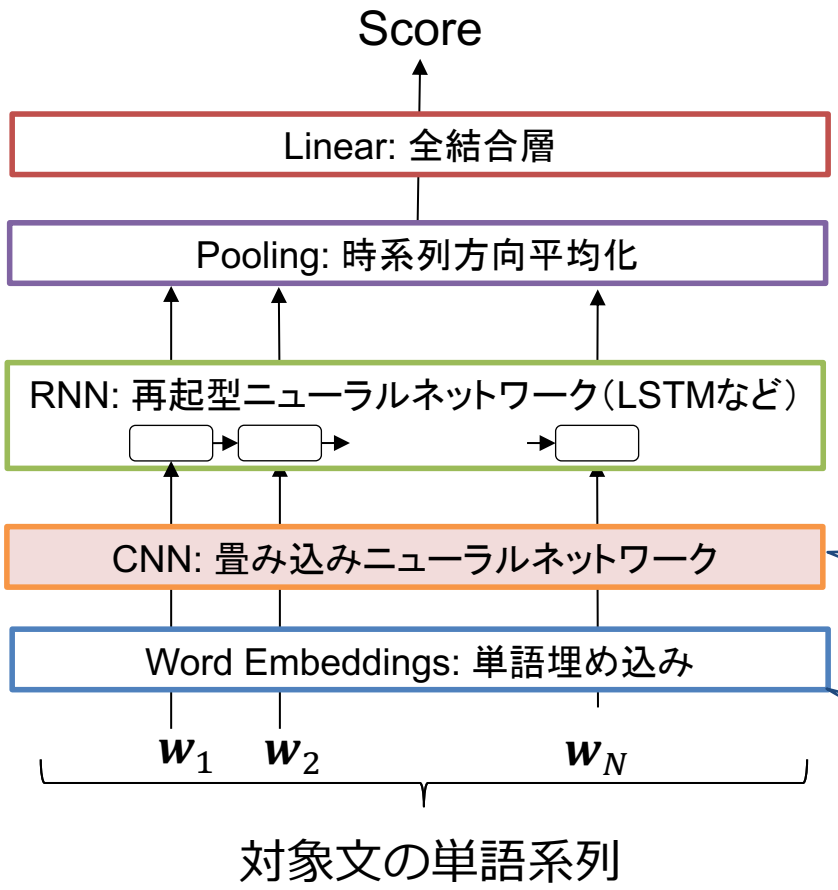


P. Sutor, Y. Aloimonos, C. Fermuller and D. Summers-Stay, "Metaconcepts: Isolating Context in Word Embeddings," *2019 IEEE Conference on Multimedia Information Processing and Retrieval.*

深層学習自動採点モデルの代表例

RNNベースモデル

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

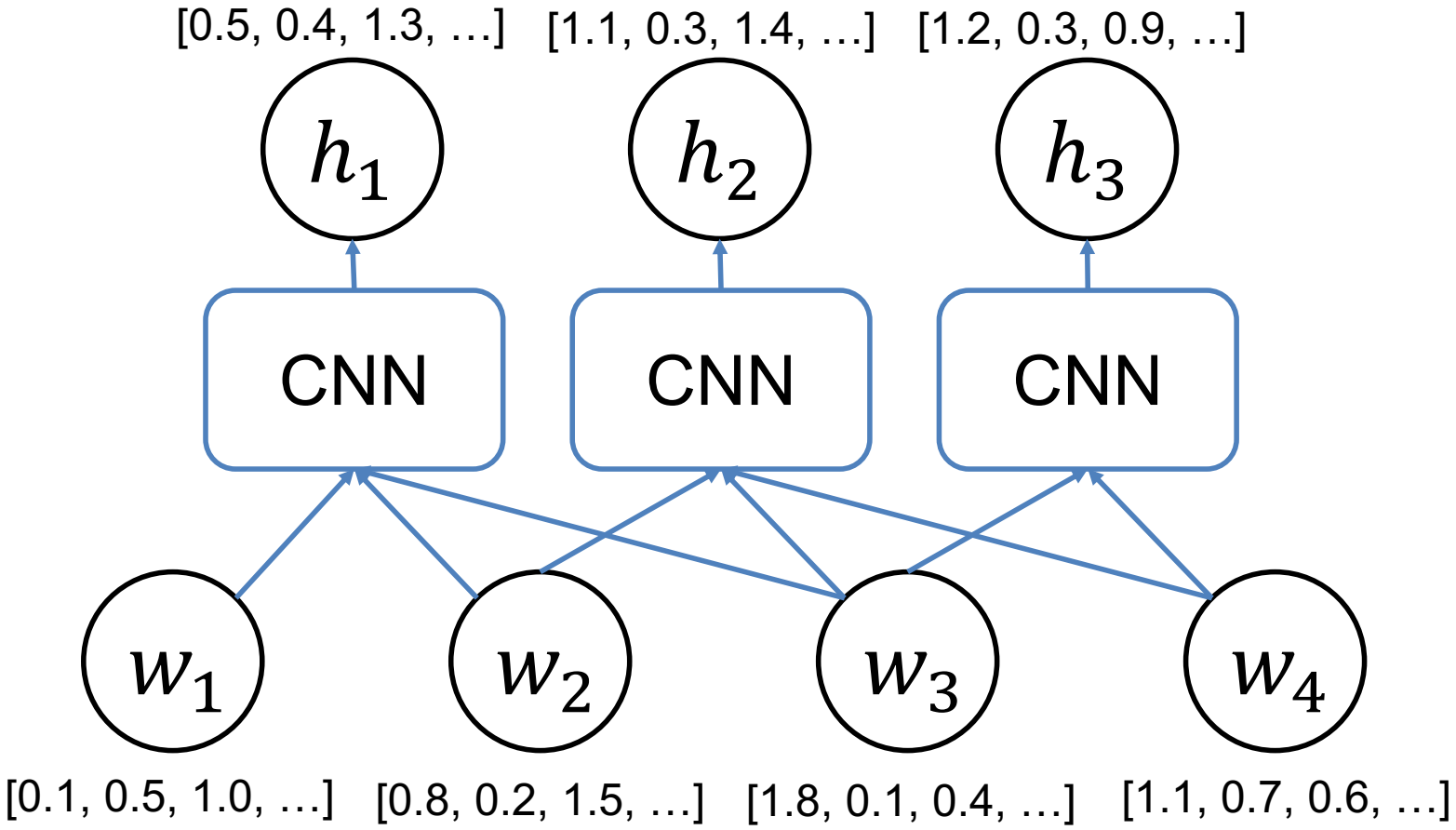


n-gramレベルの単語依存関係を特徴量化

各単語を分散表現に変換

CNN: 畳み込みニューラルネットワーク

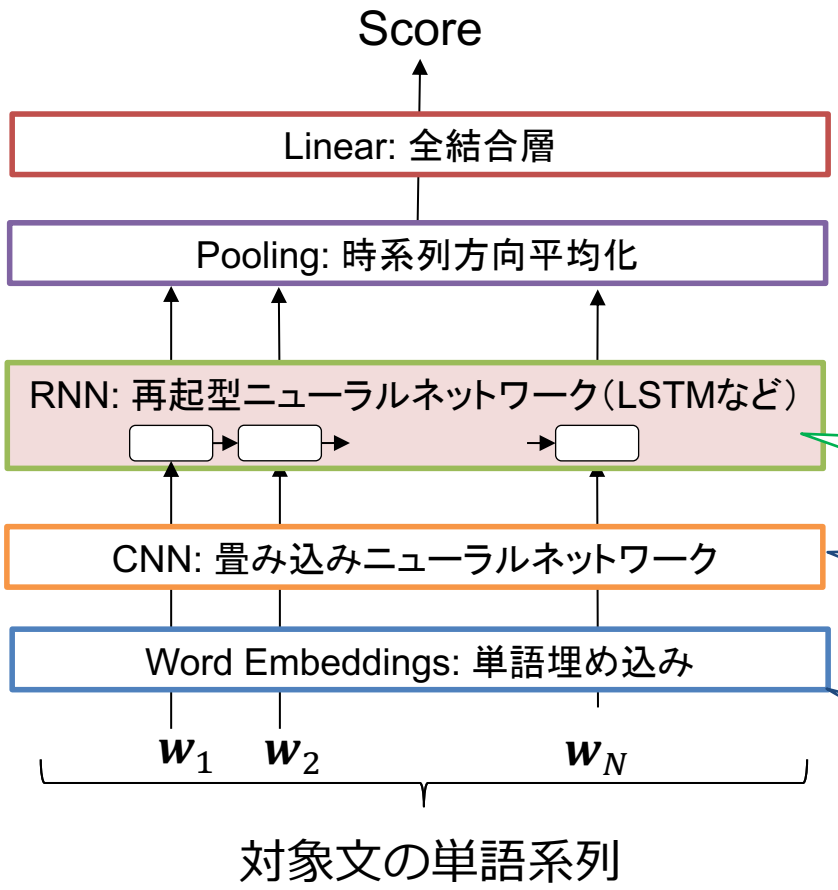
n-gramレベルの局所的な単語依存関係を特徴量化



深層学習自動採点モデルの代表例

RNNベースモデル

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



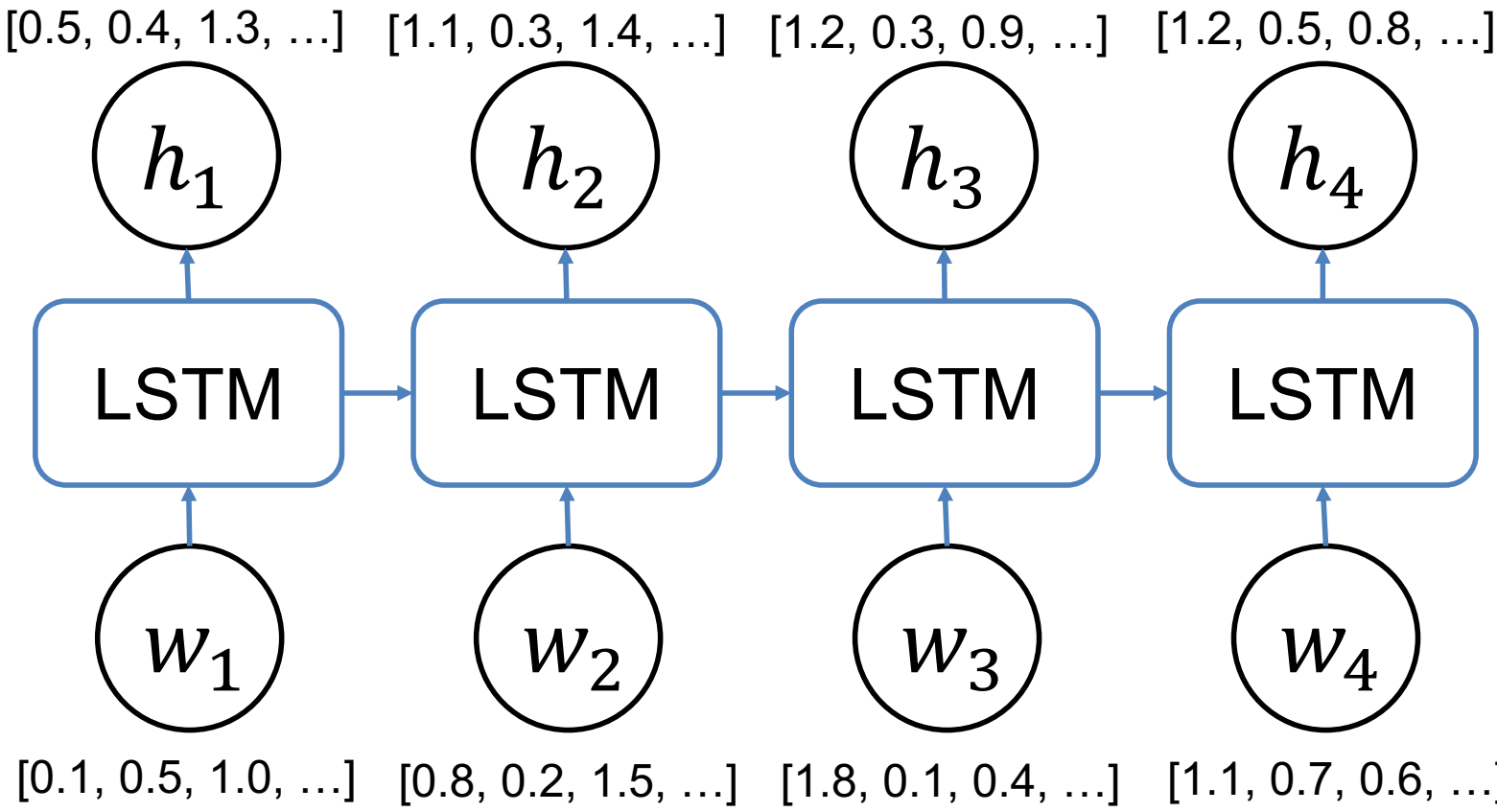
入力の時系列的な依存関係を特徴量化

n-gramレベルの単語依存関係を特徴量化

各単語を分散表現に変換

RNN: 再起型ニューラルネットワーク

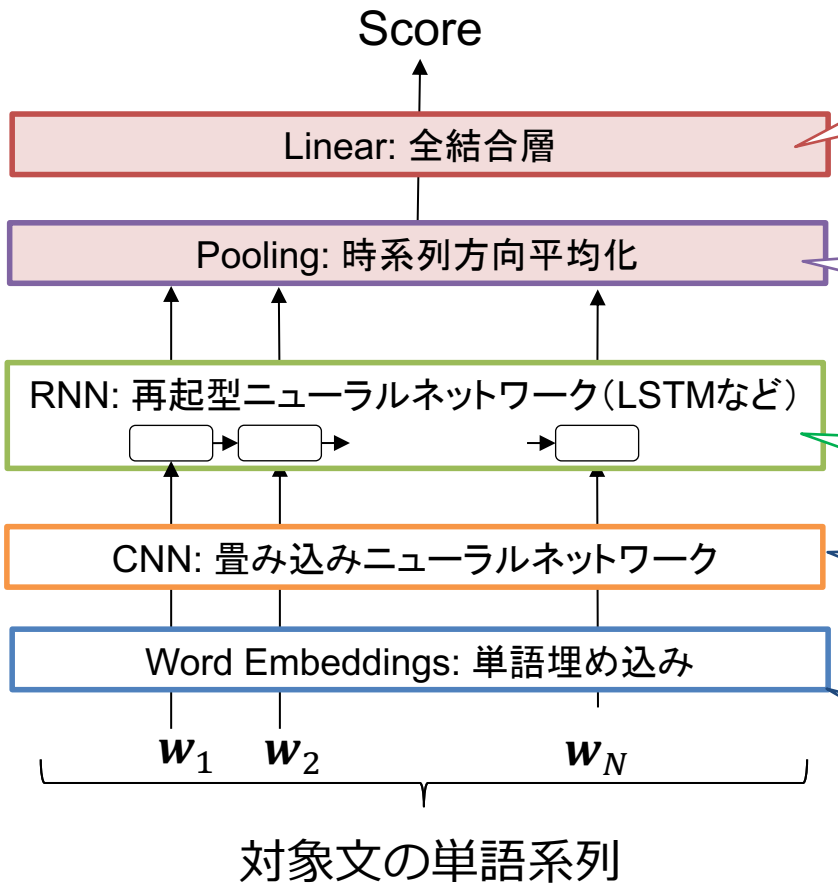
時系列の入力ベクトルを文脈を考慮した潜在変数ベクトル h に変換



深層学習自動採点モデルの代表例

RNNベースモデル

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



文章レベルの分散表現から回帰モデルで得点を予測

時系列入力を時間方向に平均化して，文章レベルの分散表現を獲得

入力の時系列的な依存関係を特徴量化

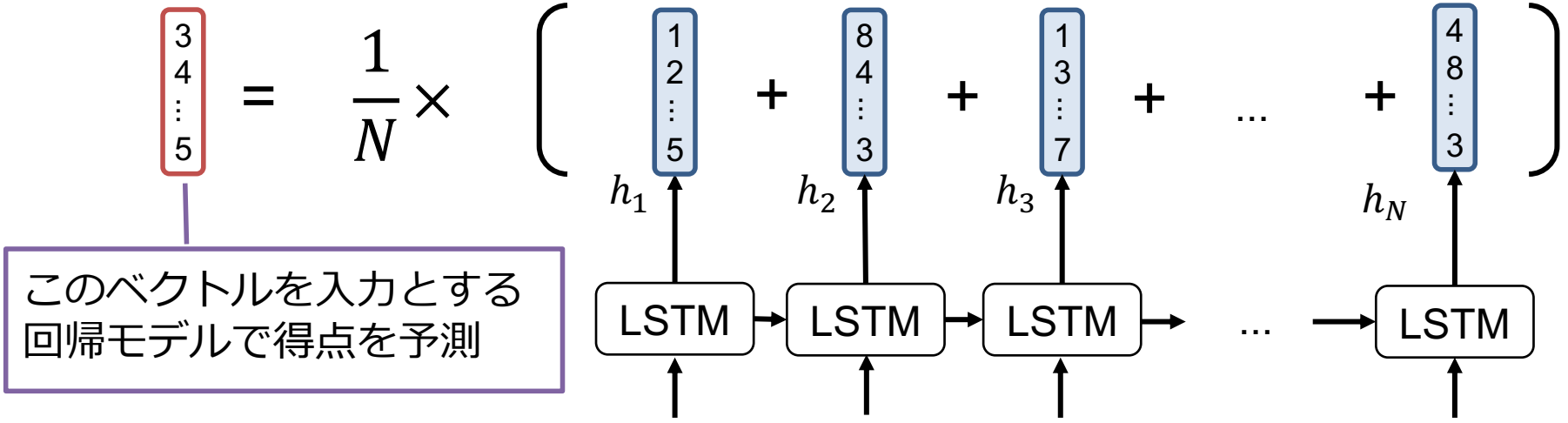
n-gramレベルの単語依存関係を特徴量化

各単語を分散表現に変換

時系列方向平均化＋回帰モデル

再起型ニューラルネットワークの出力ベクトル系列を時間方向に平均化

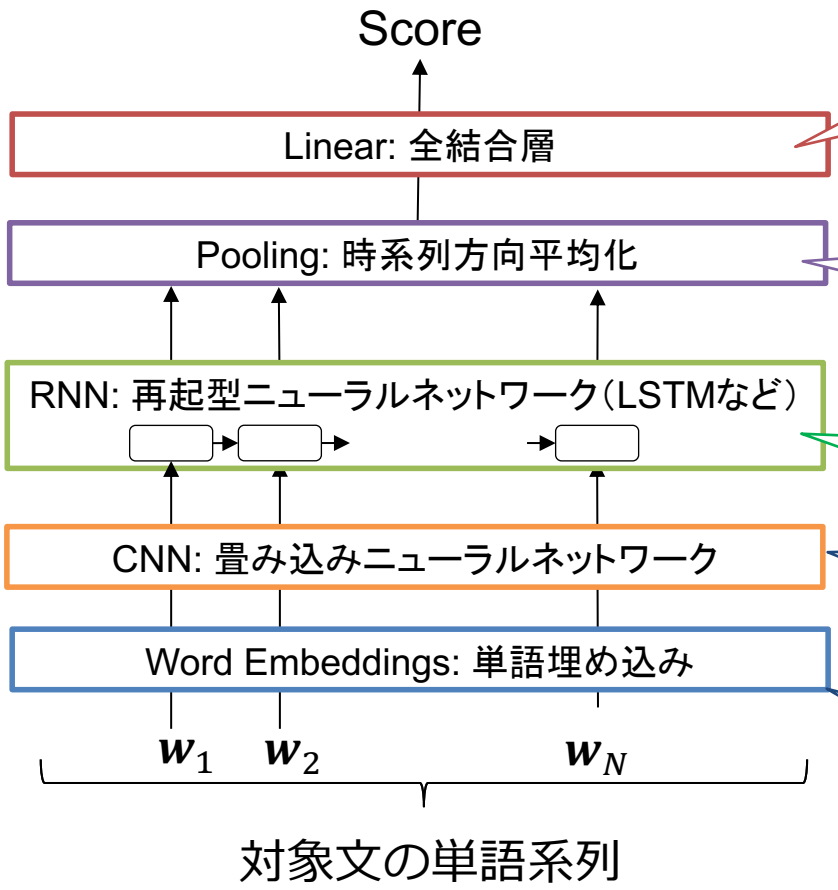
- 文脈を考慮して文章全体の特徴を縮約したベクトル表現とみなせる．
- 単語分散表現同様，類似した文章は類似したベクトル値で表される．



深層学習自動採点モデルの代表例

RNNベースモデル

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



文章レベルの分散表現から回帰モデルで得点を予測

時系列入力を時間方向に平均化して，文章レベルの分散表現を獲得

入力の時系列的な依存関係を特徴量化

n-gramレベルの単語依存関係を特徴量化

各単語を分散表現に変換

深層学習自動採点モデルの代表例

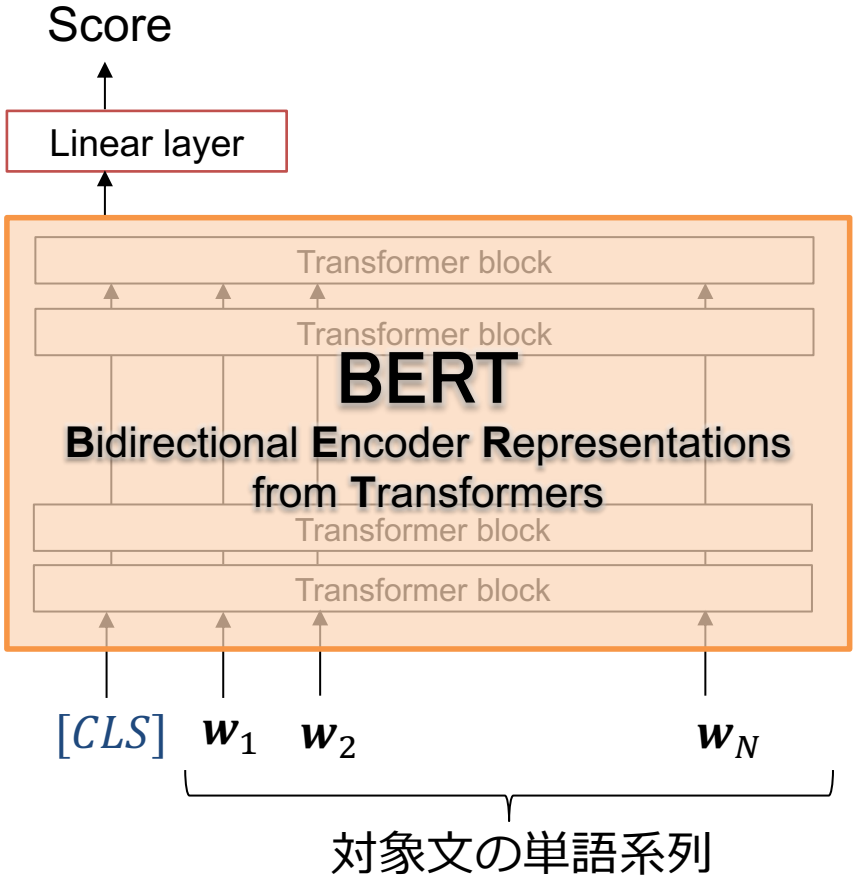
BERT

2018年にGoogleが発表した言語モデル

- 自己注意機構（Self-attention）をベースにしたTransformerネットワークで構成されるモデル
- 長距離の単語依存関係を理解可能
- 並列計算効率に優れる
- 事前学習＋ファインチューニングで高性能を達成
- 様々なタスクで最高精度を達成

BERTベースモデル

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



事前学習とファインチューニング

BERTのパラメータ数は標準(Base)モデルで1億1000万個
対象タスクのための少数の教師ありデータでは学習困難

次の方法で解決

1. 大量のラベルなし文書データで事前学習
   *BERTはWikipediaとBookCorpusの10億語以上のデータで事前学習
   *事前学習には Masked language model（単語穴埋めタスク） と
   Next sentence prediction（隣接文予測タスク）を採用

2. 事前学習で得たパラメータを初期値として，対象タスクのデータで再学習（ファインチューニング）

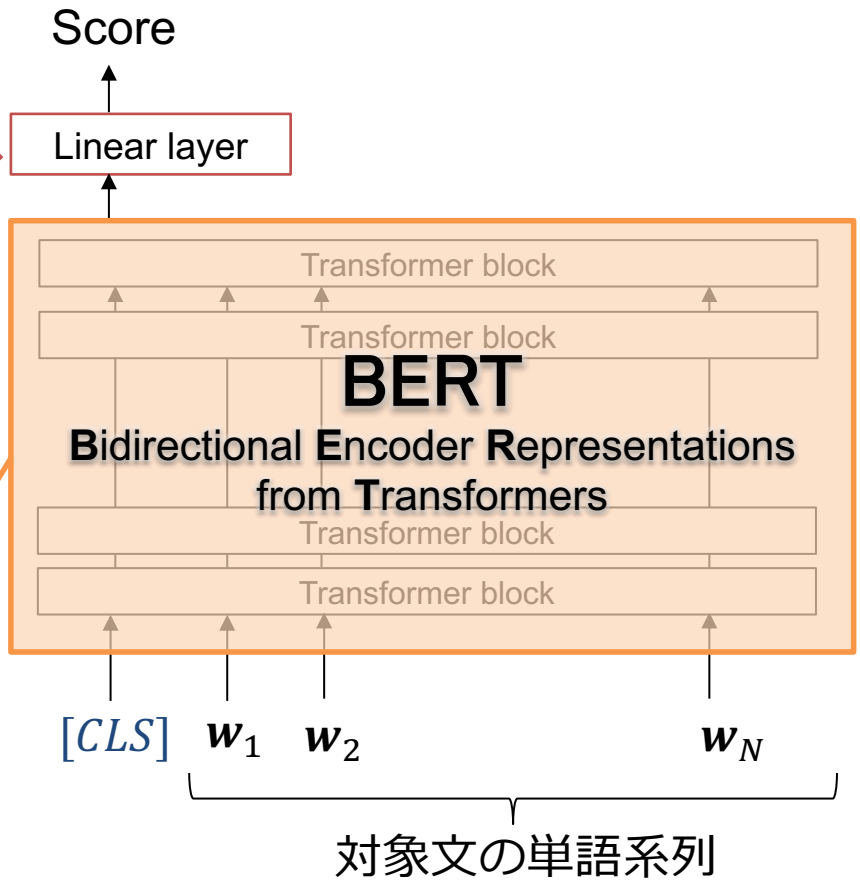
BERTによる自動採点モデル

文章レベルの分散表現から回帰モデルで得点を予測

多層のTransformerネットワークで単語系列を文章レベルの分散表現を推定

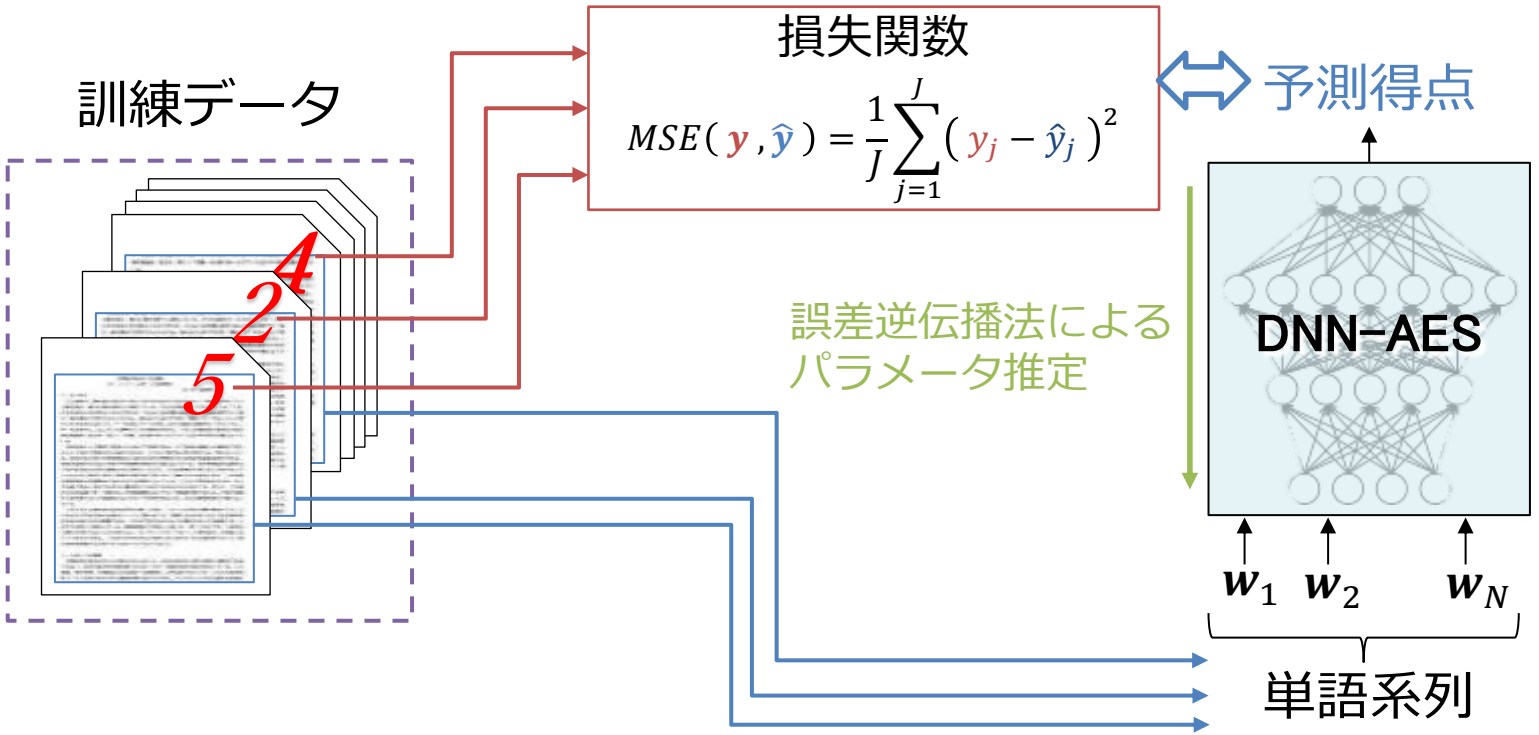
BERTベースモデル

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

大量の採点済みの答案データを用いて，損失関数を最小化するようにモデルパラメータを学習



訓練されたモデルを使用して新たな答案を自動採点

深層学習ベースのアプローチの特徴

Pros.

データセットごとの特徴量設計が不要

深層学習の知識・実装スキルのみで実装可能

Cons.

- モデルを学習するのに大規模なデータセット（採点済み答案の集合）が必要
- 採点根拠の解釈が困難

近年の自動採点技研究の方向性

1. **モデルの高度化による性能改善**
2. **総合得点のみの予測から評価観点別の得点予測へ**
3. **教師ありデータが少ない場合への対応**
   - 事前学習の導入
   - 自己教師あり学習の採用
   - 他のデータで学習したモデルの転移学習

Masaki Uto (2021) A review of deep-neural automated essay scoring models. Behaviormetrika, 48 (2) pp.459-484.

我々の研究グループの成果紹介

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).

採点者バイアスに頑健な自動採点モデル

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

項目反応理論を用いた短答記述式問題自動採点手法

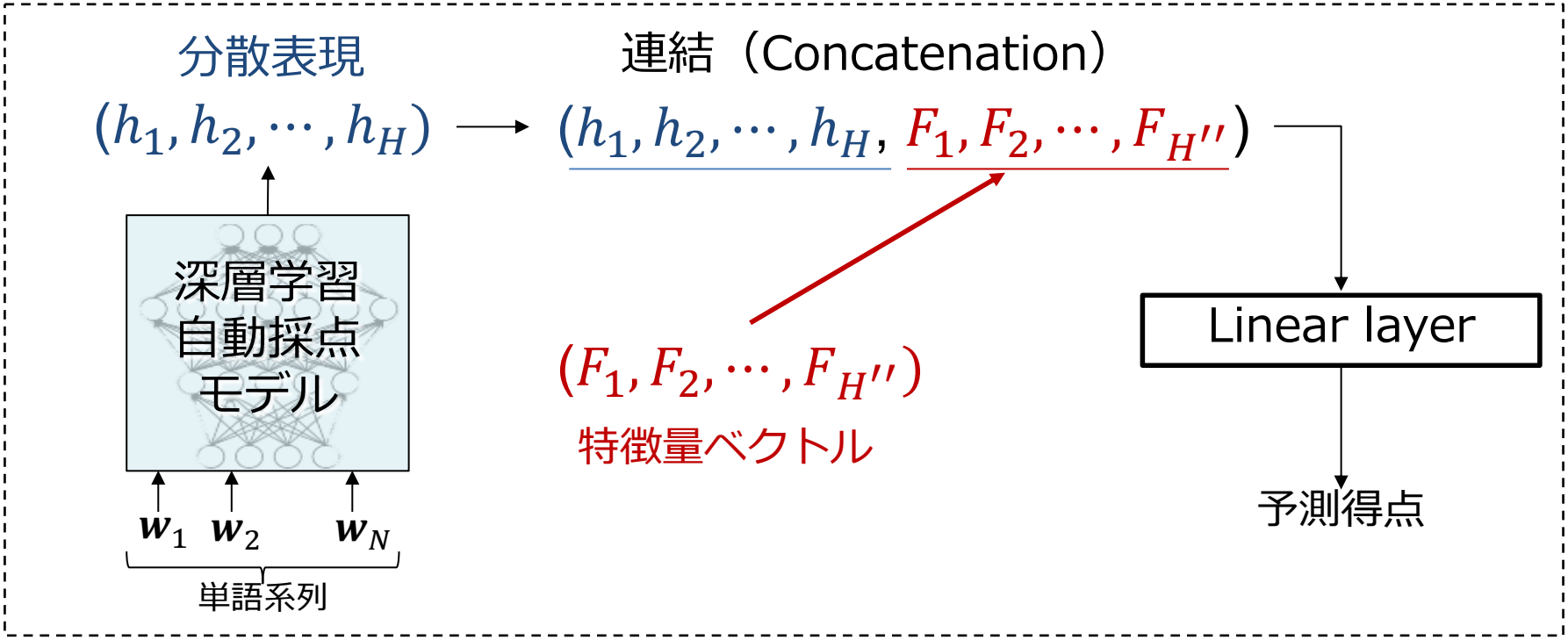
Uto & Uchida (2020) Automated short-answer grading using deep neural networks and item response theory. International Conference on Artificial Intelligence in Education (AIED).

特徴量を組み込んだ深層学習自動採点モデル

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

特徴量を組み込んだ深層学習自動採点モデル

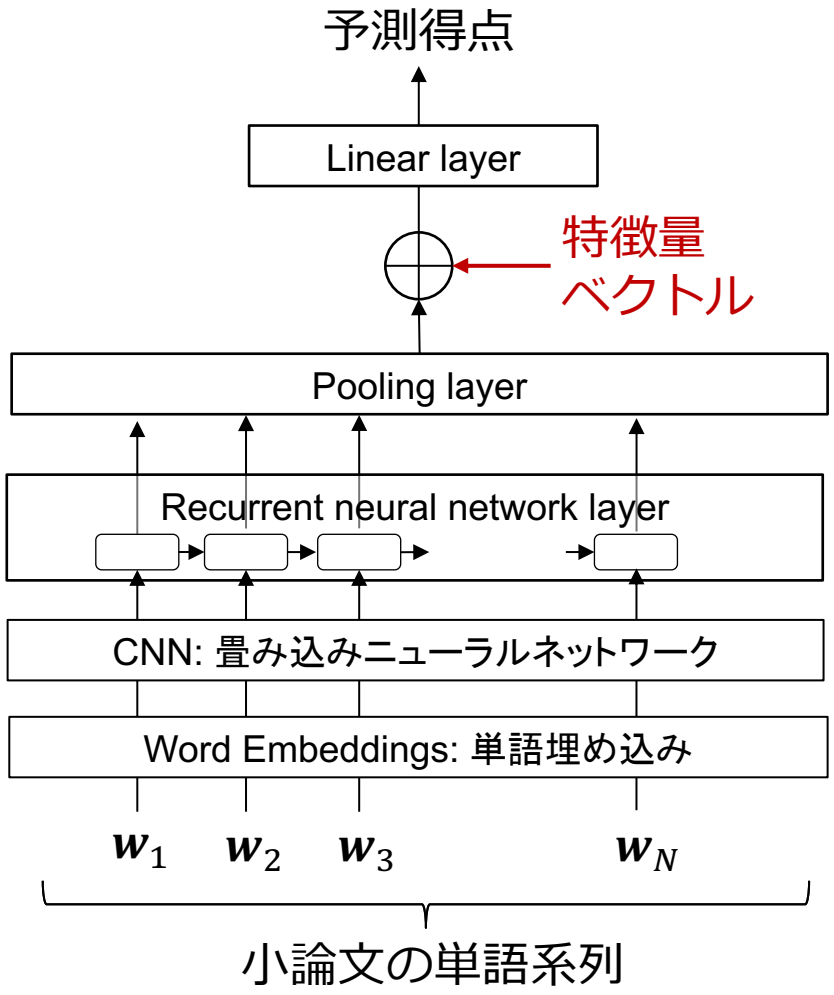
深層学習自動採点モデルに人手で設計した特徴量を統合する方法を提案



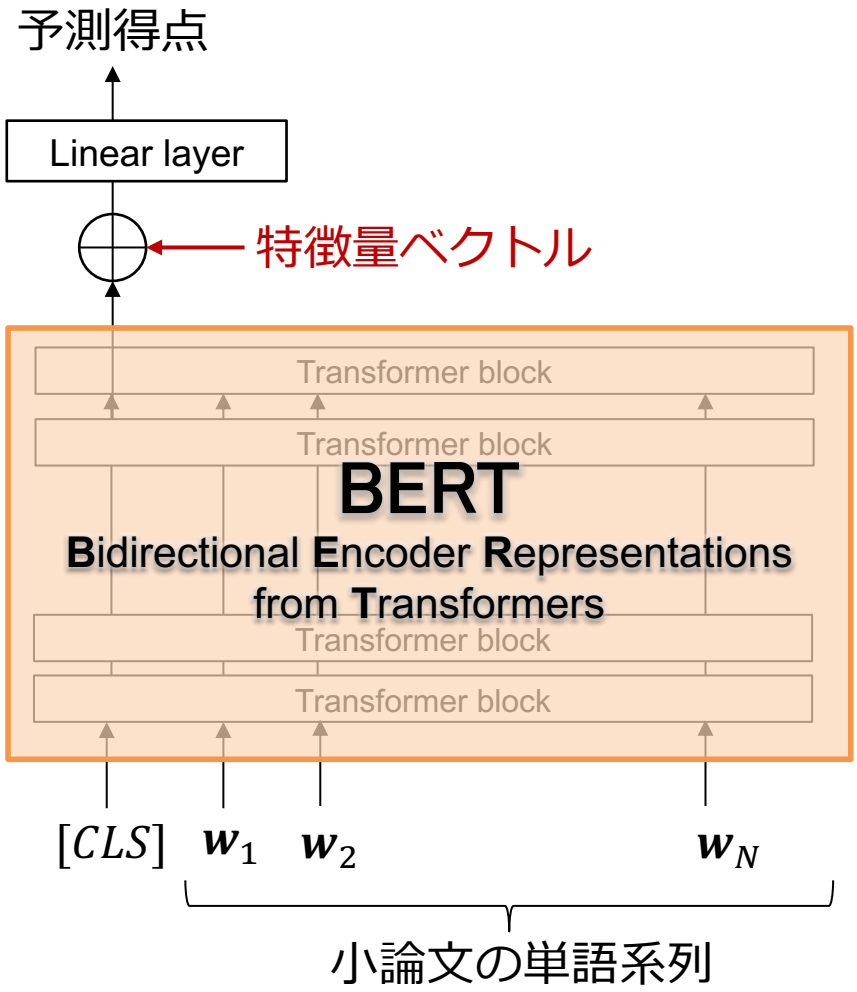
Uto, Xie & Ueno (2020) Neural Automated Essay Scoring Incorporating Handcrafted Features. Proceedings of the 28th International Conference on Computational Linguistics (COLING).

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	-

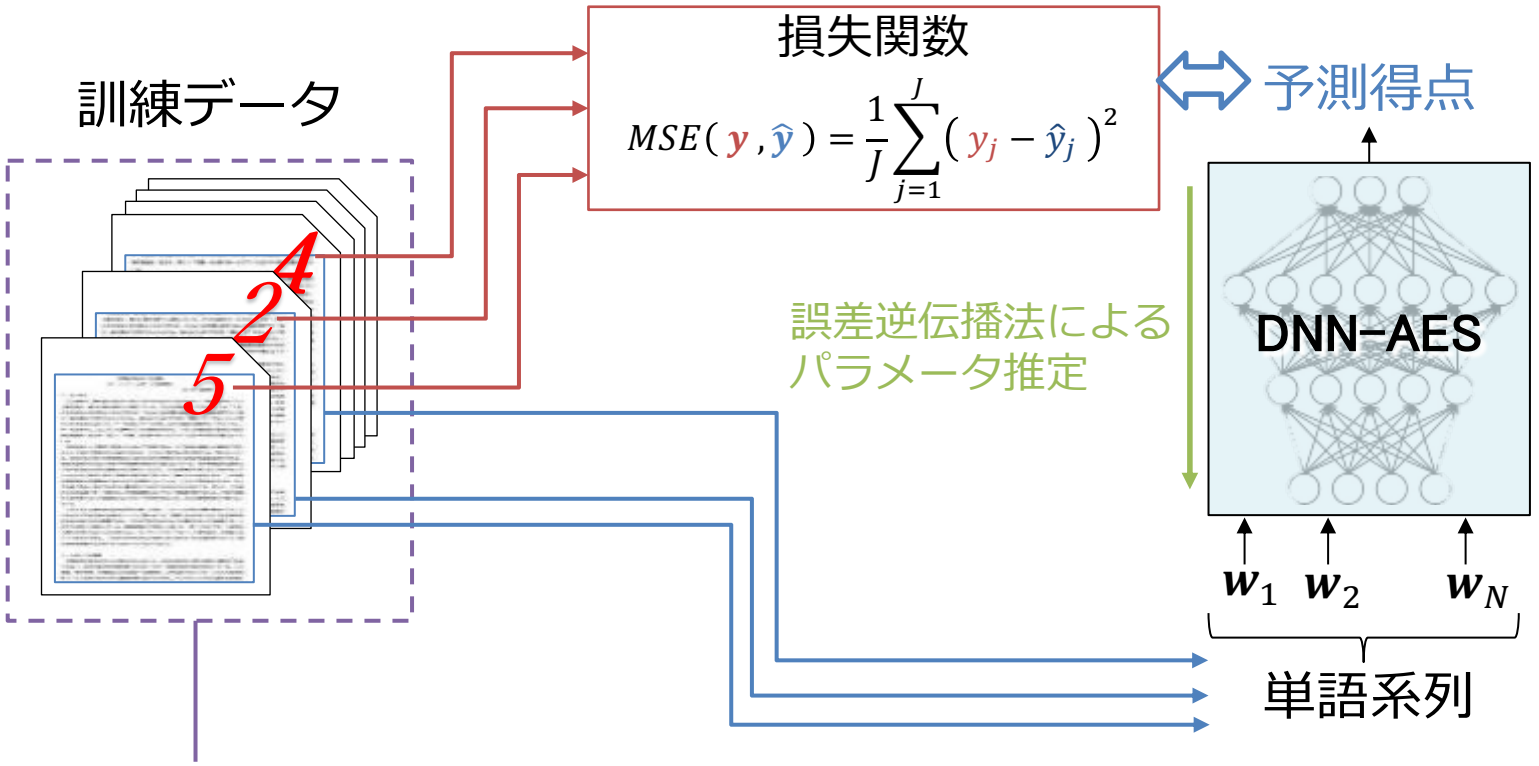
小論文自動採点タスクにおいて最高精度を達成

採点者バイアスに頑健な自動採点モデル

Uto & Okano (2020) Robust neural automated essay scoring using item response theory. International Conference on Artificial Intelligence in Education (AIED)

自動採点モデルの学習（再掲）

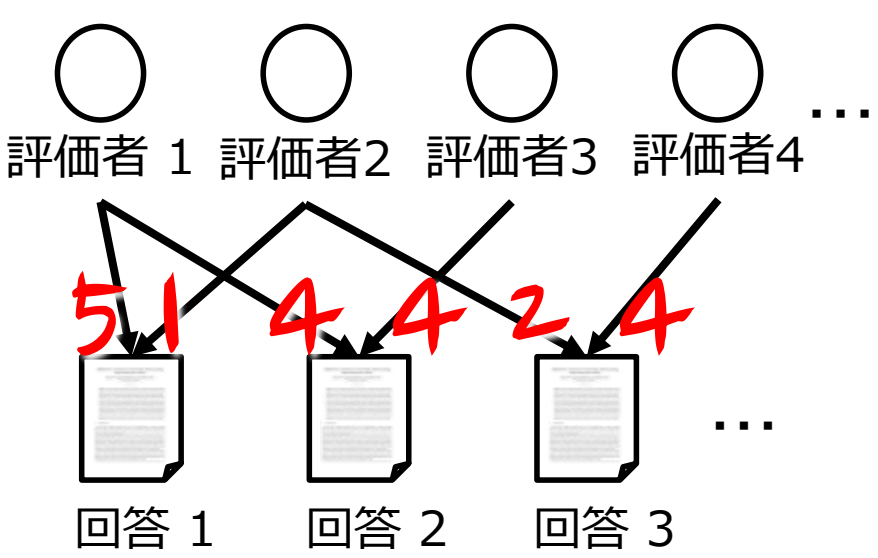
大量の採点済みの答案データを用いて，損失関数を最小化するようにモデルパラメータを学習



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

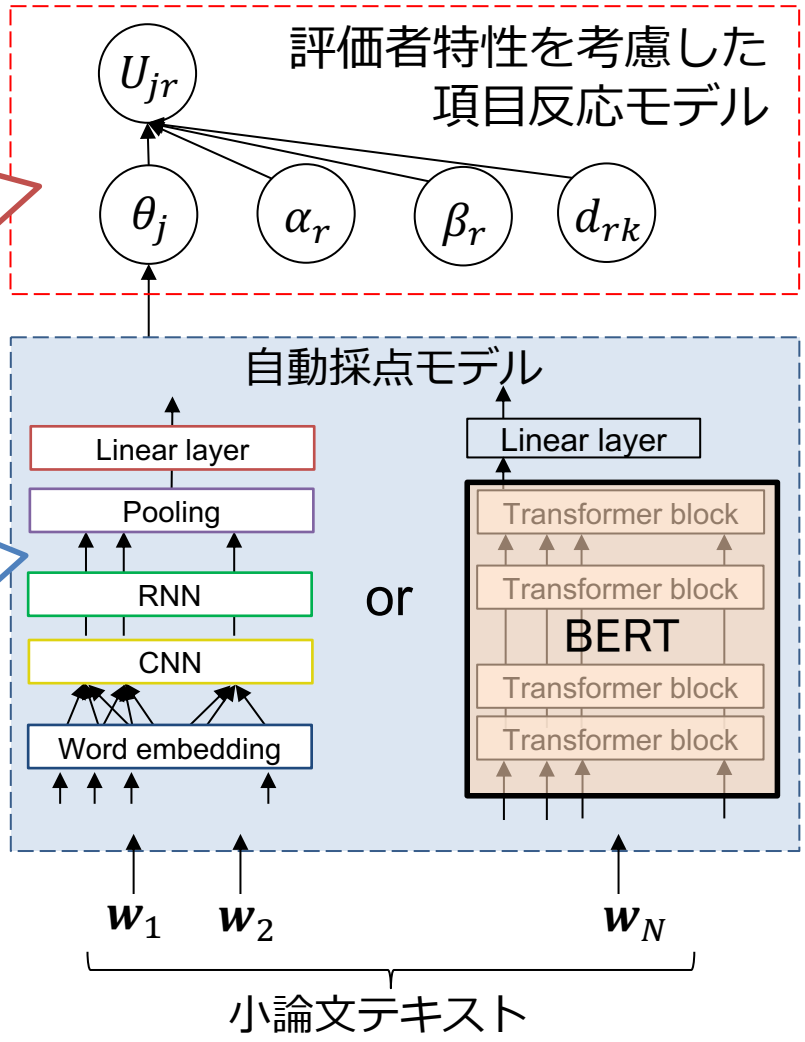
提案手法：モデル学習

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

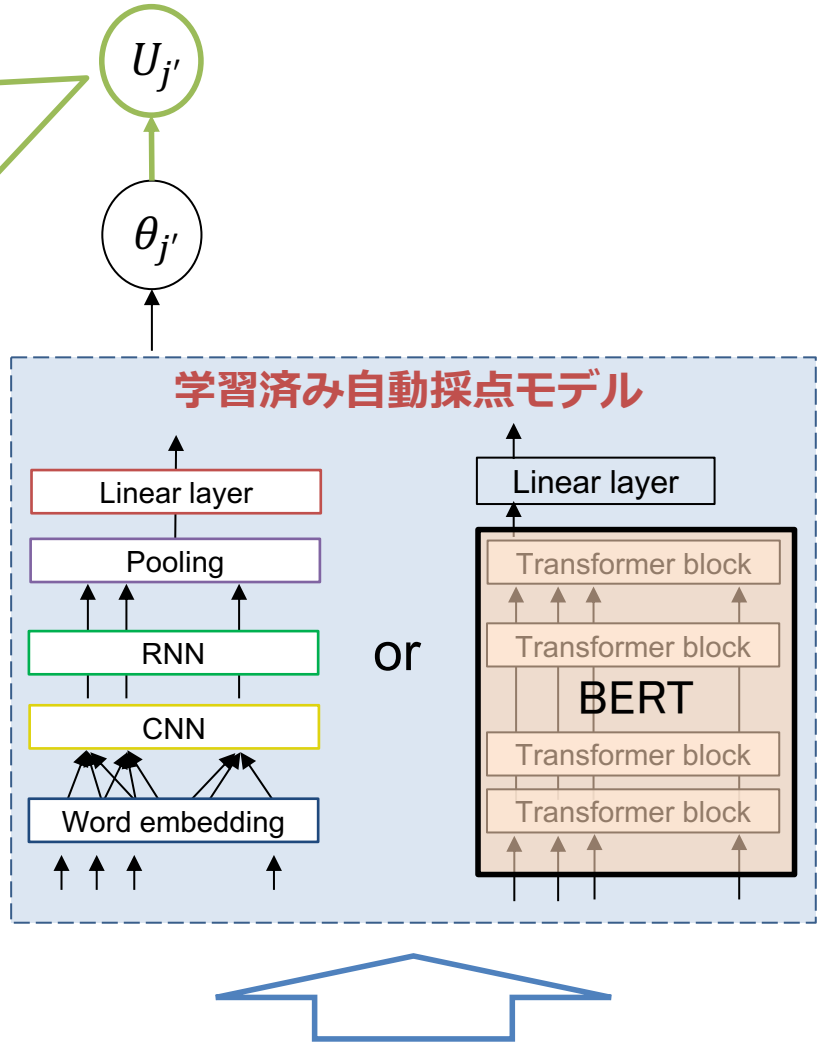


提案手法：得点予測

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

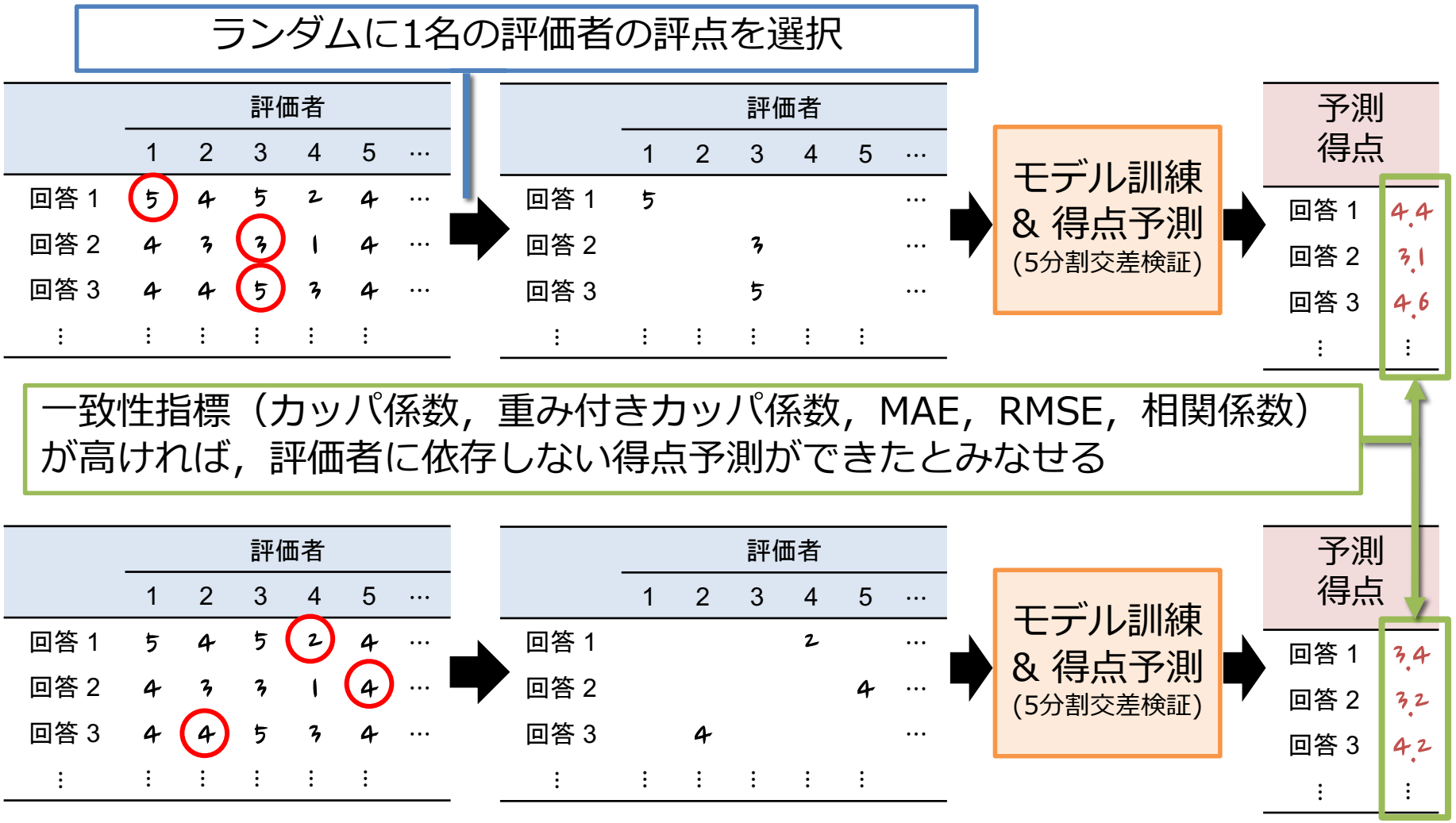
$$U_{j'} = \frac{1}{R}\sum_{r=1}^{R}\sum_{k=1}^{K} k \cdot P\left(U_{j'r} = k|\theta_{j'}\right)$$

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

個々の回答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

項目反応理論を用いた短答記述式問題自動採点手法

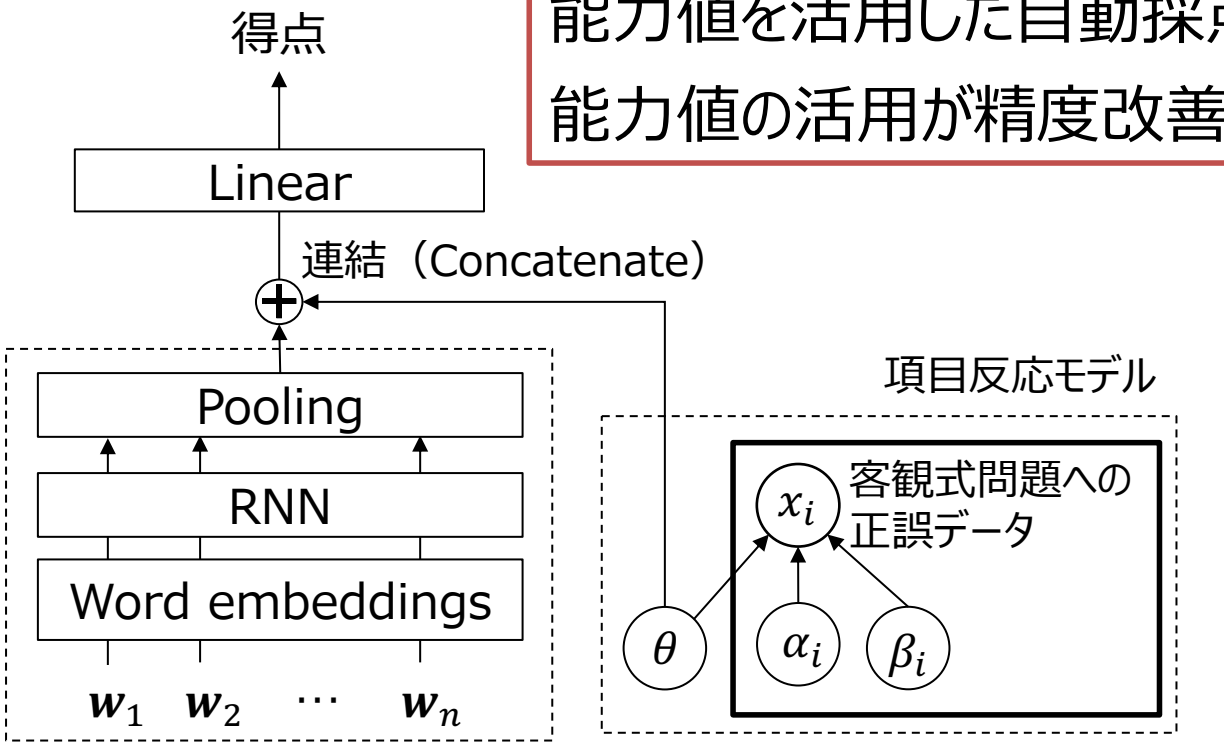
Uto& Uchida (2020) Automated short-answer grading using deep neural networks and item response theory. International Conference on Artificial Intelligence in Education (AIED).

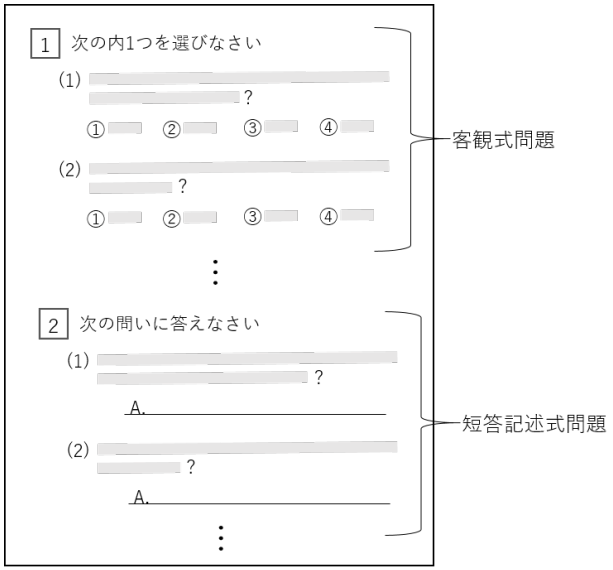
短答記述式問題の自動採点モデル

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





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を紹介
- 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

- 自動採点技術を紹介
- 項目反応理論と統合した新技術なども紹介
- 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）

まとめ

自動採点に限らず，テスティング分野におけるAI技術の開発や応用は今後ますます増加すると予測される

以下のような国際会議で活発に研究されている

人工知能の教育応用に関する国際会議例

- AIED，EDM

人工知能・言語処理に関する国際会議例

- AAAI，ACL，EMNLP，COLING