THE IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS (JAPANESE EDITION)

電子情報通信学会 論文誌 D 情報・システム

VOL. J102-D NO. 8
AUGUST 2019

本PDFの扱いは、電子情報通信学会著作権規定に従うこと。
なお、本PDFは研究教育目的（非営利）に限り、著者が第三者に直接配布することができる。著者以外からの配布は禁じられている。

情報・システムソサイエティ

一般社団法人 電子情報通信学会
THE INFORMATION AND SYSTEMS SOCIETY
THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS

論　文

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

宇都　雅輝[†a)]

IRT Topic Model for Essay Type Tests Using Rating Data and Text Information

Masaki UTO[†a)]

あらまし　近年，論理的思考力や問題解決力などの高次の能力を測定する方法の一つとして論述式試験が注目されている．他方で論述式試験の問題として，得られる評点が課題の特性だけでなく評価者の特性にも依存し，これが受験者の能力測定の精度低下を引き起こす点が指摘されてきた．この問題を解決するために，評価者と課題の特性を表すパラメータを付与した項目反応モデルが近年多数提案されている．これらのモデルは，評価者と課題のバイアスを取り除いて能力を推定できるため，素点平均などの単純な得点化法より高精度な能力測定が実現できる．しかし，これらのモデルを用いても，個々の回答文を採点する評価者の数が少なくなると能力測定の精度が低下する問題が残る．本研究では，この問題を解決するために，評価者による評点データだけでなく，受験者が執筆した回答文の内容も能力測定に利用できる新たな項目反応モデルを提案する．具体的には，評価者と課題の特性を考慮した項目反応モデルと教師ありトピックモデルを統合し，トピックモデルによって推定される各回答文のトピック分布を受験者の能力推定値に反映できるモデルを開発する．提案モデルは，回答文あたりの評価者数の減少に伴う能力測定精度の低下を緩和できるだけでなく，評点が与えられていない回答文の得点や受験者の能力を文章情報のみから推定できる利点も有する．また，本論文では，マルコフ連鎖モンテカルロ法を用いた提案モデルのパラメータ推定手法を提案し，実データ実験により提案モデルの有効性を示す．

キーワード　項目反応理論，潜在ディリクレ配分法，教師ありトピックモデル，パフォーマンス評価，論述式試験，自動採点

1. まえがき

近年，論理的思考力や問題解決力といった高次の能力を測定するニーズが高まっており，これを実現する手法の一つとして論述式試験が注目されている [1]〜[6]．一般に論述式試験は，受験者に複数の課題を与え，それらに対する回答文を数名の評価者によって採点する形式で実施される．しかし，この場合，得られる評点が評価者や課題の特性（評価者の甘さ/厳しさや課題困難度など）に強く依存し，これが受験者の能力測定の精度低下を引き起こすことが問題とされてきた [5]〜[14]．

この問題を解決する手法の一つとして，評価者と課題の特性パラメータを付与した項目反応モデルが近年多数提案されている (e.g., [6], [7], [11]〜[13])．これらの項目反応モデルでは評価者と課題の特性を考慮して受験者の能力を推定できるため，素点の合計や平均といった単純な得点化法に比べて高精度な能力測定が可能となる [6], [7], [11], [13]．

しかし，これらのモデルを用いても，個々の回答文を採点する評価者数が少なくなると高精度な能力測定は困難となる．一般に論述式試験の採点プロセスでは，評価者の負担や運用の時間的・経済的コストを軽減するために，各回答文に少数名の評価者を割り当てて採点を行わせることが多い [12], [15], [16]．

本研究では，この問題を解決するために，評価者による評点データだけでなく，受験者が執筆した回答文の内容も能力測定に利用できる新たな項目反応モデルを提案する．提案モデルは，評価者と課題の特性を考

† 電気通信大学大学院情報理工学研究科，調布市
Graduate School of Informatics and Engineering, University of Electro-Communications, 1–5–1 Chofugaoka, Chofu-shi, 182–8585 Japan
a) E-mail: uto@ai.lab.uec.ac.jp
DOI:10.14923/transinfj.2019JDP7007

 ⓒ一般社団法人電子情報通信学会 2019

慮した項目反応モデルとトピックモデルの一つである潜在ディリクレ配分法 [17] を統合したモデルとして定式化する．具体的には，潜在ディリクレ配分法を用いて個々の回答文のトピック分布を推定し，そのトピック分布を項目反応モデルにおける受験者の能力推定値に反映させるようにモデル化を行う．トピック分布の能力値への反映には，トピック分布と任意の目的変数の関係をモデル化した教師ありトピックモデル [18] のアプローチを用いる．項目反応モデルと教師ありトピックモデルを統合して受験者の能力推定に評点データと回答文情報を同時に利用する手法はこれまでに開発されておらず，本研究が新たに取り組むものである．提案モデルの利点は次のとおりである．

（1） 評価者が与える評点データに加えて，回答文の内容的な特徴も考慮して能力推定がなされるため，既存モデルより高精度な能力測定が可能であり，回答文あたりの評価者数の減少に伴う能力測定精度の低下を緩和できる．

（2） 評点が与えられていない回答文の得点と，それらの回答文を執筆した受験者の能力を文章情報のみから推定することができる．

本研究では，提案モデルのパラメータ推定手法として，周辺化ギブスサンプリングとメトロポリスヘイスティングスを組み合わせたマルコフ連鎖モンテカルロ法を提案する．更に，実データ実験により提案モデルの有効性を示す．

2. データ

本研究では，J 人の受験者 $\mathcal{J}=\{1,\cdots,J\}$ に I 個の論述課題 $\mathcal{I}=\{1,\cdots,I\}$ を与え，それらの回答文を R 人の評価者集団 $\mathcal{R}=\{1,\cdots,R\}$ が K 段階カテゴリー $\mathcal{K}=\{1,\cdots,K\}$ で採点する場合を考える．ここで，課題 $i\in\mathcal{I}$ に対する受験者 $j\in\mathcal{J}$ の回答文を e_{ij} で表し，回答文 e_{ij} に対する評価者 r の評点を U_{ijr} とすると，評点データは次式で定義できる．

$$\boldsymbol{U}=\{U_{ijr}\in\mathcal{K}\cup\{-1\}\mid i\in\mathcal{I},j\in\mathcal{J},r\in\mathcal{R}\} \tag{1}$$

ここで，$U_{ijr}=-1$ は欠測データを表す．

また，回答文集合 $\boldsymbol{E}=\{e_{ij}\mid i\in\mathcal{I},j\in\mathcal{J}\}$ に含まれる語彙集合を $\mathcal{V}=\{1,\cdots,V\}$ とすると，回答文 e_{ij} 内の単語系列は次式で定義できる．

$$\boldsymbol{W}_{ij}=\{W_{ijn}\in\mathcal{V}\mid n=\{1,\cdots,N_{ij}\}\} \tag{2}$$

ここで，W_{ijn} は回答文 e_{ij} 内の n 番目の単語を表し，N_{ij} は e_{ij} 内の単語数を表す．

本研究の目的は，これらのデータを用いて各受験者の能力を高精度に推定することである．このために本研究では項目反応理論とトピックモデルを用いる．

3. 項目反応理論

項目反応理論（IRT: Item Response Theory）は数理モデルを用いたテスト理論の一つである [19]．IRT では，受験者のテスト項目への反応を，受験者の能力を表す潜在変数と項目の特性（困難度や識別力など）を表すパラメータで定義される確率モデルで表現する．このモデルを用いることで，IRT は，1）異なる項目で構成されたテストを受験しても同一尺度上で能力を測定できる，2）個々の項目やテスト全体の能力測定精度を分析できる，3）欠測データの扱いが容易である，などの多くの利点をもつ．このような利点から，IRT は現代のテスト運用の基礎として，IT パスポート試験 [20] や医療系大学間共用試験 [21] などの大規模試験を含む，様々な評価場面で広く実用化されている．

一般的な項目反応モデルでは，テスト項目に対する受験者の反応や正誤答をデータとして扱うため [22]〜[26]，データは受験者 × 項目の 2 相データとなる．他方で，**2.** で定義したように，本研究で扱うデータは受験者 × 課題 × 評価者の 3 相データとなる．従来の項目反応モデルは，このような 3 相データに直接には適用できない．この問題を解決するために，項目反応モデルにおける項目特性パラメータを課題の特性パラメータとみなし，評価者の特性を表すパラメータを付与したモデルが近年多数提案されている [6], [7], [11], [13], [27]〜[29]．

これらの既存モデルは，異なる評価者特性パラメータと課題特性パラメータを採用しており，それぞれに異なる特徴をもつ [6], [30]．本研究では，既存モデルの中で，評価者特性を最も柔軟に捉えることができる宇都・植野のモデル [11] を基礎モデルとして採用する．このモデルは，代表的な評価者特性として知られる 1）一貫性，2）甘さ/厳しさ，3）尺度範囲の制限，を同時に考慮できる唯一のモデルであり，多様な評価者の特性を柔軟に表現でき，異質性の強い評価者が存在しても頑健な能力測定を行うことができる [11]．各評価者特性の詳細については [6], [9], [11], [12], [30] などを参照されたい．

このモデルでは，課題 i に対する受験者 j の回答文

に評価者 r が評点 k を与える確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = \frac{\exp\sum_{m=1}^{k}[\alpha_r\alpha_i(\theta_j-\beta_i-\beta_r-d_{rm})]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}[\alpha_r\alpha_i(\theta_j-\beta_i-\beta_r-d_{rm})]} \quad (3)$$

ここで，θ_j は受験者 j の能力，α_i は課題 i の識別力，α_r は評価者 r の一貫性，β_i は課題 i の困難度，β_r は評価者 r の厳しさ，d_{rk} は評価カテゴリー k に対する評価者 r の厳しさを表す．ただし，パラメータの識別性のために，$\sum_{i=1}^{I}\log\alpha_i=0$，$\sum_{i=1}^{I}\beta_i=0$，$d_{r1}=0$，$\sum_{k=2}^{K}d_{rk}=0$ を仮定する．これらのモデルパラメータと能力値は，評点データ $\boldsymbol{U}$ から推定することができる [11]．

1. で述べたように，このような項目反応モデルでは，受験者の能力を評価者や課題の特性の影響を取り除いて推定できるため，素点の合計や平均といった単純な得点化法より高精度な能力測定が可能となる [6], [7], [11], [13]．しかし，これらのモデルを用いても，個々の回答文を採点する評価者数が少なくなると，受験者あたりの評点データが減少するため，能力推定の精度が低下する問題が残る．本研究のアイディアは，この問題を解決するために，受験者の能力 θ_j の推定に，評点データだけでなく回答文の内容も利用する点にある．本研究では，回答文の内容を扱う手法としてトピックモデルを用いる．

4. トピックモデル

トピックモデルは，文書集合が与えられたとき，個々の文書が複数の潜在的な話題（トピック）をもつと仮定し，それらのトピックの出現分布を文書ごとに推定する教師なし機械学習手法である．また，トピックモデルでは，各トピックに対して語彙の出現分布を推定するため，それらの語彙分布を解釈することで個々のトピックの意味を解釈することができる．代表的なトピックモデルとしては，潜在意味解析法（LSA: Latent Semantic Analysis）[31] や確率的潜在意味解析法（PLSA: Probabilistic Latent Semantic Analysis）[32]，潜在ディリクレ配分法（LDA: Latent Dirichlet Allocation）[17] が知られている．LDA は LSA と PLSA の上位モデルであり，LSA や PLSA に比べて高精度なトピック推定が可能であることが知られており [17]，テキストを扱う様々なタスクで活用されている（e.g., [18], [33]～[37]）．そこで，本研究では，トピックモデルとして LDA を利用する．

LDA では回答文 e_{ij} 内の各単語 W_{ijn} がどのトピックから生成されたかを示す潜在変数を導入する．ここで，単語 W_{ijn} に対応するトピックを $Z_{ijn}\in\mathcal{T}=\{1,\cdots,T\}$（$T$ はトピック数）で表し，回答文 e_{ij} におけるトピック t の生起確率を ψ_{ijt}，トピック t における語彙 v の生起確率を ϕ_{tv} で表す．このとき，LDA では，各単語 W_{ijn} とトピック Z_{ijn} が以下の多項分布（$Multi(\cdot)$ と表記する）で表されるトピック分布と語彙分布に従って生起すると仮定する．

$$Z_{ijn}\sim Multi(\boldsymbol{\psi}_{ij}) \quad (4)$$

$$W_{ijn}\sim Multi(\boldsymbol{\phi}_{z_{ijn}}) \quad (5)$$

ただし，$\boldsymbol{\psi}_{ij}=\{\psi_{ij1},\cdots,\psi_{ijT}\}$，$\boldsymbol{\phi}_t=\{\phi_{t1},\cdots,\phi_{tV}\}$．

また，各分布のパラメータ $\boldsymbol{\psi}_{ij}$ と $\boldsymbol{\phi}_t$ は多項分布の共役事前分布であるディリクレ分布（$Dir(\cdot)$ と表記する）に従うと仮定する．ここで，γ と η を $\boldsymbol{\psi}_{ij}$ と $\boldsymbol{\phi}_t$ のディリクレ事前分布のパラメータとすると，$\boldsymbol{\psi}_{ij}$ と $\boldsymbol{\phi}_t$ は以下の式に従って生成すると仮定される．

$$\boldsymbol{\psi}_{ij}\sim Dir(\gamma) \quad (6)$$

$$\boldsymbol{\phi}_t\sim Dir(\eta) \quad (7)$$

LDA によって推定されるトピック分布 $\boldsymbol{\psi}_{ij}$ は，回答文 e_{ij} の内容的な特徴を T 次元のベクトルで表現したものと解釈できる [18], [35], [38]．近年では，このように文書ごとに推定されるトピック分布を他の変数の予測に利用する教師ありトピックモデル [18] と呼ばれる手法が提案されている．本研究では，トピック分布を受験者の能力値に反映させるために教師ありトピックモデルのアプローチを用いる．

5. 教師ありトピックモデル

一般に，教師ありトピックモデルでは，個々の文書 e_{ij} に対応する任意の目的変数 y_{ij} を，その文書のトピック情報を説明変数とする回帰モデルによって予測するようにモデル化する．回帰モデルには様々なモデルが利用できるが [39]，最も一般的な正規回帰モデルを想定し，変数 y_{ij} が実数値をとると仮定すると，y_{ij} の生起確率は以下のように定義される．

$$y_{ij}\sim N(\boldsymbol{\omega}^{\mathrm{T}}\bar{\boldsymbol{Z}}_{ij},\sigma_0^2) \quad (8)$$

ここで，$N(\mu,\sigma^2)$ は平均 μ，標準偏差 σ の正規分布を

表し，$\boldsymbol{\omega} = \{\omega_1, \cdots, \omega_T\}$ は目的変数に対する各トピックの重み集合を表す．σ_0^2 は目的変数の分散を表すハイパーパラメータである．また，$\bar{\boldsymbol{Z}}_{ij} = \{\bar{Z}_{ij1}, \cdots, \bar{Z}_{ijT}\}$ であり，$\bar{Z}_{ijt} \in \bar{\boldsymbol{Z}}_{ij}$ は次式で定義される．

$$\bar{Z}_{ijt} = \frac{\sum_{n=1}^{N_{ij}} \delta(Z_{ijn}, t)}{N_{ij}} \tag{9}$$

$\delta(a, b)$ は二つの値 a と b が一致するとき 1，そうでないとき 0 をとる関数とする．

教師ありトピックモデルは，個々の文書を T 次元のトピック分布パラメータで表現し，それを用いて目的変数に回帰するモデルとみなせる [39]．教師ありトピックモデルでは，各文書の内容的な意味を考慮した予測が可能となるため，単語の出現頻度ベクトルを用いた単純な回帰モデルと比べて，高い予測精度が期待できることが報告されている [18], [35], [38]〜[40]．このような利点から，教師ありトピックモデルのアプローチは，テキスト情報を予測に活用する様々な応用問題に適用され，その有効性が示されてきた (e.g., [34], [40]〜[44])．本研究でも，教師ありトピックモデルのアプローチを用いて，トピック分布を IRT モデルにおける受験者の能力推定値に反映させる．

6. 提案手法

提案モデルでは，IRT における受験者の能力値 θ_j が，その受験者の回答文のトピック分布に依存すると考えることで，文章情報を能力値に反映させる．具体的には，式 (3) における能力 θ_j の分布として次式を考える．

$$\theta_j \sim N(\boldsymbol{\omega}^{\mathrm{T}} \bar{\boldsymbol{Z}}_j, \sigma_0^2) \tag{10}$$

ここで，$\boldsymbol{\omega} = \{\omega_1, \cdots, \omega_T\}$ は能力推定値に対する各トピックの重みを表す．また，$\bar{\boldsymbol{Z}}_j = \{\bar{Z}_{j1}, \cdots, \bar{Z}_{jT}\}$ を表し，$\bar{Z}_{jt} \in \bar{\boldsymbol{Z}}_j$ は次式で定義される．

$$\bar{Z}_{jt} = \frac{\sum_{i=1}^{I} \sum_{n=1}^{N_{ij}} \delta(Z_{ijn}, t)}{\sum_{i=1}^{I} N_{ij}} \tag{11}$$

本研究の条件では，各受験者が複数の回答文を有するのに対し，目的変数は受験者ごとに一つのみ推定される能力値 θ_j となるため，通常の教師ありトピックモデルとは異なり，$\bar{Z}_{jt}$ が複数回答文のトピック情報を累積した形で定義されている点に注意されたい．また，式 (10) 中の σ_0^2 は能力値の分散を表す．IRT では，能力値に標準正規分布を仮定することが一般的であるため，本研究でも $\sigma_0^2 = 1.0$ を用いる．

式 (10) から明らかなように，提案モデルでは，文章のトピック分布から推定される能力値を，項目反応モデルにおける能力推定値 θ_j の事前分布として反映している．このとき，トピック分布と能力値の関係は，式 (10) の重み $\boldsymbol{\omega}$ によって学習される．これにより提案モデルでは，文章の内容的な特徴を能力推定に反映できるため，評点データのみを利用する IRT に比べて能力測定精度が改善されると期待できる．また，提案モデルでは，語彙分布と評価者特性，課題特性及び重みのパラメータが既知であれば，評点データが与えられていない受験者の能力を，文章情報のみを用いて推定することができる．更に，そのように推定された能力値を所与として回答文の期待得点を求めることで未採点回答文の自動評価も可能である．これらの具体的な手順は **7.3** で述べる．

7. パラメータ推定

IRT におけるパラメータ推定手法としては，EM アルゴリズムを用いた周辺最ゆう推定法やニュートンラフソン法による事後確率最大化推定法が広く用いられてきた [22]．一方で，式 (3) のような複雑な IRT モデルの場合には，マルコフ連鎖モンテカルロ（MCMC: Markov Chain Monte Carlo）アルゴリズムを用いた期待事後確率 (EAP: Expected A Posteriori) 推定法が高精度であることが示されている [7], [45]．また，LDA のパラメータ推定においては，変分ベイズ法を用いた EAP 法 [17] と MCMC を用いた EAP 法 [46] が一般的である．MCMC は変分ベイズ法に比べて計算効率は劣るものの，実装が容易であり推定精度も高いことが知られている [47]．

IRT における MCMC アルゴリズムとしては，メトロポリスヘイスティングスとギブスサンプリングを組み合わせたアルゴリズム [7], [13], [28] が一般的であり，LDA では周辺化ギブスサンプリングを用いたアルゴリズム [46] が一般に採用されている．周辺化ギブスサンプリングは，特定のパラメータ集合を周辺化することで MCMC の推定効率を高めることができる手法であり，提案モデルでも LDA と同様に利用できる．以上より，本研究では，提案モデルのパラメータ推定アルゴリズムとして，メトロポリスヘイスティングスと周辺化ギブスサンプリングを組み合わせた MCMC アルゴリズムを開発する．

提案アルゴリズムでは，トピック分布と語彙分布

のパラメータである $\boldsymbol{\psi} = \{\boldsymbol{\psi}_{ij} | i \in \mathcal{I}, j \in \mathcal{J}\}$ と $\boldsymbol{\phi} = \{\boldsymbol{\phi}_t | t \in \mathcal{T}\}$ を周辺化し，トピック $\boldsymbol{Z} = \{Z_{ijn} | i \in \mathcal{I}, j \in \mathcal{J}, n \in \{1, \cdots, N_{ij}\}\}$ と IRT のモデルパラメータ $\boldsymbol{\xi} = \{\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\alpha}_r, \boldsymbol{\beta}_r, \boldsymbol{d}, \boldsymbol{\theta}\}$，重みベクトル $\boldsymbol{\omega}$ を，それぞれの条件付き事後分布からサンプリングする．ここで，$\boldsymbol{\alpha}_i = \{\log \alpha_{i=1}, \cdots, \log \alpha_{i=I}\}$，$\boldsymbol{\beta}_i = \{\beta_{i=1}, \cdots, \beta_{i=I}\}$，$\boldsymbol{\alpha}_r = \{\log \alpha_{r=1}, \cdots, \log \alpha_{r=R}\}$，$\boldsymbol{\beta}_r = \{\beta_{r=1}, \cdots, \beta_{r=R}\}$，$\boldsymbol{d} = \{d_{11}, \cdots, d_{RK}\}$，$\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_J\}$ とする．

以降では，提案アルゴリズムの詳細について述べる．また，以降の式展開のために，提案モデルのグラフィカルモデルを図 1 に示す．図中の τ_* は添字で表されるパラメータ $*$ の事前分布のパラメータ（ハイパーパラメータ）を表す．

7.1 トピック $\boldsymbol{Z_{ijn}}$ のサンプリング

ここで，$\boldsymbol{W}^{\backslash ijn} = \boldsymbol{W} \backslash \{W_{ijn}\}$，$\boldsymbol{Z}^{\backslash ijn} = \boldsymbol{Z} \backslash \{Z_{ijn}\}$ とすると，トピック Z_{ijn} の条件付き事後分布は図 1 の構造から次のように導ける．

$$
\begin{aligned}
&p(Z_{ijn} = t | W_{ijn}, \boldsymbol{W}^{\backslash ijn}, \boldsymbol{Z}^{\backslash ijn}, \theta_j, \boldsymbol{\omega}) \\
&\quad \propto p(W_{ijn} | Z_{ijn} = t, \boldsymbol{W}^{\backslash ijn}, \boldsymbol{Z}^{\backslash ijn}) \\
&\qquad p(Z_{ijn} = t | \boldsymbol{Z}^{\backslash ijn}) p(\theta_j | \boldsymbol{\omega}, Z_{ijn} = t, \boldsymbol{Z}^{\backslash ijn})
\end{aligned} \tag{12}
$$

ここで，式 (12) の右辺第 1 項は，Z_{ijn} のサンプリング確率に依存する項のみを残すように式変形すると次のように整理できる．

$$
\begin{aligned}
&p(W_{ijn} | Z_{ijn} = t, \boldsymbol{W}^{\backslash ijn}, \boldsymbol{Z}^{\backslash ijn}) \\
&\quad \propto \int p(W_{ijn} | \boldsymbol{\phi}_t) p(\boldsymbol{\phi}_t | \boldsymbol{W}^{\backslash ijn}, \boldsymbol{Z}^{\backslash ijn}) d\boldsymbol{\phi}_t \\
&\quad = \frac{N_{tv}^{\backslash ijn} + \gamma}{N_t^{\backslash ijn} + V\gamma}
\end{aligned} \tag{13}
$$

$N_{tv}^{\backslash ijn}$ は回答文 e_{ij} の n 番目の語彙を除いたとき，語彙 v にトピック t が割り当てられた頻度を表し，N_t は $\sum_{v=1}^{V} N_{tv}^{\backslash ijn}$ を表す．

また，式 (12) の右辺第 2 項を，Z_{ijn} のサンプリング確率に依存する項のみを残すように式変形すると次のように整理できる．

$$
\begin{aligned}
&p(Z_{ijn} = t | \boldsymbol{Z}^{\backslash ijn}) \\
&\quad \propto \int p(Z_{ijn} = t | \boldsymbol{\psi}_{ij}) \cdot p(\boldsymbol{\psi}_{ij} | \boldsymbol{Z}^{\backslash ijn}) d\boldsymbol{\psi}_{ij} \\
&\quad = \frac{N_{ijt}^{\backslash ijn} + \eta}{N_{ij}^{\backslash ijn} + T\eta} \propto N_{ijt}^{\backslash ijn} + \eta
\end{aligned} \tag{14}
$$

ここで，$N_{ijt}^{\backslash ijn}$ は回答文 e_{ij} の n 番目の語彙を除外したときの e_{ij} 内のトピック t の出現回数を表し，N_{ij} は $\sum_{t=1}^{T} N_{ijt}^{\backslash ijn}$ を表す．

式 (12) の右辺第 3 項は，$\{Z_{ijn} = t\} \cup \boldsymbol{Z}^{\backslash ijn}$ を所与としたときの，式 (10) 右辺の正規分布に従う θ_j の生起確率として計算できる．

7.2 IRT パラメータのサンプリング

IRT パラメータ $\boldsymbol{\xi}$ のサンプリングは，パラメータごとにメトロポリスヘイスティングスを繰り返すことで行う．具体的には，次の手順を繰り返してサンプリングを行う．

（1） 各パラメータ $\xi \in \boldsymbol{\xi}$ に対して，現在の値を所与とした提案分布 $N(\xi, \sigma_p^2)$ から，更新先のパラメータ値の候補点 ξ^* を生成する．ここで，提案分布の標準偏差 σ_p には 0.01 などの小さい値を用いる．

（2） 以下の採択確率に基づいて候補点 ξ^* を採択する．

$$
a(\xi^* | \xi) = min\left(\frac{p(\boldsymbol{U} | \xi^*, \boldsymbol{\xi}^{\backslash \xi}) g(\xi^* | \boldsymbol{\tau}_\xi)}{p(\boldsymbol{U} | \boldsymbol{\xi}) g(\xi | \boldsymbol{\tau}_\xi)}, 1 \right) \tag{15}
$$

ここで，$\boldsymbol{\xi}^{\backslash \xi} = \boldsymbol{\xi} \setminus \{\xi\}$ を表し，$g(\xi | \boldsymbol{\tau}_\xi)$ はパラメータ ξ に対する事前分布を表す．ただし，$\xi = \theta_j \in \boldsymbol{\theta}$ の場合には，採択確率は次式で与えられる．

$$
a(\xi^* | \xi) = min\left(\frac{p(\boldsymbol{U} | \xi^*, \boldsymbol{\xi}^{\backslash \xi}) p(\xi^* | \boldsymbol{\omega}, \boldsymbol{Z}_j)}{p(\boldsymbol{U} | \boldsymbol{\xi}) p(\xi | \boldsymbol{\omega}, \boldsymbol{Z}_j)}, 1 \right) \tag{16}
$$

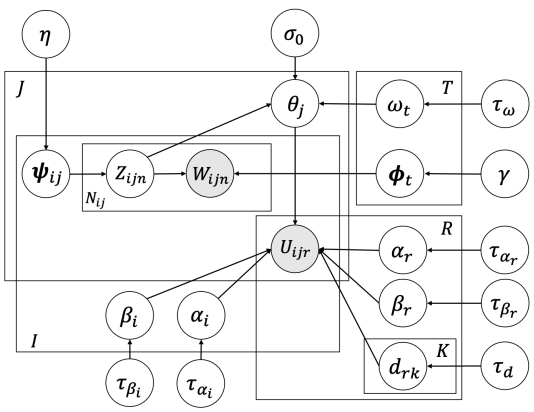

図 1 提案モデルのグラフィカル表現
Fig. 1 Graphical model representation of the proposed model.

ここで，$p(\xi|\boldsymbol{\omega}, \boldsymbol{Z}_{ij})$ は，式 (10) 右辺の分布に従う ξ の生起確率を表す．ただし，$\boldsymbol{Z}_j = \{Z_{ijn}|i \in \mathcal{I}, n = \{1, \cdots, N_{ij}\}\}$ とする．

式 (15) と式 (16) における $p(\boldsymbol{U}|\boldsymbol{\xi})$ は次式で定義できる．

$$p(\boldsymbol{U}|\boldsymbol{\xi}) = \prod_{j=1}^{J}\prod_{i=1}^{I}\prod_{r=1}^{R}\prod_{k=1}^{K}(P_{ijrk})^{x_{ijrk}}, \tag{17}$$

$$x_{ijrk} = \begin{cases} 1 : U_{ijr} = k, \\ 0 : otherwise. \end{cases} \tag{18}$$

（3） 採択確率に基づく採択・棄却の結果，候補点が採択されなかった場合には ξ^* を破棄し，元の値 ξ を次のパラメータ値として採用する．

7.3 トピックの重み $\boldsymbol{\omega}$ のサンプリング

重みパラメータ $\boldsymbol{\omega}$ のサンプリングは，IRT のモデルパラメータと同様にメトロポリスヘイスティングスとギブスサンプリングを組み合わせた手法で行う．具体的には，$\omega_t \in \boldsymbol{\omega}$ に対して，提案分布 $N(\omega_t, \sigma_p^2)$ から候補点 ω_t^* を生成し，以下の採択確率に基づいて候補点を採択する．

$$a(\omega_t^*|\omega_t) = min\left(\frac{p(\boldsymbol{\theta}|\omega_t^*, \boldsymbol{\omega}^{\backslash t}, \boldsymbol{Z})g(\omega_t^* \mid \boldsymbol{\tau}_{\omega_t})}{p(\boldsymbol{\theta}|\boldsymbol{\omega}, \boldsymbol{Z})g(\omega_t \mid \boldsymbol{\tau}_{\omega_t})}, 1\right)$$

ただし，

$$p(\boldsymbol{\theta}|\boldsymbol{\omega}, \boldsymbol{Z}) = \prod_{j=1}^{J} p(\theta_j|\boldsymbol{\omega}, \boldsymbol{Z}_{ij}) \tag{19}$$

とする．

7.4 トピック分布と語彙分布の推定

提案アルゴリズムでは，上記の手法に基づいてトピック $\boldsymbol{Z}$ とモデルパラメータ $\boldsymbol{\xi}, \boldsymbol{\omega}$ をサンプリングすると同時に，周辺消去した $\boldsymbol{\phi}$ と $\boldsymbol{\psi}$ を，トピックのサンプル $\boldsymbol{Z}$ を用いて次式で求める．

$$\phi_{tv} = \frac{N_{tv} + \gamma}{\sum_{v=1}^{V} N_{tv} + V\gamma} \tag{20}$$

$$\psi_{ijt} = \frac{N_{ijt} + \eta}{\sum_{t=1}^{T} N_{ijt} + T\eta} \tag{21}$$

ここで，N_{tv} は語彙 v にトピック t が割り当てられた回数を表し，N_{ijt} は回答文 e_{ij} におけるトピック t の出現回数を表す．また，$N_t = \sum_{v=1}^{V} N_{tv}$，$N_{ij} = \sum_{t=1}^{T} N_{ijt}$ である．

7.5 アルゴリズム

以上のサンプリングを繰り返し，得られたパラメータ・サンプルの期待値を点推定値とする．ただし，分布が収束したと推測されるまでのバーンイン期間は，パラメータの初期値の影響が残るため推定に利用しない．また，メトロポリスヘイスティングスは，サンプル間の自己相関が高いため，全てのサンプルは利用せず，一定のインターバル期間ごとに抽出したサンプルを採用する．以上のアルゴリズムの疑似コードを Algorithm 1 に示す．

Algorithm 1 MCMC algorithm for the proposed model.

```
Given maximum chain length M, burn-in period B, interval S.
Initialize parameters ξ, ω, and topic assignment Z
for loop = 1 to M do
  for each topic Z_ijn ∈ Z do
    Update Z_ijn from eq(12)
  end for
  for each ξ ∈ ξ do
    Sample ξ* ~ N(ξ, σ_p^2).
    Accept ξ* with probability α(ξ* | ξ).
  end for
  for each ω_t ∈ ω do
    Sample ω_t* ~ N(ω_t, σ_p^2).
    Accept ω_t* with probability α(ω_t* | ω_t).
  end for
  if t ≥ B and t%S = 0 then
    Calculate ψ, and φ using eq(20), (21)
    Store ξ, ω, ψ, φ
  end if
end for
return Average values of ξ, ω, ψ, φ
```

7.6 文章データのみを用いた能力値推定と得点予測

6. で述べたとおり，提案モデルでは，語彙分布と評価者特性，課題特性及び重みのパラメータが既知であれば，評点データが与えられていない受験者の能力を文章情報のみから推定することができる．具体的には，Algorithm 1 において，トピック Z_{ijn} と能力値 θ_j のサンプリング式を変更し，評価者特性と課題特性及び重みのパラメータについては更新を行わないようにしたアルゴリズムで推定できる．トピック Z_{ijn} のサンプリング式は次式で与えられる．

$$\begin{aligned} &p(Z_{ijn} = t|W_{ijn}, \boldsymbol{W}^{\backslash ijn}, \boldsymbol{Z}^{\backslash ijn}, \boldsymbol{\phi}, \theta_j, \boldsymbol{\omega}) \\ &\quad \propto p(W_{ijn}|Z_{ijn} = t, \boldsymbol{\phi}) \\ &p(Z_{ijn} = t|\boldsymbol{Z}^{\backslash ijn})p(\theta_j|\boldsymbol{\omega}, Z_{ijn} = t, \boldsymbol{Z}^{\backslash ijn}) \\ &\quad \propto \phi_{t,W_{ijn}}\left(N_{ijt}^{\backslash ijn} + \eta\right)p(\theta_j|\boldsymbol{\omega}, Z_{ijn} = t, \boldsymbol{Z}^{\backslash ijn}) \end{aligned} \tag{22}$$

表 1　実験で利用した論述式課題
Table 1　Essay tasks used in the experiment.

課題 1	高等教育における専門分野の選択は，学生本人の得意分野や興味を重視して行うべきと考える立場があります．一方で，専門分野は実用性や社会のニーズを重視して決めるべきと考える立場もあります．このテーマについてあなたの意見を述べてください．
課題 2	20 世紀には我々の生活を劇的に変化させる様々な発明がなされました．テレビや車，コンピュータなどの社会的にインパクトの大きい発明から，ボールペンやヘッドホン，電卓などの相対的にインパクトの小さな発明まであります．あなたの生活においてより重要な役割を担っているのは「大きな発明」でしょうか．それとも「小さな発明」でしょうか．このテーマについてあなたの意見を述べてください．
課題 3	メディアでは著名人や成功者を英雄（ヒーロー）のように取り上げます．しかし，あなたの身近には，日常の中で自然と素晴らしいことを為している人たちがいるでしょう．社会的に大きな偉業をなさなくとも日常の中で人々の役に立っているそうした人を真の英雄と呼べるのではないでしょうか．真の英雄についてあなたの意見を述べてください．
課題 4	科学技術の急速な進歩に伴い，私たちの生活はますます科学技術に依存するようになってきています．こうした科学技術への依存は人間自身の考える力を低下させてしまうのではないか，としばしば指摘されます．このテーマについてあなたの意見を述べてください．

このとき，語彙分布と評価者特性，課題特性及び重みのパラメータは事前に推定された値を所与とする．また，能力値 θ_j のサンプリングは事後分布 $p(\theta_j|\boldsymbol{U}_j,\boldsymbol{\xi}^{\backslash\theta_j},\boldsymbol{\omega},\boldsymbol{Z}_j) \propto p(\boldsymbol{U}_j|\boldsymbol{\xi})p(\theta_j|\boldsymbol{\omega},\boldsymbol{Z}_j)$ から行う．ここで，$\boldsymbol{U}_j = \{U_{ijr} \mid i \in \mathcal{I}, r \in \mathcal{R}\} \subset \boldsymbol{U}$，$p(\boldsymbol{U}_j|\boldsymbol{\xi}) = \prod_{i=1}^{I}\prod_{r=1}^{R}\prod_{k=1}^{K}(P_{ijrk})^{x_{ijrk}}$ とする．この分布は一般には解析的に求められないため，通常は **7.2** で説明したメトロポリスヘイスティングスに基づいてサンプリングを行う．しかし，ここでは，受験者 j の評点データが全て欠測の状況を考えているため，ゆう度項 $p(\boldsymbol{U}_j|\boldsymbol{\xi})$ は無視でき，$p(\theta_j|\boldsymbol{U}_j,\boldsymbol{\xi}^{\backslash\theta_j},\boldsymbol{\omega},\boldsymbol{Z}_j) \propto p(\theta_j|\boldsymbol{\omega},\boldsymbol{Z}_j)$ と書ける．すなわち，能力値 θ_j のサンプリングは式 (10) の正規分布に従って行えばよい．

また，提案モデルでは，このように推定された能力値を所与として未採点回答の期待得点を求めることも可能である．具体的には，文章 e_{ij} の期待得点 $\hat{U}_{ij}$ は次式で求められる．

$$\hat{U}_{ij} = \sum_{r=1}^{R}\frac{1}{R}\sum_{k=1}^{K} k \cdot P_{ijrk} \tag{23}$$

このとき，P_{ijrk} は，事前に推定された評価者・課題の特性パラメータを所与として計算する．

8. 評価実験

ここでは，実データ実験を通して提案モデルの有効性を評価する．

8.1 実データ

本研究では実データを収集するために，次の被験者実験を行った．

34 名の大学生と大学院生に対して，4 つの論述式課題を行わせ，各課題に対して提出された回答文を 10 名の評価者に採点させた（各評価者に 34 名 ×4 課題 = 136 件の回答文を全て採点させた）．本実験で利用した論述式課題を表 1 に示す．これらの課題は，National Assessment of Educational Progress（NAEP）の 2002 年 [48] と 2007 年 [49] で出題された課題を日本語に翻訳したものであり，専門知識や特別な事前知識を必要としない内容となっている．また，評価者による採点は，NAEP grade 12 [49] で使用されたルーブリックを日本語に訳して作成した 5 段階カテゴリーの評価基準を用いて行わせた．執筆された回答文の文字数は，平均が 600.41，標準偏差が 104.41 であった．

表 2　評点データの記述統計量
Table 2　Descriptive statistics for the rating data.

	平均値	標準偏差	各評価カテゴリーの出現回数 1	2	3	4	5
評価者 1	3.537	0.633	1	12	52	55	16
評価者 2	3.419	0.605	0	15	58	54	9
評価者 3	2.537	0.690	20	52	41	17	6
評価者 4	2.912	0.679	4	45	52	29	6
評価者 5	3.404	0.515	0	9	68	54	5
評価者 6	3.566	0.491	5	2	43	83	3
評価者 7	3.691	0.530	0	6	48	64	18
評価者 8	3.110	0.520	2	30	60	39	5
評価者 9	2.743	0.335	0	41	90	4	1
評価者 10	2.794	0.606	7	41	61	27	0
課題 1	3.135	0.748	14	77	124	99	26
課題 2	3.132	0.744	12	61	155	94	18
課題 3	3.126	0.786	7	69	147	108	9
課題 4	3.291	0.800	6	46	147	125	16
全体	3.171	0.887	39	253	573	426	69

ここで，評点データの記述統計量として，評価者別・課題別及び全体での評点の平均値と標準偏差，各評価カテゴリーの出現回数を表 2 に示す．表から，これらの統計量が評価者や課題ごとに異なることが確認でき，評価者と課題の特性を考慮した能力測定の必要性

が示唆される．また，これらの統計量の差異は，課題間に比べて，評価者間の方が大きい傾向が読み取れる．本研究で基礎モデルとして採用した宇都・植野のモデル [11] は，既存モデルより多様な評価者特性を表現できるため，本データのように評価者間の差異が相対的に大きい場合に適していると解釈できる．

本論文では，上記の実験で収集した評点データとテキストデータを用いて提案モデルの有効性を評価する．

8.2 能力推定精度の評価

本節では，提案モデルによる能力測定精度の評価を行う．このために，トピック数 T を $[1, 15]$ の区間で変化させながら，次の実験を行った．

（1） 実データを用いて MCMC によるパラメータ推定を行った．MCMC はバーンイン 30,000，インターバル 100，最大ループ数 50,000 とし，五つの独立のチェーンを初期値を変えて実行し，得られた結果の平均を点推定値とした．ただし，$T = 1$ のときには $\omega_1 = 0$ と固定し，ω_1 の推定は行わなかった．パラメータの事前分布とハイパーパラメータは先行研究の設定 [11], [36], [50] に合わせて次のとおりとした．

$$\log \alpha_i \sim N(0.1, 0.4) \tag{24}$$

$$\log \alpha_r \sim N(0.0, 0.5) \tag{25}$$

$$\beta_i, \beta_r, d_{rk}, \omega_t \sim N(0.0, 1.0) \tag{26}$$

$$\eta = 1/T, \gamma = 1/VT, \sigma_0 = 1.0 \tag{27}$$

回答文集合から抽出する語彙の集合としては，ストップワードを除去した名詞，動詞，形容詞，接続詞，副詞を用いた．ストップワードの判定基準は，1）全回答文のうち二つ以下の回答文でしか利用されていない，2）全回答文の半分以上の回答文で利用されている，とした．結果として，語彙数は 201 となった．

（2） 完全データとして与えられた評点データから，数名の評価者で採点を行った場合の評点データをシミュレートするために，各回答文に $n \in \{1, 2, 3, 4\}$ 名の評価者をランダムに割り当て，評価者が割り当てられていない回答文の評点データを欠測させた．

（3） 手順（2）で作成された欠測データを用いて，各学習者の能力値を MCMC により再推定した．推定は，語彙分布と評価者特性，課題特性及び重みのパラメータを所与として，**7.6** の方法で行った．

（4） 手順（3）で推定された能力値と手順（1）で推定された能力値との平均平方 2 乗誤差（RMSE: Root Mean Square Error）を計算した．

（5） 手順（2）～（4）を 10 回繰り返し，RMSE の平均を求めた．

実験結果を図 2 に示す．図の横軸はトピック数を表し，縦軸は RMSE の値を表す．また，図中の One Rater，Two Raters，Three Raters，Four Raters のプロットが，それぞれ評価者が 1 名，2 名，3 名，4 名のときの結果を表す．なお，$T = 1$ の提案モデルは，式 (3) で与えられる従来の IRT モデルと一致する点に注意されたい．

実験結果から，従来モデルに対応する $T = 1$ の場合に比べて，提案モデルでは RMSE が大幅に低下していることがわかる．これは提案モデルが，回答文の内容的な特徴を能力測定値に適切に反映できたためと考えられる．また，提案モデルでは，トピック数が 4 までは単調に RMSE が低下し，以降ではおおむね同程度の性能を示している．おおむね性能が収束したとみられるトピック数 $T \geq 4$ の提案モデルと従来モデルの性能を比較すると，提案モデルにおける評価者 n 名のときの誤差が，従来モデルにおける評価者 $n + 1$ 名のときの誤差と同等以下となっている．これは，提案モデルでは，文章情報を利用したことで，従来モデルにおいて評価者を 1 名追加した場合と同程度以上の能力測定精度の改善が達成できたことを示している．

以上の実験結果から，提案モデルでは，回答文の情報を活用することで能力測定精度を改善でき，従来モデルにおける回答文あたりの評価者数減少に伴う能力測定精度の低下を緩和できることが確認できた．

なお，ベイズ推定では，パラメータ推定値が事前分

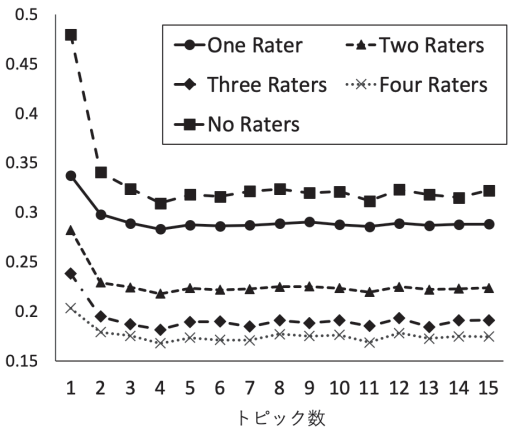

図 2 能力推定誤差の評価結果
Fig. 2 Ability estimation errors for each number of topics.

布の期待値周辺に引き寄せられることで，見た目の推定誤差が小さくなる縮小（Shrinkage）と呼ばれる現象が知られているが，提案モデルが従来モデルと比べて RMSE を低減できた主な理由は縮小ではないと考えられる．提案モデルでは，従来モデルとは異なり，受験者ごとに異なる能力値分布が仮定されるため，全ての受験者の能力推定値を特定の値周辺に偏らせるだけでは RMSE は小さくならない．**8.5** で例示するように，提案モデルでは，トピック分布から予測される能力値が受験者の妥当な順序づけを与えており，この情報が各受験者の能力値の事後分布に適切に反映されたため，従来モデルより高い能力測定精度を示したと考えられる．

8.3 文章情報のみを用いた能力測定精度

ここでは，評点データが与えられていない受験者の能力を文章情報のみから推定した場合の能力測定精度について評価する．このために，トピック数 T を $[1, 15]$ の区間で変化させながら次の手順の実験を行った．

（1） **8.2** の実験手順（1）と同様に，実データを用いて MCMC によるパラメータ推定を行った．

（2） 評点データを全て欠測させ，手順（1）で推定された語彙分布と評価者特性，課題特性及び重みのパラメータを所与として，**7.6** の方法で各受験者の能力を再推定した．この手順は，受験者の能力を文章情報のみから推定していることに対応する．

（3） 手順（1）で推定された能力値と手順（2）で推定された能力値の RMSE を計算した．

実験結果を図 2 の「No Raters」のプロットとして示した．従来モデルに対応する $T = 1$ では，評点データも文章情報も能力推定に利用できないため，能力測定誤差が著しく大きくなっている．他方で，提案モデルを利用した場合（$T > 1$ の場合）には，精度が大幅に改善していることがわかる．また，前節の実験と同様に，トピック数 $T = 4$ までは単調に RMSE が減少し，以降はおおむね同程度の性能を示している．更に，トピック数 $T \geq 4$ の提案モデルでは，評点データを利用していないにもかかわらず，従来モデルにおいて評価者 1 名の評点データを利用した場合を上回る能力測定精度を達成していることがわかる．本実験結果から，提案モデルでは，評点データが与えられていない場合でも，従来モデルを用いて評価者 1 名の評点データから推定する場合と同程度の能力測定が実現できることが示された．

8.4 未採点回答の得点予測精度

本節では，提案モデルを用いた未採点回答の得点予測の性能評価を行う．このために，トピック数 T を $[1, 15]$ の区間で変化させながら，次の手順で実験を行った．

（1） **8.2** の実験手順（1）と同様に，実データを用いて MCMC によるパラメータ推定を行った．

（2） 前節の実験手順（2）と同様に，評点データを全て欠測させたあと，手順（1）で推定された語彙分布と評価者特性，課題特性及び重みのパラメータを所与として，**7.6** の方法で各受験者の能力を推定した．

（3） 手順（2）で求めた能力推定値と手順（1）で得られた評価者と課題パラメータを用いて期待得点 $\hat{U}_{ij}$ を式 (23) を用いて求め，期待得点 $\hat{U}_{ij}$ と完全データを用いて計算した観測平均得点 $U_{ij} = \sum_r U_{ijr}/R$ との RMSE を求めた．

（4） 比較のために，各回答文に $n \in \{1, \cdots, 5\}$ 名の評価者をランダムに割り当て，割り当てた評価者の評点データから求めた各回答文の平均得点と，完全データから求めた観測平均得点 U_{ij} との RMSE を計算した．この手順は評価者の割り当てを変えながら 10 回繰り返し，RMSE の平均値を求めた．

結果を図 3 に示す．図の横軸はトピック数を表し，縦軸は RMSE の値を表す．また，図 3 では，実線のプロット（「Proposed」と表記）が提案モデルで予測した得点と完全データから求めた観測平均得点の誤差を表し，破線（「n Rater(s)」と表記）が n 名の評価

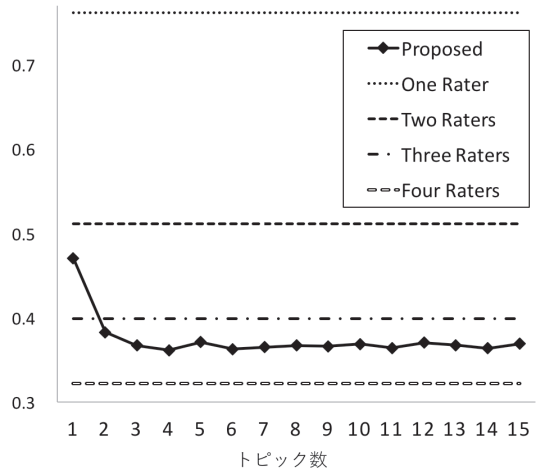

図 3 評点予測誤差の評価結果
Fig. 3 Score expectation errors for each number of topics.

者のデータのみで求めた平均得点と完全データから求めた観測平均得点の誤差を表す．

図 3 から，これまでの実験と類似した傾向として，以下の結果が読み取れる．1）従来モデルに対応する $T = 1$ では予測誤差が著しく大きい．2）提案モデルを利用した場合には精度が大幅に改善する．3）トピック数 $T = 4$ までは誤差が単調に減少し，以降はおおむね同程度の性能を示す．

更に，提案モデルによる予測得点の精度を評価者 n 名の平均得点を利用した場合の精度と比較すると，提案モデルでは，評価者 3 名の平均得点を上回る予測精度を達成したことが確認できる．この結果から，提案モデルは，未採点回答の得点予測としても妥当な結果を与えることが確認できた．

8.5 考　　察

ここでは，提案モデルにおいて性能がおおむね収束したとみなせるトピック数 $T = 4$ の場合を例として，実データ適用で得られたトピック分布や語彙分布について考察する．表 3 に受験者ごとのトピック出現確率 $\psi_{jt} = \sum_{i \in \mathcal{I}} \psi_{ijt}/I$ と能力推定値を，表 4 に各トピックに対する重みパラメータの推定値と各トピックにおいて出現確率の高かった 10 語彙を示す．ここで，表 3 における ψ_{jt} は $\bar{Z}_{jt}$ に対応する概念であり，MCMC の過程で計算されるトピック出現確率 $\bar{Z}_{ijt}$ を最終的に推定されたトピックの出現確率 ψ_{ijt} に置き換えたものである．同様に，表 3 中の $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\psi}_j$（ただし，$\boldsymbol{\psi}_j = \{\psi_{j1} \cdots \psi_{jT}\}$）は $\boldsymbol{\omega}^{\mathrm{T}}\bar{\boldsymbol{Z}}_j$ に対応する概念として解釈できる．

表 3 の ψ_{jt} の値から，受験者ごとにトピックの出現傾向に差異があることが読み取れる．例えば，受験者 6 や 10，23 はトピック 1 や 2 の出現確率が相対的に低く，トピック 3 や 4 の出現確率が相対的に高いことがわかる．反対に，受験者 12 や 33 はトピック 1 や 2 の出現確率が相対的に高く，トピック 3 や 4 の出現確率が相対的に低い傾向が読み取れる．ここで，表 4 から，各トピックの能力値への重みはトピック 1 と 2 は正であり，トピック 3 と 4 が負となっていることがわかる．したがって，提案モデルでは，トピック 1 と 2 の出現確率が高い受験者ほどトピック分布から推定される能力値 $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\psi}_j$ が高くなり，トピック 3 と 4 の出現確率が高い受験者ほどその値が低く推定される．実際，上述した受験者 6 や 10，23 は $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\psi}_j$ が相対的に低く，受験者 12 や 33 はこの値が相対的に高いことが確認できる．

表 3　$T = 4$ における受験者ごとのトピック分布と能力値
Table 3　Topic distribution and ability estimate of each examinee for $T = 4$.

j	ψ_{j1}	ψ_{j2}	ψ_{j3}	ψ_{j4}	$\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\psi}_j$	θ_j
1	0.113	0.230	0.241	0.417	0.502	0.399
2	0.179	0.213	0.159	0.449	0.585	0.800
3	0.160	0.202	0.227	0.410	0.534	0.702
4	0.190	0.201	0.205	0.405	0.585	0.888
5	0.161	0.158	0.196	0.485	0.440	0.016
6	0.140	0.144	0.193	0.524	0.369	0.230
7	0.124	0.180	0.261	0.435	0.422	1.006
8	0.132	0.153	0.263	0.452	0.381	0.673
9	0.164	0.267	0.236	0.333	0.678	0.741
10	0.109	0.165	0.178	0.549	0.351	0.416
11	0.159	0.293	0.239	0.309	0.723	0.767
12	0.128	0.395	0.187	0.290	0.873	0.698
13	0.172	0.221	0.170	0.437	0.591	0.848
14	0.135	0.290	0.217	0.358	0.670	0.271
15	0.171	0.215	0.193	0.421	0.579	0.544
16	0.165	0.139	0.204	0.493	0.408	−0.631
17	0.144	0.222	0.225	0.409	0.543	0.512
18	0.094	0.196	0.218	0.492	0.394	0.409
19	0.188	0.217	0.174	0.421	0.614	0.499
20	0.222	0.209	0.238	0.331	0.669	0.460
21	0.213	0.199	0.227	0.361	0.629	0.964
22	0.191	0.162	0.148	0.499	0.501	0.453
23	0.055	0.161	0.245	0.540	0.248	−0.371
24	0.134	0.210	0.168	0.489	0.493	0.352
25	0.131	0.163	0.215	0.492	0.393	0.477
26	0.175	0.174	0.164	0.487	0.497	0.854
27	0.104	0.186	0.190	0.521	0.387	0.390
28	0.192	0.205	0.163	0.441	0.593	0.367
29	0.161	0.196	0.169	0.475	0.516	0.316
30	0.129	0.195	0.279	0.397	0.465	0.497
31	0.105	0.190	0.271	0.433	0.408	0.796
32	0.150	0.188	0.184	0.478	0.481	0.851
33	0.213	0.259	0.241	0.287	0.756	0.857
34	0.182	0.211	0.190	0.417	0.591	0.452

表 4　$T = 4$ における各トピックの出現確率上位 10 語彙と重みパラメータ
Table 4　Top 10 most frequent vocabularies and weight parameter for each topic when $T = 4$.

t	出現確率上位 10 語彙	ω_t
1	発明，生活，重要，役割，インパクト，コンピュータ，担う，英雄，大きい，車	1.696
2	分野，専門，興味，学生，選択，社会，研究，教育，重視，ニーズ	1.876
3	技術，人間，科学，力，低下，しまう，進歩，れる，せる，自身	−0.116
4	人，思う，的，より，それ，ない，られる，できる，自分，社会	−0.220

表 4 に示したトピックごとの頻出語彙を確認すると，能力値に正に寄与するトピック 1 や 2 では課題に関連した語彙が多く出現しており，能力値に負に寄与するトピック 3 や 4 ではこれらの割合が少なく，一般的な言い回しが多い傾向が読み取れる．このことから，本実験では，1）主題に関連する語彙の利用割合が多

表 5　課題 1 への回答文例
Table 5　Answer text examples for essay task 1.

受験者 23	私はどちらの意見にも否定しません．なぜなら，学生本人たちには無限の可能性や本人たちも自覚していない得意分野が存在している可能性があるからです．学生本人たちは，その無限の可能性を生かすか生かさないかは彼らの自由であり，尊重しなければならないと思います．二つ目の意見で，専門分野は実用性や社会のニーズを重視して決めるべきと考えている人もいますが，そうすると専門分野で増える分野と減る分野に分かれてしまうと思います．また，そうすると将来その減った分野で人手不足に陥るといった社会問題も起きかねません．しかし，ここでは高等教育における専門分野に触れているので，必ずしもそれが社会にすぐに出る人を対象としている訳でもないので，そういった心配はないかと思います．高等専門学校，または専門学校大学に進学するもしないも学生本人たちの自由であり，そこで新たに自分の中に秘めていた可能性をみつけ，得意分野に生かし，社会に役立てる人材を育てていけばいいと私は思います．また，人間だれしもできる出来ないといったことで区別するのではなく，一人一人にあるオリジナルの才能やセンスを社会に生かせればよいと思います．
受験者 33	私は高等教育における専門分野の選択は，学生本人の得意分野や興味を重視して行うべきであると考えます．私がこのように考える理由は，社会のニーズは日々変化しているため，実用性などを重視した専門分野は一意に決めることがあまりに困難であると考えるためです．現在社会のニーズとして求められていると私が思う専門分野の一つに機械学習があります．この機械学習という分野はもちろん以前から研究されていた専門分野になりますが，社会のニーズとして求められているのはここ最近のことであると私は考えています．つまり，実用性などを重視させて専門分野を選択させると機械学習がまだ社会のニーズに求められていなかった時代には別の学問が選択されることになり，すぐに社会のニーズを満たすことができない分野を学んでしまうことになってしまいます．このような観点から実用性や社会のニーズを重視しようとしても，とても困難であるという考えから私は高等教育における専門分野の選択は，学生本人の得意分野や興味を重視して行うべきであると考えます．

く，2）主題と直接には関係しない表現が少ない非冗長な文章，ほど提案モデルが高い評価を与える傾向があることがわかる．ここで，回答文の例として，$\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\psi}_j$ が低い受験者 23 と，この値が大きい受験者 33 の課題 1 への回答文を表 5 に示す．これらの回答文におけるトピック 1 と 2 の頻出 10 単語の出現頻度は受験者 23 が 24 回，受験者 33 が 44 回，トピック 3 と 4 の頻出 10 単語の出現頻度は受験者 23 が 22 回，受験者 33 が 13 回であり，$\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\psi}_j$ が高い受験者 33 の方がトピック 1 や 2 の出現頻度が多く，トピック 3 や 4 の出現頻度が少なくなっている．

次に，トピック情報に基づく能力予測値と評点データも加味して推定された能力値 θ_j との関係を分析するために，表 3 における $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\psi}_j$ と θ_j の相関係数を求めた．結果として，相関係数は 0.44 となり，1%（$t = 2.81$）で有意な相関が認められた．これは，トピック分布に基づく能力予測値が受験者の妥当な順序づけを与えることを意味している．提案モデルでは，この情報を受験者の能力値に適切に反映できたため，従来モデルより高精度な能力測定が達成できたと考えられる．

8.6　トピック数の決定

提案モデルを実際に利用するためには，トピック数 T を利用者が決定する必要がある．LDA のトピック数をデータから決定する方法としてはパープレキシティが広く利用されるが，提案モデルでは文章データに加えて評点データも扱うためこの方法は単純には利用できない．他方で，Akaike Information Criterion (AIC) [51] や Bayesian Information Criterion (BIC) [52] などの情報量基準に基づくトピック数の決定もしばしば利用される．しかし，これらの基準は推定量の漸近正規性を仮定しており [53], [54]，LDA はこの性質を満たさないため，LDA や LDA を部分的に含む提案モデルではこれらの情報量基準の利用は適切ではない．漸近正規性を仮定しない情報量基準としては，対数周辺ゆう度が一般的である．LDA や提案モデルの対数周辺ゆう度を直接評価することは困難であるが，パラメータ推定に MCMC を採用した場合，この値を近似的に求めることができる [55]．具体的には，MCMC 過程でパラメータ値のサンプルが得られるたびに，その値を所与としてモデルの対数ゆう度を求め，得られた対数ゆう度の集合について調和平均を取ることで求められる．この手法はゆう度計算のみで容易に求められるため，LDA のトピック数の決定にも利用されてきた（e.g., [36], [37], [46]）

そこで，ここでは，近似対数周辺ゆう度を用いた提案モデルのトピック数推定について評価を行う．本実験では，**8.1** の実データを用いて，提案モデルの近似対数周辺ゆう度をトピック数を $[1, 15]$ の区間で変化させながら算出した．結果を図 4 に示す．図 4 では，横軸がトピック数，縦軸が近似対数周辺ゆう度の値を表す．近似対数周辺ゆう度が高いトピック数ほど望ましいと解釈される．図から，$T = 4$ までは値が急速に増加し，$T = 4$ 以降で増加量が緩慢になる傾向が読み取れる．$T = 4$ は，これまでの実験で提案モデルの性能が収束したとみなせるトピック数と一致する．この結果は，近似対数周辺ゆう度の増加量が緩慢になるト

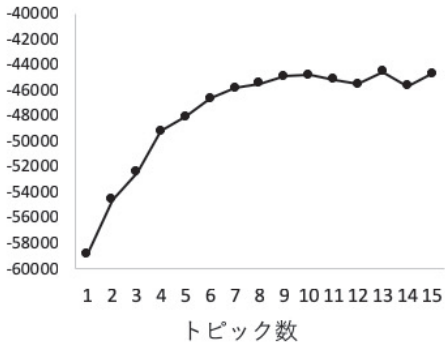

図 4　対数周辺ゆう度によるトピック数の推定結果
Fig. 4　Log marginal likelihood for each number of topics.

ピック数を採用することで，能力測定精度や評点予測精度の高いトピック数を選択できることを示唆する．なお，図 4 では，$T = 4$ 以降も近似対数周辺ゆう度値が増加傾向を示しており，増加量が緩慢になる点の選定には恣意性が残る．しかし，これまでの実験において $T \geq 4$ はおおむね同等の性能を示していたことから，$T > 4$ を採用しても性能の極端な変動はないと考えられる．

また，近似対数周辺ゆう度を利用する以外のトピック数推定法として，**8.3** と **8.4** で行った実験を利用することも考えられる．これらの実験は任意のデータセットにおいて実施できるため，これらの実験結果に基づいて，性能が高く，解釈のしやすいトピック数を決定することも可能である．

9. むすび

本研究では，評価対象物あたりの評価者数が少ない場合に IRT による能力測定の精度が低下する問題を解決するために，受験者が執筆した回答文の内容を能力測定の補助情報として利用できる新たなモデルを提案した．また，提案モデルのパラメータ推定手法として MCMC アルゴリズムによるベイズ推定法を提案した．更に，実データ実験により，提案モデルが能力測定の精度改善に有効であり，未採点の回答文をもつ受験者の能力推定とその回答文の得点予測についても妥当な結果を与えることを示した．

今後は，様々な実データへの適用を通して，提案モデルの汎用性を確認したい．なお，提案モデルでは能力分布の期待値のみが回答文の内容に依存すると仮定したが，分布の分散にも文章情報を反映させることで能力測定の精度をさらに改善できる可能性がある．本検討についても今後の課題とする．また，本研究では，トピックモデルとして LDA を活用したが，近年では様々な LDA の拡張モデル (e.g., [56], [57]) が提案されている．今後は，LDA の代わりにこれらの拡張モデルを利用することで，更なる精度改善が可能かを検証したい．

近年では，ディープラーニングを用いた自動採点技術が人工知能分野で多数提案されている (e.g., [58]〜[60])．これらの技術は，人間の評価と比べると必ずしも十分な精度を達成できてはいないが，従来の自動採点手法に比べて大幅に性能が向上している．LDA では文書内の単語出現順序に関する情報を活用できないため，文脈などの一部の情報を活用できない可能性が高いが，Long Short Term Memory に代表されるディープラーニング手法では，より詳細なテキスト情報を扱うことが可能となると期待できる．本研究のアプローチに基づいてディープラーニングと項目反応理論を統合した能力測定手法の開発も今後の課題の一つとしたい．

謝辞　本研究は JSPS 科研費 17H04726，17K20024 の助成を受けたものです．

文　献

[1] R. Schendel and A. Tolmie, “Assessment techniques and students’ higher-order thinking skills,” Assessment & Evaluation in Higher Education, vol.42, no.5, pp.673–689, 2017.
[2] Y. Abosalem, “Beyond translation: Adapting a performance-task-based assessment of critical thinking ability for use in rwanda,” Int. J. Secondary Education, vol.4, no.1, pp.1–11, 2016.
[3] Y. Rosen and M. Tager, “Making student thinking visible through a concept map in computer-based assessment of critical thinking,” J. Educational Computing Research, vol.50, no.2, pp.249–270, 2014.
[4] O.L. Liu, L. Frankel, and K.C. Roohr, “Assessing critical thinking in higher education: Current state and directions for next-generation assessment,” ETS Research Report Series, vol.2014, no.1, pp.1–23, 2014.
[5] H.J. Bernardin, S. Thomason, M.R. Buckley, and J.S. Kane, “Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability,” Human Resource Management, vol.55, no.2, pp.321–340, 2016.
[6] 宇都雅輝，植野真臣，“パフォーマンス評価のため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.55–75, 2016.
[7] M. Uto and M. Ueno, “Item response theory for

peer assessment," IEEE Trans. Learning Technologies, vol.9, no.2, pp.157–170, 2016.
[8] N.L.A. Kassim, "Judging behaviour and rater errors: An application of the many-facet Rasch model," GEMA Online Journal of Language Studies, vol.11, no.3, pp.179–197, 2011.
[9] C.M. Myford and E.W. Wolfe, "Detecting and measuring rater effects using many-facet Rasch measurement: Part I," J. Applied Measurement, vol.4, pp.386–422, 2003.
[10] T. Eckes, "Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis," Language Assessment Quarterly, vol.2, no.3, pp.197–221, 2005.
[11] 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，" 信学論（D），vol.J101-D, no.1, pp.211–224, Jan. 2018.
[12] T. Eckes, Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments, Peter Lang Pub. Inc., 2015.
[13] 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：MCMC アルゴリズムに基づく推定，" 教育心理学研究，vol.58, no.2, pp.163–175, 2010.
[14] 宇佐美慧，"論述式テストの運用における測定論的問題とその対処，" 日本テスト学会誌，vol.9, no.1, pp.145–164, 2013.
[15] G. Engelhard, "Constructing rater and task banks for performance assessments," J. Outcome Measurement, vol.1, no.1, pp.19–33, 1997.
[16] 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，" 信学論（D），vol.J101-D, no.6, pp.895–908, 2018.
[17] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," J. Machine Learning Research, vol.3, pp.993–1022, 2003.
[18] D.M. Blei and J.D. McAuliffe, "Supervised topic models," Proc. 20th International Conference on Neural Information Processing Systems, pp.121–128, 2007.
[19] F.M. Lord, Applications of item response theory to practical testing problems, Erlbaum Associates, 1980.
[20] 独立行政法人情報処理推進機構，"IT パスポート試験，" https://www3.jitec.ipa.go.jp/JitesCbt/
[21] 公益社団法人医療系大学間共用試験実施評価機構，"臨床実習開始前の「共用試験」第 14 版（平成 28 年度），" http://www.cato.umin.jp/e-book/14/index.html
[22] F.B. Baker and S.H. Kim, Item Response Theory: Parameter Estimation Techniques, Statistics, textbooks and monographs, Marcel Dekker, 2004.
[23] D. Andrich, "A rating formulation for ordered response categories," Psychometrika, vol.43, no.4, pp.561–573, 1978.
[24] G. Masters, "A Rasch model for partial credit scoring," Psychometrika, vol.47, no.2, pp.149–174, 1982.
[25] F. Samejima, "Estimation of latent ability using a response pattern of graded scores," Psychometrika, vol.34, Suppl.1, pp.1–97, 1969. https://link.springer.com/article/10.1007%2FBF03372160
[26] E. Muraki, "A generalized partial credit model," Handbook of Modern Item Response Theory, eds. by W.J. van derLinden and R.K. Hambleton, pp.153–164, Springer, 1997.
[27] 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," J. Educational and Behavioral Statistics, vol.27, no.4, pp.341–366, 1999.
[28] R.J. Patz and B.W. Junker, "Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses," J. Educational and Behavioral Statistics, vol.24, pp.342–366, 1999.
[29] M. Ueno and T. Okamoto, "Item response theory for peer assessment," Proc. IEEE International Conference on Advanced Learning Technologies, pp.554–558, 2008.
[30] 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, 2018.
[31] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," J. American Society for Information Science, vol.41, no.6, pp.391–407, 1990.
[32] T. Hofmann, "Probabilistic latent semantic indexing," Proc. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.50–57, 1999.
[33] D. Duan, Y. Li, R. Li, R. Zhang, and A. Wen, "Rank-topic: Ranking based topic modeling," IEEE 12th International Conference on Data Mining, pp.211–220, 2012.
[34] X. Li, J. Ouyang, and X. Zhou, "Supervised topic models for multi-label classification," Neurocomputing, vol.149, pp.811–819, 2015.
[35] S. Jameel, W. Lam, and L. Bing, "Supervised topic models with word order structure for document classification and retrieval learning," Information Retrieval Journal, vol.18, no.4, pp.283–330, 2015.
[36] M. Uto, S. Louvigné, Y. Kato, T. Ishii, and Y. Miyazawa, "Diverse reports recommendation system based on latent dirichlet allocation," Behaviormetrika, vol.44, no.2, pp.425–444, 2017.
[37] S. Louvigné, M. Uto, Y. Kato, and T. Ishii, "Social constructivist approach of motivation: social media messages recommendation system," Behaviormetrika, vol.45, no.1, pp.133–155, 2018.
[38] J. Zhu, A. Ahmed, and E.P. Xing, "MedLDA: maximum margin supervised topic models for regression

and classification," Proc. 26th International Conference on Machine Learning, pp.1257–1264, 2009.
[39] 奥村　学，佐藤一誠，トピックモデルによる統計的潜在意味解析，コロナ社，2015.
[40] F. Li, S. Wang, S. Liu, and M. Zhang, "SUIT: A supervised user-item based topic model for sentiment analysis," Proc. Twenty-Eighth AAAI Conference on Artificial Intelligence, pp.1636–1642, 2014.
[41] S.M. Gerrish and D.M. Blei, "Predicting legislative roll calls from text," Proc. International Conference on International Conference on Machine Learning, pp.489–496, 2011.
[42] 堂前友貴，関　洋平，"半教師ありトピックモデルにより選択した地域特徴語を用いた twitter ユーザの生活に関わる地域の推定，" 情処学論，vol.7, no.3, pp.1–13, 2014.
[43] F. Rodrigues, B. Ribeiro, M. Lourenço, and F.C. Pereira, "Learning supervised topic models from crowds," Third AAAI Conference on Human Computation and Crowdsourcing, pp.160–168, 2015.
[44] X. Zheng, Y. Yu, and E.P. Xing, "Linear time samplers for supervised topic models using compositional proposals," Proc. 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1523–1532, 2015.
[45] J.-P. Fox, Bayesian item response modeling: Theory and applications, Springer, 2010.
[46] T.L. Griffiths and M. Steyvers, "Finding scientific topics," Proc. National Academy of Sciences, vol.101, no.Suppl.1, pp.5228–5235, 2004.
[47] A. Asuncion, M. Welling, P. Smyth, and Y.W. Teh, "On smoothing and inference for topic models," Proc. International Conference on Uncertainty in Artificial Intelligence, pp.27–34, 2009.
[48] H. Persky, M. Daane, and Y. Jin, "The nation's report card: Writing 2002," Technical report, National Center for Education Statistics, 2003.
[49] D. Salahu-Din, H. Persky, and J. Miller, "The nation's report card: Writing 2007," Technical report, National Center for Education Statistics, pp.1–72, 2008.
[50] M. Taddy, "On estimation and selection for topic models," Proc. International Conference on Artificial Intelligence and Statistics, eds. N.D. Lawrence and M.A. Girolami, vol.22, pp.1184–1193, 2012.
[51] H. Akaike, "A new look at the statistical model identification," IEEE Trans. Autom. Control, vol.19, pp.716–723, 1974.
[52] G. Schwarz, "Estimating the dimensions of a model," Annals of Statistics, vol.6, pp.461–464, 1978.
[53] S. Watanabe, "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory," J. Machine Learning Research, pp.3571–3594, 2010.
[54] S. Watanabe, "A widely applicable bayesian information criterion," J. Machine Learning Research, vol.14, no.1, pp.867–897, 2013.
[55] M. Newton and A.E. Raftery, "Approximate Bayesian inference by the weighted likelihood bootstrap," J. Royal Statistical Society. Series B: Methodological, vol.56, no.1, pp.3–48, 1994.
[56] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," Proc. 22nd International Conference on World Wide Web, pp.1445–1456, 2013.
[57] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei, "Hierarchical dirichlet processes," J. American Statistical Association, vol.101, no.476, pp.1566–1581, 2006.
[58] Y. Farag, H. Yannakoudakis, and T. Briscoe, "Neural automated essay scoring and coherence modeling for adversarially crafted input," Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.263–271, Association for Computational Linguistics, 2018.
[59] D. Alikaniotis, H. Yannakoudakis, and M. Rei, "Automatic text scoring using neural networks," Proc. 54th Annual Meeting of the Association for Computational Linguistics, pp.715–725, Association for Computational Linguistics, 2016.
[60] K. Taghipour and H.T. Ng, "A neural approach to automated essay scoring," Proc. 2016 Conference on Empirical Methods in Natural Language Processing, pp.1882–1891, Association for Computational Linguistics, 2016.

（2019 年 1 月 17 日受付，3 月 29 日再受付，
4 月 19 日早期公開）

宇都　雅輝　（正員）

2013 年電気通信大学大学院情報システム学研究科博士後期課程修了．博士 (工学)．長岡技術科学大学を経て，2015 年より電気通信大学助教に着任，現在に至る．e テスティング，e ラーニング，人工知能，ベイズ統計，自然言語処理などの研究に従事．