情報処理学会論文誌 Vol.62 No.9 1586–1604 (Sep. 2021)

研究利用可能な小論文データに基づく参照文書を利用した小論文採点手法の開発

竹内 孔一[1,a)] 大野 雅幸[1,†1] 泉仁 宏太[1,†2] 田口 雅弘[2] 稲田 佳彦[3] 飯塚 誠也[4]
阿保 達彦[1] 上田 均[1]

受付日 2021年1月5日, 採録日 2021年6月7日

概要：近年，小論文の自動採点手法が英語圏において多数研究されている．この要因として研究利用可能な大規模な採点済み小論文データセットが公開されていることがあげられる．一方で，日本語では母語話者が記述した研究で広く利用されている採点済み小論文データは現段階では見受けられない．そこで，本論文ではまず模擬試験を実施し，他機関でも研究利用可能な小論文データを構築する．ルーブリックを利用して 2 名で採点することで揺れの少ない採点済み答案データを構築する．また，小論文自動採点の先行研究において，多数の採点済み答案を必要とする手法が提案されているが，たとえば試験問題など小論文課題が新規である場合，多数の答案に対して事前採点することは容易ではないと考えられる．そこで，本論文では小論文自動採点手法の枠組として，小論文課題に沿った参考文書または 1 件の模範答案のみを参照文書として利用可能な場合の採点手法について論じる．アプローチとして参照文書と小論文との類似度を利用した採点手法を複数提案し，実験的に人手による評価値と相関が高い手法を明らかにする．類似度計算法として，形態素の頻度，Wikipedia を利用した idf 値，LSI，LDA，分散表現ベクトル，BERT を用いた文書ベクトルを利用する．評価実験の結果，形態素の頻度と idf 値を利用した手法が他の手法に比べて人手による評価値と複数の課題で相関が高く有効な手法であることを明らかにする．

キーワード：小論文自動採点，採点済み答案の構築，LDA，分散表現ベクトル，BERT

Development of Essay Scoring Methods Based on Reference Texts with Construction of Research-Available Japanese Essay Data

Koichi Takeuchi[1,a)] Masayuki Ohno[1,†1] Kouta Motojin[1,†2] Masahiro Taguchi[2]
Yoshihiko Inada[3] Masaya Iizuka[4] Tatsuhiko Abo[1] Hitoshi Ueda[1]

Received: January 5, 2021, Accepted: June 7, 2021

Abstract: In recent years, many automated essay scoring methods have been developed for English essays. One of the reasons for supporting the active studies is the existence of common scored essay datasets in English that are available for research purpose. On the other hand, in Japanese, scored essay data written by Japanese native speakers is not found at this moment. Thus, we collect essays in Japanese by conducting practice tests of writing essays with permission of examinees in order that the collected essays can be available for other research institutions. Manual grading scores are elaborated by two graders using rubrics to produce reliable scores. In previous studies of automatic essay scoring, pre-scored essays are required, it must be, however, difficult to prepare a lot of pre-scored essays for automated essay scoring if the essay task is to write for a new theme in an examination. Thus, we discuss the automatic essay scoring methods that take only reference texts or a model answer of the essay task without pre-scored essays. We take an approach to automated essay scoring based on the similarity between the reference text and the essay, and then experimentally find the best method among the prepared several similarity evaluation methods that are frequency of morphemes, idf values calculated on Wikipedia, LSI, LDA, word-embedding vectors, and document vectors produced by BERT. Experimental results show that the method using the frequency of morphemes with idf values gives a highest correlation with the human annotated scores in several essay tasks.

Keywords: automatic essay assessment, construction of scored essays, LDA, word-embedding, BERT

1. はじめに

本論文では研究利用可能な日本語小論文データを作成し，課題に関連した参考文書または1件の模範答案（これらを参照文書とする）を利用した場合の有効な小論文採点手法を実験的に明らかにする．

小論文の自動採点は文献 [50] で述べられているようにすでに英語圏でシステムが開発され，実用化されている．たとえば，E-rater [3] や IntelliMetric [14] は米国の経営大学院の入学試験 GMAT（Graduate Management Admission Test）で利用された．また Intelligent Essay Assessor（IEA）[15] は Pearson Educational Technologies が提供する商用システムや試験で用いられている [50]．しかしながら自動採点のみで評価するのではなく米国の公的な試験において自動採点は人との併用で利用される [50]．

先行研究で提案されている自動採点の手法の多くは事前に採点済み答案を必要とする．先にあげた採点システム E-rater，Intelligent Essay Assessor，IntelliMetric では，重回帰分析，Latent Semantic Indexing（LSI），ルール発見の手法が利用されているが，各システムは人手による採点済み答案を利用して評価関数の内部変数を調整する [50]．また近年ではより大規模な採点済み答案を用いて機械学習 [12], [21], [28], [29] を適用する方法や，深層学習を利用した採点手法 [1], [13], [19], [23], [26], [33], [36], [38], [39] が提案されている．しかしながら，試験問題における小論文課題で少人数の答案を採点する場合，事前の採点済み答案を大量に必要とする手法は適用が困難であることが予測される*1．

入試では参考文書を読ませて考えを書かせる課題が出題されている*2．また授業で小論文課題を与える場合は授業内容に則した参考文書が想定できる．そこで，本研究では参考文書を読んで問いに答える小論文課題を仮定する*3．つまり大量の採点済み答案を仮定するのではなく，参考文書および模範答案を仮定した採点手法について検討する．

採点済み答案を利用しない手法として日本語小論文採点システム Jess が提案されている [49]．Jess は文書の内容の良さを測る特徴量と潜在的意味解析の1つである LSI を利用し，答案と参考文書との類似度を計算することで評価する．採点済み答案を必要としない利点がある一方で，他の類似度を適用した場合の採点結果を比較することができない．英語では公開されている小論文データ（5.2 節参照）で手法を実験的に比較することが可能である一方で，日本語では実験で利用している小論文は非公開であるため [52]，手法の比較が容易ではない*4．

そこで，本研究では，公開可能な*5小論文採点データを構築し，構築した小論文データを基に複数の採点手法を適用して比較することで有効な採点手法を明らかにする．採点手法として，形態素の頻度を利用する方法，Wikipedia による形態素の idf 値を利用する方法，LSI および Latent Dirichlet Allocation（LDA）[6] を利用する方法，分散表現ベクトルを利用する方法，Bidirectional Encoder Representations from Transformers（BERT）[11] を転移学習なしで利用する方法を適用する．実験の結果，形態素の頻度と idf 値を利用した手法が他の手法に比べて人手で付与した採点との相関が高いことを明らかにする．

2. 自動採点手法を評価するための小論文データの構築

試験などの採点済み小論文を利用した先行研究はある [56] が，利用されている小論文は公開されていないため，他の研究機関が試験答案を利用することは容易ではない．そこで，日本語の小論文データを構築するために模擬試験を実施し，あらかじめ参加者に了解を得て研究利用できる形で答案を収集して小論文データを構築する．小論文データとして必要な要素は，参考文書，課題，答案，人手による採点結果である．各要素における問題点と対処した方法を以下に記述する．

2.1 小論文の課題のテーマと種類

小論文の採点手法を評価するために複数のテーマと難易度の異なる課題を設定する．分野の異なる専門家2名がそれぞれテーマを設定し，ある専門用語や概念について説明を求める記述自由度の少ない課題から，解答者の考えを記

1 岡山大学学術研究院自然科学学域
Graduate School of Natural Science and Technology, Okayama University, Okayama 700–8530, Japan
2 岡山大学学術研究院社会文化科学学域
Graduate School of Humatities and Social Sciences, Okayama University, Okayama 700–8530, Japan
3 岡山大学学術研究院教育学学域
Graduate School of Education, Okayama University, Okayama 700–8530, Japan
4 岡山大学全学教育・学生支援機構
Institute for Education and Student Services, Okayama University, Okayama 700–8530, Japan
†1 現在，住友電工情報システム株式会社
Presently with Sumitomo Electric Information Systems Co. Ltd.
†2 現在，株式会社 NTT データ MSE
Presently with NTT Data MSE Corporation
a) takeuc-k@okayama-u.ac.jp

*1 少人数とはここでは約千人以下を仮定している．理由は，英語の小論文データ ASAP を利用した先行研究では約 1,000 件以上を学習データとして利用しているためである（たとえば文献 [23] 参照）．
*2 たとえば文献 [43], [44] では大阪大学の AO 入試や岡山大学での AO 入試，首都大学東京の前期日程などで文書を読んで要約や考えを問う問題が出題されている．
*3 このような課題は英語圏では Response to Text Assessment として研究されている [29], [30], [37], [39]．
*4 5.2 節に示すように日本語学習者が記述した小論文が近年構築されたが，研究利用可能な日本語母語話者が記述した採点済み小論文データ（100 文字以上）は現段階では見当たらない．
*5 言語資源協会（https://www.gsk.or.jp/）から公開を予定している．

述させる自由度の高い課題を設定する．模擬試験の形式として AO 入試などで利用される講義形式を設定する．つまり解答者は講義を聴いた後，講義内容に関する課題に対して小論文を記述する．課題で求める小論文解答の文字数を 100 字から 800 字までに設定する．

模擬試験では解答者は講義を受けて課題に関連する内容を理解するが，講義をそのまま採点システムで利用するのは難しい．そこで講義内容を書き起こした参考文書を講義ごとに作成し採点システムの入力として利用する．参考文書を入力とした採点システムを構築することで，講義形式で行われる課題だけでなく文章を読んで問いに答える形式の課題にも同様の手法で採点することが可能になる．

2.2 答案データの作成

上記の課題に対して解答者が筆記で答案を作成する．採点システムの入力には電子テキストデータが必要であるため書き起こしが必要になる．当初は OCR 文字読み取り装置による電子化を計画していたが，1) 予備実験において文字誤りが多かったこと，2) 存在しない文字などの誤りに対応できないことから人手による書き起こしを作成した．このとき，解答者が存在しない文字を記述している場合は●で表記した．自動採点システムではこの書き起こしテキストを解答者の答案として採点する．

2.3 採点基準と人手による採点

小論文の採点ではどのような側面で評価するかに関する評価軸の設定と人手による採点の揺れの軽減が課題となる．以下それぞれについて記述する．

文献 [48] では小論文を評価する側面は多岐にわたるため，評価軸を出題する側が決める必要があることが記述されている．たとえば入試であれば試験を実施する側のアドミッションポリシーに基づいて評価軸を設定する必要がある．小論文データの評価軸として理解力，論理性，妥当性，文書力を設定し，5 段階評価を行う [45]．それぞれ，「講義の内容を理解して設問に即した内容を記述しているか」，「文章の構成が論理的であるか」，「内容が専門的な知見からみて深く記述されているか」，「文章は文法的に正しく記述されているか」を評価する．通信教育会社の高校生向けの小論文指導書 [46] と比較すると，指導書では (a)「設問を読み取る」，(b)「資料を読み取る」(c)「意見を深め，固める」(d)「文章を構成する」(e)「正しい表記で書く」の 5 つの方法が提案されている．これより，理解力はおおむね (a)，(b) に対応し，妥当性，論理性，文書力はそれぞれ (c)，(d)，(e) の対応する評価と考えられる．

これらの評価軸の中で，本論文では理解力を採点する手法に焦点を当てる．まず 1 つの評価軸に絞る理由は 4 つの評価軸がそれぞれ異なる基準で評価されているため，各評価軸に対するアプローチがまったく異なることが予測されるためである．1 つの評価軸に絞ることで提案手法に対する考察が容易になると考えられる．さらに理解力に焦点を当てる理由は，理解力は設問に即した内容を記述しているかを評価するため，設問で用意される参考文書を手掛かりとした採点手法を開発することで理解力の評点に近い手法が構築できるのではないかと考えたためである．よって以降では理解力の評価軸を中心に議論する．

一方，採点によるゆれを軽減する方法として小論文採点の先行研究 [56] から評価の際にはルーブリックを作ることが提案されている．よって，ルーブリックを基に採点する．採点は各論文に対して 2 名の評価者で行い，先行研究 [59] を参考に評価値の平均を四捨五入した値を最終評点として実験で利用する．次節に構築した小論文データについて説明する．

表 1 小論文データの講義内容

Table 1 Themes and the lengths of reference texts in the essay dataset.

講義記号	講義内容	参考文書の文字数
g	グローバリゼーションの光と影	2,650
s	自然科学の構成と科学教育	2,586
e	東アジア経済の現状	6,361
c	批判的思考とニセ科学	2,526

2.4 小論文データ

小論文データは各講義ごとに，講義内容を書き言葉で整理した参考文書，課題，解答者の小論文答案，人手による採点結果からなる．まず，講義内容と参考文書について**表 1** に示す．講義は経済分野 2 件（講義 g と e）と科学教育分野 2 件（講義 s と c）である．それぞれの講義に対する参考文書は講義 e 以外 2,600 字前後である．

次に各講義における小論文課題の文字数制限と有効答案数[*6]について**表 2** に示す．各講義には複数の課題があり，文字数制限として 100 字から最長で 800 文字まで記述する．各課題の内容は付録の A.2 節に記述しているが，おおよそ課題 1 と 2 が講義に即した内容を問う課題であるのに対して，課題 3 では著者の考えなど記述内容に自由度を持たせた内容である．

表 3 に理解力に対して 2 名の評価者が付与した評点間の一致度と評点の平均を示す．一致度として 4.2 節で導入する Quadratic Weighted Kappa（QWK）を利用する．まず QWK の値では課題 c_1 で 0.885 が最も高く課題 g_3 で 0.399 が最も低い値となった．小論文答案が 100 文字以内の課題である s_1 と c_1 はどちらも QWK が 0.8 を超えており評価者間の一致度が高い．一方で，課題 3 は課題 1 と 2 に比べて QWK が低い．これは上記にも述べたように課題

[*6] 白紙など 5 文字以下の回答を排除している．

表 2 小論文の各課題と答案数
Table 2 Maximum length of characters and number of essays for each prompt.

講義記号	課題番号	文字数	小論文答案数
g	1	300	328
	2	250	327
	3	300	327
s	1	100	327
	2	400	325
	3	500 から 800	327
e	1	300	290
	2	250	288
	3	300	288
c	1	100	290
	2	400	290
	3	500 から 800	290

表 3 評価者間の比較（QWK）と各評価者の平均点
Table 3 Inter-rater agreement in quadratic weighted kappa (QWK) and their average scores of essays.

講義記号	課題番号	評価者間	平均点	
			評価者 A	評価者 B
g	1	0.636	2.784	2.723
	2	0.786	2.566	2.446
	3	0.399	3.012	3.122
s	1	0.844	4.223	4.382
	2	0.669	3.098	2.892
	3	0.468	4.034	3.786
e	1	0.671	2.959	3.059
	2	0.696	2.649	2.799
	3	0.620	2.528	2.490
c	1	0.885	3.879	3.800
	2	0.660	3.028	2.714
	3	0.517	3.224	3.093

3 の内容が他の課題よりも自由度が高いことが原因として考えられる．課題 g_3 では 300 文字以内であるが生活に即した事例の幅が大きく評価に揺れが生じて他に比べて低い QWK になったと考えられる．

また，平均点では多くの課題が 2.5 から 3 点の周辺であるが課題 s_1，s_3 および c_1 は 4 点以上または 4 点に近い評点である．これらの課題は解答者にとって容易な内容の課題と考えられる．

こうした人手による評点を基に理解力に対する最終評点を先行研究 [59] を参考に評価者が付与した評点の平均で求める．**表 4** に最終評点および最終評点と各評価者の評点との QWK を示す．

各講義で課題番号が小さい場合は QWK が高く最終評点と比較して各評価者の揺れは少ない．一方，課題番号が大きい場合は記述内容の自由度が大きくなるため QWK は下がる傾向にある．文献 [4] から上記の結果は評価者のゆれは問題になるほど大きくないと考えられる[*7]．

表 4 最終評点と評価者との比較（QWK）
Table 4 QWK scores between final scores and scores evaluated by each human rater.

講義記号	課題番号	評価者 A	評価者 B
g	1	0.802	0.880
	2	0.872	0.952
	3	0.745	0.722
s	1	0.979	0.867
	2	0.820	0.897
	3	0.783	0.801
e	1	0.894	0.840
	2	0.909	0.847
	3	0.821	0.844
c	1	0.933	0.958
	2	0.813	0.881
	3	0.757	0.838

3. 参照文書を利用した小論文答案の評価手法

本研究では事前に採点したデータを利用せずに小論文の理解力を採点する手法について検討する．各小論文課題には参照文書を利用できるため，講義に関連した内容を直接質問する設問の場合[*8]，設問を理解できている解答者の小論文は参照文書と類似することが期待される．また，自由度の高い設問の場合でも，参照文書は設問に密接した内容であるため，高い評価を得る小論文答案は類似する部分があると考えられる．

先行研究においても参照文書との類似度による評価によって採点する手法が提案されている．短答式の先行研究 [25] では 1 つの模範答案に対して余弦類似度を利用した採点手法を提案している．小論文の先行研究 [49] では良さを測る特徴量ベースの手法と LSI による類似度ベースの両方を利用している．しかしながら特徴量ベースの手法は具体的には修辞（「文章の長さ」，「漢字/カナの割合」や語彙の多様性ほか）および文の接続に関する表現（「すなわち」，「たとえば」，「だが」ほか）などを特徴として利用しており，おおむね本小論文データの「論理性」と「文書力」に対応する評価方法と考えられる．よって本論文で対象とする「理解力」に対する評価法としては取り入れない．一方で，LSI による類似度ベースの手法の考え方は本論文で対象とする「理解力」の評価法としてアプローチが一致する．そこで本研究では参照文書と小論文答案の類似度を比較し，類似度が高いほど理解力が高いと仮定して答案を評価する．

[*7] 文献 [4]（P.82）によると Kappa 値は 0.4 から 0.6 で fair，0.6 から 0.75 で good，0.75 を超えると excellent と記されている．
[*8] たとえば，用語の定義の問題や講義内容の一部を要約する設問の場合．

文書間の類似度を測定する手法が情報検索や自然言語処理分野で研究されており [22], [42], [54]，形態素の連続性を考慮した手法（n-gram など），文書どうしの共通形態素の集合を利用した手法（Jaccard 係数など），文書ベクトルを利用した余弦類似度を利用した手法が提案されている．

まず形態素の連続性を利用した手法に着目すると，要約のタスクで形態素 n-gram を利用した場合，1-gram（unigram）の方が，n=2 や 3 といった長い単位の n-gram に比べて精度が高いことが実験的に示されている [54]．また小規模ではあるが，小論文評価の先行研究 [27] においても，n が 2 以上の n-gram を利用した類似度よりも unigram の方が有効であることが示されている．よって，本研究では形態素 unigram による手法に注目する．

形態素を利用した類似度では上述のように共通形態素の集合を利用した手法が提案されているが上述の先行研究 [27] では頻度を利用した手法が有効であることが示されている．よって本研究も共通形態素の頻度を利用した手法を利用する．先行研究 [3] では小論文内容の評価において形態素の重みに対して idf 値を利用している．よって idf 値を利用した手法も取り入れる．さらに文書ベクトルを利用した手法では，先行研究 [49] では LSI を利用している．そこで，LSI も含めて LDA，分散表現ベクトル，BERT による類似度計算を取り入れる．

文献 [11] では BERT にファインチューニングを適用することで高い精度が得られることを示しているが評点が付与された答案が複数用意できない本枠組では適用することは難しい．一方で，文献 [23] では BERT の出力を文に対するベクトルとして固定して利用した場合でもファインチューニングした場合と比較して QWK に大きな違いがないことを示している．このことから BERT の出力するベクトルには文の特徴をとらえるために有効なベクトルが生成されていることが期待できる．よって事前学習済みの BERT の出力するベクトルを文の特徴量として類似度計算に取り入れる．

類似度計算の後，5 点に正規化する．正規化する方法としては文献 [25] では最大値で正規化しており，文献 [31] では最小値も考慮した正規化（min-max 正規化）を利用している．本研究では形態素を利用した手法には最大値による正規化[*9]，余弦類似度を利用した手法には最小値を考慮した正規化[*10]を適用する．5 点に正規化した後，4.2 節に示すように 1 から 5 点に離散化して評価する．これは，評点は整数であるため最終的に自動採点の結果を評点に反映させる段階で整数化する必要があるためである．先行研究ではこうした整数化の手法は示されていないため，簡素な手法として式 (28) に示すように階級にわけて整数化を行う．

以下ではまず，形態素を利用した 2 種類の類似度評価法について詳細を記述する．

3.1 共通形態素の頻度を利用した評価法

参照文書および小論文答案の両方に出現した内容語相当の形態素を抽出し，その頻度の合計を参照文書と答案の類似度とする．答案には専門用語が多く現れることから，内容語の取り出すために，NEologd 辞書[*11] [58] を利用した形態素解析器 MeCab[*12]を利用する．MeCab を適用し，名詞，動詞，形容詞の形態素を内容語として取り出し，両文書に共通して出現した形態素の頻度を類似度として利用する．

ここで類似度の評価法として 2 通りの方法を定義する．1 つは小論文答案内の形態素の頻度を加算することで類似度とする手法である．具体的には，参照文書を R，答案を E とし，内容語相当の形態素を e として参照文書に対する答案の類似度 $sim\_wd(R, E)$ を下記のように定義する．

$$sim\_wd(R, E) = \sum_{e \in E \cap e \in R} freqE(e) \quad (1)$$

$freqE(e)$ は答案 E 内での形態素 e が出現した頻度を表す．

もう一方は，参照文書における頻度も考慮する方法である．具体的には答案 E に出現する内容語相当の形態素 e の出現頻度 $freqE(e)$ と参照文書 R 内での出現頻度 $freqR(e)$ のうち，最小の頻度を合計した値を類似度 $sim\_wdm(R, E)$ として下記の式で求める．

$$sim\_wdm(R, E) = \sum_{e \in E} \min(freqR(e), freqE(e)) \quad (2)$$

ここで $freqR(e)$ は参照文書 R 内で形態素 e が出現した頻度を表している．

得られた類似度をすべての答案に対して計算し，最大値が 5 になるように正規化した値を答案の評価スコアとする．ここで，採点する答案の集合を $\mathscr{E}$ とすると答案 E の評価スコア $grade(E)$ は下記のように求める．

$$highest\_value = \max_{E \in \mathscr{E}}(sim(R, E)) \quad (3)$$

$$grade(E) = \frac{sim(R, E) \cdot 5.0}{highest\_value} \quad (4)$$

ここで式 (3) と式 (4) の $sim(R, E)$ に対して式 (1) を適用する場合は $sim\_wd(R, E)$ を代入し，式 (2) の場合は $sim\_wdm(R, E)$ を代入して評価スコアを求める．

3.2 Wikipedia による idf を利用した評価法

上述の提案手法は専門用語も一般的な言葉も同様に扱うが，小論文課題では専門分野に関する記述を求めることが考えられる．よって専門用語を重視した評価法を構築する．

形態素が専門的な内容かどうかをスコア付けするため

[*9] 3.1 節を参照．
[*10] 最小値を考慮する理由は 3.3 節に記述．
[*11] https://github.com/neologd/mecab-ipadic-neologd
[*12] http://taku910.github.io/mecab/

に，Wikipedia を利用した idf 値を計算する．Wikipedia の 1 記事を 1 文書として，ある形態素 w を含む文書数を $docf(w)$ とすると w の idf 値は下記のように求めることができる．

$$idf(w) = \log \frac{L}{docf(w)} \tag{5}$$

ここで L は Wikipedia の全文書数とする．形態素解析辞書には NEologd を利用するため，「ジニ係数」や「格差拡大」など複合語で構成された専門用語を高く評価することができる．

式 (5) の idf は上述した形態素の頻度を利用した手法に付加して用いる．つまり，式 (1) に適用する場合の類似度を $sim\_wdidf(R,E)$ とし，式 (2) に適用する場合を $sim\_wdmidf(R,E)$ としてそれぞれ文書間類似度を下記のように定義する．

$$sim\_wdidf(R,E) = \sum_{e \in E \cap e \in R} freqE(e) \cdot idf(e) \tag{6}$$

$$sim\_wdmidf(R,E) = \sum_{e \in E} \min(freqR(e), freqE(e)) \cdot idf(e) \tag{7}$$

ここで得られた sim_wdidf および sim_wdmidf を式 (3) と式 (4) の sim と置き換えて利用することで答案 E の最終的なスコア $grade(E)$ を求める．

3.3 LSI を利用した評価法

LSI は潜在的意味解析の 1 種であり [41]，低ランク近似を利用した特異値分解で文書を潜在的な特徴ベクトルで表現することができる [10], [51]．よって文書間の内容が類似していれば，各文書における LSI の特徴ベクトルどうしは類似することが期待できる．文書が D 件あり，全文書の形態素の種類数を V とする．このとき，行を形態素，列を文書にして，各文書での形態素による tf-idf 値を要素とする行列を $\mathbf{X}$ とする．ある形態素 w におけるある文書 d での tf-idf 値を $tfidf(w)_d$ とすると上記の式 (5) を利用して

$$tfidf(w)_d = tf(w)_d \cdot idf(w) \tag{8}$$

と記述できる．ここで $tf(w)_d$ は文書 d における形態素 w の出現頻度である．特異値分解を利用して $\mathbf{X}$ を下記の制約を満たす行列に分解する．

$$\mathbf{X} = \mathbf{WSD}^t \tag{9}$$

ここで $\mathbf{W}$，$\mathbf{D}$ はそれぞれ，直交行列であり，$\mathbf{WW}^t = \mathbf{I}_t$，$\mathbf{DD}^t = \mathbf{I}_d$ のように転置行列との積により単位行列となる性質を有する．$\mathbf{S}$ は対角行列で対角要素は特異値である．LSI では特異値を大きい順に k 個用いた低ランク行列 $\hat{\mathbf{S}}(k \times k)$ を利用して $\mathbf{X}$ の近似行列 $\hat{\mathbf{X}}(V \times D)$ を下記の式で求める．

$$\hat{\mathbf{X}} = \hat{\mathbf{W}}\hat{\mathbf{S}}\hat{\mathbf{D}}^t \tag{10}$$

ここで $\hat{\mathbf{W}}(V \times k)$ と $\hat{\mathbf{D}}(D \times k)$ はそれぞれ $\hat{\mathbf{S}}$ に合わせて新たに計算する．

これらの行列を利用して k 次元で特徴化した文書のベクトルを作成する．評価したい答案 E があったとき，形態素解析を適用し形態素 tf-idf ベクトル $\boldsymbol{x}_E(V \times 1)$ を作成する．このとき，答案 E に対する文書ベクトル $\boldsymbol{d}_E(1 \times k)$ は下記の式で求まる [10], [51]．

$$\boldsymbol{d}_E = \boldsymbol{x}_E^t \hat{\mathbf{W}} \hat{\mathbf{S}}^{-1} \tag{11}$$

ここで $\hat{\mathbf{S}}^{-1}$ は $\hat{\mathbf{S}}$ の逆行列を表す．

同様に式 (11) を利用して参照データ R に対する文書ベクトル $\boldsymbol{d}_R$ を求める．答案と参照データの各文書ベクトルの余弦を類似度として求める．

$$sim\_LSI(R,E) = \frac{\boldsymbol{d}_R \cdot \boldsymbol{d}_E}{||\boldsymbol{d}_R|| ||\boldsymbol{d}_E||} \tag{12}$$

ここで，$\boldsymbol{d}_R \cdot \boldsymbol{d}_E$ は $\boldsymbol{d}_R$ と $\boldsymbol{d}_E$ の内積を表す．

最大値が 5 になるように正規化する方法として余弦類似度の最小値を考慮して下記の式で最終的なスコア $grade(E)$ を計算する．

$$lowest\_value = \min_{E \in \mathscr{E}}(sim\_LSI(R,E)) \tag{13}$$

$$grade(E) = \frac{(sim\_LSI(R,E) - lowest\_value) \cdot 5.0}{highest\_value - lowest\_value} \tag{14}$$

ここで，$hightest\_value$ は式 (3) の sim を sim_LSI に置きかえて求める．式 (4) に変えて式 (14) を使う理由は予備実験において余弦類似度の最小値が課題によっては大きな値となり類似度がスコアに反映されにくい例が見受けられたためである．

実験では Wikipedia の文書データを利用して，各ページを文書として $\mathbf{X}$ を作成し，式 (10) を利用して各行列を作成する．

3.4 LDA を利用した評価法

LDA（Latent Dirichlet Allocation）[6], [17] はディリクレ分布を利用した統計的潜在意味解析の 1 種であり [41]，文書 d が K 個のトピック[*13]から構成されると考え，潜在トピックベクトル $\boldsymbol{\theta}_d = (\theta_{d1}, \cdots, \theta_{dK})^t$ を文書から推定する．よって，文書間の内容が類似していれば，各文書における LDA の潜在トピックベクトルどうしは類似することが期待できる．文書 d からトピックベクトル $\boldsymbol{\theta}_d$ を求める手法の詳細は A.1 節に記述し[*14]，ここでは LDA を利用してどのように答案を評価するかについて記述する．

[*13] K は人手で決定する（4.3 節参照）．
[*14] LDA の計算に gensim のパッケージ https://radimrehurek.com/gensim/models/ldamodel.html を利用する．

大まかなステップとして，大規模文書データ（Wikipedia）を利用して LDA の内部パラメータを学習し，その後，答案および参照データに LDA を適用して各文書に対応するトピックベクトルを獲得する．

まず Wikipedia 文書に対して形態素解析を適用し，式 (8) と同様に，各文書 d で出現する各形態素 w に対する tf-idf 値を計算する．これを A.1 節の式 (A.7) と式 (A.8) において $c_{dw} = tfidf(w)_d$ と代入することで，LDA の内部パラメータを計算する．推定して得られた $\boldsymbol{\mu}_d$，$\boldsymbol{\xi}_k$ を利用してディリクレ分布のパラメータ $\boldsymbol{\alpha}$ と $\boldsymbol{\eta}$ を更新する．更新したパラメータを $\boldsymbol{\alpha}_{wiki}$ と $\boldsymbol{\eta}_{wiki}$ とする．

次に，トピックベクトルを求めたい文書 d' に対して出現した各形態素 w に対して式 (8) で tf-idf 値を計算し，上記と同様に式 (A.7) と式 (A.8) に $c_{d'w}$ を代入する．ここでディリクレ分布のパラメータとして $\boldsymbol{\alpha}_{wiki}$ と $\boldsymbol{\eta}_{wiki}$ を利用し，文書 d' に対する $\boldsymbol{\mu}_{d'}$ を求める．$\boldsymbol{\mu}_{d'}$ は文書 d' に対するトピックベクトル $\boldsymbol{\theta}_{d'}$ を生成するディリクレ分布のパラメータである（式 (A.4) 参照）．そこで，下記のように平均を求めることで $\boldsymbol{\theta}_{d'}$ を得る．

$$\mathbf{E}[\theta_{d'k}] = \frac{\mu_{d'k}}{\sum_{i=1}^{K} \mu_{d'i}} \tag{15}$$

$$\boldsymbol{\theta}_{d'} = (\mathbf{E}[\theta_{d'1}], \cdots, \mathbf{E}[\theta_{d'K}])^t \tag{16}$$

上記の手法で答案 E のトピックベクトル $\boldsymbol{\theta}_E$ と，参照データ R のトピックベクトル $\boldsymbol{\theta}_R$ を求め，ベクトルの余弦類似度を下記のように定義する．

$$sim\_LDA(R,E) = \frac{\boldsymbol{\theta}_R \cdot \boldsymbol{\theta}_E}{||\boldsymbol{\theta}_R|| ||\boldsymbol{\theta}_E||} \tag{17}$$

3.3 節と同様に式 (14) で答案 E の点数 $grade(E)$ を求める．このとき式 (3) の sim および式 (13) と式 (14) の sim_LSI を sim_LDA に置きかえて計算する．

3.5 分散表現ベクトルを利用した評価法

形態素の分散表現ベクトルを利用して文書間の類似度を測定する．具体的には文書 d に出現した各形態素 w_i に対して分散表現ベクトル $\boldsymbol{w}_i$ を求め，文書に出現した形態素ベクトルに対して平均して文書ベクトル $\boldsymbol{v}_d$ を求める．

$$\boldsymbol{v}_d = \frac{\sum_i^{N_d} \boldsymbol{w}_i}{N_d} \tag{18}$$

ここで N_d は文書 d における形態素数である．答案 E の文書ベクトル $\boldsymbol{v}_E$ と，参照データ R の文書ベクトル $\boldsymbol{v}_R$ を求めて文書間類似度を下記のように定義する．

$$sim\_vec(R,E) = \frac{\boldsymbol{v}_R \cdot \boldsymbol{v}_E}{||\boldsymbol{v}_R|| ||\boldsymbol{v}_E||} \tag{19}$$

3.4 節と同様に，式 (3) の sim および式 (13) と式 (14) の sim_LSI を sim_vec に置きかえて答案 E の点数 $grade(E)$ を求める．

3.6 BERT の文ベクトルを利用した評価法

日本語の Wikipedia で事前学習した BERT の文ベクトルを利用して文書間の類似度を測定する．事前学習モデルとして，MeCab の IPADIC に基づく WordPiece により Wikipedia をトークン化して学習したモデルを利用する[*15]．本論文では多数の事前採点済み答案を利用しない手法を明らかにする目的から，転移学習を利用せず BERT を文書ベクトルを獲得するモデルとして利用する[*16]．

BERT を利用して文書ベクトルを獲得する方法について説明する．入力として文書 d の前後に [CLS] および [SEP] のトークンを付与し，BERT に最初の文として入力する．2 文目は空として入力し，BERT の出力層における [CLS] に対応するユニットを文書ベクトル $\boldsymbol{t}_d$ とする．ここで参照データでは 2,000 文字を超えるものがあるため BERT の最長トークン数である 512 を超える場合がある．そこで文書内の句点を利用して 1 文ごとに区切り，各文をトークン化して文単位で連結し，512 トークンを以下になるように文書を分割する．

文書 d を 512 トークン以下に分割した M_d 個の文書を $d = d_1, d_2, \ldots, d_{M_d}$ とする．分割した文書 d_i に対して文書ベクトル $\boldsymbol{b}_{d_i}$ を BERT の [CLS] に対応する出力層から取り出し，平均することで最終的な BERT による文書ベクトル $\boldsymbol{t}_d$ を得る．

$$\boldsymbol{t}_d = \frac{\sum_i^{M_d} \boldsymbol{b}_{d_i}}{M_d} \tag{20}$$

答案 E の文書ベクトル $\boldsymbol{t}_E$ と，参照データ R の文書ベクトル $\boldsymbol{t}_R$ を求めて文書間類似度を下記のように定義する．

$$sim\_BERT(R,E) = \frac{\boldsymbol{t}_R \cdot \boldsymbol{t}_E}{||\boldsymbol{t}_R|| ||\boldsymbol{t}_E||} \tag{21}$$

3.5 節と同様に，式 (3) の sim および式 (13) と式 (14) の sim_LSI を sim_BERT に置きかえて答案 E の点数 $grade(E)$ を求める．

4. 実験

参照文書を利用して提案した各手法がどの程度人手による採点結果に近いかを明らかにする．参照文書として講義の書き起こし文書である参考文書を利用した場合と，模範答案を利用した場合について実験する[*17]．まず，実験で利用する模範答案，評価の枠組および実験設定について記述する．

[*15] https://github.com/cl-tohoku/bert-japanese

[*16] 転移学習およびファインチューニングを利用する研究例では異なるタスクの文書の良さに関する学習データを利用して精度向上を試みるなど目的とするタスク以外の学習データで採点精度を向上させる例が試されている（5.4 節参照）．こうした研究は今後の課題としたい．

[*17] 参考として付録 A.3 節に先行研究の Jess [49] による評価の結果，および A.4 節に模範答案を利用した識別モデルによる評価の結果を示す．

4.1 模範答案の抽出

模範答案は採点された小論文答案の中から取り出す．評価軸として理解力を優先して合計の最終評点が高い答案を1つ選択する．答案 d の理解力，論理性，妥当性，文書力の最終評点を c_d，l_d，v_d，g_d とする．下記の式で最もスコア $score_d$ が高い答案集合 E' を求め，E' 中の1件を模範答案として選択する．理解力以外の評価軸も考慮する理由は模範答案は他の答案と比較に利用するために論理性，妥当性，文書力も良い完成度の高い答案であることが望ましいと考えられるためである．

$$score_d = 2.0c_d + 1.5l_d + 1.5v_d + 1.0g_d \tag{22}$$

$$E' = \arg\max_{d \in \mathscr{E}}(score_d) \tag{23}$$

各評価軸に対する式 (22) の重み付けは理解力の次に文書の良さを評価する論理性と妥当性を重視し，誤字脱字など文の外形的な評価である文書力は理解力との関係が薄いため重みを低く設定する．

選ばれた模範答案は評価データから除外する．よって表 2 から 1 件引いた答案数で参考文書および模範答案を利用した場合の各手法の評価を行う．

4.2 評価の枠組

人手で付与した最終評点を基に各手法による評価スコアを QWK を利用して評価する [13]．QWK は-1 から 1 を取る値で 1.0 が最も一致していることを表す．QWK は偶然による一致を排除し，さらに評点とスコアの差に応じて重みを掛けて値を割り引く計算を行う．

小論文答案に対して付与された最終評点と手法が出力した点数を 1 から 5 点の件数で集約した分割表を作成する．最終評点を $i = \{1, \ldots, 5\}$，手法の点数を $j = \{1, \ldots, 5\}$ とし，各点数に該当する答案の数を $num_{i,j}$ とする．このとき，一致率 $O_{i,j}$ と偶然の一致率 $E_{i,j}$ を以下のように計算する．

$$O_{i,j} = \frac{num_{i,j}}{\sum_{i,j} num_{i,j}} \tag{24}$$

$$E_{i,j} = \frac{(\sum_i num_{i,j})(\sum_j num_{i,j})}{(\sum_{i,j} num_{i,j})^2} \tag{25}$$

評点 i, j の差に対する重み $W_{i,j}$ を下記のように計算する．

$$W_{i,j} = \frac{(i-j)^2}{(R-1)^2} \tag{26}$$

ここで R は最大評点の 5 である．これにより QWK を下記のように計算する．

$$QWK = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \tag{27}$$

ここで，QWK で評価を行う場合，各手法の評価スコア $grade(E)$ を下記のように 1 から 5 点に整数化する．

$$score(E) = \begin{cases} 1\,(0.0 \leq grade(E) \leq 1.0) \\ 2\,(1.0 < grade(E) \leq 2.0) \\ 3\,(2.0 < grade(E) \leq 3.0) \\ 4\,(3.0 < grade(E) \leq 4.0) \\ 5\,(4.0 < grade(E) \leq 5.0) \end{cases} \tag{28}$$

各手法の $score(E)$ を用いて QWK を計算する．

4.3 実験設定

3.3 節の LSI，および 3.4 節の LDA におけるトピック数 K は 200 に設定した．LDA の $\boldsymbol{\alpha}$ と $\boldsymbol{\eta}$ の初期値はそれぞれ等確率とした．LSI と LDA で利用する形態素解析器は 3.1 節と同様に NEologd 辞書を用いる．

3.5 節の分散表現ベクトルを利用した手法では形態素の分散表現ベクトルとして国立国語研究所が作成した nwjc2vec [2], [53]（300 次元の skip-gram）を利用する．分散表現ベクトル作成に利用する形態素解析器の辞書として UniDic 辞書[*18]を利用する．

3.6 節で利用する BERT は隠れ層の内部状態は 768 ユニット，12 層で 12 ヘッドのモデルを利用する．

4.4 参考文書を利用した場合の実験結果と評価

参照文書として参考文書を利用した場合について各手法に対する QWK を**表 5** に示す．ここで，各手法の記号として 3.1 節の式 (1) の類似度を利用する手法を wd，式 (2) の手法を wdm，3.2 節の手法を加えた場合は idf を付加して記述し，3.3 節から 3.6 節の各手法はそれぞれ，LSI，LDA，wdvec，BERT と記述する．

表 5 の結果から，最も QWK の平均値が高い手法は参考文書と答案に共通して出現した形態素の最小頻度を利用した wdm であることが分かる．先行研究 [49] で利用された LSI やより洗練された LDA，分散表現ベクトルを利用した手法は wdm と比較して平均 QWK が高くなかった．対応のある t 検定を行うと，LDA に対して wdm は $p=0.004 < 0.05$（両側検定）[*19]を示し，有意水準 5%で差があることが分かる．また他の共通形態素を利用した手法もすべて有意水準 5%で差があり，共通形態素の頻度を利用した手法が LDA に対して効果があることが分かる[*20]．

次に，各課題について比較すると，課題によって QWK が大きく異なることが分かる．たとえば，s_1，s_2 および c_1 は形態素の頻度を利用した場合，QWK が他の課題と比較して高い値が得られている．これらの課題に対する正解が講義内で述べられているため参考文書内に正解に近い記述が存在することが理由である．しかしながら，s_1，s_2，c_1 に対してベクトルによる文書間類似度を用いた手法（すなわ

[*18] https://unidic.ninjal.ac.jp/
[*19] 以下すべて両側検定の値を示す．
[*20] 表 5 では QWK の平均値について LDA に対して有意水準 5%で差があった手法について*印を付与している．

表 5 参考文書を利用した手法の比較（QWK）
Table 5 QWK scores of the proposed methods using a reference text for each prompt.

手法	課題												平均
	g_1	g_2	g_3	s_1	s_2	s_3	e_1	e_2	e_3	c_1	c_2	c_3	
wd	0.027	0.134	0.279	0.638	0.429	0.247	−0.053	0.161	0.099	0.578	0.106	**0.368**	0.251*
wdidf	0.034	**0.200**	0.254	0.553	0.517	0.173	−0.092	0.264	**0.165**	0.627	**0.158**	0.120	0.248*
wdm	0.047	0.135	0.267	**0.686**	0.438	0.247	0.015	0.185	0.111	0.623	0.004	0.346	**0.259***
wdmidf	**0.064**	0.169	**0.300**	0.544	**0.530**	0.143	−0.128	**0.276**	0.134	**0.629**	0.072	0.123	0.238*
LSI	−0.093	0.060	0.063	0.237	0.063	0.194	−0.029	0.006	0.046	0.047	0.017	0.002	0.051
LDA	0.000	0.026	0.077	0.292	0.122	**0.261**	−0.200	0.066	0.035	0.029	0.046	−0.016	0.062
wdvec	−0.008	0.035	0.076	−0.028	0.074	0.050	**0.050**	0.055	0.028	−0.099	0.015	0.142	0.033
BERT	−0.042	0.014	0.068	−0.095	0.084	−0.042	−0.262	0.046	0.100	−0.108	0.006	−0.008	−0.020

表 6 模範答案を利用した手法の比較（QWK）
Table 6 QWK scores of the proposed methods using a reference essay with the highest score for each prompt.

手法	課題												平均
	g_1	g_2	g_3	s_1	s_2	s_3	e_1	e_2	e_3	c_1	c_2	c_3	
wd	0.180	0.167	0.220	0.686	0.411	0.218	0.131	0.402	0.165	0.835	0.233	**0.336**	0.332*
wdidf	**0.248**	0.209	0.285	0.620	**0.434**	**0.253**	0.206	0.425	0.161	0.835	0.371	0.227	0.356*
wdm	0.104	0.146	0.249	0.673	0.336	0.244	0.144	0.421	0.177	**0.850**	0.233	0.269	0.321*
wdmidf	0.228	**0.215**	**0.294**	**0.688**	0.431	**0.253**	0.318	**0.547**	**0.256**	0.833	**0.406**	0.254	**0.394***
LSI	0.137	0.058	0.085	0.508	0.080	0.168	0.192	0.050	0.026	0.335	0.300	−0.001	0.162
LDA	0.002	0.061	0.103	0.176	0.047	0.178	**0.324**	0.082	0.040	0.374	0.154	−0.028	0.126
wdvec	0.005	0.062	0.096	0.683	0.156	0.050	0.059	0.067	0.034	0.568	0.014	0.075	0.156
BERT	0.075	0.027	0.230	0.256	0.118	0.084	0.132	0.153	0.235	0.124	0.093	0.096	0.135

ち LSI，LDA，wdvec および BERT）では QWK が低く，参考文書の部分的な表現の類似性をとらえることができていない．これは参考文書が約 2,600 文字以上と答案に対して大きく，課題と無関係な部分が文書間の類似度計算に影響を与えたためと考えられる．

上記以外の課題（講義 g と e の全課題および s_3，c_2，c_3）における QWK はどの手法も低く有効に働いていないことが分かる．課題 g_2，g_3，e_1，c_3 は具体的な事例を解答者に求めるため参考文書に記載がない．また，課題 g_1，s_3，e_2，e_3，c_2 は講義の中で説明があるため参考文書内に関連する記述があるが，文書の抜き出し表現では正解にならず表現をいい換えた文書を作成する必要があるため参考文書が有効に働かないことが原因と考えられる．

一方で，課題 s_1 など記述すべき内容が参考文書内にある場合は，wdm で QWK が 0.6 を超える値が得られており，単純な手法であるにもかかわらず比較的高い値[*21]が得られている．このことから，課題を適切に設定し参考文書を用意することができれば比較的高い QWK を得られることが分かる．

4.5 模範答案を利用した場合の実験結果と評価

模範答案を参照文書として利用した場合の結果を**表 6** に示す．すべての手法で平均 QWK が向上した．最も高い平均 QWK を示した手法は形態素の最小頻度と idf 値を利用した wdmidf である．一方で，ベクトルを利用した手法の中でもっとも平均 QWK が高い手法は LSI であるが，形態素の頻度を利用した各手法の方が LSI よりも平均 QWK が高い値を示している．対応のある t 検定で評価すると，LSI に対して wdmidf では p=0.0002，wdm では p=0.009 を示しどちらも有意水準 5%で差があることが分かる[*22]．

各課題について比較すると課題 s_1 と c_1 では形態素を利用した手法がベクトルを利用した手法に比べて高い QWK を示した．課題 c_1 に対して wdm を適用した場合に QWK は 0.850 を示し，参考文書を利用した場合（表 5）と比較して約 0.2 ポイント改善し高い値となった．課題 s_1 に対して wdmidf を適用した場合に QWK が 0.688 を示し，参考文書を利用した場合と比較して向上したが参考文書を利用した場合の wdm で QWK が 0.686 とすでに高く，この値と比較すると QWK は大きく変わっていない．また wdm は模範答案を利用した場合に QWK が低下した．これは模範答案ではとらえられていない正解とすべき表現があり，それを含んだ答案に対して模範答案が有効に働かなかったと考えられる．一方で課題 c_1 では正解とすべき表現が模範答案でほとんど網羅されていたと考えられる．

[*21] 脚注の*7 の評価では good となる．また脚注*32 も参照．

[*22] 表 6 では QWK の平均値について LSI に対して有意水準 5%で差があった手法について*印を付与している．

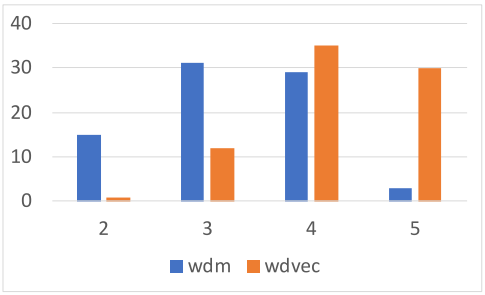


図 1 課題 c_1 に対する最終評点が 3 点の答案に対する wdm と wdvec の評点分布

Fig. 1 Frequency distributions of output scores generated by wdm and wdvec for essays with a final score of three in the c_1 prompt.

課題 s_1 と c_1 においてベクトルを利用した手法間で比較すると wdvec が他のベクトルを利用した手法より高い QWK を示した．

特に課題 s_1 では課題 c_1 に比べて wdmidf に近い QWK を示し，wdvec が有効に働いていることが分かる．具体的に例をあげて説明する．下記は課題 s_1 の模範答案（模範 1）と最終評点が 5 点でかつ wdmidf が 4 点と評価し wdvec が 5 点と正しく評価した小論文（答案 1）である．

模範 1 仮説が観察・実験などによって検討できる実証性、同一の条件下では同一の結果が得られる再現性、多数の人々によって承認され、公認される客観性の 3 つの条件をみたす必要がある。

答案 1 科学的であることの条件としては、仮説が検討可能であるという実証性、同一条件下で同一の結果が得られる再現性、広く承認され公認される客観性を満たす必要がある。

「仮説」や「実証性」など形態素の一致が多く見られる一方で，「みたす」と「満たす」という表記の違いのほかに，「多数の人々によって」と「広く」など表現の異なりがあるため，この部分での差異を wdvec が吸収して高い QWK を得たと考えられる．このように wdmidf が 4 点と推定した答案に対して wdvec が正しく 5 点を付与した例は 44 例観測された．

一方で課題 c_1 では wdvec は wdm に比べて低い QWK を示している．ここで**図 1** に最終評点が 3 点の答案（78 件）に対する場合の wdm と wdvec の評点分布を示す．

図 1 から wdm は 3 点と正しく判定した答案が多いが wdvec では多くの答案を 5 点と誤って評価していることが分かる．このことから wdvec の特徴である表現の異なりを吸収する働きによって模範答案とは類似していない答案に対しても過剰に類似として評価していることから低い QWK を示すことになったと考えられる．

課題 g_1，e_1，e_2，e_3 および c_2 では模範答案を利用することで QWK の向上が見られた．課題 e_2 は用語の意味と解決方法の説明，課題 g_1，e_3，c_2 はある現象に対する説明を求めているがこれらは記述すべき内容が限られているため模範答案が有効に働いたと考えられる．課題 e_1 は例をあげて説明する課題であり LDA が最も高い QWK を示しているが形態素を利用した wdmidf も近い値を示しており，模範答案に含まれる専門的な内容の形態素が有効に働いたと考えられる．

一方，課題 g_2，g_3，s_2，c_3 では模範答案を利用したにもかかわらず QWK が低下もしくはほとんど変わらない値を示した．これは課題 g_2，g_3，c_3 は具体例をあげて説明する課題であるため模範答案に含まれていない内容を評価できないことが原因と考えられる．課題 s_2 は講義内で説明した内容をまとめて記述するため記述内容は定まっているが 400 字以内と文字数が長いため，模範答案の利用だけでは表現の多様性を正しく評価できなかったと考えられる．

次に，形態素の頻度を利用した手法について事例を取り上げて分析する．下記は課題 g_1 の模範答案（模範 2），および最終評点が 5 点でかつ wdidf が高く評価した答案 2（wdidf が 4 点，wd が 3 点）と，低く評価した答案 3 (wdidf が 3 点，wd が 2 点) の一部を示す．

模範 2 **グローバリゼーション**は世界全体の**所得格差**を**縮小**させた。（中略）しかしながら、各国の**所得格差**は**拡大**させた。たとえば日本の**ジニ係数**を見ると、（以下略）

答案 2 **グローバリゼーション**は、全世界的に見れば、**所得格差**を**縮小**させているが、一方では、先進国を中心に、各国内における**所得格差**を**拡大**させている。実際に、**グローバリゼーション**が進展していった 20 世紀後半において、東アジアにおける貧困層は減少しているが、先進国における**所得格差**を表す**ジニ係数**は大きくなっている。（以下略）

答案 3 **グローバリゼーション**は、世界における最貧層を減少させたが、各国に目を向けると、富裕層により富が独占され、格差が**拡大**しているといえる。まず、最貧層の減少は、グローバル化により経済活動が世界規模に**拡大**し、（以下略）

太字は各答案において模範答案と一致した形態素の中で評価に関わる形態素を取り上げて示している．課題 g_1 で高く評価される答案は「グローバリゼーションで国家間の格差は縮小したが国内では格差が拡大した」という内容を「ジニ係数」など根拠を示して論を展開するものである．答案 2 では「グローバリゼーション」，「所得格差」，「ジニ係数」，「縮小」，「拡大」といった形態素が模範答案と同じであり，頻度が高くなることから高く評価される．これは wd も wdidf も同じであるが，wdidf では「グローバリゼーション」，「所得格差」，「ジニ係数」は専門分野で現れるため idf 値が高く[*23]，これらの形態素が高く評価されること

[*23] 「グローバリゼーション」の idf 値は 8.16，「所得格差」は 9.15，「ジニ係数」は 10.20，「格差」は 6.75，「拡大」は 3.77，「縮小」は 4.92 であった．

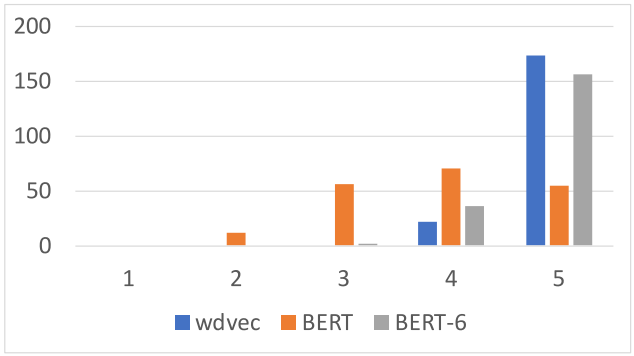


図 2 課題 s_1 に対する最終評点が 5 点の答案に対する wdvec, BERT, BERT-6 の評点分布

Fig. 2 Frequency distributions of output scores generated by wdvec, BERT and BERT-6 for essays with a final score of five in the s_1 prompt.

で，全体的に高い QWK が得られたと考えられる．

一方，答案 3 では「グローバリゼーション」は模範答案と一致するが，「所得格差」は答案 3 では「格差」と表現されるため模範答案と一致しない．よって，wd および wdidf とも答案 3 を低く評価している．これは形態素解析器の辞書として NEologd を利用したため専門用語などが取り出せる一方で，短い表現（この場合は「格差」）は形態素が異なるため文字列としては模範解答に含まれていてもまったく評価されないことが原因である．

分散表現ベクトルを利用した wdvec は先に示した課題 s_1 と c_1 以外の課題に対しては他のモデルと比べて低い QWK を示した．BERT は記述の自由度が高い一部の課題（g_3 と e_3）で他のモデルと比較して若干高い QWK を示したが全体として QWK が低く，課題 s_1 や c_1 において wdvec と比較して低い QWK を示した．そこで課題 s_1 を対象に wdvec と BERT の評価結果の異なりに着目して分析する．**図 2** は課題 s_1 の中で最終評点が 5 点の答案（195 件）に対する wdvec と BERT の点数の頻度分布を表している[*24]．横軸は各手法の評点で，縦軸が答案数である．wdvec はほとんど 5 点と正しく分類している一方で，BERT は 3 から 5 点の間で評価した答案数が大きく変わらない．つまり 3 点や 4 点と評価した答案と模範答案は類似していないと分類されていることになる．具体的な例で考察する．下記の答案 4 および答案 5 は最終評点が 5 点で，それぞれ BERT が 5 点および 3 点と評価した答案である[*25]．

答案 4 仮説が観察実験などによって検討できる「実証性」、同一の条件下では同一の結果が得られる「再現性」、そして、多数の人々（科学コミュニティ）によって承認され公認される「客観性」の 3 条件をみたす必要がある。

答案 5 仮説が実際に検証できる実証性と同一条件のもとで同一結果が得られる再現性と多数の人に承認・公認される客観性の 3 つの条件。

[*24] BERT-6 については 4.6 節で説明する．
[*25] wdvec はどちらも 5 点と評価している．

模範 1 と比較すると答案 4 では「などによって検討できる」や「によって承認され」という部分文字列が含まれており，模範答案と一致する一方で，答案 5 では内容はほとんど同じであるがこれらの表現が省略されている．よって BERT では「によって」など機能語を含めた表現も文の類似度に考慮されたため答案 5 の方が文の類似度が低いと判定されて QWK が低い値になったと考えられる．これは BERT がトークン列を考慮した文の類似度を計算する特性から生じる結果である．よって，トークン前後の依存情報を軽減した文ベクトルを構築できれば精度が向上する可能性がある．そこで 4.6 節で BERT で利用する層を変更したモデルについて検討する．

4.6 BERT で利用する層の異なりに関する考察

近年，BERT のような Transformer を利用したモデルの各層に対する分析から，最終層以外の層の利用が先行研究において提案されている [32], [35]．本研究では 12 層の BERT を利用しているが層を重ねるごとに Attention による文脈情報が加算されるためより入力に近い層を利用することでトークン前後の依存情報を軽減したモデル化が可能であると考えられる．そこで 4.5 節で利用した BERT の最終層から n 層削除した BERT-n モデルを作成し，模範答案を参照文書として利用した場合の QWK を**表 7** に示す．

表 7 において課題全体に対する平均 QWK を比較すると BERT-2 が高い QWK を示しているが BERT と比較して有意な差は認められなかった（有意水準 5%）．また，各課題に対して最も高い QWK を示した手法を比較すると半分の課題は BERT が最も高い QWK を示した．課題 s_1 と c_1 は BERT の出力層を削除したモデルの方が BERT よりも高い QWK を示している．課題 s_1 に対しては BERT-6, c_1 に対しては BERT-10 が最も高い QWK を示した．これは浅い層を利用することでトークン前後の依存情報が減少し，機能語などの表現の異なりを吸収する文ベクトルが作成されたためと考えられる．

課題 s_1 に対する最終評点が 5 点の答案に対する BERT-6 の頻度分布を図 2 に示す．頻度分布の比較から BERT-6 は BERT と異なり多くの答案を 5 点と正しく評価することが分かる．また 4.5 節の答案 5 に対して BERT では 3 点であったが BERT-6 では 4 点と評価し改善が見られた．しかし課題 s_1 および c_1 について wdvec と比較すると層を削除した BERT よりも依然として wdvec が高い QWK を示している．この原因として 2 つの要因が考えられる．1 つは 4.5 節ですでに述べたように課題の特性が考えられる．課題 s_1 および c_1 は 100 字程度で正解とすべき記述内容の表現が限られており模範答案で出現した形態素およびそのいい換えで文の類似度がとらえられるため，形態素の並びに関する情報を使わなくても評価できる課題である．

もう 1 つの要因として学習データ量の異なりが考えられ

表 7 BERT で利用する層の異なりによる QWK の違い（模範答案を利用した場合）
Table 7 QWK scores of BERT-based models utilizing different layers using a reference essay with the highest score for each prompt.

手法	課題												平均
	g1	g2	g3	s1	s2	s3	e1	e2	e3	c1	c2	c3	
BERT	**0.075**	0.027	**0.230**	0.256	**0.118**	0.084	0.132	0.153	**0.235**	0.124	**0.093**	0.096	0.135
BERT-1	0.051	0.041	0.127	0.366	0.052	0.135	0.117	**0.159**	0.181	0.116	0.029	**0.147**	0.127
BERT-2	0.025	**0.075**	0.132	0.440	0.076	**0.137**	0.114	0.107	0.124	0.362	0.022	0.109	**0.144**
BERT-4	0.029	0.034	0.085	0.553	0.063	0.130	0.088	0.076	0.064	0.412	0.015	0.089	0.137
BERT-6	0.036	0.022	0.066	**0.635**	0.071	0.119	0.103	0.039	0.051	0.459	0.015	0.060	0.140
BERT-8	0.045	0.020	0.044	0.541	0.040	0.119	0.040	0.041	0.007	0.563	0.014	0.067	0.128
BERT-10	0.037	0.028	0.045	0.535	0.041	0.084	0.092	0.041	0.015	**0.580**	0.014	0.134	0.137

る．利用した BERT[*12] では Wikipedia の約 1,800 万文で学習しているのに対して wdvec で利用している分散表現ベクトル（nwjc2vec）は約 14 億文 [53] で学習しており，大きく異なっている．wdvec と比較して利用した BERT は事前学習の量が少ないためトークン列の並びに対して同じとすべき内容を異なる表現として処理した可能性が考えられる．事前学習を増やした BERT を利用すると上記の結果が改善される可能性がある．

5. 関連研究

多様な視点から多くの先行研究が行われている．それぞれの特徴について以下に分類して説明する．

5.1 実用システム

英語圏では E-rater や Intelligent Essay Assessor，IntelliMetric など実用システムが開発されている．それぞれのシステムでは重回帰分析，LSI，ルール発見などの手法が適用されているが内部パラメータを調整するために人手による採点済み答案を必要とする [50]．

5.2 採点済み小論文答案データ

英語圏では 2012 年に Automated Student Assessment Prize（ASAP）[*26]が作成された．ASAP は 8 個の小論文課題に対して学生（7 年生から 10 年生[*27]）が記述した答案とその採点結果である．答案数が各課題で約 700 件から 1,800 件程度あり学習およびテストデータとして利用されている．

参考文章を読んで問いに答える形式で学生（4 年生から 8 年生[*28]）が記述した小論文答案データ（2,970 件と 2,076 件）が作成されている [9]．このデータは Response to Text Assessment（RTA）データとして文献 [29], [30], [37], [39] で利用されている．ほかに非母語話者が記述した TOEFL の小論文データ 12,100 件が Linguistic Data Consortium（LDC）から配付されており [5]，文献 [26] で使用されている．

一方で母語話者が記述した日本語の小論文で研究利用可能な採点済み答案データは著者が知る限り見当たらなかった．よって本論文では研究利用可能な形で小論文課題と答案および採点結果を作成し評価実験に用いた．また，近年では日本語学習者が記述した小論文データ GoodWriting[*29]が構築され，文献 [16], [59] で利用されている．

5.3 深層学習の利用

深層学習モデルに採点済み答案を学習させることで小論文を採点するモデルが多数提案されている．Alikaniotis ら [1] は Bi-LSTM モデル，Taghipour と Ng [33] は CNN（Convolutional Neural Network）と LSTM を利用したモデル，Dong ら [13] は CNN と LSTM に Attention を取り入れたモデル（CNN+LSTM+ATT），Tay ら [36] は SkipFlow という機構を取り入れた LSTM モデルを提案した．また，Zhang と Litman [38] は参考文書と答案との相互 Attention を利用する CNN+LSTM モデルを提案し，Dong らのモデルよりも高い QWK を示した．

Mayfield と Black [23] は Bag-of-Words，分散表現ベクトル，BERT および DistilBERT の転移学習およびファインチューニングを用いたモデルを提案した．複数の課題に対する評価実験では BERT を利用したモデルよりも下記に述べる Cozma らのモデル [21] の方が若干高い QWK を示す課題が多いことを明らかにした．このことから約 1,100 件の採点済み答案を利用して BERT の転移学習およびファインチューニングを適用しても既存手法を大きく上回ることは容易ではないことが分かる[*30]．

日本語では平尾ら [16], [59] が日本語学習者が記述した GoodWriting の答案データに対して BERT のファインチューニングを適用し，Random Forest や LSTM より高い QWK を示すことを明らかにした．このように BERT が日本語の小論文採点に有効に働く結果が示されている一

[*26] https://www.kaggle.com/c/asap-aes/data
[*27] 日本の中学 1 年から高校 1 年に相当する．
[*28] 日本の小学 4 年から中学 2 年に相当する．
[*29] GoodWriting. https://goodwriting.jp/wp/system-ml/
[*30] テストの答案は 360 件である [23]．

方で，上述のとおり母語話者が記述した英語の小論文課題では BERT の転移学習およびファインチューニングを利用したモデルより優れたモデルが示されている [23].

5.4 他の文書に関する評価結果の利用

他の文書の良さに関するデータを学習に利用し，その後目的とする採点済み答案を学習して精度を向上する手法が提案されている．Jin ら [19] は他の小論文課題で採点済みの答案から抽象的な特徴量を利用して答案の良し悪しを識別する RankSVM を作成し，目的とする課題の答案の中で評価の高い答案と低い答案のみを取り出してから深層学習モデルで学習する手法を提案した．この手法により採点済み答案のみを学習する CNN+LSTM+ATT よりも高い QWK が得られることを示した．

また Nadeem ら [26] はマルチタスク学習による採点手法を提案した．具体的には Stanford 大学の Natural Language Inference データ [7] および談話情報付与データ [40] を利用して 2 文の接続タイプに関する識別問題を LSTM+ATT モデルで事前に学習し，最後に目的とする課題の答案を学習手法を提案した．しかしながら小論文課題に対しては Liu ら [20] の手法がわずかに高い QWK を示した．Liu らは採点済み答案以外に点数付けされた文書を LSTM モデルに学習させ，これらの出力を最終的に XGboost [8] の特徴量として利用する．文献 [20] では提案手法が CNN+LSTM より多くの課題（5/8）で高い QWK が得られることを示した．

日本語では水本ら [24], [55] が短答式問題（70 文字以下）に対して採点の根拠となる情報を付与したデータを作成し，採点済み答案と同時に学習することで採点精度の向上と採点の根拠を示すモデルを構築している．これらは短答式課題に対する結果であるが上記のように英語の小論文に対しても効果があることから，学習モデルを利用した日本語小論文採点手法を研究する場合には考慮すべき方法と考えられる．

目的とする課題の採点済み答案を低減させる手法

目的とする課題の採点済み答案をなるべく少なくする手法が提案されている．Phandi ら [28] は分野適用を利用して，他の小論文課題の採点済みの答案で学習した採点モデルを目的とする課題の答案評価に適用する実験を行った．採点モデルとして Bayesian Linear Ridge Regression を利用し，ASAP の小論文課題について異なる課題の採点済み答案に加えて目的とする課題の採点済み答案 10 件のみを利用して 0.484 から 0.649 の QWK を得た．同様の課題に対して Dong と Zhang [12] は CNN を利用したモデルを提案し，目的とする課題の採点済み答案 10 件で学習したモデルが 0.546 から 0.647 の QWK を得た．

Cozma ら [21] は string kernel と bag of super word embedding を特徴量として利用し，ν-Support Vector Regression（ν-SVR）により採点するモデルを提案した．先行研究と同じ小論文課題に対して目的とする課題の採点済み答案 10 件を利用した場合において 0.586 から 0.734 の QWK を示し，上記の Phandi らおよび Dong と Zhang のモデルを超える結果を示した．興味深いことに Cozma らは目的とする課題の採点済み答案を利用しない場合の QWK を示しており 0.542 から 0.728 の値を得ている．

上記の小論文課題はすべて ASAP の 8 つの課題内での実験結果であり，課題間の類似性が採点精度に影響を与えることが考えられる．しかしながら採点済み答案を用いない手法であり本研究の方針とも一致する手法である．本論文では学習モデルを利用していない手法を選択したが今後の課題として，学習モデルを利用する場合に検討する必要がある手法と考えられる．

5.5 ルーブリックを基にした表現の利用

文書からではなく評価の観点から人が採点する際に基準とするルーブリックを基に各課題に対して評価すべき表現を設定する手法が提案されている．Rahimi ら [29], [30] は課題に対して記述すべき表現リストを人手で構築し，それらを基にした特徴量を Random Forest に取り入れて採点済み答案で学習する方法を提案した．Zhan と Litman [37] は Rahimi らの手法に skip-gram の分散表現ベクトルを導入して QWK を向上させた．さらに文献 [39] では Attention を利用して上記の人手で作成していた課題ごとの表現リストを自動で取り出す手法を提案し，人手で構築した場合よりも高い QWK を示す結果を示した．

日本語の小論文に対して Ishioka ら [34] は Rahimi らの手法と同様に課題で評価すべき表現を人手で設定し，それらを特徴量として Random Forest により識別する手法を提案している．国立情報学研究所が主催する NTCIR-13 の QALab-3 タスク[*31]における東京大学の小論文課題（450 から 600 字以内）に適用し，人手による評価と比較して良好な結果であることが述べられている．

本論文では評価すべき表現集合を収集していないが，評価すべき表現を参考文書や模範答案で代替して与えている手法ととらえることができる．つまり本論文での形態素の頻度を基にした手法は評価すべき表現の頻度を基に採点した結果と考えられる．簡素な方法にもかかわらず，課題によっては 0.6 以上の QWK が得られていることから[*32]表現集合を利用した採点手法は本課題でも有効であると考えられる．

5.6 採点済み答案を利用しない手法

採点済み答案を利用しない小論文評価手法として日本語

[*31] http://research.nii.ac.jp/qalab/task.html
[*32] 文献 [23] では試験目的では QWK は 0.6 から 0.8 が最低求められる値であると記述されている．

では Jess が開発されている [18], [51]．Jess では文書の良さに関する特徴量を利用するともに参考文書と答案との内容の類似度評価法として LSI を利用した．本論文では答案の内容に対する評価軸である理解力に対して参考文書と答案との類似度評価による採点手法を提案しているため Jess の手法と対応していると考えられる．本論文では 4.4 節に示すように本研究対象の小論文課題では LSI よりも形態素の頻度を利用した手法が高い QWK を示すことを明らかにした．

6. まとめ

本論文ではまず模擬試験を実施し，解答者に了解を得て研究利用可能な小論文データを構築した．小論文課題は講義を聞いて課題に答える形式を採用し，合計で 12 課題（4 講義で各 3 課題），各課題で約 300 件前後の答案を収集し人手による採点を付与した．

採点手法として参照文書と答案の文書類似度で評価する手法を仮定し，文書類似度として形態素の頻度，Wikipedia を利用した idf 値，LSI，LDA，分散表現ベクトル，BERT を用いた文書ベクトルを利用する手法を提案した．上記の構築した採点済み小論文答案を利用して評価実験を行った結果，模範答案を参照文書として使用した場合における形態素頻度と idf 値を適用した手法が他の手法に比べて高い平均 QWK を示し，有効な手法であることを明らかにした．

本論文では試験問題など新規の小論文課題を想定して，採点済み答案を必要とする機械学習を使わない手法を提案した．しかしながら近年，目的とする採点済み答案は利用せずに，他の採点済み答案を機械学習モデルに適用することで高い QWK を示す手法が提案されている．また理研 AIP から短答式（80 文字以内）の記述問題採点データセットが国立情報学研究所情報学研究データリポジトリ（IDR）から配付されている[*33]．今後こうした採点済み答案を利用した手法を検討することが課題である．

謝辞 本研究の遂行にあたって岡山大学運営費交付金機能強化経費「小論文、エッセイ等による入学試験での学力の三要素を評価するための採点評価支援システムの開発導入」の助成を受けた．また，BERT の利用に当たって竹内研究室の江島知優さんにご協力をいただいた．Jess の Web サイトを利用した．査読者の皆様には大変有益なコメントをいただいた．ここに記して感謝申し上げる．

参考文献

[1] Alikaniotis, D., Yannakoudakis, H. and Rei, M.: Automatic Text Scoring Using Neural Networks, *Proc. 54th Annual Meeting of the Association for Computational Linguistics*, pp.715–725 (2016).

[2] Asahara, M.: NWJC2Vec: Word Embedding Dataset from 'NINJAL Web Japanese Corpus', *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, Vol.24, No.1, pp.7–22 (2018).

[3] Attali, Y. and Burstein, J.: Automated Essay Scoring with e-rater V.2, *The Journal of Technology, Learning, and Assessment*, Vol.4, No.3, pp.1–30 (2006).

[4] Bakeman, R. and Gottman, J.M.: *Observing Interaction*, Cambridge (1986).

[5] Blanchard, D., Tetreault, J., Higgins, D., Cahill, A. and Chodorow, M.: TOEFL11: A Corpus of Non-Native English, *ETS Research Report Series*, pp.13–22 (2013).

[6] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).

[7] Bowman, S.R., Angeli, G., Potts, C. and Manning, C.D.: A Large Annotated Corpus for Learning Natural Language Inference, *Proc. 2015 Conference on Empirical Methods in Natural Language Processing*, pp.632–642 (2012).

[8] Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785–794 (2016).

[9] Correnti, R., Matsumura, L.C., Hamilton, L. and Wang, E.: Assessing Students' Skills at Writing Analytically in Response to Text, *The Elementary School Journal*, Vol.114, pp.142–177 (2013).

[10] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R.: Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, Vol.41, No.7, pp.391–407 (1990).

[11] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018).

[12] Dong, F. and Zhang, Y.: Automatic Features for Essay Scoring – An Empirical Study, *Proc. 2016 Conference on Empirical Methods in Natural Language Processing*, pp.1072–1077 (2016).

[13] Dong, F., Zhang, Y. and Yang, J.: Attention Based Recurrent Convolutional Neural Network for Automatic Essay Scoring, *Proc. 21st Conference on Computational Natural Language Learning*, pp.153–162 (2017).

[14] Elliot, S.: IntelliMetric: From here to validity, *Automated essay scoring: A cross-disciplinary perspective*, Shermis, M.D. and Burstein, J. (Eds.), pp.71–86, Lawrence Erlbaum Associates (2003).

[15] Hearst, M.A.: The Debate on Automated Essay Grading, *IEEE Intelligent Systems and their Applications*, Vol.15, No.5, pp.22–37 (2000).

[16] Hirao, R., Arai, M., Shimanaka, H., Katsumata, S. and Komachi, M.: Automated Essay Scoring System for Nonnative Japanese Learners, *Proc. 12th Conference on Language Resources and Evaluation*, pp.1250–1257 (2020).

[17] Hoffman, M.D., Blei, F.M. and Bach, F.: Online Learning for Latent Dirichlet Allocation, *Proc. 23rd International Conference on Neural Information*, pp.856–864 (2010).

[18] Ishioka, T. and Kameda, M.: Automated Japanese Essay Scoring System based on Articles Written by Experts, *Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL* (2006).

[19] Jin, C., He, B., Hui, K. and Sun, L.: TDNN A Two-stage Deep Neural Network for Prompt-Independent Auto-

[*33] https://www.nii.ac.jp/dsc/idr/rdata/RIKEN-SAA/

mated Essay Scoring, *Proc. 56th Annual Meeting of the Association for Computational Linguistics*, pp.1088–1097 (2018).
[20] Liu, J., Xu, Y. and Zhu, Y.: Automated Essay Scoring based on Two-Stage Learning, *CoRR*, arXiv:1901.07744 (2019).
[21] Cozma, M., Butnaru, A.M. and Ionescu, R.T.: Automated Essay Scoring with String Kernels and Word Embeddings, *R.E. Asher* (*Editor-in-Chief*), *The Encyclopedia of Language and Linguistics, Vol.6, Oxford: Pergamon Press*, pp.3168–3171 (1994).
[22] Manning, C.D., Raghavan, P., Schütze, H., 岩野和生，黒川利明，濱田誠司，村上明子（訳）：情報検索の基礎，共立出版 (2012).
[23] Mayfield, E. and Black, A.W.: Should You Fine-Tune BERT for Automated Essay Scoring?, *Proc. 15th Workshop on Innovative Use of NLP for Building Educational Applications*, pp.151–162 (2020).
[24] Mizumoto, T., Ouchi, H., Isobe, Y., Reisert, P., Nagata, R., Sekine, S. and Inui, K.: Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring, *Proc. Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp.316–325 (2019).
[25] Mohler, M. and Mihalcea, R.: Text-to-text Semantic Similarity for Automatic Short Answer Grading, *Proc. 12th Conference of the European Chapter of the ACL*, pp.567–575 (2009).
[26] Nadeem, F., Nguyen, H., Liu, Y. and Ostendorf, M.: Automated Essay Scoring with Discourse-Aware Neural Models, *Proc. 14th Workshop on Innovative Use of NLP for Building Educational Applications*, pp.484–493 (2019).
[27] Ohno, M., Takeuchi, K., Motojin, K., Taguchi, M., Inada, Y., Iizuka, M., Abo, T. and Ueda, H.: Construction of Open Basic Data for Automatic Scoring of Essay and Evaluation of Automatic Scoring Method at Current Stage, *Proc. 15th International Conference of the Pacific Association for Computational Linguistics*, pp.215–220 (2017).
[28] Phandi, P., Chai, K.M.A. and Ng, H.T.: Flexible Domain Adaptation for Automated Essay Scoring Using Correlated Linear Regression, *Proc. 2015 Conference on Emperical Methods in Natural Language Processing*, pp.143–177 (2015).
[29] Rahimi, Z., Litman, D., Correnti, R., Wang, E. and Matsumura, L.C.: Assessing Students' Use of Evidence and Organization in Response-to-Text Writing: Using Natural Language Processing for Rubric-Based Automated Scoring, *International Journal of Artificial Intelligent Education*, Vol.27, pp.694–728 (2017).
[30] Rahimi, Z., Litman, D.J., Correnti, R., Matsumura, L.C., Wang, E. and Kisa, Z.: Automatic Scoring of an Analytical Response-To-Text Assessment, *Proc. International Conference on Intelligent Tutoring Systems*, pp.601–610 (2014).
[31] Roy, S., Dandapat, S., Nagesh, A. and Narahari, Y.: Wisdom of Students A Consistent Automatic Short Answer Grading Technique, *Proc. 13th International Conference on Natural Language Processing*, pp.178–187 (2016).
[32] Sajjad, H., Dalvi, F., Durrani, N. and Nakov, P.: Poor Man's BERT: Smaller and Faster Transformer Models, *CoRR*, arXiv:2004.03844 (2020).
[33] Taghipour, K. and Ng, H.T.: A Neural Approach to Automated Essay Scoring, *Proc. 2016 Conference on Empirical Methods in Natural Language Processing*, pp.1882–1891 (2016).
[34] Ishioka, T., Yamaguchi, K. and Mine, T.: Rubric-based Automated Japanese Short-answer Scoring and Support System Applied to QALab-3, *Proc. 13th NTCIR Conference on Evaluation of Information Access Technologies*, pp.152–158 (2017).
[35] Voita, E., Sennrich, R. and Titov, I.: The Bottom-up Evolution of Representations in the Transformer: A Study with Machine Translation and Language Modeling Objectives, *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (2019).
[36] Tay, Y., Phan, M.C., Tuan, L.A. and Hui, S.C.: SkipFlow: Incorporating Neural Coherence Features for End-to-End Automatic Text, *Proc. 32nd AAAI Conference on Artifical Intelligence*, pp.5948–5955 (2018).
[37] Zhang, H. and Litman, D.: Word Embedding for Response-To-Text Assessment of Evidence, *Proc. 55th Annual Meeting of the Association for Computational Linguistics-Student Research Workshop*, pp.75–81 (2017).
[38] Zhang, H. and Litman, D.: Co-Attention Based Neural Network for Source-Dependent Essay Scoring, *Proc. 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pp.399–409 (2018).
[39] Zhang, H. and Litman, D.: Automated Topical Component Extraction Using Neural Network Attention Scores from Source-based Essay Scoring, *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, pp.8569–8584 (2020).
[40] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A. and Fidler, S.: Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books, *Proc. IEEE International Conference on Computer Vision*, pp.19–27 (2015).
[41] 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).
[42] 難波英嗣：テキスト間の類似度の測定，情報の科学と技術，Vol.70, No.7, pp.373–375 (2020).
[43] 旺文社：蛍雪時代 7 月臨時増刊号全国大学推薦・AO 入試合格対策号（2020 年入試対策用）（旺文社蛍雪時代）(2019).
[44] 旺文社：全国大学入試問題正解 2020 年受験用 8 (2019).
[45] 大野雅幸，竹内孔一，泉仁宏太，小畑友也，田口雅弘，稲田佳彦，飯塚誠也，阿保達彦，上田　均：参照データと idf を利用した事前採点不要な小論文評価手法，電子情報通信学会言語理解とコミュニケーション研究会，pp.103–108 (2018).
[46] 株式会社ベネッセコーポレーション：基礎小論文ハンドブック I (2017).
[47] 岩田具治：トピックモデル，講談社 (2015).
[48] 石川　巧：「いい文章」ってなんだ？入試作文・小論文の思想，ちくま新書 (2010).
[49] 石岡恒憲：日本語小論文の自動採点および作文支援システムの開発，科学研究費補助金研究成果報告書 (2007).
[50] 石岡恒憲：コンピュータ上で実施する記述式試験—エッセイタイプ，短答式，マルチメディア利用について，電子情報通信学会誌，Vol.99, No.10, pp.1005–1011 (2016).
[51] 石岡恒憲，亀田雅之：コンピュータによる小論文の自動採点システム Jess の試作，計算機統計学，Vol.16, No.1, pp.3–19 (2007).

[52] 石岡恒憲，亀田雅之，劉 東岳：人工知能を利用した短答式記述採点支援システムの開発，電子情報通信学会技術研究報告．NLC, 言語理解とコミュニケーション，pp.87–92 (2016).
[53] 浅原正幸，岡 照晃：nwjc2vec：『国語研日本語ウェブコーパス』に基づく単語の分散表現データ，言語処理学会第 23 回年次大会，pp.94–97 (2017).
[54] 浅原正幸，加藤 祥：文書間類似度について，自然言語処理，Vol.23, No.5, pp.463–499 (2017).
[55] 水本智也，磯部順子，関 根聡，乾健太郎：採点項目に基づく国語記述式答案の自動採点，言語処理学会第 24 回年次大会発表論文集，pp.552–555 (2018).
[56] 柴山 直，前田忠彦介：複数採点者の小論文評価に関する方法論的検討，pp.119–131, 商事法務 (2007).
[57] 須山敦志：ベイズ推定による機械学習，講談社 (2017).
[58] 佐藤敏紀，橋本泰一，奥村 学：単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討，言語処理学会第 23 回年次大会，pp.875–878 (2017).
[59] 平尾礼央, 新井美桜, 嶋中宏希, 勝又 智, 小町 守：複数項目の採点を行う日本語学習者の作文自動評価システム，言語処理学会第 26 回年次大会発表論文集，pp.1181–1184 (2020).

付 録

A.1 LDA による内部パラメータの計算

本研究では LDA の計算に gensim[*34]のパッケージを利用する．gensim では文献 [17] に従い，変分ベイズ推定により近似した事後分布を仮定して，LDA の内部パラメータを求める．

変分ベイズによるパラメータ推定の学習の詳細は文献 [6], [17], [41], [47], [57] に譲るとして，ここでは文献 [17] に従い，大規模文書からパラメータを獲得する際の数式的な枠組で，本研究に関連する部分を中心に記述する．

LDA では文書 d が K 個のトピックから構成されると考え，トピック分布 $\boldsymbol{\theta}_d = (\theta_{d1}, \cdots, \theta_{dK})^t$ を仮定する[*35]．文書 d が n 個の形態素から成り立っている $(\boldsymbol{w}_d = (w_{d1}, \cdots, w_{dn})^t)$ としたとき，各形態素 w_{di} に対してトピック $z_{di} \in \{1, \cdots, K\}$ が割り当てられるとする．このとき，z_{di} は θ_d に基づいて生成される．

$$z_{di} \sim \boldsymbol{\theta}_d \tag{A.1}$$

選ばれたトピック z_{di} に基づいて各形態素 w_{di} は $\pi_{z_{di}}$ の分布に基づいて生成される．

$$w_{di} \sim \boldsymbol{\pi}_{z_{di}} \tag{A.2}$$

ここで形態素の語彙数を V 個とするとあるトピック k における形態素の分布は $\boldsymbol{\pi}_k = (\pi_{k1}, \cdots, \pi_{kV})^t$ である．

確率分布 $\boldsymbol{\theta}_d$ と $\boldsymbol{\pi}_k$ の事前分布はそれぞれ定数として与えるパラメータ $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)^t$ および $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_V)^t$ のディリクレ分布に従うと仮定する（初期値は 4.3 節参照）．

[*34] https://radimrehurek.com/gensim/models/ldamodel.html
[*35] K は人手で与える．

文書 d の i 番目の形態素のトピックが k であるとき変分ベイズにより因子分解した近似事後分布を下記のように仮定する．

$$q(z_{di} = k) = \phi_{dw_{di}k} \tag{A.3}$$

$$q(\boldsymbol{\theta}_d) = Dir(\boldsymbol{\theta}_d | \boldsymbol{\mu}_d) \tag{A.4}$$

$$q(\boldsymbol{\pi}_k) = Dir(\boldsymbol{\pi}_k | \boldsymbol{\xi}_k) \tag{A.5}$$

ここで $\boldsymbol{\mu}_d$ は K 次元のベクトルであり，各 μ_{dk} は正の実数である．また，$\boldsymbol{\xi}_k$ は V 次元のベクトルであり，各 ξ_{kv} は正の実数である．

変分ベイズを利用したパラメータの更新式は下記のようになる．

$$\phi_{dwk} \propto \exp\{\mathbf{E}_q[\log\theta_{dk}] + \mathbf{E}_q[\log\pi_{kw}]\} \tag{A.6}$$

$$\mu_{dk} = \alpha_k + \sum_{w=1}^{V} c_{dw}\phi_{dwk} \tag{A.7}$$

$$\xi_{kw} = \eta_w + \sum_{d=1}^{D} c_{dw}\phi_{dwk} \tag{A.8}$$

ここで，ϕ_{dwk} は文書 d 内で形態素 w がトピック k で出現した場合の確率分布を表している．また c_{dw} は文書 d 内で形態素 w が出現した回数を表し[*36]，D は全文書の件数を表す．

ここで求まる $\boldsymbol{\mu}_d$ と $\boldsymbol{\xi}_k$ を利用して下記を求める．

$$\mathbf{E}_q[\log\theta_{dk}] = \Psi(\mu_{dk}) - \Psi\left(\sum_{i=1}^{K}\mu_{di}\right) \tag{A.9}$$

$$\mathbf{E}_q[\log\pi_{kw}] = \Psi(\xi_{kw}) - \Psi\left(\sum_{i=1}^{W}\xi_{ki}\right) \tag{A.10}$$

ここで Ψ はディガンマ関数 [41] を表す．これらの式を利用して EM アルゴリズムにより $\boldsymbol{\mu}_d$ と $\boldsymbol{\xi}_k$ を求める [17]．

A.2 小論文の各課題内容

4 章で利用した小論文の各課題内容について**表 A·1** に記述する．どの講義でも課題 1 は講義で述べている内容を尋ねているのに対して，課題 3 では解答者の考えを聞く内容になっている．

A.3 Jess の内容評価を利用したモデル

日本語小論文評価採点システム Jess [49] を本小論文データに適用した場合の評価結果を記述する．Jess は小論文に対して「修辞」,「論理構成」,「内容」についてそれぞれ評点を出力する．Jess の「内容」に対する評価が本論文の「理解力」に対応していると考えられる．Jess は Web 上で利用することができる[*37]．Web 上の Jess では「質問文」と

[*36] 3.4 節で述べたように文書内の形態素の頻度を tf-idf 値に置きかえて入力するため，c_{dw} は文書 d 内での形態素 w の tf-idf 値となる．
[*37] http://tk2-203-11024.vs.sakura.ne.jp/jess/

表 A·1 各課題
Table A·1 Essay prompts.

講義記号	課題番号	課題内容
g	1	グローバリゼーションは、世界、または各国の所得格差をどのように変化させましたか。また、なぜ所得格差拡大、または縮小の現象が現れたと考えますか。300 字以内で答えなさい。
	2	多国籍企業は、グローバリゼーションの進展の中でどのような役割を果たしましたか。多国籍業の具体例をあげて、250 字以内で答えなさい。
	3	文化のグローバリゼーションは、私たちの生活にどのような影響を与えましたか。また、あなたはそれをどのように評価しますか。具体例をあげて、300 字以内で答えなさい。
s	1	「科学的」とはどのような条件をみたす必要があるのか 100 字以内で答えよ。
	2	講義で解説した自然科学の二つの側面を参考に、自然科学が果たす役割について 400 字以内で論ぜよ。
	3	「Scientific and Technological Literacy for All」の狙いを考慮し、これからの科学教育はどうあるべきか 500 字以上 800 字以内で論ぜよ。
e	1	日中韓の相互依存の強さを、データを示して簡潔に述べなさい。また、相互依存を示す経済協力・協業の具体例をあげ、合わせて 300 字以内で答えなさい。
	2	「中所得国の罠」の概略を説明し、どうしたらそれを乗り越えることができるか 250 字以内で説明しなさい。
	3	日中韓には少子化や環境問題など 3 国に共通する経済問題がある一方、それぞれの国に特有の課題も多くあります。それぞれの国が抱えている特徴的な経済問題をあげ、東アジアにおける協調と対立の構造を 300 字以内で説明しなさい。
c	1	「批判的思考」の定義に関連して、「批判的思考」に関する研究で共通に見出される「批判的思考」の 3 つの観点を述べなさい。100 文字
	2	講義で紹介した右のグラフを根拠に「長生きするためにはカラーテレビを多く所有すれば良い」と主張することが妥当ではない理由を 400 字以内で述べなさい。ただし、このようなグラフが形成される理由の説明を加えること。
	3	各自で「ニセ科学」の可能性があると思う実例を挙げ、その実例が「ニセ科学」であることを証明するためには、どのような方法で、どのような証拠を得て、どのように説明する必要があるのかを論じなさい。また、その実例がニセ科学でも信じてしまいやすい要因は何かについても考察し、説明しなさい。ただし、講義で扱った事例以外のものを挙げること。500 字以上 800 字以内。

表 A·2 参考文書を利用した Jess による評価 (QWK)
Table A·2 QWK scores of Jess using a reference text for each prompt.

課題						
g_1	g_2	g_3	s_1	s_2	s_3	
0.082	0.011	0.046	−0.111	0.035	0.107	
e_1	e_2	e_3	c_1	c_2	c_3	平均
0.098	0.039	0.024	0.260	0.087	0.159	0.070

表 A·3 模範答案を利用した Jess による評価 (QWK)
Table A·3 QWK scores of Jess using a reference essay with the highest score for each prompt.

課題						
g_1	g_2	g_3	s_1	s_2	s_3	
0.009	0.020	0.109	0.333	0.041	0.079	
e_1	e_2	e_3	c_1	c_2	c_3	平均
0.062	0.024	0.026	0.136	0.037	0.216	0.091

「解答文」を入力することで採点が実行される．そこで参照文書を「質問文」に入力し，答案を「解答文」に入力することで評価スコアを得る[*38]．

「内容」の評価スコアを取り出し，本手法の余弦類似度の場合と同様に式 (3) の *sim* および式 (13) および式 (14) の *sim_LSI* を Jess の出力の値に置きかえて 5 点に正規化した値を求める．さらに，式 (28) で 1 から 5 点に階級化して QWK で評価する[*39]．**表 A·2** に参考文書を利用した場合，**表 A·3** に模範答案を利用した場合の QWK を示す．

表 A·2 から，参考文書を利用した場合 Jess が低い QWK を示すことが分かる．これは 4.4 節で記述したとおり，参考文書が答案に対して大きく，課題と無関係な部分が文書間の類似度計算に影響を与えたためと考えられる．

模範答案を利用した場合，表 A·3 から課題 s_1 で QWK が向上する結果が得られた．この傾向は本手法の LSI も同様である．一方，課題 c_3 では Jess が LSI に比べて高い QWK を示した．QWK の値が本手法の LSI と異なる原因の 1 つとして学習データが違うことがあげられる．本手法の LSI の学習には日本語 Wikipedia を利用しているが Jess では新聞記事を利用している．こうした学習データと評価する小論文答案との関係性については今後の課題である．

[*38] このとき，字数制限も入力する．また質問文の中に課題文は入力しなかった．理由としては Jess の内容評価は LSI を利用しており，「質問文」は本論文での「参照文書」に対応すると考えられる．本手法では課題文は参照文書に入れていない．よって Jess の場合も同様に課題文を入力せずに実験した．

[*39] 「内容」配点は 3 で出力した．予備実験で配点を 5 に変更して正規化を行わず整数化のみ行って平均 QWK を求めた場合，配点 3 の場合よりも平均 QWK が低くなったため配点は変更しなかった．

A.4 模範答案を利用した識別モデルによる小論文評価

各課題で模範答案を利用して，識別モデルを利用することで小論文を評価することを試みる．各課題で評価の高い小論文は模範答案に類似すると考えられる．よって模範答案で 12 個の課題（$\{g_1, g_2, .., c_3\}$）を分類する識別モデルを学習で求めて，評価したい小論文答案に対して正解となるクラスの識別関数の出力値を評価スコアとして利用する．

識別モデルとして 3.6 節の BERT を利用する．3.6 節と同様に文書 d を BERT に入力し，[CLS] に対応するユニットの文書ベクトルを $\boldsymbol{t}_d \in \mathbb{R}^{vec}$ とする．ここで vec は [CLS] のユニット数とする．下記に示すように課題に対応した 1 層ニューラルネットワークを結合して，文書 d に対するベクトル $\boldsymbol{y}(d) \in \mathbb{R}^{12}$ を得る．

$$\boldsymbol{y}(d) = f(\boldsymbol{W}\boldsymbol{t}_d + \boldsymbol{q}) \tag{A.11}$$

ここで，非線形関数 f は softmax 関数であり，$\boldsymbol{W} \in \mathbb{R}^{12\times vec}$ および $\boldsymbol{q} \in \mathbb{R}^{12}$ はユニットに付随する重みである．答案 E に対してベクトル $\boldsymbol{y}(E)$ を求め，そのうち答案 E の属する p 番目の課題に対応する出力を E の評価スコア $scr(E)$ とする．

$$\boldsymbol{y}(E) = [y(E)_1, y(E)_2, \ldots, y(E)_{12}]^T \tag{A.12}$$

$$scr(E) = y(E)_p \tag{A.13}$$

QWK の比較では 3.6 節と同様に式 (3) の sim および式 (13) と式 (14) の sim_LSI を scr に置きかえて答案 E の点数 $grade(E)$ を求める．

学習時には各模範答案を 1 文に分割して事例数を増加させた．学習時は $\boldsymbol{W}$ と $\boldsymbol{q}$ および BERT の重みも学習するファインチューニングを適用した．学習時のパラメータとしてはバッチサイズは 16，学習回数は 7 回とした．各課題に対する QWK の値を**表 A·4** に示す．

表 A·4 では平均 QWK は表 6 と比較して学習を適用しなかった BERT よりも低い値となった．しかしながら課題 s_1 では QWK が 0.460，g_1 と e_2 では QWK が 0.3 付近となり学習により向上する場合があることが明らかになった．この学習では同じ課題で評価の低い答案が学習データに入っていないため，より高い QWK を得るためには負例をどのように集めるかが鍵となると考えられる．この結果からも模範答案が利用できる場合は形態素を利用した手法が有効であると考えられる．

表 A·4 模範答案を利用した識別モデルによる評価（QWK）
Table A·4 QWK scores of a fine-tuned BERT-based model trained on reference essays of the highest score for each prompt.

課題						
g_1	g_2	g_3	s_1	s_2	s_3	
0.301	−0.064	0.106	0.460	−0.011	0.008	
e_1	e_2	e_3	c_1	c_2	c_3	平均
−0.037	0.302	0.171	0.180	0.168	0.002	0.132

竹内 孔一 （正会員）

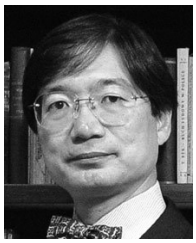
1998 年奈良先端科学技術大学院大学博士後期課程修了．博士（工学）．同年学術情報センター助手．2000 年国立情報学研究所助手．2003 年岡山大学工学部情報工学科講師．2021 年同大学学術研究院自然科学学域准教授．現在に至る．主に，専門用語研究，述語項構造の言語資源構築と解析に従事．言語処理学会，人工知能学会，電子情報通信学会，ACM 各会員．

大野 雅幸

2017 年岡山大学工学部情報系学科卒業．2019 年岡山大学大学院自然科学研究科博士前期課程修了．在学中，採点支援システムの開発に従事．住友電工情報システム株式会社入社．

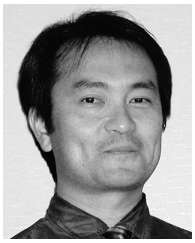
泉仁 宏太

2017 年岡山大学工学部情報系学科卒業．2019 年岡山大学大学大学院博士前期課程終了．在学中，採点支援システムの開発に従事．株式会社 NTT データ MSE 入社．



田口 雅弘

1988 年京都大学大学院経済学研究科博士課程後期単位取得満期退学．2006 年京都大学博士（経済学）．2007 年岡山大学大学院社会文化科学研究科教授．2021 年同大学学術研究院社会文化科学学域教授．現代ポーランド経済研究に従事．ロシア・東欧学会会員．



稲田 佳彦

1995 年東北大学大学院理学研究科物理学第二専攻博士後期課程修了．博士（理学）．同年大阪大学理学部物理学科助手．2002 年岡山大学大学院教育学研究科助教授．2009 年同大学院教授．2021 年同大学学術研究院教育学域教授．主に，強相関電子物性，超伝導，科学教育，創造性教育に従事．日本物理学会，応用物理学会，日本物理教育学会，日本教育工学会，日本理科教育学会，各会員．

飯塚 誠也

1997 年東海大学大学院博士課程前期修了．1999 年岡山大学大学院博士後期課程中途退学．博士（理学）．2013 年岡山大学アドミッションセンター教授，2016 年同大学全学教育・学生支援機構教授．計算機統計学の研究に従事．日本統計学会，日本計算機統計学会，日本分類学会，日本行動計量学会各会員．

阿保 達彦

1994 年東京大学大学院博士課程修了．博士（農学）．同年ラホヤ癌研究所（アメリカ合衆国），1996 年コーネル大学（アメリカ合衆国）博士研究員．1998 年名古屋大学理学部助手．2002 年岡山大学理学部助教授．2019 年同大学大学院自然科学研究科教授．2021 年同大学学術研究院自然科学学域教授．

上田 均

1985 年東北大学大学院農学研究科博士課程後期修了．農学博士．同年 National Institutes of Health（アメリカ合衆国）Visiting fellow．1987 年国立遺伝学研究所助手．1998 年同助教授．2004 年岡山大学大学院自然科学研究科教授．2021 年同大学学術研究院自然科学学域教授．