View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Okayama University Scientific Achievement Repository

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

信学技報
IEICE Technical Report
NLC2018-33 (2018-12)

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

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†††††

† 岡山大学大学院自然科学研究科　〒700–8530 岡山市北区津島中 3 丁目 1 番 1 号
†† 岡山大学院社会文化科学研究科
††† 岡山大学院教育学研究科
†††† 岡山大学全学教育・学生支援機構
E-mail: †{pw2z9792,pm9n6cei,pbgn8vxd}@s.okayama-u.ac.jp, ††koichi@cl.cs.okayama-u.ac.jp

あらまし　大学入試において 2020 年から記述式問題が導入されることから記述式の問題を自動で採点する手法の開発が求められている．本論では，エッセイタイプの小論文課題を対象に，課題に関連する参照データと Wikipedia 全文から作成した idf を利用した事前採点不要な自動採点手法を提案する．先行研究において，日本語小論文を対象とした自動採点では，多くの事前採点が必要となり，実際の数百人規模の試験では利用することが難しいと考えられる．そこで本研究では，事前採点が不要な小論文採点手法を提案する．また，小論文の模擬試験を実施して小論文データを構築する．構築した小論文データに対して採点手法を用い，実験を行い評価する．また小論文データの人手による採点に対しても評価を行う．評価実験の結果 neologd 辞書を利用した形態素解析器を用いて， idf 値を利用した形態素の一致数が，人手の評価値と相関が高いことを示す．

キーワード　自動採点，アノテーション，採点支援，idf，neologd

Proposing an Unsupervised Approach to Evaluate Essays Using IDF on Reference Data

Masayuki OHNO†, Koichi TAKEUCHI†, Kota MOTOJIN†, Yuya OBATA†, Masahiro TAGUCHI††, Yoshihiko INADA†††, Masaya IIZUKA††††, Tatsuhiko ABO†††††, and Hitoshi UEDA†††††

† Graduate School of Natural Science and Technology, Okayama University
3–1–1 Tushimanaka, Kita-ku, Okayama,
†† Graduate School of Humanities and Social Science, Okayama University
††† Graduate School of Education, Okayama University
†††† Institute for Education and Student Services, Okayama University
E-mail: †{pw2z9792,pm9n6cei,pbgn8vxd}@s.okayama-u.ac.jp, ††koichi@cl.cs.okayama-u.ac.jp

Abstract　In this paper, we describe an on-going study of developing an automatic essay-scoring system in Japanese. Essay scoring systems have already been developed and used mainly in English, while not many previous studies have been done on Japanese essay evaluations. Most of the methods and systems of automatic essay evaluation need not small number of previously human-graded essays for calibrating the parameter of regression functions or parameter of machine learning. The previous studies show the high performance for essay evaluation task, however, it must be not easy to assume large graded essays in, for example, actual tests or entrance examinations. Thus, we take a approach to evaluate Japanese essays without previously human-graded essays but with assuming reference data related to essay questions. The proposed method is a simple one, that is, evaluating the essays with co-occurrences with the reference data in their words or morphemes. In the method technical terms would be given high scores using neologd dictionary and idf values. Experimental results show that the proposed method works well in our developing Japanese mock trial writing tests.

Key words　automatic scoring of essays, human annotation, supporting system of essay evaluation, idf, neologd,

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2018　by　IEICE

1. はじめに

本研究では日本語の小論文に対して，自動的に採点する手法を検討するとともに，必要となる研究用の模擬試験データの構築，ならびに，自動採点手法を採点時に採点者に提示し，採点の負担を軽減する採点システムを構築している．

既に英語圏では e-rator や IntelliMetrics などアメリカ経営大学院の入学試験 GMAT（Graduate Management Admission Test）で使われている実用の段階である [1]．しかし自動採点システムは小論文のすべての要素を評価できるわけではないため，公的な採点では自動採点は人との併用で使われている．本研究でも同様に人と自動採点システムの併用を想定して構築する．

先行研究で提案されている小論文の自動採点手法の多くは事前に人手による採点データを必要とする．先に述べた e-rater [2] では，ルールの発見の手法が利用されているが，それは人手による採点データを利用してパラメータを調整する．また近年では，Hewlett Foundation による小論文データが構築され(注1)より大規模な人手による採点データを使用が可能になり，機械学習 (例えば [3]) および深層学習 [4]〜[6] を利用した採点手法が提案されている．また，日本語でも短答式に対して深層学習を適用している研究が発展している [7], [8]．

しかしながら，事前に採点データを必要とする手法は大学入学試験の小論文採点や授業でのレポートでの採点では利用することは容易ではないと考えられる．採点データを事前に必要としない手法として，LSI を利用した手法 Jess [9]，キーワードや短い表現を利用して評価する手法 [9], [10]，もしくは，既に採点したデータを蓄積しながら漸近的に評価する手法 [11] が提案されている．

そこで，本研究では小論文課題に関連した参照データを利用して事前採点不要な自動採点手法を構築を目指す．参照データ内の専門内容に近い言葉を評価するために neologd 辞書を利用して形態素解析を行い，参照データと小論文との一致度を利用する．さらに，評価の際に Wikipedia から獲得した idf 値を利用する手法を提案する．研究室で構築している小論文採点データ (公開予定) に対して比較実験を行ったところ，neologd 辞書を利用した形態素解析システムによる，内容語の一致数に Wikipedia から作成した idf 値を重みとして利用した手法が先行研究 [12] に比べ，人手による評価に近いことを報告する．また，自動採点手法をとりいれた採点支援システムを構築している．システムについて記述し，具体的にどのように採点者を補助するかについて説明する．

2. 小論文採点支援システムと模擬試験データ

本章では構築を目指している小論文採点支援システムの概要，および構築中の小論文データ，システムの現状について述べる．

2.1 小論文の評価軸

先行研究で述べているとおり [13], [14]，小論文の採点の標準的な評価基準は存在しない．そのため本研究では課題に対する理解力，文章の論理性，論理的な妥当性，誤字脱字など誤りの少ない文書を記述する文法力の 4 つの基準を評価軸として策定し，自動採点手法および，採点支援システム，小論文データを構築する．

2.2 小論文データの構築

日本語の小論文採点研究において問題となるのは，研究利用可能な小論文答案および，採点結果が存在しないことである．そこで本研究プロジェクトでは 2016 年度から 2018 年度にかけて毎年，大学生を対象に小論文の模擬試験を行い，研究利用許諾を受けつつ答案データを作成している．さらに，答案に対してルーブリックを作成し，基準をたてた上で複数人で採点を進めている．

2016 年と 2017 年では受講者が講義を受け，その講義内容に関する小論文課題を解く形式を採用した．講義の内容を整理して記述したデータを参照データとして用意し，自動採点に利用する．講義内容は 2 種類あり，各講義について 3 課題出題した．一部を除き，解答者はマス目の用紙に対して筆記により小論文を記述する．記述された答案を人手により電子化したテキストデータを採点システムは利用する．その際文字間違いなどで存在しない文字を書いている場合は●（黒丸）で表した．例えば部首を間違えて記述し，存在しない漢字になった場合などがこれに該当する．例として”負荷”という言葉を記述する際，誤って’荷’という漢字を「くさかんむり」ではなく「たけかんむり」にしてしまい存在しない文字となった場合が挙げられる．図 1 に講義 1（4.1 節参照）の課題 1 に対する解答例を示す．図でわかるように文字は小さくまた筆記であるため既存の OCR などでは精度高く取り込むことができなかった．マス目内で改行する記述方法も指導されていることが多く，誤りではない．よって後の人手の評価における指定文字数の充足度において，単にマス目だけでなく行数も考慮して充足率による人手の評価を行っている．本研究ではこれらのデータのうち，整理が終了している講義 1 の課題 1 から 3 を利用する．

図 1　小論文の例

上記の小論文答案に対して人手による採点を行っている．上記で記述した理解力，論理性，妥当性，文法力のそれぞれについて採点基準をルーブリックとして整理した．各評価軸で点数は 1 から 5 点とした．ルーブリックを元に各小論文に対して 2 名の採点者が独立に採点する．この 2 人分の採点結果や採点中のメモなども研究利用可能な形で公開する予定である．

2.3 小論文採点支援システム

構築中の小論文採点支援システムについて述べる．小論文採

(注1)：https://www.kaggle.com/c/asap-sas

点では，小論文データの整理，及び採点後の一貫性を確認するための並べ変え，など基本的な操作が可能なように構築されている．自動採点はこの一部の機能として取り込んでいる．仮定する利用法として，機械的な手法が一貫して計算する点数を提示することで，人間が採点する際に揺れているかどうか自己確認するというものである．また周辺機能として，小論文問題 (ここでは講義が 1 単位) や答案をシステムに登録する部分を構築している．小論文課題や参照データ，答案を指定された excel 形式で作成するとブラウザベースで upload することが可能で，システムの詳細をしらなくても小論文などの採点補助ができるわくぐみになっている 図 2 に小論文採点支援システムのメ

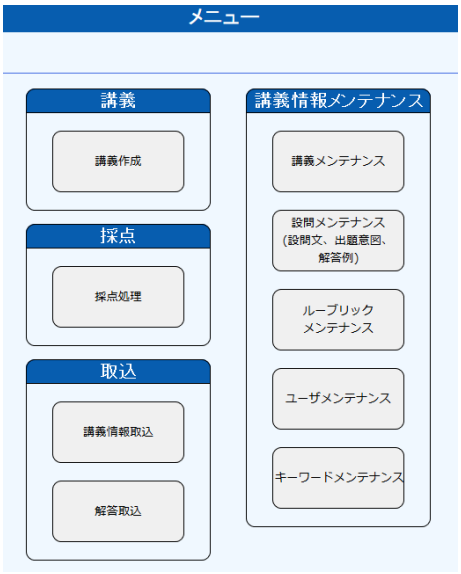

図 2　小論文採点支援システムのメニュー

ニュー画面を表示する．講義を作成する機能，解答を取り込む機能，ならびに採点処理を行う機能があることがわかる．図に

図 3　システムによる採点を参照し人手による採点を行う画面

小論文採点の際の操作画面の例を表示する．操作画面では受講者の小論文データが番号順で並び，システムで自動評価された点数が同時に表示される．採点者は自動採点結果を勘案しつつ最終的な評価を付与する．また付与した結果を点数で並び変えることが可能である．これにより同じ点数をつけた小論文の集団を確認することで，自分の評価の一貫性がくずれていないかを確認することが可能になる．並べ変えは自動採点のスコアでも可能で，機械的な手法による評価結果での小論文の並びも確認することができる．

3.　評価モジュール

4 つの評価軸にて評価を行うため，採点モジュールに関しても 4 つの値を出力する必要がある．しかし本論ではそのうちの理解力評価のために作成されたモジュールについて述べていく．他の 3 つの評価を行うモジュールについても現在構築が進められている(注2)．

講義の内容に対して，高い理解力を持った受講生によって書かれた小論文回答には，講義の内容等の課題に関するデータとよく似た文章にになると考えられる．そのため，講義内容と回答の類似度によって評価を行う．

3.1 節では内容語の形態素類似度評価について述べ，3.2 節で各単語に idf 重みをつけた場合について述べる．ここで形態素 n-gram による語の並びを考慮した評価を行わず，単純な形態素の一致を測るのは n-gram の一致数よりも形態素の一致数の方が人手に近い相関が出ることが分かっているからである [12].

3.1　内容語の形態素類似度評価

内容語の一致数で評価する関数を構築する．文や採点に使用するデータの形態素区切りを行うために本研究では，形態素解析器 CaboCha(注3) を用いた．実験当初はデフォルトの辞書を使用していたが，本システムは大学の小論文入試や大学の講義レポートでの利用を考えているため，そのような課題で使われるような言葉が 1 形態素として認識されなかった．例えば”ジニ係数”という言葉が課題 1 では多く使用されるが，デフォルトの辞書を用いた場合”ジニ”と”係数”という 2 形態素に分かれて解析される．そのため本研究ではそのような専門用語をカバーするため，専門用語を単語として解析することができる mecab-ipadic-NEologd(注4) の 2017 年 6 月 27 日のものを利用した．これによって”ジニ係数”はいつの形態素として認識されるようになった．また形態素解析を行った結果の中から名詞，動詞，形容詞で自立語の内容語とそのほかの機能語に分け，文章の意味を担わない機能語は無視し，文章の意味を表す内容語のみで評価を行う．このように小論文と課題に関するデータから内容語を抽出し一致数を評価とする．文書 A, B に出現するそれぞれの内容語を a, b とすると内容語が一致しているかどうかを返す関数 $sim(a, b)$ は式 (1) で表される．それを利用した文書 A, B の内容語の一致数 $score\_match(A, B)$ は式 (2) となる．

$$sim(a,b) = \begin{cases} 1 & (a = b) \\ 0 & (a \neq b) \end{cases} \tag{1}$$

(注2)：例えば妥当性モジュールの評価法に関してはこの文献 [15] で発表している．
(注3)：https://taku910.github.io/cabocha/
(注4)：https://github.com/neologd/mecab-ipadic-neologd/blob/ master/README.ja.md

$$score\_match(A,B) = \sum_{a \in A, b \in B} sim(a,b) \quad (2)$$

以下に例を示す．

例文 1　グローバリゼーションに伴い世界的格差は徐々に縮小を見せる一方で各国内での格差は拡大した。

例文 2　グローバリゼーションによって、先進国と発展途上国の所得格差はそれぞれ変化した。

これらの例文の類似度を評価するため，まず形態素解析を行う．行った結果が以下である．

例文 1　グローバリゼーション/に/伴い/世界/的/格差/は/徐々に/縮小/を/見せる/一方/で/各国/内/で/の/格差/は/拡大/し/た/。

例文 2　グローバリゼーション/によって/、/先進/国/と/発展/途上/国/の/所得/格差/は/それぞれ/変化/し/た/。

続いて，両方の文に現れている形態素は「グローバリゼーション」,「格差」,「は」,「し」,「た」,「。」そのうち内容語は「グローバリゼーション」,「格差」の 2 つなので一致数は 2 と評価される．

3.2　*idf* による重みづけ

Wikipedia の全文書 (2016/10/1 最新版) を用いて idf 重みを計算した．そして 1386126 単語の idf 重みを得た．この重みを 3.1 節の手法で抽出した内容語に重みづけし，その結果で評価を行った．その際，Wikipedia から作成した idf 重みに存在しない形態素に関しては，ノイズとして取り除いて評価を行った．ここで取り除いたものは全角数字や漢数字であった．先ほどと同様に各文章の内容語を a, b とすると内容語が一致した場合，その形態素の idf 重みを返す関数 $sim_{idf}(a,b)$ は式 (3) で表される．それを利用した文書 A, B の内容語の一致数 $score\_match\_idf(A,B)$ は式 (4) とする

$$sim_{idf}(a,b) = \begin{cases} w_{idf} & (a = b) \\ 0 & (a \neq b) \end{cases} \quad (3)$$

$$score\_match\_idf(A,B) = \sum_{a \in A, b \in B} sim_{idf}(a,b) \quad (4)$$

前節の例で説明すると，Wikipedia 全文から作成した idf 重みは「グローバリゼーション」は約 8.16,「格差」は約 6.75 なのでスコアは約 14.91 となる．

4. 評価実験

評価実験を行うために，まず利用した小論文課題 (講義と課題内容)，人手による採点，評価関数について記述する．その後，提案手法による評価結果を示し，先行研究の結果 [12] と比較する．

4.1　小論文課題について

2016 年度に行われた講義 1(約 300 件) を利用する．講義に対して課題 1 から 3 が設定されており，受講者は 30 分講義を受けた後 1 時間で 3 問の課題について小論文を作成している．講義のタイトルと各課題は下記の通りある．

講義 1 のタイトル：　グローバリゼーションの光と影

課題 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。

課題 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。

課題 3：　文化のグローバリゼーションは、私たちの生活にどう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。

課題 1 は講義で説明した内容から単に整理して記述するだけの課題であるが，課題 2，課題 3 となるにつれて解答者自身が考えて記述する必要がある課題になっている．よって課題の性質として，課題 1 は講義内容に則した解答が評価される一方で，課題 3 は背景知識などが必要となる問題である．

この講義課題に対して 1 つの参照データを構築した．参照データは講義内容を整理して記述したテキストで約 2600 文字で記述されている．内容の一部は下記の通りである．

> 政治、経済、文化などの分野で、様々な現象が地球規模で展開していくことをグローバリゼーションといいます。たとえば、市場経済が世界の隅々に行き渡ること、同じ金融のルールが世界中で共有され資金がより広い範囲で流通するようになること、インターネットの普及により世界のどこからでもいち早く同じ情報を共有できるようになることなどは、グローバリゼーションの典型的な側面です。また、地球環境問題、世界の食糧問題など、様々な課題に対する対策の地球規模化もグローバリゼーションの一側面です。(続く)

自動採点手法はこの参照データを利用するので，事前の採点を必要とせず評価する．よって，小論文課題に即した内容であれば自動採点手法を利用することが可能になる．

また，答案データは後の表 2 に示すように各課題に対して，得られた小論文は 328 件，327 件，293 件であった．各課題に対して小論文数が異なるのは白紙解答を差し引いているためである．

4.2　人手による採点

上記の小論文課題に対する受講者の答案に対して人手で採点した．採点は上記の 4 軸 (理解力，論理性，妥当性，文法力) を 1 から 5 点で評価し，2 名の作業者で独立に付与した．人手による採点の揺れは法学における論述で既に述べられるている [16]．そこで，2 名の評価者の異なりについ一部のデータに対して第 4.3 節で説明されている評価関数を利用して評価の違いを調べる．また採点は理解力，論理性，妥当性，文法力について付与されているが，以下では，理解力に対する評価値を利用する．結果を表 1 に示す．

表 1 の結果から，課題 1 および課題 2 では相関係数が 0.7 を越えており，とても高く一致した結果であることが分かる．これは課題内容に対する書くべき内容がはっきりしていることから 300 字の字数であっても，人手による判断では揺れが少ない

表 1　2 名の評価者の異なりの結果

課題	件数	相関係数	accuracy	QWK	RMSE
1	83	0.808	0.482	0.743	0.744
2	83	0.743	0.446	0.742	0.836
3	83	0.366	0.422	0.317	1.340
average		0.639	0.450	0.601	0.973

ことが示されている．一方で，解答に自由度がある課題 3 では，人手の相関が低くなってきている．自由度が高い分，人による評価が一致せず，低い値になった．本実験ではシステムと人手による採点者 1 人の採点結果を比較する．選択した方の採点者は採点経験が長く他の講義課題についても最も多く採点している作業者の結果を利用した．

4.3 評 価 関 数

比較方法として相関係数，accuracy，Quadratic Weighted Kappa (QWK)，Root Mean Squared Error (RMSE) の 4 つの評価を行う．その際，相関係数以外の評価尺度で測るためにシステムの採点と人手の採点が同じ n 値分類でされている必要がある．人手による採点では 1 から 5 点の 5 値分類で採点を行っているのに対して，システムはスコアを加算していく計算になっているため，システムの採点結果を 5 点に丸める必要がある．そのためシステムの採点の結果集合 S の最大値 S_{max} と天井関数を用いてシステムの採点結果を 5 値分類に丸める．各採点結果のスコア x とすると 5 点に丸めたスコア x_{round} は以下の式 (5) となる．

$$x_{round} = \lceil \frac{5 \times x}{S_{max}} \rceil \tag{5}$$

以降，相関係数以外の評価項目に関してはここで丸めた値を用いて評価を行う．

評価を行う 2 つのスコアをそれぞれ m，n とし，m，n と採点された回数を $ob(m,n)$，偶然 m，n と採点される確率を $ex(m,n)$ とすると，QWK は以下の式（6）で表される．

$$QWK = 1 - \frac{\sum_{m,n=1}^{5} ob(m,n) \times |m-n|^2}{\sum_{m,n=1}^{5} ex(m,n) \times |m-n|^2} \tag{6}$$

これは 1 に近いほど一致度が高いと言える．

採点した小論文の数を t，l 番目の採点結果をそれぞれ m_l，n_l とすると $RMSE$ は以下の式（7）で表される．

$$RMSE = \sqrt{\frac{\sum_{l=1}^{t} |m_l - n_l|^2}{t}} \tag{7}$$

これは 0 に近いほど誤差が少ないと言える．

4.4 実験結果と考察

2 種類の評価手法について上記の課題 1 から 3 に対する評価結果を示す．まず，表 2 に参照データと各小論文との形態素の一致度を利用した手法を適用した場合について人手の評価値との差を示す．結果はそれぞれ小数第四位を四捨五入している．次に，Wikipedia の idf 値を利用した手法を適用した結果を表 3 に示す．まず表 2 について検討する．課題 1 では講義内容に即した形の課題であるにも関わらず，人手との相関係数が低く，

表 2　内容語のマッチ数の結果

課題	件数	相関係数	accuracy	QWK	RMSE
1	328	0.104	0.207	0.036	1.509
2	327	0.233	0.183	0.107	1.478
3	293	0.379	0.287	0.287	1.147
average		0.239	0.226	0.127	1.378

表 3　内容語の idf 重み手法の結果

課題	件数	相関係数	accuracy	QWK	RMSE
1	328	0.093	0.314	0.059	1.293
2	327	0.229	0.220	0.130	1.399
3	297	0.433	0.433	0.260	1.214
average		0.252	0.322	0.150	1.302

accuracy および QWK が低い値になっている．また値そのもののずれを示す RMSE も大きい．これは回答内容が 300 字であり，書き方に自由度があるため，選択する単語の幅が広がり，参照データだけでは捉えられない表現が多く正解として人でで評価されていることが原因として考えられる．課題 1 の模範的な解答は，「国家間では格差は縮小傾向」である一方で，「国内では格差が広がる」というものである．これらの言い換えの幅は広く，人間にとっては簡単な言い換えも単語だけの観測ではうまく捉えられていない．また，単純に形態素の品詞で内容語を全て数えたため，「する」など意味の無い言葉も多く数えられたのが精度が低い原因である．

課題 2，課題 3 に向けて解答者の考えを聞いている問題であるが，単純な手法であるにも関わらす，相関係数および他の評価指標も改善する傾向が見られた．これは本来，課題提案者はより幅広い事例に関して，解答者が答えることを期待していたが，実際には事例がほとんど講義で触れた内容に即しており，また，課題で講義の事例を利用することは排除していないことから講義の内容に即した解答が高い評価を得る結果を得ている．その結果，参照データとの形態素の一致数をみた手法でもある程度評価することが可能になった．

続いて表 3 の idf 値を利用した場合の結果と比較する．表 2 と比較して課題 1 から 3 の全てで評価値が改善している．課題 1 に関して相関係数は少し下がっているが QWK が上昇しているため，ずれ方に関してより良い方向に変化している異が分かる．よって専門用語などに対して重みを与えた方が，人手の評価値とよく合うことが分かる．

idf 値の値で課題 3 では相関係数は 0.433 と人手の評価と相関が大きくなってきている．また RMSE も低く，QWK の値も上がっている．単純な方法ではあるが重要語句を認識して重みを与えることで，事前の人手による正解データがなくても，小論文の内容の良さに関してある程度評価可能であることが分かる．

次に先行研究 [12] との比較を行う．先行研究では，単語の頻度と n-gram，さらに，重要語を人手で指定した手法が最も良い値を示していた．そこで先行研究の手法を本実験データの課題 1 に適用して本提案手法との比較を行う．具体的には先行研究

の手法は参照データとの単語の一致数に加えて，形態素 n-gram（1-gram から 4-gram），重要語として”ジニ係数”を選び重みを 2 倍にした．また形態素は MeCab の IPAdic を利用している．適用した結果を表 4 に示す．

表 4　先行研究での単語と n-gram，内容語を利用した手法の結果

課題	件数	相関係数	accuracy	QWK	RMSE
1	328	0.009	0.259	0.008	1.303

結果は相関係数が 0.009, accuracy が 0.259, QWK が 0.008, RMSE が 1.303 となった．これは相関係数が低く，得点の高い小論文を逆に低く評価している場合が多いことを示している．一方で，accuracy は少し高いのは偶然評価値が一致していることを示している．

この結果から，課題 1 に対して先行研究の手法では小論文評価をうまく捉えられていないことがわかる．本稿が提案する idf 値を利用した手法が相関係数，accuracy， QWK，RMSE の全てにおいて勝っていることから，提案手法が優位であることが示された．

5. おわりに

本論文では記述式問題の中でも長文である小論文に対する自動採点手法を提案し，実験による手法の有効性について議論した．また自動採点手法を作り上げる上で，必要となる小論文データの構築について記述した．さらに，自動採点手法を取り入れた，採点支援システムについて記述し，提案手法をどのように実際の状況で利用するかについて現状を明らかにした．

本研究では 4 つの評価軸のうち，理解力の評価に関する自動採点手法を提案した．手法の特徴として neologd を利用した形態素解析を行うこと，さらに，形態素と参照データの一致を Wikipedia を利用した idf 値を利用して評価する手法を提案し，単純に単語の頻度による評価よりも人手によるスコアとの一致が高いことを実験的に示した．

実験対象として，講義内容などで説明したことを解答者に答えさせる容易な課題から自分で考えさせる課題があったが，容易な課題の場合に提案手法はうまく機能しなかった．これは先行研究などの 100 字以下の短答式とは異なり，300 字の余裕があるため，表現に幅が出て，参照データでは捉えきれない異なる形態素を利用して正しく解答した小論文が多かったためである．

今後の課題として，こうした幅広い言い換えに関して捉えることのできる言語モデルを取り込むことで適切に同様の表現を評価できる手法の開発を目指したい．

6. 謝　辞

模擬試験の実施ならびに研究の遂行にあたり岡山大学学務部にご協力いただきました．深く感謝いたします．

文　献

[1] 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.

[2] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning, and Assessment*, Vol. 4, No. 3, pp. 1–30, 2006.

[3] Hongbo Chen and Ben He. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1741–1752, 2013.

[4] Fei Dong and Yue Zhang. Automatic features foressay scoring âĂŞ an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1072–1077, 2016.

[5] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. In *Proceedings of the the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 715–725, 2016.

[6] Madalina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 503–509, 2018.

[7] 寺田凛太郎, 久保顕大, 柴田知秀, 黒橋禎夫, 大久保智哉. ニューラルネットワークを用いた記述式問題の自動採点. 第 22 回言語処理学会年次大会発表論文集, pp. 370–373, 2016.

[8] 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.

[9] Tshuneori Ishioka, Kohei Yamaguchi, and Thuneori Mine. Rubric-based Automated Japanese Short-answer Scoring and Support System Applied to QALab-3. In *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, pp. 152–158, 2017.

[10] 中島功滋. 短答式記述答案の採点支援ツールの開発と評価. 言語処理学会第 17 回年次大会発表論文集, pp. 611–614, 2011.

[11] 高井浩平, 竹谷謙吾, 森康久仁, 須鎗弘樹. シーケンスアライメントを用いた記述式問題の採点支援システムの提案. 2018 年度人工知能学会全国大会論文集, 2L4-05, 2011.

[12] Masayuki Ohno, Koichi Takeuchi, Kota Motojin, Masahiro Taguchi, Yoshihiko Inada, Masaya Iizuka, Tatsuhiko Abo, and Hitoshi Ueda. Construction of open basic data for automatic scoring of essay and evaluation of automatic scoring method at current stage. 2017.

[13] 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.

[14] 石川巧. 「いい文章」ってなんだ？—入試作文・小論文の歴史. 筑摩書房, 2010.

[15] 泉仁宏太, 竹内孔一, 大野雅幸, 田口雅弘, 稲田佳彦, 飯塚誠也, 阿保達彦, 上田均. 小論文採点支援のための関連文書取得法の考察. 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. 47–51, 2017.

[16] 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, 法科大学院統一適正試験テクニカルレポート 2006, pp. 119–131. 商事法務, 2007.