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

平尾 礼央　新井 美桜　嶋中 宏希　勝又 智　小町 守
首都大学東京
{hirao-reo@ed., arai-mio@ed., shimanaka-hiroki@ed.,
katsumata-satoru@ed., komachi@}tmu.ac.jp

1　はじめに

作文は言語能力を測定するのに有効だが，採点に時間がかかり，採点可能な教師の数が限られているという問題がある．そのため，教師の代わりに機械によって作文を評価する作文自動評価の研究が盛んに行なわれている．既存の作文評価システムの多くは作文全体の点数は提示できるが，どこが悪いのかを指摘できないという問題点がある．特に学習者にとっては，語学教師無しでどこが悪いのかを予想するのは困難である．

自動評価において，より多くのフィードバックを提供するため，作文を複数の項目で評価する研究 [1, 2] が近年盛んに行なわれており，英語では複数項目に対して採点したデータセットが作成・公開されている．一方で，日本語作文のデータセットはほとんど公開されておらず，日本語学習者の作文データセットはさらに少ない．この問題を解決するため，田中と久保田 [3] は，作文全体の点数に加え，内容・構成・言語の三つの作文能力に関する項目について採点された日本語学習者の作文データセットを作成した．

本研究では，複数項目について採点されたデータセットを使用し，日本語学習者に焦点を当てた作文自動評価システムを作成した．作文自動評価システムは，素性を用いた手法とニューラルネットワークを用いた手法を使って大きく分けて 2 種類のモデルを開発した．素性を用いた手法では，李と長谷部 [4] によって提案された素性を用いた．これらの素性が点数の予測に十分かどうかは明らかではないため，素性を作成せずに点数を予測するニューラルネットワークを用いた手法を使用し，点数の予測を行なった．ニューラルネットワークを用いた手法では，BERT（Bidirectional Encoder Representations from Transformers）[5] を使用したモデルを作成した．本研究では，作成したモデル[1]の性能を評価し，実際の出力の分析を行なった．

2　関連研究

近年，英語作文における複数の項目を評価するシステムを含む様々な作文自動評価システム [6] が開発されているが，日本語の作文自動評価はデータの不足と入手が難しいことから研究があまり行なわれていない．

既存の日本語作文自動評価システムとして，Jess [7] と jWriter [4] が挙げられる．大学入学試験における小論文採点のために開発された Jess（Japanese essay scoring system）は，統計的な手法を使用し，全体の点数と内容・構成・修辞の点数を予測するシステムである．

Jess は毎日新聞のデータを基準とし，文長や漢字/ひらがな率のような素性を用いて違いを測定している．jWriter は日本語学習者に焦点を当てたシステムであり，全体の点数を 3 段階で予測する．このシステムは文長や品詞率などの素性を用いた重回帰分析によってスコアの予測を行なっている．限られた素性を使用することで，少ないデータで頑健性を獲得した一方で，統計量に基づいた素性しか使用していないため，素性が複数の項目の予測をするのに不十分である可能性が高い．

近年研究が盛んなニューラルネットワークを使用した手法 [8] は，素性を作成する必要がなく，作文自動評価の複数のデータセットにおいて高いスコアを獲得している [9]．しかし，ニューラルネットワークを使用した日本語の作文自動評価システムは開発されておらず，BERT を使用した複数項目を予測する作文自動評価システムは他の言語においてもまだ開発されていない．

3　使用したデータセット

本研究では，GoodWriting データセット[2]を使用した．このデータセットは日本語学習者によって書かれた 800 以上の作文とそれぞれの作文にあらかじめ与え

[1] https://github.com/reo11/aes-for-japanese-learner

[2] https://goodwriting.jp/wp/system-data

ID	A	B	C	D	E	F	ave.
κ	0.57	0.61	0.59	0.45	0.60	0.53	0.56

表 1: 最終的なスコアと各アノテータ間のカッパ係数

られたお題で構成され，それぞれの作文は 3 人のアノテータによって付けられた点数の中央値が点数として付けられている．全体の点数は 886 作文全てに付けられており，内容・構成・言語の項目の点数が付いた作文は 148 作文ある．全ての評価項目において，1 から 6 の 6 段階のスコアが付いている．

GoodWriting データセットは I-JAS データ，TK データ，EU データから構成される．I-JAS は 12ヵ国の日本語学習者，TK データは英語と中国語を母語とする日本語学習者，EU データは 11ヵ国の日本語学習者が書いた作文が使用されている．

3.1 作文の評価項目

GoodWriting データセットは，全体の点数と内容・構成・言語の項目における点数が付けられており，採点を行なう際に，アノテータ間の差異を少なくするため，採点用のフローチャート[3]に従ってアノテーションが行なわれた．

全体 全体評価では，細かな誤りではなく，全体の印象を重視し，お題に沿った議論がなされているかが評価される．その他の項目においても一定以上のレベルが達成される度に全体評価の点数が上昇する．

内容 内容評価は,「目的」と「内容」で評価される．「目的」は，お題にあった比較がなされ，意見が述べられているかを評価し,「内容」はメインアイディアの一貫性と根拠の妥当性を評価する．

構成 構成評価は,「構成意識」,「段落意識」,「結束性」で評価される.「構成意識」は話の順序を考える意識であり,「段落意識」は 1 つのまとまった話を同じ段落に書こうとする意識，マクロ構成（序論・本論・結論）を評価し,「結束性」は段落間の繋がりを評価する．

言語 言語評価は,「可読性」,「妥当性」,「多様性」で評価される.「可読性」は語学教師が推測によって辛うじて読めるというところから，一定のレベルに上がる度に点数が上昇する.「妥当性」は単語のスタイルや使われ方が妥当かどうか,「多様性」は語彙などの言語知識を評価する．

[3]https://goodwriting.jp/wp/system-flowcharts

3.2 注釈の詳細

全ての作文は全 6 人のアノテータの中からランダムに選ばれた 3 人によって点数が付けられ，3 人の点数の中央値が最終的な作文の点数となっている．全てのアノテータは日本語母語話者であり，作文採点の経験がある．表 1は，各アノテータが付けた点数と最終的な作文の点数の間のコーヘンのカッパ係数を表す．6 人の平均したカッパ係数は 0.56 であり，これは中程度の一致を意味する．

4 作文自動評価システム

整数値の点数を予測する作文自動評価タスクは，回帰を用いて点数を実数値で予測した後に閾値を設定し，整数値に変換する回帰タスクと各整数値の点数をクラスとしたクラス分類タスクのいずれかに置き換えることができる．本研究で作成した作文自動評価システムは，回帰タスクとしてこの問題に取り組んだ．回帰タスクに置き換える利点として，以下の 2 点が考えられる．1 つ目は，最先端の作文自動評価システムが回帰モデルを使用しており，そちらの方が高い精度が望めること [9]．2 つ目は，偏りがある少数のデータセットに対して，クラス分類モデルよりも回帰モデルの方が過学習しないと考えられることである．本研究では，2 つの素性を使用した回帰モデルと 1 つのニューラルネットワークを使用した回帰モデルを作成した．

4.1 素性を使用した手法

本研究では，先行研究で提案された素性 [10] と jWriter [4]，GoodWriting Rater[4]の素性を使用した．予測用のモデルとして，先行研究でも使用された線形 SVR（Support Vector Regression），線形回帰を使用した．

4.2 ニューラルネットワークを使用した手法

BERT は，Transformer を利用した汎用言語表現モデルであり，マスクされた言語モデルと隣接文予測の 2 つのタスクで訓練されている．BERT は近年作文の自動評価にも利用され始めた [11]．図 1に作成した BERT モデルの構造を示す．BERT モデルでは，作文のお題と本文が符号化され，入力される．符号化する際に冒頭に [CLS] トークンが挿入され，お題の終わりと本文の終わりにそれぞれ [SEP] トークンが挿入される．符号化された本文の長さがモデルが読み取れる系列長を超えた場合，本文の最大系列長を超えた部分は削除される．図のように Transformer Encoder に入力された

[4]https://goodwriting.jp/wp/system-ml

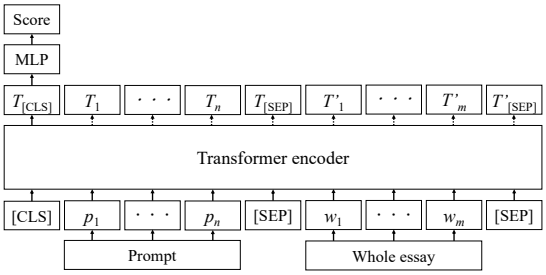

図 1: BERT を使用した自動作文評価システム．あらかじめ与えられるお題のトークンを p で表し，作文のトークンを w で表している．また，n はお題のトークン数，m は作文のトークン数を表す．Transformer Encoder の出力である埋め込みベクトルは T と T' で表される．

系列は，対応する系列として出力される．隠れ層において [CLS] トークンに対応するベクトル $T_{[CLS]}$ はお題と作文の対表現を表す．最終的な実数値は $T_{[CLS]}$ のベクトルを MLP（Multi Layer Perceptron）を使って集約することで得られる．図中の Transformer Encoder は MLP を訓練するときに同時に再訓練が行なわれる．

5 実験

5.1 評価指標

本研究では，評価指標として，重み付きカッパ係数（QWK: Quadratic Weighted Kappa）を使用した．QWK は作文自動評価の評価指標として一般的な指標であり，予測値が実際の値に近いほど小さいペナルティを与え，違いが大きいほど大きいペナルティを与える指標である．本研究で作成した作文自動評価システムでは，実数値を出力し，閾値を決めて整数値とする．最終的な整数値を決定する際には，開発データにおいて QWK が最大になるような閾値を設定した．全てのモデルのスコアは 5 分割の交差検証によって計算し，5 つのシード値の平均を最終的なスコアとした．

5.2 実験設定

素性を作成する際の形態素解析器として MeCab（ver. 0.996）[5]を使用し，辞書として IPAdic（ver. 2.7.0）を使用した．素性を利用したモデルのハイパーパラメータの最適化には，optuna[6]を使用した．

BERT モデルでは，BERT の base アーキテクチャ [5] を使用した．BERT の言語モデルは訓練に時間がかかるので，SentencePiece を利用した日本語訓練済みの BERT[7]を使用した．この訓練済みモデルは公開されている日本語訓練済み BERT の中で最も長い最大シーケンス長（512）に対応している．モデルの語彙サイズは 32,000 であり，学習率を 5e-5，バッチ数を 4，最大シーケンス長を 512 とした．全ての層でドロップアウトを使用し，ドロップアウト率を 0.1 に設定した．

5.3 実験結果

それぞれのモデルのスコアを比較し，結果の分析を行なった．表 2はそれぞれのモデルの QWK スコアを表す．平均の列は全ての評価項目におけるスコアの平均値を表す．結果より，BERT モデルが構成以外の項目で最も高くなった．

QWK スコアにおいて，全体評価の項目と他の項目（例えば内容の項目）では，最大 0.12 程度の差が生まれた．これは，全体評価で使用したデータ数が 800 以上であり，他の項目では 150 程度であるためであると考えられる．

5.4 分析

素性を使用した手法 表 2では，SVR を使用したモデルが構成の項目で最も高い QWK スコアを記録した．また，他の素性を利用したモデルも BERT モデルより高いスコアを記録した．素性を使用したモデルのスコアが BERT モデルより高い理由を分析するため，SVR の全体評価と構成評価における重みの絶対値を順位付けし，比較を行なった．

全体評価と構成評価において，重要度上位にある素性の多くは作文の長さに関係するものであった．一般的な作文自動評価システムでは，作文の長さと点数に相関があることが知られている [12]．構成評価の重みの中で特徴的であった素性は「段落数」であり，負の重みが重要度の上位に来ていた．適切な段落で区切られた作文は，序論，本論，結論を評価する構成評価では高い評価を得る傾向がある．元データの分析を行なうと，ほとんどの作文は 6 段落程度で構成されていたが，10 以上の段落で構成される作文も存在した．10 段落以上で構成され，少しの文しか書かれていない作文は 1 点や 2 点など低い構成スコアが付けられていた．以上のことから，不必要な段落の分割は構成スコアの低下を招くことがわかった．BERT モデルのスコアが低かった原因は，ニューラルネットワークを使用したモデルは段落数のような数値情報を正確に取得できないためであると考えられる．

[5]https://taku910.github.io/mecab
[6]https://github.com/pfnet/optuna
[7]https://github.com/yoheikikuta/bert-japanese

モデル名	全体	内容	構成	言語	平均
SVR	0.479 ± .0049	0.389 ± .0268	**0.583 ± .0081**	0.579 ± .0147	0.508 ± .0197
Linear	0.528 ± .0010	0.384 ± .0219	0.554 ± .0164	0.589 ± .0079	0.514 ± .0176
BERT	**0.627 ± .0008**	**0.497 ± .0187**	0.516 ± .0107	**0.623 ± .0171**	**0.566 ± .0151**

表 2: 作成した作文自動評価モデルの QWK スコア

ニューラルネットワークを使用した手法　BERT モデルは他のモデルと比べて良いスコアを出しているが，問題点も存在する．図 2は作文の全体評価の点数と BERT モデルで予測された点数の混同行列を表している．図に示すように，予測の誤りのほとんどは ±1 に収まっているが，1 点と 6 点と予測された作文はほぼ存在しない．このように，BERT モデルは平均に近い点数を予測してしまう傾向がある．これは，素性を利用したモデルと異なり，各ベクトルの影響がエンコーダによって緩和されるため，点数を大きく変化させることが難しく，損失関数である二乗誤差が低くなるように学習しているためであると考えられる．また，最大/最小のスコアがついたデータが少ないというのも原因の 1 つであると考えられる．

この問題を解決する手法として，2 通り考えられる．1 つ目は，少ないデータセットでも過学習せずに予測ができるようなシステムを開発することである．2 つ目は，新たなデータセットを作成することである．作文の自動評価で利用可能なデータセットはとても少なく，複数項目のデータはさらに少ないので，もっと多くのデータセットを作成する必要がある．

6　おわりに

本研究では，日本語学習者を対象とした作文能力自動評価システムを開発した．開発したシステムは，作文全体の点数だけでなく，作文の内容，構成，言語の点数も予測する．複数のモデルを作成し，比較を行なった結果，BERT を利用したモデルが最も良いスコアを記録した．作成した BERT モデルの出力を確認するため，実際に作文を入力すると，BRET モデルが素性を利用したモデルと比べて頑健であることがわかった．

謝辞

データセットを作成した，GoodWriting グループの諸氏に感謝する．本研究の一部は JSPS 科研費（研究課題番号: 26284074）の助成を受けたものである．

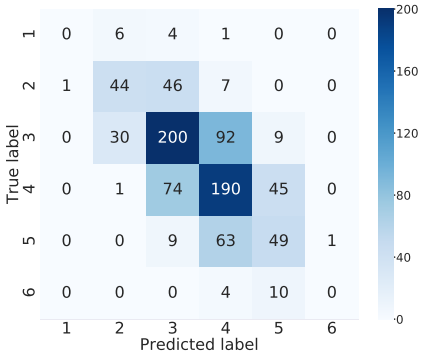

図 2: 全体評価における実際の点数と BERT モデルで予測された点数の混同行列

参考文献

[1] Isaac Persing and Vincent Ng. Modeling prompt adherence in student essays. In *ACL*, 2014.

[2] Isaac Persing, Alan Davis, and Vincent Ng. Modeling organization in student essays. In *EMNLP*, 2010.

[3] 田中真理, 久保田佐由利. 日本語のアカデミック・ライティングに規範は必要ないか :「構成」面の分析に基づく提案. In *CAJLE*, 2016.

[4] 李在鎬, 長谷部陽一郎, 迫田久美子. 人工知能の仕組みを利用した学習者作文評価システム「jWriter」–I-JAS を利用した試み–. 2017 年度日本語教育学会秋季大会予稿集, 2017.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[6] Mark D Shermis and Jill C Burstein. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge, 2013.

[7] Tsunenori Ishioka and Masayuki Kameda. Automated Japanese essay scoring system based on articles written by experts. In *ACL*, 2006.

[8] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *EMNLP*, 2016.

[9] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. In *IJCAI*, 2019.

[10] Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. Task-independent features for automated essay grading. In *BEA*, 2015.

[11] Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. Automated essay scoring with discourse-aware neural models. In *BEA*, 2019.

[12] Mark D Shermis and Jill C Burstein. *Automated essay scoring: A cross-disciplinary perspective*. Routledge, 2003.