D-43

小論文自動採点のための文書力スコア推測

中本さや香，嶋田和孝
（九州工業大学情報工学部知能情報工学科）

1 はじめに

高校・大学入試などにおける小論文問題では一般的に手作業で採点が行われる．しかし，手作業での採点は莫大な時間がかかる．また，すべての答案に平等な評価をするのは大変難しいのが現状である．

そこで，小論文に対する自動採点が注目されている．2021 年に竹内らによって小論文データセットが開発された [1]．このデータセットには理解力・妥当性・論理性・文書力の 4 つの評価観点における採点結果が付与されている．理解力・妥当性のスコアに関する研究は発表されているが，論理性や文書力のスコアに関する研究はまだ進んでいない．これら 4 つの評価観点の中でも文書力スコアは，評価基準が小論文のテーマや課題に大きく左右されない．また，他評価観点に比べ具体的な基準が多いため，それらを特徴量に反映することで高精度の自動採点システムの開発が可能だと考える．

本研究では機械学習を用いた文書力スコアの推測システムを開発し，その結果について報告する．

2 データセット

前述の竹内らのデータセットを使用する．データセットには，計 23 問の問題文，ルーブリック，答案・採点結果などが提供されている．表 1 からわかるように，今回着目する文書力スコアは字数及び誤字数に加え，文章の質も評価の対象となっている．

実験データの設定は先行研究 [1] に倣う．具体的にはモデルの学習及び評価に利用する正解スコアは先行研究で用いられた信頼できる二名の人手採点結果の平均値である．先行研究が実験で排除している答案は同様に排除する．加えて，我々の分析で，ルーブリックと採点結果が一致しない答案が存在することが分かったため，それらも排除している．

3 提案手法

本研究では，以下に挙げる特徴量を結合してランダムフォレスト回帰モデルに入力し，文書力スコア (1 点〜5 点) を予測する．

- 文字数割合
- 漢字含有率 (漢字の文字数/答案文の長さ)
- 誤字数 (誤字の個数)
- 漢字の画数ベクトル

ここで，文字数割合は式 1 を適用して変換する．式 1 において len は小論文答案文の長さ，lim は各問題の文字数制限を表す．

$$contentLen = \tanh(\frac{len}{lim} - 0.7) \quad (1)$$

文字数割合，漢字含有率，誤字数はそれぞれ 1 次元ずつの特徴量である．画数ベクトルは 1〜36 画に対応させるため，36 次元となっている．各要素には文中にその画数の漢字が表れた個数が入る．以上の特徴量を結合した 39 次元のベクトルを正規化し，モデルへの入力とする．

表 1. 文書力評価ルーブリック

点数	評価基準
5	誤字なし．指定字数をほぼ埋めている． 論理的で無駄のない文
4	誤字 1 個．指定字数をほぼ埋めている．読みやすい文
3	誤字 2 個．わかりやすい文章
2	指定字数の 85%以下．誤字が多く冗長な文
1	指定字数の 60%以下．誤字が多く冗長な文

表 2. 文書力スコアの予測精度

使用特徴量	QWK
特徴量全て	**0.913**
文字数割合を除く	0.385
漢字含有率を除く	0.911
誤字数を除く	0.910
画数ベクトルを除く	0.904

4 実験

評価指標として QWK (重み付きカッパ係数) を導入し，モデルの出力した文書力スコアを評価する．QWK は，確率分布を利用したランダムな予測をするモデルに対してどの程度良いモデルであるかを示す．ランダムな予測モデルと同程度であれば 0 に近づき，良いモデルであれば上限値 1 に近づく．Landis and Koch [2] によるカッパ係数の評価目安によると，0.81〜1.00 はほとんど完全であると記されている．

本研究では特徴量の要素それぞれの有効性を確認するため，各要素を削除した特徴量を入力としたモデルで実験を行う．それぞれ 5-fold 交差検証をした結果を表 2 に示す．

4 つの特徴量すべてを使用したモデルが最大の精度を示しており，本手法の有効性が確認できた．QWK は目安となる 0.81 を大きく上回ったため，本手法の精度は十分高いといえる．

また，ルーブリックに明記されている文字数割合は削除すると顕著に精度が低下したため，重要な特徴量であると考えられる．漢字含有率や漢字の画数はルーブリックに明記されている特徴量ではないが，特徴量から削除したときに精度が低下した．漢字の多さや，画数が多く難しい漢字を使うことが文の質を上げ，採点に寄与した可能性がある．

5 おわりに

本論文では，日本語小論文データセットを用いた文書力自動採点モデルを作成した．特徴量を用いることで文書力スコアを求めることができ，文書力の自動採点化に期待が持てる結果となった．今後はその他の評価基準に関するモデル構築についての研究も進めていきたい．

参考文献

[1] 竹内孔一, 大野雅幸, 泉仁宏太, 田口雅弘, 稲田佳彦, 飯塚誠也, 阿保達彦, 上田均：“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”, 情報処理学会論文誌, **62**, 9, pp. 1586–1604 (2021).

[2] J. R. Landis and G. G. Koch: “The measurement of observer agreement for categorical data”, Biometrics, **33**, 1, pp. 159–174 (1977).