2G-06

教員評価が類似するレポートの特徴分析

山本　恵†

名古屋外国語大学†

1. はじめに

ルーブリックに基づく自動採点システムを提案し構築している[1]．現段階では総合得点を算出するための分類精度が低いため，評価項目の改善を目指している．6 名の教員によるレポートの手動採点結果を分析したところ，不要な評価項目が明らかになった．これらを除く評価項目については自動採点で用いている評価項目と総合得点との相関を確認できた．さらに，総合得点が低いレポートは教員間の評価が概ね共通しており，総合得点が高いレポートは評価値に差があることが明らかになった[2]．

そこで本研究では，評価値が類似するレポートを分析して特徴を明らかにし，自動採点システムの改善案を示す．

2. ヒューマンスコア収集方法

ヒューマンスコア収集にあたっては，2000 字以上を条件に出題された学生レポート 25 件を対象とした．A4 用紙 2〜3 ページ程度である．基本統計表を表 1 に示す．

表 1　分析対象レポートの基本統計量

分析対象レポート	25	件
本文文字数平均	2284.3	文字
本文平均抽出語数	1318.6	語
本文平均異なり語数	336.1	語

大学教員 6 名に対し，提案しているルーブリック（表 2）に基づく手動採点を依頼し，採点結果と使用したルーブリックの評価項目に関する意見を収集する．できるだけ条件を統一するために 1 時間に 3 件程度を目安に採点している．

5 つの評価観点について 0-10（10 点満点）で評価する．ルーブリックのマイルストーンは 5 段階に分けて記述してある．例えば評価観点[Content]では，「記述内容が，課題とは無関係である」を 0（ゼロ），「的確な解答である．改善の余地はない．」を 10 として，スコアを入力する．図 2 に示すように，細分化した評価項目を，学生に示すべき改善点として平易な表現に変えてある．採点時にこれらの評価項目に該当すると気づいた場合はチェックする．

表 2 レポートの採点に使用したルーブリック※

評価観点	D(0)		A+(10)	学生に示す改善すべき点
[Content] 課題の理解度と記述(解答)内容の妥当性	記述内容が，課題とは無関係である．	～	的確な解答である。改善の余地はない．	論題と内容が合致していない
				キーワードなど主要な関連語が含まれていない
				出題意図を理解していない
				内容に妥当でない部分がある
				授業などの学修内容が反映されていない
[Structure] 論理的な展開	記述内容にまとまりがなく，何を言おうとしているかわからない．	～	意見・主張があり，論理の展開に矛盾がない．説得力があり，改善の余地はない．	順序立てて述べる
				全体の構成を再考する
				意見・主張を持つ
				事実と意見を区分けする
				説明に矛盾がある
[Evidence] 資料や根拠(エビデンス)の妥当性	根拠を示していない．	～	根拠の内容や示し方が妥当であり，改善の余地はない．	資料・文献の水準が低い
				関連性のない資料・文献を参照している
				根拠に妥当性が見られない
				図表の説明がない
				根拠が十分でない
[Style] 文章作法の遵守と適切な推敲	文章が全く推敲されていない．誤りが多い．	～	文章作法を遵守し，よく推敲してある．改善の余地はない．	文体が統一されていない
				誤字・脱字がある
				文章のねじれがある
				二重否定がある
				修飾語と被修飾語が離れている
				主語と述語が対応していない
				句読点を効果的に用いていない
				冗長な表現がある
				表記ゆれがある
[Skill] 読みやすさ・表現の巧みさ	文章が読み辛い．明らかに文章スキルがない．	～	文章が読みやすく，表現が巧みで，語彙が豊富である．改善の余地はない．	漢字の使用が適切でない
				文章が長すぎたり短すぎたりする
				語彙が少ない
				分り辛い表現がある
				適切な言葉に置き換える

※表 2 ではマイルストーンの一部を省略している

3. ヒューマンスコアの分析結果

表 3 は各採点者（A〜F と表記）の総合得点一覧である．濃い網掛けは各採点者の低いスコア 10%を示す．薄い網掛けは高いスコア 10%である．低いスコアは比較的共通しているが，高いスコアは採点者ごとにバラバラであり，高く評価する観点が教員により大きく異なることがわかる[2]．

4. 評価が低いレポートの特徴

表 3 の結果から，複数教員が下位 10%に含まれる評価をしているレポート No. 2, 5, 19 を低評価レポートの代表，複数教員が上位 10%に含まれる評価をしているレポート No. 13, 19, 21 を高評価レポートの代表とする．

Feature Analysis of Essays with Similar Evaluation Scores
†Megumi Yamamoto・Nagoya University of Foreign Studies

表 3 採点結果の総合得点一覧

ﾚﾎﾟｰﾄNo	採点者					
	A	B	C	D	E	F
1	45	29	17	29	20	30
2	15	25	9	28	33	36
3	48	45	17	28	13	34
4	**49**	49	24	13	25	40
5	23	25	10	5	18	20
6	42	48	15	22	12	37
7	**50**	48	16	16	11	40
8	30	32	**27**	7	16	35
9	35	**50**	9	16	17	29
10	23	34	13	0	8	33
11	44	**50**	25	26	25	32
12	43	41	22	19	36	39
13	48	**50**	**38**	22	23	42
14	41	42	20	20	30	33
15	32	31	11	9	9	40
16	47	41	15	11	23	31
17	33	**50**	23	14	28	38
18	39	33	16	14	22	27
19	47	**50**	17	**37**	30	39
20	46	48	12	**36**	17	34
21	47	48	24	28	**42**	**43**
22	46	38	12	7	10	**43**
23	45	49	16	8	26	33
24	36	45	16	7	**39**	35
25	42	45	11	17	22	41

表 4 は教員の改善指摘箇所を比較したものである．低評価レポートの共通した特徴の１つとして，網掛け箇所があげられる．文体が統一されていない文書は，スキルがかなり低く評価される．また内容やレポートの構成ができていると，高い評価が得られることがわかる．すなわち全般的に渡り評価が類似する．

表 4 改善指摘箇所の比較

	改善指摘数	
評価項目	低評価レポート	高評価レポート
適切な言葉に置き換える	10	7
全体の構成を再考する	9	0
分り辛い表現がある	9	9
内容に妥当でない部分がある	8	4
文体が統一されていない	8	0

5. レポート類似度の自動採点への導入可能性

文体の誤りについてはすでに自動採点の評価項目として計算しているが，前節の結果を受け総合評価を算出する際の重みづけを調整することとする．全体の構成や内容に妥当でない点を自動採点に反映させる方法は，レポートのテーマにより異なる可能性が高く自動計算するのは困難である．そこで一案として低評価レポートとされたデータを蓄積し，それらとの類似度を計算した結果を参入する可能性を調査した．

図 1 は KH Corder 3 を用いた階層的クラスター分析による文書の類似度を示したものである．併合水準（非類似度）を確認したところ，クラスター数 5 より非類似度が高くなることから，クラスター数 6 としている[3]．●は前節の低評価レポートで，3 件のレポートは隣接している．一方○は高評価レポートで，ここでは類似度がみられない本結果は低評価レポートに分類する際の自動計算で，蓄積した低評価レポートを活用する可能性を示唆している．

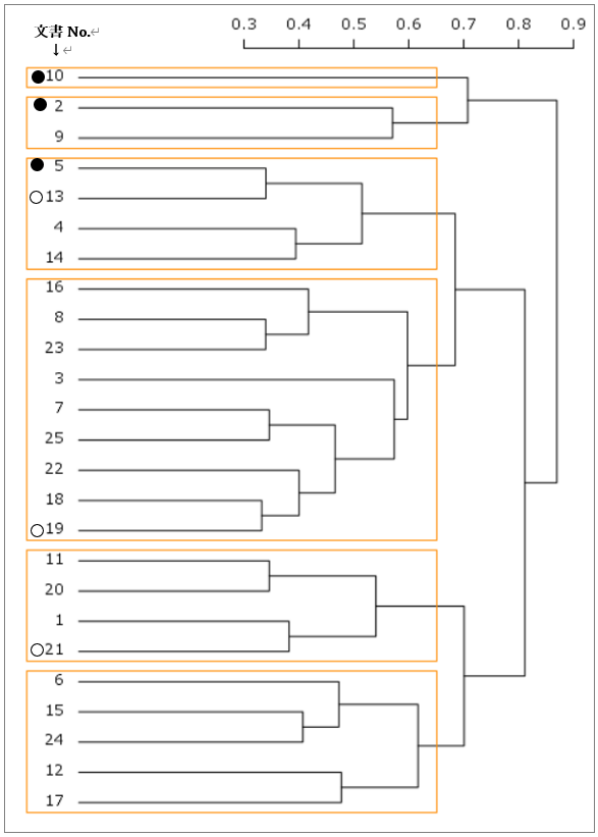

図 1 レポート間の類似度

6. まとめと今後

低評価レポートは採点結果が文書スキル，内容，全体構成など全般に類似する傾向がある．文書の構成など自動計算し辛い部分は，蓄積したデータとともにクラスタリングすることで低評価の位置付けができる可能性があり，今後検証を進めたい．

謝辞

本研究は JSPS 科研費 18K11589，17K00432 の助成を受けたものである．

参考文献

[1] 山本恵，梅村信夫，河野浩之：レポート自動採点プラグインの開発と評価，Proceedings of Moodle Moot Japan 2017 Annual Conference，pp.16-21 (2017).
[2] 山本恵：ルーブリックに基づくレポート自動採点システムの評価項目の改善,情報処理学会 第 84 回全国大会抄録 (2022).
[3]樋口耕一：R を用いた多変量解析と可視化 https://khcoder.net/scr_r.html（最終閲覧日:2023 年 1 月 10 日）