

ループ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

山本 恵†

名古屋外国語大学†

1. はじめに

レポート作成指導，教員の採点の厳正化，負担軽減を目的として，ループリックに基づく自動採点システムを構築している．具体的には5つの評価観点からなる教員の手動採点用ループリックを作成し，それらを細分化した評価項目を設け，評価項目ごとに自動採点するものである．またこれらの自動採点結果からサポートベクターマシン（SVM）により総合成績レベルの分類を行っている[1][2]．現段階では分類精度が55%前後にとどまり低い．これは分類器作成のための学習データ件数が少ないこと，細分化した自動採点評価項目の中に精度が高くない採点方法があることが原因であると推測し，これまで計算方法の見直しや語彙水準計測のための辞書の構築などを進めてきた．さらに別の要因として，評価項目そのものの必要性や重みづけをあらためて検討する必要があると考える．

そこで本研究では，教員の手動採点結果（ヒューマンスコア）をあらためて収集・分析し，評価項目の改善を目指すこととした．その結果，評価項目と採点結果である総合得点との相関を確認できた．しかしながら教員により，評価時に重視する項目がかなり異なること，総合得点に差異が見られるなどの知見が得られたので報告する．

2. 自動採点モデルの概要

採点システムの中核となるループリックは全米カレッジ・大学協会（AAC&U）が作成したバリユーループリックをもとに作成している．はじめに手動採点向けに5つの評価観点（Contents：課題の理解度と解答内容の妥当性，Structure：論理的な展開，Evidence：資料と根拠の妥当性，Style：文章作法の遵守と適切な推敲，Skill：読みやすさ・表現の巧みさ）を持つレポート採点用ループリックを作成し，これらの評価観点を機械で処理するために，25項目に細分化し自動採点用ループリックを作成した．

図1にその一部を示す．細分化した評価項目のうち機械による計算に置き換えているのは図中の網掛けの項目のみである．Structure, Evidence については，レポートの題材や教員の主観が大きく影響すると考えられる．またContents についても，授業達成度や課題提示の目的によって異なるため，現段階では論題とレポート内容の合致度のみコサイン尺度を用いて算出している．[Style]や[Skill]はレポートの読みやすさに大きくかわることから，レポートの最終評価に大きく影響すると閑雅へ，これらの評価項目の計算結果からA+, A, Bなどの総合成績レベルを分類（予測）するものである．

評価視点	採点方法	評価項目
[Content] 課題の理解度と解答(記述) 内容の妥当性	一部自動	1 論題と記述の合致度
	手動採点	2 主要な関連語の存在
15 文体の統一性		
16 誤字・脱字の排除		
17 構文の妥当性		
18 主述関係の妥当性		
19 句読点の妥当性		
[Evidence] 資料と根拠(エビデンス) の妥当性	自動採点	20 冗長さ・二重否定の排除
[Style] 文章作法の遵守と適切な推敲		21 表記ゆれ・曖昧さの排除
		22 漢字の使用率
[Skill] 読みやすさ・表現の巧みさ		23 文長の妥当性
		24 語彙の豊富さ
		25 語彙の水準

図1 評価観点を細分化した評価項目(抜粋)

3. ヒューマンスコア収集方法

大学教員6名に対し，提案しているループリックに基づく手動採点を依頼し，採点結果と使用したループリックの評価項目に関する意見を収集する．学生レポートは25件で，平均2100文字（A4用紙2～3ページ）である．25件について1時間に3件程度を目安に採点する．

5つの評価観点について0-10（10点満点）で評価する．ループ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．

評価項目		手動採点時の表現 (学生へ提示)
Style	15 文体の統一性	文体が統一されていない
	16 誤字・脱字の排除	誤字・脱字がある
	17 構文の妥当性	文章のねじれがある
	18 主述関係の妥当性	二重否定がある
	19 句読点の妥当性	修飾語と被修飾語が離れている
	20 冗長さ・二重否定の排除	主語と述語が対応していない
Skill	21 表記ゆれ・曖昧さの排除	句読点を効果的に用いていない
	22 漢字の使用率	冗長な表現がある
	23 文長の妥当性	表記ゆれがある
	24 語彙の豊富さ	漢字の使用が適切でない
	25 語彙の水準	文章が長すぎたり短すぎたりする

図2 手動採点時に示した評価項目(抜粋)

4. ヒューマンスコアの分析と考察

図3は、5つの評価観点の合計を総合得点とし、採点時にチェックされた改善すべき項目数の合計との相関を求めた結果である。相関係数は0.8323と高く、改善すべき点が多いほど、得点が低いことは明らかである。したがって細分化した評価項目は概ね妥当であると言える。

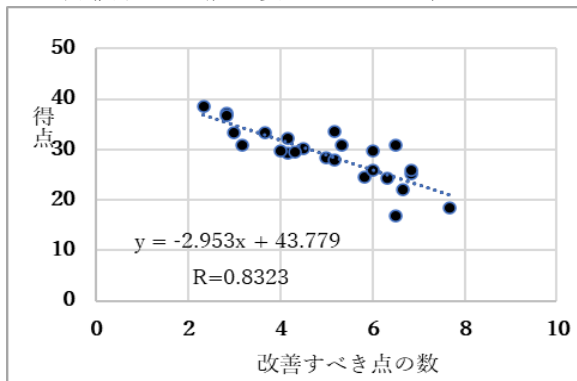


図3 総合得点と改善すべき点のチェック数との相関

表1は、チェック数が多い上位15項目を示したものである。[Style]と[Skill]の項目が多く、特に読みやすさにかかわる[Skill]項目はすべて上位に含まれている。最もチェックが多かったのは「適切な言葉に置き換える」である。「語彙が少ない」「分かり辛い表現がある」も含め、自動採点では「語彙の豊富さ」と「語彙水準」を計算して求めている。実際に教員が採点する際に言葉の選び方を重視することから、「語彙の豊富さ」と「語彙水準」の計算方法を再検討する必要がある。また最も少なかったのは「二重否定がある」である。採点終了後のアンケートにこの項目と「修飾語と被修飾語が離れている」が必要ないとの回答があった。しかし読みやすさ、あるいは内容の正確な把握に重要であると考え、当面この項目は排除せず、今後の検討課題とするにとどめる。

表2は各採点者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。

表1 評価観点 Style, Skill の細分化した評価項目

評価観点	学生に示す改善点	チェック数合計
[Skill]	適切な言葉に置き換える	82
[Skill]	分かり辛い表現がある	56
[Style]	誤字・脱字がある	52
[Content]	内容に妥当でない部分がある	45
[Evidence]	根拠が十分でない	44
[Style]	冗長な表現がある	41
[Style]	句読点を効果的に用いていない	38
[Structure]	全体の構成を再考する	33
[Evidence]	資料・文献の水準が低い	32
[Skill]	文章が長すぎたり短すぎたりする	32
[Style]	主語と述語が対応していない	32
[Skill]	語彙が少ない	30
[Structure]	説明に矛盾がある	29
[Style]	文体が統一されていない	29
[Skill]	漢字の使用が適切でない	23

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

表2 総合得点一覧

しているが、高いスコアは採点者頃にバラバラであり、教員ごとに高く評価する観点が大きく異なることがわかる。

5. まとめと今後

評価項目と採点結果である総合得点との相関を確認できた。しかしながら教員により、重視する項目がかなり異なることがわかり、今後評価項目をさらに精査する必要がある。

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, 梅村信夫, 河野浩之: LMS 上の自動採点システム構築による自由記述文評価の取り組み事例、平成 29 年度 ICT 利用による教育改善研究発表会資料集, pp.154-157 (2017).