日本語小論文自動採点システムの評価指標の改善に関する研究

楊華†　　藤田彬‡　　前川眞一††　　田村直良‡‡
†．††東京工業大学大学院社会理工学研究科
‡横浜国立大学大学院環境情報学府　　‡‡横浜国立大学大学院環境情報研究院
†you.k.aa@m.titech.ac.jp　‡actica@tamlab.ynu.ac.jp
††mayekawa@hum.titech.ac.jp　‡‡tam@ynu.ac.jp

1. はじめに

本研究では，評定値が順序尺度であることを前提とする日本語小論文自動採点システムを作成し，その中で用いられる評価指標の改善を試みる．

近年，学校教育や会社採用など様々な場面で小論文を書かせることが増加してきた．その理由としては，自由記述式問題が多肢選択式問題より書き手の能力，思考や表現などを多角的に評価することが可能であると考えられるからである．従って，自由記述式問題の答案も多肢選択式問題の答案のようにスピーディー，効率的かつ公平に評価することが求められてきている．このような目的のために利用可能な日本語小論文自動採点システムは既に存在するが，そこで用いられている各小論文の特徴を表す評価指標の優劣について様々な議論がされている.また,それらのシステムでは，評定値は連続量であると仮定されているが，実際に数段階の評価尺度（順序尺度）であることが多い．

そこで本研究では，特殊な判別分析である多項ロジットモデル分析を用いて小論文の採点を行う．そのためにまず，評価指標を集め，選別し，評価指標セットを定める．そして，評価済みの小論文のデータを，教師データと評価データの 2 セットに分け,クロスバリデーションを行う.すなわち，教師データを用いて多項ロジットモデルの係数を推定し，それを評価データに適用してそこに含まれる小論文の評定値の予測値を算出する．

2. 評価指標

既存の小論文自動採点システムの評価指標（features）は多くのシステムで未公開である．本研究では，公開されている評価指標を取り入れ，日本語小論文を評価するための指標を4つの観点から合計76項目にまとめた．詳細は以下のとおりである．

I.　基本情報（37）
II.　基本文体と表現（4）
III.　構造（11）
IV.　内容（24）

2.1 基本情報

文章の基本情報とは文章の表層的な特徴情報のことを指している．文字数，段落数，各品詞の数などの指標以外に，Jess[1]で「修辞」を評価するために用いられている指標，語彙密度[2]とシャノンの情報量（エントロピー）等の合計37項目がある．情報源全体の不確実性を測る尺度であるエントロピーは文章 X の情報量 $H(X)$ を以下のように算出する．ただし，x_i は文章に含まれている i 番目の文字列（セグメンテーション単位）である．

$$H(X) = -\sum_i P(x_i)\log_2 P(x_i)$$

2.2 基本文体と表現

小論文は作文と異なり，書き手は客観的に事象を判断し，その判断の正当性の根拠を挙げて主張する文章である．本研究では，以下の4つの文体と表現の仕方が小論文を判断する基準になるべきであると考える．

I.　口頭的表現（逆項目）
II.　主観的主語（逆項目）
III.　（能動態より）受動態
IV.　である調

2.3　構造

順序良く議論を展開していくことで，読み手の理解に助け，文章は分かりやすくなる．Jess[1]では，接続表現を検出することで，文章の論理構造の評価をしている．また，藤田ら[3]は文章の論理構造を多角的に評価するため，文章の構造解析を行った．本研究では，接続表現と構造解析の両方から小論文の構造を評価する．

2.4　内容

小論文自動採点システムであるE-rater,Intelligent Essay Assessor(IEA)とJessは書かれている小論文がトピックとの関連性を判断するときに，コサイン類似度を用いている．コサイン類似度の計算には，文字列単位と文字列特徴指標の2つの要素が必要である．本研究では，文字列単位は形態素とバイグラムの2つ，文字列特徴指標はタームの出現頻度とタームの重みの2つがあると考える．つまり，コサイン類似度のパターンは4つがある．

I.　形態素×出現頻度
II.　形態素×重み
III.　バイグラム×出現頻度
IV.　バイグラム×重み

本研究では，この4通りのコサイン類似度の妥当性を検証することをサブ目的としている．

小論文データを教師データと評価データの2セットに分け，教師データを定めたカテゴリに沿ってグルーピングをする．すべての小論文は各カテゴリに属している小論文群とのコサイン類似度を計算する．すなわち，1つの小論文に対して（4パターン×カテゴリ数）個のコサイン類似度指標が得られる．

3. 分析および結果

本研究では，宇佐美[4]が収集した「小学校の授業における，英語の早期教育は必要であるか否かに対して意見を述べよ」（課題A）と「グラフと説明文を読み，日本人の子育ての態度に関してどのような特色が読み取れるかに関して述べよ」（課題B）の2種類の課題，合計594編の小論文を用いて課題別に分析を行った．課題ごとに小論文を教師データと評価データの2セットに分けた後，

表 1　小論文データの使い分け

		教師データ	評価データ	合計
課題	A	250 編	44 編	295 編
	B	250 編	50 編	300 編

Perl で評価指標の抽出プログラムを作成し，594編の小論文の特徴を表す指標を抽出した．

3.1　カテゴリの策定

小論文は4人（専門家2人，国語教師2人）により10点満点で採点されている．国語教師2人の採点結果の相関は高く（課題Aは0.77，課題Bは0.75），その合計点をとり元評定値（連続量）として使用する．本研究では，評定値は連続量ではなく順序尺度（カテゴリデータ）とみなすため，課題別に小論文の評定値がほぼ正規分布になるように4つのカテゴリ(A,B,C,D)を定めた．

3.2　評価指標の選別

76項目の評価指標の中には類似する項目があるため，評価指標間の相関関係から指標を選別した．相関係数は0.8以上の項目をまとめ，代表的な項目を残すアプローチを用いた結果，相関の高い評価指標セットは14個であり，選別した結果は残った評価指標は49項目であった．課題Aと課題Bはほぼ同じ結果が得られた．主成分分析と因子分析を行い，その妥当性を検証した．

相関の高い評価指標セットには以下の2つがある．これらをセットαとセットβと呼ぶ．

セットα

No.51：カテゴリAとのコサイン類似度（形態素×出現頻度）
No.55：カテゴリBとのコサイン類似度（形態素×出現頻度）
No.59：カテゴリCとのコサイン類似度（形態素×出現頻度）
No.63：カテゴリDとのコサイン類似度（形態素×出現頻度）

表 2　（課題 A）セット α の相関行列

	No.51	No.55	No.59	No.63
No.51	1			
No.55	0.983	1		
No.59	0.957	0.978	1	
No.63	0.986	0.985	0.967	1

表 3 （課題 B）セット α の相関行列

	No.51	No.55	No.59	No.63
No.51	1			
No.55	0.989	1		
No.59	0.987	0.990	1	
No.63	0.965	0.980	0.973	1

セットβ

No.53：カテゴリAとのコサイン類似度（バイグラムー出現頻度）

No.57：カテゴリBとのコサイン類似度（バイグラムー出現頻度）

No.61：カテゴリCとのコサイン類似度（バイグラムー出現頻度）

No.65：カテゴリDとのコサイン類似度（バイグラムー出現頻度）

表 4 （課題 A）セット β の相関行列

	No.53	No.57	No.61	No.65
No.53	1			
No.57	0.968	1		
No.61	0.958	0.978	1	
No.65	0.900	0.854	0.810	1

表 5 （課題 B）セット β の相関行列

	No.53	No.57	No.61	No.65
No.53	1			
No.57	0.992	1		
No.61	0.980	0.991	1	
No.65	0.954	0.971	0.969	1

課題Aと課題Bとも，評価指標No.51，No.55，No.59，No.63は相関係数が1近くであり，評価指標No.53，No.57，No.61，No.65も相関係数が0.80以上である．理由としては，各カテゴリの小論文で出現頻度の高い文字列が類似していると考えられる．課題Aの各カテゴリで出現頻度の高い形態素を表6で示す．バイグラムも同様である．

表 6 (課題 A)各カテゴリで出現頻度の高い形態素

カテゴリ	A	B	C	D
1	英語	英語	英語	英語
2	こと	する	する	する
3	する	こと	こと	こと
4	思う	の	いる	の
5	いる	いる	の	教育
6	教育	教育	思う	いる
7	なる	思う	なる	なる
n	…	…	…	…

つまり，出現頻度を用いて計算されるコサイン類似度がカテゴリと関係なく，意味のない内容評価指標である．この8項目の評価指標を削除することにした．コサイン類似度を計算するときにTF・IDF法で計算される重みを用いるべきだといえる．

3.3 多項ロジットモデル分析（MLA）

評定値が順序尺度であるとき，採点という作業は各小論文がどのカテゴリに属するかを判断することになる．統計では「判別」と呼ぶ．カテゴリは4つがあるので，特殊な判別分析である「多項ロジットモデル分析」を用いる．

多項ロジットモデル分析は分析対象を各カテゴリに判定される確率を計算し，確率の高いカテゴリに判定する．x_i という属性を持った対象 i がカテゴリ j に判定される確率 π_{ij} を次のように表す．

$$P(y_i = j | x_i) = \pi_{ij}$$

4つのカテゴリのうち，3つが決まれば4つ目は自然に決まるので，本研究ではJ＝1を基準とする．各カテゴリに属する確率の計算は以下のように定式化できる．

$$\pi_{ij} = \frac{\exp(x_i' \beta_j)}{\sum_{r=1}^{4} \exp(x_i' \beta_r)} \cdots\cdots (\beta_1 = 0)$$

教師データを用いて多項ロジットモデルの係数を推定し，ステップワイズ法でさらに変数選択をする．選別された評価指標が課題Aは7項目，課題Bは4項目であった．

表 7 （課題 A）MLA で選別された評価指標

No.	評価指標
No.19	K特性値
No.37	である調以外の語尾（逆項目）
No.39	語彙密度
No.41	主張文の個数

No.54	カテゴリAとのコサイン類似度（バイグラム－重み）
No.56	カテゴリBとのコサイン類似度（形態素－重み）
No.60	カテゴリCとのコサイン類似度（形態素－重み）

表 8　（課題 B）MLA で選別された評価指標

No.	評価指標
No.54	カテゴリAとのコサイン類似度（バイグラム－重み）
No.58	カテゴリBとのコサイン類似度（バイグラム－重み）
No.62	カテゴリCとのコサイン類似度（バイグラム－重み）
No.66	カテゴリDとのコサイン類似度（バイグラム－重み）

こちらの評価指標の組み合わせで教師データのフィットが非常によく，課題Aの誤判別率が0.8%，課題Bの誤判別率が0.4%であった．しかし，推定した係数を評価データに適用して算出できた評定値の予測値は評定値とのズレが見られた．課題Aにおいて，正確に判定されるのが22.73%であり，カテゴリが1つずれているのが45.45%である．モデルのAICは53.67である．課題Bにおいて，正確に判定されるのが38%であり，カテゴリが1つずれているのが50%である．モデルのAICは30である．

一般的に，頑強性を考慮するため，変数選択を行う．しかし，教師データにおいてはあまりにも判別率が良すぎるためか，変数選択の機能がうまく働いていないと考えられるため，本研究では，変数選択せずに係数を推定し，評価データの予測値を算出し，クロスバリデーションを行った．その結果は表9と表10で示す．

表 9　（課題 A）評価データのクロスバリデーション結果

		評定値（人間の採点結果）			
		A	B	C	D
予測値	A	0	1	2	1
	B	2	10	6	0
	C	2	1	11	4
	D	0	2	1	1

表 10　（課題 B）評価データのクロスバリデーション結果

		評定値（人間の採点結果）			
		A	B	C	D
予測値	A	1	1	1	2
	B	3	11	5	1
	C	0	9	8	2
	D	1	1	1	3

4. まとめ・今後の展望

本研究では，多項ロジットモデル分析を用いて日本語小論文自動採点のモデルを提案した．

石岡[1]によると E-rater は GMAT の小論文の採点に実用されている．ひとつの答案は人間とコンピュータが独立に採点し，その結果，得点差が 6 点満点中 2 点以上あった場合に別の人間の評定者が最終的な得点を決定する．すなわち，コンピュータと人間の間で 2 点差が許容範囲である．本研究ではカテゴリが 4 つであり，1 点差が許容範囲であると考える．49 項目の評価指標のモデルを使用する場合，課題 A は 1 点差以上の評価データが 7 件であり，全体の 16%である．課題 B は 1 点差以上の評価データは 6 件であり，全体の 12%である．この結果，ほぼ実用に供せるシステムを構築することの可能性が示されたと考えられる．

今後，変数選択のプロセスを踏んでからモデルの妥当性を検証するため，より多くの小論文データの収集が必要とされる．評価指標の吟味も今後の課題として挙げられる．

謝辞

本研究は株式会社ベネッセコーポレーションの「研究テーマ設定型 E(Educational)奨学金」のもとで行われました．厚く御礼申し上げます．

研究で用いたデータの使用を許可してくださった東京大学大学院教育学研究科の宇佐美慧氏に感謝の意を表します．

参考文献

[1] 石岡恒憲：記述式テストにおける自動採点システムの最新動向.行動計量学会誌,Vol.31，No.2（通巻 61 号）,pp.67-86,2004.

[2] 佐野大樹,丸山学彦：システミック文法に基づく書きことばの複雑さ測定.言語処理学会,第 14 回年次大会,発表論文集,pp.1097-1100,2008.

[3] 藤田彬，五條善雅，田村直良：文章構造解析に基づく小論文の自動評価.言語処理学会,第 14 回年次大会,発表論文集,pp.576-579,2008.

[4] 宇佐美慧：未公開論文