オンライン自由記述問題の解答に対する即時フィードバックのための自動評価手法

Automatic evaluation method for immediate feedback on answer to online free description problem

生田 寛[*1], 槇原 竜之輔[*1], 杉谷 賢一[*1], 久保田 真一郎[*1], 中野 裕司[*1]
Kan IKUTA[*1], Ryunosuke MAKIHARA[*1], Kenichi SUGITANI[*1], Shin-Ichiro KUBOTA[*1], Hiroshi NAKANO[*1]
[*1] 熊本大学
[*1]Kumamoto University
Email: c4709@st.cs.kumamoto-u.ac.jp

あらまし：オンライン小テストの記述問題は解答が多様なため自動的に評価して，フィードバックを返すことが困難である．本研究では，解答に含めるべき文が不定の自由記述問題に対して過去の解答例をもとに潜在的意味解析を用いて学習者の解答文を自動評価できる仕組みを提案する．

キーワード：自動評価，フィードバック，潜在的意味解析

1. はじめに

LMSの小テスト機能の記述式問題では解答が多様であり，自動評価が困難で，採点者は評価に多大な負担を要す．中島の研究(1)は，文字数制限があり解答に含めるべき文が定まる記述問題において，複数の採点済みの解答を学習データとし，機械学習手法を用いて自動評価のための自動識別を試みている．その結果，相応の識別力を確認したが，判定を失敗する解答文があることを確認している．このように，解答に含めるべき文が定まっていたとしても，その自動識別は困難となっている．石岡らの研究(2)では，新聞の社説やコラムを学習データとし，文章の構造，論理構成，内容をもとに，小論文を自動採点するシステムを開発している．このシステムは，800字から1600字程度の小論文に対して有効であるとされ，短文で解答するような記述式問題では，構造や論理構成の評価ができるとは限らない．また，小論文の内容を潜在的意味解析手法により評価しており，内容判定の問題点として，学習データにない文章の内容をうまく評価できないことを指摘している．

2. 目的

中島の研究では解答に含めるべき文が定まっていない自由な記述が許される問題設定の場合，学習データに存在しない文章を含む可能性が高く，十分な結果が得られないと考えられる．また，石岡らの研究では，潜在的意味解析手法による内容判定は特定のテーマが設定された記述問題であれば，短文であっても判定に有効と考えられる．そこで，本研究では，オンライン記述問題のうち，特定のテーマ設定がされた短文の自由記述問題に対して自動評価する手法について提案する．

3. 提案手法

研究対象となる問題は，特定のテーマ設定がされた短文の自由記述問題である．次の手順によって自動評価する手法を提案する．

(1)研究対象とする自由記述問題の過去の答案に，評価者がフィードバックの必要があると判定する文(悪い文)と正解とする文(良い文)とに分類する．

(2)評価者が良い文と判定した文章を使って単語文書行列を作成し，潜在的意味解析手法により要約された単語文書行列 X を生成する．

(3)評価対象となる解答文を単語文書行列と同じ手法でベクトル V とする．

(4)要約された単語文書行列 X の各行と評価対象となる解答文のベクトル V との類似度を求める．

(5)良い文は高い類似度，悪い文は低い類似度を示すと考えられるため，類似度をパラメータとして適切な閾値を設定し，良い文かを判定する．

4. パラメータ決定の検証実験と評価実験

特定のテーマに関する短文の自由記述問題として，2017 年 10 月から 12 月に開講された授業で実施されたオンライン記述式問題の解答文を対象として検証実験および評価実験を行った．図 1 に問題文を示す．

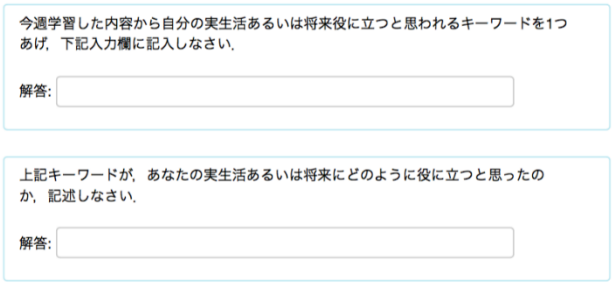

図 1 オンライン記述式問題の問題文

図 1 に示すように，まず学習したキーワードを答える問題が設けられ，そのキーワードに従って自由記述問題が設けられている．評価者が評価した結果，キーワード W_1 に対して良い文 25 個，悪い文 31 個があった．これらのデータをもとに交差検証を行い，パラメータを決定した．良い文を 5 個ずつの 5 組にわけ，5 組から 2 つの組を選び，2 組の 10 個の文章から潜在的意味解析を行った単語文書行列 X を作成し，行列に使用しなかった 3 組の良い文 15 個，悪い文 31 個の中から選出した悪い文 15 個の合計 30 個をテストデータとした．テストデータの 1 個 1 個と行列 X の各行との類似度をそれぞれ算出し，算出される 10 個の類似度をもとに，悪い文か，良い文かを評価する．10 個の類似度の代表値として平均値，最大値，最小値，中央値を扱い，閾値を 0.05 刻みで変化させ，それぞれの代表値が閾値の類似度より低い場合，悪い文と評価する．評価者が悪い文と判定した解答文を，類似度の代表値によって悪い文と判定できたか否かにより精度の指標である F 値を求めた．この F 値の算出を 5 組から 2 つの組を選ぶ全ての組み合わせ(10 パターン)で行い，各パターンで F 値が最大となる代表値を扱うことにした．今回の実験では，最大値と中央値の両方の値を扱う場合に，各パターンで F 値が最大となったので，判定の指標として，類似度の最大値と中央値を扱うことにした．次に，閾値を決定するために，最大値を指標として閾値を 0.05 刻みごとに変化させ 10 パターンの F 値を積み上げた値をプロットし，その値が最大となる 0.6 を閾値とした．また，中央値を指標とした場合についても同様にプロットし，その値が最大となる 0.5 を閾値とした．これらの検証実験によって，評価対象の解答文のベクトル V と単語文書行列 X の各行との類似度から，その最大値および中央値を使い，最大値の閾値を 0.6，中央値の閾値を 0.5 として判定を行うこととした．

この提案手法の有効性を確かめるために，別のキーワードの解答文をもとに評価実験を行った．キーワード W_2 をもとに解答された解答文は，良い文 15 個，悪い文 74 個であった．15 個の良い文をもとに単語文書行列 X を生成し，その他の悪い文すべてに対して提案手法による判定を行った．その結果を表 1 に示す．本提案手法により悪い文 74 個のうち 69 個を悪い文と判定でき，5 個を判定できなかった．

表 1 キーワード W_2 の悪い文 74 個の評価

	悪い	悪くない	再現率
最大値と中央値	69	5	0.93

5. まとめ

テーマが設定された短文の自由記述問題に対する過去の解答文をもとに単語文章行列を作成し，潜在的意味解析手法により生成される新たな単語文書行列を使って類似度を求め，類似度をもとに良い文と悪い文を判定する手法を提案した．パラメータを決める検証実験をもとに閾値を決定し，提案手法の妥当性を検証する実験を行い，高い再現率を実現した．しかし，正しく判定できなかった解答文もあり，これらの解答文に対してさらに考察することで，さらに再現率を向上することができると考えている．

参考文献

(1) 中島 功滋，機械学習を利用した短答式記述答案の自動識別，日本教育工学会 第26回全国大会，pp639-640，2010 年

(2) 石岡 恒憲，亀田 雅之，コンピュータによる小論文の自動採点システム Jess の試作，計算機統計学 第 16 巻 第 1 号，pp3−19，2003