短答式記述答案の採点支援ツールの開発と評価

中島　功滋
ベネッセコーポレーション／教育テスト研究センター
nakaji@mail.benesse.co.jp

1　はじめに

短答記述式の解答形式は学校現場では広く用いられている一方で，選択式解答と比較すると採点に多くの人的・時間的コストを必要とする。また，大規模テストの記述式解答採点では複数採点者間で採点結果の不一致が生じることが問題となりうる。採点結果の不一致を少なくし，大量の答案を効率的に採点するためには記述式解答に対してコンピュータを利用した自動採点・採点支援を行うことが有用である。

自動採点・採点支援の先行研究 [4, 7] では，テキストデータを入手しやすいレポートや小論文が研究の対象となっていた。学校教育現場の多くが紙と鉛筆による試験を行っている現状では，テキスト処理を行う前に，手書きの答案をテキストデータに変換する必要がある。しかし，今後コンピュータを使用した試験の普及や入力デバイスの進歩によって，受験者がキーボードや手書きで入力した記述解答が直ちにテキストデータとして扱える環境が一般的になるのはそう遠くないと考える。

また，様々なテキストマイニングのソフトウェアが販売もしくは無償で頒布されており (例えば KH Coder[2]) 簡単な操作で様々な分析を行うことができる。しかし，これらのツールを使用するには分析手法の理解や目的に合った分析手法の選択等，使用者にある程度の知識が必要であり，採点支援にそのまま使用できるとは言い難い。

このような状況を踏まえ，本研究では開発中の短答式記述答案のコンピュータによる採点支援ツールを報告する。多くの機能は一般的なテキストマイニングツールと重複するが，問題作成者・試験採点者が使うことを想定し，GUI による操作と正答例を直接入力して答案と照合する機能を実装した。

本稿では，まず本研究が支援する記述答案採点処理を規定し，次にツールの各機能，特に解答例との照合機能を説明する。最後に実データへの適用例を紹介し，ツールの有用性について議論する。

2　ツールの開発

2.1　短答式記述答案採点の特徴

本研究ではいわゆる「正解」が存在する短答式記述問題（読解・知識・説明等）を扱う。このような問題の採点では正解と意味的に等価な内容が答案に書かれているかどうかの照合が作業の中心となる。

試験の実施に当たって、採点の不一致やあいまいさを回避するために,(特に大規模試験や複数採点者で採点する場合）あらかじめ採点基準（正答例・誤答例・判断基準等）が作成されるのが一般的である。しかし，この採点基準が十分に精緻化されていなかったり，採点者の訓練が不十分な場合，採点基準の追加変更と採点済み答案の再採点を余儀なくされる場合がある。

また，このような記述問題採点では正答・誤答の判断だけではなく，記述内容に応じて中間点や部分点を与えることが一般的である。「正解」との照合の結果，中間点や部分点を与える場合の判断基準は各設問に依存している。典型的なものを挙げると，

文字・文法 誤字・脱字・文法的誤りに対する減点

語句 定められた語句が含まれていれば加点（なければ減点）

複数照合 複数の解答例と答案を照合し，最も近い解答例に付与されている得点を与える

部分照合 正答例の一部分と同じ内容が答案に書かれていればその部分に応じた得点を与える（欠落していればその分減点する）

実際の採点基準は上記のような基準を組み合わせて作成されていると考えられる。

2.2　GNU R による実装

一般利用者向けのツールの開発においてはツール本来の機能と GUI 両方の実装が不可欠である。本研究

ではこれらを GNU R[1]（以下 R）上で実装した。

R は統計パッケージとして一般に認知されているが，プログラミング言語として利用することも可能である。R 言語ではベクトルや行列等の操作・計算を直感的に記述することができる。全ての処理手続きを自らコーディングする必要はなく，各種解析用の追加パッケージが CRAN ミラーサイト等で公開されている。R で実装・実行が難しい処理や高速処理が必要な場合は，system() 関数を使用して外部プログラムを実行し，実行結果を R 内で処理することで対応できる。

GUI に関しては，R のグラフィックス機能と R の標準構成に含まれる tcltk パッケージ (tk ツールキットのラッパー) を使用することで作成できる。

R は主要な OS(Windwos, MacOSX, Linux 等) で使用できるので，実装時に OS 間の差異をほとんど意識する必要がない。また，これまでは様々プログラムの組み合わせで行っていた処理や分析を (基本的に)R 言語のみで記述できることが開発効率上有利な点である。

本ツールでは各機能を実現するために以下の R 拡張パッケージと外部プログラムを使用した。RMeCab[3]（形態素解析），RCaBoCha（係り受け解析），e1071（サポートベクターマシーン），topicmodels（LDA:Latent Dirichlet Allocation），bayon[2][1]（クラスタリング）

2.3 本ツールの主な機能

本稿ではツールの機能のうち，答案と解答例の照合機能（複数照合と部分照合）に焦点を当てて紹介する

2.3.1 BLEU を利用した照合機能

BLEU[6] は機械翻訳の自動的評価の指標として提案された。機械翻訳と人手による翻訳（参照訳）において n-gram 正解率の幾何平均と短い機械翻訳に対するペナルティ (BP) を元に算出される。

$$BLEU = BP \cdot \left(\prod_{i=1}^{n} p_i \right)^{\frac{1}{n}} \quad (1)$$

$$BP = \exp(1 - \max\{1, r/c\}) \quad (2)$$

p_n は n-gram 正解率,r は参照訳の長さ，c は機械翻訳の長さである。

本ツールでは，機械翻訳を受験者の答案，参照訳を採点者が用意する解答例とみなし照合を行う。通常 BLEU の利用においては，n-gram 正解率をすべての機械翻訳と対応する参照訳から算出し，システムや手法全般の評価値として１つの値を返す。一方，採点支援においては各答案に対する照合結果が必要である。そこで，１つの答案と１つの解答例から n-gram 正解率を算出し，BLEU を計算して照合のスコアとして使用する。

図１は照合機能のスクリーンショットである。利用者は画面上部のテキストエリアに解答例を入力し，n の値（幾何平均を構成する ngram）を問題に合わせて指定する。計算を指示すると，解答例に対して形態素解析などの前処理を行い，各答案ごとの BLEU を計算し答案とともに画面下部に出力する。複数照合が必要な場合は，利用者は照合したい解答例を複数入力し計算を指示すると，BLEU が最大値になる解答例を照合結果として各答案ごとに出力する。部分照合が必要な場合は，利用者は画面上で区分けされた領域それぞれに解答例の部分文字列を入力[3]すると，答案と部分文字列から BLEU が算出され，照合箇所ごとに値が出力される。

BLEU は 0 以上 1 以下の値をとり，大きいほどよいとされる。しかし，出力値を解釈する際は，BLEU が答案及び解答例双方の文字列長に影響される指標であることに留意する必要がある。もっとも，本研究が対象とする短答式記述答案では解答字数制限のある設問が多く，このような設問については，

> 制限を越えるような長い答案や解答例は存在しない
>
> 短い答案を誤答扱いにする採点基準が作成されることが多い

という特徴があるため，他の種類のテキストより文字列長の影響は少ない可能性がある。

また，BLEU が 0 になる答案は,（n が適切に設定されていれば）どの解答例にも全く合致していないと解釈できる。意味的に等価な正答例を十分用意して計算することで，誤答答案を検出できる可能性が高い。

2.3.2 その他の機能

前処理　前処理として記述答案データに対して RMeCab による形態素解析を行う。利用者は解析対象にする品詞・ベクトルの素性・出力値を指定して解析を実行する。ベクトルの素性は以下のものが選択可能である。

文字 n グラム ($1 \leq n \leq 7$)

[1] http://cran.md.tsukuba.ac.jp/ (CRAN 国内ミラー)
[2] 外部プログラムとして利用
[3] 図１では領域は３箇所で各領域に複数の解答例が入力可能である

図 1: 照合機能の画面例

形態素 n グラム ($1 \leq n \leq 3$)

文節 n グラム ($1 \leq n \leq 2$)

文節 n グラムは RCaBoCha の機能を使用し，通常のバイグラムに加えて，係り－受けの関係を持つバイグラムを抽出することができる。出力値は 2 値，頻度，TF/IDF から選択する。

素性の一覧を利用者が確認し，形態素の分割が不適当と判断した場合は，ユーザー辞書に登録し再解析することで登録した語を単独の形態素として切り出すことができる。

解析結果はＲオブジェクトの要素として登録し，後の解析に使用する。

グラスタリング　クラスタリングによって類似答案をグループ分けできる。利用者はクラスタ数とクラスタリング手法を指定して分類を実行する。クラスタリング手法は Repeated Bisection(bayon が実装したアルゴリズム，高速) か LDA(トピックモデル) から選択する。画面上で各クラスタごとの答案を閲覧することができ，解答傾向の把握が容易になることが期待される。

得点予測　答案が採点済みであったり，過去の採点済み答案がある場合，これらを機械学習の訓練データとして利用し，残りの答案の得点予測を試みる。利用者は学習に用いる行列と機械学習の手法 (SVM, NaiveBayes) を選び実行する。先行研究 [5] で以下の 3 点を確認している。

予測の分類正解率は問題に依存し 50%から 90%程度である

訓練データのランダムリサンプリングによる多数決で，予測が一貫している答案（全体の約 3 割）について 90%以上の分類正解率が得られる

問題に応じてベクトルの素性を考慮することで分類正解率を上げることができる

3　実データへの適用例

本稿で焦点を当てた照合機能の適用例として，国語の記述答案データに適用した結果を報告する。

3.1　データの概要

データは，高校 3 年生向け通信教育教材に同梱されている模擬試験の採点済み短答式記述解答である。受講生が自宅等で解答・返送した答案に採点が行われたものの一部を電子化した。本稿では国語の 1 設問への適用例を報告する。設問は本文内の下線部を説明するよう求めたもので，解答は 50 文字以内で記述するよう求められ，正答に含まれる 3 つの内容（以下観点 A, 観点 B, 観点 C）がすべて記述されていれば正答で，それぞれ欠落している場合は減点される。

得点分布より典型的な得点パターンとして 8 点（満点：すべての観点を満たす），6 点（観点 B のみ満たさない），3 点（観点 A, 観点 C いずれか一方のみ満たす），0 点（誤答：すべての観点を満たさない）の 4 種類を抽出した。抽出総数は 2419 件（内訳，8 点 55 件，6 点 820 件，3 点 1343 点，0 点 201 点）であった。

3.2　適用手続き

満点の答案にはすべての観点について意味的に等価な記述が含まれていると考えられる。そこで，満点の

答案の文字列から各観点に対応すると思われる文字列を選び，解答例として入力し，BLEU を計算した。計算の結果，満点であるにもかかわらず BLEU の値が 0 の答案がある場合は，その答案から記述の一部を意味的に等価な表現として解答例に追加し，BLEU を再計算した。満点の答案ほとんどの BLEU が非ゼロになるまでこれを繰り返した。同様の操作を総ての観点について行い，３つの観点における BLEU 値を算出した。BLEU の計算には形態素ユニグラム，形態素バイグラムを用いた。

3.3 結果と考察

観点 A に関して 4 種類，観点 B について 9 種類，観点 C について 3 種類の意味的等価表現を入力した。

表 1 は BLEU 値が 0 より大きい観点の組み合わせと採点済みの得点とのクロス集計結果である。

表 1: BLEU 値の組み合わせと得点分布

	得点				
BLEU>0	8 点	6 点	3 点	0 点	計
A,B,C [8]	52	24	8	0	84
A,C [6]	2	613	132	2	749
A xor C[3]	1	165	1063	32	1261
なし [0]	0	9	74	60	143
その他	0	9	66	107	182
計	55	820	1343	201	2419

*[ ] 内の値はパターンに対応する得点

BLEU のパターンと実際との得点に関して，今回は 8 点における再現率をできるだけ高くなる (0.95) ように操作している。他の得点の再現率は，0.75(6 点), 0.79(3 点), 0.30(0 点) で，精度は，0.62(8 点), 0.82(6 点),0.84(3 点), 0.45(0 点) であった。全ての得点を込みにした分類正解率は 0.73 であった。

本ツールの適用によって，判断が難しい 3 点と 6 点の答案に対して良好な精度と再現率が得られた。また，0 点の精度や再現率が低いが，BLEU の計算にバイグラムまでしか使用していないため，0 点の答案でも部分的に合致することで，BLEU が 0 より大きくなりやすかったと考えられる。

照合結果と得点が一致していない答案を精査することで，問題ごとのチューニングを行ったり，人手による採点の精度の確認などに利用できると考える。

4 まとめ

本ツールは開発段階であり，指標の有用性及びユーザーにとっての利便性双方の評価・検証が不十分である。様々な記述答案データに対して評価を行い，ツールの完成度を高めるのが今後の課題である。

謝辞

本研究の一部は日本学術振興会科学研究費補助金（挑戦的萌芽研究）「課題番号 20650150」の助成を受けて行った。また，ツールの実装に関しては青山学院大学大学院生の鈴木龍一氏の協力を得られた。ここに感謝の意を表する。

参考文献

[1] 藤澤瑞樹. 軽量データクラスタリングツール bayon. http://code.google.com/p/bayon/.

[2] 樋口耕一. テキスト型データの計量的分析— 2 つのアプローチの峻別と統合. 理論と方法, Vol. 19, No. 1, pp. 101–115, 2004.

[3] 石田基広. R によるテキストマイニング入門. 森北出版, 2004.

[4] 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, Vol. 16, No. 1, pp. 3–18, 2003.

[5] 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 回全国大会講演論文集, pp. 639–640, 2010.

[6] K Papineki, et al. A bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of ACL*, pp. 311–318, 2002.

[7] 椿本弥生, 赤堀侃司. 主観的レポート評価の系列効果を軽減するツールの開発と評価. 日本教育工学会論文誌, Vol. 30, No. 4, pp. 275–282, 2004.