短答記述式問題解答文の採点支援システム JS^4 の試作

亀田 雅之 † 石岡 恒憲 ‡ 劉 東岳 ¶
†‡大学入試センター 研究開発部 {†付} ¶学研教育総合研究所
† m-kamedax@nifty.com ‡ tunenori@rd.dnc.ac.jp ¶ DL.Liu@mmf.gakken.co.jp

1 はじめに

大学入試センター試験に代わり、2020 年度に始まる新共通テストでは、記述式問題が導入される [1]。しかし、解答文が 40~80 字の案が出ているが、出題や採点に関する詳細は検討中である (2017 年 1 月現在)。

記述式問題の解答の採点では、「正解のない」エッセイについて、自動評価採点システム [2, 3] が先行しているが、「正解」に対する自動採点は、近年、取り組まれ始めたところである [4, 5]。後者では、正解と解答の意味関係や含意関係レベルの照合が必要となるが、このための技術は、「東ロボ」プロジェクトでの歴史科目の自動解答の取り組み [6] や筆者 (石岡) らが進めている科研費研究 [7] でも、未だ技術的な課題/限界が多いことが明らかになってきている。

こうした背景のもと、我々は短答式記述問題の解答文の採点システム JS^4（Japanese Short answer Soring Support System) の Web 上での検討・試作に着手した。当初は、自動採点システムとはせず、システムの提示する採点支援情報や採点案を採点者が採点を確認・変更した上で最終確定する採点支援システムとした。

新共通テストでは、記述式を国語と数学に導入することになっているが、いずれの科目であっても、特有の困難な課題が多いため、現時点では比較的取り組みやすい社会科の記述式問題を対象にした。

我々は、前報 [8] で「機械学習 (RandomForest :RF[9]) による採点予測」を中心に報告した。本稿では、「採点基準に基づく採点」に重点を置いて報告する。

2 サンプル問題

試作では、2015 年度の「学研全国総合模試」で出題された地理 B, 日本史 B1/B2, 世界史 B2 の計 8 問と 30-60 字程度の各々100 程度の解答文とその採点結果をサンプルとして用いた。本稿では、世界史 B2-3 の問題とその解答文 (総数 102 文) を例にして説明する。

世界史 B2-3 では、7 世紀から 11 世紀にかけてのイスラーム世界に関する 500 字程度の記載があり、その一部の「アッバース朝の租税制度」の記載部分を引用した記述式解答を求める問 5 を対象とした。

世界史 B2-3 問 5

アッバース朝の時代には，具体的にどのような変更が加えられたのか。次の語句を必ず使用して，６０字以内で記述せよ。[ジズヤ、ハラージュ]

3 採点支援システム JS^4

3.1 システム構成

図 1 に JS^4 のシステム構成図を示す。

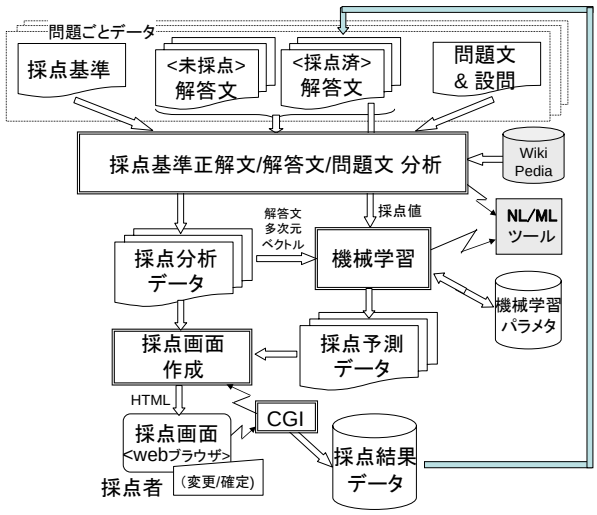

図 1. JS^4 システム構成図

JS^4 では、問題に応じて、対象解答文そのものとともに、解答文と問題文/設問、採点基準 (3.2) に記載の模範解答文などとの相互間について、文字列処理や言語処理ツール、機械学習ツールなどを活用・分析し、その結果を採点参考情報や採点案として、採点者が操作する採点画面（図２）に次のように表示する。

・中央の採点基準の表には、基準項目ごとに計算した「適合度」と、適合度に基づき判定した基準項目の該当・非該当を「チェックボタン (Y/1・N/0)」にデフォルトで設定・表示し、

・下部の表には、採点案として、配点と該非/適合度で計算した「チェック得点」/「適合度得点」とともに、採点付き解答文の多次元特徴ベクトルの機械学習結果を用いて予測した確率を付与し、最大予測確率の得点をデフォルトで設定した「得点決定ボタン」と予測確率に基づいた「期待値得点」を表示する。

図 2. JS[4] 採点画面

採点者は、採点基準項目ごとの「適合度」も参考にして、「チェックボタン」で該非を変更し、「再計算」ボタンで「チェック得点」を再計算することができる。

最終的に、「適合度得点」や「期待値得点」も参照しつつ、「チェック得点」、得点ごとの予測確率を参考に、「得点決定ボタン」で設定した得点を「確定」し、次の解答文の採点に移る。最大確率のデフォルト得点に同意するなら、直ちに得点を確定することもできる。

3.2 採点基準と採点基準ファイルの形式

記述式解答を求める設問には、模範解答文（正解）/配点などからなり、採点者が採点する際に拠り所とする「採点基準」がある。その「採点基準」に準じ、適宜、必要な情報を追加/拡張し、システムが参照するために形式化した「採点基準ファイル」を設計した。

図 2 は、世界史 B2-3 の設問 5 の一つの解答文の採点画面であるが、中央の表の「採点基準ファイルの内容」欄に「採点基準ファイル」の内容を確認できる。

0) syno **同義置換文字列**

必ずしも明示されていないが、解答文中に現れる同義語、言い換え可能な表現等を同等に扱うため、標準表現と置換対象文字列の正規表現との対で指定する。

1) gold **模範解答文とその配点**

いわゆる「模範解答 (文)」の記載で、通常、満点が配点される。複数指定がある場合は、最もよく適合している模範解答文による得点を採用する。

2) part{G} **部分点解答文とその配点**

いわゆる「部分点」を与える解答内容の記載である。「採点基準」に明示はないが、同一観点ごとに同じ G でグループ化し、最もよく適合する部分点解答文の得点を採用し、排他的な各グループの部分得点を合算する。

3) lack{G} **必須語とその欠落に応じた配点 (減点)**

解答文に指定の必須語が欠落している場合、負の配点により減点となる。G は「part{G}」の G と同様にグループとして扱う。「採点基準」では、語以外に、句や複数の同時欠落の指定もあるので、必須文字列の正規表現の and 並びの指定とした。

4) vol **制限文字数と違反に対する配点（減点）**

(原) 解答文の字数制限の指定であり、制限違反の場合、負の配点により減点となる。「採点基準」では、最大字数の指定だけだが、最小字数の指定も可能とした。

5) nons **無意味文に対する配点 (減点)**

解答文が意味不明の場合の減点の配点を指定する。通常、満点分の減点を指定する。

6) goji **誤字 1 字あたりの配点（減点）**

誤字 1 字あたりの負の配点により誤字数分の減点となる。手書き解答の場合、人手入力や文字認識の過程で誤字扱いとなる文字が対象となる。判読不能文字「=」をデフォルト設定としている。

7) type **解答文型と違反に対する配点（減点）**

設問が、「reason(理由)」、「noun(名詞 (句))」の解答を要求する場合を想定し、追加した（図 2 にはない）。指定型に合わない場合、負の配点により減点となる。

3.3 採点基準の判定：適合度と該非初期値

実際の採点では、採点者は、採点基準項目ごとに判定した該非（1/0）と配点により項目ごとの得点を得、最終的に項目ごとの得点の集計で設問の得点を得る。

本システムでは、この作業支援として、項目ごと「チェックボタン」に自動判定した該非を初期値設定し、確認の上、同ボタンで修正できるようにした。

項目ごとの該非の自動判定は、高い精度が求められるが、現状では、解答文の採点基準の各項目に対して、下記のように計算した「適合度」に基づき、その項目の該非（1/0）を得て、初期値として「チェックボタン」に設定する。「適合度」は該非初期値の根拠として表示し、採点者が確認できる。

1) gold **模範解答文**， 2) part **部分点解答文**

採点者による判定では、解答文と模範解答文/部分点解答文との構文・意味的同一性や含意性まで踏み込んでいると思われるが、現状では、困難な課題である。そこで、まずは、表層レベルだが、文内のキーワード群同士の関連度を適合度と扱った（機械学習では、意味的同一性の指標として、wikipedia も使用して、gensim による文間一致度も特徴量に加えている [8]）。

ここでは、筆者が重要キーワード/重要文/関連文の抽出のために提案した、キーワード候補間の関連度の手法 (擬似キーワード相関法)[10] に基づき、文内のキーワード候補群間で定義した参照関連度と被参照関連度 [11] 及びそれらの調和平均 (F 値) を活用した。

解答文の適合度 (0~1) は、関連度の特性から、模範解答文に対しては、キーワード候補群間の「２関連度の F 値」、部分点解答文に対しては、部分点解答文のキーワード候補群が解答文のキーワード候補群によりどの程度再現されているかを示す「参照関連度」を使い分けた（参照/被参照関連度の関係は、再現率と正解率の関係と同じ）。該非 (初期値) は、当初、単純に適合度を四捨五入して、1(該当)/0(非該当) とした。

3) lack **必須語欠落**

指定された必須文字列正規表現の and 並び個数 n のうち、照合個数 m の割合を r (=m/n) とし、下記のように、適合度 R(=1-r) と R に基づく該非を設定する。

必須語 and 並び	照合割合 r	適合度 R	lack 該非
全てある	1	0	0 (非該当)
全てはない	0≤ r<1	1≥ R> 0	1 (該当)

4) vol **制限文字数**

syno 置換前の原解答文の文字数 m を計数し、制限最大字数 max を超えていたら m/max、制限最小文字数 min 未満であれば 1-m/min、それ以外は 0 を適合度とする。該否 (初期値) は、適合度 0 以外で 1(該当:字数制限違反)、適合度 0 で 0(非該当:違反なし) とする。

5) nons **無意味文**

通常、構文解析ツールは、解析を成功させるよう動作するため、無意味文の判定は難しく、充分な検討には至っていない。現状、原則、適合度は 0 とし、構文解析が失敗したり、係り受け関係が適切に得られないなどの場合に、適合度 1 とする。該否 (初期値) は、適合度をそのまま 1(該当:無意味文)、0(非該当) とする。

6) goji **誤字**

指定された誤字の字数を計数し、字数を適合度として設定する。該否 (初期値) は、適合度 1 以上で 1(該当:誤字あり)、適合度 0 で 0(非該当:誤字なし) とする。

7)type **解答文型違反**

「reason」、「noun」が指定されている場合、解答文の構文解析結果を参照し、文末文節の構成等により、指定型に整合した表現か否か判定し、非整合なら適合度 1、整合なら適合度 0 とし、そのまま該非 (初期値) を 1(該当:形式違反)、0(非該当:形式整合) とする。

3.4 採点基準に基づく採点計算

採点計算では、まず、採点基準項目ごとに { 配点 × 該非 (0/1)}（goji では、さらに {× 適合度 (=誤字数)}）で該非による得点 (減点) を得る。該非 (0/1) の代わりに適合度 (0~1) を用いると、適合度による得点 (減点) となる。さらに、採点基準の種別ごとに、以下のように計算を進める。

gold は、複数あった場合は一つでも該当 1 があれば、gold 配点の満点（あるいは最大得点）が得点となり、part は、同一グループごとに最大の得点を得た上で、排他的グループごとの得点の合計が得点となる。gold と part の両得点の最大値が正の得点となる。gold が 0 であっても、part で得点を得ることができる。

lack は、同一グループごとに最大減点を得て、排他的グループごとの得点の合計が得点 (減点) となる。lack 得点 (減点) と vol、goji、nons、type のそれぞれの得点 (減点) の合計が減点分の全体となる。

最終的に、正得点と減点分の合計を（負なら 0、満点を超えたら満点におさめた上で、）「チェック得点」あるいは「適合度得点」とする。

「チェック得点」は、最初は、適合度から自動判定した「チェックボタン」に初期設定された該非初期値による計算結果であるが、「再計算」ボタンにより、採点者が変更した「チェックボタン」の状況で再計算する。

4　考察

図 2 の採点画面で採点対象となっている解答文の実際の採点は 0 点であり、機械学習による最大確率の得点は 0 点と正解と合致している。(現状では、サンプル 8 問のどれも学習データが 100 文程度と少ないため、機械学習では過学習状態となり、最大確率では 100%近い正解率となる)

一方、「チェック得点」は 4 点と正解 0 点から大きく解離し、また、この問題の全解答文 102 文中での正解率も 14%、±1 点誤差内で 31%と極めて低い。これは、得点に大きく影響する gold と part の適合度がキーワードの表層レベルである他に、該非を単純に適合度の四捨五入で得ているのが要因と考えた。

図 3 は、適合度 (0~1) から該非 (0/1) を得る切上げ/切下げ閾値 (0.5~1) に対する正解率の推移グラフ [(下)±0（or 0.1 内) , (上)±1 内] に、さらに、1) 左端 (閾値 0 付近) に「適合度得点」の正解率、2) 閾値 1 の右側に機械学習 RF ツールが出力する正誤交差行列による正解率、3）右端に機械学習の確率付き得点予測による「期待値得点」と最大確率の得点の正解率 (≃100%) を加えた図である。

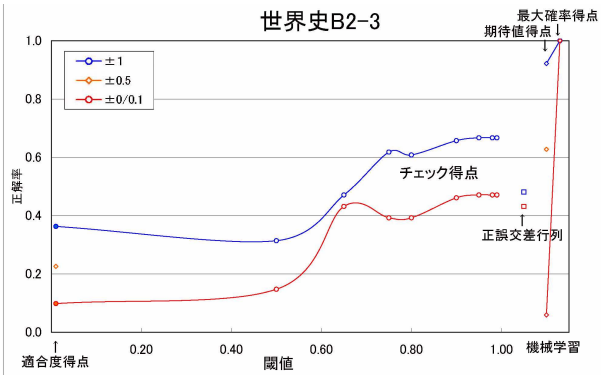

図 3. 採点案の正解率

「チェック得点」の正解率は、閾値が 0.9 なら、46%、1 点差内で 66%と大きく向上する。他の問題では、機械学習部分も含め、グラフの振る舞いに相違はあるが、閾値 0.9 は比較的よい正解率を導く。正解文と充分に一致して、ようやく該当とすべき、という直感に合う。

5　まとめ

新共通テストへの記述式問題の導入を控え、採点支援、さらに自動採点を念頭に、入手できたサンプル問題と解答文＋採点データを用いて「短答記述式問題解答文の採点支援システム JS[4]」の試作に着手した。

設問に付随する「採点基準」に沿った採点計算、採点データの機械学習による採点予測、採点作業の支援形態、自動採点などの全体枠組みや可能性を示したが、

1) 正解文と解答文との適切な照合技術

2) 採点基準ごとの該非データがない

3) 機械学習用の大量の採点付き解答文が得にくい

といった多くの課題を確認した。

特に、1）では、言語処理として、キーワード～構文レベルから、さらに意味/文脈理解レベル、含意関係認識に踏み込み、正解文と解答文との関係を考察した検討が必要である。

また、現在の採点作業の支援の上での離散的な得点とともに、自動採点システムでは、「適合度得点」などで示した連続値得点を与える途もありうると考える。

謝辞

本試作にあたり、サンプル問題一式及び貴重なコメントをいただいた学研グループに感謝申し上げます。

参考文献

[1] 文部科学省. 高大接続システム改革会議「最終報告」の公表について, Mar.2016, http://www.mext.go.jp/b_menu/shingi/chousa/shougai/033/toushin/1369233.htm

[2] J. Burstein, J. Tetreault and N. Madnani. The E-rater Automated Essay Scoring System, in Handbook of Automated Essay Evaluation edited by M. D. Shermis and J. Burstein, pp. 55–67, 2013.

[3] T. Ishioka and M. Kameda. Automated Japanese essay scoring system based on articles written by experts, Coling-ACL 2006, no. P06-1030, pp.233–240, 2006.

[4] 中島巧滋. 短答式記述答案の採点支援ツールの開発と評価, 言語処理学会, 第 17 回年次大会 発表論文集, pp611–614, 2011.

[5] 寺田凛太郎, 久保顕大, 柴田知秀, 黒橋禎夫, 大久保智哉. ニューラルネットワークを用いた記述式問題の自動採点, 言語処理学会 第 22 回年次大会 発表論文集, pp370–373, 2016.

[6] 狩野芳伸, 川添愛, 渋木英潔, 藤田彬. 「ロボットは東大に入れるか」 歴史科目の自動解答, 人工知能 Vol.31 No.5, pp813–819, 2016.

[7] 短答式記述テストにおける実用的な自動採点システムの開発, 研究課題:23650558, 科学研究費助成事業データベース

[8] 石岡恒憲, 亀田雅之, 劉東岳. 人工知能を利用した短答式記述採点支援システムの開発, 信学技報告, 言語理解とコミュニケーション研究会 (NLC), 2016.12.22

[9] Breiman L, Random Forests. Machine Learning, 45 (1), pp.5–32, 2001.

[10] 亀田雅之. 擬似キーワード相関法による重要キーワードと重要文の抽出, 言語処理学会 第 2 回年次大会 発表論文集, pp.97–100. 1996.

[11] 亀田雅之. 段落間及び文間関連度を利用した段落シフト法に基づく重要文抽出, 情報処理学会 自然言語処理研究会 121-17, 情報学基礎 47-9 , pp.119–126, 1997.