共通テストの試行調査国語記述解答データを用いた自動採点のアルゴリズムとその評価

○石岡 恒憲*，岡 知樹**，Nguyen Tuan Hung***，Nguyen Tuan Cuong***，中川 正樹***
*大学入試センター，**東京大学，***東京農工大学

1 はじめに

記述式試験は一般に多肢選択に比べ、いわゆる正当的（authentic）な学力を測ると広く考えられており、その採点が適切に、かつ受容可能な時間で処理できるなら、その潜在的な利用の可能性は測りしれない。そのため、エッセイタイプ問題に比べその採点が難しいとされる、短答式記述採点においても多くの研究（たとえば Alikaniotis,、2016 など）が行われている。わが国においても、予備校の模試での記述採点データが公開され、これを用いた研究（Mizumoto, 2019 など）がある。

本稿では、平成 29 年（2017 年）と平成 30 年（2018 年）に実施された 2 つの試行調査による国語の記述問題の手書き文字データを、現在の最先端の技術で採点、評価することを試みる。文字認識には畳み込みニューラルネットによるアンサンブル学習と N グラムを用いた。自動採点には日本語 Wikipedia を事前学習した BERT (Devlin, 2019)（**B**idirectional **E**ncoder **R**epresentations from **T**ransformers の略、「Transformer による双方向のエンコード表現」と訳される）を用いる。既存研究にはない本稿の技術的なアドバンテージは以下の通りである。

1) 各年とも 6 万件を超える教育用の文字解答採点済みデータはかつてなく、おそらく世界最大の大量のデータを取り扱う。
2) 機械学習による完全な自動採点を行う。手書き文字解答データについては判別不能な文字や、人では判別可能であっても機械が誤認識する文字を含むが、これを人手で修正することをしない。また採点エンジンにおいても、既存研究では性能を向上させるために文字読み取りデータに採点を補助するためのアノテーションを人手で行うことをすることが多いが、これを一切しない。

2 試行調査国語の記述問題の外形的仕様

2.1 出題形式

平成 29 年の国語の記述問題は与えられた「生徒会部活動規則」を踏まえ、5 人の生徒の会話、およびその執行部会で使用された 3 つの資料をもとに、3 つの設問についてそれぞれ 50 字、25 字、（80 字以上）120 字以内で解答する。平成 30 年度の記述問題は「ヒトと言語」についての探求レポートを書くときに参考とした 2 つの資料をもとに、「指さし」に関する 3 つの 3 つの設問に対しそれぞれ 30 字、40 字、（80 字以上）120 字以内で解答する。表記のバラツキを減ずるために一部、書くべき文章の文頭や文末を指定している。試験問題や正解表については大学入試センターの Web ページに公開されている。

2.2 採点方法

採点者による評点のバラツキを減ずるために、予め「正当の条件」を 3～4 個を用意し、その条件への適否の組み合わせによって評定するという方式が取られた。

3 採点のための要素技術

3.1 手書き文字認識

手書き文字認識のために JIS 第 1 水準手書漢字データベース ETL9(斉藤ほか 1985)を用いる。100 万を超える手書き文字を畳み込みニューラルネットに基づき学習する。我々のグループが用いるのはニューラルネットの階層を 16, 24, 34, 50 に変えて、それぞれの結果に対して、最終的な予測が最大になるようにチューニングされた重みづけを行なう。こ

のアンサンブル学習した重みを予測に適用する。この結果は字形の予測をある確率で与えるものであるが、これに日本語 Wikipedia で学習した 5 グラムモデルを用い、それらの 2 つの確率（すなわち字形の確率と繋がり易さの確率）の積が最大となるような字形の並びを抽出した。

3.2 自動採点モデル

日本語 Wikipedia で事前学習した BERT を用いる。そこに答案の情報を上乗せする形で学習することで、 Wikipedia の大規模な情報から得られた知識をも用いて答案の採点をおこなうことができる。手順は以下の通り：

1) テキストデータ化された答案を形態素に分割する。形態素はトークンと呼ばれそれぞれに ID が振られる。文章の先頭には CLS と呼ばれるスペシャルトークンを付与する。
2) 各トークンに割り振られた ID を（日本語 Wikipedia で事前学習した結果から）768 次元のベクトルに変換する。
3) BERT における隠れ 12 層のうち、最終 4 層の CLS トークンのベクトルを獲得する。
4) 取り出した 768 次元のベクトルを列方向に結合する。これで 768*4 の行列を作る。
5) この行列を線形層で次元変換し、その後、ソフトマックス関数にかけて最も確率の高いカテゴリを出力する。

BERT での推定にあたり全データを 3 : 1 : 1 (すなわち 60%, 20%, 20%)に分割し、それぞれ訓練データ、開発データ、検証データとした。モデルの最適化は Adam を用い、バッチサイズは 16、epoch 数は 5 で学習を行った。

4 予測の評価

2017 年および 2018 年の各 3 問ずつ計 6 問の採点における統計量を表 2 に示す。順に、回答数、正答の条件数、満点（採点カテゴリの数から 1 を減じる）、平均、標準偏差、QWK（2 次の重み付き κ 係数）、字数制限を示す。段階別に評価されたスコアの平均や標準偏差を計算するために、例えば a-d の 4 段階評価については最悪の評価(d)がゼロになるよう、a→3、b→2, c→1, d→0 とした。表の満点および平均の欄から分かるように 2017#1 が最も易しく、2017#3 が最も難しい。

5 まとめ

大規模な試験データを用いて、手書き文字データと人手による採点結果のみから現在の技術水準における機械学習による採点の精度を検証した。これより以下がわかった。

1) データの数が十分にあれば、すなわち大規模試験においてはアノテーション作業なしに推定精度をそれなりに保証できる。
2) 25 字から 120 字程度の記述問題においては 5 万個程度のデータであっても学習は通常の場合、収束しない。それだけ記述あるいは解答のバリエーションが多様であることの証左であろう。
3) 手書き文字認識のエラーが現状程度（5%程度）にあっても、ある程度の採点の精度は保証される。

連絡先：tunenori@rd.dnc.ac.jp

表 2：各問における統計量

問題	回答数	正答の条件数	満点	平均	標準偏差	QWK	字数制限
2017#1	62,222	4	6	4.46	1.67	0.978	～50
2017#2	61,777	3	2	1.51	0.87	0.963	～25
2017#3	59,791	4	5	0.43	1.10	0.866	80～120
2018#1	67332	3	3	2.51	0.88	0.976	～30
2018#2	66,246	3	3	1.87	1.14	0.954	～40
2018#3	58,159	5	3	0.76	1.07	0.944	80～120