E-016

クラスタリングを用いた学習データの選定による自動採点システムの精度向上

Accuracy Improvement of Automatic Scoring System by Selection of Learning Data with Clustering

作田 航平† 横尾 拓未† 早川 純平† 森 康久仁‡ 須鎗 弘樹‡
Kohei Sakuta Takumi Yokoo Jumpei Hayakawa Yasukuni Mori Hiroki Suyari

1. はじめに

教育改革の 1 つである「大学入試改革」では，2020 年度より「大学入試センター試験」に代わって「大学入試共通テスト」を実施する予定である．この「大学入試共通テスト」では思考力や表現力を測ることを目的に，記述式問題の導入を予定していたものの延期となった．その理由として 50 万人規模の答案を 20 日間程度で正確に採点する必要があるため多くの採点者が必要とされ，アルバイトの起用による採点のミスやぶれへの懸念があるからである．ミスやぶれを少なくし，時間や人件費の削減のために，記述式問題の自動採点や採点支援システムが期待されている．また，新型コロナウィルスの影響で，学生は，オンライン学習を求められることが多くなった．選択式問題の採点は簡単であるが，記述式問題については，採点が難しく，減点が伴う解答では，なぜ減点なのか，理由を求められる．こうした背景から，記述式問題の自動採点や採点支援の研究が近年増えてきている．

高井らは，中学生を対象に行われた模試の記述式問題 3 科目分の解答データを対象として，bidirection-LSTM と self-attention を用いた文書分類モデルを提案している [1]．機械学習を用いた自動採点では，学習データとして人の手で採点された解答が必要であり，高井らは人の手で採点する解答をランダムに選んでいる．この手法での採点精度は，国語と理科は約 80%，社会は約 90%という結果になっている．また，self-attention により，予測の際にどの単語に注目しているかを可視化できるため， 採点支援への活用を示唆している．しかしながら，自動採点と採点支援を並行して行える可能性があるものの，採点精度が十分に高いとは言えない．その理由の１つとして，人の手で採点する解答をランダムに選んでいる点が挙げられる．

鈴木らは国語の記述式答案を対象として，Doc2Vec を用いた記述式答案を評価する手法を提案している [2]．記述式答案と模範解答を Doc2Vec でベクトルに変換し，それらの cos 類似度の値で評価している．ゆえに， 採点済み答案を必要としないため他の手法よりも簡単に記述式問題を評価できる可能性がある．しかし，答案と模範解答に含まれる単語の一致度合いで答案の良さを評価できる場合においては有効な手法であるが，模範解答の言い換えが多い答案では評価が困難になる問題がある．

これらの手法では，表現が幅広い解答や珍しい表現を用いた解答に対して正しい採点や評価をすることが困難である．ゆえに，このような種類の解答を学習することができれば採点精度が向上すると考えられる．

本研究の目的は，採点精度が向上するように人の手で採点する解答を適切に選ぶことであり，その手動で採点するべき解答を k-means 法でクラスタリングして選ぶ方法を提案する．自動採点システムには，汎用言語表現モデルである BERT と RNN の拡張モデルである LSTM を組み合わせたネットワークを用いる．このシステムでは単語の順番を考慮できるため，高い精度で採点予測することを可能にした．人の手で採点する解答をランダムに選んだ場合と解答文を k-means 法でクラスタリングし，全てのクラスタから 1 つづつ解答を選んだ場合とを比較し，採点における学習データの選び方を考察する．

また，Zesch らは，クラスタリングを用いることで学習データ数を削減できることを示している[3]．しかし，長い解答に対しては効果が期待できない．したがって，本論文では日本語の比較的短い解答に対して，クラスタリングを用いる手法がどの程度有効であるか示す．

† 千葉大学大学院融合理工学府 Graduate School of Science and Engineering, Chiba University
‡ 千葉大学大学院工学研究院 Graduate School of Engineering, Chiba University

2. 提案手法

本研究では，文章をベクトル化する際に単語の順番を考慮できる BERT と音声や文章などの時系列データを学習することができる LSTM を用いて自動採点を行う．人の手で採点をする解答をランダムに選ぶ手法をランダム手法（図１），解答文をクラスタリングし，全てのクラスタから１つづつ解答を選ぶ手法を提案手法と呼ぶ（図２）．

2.1 文章のベクトル化

記述式問題を自動採点するためにまず最初に解答文をベクトルにする必要がある．解答の文章を JUMAN++を用いて形態素に分割した後，京都大学の黒橋・河原研究室が公開している BERT の日本語 pretrained モデル[4]を使用して，文章ベクトルと単語ベクトルを得る．このモデルは，日本語 Wikipedia の全ページを学習している．BERT を用いてベクトル化をしているため，単語の順番を考慮した文章ベクトルを得ることができる．

2.2 学習データの選び方

ランダム手法では，ランダムに選んだ解答を採点する解答とする．一方，提案手法では，その文章ベクトルに対して k-means 法を用いてクラスタリングをする．クラスタ内の各文章ベクトルの平均を取り，それをそのクラスタの中心とする．そのクラスタの中心と各文章ベクトルのユークリッド距離を計算し，最も近いものをそのクラスタの代表として採点すべき解答とする．

その後，手動採点をした解答を学習データ，残りの解答をテストデータとする．

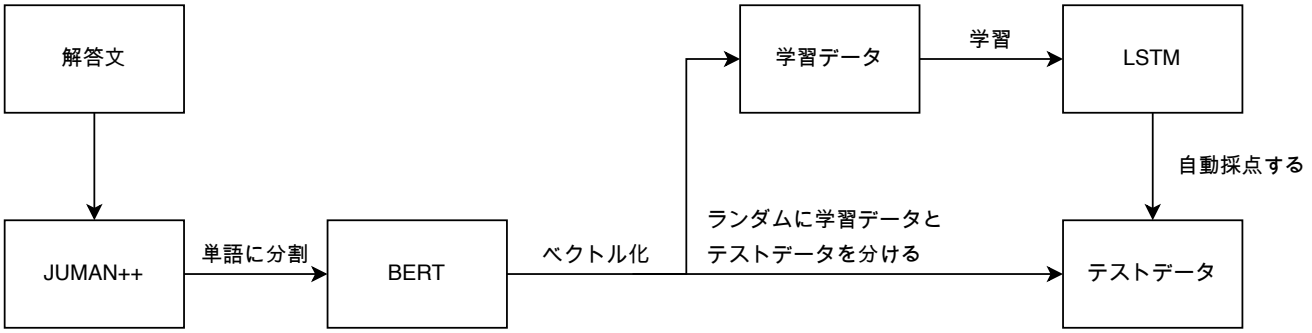

図 1　ランダム手法の概要図

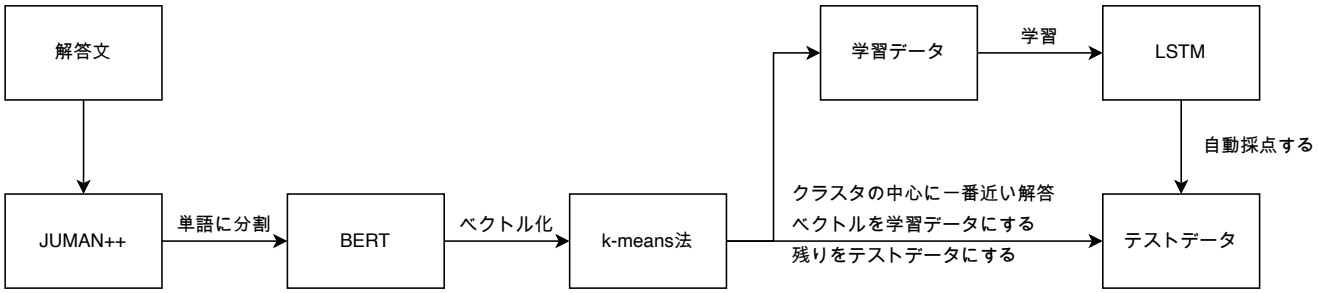

図 2　提案手法の概要図

2.3 学習モデル

BERT で得た学習データの単語ベクトルを入力として，LSTM で学習させると単語の順番を考慮した文章ベクトルを得ることができる．このベクトルを入力として 1 層のニューラルネットワークで，正解・不正解の分類を行う．このとき，解答からランダムに学習データを選んだ場合とクラスタリングを用いて選んだ場合を比べて採点精度がどのように変化するのか考察する．

3. 評価実験

3.1 実験データ

株式会社進学研究会から提供して頂いた「中学生を対象に行われた記述式問題の解答データ」に対してシステムによる自動採点を行う．国語，社会，理科の 3 科目について各 1 題ずつをデータセットとして用いた．本研究では，解答が空欄のものはデータとして適切ではないため取り除いた．これらのデータには人による採点結果のラベルがついており，各科目の正解・不正解の割合は表 1 の通りである．理科に関しては，正解の解答が多く偏りがある．

表 1　解答データの内訳

科目	正解数(割合)	不正解数(割合)	全解答数
国語	420(44.8%)	518(55.2%)	938
社会	616(62.7%)	367(37.3%)	983
理科	863(75.6%)	278(24.4%)	1141

各科目の問題の特徴を表 2 に示す．国語と社会は穴埋め形式であるため，解答する際に前後の文章を考慮する必要がある．また，各科目の具体的な問題と模範解答を図 3，図 4，図 5 に示す．

表 2　各問題の特徴

科目	形式	文字数制限	指定単語数
国語	穴埋め	20 字以内	1 単語
社会	穴埋め	25 字以内	3 単語
理科	自由記述	なし	なし

問題
　文章中に「私は口を尖らせた」とあるが，このときの「私」の気持ちについて述べた次の文の[　]に入る言葉を，「おばあちゃん」という言葉を使って 15 字以上，20 字以内で書きなさい．

本が見つかるのと，本が見つからないままなのと，どちらがいいことなのか本気で悩むと同時に，必死で探しても本が見つからず，[　]ことに不満を感じている．

模範解答
　おばあちゃんが苦労を理解しようとしない

図 3　国語の問題と模範解答

問題
　次の文章は，日米修好通商条約について述べたものである．文章中の[　]に当てはまる適当なことばを，「関税の率」「権利」「日本」の 3 つの語を用いて，25 字以内で書きなさい．

日米修好通商条約は，日本にとって不利な内容をふくんだ不平等条約であったが，江戸幕府は，アメリカに次いで，オランダ，ロシア，イギリス，フランスとも同様の条約を結んだ．条約によって自由な貿易が始まると，不平等な内容の一つである[　]ことから，イギリスから安い絹織物や綿糸が大量に輸入されて，国内の産地は大きな打撃をうけた．

模範解答
　輸入品にかける関税の率を決める権利が日本になかった

図 4　社会の問題と模範解答

問題
　イモリやカエルなどのなかまは，子と親とで呼吸のしかたが異なる．子と親の呼吸のしかたを，それぞれ簡潔に書きなさい．

模範解答
　子はエラで呼吸し，親は肺で呼吸する．

図 5　理科の問題と模範解答

各科目の解答文の総単語数と解答文 1 つあたりの平均単語数を表 3 に示す．重複している単語は総単語数として数えていないので，国語は他の科目に比べて使われている単語が多く，様々な表現の解答があると言える．

表 3　解答データの単語数

科目	総単語数	平均単語数
国語	597	8.4
社会	423	12.5
理科	247	15.0

3.2 実験方法

本研究では，LSTM のハイパーパラメータを変更せずに 2 つの手法を比較する．手動で採点する解答はできるだけ少ない方が良いが，手動で採点する解答が多いほど採点精度が良くなることが予想される．したがって，学習データの数を 100，300，500，700 とし，この 2 つの手法を比較する．バッチサイズの関係上，理科の解答のみ学習データを 100，300，501，700 にしている．

4. 結果と考察

解答のデータは人による採点結果のラベルがついているため，テストデータ全てに対して採点予測した結果とそのラベルを照合して正しく採点できた割合を採点精度とする．採点精度が高ければ高いほど採点予測が適切であると言える．各科目の採点予測を 10 回行い，その採点精度の平均を図 6，図 7，図 8 に示す．

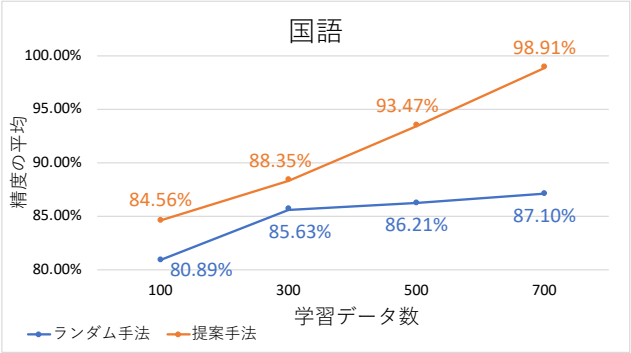

図 6　国語の採点結果

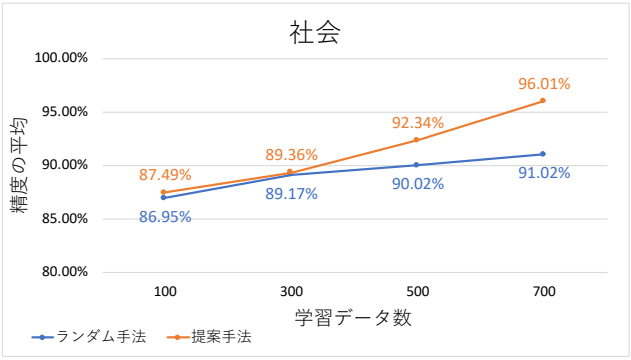

図 7　社会の採点結果

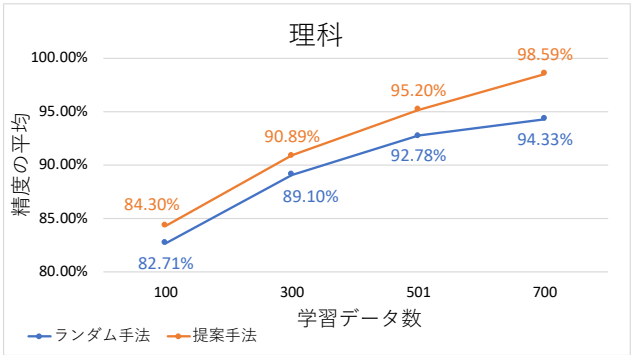

図 8　理科の採点結果

また，各科目の 10 回の採点精度を箱ひげ図として図 9，図 10，図 11 に示す．×は 10 回の採点精度の平均値を表している．

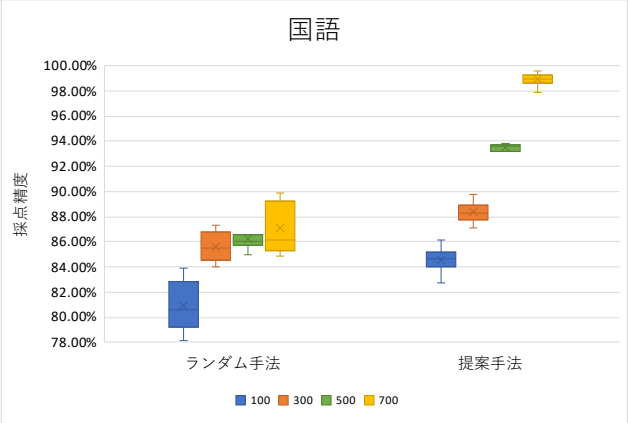

図 9　国語の箱ひげ図

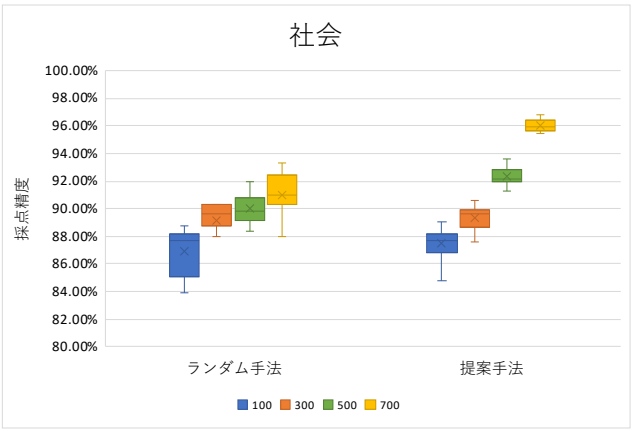

図 10　社会の箱ひげ図

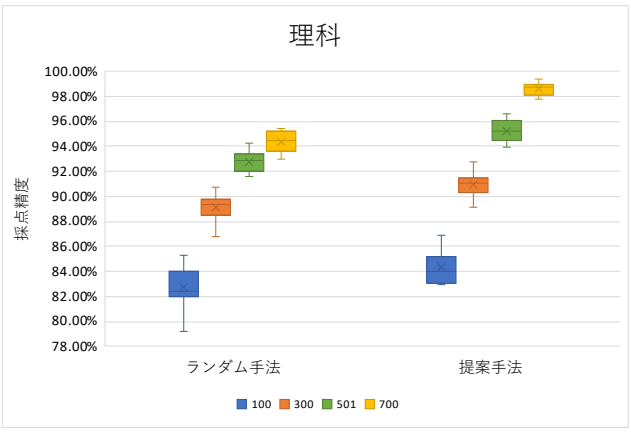

図 11　理科の箱ひげ図

折れ線グラフを見ると，3 科目ともランダム手法より提案手法の方が採点精度が高いことが分かる．これはクラスタリングにより幅広い解答を学習したことで，テストデータにある同じような表現の解答に対して，正しく予測を行えるため採点精度が向上したと考えられる．また，学習データが多い方が採点精度が高くなり，ランダム手法と提案手法の精度の差が開いた．これはクラスタリングをすることで採点予測をするのが困難な解答を人の手で採点する学習データに選ぶことができるからであると考えられる．

社会と理科は解答に出現する単語がほぼ同じであるため，ランダム手法と提案手法による精度の差は小さい．国語においては解答の表現の幅が広いので，ランダム手法と提案手法による精度の差が大きくなったと考えられる．また，箱ひげ図を見ると，ランダム手法より提案手法の方が採点精度のブレが小さい．これらの結果より，提案手法の方が優れていると言える．

5. 結論

短い解答を自動採点する場合はクラスタリングで学習データを選ぶという提案手法は有効であると言える．また，問題のタイプによって，提案手法の効果が異なることが分かった．国語のような表現の幅が広くなる問題や指定単語がない問題においては，提案手法によって採点精度が向上しやすいが，社会や理科などの解答に出現する単語が同じになりやすい問題や指定単語が多い問題においては採点精度が向上しづらいと考えられる．

提案手法によって採点精度の向上を達成したものの，実用化するには 100%に限りなく近い精度が求められる．テキストクリーニングなどの前処理やモデルの改善などを行い，さらなる精度向上を目指す．

謝辞

本研究を進めるにあたり，評価実験を行うために必要な解答データを提供してくださった株式会社進学研究会に心から御礼申し上げます．

参考文献

[1] 髙井 浩平, 竹谷 謙吾, 早川 純平, 森 康久仁, 須鎗 弘樹, “LSTM と Attention を用いた自動採点及び採点支援の実用化に向けて”, 人工知能学会 第 33 回全国大会, 2019
[2] 鈴木 千尋, 佐藤 直行, “Doc2Vec を用いた国語記述式答案の自動評価”, 情報処理学会 第 81 回全国大会, 2019
[3] Torsten Zesch, Michael Heilman, Aoife Cahill, “Reducing Annotation Efforts in Supervised Short Answer Scoring “, Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 124–132
[4] BERT 日本語 Pretrained モデル -KOROHASHI LAB : http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT 日本語 Pretrained モデル