E-019

アンサンブル学習による記述式問題自動採点システムの改善
Improvement of scoring system for descriptive problem with ensemble

早川 純平† 高井 浩平† 竹谷 謙吾† 森 康久仁‡ 須鎗 弘樹‡

Jumpei Hayakawa Kohei Takai Kengo taketani Yasukuni Mori Hiroki Suyari

1. はじめに

2020 年度より大学入試センター試験に短答記述式問題が導入される予定であり，それに伴い採点における人件費や採点時間のコストが大きく増えることが予想される．竹谷らは模範解答を基に作成されたキーフレーズを解答文と比較することにより自動採点を行うシステムを提案し，高精度で自動採点を行うことを可能にした[1]．以後，このシステムを従来システムと呼ぶ．従来システムでは機械学習を使用しないため，事前に学習するための学習データ，つまり人の手で採点され，正解または不正解のラベルが付与された解答データを必要としない．その一方で，多種多様な解答に対してそれらを網羅するキーフレーズのパターンを十分に設定することは難しく，自動採点不可能なデータは全解答の約 4 割に及び，それらは人の手によって改めて手動採点する必要がある．

本研究では，アンサンブル学習を導入し，従来システムでは自動採点不可とされた解答を自動採点する手法を提案する．従来システムによって自動で採点されたデータを学習データとして学習モデルに学習させることで正解・不正解ごとの特徴を学習し，従来のシステムでは採点不可能とされたデータを学習モデルに改めて採点させる．また，アンサンブル学習の構造を利用し，提案手法による自動採点の信頼性が高いとされる解答を自動で判別し，その解答のみ自動採点を行う．本研究は提案手法により従来よりもどれだけ多くの解答が自動採点可能となり，また精度がどれだけ保証されるのかを検証することを目的とする．

2. 関連研究

記述式問題の自動採点の研究は近年増えつつある．水本らは LSTM を含んだ深層学習のモデルを拡張しアテンション機構を用意することで，記述式問題を項目ごとに採点し学習者へのフィードバックを可能とした学習支援システムを提案している[2]．しかし，学習データとして事前に手動で採点しラベルを付与した解答データを用意しなければならないという問題がある．寺田らは畳み込みニューラルネットワークを用いた自動採点について発表している[3]．これは人による手動採点を行った後にシステムによる自動採点を 2 人目・3 人目の採点者とすることで学習データの確保をしている．

機械学習を用いた自動採点をするためには事前に人による手動採点で学習データを用意しなければならないことが課題となっている．しかしながら入学試験の性質上，学習データとして用いる解答をあらかじめ大量に用意することは不可能である．本研究ではシステムが全ての解答を自動採点する場合と，システムが 1 人目の採点者として自動採点を行い，その上でシステムによる自動採点が難しい解答を人が 2 人目の採点者として採点を行う場合の 2 通りについて検証する．

3. 提案手法

従来システムの概要図を図 1 に示す．従来システムでは自動採点できない解答は人の手で手動採点される．提案手法の概要図を図 2 に示す．本手法ではアンサンブル学習を用いて，従来システムでは自動採点できない解答を自動採点する．

3.1 事前処理

従来システムにより全ての解答を自動採点可能・自動採点不可能に分類する．このとき，自動採点可能な解答は正解・不正解のラベルが付与される．自動採点可能な解答を学習データとして使用し．自動採点不可能な解答をテストデータとする．

学習データ・テストデータを TF-IDF を用いて単語の出現頻度に基づき，分散表現に変換する．また従来システムにおいて自動採点された解答のうち無解答のため不正解，指定文字制限より文字数が少ないため不正解と自動採点された解答は学習データとしては不適切であるため学習データからは削除する．

3.2 学習モデル

本研究では SVM(線形カーネル)，SVM(RBF カーネル)，確率的勾配降下法，ランダムフォレスト，ロジスティック回帰，K 近傍法，ニューラルネットワークの 7 つの分類器を使用したアンサンブル学習を行う．各分類器は scikit-learn のモデルを使用し，ハイパーパラメータは scikit-learn の GridSearchCV で最適化したものを使う．

それぞれの分類器は前述の学習データで学習した後にテストデータを正解・不正解の 2 値に分類し，これら 7 つの採点結果について解答ごとの多数決を取り，最終的な採点結果とする．

3.3 採点結果の信頼度

アンサンブル学習で自動採点された解答は次の 2 通りに分類することができる．

A:ある分類器では正解と採点され，他の分類器では不正解と採点されたが，多数決をもって採点結果を決めた解答．

B:全ての分類器での採点結果が一致し，採点結果を決めた解答．

このとき，B の解答の方が精度が高いと予想できる．実験では A，B ともに採点した場合を実験(a)，B のみを採点した場合を実験(b)として精度と採点率を比較する．

† 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

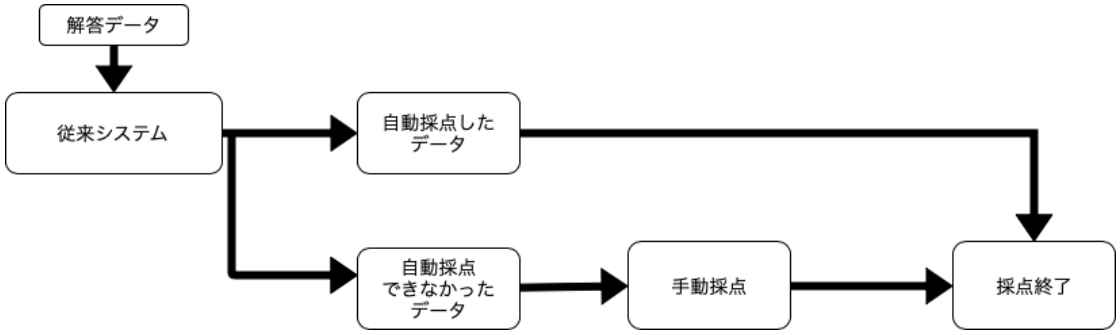

図 1: 従来システムの概略図

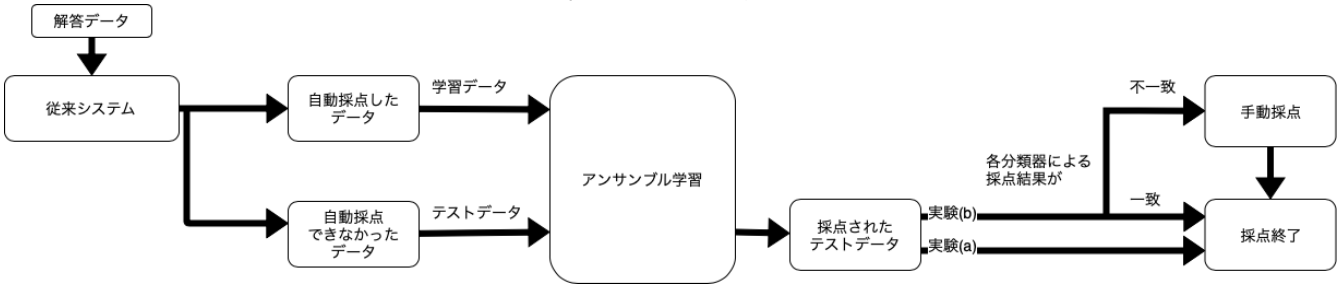

図 2: 提案手法の概略図

4. 実験

4.1 実験設定

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従来システム及び，本システムによる採点を行う．国語，社会，理科の 3 科目について 1 題ずつを使用する．各科目の問題文と模範解答をそれぞれ図 3，図 4，図 5，に示す．これらの問題に対して，それぞれ約 1200 人分の解答データがあり，人による採点結果のラベルもついている．解答方法については以下の特徴を持つ．

・国語（穴埋め形式，15 字以上 20 字以内，1 単語指定あり）
・社会（穴埋め形式，25 字以内，3 単語指定あり）
・理科（自由記述，20 字程度）

また，従来システムによる採点結果を表 1 に示す．自動採点を行った解答をその採点結果とともに学習し，自動採点不可となった解答を提案手法で自動採点する．

国語

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: 国語の問題文と模範解答

社会

文章の［　　］にあ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含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, イギリスから安い綿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受けた。

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: 社会の問題文と模範解答

理科

イモリやカエルなどのなかまについて，次の (a),(b) の問いに答えなさい。

(a) 略
(b) これらのなかまは，子と親で呼吸の仕方が異なる。子と親の呼吸の仕方を，それぞれ簡潔に書きなさい

模範解答 : 子はえらで呼吸し，親は肺と皮ふで呼吸する。

図 5: 理科の問題文と模範解答

4.2 実験(a)

従来システムで採点不可能とされた解答を全て提案手法で自動採点し，従来システムによる採点と合わせた結果の自動採点率とその採点精度を検証する．この場合，全ての解答を自動採点するため自動採点率が 100%になる．その状況での自動採点精度を検証する．

4.3 実験(b)

3.4 節で述べたようにアンサンブル学習での採点結果が全ての分類器で一致した解答のみを提案手法での採点結果とし，従来システムと合わせた結果の自動採点率と採点精度を検証する．この場合，一部の解答を採点しないため採点率が 100%に満たない．しかし実験(a)よりも高い精度での採点に期待できる．

表 1:従来システムによる自動採点結果

	解答数	自動採点数	自動採点不可数	自動採点率
国語	1198	707	491	59.0%
社会	1174	740	434	63.0%
理科	1195	826	369	69.1%

5. 結果・考察

5.1 実験(a)の結果

実験(a)の結果を表 2 に示す．結果の値は実験を 5 回行った際の平均値となっている．表 2 よりアンサンブル学習でテストデータのすべての解答を自動採点し，従来システムと合わせると精度が約 9 割まで下がった．

また，表 3 に各学習器による採点とそれらのアンサンブル学習での採点結果を示す．アンサンブル学習の精度は国語では 1 位，社会では 1 位，そして理科では 5 位となっているが．3 科目平均では 1 位となっている．このことから問題によって精度の良い分類器が異なることがわかる．またアンサンブル学習は個々の問題に対しては必ずしも有効でないことも確認できた．一方でアンサンブル学習は 3 科目の平均での精度が最も高いことが確認できる．

表 2:実験(a)の結果

	従来システム		実験(a)	
	自動採点率	自動採点精度	自動採点率	自動採点精度
国語	59.0%	99.3%	100.0%	89.4%
社会	63.0%	97.4%	100.0%	91.8%
理科	69.1%	100.0%	100.0%	90.1%

表 3:アンサンブル学習とその分類器ごとの自動採点精度

	国語	社会	理科	3 科目平均
SVM(線形)	89.1%	92.0%	89.1%	90.1%
ロジスティック回帰	88.6%	91.1%	88.6%	89.5%
K 近傍法	88.4%	88.1%	93.0%	89.8%
決定木	86.3%	89.4%	90.4%	88.7%
ニューラルネットワーク	88.6%	91.1%	90.5%	90.1%
SVM(RBF)	88.3%	91.2%	89.2%	89.6%
確率的勾配降下法	89.3%	91.2%	90.5%	90.3%
アンサンブル学習	89.4%	91.8%	90.2%	90.5%

5.2 実験(b)の結果

実験(b)の結果を表 4 に示す．結果の値は実験を 5 回行った際の平均値となっている．採点率は従来よりも約 20%増加し，採点精度は 95%を上回る結果となった．また，実験(a)との比較として，従来システムの結果を含めずにアンサンブル学習での採点のみの結果についての自動採点率と採点精度を表 5 に示す．実験(b)では実験(a)よりも採点精度が平均して 11.9%向上しており，アンサンブル学習での自動採点のうち，より精度の高いものを選ぶことに成功している．

表 4:実験(b)の結果

	従来システム		実験(b)	
	自動採点率	自動採点精度	自動採点率	自動採点精度
国語	59.0%	99.3%	81.3%	95.8%
社会	63.0%	97.4%	83.9%	95.4%
理科	69.1%	100.0%	85.1%	97.3%

表 5:実験(a),実験(b)の比較

	実験(a)		実験(b)	
	自動採点率	自動採点精度	自動採点率	自動採点精度
国語	41.0%	75.2%	22.3%	85.8%
社会	37.0%	82.3%	20.9%	90.0%
理科	30.9%	68.1%	15.9%	85.4%

最後に実験(b)では採点できなかった解答を人の手で採点を行うとした時，解答全ての採点精度と，実際に人が採点する必要のある解答の割合を表 6 に示す．ここで，人の手

で採点された解答は精度 100%で採点されるものとする．表 6 によると今回使用した記述式問題では手動採点を全解答の 16.6%行うことで，全ての解答を精度 96.8%で採点することが可能であることが確認できた．

表６:実験(b)に手動採点を加えた結果

	手動採点が必要な割合	精度
国語	18.7%	96.6%
社会	16.1%	96.1%
理科	14.9%	97.7%
3 科目平均	16.6%	96.8%

6. まとめ

本研究では国語，社会，理科の 3 科目それぞれ 1200 人分の解答データに対して，従来システムを利用して作成した学習データを用いて学習を行い，自動採点不可とされた解答の自動採点を行った．実験(a)では全ての解答を自動採点し，その精度は 90.5%となった．実験(b)では全解答のうち 16.6%を人の手で採点することになるが，精度 96.8%での採点が可能となった．

実験から機械学習を使うことにより，より幅広い表現を持つ解答を自動採点することができた．しかし，それでもなお自動採点が困難な解答も存在し，全ての解答を高い精度で自動採点することは難しい．これは従来システムが自動採点する解答は模範解答をベースに作成されたキーフレーズに近いものが多くを占めており学習データとして十分とは言えないことも一因として考えられる． 機械学習を使用するにあたっては，より多くの学習データを得ることや解答の構文に注目し，より良質な特徴量を得ることが必要である．

謝辞

本研究にて評価実験を行うにあたり，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感謝申し上げます．

参考文献

[1] 竹谷 謙吾, 高井 浩平, 清水 杏奈, 早川 純平, 森 康久仁, 須鎗 弘樹, “大規模実データにおける記述式問題自動採点システムの検証”, 言語処理学会第 24 回年次大会発表論文集. 2019, p.880-881.

[2] 水本 智也, 磯部順子, 関根 聡, 乾 健太郎,“採点項目に基づく国語記述式答案の自動採点”, 言語処理学会第 24 回年次大会発表論文集. 2018, p.552-555.

[3] 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 24 回年次大会発表論文集. 2016, p.370-373.