The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

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

Automatic Scoring and Scoring Support System using LSTM and Attention

高井 浩平 *1 Kohei Takai　竹谷 謙吾 *1 Kengo Taketani　早川 純平 *2 Junpei Hayakawa　森 康久仁 *3 Yasukuni Mori　須鎗 弘樹 *3 Hiroki Suyari

*1千葉大学 大学院融合理工学府 情報科学コース
Department of Applied and Cognitive Informatics, Graduate School of Science and Engineering, Chiba University

*2千葉大学 工学部 情報画像学科
Department of Informatics and Imaging Systems, Faculty of Engineering, Chiba University

*3千葉大学 大学院工学研究院
Graduate School of Engineering, Chiba University

In the university unified entrance examinations from 2020, a new kind of question to require answers in description style will be introduced. In this paper, we apply the scoring using LSTM and attention. We use about 1,200 score sheets and prepare two datasets((a), (b)) in the evaluation. (a) is randomly generated and (b) is generated using an automatic scoring system. As a result, the accuracy is 0.91 for (a) and 0.73 for (b).

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，採点には多大なコストがかかると予想されている．そこで，竹谷らは自動採点をメインとしたシステムを目指し開発を行っている [竹谷 19]．その竹谷らの採点システムでは採点精度を重視し，機械学習手法を取り入れていない．

本研究では，自動採点で機械学習手法が採点においてどれほど効果を発揮するかを検証するため，bidirection-LSTM と self-attention を用いた文書分類モデルを構築し，採点を行った際の精度を検証する．また，self-attention を利用することで予測理由の可視化を行い，採点の際に何に注目しているかを確認する．実験では，実際に中学生が解答した約 1,200 人分のデータを用い，採点は正解/不正解の 2 値分類とした．全データをランダムに分割して作成したデータセットと，竹谷らの採点システムにおいて採点を行った結果をデータセットとした 2 つのデータセットに対して実験を行い，結果を比較する．

2. 関連研究

竹谷らの自動採点システムは，シーケンスアライメントアルゴリズムを用いたルールベースのシステム（図 1）となっており，国語・理科・社会の 3 科目それぞれ平均して 1,188 人分の解答データに対して自動採点率約 68%，採点精度約 99.8%という結果を実現している（表 1）．

表 1: 実験結果 [竹谷 19]

	国語	理科	社会	平均
データ数	1198	1195	1174	1188
自動採点率 (%)	64.5	76.4	63.2	68.0
採点精度 (%)	99.9	100	99.6	99.8

連絡先: 高井 浩平，千葉大学 大学院融合理工学府 須鎗・森研究室，263-8522 千葉市稲毛区弥生町 1-33，e-mail : k.takai@chiba-u.jp

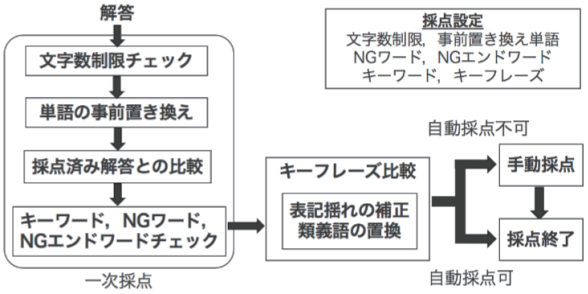


図 1: システム構成図 [竹谷 19]

ニューラルネットワークを用いた採点に関する研究では，寺田らが SVM，CNN の各手法の精度を検証している [寺田 16]．高校学習範囲の試験問題を大学生に解かせたものをデータとして，正解/不正解の 2 値分類を行い，どちらの手法も 90%近い精度で分類可能であることを確認している．

また，水本らが国語記述式答案を対象として，採点項目ごとの採点を行うモデルを提案している [水本 18]．モデルは LSTM と複数の attention を用いて構築されており，どの問題においても高い性能で予測が可能である．複数の項目点を予測することで学習者に対するフィードバックを可能にし，attention の分析によって学習支援に活用する可能性を示唆している．

しかしながら，どちらの手法も手動によるラベル付けが必要であり，全自動で採点を行った際の有効性を考慮していない．本研究では，手動によるラベル付けを行って実験した結果を踏まえた上で，採点システムを利用して全自動で採点を行った場合の精度を検証する．

3. self-attention を用いた文書分類モデル

本研究では，Zhouhan らの提案したモデル [Zhouhan 17] を参考に，日本語のテキストを入力として，正解/不正解の 2 値ラベルを出力する文書分類モデルを構築した．

bidirection-LSTM にテキストデータを与え，各単語に対応

する隠れ層から，ラベル予測の際にどの単語に注目するべきかを表す確率（self-attention）を予測する．その後，self-attention の重みを与えた各単語に対応する隠れ層を足し合わせ，正解/不正解の 2 値ラベルの予測を行う．このとき，self-attention の重みを可視化することで，予測の際にどの単語に注目したのかを確認することができる．本研究の実験では epoch 数を 300 に設定して学習を行い，self-attention の重みが大きい単語は赤色で強調される仕様になっている．

ラベルが付与されたデータに対し構築したモデルのみで採点を行うだけでなく，竹谷らの採点システムにモデルを導入して全自動で採点を行うことで精度を比較する．

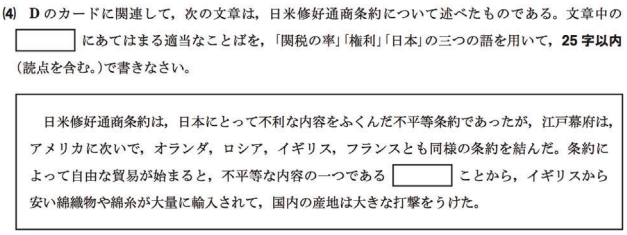
4. 評価実験

4.1 実験設定

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模試の記述式問題 3 科目分の解答データ」を検証データとして用いた．ここで，解答が空欄のものはデータとして適切ではないため除外した．また，各解答には正解/不正解の 2 値ラベルが付与済み（採点済み）である．各科目の問題の特徴は表 2 のようになっており，例として図 2 に社会の問題文を示す．

表 2: 各科目の問題の特徴

国語	20 字以内 穴埋め形式
理科	字数制限なし 自由記述形式
社会	25 字以内 穴埋め形式


(4) D のカードに関連して，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［　　］にあてはまる適当なことばを，「関税の率」「権利」「日本」の三つの語を用いて，**25 字以内**（読点を含む。）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うけた。

図 2: 社会の問題文

4.1.1 実験 (a)：ランダム

全解答データに正解/不正解のラベルが付与済み（採点済み）であることを前提として，文書分類モデルでの採点を行う．

使用するデータセットは，全データをランダムに分割して，約 7 割の解答をトレーニングデータ，残りをテストデータとして設定する（表 3）．

表 3: 実験 (a) データセット 内訳

	国語	理科	社会
トレーニングデータ数	781	951	819
テストデータ数	157	191	164

4.1.2 実験 (b)：竹谷らの採点システムを利用

全解答データに正解/不正解のラベルが付与されていなかった場合を想定し，竹谷らの採点システムに構築した文書分類モデルを導入することで全自動での採点を行う．

データセットとして，竹谷らの採点システムで自動採点されたものをトレーニングデータ，自動採点できなかったものをテストデータとして設定する（表 4）．

表 4: 実験 (b) データセット 内訳

	国語	理科	社会
トレーニングデータ数	447	773	549
テストデータ数	491	369	434

4.2 結果・考察

4.2.1 実験 (a)

実験 (a) の結果を表 5 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おり，データセットは実験毎にランダムに作り直している．

また，最も精度の高かった社会のなかで，正解の解答を正しく予測したものと，正解の解答を正しく予測できなかったものの self-attention を可視化した例を図 3，図 4 に示す．

表 5: 実験 (a) 結果

	国語	理科	社会
accuracy	0.83	0.81	**0.91**
precision	0.80	0.84	**0.92**
recall	0.89	0.93	**0.95**
F-measure	0.84	0.88	**0.93**

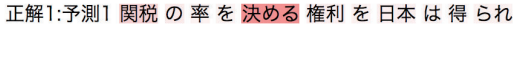


図 3: 実験 (a) にて正解の解答を正しく予測した例

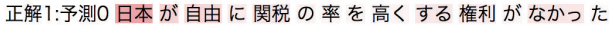


図 4: 実験 (a) にて正解の解答を正しく予測できなかった例

実験 (a) では，すべての科目で精度 0.8 以上を記録し，社会に関しては精度 0.9 以上に達することが確認できた（表 5）．各科目の精度を比較してみると，理科の精度が 0.81 と最も低かったが，これは問題形式が「字数制限なし 自由記述形式」だったことが原因であると考えられる．実際に解答データを確認してみても，生徒の解答内容はバラバラであり，中には単語のみで解答しているものもあった．よって，トレーニングデータを使って学習してもテストデータに対応できなかったのだろうと考えられる．

一方，0.91 と最も精度の高かった社会の問題形式は「25 字以内 穴埋め形式」であり，解答も全体的に似通った答え方のものが多かった．self-attention を可視化した図 3 を見ると，「決める」「なかった」という指定語以外の解答条件に沿った部分に注目できていることが確認できた．しかし，図 4 を見ると，「高くする」という表現に対応できず予測を誤っていることがわかる．この解答は正解なのだが，「高くする」という表現を使っている解答が他にないため，予測を誤ってしまったのだと考えられる．

4.2.2 実験 (b)

実験 (b) の結果を表 6 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．竹谷らの採点システムでは採点毎に結果が変化しないため，データセットは実験毎に変更していない．

実験 (a) と比較するため，社会のなかで，正解の解答を正しく予測したものと，正解の解答を正しく予測できなかったものの self-attention を可視化した例を図 5，図 6 に示す．

表 6: 実験 (b) 結果

	国語	理科	社会
accuracy	**0.73**	0.67	0.67
precision	0.79	**0.94**	0.91
recall	0.57	**0.67**	0.53
F-measure	0.66	**0.78**	0.67

正解1:予測1 輸入 品 の 関税 の 率 を 決める 権利 が 日本 に は ない

図 5: 実験 (b) にて正解の解答を正しく予測した例

正解1:予測0 関税 の 率 を 日本 で 決め られる 権利 が なかっ た

図 6: 実験 (b) にて正解の解答を正しく予測できなかった例

実験 (b) では，全ての科目で精度が下がり，最高精度も 0.73 ポイントという結果になった（表 6）．これは，竹谷らの採点システムがルールベースで採点を行っていることから，自動採点ができた解答（トレーニングデータ）に似た表現の解答が集まり，手動採点が必要な解答（テストデータ）に対応できなかったことが原因であると考えられる．また，3 科目ともに precision 値が高く，recall 値が低いという共通点が見られた．つまり，文書分類モデルが正解と予測したものが実際に正解である精度は高いが，実際は正解の解答を誤って不正解と予測しているケースが多いということである．以上より，実験 (a) でも述べたことであるが，やはり稀有な表現が用いられている解答に対しては正しい予測が困難であると考えられる．

5. おわりに

本研究では，国語，理科，社会の 3 科目それぞれ約 1,200 人分の解答データに対して，(a) ランダムに作成したデータセットと，(b) 竹谷らの採点システムを利用して作成したデータセットの 2 つのデータセットを用意し，ニューラルネットワークを用いた文書分類モデルによる評価実験を行った．その結果，(a) の実験では精度 0.91 ，(b) の実験では精度 0.73 という結果が得られた．

実験から，ニューラルネットワークを用いた採点では，ある程度解答の表現に制限を設けられる問題形式のものに関しては高い精度で採点を行えるが，解答の表現が幅広い自由記述形式の問題では高い精度で採点を行うことは難しいことがわかった．また，竹谷らの採点システムの出力を利用して採点を行ったが，採点精度が約 7 割という事実は実用上問題があり，本アルゴリズムをシステムに導入するのは難しく，全ての解答を自動で採点するにはまだまだ解決すべき問題がある．しかし，self-attention を可視化することで注目した単語を確認することが可能になるので，採点支援の観点では利用する価値がある．

6. 謝辞

本研究にて評価実験を行うにあたり，データを提供して下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

[竹谷 19] 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 25 回年次大会，2019.

[寺田 16] 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，2016.

[水本 18] 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.

[Zhouhan 17] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, Yoshua Bengio : A Structured Self-attentive Sentence Embedding，Conference paper in 5th International Conference on Learning Representations，2017.