Softmax 関数の出力を利用した 信頼度推定による記述式問題の誤採点の検出

横尾 拓未 [1]　作田 航平 [1]　早川 純平 [1]　倉田 基成 [2]　森 康久仁 [3]　須鎗 弘樹 [3]

[1] 千葉大学大学院融合理工学府
[2] 千葉大学工学部
[3] 千葉大学大学院工学研究院

1 はじめに

2020 年度から大学入試センター試験が廃止され，大学入学共通テストが開始された．大学入学共通テストに導入予定であった記述式問題だが，誤採点の可能性を完全に解消することはできないなどの理由から，導入は見送られた．記述式問題の誤採点という課題は大学入学共通テストに限った話ではなく，企業の実施する模試などにおいても不可避な問題の 1 つである．そこで本研究では，部分点の無い正解・不正解の記述式問題のデータを扱い，機械学習を用いて，採点済みの解答から誤採点を検出するシステムを提案する．

2 関連研究

近年，機械学習を用いた記述式問題における自動採点の研究は増えつつある．寺田らは SVM や畳み込みニューラルネットワーク (CNN) を用いた自動採点を提案した [1]．化学 1 題，生物 1 題，世界史 1 題，国語 4 題からなる計 7 題の問題に対してそれぞれ約 250 から 460 解答の採点済みデータを用意し，SVM および n-gram を利用した CNN で leave-one-out 交差検定を用い，正解・不正解の二値分類を行った．その結果，全体で 90%前後の精度，最も良い結果で 98.4%でのクラス分類に成功している．高井らは LSTM と Attention を用いた自動採点を提案している [2]．国語，社会，理科の計 3 題それぞれ約 1200 解答について約 7 割を学習データ，約 3 割を検証データとし，正解・不正解の 2 値分類を行なった結果，約 80%～90%の精度のクラス分類に成功した．水本らは部分点を持つ 50 字から 70 字程度の論筆 3 題，随筆 2，小説 1 題の計 6 題の国語の記述式問題の項目点予測に取り組んだ [3]．LSTM および Attention 機構を用い，項目点のついたデータを学習することで項目点を予測し，また全体点のデータを追加で学習することで項目ごとの性能の向上に成功した．しかし，自動採点の実用性を考えた時，現場では採点精度が限りなく 100%に近い精度を求められるが，いずれの研究も十分な精度とは言えず，依然として実用レベルに達していない．

機械学習による自動採点では，あらかじめ人間が採点した解答を学習データとして用いる．実際の現場では人によって採点のズレが生じたり，採点を誤る場合がある．採点に揺れが存在する解答を学習データとすることによって，自動採点の精度向上の妨げとなる可能性がある．そのため，モデルが高精度であればあるほど，100%の採点精度に届かない理由が学習データへの信頼性が不十分であるという問題が浮かび上がる．自動採点で精度を向上させる観点からも，また，採点結果をチェックする観点からも誤採点の検出は必須であると言える．

3 提案手法

本節では，採点済みの解答から誤採点を検出するアルゴリズムを提案する．はじめに，本研究では正解・不正解の二値分類となる記述式問題から人間の誤採点を検出することを考える．人間による採点のついた解答 (A_{human}) を用意し，4-fold cross validation にてモデルの学習を行う．学習したモデルによってテストデータに正解・不正解の分類結果 (A_{model}) が付与される．その後，A_{human} と A_{model} を比較し，$A_{human} \neq A_{model}$，すなわち，人間の採点とモデルの分類が異なった場合に，その解答を誤採点予測データとして抽出する．モデルの解答の分類では，softmax 関数の出力が [正解, 不正解] = [0.8, 0.2] のように出力され，出力値の大きい方の正解・不正解をモデルの結果としている．本手法では，以降，各解答ごとの softmax 関数の出力の大きい方を softmax

値と呼び，擬似的に信頼度として扱い，誤採点予測データの絞り込みに利用する．抽出した誤採点予測データには，表 1 のように，実際に人間の採点が誤っており，誤採点である解答 (真陽性) だけでなく，モデルが間違って採点した解答 (偽陽性) が存在する．

表 1　誤採点予測データの詳細

人間の採点	結果	
正しい	誤採点でない解答を検出	偽陽性
誤っている	誤採点を正しく検出	真陽性

本手法では機械学習のモデルに BERT-NN モデルと LSTM-NN モデルの 2 種類のモデルを用い，実験にてモデルの評価を行った．以下にそのモデルを示す．

- BERT-NN モデル

  BRET-NN モデルのモデル図を図 1 に示す．BERT の日本語 Pretrained モデルを用い，文章ごとにベクトルを与える．その後文章ベクトルを入力として学習を行い，正解/不正解の分類を行う．

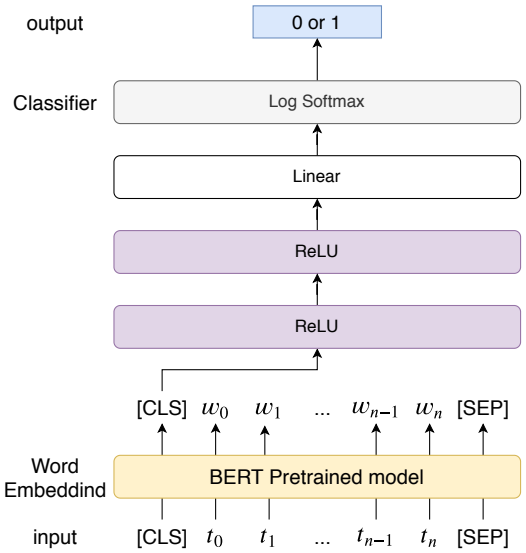

図 1　BERT-NN のモデル構造

- LSTM-NN モデル

  LSTM-NN モデルのモデル図を図 2 に示す．文章を Juman++によって形態素解析し，BERT の日本語 Pretrained モデルを用い，それぞれの単語にベクトルを付与する．その後，解答ごとに単語ベクトル群を入力とし，LSTM ネットワークで学習を行い，正解/不正解の分類を行う．

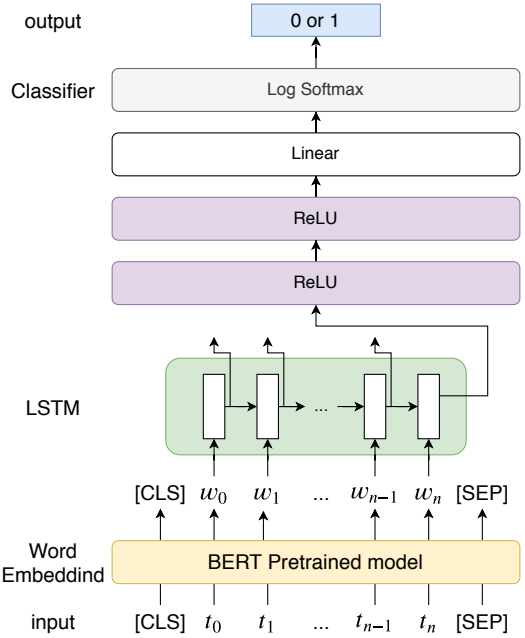

図 2　LSTM-NN のモデル構造

4　実験

提案手法の有効性を確認するために，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実験を行った．

4.1　実験データ

本研究では国語，社会，理科の 3 科目の問題計 3 問についてそれぞれ約 1200 人分の解答を用いた．これらの解答には人間による正誤の採点結果のラベルがついている．また各教科のデータの詳細を表 2 に示す．ここで解答数については空欄のものを削除した数となっている．

4.2　実験方法

本実験では，データセットからランダムに選んだ 10 個の解答に対して，人間が採点した正解/不正解のラベルを反転させ，意図的に誤採点を作成した．誤採点とする解答に偏りをなくすため，10 回実験を行い，モデルが誤採点と指摘した解答数，また実際に検出できた誤採点の解答数を比較し各モデルの評価を行った．評価指標は以下の 2 つを用いた

- 検出率

  全 10 問の誤採点の内，実際に検出できた割合
- 指摘精度

  誤採点と指摘した解答の内，実際に誤採点の解

表 2　実験データの詳細

科目	解答形式	字数制限	指定語句	解答数	正答：誤答の比 (%)	総単語数	解答文の平均単語数
国語	穴埋め	15 字〜20 字	1 語句	938	45:55	597	8.4
社会	穴埋め	25 字以内	3 語句	983	63:37	423	12.5
理科	自由記述	なし	なし	1142	76:24	247	15.0

答の占める割合

$$\text{指摘精度} = \frac{\text{真陽性数}}{\text{真陽性数} + \text{偽陽性数}} \quad (1)$$

誤採点と指摘する解答数を増やせば，一定量の偽陽性を認めることで検出率をあげることができる．したがって誤採点と指摘する解答数を減らした時に，検出率を維持し，指摘精度を向上できることが理想的なモデルと言える．本実験では，BERT-NN モデルと LSTM-NN モデルの 2 つの機械学習のモデルについて，以下に示す実験 1，またその結果からモデルを改善した実験 2 を行った．

4.3　実験 1

誤採点予測データそのままでは偽陽性が多く存在するため，モデルの出力結果の信頼度が高い解答を選ぶことを考えた．softmax 値があるしきい値以上の確率の解答についてのみを抽出することで絞り込みを行う．softmax 値のしきい値を変え，誤採点と指摘する解答数を変化させた時の，検出率と指摘精度の関係を見た．

4.3.1　結果と考察

国語，社会，理科に対する，誤採点と指摘する解答数を変化させた時の検出率と指摘精度の関係を図 3，図 4，図 5 にそれぞれ示す．最高の検出率は全教科で 80%を超えていることが見て取れる．また，LSTM-NN モデルの方が BERT-NN モデルよりも最高の検出率が若干高い．

最高の指摘精度は BERT-NN モデルが社会で 30.9%，LSTM-NN モデルが国語で 22.1%とともに低い．大きい softmax 値を持つ解答を抽出することが，モデルの出力結果の信頼度が高い解答を抽出することにはならず，偽陽性の削減に有効でないことが分かった．また，どの科目でも，LSTM-NN モデルの方が BERT-NN モデルより指摘精度が上がっていないことがわかる．BERT-NN モデルが BERT の学習済みモデルによって均一な方法でベクトル化されるのに対して，LSTM-NN モデルは LSTM の学習によって LSTM 内部の重みが決定され，その重みによって偏ったベクトル化が行われる．それによって解答間で softmax 値に偏りができ，softmax 値を信頼度として扱うことに適さなかったと考えられる．

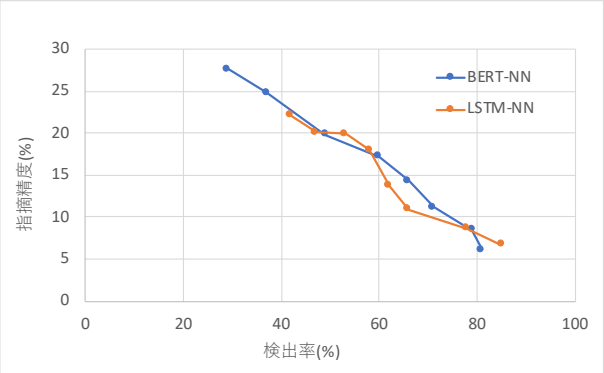

図 3　国語：検出率と指摘精度の関係 (実験 1)

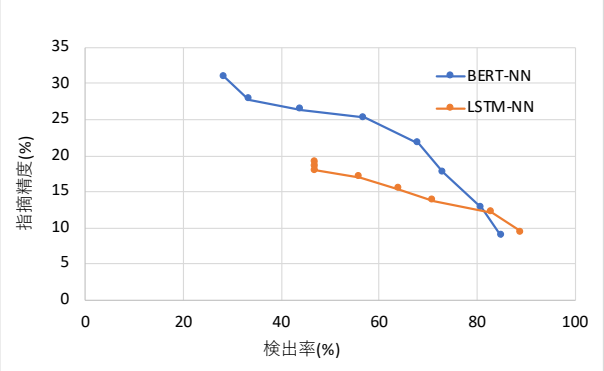

図 4　社会：検出率と指摘精度の関係 (実験 1)

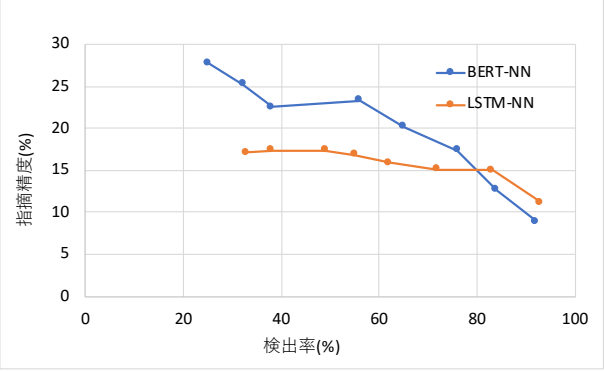

図 5　理科：検出率と指摘精度の関係 (実験 1)

4.4 実験 2

実験 1 の結果を受け，等しい条件で softmax 値を評価することを考える．均一にベクトル化できる BERT-NN モデルを用い，学習データに依存した softmax 値を避けるために，4-fold cross validation ではなく，全ての解答でモデルを学習させ，その際の各解答の softmax 値を評価する．検出率が高い LSTM-NN モデルを用いて誤採点予測データを用意し，BERT-NN モデルで全ての解答を学習させた時の，誤採点予測データの解答それぞれに対応する softmax 値を利用して解答を絞り込む．誤採点の解答を学習するということは本来したい分類の逆の学習であり，正しい分類の解答の影響を受け，誤採点の解答の softmax 値は低くなると予想される．これを利用し，あるしきい値以下の softmax 値を持つ解答を抽出することで誤採点以外の解答を除外し，指摘精度を向上させる．しきい値となる softmax 値を変え，指摘する解答数を変化させた時の検出率と指摘精度の関係を見る．

4.4.1 結果と考察

国語，社会，理科に対し，誤採点と指摘する解答数を変化させた時の検出率と指摘精度の関係を実験 1 の結果とともに，図 6，図 7，図 8 に示す．実験 2 の手法を手法 2 とする． 図 6 から国語は指摘精度の

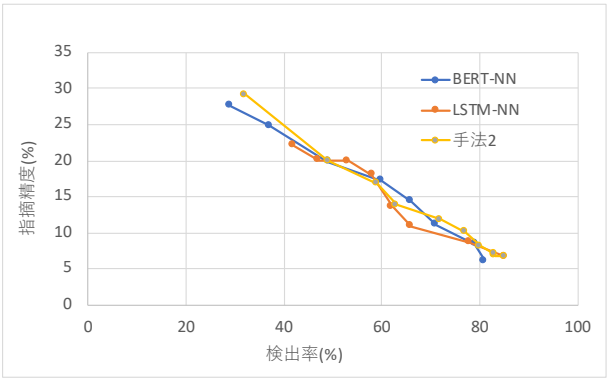

図 6 国語：検出率と指摘精度の関係 (実験 2)

向上がほとんど見られない．これは国語の解答が表 2 からわかるように総単語数が 597 と多く，解答ごとに様々なベクトル化がされるためと考えられる．全解答を学習した際，誤採点の解答が周りの解答の影響を受けることなく誤採点のまま学習されたため，softmax 値が低くならず，しきい値を設けても誤採点を抽出することに繋がらなかったと考えられる．図 7，図 8 から社会と理科で実験 1 に比べて，

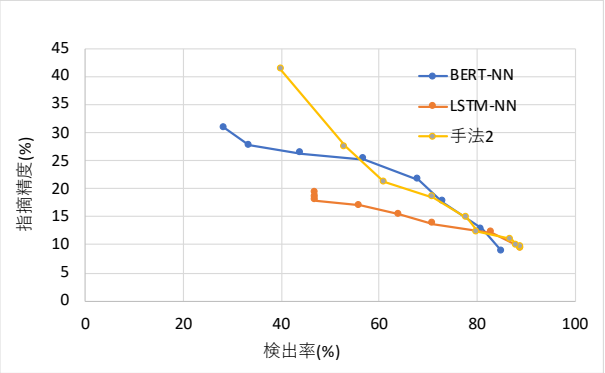

図 7 社会：検出率と指摘精度の関係 (実験 2)

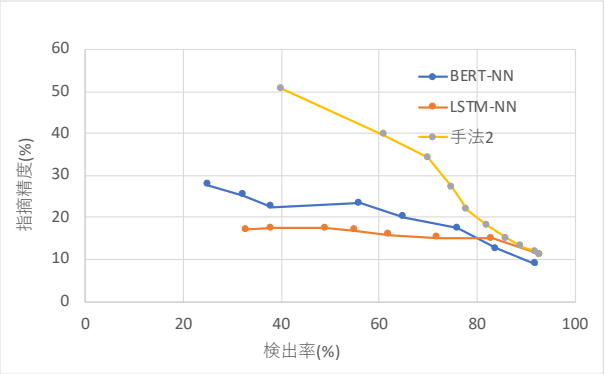

図 8 理科：検出率と指摘精度の関係 (実験 2)

指摘する解答数を減らした際に，検出率を維持したまま指摘精度を向上させることができた．

5 まとめ

実験 2 によって指摘精度の向上に成功した．本研究では全教科で約 90%の高い誤採点の検出率を実現し，また理科と社会においては誤採点と指摘する解答数を絞り込んだ際に指摘精度を理科で 51%，社会では 41%まで向上させることができた．今後の課題として検出率向上のためのモデルの改善，また，依然残っている偽陽性の削減などが上げられる．また誤採点となる解答を今回はランダムに発生させたが，実際に誤採点しやすいような解答を誤採点のデータセットとするような条件下で再実験する必要があると考える．

謝辞

本研究を進めるにあたり，研究を行うための貴重なデータセットを提供してくださった株式会社進学研究会に深く感謝し，心より御礼申し上げます．

参考文献

[1]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，pages 370-373，2016

[2]高井浩平，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，2I5-J-9-05(3 pages)，人工知能学会第 33 回全国大会論文集，2019

[3]Tomoya Mizumoto，Hiroki Ouchi，Yoriko Isobe，Paul Reisert，Ryo Nagata，Satoshi Sekine，Kentaro Inui ："Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring"，Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications，pages 316-325，2019