記述式答案自動採点のための確信度推定手法の検討

舟山弘晃 [1,2]　佐々木翔大 [2,1]　水本智也 [3,2]　三田雅人 [2,1]
鈴木潤 [1,2]　松林優一郎 [1,2]　乾健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所　[3] フューチャー株式会社
{hiroaki,jun.suzuki,,inui}@ecei.tohoku.ac.jp {y.m}@tohoku.ac.jp
{shota.sasaki.yv,tomoya.mizumoto,masato.mita}@riken.jp

1　はじめに

記述式問題の自動採点は，事前に人手で作成された採点基準について，入力された文章が採点基準を満たしているか評価し点数として出力するタスクである．主に，大規模な試験において低コストかつ公平な採点を実現するための採点者支援や，教育現場における学習支援を目的に研究されてきた [1–4]．深層学習に基づく自動採点モデルの登場により，近年自動採点システムの性能が向上している [3, 4] ものの，実際の教育現場での運用に耐える性能であるとは言い難い．

本研究では自動採点モデルの予測の信頼性を表す確信度を導入することによって，この問題の解決を試みる．予測の信頼性がうまく推定できれば，予測が十分信頼できる場合のみモデルの予測結果を採用し，信頼できない場合は人間の採点者に照会すると言った運用が可能になり，全体として採点の信頼性を担保することができる．また，日常的な教育における学習支援では，採点予測の信頼性の情報そのものを学習者に開示することによって採点誤りによる混乱を緩和するといった運用も考えられる．しかしながら，我々の知る限り自動採点における確信度推定に関する先行研究は存在しない．一方で，汎用的な深層学習を基にしたモデルの確信度を推定する試み自体はこれまでにも行われてきた．最も一般的な方法は，モデルの softmax 層からの出力，すなわち事後確率を用いるアプローチである [5]．また，事後確率を確信度とする手法以外にも確信度を推定する手法はいくつか提案されている [6, 7]．そこで，これら既存の確信度推定手法が自動採点タスクにおいて有効に機能するかを明らかにしたい．

本研究では事後確率を用いる場合と，モデルの中間層のベクトルを用いる場合の二つのアプローチに焦点を当てる．それぞれの確信度の振る舞いについて実験により検証し，記述式問題の自動採点における確信度推定手法としての有効性について議論する．具体的には，(1) 十分高い精度でどれだけ多くの回答を採点することが可能か，(2) 確信度を用いることにより重大な採点誤りを除くことが可能か，という 2 つの観点から調査を行う．国語長文読解問題データセットを用いた評価実験により，事後確率は確信度として機能するものの，重大な採点誤りを防ぐことができないことを明らかにした．また，モデルの中間層のベクトルを使う手法によって，事後確率より効果的に確信度を推定できることを確かめた．

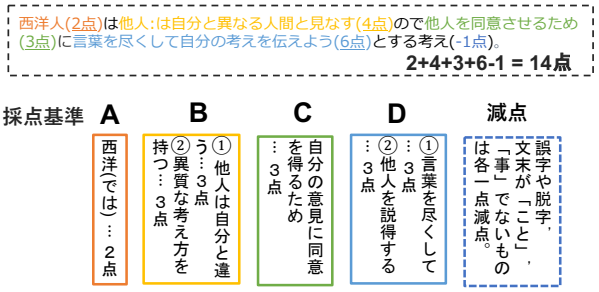

図1: 国語長文読解問題データセットの答案と採点基準の例．4.1節にて詳細に説明する．

2　記述式答案の自動採点

本節では，我々が取り扱う記述式答案の自動採点タスクと自動採点モデルについて説明する．

2.1　タスク設定

本研究における記述式答案の自動採点タスクは解答者の答案テキストを入力として受け取り，その答案に対する点数を出力するタスクである．本研究で対象とする記述式問題は，小論文記述問題のような採点基準が厳密に定義されていない問題ではなく，採点基準を満たす内容が答案中に書かれているかどうかで点数が決まる記述式問題を対象とする．図1に例を示す．この例では採点基準が A-D，減点の 5 つありそれぞれを満たすかどうかで，項目ごとに点数が付与され，その合計として全体点が計算される．本研究では，全体点を出力するタスクを扱う．

2.2　自動採点モデル

本研究で用いる自動採点モデルとして，入力される答案テキスト $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ に対して，$\mathbf{x}$ の得点 $s \in C$ を出力する分類モデルを説明する．ここで，$C := \{0, 1, ..., N\}$ は，配点が N である時のラベルである．

はじめに答案テキスト $\mathbf{x}$ を，トークンごとに embedding 層によって分散表現ベクトルに変換する．次にこれを Bi-LSTM に入力し，次元数 D の n 個の隠れベクトル $\{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n\}$ を得た後，これらの平均ベクトルを計算し，文ベクトル $\widetilde{\mathbf{h}}$ を得る．

$$\widetilde{\mathbf{h}} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{h}_t \tag{1}$$

最後に，ラベルの予測分布を以下の式により得る．

$$p(y|\mathbf{x}) = \mathtt{softmax}(\mathbf{W}\widetilde{\mathbf{h}} + \mathbf{b}) \tag{2}$$

ただし，$\mathbf{W} \in \mathbb{R}^{N\times D}, \mathbf{b} \in \mathbb{R}^N$ はパラメータである．

3 自動採点における確信度の推定

本節では，確信度推定手法として分類モデルの事後確率を用いる手法と Trust Score [6] について説明する．

3.1 事後確率

一つ目の確信度推定手法として，分類モデルの事後確率を用いる手法を以下のように定義する．

$$P = \max_{y\in C} p(y|\mathbf{x}) \tag{3}$$

分類問題において確信度を推定する際には事後確率を使うのが一般的である [5] が，一方でその有効性には懐疑的な見方を示す研究も存在する [8]．したがって，自動採点タスクにおいて事後確率が有効に働くかどうかは検証の必要がある．

3.2 Trust Score

二つ目の確信度推定手法として，文献 [6] で提案された Trust Score を用いる．Trust Score は推論時の中間層のデータ点が，予測ラベルを教師信号に持つ学習データ点と近く，別のラベルを教師信号に持つ学習データ点と遠いほど，予測の信頼性は高いという仮説に基づいて予測の信頼性を測る指標である．具体的には，推論時の中間層のデータ点から予測されたラベルを教師信号に持つ学習データ群を除いた時の最近傍のデータ点への距離と予測されたラベルを教師信号に持つ最近傍の学習データ点への距離の比として算出する．

Trust Score の算出法を説明する．m 個の学習データ $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$ をそれぞれ自動採点モデルに入力し，式 1によって得られる文ベクトルの集合を $\mathcal{H} := \{\widetilde{\mathbf{h}}_1, ..., \widetilde{\mathbf{h}}_m\}$ とする．あるテストデータ $\mathbf{x}_{\text{test}}$ を入力した時に式 1によって得られる文ベクトルを $\widetilde{\mathbf{h}}_{\mathbf{x}_{\text{test}}}$，モデルの予測 $s = \arg\max_{y\in C} p(y|\mathbf{x}_{\text{test}})$ に対して，その予測クラス s に属するデータのみの文ベクトルを集めた集合 $\mathcal{H}_s = \{\widetilde{\mathbf{h}}_k \in \mathcal{H} | 1 \le k \le m \wedge y_k = s\}$ とする．このとき，あるテストデータ $\mathbf{x}_{\text{test}}$ に関する Trust Score $T(\mathbf{x}_{\text{test}}, \mathcal{H})$ は以下の式で算出される．

$$T(\mathbf{x}_{\text{test}}, \mathcal{H}) = \frac{d_c(\mathbf{x}_{test}, \mathcal{H})}{d_p(\mathbf{x}_{\text{test}}, \mathcal{H}) + d_c(\mathbf{x}_{\text{test}}, \mathcal{H})}, \tag{4}$$

表1: データの統計情報

問題	評論 1	評論 2	評論 3	評論 4	随想	小説
字数制限	70	70	50	70	50	60
配点	16	15	15	16	12	12
平均点	6.78	5.44	4.60	6.91	4.00	5.26
標準偏差	3.50	2.71	2.67	3.78	1.92	2.09

ただし，

$$d_p(\mathbf{x}_{\text{test}}, \mathcal{H}) = \min_{\widetilde{\mathbf{h}}\in\mathcal{H}_s} d(\widetilde{\mathbf{h}}_{\mathbf{x}_{\text{test}}}, \widetilde{\mathbf{h}}), \tag{5}$$

$$d_c(\mathbf{x}_{\text{test}}, \mathcal{H}) = \min_{\widetilde{\mathbf{h}}\in(\mathcal{H}\setminus\mathcal{H}_s)} d(\widetilde{\mathbf{h}}_{\mathbf{x}_{\text{test}}}, \widetilde{\mathbf{h}}) \tag{6}$$

であり，$d(\widetilde{\mathbf{h}}_{\mathbf{x}_{\text{test}}}, \widetilde{\mathbf{h}})$ は $\widetilde{\mathbf{h}}_{\mathbf{x}_{\text{test}}}$ から $\widetilde{\mathbf{h}}$ へのユークリッド距離を表す．

4 実験

本節では，事後確率を用いた確信度推定手法と Trust Score を用いた確信度推定手法が自動採点タスクにおいてどのくらい有効に機能するかを検証する．

4.1 国語長文読解問題のデータセット

本研究では，代々木ゼミナールの国語長文読解問題データセットを用いる[*1]．このデータセットは，各受験者の答案テキストと採点者によって付与された点数のペアのデータで構成される．本データセットでは，各問題に対して複数の採点項目が存在し項目点が付与されている．採点項目は複数の加点項目に加え，誤字・脱字，主述のねじれなどを対象とした減点項目から構成されているが，本実験では，加点項目のみの合計を解答の得点とした．また，実験に使用するデータの統計量を表 1に示す．なお，解答数はそれぞれ 2000 件である．

4.2 実験設定

自動採点モデルの embedding 層には，文字単位の事前学習済み BERT [9] を使用した[*2]．訓練セットとして，実験により 1600 件を使用し，開発，評価セットとして，それぞれ 200 件を使用した．採点精度の評価尺度として，Quadratic Weighted Kappa (QWK) を使用し，訓練中に開発セットに対して最も高い QWK を示した時点のモデルを評価に使用した．なお，実験結果として，5 つのランダムシードを用いて訓練したモデルの性能の平均値および最大値と最小値を報告する．

4.3 実験結果

図 2に，事後確率および Trust Score それぞれについて，確信度が高い順に評価対象に加えた時の QWK の推移を示す．ここで，横軸が 100% の時の値は確信度を用

[*1] 当データセットは以下の URL で公開予定である：https://aip-nlu.gitlab.io/resources/sas-japanese

[*2] 事前学習済み BERT は以下の URL の物を使用した：https://github.com/cl-tohoku/bert-japanese

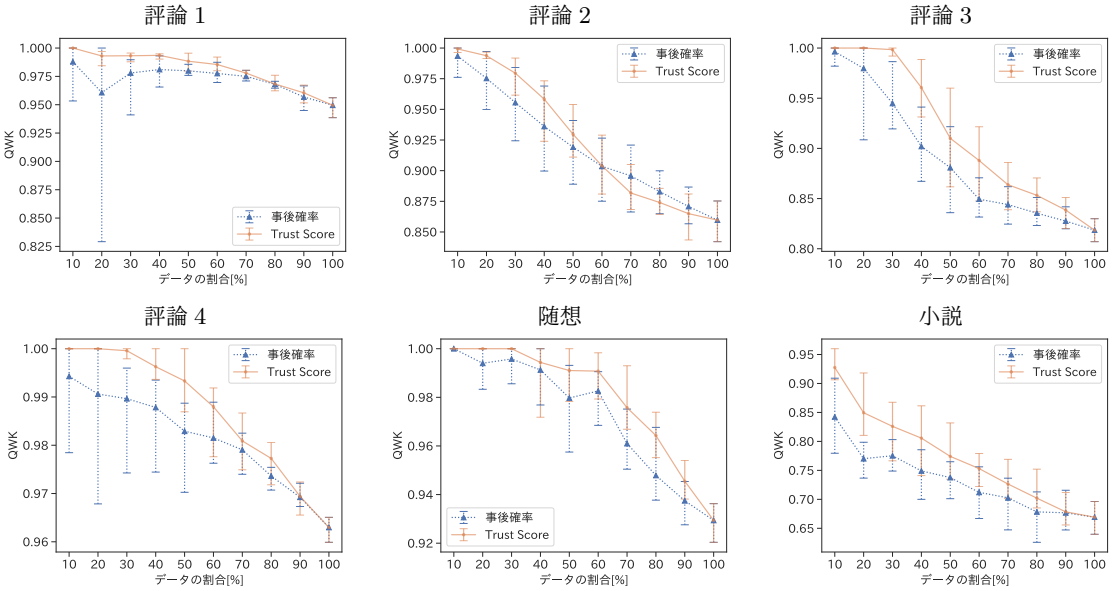

図2: Trust Score と事後確率を用いて確信度が高い順に評価対象に加えていった際の QWK の変化．点は 5 回の試行の平均値を表し，最大値と最小値を高低線で表しており，高低線が長いほど分散が大きいと考えられる．

いなかった場合の値，すなわちモデルの素の性能を示している．事後確率および Trust Score のどちらを用いた場合においても，大半の問題について確信度の高い解答群では採点精度が高く，確信度の低い解答群では採点精度が低い値となっており，どちらも確信度としてある程度機能していることがわかる．また，自動採点が難しい種類の問題に対しても確信度は有効に働いていることがわかる．図2の小説は確信度を用いない場合，他の問題より QWK が 0.2 程度低く自動採点が難しい．しかし，Trust Score を用いて．確信度上位 50% の解答に絞ることで QWK が 0.15 程度高くなる．したがって，自動採点が難しい問題に対しても確信度は有効である．

より詳細に見ていくと，事後確率は確信度が高い解答群に対しても採点精度が低下する場合があり（例．評論 1 における上位 20%），一部のデータにおいては確信度として機能しない場合があることがわかる．一方，Trust Score は事後確率よりも上位 50% 以上の解答群に対する採点精度が全ての問題において高いことから，より予測精度の高い解答と低い解答を分離する能力が高いことがわかる．さらに，Trust Score は事後確率に比べて採点精度の分散が小さいため，パラメーターの初期値のランダム性に対して頑健であると言える．

5 分析

自動採点のシステムの実応用に向けて，システムが満たすべき要請は次の 2 点であると我々は考えている．

- 重大な採点誤りを起こさないこと
- 学習に使用可能なデータが少ない状況下でも，妥当な精度で採点可能であること

そこで本節では，事後確率や Trust Score を用いることによって，これらの要請を満たすことが可能かどうか，分析を行う．また，実際に確信度を導入する際には，ある閾値を設定し，それより確信度の高い解答に対するモデルの予測結果のみを信頼する，という状況が想定される．したがって，そのような設定に基づいた分析も行う．ここでは議論を簡単にするために，対象を『評論 4』のみに絞って検証を行う[*3]．

□ 確信度による重大な採点誤りの検出

まず，重大な採点誤りを確信度を用いることによって検出できるか検証した．図3に結果を示す．いずれの確信度も上位 10% の解答においては，重大な採点ミスを除くことができている．しかし，事後確率は上位 20% までみた時点で重大な採点ミスを含んでしまうことがわかる．一方，Trust Score は上位 40% の範囲まで重大な採点ミスを取り除くことができている．

□ 学習データが限られた状況下における自動採点

自動採点システムの日常的な教育における学習支援への応用を考える上で，学習に利用可能なデータが少ない場合においても採点の信頼性を確保することが重要である．そこで，学習データとして 200 件の解答を用いた時の採点精度について，確信度を用いた時の QWK の変化を検証した．図4に結果を示す．

Trust Score を用いることによって，学習に利用する

[*3] なお，本稿で扱った 6 つの問題では，傾向は類似しており，その分析結果は以下の URL 内の本稿に関するページで公開する予定である:https://aip-nlu.gitlab.io/projects/sas-j

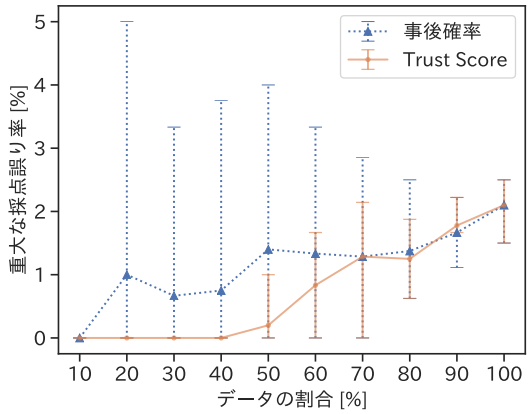

図3: 評論 4 について，Trust Score と事後確率を用いて確信度が高い順に評価対象に加えていった際の重大な採点誤りの変化

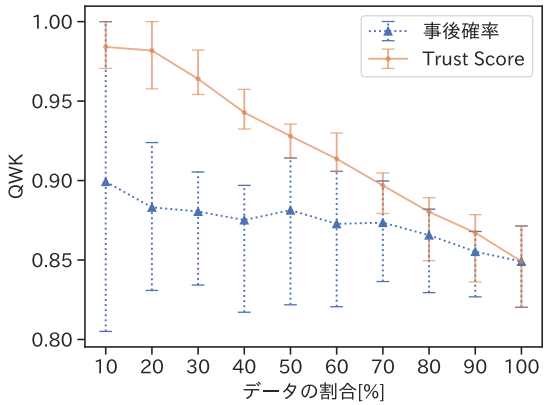

図4: 評論 4 について，Trust Score と事後確率について，確信度の高い回答から評価対象に加えた時の QWK の変化. 学習には 200 件のデータを用いた

データを 8 分の 1 に減らしても，確信度上位 40% 程度の解答に対して元の採点精度を維持していることがわかる．一方, 事後確率上位 10% の解答を用いた時の QWK と全ての解答を用いた時の QWK の差は 0.05 以内に収まっており，Trust Score に比べて，予測の正しい解答と予測の誤った解答の分離が難しくなっていることがわかる．さらに，事後確率の分散が 1600 件の時より大きく，不安定であることがわかる．学習に利用可能なデータが少ない場合に，特に Trust Score の頑健性は顕著である．

□ 閾値による低信頼度予測のフィルタリング

TrustScore を用いて閾値を使ってフィルタリングを行った時の採点誤り率と解答の割合を算出した．その結果を図5に示す．実線は重大な採点誤り率を表す．閾値を 0.6 付近に取ることで．40% 弱の解答にたいして，重大な採点誤りを完全に取り除くことが可能である．

6 おわりに

実際の教育現場に自動採点システムを導入するうえで，予測の信頼性の担保が課題になっている．本研究で

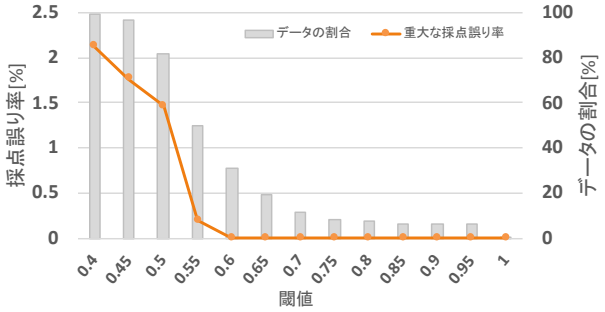

図5: 評論 4 において，Trust Score を用いて閾値を使って解答をフィルタリングした時の，解答の割合と重大な採点誤り率の変化．第 1 縦軸が採点誤り率 (実線) であり．第 2 縦軸はデータの割合 (棒グラフ) を表す．当図では最大値と最小値を掲載していない．

は予測の確信度の導入という観点からこの問題に取り組んだ．具体的には，自動採点システムの確信度を推定するにあたり，モデル自体の出力する事後確率と，学習時と推論時の中間層の情報を使う手法である Trust Score について実験を行い，その振る舞いを検証した．検証の結果，分類モデルの出力する事後確率よりも中間層のベクトルを用いた確信度の推定手法である Trust Score の方が効果的に確信度を推定できることを確かめた．

謝辞 本研究は JSPS 科研費 JP19H04162 の助成を受けたものです。また，実際の模試データを提供していただいた学校法人高宮学園代々木ゼミナールに感謝します．

参考文献

[1] Peter Foltz, Darrell Laham, and T. Landauer. “The Intelligent Essay Assessor: Applications to Educational Technology”. In: *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning* (1999).

[2] Yigal Attali and Jill Burstein. “Automated Essay Scoring with E-rater v.2.0”. In: *Journal of Technology, Learning, and Assessment* (2006).

[3] Tomoya Mizumoto et al. “Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring”. In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2019.

[4] Kaveh Taghipour and Hwee Tou Ng. “A Neural Approach to Automated Essay Scoring”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016.

[5] Dan Hendrycks and Kevin Gimpel. “A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks”. In: *CoRR* (2016).

[6] Heinrich Jiang et al. “To Trust Or Not To Trust A Classifier”. In: *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018.

[7] Philipp Oberdiek, Matthias Rottmann, and Hanno Gottschalk. “Classification Uncertainty of Deep Neural Networks Based on Gradient Information”. In: *CoRR* (2018).

[8] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. “Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images.” In: *CVPR*. IEEE Computer Society, 2015.

[9] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019.