確信度と得点の予測精度を両立する論述回答自動採点モデル

高橋 祐斗 [1]　宇都 雅輝 [1]
[1] 電気通信大学大学院
{takahashi, uto}@ai.lab.uec.ac.jp

概要

近年，深層学習を用いた論述回答自動採点モデルが高精度を達成し，実用化が期待されている．しかし，高性能な自動採点モデルであっても，得点予測を誤る可能性は依然として残っており，このことがハイステークス試験における自動採点導入の妨げの一つとなっている．この問題を解決する方法の一つとして，得点予測に加えて，予測した得点に対する確信度も出力できる深層学習自動採点モデルを用いて，予測誤りの検出を試みる研究がなされている．本研究では，確信度を推定可能な従来の深層学習自動採点モデルを拡張することで，確信度推定と得点予測の両面における性能向上を目指す．

1　はじめに

近年，学校教育では思考力・判断力・表現力といった能力が重視されるようになり，これらの能力を評価する方法の一つとして論述式試験の活用が注目されている [1, 2, 3]．一方で，論述式試験には，人手採点に伴うコストや採点の公平性担保の困難さといった問題が残る [4, 5]．このような問題を解決する手法の一つとして，深層学習を用いた自動採点手法は多数提案され，実用化が期待されている (e.g.,[6, 7, 8, 9, 10])．近年では，特に BERT（Bidirectional Encoder Representations from Transformers）[11] などの事前学習済み Transformer モデル [12] を基礎モデルとした深層学習自動採点モデルが広く活用され，高精度を達成している (e.g.,[13, 14, 15, 16, 17, 18])．

他方で，このような高性能な自動採点モデルであっても，得点予測を誤る可能性は依然として残っており，このことが自動採点導入の妨げの一つとなっている．この問題を解決する方法の一つとして，得点予測に加えて，予測した得点に対する確信度も推定できる深層学習自動採点モデルが提案されている [19]．この手法では，確信度が低い，すなわち予測得点が誤っている可能性が高い回答文を検出して，そのような回答文を人間の評価者が採点することで，採点コストの増加を抑えつつ，採点精度の向上を目指している．

確信度を推定できる従来モデル [19] は，BERT に基づく深層学習自動採点モデルを分類器として設計したモデルとなっている．分類器として設計した従来モデルは，交差エントロピー誤差を損失関数とするため，真得点と予測得点が異なる訓練データに対しては，その違いの大小によらず等しく誤りとみなして学習が進められる．他方で，一般的には深層学習自動採点モデルは回帰の枠組みで設計されることが多い．回帰の枠組みで設計したモデルでは，二乗誤差を損失関数とするため，真得点と予測得点の差の大きさを考慮してモデルの学習が進む．そのため，回帰の枠組みで設計した方が，分類に基づく手法よりも高精度な得点予測が期待できる．また，得点予測の精度向上は確信度の推定精度向上にも寄与すると期待できる．

そこで，本研究では回帰として設計した深層学習自動採点モデルに基づいて得点予測と確信度推定を行う手法を提案する．しかし，実データ実験の結果，提案モデルは得点の予測精度では優れる傾向が見られたが，確信度の推定精度は総じて分類器として設計した従来モデルに劣る傾向が確認された．この結果は，得点予測を回帰で行い，確信度推定を分類で行うことで，双方の観点において平均的に優れた性能が得られる可能性を示唆している．

そこで，本研究ではさらに，得点予測を回帰で行い，確信度推定を分類で行うことができる深層学習自動採点モデルを提案する．具体的には，BERT ベースの深層学習自動採点モデルを回帰と分類のそれぞれの方法で得点を予測する出力層を加えた二出

力型モデルとして設計し，マルチタスク学習の枠組みで訓練する．実データ実験の結果，分類と回帰のハイブリッド型提案モデルが，得点予測と確信度推定の両面において，分類または回帰として設計したモデルと同等以上の性能を達成できることが確認された．

2 従来手法

従来手法 [19] では，BERT に基づく深層学習自動採点モデルを分類器として設計したモデルにより得点予測と確信度推定を実現している．具体的には，冒頭に特殊タグ [CLS] を加えた単語系列で表される回答文 $\boldsymbol{x}$ を BERT に入力することで，[CLS] に対応する BERT の出力ベクトルである分散表現ベクトル $\boldsymbol{h}$ を取得する．次に，分散表現ベクトル $\boldsymbol{h}$ を得点段階数と同長のベクトルに $\boldsymbol{u} = \boldsymbol{W}_c\boldsymbol{h} + \boldsymbol{b}_c$（ここで，$\boldsymbol{W}_c$，$\boldsymbol{b}_c$ はパラメータ）と変換し，ソフトマックス関数に通すことで，得点 k に対する分類確率を $P_k = \exp(u_k)/\sum_{i=0}^{K-1}\exp(u_i)$（ここで，$K$ は得点段階数，u_k はベクトル $\boldsymbol{u}$ の k 番目の要素）と求める．このモデルでは，回答文 $\boldsymbol{x}$ に対する予測得点を $\hat{y} = \underset{k}{\mathrm{argmax}} P_k$ で求め，$\hat{y}$ に対応する確信度を $P_{\hat{y}}$ で評価する．

モデル学習時の損失関数には次式で定義される交差エントロピー誤差を用いる．

$$\mathcal{L}_{CE} = -\sum_{n=1}^{N}\log P_{y_n}(\boldsymbol{x}_n) \tag{1}$$

ここで，N は訓練データの数，$\boldsymbol{x}_n$ は訓練データ集合の n 番目の回答文，y_n は $\boldsymbol{x}_n$ に対して人間評価者が与えた得点，$P_{y_n}(\boldsymbol{x}_n)$ は回答文 $\boldsymbol{x}_n$ を入力したときの得点 y_n に対応する分類確率を表す．

このモデルでは，学習時の損失関数として交差エントロピー誤差を用いているため，真得点と予測得点が異なる訓練データに対して，その差の大小によらず等しく誤りとみなしてモデルの学習が進められる．一方で，一般的には深層学習自動採点モデルは回帰の枠組みで設計されることが多い．回帰では，二乗誤差を損失関数とすることにより，真得点と予測得点の差の大きさもモデル学習に活用することができる．そのため，回帰に基づくモデルを利用した方が，分類に基づく手法よりも高精度な得点予測が期待できる．また，得点予測の精度向上は確信度の推定精度向上にも寄与すると期待できる．

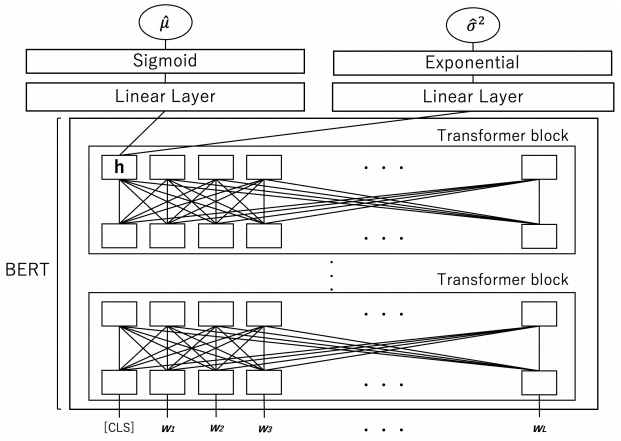

図 1 回帰の枠組みで設計した深層学習自動採点モデル図

3 回帰モデルとして設計した深層学習自動採点モデル

そこで本研究では，回帰の枠組みで設計した深層学習自動採点モデルに基づき，予測の確信度を推定するモデルを提案する．概念図を図 1 に示す．なお，基礎モデルには BERT を採用している．

提案モデルでは，出力層を正規線形回帰モデルとして設計する．具体的には，回答文 $\boldsymbol{x}$ を BERT に入力して得られる分散表現ベクトル $\boldsymbol{h}$ を 2 つの線形層に通すことで，正規線形回帰モデルの平均項を $\hat{\mu} = \mathrm{sigmoid}(\boldsymbol{W}_m\boldsymbol{h} + b_m)$（$\boldsymbol{W}_m, b_m$ はパラメータ，$\mathrm{sigmoid}(\cdot)$ はシグモイド関数を表す）と求め，分散項を $\hat{\sigma}^2 = \exp(\boldsymbol{W}_v\boldsymbol{h} + b_v)$（$\boldsymbol{W}_v, b_v$ はパラメータ）と求める．$\hat{\mu}$ は 0 から 1 の値を取るため，$(K-1)\hat{\mu}$ と 1 次変換した値を予測得点とする．また，予測得点に対する確信度は $-\hat{\sigma}^2$ で評価する．

モデル学習時の損失関数は次式で定義される．

$$\mathcal{L}_{NL} = \sum_{n=1}^{N}\left\{\frac{\|z_n - \hat{\mu}(\boldsymbol{x}_n)\|^2}{2\hat{\sigma}^2(\boldsymbol{x}_n)} + \frac{1}{2}\log 2\pi\hat{\sigma}^2(\boldsymbol{x}_n)\right\} \tag{2}$$

ここで，$\hat{\mu}(\boldsymbol{x}_n)$ と $\hat{\sigma}^2(\boldsymbol{x}_n)$ はそれぞれ，回答文 $\boldsymbol{x}_n$ を入力して得られる平均項と分散項の予測値であり，z_n は真得点 y_n を 0 から 1 の尺度に変換した値である．

3.1 評価実験

ここでは，実データ実験により，提案モデルと従来モデルを比較する．本研究では，実データとして，論述回答自動採点の研究で利用される ASAP（Automated Student Assessment Prize）データセットを使用した．ASAP は，8 つの異なる論述式問題に対して英語を母語とする米国の学生が記述した回答文と，それらの回答文に対して人間評価者が付与した

スコアで構成される．ASAP データセットの基礎統計量を付録表 3 に示す．本実験では，5 分割交差検証法で得点予測と確信度推定の性能評価を行った．得点予測の精度評価では評価指標に 2 次重み付きカッパ係数（Quadratic Weighted Kappa;QWK）と相関係数を用いた．QWK と相関係数はどちらも値が 1 に近いほど得点予測の精度が高いと言える．確信度推定の性能評価には，Reversed Pair Proportion（RPP）[20] と Area Under the Receiver Operating Characteristic（ROC-AUC）を用いた．RPP は数値が小さいほど，ROC-AUC は数値が大きいほど確信度の推定性能が高いと言える．評価指標の詳細は付録 B に示す．

3.1.1 評価結果

得点予測精度の評価結果を表 1 に示す．また，確信度推定性能の評価結果を表 2 に示す．表では，従来モデルの結果を「分類モデル」の行，提案モデルの結果を「回帰モデル」の行に示した．また，全問題についての平均性能を「平均」列に示した．得点予測精度の評価では，平均性能に着目すると，提案モデルは従来モデルより高性能であることがわかる．一方で，確信度推定性能の評価では，提案モデルの平均性能は，従来モデルに劣る結果となった．

4 回帰と分類のマルチタスク学習を組み込んだ深層学習自動採点モデル

3.1 節の実験結果から，提案モデルにより得点予測の精度は向上する傾向がみられた一方で，確信度推定の性能については従来モデルに基づく手法が高性能を示す傾向があることが明らかとなった．この結果は，得点を回帰で予測し，確信度を分類器の分類確率により推定できれば，双方の観点で平均的に優れた性能が期待できることを示唆している．そこで，本研究ではさらに，回帰と分類の二つの出力層を有し，得点を回帰で予測し，確信度を分類確率により推定する深層学習自動採点モデルを提案する．

4.1 モデル設計

回帰と分類のハイブリッド型自動採点モデルの概念図を図 2 に示す．本モデルは，BERT によって得られる分散表現ベクトル $\boldsymbol{h}$ を，シグモイド関数を活性化関数とする回帰層と，ソフトマックス関数を介して分類確率を計算する分類層の二種類の出力層に入力する構造となっている．本モデルにおける回帰層は次式となる．

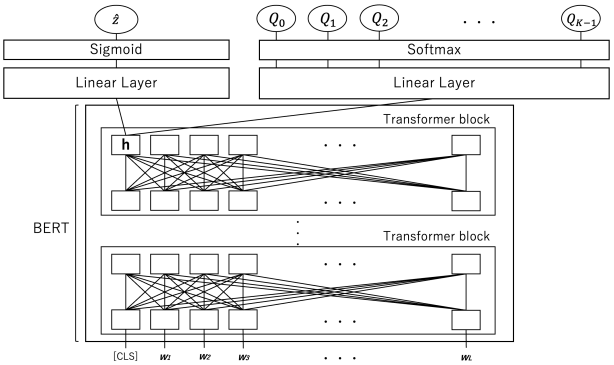

図 2 回帰と分類のハイブリッド型自動採点モデル図

$$\hat{z} = \mathrm{sigmoid}(\mathbf{W}_s\mathbf{h} + b_s) \tag{3}$$

ここで，$\mathbf{W}_s$，b_s は学習されるパラメータである．

また，分類層では，$\boldsymbol{h}$ を得点段階数と同長のベクトルに $\boldsymbol{v} = \boldsymbol{W}_t\boldsymbol{h} + \boldsymbol{b}_t$（ここで，$\boldsymbol{W}_t$，$\boldsymbol{b}_t$ はパラメータ）として変換し，ソフトマックス関数を通して得点 k に対する分類確率を次式で計算する．

$$Q_k = \frac{\exp(v_k)}{\sum_{i=0}^{K-1}\exp(v_i)} \tag{4}$$

ここで，v_k はベクトル $\boldsymbol{v}$ の k 番目の要素である

ハイブリッド型の提案モデルを用いて任意の回答文 $\boldsymbol{x}$ に対する予測得点を求める場合には，回帰層から得られる $\hat{z}$ を用いる．ただし，$\hat{z}$ は 0 から 1 の値として得られるため，$\hat{z}$ を元の K 段階尺度に 1 次変換した値を予測得点とする．予測得点に対する確信度には，分類層から得られる分類確率を用いる．具体的には，$\hat{z}$ を元の段階得点尺度に 1 次変換して四捨五入した値を $\hat{s}$ とすると，その得点に対する分類確率 $Q_{\hat{s}}$ が確信度となる．

4.2 モデル学習

本モデルはマルチタスク学習の枠組みで訓練を行う．マルチタスク学習とは複数のタスクを一つの機械学習モデルで同時に学習することで，双方のタスクの精度を高めることを目指した手法である [21]．マルチタスク学習では，一般に複数タスクの損失関数の重み付き線形和を全体の損失関数として設計する [22, 23]．提案モデルの学習においても，回帰の得点予測に基づく二乗平均誤差 $\mathscr{L}_{reg}$ と分類の得点予測に基づく交差エントロピー誤差 $\mathscr{L}_{class}$ の重み付き線形和として全体の損失関数を定義する．なお，$\mathscr{L}_{reg}$ と $\mathscr{L}_{class}$ は次式で定義される．

$$\mathscr{L}_{reg} = \frac{1}{N}\sum_{n=1}^{N}(z_n - \hat{z}(\boldsymbol{x}_n))^2 \tag{5}$$

表 1　得点予測精度の評価結果

	モデル	問題 1	問題 2	問題 3	問題 4	問題 5	問題 6	問題 7	問題 8	平均
QWK	分類モデル	0.737	*0.650*	**0.696**	0.802	*0.802*	0.777	0.757	0.543	0.720
	回帰モデル	*0.790*	**0.655**	0.662	*0.811*	0.789	*0.794*	**0.829**	*0.696*	*0.753*
	ハイブリッド	**0.812**	0.645	*0.683*	**0.812**	**0.804**	**0.806**	*0.822*	**0.724**	**0.764**
相関係数	分類モデル	0.759	**0.667**	**0.698**	*0.807*	**0.810**	0.797	*0.771*	*0.609*	0.740
	回帰モデル	*0.809*	**0.667**	0.680	**0.815**	0.802	*0.808*	**0.841**	**0.746**	*0.771*
	ハイブリッド	**0.817**	*0.659*	*0.694*	**0.815**	*0.809*	**0.809**	**0.841**	**0.746**	**0.774**

表 2　確信度推定性能の評価結果

	モデル	問題 1	問題 2	問題 3	問題 4	問題 5	問題 6	問題 7	問題 8	平均
RPP	分類モデル	*0.084*	**0.081**	*0.078*	*0.074*	**0.082**	**0.082**	**0.061**	0.062	*0.076*
	回帰モデル	0.107	0.101	0.104	0.094	0.104	0.098	0.071	*0.051*	0.091
	ハイブリッド	**0.079**	*0.082*	**0.076**	**0.070**	*0.085*	*0.085*	*0.062*	**0.033**	**0.072**
ROC-AUC	分類モデル	*0.664*	*0.632*	*0.637*	*0.640*	**0.625**	**0.639**	**0.625**	*0.660*	**0.640**
	回帰モデル	0.571	0.544	0.537	0.543	0.519	0.558	0.491	0.482	0.531
	ハイブリッド	**0.682**	**0.635**	**0.656**	**0.669**	*0.610*	*0.621*	*0.550*	**0.671**	*0.637*

$$\mathcal{L}_{class} = -\sum_{n=1}^{N} \log Q_{y_n}(\boldsymbol{x}_n) \tag{6}$$

ここで，$\hat{z}(\boldsymbol{x}_n)$ は回答文 $\boldsymbol{x}_n$ を入力して回帰層から得られる値であり，$Q_{y_n}(\boldsymbol{x}_n)$ は回答文 $\boldsymbol{x}_n$ に対して分類層で計算される得点 y_n に対応する分類確率を表す．

$\mathcal{L}_{reg}$ と $\mathcal{L}_{class}$ の重み付き線形和における重みの調整では，学習中に逐次的に重みの値を決定するアルゴリズムである Loss Scale Balancing（LSB）[23] を用いる．LSB では，各タスクの損失関数のスケールを近づけるための重み a と，各タスクの難しさに応じて値が決定される重み d の二つの調整パラメータが定義される．これらの重みの値は，学習中，逐次的に調整される．具体的には，t エポック目における全体の損失 $\mathcal{L}_M^t$ は次の式で与えられる．

$$\mathcal{L}_M^t = \lambda^t (a_{reg}^t d_{reg}^t \mathcal{L}_{reg}^t + a_{class}^t d_{class}^t \mathcal{L}_{class}^t) \tag{7}$$

ここで，$\mathcal{L}_{reg}^t$ と $\mathcal{L}_{class}^t$ は t エポック目の $\mathcal{L}_{reg}$ と $\mathcal{L}_{class}$ の値を表す．また，a_{reg}^t と d_{reg}^t は t エポック目における L_{reg} に対する二種の重み，a_{class}^t と d_{class}^t は t エポック目における $\mathcal{L}_{class}$ に対する二種の重み，λ^t は全体の損失関数のスケールを調整する値である．重みの詳細な計算方法は付録 C に示す．

4.3　評価実験

ここでは，回帰と分類のハイブリッド型の提案モデルの性能を，従来モデルおよび 3 章で提案した回帰モデルとして設計したモデルと比較する．具体的には，3.1 節と同様の実データ実験をハイブリッド型モデルに対して行った．

得点予測の精度評価結果を表 1 の「ハイブリッド」行に示す．平均性能に着目すると，QWK と相関係数の双方においてハイブリッド型の提案モデルが最も高い得点予測精度を示したことがわかる．また，確信度推定の性能評価結果では，表 2 から，RPP においてはハイブリッド型の提案モデルの平均性能が最も高く，ROC-AUC においても 2 番目に良い結果を示している．また，ROC-AUC では，最高性能の手法との差は 0.003 と微小であった．

以上から，回帰と分類のハイブリッド型モデルは，回帰か分類のいずれかのモデルを利用する場合と比べて，得点予測と確信度推定の双方において，平均的に同等以上の性能を示すことが確認できた．

5　まとめ

本研究では，分類器として設計された従来モデルを，回帰の枠組みに拡張したモデルを提案した．しかし，実験から，このモデルは得点予測の精度については改善の傾向がみられたものの，確信度推定の性能は従来手法に劣ることが確認された．この知見を踏まえ，本研究ではさらに，回帰に基づいて得点を予測し，分類に基づいて確信度を推定する深層学習自動採点モデルを提案した．実験から，回帰と分類のハイブリッド型の深層学習自動採点モデルは，回帰か分類のいずれかのモデルを利用する場合と比べて，得点予測と確信度推定の双方において，平均的に同等以上の性能を示すことが確認できた．

謝辞

本研究は JSPS 科研費 23K17585，21H00898，19H05663 の助成を受けたものです．

参考文献

[1] Yousef Abosalem. Assessment techniques and students' higher-order thinking skills. **Journal of Secondary Education**, Vol. 4, No. 1, pp. 1–11, 2016.

[2] H John Bernardin, Stephanie Thomason, M Ronald Buckley, and Jeffrey S Kane. Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability. **Journal of Human Resource Management**, Vol. 55, No. 2, pp. 321–340, 2016.

[3] Mohamed Abdellatif Hussein, Hesham Hassan, and Mohammad Nassef. Automated language essay scoring systems: A literature review. **Journal of PeerJ Computer Science**, Vol. 5, , 2019.

[4] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. In **Proceedings of AAAI Conference on Artificial Intelligence**, Vol. 19, pp. 6300–6308, 2019.

[5] Masaki Uto. A Bayesian many-facet Rasch model with Markov modeling for rater severity drift. **Journal of Behavior research methods**, Vol. 55, No. 7, pp. 3910–3928, 2023.

[6] Masaki Uto. A review of deep-neural automated essay scoring models. **Journal of Behaviormetrika**, Vol. 48, No. 2, pp. 459–484, 2021.

[7] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In **Proceedings of Empirical Methods in Natural Language Processing**, pp. 1882–1891, 2016.

[8] Fei Dong and Yue Zhang. Automatic features for essay scoring–an empirical study. In **Proceedings of Conference on Empirical Methods in Natural Language Processing**, pp. 1072–1077, 2016.

[9] Mohsen Mesgar and Michael Strube. A neural local coherence model for text quality assessment. In **Proceedings of Conference on Empirical Methods in Natural Language Processing**, pp. 4328–4339, 2018.

[10] Takumi Shibata and Masaki Uto. Analytic automated essay scoring based on deep neural networks integrating multidimensional item response theory. In **Proceedings of International Conference on Computational Linguistics**, pp. 2917–2926, 2022.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional Transformers for language understanding. In **Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4171–4186, 2019.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of International Conference on Neural Information Processing Systems**, 2017.

[13] Jianwei Li and Jiahui Wu. Automated essay scoring incorporating multi-level semantic features. In **Proceedings of International Conference on Artificial Intelligence in Education**, pp. 206–211. Springer, 2023.

[14] Haruki Oka, Hung Tuan Nguyen, Cuong Tuan Nguyen, Masaki Nakagawa, and Tsunenori Ishioka. Fully automated short answer scoring of the trial tests for common entrance examinations for Japanese university. In **Proceedings of International Conference on Artificial Intelligence in Education**, pp. 180–192. Springer, 2022.

[15] Masaki Uto, Itsuki Aomi, Emiko Tsutsumi, and Maomi Ueno. Integration of prediction scores from various automated essay scoring models using item tersponse theory. **Journal of IEEE Transactions on Learning Technologies**, pp. 1–18, 2023.

[16] Yongjie Wang, Chuan Wang, Ruobing Li, and Hui Lin. On the use of BERT for automated essay scoring: Joint learning of multi-scale essay representation. In **Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics**, pp. 3416–3425, 2022.

[17] Misato Yamaura, Itsuki Fukuda, and Masaki Uto. Neural automated essay scoring considering logical structure. In **Proceedings of International Conference on Artificial Intelligence in Education**, pp. 267–278. Springer, 2023.

[18] Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In **Proceedings of Conference on Empirical Methods in Natural Language Processing**, pp. 1560–1569, 2020.

[19] Hiroaki Funayama, Tasuku Sato, Yuichiroh Matsubayashi, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. Balancing cost and quality: an exploration of human-in-the-loop frameworks for automated short answer scoring. In **Proceedings of International Conference on Artificial Intelligence in Education**, pp. 465–476. Springer, 2022.

[20] Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. The art of abstention: Selective prediction and error regularization for natural language processing. In **Proceedings of Annual Meeting of Association for Computational Linguistics and International Joint Conference on Natural Language Processing**, pp. 1040–1051, 2021.

[21] Richard A. Caruana. Multitask learning: A knowledge-based source of inductive bias. In **Proceedings of International Conference on Machine learning**, pp. 41–48. 1993.

[22] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In **Proceedings of IEEE International Conference on Computer Vision**, pp. 2650–2658, 2015.

[23] Jae-Han Lee, Chul Lee, and Chang-Su Kim. Learning multiple pixelwise tasks based on loss scale balancing. In **Proceedings of IEEE/CVF International Conference on Computer Vision**, pp. 5107–5116, 2021.

A　実験設定

3.1 節の実験で使用した ASAP データの基礎統計量を表 3 に示す．また，本実験では，全てのモデルを PyTorch で実装した．また，BERT モデルは Huggingface で公開されている英語テキストで事前学習されたモデルである bert-base-uncased を利用した．学習アルゴリズムは AdamW（学習率 0.00005）を利用し，最大エポック数は 30，バッチサイズは 8 とし，検証データに対する損失関数の値に基づいてアーリーストッピングを行った．

表 3　ASAP データセットの基礎統計量

	問題 1	問題 2	問題 3	問題 4	問題 5	問題 6	問題 7	問題 8
受検者数	1783	1800	1726	1771	1805	1800	1569	723
回答文の平均単語数	365.7	380.7	108.6	94.4	122.1	153.3	168.1	604.9
得点段階数	11	6	4	4	5	5	31	61

B　評価指標

3.1 節における実験で利用した評価指標について説明する．

得点予測の精度評価で利用した QWK は予測得点と真得点のずれの大きさを考慮した一致度を示すカッパ係数の一つであり，自動採点モデルの性能評価に広く利用されている．

確信度推定の性能評価で利用した RPP は，「確信度が高いデータは予測得点と真得点の誤差が小さいはずである」という前提を反映した指標であり，次式で定義される．

$$RPP = \frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}\mathbb{I}[\hat{c}_i < \hat{c}_j, e_i < e_j] \tag{8}$$

ここで，B は評価データのデータ数，$\hat{c}_i$ は i 番目の評価データに対する確信度，e_i は i 番目の評価データにおける予測得点と真得点の二乗誤差，$\mathbb{I}[\cdot,\cdot]$ は 2 つの条件式が満たされるときに 1，そうでないときに 0 を返す関数である．RPP は数値が小さいほど確信度の推定性能が高いことを意味する．

確信度推定の性能評価で利用した ROC-AUC は，確信度に閾値を定めて各評価データに対する予測の正誤を判定した際の True Positive Rate（TPR）と False Positive Rate（FPR）に基づいて計算される指標である．ここで，TPR は「予測得点が実際に正しい評価データのうち，確信度に基づいて予測が正しいと判定できた割合」を意味し，FPR は「予測得点が誤っていた評価データのうち，確信度に基づいて予測が正しいと誤判定してしまった割合」を意味する．ROC は，予測が誤っているか否かを区別する確信度の閾値を，評価データ中の確信度の最大値から最小値まで変えながら TPR と FPR を求め，横軸が FPR，縦軸が TPR となるように描画することで得られる．ROC-AUC はこのようにして描いた ROC の曲線下面積に対応している．ROC-AUC は数値が大きいほど確信度の推定性能が高いといえる．

C　ハイブリッド型提案モデルの損失関数における重みの計算

式 (7) で表されたハイブリッド型提案モデルの損失関数における，重みの具体的な計算方法について説明する．式 (7) における $\mathcal{L}^t_{reg}$ に対する 2 種類の重み a^t_{reg}，d^t_{reg} と $\mathcal{L}^t_{class}$ に対する 2 種類の重み a^t_{class}，d^t_{class} はそれぞれ以下のように計算される．

$$a^t_{reg} = \frac{\mathcal{L}^{t-1}_M}{2\mathcal{L}^{t-1}_{reg}} \qquad d^t_{reg} = \frac{\mathcal{L}^{t-1}_{reg}/\mathcal{L}^{t-2}_{reg}}{\mathcal{L}^{t-1}_M/\mathcal{L}^{t-2}_M} \tag{9}$$

$$a^t_{class} = \frac{\mathcal{L}^{t-1}_M}{2\mathcal{L}^{t-1}_{class}} \qquad d^t_{class} = \frac{\mathcal{L}^{t-1}_{class}/\mathcal{L}^{t-2}_{class}}{\mathcal{L}^{t-1}_M/\mathcal{L}^{t-2}_M} \tag{10}$$

また，式 (7) における λ^t は全体の損失関数のスケールを調整する値であり，次式で計算される．

$$\lambda^t = 0.5(d^t_{reg} + d^t_{class}) \tag{11}$$