JGLUE データを用いた模範解答との差異に基づく汎用採点モデルの構築

齊藤隆浩 [1]　古宮嘉那子 [1]　石岡恒憲 [2]　中川正樹 [1]
[1] 東京農工大学　[2] 大学入試センター研究開発部
s206216x@st.go.tuat.ac.jp　kkomiya@go.tuat.ac.jp
tunenori@rd.dnc.ac.jp　nakagawa@cc.tuat.ac.jp

概要

本研究では、日本語学習済み BERT モデルを用いて短答式記述問題の採点システムを作成した。短答式記述問題の採点に関しては、個別の問題に特化してモデルを学習した岡ら [1] の研究などがあるが、本研究では答案と模範解答の両方を入力し、それらの差異をとるようにモデルを学習することで、一度の学習で複数の設問に対応することができ、かつ限られた数の答案でも採点の正解率を高める手法を提案する。学習・評価データには中学生の国語ドリルの答案データを使用した。学習データに JGLUE: 日本語理解ベンチマーク [2] のコーパスを加えたところ、国語ドリルのみで学習した場合に比べて、採点結果が大きく改善し、文ペア分類のコーパスによる学習が汎用的な採点モデルの構築に効果があることが分かった。

1　はじめに

近年、教育現場の過大な負担が社会問題として認知され、自然言語処理モデルを活用した採点業務の自動化が模索されている。採点タスクに関する研究としては、手書き文字認識と自然言語処理を組み合わせることで短答式記述問題を完全自動採点した岡らの研究 [1]、自動採点結果に説明性や整合性を持たせた佐藤ら、浅妻らの研究 [3][4]、暗黙的な表現を明示的な表現に変換することで結果が改善することを示した Bexte らの研究 [5] など様々なものがあるが、これらは全て設問ごとにモデルを作成して採点を行っている。その場合、各設問ごとに大量の学習データが必要になるため、活用場面は大規模テストなどに限られ、学校現場での導入が難しいという課題がある。モデルの学習に使用した設問と異なる設問を採点する試みとしては、1 つの模範解答と他設問の答案をもとに One-shot learning を実施した江島らの研究 [6] が挙げられるが、これは小論文の採点に適用したものであり、文字数の多い設問では精度が悪かったことが報告されている。また、小論文の正解は 1 通りではないため、模範解答と異なる解答に対応することが難しいという課題も考えられる。

本研究では短答式記述問題を題材に、BERT[7] をベースとして模範解答との差異に基づく汎用採点モデルを構築した。短答式記述問題は明示的な正解が定まっている場合が多く、模範解答との差異に基づく採点が有効であると考えられる。さらに、模範解答との差異を出力するタスクが文ペア分類タスクに類似していることに着目し、JGLUE: 日本語理解ベンチマーク [2] のコーパスに含まれる文ペアを使って学習を行うことで、学校現場で用意できる学習データの不足を補いモデルの性能を底上げすることを狙った。

2　関連研究

本研究で使用した BERT[7] は 2019 年に発表された大規模事前学習モデルである。$\text{BERT}_{\text{base}}$ の場合、入力された各トークンはまず、Input Embedding 層で 768 次元のベクトルに変換された後、12 層の Transformer[8] の Encoder ブロックで双方向の文脈を考慮した埋め込みベクトルに変換される。中間層の出力ベクトルも抽出することが可能で、下層、中層、上層のベクトルはそれぞれ表層特徴、構文特徴、意味特徴の情報を含むとされている [9]。また、トークン列の先頭には [CLS]、文末や 2 文の境界には [SEP] といった特殊トークンが付加され、とくに [CLS] は文全体の意味情報を含むとされている。

JGLUE[2] は、2022 年に提唱された日本語の言語理解ベンチマークである。2018 年に提唱された英語ベンチマークの GLUE[10] のように、一般的な言

語理解能力を測ることを目的として作成された。文章分類、文ペア分類、QA の 3 つのタスクと、それぞれに対応する 2 種類ずつ、計 6 種類のデータセットから構成される。

本研究と同様に日本語の短答式記述問題を採点した研究には岡らの [1] がある。岡らは $\text{BERT}_{\text{base}}$ を使用し、6 設問 × 約 6 万サンプルの答案データを用いて、各設問ごとに特化したモデルを作成した。その結果、全ての設問において高い正解率を達成した一方で、6 万のサンプル数でも正解率が収束しないことや、BERT の上層 (9〜12 層) の情報を抽出した場合に最も良い採点結果が得られることなどが示された。

その他の採点に関する研究として、Cahill らの研究 [11] がある。Cahill らは、答案に数式と文章の両方が含まれる数学の文章題に対する採点システムを作成した。人間に迫る水準の採点性能を達成しただけでなく、採点ルーブリックを導入することで説明可能なモデルを構築することに成功した。

3 データ

3.1 国語ドリルデータ

中学生向けの国語ドリル 3 冊に含まれる 25 問の短答式記述問題を抜き出して使用した。答案サンプルは各設問につき 55〜66 件、計 1,547 件あるが、実際に使用したのは白紙答案などを除外した 1,372 件である。各答案は (模範解答, 生徒の解答, 正誤ラベル) の組である。ただし、生徒の解答は手書き答案を人手で文字起こししたものであり、正誤ラベルは生徒の保護者による採点である。答案収集は、本学における人を対象とする研究に関する倫理審査委員会の承認を得て実施した (No.220707-04111)。

学習データに含まれる設問と異なる設問に対する採点精度を評価するため、全データを 5 設問ずつ 5 つの問題セットに分割し、それらを 3:1:1 の割合で学習データ、検証データ、評価データとする。

3.2 JGLUE データ

JGLUE[2] で公開されているデータセットのうち、文ペア分類タスクのための 2 種類のデータセット (JSTS、JNLI) を、追加の学習データとして使用した (検証データ、評価データには用いていない)。ただし次のようにして国語ドリルデータと同様の形式 (模範解答, 生徒の解答, 正誤ラベル) に変換した。

JSTS は同格な 2 文の類似度を表す 0.0〜5.0 の連続値ラベルを予測するタスクであり、データセットは (文 1, 文 2, ラベル値) からなる。(文 1, 文 2) はそのまま (模範解答, 生徒の解答) とみなす。ラベル値は表 1 のように正誤ラベルに変換する。なお事前実験では、類似度が中間値 (0.5〜4.5) のデータを使用する場合に比べて、使用しない場合の方が、学習データ数の減少に反して性能の向上が見られた。使用したデータ数は 2,618 件である。

表 1 JSTS データセットのラベル変換

変換前ラベル	変換後ラベル	用例数
0.0〜0.5	誤	2270
0.5〜4.5	(使用しない)	9833
4.5〜5.0	正	348

JNLI は前提文と仮説文の関係性を entailment(含意)、neutral(中立)、contradiction(矛盾) の 3 つに予測するタスクであり、データセットは (前提文, 仮説文, ラベル) からなる。(前提文, 仮説文) はそのまま (模範解答, 生徒の解答) とみなす。ラベルは表 2 のように正誤ラベルに変換する。なお事前実験では、neutral のデータを使用しない場合に比べて、使用した場合の方がモデルの性能が向上した。また、順序を変えて (前提文, 仮説文) を (生徒の解答, 模範解答) に対応させる事前実験も行ったが、モデルの性能に有意な差はみられなかった。使用したデータ数は 20,073 件である。

表 2 JNLI データセットのラベル変換

変換前ラベル	変換後ラベル	用例数
entailment (含意)	正	2876
neutral (中立)	誤	11193
contradiction (矛盾)	誤	6004

4 模範解答との差異に基づく汎用採点モデルの構築

本研究では、模範解答と生徒の解答のペアを入力して、答案を正と誤に分類するシステムを作成した。まず、各解答のペアを連結して BERT の Tokenizer に入力し、単語 ID の系列に変換する。その際、模範解答と生徒の解答の間には [SEP] トークンを挿入する。得られた単語 ID の系列を BERT モデルに入力することで、1 単語当たり 768 次元の埋め込みベクトルが出力される。岡ら [1] に従い、第 9〜12 層の [CLS] トークンのベクトルを抽出して concat し、3,072 次元のベクトルを得る。これを分類器 (3,072→2 次元の線形層と Softmax) に入力し、出

力が閾値 (通常は 0.5) 以上の場合に正解、それ未満の場合に不正解と判定する。

5 実験

5.1 学習の方法

学習のフェーズでは、日本語事前学習済み BERT[1] を fine-tuning することによって採点モデルを作成する。なお、下流の分類器のパラメータも同時に学習する。実験は次に示す 3 パターンを行う。

国語

国語ドリルデータのみを使って学習する。

JSTS+国語

最初に JSTS データ、その後に国語ドリルデータで学習する。

JNLI+国語

最初に JNLI データ、その後に国語ドリルデータで学習する。

いずれのデータによる学習でも、モデルの最適化アルゴリズムには SGD、損失関数には CrossEntropyLoss を使用し、それぞれ 8 エポックの学習を行う。学習率とバッチサイズは 5.2 節に述べるように Grid Search を行い決定する。また、国語ドリルデータは 3.1 節で述べたように問題別に分割し、五分割交差検定を行う。

5.2 Grid Search によるハイパーパラメータの決定

学習の際のハイパーパラメータは、次の手順で Grid Search を行い決定する。すべての手順において、国語ドリルデータは 3.1 節で述べたように五分割交差検定を行う。ハイパーパラメータの評価には検証データを用い、5 回の平均の AUC が最も高かったハイパーパラメータを採用する。

国語

国語ドリルでの学習時のハイパーパラメータを、表 3 の範囲で Grid Search を行い求める。

JSTS+国語, JNLI+国語

1. 国語ドリルでの学習時のハイパーパラメータを仮の値[2] に固定し、JGLUE データでの学習時のハイパーパラメータを、表 4 の範囲で Grid Search を行い求める。
2. JGLUE での学習時のハイパーパラメータを手順 1 で求めたものに固定し、国語ドリルでの学習時のハイパーパラメータを、表 3 の範囲で Grid Search を行い求める。

表 3 国語ドリルデータによる学習時の Grid Search の範囲

学習率	0.00005, 0.0001, 0.0002, 0.0004, 0.0008, 0.0016, 0.0032, 0.0064, 0.0125, 0.025
バッチサイズ	4, 8, 16, 32, 64

表 4 JGLUE データによる学習時の Grid Search の範囲

学習率	0.00001, 0.00003, 0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3
バッチサイズ	4, 8, 16, 32, 64

上記の方法により、ハイパーパラメータを表 5 のように決定した。

表 5 決定したハイパーパラメータ

実験	データセット	学習率	バッチサイズ
国語	国語ドリル	0.0125	16
JSTS	JSTS	0.01	64
+国語	国語ドリル	0.0032	8
JNLI	JNLI	0.0001	4
+国語	国語ドリル	0.0064	64

5.3 モデルの評価

モデルの評価指標には ACC(正解率) と AUC を併用する。

まず ACC(正解率) は、全ての問題数のうち、モデルが正誤を正しく予測した数の割合である。

$$\mathrm{ACC} = \frac{\text{正解数}}{\text{全問題数}} \tag{1}$$

また、AUC は真陽性率、擬陽性率、ROC 曲線によって説明できる。

真陽性率 (TPR) は、すべての正答答案のうち、モデルが正答と予測した数の割合である。

$$\mathrm{TPR} = \frac{\text{モデルが正答と予測した正答答案数}}{\text{すべての正答答案数}} \tag{2}$$

擬陽性率 (FPR) は、すべての誤答答案のうち、モデルが正答と予測した数の割合である。

$$\mathrm{FPR} = \frac{\text{モデルが正答と予測した誤答答案数}}{\text{すべての誤答答案数}} \tag{3}$$

1) https://huggingface.co/cl-tohoku/bert-base-japanese-v3

2) 学習率 0.0025、バッチサイズ 16

4 節で述べたように、本システムは出力が閾値 (通常は 0.5) を超えた場合に正答と判定するが、ここでは閾値を $t \in [0, 1]$ の範囲で変化させる。グラフの x 軸に FPR、y 軸に TPR をプロットし、滑らかな曲線で結んだものが ROC 曲線である。

AUC は、ROC 曲線の下側の面積である。完全なモデルでは値は 1.0 になり、ランダムな推測を行うモデルでは 0.5 に近づく。

$$\mathrm{AUC} = \int_{x=0}^{x=1} \mathrm{ROC}(x)\,dx \tag{4}$$

6 実験結果

実験はすべて、表 5 に示したハイパーパラメータを用いて行った。三つの実験の結果は表 6 のようになった。最も性能が高かった実験の値を太字で示した。なお、3 回実験を行った平均を示した。

表 6 実験結果

	国語	**JSTS+国語**	**JNLI+国語**
ACC	0.714	0.749	**0.769**
AUC	0.828	0.843	**0.861**

いずれの手法も 71%以上の ACC、82%以上の AUC で正解を予測できたが、JGLUE データを使用した場合により良い採点結果が得られ、特に JNLI データを使用した手法では ACC が 76.9%、AUC が 86.1%に達した。

7 考察

表 6 から、模範解答と生徒の解答のペアを入力する採点システムを用いることで、同じ問題の採点結果が学習データに含まれていないときでも ACC が 71%、AUC が 82%以上の高水準の採点が行えることが分かる。さらに JGLUE ベンチマークのコーパスに含まれる文ペアを使って学習を行うことで、最大で ACC が 76.9%、AUC が 86.1%に達した。

三つの実験結果に対しカイ二乗検定を行ったところ、**国語**と **JSTS+国語**、及び**国語**と **JNLI+国語**の ACC の差は、有意水準 $\alpha = 0.01$ で有意であった。このことから、JGLUE ベンチマークに含まれる JSTS または JNLI データセットを使って学習を行うことにより、モデルの性能が向上することが明らかになった。また、**JSTS+国語**と **JNLI+国語**の間では、本実験では JNLI の方が高い値を示しており、その差は有意水準 $\alpha = 0.05$ で有意である (有意水準 $\alpha = 0.01$ では有意ではない)。JSTS データよりも JNLI データを使用した方が効果が大きい可能性がある。

8 まとめ

本研究では、日本語学習済み BERT モデルを使用して短答式記述問題の採点タスクを行った。模範解答と生徒の解答と両方を入力することにより、採点対象の設問とは異なる、限られたデータ数で学習した場合でも概ね高い採点結果が得られた。JGLUE ベンチマークに含まれる JSTS または JNLI データセットを学習データに含めたところ、採点の性能が大幅に上がった。これらのデータが学習データ数の不足を補い、採点モデルの性能を向上させるのに有効であることが分かった。また、JSTS データよりも JNLI データを使用した方が効果が大きい可能性が示唆された。

今後の展望としては、Sentence BERT モデルを使用して模範解答と生徒の解答の差異を明示的に取得して採点を行い、今回の BERT モデルでの実験結果と比較することを考えている。また、大学入試共通テストの試行調査のデータを使用した実験も行う予定である。

謝辞

本研究は、科研費 23H03511 の助成を受けたものです。

参考文献

[1] 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. **AIED (1)**, pp. 180–192, 2022.

[2] 栗原健太郎, 河原大輔, 柴田知秀. JGLUE: 日本語言語理解ベンチマーク. 自然言語処理, Vol. 30, No. 1, pp. 63–87, 2023.

[3] 佐藤汰亮, 舟山弘晃, 塙一晃, 浅妻佑弥, 乾健太郎. 根拠箇所に基づく自動採点結果の説明. 言語処理学会第 28 回年次大会, pp. 459–464, 2022.

[4] 浅妻佑弥, 舟山弘晃, 松林優一郎, 水本智也, 乾健太郎. 記述式答案採点モデルの採点基準に対する整合性の検証. 言語処理学会 第 29 回年次大会, pp. 1868–1873, 2023.

[5] Marie Bexte, Andrea Horbach, and Torsten Zesch. Implicit phenomena in short-answer scoring data. In Michael Roth, Reut Tsarfaty, and Yoav Goldberg, editors, **Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language**, pp. 11–19, Online, August 2021. Association for Computational Linguistics.

[6] 江島知優, 竹内孔一. 模範答案のみを利用した日本語小論文採点支援システム. 言語処理学会 第 28 回年次大会, pp. 664–668, 2022.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding". **NAACL-HLT**, pp. 4171–4186, 2019.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

[9] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupa la, and Afra Alishahi, editors, **Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[11] Aoife Cahill, James H Fife, Brian Riordan, Avijit Vajpayee, and Dmytro Galochkin. Context-based automated scoring of complex mathematical responses. In **Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 186–192, 2020.