大規模言語モデルによる和文英訳問題の自動採点

三浦直己 [1,2]　舟山弘晃 [1,2]　松林優一郎 [1,2]　岩瀬裕哉 [1,2]　乾健太郎 [3,1,2]
[1] 東北大学　[2] 理化学研究所　[3]MBZUAI
{miura.naoki.p6, h.funa, yuya.iwase.t8}@dc.tohoku.ac.jp
y.m@tohoku.ac.jp　kentaro.inui@mbzuai.ac.ae

概要

本研究では，L2 学習においてよく用いられる和文英訳問題を対象とし，大規模言語モデル (LLM) を用いた記述式答案の自動採点手法を提案する．LLM による few-shot の文脈内学習を行うことで，既存採点モデルの運用面での課題の解決を試みた．モデルの採点性能を測定するため，実験では，モデルの予測得点と人手の採点結果の一致率を算出した．実験の結果，LLM の採点性能は少量の学習データを用いた教師あり学習に基づく既存手法を下回った．さらにエラー分析によって，LLM が採点タスクを適切に解釈できていない可能性も示唆された．これらの結果は，現行の最先端 LLM が本タスクにおける難易度と挑戦性を有していることを明らかにした．

1　はじめに

文章の翻訳問題は，第二言語 (L2) 学習の初期段階において，特に言語学的に離れた言語間で，教育ツールとしてよく利用される [1, 2]．国内の英語教育では，和文英訳問題として初期段階の英語学習に活用されている．図 1 に，我々が作成したデータセットにおける和文英訳問題の例を示す．学習者は，母国語（L1）の短い文章を学習中の言語（L2）に翻訳する．その後，この翻訳が E3 や G4 などの採点基準表内の各項目ごとに採点される．各採点項目は，特定の文法項目や語彙表現に対応しており，答案は項目ごとに分析的に採点されるため，学習者は詳細なフィードバックを受け取ることができる．

この問題形式は，L1・L2 言語間の類似点や相違点の認識を促し，相応しい表現方法への理解を深めることができる．そのため，言語学習の初期段階において学習者が基本的な文法や表現を習得するのに特に効果的である [1]．しかし，和文英訳はその問題の性質上，記述式回答となるため，答案の採点やフィードバックの返却は教育者にとって大きな負担

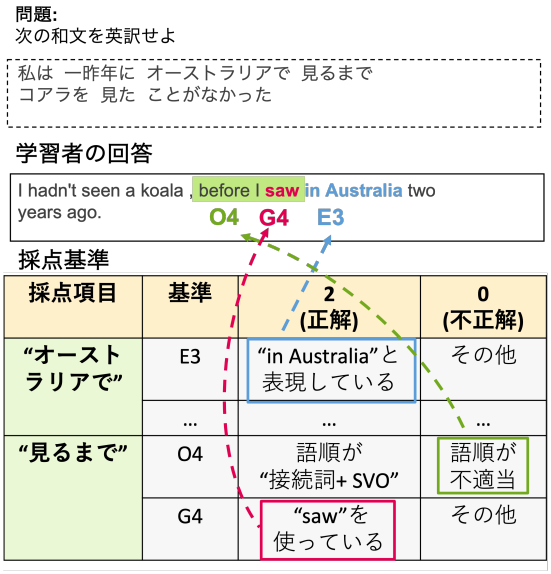

図 1　和文英訳問題の具体例（Q11）.”E”,”O”,”G”は各分析基準のカテゴリーを示し，それぞれ「表現」,「語順」,「文法」を意味する．

となる．結果として，言語習得においては反復練習が重要であるにも関わらず [3]，実施の頻度が限定されてしまう．

本研究では，和文英訳問題における添削作業の負荷削減を目的として，図 1 に示したような部分採点項目の自動採点技術を構築する．本研究で扱うタスクと密接に関連するタスクとして，英文の文法的な正しさを評価する文法誤り訂正（GEC）や機械翻訳（MT）がある．しかしながら和文英訳問題は，実際の教育現場において，教師が設定した明確な学習目標の下で活用され，教育者の意図を学習に反映しているという点で，これらのタスクとは大きく異なっている．教育者の意図を反映する必要性は，GEC システムの純粋な使用だけでは学習者の学習への取り組みを効果的に引き出せないという報告によっても支持されている [4, 5]．

本研究と同様の目的のもと行われた先行研究とし

て，[6] が挙げられる．先行研究では，和文英訳問題データセットを作成し，項目採点モデルを構築し採点を自動化した反面，不正解答案に対する採点性能の低さやモデル構築コストの高さが課題であった．

本研究では，和文英訳データセットを拡張しつつ，和文英訳問題の自動採点に大規模言語モデル (LLM) である GPT を用いることで，先行研究 [6] で明らかとなった課題に対応することを目指した．実験では，LLM の和文英訳問題における採点性能を導出するため，GPT モデルを使用し，モデルの予測結果と人手による採点結果との一致率を算出した．

実験の結果，GPT-3.5,GPT-4 の両モデルはベースラインとして設定した BERT モデルの性能を下回る結果を示し，最先端の LLM を以てしても難しいタスクであることが明らかになった．またエラー分析によって，採点モデルが和文英訳問題の採点タスクを適切に理解できていないことが示唆された．

2 和文英訳答案の自動採点

和文英訳問題の答案評価において重要なことの一つは，生徒が教師の設定した学習目標をどの程度達成しているのかを判定することである．教員は学習目標の達成度を効率よく評価するために，問題文と採点基準を設計する．各問題の採点基準は，設問内で核となる学習目標を問うような採点項目ほか，語彙表現や時制などの基本的な文法に関する採点項目を持つ．本研究は，先行研究 [6] に従い，和文英訳問題の採点を複数の採点項目を持った Short Answer Scoring (SAS) タスクとして定式化する．

採点項目に対する得点予測： 翻訳問題に対する採点項目の集合を C とする．入力された答案 $(w_1, w_2, ..., w_n)$ に対し，モデルは与えられた採点項目 $c \in C$ における得点 $s_c \in \{2, 1, 0\}$ を出力する．2, 1, 0 はそれぞれ「正解」,「部分的正解」,「不正解」に対応している．採点は採点項目ごとに行い，答案に対する総合評価は行なわない．

3 和文英訳問題データセット

実験では，[6] で作成された和文英訳データセットを，本タスクの評価データとして使用する．このデータセットは全 7 問のデータセットであり，評価データとして小規模であった．またデータ内の答案は，1 名のアノテーターによってアノテーション（採点）されたものであるため，採点の一貫性や信頼性が十分に確保されていない問題が存在した．これらの課題に対処するため，問題数を拡充した上で，一部の答案について別のアノテーターによる再採点を行い，２名の採点の一致度を計測した．

問題数の拡充： LLM による採点性能の適切な評価のために，先行研究とおおよそ同様の設計で，新たに 14 問を作成・追加し，計 21 問のデータセットに拡張した．データ全体の統計値は付録 A に示す．

採点者間の一致率の測定： 採点の一貫性や信頼性を担保するために，拡充した問題を含めた 21 問の中からランダムに 10 問を選び，各問題に対して 20 答案を別の採点者に採点させることで，採点者間の一致率を測った．各採点項目に対する採点結果 (2：正解, 1: 部分正解, 0: 不正解) の一致度には Cohen カッパ係数 [7] を、採点結果の根拠となる答案内の単語列（根拠箇所）の一致率には F_1 値を用いた．

計測の結果，全採点項目に対する採点結果の Cohen カッパ係数は 0.74 であり，採点者間で実質的一致 [8] を示した．また採点結果の根拠箇所の一致度は F 値で 0.92 となり，高い一致度 [9, 10] を示した．これらの結果から，採点結果は異なる採点者間でも高い一致率であることがわかるため，このデータセットは評価データとして一定の一貫性や信頼性がある．

4 提案手法

4.1 GPT few-shot モデル

提案手法として，GPT-3.5,GPT-4[11] の few-shot 文脈内学習 [12] によって採点を行うモデルを構築する．この方法により，各採点項目ごとに採点モデルを構築するコストと，モデルの微調整に必要な訓練データの大幅な削減が期待できる．さらに，GPT モデルは翻訳や要約のようなタスクで優れた性能を示していることから [13, 14], 和文英訳問題の採点に必要な文法や語彙知識を潜在的に備えていることも期待される．

図 2 に GPT モデルへの入力テンプレートを示す．入力は大きく 2 つの部分に分けられる．まずは事前入力にあたる部分である．ここでは, 採点タスクに関する指示文, 出力形式の説明, 翻訳される母語の原文, 採点項目と採点基準, その基準に対応する採点例が入力される．採点項目と採点基準は図 1 の採点基準に記載された部分を 1 行ごとに文字起こししたものである．さらに，採点基準とその採点例（出力例）を説明するために, 各得点ごとに 2 つの採点例を提供する．

表 1　採点項目のカテゴリー（E: 表現，O: 語順，G: 文法）ごとの各モデルの F_1 値と標準偏差．語順の採点基準には部分正解の基準が含まれていないため，対応する値は”nan” とされている．

採点項目 (項目数)	BERT			GPT-3.5			GPT-4		
	正解	部分正解	不正解	正解	部分正解	不正解	正解	部分正解	不正解
E : (96)	0.92 ± 0.15	0.64 ± 0.36	0.82 ± 0.24	0.83 ± 0.12	0.80 ± 0.23	0.60 ± 0.20	0.91 ± 0.09	0.80 ± 0.15	0.78 ± 0.20
O : (42)	0.95 ± 0.05	nan	0.79 ± 0.25	0.78 ± 0.11	nan	0.53 ± 0.21	0.87 ± 0.08	nan	0.65 ± 0.21
G : (45)	0.94 ± 0.11	0.81 ± 0.21	0.88 ± 0.13	0.82 ± 0.13	0.48 ± 0.11	0.64 ± 0.25	0.90 ± 0.08	0.62 ± 0.37	0.77 ± 0.24
全体平均	0.93	0.68	0.83	0.81	0.73	0.59	0.89	0.76	0.73

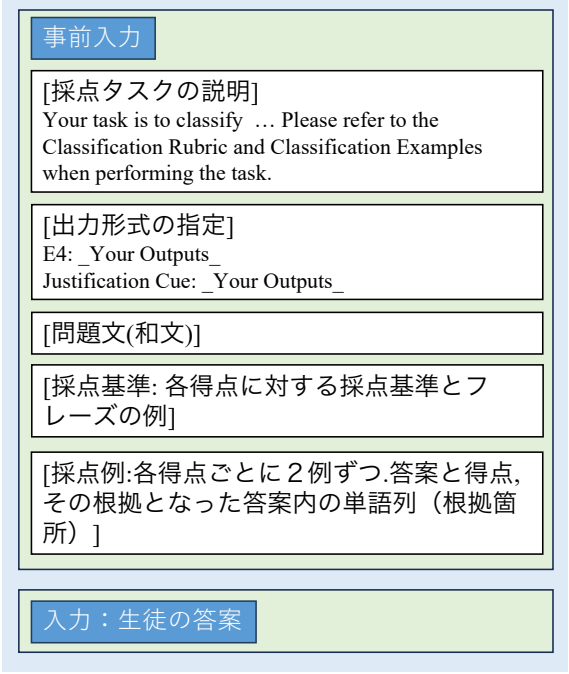

図 2　GPT few-shot モデルへの入力テンプレート

次に生徒の答案が入力される．GPT モデルはこれら 2 つの入力を利用して，指定された採点項目に対する得点と，その採点の根拠となった答案内の単語列を根拠箇所として出力する．GPT モデルへの詳細な入力例は付録 B に記載する．

4.2　BERT モデル

和文英訳答案自動採点のタスクにおけるベースラインとして，BERT[15] ベースの採点モデルである先行手法 [6] の項目採点モデルを用いた．実験では，モデルのエンコーダーに BERT[1)] を導入し，教師あり学習による手法で採点を行う．モデルの微調整では，最適化関数に Adam [16] を使用し，学習率は 0.001 に設定した．Bi-LSTM の隠れ状態の次元は 128 に設定し，学習に用いるデータ量が限られているためバッチサイズは 10 とした．

1)　https://huggingface.co/bert-base-uncased

5　実験

和文英訳答案の自動採点におけるモデルの採点性能を評価するため，実験では BERT モデルと GPT-3.5，GPT-4 を使用して答案に対する得点予測を行った．得られた予測得点と人間による採点結果との F_1 スコアを用いて評価した．

5.1　設定

実験では，和文英訳問題自動採点の先行研究 [6] に従い，モデルの性能を F_1 スコアで評価した．各問題のデータセットを訓練セット，開発セット，評価セットに 3:1:1 の比率で分割し，5 分割交差検証を行った．BERT モデルは各訓練セットにおいて 50 エポックでモデルの微調整を行い，開発セットにおいて最も良い性能を示した時のパラメータを用いた．GPT モデルについては，訓練セットから各得点ごとにランダムに 2 つの採点例を選定した．

また，一部の採点基準は，学習者にとって難易度が低く，不正解の答案が非常に少なかった．そのため，採点モデルの性能を適切に評価するために，10%以上の不正解の答案事例数を持つ採点項目だけを評価に使用した．

5.2　結果

表 1 は，評価セットにおける BERT，GPT-3.5，GPT-4 の性能を、各カテゴリー (E: 表現，O: 語順，G: 文法) の F_1 値の平均と標準偏差で示したものである．

本研究では，LLM の他タスクでの性能の高さや獲得している言語知識量から，和文英訳問題の採点において GPT モデルが優れた性能を示すという仮説を立てた．しかしながら，BERT モデルが GPT モデルを上回る結果となった．全てのモデルにおいて，正解の答案の採点精度は比較的高い性能を示し，GPT-4 は BERT モデルに匹敵する精度であった．

その一方で，不正解答案に対する採点性能に関しては，GPT-3.5 の採点性能は顕著に低く、GPT-4 でも BERT モデルに劣る性能であった．ただ興味深いことに，部分正解の答案に関しては，GPT モデルが既存手法より高い採点性能を示した．これは，部分正解の答案に対する BERT モデルを微調整するためのデータサイズが限られているためであると推測される．

また，ほぼすべての結果で標準偏差が 0.10 を超えていることが明らかになった．これは，各採点項目における採点性能にかなりのばらつきがあることを意味し，中にはかなり悪い採点精度を示した採点項目があることを示唆する．以上より，いくつかの採点項目や採点基準が，両モデルにとって困難であったことがわかる．

5.3 エラー分析

不正解答案に対する採点精度: 5.2 節で述べたように, 不正解答案に対する採点性能は, 両モデルとも正解答案への採点性能に比べて著しく低かった. これは, 正答と誤答のパターン数の違うことが原因であると推測される. 採点基準では，正解パターンの表現は少なくなるように設定されている．その一方で不正解の答案のパターンにはかなりの幅がある．そのため，不正解と分類される答案のあらゆるパターンを包含する必要性がある．しかしながら，訓練データは，正解パターンの大部分を網羅している一方で，潜在的な誤答をすべて網羅することはできなかった．また我々は GPT モデルに，内部の言語知識を活用し，不正解答案に対して包括的に対応することを期待したが，BERT モデルと比べ，良い精度を示さなかった．

GPT モデルのエラー例： 表 2 は GPT-4 による採点ミスを示している．GPT モデルが出力した根拠箇所を見ると，採点基準に直接的な関係のない部分を出力していることがわかる．このような採点ミスが GPT-4 の採点ミスの大部分を占めていた．この原因として，和文英訳問題の採点タスクについての説明や独自の採点基準に関する入力を LLM が十分に解釈できず，fewshot で入力された採点例を真似る形で出力が行われたためでないかと考えられる．そのため，採点例として入力されなかったパターンの不正解答案が現れた際，表 2 で見られるような採点ミスが発生したと推測する．事前にウェブサイト等から膨大な量の学習を行っている LLM でも，和文英訳問題の採点のようなタスクの経験がなかった．実際の教育現場を考えると，採点基準や数例の採点済み答案のみから採点を自動化することが，既存手法よりも望ましい形態であると考えられる．しかしながら，今回の実験から，LLM の内部パラメータをそのまま活用し，和文英訳問題の採点を行わせることは難しいことが示唆された．

表 2 GPT-4 の出力エラー例

入力（概要）
問題文（和文）: 私は一昨年にオーストラリアで見るまでコアラを見たことがなかった 採点項目: G1 (時制) - Past tense ”saw” is used as a verb 生徒の回答: I had never seen a koala before I have seen it two years ago in Australia .
GPT の出力 & (人手採点)
GPT の得点予測: 2 (人手採点: 0) 予測した根拠箇所: I had never (人手採点の根拠箇所: seen)

6 おわりに

本研究では，先行研究の課題を受け，和文英訳問題の自動採点タスクに LLM の活用を提案した．また LLM の採点性能の評価を充実させるため，先行研究で作成された和文英訳データセットを拡張し，その品質を担保した．LLM に採点基準や数例の採点済みデータを入力することで採点を行わせる手法は，先行研究のコストや汎用性の課題に対処するものであり，実際の教育現場での応用が望めるものであった．しかしながら，実験では最先端の LLM をもってしても，教師あり学習による手法で構築された既存の採点モデルの性能に劣る結果となった．このことから，和文英訳問題の自動採点において，現行最先端の LLM が難易度と挑戦性を有していることを明らかになった．

今後の展望として，GPT モデルを和文英訳問題の採点タスクに合わせて微調整するほか，採点タスクを GEC タスク，意味の整合性の確認，採点基準との整合性確認等に分解し，LLM の優位性を活用できるようなタスク設計を行っていくことが考えらえれる．

謝辞

本研究は JST 次世代研究者挑戦的研究プログラム JPMJSP2114 の助成を受けたものです.

参考文献

[1] Guy Cook. **Translation in Language Teaching: An Argument for Reassessment**. Oxford University Press, Oxford, March 2010.

[2] Wolfgang Butzkamm and John Caldwell. **The Bilingual Reform. A Paradigm shift in Foreign Language Teaching**. Narr Dr. Gunter, 01 2009.

[3] Diane Larsen-Freeman. On the roles of repetition in language teaching and learning. **Applied Linguistics Review**, Vol. 3, No. 2, pp. 195–210, 2012.

[4] Svetlana Koltovskaia. Student engagement with automated written corrective feedback (AWCF) provided by grammarly: A multiple case study. **Assessing Writing**, Vol. 44, p. 100450, April 2020.

[5] Jim Ranalli. L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. **Journal of Second Language Writing**, Vol. 52, p. 100816, June 2021.

[6] 菊地正弥*, 尾中大介*, 舟山弘晃*, 松林優一郎, 乾健太郎. 項目採点技術に基づいた和文英訳答案の自動採点. 言語処理学会　第 27 回年次大会 発表論文集, p. 691, 2021.

[7] Jacob Cohen. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, Vol. 20, No. 1, pp. 37–46, 1960.

[8] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. **Biometrics**, Vol. 33, No. 1, pp. 159–174, 1977.

[9] Tomoya Mizumoto, Hiroki Ouchi, Yoriko Isobe, Paul Reisert, Ryo Nagata, Satoshi Sekine, and Kentaro Inui. 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**, pp. 316–325, Florence, Italy, August 2019. Association for Computational Linguistics.

[10] Tasuku Sato, Hiroaki Funayama, Kazuaki Hanawa, and Kentaro Inui. Plausibility and faithfulness of feature Attribution-Based explanations in automated short answer scoring. In **Artificial Intelligence in Education**, pp. 231–242. Springer International Publishing, 2022.

[11] OpenAI. Gpt-4 technical report, 2023.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[13] Serge Gladkoff, Gleb Erofeev, Lifeng Han, and Goran Nenadic. Predicting perfect quality segments in mt output with fine-tuned openai llm: Is it possible to capture editing distance patterns from historical data?, 2023.

[14] Abdulkader Helwan, Danielle Azar, and Dilber Uzun Ozsahin. Medical reports summarization using text-to-text transformer. In **2023 Advances in Science and Engineering Technology International Conferences (ASET)**, pp. 01–04. IEEE, 2023.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

A　拡張後のデータセットの統計値

本研究では問題番号 8 から 21 を追加した．先行研究における作成時に 1（部分的正解）に該当する答案の収集が難しく，評価データとして扱うのが難しかったため，追加分からは 2（正解）と 0（不正解）の 2 値分類にしている．そのため問題番号 8 以降では 1（部分的正解）に該当する事例数がないため，記載されていない．

表 3　和文英訳問題データセットの統計値

	答案数	採点項目数	2	1	0
Q1	159	9	923	114	235
Q2	172	8	652	98	454
Q3	77	8	357	40	142
Q4	69	9	356	76	120
Q5	102	9	387	161	268
Q6	79	12	701	14	154
Q7	90	10	534	72	204
Q8	200	6	856		343
Q9	200	10	1324		676
Q10	200	9	1197		612
Q11	200	10	1285		715
Q12	200	8	1175		425
Q13	200	7	850		550
Q14	150	8	847		353
Q15	200	11	1347		853
Q16	200	10	1565		435
Q17	200	11	1082		1118
Q18	200	9	1220		580
Q19	200	12	1671		729
Q20	200	8	1064		536
Q21	200	12	1538		862

B　GPT モデルへの入力の詳細

表 4 に問題 11 の採点において，GPT モデルに入力されたプロンプトの詳細を示す．和文の「見るまで」に対応する採点項目を採点するためのプロンプトである．このように，GPT モデルへは，採点基準の各採点項目ごとに入力し，各得点ごとに 2 件ずつの採点例を入力した．

表 4　GPT モデルへの入力の詳細 (Q11)

PROMPT(SYSTEM)

Your task is to classify the labels corresponding to the analytic criterion from the input response. Please refer to the Classification Rubric and Classification Examples when performing the task.

*_Your Outputs_*
*E4: _Your Outputs_*
*Justification Cue: _Your Outputs_*

*_ Question_*
”私は一昨年にオーストラリアで見るまでコアラを見たことがなかった。”

*_Analytic criterion_*
E4:Tense of expressions corresponding to ” 見るまで*”*

E4: 2 -Express ” 見るまで*” as ”before I saw one(s)” , ”before I saw some”, ”before I saw them”*
E4: 0-Using ”it” instead of ”one(s)”. Otherwise.

*_Classification Examples_*
Ans：I have not seen koalas before I saw them in Australia 2 years ago .
E4： 2
justification cue：before I saw them

Ans: I had never seen koalas before I saw ones in Australia two years ago .
E4: 2
justification cue: before I saw ones

Ans: I never see koala before I saw that at Australia last year .
E4: 0
justification cue: before I saw that

Ans: I had never seen a koala until I saw it in Australia in the year before last .
E4: 0
justification cue: until I saw it

Input student response

I had never seen a koala before I saw one in Australia the year before last.