自動アノテーションを導入した G-Eval による英文要約課題評価

藤田晃輔 [1]　山田寛章 [1]　徳永健伸 [1]　石井雄隆 [2]　澤木泰代 [3]
[1] 東京科学大学　[2] 千葉大学　[3] 早稲田大学
{fujita.k.au@m,yamada@c,take@c}.titech.ac.jp
yishii@chiba-u.jp, ysawaki@waseda.jp

概要

英語学習者向け英文要約課題の自動評価のため，大規模言語モデル (LLM) を活用し，要約の内容に基づいた評価を実現する新たな手法を提案する．本研究では，Few-shot 学習，採点基準の自動展開，要約内の重要な概念や表現の自動アノテーションを組み合わせることで，要約内容に関する質の高い評価を可能にした．

1　はじめに

教育現場において，採点作業は多くの場合，教師が手作業で行うため，多大な時間と労力を要する．本研究では要約の評価を対象に，その自動化に取り組む．従来の要約の自動評価では，語彙や文法などの表層的な要素に焦点が当てられる傾向があり，要約の内容に関する評価は十分に行われていない [1]．本研究は要約内の重要な概念や要点を明確化し，採点基準に基づいて内容を評価する手法を提案する．

近年，LLM が思考過程を生成しながら問題を解決する Chain-of-Thought (CoT) [2] を取り入れて自動評価の質を向上させる手法が注目されている．Yang らは，LLM 自身が採点基準を詳細な採点ステップに展開する自動 CoT を導入し，人間の採点により近い評価が可能な G-Eval [3] を提案している．本研究では G-Eval を拡張し，自動 CoT での採点ステップ生成時に要約内の重要な概念や要点を特定するタグを定義させ，採点ステップ実行時にそのタグを用いたアノテーションを併用することで，採点基準をより的確に反映した自動評価手法を提案する．

2　関連研究

近年，LLM を用いたテキストの品質評価の研究が進められている [4]．Cheng-Han [5] らは，LLM に評価の根拠を説明させることで，評価性能が向上することを示した．これにより，LLM に採点の根拠を説明させることで，評価性能が高まることが期待できる．Lee ら [6] は，LLM を用いてエッセイを複数の特徴に分解することで，自動英文採点の質が向上することを示している．また，Yancey ら [7] は，LLM が自動英文採点において採点基準を詳細化したり，採点前に根拠を作成したりすることが有効であることを示している．これらの研究は，LLM が英文自動採点においても信頼性のある評価器として活用できる可能性を示している．吉田ら [8] は，英文自動採点において Few-shot 学習が効果的であることを示している．これにより，Few-shot 学習を導入することで，教師が意図する採点基準や評価基準をモデルにより正確に反映することが期待できる．

3　提案手法

G-Eval は，LLM 自身が採点基準を詳細な採点ステップに展開する自動 CoT の第 1 段階，そのステップに基づいて採点を実行する第 2 段階から構成される．第 1 段階では，採点基準と要約対象のテキストを LLM に入力し，より詳細に展開された採点ステップを出力する．第 2 段階では，得られた採点ステップと採点対象の要約を LLM に入力し，採点結果としてスコアを得る．

英文要約課題の人手採点を分析すると，要約内容の適切性について細分化した採点基準を用いた採点をする際，その適用の可否は採点対象要約の特定部分に対して判断しているケースが多いことがわかった．採点ステップ中に採点基準の適用対象部分を特定する注釈付けを組み込むことができれば，より人間の評価に近い採点を実現できると考えられる．そこで，G-Eval での採点ステップ生成時に要約内の重要な概念や要点を特定するタグを自動で定義させ，

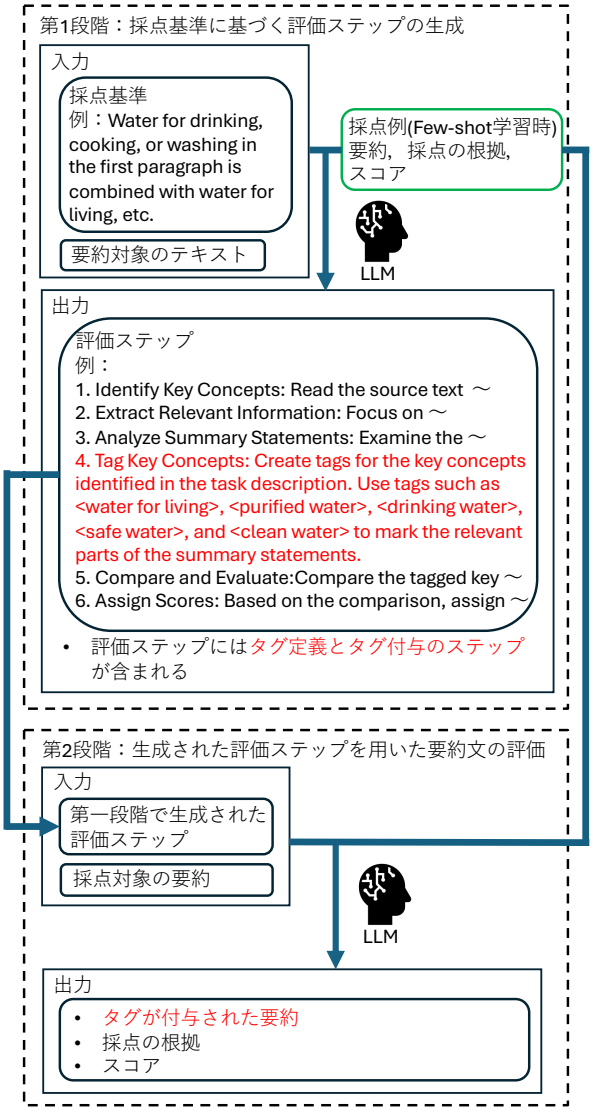

図 1　提案手法．Dynamic Tags，Few-shot (要約，採点の根拠，スコア) の場合．(4.2 参照)

採点ステップ実行時にそのタグを用いたアノテーションを併用する新たな手法（図 1）を提案する．

提案手法における第 1 段階では，G-Eval と同様の入力をとる．ただし，自動 CoT の結果得られる採点ステップ中に，タグ定義とタグを用いたアノテーションが含まれるようなプロンプト（付録 A.1）に変更する．この手法を Dynamic Tags と呼ぶ．得られる採点ステップでは，入力された採点基準に応じて LLM 自身が定義したタグと，そのタグを用いたアノテーションの指示が含まれる（図 1 赤字部分）．例えば，「飲料水，またはそれを言い換えた表現が含まれているかどうか」のような採点基準が与えられた場合，飲料水，あるいはそれに相当する表現が見つかった場合は，その箇所を ⟨drinking water⟩ タグで囲むような指示を含むステップを生成させる．

比較対象として，タグをモデルに定義させず既定のタグのみを用いる Static Tags を用意した．Static Tags では，加点対象箇所を示す ⟨bonus point⟩ と減点対象箇所を示す ⟨penalty point⟩ のタグをあらかじめ定義しておき，アノテーション時にはこの 2 種類のタグのみを用いて採点する．

提案手法の第 2 段階においても，入力は G-Eval と同様，採点ステップと採点対象の要約である．一方，出力にはスコアに加えて，第 1 段階で定義された採点ステップに則ったアノテーション結果と採点の根拠が含まれるようになる．採点の根拠は，要約が採点基準にどのように適合しているかを詳細に記述した出力である．例えば，「第一段落の ‘water for drinking, cooking, or washing’ を ‘water for living’ という表現にまとめている」という採点基準の場合，要約内の該当部分を特定し，採点基準の意図する言い換えや要約に該当していることを説明する．Cheng-Han [5] らが，採点の根拠，スコアの順に出力させることで性能が向上することを報告していることから，提案手法では，1) タグによるアノテーション済みの要約，2) 採点の根拠，3) スコアの順序で LLM が出力するようにプロンプト A.2 を設計した．

なお，元の G-Eval において Few-shot は用いられていないが，本提案手法の第二段階では，要約と，要約に対応する採点の根拠・スコアを事例[1)]として与えて文脈内学習を行う．

4　実験設定

4.1　データセット

本研究では，環境技術 Cycloclean を題材とした英文 (391 語) を東京都内の大学 (A，B) に通う学生 97 名が 80 語程度に要約したデータセットを用いる．その内訳は，2020 年に大学 A の学生が作成した 40 件（Gecchele ら [9] によるデータセットからサンプリングした 20 件を含む），2023 年に同大学の学生が作成した 25 件，および 2021 年に大学 B の学生が作成した 32 件 [10] である．97 件の要約のうち，ランダムで抽出された 25 件は二人の英語教師 (二人のうち一人は共著者)，残りの 72 件は一人の英語教師 (共著者) によって採点されたものであり，前者は最終的に統一されたスコアが出されている．本研究で

1)　厳密にはタグが付与された要約の例も必要だが本研究では与えていない．

は，評価の客観性と信頼性を確保するために，二人の英語教師によって評価された 25 件の要約を評価データとして使用し，残りの 72 件を開発データとして使用した．

採点基準　このデータセットには，要約内容の適切性を多角的に評価するための 13 個の採点基準が含まれている．

要約　要約は，要約対象のテキストを読み，その主要なアイデアと重要な詳細を約 80 語で要約するという授業課題の一環として収集された．

採点　各採点基準に対して 0，1，2 の 3 段階評価が行われている．

Few-shot 学習用の事例　Few-shot 学習に使用するデータは，開発データの中から各採点基準においてスコア 0，1，2 を代表する要約を選定し，それらの要約に対して採点の根拠を作成したものである．これら 3 組の要約，採点の根拠，スコアを使用する．

4.2　実験条件

本研究では，提案手法の有効性を検証するため，以下の条件で実験を行う．

入力　本実験では，13 個の採点基準と，二人の英語教師により評価された 25 件の要約を入力として使用する．

出力　出力として，採点基準に基づいてタグ付けされた要約，採点の根拠，および 0，1，2 の 3 段階のスコアの三つを生成する．

使用モデル　実験には GPT-4o [11]，Llama 3.1 70B Instruct [12]，Llama 3.1 405B Instruct [12]，Claude 3.5 Sonnet [13] の 4 つの LLM を用いる．

実験では，文脈内学習に関する以下の 3 通りの異なる条件を設定し，事例を与えることの有効性を検証する．

1. Zero-shot: モデルに例を与えない．
2. Few-shot (要約，スコア): 採点対象の要約とそのスコアのみをモデルに例として与え，採点の根拠を与えない．
3. Few-shot (要約，採点の根拠，スコア): 要約，採点の根拠，スコアをモデルに例として与える．

上記 3 条件に以下の 4 つの手法を組み合わせて，合計 12 種類の実験をおこなった．

1. Normal: 採点基準をそのままモデルに入力し，スコアを生成させる．
2. G-Eval: 採点基準を LLM 自身が展開して具体的な採点ステップを作成し，そのステップに従ってスコアを生成する．
3. Dynamic Tags: 重要な概念を特定するタグを自動定義させ，採点ステップ中にそのタグを用いたアノテーションうステップを組み込み，スコアを生成する．
4. Static Tags: 採点ステップ中に，事前に用意したタグ (⟨bonus point⟩，⟨penalty point⟩) を用いたアノテーションを行うようステップを組み込み，スコアを生成する．

4.3　評価方法

本研究では，0，1，2 の三段階採点を多クラス分類問題とみなし，英語教師の採点結果を正解とした場合のモデル出力の Micro-F1 スコアを評価指標として用いる．具体的には，13 個の採点基準ごとに Micro-F1 スコアを算出し，それらの平均値を全体評価とする．また，各実験は 3 回ずつ実施し，その平均値を最終スコアとする．

5　実験結果と考察

5.1　実験結果

各組合せの Micro-F1 の値を表 1 に示す．最も高い性能を示したのは，Dynamic Tags と Few-shot (要約，採点の根拠，スコア) を用いた GPT-4o モデルであった．この結果は，G-Eval で提案された採点基準の展開に加え，本研究で提案した動的なタグの定義・付与，Few-shot 学習の組み合わせが採点基準に基づく要約の評価に有効であることを示している．

5.2　手法間の比較

手法間で比較すると，Zero-shot 条件では最も高い性能を示した手法がモデルごとに異なった．一方，Few-shot 条件では，Llama 3.1-405B Instruct の Few-shot (要約，採点の根拠，スコア) を除くすべてのケースで Dynamic Tags が他の評価手法よりも高い性能を達成した．タグを定義する手法間で性能に差が出た理由として，Static Tags は採点基準に対して事前にタグを定義しているため，特定の採点基準において有効でない可能性がある．一方，Dynamic Tags は採点基準の内容に基づいてタグを動的に定義するため，採点基準に対して柔軟な対応が可能である．以上が，Dynamic Tags が高い性能を示した要因と考えられる．

モデル	手法	Zero-shot	Few-shot (要約，スコア)	Few-shot (要約，採点の根拠，スコア)
GPT-4o	Normal	0.479	0.665	0.712
	G-Eval	0.553	0.707	0.722
	Dynamic Tags	**0.589**	**0.714**	**0.763**
	Static Tags	0.562	0.652	0.728
Llama 3.1 70B Instruct	Normal	0.573	0.594	0.610
	G-Eval	**0.610**	0.604	0.628
	Dynamic Tags	0.572	**0.617**	**0.631**
	Static Tags	0.558	0.579	0.602
Llama 3.1 405B Instruct	Normal	**0.615**	0.631	0.656
	G-Eval	0.577	0.653	0.679
	Dynamic Tags	0.587	**0.667**	0.679
	Static Tags	0.604	0.658	**0.684**
Claude 3.5 Sonnet	Normal	**0.693**	0.642	0.686
	G-Eval	0.560	0.640	0.707
	Dynamic Tags	0.595	**0.670**	**0.719**
	Static Tags	0.657	0.667	0.711

表 1 提案手法の各条件 (Zero-shot，Few-shot (要約，スコア)，Few-shot (要約，採点の根拠，スコア)) と評価手法 (Normal，G-Eval，Dynamic Tags，Static Tags) における各モデルの Micro-F1 スコアの比較結果．太字で示された数値は，モデルごとの各条件における最も高い Micro-F1 スコアを表している．

5.3 Few-shot 学習の有効性

Zero-shot，Few-shot (要約，スコア)，Few-shot (要約，採点の根拠，スコア) の 3 つの結果を比較すると，Claude 3.5 Sonnet を除くモデルでは，Few-shot (要約，採点の根拠，スコア)，Few-shot (要約，スコア)，Zero-shot の順に性能が高くなる傾向がある．Claude 3.5 Sonnet においても Few-shot (要約，採点の根拠，スコア) 条件が最も高い性能を示しており，例を与えることでモデルが採点基準を理解しやすくなると同時に，採点の根拠を追加することで評価要件をさらに深く把握し，性能が向上したと考えられる．

5.4 モデル間の比較

モデル間で比較すると，GPT-4o が全体的に最も高いスコアを記録し，Claude 3.5 Sonnet が次点として良好な性能を示した．一方，Llama 3.1 70B Instruct や Llama 3.1 405B Instruct は Zero-shot 条件ではやや高いスコアを示す場合があったものの，Few-shot 条件では GPT-4o や Claude 3.5 Sonnet に劣る結果となった．さらに，Dynamic Tags と Static Tags の G-Eval に対する向上幅を比較したところ，GPT-4o と Claude 3.5 Sonnet モデルで特に大きな向上が観察された．具体的には，GPT-4o では Few-shot (要約，採点の根拠，スコア) 条件で G-Eval と比べ Dynamic Tags は 0.041，Static Tags は 0.006 の向上が見られ，Claude 3.5 Sonnet では同条件で Dynamic Tags は 0.012，Static Tags は 0.004 向上した．一方で，Llama 系モデルでは，Dynamic Tags と Static Tags の性能差は比較的小さかった．以上から，モデルごとの特性に応じた条件の選択が必要であると言える．

6 おわりに

本研究では，教育現場における要約の自動採点を効率化し，評価性能を向上させるため，LLM を活用した新たな手法を提案した．提案手法は G-Eval を拡張し，要約内の重要な概念や要点を示すタグを自動定義し，付与するとともに，Few-shot 学習と組み合わせることで，従来手法が抱えていた表層的要素への依存や，教師とモデルの採点基準との認識のずれといった課題を解消することを目指した．

実験の結果，提案手法は要約の内容を適切に反映した評価を可能にし，従来手法を上回った．特に，Dynamic Tags を活用することで採点基準をモデルに的確に伝達し，要約評価の質の向上が確認できた．また，Few-shot (要約，採点の根拠，スコア) では採点の根拠を含めた事例の提供により，モデルが採点基準をより正確に理解し，教師の採点基準に近い形で評価できた．

今後は，学習者への迅速で具体的なフィードバック提供，モデル出力の一貫性の向上，他分野への適用可能性を検討することで，英文要約評価のさらなる自動化と実用性の向上に取り組む．

謝辞

本研究は JSPS 科研費 JP24K00095 の助成を受けたものです.

参考文献

[1] Sean Xin Xu Kunpeng Zhang Yufang Wang Qi Fu Changrong Xiao, Wenxing Ma. From automation to augmentation: Large language models elevating essay scoring landscape. **arXiv preprint arXiv:2401.06431**, 2024.

[2] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. **arXiv preprint arXiv:2201.11903**, 2022. Version 6, last revised 10 Jan 2023.

[3] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. **arXiv preprint arXiv:2303.16634**, 2023.

[4] Cheng-Han Chiang and Hung yi Lee. Can large language models be an alternative to human evaluations? In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15607–15631, Toronto, Canada, 2023. Association for Computational Linguistics.

[5] Cheng-Han Chiang and Hung yi Lee. A closer look into using large language models for automatic evaluation. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 8928–8942, Singapore, 2023. Association for Computational Linguistics.

[6] Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. Unleashing large language models' proficiency in zero-shot essay scoring. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 181–198, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[7] Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. Rating short L2 essays on the CEFR scale with GPT-4. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, **Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)**, pp. 576–584, Toronto, Canada, July 2023. Association for Computational Linguistics.

[8] L. Yoshida. The impact of example selection in few-shot prompting on automated essay scoring using gpt models. In **International Conference on Artificial Intelligence in Education**, pp. 61–73, Cham, 2024. Springer Nature Switzerland.

[9] Marcello Gecchele, Hiroaki Yamada, Takenobu Tokunaga, Yasuyo Sawaki, and Mika Ishizuka. Automating idea unit segmentation and alignment for assessing reading comprehension via summary protocol analysis. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 4663–4673, Marseille, France, June 2022. European Language Resources Association.

[10] Author(s) of the article. Title of the article. **Language Testing in Asia**, 2024.

[11] OpenAI. Gpt-4o system card, 2024.

[12] Llama Team. The llama 3 herd of models, 2024.

[13] Anthropic. Claude 3.5 sonnet model card addendum, 2024. Addendum to the Claude 3 Model Card.

A　プロンプト

本節には，Few-shot (要約，採点の根拠，スコア) の実験で使用したシステムプロンプトを示す．

A.1　採点ステップ生成プロンプト

表 2 に採点ステップ生成プロンプトを示す．

表 2　採点ステップ生成のプロンプト

System
You are an expert in evaluating the quality of generated texts. Based on the following information and scoring example, generate detailed evaluation steps. Also, When creating the evaluation steps, include a step to come up with appropriate tags surrounding key concepts that are important in scoring based on the task description. Tags should be created to meet the requirements of clarity and conciseness. Also, tags should be enclosed in <>. Source Text: {source} Score range: 0,1,2 Example(Examples of summary statements and their scores and reasons. Also, Do not use the following examples directly in the evaluation step): Example1 Summary:{text[0]} Reason:{reason[0]} Score:{score[0]} Example2 Summary:{text[1]} Reason:{reason[1]} Score:{score[1]} Example3 Summary:{text[2]} Reason:{reason[2]} Score:{score[2]} Output format(Output only the following statements): Evaluation steps: 1.

A.2　要約評価プロンプト

表 3 に要約評価プロンプトを示す．

表 3　要約評価のプロンプト

System
You will be given one summary. Your task is to evaluate that summary based on the following information. Evaluation Criteria: {checklist} Score range: 0,1,2 Evalation steps: {evaluation_steps} Sorce text: {sorce} Evaluation Form: -Annotated Summary: (Annotations listed in the evaluation steps added to the original summary statement) Analyze:(Summary statement analysis) - Score:0,1,2