大規模言語モデルを用いた関連研究セクション生成

栗原龍生[1] 杉山弘晃[2] 堂坂浩二[3] 田中陸斗[3] 平博順[1]
[1]大阪工業大学大学院 [2]NTT コミュニケーション科学基礎研究所 [3]秋田県立大
{m23a10,hiroto.shitaira}@oit.ac.jp h.sugi@ieee.org
{dohsaka, m24p010}@akita-pu.ac.jp

概要

科学論文が増加している中，論文執筆支援に注目が集まっている．論文執筆支援における部分タスクとして，関連研究セクションの文章生成がある．関連研究セクションの文章生成に関する研究は，言語モデルの入力トークン数の制限から，そのほとんどが単一の引用文生成に焦点を当てていた．本研究では，入力トークン数の上限が上がった最新の大規模言語モデルを用いて，関連研究セクション全体の文章生成を試みた．また，人手による文章品質の評価と相関が高いと言われている G-EVAL と呼ばれる評価手法により，文章生成結果についての品質評価および分析を行った．

1 はじめに

科学論文は年々増加傾向にあり，自分の研究に関連する論文をすべて網羅して研究を遂行したり，論文執筆を行ったりすることが困難になりつつある．自然言語処理技術などの情報処理技術を利用した研究遂行支援や論文執筆支援への期待は以前よりも高まっている．

論文執筆支援に関しては Narimatsu ら[1]がいくつかの部分タスクに分け，各部分タスクの定義をおこなっている．本研究では，その 1 タスクである「関連研究セクション自動生成タスク」について取り上げる．このタスクは，引用予定の参考文献情報を入力として，執筆予定の論文における「関連研究(Related works)セクション」の文章を自動作成するタスクである．関連研究セクションの自動作成手法としては，大規模言語モデル（LLM）を用いた手法がある[2, 3]．しかし，これまで LLM に入力可能なトークン数には大きさの制限があり，与えられた 1 つの参考文献の引用を含んだ文（引用文）を生成する研究が多く，関連研究セクション全体を生成させ，セクション全体として生成された文章に対して結束性などの文章品質の評価を行ったものはほとんどなかった．自然言語生成の研究における生成文章の評価については，人手評価によるコストを避けるため，ROUGE や BLEU といった参照正解文との単語 n-gram を基本とした自動評価指標が広く用いられている [4]．一方，これらの単語 n-gram を基本とした自動評価指標については，その限界も指摘されている．Reiter と Belz [5]は，ROUGE や BLEU などの自動評価指標が，必ずしも文章全体の品質を評価したものにはなっていないことを明らかにしている．また，Stent ら[6]もこれらの自動評価指標では，文章の流暢性を正しく評価できないと主張している．

一方，Liu ら[7] は，大規模言語モデルの一つである GPT を利用して文章品質を評価する G-EVAL と呼ばれる評価手法を提案している．G-EVAL を用いて，要約タスクや対話生成タスクにおける文章生成結果を評価したところ，結束性，一貫性，流暢性，関連性などの様々な観点による評価に関して，ROUGE や BLEU といった既存の自動評価指標と比較して，人手評価との相関が高いことを示されている．

そこで本研究では，入力トークン数の上限が上がった最新の LLM を用いて，関連研究セクション全体の文章を自動生成するとともに，G-EVAL による生成文章の評価を行い，関連研究セクション文章生成タスクにおける LLM 利用の有効性について，評価・分析を行う．

2 関連研究

2.1 自動生成された文に関する評価

論文中の関連研究セクション文章を自動生成する研究については，生成型要約モデルを用いて，与えられた 1 つの参考文献から引用文を生成する研究がある．AbuRa’ed ら[2]は，Pointer-Generator Network と Coverage Mechanism を組み合わせた引用文の生成手法を提案している．Pointer-Generator Network は，エンコードされる前のテキスト中の単語を参照して

生成テキストに含めることができるため，未知語を含む文に対しても高品質な文生成が可能になっている．また，単語の繰り返しにペナルティを与えるCoverage Mechanism により，同じ単語が何度も繰り返されるような文の生成を抑制している．さらにXing ら[3] は，引用文前後の文脈も考慮可能なPTGEN-Cross と呼ばれる手法を提案している．これらの手法はいずれも，与えられた 1 つの参考文献に対する引用文を生成する手法である．

一方，実際の関連研究セクション文章に近い文章生成を行う研究として，Chen ら[8] の研究がある．Chen らは，グラフエンコーダを用いてターゲット論文と参考文献の関係をモデル化する手法を提案している．この研究では複数の引用を含む文生成も行っている．しかし，使用しているデータでは，参考文献の引用マーカーが一律に同一記号（&）で置き換えられているため，評価の際，生成された文が，どの参考文献についての引用文であるか分からないという課題があった．

2.2 生成文章に対する自動評価手法

先に述べたように，生成文章に対する評価手法としてはROUGE や BLEU が使われることが多いが，人手評価との相関は高くないと言われている．最近では，要約タスクなどのタスクで，BERTScore[9]のような，ニューラルネットワークによる，トークンの意味的類似度を利用したスコアも使用されるようになってきた．こうした正例と生成文との類似度に基づく評価手法は，正例の空間が狭い問題に対しては有用である．一方，関連研究セクション生成タスクのように，正例の空間が広いタスクに対しては，適切に評価することは，難しいと考えられる．そこで最近では，LLM を利用して，正解の生成結果を用意しなくても生成文章を評価する手法が提案されている． Chiang ら[10]は，物語生成タスクと敵対的攻撃生成タスクの結果に対して評価を行い，専門家による評価と評価結果に相関が高いことを示している．また，Liu ら[5] は，要約タスクおよび対話生成タスクにおける文章生成結果に対し，LLM の一つであるGPT を使って正解生成文章なしに評価を行うG-EVAL と呼ばれる評価手法を提案し，G-EVAL がその他のほとんどの自動評価手法より，人手評価との相関が高いことを示している．

3 関連研究セクション生成

3.1 文章生成モデル

ベースラインモデル（Merged-Abs）：Related Work セクションで引用された全ての参考文献の Abstract を単純に連結して出力する．ただし，”We” といった引用文中で使用すると不適当になってしまう単語については，正規表現による置き換えを行っている．
Llama2：大規模言語モデルの一つである Llama2 [11] を使用した．7 億から 70 億のパラメータまでの規模のモデルのうち，最も Instruction の理解性能が高い Llama2-Chat 70B をベースとして使用した．このモデルに対し，LongLoRA[12] を利用して長文に対応したファインチューニングを行ったモデルを最終的に使用した．
GPT-3.5 Turbo, GPT-4 Turbo：高いパフォーマンスを達成しているクローズドな大規模言語モデルとして GPT を使用した．具体的には，GPT-3.5 Turbo (gpt-35-turbo-16k-0613) と GPT-4 Turbo (gpt-4-1106-Preview) を使用して実験を行った．

3.2 文章生成の流れ

ターゲット論文と参考文献の内容を LLM に入力し，関連研究セクションを生成する．各論文の内容として，本研究では各論文の Abstract を用いる．文章生成モデルとしては，上記の 4 つのモデルのいずれかを使用した．Llama2, GPT-3.5 Turbo，GPT-4 Turbo を使用した際には，適切であると考えられる Instruction をプロンプトとして与えた（付録 A 参照）．

4 評価実験

4.1 評価用データ

評価実験に使用したデータは，プレプリントサーバ arXiv に登録されている自然言語処理に関連し，Related Work セクションが存在する文献から任意に選択して独自に作成した．各文献（ターゲット文献）について，a) 文献の Abstract，b) Related Work セクションまたは Related Work セクションの下のサブセクションの文章, c) b)の文章に含まれる参考文献すべてについての Abstract の文章，を収集して作成した．収集したターゲット文献の数は 30 である．

4.2 評価手法

生成文章に対する評価手法として，ROUGE および G-EVAL を使用した．G-EVAL については GPT-4 Turbo を用いて，関連研究セクションに適応させたものを用い，以下の 3 つの項目それぞれで，5 が最も高く，1 が最も低いとする 5 段階評価を行った．

文法 (Grammar)：生成文章に対する文法的な正しさに対する評価．

事実性 (Factuality)：生成文章が，ターゲット文献の Abstract や参考文献の Abstract の内容に反する事実が書かれていないか，不確実な内容，誤解を与える内容が書かれていないかについての評価．入力には評価対象の生成文章とともに，ターゲット文献の Abstract，および参考文献の Abstract を与えた．

一貫性 (Coherence)：生成文章が，ターゲット論文の主張内容を正しくサポートしつつ，論理的に適切な順序で参考文献を引用する文章になっているかについての評価．入力には評価対象の生成文章とともに，ターゲット論文の Abstract を与えた．

4.3 実験結果

表 1 に，4 つのモデルそれぞれで生成した，関連研究セクション文章に対する，G-EVAL による評価スコアと ROUGE スコアの値を示す．一般的には，GPT-3.5 Turbo よりの GPT-4 Turbo の方が文章生成品質は高いとされているが，ROUGE スコアで見ると，ROUGE-1, 2, L のすべてで，逆の評価になっている．一方，G-EVAL による評価では，Grammar, Coherence の項目で, GPT-3.5 Turbo より GPT-4 Turbo による生成結果の方が高い評価となっており，G-EVAL により GPT-3.5 Turbo と GPT-4 Turbo の性能の優劣を正しく評価できていることが推察できる．また, Marged-Abs モデルによる生成文章に対しては，Factuality スコアが高く，Coherence スコアは低く評価されている．Marged-Abs モデルは，参考文献間の関連性について説明した文を生成しておらず，ターゲット論文の内容をそのまま用いている．誤った内容は含まれないため，適切な評価がされているといえる．Marged-Abs モデルの生成文章に対する Grammar スコアは, 他のモデルより低くなっている．これは，Marged-Abs モデルにおいて参考文献の Abstract を連結する際，一部の参考文献について，OCR の文字認識誤りによって Abstract の読み取りに失敗した文字列が含まれていたためと考えられる．

図 1 に関連研究セクション文章生成結果の一つを示す．この結果は，文献[5]をターゲット論文としたときの Related Work セクション中の Ngram-based Metrics のサブセクションの文章を Llama2, GPT-3.5 Turbo, GPT-4 Turbo モデルで生成したときの結果である．生成時に，Instruction ととも与えた参考文献の Abstract の順番は，Papineni et al., 2002, Stent et al., 2005, Reiter and Belz, 2009, Kasai et al., 2021, Lin,2004 の並びである．

Llama2 の生成例を見ると，参考文献の出版年が欠けていることが分かる．これはほとんどの生成例にみられ，表 1 における Grammar スコアや Factuality スコアを低下させる原因の一つになっていると考えられる．また，“Papineni et al. (202) which proposes” は三人称複数形であるのに対して三人称単数形の proposes が使用され誤っている．これらは，G-EVAL で正しく指摘されている．さらに，事実性の観点では，参考文献中で使用されているデータセットに関する統計情報，達成した精度など具体的な数値が間違っている場合や，ターゲット論文の内容が，参考文献の内容と混同されているような記述がある例も確認された．その結果，Llama2 の生成文章に対する Factuality スコアが GPT-3.5 Turbo, GPT-4 Turbo と比べて低くなっていることが考えられる．また，一貫性の観点では，参考文献を紹介しているが，ターゲット論文とどのように関連しているかを示さず，参考文献を，与えた順番通りに説明しているため，一貫性のスコアが低いと考えられる．次に，GPT-3.5 Turbo モデルによる生成文章を確認すると，文法的には問題がなくまた流暢な文章であることが分かった．また，与えた参考文献に関する記述にも誤りが

表 1　実験結果

Models	Grammar	Factuality	Coherence	ROUGE-1	ROUGE-2	ROUGE-L
Marged-Abs	3.7	**4.6**	2.1	17.89	4.08	16.84
Llama2	3.9	2.5	2.5	23.13	5.23	21.53
GPT-3.5 Turbo	4.4	3.8	2.7	**25.37**	**6.41**	**23.67**
GPT-4 Turbo	**4.8**	3.8	**3.8**	24.28	6.07	22.63

Llama2:

The proposed G-EVAL framework is related to the work by Papineni et al. (202) which proposes a method of automatic machine translation evaluation that is quick, inexpensive, and language-independent, that correlates highly with human evaluation and that has little marginal cost per run. Stent et al. (05) compare the performance of several automatic evaluation metrics using a corpus of automatically generated paraphrases and show that these metrics can at least partially measure adequacy (similarity in meaning) but are not good measures of fluency (syntactic correctness). Reiter and Belz (09) review previous work on NLG evaluation and on validation of automatic metrics in NLP and then present the results of two studies of how well some metrics which are popular in other areas of NLP (notably BLEU and ROUGE) correlate with human judgments in the domain of computer-generated weather forecasts. Their results suggest that, at least in this domain, metrics may provide a useful measure of quality, although the evidence for this is not as strong as they would ideally like to see; however, they do not provide a useful measure of content quality. Lin (04) introduces ROUGE, a package for evaluating summarization systems that includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans.

GPT-3.5 Turbo

Papineni et al. (2002) propose an automatic machine translation evaluation method that is quick, inexpensive, and correlates highly with human evaluation. Stent et al. (2005) compare the performance of automatic evaluation metrics for generation systems that produce syntactic variation. Reiter and Belz (2009) review previous work on NLG evaluation and the validation of automatic metrics, specifically in the domain of computer-generated weather forecasts. Kasai et al. (2021) argue for a generalization of leaderboards, called BILLBOARDs, that track progress in language generation models and metrics for evaluation. They propose an ensemble metric that selects and combines multiple metrics based on a global analysis across generators. Lin (2004) introduces the ROUGE measures for summarization evaluation, which compare computer-generated summaries to ideal summaries created by humans.

GPT-4 Turbo

Papineni et al. (2002) introduced BLEU, an automatic, quick, and inexpensive metric for machine translation evaluation that correlates with human evaluations. Similarly, Lin (2004) proposed ROUGE, a set of metrics for evaluating the quality of summaries by comparing them to human-generated reference summaries. Both BLEU and ROUGE have been widely adopted in the field, but they primarily rely on n-gram overlap, which may not fully capture the quality of generated text, especially in tasks requiring creativity and diversity.Stent et al. (2005) and Reiter and Belz (2009) have both highlighted the limitations of automatic metrics like BLEU and ROUGE, particularly in their ability to measure fluency and content quality, respectively. Stent et al. (2005) focused on the challenge of evaluating syntactic variation in generated text, while Reiter and Belz (2009) examined the correlation of these metrics with human judgments in the domain of computer-generated weather forecasts, finding that they may not provide a useful measure of content quality.Kasai et al. (2021) addressed the limitations of evaluation methodologies by proposing bidimensional leaderboards (BILLBOARDs) that track progress in both language generation models and evaluation metrics. Their approach involves creating an ensemble metric that combines various metrics based on their correlation with human judgments, acknowledging the need for metrics to evolve as generation models improve.

図 1　関連研究セクション生成の例

ないことを確認した．しかし，Llama2 モデルによる生成文章と同様，与えられた参考文献の Abstract の順番通りに引用が行われた上，ターゲット論文との差異についての文は含まれておらず，ターゲット論文の主張をサポートした文章となっていない．最後に GPT-4 Turbo の生成例を見ると，GPT-3.5 Turbo 同様，文法には問題がなく，事実性に関しても誤りがないことが確認された．一貫性に関しては，Llama2, GPT-3.5 Turbo による生成文章とは異なり，参考文献を与えられた順番通りに引用するのではなく，BLEU, ROUGE について紹介した後，BLEU, ROUGE の評価手法の制限を引用し，Kasai らの複数メトリックスを使用した手法について順序立てて引用していることが分かった．このような例は他にも見られ，GPT-4 Turbo による生成文章の方が，Llama2, GPT-3.5 Turbo による生成文章よりも，ターゲット論文の主張をサポートするように文章が構成されていることが分かった．

5 おわりに

Llama2，GPT-3.5 Turbo，GPT-4 Turbo などの大規模言語モデルにおいて，関連研究セクション全体の文章生成を行い，引用文献の内容をカバーする一定レベルの生成は行えることを確認した．また，GPT-4 Turbo が，特に一貫性の観点で他のモデルより優れていることが分かった．さらに，G-EVAL を用いた生成文章の品質評価を行い，生成結果と評価結果を比較したところ，GPT-3.5 Turbo と GPT-4 Turbo の優位性を妥当に評価し，G-EVAL が関連研究セクション生成のタスクにおいても有用である可能性が高いことが分かった．今後はこれらの評価手法をベースに，より良い関連研究セクション生成手法について提案を行っていく予定である．

参考文献

[1] Hiromi Narimatu, Kohei Koyama, Kohji Dohsaka, Ryuichiro Higashinaka, Yasuhiro Minami, and Hirotoshi Taira. Task Definition and integration for scientific document writing support. **In Proceedings of the Second Workshop on Scholarly Document Processing**, pp.18–26, Online, June 2021. Association for Computational Linguistics.

[2] Ahmed AbuRa'ed, Horacio Saggion, Alexander Shvets, and Alex Bravo. Automatic related work section generation: experiments in scientific document abstracting. **Scientometrics** 125, 3, pp. 3159–3185. 2020.

[3] Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. Automatic generation of citation texts in scholarly papers: A pilot study. **In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.** pp. 6181–6190. 2020.

[4] Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander R Fabbri, Yejin Choi, and Noah A Smith. Bidimensional leaderboards: Generate and evaluate language hand in hand. **arXiv preprint** arXiv:2112.04139, 2021.

[5] Ehud Reiter and Anja Belz. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. **Computational Linguistics**, pp. 529–558. 2009.

[6] Amanda Stent, Matthew Marge, and Mohit Singhai. Evaluating evaluation methods for generation in the presence of variation. **In Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing**, pp. 341–351. 2005.

[7] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. **In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 2511–2522, 2023. Singapore. Association for Computational Linguistics.

[8] Chen, Xiuying, Hind Alamro, Mingzhe Li, Shen Gao, Rui Yan, Xin Gao, Xiangliang Zhang. Target-aware Abstractive Related Work Generation with Contrastive Learning. **Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.** pp. 373-383. 2022.

[9] Tianyi Zhang and Varsha Kishore and Felix Wu and Kilian Q. Weinberger and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT, **arXiv preprint** arXiv: 1904.09675v3, 2020.

[10] 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. Association for Computational Linguistics. 2023.

[11] Touvron et al. Llama 2: Open Foundation and Fine-Tuned Chat Models, **arXiv** preprint arXiv: 2307.09288, 2023.

[12] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, Jiaya Jia. LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models, **arXiv preprint** arXiv:2309.12307, 2023

付録 A　生成プロンプト例

Generate a related work section based on the provided summaries of the reference papers and the target paper. The references include {[ここに論文のリストを入力]}. In the related work section, compare these papers with the target paper and succinctly describe the positioning of the target paper in one paragraph. Group together references with similar content if applicable. Provide a concise response.
The abstract of target paper: [ここにターゲット論文のAbstractを入力]
The abstract of {[ここに参考文献にの引用を入力]}:　[ここに参考文献のAbstractを入力]
The abstract of {[ここに参考文献にの引用を入力]}:　[ここに参考文献のAbstractを入力]
…
related work:

付録 B　Llama2 による生成文章に対する G-EVAL 評価の出力例

Grammar
Specific instances of grammar errors:
1. The citation format is inconsistent (e.g., "Papineni et al. (202)" should likely be "Papineni et al. (2002)" and "Stent et al. (05)" should likely be "Stent et al. (2005)"). However, this is more of a formatting issue than a grammar error.
2. The phrase "the authors also propose analysis on the behavior" would be more grammatically correct as "the authors also propose an analysis of the behavior.“
…（略）
Score for Grammar: 4
The paragraph is well-written with only minor grammatical issues that do not significantly impede comprehension. The errors noted are relatively minor and do not detract from the overall quality of the writing.
Factuality
Specific instances of factuality errors:
1. The citation for Papineni et al. is incorrect in the paragraph. It should be Papineni et al. (2002) instead of Papineni et al. (202).
2. The citation for Stent et al. is incorrect in the paragraph. It should be Stent et al. (2005) instead of Stent et al. (05).
…（略）
Score for Factuality: 3
The paragraph is generally accurate in describing the context and findings of the target paper and related work, but there are several citation errors.
Coherence
Specific instances of coherence errors:
1. The paragraph introduces the work of Papineni et al. (202), Stent et al. (05), Reiter and Belz (09), and Lin (04) without clearly connecting how these works relate to the proposed G-EVAL framework, which can confuse the reader about the progression of ideas.
2. The transition between discussing the limitations of existing metrics and the introduction of previous studies is abrupt, lacking a smooth narrative that guides the reader through the evolution of the field.
…（略）
Score for Coherence: 3
The paragraph generally integrates references coherently, but there are instances where the connection between the target paper's claims and the cited works could be improved. The abrupt introduction of related works without clear connections to the proposed framework detract from the overall coherence.