科学文書における「間接的」引用についてのハルシネーション検出の評価

桒原龍生 [1]　杉山弘晃 [2]　堂坂浩二 [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] 秋田県立大学
{m23a10,hirotoshi.taira}@oit.ac.jp　h.sugi@ieee.org
dohsaka@aktia-pu.ac.jp

概要

近年，科学論文執筆支援技術の研究が進み，科学論文における関連研究セクションの自動生成が試みられている．関連研究セクションの自動生成を構成する技術の一つに，与えられた引用論文と作成対象の論文との関係性を考慮しつつ，引用論文の内容を簡潔に表現した引用文を自動生成する引用文生成技術がある．近年の引用文生成技術では，大規模言語モデル（LLM）が用いられることが増えているが，事実と異なる文章の生成（ハルシネーション）が課題となっている．本研究では，LLM による引用文のハルシネーション検出について，評価用ベンチマークデータを構築し，引用方法の観点から分析を行った．

1　はじめに

科学論文の数は年々増加している．この増加に伴い，研究者が自身の研究に関連する論文をすべて網羅することは，ますます困難になりつつある．この困難さは，研究の遂行や論文の執筆において特に顕著である．その結果，自然言語処理技術をはじめとする情報処理技術を活用した支援ツールへの期待が高まっている．研究遂行支援や論文執筆支援の分野では，LLM の活用が注目されている．LLM を利用することで，自動的に文章を生成することが可能となり，利用者の負担を軽減できる．しかし，このような技術には課題も存在する．その一つが，事実と異なる内容の文章を生成してしまう現象である．この現象は「ハルシネーション」と呼ばれ，利用者の信頼を損なう重大な問題となっており，生成された文章の適切性を正確に評価する方法が求められている．従来の文章生成の評価では，ROUGE [1] がよく用いられてきた．これらの指標は生成された文章と参照文章の類似度を数値化するものである．しかし，ROUGE は文法や内容の正確さを評価する能力が限られている．そのため，たとえ生成された文章にハルシネーションが含まれていたとしても，ROUGE にその問題が反映されない場合があるため，ハルシネーションを適切に検出し，評価する必要性がある．

引用文とは，ある科学論文や研究成果を参照し，その内容を要約や言及するために記述される文または，文の集合を指す．引用文は科学論文の中で重要な役割を果たしている．引用文は，一つまたは複数の引用論文を短い文章，時には一文や数文のみで要約して表現することが特徴である．その際，引用元の内容を抽象的に説明したり，引用元の論文に記載されている内容と異なる言い回しや構造を用いることが頻繁に見られる．このような特性のため，引用文は一般的な文章とは大きく異なり，その正確性を評価する際に特有の課題を伴う．ハルシネーション検出のためのベンチマークは既にいくつか存在している [2, 3]．これらのベンチマークは主に一般的な文章生成タスクを対象としており，ニュース要約や対話生成における誤情報の検出に焦点を当てている．しかし，引用文特有の特性を考慮したベンチマークは，我々が調査した限りでは存在していない．従来のベンチマークでは，引用文が持つ抽象的な表現や異なる言い回しを適切に扱うことが難しく，引用文の正確性を正確に評価するには不十分であることが考えられる．このような背景から，本研究では，引用方法の観点が異なる 2 種類のベンチマークを構築し，ハルシネーション検出の評価を行った．

2　先行研究

ハルシネーション検出のための様々なベンチマークが提案されている [2, 3]．これらのベンチマークは，生成された文章に含まれる潜在的な事実誤認や不正確な情報を評価し，それを分析するために設計されており，研究者にとって重要なツールとなっている．たとえば，Honovich ら [2] は，ニュース要約，対話，パラフレーズ，および事実検証を含む異なる 11 種類のハルシネーション検出ベンチマークを提案しており，ラベルが異なるベンチマークを，二項分類（事実と一致するか否かを判断）の枠組みで扱う方法を提示した．また，Li ら [3] は，ハルシネーション検出のベンチマーク作成において，サンプリングとフィルタリングという二段階のフレームワークを提案した．この方法を用いて，質問応答，対話，ニュース要約といった異なる分野におけるハルシネーション検出のためのベンチマークを作成した．

3　ベンチマークの構築

本研究では，引用文が引用論文と異なる記述を用いた際の，ハルシネーション検出の正確性を評価するベンチマークを構築する．このベンチマークは，引用文が引用元の論文に記載されている表記と異なる言い回しを用いる場合と，そうでない場合を評価できるように設計する．また，ハルシネーションが発生する箇所を制御し，より正確な評価を可能にするため，ベンチマークでは，引用文中の「手法や方法論」を表す名詞句に限定して，これを LLM に再生成させる形で，引用方法の観点が異なる 2 つのベンチマークを作成する．

ベンチマークの構築は，以下の 3 ステップの処理に分けられる．

1. 科学論文から，引用文を抽出する．
2. 引用文から，名詞句を抽出する．
3. 名詞句の箇所をマスク・アンマスク化し，引用方法の観点に応じて 2 種類のベンチマークに分類する．

3.1　引用文抽出

科学論文における自然言語処理やテキストマイニングタスクの大規模な評価用データセットの一つに S2ORC [4] がある．S2ORC は，8110 万本の英語で書かれた科学論文からなるデータセットであり，豊富なメタデータ，アブストラクト，参考文献データが含まれるとともに，そのうち 810 万件はオープンアクセス論文から取られた，完全に構造化された論文データとなっている．本研究では，この S2ORC データセットに含まれる論文の内，ACL Anthology から取られた 4 万 2000 件の論文を使用した．そして，これらの論文中の導入セクションと関連研究セクションに含まれる，他の論文を引用して説明を行っている引用文を全て抽出した．

3.2　名詞句抽出

引用文における「手法や方法論」を表す名詞句が，引用論文に対して，異なる表現で言い換えられることが多く，引用元との内容一致やハルシネーション検出において重要な要素となると考える．この名詞句に着目することで，引用方法の観点が異なる 2 つのベンチマークを作成することができる．具体的には，3.1 節で抽出された引用文と，各引用文と対応する引用論文を用い，GPT-4o を活用して「手法や方法論」を表す名詞句の抽出を行った．

3.3　マスク・アンマスク処理と分類

引用文中の名詞句を再生成させるため，3.2 節で抽出された名詞句を用いて，引用文から抽出された名詞句をマスクした．これにより，名詞句が一部隠された状態の引用文が作成される．この引用文を用いて，引用内容の推定を行うために，ローカル LLM を用いて推定を行う．ローカル環境で動作する LLM を使用することで、モデルの推論プロセスを完全に制御することが可能であり，推定結果の再現性が保証される．また，ローカル環境で推定された結果を詳細に分析することで，LLM が引き起こす可能性のあるハルシネーションの発生頻度やパターンを検出し，これを引用文の正確性評価に役立てることができる．本研究ではその中でも，比較的性能が高く，広く使用されている Llama 3.1 8B モデルを選択した．マスクされた引用文の中で隠された部分を推定するために，Llama 3.1 8B モデルを使用し，異なる名詞句が推定されるまで，推定を最大で 10 回繰り返した．推定された名詞句は，引用論文中で使用された表現が選ばれる場合と，そうでない場合があるため，本論文では，推定名詞句が引用論文中に含まれるか否かに応じて，得られた引用文を「直接的引用（引用論文の本文テキストに 1 回以上現れた場合）」と「間接的引用（本文テキストには現

表 1　モデルの性能比較 (/の左が直接的引用の値，右が間接的引用の値)

モデル名	プロンプト	正解率	適合率	再現率	F1 値
Meta-Llama-3.1-8B-Instruct	Vanilla	69.5 / 63.5	86.8 / 77.6	46.0 / 38.0	60.1 / 51.0
	CoT	72.5 / 68.5	68.9 / 66.4	82.0 / 75.0	74.9 / 70.4
Qwen2.5-7B-Instruct	Vanilla	86.0 / 78.5	81.6 / 74.4	93.0 / 87.0	86.9 / 80.2
	CoT	82.0 / 70.5	85.6 / 71.6	77.0 / 68.0	81.1 / 69.7
Qwen2.5-1.5B-Instruct	Vanilla	74.5 / 74.0	74.3 / 72.2	75.0 / 78.0	74.6 / 75.0
	CoT	57.0 / 65.5	61.3 / 71.8	38.0 / 51.0	46.9 / 59.6
Ministral-8B-Instruct-2410	Vanilla	88.0 / 78.0	82.8 / 71.2	96.0 / 94.0	88.9 / 81.0
	CoT	80.0 / 67.5	76.3 / 63.8	87.0 / 81.0	81.3 / 71.4

れない場合)」に分類した．

3.4　ハルシネーション判定付与と正規化

GPT-4o を用いて，3.3 節で作成された引用文と引用論文を与え，「ハルシネーション」と「非ハルシネーション」の自動分類を実施した．この分類作業では，推定された引用文が事実に基づいた正確な内容を含むかどうかを判別することを目的とした．「ハルシネーション」に該当するものは，事実に反する情報を含むか，あるいは元の文献の内容にはない情報を含む引用であると判断される．一方，「非ハルシネーション」と分類されるものは，元の文献と一致する正確な情報を反映した引用であると見なされる．さらに，引用の種類に応じてデータを整理する．具体的には，直接的引用と間接的引用に分け，さらにそれぞれのカテゴリにおいて，ハルシネーションと非ハルシネーションの事例を各 100 件ずつ選定する．このようにして，引用の種類ごとに均等なラベル分布が得られるようにデータを整形し，正規化を行った．

4　評価実験

本研究で作成した直接的引用と間接的引用のベンチマークを用いて，LLM のハルシネーション検出の評価を行う．

4.1　実験設定

評価 LLM

評価には，以下の 4 つの LLM を使用する．

1. Meta-Llama-3.1-8B-Instruct [5]
2. Qwen2.5-7B-Instruct [6]
3. Qwen2.5-1.5B-Instruct [6]
4. Ministral-8B-Instruct-2410 [7]

プロンプト設計

ハルシネーションの検出には、Dong ら [8] が提案したプロンプトを使用する．さらに、同じプロンプトを Chain-of-Thoughts（CoT）形式に拡張することで、検出精度の向上を図る．このアプローチでは、単に結果を得るだけでなく、思考の過程を明示化することによって、ハルシネーションの兆候をより正確に捉えることが可能となる．

評価指標

評価指標としては，正解率，適合率，再現率，F1 値を使用する．これらの指標は，ハルシネーション検出の精度を定量的に測るために重要な役割を果たす．正解率は，モデルが「ハルシネーション」と「非ハルシネーション」の 2 値分類において，正しく判定できた引用文の割合を示す．この指標は，全体の予測に対して正解した割合を評価するものである．次に，適合率について述べる．適合率は，モデルが「ハルシネーション」と予測した引用文が，実際に「ハルシネーション」であった割合を示す．この指標は，モデルが予測した「ハルシネーション」の予測結果のうち，どれだけ正確だったかを測るものである．再現率は，正解が「ハルシネーション」である引用文の中で，モデルが正しく「ハルシネーション」と予測した割合を示す．この指標は，実際の「ハルシネーション」の引用文をモデルがどれだけ見逃さずに検出できたかを測るものである．最後に，F1 値について説明する．F1 値は，適合率と再現率の調和平均であり，両者のバランスを取る指標である．F1 値が高いほど，モデルの予測が全体的に

優れていることを意味する．

4.2 実験結果

実験結果は表 1 に示す．結果，間接的引用の精度は，直接的引用に比べて多くの LLM において低いことが分かる．間接的引用では，引用論文の情報が抽象化されるため，LLM が正確にハルシネーションを検出する能力が低下することに起因していると考えられる．

プロンプトの違いが精度に与える影響についても考察を行った．結果，Llama-3.1 を除く他のモデルでは，CoT を使用した場合に精度が低下する傾向が確認された．この結果について，数学のように本質的に段階的な推論を必要とするタスクでは，CoT によるプロンプトが精度の向上に寄与するが，ハルシネーション検出タスクにおいては，明示的な推論を必要とせず直感的な判断で十分である場合が多いため，過剰な推論（overthinking）による精度低下が発生した可能性が考えられる．

5 おわりに

本研究では，引用文が引用論文とは異なる言い回しで表現された場合におけるハルシネーション検出に焦点を当てた．直接的引用と間接的引用に基づくベンチマークを構築し，その性能を LLM を用いて評価した．この評価を通じて，異なる記述が用いられている場合に検出が不正確になることが明らかとなった．

参考文献

[1] Chin-Yew Lin. Ttext summarization branches out. **Association for Computational Linguistics**, pp. 74–81, 2004.

[2] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. **Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering**, pp. 161–175, 2022.

[3] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. **Association for Computational Linguistics**, pp. 6449–6464, 2023.

[4] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4969–4983, 2020.

[5] Meta. Llama 3.1 - 8b instruct, (2024-12 閲覧). https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

[6] Qwen Team. Qwen2.5: A party of foundation models, (2024-12 閲覧). https://qwenlm.github.io/blog/qwen2.5/.

[7] Mistral AI. Ministral-8b-instruct-2410, (2024-12 閲覧). https://huggingface.co/mistralai/Ministral-8B-Instruct-2410.

[8] Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. BAMBOO: A comprehensive benchmark for evaluating long text modeling capacities of large language models. **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 2086–2099, 2024.

A ハルシネーション判定付与のプロンプト

ベンチマーク構築でハルシネーション判定付与に使用したプロンプトを表 2 に示す．LLM に推定された名詞句は＆のプロフィックスがつけられている．

表 2　ハルシネーション判定付与のプロンプト

ソース文書と仮説文が与えられます．あなたのタスクは，その仮説文がソース文書に基づいて，正しいかどうかを＊理由付き＊で判断することです．ソースに基づいており，事実に忠実な場合を「@@正確@@」，ソースにない情報を含む，または事実に反する場合を「@@幻覚@@」とします．" - **ソース文書**: {ここに引用論文を入れる} - **仮説文**: {ここに推定された名詞句を含む引用文を入れる} 特に，&で囲まれた部分には，注意を払ってください． **理由**: