Fact-checking のための補足コンテキストによる情報の拡充

籾井裕貴　滝口哲也　有木康雄
神戸大学大学院システム情報学研究科
235x075x@stu.kobe-u.ac.jp　{takigu, ariki}@kobe-u.ac.jp

概要

近年では，SNS や大規模言語モデルの普及に伴い，誤った情報が拡散しやすくなっている．誤った情報の影響を抑えるには，Fact-checking が重要となる．Fact-checking は，判定対象が誤っていないか，外部のデータベースに基づいて判定するタスクである．このタスクでは，判定に必要な情報を外部データベースから検索することが求められる．本研究では，情報検索を行った後に，補足コンテキストの生成を行うことで，より多くの判定に有用な情報の取得を試みた．

1　はじめに

近年では，SNS やオンラインを通じて誤った情報が拡散しやすくなった [1, 2]．また，大規模言語モデル (LLM: Large Language Model) の発展に伴い，多くの対話システムに LLM が活用されているが，ハルシネーションのような誤った情報を含む応答を生成する危険性がある [3]．そこで，文に誤った情報が含まれていないかを自動で判定する，Automated Fact-checking が重要となる．Automated Fact-checking は，文に誤った情報が含まれていないか，外部データベースから判定に必要な情報 (evidence) を検索し，検索した evidence に基づいて判定するタスクである [1]．これにより，真偽を確認したい情報や対話システムの応答文を検証することで，不正確な情報の拡散を防ぐことができる．以降は，Automated Fact-checking を単に Fact-checking と呼ぶ．

Fact-checking に関する様々な Shared Task も開催されており [4, 5]，AVeriTec Shared Task[6] において，我々は図 1 に示すシステムを開発した [7]．このシステムは，RAG-fusion[8] を参考にし，判定対象 (claim) から，evidence 検索用のクエリを生成することで，多様な観点からの evidence 検索を目的としている．しかし，検索される evidence が文単位となっており，前後の文脈が無視されてしまうという欠点がある．そこで，本研究では，evidence となる文を取得した後に，前後の文脈を含めた補足コンテキストを生成し，evidence の補足を試みた．このコンテキストにより，より多くの情報に基づいて判定を行うことが可能となる．

検索単位をチャンクにすることでも文脈を利用できる [9]．しかし，実世界での Fact-checking では，claim に応じて web から情報を取得する場合が想定される．つまり，情報検索対象となる文書を事前に固定できず，判定のためにチャンクを作成し直すこととなる．判定を行うたびにチャンクを作成する必要がある以上，より claim と関連するチャンクを得るため，今回は既に取得した文を補足するというアプローチをとった．実験結果では，補足コンテキストを最終判定モデルに与えることで，判定精度が向上する可能性を示した．

2　関連研究

2.1　AVeriTec Shared Task

AVeriTec Shared Task[1] は，AVeriTec データセット [6] を使用した Fact-checking に関する Shared Task であり，EMNLP 2024 のワークショップ[2] で開催された．本研究で扱う Fact-checking は，この Shared Task で設定された枠組みに基づいている．

まず，evidence の取得は，与えられた claim を用いて得られる web 検索結果の文書から行う．その evidence に基づいて，claim が “Supported”, “Refuted”, “Not Enough Evidence (NEE)”, “Conflict” であるかを判定する．これら 4 つはアノテーションされた claim の判定ラベルである．“NEE” は判定に必要な情報が不足している場合を示し，“Conflict” は複数 evidence 間で，Support されるか Refute されるか異なる場合を示す．evidence は，質問とその回答という形式である．例えば，“*In a letter to Seteve Jobs,*

1） https://fever.ai/task.html
2） https://fever.ai/workshop.html

Sean Connery refused to appear in an apple commercial.” という claim は，evidence として，{Question1: “*Where was the claim first published first*”，Answer1: “*It was first published on Soccopertino*”}，{Question2: “*What kind of website is Scoopertino*”，Answer2: “*Sc-oopertino is an imaginary news organization devoted to ferreting out the most relevant stories in the world of Apple*”}が与えられた場合，判定ラベルは “Refuted” である．

なお，web 検索を行って得られた文書集合はあらかじめ与えられており，文書集合から，判定に有用な evidence を取得する．

2.2 Contextual Retrieval

チャンクに対してコンテキスト生成を行う手法としては，Contextual Retrieval[10] が存在している．これは，Retrieval-Augmented Generation (RAG)[11] において，応答に必要な情報の検索精度を向上させるための手法である．検索対象の文書をチャンクに分割した後，各チャンクと分割前の文書全体を LLM に入力し，チャンクの説明コンテキストを生成させる．各説明コンテキストを対応するチャンクの前に付与することで，必要な情報の検索精度が向上することを示している．

本論文のアプローチとは，コンテキストを作成するタイミングが異なる．Fact-checking を行う場合，claim に応じて web 検索結果が変化し，evidence を抽出する対象も変化するので，事前に説明コンテキストを作成できない．そこで，今回は検索後に補足コンテキストを作成する．

2.3 質問生成による Fact-checking システム

図 1 に，我々が AVeriTec Shared Task で開発した Fact-checking システムを示す．このシステムは，(1) 文書検索 (2) 質問生成+文検索 (3) 最終判定の 3 段階のステップで構成される．文書検索では，埋め込みベクトルを使用して，claim と関連のあるドキュメントを 50 件取得する．質問生成+文検索では，claim から情報取得のための検索クエリを最大 3 件生成し，各クエリに対する回答候補の文を 3 件検索する．claim から検索用の質問を生成することで，多様な観点からの evidence 検索を行う．最終判定では，生成した検索クエリと回答候補の文を GPT-4o に入力し，最終判定ラベルを生成する．GPT-4o に与えるプロンプトには，誤った情報の検出率を向上させるため，“Refute の可能性がわずかでもあれば，

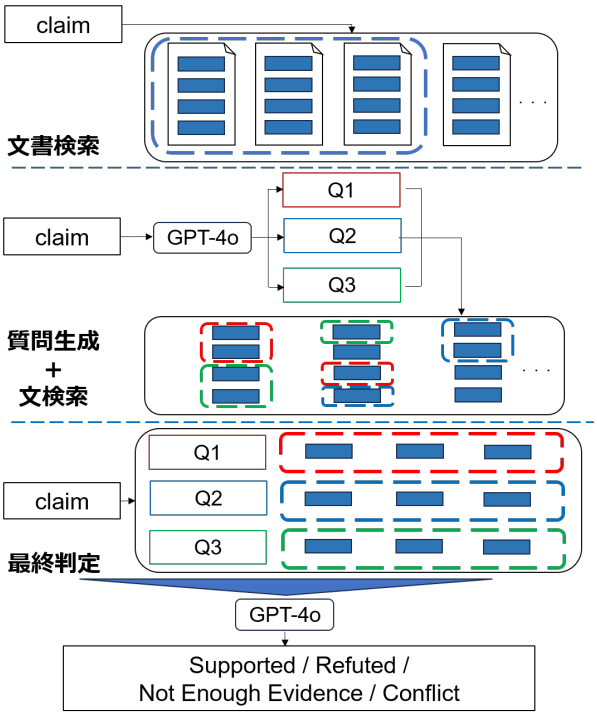

図 1 質問生成による Fact-checking システム: 文書検索では 50 個の文書を取得する．質問生成+文検索では，最大で 3 件の質問が生成され，各質問ごとに 3 つの回答候補の文を取得する．

Refute と判定せよ” という内容を含めている．

このシステムでは，最終的な evidence は文単位で取得される．文単位での検索では，チャンク単位での検索と比べて，より重要な情報に絞って検索が行えるが，文脈の情報が利用できない．そこで，文単位での evidence 検索を行い，その後，補足コンテキストを生成することで，文脈情報の利用を試みた．

3 evidence 補足コンテキストの生成

図 1 の質問生成+文検索のステップで取得した文単位の evidence に対して，補足コンテキストを生成する．具体的な補足コンテキストの生成手順を図 2 に示す．まず，取得した文がどの文書に含まれていたかを特定する．今回は，使用したデータセットの文書ごとに付与されている URL を利用した．図 1 の文検索時，各文がどの URL の文書から得られたかを記録することで，取得元の文書を特定できる．特定した文書から，既に検索した文の周辺の文を取得し，周辺チャンクを得る．次に，既に検索していた文と，周辺チャンクを LLM に入力し，既に取得した evidence の補足コンテキストを生成する．

生成した補足コンテキストは，Contextual Retrieval [10] に倣い，既に取得した文の前に追加する

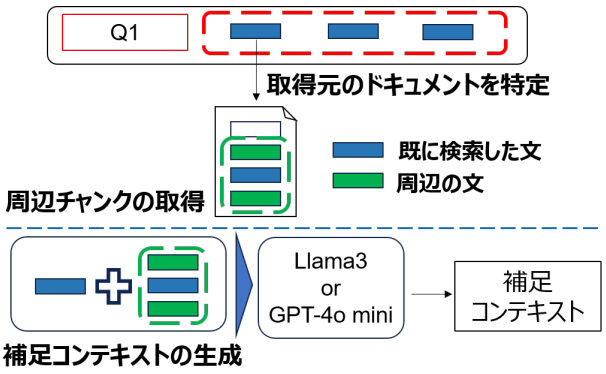

図 2　補足コンテキストの生成手順

ことで，その後の最終判定のステップに用いる．

今回は，補足コンテキストを，evidence の補足だけではなく，リランキングに用いた場合の実験も行った．図 1 では，各質問に対して上位 3 件の回答候補を取得していたが，上位 10 件を取得し，各文の補足コンテキストを生成する．その後リランキングを行い，スコアが高い上位 3 件を取得する．

4　実験

4.1　実験設定

- **データセット**: AVeriTec データセット [6] の検証データ 500 件を使用した．
- **ベースライン**: 図 1 の判定システム
- **モデル**: 質問生成と最終ラベル判定には GPT-4o を，文書検索と文検索には stella_en_400M_v5[3] による埋め込みベクトルの類似度を利用した．補足コンテキスト生成には Llama3-8b-instruct[12] と GPT-4o mini の 2 種類を使用し，リランカーには ms-marco-MiniLM-L-12-v2[4] を使用した．実験の再現性を高めるため，Llama3-8b-instruct は do_sample=False で，GPT-4o と GPT-4o mini は temperature=0, top_p=1.0 で用いた．（GPT-4o と GPT-4o mini は OpenAI の API[5] 経由で利用した．）
- **評価指標**: [6] で提案された Evidence 評価スコアと，最終判定ラベルの正解率を用いた．Evidence 評価スコアは以下の式 1 で算出される．

$$u_f(\hat{Y}, Y) = \frac{1}{|Y|} \max \sum_{\hat{y} \in \hat{Y}} \sum_{y \in Y} f(\hat{y}, y) X(\hat{y}, y) \quad (1)$$

ここで，Y は正解の evidence のシーケンス，$\hat{Y}$ はシステムが取得した evidence のシーケンスである．また，f は，pairwise scoring function で，$\hat{Y} \times Y \to \mathbb{R}$ と定義される．今回は，[6] の実装に従い，METEOR[13] を使用した．また，Hungarian Algorithm[14] により，正解と取得したシーケンスの最適な対応を求めている．これは，形式的には，X: $\hat{Y} \times Y \to \{0, 1\}$と定義され，evidence 取得の再現率に近い式となっている．Evidence 評価スコアは，質問のみを用いた場合，質問と回答を両方用いた場合の 2 通りで計算する．

最終的な判定ラベルの評価は，正解ラベルとの一致度に基づく Label Accuracy で行う．また，判定が正しい evidence に基づいているか検証するため，予測ラベルが正しく，かつ Evidence 評価スコアが一定の閾値以上の場合のみ，正しく判定が行えたとする Label Accuracy も報告する．

なお，周辺チャンクは，対象の文の前後 5 文を取得した．ただし，計算リソースの都合上，Llama3-8b-instruct の tokenizer を使用し，補足コンテキスト生成時の入力の上限を 768 トークンに制限している．

4.2　実験結果

4.2.1　補足コンテキストによる文脈補足の影響

表 1 の 1~3 行目に，ベースラインと，ベースラインに図 2 で生成した補足コンテキストを追加した場合の実験結果を示す．補足コンテキストを追加することで，Evidence 評価スコア，Label Accuracy が向上しており，補足コンテキストが判定に有用な情報を提供できている．

また，表 1 の 4~6 行目に，リランカーを用いた場合の実験結果を示す．4 行目の判定では，補足コンテキストは生成せず，ベースラインの質問生成+文検索において，各質問につき回答候補を 10 件取得し，その後リランカーでスコアが高かった 3 件を取得した．5~6 行目では，リランカーで取得された 3 件に対して補足コンテキストを生成した．補足コンテキストにより，evidence に含まれる情報量が増加し，Evidence 評価スコアは向上したが，Label Accuracy は向上しなかった．

さらなる解析のため，ベースラインに対してリランカーを加えた場合と，GPT-4o mini で生成した補足コンテキストを追加した場合の予測結果の混合行列

3）https://huggingface.co/dunzhang/stella_en_400M_v5
4）https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2
5）https://platform.openai.com/docs/models/

表 1 Evidence 評価スコア (質問のみの場合と Q 質問と回答の場合)，Label Accuracy と，Q+A の Evidence 評価スコアが λ=(0.1, 0.2, 0.25) を超えた場合のみ正しく判定できたとする Label Accuracy

Method	Evidence 評価スコア Q	Evidence 評価スコア Q+A	Label Accuracy	Label Accuracy (.1, .2, .25)		
ベースライン	0.3788	0.2703	0.688	0.674	0.504	0.362
+補足コンテキスト追加 (Llama3)	0.3788	0.2935	0.700	0.694	0.588	0.45
+補足コンテキスト追加 (GPT-4o mini)	0.3788	0.2972	0.704	0.702	0.618	0.462
ベースライン+リランカー	0.3788	0.2781	**0.718**	0.706	0.556	0.400
+補足コンテキスト追加 (Llama3)	0.3788	0.2977	0.712	0.708	0.606	0.478
+補足コンテキスト追加 (GPT-4o mini)	0.3788	**0.2993**	**0.718**	**0.718**	**0.642**	**0.48**
リランキングに補足コンテキスト利用 (Llama3)	0.3788	0.2754	0.700	0.690	0.522	0.388
リランキングに補足コンテキスト利用 (GPT-4o mini)	0.3788	0.2746	0.692	0.678	0.506	0.366

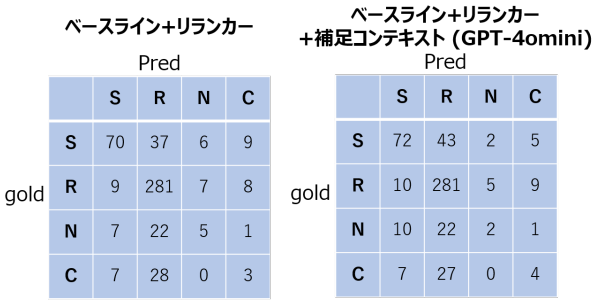

図 3 リランキングに補足コンテキストを利用した場合の予測ラベルの変化．(S: “Supported”，R: “Refuted”，N: “Not Enough Evidence”，C: “Conflict”)

を図 3 に示す．この混合行列から，最終判定モデルが “Not Enough Evidence” と予測する件数が減少していることが分かる．つまり，補足コンテキストにより，最終判定モデルがより高い確信をもって判定を行えるようになったと推察される．そこで，図 1 における最終判定モデルのプロンプトに含めていた，“Refute の可能性がわずかでもあれば，Refute と判定せよ” という指示を除外して判定を行った．Label Accuracy は，リランカーのみの場合は 0.572，補足コンテキストを追加した場合は 0.584 となった．この結果からも，補足コンテキストがある場合，モデルがより高い確信をもって判定することがわかる．

4.2.2 補足コンテキストを用いたリランキング

表 1 の 4~5 行目では，既にリランキングが行われた文に対して補足コンテキストを生成していた．表 1 の 7~8 行目では，回答候補の上位 10 件に対して補足コンテキストを生成し，リランカーでスコアが高かった 3 件を取得している．ただし，最終判定時には生成した補足コンテキストは入力せず，Evidence 評価スコア計算時にも補足コンテキストは予測 evidence に含めていない．

表 1 の 4 行目に比べて，Label Accuracy が減少してしまった．このことから，今回作成した補足コンテキストはリランキングには有効でないと分かった．リランカーにとって有用な補足コンテキストを生成するには，周辺チャンクのサイズを大きくすることが考えられる．補足コンテキストに含まれる情報量を増やすことで，リランカーにとって有効な情報が取得されやすくなり，さらに最終判定モデルもより多くの情報を受け取ることができる．

5 おわりに

本研究では，質問生成による Fact-checking システムのさらなる精度向上を目的として，文単位での evidence 検索を行った後に，補足コンテキストを生成するアプローチに取り組んだ．実験の結果から，補足コンテキストは，より判定に必要な evidence を提供できていることが分かった．一方で，補足コンテキストをリランキングに用いた場合，逆に検索精度が低下してしまうことが分かった．

今後の課題として，補足コンテキストを最大限活用できる最終判定プロンプトを作成する必要がある．今回は，補足コンテキストを用いた場合でも，最終判定モデルに与えるプロンプトを変更しなかった．実際には，補足情報と既に検索した情報を明示的に区別し，必要に応じて参照させるプロンプトのほうが適切であると考えられる．

謝辞

本研究の一部は，JSPS 科研費 JP23K20733 の支援を受けたものである．

参考文献

[1] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 178–206, 2022.

[2] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[3] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. **ACM Transactions on Information Systems**, 2023.

[4] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The Fact Extraction and VERification (FEVER) shared task. In **Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)**, 2018.

[5] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The FEVER2.0 shared task. In **Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)**, 2018.

[6] Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. Averitec: A dataset for real-world claim verification with evidence from the web. In **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

[7] Yuki Momii, Tetsuya Takiguchi, and Yasuo Ariki. RAG-fusion based information retrieval for fact-checking. In Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos, editors, **Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)**, pp. 47–54, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[8] Zackary Rackauckas. Rag-fusion: A new take on retrieval augmented generation. **International Journal on Natural Language Computing**, Vol. 13, pp. 37–47, 02 2024.

[9] Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In **The Twelfth International Conference on Learning Representations**, 2024.

[10] Anthropic. Introducing contextual retrieval, 2024. https://www.anthropic.com/news/contextual-retrieval, [Accessed: Dec. 15, 2024].

[11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 9459–9474, 2020.

[12] AI@Meta. Llama 3 model card. 2024.

[13] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[14] H. W. Kuhn. The hungarian method for the assignment problem. **Naval Research Logistics Quarterly**, Vol. 2, No. 1-2, pp. 83–97, 1955.

A　プロンプト一覧

Claimからの質問生成

You will be given a text. Your task is to generate up to 3 questions that are necessary to verify the accuracy of the information contained in the text.

Example:
Text: Why should you pay more taxes than Donald Trump pays? And that's a fact. $750. Remember what he said when that was raised a while ago, how he only pays… He said, 'Because I'm smart. I know how to game the system.'
Questions: 1. What was Trump's tax return in 2017
2. When did Trump say he was smart for not paying taxes

最終判定

Classify the given claim into four labels: "Supported", "Refuted", "Not Enough Evidence" or "Conflicting Evidence/Cherrypicking".
Your predictions must be based on the given evidence.
The evidence includes questions and three pieces of related information for each question.
If there is even the slightest possibility that it is incorrect, output "Refuted".

Output Format:
"Supported", "Refuted", "Not Enough Evidence" or "Conflicting Evidence/Cherrypicking"

補足コンテキスト生成

<document>
{CHUNK_CONTENT}
</document>
Here is the chunk we want to situate within the whole document
<chunk>
{ALREADY_SEARCHED_EVIDENCE_SENTENCE}
</chunk>
Please give a short succinct context to situate this chunk within the overall document for the purposes of improving search retrieval of the chunk. Answer only with the succinct context and nothing else.

図 4　実験に使用したプロンプト一覧．補足コンテキスト生成のプロンプトは Contextual Retrieval[10] を参考に作成した．