大規模言語モデルによる時系列を考慮したフェイクニュース生成

原悠貴 [1]　Yin Jou Huang[2]　Fei Cheng[2]
[1] 京都大学工学部　[2] 京都大学大学院情報学研究科
{hara,huang}@nlp.ist.i.kyoto-u.ac.jp　feicheng@i.kyoto-u.ac.jp

概要

大規模言語モデルの登場により，フェイクニュース検出タスクに新たな課題が生じている．既存の検出タスクの多くは，その F1 スコアが 90%を超えている．これは 1 つのニュース記事だけに着目し，その言語的特徴等を手掛かりにした，簡易的なタスク設定が原因と考える．そこで本研究では，大規模言語モデルを用いて，既存のフェイクニュース検出器に検出されにくいフェイクニュースの生成を試みた．具体的には，時系列順に並んだ複数の記事に着目し，論理的に一貫したフェイクニュースを生成することを目標にした．

1　はじめに

ChatGPT [1] や GPT-4 [2] をはじめ，最近の大規模言語モデル (以下 LLM) は，人間が書く文章によく似た文章を生成することができ，ある文章が人間が書いたものなのか LLM が生成したものなのかを判断するのが難しい場合が多い．この技術は非常に便利である一方で，これを悪用することで，間違った情報 (フェイクニュース) が拡散されるリスクがある．フェイクニュース検出に関する研究は以前から行われていたが，LLM の登場により新たな課題が生じている．その内の 1 つが，既存のフェイクニュース検出器は LLM が意図的に生成したフェイクニュースをどの程度検出できるのかという課題だ．

既存の検出タスクの多くは，1 つのニュース記事だけに着目している．タイトルの言語的特徴や内容が風刺的か [3]，あるいは内容が誹謗中傷を含むか [4] といった，ニュースの表面的な側面に注目しており，これは簡易的なタスク設定だと言える．その結果，F1 スコアはどれも 90%を超えている [4, 5]．

そこで本研究では複数のニュース記事に着目し，LLM を用いることで，既存のフェイクニュース検出器に検出されにくい，文脈を意識したフェイクニュースデータセットを構築する手法を提案する．図 1 のように，あるトピックに関する複数のニュース記事が時系列順に並んだタイムラインを作成し，1 つのニュース記事を置換する形でフェイクニュースを生成する．生成時に文脈や時系列的なつながりを考慮することで，他の記事との論理的一貫性を保つ．

また，本手法により生成されたフェイクニュースについて，関連性・矛盾性・一貫性という 3 つの評価指標に基づいて人間による評価を行った．関連性と矛盾性は 100%, 90.91%と高スコアだったが，一貫性はタイムラインの質が原因で 27.27%と低いスコアに留まった．

2　データセット構築手法

時系列を考慮したフェイクニュース生成の全体像を図 1 に示す．まずニュース記事のデータセットから，キーワードグループを抽出する (2.2 節)．次に，得られたキーワードグループを含むニュース記事からストーリーを生成し，それを用いてタイムラインを作成する (2.3 節)．最後にタイムラインの中の 1 つのニュース記事を置換する形で時系列を考慮したフェイクニュースを生成す (2.4 節)．

2.1　使用するニュース記事の要素

本研究で扱うニュース記事は以下の要素を用いる．

- headline: ニュース記事のタイトル
- short_description: ニュース記事の要約
- date: 投稿された年月日
- content: ニュース記事の内容

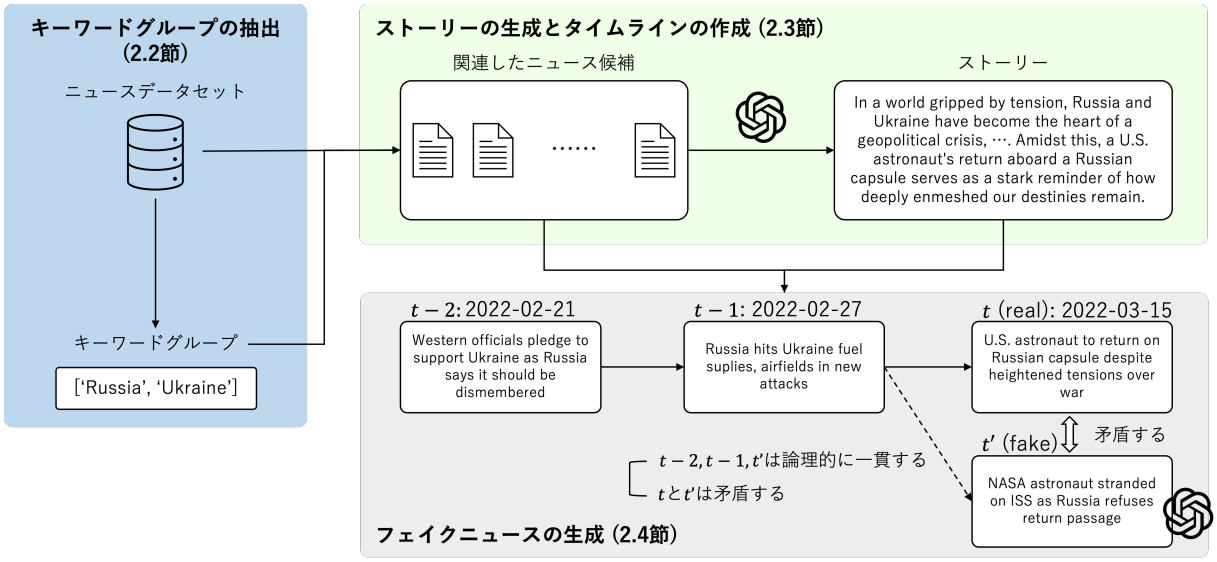

図 1　提案手法の全体像

2.2　キーワードグループの抽出

タイムラインの作成にあたって，似たようなトピックについてのニュース記事を取得する．ここでは同じキーワードを持つ記事は関連したものと仮定する．したがってニュースデータセット全体するよく共起するキーワードグループを抽出する．

各ニュース記事の headline と short_description を連結した文字列について，レマタイズ，ストップワードの除去，1 文字の単語の除去，名詞以外の単語の除去，BM25[6] を用いた不必要な単語の除去し，残りのものをキーワードとする．次に，得られた複数の名詞群に対して FP-growth (Frequent Pattern-growth)[7, 8] を適用することで，['Russia', 'Ukraine'] のような二単語以上からなるキーワードグループを抽出することができる．

2.3　タイムラインの作成

タイムラインとは，あるトピックについてのニュース記事が時系列順に並んだものである．タイムラインに含まれる記事同士が関連性の高いものとなるように，2 段階的にタイムラインを作成する．

まず，タイムラインを作成するのに必要なストーリーを作成する．あるキーワードグループを headline もしくは short_description に含む記事を全て取得することで関連したニュース候補を取得し，それらを LLM に入力することでストーリーを生成する．あるキーワードグループを含むというだけではそれらのニュース記事が同じ話題についての内容ではない可能性があり，タイムラインを作成できない．そこでこのようにストーリーを作成し，そのストーリーに沿ってニュース記事を選択することで記事同士の関連性が高いタイムラインを作成することができる．

次にこのストーリーを用いてタイムラインを作成する．タイムラインに関連したニュース候補の中からニュース記事を追加した際，そのタイムラインと生成したストーリー間の ROUGE-2 スコア [9] が最も高くなるニュース記事を，正式にタイムラインに追加する．このように貪欲にニュース記事を選択することで関連性の高いニュース記事を選ぶことができ，ストーリー性のあるタイムラインを作成できる．

2.4　フェイクニュースの生成

ここまでで作成したタイムラインデータセットと LLM を用いて，1 つのタイムライン中に 1 つのフェイクニュースを生成する．フェイクニュースは時系列的に前 2 つのニュース記事とは論理的に一貫するが，置換するニュース記事とは矛盾するようにしたいため，2 段階的に生成する．

まずある 1 つのタイムラインに含まれるニュース記事を全て LLM に入力し，フェイクニュースを生成する位置を決めてもらう．次に，文脈を意識し論理的に一貫したフェイクニュースを生成するため，フェイクニュースを生成する位置の前 2 つのニュース記事と，矛盾させるニュース記事を LLM に入力する．

表 1　使用するデータセット

year	2019	2020	2021	2022	計
count	2005	2054	2066	1398	7523

また基本的な条件として次の 2 つを設ける．1 つ目は，フェイクニュースを生成する際は， headline, short_description, date を使う．また矛盾させるニュース記事のみ content も使うこと．2 つ目は，フェイクニュースの位置は，タイムラインにおいて 3 つ目以降とすること．2 つ目の条件はフェイクニュースの前に 2 つのニュース記事があることを保証するために必要なものである．

3　実験の結果と分析

2 章で説明した手法を用いて，フェイクニュースデータセットを作成する．

3.1　実験設定

LLM には GPT-4 を採用し，今回は API で利用できる gpt-4-1106-preview を使用する．Temperature は 0.8 とした．

ニュースデータセットには News Category Dataset[10, 11] を使用する．News Category Dataset には content が含まれないため，ニュース記事の URL からスクレイピングでニュース記事の内容を取得した．また，使用するデータセットの年毎のデータ数を表 1 に示す．今回は 2019 年以降に投稿された 7523 個のニュース記事を用いる．

次にタイムラインの作成に関する設定について説明する．タイムラインに含まれる記事の数は $n_{\min}$ 以上 $n_{\max}$ 以下とし，それぞれのキーワードグループ i において，作成するタイムラインの数 k_i は作成可能なタイムライン数の最小値の半分以下とする．つまり，キーワードグループ i について関連したニュース候補の数が l_i の時，以下の式が成り立つ．

$$n_{\min} \leq n_{k'} \leq n_{\max} \text{ for } \forall k', 1 \leq k' \leq k_i \tag{1}$$

$$k_i \leq \left[\left[\frac{l_i}{n_{\max}}\right] \times 0.5\right] \tag{2}$$

今回は，$n_{\min} = 4, n_{\max} = 10$ として実験を行う．

ニュース記事をタイムラインに追加する際の停止条件としては，次の 2 つを用いる．1 つ目はタイムラインに含まれる記事の数が最大数 ($n_{\max}$) に達したとき．2 つ目はタイムラインに記事を追加する前後における ROUGE-2 スコア (R) の増加量が，どのニュース記事においても閾値 (ϵ) を下回ったとき．つまり $R_{j+1} - R_j < \epsilon$ となるとき，$j+1$ 番目の記事はタイムラインには追加しない．今回は $\epsilon = 0.007$ として実験を行う．

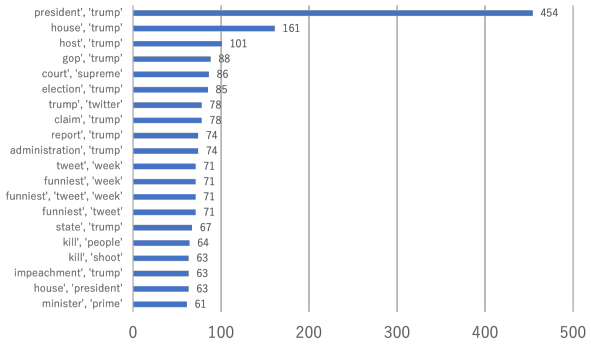

図 2　キーワードグループの上位 20 個

3.2　統計情報

次に得られたキーワードグループについて分析する．全てのニュース記事がキーワードグループの単語を含むわけではないため，この時点で使用されないニュース記事が出てくる．どのキーワードグループも含まない記事は 4492 個存在しているため，以降使用されるニュース記事の数は 7523 個中 3031 個となる．またキーワードグループは合計で 262 個取得でき，そのうちそのキーワードグループを含む記事の数が多いものの上位 20 個を図 2 に示す．上位には主に ‘Trump’ や ‘president’ といった政治的な単語が入っている．

また $\epsilon = 0.007$ のときの，タイムラインに含まれるニュース記事の数の分布を表 2 に示す．作成されたタイムラインの数は合計で 359 個であり，4 個のニュース記事からなるタイムラインが最も多い．

3.3　生成したフェイクニュースの量的分析

生成したフェイクニュースを含むタイムラインを 22 個ランダムにサンプリングし，人間による評価を行う．評価を行う際は，以下の 3 つの評価指標により評価する．

1. 生成されたフェイクニュースは，置換された記事と関連した内容か (関連性).
2. 生成されたフェイクニュースは，置換された記事と矛盾しているか (矛盾性).
3. 生成されたフェイクニュースは，時系列的に前 2 つの記事とは論理的に一貫しているか (一貫性).

表 2 タイムラインに含まれるニュース記事の数の分布

ニュース記事	4	5	6	7	8	9	10	計
タイムライン	127	116	63	20	14	8	11	359

表 3 人間による評価の結果

	1. (関連性)	2. (矛盾性)	3. (一貫性)
スコア	100% (22/22)	90.91% (20/22)	27.27% (6/22)

表 4 生成したフェイクニュースの例 (上: 良い例, 下: 悪い例)

$t-2$	Western Officials Pledge To Support Ukraine As Russia Says It Should Be Dismembered
$t-1$	Russia Hits Ukraine Fuel Suplies, Airfields In New Attacks
t (real)	U.S. Astronaut To Return On Russian Capsule Despite Heightened Tensions Over War
t' (fake)	NASA Astronaut Stranded on ISS as Russia Refuses Return Passage
$t-2$	Twitter Users Bury Thankless Trump Over Ugly John McCain Funeral Slam
$t-1$	Ivanka Trump Mocked By Twitter Users While Promoting ‘Skills-Based Hiring’
t (real)	Trump Lashes Out At The Media But Twitter Users Push Back With Fierce Replies
t' (fake)	President Trump Announces COVID-19 Cure; Credits Personal Genius

人間による評価の結果を表 3 に示す．1 つ目の評価指標については 100%となっており，GPT-4 が生成するフェイクニュースは，置換されたニュース記事に関するものとなっていることがわかる．また 2 つ目の評価指標については 90.91%となっており，ほとんどのフェイクニュースが置換されたニュース記事と矛盾するように生成されたことがわかる．一方で 3 つ目の評価指標については 27.27%とスコアが低くなってしまっている．その理由について次の節で分析する．

3.4 生成したフェイクニュースの質的分析

ケーススタディも交えつつ，3 つ目の評価指標におけるスコアが低い点について分析する．生成したフェイクニュースの例を，headline に限って表 4 に示す．

まず良い例について分析する．これは [‘Russia’, ‘Ukraine’] というキーワードグループから作成されたタイムラインの一部である．時系列的に前 2 つのニュース記事 ($t-2$ と $t-1$) では，ロシアとウクライナの情勢に関するものとなっており，置換されたニュース記事はその情勢の中アメリカの宇宙飛行士がロシアのカプセルで帰還することについての記事である．生成されたフェイクニュースは NASA の宇宙飛行士がロシアのカプセルに搭乗できず ISS に取り残されたという記事であり，３つの評価指標の全てを満たしていることがわかる．

次に悪い例について分析する．これは [‘Trump’, ‘users’] というキーワードグループから作成されたタイムラインの一部である．置換された記事はトランプ氏が COVID-19 を軽視する立場であることを示す記事であり，生成されたフェイクニュースはトランプ氏が COVID-19 の治療法を開発したという記事なので，1 と 2 の評価指標は満たしている．一方で，時系列的に前 2 つの記事は COVID-19 に関する記事ではなく，明らかに 3 つ目の評価指標を満たしていない．これはフェイクニュース生成前のタイムラインの質が原因だと考えるが，これには 2 つの理由が考えられる．

1 つ目はキーワードグループの質だ．悪い例には ‘users’ というあまり重要ではない単語が含まれている．このような単語がキーワードグループに含まれることで，関連性の低いニュース記事同士がタイムラインに含まれる可能性を高めていると考える．

2 つ目はストーリーの質だ．タイムラインはストーリーをカバーするように，関連したニュース候補から選択されるため，GPT-4 が生成するストーリーがストーリー性をもたず，複数の話題について言及している場合，得られるタイムライン中のニュース記事も関連性の低いものとなると考える．実際，悪い例のタイムラインを作成する際のストーリーは一貫性のないものとなっていた．

4 おわりに

本研究では大規模言語モデルによる時系列を考慮したフェイクニュース生成の手法について提案した．また，3 つの評価指標を用いて，生成したフェイクニュースの人手評価を行った．時系列的に前 2 つの記事と論理的に一貫したフェイクニュースは 27.27%程度に留まった．将来的にはフェイクニュースの質を向上させ，既存のフェイクニュース検出器がどれほど検出できるのかを確認したい．

謝辞

本研究は JSPS 科研費 23K16946, JST ACT-X JPMJAX23CP の助成を受けたものです.

参考文献

[1] OpenAI. Introducing chatgpt., 2023. https://openai.com/blog/chatgpt.

[2] Gpt-4 technical report, 2023.

[3] Benjamin D. Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, 2017.

[4] Archita Pathak and Rohini Srihari. BREAKING! presenting fake news corpus for automated fact checking. In Fernando Alva-Manchego, Eunsol Choi, and Daniel Khashabi, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop**, pp. 357–362, Florence, Italy, July 2019. Association for Computational Linguistics.

[5] Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. Welfake: Word embedding over linguistic features for fake news detection. **IEEE Transactions on Computational Social Systems**, Vol. 8, No. 4, pp. 881–893, 2021.

[6] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. **Found. Trends Inf. Retr.**, Vol. 3, No. 4, p. 333–389, apr 2009.

[7] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. **SIGMOD Rec.**, Vol. 29, No. 2, p. 1–12, may 2000.

[8] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In **Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data**, SIGMOD '00, p. 1–12, New York, NY, USA, 2000. Association for Computing Machinery.

[9] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[10] Rishabh Misra. News category dataset. **arXiv preprint arXiv:2209.11429**, 2022.

[11] Rishabh Misra and Jigyasa Grover. **Sculpting Data for ML: The first act of Machine Learning**. 01 2021.

A　参考情報

参考となる情報を記す.

A.1　Prompt

フェイクニュースを生成する際のプロンプトを表 5 に示す.

表 5　フェイクニュースを生成する際のプロンプトの例

INSTRUCTIONS
Generate ONE fake news based on the following constraints and input documents by following criteria:
- It needs to contain headline, short_description, date (YYYY-MM-DD),
and content properties, and please strictly adhere to around 200 words for the content you generate.
- Additionally, explain how the fake news contradicts the real ones.

criteria1: Ensure that the headline and content of the fake news you generate are contradict to
headline and content of the following document.
document ID. 1102
headline: Russia Hits Ukraine Fuel Suplies, Airfields In New Attacks
short_description: Russian forces blew up a gas pipeline in Kharkiv and an oil depot near the Zhuliany airport,
according to the office of Ukrainian President Zelenskyy.
date: 2022-02-27
content: ...

criteria2: Please generate fake news that is a possible consequence of the following two documents.
document ID. 1666
headline: Kremlin Denies Troop Buildup Near Ukraine Border Signals Plan To Invade
short_description: “Russia doesn’t threaten anyone,” said a Kremlin spokesman.
date: 2021-11-13

document ID. 1136
headline: Western Officials Pledge To Support Ukraine As Russia Says It Should Be Dismembered
short_description: Faced with a Russian attempt to redraw borders in Europe, the U.S. is preparing sanctions
and European allies are signaling solidarity with the government in Kiev.
date: 2022-02-21

- The date of the fake news you generate should be 2022-02-27.

A.2　パラメータ

キーワードグループの抽出に関する設定について説明する.
BM25 のパラメータである k_1 と b については，$k_1 = 1.2, b = 0.75$ とし，閾値を 0.5 とした.
また FP-Growth のパラメータである minsup と minconf については，minsup = 0.003, minconf = 0.6 とした.