フレーム意味論に基づく 言及先情報を含む SNS 投稿の事実忠実度のアノテーション

遠田哲史 [1]　吉永直樹 [2]　豊田正史 [2]
[1] 東京大学大学院情報理工学系研究科　[2] 東京大学生産技術研究所
tohda@tkl.iis.u-tokyo.ac.jp　{ynaga,toyoda}@iis.u-tokyo.ac.jp

概要

SNS 投稿には他の投稿や外部サイトの情報に対して言及するものがあるが，言及先の情報が必ずしも投稿内に忠実に記述されるとは限らない．このような情報の変容は偽・誤情報の蔓延に繋がるため，投稿に含まれる事実表現が言及先に忠実であることを自動的に確認する手法の開発が大きな課題となっている．本研究では，SNS 上の情報伝播時における事実表現の累積的な改変の解明に向けて，フレーム意味論に基づいた自動要約システムのアノテーション枠組みを応用し，SNS 投稿内に含まれる事実表現に関する誤りの種類の分類を行う．データセットとしてニュース記事に言及する日本語 X 投稿を人手でアノテーションし，分析を行う．結果として，全体の約 1/3 に事実忠実度の観点での誤りが認められた．

1　はじめに

多数のユーザが自由に情報の投稿・拡散を行うソーシャルメディア上では，他の投稿や外部の情報源などに言及する投稿が多く存在する．このような投稿を介した情報入手は自身の知らない知識を獲得する手段として有用であるものの，事実に関する誤りを含む誤情報を誤って入手・拡散してしまう恐れがあり，インフォデミック[1)]等の悪影響が懸念されている．

このように，ソーシャルメディア上で拡散されてしまう多種多様な信憑性の主張に対し，根拠となる情報と照らし合わせて検証を行うことで，誤情報や偽情報を検出する自動ファクトチェックの研究が行われている [1][2][3]．しかし，ファクトチェックでは個別の主張の客観的な信憑性を評価するのに対し，ソーシャルメディア上の情報共有活動では，信憑性の判断が根本的に困難な新規情報や，信憑性の判断が一意に定まらないような主張も存在する．そこで，ソーシャルメディア投稿と言及先の文章を比較し，投稿内の事実表現がどの程度言及先の文章に忠実であるか，という観点で評価することで，情報の拡散過程における情報改変現象を適切かつ迅速に検出することが可能になると考えられる（図 1）．

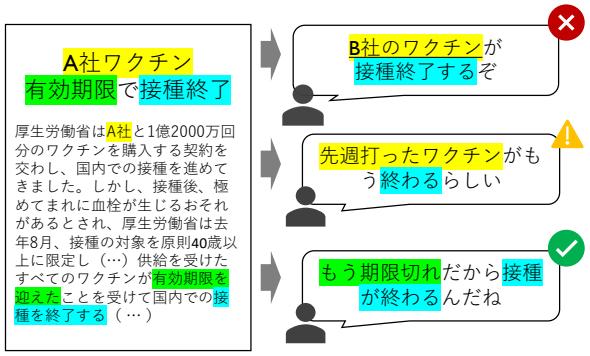

図 1　本研究で扱う，言及先情報を含む SNS 投稿の例．右側の SNS 投稿は左側のニュース記事に言及している．右下の投稿から，言及先の記事に忠実な投稿，忠実でないが矛盾もしていない投稿，矛盾している投稿の例を示す．ハイライト部分は文章間の対応関係を示す．

本研究では，SNS 上の情報伝播時における事実表現の累積的な改変の解明に向けて，言及先情報を含む SNS 投稿の事実忠実度をフレーム意味論 [4][5] に基づいて行うアノテーション基準を整備し，少数の投稿について実際に注釈づけを行い，注釈基準の妥当性の評価と，注釈に基づく分析を行う．事実忠実度のアノテーションにあたっては，まず各投稿について，投稿に含まれる事実表現が言及先の文章と異なるか判定を行う．異なる場合，その差異の性質を「述語に関する誤り」「エンティティに関する誤り」「付帯状況に関する誤り」「その他の誤り」の 4 ラベルから選択し，誤りの内容が言及先の文章と矛盾するか否かを選ぶ．データセット構築にあたっては，2019 年 7 月 1 日に X に投稿された日本語投稿のうち，外部ニュースサイトに言及しているものを 80 件選択した．アノテーションの結果として，全体の

1）https://www.who.int/health-topics/infodemic (2025 年 1 月 7 日参照)

約 1/3 に事実忠実度の観点での誤りが認められた．ただし，2 件を除いて，ニュースサイトに掲載されている情報との矛盾は無く，多くの場合は投稿が追加的な情報を加えるのみに留まることが判明した．

2 事前知識: フレーム意味論

フレーム意味論では，単語の意味が背景知識の構造（フレーム）と結びついていると考える．例えば商取引フレーム [6] においては，フレームを喚起させる単語として「売る」や「買う」などの動詞があり，これらの述語の項として「売り手」「商品」「買い手」「代金」などが存在する．これらフレームに固有の要素は中核フレーム要素 (core frame element) と呼び，一方で「時間」「場所」など固有でない要素については周辺フレーム要素 (peripheral frame element) と呼ぶ [7]．本研究におけるフレーム意味論の適応については，Pagnoni ら [8] の抽象的要約システムの評価における事実忠実度のアノテーションを参考に，SNS 投稿の事実忠実度の評価を行うアノテーション基準の設計を行った．

3 関連研究

3.1 自動ファクトチェック

SNS 投稿の事実表現の正しさを扱う研究としては，SNS 投稿を対象とした自動ファクトチェックの取り組みである RumourEval [9] や PHEME [10, 11] が挙げられる．これらの研究では，特定の噂をめぐって議論が行われるツリー構造のスレッドに含まれる各ツイートを分類タスクの対象として扱っており，各ツイートは単一の主張のみを含むように制限されている．他の研究では，ソーシャルメディア上の投稿を用いてファクトチェック用の学習データを拡充しており，たとえば検証記事へのリンクを含む返信を正しいものとして扱う手法 [12] がある．

近年では，大規模言語モデル（LLM）の推論能力が自動ファクトチェックのタスクに応用されており，主張をより細かく分割して個別に検証することで，重要な要素を見落とすリスクを低減している [13][14]．

これらの研究は SNS 投稿の客観的な信憑性を判断することを目的とするが，本研究で扱う情報忠実度の観点に最も近いのは MANITWEET [15] である．この研究では，ニュースに関連する SNS 投稿が，本来のニュースと異なる情報を含む場合に着目し，これを検出するタスクを提案している．ただし，投稿内に含まれるエンティティを人為的に操作した投稿をデータとしてまとめているため，エラーの種類が限定される．これに対して本研究では，外部のニュース記事に言及する実際の SNS 投稿を対象に分類を行うことで，事実忠実度が損なわれる場面を網羅的に分析する．

3.2 LLM 幻覚

SNS 投稿は必ずしも言及先の情報を忠実に反映しないが，類似の課題に，LLM が元文章に含まれていない情報を要約タスクなどで生成してしてしまう LLM 幻覚の問題がある [16][17][18]．本研究で扱う人の投稿の事実忠実度は，いわば人の言語能力における幻覚を評価することに相当するため，要約の事実忠実度判断にフレーム意味論を用いた研究 [8] を参考に忠実度のアノテーション基準を設計した．

ただし，SNS 投稿は言及先の情報以外の多様な情報を含むことが多く，投稿者の個人的な意見や体験，あるいは他の情報を導入する場合も見られる．本研究では，これらの違いを考慮した SNS 投稿の注釈付けに必要な補足的な基準を設け，また，記事の内容との直接的な矛盾の有無という観点でのラベル付けを新たに導入した．

4 投稿の事実忠実度の注釈付け

本稿では，フレーム意味論に基づいた事実忠実度の分類およびアノテーションの具体的な手法を説明する．分類の対象となるのは，他の投稿や文章に言及している SNS 投稿であり，以下のステップの通りにアノテーションを進める．

1. 言及先の文章および投稿を読む．
2. 投稿内の事実表現が，すべて言及先の文章に含意されるか判断する．
3. 言及先の文章に含まれない事実表現が投稿に存在する場合，その差異の種類を以下の誤りの分類から選ぶ．(複数可)
4. それぞれの誤りについて，「言及先と矛盾するか」「言及先と矛盾しないか」を選ぶ．[2)]

誤りの分類については，Pagnoni ら [8] が提案した

2) 事実忠実度が完全ではなく，誤りが存在する場合でも，言及先の文章と矛盾しない場合はある．例えば「朝にパンを食べた」など，言及先の文章と関係のない情報が投稿に含まれている場合，忠実ではないが矛盾もしない．

述語の意味フレームに着目した分類を用いて，以下の通りに行う．なお，一つの投稿には複数の文が含まれることを鑑みて，複数選択することが可能である．

述語に関する誤り: 述語（日本語においては動詞，イ形容詞，ナ形容詞＋だ，名詞＋だ，など）が言及先の文章に含まれていない場合は，これを選択する．フレーム意味論におけるフレームの誤りに相当する．[3)]

例）投稿内に「2 月 1 日に，総理大臣が被災地を訪れた」という記述があるが，訪れたことが言及先の文章に書かれていない場合など．

エンティティに関する誤り: 言及先の文章にも記述のある述語について，その項となるエンティティ（行為に関する「誰が」「何を用いて」「誰に」等を表す名詞・名詞句）が言及先の文章に含まれていない場合は，これを選択する．フレーム意味論における中核フレーム要素の誤りに相当する．

例）投稿内に「2 月 1 日に，総理大臣が被災地を訪れた」という記述があるが，言及先の文章に書かれてる人物が異なる場合など．

付帯状況に関する誤り: 言及先の文章にも記述のある述語について，その項となる周辺的な状況（場所，時間，日時，形容，程度など）が言及先の文章に含まれていない場合は，これを選択する．フレーム意味論における周辺フレーム要素の誤りに相当する．

例）投稿内に「2 月 1 日に，総理大臣が被災地を訪れた」という記述があるが，言及先の文章に書かれてる日時が異なる場合など．

その他の誤り: 上記に該当しないが，言及先の文章に含まれない記述が投稿内にある場合．（一例として，複数の文にまたがる表現など．）

例）投稿内に「2 月 1 日．その日，総理大臣が被災地を訪れた．」という記述があるが，「その日」が指す日にちが言及先の文章と異なる．

ここで，Pagnoni ら [8] が事実忠実度の分類をした抽象的要約文と大きく異なり，SNS 投稿には投稿者の意見や投稿行為そのものに対する言及も含まれる．また，要約文においては，それに含まれる事実表現が全て元の文章に含まれている必要がある一方で，SNS 投稿にはそのような制約は原則としてない．したがって，本研究では新たに設けた以下の基準を参考に誤りの選択を行う．

- 投稿者の意見や指示であることが明示された表現は考慮しない（「面白い」，「〜だと思う」，「〜しろ」など）．
- 言及先に関するメタ言語表現は，言及先の文章に含まれない情報として扱う（「こんな報道が最近多い」など）．
- 発言などの引用については，発言行為フレームと引用内容の述語に関するフレームをそれぞれ独立のものとして扱う．

3) Pagnoni らの論文には明示的な言及がないが，述語が異なっていても同じフレームを喚起させる場合は，この限りではない（商取引フレームと結びつく「買う」「売る」など）

表 1 フレーム意味論に基づいた事実忠実度の分類結果

ラベル	投稿数	割合 (%)
完全に忠実ではない	27	33.75
述語に関する誤り	23	28.75
エンティティに関する誤り	4	5.00
付帯状況に関する誤り	3	3.75
その他の誤り	0	0.00
完全に忠実	53	66.25
合計	80	-

5 事実忠実度データセット

本稿では，言及先情報を含む SNS 投稿を対象に，先述のフレーム意味論に基づいたデータセットの構築を実際に行った際の手順を紹介し，データセットの分析と考察を行う．

5.1 データ収集

アノテーションを行う対象の投稿データは，2019 年 7 月 1 日に投稿された日本語 X 投稿のうち，ウェブニュース記事 (NHK NEWS Web[4)] および Yahoo! News[5)] に限定) へのリンクを本文中に含むものを選択した．これらを同じウェブニュース記事へのリンク別に整理し，投稿が 2 個以上あるウェブニュース記事について，各記事について 5 つまでの投稿を無作為に取得した．また，記事のタイトルのみ含む投稿や，重複している投稿を除く処理を行った．

5.2 統計情報

データセットの統計情報を表 1 に示す．本実験で収集した投稿のうち，約 1/3 に言及先の記事に含まれていない事実表現があることが判明した．その差異の種類としては，述語に関する誤りが圧倒的に多く，フレームそのもののずれが投稿と言及先の記事とで多く発生していることが読み取れる．これに対して，エンティティに関する誤りや付帯状況に関する誤りは少ない．

完全に忠実でない投稿のうち，投稿に書かれている事実表現が言及先の記事と矛盾するものは 2 件のみであった．それぞれ「述語に関する誤り」「エンティティに関する誤り」について付与されたものである．

5.3 考察

まず，SNS 投稿を介した情報伝播においては，投稿が言及先の事実表現に「完全に忠実ではない」ケースが多くみられることが判明した．しかし言及先と矛盾する事実表現を採用しているものは少なく，むしろ新たな情報を付け足すものが殆どである．内訳としては，投稿者自身の考えや体験に関する記述に加え，関連する外部の情報の導入などが行われている投稿がみられた．

ただし，言及先と直接矛盾するような事実表現を含まない場合でも，言及先の情報についての誤解を生じさせる恐れは依然として存在することに留意が必要である．例えば，ニュースに含まれていない補足情報を付け足す投稿を読んで，その補足情報がニュースに含まれていると誤解してしまうケースなどが考えられる．

また，「完全に忠実ではない」投稿の誤り分類については，「その他の誤り」を除いた他 3 種のカテゴリで投稿の 100%を網羅しているため，分類の妥当性が確認された．

6 おわりに

本研究では，SNS 投稿が他の情報に言及するような形で情報が伝播する場面に着目し，X 投稿に含まれる事実表現の忠実度についてフレーム意味論に基づいた分類を実施した．結果として一定数の投稿に事実表現が「忠実でない」投稿がみられた一方で，投稿の言及先である記事と明確に矛盾する投稿は少数であり，言及先の情報に新たな情報が付け足される形で X 投稿が行われる現象が明らかになった．

4) https://www3.nhk.or.jp/news/
5) https://news.yahoo.co.jp/

謝辞

本研究は JSPS 科研費 JP24KJ0736，および NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託業務（JPNP23031）の結果得られたものです．

参考文献

[1] Terry Flew, Christina Spurgeon, Anna Daniel, and Adam Swift. The promise of computational journalism. **Journalism Practice**, Vol. 6, No. 2, pp. 157–171, 2012.
[2] Lucas Graves. Understanding the promise and limits of automated fact-checking. **Reuters Institute for the Study of Journalism**, 2018.
[3] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In **Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science**, pp. 18–22, Baltimore, MD, USA, June 2014. Association for Computational Linguistics.
[4] Charles J. Fillmore. Frame semantics and the nature of language. **Annals of the New York Academy of Sciences**, Vol. 280, No. 1, pp. 20–32, 1976.
[5] Charles J. Fillmore. Frames and the semantics of understanding. **Quaderni di Semantica**, Vol. 6, No. 2, pp. 222–254, 1985.
[6] Charles J. Fillmore. The case for case reopened. **Syntax and Semantics**, Vol. 8, pp. 59–82, 1977.
[7] Charles J. Fillmore and Collin Baker. 313 a frames approach to semantic analysis. In **The Oxford Handbook of Linguistic Analysis**. Oxford University Press, 12 2009.
[8] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4812–4829, Online, June 2021. Association for Computational Linguistics.
[9] Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In **Proceedings of the 13th International Workshop on Semantic Evaluation**, pp. 845–854, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.
[10] Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, and Rob Procter. Pheme dataset of rumours and non-rumours, October 2016.
[11] John Dougrez-Lewis, Elena Kochkina, Miguel Arana-Catania, Maria Liakata, and Yulan He. PHEMEPlus: Enriching social media rumour verification with external evidence. In Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos, editors, **Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)**, pp. 49–58, Dublin, Ireland, May 2022. Association for Computational Linguistics.
[12] Momchil Hardalov, Anton Chernyavskiy, Ivan Koychev, Dmitry

Ilvovsky, and Preslav Nakov. CrowdChecked: Detecting previously fact-checked claims in social media. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, **Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 266–285, Online only, November 2022. Association for Computational Linguistics.

[13] Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. Towards fine-grained reasoning for fake news detection. In **Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022**, pp. 5746–5754. AAAI Press, 2022.

[14] Xuan Zhang and Wei Gao. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, **Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 996–1011, Nusa Dua, Bali, November 2023. Association for Computational Linguistics.

[15] Kung-Hsiang Huang, Hou Pong Chan, Kathleen McKeown, and Heng Ji. Manitweet: A new benchmark for identifying manipulation of news on social media, 2023.

[16] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics.

[17] Meng Cao, Yue Dong, and Jackie Cheung. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3340–3354, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[18] Taiji Li, Zhi Li, and Yin Zhang. Improving faithfulness of large language models in summarization via sliding generation and self-consistency. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 8804–8817, Torino, Italia, May 2024. ELRA and ICCL.