LLM の相対評価に基づく ニュース記事要約の誇張度スコアリング手法の提案

岩本圭介 [1]　嶋田和孝 [1]
[1] 九州工業大学大学院
iwamoto.keisuke629@mail.kyutech.jp　shimada@ai.kyutech.ac.jp

概要

SNS 上では，ニュース記事本文よりも要約が単独で拡散されることが多く，本文と矛盾しないまま特定の要素を強調する誇張要約が含まれる場合もある．本研究では，要約に含まれる誇張の程度を定量的に扱うため，LLM の相対評価に基づく誇張度スコアデータセット構築手法を提案する．提案手法では，LLM の相対評価を比較関数としたソートにより誇張度の順位を推定し，その順位を連続値スコアへと写像する．さらに，LLM の出力確率を比較に利用することで，ソート結果の安定性を向上させる．実例のコーパスを利用した実験により，提案手法が一貫した誇張度スコアを付与できることを確認した．

1　はじめに

近年，SNS の普及により，ニュース記事に対する意見や解釈が即座に共有される環境が一般化している．特に，文字数制限のある SNS プラットフォームにおいては，記事本文そのものよりも，短い要約や見出しが用いられることが多く，それらが単独で拡散される傾向にある．これらの要約は，記事の著者本人によって作成される場合に加え，第三者や自動生成システムによって作成される場合も多い．しかし，これらの要約の中に，記事本文の一部を過度に強調した誇張的な表現を含むものも存在する．誇張的な要約は，本文と完全に矛盾していないとしても，記事全体の内容に対し過度に強い印象を読者に感じさせる可能性がある．さらに，そのような要約を基にした意見や解釈が二次的に共有されることで，読者の理解が偏る可能性も生じ得る．

誇張的な要約が持つ問題点は，単なる情報の不正確さに限らない．誇張表現は，一部の情報を過度に強調することにより，読者の感情に訴えやすい表現となる場合がある．SNS 上においては，事実を中立的に表現した情報よりも，感情的あるいはセンセーショナルな表現を含む情報の方が拡散されやすいことが指摘されている [1, 2]．その結果，事実に即した要約と比較し拡散されやすくなり，特定の解釈が強調された情報流通を助長する可能性がある．そのため，誇張要約の拡散は，読者のニュース内容に対する理解を誤らせる要因の一つであると考えられる．

このような背景のもと，本研究では，誇張された要約によって生じ得る誤った情報理解を抑制し，それに基づく世論形成を防ぐことを目的とする．この目的を達成するため，要約がどの程度誇張されているかを把握できる手法を構築する．情報が高速に流通する SNS 環境下においては，どの程度誇張されているかを迅速かつ直感的に把握できることが求められる．そこで，誇張の程度を定量的な指標として示すことが有効であると考えられる．文章による定性的な説明は，誇張がどのような点において表出しているかを詳細に伝えられる一方で，読むための時間や認知負荷が高く，SNS 環境においては適さない．これに対し，数値として提示される指標は視覚的に把握しやすく，読者が即時に注意の必要性を判断することを可能にする．

SNS 上は大量の情報が流通しており，人手による逐一の確認は現実的でないため，指標を自動的に算出できることが望ましい．そこで，本研究では，ニュース記事の本文と要約のペアに対し，誇張の度合いを表す数値スコアを付与したデータセットの構築を行う．このデータセットは，将来的に誇張度を自動推定する機械学習モデルの学習に利用されることを想定している．本研究では，ニュース記事要約に対する誇張度スコア付きデータセットを構築する手法を提案するとともに，大規模言語モデル (LLM) を利用し，誇張度を安定的かつ一貫した形で付与する方法を検討する．

2 関連研究

2.1 要約における正確性評価

ニュース記事要約の研究においては，要約が元記事の内容と整合しているかを評価する手法が数多く提案されてきた．FactCC [3] は要約と本文が意味的に一致しているかを分類問題として扱い，要約に含まれる事実誤りを検出する手法である．また，BERTScore [4] は事前学習済み言語モデルの埋め込み表現を利用し，本文と要約の意味的類似度を測る指標である．一方，BARTScore [5] は要約生成モデルに基づき，本文を条件とした要約の生成確率を利用し，要約の妥当性を評価する手法である．

これらの手法は，要約に含まれる事実誤りや不整合を検出するうえで有効である．一方で，いずれも本文と矛盾しているか否かといった観点に焦点を当てており，本文と整合していても特定の要素を強調するような誇張表現を直接的に扱うものでない．

2.2 誇張要約に関する研究

誇張要約に着目した研究は，正確性評価に関する研究と比較すると多くはない．Iwamoto ら [6] は，LLM を利用し通常の要約を誇張的なものになるように書き換えを行うことで誇張要約データセットを構築し，誇張要約検出の検証を行った．しかし，このデータセットに含まれる誇張要約は人工的に生成されたものであり，実際のニュース記事要約を十分に再現していると限らない．また，ラベルは二値分類に基づいており，誇張の程度の違いを扱うことはできない．

これに対し，Iwamoto ら [7] は，実際のニュース記事要約に対し誇張の程度を表す連続的な数値が付与された誇張度スコアデータセットの構築手法を提案した．この研究では，LLM により複数の本文-要約ペアを比較し，要約の誇張度の大小関係に基づきソートを行ったうえで，得られた順位に基づき，各要約に連続的な誇張度スコアを付与する枠組みが示されている．この枠組みにより，誇張要約を二値でなく連続的な尺度で扱うことが可能になった．一方で，誇張度スコアが複数回の比較結果に基づき決定されるため，比較過程における LLM の判断のばらつきが，ソートや最終的なスコアに影響することが報告されている．実際に，ペアの初期配置や実行回によりソート結果が変動する傾向が確認されており，誇張度スコアの安定性に依然として課題が残されている．

そこで，本研究では，Iwamoto ら [7] により提案されたデータセット構築の枠組みを踏襲しつつ，誇張度スコア付与の安定性を向上させる手法について検討する．

3 提案手法

本節では，先行研究 [7] で提案された誇張度スコアデータセット構築の枠組みを踏襲し，誇張度スコアの付与を行う手法について説明する．特に，先行研究で指摘されている比較過程における LLM の判断のばらつきに着目し，3.3 節において，誇張度スコア付与の安定性向上を図る．

3.1 絶対評価と相対評価

LLM を利用し，ニュース記事要約の誇張度を評価する方法として，絶対評価と相対評価の 2 つが考えられる．

絶対評価では，LLM に対し単一の本文-要約ペアを入力し，要約が本文に対しどの程度誇張しているかを数値として直接出力させる．この方法は直観的である一方で，同一の入力に対し出力が変動しやすく，評価結果の一貫性が低いという問題がある [8]．これは，LLM が内部に明確な絶対的評価基準を持たないことに起因すると考えられる．

これに対し，相対評価では，2 つの本文-要約ペアを同時に提示し，どちらの要約がより誇張的であるかを比較させる．LLM は単一の事例に対する絶対的な数値判断よりも，2 つの事例を比較するような判断において，より安定した出力を示す傾向がある．ただし，相対評価からは二項的な大小関係しか得られないため，そのまま連続的な誇張度スコアを得ることはできない．

そこで，本研究では，相対評価が直接的な数値スコアを与えられないという制約を踏まえつつ，その一貫性の高さを活用するため，多数の相対比較結果を集約し，全体における順位として整理することで，誇張度を連続的な数値として表現する手法を採用する．

3.2 手法の全体像

誇張度スコアデータセット構築手法の全体像を図 1 に示す．本手法では，ニュース記事の本文-要約ペア集合に対し，それらを誇張度の低いものから高

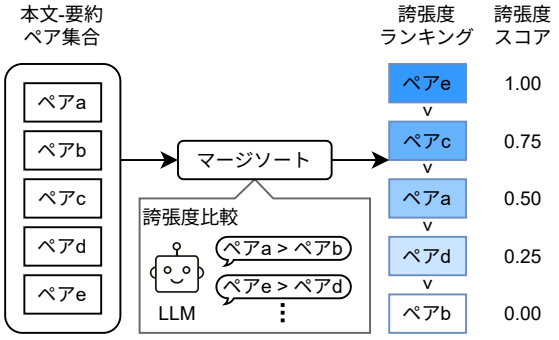

図 1　誇張度スコアデータセット構築手法の全体像.

いものへとソートしたうえで，その順位を基に連続値スコアを付与することを目指す．まず，図 1 の左側に示すように，ペア a，ペア b などの複数の本文-要約ペアを収集する．次に，これらの本文-要約ペアを誇張度順にソートする．このランキングを得るため，マージソートアルゴリズムを用い，その比較関数として LLM を利用する．具体的には，ソートの過程で 2 つの本文-要約ペアを入力として LLM に与え，どちらの要約が本文に対しより誇張的であるかを相対的に判断させる．LLM は各比較において，2 つのペア間の誇張度の大小関係を出力し，この結果がソート処理における比較結果として用いられる．

このようにして得られたソート結果から，誇張度スコアを導出する．具体的には，最も誇張度が低いペアにスコア 0 を，最も誇張度が高いペアにスコア 1 を付与する．残りの要約については，ソート結果における順位に応じ，0 から 1 の範囲で線形補間されたスコアの割り当てを行う．この線形割り当ては，大規模な本文-要約ペア集合を対象とする場合，ランキングの両端に位置する要約が，実データにおける誇張度の最小値及び最大値を近似すると見做せるという仮定に基づく．即ち，順位 1 位及び N 位に位置する要約は，観測可能な誇張度の下限及び上限を代表すると解釈できる．この統計的仮定の基で，相対評価に基づくランキングを 0 から 1 の連続値区間へ写像することにより，解釈可能な数値スコアとして付与することが可能になる．

3.3　出力確率を利用した相対評価

LLM を比較関数として利用する場合，評価対象の提示順序が判断結果に偏りを生じさせる位置バイアスの存在が知られている [8]．このバイアスは，2 つの本文-要約ペアの誇張度に大差が無い場合に特に顕著である．Iwamoto ら [7] は，この問題に対処するため，同一の 2 つのペアについて，提示順序を

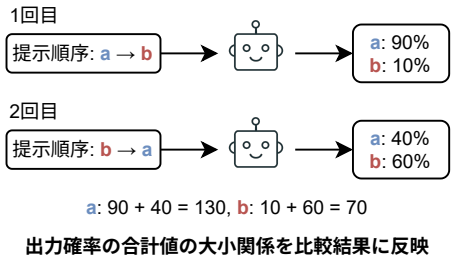

図 2　提示順序を変化させた相対比較における出力確率の統合方法．本文-要約ペア a と b に対し，提示順序を a → b 及び b → a とした 2 通りの比較を行い，各比較において最終的な判断結果と対応する出力確率を得る．本図の例では，離散的な判断のみを見ると比較結果が競合しているが，本手法では，出力確率を統合することで両提示順序を通し確信度の高いペア (本図では a) を選択可能にする．

入れ替え 2 回の比較を行い，両者の判断結果が一致した場合のみを有効な比較結果として採用する手法を用いている．この方法は位置バイアスを低減する一方で，LLM が出力時に保持している確率的情報を考慮していない．

LLM は，各比較において 2 つの選択肢のいずれかを最終出力として生成するが，その際に各選択肢に対する確率分布が計算されている．提示順序を変化させた 2 回の比較において最終的な出力値が一致しない場合であっても，対応する出力確率を比較すると，一方の判断が他方よりも高い確信度を伴っている場合がある．実際に，Wang ら [9] は，LLM-as-a-Judge に関する研究において，離散的な最終出力のみを利用した評価よりも，出力確率分布を考慮した評価の方が，人手評価，即ち真値に近くなることを報告した．これは，LLM の判断をより正確に反映するうえで，確率情報が有効であることを示唆している．

そこで，本手法では，LLM の出力トークンの確率値を比較に利用する．具体的には，2 つの本文-要約ペア a と b を比較する際，提示順序を入れ替え 2 通りの比較を行う．

図 2 に示すように，提示順序 a → b の場合と b → a の場合のそれぞれの比較で得られた出力確率を，各本文-要約ペアごとに合算する．そのうえで，合計確率の大小関係を基に，どちらのペアがより誇張的であるかを決定する．

このように，提示順序を変化させた複数回の比較結果を確率値として統合することで，離散的な出力に依存しない，誇張度の相対比較が可能になる．その結果，比較結果のばらつきが抑制され，ソート結果の安定性向上に寄与すると考えられる．

4　実験

本節では，提案手法をニュース記事コーパスに適用し，誇張度スコアデータセットを構築したうえで，その安定性を検証する．また，構築後の誇張度スコアデータセットに含まれる要約の具体例を B 節に示し，付与されたスコアの内容的な妥当性について補足的に確認する．

4.1　実験設定

本実験では，ニュース記事の本文-要約ペアを多数収録した Newsroom コーパス [10] のうち，1,000 件のペアを利用する．本文-要約ペア間の相対評価用 LLM として `Mistral-7B-Instruct-v0.3`[1] を用い，温度パラメータを 0.8 に固定する．なお，LLM への入力に利用したプロンプトを A 節の図 4 に示す．Newsroom コーパスに対し本手法を適用し，誇張度スコアデータセットを構築する．

構築したデータセットに対し，LLM の相対評価によるソート結果が，初期配置や実行回に左右されることなく安定しているかどうかといった観点から評価を行う．比較対象として，出力確率を利用した提案手法に加え，先行研究 [7] に基づく，LLM の最終出力である離散的な判断結果のみを利用する手法をベースラインとして設定する．

4.2　ソート結果の安定性

本節では，LLM を比較関数として利用した相対評価に基づくソート結果が，初期配置の違いに対しどの程度安定しているかを検証する．特に，相対評価において LLM の出力確率を利用する設計が，ソート結果の安定性にどの程度寄与しているかを検証する．

本検証では，初期配置をランダムに変更した状態でマージソートを 4 回実行し，各本文-要約ペアについて得られた順位の標準偏差を算出することで，順位の変動幅を評価する．順位の標準偏差が小さいほど，初期配置の違いに左右されることなく同様の順位が得られていることを意味し，ソート結果が安定していると解釈できる．

検証の結果，4 回のソートにおける各本文-要約ペアの順位の標準偏差の平均値は，提案手法では 148.0 位，ベースライン手法では 177.2 位であった．

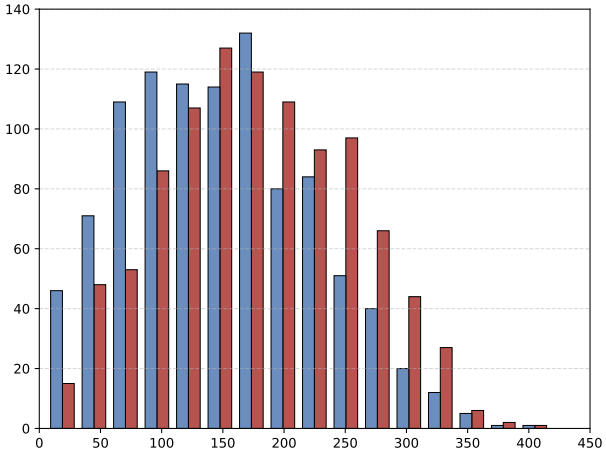

図 3　4 回のソート結果における各本文-要約ペアの順位の標準偏差の分布．横軸は 4 回のソート結果における順位の標準偏差を，縦軸は該当する本文-要約ペア数を表す．青色のヒストグラムは出力確率を利用した提案手法，赤色のヒストグラムは離散的な判断結果のみを利用した場合の結果を示す．

出力確率を利用した提案手法は，離散的な判断結果のみを利用するベースライン手法と比較し，ソート結果のばらつきが低減できているといえる．

図 3 に，両手法における，4 回のソート結果における順位の標準偏差の分布を表すヒストグラムを示す．図中の青色のヒストグラムは提案手法の結果を，赤色のヒストグラムはベースライン手法の結果を表す．提案手法の青色ヒストグラムについて，ベースライン手法の赤色ヒストグラムと比較すると，分布全体が左側に寄っており，多くのペアについて順位の変動が比較的小さいことが確認できる．

この結果は，出力確率を利用し比較結果を統合することで，比較過程における LLM の判断のばらつきが低減され，ソート結果の安定性が向上することを示唆している．

5　おわりに

本研究では，LLM の相対評価に基づく誇張度スコアデータセット構築手法を提案した．LLM の相対評価を比較関数として利用したソートにより，一貫した誇張度スコアを付与できることを確認した．また，LLM の出力確率を比較に利用することで，ソート結果の安定性が向上することを確認した．今後は，構築したデータセットを利用した誇張度推定モデルの構築を通じ，本手法の有効性を検証する．

1）https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

参考文献

[1] William J. Brady, Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. Emotion shapes the diffusion of moralized content in social networks. **Proceedings of the National Academy of Sciences**, Vol. 114, No. 28, pp. 7313–7318, 2017.

[2] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. **Science**, Vol. 359, No. 6380, pp. 1146–1151, 2018.

[3] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 9332–9346, Online, November 2020. Association for Computational Linguistics.

[4] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In **8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020**. OpenReview.net, 2020.

[5] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, **Advances in Neural Information Processing Systems**, Vol. 34, pp. 27263–27277. Curran Associates, Inc., 2021.

[6] Keisuke Iwamoto and Kazutaka Shimada. Dataset construction and verification for detecting factual inconsistency in japanese summarization. In **2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)**, pp. 243–248, 2024.

[7] Keisuke Iwamoto and Kazutaka Shimada. Exaggeration scoring of news summaries through llm-based relative judgments. In **Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation**, Hanoi, Vietnam, December 2025.

[8] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[9] Victor Wang, Michael JQ Zhang, and Eunsol Choi. Improving LLM-as-a-judge inference with the judgment distribution. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, **Findings of the Association for Computational Linguistics: EMNLP 2025**, pp. 23173–23199, Suzhou, China, November 2025. Association for Computational Linguistics.

[10] Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

A　本研究で利用したプロンプト

本節では，本研究の実験において，LLM による相対評価に利用したプロンプトを示す．

```
Please act as a fair judge and evaluate the level of exaggeration in the summaries of two news article pairs (A and B).
Each pair includes a full article and its summary.
Your task is to evaluate which summary is more exaggerated relative to the content of the original article.

When evaluating, consider the following:
– Exaggerated numbers or statistics
– Making the subject or event seem more important than it is in the original article
– Use of meiosis (understatements) or litotes (negation–based softening)
– Stylistic choices that may change the overall tone or impression of the article

Be as objective as possible.
Do not be influenced by the length or order of the summaries.
After your evaluation, provide your final judgment ONLY a single character: A or B.
Do NOT output anything else.
```

図 4　LLM の相対評価に利用したプロンプト．

B　誇張度スコアデータセットの例

本節では，本研究で構築した誇張度スコアデータセットに含まれる要約例を示す．誇張度が低いと考えられる要約 (誇張度スコア 0) 及び，高いと考えられる要約 (誇張度スコア 1) をそれぞれ取り上げ，各本文と要約の対応関係について説明する．

図 5 に，誇張度スコア 0 の要約例を示す．本文では，デビスカップ準々決勝における試合結果が記述されている [2]．要約では，「ジョン・イズナーがジル・シモンを 6-3, 6-2, 7-5 で破った (John Isner ... dispatched Gilles Simon 6-3, 6-2, 7-5)」といった形で，本文中の主要な事実のみが簡潔にまとめられている．試合の重要性や選手の評価について過度な強調はなく，本文の内容を中立的に反映した要約であることから，誇張度は低いと考えられる．

図 6 に，誇張度スコア 1 の要約例を示す．本文では，アルカイダの指導者交代について，実質的な変化は大きくないという論調で分析が行われている．一方，要約では，「アルカイダの新 CEO から何を期待すべきか？ (What can we expect from the new CEO of Al Qaeda?)」といった表現が用いられ，指導者交代による大きな変化を想起させる構成となっている [3]．本文の主張自体と矛盾はしないものの，論点の置き方により変化のインパクトが強調されており，相対的に誇張度が高い要約であると解釈できる．

```
John Isner used his brutal forehand to dispatch Gilles Simon 6–3, 6–2, 7–5 as the United States drew level with France in their Davis Cup quarterfinal on Friday.
```

図 5　誇張度スコア 0 の要約例．

```
What can we expect from the new CEO of Al Qaeda? Will Zawahiri be radically different from Bin Laden?
```

図 6　誇張度スコア 1 の要約例．

2）https://web.archive.org/web/20120407095954/http://nbcsports.msnbc.com/id/46950872/ns/sports-tennis/

3）https://web.archive.org/web/20240402085157/https://www.foxnews.com/opinion/zawahiri-is-al-qaedas-new-ceo-will-number-two-try-harder