D-009

大規模言語モデルを用いた賛否両論ある検索トピックに関するウェブページのスタンス判定

Stance Detection of Web Pages on Controversial Search Topics Using Large Language Models

萩原 諒[1)]　山本 岳洋[1)]　湯本 高行[1)]
Ryo Hagiwara　Takehiro Yamamoto　Takayuki Yumoto

1　はじめに

今日，ウェブ検索を用いて様々な重要な意見形成が行われている．例えば，「緑茶はがん予防に効果があるのか」などの健康情報を検索して自身の意見を形成することがある．2020 年の Moz の報告によれば，1100 人の回答者のうち約半数が Google を用いて頻繁に医療の重要な意見形成を行っていることが明らかになっている [13].

一方で，「小学生に宿題は必要か」や「学生服は必要か」などのような，真偽が明らかになっていなく賛成意見と反対意見が存在し議論になっているトピックもある．そのようなトピックを本研究では賛否両論あるトピックと定義する．このようなトピックは真偽が明らかになっていないため，様々な情報から慎重な意見形成を行うことが望ましい．しかし，様々な問題点が報告されている．例えば，賛否両論あるトピックの検索結果は偏る傾向にあり，また，偏った検索結果で意見形成を行った場合，その意見形成も偏ってしまうことが明らかになっている．さらに，検索結果の上位に筆者の意見が書かれていないウェブページが見つからないことが多々ある．意見が書かれているウェブページは共感や納得感があり，自身の意見を決めるうえでとても参考になると考えられる．

そこで本研究では，賛否両論あるトピックの検索結果多様化のために，ウェブページがどのような意見を主張しているのかのスタンス判定に取り組む．具体的には，賛否両論あるトピックとウェブページ本文を入力し，4 つのスタンス（支持，支持しない，中立，関係ない）のいずれかを出力するモデルを構築する．ウェブページのスタンス判定が可能となれば，ウェブページの再ランキングなどを行うことで，意見の多様性を考慮した検索結果が実現でき，検索者は様々な情報から慎重な意見形成を行うことができると考えられる．

本研究ではウェブページのスタンス判定を行うため，ChatGPT と BERT を用いた．ChatGPT とは，OpenAI が開発した大規模言語モデルである．大量のテキストデータを学習しているため，ファインチューニングを行わなくてもプロンプトだけで良い精度が出るとされている．プロンプトは，zero-shot プロンプトと CoT プロンプトを用いる．BERT とは，Google が開発した自然言語処理モデルの 1 つである．ファインチューニングを行うことで，特定のタスクで高い性能を発揮できるため，文書分類などの自然言語処理の課題で広く使われている．本研究では，ChatGPT を用いて生成した疑似データでファインチューニングを行った．

構築したモデルの有効性を確かめるため，アノテーションで得られたデータを用いて評価を行った．その結果，CoT プロンプトを用いた ChatGPT が最も高い精度を示した．

2　関連研究

本節では関連研究について述べる．まず賛否両論あるトピックに関する研究について述べ，次に大規模言語モデルを用いたスタンス判定に関する研究について述べる．最後に，疑似データ生成を用いた研究について述べる．

2.1　賛否両論あるトピック

賛否両論あるトピックに関する研究は様々行われている．例えば，クエリが賛否両論あるトピックかどうかを判定する研究 [1][2] や，ユーザの検索行動 [4]，検索結果 [3] に関する研究などがある．Karl らは，クエリ補完機能を用いることでクエリが賛否両論あるトピックかどうかを判定することが可能であることを明らかにしている [2]．また，Gizem らは賛否両論あるトピックの検索結果は偏る傾向にあることを明らかにしている [4].

2.2　スタンス判定

スタンス判定に関する研究が行われている [9][10][11][12]．Tim らは，自動スタンス判定が可能なことを明らかにし，いくつかの手法でユーザに納得のいく説明が提供できることを明らかにした [9]．また，Chuang らは，ファインチューニングを行っていない ChatGPT のプロンプト作成のみで高精度のスタンス判定ができることを明らかにした [12].

2.3　疑似データの生成

疑似データを生成することで，コストの低下や精度の向上が可能になることを明らかにした研究が行われている [5][6][7][8]．Shushkevich らは，データセットをバランスよくするため ChatGPT を用いて疑似データを生成し，データを拡張させたことで精度が上がったことを明らかにした [5]．銭本らは，ChatGPT を用いて作成したアンケートの疑似回答データで作成したモデルとと人手でアノテーションして作成したモデルの精度を比較すると，同等の精度であったことを明らかにした [8].

3　データセット

本節では，本研究で扱った賛否両論あるトピックの選定を述べ，その後テストデータの作成について詳細に述べる．

3.1　トピックの選定

本研究では，Procon.org[1)]に記載されている賛否両論あるトピックの中から 5 つのトピックを用いた．Procon.org とは，賛否両論あるトピックをまとめているサイトである．このサイトから，小学生に宿題は必要か，子どもにワクチンを打つべきか，ベジタリアンになるべきか，学生服は必要か，大学の無償化は必要かの 5 つのトピックを選んだ．以後，「小学生に宿題は必要か」

1）　兵庫県立大学 University of Hyogo

1）　https://www.procon.org/

Copyright © 2024 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

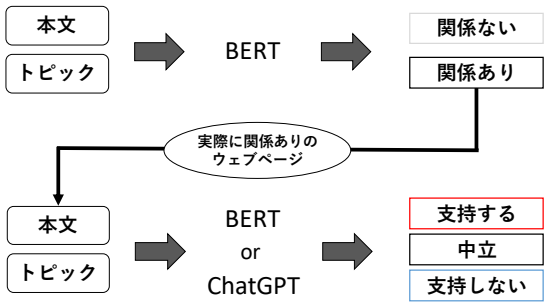

図 1　スタンス判定の手順．

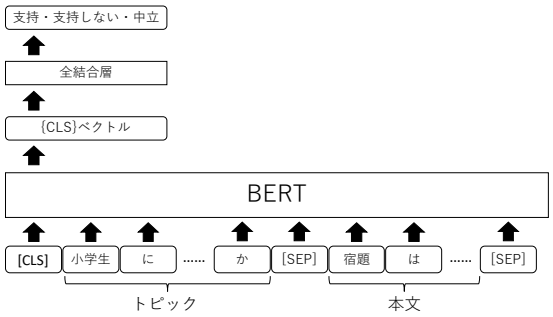

図 2　スタンス判定で用いた BERT モデルの図．

のタスクを宿題タスク，「子どもにワクチンを打つべきか」のタスクをワクチンタスク，「ベジタリアンになるべきか」のタスクをベジタリアンタスク，「学生服は必要か」のタスクを学生服タスク，「大学の無償化は必要か」のタスクを大学無償化タスクと呼ぶ．

3.2　テストデータの作成

テストデータは，実際のウェブページに対し著者を含む 3 人のアノテーターがラベル付けを行い作成した．

3.2.1　ウェブページの取得

ある賛否両論あるトピックに対して，支持する，支持しない，中立なウェブページを取得するため，支持するクエリを 2 つ，支持しないクエリを 2 つ，中立なクエリを 1 つ用意した．具体的には，支持するクエリは，「〇〇はなぜ必要か」と「〇〇のメリット」，支持しないクエリは，「〇〇はなぜ不要か」と「〇〇のデメリット」，中立のクエリは，「〇〇のメリットとデメリット」の 5 つのクエリを用意し，トピックによって〇〇の箇所を変化させた．また，1 クエリに対して上位 30 件のウェブページを取得した．取得したウェブページで重複しているものは 1 つのものとカウントし，PDF と YouTube はあらかじめ除外した．

3.2.2　アノテーション

アノテーションでは，得られたウェブページに対して，そのウェブページの本文が，あるトピックに対して支持，支持しない，中立，関係ないのいずれのスタンスを結論付けているのかのラベル付けを行った．ラベル付けの結果は，表 1 に示す．

4　検索トピックに関するスタンス判定

本節では，2 つの提案手法について述べる．まず ChatGPT によるスタンス判定手法について述べ，その後 BERT によるスタンス判定手法について詳細に述べる．

モデルへの入力は，賛否両論あるトピックとウェブページ本文である．ウェブページ本文を全て入力すると最大トークン数を超えてしまうので，本文の後ろから 400 文字を抽出し，本文データとした．後ろから抽出したのは，本文の最後に筆者の主張が書かれることが多いと考えたためである．

4.1　ChatGPT によるスタンス判定

ChatGPT は，OpenAI が開発した大規模言語モデルである．大量のテキストデータを学習しているため，ファインチューニングを行わなくてもプロンプトだけで高い精度を達成している．プロンプトは様々存在しており [15][16][18]，本研究では zero-shot プロンプトと CoT プロンプトでそれぞれ実装した．

4.1.1　zero-shot プロンプトを用いたスタンス判定

zero-shot プロンプトとは，具体例などを与えず質問のみを投げるプロンプトである．本研究では，具体例を与える few-shot プロンプトを用いた場合，具体例の影響を強く受け，汎用性がなくなると考えたため zero-shot プロンプトを採用した．

4.1.2　CoT プロンプトを用いたスタンス判定

CoT（Chain of Thought）プロンプトとは，中間的な推論ステップを挟むことで，複雑な推論能力を向上させるプロンプトである [15]．本研究では，出力の際にその出力になった理由を表示させることで，CoT の考え方を用いている．以下のプロンプトは，実際の CoT のプロンプトである．

```
# 説明文
入力されたトピックに対して，入力された本文が最終的にどのような意見の立場として書かれているのかを判定してください．
回答は出力形式でお願いします．
また，なぜその分類になったのかの理由を 100 字以内で教えてください．

# 出力形式
判定は’支持する’, ’支持しない’, ’中立’ のいずれかで回答してください．
評価基準は以下の通りです．
支持する：与えられたトピックに対して本文全体の意見が支持している内容になっている
支持しない：与えられたトピックに対して本文全体の意見が支持しない内容になっている
中立：与えられたトピックに対して本文全体の意見が記載されていない
また，理由は入力本文を踏まえて回答してください．

# 入力
・トピック：{topic}
・本文：{text}

# 出力
・スタンス：
・理由：
```


Copyright © 2024 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

表 1　ラベル付けの結果．

	支持	支持しない	中立	関係ない
宿題タスク	8	7	8	16
ワクチンタスク	32	3	37	7
ベジタリアンタスク	19	2	23	8
学生服タスク	5	2	30	19
大学無償化タスク	10	7	9	32
計	74	21	107	82

表 2　関連の有無の分類精度．

	関係あり	関係ない
関係あり	177	25
関係ない	64	18

4.2　BERT によるスタンス判定

BERT （Bidirectional Encoder Representations from Transformers）[14] は，Google が開発した自然言語処理モデルの 1 つである．ファインチューニングを行うことで，特定のタスクで高い精度を達成しているため，文書分類などの自然言語処理の課題で広く使われている．本研究では，東北大学乾研究室が公開している BERT モデルを用いたファインチューニングを行うことで，スタンス判定モデルを構築した．

4.2.1　疑似データの作成

ファインチューニングで用いるデータセットは，ChatGPT で生成した疑似データを用いる．疑似データの生成に用いたプロンプト例は，以下に示す．

> # 説明文
>
> {topic}のトピックに対して，中立な意見を結論付けた文章を生成して下さい．中立な意見とは，支持も不支持もしていなく，読者に結論を委ねているウェブページのことである．回答は，ニュース記事風な書き方で生成してください．文字数は 100 文字以上 200 文字以内にし，多様な意見を含めてください．
>
> # 回答

各トピックに対して，支持，支持しない，中立のウェブページ本文をそれぞれ 200 件ずつ生成した．また，可能な限り多様な意見を含むデータセットにするため，本文の書き方のスタイルを 10 種類用意し，1 つのスタイルに 20 件のウェブページを生成した．関係ないのデータは，他の 4 トピックで生成したウェブページ本文を用いて作成した．最終的に，1 つのトピックに対して，支持するウェブページを 200 件，支持しないウェブページを 200 件，中立なウェブページを 200 件をそれぞれ生成した．

4.2.2　BERT のファインチューニング

図 1 に 4 段階のスタンス判定の流れを示す．4 段階のスタンス判定を行う際，まず，あるトピックが入力されたトピックに対して，関係ありか関係ないかの 2 値分類を行う．関係ありのデータは，各トピックの支持，支持しない，中立のデータから無作為に選んだデータである．その後，関係ありと分類された本文の中で，実際に関係ありのデータに対して，3 段階のスタンス判定を行う．

図 2 に 3 段階のスタンス判定で用いた BERT モデルの図を示す．本モデルのファインチューニングには，前述した支持，支持しない，中立の疑似データをトークン化したものを用いる．入力は，賛否両論あるトピックと，疑似データを入力する．その際，トピックの頭に [CLS] 特殊トークンを付与し，トピックと疑似データの間に [SEP] 特殊トークンを付与する．その後，BERT を適用しベクトル化を行う．そして，[CLS] トークンを用いて支持，支持しない，中立のいずれかを出力させる．

ファインチューニングのデータセットは，4.2.1 節で述べたデータを用いる．具体的には，各トピックに対して，訓練データが 600 件，検証データが 200 件である．

表 3　2 値分類の分類精度．

正解率	0.69
適合率	0.88
再現率	0.73
F1 値	0.80

表 4　BERT を用いたスタンス分類の混同行列．

	支持	支持しない	中立
支持	10	9	46
支持しない	0	7	10
中立	3	17	75

5　実験

本節では，まず実験方法について述べる．次に，モデルの評価について述べる．評価指標として，正解率，マクロ適合率，マクロ再現率，マクロ F1 値を用いた．

5.1　実験方法

本研究では，グループ化交差検証を行い評価を行った．具体的には，3 節で述べた 5 つのトピックのうち 3 つのトピックを学習データ，1 つのトピックを検証データ，そして残りの 1 つのトピックをテストデータとして評価を行った．5 つすべてのトピックのテストデータで，それぞれ評価を行い 4 つの評価指標で評価を行った．

5.2　関連の有無の分類結果

2 値分類の混同行列と分類精度を表 2 と表 3 に示す．分類の結果，正解率が 0.69 であった．特に，関係ありのデータを関係ありとして出力した割合は 88%と高い精度で分類することができた．その結果，202 件の関係ありデータが 177 件となり，このデータを 3 値分類のデータとして用いた．

5.3　スタンス分類の分類結果

表 4，表 6，表 7 に混同行列，表 5 にモデルの分類精度を示す．スタンス判定の結果，CoT プロンプトを用いた ChatGPT の正解率が 0.68 と最も高い精度という結果になった．また，ChatGPT と BERT を比較した場合，すべての評価指標において ChatGPT が上回る結果となった．

5.4　議論

本研究における 2 つ限界点について述べる．本研究では，モデルに入力する本文として，ウェブページ本文の後ろから 400 文字を抽出した．これは，ウェブページの最後の方に筆者の意見が記載されていることが多いと考えたためである．しかし，実際にはウェブページの最初の方に意見が記載されていることや，見出しや題名に意見が強く反映されていることがあった．そのため，入力

Copyright © 2024 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

表 5　3 段階スタンス判定の分類精度．

モデル	正解率	マクロ適合率	マクロ再現率	マクロ F1 値
ChatGPT（zero-shot プロンプト）	0.64	0.79	0.49	0.52
ChatGPT（CoT プロンプト）	0.68	0.62	0.66	0.63
BERT	0.52	0.52	0.45	0.40

表 6　zero-shot プロンプトを用いたスタンス分類の混同行列．

	支持	支持しない	中立
支持	20	0	45
支持しない	0	4	13
中立	6	0	89

表 7　CoT プロンプトを用いたスタンス分類の混同行列．

	支持	支持しない	中立
支持	38	6	21
支持しない	2	11	4
中立	15	9	71

する本文として見出しを追加することや，段落の最初の行を追加することで，より精度の高いスタンス判定が可能になると考えられる．

また，本研究で用いた ChatGPT はファインチューニングを行っていない．ChatGPT はインタネット上のテキスト情報を大量に学習している．そのため，中立な文章の情報をたくさん学習し，中立と出力することが多くなった可能性がある．これを改善するために，様々なスタンスの賛否両論あるトピックに関するデータでファインチューニングを行うことで，より精度の高いスタンス判定が可能になると考えられる．

6　まとめと今後の課題

本研究では，賛否両論ある検索トピックに関するウェブページがどのような意見を主張しているのかのスタンス判定に取り組んだ．提案モデルとして，zero-shot プロンプトを用いた ChatGPT と CoT プロンプトを用いた ChatGPT，そして ChatGPT で生成した疑似データを用いてファインチューニングを行った BERT をそれぞれ構築し評価を行った．スタンス判定の結果，最も高い精度を示したのは CoT プロンプトを用いた ChatGPT であった．また，自然言語処理タスクで高い性能を出す BERT はあまり精度が高くない結果となった．

今後の課題として，2 つのことが考えられる．1 つ目が，入力するウェブページ本文の抽出方法である．本研究では，ウェブページの最後の方に筆者の意見が記載されているという考えの元，ウェブページ本文の後ろから 400 文字を抽出した．しかし，ウェブページによっては最初の方に筆者の意見が記載されているケースや，見出しに筆者の意見が強く反映されているケースが存在する．そのため，ウェブページの本文の抽出方法を改善することで，より高い精度のスタンス判定が可能になると考えられる．2 つ目が，ChatGPT のファインチューニングである．ChatGPT はインターネット上のテキストを大量に学習しているため，ファインチューニングを行わない場合中立のスタンスをとる可能性が高くなることが考えられる．そのため，ファインチューニングを行い賛否両論あるトピックのスタンス判定に特化させた ChatGPT モデルを構築することで，より高い精度のスタンス判定が可能になると考えられる．

どのモデルも共通して，中立と誤判定することが多い結果となった．これは，ウェブページに支持するメリットとデメリットが記載されていて意見を正確に抽出できなかったからだと考えられる．

謝辞

本研究は JSPS 科学研究費助成事業 JP24K03228，JP21H03775，JP22H03905，による助成を受けたものです．ここに記して謝意を表します．

参考文献

[1] Chelaru, Sergiu and Altingovde, Ismail Sengor and Siersdorfer, Stefan and Nejdl, Wolfgang, Analyzing, detecting, and exploiting sentiment in web queries, ACM Transactions on the Web, Vol. 8, No. 1, pp. 1–28, 2013.

[2] Gyllstrom, Karl and Moens, Marie-Francine, Clash of the typings: Finding controversies and children's topics within queries, In Proceedings of the Advances in Information Retrieval - 33rd European Conference on IR Research., pp.80–91, 2011.

[3] Draws, Tim and Tintarev, Nava and Gadiraju, Ujwal and Bozzon, Alessandro and Timmermans, Benjamin, This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics, In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 295–305, 2021.

[4] Gizem Gezici and Aldo Lipani and Yücel Saygin and Emine Yilmaz, Evaluation metrics for measus results, Information Retrieval Journal, Vol. 24, pp. 85–113, 2021.

[5] Shushkevich, Elena and Cardiff, John, Tudublin at CheckThat! 2023: Chatgpt for data augmentation, Working Notes of CLEF, 2023.

[6] Kieser, Fabian and Wulff, Peter and Kuhn, Jochen and Küchemann, Stefan, Educational data augmentation in physics education research using ChatGPT, Physical Review Physics Education Research, Vol. 19, No. 2, pp. 020150, 2023.

[7] Van Nooten, Jens and Daelemans, Walter, Improving Dutch vaccine hesitancy monitoring via multi-label data augmentation with GPT-3.5, In Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, July 2023; Toronto, Canada, Vol. 1, pp. 251–270, 2023.

[8] 銭本友樹, 長谷川遼, 宇津呂武仁, 大規模言語モデルにより生成した疑似データを用いた自由記述アンケートの自動集約, 言語処理学会第 30 回年次大会 2024.

[9] Draws, Tim and Natesan Ramamurthy, Karthikeyan and Baldini, Ioana and Dhurandhar, Amit and Padhi, Inkit and Timmermans, Benjamin and Tintarev, Nava, Explainable cross-topic stance detection for search results, In Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, pp. 221–235, 2023.

[10] Shalmoli Ghosh and Prajwal Singhania and Siddharth

Copyright © 2024 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

Singh and Koustav Rudra and Saptarshi Ghosh, Stance Detection in Web and Social Media: A Comparative Study, ArXiv, 2019.
[11] Küçük, Dilek and Can, Fazli, Stance detection: A survey, ACM Computing Surveys (CSUR), Vol. 53, No. 1, pp. 1–37, 2020.
[12] Chuang, Yun-Shiuan, Tutorials on Stance Detection using Pre-trained Language Models: Fine-tuning BERT and Prompting Large Language Models, arXiv preprint arXiv:2307.15331, 2023
[13] Moz Inc., 2020 Google Search Survey: How Much Do Users Trust Their Search Results?, https://moz.com/blog/2020-google-search-survey, 2024 年 6 月 11 日閲覧
[14] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186, 2019.
[15] Wei, Jason and Wang, Xuezhi and Schuurmans, Dale and Bosma, Maarten and Xia, Fei and Chi, Ed and Le, Quoc V and Zhou, Denny and others, Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems, Vol. 35, pp. 24824–24837, 2022.
[16] Jiacheng Liu and Alisa Liu and Ximing Lu and Sean Welleck and Peter West and Ronan Le Bras and Yejin Choi and Hannaneh Hajishirzi, Generated Knowledge Prompting for Commonsense Reasoning, Annual Meeting of the Association for Computational Linguistics, 2021.
[17] Zhang, Bowen and Fu, Xianghua and Ding, Daijun and Huang, Hu and Li, Yangyang and Jing, Liwen, Investigating chain-of-thought with chatgpt for stance detection on social media, arXiv preprint arXiv:2304.03087, 2023.
[18] Wang, Xuezhi and Wei, Jason and Schuurmans, Dale and Le, Quoc and Chi, Ed and Narang, Sharan and Chowdhery, Aakanksha and Zhou, Denny, Self-consistency improves chain of thought reasoning in language models, arXiv preprint arXiv:2203.11171, 2022.

Copyright © 2024 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.