人工知能学会第二種研究会資料
市民共創知研究会
SIG-CCI-012-08

LLM を用いたファクトチェック機能の試作と Web 議論における関連情報推薦システムへの応用

Prototype of a fact-checking function using LLM and its application to a related information recommendation system for Web discussions

木下 良輔[1]　櫻井 崇貴[1]　白松 俊[1]

Ryosuke Kinoshita[1], Takayoshi Sakurai[1], and Shun Shiramatsu[1]

[1]名古屋工業大学
[1]Nagoya Institute of Technology

Abstract: In Web discussions, both those who are familiar with the content of the discussion and those who have no background knowledge at all can participate. Therefore, there are people who are reluctant to speak up due to differences in the amount of information they have or their level of understanding of the discussion, and this hinders the formation of consensus for the entire discussion. In this study, therefore, we developed a system that automatically recommends information relevant to the discussion, which we believe will help participants understand the content and progress the discussion. In addition, we thought that it would be possible to support the formation of consent based on evidence by simultaneously conducting fact-checking using LLM when presenting information.

1. はじめに

本研究では，公的な社会問題に関する市民参加型の，いつでもどこからでも参加可能な Web 議論を想定したファシリテーション機構の開発を目指す．Web 議論では，議論内容に精通している人も，全く背景知識を持っていない人も参加可能であるため，持っている情報量や議論の理解度の差によって発言しづらい人が存在し，これが議論全体の合意形成の妨げとなっている．そこで議論に関連した情報を推薦することで，参加者の内容理解や議論進行に役立つと考えた．

近年では，GPT-4[1]等の大規模言語モデル (LLM) の発展により，単に情報提示させるだけでなく、そこから導いた意見を投稿させることも可能になった．しかし，科学的知見やデータ等の事実情報を用いずに生成された意見や議論は，ハルシネーションや誤情報を含む可能性がある．これが議論参加者の困惑や議論の停滞を招き，最終的な合意形成の方向性を捻じ曲げる恐れがあると考えた．そこで，LLM を用いた情報提示を行う際に，同時に LLM を用いてファクトチェックも行うことによって，根拠に基づく同意形成支援を行えるのではないかと考えた．

また，情報提示の必要性や検索可能性などの，情報提示要求を抽出する櫻井ら[2]との共同研究を通じて，情報提示のタイミングについても着目する．

2. 関連情報推薦システムの設計・開発

まずは，提案手法のシステムフローを図 1 に示す．

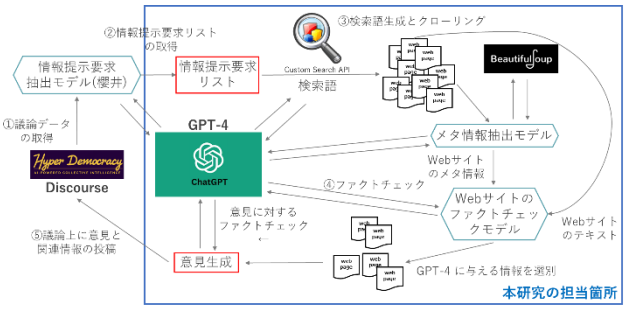

図 1:システムフロー

関連情報推薦システムは，まず初めに Discourse という Web 議論システム上で行われている議論から情報を取得し，櫻井らのシステムによって情報提示の必要性がある投稿を決定する．その後，関連情報をクローリングする際の検索語の生成を行い，検索結果の上位 5〜10 件のサイトを取得する．後述する議論実験では，処理時間短縮のため 3 件のみ取得した．取得した各サイトについて，GPT-4 によるファクトチェックを行い，誤情報を含まない信頼性の高い情報を選別する．Web サイトの選別後，信頼性の高いサイトの情報を GPT-4 に与え，情報提示が必要な投稿への返信を生成する．GPT-4 の発言に対してもファクトチェックを行い，発言の信頼性が確

認出来たらDiscourse上に関連情報の投稿を行う. 本システムでは以上の流れで関連情報推薦を行う.

3. LLMを用いたファクトチェック

3.1 Webサイトのファクトチェック

Webサイトのファクトチェックには, GPT-4を利用する. 本研究では, OpenAI API経由で"gpt-4-1106-preview"モデルを使用した. このモデルによって, 処理できるテキストの量が以前までの8000トークンから12万8000トークンになった. これにより, 複数のWebサイトの内容をGPT-4に与え, その情報を参考にして引用などを行った情報提示が可能になった. しかし, 誤情報や古い情報を含むWebサイトも存在するため, Webサイトのファクトチェックを行い, Webサイトの信頼度を算出することで, GPT-4に与える情報を選別する処理を行う必要がある.

3.2 Webサイトのメタ情報の抽出

Webサイトのファクトチェックを行う際には, Webサイトのメタ情報も一緒に与える. 各サイトのメタ情報は, リアルタイムでスクレイピングする. 取得するメタ情報は, 発行日・更新日, 著者・発行元, ドメイン拡張子, 引用総数, ドメイン別集計とする. 以下にメタ情報の取得例を示す.

サイトのメタ情報の抽出例:

- 発行日: 2019-11-27
- 更新日: 2023-02-01
- 著者・発行元: 明治大学名誉教授の市川 宏雄
- ドメイン拡張子: jp
- 引用総数: 124
- ドメイン別集計:

{'member.vortex-net.com': 5, '100years-company.jp': 114, 'www.vortex-net.com': 3, 'www.facebook.com': 1, 'twitter.com': 1}

ここで, Webサイトの発行日, 更新日については古い情報ではないかの確認や, サイトの中で出てくる日付との整合性の確認に使用する. ドメイン拡張子は, 明らかに怪しいWebサイトを排除するために使用する. 引用に関する部分は, 特定の偏った情報からのみの引用が行われていないか等の確認に使用する.

また, Webサイトの発行元のタイプによってもある程度の信頼度の目安を設けた. これはChatGPTに出力させた値に, 実際に推薦したいサイトのタイプなども加味して値を決定させた. 以下にそれらの値の一覧を示す.

Webサイトの発行元のタイプの信頼度の目安:

- 個人ブログ・ウェブサイト: 30〜70
- ニュースメディア: 50〜90
- 学術機関: 80〜95
- 政府機関: 70〜90
- 非営利団体: 60〜85
- 商業ウェブサイト: 40〜75
- オンラインコミュニティ・フォーラム: 20〜60
- 専門・業界団体: 60〜85
- ウィキ・データベース: 55〜80
- ソーシャルメディア: 10〜50
- エンターテイメント: 25〜60
- 教育・啓発: 60〜85
- レビュー・評価サイト: 40〜70
- 雑誌・定期刊行物: 45〜75
- 医療・健康関連: 50〜90

この信頼度の目安の値は, あくまで関連情報として使用したいサイトの種類の情報を推薦させやすくするための数値として使用しているため, 絶対にこの値の範囲内に収まるような出力をさせるためのものではない.

3.3 GPT-4の発言のファクトチェック

ファクトチェックを行って, 信頼性の高いWebサイトから生成した意見でも, GPT-4の事前知識からハルシネーションや誤情報が生じる可能性がある. そのため, GPT-4の発言に対してもファクトチェックを行う. ここでは, 参考にしたサイトに書かれている内容との差異はないか, そこから予測した未来予測を含む仮定がある場合, その仮定は現実的に妥当かなどの観点から信頼度を算出する. これらの処理により, Webサイトと発言の両方の信頼度が設定した閾値を超えた場合に投稿に対する返信が確定し, Discourse上に発言と情報を推薦する.

4. 評価・考察

提案手法の有用性を検証するために, 関連情報推薦システムを介入させた議論実験を実施し, 実験後にアンケートによる評価を行った. 議論実験は参加者10名を5人ずつのグループAとBの2つに分け, 以下のテーマで実施した. それぞれのグループの議論テーマは表1に示す.

表1:議論実験の議論テーマ

グループA	遺伝子編集と倫理問題
グループB	仮想通貨と金融規制

この実験を通してグループAでは59個, グループBでは50個の合計109個の発言が得られた.

まずは, 議論実験の中でシステムが実際に推薦した関連情報の例を示す. 以下の図2は関連情報推薦システムの発言である.

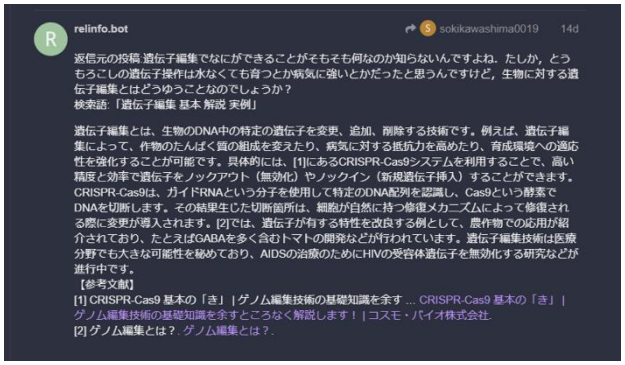

図 2:システムの情報提示例

この例はグループ A の「遺伝子編集と倫理問題」のテーマの議論で行われた情報提示である．直近の発言をした参加者は，生物に対する遺伝子編集に関する根本的な疑問を提示しており，この発言に対して遺伝子編集技術の基礎知識を解説しているサイトから具体例を交えながら的確に解説することに成功している．そのため，この例では議論参加者にとって有意義な情報提示が行えていることが確認できる．

次に実験後のアンケート結果からシステムの評価を行う．議論テーマの難易度や理解度の差によってシステムの評価が大きく異なることが予想されたので，グループ A の 5 人の参加者とグループ B の 5 人の参加者で別々にアンケート結果を集計した．

まずは，推薦された情報は実際に参加者にとって有益だったのかを表すアンケート結果を評価する．以下の表 2 に，先行研究[3]と本研究における，推薦した情報の議論への関連度の平均値と，議論への有益度の平均値を示す．評価はいずれも 1〜7 の 7 段階評価のものである．

表 2: 議論への関連度と有益度の平均値の比較

アンケート	先行研究	本研究
議論への関連度の平均値 (グループ A)	5.4	6.4
議論への関連度の平均値 (グループ B)	5.4	4.6
議論への有益度の平均値 (グループ A)	5.1	6.4
議論への有益度の平均値 (グループ B)	5.1	6.4

これらの結果から，グループ B での推薦した情報の議論への関連度以外は評価が向上したことが確認できた．グループ A と B で評価が異なった理由としては，設定した議論テーマの難易度や，そのテーマについての背景知識によるものであると考えている．今回のグループ B では，システムは想定通りの適切な情報提示は行えていた．しかし，テーマが難しいものであったため，あまり理解が追い付いていない参加者には複雑な専門用語が多く含まれており，余計に混乱させてしまっていた．そのため，関連していたかについて評価が難しかったのではないかと考えられる．一方で，グループ A も難しいテーマではあったが，背景知識を持った参加者が数人いたため，本システムが推薦する情報がとても有効であった．

また，有意差を確かめるために各項目について有意水準 5%でマン・ホイットニーの U 検定を行った．その結果，グループ A での推薦した情報の議論への有益度の向上についてのみ，統計検定量が 8.0 で棄却限界値 8.0 以下となったため，統計的に有意なアドバンテージがあることが確認できた．しかし，それ以外の項目では有意差は確認できなかったため，今後データ数を増やして再検証する必要がある．

次に，システムの情報提示による議論参加者の発言量の変化を評価する．推薦された情報を参考にして投稿した発言の個数については，グループ A では平均 1.2 個，グループ B では平均 1.0 個という結果だった．本システムを議論に介入させたことによる議論全体での発言量の変化，有意水準 5%で両側検定の t 検定を行った結果を以下の表 3 に示す．

表 3: 発言量の変化および t 検定の結果

	発言量の変化	P 値	有意差
グループ A	+30.0%	0.028	あり
グループ B	+29.4%	0.070	無し

グループ A，B ともに発言量は約 30%増加した．しかし，表 3 より統計的に有意であるといえるのは，p 値が 0.05 を下回ったグループ A のみであることが分かった．これはシステムの情報提示によって，グループ A ではほぼ全員が平均的に 1〜2 個程度の発言が増加したことに対し，グループ B ではもともと発言が多かった参加者の発言がさらに増え，分散が大きくなったためであることが考えられる．またアンケートの結果の中には，「知識が補完され，議論内容を理解できるようにはなるが，それだけでは議論に参加することは難しかった」という意見があった．

このように，本システムの情報提示によって発言しやすくなる参加者は一定数いるが，これだけでは議論に参加することが難しい参加者も存在すること

が分かった．そのため，本システムで着目していた推薦した情報の有益度という観点とは別に，議論に参加できるようなほかの観点も必要であると分かった．

5. おわりに

本研究では Web 議論において，議論に関連する情報を推薦することで議論を活性化させ，議論の合意形成支援を行うこと目的としてきた．議論実験による評価を行った結果，本研究で開発した関連情報推薦システムは，先行研究よりも参加者にとって有益な情報提示が可能になったといえる．しかし，有益な情報を提示するだけでは議論に参加することが難しい参加者も見られた．そこで今後は，推薦した情報の有益度という観点とは別に，議論に参加できるような別の観点も明確にしたい．

謝辞

本研究を進めるにあたり，ご支援ご協力を賜りました皆様に深謝します． 本研究の一部は，JST CREST（JPMJCR20D1）およびNEDO（JPNP2000S）の支援を受けたものです．

参考文献

[1] OpenAI: GPT-4 Technical Report, arXiv, pp. 2303–08774,(2023)

[2] 櫻井 崇貴, 白松 俊: 多人数対話における LLM を用いた合意形成支援の必要性判定手法, 情報科学技術フォーラム講演論文集, Vol. 22, No. 3, pp. 443-444, (2023)

[3] Ryosuke Kinoshita, Shun Shiramatsu: Agent for Recommending Information Relevant to Web-based Discussion by Generating Query Terms using GPT-3, IEEE International Conference on Agents, pp. 24-29, (2022)