ファクトチェック演習のための AI 生成文書中の要注意箇所推定の試み
An Attempt to Estimate Unreliable Portions in Outputs of Generative AI for Fact-Checking Exercise

内田 翔梧[*1], 大沼 亮[*1], 中山 祐貴[*1], 神長 裕明[*1], 宮寺 庸造[*2], 中村 勝一[*1]
Shogo UCHIDA[*1], Ryo ONUMA[*1], Hiroki NAKAYAMA[*1], Hiroaki KAMINAGA[*1]
Youzou MIYADERA[*2], Shoichi NAKAMURA[*1]
[*1] 福島大学 共生システム理工学類/共生システム理工学研究科
[*1]Faculty of Symbiotic Systems Science, Fukushima University
[*2] 東京学芸大学 教育学部
[*2]Faculty of Education, Tokyo Gakugei University
Email: uchida@cs.sss.fukushima-u.ac.jp, {hnakayama, onuma, kami}@sss.fukushima-u.ac.jp, miyadera@u-gakugei.ac.jp, nakamura@sss.fukushima-u.ac.jp

あらまし：生成 AI の急速な普及に伴い，その適切な利用法を育成することが昨今の課題になっている．ユーザは AI 生成文書中の危ういファクタを自ら注意・検証すべきであるが，未熟者には難しい作業と言える．本研究では，AI 生成文書の要注意ファクタを主体的にチェックさせる演習のための手がかり情報生成手法の開発を目指す．本稿では，主に，AI 生成文書中の要注意ファクタの見極めのための手がかり情報生成手法の概要を示す．
キーワード：機械学習，フェイクニュース検出，情報フィルタリング，情報探索支援

1. はじめに

近年，生成 AI の発達により，一見すると質の高い文章を一般の人でも容易に作成することが可能になってきている．生成 AI はインターネット上の多様な情報を学習データとして活用していることが強みであるが，同時に，ユーザが無意識のうちに著作権などの権利侵害やトラブルに陥る可能性がある．ユーザは生成 AI の利用に際して，危ういファクタを自ら注意・検証することが重要である．しかしながら，生成 AI の利用経験が浅いユーザにとって，事実と異なる情報の混入や記述内容の情報源などの要注意ファクタを自ら発見・吟味することは難しい．

これに対し，生成 AI により自動生成された人物肖像の利用による肖像権侵害に関する研究[1]が報告されている．しかし，この研究では生成 AI によって自動生成された人物肖像のみに焦点を当てており，自動生成された文章については言及されていない．また，ファクトチェックアラートの有効性に関する研究[2]では，SNS 上における不明確な情報に対してファクトチェックアラートを付けることが誤情報の拡散の予防策として有効であると報告されているが，予備の情報で実験が行われており，ユーザ自身が現実的にファクトチェックを試みることができない．

本研究では，生成 AI の要注意ファクタをユーザが主体的にチェックする演習のための手がかり情報生成手法の開発を目指す．これにより，生成 AI の活用スキル育成方法の新たな可能性を示す．

2. 問題点とアプローチ

2.1 問題点

本研究では，生成 AI の利用に関する問題点として，以下の 3 点に着目する．

（問題点 1）生成物に事実と異なる情報の混入を防ぐことが重要であるが，容易ではない．

（問題点 2）生成物の情報源の確認が必要なケーズが多いが，その確認が困難である．

（問題点 3）注意を要する要素を自らチェックするきっかけを得ることが困難である．

2.2 アプローチ

本研究による支援の概要を図 1 に示す．本研究では，生成 AI の利用経験が浅いユーザが指導者から与えられたテーマに沿って生成 AI を用いて文章を生成し，生成物中の要注意ファクタを主体的にチェックする演習を想定する．

生成 AI による生成物の中で要注意ファクタとなり得るものとして，情報の真偽を疑うべき箇所を推定する手法(問題点 1 に対応)，情報源の確認が必要な箇所を推定する手法(問題点 2 に対応)を開発する．

開発した手法に基づいて，生成 AI による生成物中の要注意ファクタ吟味のための手がかりをユーザに提示することで，要注意ファクタをユーザが主体的にチェックする演習を支援する(問題点 3 に対応)．

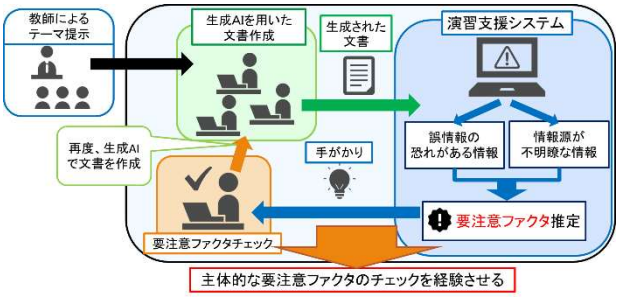

図 1：演習支援の概要

3. 要注意ファクタの推定手法

3.1 誤情報の恐れのある情報の抽出手法

生成 AI により出力された文書から，真実でない情報を含む恐れのあるものを以下の手順で抽出する．

（1）特徴フレーズの抽出
まず，生成された文章から日本語 NLP ライブラリ GiNZA と EmbedRank を用いて重要な語句を選出し，特徴フレーズを抽出する．
（2）AI による生成物中の主要な文の抽出
生成された文章に対し，Python ライブラリの sumy を用いて抽出型要約を行う．要約結果を生成文書の主旨を含む主要な文として抽出する．
（3）主要な文の主旨を表す情報の抽出
（3－1）AI による生成物の主旨の形成
（2）で抽出された主要な文に形態素解析を行い，（1）の特徴フレーズと合わせて以下のような形態素セットを形成する．

特徴フレーズ＋形容詞＋末尾表現
特徴フレーズ＋動詞を含む末尾表現

（3－2）特徴フレーズが同様の Web 上の記事
（1）で抽出された特徴フレーズと，サジェストワードを使い合わせてクエリを準備する．このクエリを用いて検索エンジンを介して Web ページ群を取得する．取得した Web 上の記事から特徴フレーズまたは，特徴フレーズの類語を含む文を抽出する．抽出した文から特徴フレーズまたは特徴フレーズの類語を含んだ形態素セットを形成する．
（4）誤情報の恐れがある情報の推定
誤情報の恐れがある情報の抽出方法を図 2 に示す．（3－1），（3－2）で抽出した形態素セットを比較し，AI 生成文書の主旨と Web 上の記事の一致状況の分析を行う．一致率が一定未満の場合に，当該生成物の内容には誤情報（事実でない情報）が含まれる可能性があると推定する．

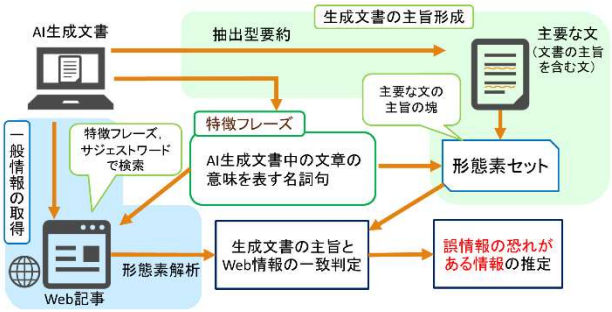

図 2 誤情報の恐れがある情報の抽出

3.2 著作権侵害の恐れのある情報の抽出手法

本研究では，情報源の確認が必要な箇所として主に著作権的な観点に注視し，著作権侵害の恐れがある情報を抽出する．
剽窃チェックソフト CopyContentDetector を用いて生成された文章とインターネット上のドキュメントを比較する．文章単位で似ている文章がないかを表す「類似度」と，単語単位でどれだけ一致しているかを表す「一致率」を分析する．これらを用いて，著作権侵害の恐れがある情報を抽出する．
類似度，一致率が高い Web サイトが見つかった場合には生成物中の著作権侵害の恐れがある情報として抽出し、その部分と対応していると検出された Web サイト情報を提示する．

3.3 要注意ファクタを見極める手がかりの示唆

抽出された誤情報の恐れがある情報と著作権侵害の恐れがある情報の 2 点を分析し，生成 AI による生成物中の要注意ファクタを推定する．
特に，誤情報の恐れがある情報かつ著作権侵害の恐れがある情報であると判定された箇所は厳密な注意が必要になるため，最要注意ファクタとして最優先にして抽出する．

4. 支援システムの概要

各手法に基づいて，ユーザが主体的に要注意ファクタをチェックする演習の支援システムを開発する．
演習では，教師から生成 AI で生成する文書のテーマを学習者に提示する．各学習者は生成 AI を用いてテーマに沿った文書を出力する．学習者は生成された文章をシステムに入力する．
システムは生成物中で著作権侵害の恐れがある情報と判定された部分をハイライト表示し，それらに対応した生成 AI に参考にされた可能性が高い Web サイト情報を提示する．さらに，生成物中から，情報の真偽の確認が必要であると推定された箇所を危険フレーズとして抽出することで要注意ファクタの手がかりをユーザに提示し，ユーザの主体的なファクトチェックを支援する．チェック後，再度同様の手順で要注意ファクタの検出，ファクトチェックを行うことで，生成 AI の活用スキルの育成を図る．

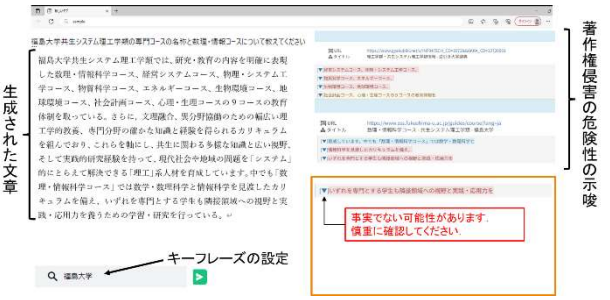

図 3：支援システムのインターフェース

5. おわりに

本論文では，AI 生成文書中の要注意ファクタをユーザが主体的にチェックする演習のための手がかり情報生成手法について述べた．支援システムによる演習支援の概要を示した．
今後は，実際の生成 AI 文書を用いた実験を重ね，提案手法の検証と改善を進めたい．

参考文献

(1) 柿沼太一，“画像生成 AI をめぐる著作権法上の論点”，法律のひろば ぎょうせい編, vol. 76, pp.19-30, 2023.
(2) 竹田悠人，滝口桐矢，西谷陽佳，前田ひなた，森川祐介，“ファクトチェックアラートの有効性の検証”，第 35 回人工知能学会全国大会論文集，2021.
(3) Jürgen Rudolph, Samson Tan, Shannon Tan, “ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?”, Journal of Applied Learning & Teaching, vol. 6, no. 1, pp.342-363, 2023.