日本語小説をもとに自動生成された文章における著者推定の検証

C0119290　　松原　拓未

1. はじめに

近年、情報化が進み、電子の文章のやり取りがメールやSNS等で行われている。匿名の文章が増えたことにより、なりすましや著作物の複製が問題視されている。この問題に対して著者推定の研究がなされてきたが、 近年 AI の技術の進歩により高精度になったフェイク文章に対する有効性は示されていない。したがって本研究では著者推定のモデルを作成し、自動生成した文章を正しく推定できるか検証する。

2. 関連研究

2.1　日本語小説の著者推定に関する研究

清水はDoc2Vec と BERT を用いた日本語小説の著者推定を行った[1]。古くから行われている著者推定の研究の文脈から研究を行い、Doc2Vec においては84.89%、BERT においては55.43%の正解率を達成した。この研究では自動生成された文章に対する判定は行われていない。

2.2　フェイクニュースに関する研究

柳・田原らは、画像付きフェイクニュースとジョークニュースの検出・分類に向けた機械学習モデルの検討を行った[2]。SNS でのフェイクニュースにより、無実の人を犯人と誤認し実害を与えたことを問題として、SNS の投稿からテキスト CNN、VGG-19 を用いて正しい情報、ジョーク情報、フェイク情報の 3 カテゴリに分類した。その結果、3 カテゴリでもマクロ F 値が 0.93 と良好な数値を示した。この研究では著者推定は行われていない。

以上のことより、本研究ではAI のなりすましによるフェイク文章の解決に著者推定の手法を適用し解決を目指す。

3. 自動生成した文章を著者推定するシステムの概要

本研究では、GPT-3 を用いて小説をもとに自動生成した文章を、著者推定できるかを検証する。小説の文章をDoc2Vec で数値表現し、Keras でニューラルネットワークの学習を行う。自動生成する文章はGPT-3 を用いて小説の一節をもとに作成する。Keras で学習したモデルを使い、自動生成した文章の著者を推定する。その後評価する。

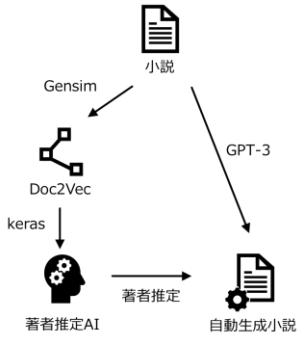

図1. 本研究で構築するシステムの概要図

4. 研究計画

表1. 本研究の計画表

	8月	9月	10月	11月	12月	1月
データの作成						
著者推定システムの開発						
結果の分析と考察						
論文執筆						
プレゼンの準備						

5. 進捗状況

・自然言語処理の学習

形態素解析、固有表現抽出、スクレイピング、GPT-3 の概要及びその実装

・関連研究の調査

著者推定に関する論文の調査

6. おわりに

AI 技術の進歩に伴い、フェイク画像・動画が社会的に問題となっているため、今後はAI のなりすましによるフェイク文章への対応も重要になってくる。このような問題意識のもと、日本語小説から自動生成された文章を対象として、著者を推定するシステムの概要について述べた。

参考文献

[1] 清水大志, “Doc2Vec と BERT を用いた日本語作品の著者推定”, 人工知能学会全国大会論文集, 第32 回, 2020.

[2] 柳裕太, 田原康之, 大須賀昭彦, 清雄一, “画像付きフェイクニュースとジョークニュースの検出・分類に向けた機械学習モデルの検討”, 情報処理学会, vol.2019-ICS-193, No. 11, 2019.