F-010

半教師あり学習による AI 生成文の識別

2022006　今井　悠斗
指導　森　博彦　教授

1　背景

近年 AI 技術の発展が私たちの生活を豊かにする一方で，AI 生成物の見分けがつかないという問題が発生している．ディープフェイクと呼ばれる動画編集技術は，人が実際に話していない言動を話しているかのような動画を作成することができる．また，課題レポートを AI に書いてもらう学生も散見され，学校は個々に対応を取らざるを得ない状況となっている．玉石混交ではあるが画像，音声，テキストなど様々な形で AI は本物に近いものを作成できる．その中でも画像や動画の識別は様々な手法で高い識別精度が見られるが，テキスト情報のなりすまし AI 判別や AI と人間の文章識別は画像や動画に比べて難航している．

2　先行研究

2．1　様々な手法によるフェイクニュース識別

岡山ら[1]は米国のフェイクニュースとフェイクではない政治ニュースの識別を SVM による分類で識別精度ほぼ 100%のモデルを作成した．

また，柳ら[2]は Twitter の投稿をテキストの特徴量と画像の特徴量を TextCNN と VGG19 を用いて抽出し，“リアル”“フェイク”“ジョーク”の 3 つに分類するモデルを作成し，90%ほどの精度が確認できた．

2．2　文書識別，著者推定

文書識別，著者推定の分野で神田ら[3]は事前学習済みモデルの BERT に様々な学習データによる転移学習を行い 10 名の識別で 65%程度のモデルを作成した．

また，岩田[4]は教師無し学習による著者推定モデルを作成し，教師ありモデルと比較を行っている．岩田[4]は敵対生成ネットワーク（以下 GAN と称する）の 1 つである seqGAN の識別器によって夏目漱石の書いた文章であるか否かの識別を試みた．それと同時に精度比較として 3 つの教師ありアルゴリズム（LSTM，CNN，TextCNN）による識別モデルも作成した．その結果，教師あり学習で 70%～80%の精度が出ている一方，GAN の識別器は 50%を下回る識別精度となった．岩田[4]は generator モデルの生成文の品質の低さと，不正解の特徴を学習できなかったことに原因があると考察している．

3　研究目的

テキスト識別が画像識別に後れを取っているとはいえ，SVM や CNN 等の教師あり学習であれば 7 割～9 割以上の精度でフェイクを識別できている．学習に大量のデータが必要な教師あり学習は識別精度が高い反面，新しい技術ができるたびにそのフェイク生成物のデータを集めなおしてモデルを作り直すというイタチごっこになってしまう．したがって，AI 生成文データを必要としない教師無し学習が望ましい．既存の論文で教師無し学習で著者識別を行っている論文は少なくその精度も低い．岩田[4]の考察から，フェイクデータを学習に用いることで識別器が不正解の特徴量を学習できるうえに，そのフィードバックを受けた生成器がそれを踏まえた偽文章を作成でき，相互作用的に識別精度が向上するのではないかと考える．その際に必要なデータ数が少なければ教師無し学習であるメリットを損なうことなくモデルを作成できる．上記の仮説から本研究では既存の人間の文章と少量の AI 文章を元に AI 文章識別モデルを作成することを目的として，教師無し学習である GAN に少量の AI データを組み込んだ半教師あり学習による AI の生成文章と人間の書いた文章の識別を試みる．その後，学習に使っていない言語生成 AI を新しい AI と見立て，追加学習による新しい AI への対応シミュレーションを行う．

4　提案手法

本節では半教師あり学習の具体的な手法を述べる．アルゴリズムや前処理に関しては先行研究で紹介

した岩田[4]の手法を参考にする.

４．１　アルゴリズム

基本的なアルゴリズムは教師無し学習の中で自然言語を扱うことのできる GAN を使用する．本来 GAN では生成器が偽のデータを生成して，そのデータに不正解ラベル，本物の真のデータに正解ラベルを貼り，識別器がそれを識別し，生成器にフィードバックを与えて，2 つを競わせて互いに学習していく．したがって本来は真のデータのみで学習を行うが，本研究では識別器の方に不正解ラベルを貼った AI 生成文を少量組み込む.

４．２　データセット

今回必要となるデータは大量の人間の書いた文章と少量の AI 生成文章である．人間の書いた文章データはテレ朝ニュースのウクライナ状勢に関するネットニュース 3419 記事の本文を利用する．AI 生成文は OpenAI 社の chatGPT3.5 にテレ朝ニュースの記事のタイトルを元にウクライナ状勢に関する文章を生成させ，その生成文 1000 文を利用する．また，新しい AI 役として GPT4 および Google 社の Bard でも同様に文章を生成させる．こうして集めた文章のタイトル行や改行などの余分な情報を削除し BERT モデルを使って 768 次元のベクトルに変換しデータセットとする.

５　結果（現時点）

今回，通常の GAN および DCGAN を使った学習で人間の生成文と AI 生成文の識別を試みた.パラメータの調整を繰り返す中で AI データを学習に取り入れない通常の学習では正解率 0.57，適合率 0.59，再現率 0.44，F 値 0.50 なのに対し，通常の学習に AI データを 1 個取り入れるだけで正解率 0.67，適合率 0.67，再現率 0.66，F 値 0.66 まで向上し，10 個取り入れた場合，正解率 0.94，適合率 0.88，再現率 1.0，F 値 0.94 まで上昇した.その際に必要となった学習回数は 3000 回程度であった.

６　追加学習による新しい AI への対応

GPT3.5 データを少量組み込んで精度が 9 割程度まで上がったモデルを使って新しい AI と見立てた Bard データと GPT4 データを識別させると GPT4 データへの正解率は 0.96 である一方，Bard データへの正解率は 0.52 であった．そのモデルに対して組み込むデータを Bard データに変えて GAN の追加学習を行うと，Bard データに対しての正解率は 0.89 まで上昇したが，0.9 だった GPT3.5 データへの正解率は 0.66 まで低下した.続いて GPT3.5 データと Bard データ両方を組み込んで追加学習を行ったところ，2000 回程度の学習ループで人間データは 1.0，GPT3.5 データは 0.92，GPT4 データは 0.96，Bard データは 0.94 の正解率が得られるモデルとなった.組み込むデータ数としては 1 個では大きな変化は見られず，3～5 個程度からそれぞれの正解率に変化が見られ，GPT3.5 と Bard データそれぞれ 10 個程度組み込んで追加学習を行うことで全てのデータに対して 0.9 程度の正解率が得られた.

５　今後の取り組み

今回のデータセットは 200 文字以上 500 文字以下のウクライナに関する文章に限定した.その条件を拡張したときの識別精度や学習に必要な AI 文章の個数などを検証していきたい.また追加学習手法について，他のモデルへの適応も可能ではないかと考えており，その検討も続けていきたい.

参考文献

[1]岡山光平，石川博，廣田雅春：「フェイクニュース分類器を用いた口コミサイトのレビューの分析」DEIM Forum 2018 P3-5

[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」，研究報告知能システム(ICS)2019-ICS-193 11 号 P1-8 (2019-02-19)

[3]神田泰誠，柳烨佳，金明哲「C41D - 2 著者推定における異なる事前学習データを持つ日本語版 BERT の性能比較分析」，日本行動計量学会大会抄録集 C41D-2 (2022)

[4]岩田一樹「敵対生成ネットワークによる文書分類」感性福祉研究所年報 22 号 p19-31(2021-03-31)