5V-02

BERTを用いたSNS上における攻撃的文章訂正システム

吉田　基信 †　　松本　和幸‡　　吉田　稔‡　　北　研二 ‡

徳島大学理工学部 †　　徳島大学院社会産業理工学研究部 ‡

1　はじめに

現代社会においてSNSの発展は凄まじく，ほとんどの人が使用しているという状況に置かれている．それに伴い，誹謗中傷や炎上などの被害にあう件数も増加している．

2020年には，某テレビ番組における出演者の行動や言動に対して，Twitter上で誹謗中傷が多数投稿されたことから，自殺に追い込んでしまうといった事件が起きている．

本研究では，SNSユーザが発言を投稿する前に，投稿文の攻撃性を判定し，攻撃性を緩和した文章に変換して提示するシステムの構築を目指す．

2　関連研究

大西ら [1] は，入力文章に対して，SVMを用いて炎上可能性を判定した後，炎上すると判断された単語に対して，日本語ツイートで学習したword2vecを用いてテキストを訂正している．この研究では炎上検知度についてF値0.74と良好な結果が得られている．しかし，彼らの手法では，日本語として意味を成さない文が訂正後の文として出力される場合がある点において改善の余地がある．また山腰ら [2] は，意味や読みが互いに類似する法令用語に対して厳密に書き分けるための分類器をBERTを用いて作成し，正しい表現に校正する予測モデルを構築している．

本研究では，文の意味をとらえることに適しているモデルとしてBERTを用い，攻撃性の有無の分類およびテキストの訂正を行う．

3　提案手法

本節ではシステムの機能について説明する．システムは次にあげる4つの機能によって，入力文の攻撃性の判別とその訂正を実現する．これらすべての機能はBERTに基づき構成する．

(1) 入力テキストの前処理
(2) 入力テキストに対する攻撃的文章の判定
(3) 攻撃的表現を緩和したテキストの提案
(4) 元の文章との類似度の表示

3.1　入力テキストの前処理

(1) では，入力テキストに対し，前処理ライブラリであるTweetL[3] を基に文字列の正規化を行う．TweetLでは，ハッシュタグやメンションの除去，全角英数字は半角，半角カタカナは全角にするなどの処理が行われている．

3.2　BERTによる攻撃的文章の判定

(2) では，入力テキストに対する攻撃性の判定を行う．使用する分類器は，株式会社レトリバ [4] が国立国語研究所と共同で開発した話し言葉コーパスに基づき学習されたパラメータを初期パラメータとし，安全・攻撃的・スパムに分類するためのタグ付きテキストデータをファインチューニングに用いる．

このテキストデータは，Twitterから収集した2000件のツイートを成人男性4人により安全・攻撃的・スパムの3種類のタグを付与したデータであり，このうち，学習データとして1600件を用いる．残り400件のデータはテストデータとして用いる．

タグ付けにおいて，意味不明なものや単語の羅列，広告文章，誘い文句などをスパムとして分類する．スパム文章は内容そのものが不適格なものであるため，提案システムにおける変換処理から除外するものとする．

3.3　攻撃的表現を緩和したテキストの提案

表 1．MASK 変換例

マーチ関関同立は[f ラン]，低学歴でしょうに
↓
マーチ関関同立は [MASK]，低学歴でしょうに
↓
マーチ関関同立は当然，[低学歴]でしょうに
↓
マーチ関関同立は当然，[MASK] でしょうに
↓
マーチ関関同立は当然，不可能でしょうに

(3) では，(2) で攻撃的と判断されたものに対して，表1のような変換処理を行う．BERTの中にあるTransformerの第12層目のAttentionの重みを抽出し，文章内における重みが閾値以上の表現に対して変換処理を行う．この変換処理を行った後，事前に収集した危険単語を含んでいた場合，その表現に対してもBERTによるMASK予測変換処理を行う．該当する表現を一度にMASK予測変換処理するのではなく、該当箇所を一度抽出した後、文章に先に出てくる順番に並び替えてから一つずつ変換を行う。

The system to correct aggressive text with BERT on SNS

† Yoshida Motonobu, Tokushima University

‡ Matsumoto Kazuyuki, Yoshida Minoru, Kita Kenji, Tokushima University, Graduate School of Technology, Industrial and Social Sciences

表 2. 文章訂正手法

変換前	attention 語句	危険語句	変換後	類似度
鉄オタは新左翼と同じくらい内ゲバが好き	鉄オタ・内ゲバ・好き	新左翼	鉄人は人間と同じくらい数が多い	0.77
左翼は頭いかれたやつしかおらんのか？	頭	左翼	これはもういかれたやつしかおらんのか?	0.85

3.4 元の文章との類似度比較

(4) では，変換前と変換後で意味の変化が大きく変化していないかを BERTscore を用いて判断する．BERTscore は，Zhang[5] らが提案している手法であり，2 つの文章に対するベクトル表現を利用して，各トークン間のコサイン類似度を用いて文章間の類似度を測るものである．

4 評価実験

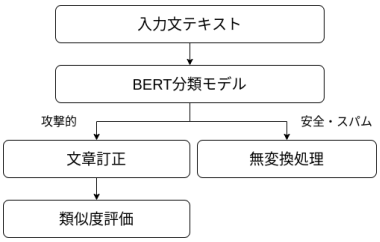

図 1. 評価手順

テストデータ 400 ツイートを用いて提案モデルの評価実験を行う．400 ツイートにタグ付けされた文章の個数の内訳は，安全：攻撃的：スパムが 213:153:34 になっている．攻撃的文章を判定する精度の評価と，訂正した文章の提案，元の文章と訂正した文章の類似度に対する評価を行う．

4.1 分類精度の評価

安全：攻撃的：スパムに対する F 値は表 3 のようになった．「安全」に関しては分類精度が比較的高く，「攻撃的」や「スパム」に対する分類精度は低い．

表 3. 分類精度

分類	安全	攻撃的	スパム
F 値	76%	70%	63%

4.2 訂正文に対する評価

元の文章と比較した例を表 2 に示す．変換後の文章では攻撃性は無くなっているものの，元の文章と比較して意味が大きく変化してしまっている．この要因としては，MASK 単語予測の際に意味的に異なる単語への変換を行ってしまったことがあげられる．

4.3 変換後の文章の意味変化に関する評価

変換前と変換後の文章間の類似度を，図 2 に示す．この図より，攻撃的と判断されたものの 8 割以上は，85% 以上は意味が似ていると判断されているが，残りの文章はそれ以下と意味が変化してしまっていることがわかる．

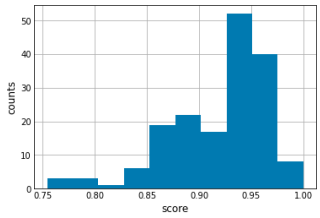

図 2. 文章類似度の分布

5 考察

攻撃的な文章の分類精度が低くなっている要因に，ファインチューニングで用いた 1600 個のデータの内容が限定的に加えて，数が少ないことが要因に挙げられる．また，訂正文に関しては，変換する用語の品詞を考慮していないため，意味が大きく変わったり，不自然な文章になってしまっていると考えられる．そして，誹謗中傷や差別表現など，BERT の事前学習に用いられているコーパスに含まれないような表現への対応が困難であることも原因の一つと考えられる．さらに，類似度評価の低いものは短文であることが多く，長文と比べて変換した際の影響の大きさが出ている．

6 終わりに

本研究では，BERT を用いて攻撃的文章の訂正を試みた．攻撃性の判別精度は，3 値分類でありながら 70% を達成できた．文章訂正については改善の余地があるが，今後，ファインチューニングに用いるコーパスの量を増やすことにより，さらなる精度向上が期待できる．

謝辞

本研究は JSPS 科研費 JP20K12027 の助成を受けたものです。

参考文献

[1] 大西 真輝, 澤井 裕一郎, 駒井 雅之, 酒井 一樹, 遠藤裕之, ツイート炎上抑制のための包括的システムの構築, The 29th Annual Conference of the Japanese Society for Artificial Intelligence,pp.3O1-3in (2015)

[2] 山腰貴大, 駒水孝裕, 小川泰弘, 外山勝彦, 事前学習モデル BERT による法令用語の校正, The 34th Annual Conference of the Japanese Society for Artificial Intelligence ,pp.4P3-OS-8-05(2020)

[3] TweetL,https://github.com/deepblue-ts/Tweetl(参照 2021-12-22)

[4] CSJ を用いた日本語話し言葉 BERT の作成, 勝又智, 坂田大直, The 27th Annual Conference of the associTheation for Natural Language Processing, pp.805-810(2021)

[5] BERTSCORE: EVALUATING TEXT GENERATION WITH BERT,Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger,Yoav Artzi,ICLR(2020)