E-001

攻撃的文章判定を用いた不適切文章の訂正

Correction of inappropriate sentences using toxic sentence judging

吉田 基信† 松本 和幸† 吉田 稔†
Motonobu Yoshida Kazuyuki Matsumoto Minoru Yoshida

1. はじめに

近年, SNS 利用者が急増しており, 8 割以上の人が使用しているという統計データもある. これに伴い, SNS 上で誹謗中傷や炎上などの被害に遭う件数も増加している.

2020 年には某番組内でタレントがとった行動や言動への Twitter 上での誹謗中傷が, 結果的に自殺を幇助してしまった事例もある. 2022 年には回転寿司店での不適切行為に対して不特定多数の人たちが非難を浴びせる「私刑」と呼ばれる現象も問題となった.

このような事件を回避する方法として我々は, 攻撃的な内容を含んだ文章を投稿前に検出し, 適切な文章に訂正するシステムを提案する.

2. 関連研究

大西ら[1]は, 入力された文章に対して 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.

また Reid ら[2]は, Lewis という編集・合成フレームワークを用いて多くの文章タイプに依存しない文章変換を試みている. 変換対象となる文章に対して挿入・変換・削除といったどのプロセスが適切かを判断し, 該当箇所を SLOT に置き換えることにより教師なしデータとして処理できる普遍的なスタイルを生成する. そしてそのスタイルの SLOT の部分に対して事前学習済みモデルを用いて予測を行い, 適切な文章変換を試みている.

本研究では, 吉田ら[3]の分類モデルを使用しつつ, ChatGPT をはじめとする生成 AI により注目を集めている大規模言語モデルの一種である GPT を使用することで適切な文章の生成を試みる.

3. 提案手法

本章では提案手法の流れについて説明する. 実際に行った手順は以下のとおりである.

Step 1. 攻撃的文例の判断・収集
Step 2. 訂正文生成システムの構築
Step 3. 訂正文生成・評価

以上の流れを詳細に説明していく.

3.1. 攻撃的文例の判断・収集

Step1 では, Twitter API を用いて無作為に収集した文章約 300 万件を吉田ら[3]が作成した分類モデルを用いて攻撃性判定を行い, 約 3 万件まで数を絞り込んだ. その後, 目視により攻撃的と思われる文章を 120 件抽出した.

この抽出した 120 件の文章に対し, 攻撃的要素を最小限にした適切な文章を手動で生成し, 実験に用いる訓練データとする.

3.2 訂正文生成システムの構築

Step2 では, 入力文に対して適切な文に訂正するシステムを構築する. 構築するシステムには rinna 社が公開している GPT-2[5]及び GPT-NeoX[6]の事前学習済みモデルを, 3.1 節で述べた訓練データを用いてファインチューニングしたものを用いる.

ファインチューニングする際は, 表 1 に示すテンプレートを用いてデータ形式を変換する.

表 1：訓練データの形式

データの項目	データの内容
Category	Brainstorming
Instruction	次の文章は攻撃的です。 「<攻撃的文章>」。 この文章を適切な文章に変換してください。
Output	<適切な訂正文>

<攻撃的文章>及び<適切な訂正文>にはそれぞれに対応する手動の生成文が入る. 文章を生成する際に, GPT-NeoX では instruction の形式で入力を与える.それ以外の手法では Instruction の語尾を変化させた次のような形式で入力を与える.

・” 次の文章は攻撃的…変換すると「 ”

3.3 訂正文生成・評価

Step3 では, Step2 で作成したモデルを用いて訂正文の生成を行い, 評価を行う. 吉田ら[3]のシステムが生成した文章と比較し, 文章構成や文章全体の意味の変化度にどのように違いが出たかを比較する.

4. 評価実験

分類モデルで攻撃的と判断された文章を 10 件程度用意し, 各訂正文生成システムにより訂正を行った.

今回実験に用いる訂正文生成システムは以下の 3 種類である.

(1) 吉田ら[3]の訂正システム
(2) GPT-2 を用いた文章生成システム

†徳島大学 Tokushima University

(3) GPT-NeoX を用いた文章生成システム

簡潔にシステムと結果を説明する.

4.1 吉田ら[3]の訂正システム

この訂正システムでは BERT の単語予測を用いることで, Transformers[7]の Attention の値や危険語句辞書に基づき抽出した攻撃的と判断された単語の変換を行う.

4.2 GPT-2 を用いた文章生成システム

この生成システムは, 事前学習モデルに rinna-japanese-gpt2-medium を使用し, 3.1 節で述べた訓練データセットを用いてファインチューニングする. 3.2 節で述べたデータ形式に変換してファインチューニングを行う.

4.3 GPT-NeoX を用いた文章生成システム

この生成システムは, 事前学習モデルに rinna-japanese-gpt-neox-3.6b を使用し, 4.2 節と同様の方法でファインチューニングを行う.

4.4 各モデルの比較

各モデルを用いて, 攻撃的文章を適切な文章に変換した. 結果の一例を表 2, 3 に示す. 太字は攻撃的と判断できる箇所である.

表 2：各モデルの文章変換結果

例文	**左翼**は**頭いかれたやつしかおらん**のか?
吉田ら	これはもう**いかれたやつしかおらん**のか?
GPT-2 (normal)	お前は**左翼**じゃないだろう!
GPT-NeoX (normal)	(**左翼**は**頭いかれたやつしかおらん**のか?)と、思われているのではないでしょうか?
GPT-2 (Fine-tune)	**左翼**は不謹慎だけど**左翼**も**頭イかれた**んだよね
GPT-NeoX (Fine-tune)	私は**左翼嫌い**だけどそういうのしかいないんだろうな...

表 3：各モデルの文章変換結果 2

例文	人間を**馬鹿**にして**踏み潰すクソ**は**死ね。頭イカレてんの**か
吉田ら	人間を前にして踏み潰す人間は**死ね**。かどう生きてんのか
GPT-2 (normal)	僕と君は、喧嘩した覚えがない
GPT-NeoX (normal)	人間というモノを踏み潰せる貴様は**死ね**
GPT-2 (Fine-tune)	**頭がイカレてんな**(略)
GPT-NeoX (Fine-tune)	人間じゃねぇよ**クズ野郎**

5 考察

吉田らのモデルの問題点であった文構造の崩れは軽減できたものの, 依然として攻撃性がみられる文が生成されることがあった. しかし, 全体としてわずかであるが攻撃性が抑制されており, GPT による訂正手法は有効であると考えられる.

6 おわりに

本研究では, 攻撃的文章を適切な文章に変換するために, 生成モデルをファインチューニングして自然な構造かつ可読性の高い文章への変換を目指した. 実験の結果, 自然な文章の生成がみられたが, 攻撃性をより小さくする点に課題が残る結果となった. 今後, ファインチューニングの量を増やすことや, 追加で処理を行う工夫を行うことでさらなる適切な文章の生成を目指したい.

謝辞

この研究は JSPS 科研費 20K12027, 21K12141 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, 人工知能学会第 29 回全国大会, pp.3O1-3in (2015)

[2]M. Reid and V. Zhong: LEWIS: Levenshtein Editing for Unsupervised Text Style Transfer. In: Findings of the Association for Computational Linguistics (ACL-IJCNLP), pp.3932-3934 (2021)

[3]吉田基信, 松本和幸, 吉田稔, 北研二, BERT を用いた SNS 上における攻撃的文章訂正システム, 情報処理学会第 84 回全国大会, pp.725-726 (2022)

[4]趙天雨, 沢田慶, 日本語自然言語処理における事前学習モデルの公開, 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol.93, pp.169–170 (2021)

[5]A. Redford, J, Wu, R,Child, D. Luan, D. Amodel, I. Sutskever. Language Model are Unsupervised Multitask Learners, Open AI blog, 1(8), 9. (2019)

[6]S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leady, K. McDonell, J. Phang, M. Pieler, USVSN S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, S. Weinbach. GPT-NeoX-20B: An Open-Source Autoregressive Language Model, arXiv:2204.06745

[7]A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. Attention is all you need. In Advances in neural information processing systems, pp5998-6008 (2017)

[8]勝又智, 坂田大直, CSJ を用いた日本語話し言葉 BERT の作成, 言語処理学会第 27 回年次大会発表論文集, pp.805-810(2021)