5ZA-09

誹謗中傷表現やネガティブ表現を別の言い回しに変換するユーザー参加型変換システムの提案

上北真也† 塚田晃司†
和歌山大学システム工学部†

1. はじめに

近年インターネット上でのトラブルが顕著になり，こころない投稿による事件が発生している．

総務省では2020年8月，「インターネット上の誹謗中傷への対応の在り方に関する緊急提言」を公表する[1]など，ネット上での誹謗中傷問題はすでに大きな社会問題となっている．

誹謗中傷を書き込む原因のひとつとして，ユーザーの情報モラル・ICT リテラシーの欠如が考えられる．

そこで本研究では，事業者による投稿削除やペナルティを科すといった方法ではなく，①ユーザー自身の誹謗中傷表現に対する意識を向上させること②誹謗中傷表現を吐き出せる場所を提供することを目的とし，誹謗中傷表現を含む文章を柔らかな表現の文章へ翻訳する，ユーザー参加型システムを提案する．

2. 関連研究・関連サービス

誹謗中傷に関するサービスとしては，ネガティブバスター[2], SNS PEACE[3], しずかったー[4], matte[5]などがある．ネガティブバスターは誹謗中傷ワードを自動でポジティブワードに変換して表示するサービス．SNS PEACE は Twitter 上での誹謗中傷メッセージや不快な画像の自動ミュート/非表示にするサービス．しずかったーは暴言・悪口・中傷を、綺麗な言葉やオブラートな言い方に翻訳するアプリであった．matte は投稿者がトラブルの元となる不適切な投稿内容を SNS 等インターネット上に投稿する前に検知し，内容再考の機会を促すアラートを出すサービスである．

matte を除くといずれの研究･サービスもユーザーの情報モラル・ICT リテラシーの向上といった根本的な解決方法には貢献できないという課題点がある．また今回提案する手法は matte と併用することでより効果的なサービスになると考えられる．

A proposal for a collaborative translation system that converts slanderous and negative expressions into less offensive ones.
†Shinya Uekita , Koji Tsukada
†Faculty of Systems Engineering Wakayama University

3. 提案手法

3.1 システムの構成

図1に本システムの構成を提示する．本システムは，データベースと連携した web ページをブラウザ上に表示することで実現する．開発環境は windows10 で XAMPP を用いた．XAMPP を用いることで，誹謗中傷的な表現の単語（以下 NG ワード）とその単語の中傷表現を抑えた言いまわし（以下変換後の表現）のリストを管理するデータベースである MySQL と，web サーバーである Apache を用意した．また形態素解析には日本語形態素分析[6]を用いた．

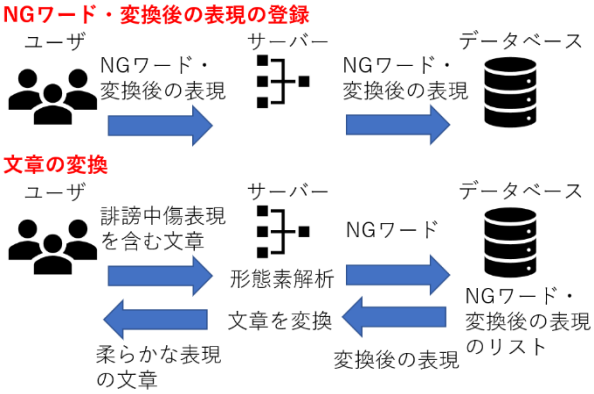

図1：システム構成図

3.2 システムの処理

まず，ユーザーに NG ワードと変換後の表現を入力してもらい，入力された情報をデータベースに挿入する． 次に，ユーザーに誹謗中傷表現を含む文章が入力された時，文章を形態素分析し，データベースと一致する NG ワードがあれば，その単語を変換後の表現に置き換えて出力する．このシステムにより，誹謗中傷的表現を含む文章を柔らかな表現の文章へ翻訳することができる．

またユーザーに NG ワードと変換後の表現を登録してもらう過程で，誹謗中傷的表現を用いない言い回しをユーザーが考える機会が生まれ，ユーザーの情報モラル・ICT リテラシーが向上することが期待できる．

例として NG ワードと変換後の表現を表1に，誹謗中傷表現を含む文章と・変換後の文章を比較したものを表2にまとめる．

表 1：NG ワードと変換後の表現の比較

NGワード	感情値	変換後の表現	平均感情値
馬鹿	-0.964567	天才と紙一重	0.197414
無愛想	-0.995069	クール	-
あほう	-0.823749	どじっ子	-0.740048
嫌い	-0.629629	これから好きになる	0.257467
死ぬ	-0.999999	人生のゴールテープを切る	-0.445304

赤文字・下線部は単語感情極性表と一致したものを示す.

表 2：誹謗中傷表現を含む文章と変換後の文章の比較

誹謗中傷表現を含む文章	平均感情値
馬鹿は嫌い．早く死ぬことを願う.	-0.717099
変換後の文章	**平均感情値**
天才と紙一重はこれから好きになる．早く人生のゴールテープを切ることを願う.	-0.1089317

赤文字・下線部は単語感情極性表と一致したものを示す.
波線部は NG ワードを変換した部分を示す.

4. 評価・考察

単語・文章ごとにどれほどポジティブ・ネガティブであるかを示した数値である感情値・平均感情値を算出することで，今回の提案システムを評価する．感情値は単語感情極性表[7]を用いて求める．感情値は -1～1 で表され，-1 に近いほどネガティブ，1 に近いほどポジティブであると表現できる．

文章中に含まれる単語のうち，単語感情極性表と一致した単語を w1,w2,…,wn とし，それに対応する感情値を f(w1),f(w2),…,f(wn)と定義する．平均感情値 F は，式(1)を用いて求められる．

$$F = \Sigma \frac{f(wn)}{n} \quad (1)$$

今回の提案システムを用いて変換した結果は，NG ワードと変換後の表現の比較（表 1）・誹謗中傷を含む文章と変換後の文章の比較（表 2）のいずれにおいても，平均感情値の値が 1 に近づいていることから，よりポジティブな表現に変換できていることがわかる．

しかし実際に誹謗中傷表現を含む文章より変換後の文章の方が人を傷つけない表現になっているかどうか，このシステムが誹謗中傷表現を吐き出せる場所を提供できているか，は今後アンケート調査をもとに確かめる必要がある．

また，ユーザーに NG ワードと変換後の表現を登録してもらう過程で，ユーザーの情報モラル・ICT リテラシーが向上するかどうかについても調査が必要である．

5. おわりに

本稿では，誹謗中傷的表現を含む文章を柔らかな文章に翻訳するシステムを提案し，実装した．ユーザー自身の誹謗中傷表現に対する意識を向上させることと，誹謗中傷表現を吐き出せる場所を提供することを目標としている．

今後の課題として，継続的にこのシステムをユーザーに使わせる工夫が必要である．具体的には，ユーザーが登録した変換表現をほかのユーザーが評価するシステムや，ユーザーが入力した誹謗中傷表現を含む文章と，変換後の文章それぞれの感情値を極性分析で数値化し，表示するシステムなどを導入する．

また，このシステムを使うことで，ユーザーの情報モラル・ICT リテラシーが向上するか，誹謗中傷表現を含む文章より変換後の文章の方が人を傷つけない表現になっているか，変換後の文章の意味が伝わるか，誹謗中傷表現を吐き出せる場所になっているかについてもアンケート調査が必要である．

参考文献

[1] 総務省：「インターネット上の誹謗中傷への対応の在り方に関する緊急提言」及び意見募集の結果の公表，入手先 <https://www.soumu.go.jp/menu_news/s-news/01kiban18_01000092.html> (参照 2020-12-27).
[2] nanka：ネガティブバスター，入手先 <https://nankasince2016.jimdofree.com/>(参照 2020-12-21)
[3] GMO インターネット株式会社：SNS PEACE，入手先<https://sns-peace.com/>(参照 2020-12-21)
[4] TOYOTA MOTOR CORPORATION:しずかったー
[5] アディッシュ株式会社：matte，入手先 <https://www．adish.co.jp/news/20200915_spd/>(参照 2020-12-21)
[6] Yahoo!デベロッパーネットワーク：日本語形態素解析，入手先 <https://developer.yahoo.co.jp/webapi/jlp/ma/v1/parse.html>(参照 2021-01-08)
[7] 高村大也：単語感情極性対応表，入手先 <http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html>（参照 2020-12-27）