誹謗中傷表現を言い換える変換システムの提案と評価

上北真也[1]　塚田晃司[2]

概要：近年ネット上での誹謗中傷が社会問題となっている．そこで本研究では，複数のユーザーに誹謗中傷表現とその言い換え表現を辞書に登録してもらい，そのデータを元に言い換えを行うシステムを提案する．本提案システムを用いることで，誹謗中傷に対するモラルの向上や，誹謗中傷表現を吐き出せる場所を提供するといったアプローチから，誹謗中傷問題の解決に貢献する．本研究では，言い換え結果を単語感情極性対応表とアンケート調査にて評価を行った．また，システム全体についてもアンケート調査にて評価した．

キーワード：誹謗中傷，言い換え，翻訳，集合知

Proposal and Evaluation of a Conversion System for Paraphrasing Slanderous Expressions

SHINYA UEKITA[†1]　KOJI TSUKADA[†2]

Abstract: In recent years, slander on the Internet has become a social problem. In this study, we propose a system that allows multiple users to register slanderous expressions and their paraphrases in a dictionary, and paraphrase them based on the data. The proposed system will contribute to solving the slander problem by improving the morality against slander and by providing a place where slanderous expressions can be vented. In this study, we evaluated the paraphrase results using a word-emotion polarity correspondence table and a questionnaire survey. We also evaluated the system as a whole using a questionnaire.

Keywords: Slander, defamatory language, rephrasing, translation, wisdom of crowds

1. はじめに

近年スマートフォンやパソコンの普及が進み，社会生活の場だけでなく，インターネット上も交流の場として盛んに利用されている．しかし，匿名で自由に投稿できる特徴から，ユーザー同士の交流を促進している反面,こころない投稿によって,有名人・学生などが自ら命を絶ってしまうといった事件が発生するなど，インターネット上でのトラブルが顕著になっている．総務省においても 2020 年 9 月，「インターネット上の誹謗中傷への対応に関する政策パッケージ」[1]を公表するなど，インター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．

本研究では，ユーザー自身に誹謗中傷表現やネガティブな表現を含む言葉（以下 NG ワードと表記する）と誹謗中傷表現やネガティブな表現を抑えた表現（以下言い換え表現と表記する）を考えてもらい，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，また，誹謗中傷の投稿というネガティブな情報をシェアするのではなく，その言い換え表現を投稿することでポジティブな情報としてシェアすることを通じて，誹謗中傷表現を吐き出せる場所を提供することいったアプローチから誹謗中傷問題の解決を支援することを目的とする．本研究では，ユーザーに誹謗中傷表現とその言い換え表現を登録してもらい，そのデータを元に言い換えを行うシステムの開発と評価を行う．

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．

誹謗中傷表現を集めた辞書を作成している研究として，いじめ表現辞書を用いた Twitter 上のネットいじめの自動検出[6]がある．この研究では Twitter 上のテキストを対象とし，いじめ表現辞書を構築し,各単語にいじめ度を付加している．いじめ度とは，単語とその単語がどれだけいじめ

1 和歌山大学　大学院システム工学研究科
Graduate School of Systems Engineering, Wakayama University
2 和歌山大学　システム工学部
Faculty of Systems Engineering, Wakayama University

と関連するかの程度を数値で表すものである．辞書に登録する単語は，特定の単語を使って収集したツイートに含まれている単語とし，その単語につける値は，SO-PMI というものを用いて算出している．また,構築したいじめ表現辞書を含む複数の特徴量を用いて複数の機械学習手法と組み合わせ,ネットいじめの自動検出に最適なモデルの構築を図っている.構築したモデルを用いていじめ文,非いじめ文の分類の評価を行ったところ，多くの機械学習手法でいじめ表現辞書が正しい検出に貢献すること,さらに最も良かったモデルでは90%を超える評価を得たことを述べている．

さらに，日本語の言い換えに関する研究として，感性を考慮した日本語俗語の標準変換[7]がある．この研究では，アンケートによって得たデータをもとに，若者言葉を多次元の印象軸（感性評価軸）と，意味（概念）ベクトルによって表現することで，意味的にも感性的にも類似した標準語に変換する手法を提案している．

3. 提案手法

提案手法の概要（図 1）について述べる．

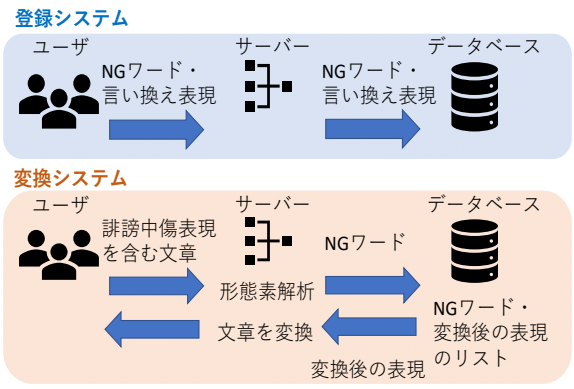

図 1　システム構成図

Figure 1　System configuration diagram.

インターネット上では誹謗中傷をするユーザーが存在する．そこで，誹謗中傷表現を含む文章をそのまま投稿するのではなく，NG ワードと言い換え表現をユーザー自身に考えさせ，それを投稿し変換表現をシェアできるシステムを作成する．登録システムでは，ユーザーが考えた NG ワードと，その言い換え表現をデータベースに登録する．変換システムではこのデータを使って誹謗中傷表現を含む文章を入力したときに，その文章をべつの言い回しに変換する．変換システムでは，誹謗中傷を含む文章が入力された時，文章を形態素解析し，単語ごとに分ける．その後登録システムで作成した，NG ワードとその言い換え表現がセットになったリストを用いて，誹謗中傷表現を言い換え表現に変換して出力する．

変換システムでは，登録システムで登録されていない NG ワードを言い換えることができない．しかし，うまく変換されないときは，ユーザーが登録システムを使い，NG ワードと言い換え表現を登録することで，少しずつ言い換えられる表現が増えていく．このことを通して，登録システムを何度もユーザーに使ってもらうことが期待できる．

図 2 を用いて具体的に説明する．NG ワードと言い換え表現の登録システムを用いて NG ワードに「馬鹿」・「嫌い」・「死ぬ」を，それぞれと対応する言い換え表現として「天才と紙一重」・「これから好きになる」・「人生のゴールテープを切る」が事前に登録されていたとする．その後，変換システムの方で，「馬鹿は嫌い．早く死ぬことを願う．」という文が入力されると，文章を形態素解析し，単語ごとに分ける．分けられた単語の中で，登録されている NG ワードと一致するものがあれば，言い換え表現に置き換えて出力する．この一連の流れによってシステムを実現した．

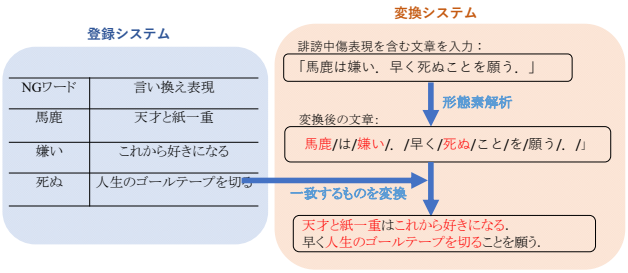

図 2　システムの概要図

Figure 2　System overview diagram.

4. 実装

想定環境を図 3 に示す．本研究では，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した．本研究の提案システムは，将来的に，不特定多数のユーザーが web ページ上で提案手法を用いたサービスを継続的に利用できる「ユーザー参加型」のコンテンツとして作り上げることを想定している．しかし，今回は，試験的に localhost 内でのみ使えるサービスとして構築した．

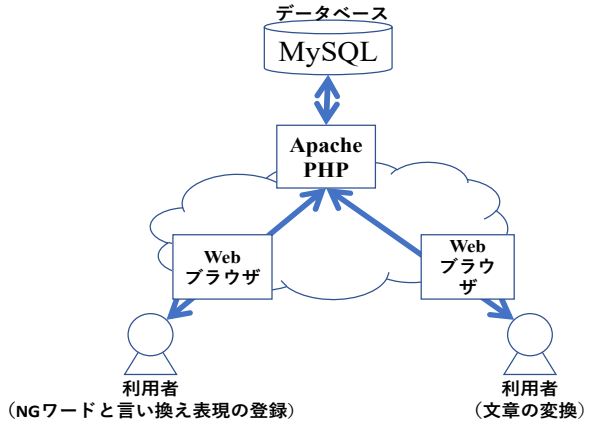

図 3　想定環境

Figure 3　Assumed environment

4.1 開発環境

今回提案手法のためのプログラムを作る手立てとして PHP というプログラミング言語を使用した．また，情報を格納するデータベースも必要なので SQL も使用した．開発環境としては windows10 を用いて XAMPP で PHP を扱うために XAMPP を使用した．XAMPP は PHP を用いるための Apache がついている，また，データベースと SQL を扱うための MySQL もついている．データベースの操作は phpMyAdmin をネットワーク上から使用して行い，PHP は localhost からネットワークを用いて表示した．

4.2 システムの処理

システムの処理の大まかな流れを図 4 に示す．

まず，登録システムの処理について述べる．Web ページにて入力された NG ワードとその言い換え表現は，mysql の NG ワード言い換え表現管理テーブルに追加する．その後登録した NG ワードとその言い換え表現，すでに登録されている NG ワードとその言い換え表現が確認できる画面を表示する．NG ワードと言い換え表現を管理テーブルの詳細を以下の表 1 にまとめる．

次に登録システムの処理について述べる．Web ページにて入力された誹謗中傷表現を含む文章を，日本語形態素解析[17]を使うことで，単語ごとに分ける．分けられた単語のうち，NG ワード言い換え表現管理テーブルの NGword カラムと一致するものがあれば，hennkanngo カラムにある言い換え表現に置き換えて出力する．

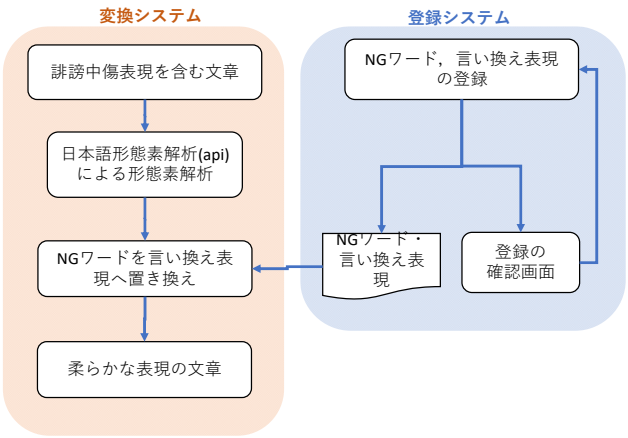

図 4　システムの処理の流れ

Figure 4　Processing flow of the system

4.3 システムの機能

この節ではシステムに実装した機能を紹介する．

4.3.1 NG ワードと言い換え表現の登録・表示

システム起動時に，NG ワードと言い換え表現を登録できるページを表示する．図 5 にそのページ画面を示す．ユーザーは誹謗中傷ワードと書かれているテキストボックス内に NG ワードを，オブラート変換と書かれているテキストボックスに，言い換え表現を記入し辞書登録ボタンを押す．この動作を行うことで，NG ワードと言い換え表現を登録できる．また，登録ボタンを押すと，登録後の確認画面のページが表示される．

上記の動作を行うことで，NG ワード言い換え表現管理テーブル（表 1）の NGword カラムに NG ワードが，hennkanngo カラムに言い換え表現が挿入される．

図 5　登録画面

Figure 5　Register NG words and paraphrased expressions

表 1　テーブルの詳細

Table 1　Table Details

カラム名	保存するデータ	役割
number	整理番号	登録された NG ワードや言い換え表現を整理するための通し番号.
NGword	誹謗中傷表現である単語(NG ワード)	NG ワードを管理する. 変換システムでは，誹謗中傷表現を含む文章と一致する単語があれば使用する.
hinsi	NG ワードの品詞	今回のシステムでは未使用．今後システムを拡張し，自然言語処理を扱う際に使用する予定.
hennkanngo	NG ワードを柔らかい表現に言い換えた単語	登録された言い換え表現を管理する 変換システムでは NG ワードを言い換え表現に変換するときに使用する
user	アカウント名	今回のシステムでは未使用. 今後システムを拡張し，アカウント認証等を扱うときに使用する.
good	言い換え表現の「いいね」数	今回のシステムでは未使用. 今後言い換え表現の「いいね」等を他社が評価できるようにする際に使用する予定.
bad	言い換え表現の「悪いね」数	今回のシステムでは未使用. 今後言い換え表現の「悪いね」等を他社が評価できるようにする際に使用する予定.
daytime	登録日時	NG ワード・言い換え表現が登録された日時を管理する.

4.3.2 NG ワードと言い換え表現の登録・表示

この項では誹謗中傷表現を含む文章が入力された時，それを柔らかい表現に変換して表示する機能について詳しく述べる．

システム起動時に，誹謗中傷表現を含む文章が入力された時，それを柔らかい表現に変換して表示するページを表示する．図 6 にそのページを示す．ユーザーは上側にあるやわらか翻訳テキストボックス内に誹謗中傷表現を含む文章を記入し翻訳ボタンを押す．この動作を行うことで，登録されている NG ワードと言い換え表現をもとに，誹謗中傷表現を含む文章をやわらかな表現に変換した文章が出力される．

図 6　柔らかい表現に変換するページの表示画面
Figure 6　Screen for converting to softer expressions.

誹謗中傷表現を含む文章が入力された時，日本語形態素解析を使うことで，文章を単語ごとに分ける．分けられた単語のうち，NG ワード言い換え表現管理テーブル（表 1）の NGword カラムと一致するものがあれば，hennkanngo カラムにある言い換え表現に置き換えて出力する．

5. 評価

本章では開発したシステムの動作や使いやすさ，効果などに対する調査を行い，それらの結果に対する考察を行う．

誹謗中傷表現を含む文章を別の表現に言い換えて表示するシステムについては 5.1 節にて，単語感情極性対応表[8]を使うことで数値的に評価する．システム全体を評価は，5.2 節にて，このシステムについての説明を受けた 20 代を対象としたアンケートを行うことで評価する．

5.1 単語感情極性対応表による分析

この節では Twitter 上で実際にあった誹謗中傷を含む文章を，誹謗中傷を言い換えるシステムを使うことで，柔らかい表現に変換した文章に変換した（表 2）．それぞれの文章について単語感情極性対応表を用いてどの程度ポジティブな文章に変換できたのかを定量的に評価する．

まず，感情値について説明する．単語がどれほどポジティブ・ネガティブな意味を含んでいるかを示した数値を感情値と呼ぶ．感情値は単語感情極性対応表を用いて求める．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．

次に単語感情分析を使った評価方法について説明する．文章中に含まれる単語のうち，その単語の表記・基本形・品詞の 3 つが単語感情極性対応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．

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

例えば表 2，例文 1 の誹謗中傷表現を含む文章では，「馬鹿」という単語の感情値が-0.967567，「嫌い」の感情値が-0.629629，「死ぬ」の感情値が-0.999999，「願う」の感情値が-0.274203 であるため，その平均よりこの文章の平均感情値は，-0.717099 となる．

表 2　誹謗中傷表現を含む文章と変換後の文章との比較
Table 2　Comparison of sentences containing slanderous expressions and sentences after paraphrasing.

例文	誹謗中傷表現を含む文章	変換後の文章
1	馬鹿は嫌い．早く死ぬことを願う．	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．
2	お前嫌いだわ～、マジで	お前これから好きになるだわ～、マジで
3	性格ブスすぎ www 気持ち悪いよ	性格趣のあるお顔すぎ www そりが合わないよ
4	出て行けクソ女無責任野郎	出て行けとても女性無責任ヒーロー男
5	見てて腹立つんだよピンクゴリラ。早く出てって動物園に帰りやがれドブス	見ておへそが茶を沸かすんだよピンク霊長目ヒト科ゴリラ属。早く出てって動物園に帰りなされ印象的な方
6	自分で取り忘れた、ゴミ女 おわってんな 辞めちまえ	自分で取り忘れた、資源女性 かわってんな 肩の荷をおろしてはいかがでしょうか
7	メンヘラでキモっ	繊細さんで自分の感覚にあわなっ
8	ブス過ぎる。きもい	趣のあるお顔過ぎる。不思議な気持ち
9	てか死ねやくそが	てか仏さまになろうやうんちが
10	本当に最低です。話し方下品。手も出す。もうさっさと卒業お願いします。	本当にチョベリバです。話し方さもしい。手も出す。もう早めに卒業お願いします

表 2 で提示したそれぞれの文章について式（1）を用いることで平均感情値を求めた．その結果を表 3 に示す．値は小数点第 6 位まで求めた．

「ブス」・「クソ」・「メンヘラ」などの誹謗中傷表現が，単語感情極性対応表に載っていないなどの問題点があるが，誹謗中傷表現を含む文章の平均感情値の平均値は-0.616126 であった．一方，変換後の文章の平均感情値の平均値は-0.150700 であった．このことから，誹謗中傷表現を含む文章を別の表現に言い換えて表示するシステムは，誹謗中傷表現を含む文章をポジティブな表現に変換できたことが読み取れる．

表 3　平均感情値の比較

Table 3　Comparison of mean emotional values

例文	誹謗中傷表現を含む文章の平均感情値	変換後の文章の平均感情値
1	-0.717100	-0.127300
2	-0.629630	0.932040
3	-0.242544	-0.477496
4	-0.714652	-0.369700
5	-0.226015	-0.421250
6	-0.534227	-0.441569
7	一致する単語なし	-0.703120
8	-0.792350	-0.400906
9	-0.999999	0.991565
10	-0.688620	-0.489721
平均値	-0.616126	-0.150700

5.2 システムに対するアンケート評価

提案システムは，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ができているのか，また誹謗中傷の投稿というネガティブな情報をシェアするのではなく，その言い換え表現を投稿することでポジティブな情報としてシェアすることでストレス解消に役立つシステムになっているのかを調べる目的で，知人へのアンケート調査を行った．さらに，提案システムは将来的にユーザーに何度も使ってもらう想定であるため，各システムの見やすさや，おもしろさ，継続的に使いたいかどうか等についても調査した．

5.1 節では単語感情極性対応表を用いて誹謗中傷表現を含む文章を別の表現に言い換えて表示するシステムについて，定量的に評価した．しかし単語感情極性対応表だけでは，意味が通じる文章人っているか，また実際にユーザーがどのように感じるかを調べることができないため，アンケート評価にてこのことを調査した．表 4 にアンケート内容をまとめる．

表 4　アンケート調査の概要

Table 4　Overview of the survey.

アンケート調査の内容	提案システム全体について(5.3節) ① 見やすさ･分かりやすさ ② 継続的に使いたいか ③ おもしろさ ④ モラルの向上へ役立つか ⑤ ストレス解消に貢献できるか 誹謗中傷を言い換えるシステムについて(5.4節) ① 人を傷つけない表現に変換できているか ② 意味が通じる文章になっているか ③ 面白い表現と感じられる表現になっているか
被験者	15人
評価の仕方	5段階評価で(5が最も設問に同意とする) 良い+- 5- +- 4- +- 3- +- 2- +- 1-¦ 悪い

5.2.1 NG ワードと言い換え表現の登録・表示

各システムのページの見やすさ・分かりやすさの観点からアンケート調査を行った．5 段階(5 が最も設問に同意とする)で評価して頂いた調査結果を，表 5 に示す．各設問番号と質問内容は以下の通りである．

表 5　見やすさ・分かりやすさに関する調査結果

Table 5　Results of a survey on ease of viewing and understanding.

	分布					平均値	中央値	最頻値
	1	2	3	4	5			
設問1	0	0	0	6	9	4.6	5	5
設問2	0	1	1	7	6	4.2	4	4
設問3	0	0	0	7	8	4.53	5	5
設問1～3						4.44		

・設問 1 「図 4.2 のような，NG ワードと言い換え表現の登録画面について，見やすい・わかりやすいか.」
・設問 2 「図 4.3 のような，登録後の確認画面について，見やすい・分かりやすいか.」
・設問 3「図 4.4 のような柔らかい表現に変換するページの表示画面について見やすい・分かりやすいか.」

アンケート結果の平均値，中央値，最頻値は，いずれも 5 段階中 4 以上となっており，見やすさ・使いやすさの観点では，高評価であった．

5.2.2 継続的に使いたいか

システムを継続的に使いたいかどうかについてアンケート調査を行った．その結果を表 6 に示す．設問番号と質問内容は以下のとおりである．
・設問 4「システム全体を通して継続的にこのシステムを使いたいと思うか.」

表 6　継続的に利用するかに関する調査結果

Table 6　Survey results on continued use

	分布					平均値	中央値	最頻値
	1	2	3	4	5			
設問4	0	0	7	4	4	3.8	4	3

アンケートの結果は，平均値 3.8，中央値 4，最頻値 3 であり，継続的に使いたいかどうかについては，継続的に使いたいという意見がやや多い結果になった．

5.2.3 おもしろさ

システムをおもしろいと感じるかについてアンケート調査を行った．その結果を表 7 に示す．設問番号と質問内容は以下のとおりである．
・　設問 5「システム全体を通しておもしろいシステムだと思うか.」

表 7　面白さ関する調査結果

Table 7　Survey Results on Interestingness

	分布					平均値	中央値	最頻値
	1	2	3	4	5			
設問5	0	0	0	3	12	4.8	5	5

アンケート結果は，平均値 4.8，中央値・最頻値ともに 5 であり，非常にこのシステムをおもしろいと感じるユーザーが多いことがわかる．

5.2.4 モラルの向上

システムを通じて誹謗中傷表現に対する意識が向上すると感じるかについてアンケート調査を行った．その結果を表 8 に示す. 設問番号と質問内容は以下のとおりである．・設問6「システムを通して情報モラル・ICT リテラシーが向上すると思うか.」

表 8　モラルの向上調査結果

Table 8　Survey results on moral improvement

	分布					平均値	中央値	最頻値
	1	2	3	4	5			
設問6	1	3	3	7	1	3.27	4	4

アンケート結果は，平均値 3.27，中央値・最頻値がともに 4 であった．平均値・中央値・最頻値を元に考えると，モラルの向上に役立つことが期待できる．しかし，アンケート結果をみてみると，分布にばらつきがでており，個人差が大きい結果となった．

5.2.5 ストレス解消

このシステムが誹謗中傷表現を吐き出せる・鬱憤を晴らせるシステムになっているかについてアンケート調査を行った．その結果を表 9 に示す．設問番号と質問内容は以下のとおりである．

・設問 7「誹謗中傷表現を吐き出せる・鬱憤を晴らせるシステムになっているか」

表 9　ストレス解消に関する調査結果

Table 9　Survey results on stress reduction

	分布					平均値	中央値	最頻値
	1	2	3	4	5			
設問7	1	3	0	10	1	3.47	4	4

アンケート結果は，平均値が 3.47，中央値・最頻値がともに 4 であった．このことから提案システムが誹謗中傷表現を吐き出せる・鬱憤を晴らせるシステムになっていること．つまりネット上の誹謗中傷を削減する効果があることが期待できる．

5.3 言い換え結果に対するアンケート評価

この節では，誹謗中傷を言い換えるシステムを対象とするアンケート調査の結果を述べる．5.1 節で用いた，誹謗中傷表現を含む文章とそれを誹謗中傷を言い換えるシステムを使うことで変換した文章（表 2）を比較してもらった．

5.3.1 人を傷つけない表現に変換できているか

表 2 の文章を見て，誹謗中傷を言い換えるシステムが，人を傷つけない表現に変換できているかについてアンケート調査を行った．その結果を表 10・表 11 に示す．各設問番号と質問内容は以下のとおりである．

・設問 1「例文 1〜10 についてそれぞれ，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と思うか.」

・設問 2「例文 1〜10 についてそれぞれ，変換後の文章は，受け流せるレベルの表現になっていると思うか.」

表 10　設問 1 に関する調査結果

Table 10　Survey results for question 1

設問1	分布					平均値	中央値	最頻値
	1	2	3	4	5			
例文1	0	1	4	6	4	3.87	4	4
例文2	1	1	2	2	9	4.13	5	5
例文3	3	3	0	5	4	3.27	4	4
例文4	3	3	4	3	2	2.87	3	3
例文5	4	2	5	3	1	2.67	3	3
例文6	0	3	2	8	2	3.6	4	4
例文7	0	1	3	6	5	4	4	4
例文8	2	0	2	5	6	3.87	4	5
例文9	1	1	8	2	3	3.33	3	3
例文10	0	0	9	4	2	3.5	3	3
例文1〜10						3.51		

表 11　設問 2 に関する調査結果

Table 11　Survey results for question 2

設問2	分布					平均値	中央値	最頻値
	1	2	3	4	5			
例文1	3	2	5	2	4	3.13	3	3
例文2	2	1	1	4	7	3.87	4	5
例文3	4	2	2	3	4	3.07	3	1･4
例文4	5	2	4	3	1	2.53	3	1
例文5	4	3	4	3	1	2.6	3	1･3
例文6	2	4	4	4	1	2.87	3	2･3･4
例文7	1	1	3	4	6	3.87	4	5
例文8	2	2	1	4	6	3.67	4	5
例文9	3	5	2	3	2	2.73	2	2
例文10	2	0	7	3	3	3.33	3	3
例文1〜10						3.17		

設問 1 に対する平均値は 3.51，設問 2 に対する平均値は 3.17 であった．このことから，どちらかといえば人を傷つけない表現に変換できたことがわかる．特に例文 2・7・8 については人を傷つけない表現に変換できた．逆に，例文 4・5・6 については人を傷つけない表現に，あまりできていない事が分かる．また，例文 3・4・5・9 など人によって感じ方のばらつきが大きい文章もあった．

5.3.2 意味の分かる文章になっているか

表 2 の文章を見て，誹謗中傷を言い換えるシステムが，意味が分かる文章に変換できているかについてアンケート調査を行った．その結果を表 12 に示す．設問番号と質問内容は以下のとおりである．

表 12　設問 3 に関する調査結果

Table 12　Survey results for question 3

設問3	分布					平均値	中央値	最頻値
	1	2	3	4	5			
例文1	3	5	2	2	3	2.8	2	2
例文2	0	1	5	6	3	3.73	4	4
例文3	2	2	2	3	6	3.6	4	5
例文4	4	9	2	0	0	1.89	2	2
例文5	5	6	3	0	1	2.07	2	2
例文6	6	3	6	0	0	2	3	1･3
例文7	0	0	1	7	7	4.4	4	4･5
例文8	1	0	3	5	6	4.07	4	5
例文9	3	5	1	4	2	2.8	2	2
例文10	4	5	3	3	0	2.33	2	2
例文1〜10						2.97		

・設問 3「例文 1〜10 についてそれぞれ，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と思うか.」

設問 3 に対する平均値は 2.97 であった．この結果から誹謗中傷を言い換えるシステムでは，意味が少しわかりにくい文章に変換される事が分かる．特に例文 4・5・6・10 は分かりにくい文章に変換される．逆に例文 2・3・7・8 についてはかなり分かりやすい文章に変換できた．また，例文 1・3・9・10 のように感じ方の個人差がある文章もあった．

5.3.3 おもしろい文章になっているか

表 2 の文章を見て，誹謗中傷を言い換えるシステムが，面白い文章に変換できているかについてアンケート調査を行った．その結果を表 13 に示す．設問番号と質問内容は以下のとおりである．

・設問 4「例文 1〜10 についてそれぞれ，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と思うか.」

表 13　設問 4 に関する調査結果

Table 13　Survey results for question 4

設問4	分布					平均値	中央値	最頻値
	1	2	3	4	5			
例文1	0	2	3	4	6	3.93	4	5
例文2	1	3	3	3	5	3.53	4	5
例文3	0	1	5	3	6	3.93	4	5
例文4	1	3	4	4	3	3.33	3	3･4
例文5	1	2	2	3	7	3.87	4	5
例文6	2	2	6	2	3	3.13	3	3
例文7	0	2	4	1	8	4	5	5
例文8	0	1	2	2	10	4.4	5	5
例文9	1	1	2	6	5	3.87	4	4
例文10	1	0	6	4	4	3.67	4	3
例文1〜10						3.67		

設問 4 に対する平均値は 3.67 であった．例文 1〜10 の各々についていずれも面白可笑しいと感じられる文章に変換できた．

6. 考察

この節では単語感情極性対応表を用いた分析・アンケート評価の結果から，提案システムについて考察する．

6.1.1 言い換えシステムについて

単語感情極性対応表を用いて評価を行った結果，誹謗中傷表現を含む文章の平均感情値の平均値が-0.616126 だったのに対し，変換後の文章では-0.150700 と，＋0.465426 ポイント上昇していることから，ポジティブな文章に変換できたことが示された．

5.3 節のアンケート調査の結果を表 14 にまとめ直す．それぞれの評価項目の平均値について，面白可笑しい文章に変換できたが 3.76，人を傷つけない表現にできた・受け流せるレベルの表現に変換できたが，それぞれ 3.51，3.17，意味の分かる文章に変換できたが 2.97 であった．このことから面白おかしい表現に変換することができたこと，人を傷つけない柔らかな表現に変換できたかについてはややできたこと，意味の分かる文章にはうまく変換できなかったことが示された．また，それぞれの項目についての感じ方に個人差がある文章が多々見受けられた．

例文 1〜10 の文章ごとに比較すると，平均値が 3.5 を超えて評価の高い変換結果となった例文 2・7・8 は，字数が少ない．逆に，平均値が 3 を下回り評価の低い変換結果となった例文 5・6 については字数がやや多い傾向がある．例文 4 は字数が少ないのに評価が低い変換結果となった．

アンケート参加者の感じ方は個人ごとに差が激しく，自由記載欄に回答してくれた意見として「意味の分かる文章に変換されるとさらに良い」という参加者や，逆に「文章がある程度破綻している方がおもしろく，このままでも良い」という参加者もいた．また,「皮肉のような表現に変換されてしまい，かえって受け流しにくい文章になったと感じる」という参加者もいれば,「変換後の文章を見るのが面白いので使っているうちに負の感情が少なくなっていくのかなと思った。」という参加者もいた．

誹謗中傷表現を別の表現に言い換えるシステムは，NG ワードと言い換え表現のリストデータに依存している．今回用意した NG ワードと言い換え表現のリストデータは，私個人が一人で作ったものである．今後たくさんのユーザーが言い換え表現を考えたり，ユーザーが考えた言い換え表現を別のユーザーが評価できるようにしたりするなど，ユーザー参加型のシステムにすることで，誹謗中傷表現を別の表現に言い換えるシステムを高品質なものにできると考えられる．

表 14　言い換えに関するアンケート調査のまとめ

Table 14　Summary of the survey on paraphrasing

評価項目	平均値										
	例文1	例文2	例文3	例文4	例文5	例文6	例文7	例文8	例文9	例文10	例文1〜10
人を傷つけない表現に変換できているか	3.87	4.13	3.27	2.87	2.67	3.6	4	3.87	3.33	3.5	3.51
受け流せるレベルの表現に変換できているか	3.13	3.87	3.07	2.53	2.6	2.87	3.87	3.67	2.73	3.33	3.17
意味の分かる文章に変換できているか	2.8	3.73	3.6	1.87	2.07	2	4.4	4.07	2.8	2.33	2.97
面白可笑しい文章に変換できているか	3.93	3.53	3.93	3.33	3.87	3.13	4	4.4	3.74	3.67	3.76
平均値	**3.4325**	**3.815**	**3.4675**	**2.65**	**2.8025**	**2.9**	**4.0675**	**4.0025**	**3.15**	**3.2075**	**3.3525**

7. おわりに

本章では，本研究のまとめと今後の課題について述べる．

7.1 まとめ

本研究では，ユーザーに誹謗中傷表現とその言い換え表現を登録してもらい，そのデータを元に言い換えを行うシステムの開発と評価を行った．結果として，本研究目的である，誹謗中傷に対するモラルの向上や，誹謗中傷表現を吐き出せる場所を提供することにおいて，一定の効果が期待できることが確認できた．また，本研究の提案システムに対し，「面白い・継続的に使いたい」と感じることも示されており，将来的にユーザー参加型のコンテンツとして，不特定多数のユーザーが使う際には，たくさんのユーザーが，継続的に使用することが期待できる．

7.2 今後の課題

本研究では，誹謗中傷に対するモラルの向上や，誹謗中傷表現を吐き出せる場所を提供することにおいて，一定の効果が期待できることが確認できた．しかし，本システムは，使うユーザー数が多ければ多いほど，質が向上する仕組みになっている．そのため，今後不特定多数のユーザーが提案手法を用いたサービスに参加できる，ユーザー参加型のシステムとして作り上げることが今後の課題である．

具体的には，本研究ではXAMPPを用いることでlocalhost内でのみ使えるサービスを試験的に構築した．今後ドメインを取得し，サーバーを立ち上げ，サーバーにプログラムをアップロードする等の手順を経て，不特定多数のユーザーが使えるシステムにする必要がある．また本システムは，使うユーザー数が多ければ多いほど，質が向上する仕組みになっている．そのため，ログイン機能や，言い換え表現を不特定多数のユーザーが評価できる「いいね機能」を導入し，他のユーザーの評価を考慮して変換できるようにするなど，誹謗中傷表現を含む文章をより上質な言い換え表現で変換して表示するシステムにすることで，多くのユーザーが継続的に使うようなサービスにすべきである．

また，意味の伝わりづらい文章に変換されてしまうことも課題である．今後は単純に置き換えるのではなく，接続助詞や，動詞の活用形などにも合わせた変換を行えるようにする必要がある．

今回の評価に関わったアンケート参加者は，全員 20 代であり，かつ研究室関係者が多かった．今後ユーザー参加型のコンテンツにするに伴い，評価を不特定多数のユーザーで実施する必要がある．

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に関する政パッケージ」，
<https://www.soumu.go.jp/main_sosiki/kenkyu/information_disclosure/02kiban18_02000105.html> (参照 2021-10-24).
[2]nanka：ネガティブバスター，
<https://nankasince2016.jimdofree.com/>(参照 2021-10-24)
[3]GMO インターネット株式会社：SNS PEACE，入手先
<https://sns-peace.com/>(参照 2021-10-24)
[4]TOYOTA MOTOR CORPORATION:しずかったー
[5]アディッシュ株式会社：matte
＜https://matte.ai/>(参照 2021-10-24)
[6]大友 泰賀ほか:いじめ表現辞書を用いた Twitter 上のネットいじめの自動検出，DEIM2020 C7-1(day2 p22)
[7]松本和幸ほか：感性を考慮した日本語俗語の標準変換，
web インテリジェンスとインタラクション 2017，
人工知能学会論文誌 32 巻 1 号 WII-A，pp.1-12(2017 年)
[8]高村大也ほか：スピンモデルによる単語の感情極性抽出，情報処理学会論文誌ジャーナル，Vol.47 No.02 pp627-637（2006）

