様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１４　日現在

機関番号：　１２１０２
研究種目：　若手研究(B)
研究期間：　2015～2017
課題番号：　１５Ｋ２０８８４

研究課題名（和文）言語表現の使用実態を踏まえたソーシャルメディア上の誹謗中傷行為の検出に関する研究

研究課題名（英文）A Study of Detecting "bad mouth" Comments in Social Media based on Actual Usage of Language Expressions

研究代表者

乾　孝司（INUI, Takashi）

筑波大学・システム情報系・准教授

研究者番号：６０３９７０３１

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：本研究課題では，「いじめ」につながる誹謗中傷行為について，ソーシャルメディア上での言語表現レベルでの詳細な実態調査をおこない，その結果を踏まえた言語解析技術に基づく誹謗中傷行為の高精度な自動検出技術の開発を目的とする．主な研究成果として以下が挙げられる：（１）およそ5,000件の投稿データを対象にした言語表現の使用実態調査をおこない，誹謗中傷表現および被害者／加害者表現の出現傾向をある程度明らかにした．（２）上記傾向を考慮したSVMベースの検出モデルを提案し，その有効性を示した．（３）被害者情報の自動取得では知識ベース素性が大きく貢献することを明らかにした．

研究成果の概要（英文）：In this research project, we conduct a detailed survey of language expressions of libelous acts that lead to "bullying" on social media, and developed an automatic detection method of slander act based on the survey results. The main research results are as follows: (1) Survey on actual usage of language expression targeting about 5,000 text data, revealed appearance trend of slander expression and victim / perpetrator expression. (2) SVM based detection model considering the above tendency was proposed and its effectiveness was shown. (3) Knowledge-based features greatly contribute to automatic acquisition of victim information.

研究分野：　自然言語処理

キーワード：　自然言語処理　有害情報　ソーシャルメディア

１．研究開始当初の背景

近年のインターネット関連技術の発達は目覚ましく，老若男女を問わず誰しもが容易にインターネットにアクセスでき，社会生活を豊かにする多くの情報・サービスを享受できる環境が整ってきた．このことは青少年にも該当し，青少年にとってインターネットは以前にも増して身近な存在となっている．

しかし，インターネットには負の側面もあり，青少年の健全な成長・育成にとって有害となる情報に青少年が接する機会も残念ながら増えている．また，配慮すべき有害情報の種類も広がりを見せており，2000 年代後半から急速に普及が進んだソーシャルメディアを利用することで，知らない大人との不適切な出会いや，悪口，からかいを含む誹謗中傷行為といった，ユーザ間のコミュニケーションに基づく新しいタイプの有害情報も増長している．

このような社会背景に対して，インターネット上の有害情報から青少年を守るための対策が続けられており，近年では機械学習アルゴリズムを用いることによって，NG ワードに基づく手法よりも高精度なフィルタリングの実現を目的とした研究開発も進んでいる．しかし，現状のところ，実用化の目処が立つ程の精度が得られているとは言い難い．

２．研究の目的

本研究課題では，インターネット上の投稿テキストに含まれる，ある特定の有害情報カテゴリに特化して言語表現の使用実態の十全な調査・分析をおこない，その結果に基いて，ある特定の有害情報カテゴリに特化した高精度な有害情報検出モデルを開発することを目的とする．

(1) 特定の有害情報として，青少年の間でのいじめと密接に関連している「誹謗中傷」発言の自動検出を目指す．各種の統計調査によって，中学生・高校生の間でのネットいじめ発生件数は増長していることが報告されており，本研究課題を遂行することで，いじめ問題の解消につながる技術的な足がかりの構築を目指す．

(2) インターネット上の投稿テキストデータを対象として，どのような表現で誹謗中傷行為がなされているかを把握するため，誹謗中傷行為に使用される言語表現の実態調査をおこなう．また，誹謗中傷行為がその他の有害情報と比較して持ちやすい言語的な特徴として，「○○は×××だ」の○○ように，被害者／加害者を示す言語表現が誹謗中傷表現に近接して表記されやすい．そこで，被害者／加害者および誹謗中傷の対要素を実態調査における中心的な調査項目とする．

(3) 上述の言語使用の実態調査結果を踏まえ，言語解析技術に基づく誹謗中傷行為の高精度な検出モデルを開発する．検出モデルの開発時に考慮すべき中心的な特徴量は実態調査から得られる言語表現となるため，申請者の主研究領域である言語解析技術を検出モデル開発における技術的な核と捉えている．

３．研究の方法

本研究では，まず，インターネット上の投稿テキストデータを対象として，どのような表現で誹謗中傷行為がなされているかを把握するため，誹謗中傷行為に使用される言語表現の実態調査をおこなう．その後，調査結果から得られる特徴に基づいて誹謗中傷検出モデルを構築，評価する．

誹謗中傷行為に使用される言語表現の実態調査は，調査対象となるテキストデータに対する言語表現アノテーションを実施し，アノテーション結果を集計することで実現する．誹謗中傷行為の定義付けにあたっては，インターネット・コンテンツに対する既存のレーティング基準（SafetyOnline3）等を参考にしつつ，言語テストと外延を併用した定義策定をおこなう．

検出モデル開発の方針として，基本となる機械学習アルゴリズムは先行研究に従って Support Vector Machine(SVM)を採用することとし，主に学習時に利用する特徴量の設計および特徴量の抽出手法を重点的に検討する．特徴量として，実態調査の中心に据えた誹謗中傷表現および被害者／加害者表現に焦点をあて，各表現およびその組合せに関して有用性を検証する．

４．研究成果

およそ 5,000 投稿のテキストデータを対象にした言語表現の使用実態調査では，誹謗中傷表現は動態述語でなく状態述語としてあらわれ，特に形容詞，形容動詞が多いこと，被害者表現は名詞として出現する傾向が強く，特に一般名詞，固有名詞，代名詞が多いことを確認した．加害者については投稿 ID のような SNS 環境に特化したものを除けば投稿内であらわれることは稀であることを確認した．また，被害者表現がもつ格関係情報については，被害者表現は後続する助詞が省略されやすいこと，省略されない場合は助詞「は」およびそれの話し言葉調で崩れた形である「って」となる傾向があることを確認した．

誹謗中傷が書かれていると判定できる投稿の一部を用い，誹謗中傷の直接的な表現の有無について調査したところ，およそ６割の投稿には直接的な表現が含まれているが，残り４割には直接的な表現が含まれない事を確認

した（表１）.

表１

| 悪口表現の内訳 | 投稿数 | 割合 (%) |
|---|---|---|
| 悪口表現が明示的に存在する | 334 | 57.19 |
| 悪口表現が明示的に存在しない | 250 | 42.81 |
| 合計 | 584 | 100 |

表現位置に関して，誹謗中傷検出モデルの構築にとって有効と考えられる言語表現は投稿内の限られた一部の箇所にのみ限定的に出現する傾向があることがわかった．誹謗中傷表現と被害者表現の相対的な位置関係について，被害者表現でない名詞とのあいだの平均単語数は 13.2 であるが，被害者表現とのあいだの平均単語数は 6.54 であり，誹謗中傷表現と被害者表現は投稿内で相対的に近い位置に現れることを確認した．

上記の調査結果を踏まえ，SVM に基づく誹謗中傷検出モデルのプロトタイプを開発した．プロトタイプモデルでは，特に被害者情報を特別に考慮するために被害者の種別に応じて複数ある検出モデルから適切なモデルが自動選択される機構を実装した（図１）．評価実験の結果，被害者の種別に注目したモデル選択機構が検出性能の向上（F 値尺度で 0.52）に貢献することを確認した．

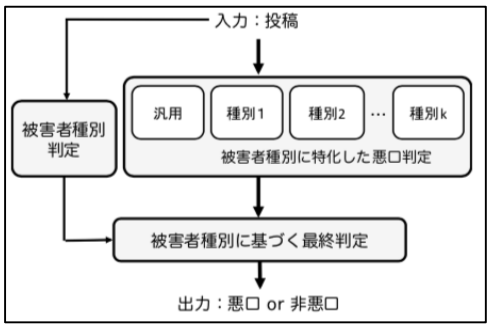


図１

続いて，検出モデルの構築時に利用する教師事例の不足を補うための仮想教師事例自動生成手法を開発した．自動生成の既存研究では仮想教師事例を生成する際に文章中の一部分を編集する操作をおこなうが，実態調査の結果で述べた知見を考慮しないため誹謗中傷行為に直接関与する単語も編集してしまう致命的な問題があった．この問題に対して，事例プール法および単語に関する出現分布法という 2 つの手法を提案し，モデルを改良した．評価実験を通した検証した結果，多くの実験設定において提案手法が有効であることを確認した．

さらに，誹謗中傷行為を受けている被害者情報を自動取得するため，投稿内から人物（および組織）表現を自動抽出する方法に対する現状の問題点を整理した．固有表現の要素となる割合および固有表現の要素となる場合のその固有表現の異なり度合いに基づいて言語表現毎に抽出難易度を測定した結果，抽出難易度のレベルと抽出性能（F 値）に相関関係が認められること（表２），抽出難易度の高い事例は表現自体が短く，その原因として省略表現が多用されていること等を明らかにした．

表２

| レベル | PER | ORG | GPE | GOE | P_VOC |
|---|---|---|---|---|---|
| 1 | 0.945 | 0.898 | 0.900 | 0.867 | 0.895 |
| $2_{NR>MR}$ | 0.786 | 0.741 | 0.885 | 0.450 | 0.666 |
| $2_{NR<MR}$ | 0.846 | 0.705 | 0.760 | 0.585 | 0.768 |
| 3 | 0.667 | 0.602 | 0.780 | 0.601 | 0.464 |

被害者情報を自動取得する手法として，投稿テキストから人物をあらわす表現を自動認識する手法および，人物表現と実世界の人物エンティティとの対応付け手法を開発した．投稿テキスト中の人物表現を自動認識する手法については，条件付確率場に基づく既存手法に対して，新たに知識ベースに基づく素性および文節情報に基づく素性を取り込む手法を提案し，人物表現認識性能の向上を達成した．難易度指標を用いた分析の結果，知識ベースに基づく素性はすべての難易度の事例群に対して有効であることがわかった．文節情報に基づく素性では，難易度の高い事例群に対して文節主辞の情報に加えて格情報を考慮することが重要であることがわかった（表３）．

表３

| 追加素性 | 1 | $2_{NR>MR}$ | $2_{NR<MR}$ | 3 |
|---|---|---|---|---|
| （ベースライン） | 0.918 | 0.654 | 0.739 | 0.626 |
| Gazetteer 素性 | 0.926 | 0.659 | 0.771 | 0.651 |
| ＋所属文節主辞素性 | **0.927** | 0.668 | 0.776 | 0.667 |
| ＋係り先文節主辞素性 | 0.919 | 0.656 | 0.760 | 0.641 |
| ＋所属文節格助詞素性 | 0.926 | 0.659 | 0.773 | 0.654 |
| －所属文節主辞素性 | 0.919 | 0.656 | 0.760 | 0.639 |
| －係り先文節主辞素性 | 0.926 | 0.668 | **0.777** | **0.670** |
| －所属文節格助詞素性 | 0.923 | **0.671** | 0.768 | 0.660 |
| 全素性 | 0.922 | **0.671** | 0.767 | 0.661 |

人物表現と実世界の人物との対応付け手法について，固有表現クラス情報に基づくフィルタリングを導入することで不要なエンティティ候補を削減し（図２），これによって人物表現に関する実世界の人物エンティティとの対応付け正解率が向上することを確認した．

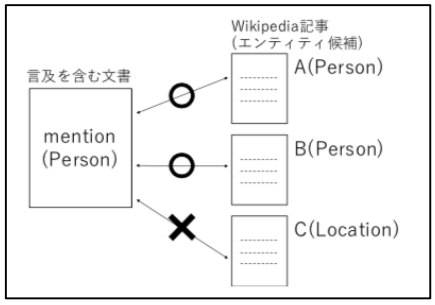


図２

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計６件）

① 三好聖子、仲野友規、乾孝司、Wikification における固有表現クラス情報を用いた候補削減、言語処理学会第24回年次大会、pp.1100-1103、2018
② 仲野友規、乾孝司、人物・組織エンティティに特化した固有表現抽出器の開発、言語処理学会第24回年次大会、pp.925-928、2018
③ 仲野友規、乾孝司、人物・組織エンティティに対する固有表現抽出課題の難易度評価、言語処理学会第23回年次大会、pp.246-249、2017
④ 河原裕樹、乾孝司、悪口投稿検出におけるソーシャルテキストの特性を考慮した仮想教師事例作成、電子情報通信学会思考と言語研究会(TL2017-3)、pp.13-18、2017
⑤ 仲野友規、河原裕樹、乾孝司、SVMを用いた誹謗中傷・悪口投稿からの被害者表現の自動抽出、電子情報通信学会思考と言語研究会（TL2016-6）pp.23-28、2016
⑥ 河原裕樹、乾孝司、山本幹雄、被害者種別を考慮したソーシャルテキストからの悪口検出、電子情報通信学会思考と言語研究会（TL2015-12)、pp.67-72、2015

６．研究組織
(1)研究代表者
乾　孝司（INUI, Takashi）
筑波大学・システム情報系・准教授
研究者番号：６０３９７０３１