学校非公式サイトにおける有害情報検出を目的とした極性判定モデルに関する研究

松葉　達明†　桝井　文人‡　河合　敦夫†　井須　尚紀†

†三重大学大学院工学研究科　‡北見工業大学情報システム工学科

{matsuba,kawai,isu}@ai.info.mie-u.ac.jp,f-masui@mail.kitami-it.ac.jp

1　はじめに

「ネット上のいじめ」が新しい「いじめ」の形態として問題となっている.「ネット上のいじめ」とは，携帯電話やパソコンを通じてインターネット上のいわゆる学校非公式サイトの掲示板などにおいて，特定の子どもの悪口や誹謗・中傷を書込んだり，メールを送信するなどして，有害情報によるいじめを行うもの[1]である.

学校非公式サイトの掲示板は,複数のユーザが相互に発言を行い，情報交換をする場である．このようなサイトでは，議論の食い違いや学校での諍いなどが発端となり，他のユーザが不快と感じる発言や特定個人を誹謗中傷する発言が書込まれるケースも頻繁に発生する[2].

これらの有害情報は，ネットパトロールにより監視されている．ネットパトロールとは，文字通り学校非公式サイトの掲示板などを人手でつぶさにチェックを行う監視作業である．ネットパトロールによって有害であると判断された書込みは，該当掲示板の管理人あるいはプロバイダに削除依頼がなされる.

しかしながら，現状では，ネットパトロールにおける書込みの確認作業が最も負担が大きく，増大し続ける学校非公式サイトを監視するのは困難となる．そこで，本研究では学校非公式サイトの掲示板に書込まれる有害情報を検出するシステム構築を目指す．これにより，ネットパトロール活動の一部を自動化し，担当者の負担を軽減することができ,有害情報の早期発見・早期対応の支援にもつながる.

そこで本論文では，有害情報を「掲示板やブログなど自由に書込み可能なネット上のサイトにおいて，個人情報の流出や誹謗中傷などを通して，特定個人の実生活に悪影響を及ぼす可能性を内在する情報」（図 1）[3]と定義し，これらを判別する手法を提案する．提案手法は，Turney による PMI-IR 手法を拡張した手法[6]であり，有害語とそれらとの係り受け関係を利用して判別を行う.

・商業の１年のガキ！！しんでほしい

・abcdef@docomo.ne.jp マツバのメールアドレス

・名出し禁止，松葉達明は良い奴だよ

図 1 有害情報を含む書込み例

以下，2 章で関連研究について述べ，3 章で分類実験前の準備，4 章で提案手法，5 章で分類実験と考察，6 章でまとめと今後の課題について述べる.

2　関連研究

有害情報の抽出に関する先行研究としては，石坂らの研究[4]と池田らの研究[5]があげられる.

石坂らの研究では，巨大電子掲示板「2 ちゃんねる」[1]を対象とし，悪口表現辞書を構築している．石坂らは，悪口表現を「バカ」や「マスゴミのクズ」などの特定他者に対して直接侮辱や誹謗中傷している単語，句と定義している．そして，悪口表現を n-gram 確率を使用し，周辺単語列から抽出することを試みている．しかし，悪口表現にのみ連接しやすい単語列は少なく，悪口表現自体も定型的に存在するわけではないことを報告している.

池田らの研究では，人手により有害と無害に分けられた学習用文書を用いて，形態素の出現頻度の偏りによる有害判定キーワードリストを構築している．さらに，形態素の係り受け関係を用いて，有害無害の判定性能を向上させた．しかし，ウェブ文書では「爆破」と「爆ー破」のように少しだけ文字を変えた表現も多く，日々増え続ける新しい表現に，次第に対応しきれなくなる問題がある.

本研究では，周辺単語列を考慮せず有害情報候補単語列を有害無害に判定する．また，ウェブ検索ヒット数を判定基準に利用するため，学習用文書を構築する必要も無い.

3　準備

まず，有害情報と無害情報，それぞれの書込みを構成する形態素を主な分析対象として言語表現の分析をした．掲示板の書込みに対し形態素解析，係り受け解析を行うため，形態素解析に Chasen(ver2.3.3,日本語辞書 ipadic-2.7.0)，係り受け解析に Cabocha(ver0.53)を利用した．これらの解析器は，既存の新聞記事など向けに作成された解析器であり，方言や若者言葉，表記のゆれなどの様々な表現方法が含まれる掲示板の書込みに対しては，解析精度が低くなってしまう．そこで，一定の解析精度を保つために有害単語の辞書登録と書込みデータの修正を人手で行った．以下，施した処理について詳述する.

3.1　有害単語の辞書登録

有害情報を構成する一つの要素として「キモい」や「チャラ男」など特有の単語(有害単語)が含まれる.有害単語は，既存の解析器に用意された日本語辞書には含まれていないため，未知語判定，もしくは解析ミスをしてしまう可能性が高い．そこで，有害単語を人手で収集して整理し，あらかじめ形態素辞書に登録した．有害単語であるかどうかの判断は，文部科学省の「サイト・スレッドの書込み類型化」[1]に基づいて行った．収集した書込みデータは，実際にネ

[1] (http://www.2ch.net/)

ットパトロールによって収集された掲示板への書込みデータと，筆者らが独自に収集した書込みデータ(三重県域に限定されたサイトから収集したもの)2998 件である．この書込みデータから「ｷｯｼｮｲ」や「うざっこい」など，239 単語を収集した．

3.2 掲示板書込みの標準化

電子掲示板は，誰でも手軽に書込むことができ，閲覧する対象者も不特定多数である．そのため，伏字や方言，若者言葉などを含み，正しい文章や標準語で書かない,くだけた書込みも多い．それらの書込み例を図 2 に示す．

- あいつはヤ○マン.
- 今も見とるんやろな．でてこやんかな.
- サンクス．ちょっといってみるわ.

図 2 くだけた書込み例

このような書込みは，形態素解析器で解析ミスをしてしまう可能性が高い．そこで，用意した掲示板の書込みデータ 2998 件中 1430 件を人手で標準語に修正した．

3.3 言語表現の分析

有害情報と無害情報における言語表現の差異を調べるために，形態素の出現頻度を用いて分析した．まず，修正した有害情報書込み 1508 件，無害情報書込み 1490 件を形態素解析した．そして，有害と無害間で重複している形態素を除去し，品詞別[2]に出現頻度順ランキングを作成した．その結果，名詞，動詞，形容詞で有害と無害の上位を占める形態素には，品詞により出現傾向の違いが見られた．名詞では個人名や「バカ」などの誹謗中傷語や卑猥語が目立ち，動詞では「死ね」などの暴力誘発語，形容詞では「キモイ」などの誹謗中傷語が支配的であった（図 3）．

この結果は，上述した文部科学省の「サイト・スレッドの書込み類型化」にほぼ対応していることが分かった．

さらに，有害情報の係り受け関係を調べたところ，以下のように特定の要素が組み合わされるという条件によって有害化する傾向が見られた．

- 係り受け関係にある名詞と名詞の組
  (例)ゴリラ風の顔 → **「対象をある名詞で例える」**
  (例)臆病者の松葉 → **「対象の修飾」**
- 係り受け関係にある名詞と動詞の組
  (例)松葉を殺す → **「対象に行動する」**
- 係り受け関係にある名詞と形容詞の組
  (例)性格が悪い → **「対象を表現する」**

例えば，「性格が悪い」や「胸がでかい」などの有害表現は，「性格–悪い」「胸–でかい」という係り受けで構成されている．しかし，その構成要素である「性格」や「胸」，「悪い」，「でかい」のみが単独で出現したとしても有害性を持たず，これらの要素が係り受け関係を持って共起することによって，はじめて有害性を持つのである．

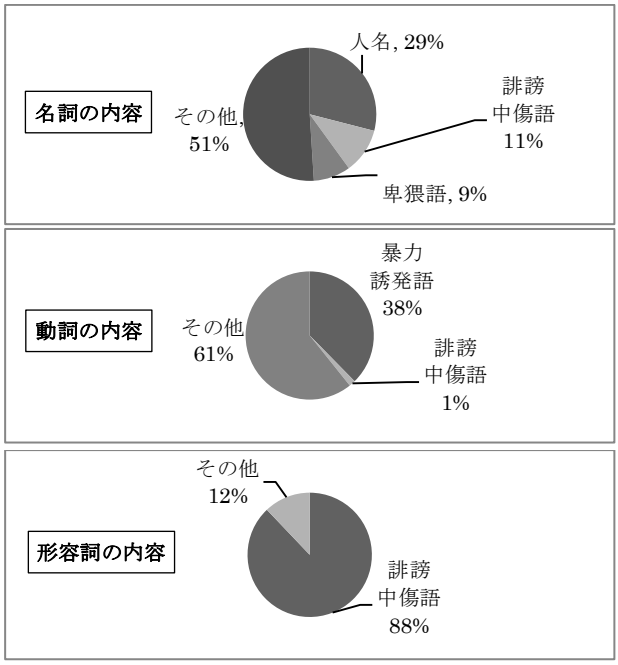

図 3 有害情報における品詞別の内容

よって，このような係り受け関係を持つ要素の組を有害性判定の素性として用いることは，有害表現の判別に大きく寄与すると考えられる．

4 提案手法

提案手法について述べる．提案手法は，（1）有害情報候補単語列の抽出，（2）極性単語の参照，（3）拡張 PMI-IR による有害表現判別，という三つのステップで処理を実施する（図 4）．以下，各ステップについて説明する．

4.1 有害情報候補単語列(phrase)の抽出

有害情報書込みにも，無害な表現は含まれ，有害無害表現が混在している．よって，判別要素に無害な表現が含まれることを最小限にするため，有害情報候補となる単語列（図 4 では”phrase”と表記）を抽出する．

まず，入力として与えられた書込み文を係り受け解析する．解析の結果から，「名詞–名詞」，「名詞–動詞」，「名詞–形容詞」のいずれかの係り受け関係を持つ形態素の組を，有害情報候補の単語列として抽出，保持する．

4.2 極性単語

Turney[6]の研究では，ポジティブとネガティブの 2 極性を持つそれぞれの単語との関連度を算出している．有害性の判定においても同様に，有害と無害の 2 極性になる．

有害な極性単語は，疑わしいものは全てチェックという観点から有害単語を極性単語とすればよい．しかし，無害な極性単語は「面白い」という形態素を例に挙げると，「面

[2] 連体詞，接頭詞，名詞，動詞，形容詞，副詞，接続詞，助詞，助動詞，感動詞

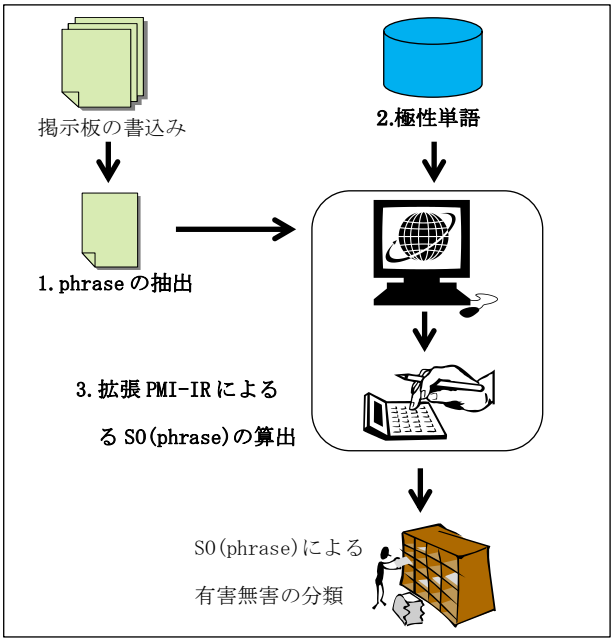

図 4 提案手法の処理の流れ

白い髪形」と「面白いゲーム」のように，明確に無害だと判別可能な極性単語はほとんど無いと考えられる．

そこで，極性単語は有害な極性単語のみとした．極性単語には，3.3 節で作成した有害情報の形態素出現頻度順ランキングにおける，卑猥語，暴力誘発語，誹謗中傷語に該当する上位 3 件ずつとした．有害極性単語を以下に示す．

・卑猥語　　:セッ○ス，ヤ○マン，フ○ラ[3]
・暴力誘発語　　:死ね，殺す，殴る
・誹謗中傷語　　:きもい，うざい，不細工

4.3 拡張 PMI-IR による SO(phrase)の算出

PMI-IR とは，ウェブ検索ヒット件数を利用した共起度判定手法である．Turney の研究では，レビューサイトのコメントを”excellent”(ポジティブ)か”poor”(ネガティブ)の 2 極に自動分類する研究を行った．PMI-IR によって，コメントと”excellent”,”poor”との関連度を算出し，どちらに強く関連しているかによって，分類を行っている．

我々は，PMI-IR 手法を phrase と選択した極性単語との関連度の強さを算出するように拡張した．以下，拡張 PMI-IR 手法について述べる．

まず，PMI(1)は二つの word の関連度を示す．

$$PMI(word_1, word_2) = log_2\left\{\frac{p(word_1 \ \& \ word_2)}{p(word_1)p(word_2)}\right\} \quad (1)$$

そして，PMI-IR(2)は IR(ウェブ検索ヒット件数)によって与えられる．

$$PMI\text{-}IR(word_1, word_2) = log_2\left\{\frac{hits(word_1 \ \ AND \ \ word_2)}{hits(word_1)hits(word_2)}\right\} \quad (2)$$

hits(word$_1$)は，word$_1$ を検索単語にしたときのウェブ検索ヒット件数である．AND はウェブ検索エンジンの AND 演算子を指し，ウェブ検索エンジンには yahoo!JAPAN[4]を利用した．そして，意味の方向性 : SO(phrase)を式(3)のように算出する．

$$SO(phrase) = PMI\text{-}IR\{"phrase","セッ○ス"\} + \cdots + PMI\text{-}IR\{"phrase","不細工"\} \quad (3)$$

つまり，SO(phrase)は極性単語とどれだけ関連度が強いかを示している．また，一つの書込みには複数の phrase が存在する場合もあるので複数の SO(phrase)が得られる．疑わしいものは全てチェック対象という観点から，SO(phrase)が最大のものをその書込みの SO(phrase)とした．

5 評価と考察

提案した手法を用いて有害情報と無害情報を入力し，全ての書込みの SO(phrase)を算出して，SO(phrase)による有害無害の分類を行った．全ての書込みを SO(phrase)順にランキングし，上位 n 件以上を有害情報と判定する，閾値 n を設定した．評価方法は適合率(4)と再現率(5)で行った．

$$適合率 = \frac{閾値以上の有害情報書込み件数}{閾値以上の書込み件数} \quad (4)$$

$$再現率 = \frac{閾値以上の有害情報書込み件数}{全ての有害情報書込み件数} \quad (5)$$

5.1 比較実験の設定

比較実験として，4 種類の対象実験を用意した．以下の各対象実験について述べる．

1. 前方参照手法　「する」や「の」などのそれ単体では意味が不十分な形態素の場合，形態素の 2 個前まで取得する(例:する→抗議をする)．

2. 品詞別手法　3.3 節の分析結果を利用し，「名詞と名詞」の phrase には卑猥語と誹謗中傷語の極性単語を，「名詞と動詞」の phrase には暴力誘発語の極性単語，「名詞と形容詞」の phrase には誹謗中傷語の極性単語の組み合わせで SO(phrase)を算出する．

3. 有害単語辞書手法　有害単語辞書は，3.1 節で構築した有害単語 239 語で，phrase にそれらが含まれる場合に重み付けを行う(6)．

4. 単語感情極性辞書手法　単語感情極性辞書は，高村らが構築した辞書である[7]．単語には-1～+1 の値を付与されており，-1 に近いほどネガティブ，+1 に近いほどポジティブとしている．重み付けは,式(6)のように行う．

$$SO(phrase) = log_2\left\{\frac{hits(phrase \ AND \ 死ね)}{hits(phrase)hits(死ね)} \times \alpha\right\} \quad (6)$$

$$\alpha = 1 + (有害辞書単語を含むなら + 1 \quad or \quad 感情極性辞書で度合が - 0.99 以下の単語を含むなら + 1)$$

実験データセットとして 3.2 節で述べた修正後のデータ 2998 件(有害情報書込み 1508 件，無害情報書込み 1490 件)を用いた．また，上述した 4 種類の設定を加えていない実験を実験 1 とし，設定を加えた実験を実験 2 とする．

[3] 実際の卑猥語には伏字は無い

[4] (http://www.yahoo.co.jp/)

5.2 実験結果

phrase を抽出した結果, 有害情報は 1508 件中 1034 件, 無害情報は 1490 件中 1022 件抽出できた．計 2056 件を対象に実験 1，実験 2 を行ったところ，適合率が図 5，再現率が図 6 となった．

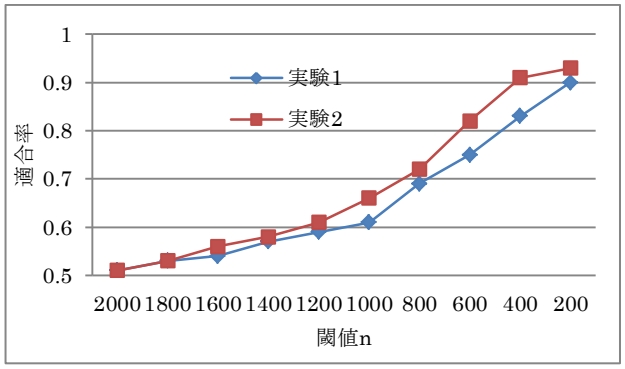

図 5 適合率

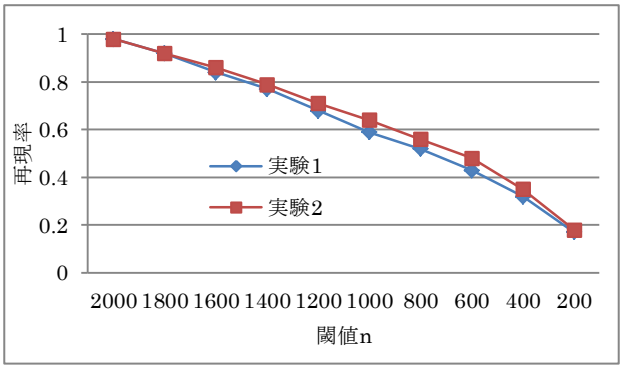

図 6 再現率

5.3 考察

SO(phrase)の高い phrase を調べると，有害情報では「きもい－顔」や「馬鹿－男」などの有害単語が含まれる phrase が多く占め，提案手法が上手く有害情報を抽出できていることが分かる．さらに，「眉毛－濃い」や「乳－でかい」などの単体では無害な phrase も上位に位置していた．逆に，SO(phrase)の低い phrase は「名張－松葉」や「三重－2」などの，悪口を含まない個人情報の流布や，意味の取れない phrase などが占めていた．無害情報では，SO(phrase)の高い phrase は「あそこ－美味い」や「のど－痛い」などが占めていた．これは，「あそこ」は指示代名詞であり，卑猥語と共起が高くなったり，「痛い」は暴力誘発語と共起が高くなってしまったからである．SO(phrase)の低い phrase は，「ジンギスカン－食べ」や「的－考え」などの無害なものや，意味の取れないものであった．

比較実験に設定した実験 2 では，どの閾値においても実験 1 より分類性能が良かった．そこで，実験 2 で追加した設定を一つ一つ実験しなおし，それぞれの設定がどのような効果をもたらしたのか分析を行った．

まず，前方参照手法において，有害情報では「先生－し」という組が「先生－ワイセツ行為し」のように上手く取れるようになっていた．しかし，無害情報では，「今月－さ」というのが「今月－首宣告さ」のように極性単語と共起の強い形態素が現れ，SO(phrase)を引き上げてしまっていた．この結果から，極性単語に選んだ単語がまだ，有害無害どちらにも関連の強い語を選んでいる可能性がある．

品詞別手法は，あまり効果がなかった．このことは，有害情報の類型は有害表現を規定する有効な軸とはなり得ないことを示唆している．

単語感情極性辞書による重み付けは，品詞別手法と同じく効果がなかった．これは，ポジティブ／ネガティブの極性が必ずしも有害性の判定には寄与しないことを示している．例えば，「事故」という言葉はネガティブな言葉だが，有害な方向にのみ強く関連する単語ではない．このことを証明するように，有害単語辞書による重み付けは大きな効果があった．

6 おわりに

「ネット上のいじめ」における有害情報の分析を行い，PMI-IR を応用して有害無害の分類を試みた．まず，形態素の出現頻度順ランキングを作成し分析を行った．結果，有害情報の名詞には人名や誹謗中傷，卑猥な語が見られ，動詞には「死ね」などの暴力を誘発する語，形容詞には「キモイ」などの誹謗中傷語が頻出することが分かった．また，係り受け解析による分析も行ったところ，「名詞と名詞」，「名詞と動詞」，「名詞と形容詞」の組において有害情報の出現パターンがあることが分かった．そして，PMI-IR による有害無害の分類を試みた．分類実験の結果，閾値が高いと有害情報は分別可能だが，閾値が低いところでは人名や，意味の取れない phrase が出現し分類困難となる．比較実験の結果からは，極性単語の選択も再考慮の必要性があり，一般的なポジティブネガティブは有害無害と性質が違うことが分かった．

今後の課題として，適切な極性単語の選択，適切な phrase の抽出パターンを考慮する必要がある．

参考文献

[1] 文部科学省.「ネット上のいじめ」に関する対応マニュアル事例集(学校・教員向け).文部科学省,2008.

[2] 渡辺凡,砂山渡:電子掲示板におけるユーザの性質の評価.電子情報通信学会技術研究報告,No.652 in 2006-KBSE,pp.25-30,2006.

[3] 松葉達明, 桝井文人ら: 学校非公式サイトにおける有害情報検出,言語処理学会第 16 回年次大会, pp.383-386 (2010.3)

[4] 石坂 達也,山本 和英: 2 ちゃんねるを対象とした悪口表現の抽出,言語処理学会第 16 回年次大会, pp.178-181 (2010.3)

[5] 池田和史, 柳原正ら: 格要素の抽象化に基づく違法・有害文書検出手法の提案と評価, 情報処理学会第 72 回全国大会,pp.71-72(2010.3)

[6] Peter D. Turney : Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, pp.417-424(2002,7)

[7] 高村大也, 乾孝司, 奥村学"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 pp. 627--637, 2006.