Web上の誹謗中傷を表す文の自動検出

石坂 達也　　山本 和英
長岡技術科学大学 電気系
{ishisaka, yamamoto}@jnlp.org

1　はじめに

近年，Webサービスの充実化により誰もが容易に情報を発信できるようになった．それに伴い，Web上では他者を誹謗中傷する書き込みも増加している．さらに，小中学生のような低年齢層が誹謗中傷することも増えており，被害者が自殺する事例もある．この問題はこれからインターネットサービスが成長するうえで解決しなければいけない大きな問題である．

現状，企業や自治体による人手の監視が主な解決策とされている．監視するWebサイトを限定しても毎日更新される大量の文があり，これらを一つ一つ読んで有害かどうかを判断するのは時間的，作業量的に監視者の負担が大きい．この作業を効率化し負担を軽減させるための方法としては半自動化がある．そこで，本稿では誹謗中傷を表す自動検出手法を提案する．

本手法では，誹謗中傷に使用される単語であるか否かを示す悪口度を単語に与える．この悪口度をもとに文が誹謗中傷を表しているか否かをSVMを使って2値分類する．

1.1　誹謗中傷の定義

本稿では誹謗中傷を「悪口」として次のように定義する．本稿における「悪口」とは，特定の他者に対して侮辱や批判する表現とする．そして，悪口単語とは単語単独で他者への批判・中傷ができる単語とし，悪口となる単語や句を含み，文として悪口を表現している文を悪口文とする．本研究では皮肉のような文脈や他の情報を必要とする中傷は対象としない．

1.2　本研究で扱うデータ

本研究では巨大電子掲示板サイト“2ちゃんねる(1)”の書き込みを入力文や学習データに使用する．

誹謗中傷が行われる現場は主に電子掲示板であり，2ちゃんねるは国内最大級の電子掲示板である．そして多くの書き込みが存在し，頻繁に誹謗中傷が行われることで社会的にも認知されている．また，2ちゃんねるには2ちゃんねる特有の表現から一般的な表現まで多くの言語表現を含んでいるため，2ちゃんねるの誹謗中傷の言語表現に対応できればWeb上の多くの誹謗中傷に対応できると予想した．

2　関連研究

Web上の有害情報の検出に関する研究はいくつか存在する．松葉ら[3]は学校非公式サイトからの有害情報の検出を行っている．検出のために，学校非公式サイトの文字列と品詞を素性としてSVMで有害情報か否かを分類している．

池田ら[2]は違法・有害文書の文書検出手法を提案している．有害文書から係り受け関係にある文節の組を抽出し，その文節組が違法・有害文書に偏って出現する度合いを算出し，学習している．

また，Hidalgo et al.[1]はEU委員会のプロジェクトの1つとして，Webページがわいせつ文書かどうかを判定するための分類器を作成した．この手法は手がかり語と周辺単語に着目し文書を分類している．

これらのどの手法も有害情報と保証された文書集合を利用している．我々はWeb全体を1つの大規模なコーパスとしているため，提案手法は文書集合の事前準備が不要であり，再現が容易という利点をもつ．

単語が悪口単語かどうかを判別する方法について現在ほとんど議論されていない．しかし，特定の分野の専門用語を抽出する研究やある単語の関連用語を収集する研究は盛んに行われている．その中でも評判表現の評価極性を認識する研究が本研究に最も類似している．悪口は人に対しての不評表現だと考えることもできるからである．

評価極性を認識する研究には，Turney and Littman[4] の研究がある．“excellent” と “poor” を肯定/否定極性を表す代表的な語 (基本単語) として，ある単語がどちらと偏って共起しているか求めることで評価極性を判定している．Wang and Araki[5] はこの手法を日本語用に改良した．基本単語を “素晴らしい” と “不良” とした場合に，“不良” のほうが圧倒的に検索ヒット数が少ないため，正確な評価極性を算出できないとして，Web 検索ヒット数の比率を考慮する要素を追加して精度を向上させた．

我々は Wang and Araki の手法をもとに悪口度を算出している．

3 提案手法

本手法は，単語悪口度の算出と悪口文/非悪口文の文分類の 2 つにより構成される．

3.1 単語悪口度を算出

誹謗中傷が問題として取り上げられるのは悪口単語が入っている文がほとんどである．そのため，悪口単語を認識できれば多くの誹謗中傷を検出できると期待する．

本節では単語に悪口文に使用される単語かどうかの可能性を示す悪口度を付与する方法について述べる．

我々は Wang and Araki が提案した SO-PMI (Semantic Orientation Using Pointwise Mutual Information) を使用し，この値を悪口度とする．悪口度は高いほど悪口単語である可能性が高いという意味のみを持ち，悪意の強さは意味しない．

SO-PMI は事前に 2 つの基本単語を用意し，対象の単語がその 2 つのどちらと文書内共起しやすいかを計る．この手法を使用する理由は 2 つある．1 つ目は悪口単語同士は同一 Web ページ内において共起しやすいという性質を持っているからである．電子掲示板では悪口が書き込まれやすい性質のスレッド (題目) が存在し，そのページは多くの悪口単語を含んでいる．2 つ目は悪口文と非悪口文に分類された集合を必要としない点である．手法によってはあらかじめ悪口/非悪口に分類された大規模な集合を用意しなければならない．しかし，SO-PMI は Web 検索ヒット件数を使用するため，基本単語を 2 つを用意だけで実装できる．

Wang and Araki が使用した SO-PMI の式を以下に示す．

$$C(w) = \log \frac{hit(w, w_p) * hit(w_n)}{hit(w, w_n) * hit(w_p)} \quad (1)$$

$$f(a) = a * \log \frac{hit(w_p)}{hit(w_n)} \quad (2)$$

$$SO\text{-}PMI(w) = C(w) + f(a) \quad (3)$$

この式における w_p と w_n は極性を示す代表的な単語 (基本単語) である．評判分析をした Wang and Araki は w_p を「素晴らしい」「良い」など，w_n を「不良」「悪い」など評価極性が逆となる 2 つの単語を基本単語としている．

hit 関数は Web 検索ヒット件数を求める関数である．例えば，$hit(w)$ は w をクエリとした時の Web 検索ヒット件数である．C 関数では w が w_p と w_n のどちらと共起しやすいかを求める．f 関数は w_p と w_n の検索ヒット数の差による優位性を解消するための関数であり，a は f 関数の重みを設定するための定数である．本研究では Wang and Araki が示す結果を参考に a を 0.9 とした．SO-PMI は C 関数と f 関数の和により算出される．なお，Web 検索ヒット件数を求めるために Google[1] の検索エンジンを使用する．

SO-PMI において基本単語となる w_p と w_n をどのような単語にするかが重要である．Wang and Araki は人手によって基本単語を選択していた．我々は悪口極性を「死ね」「ウザい」などの悪口単語を用いた．また，悪口の逆極性として「イケメン」「可愛い」などの賞賛単語や,「チューリップ」「机」などの人が悪口とは関係がないと連想する単語を恣意的に選出した．事前に評価実験を行った結果，基本単語によって大きく精度が異なった．我々はより高い精度を得るために，悪口度算出に適切な基本単語の選定を行う．

3.2 基本単語の選定

本節では基本単語の選定方法について説明する．SO-PMI において最も重要な部分は式 (1) である．式 (1) は w_p と w，w_n と w の相互情報量 (MI) の除算から構成されている．w を悪口単語とした時，MI を大きくする w_p と MI を小さくする w_n を基本単語とすればより正確に悪口度となる SO-PMI を算出できるはずである．悪口単語との MI が大きい単語と小さい単語を次の方法で見つけ出す．

悪口単語と単語 7-gram[(4)] 内共起している単語のみを対象に MI を求める．悪口単語は人手で用意した 110

[1] http://www.google.com/

単語を使用する．悪口単語との MI の合計が高く，多くの悪口単語[2]と共起している単語が w_p として適切であるとする．一方，非悪口単語は悪口単語との MI が算出できない語，すなわち共起しない語が適切である．その中でも単独での出現頻度が高い場合に，悪口度が大きくなる．そこで，悪口単語と一度も 7-gram 内共起せず，単独での出現頻度が多い語が w_n として適切であるとする．対象とする単語の品詞は動詞-自立，名詞-一般，形容詞に限定した．単語切り出しには形態素解析器「MeCab(2)」を用い，品詞体系はこれに準ずる．また，単語は原形に直さず表層形のまま扱った．

多くの悪口単語と共起し，合計の MI が高かった上位 5 単語を例 1 に示す．また，悪口単語と共起せずに出現頻度が多い単語の上位 10 単語を例 2 に示す．

例 1) 悪口極性の基本単語候補

死ね，消えろ，蛆虫，カス，死ねよ

例 2) 非悪口極性の基本単語候補

引換，買い上げ，絞り込み，降順，振替

これらの単語を基本単語に使用して，悪口単語と非悪口単語に悪口度を与える実験を行い，悪口度が高い単語の中に悪口単語がいくつ含まれているかを検証した．その結果，3.1 節で恣意的に選出した賞賛単語や人が悪口と無関係だと連想した語を基本単語とした時よりも多くの悪口単語を認識できた．特に，賞賛単語は認識精度が低かった．この結果より，悪口度を与える場合は「賞賛」のような意味的な逆の考慮は必要がなく，単純に MI の大小関係のみを考慮すれば良いことが分かった．人が悪口と無関係だと連想する語は経験則的に悪口単語と共起しない単語を選出しており，本節で求める非悪口極性の適切な単語と同じである．しかし，認識精度に違いが発生した．本節では，悪口単語と共起しないという条件に加えて単独の出現頻度が高いという条件を加えた．この条件が精度に影響を与えたと考える．

例 1 と例 2 の組み合わせの中では，悪口極性に「消えろ」，非悪口極性に「振替」とした時に最も正確に悪口度を付与できた．そのため，本研究では「消えろ」と「振替」を基本単語に使用する．

[2]本稿では 20 語以上とした．

4 悪口文文分類

本節では悪口度を用いて，文を悪口文と非悪口文に分類する手法について説明する．

分類には SVM(5) を使用し，悪口度を用いた素性選択手法となる．文を構成する単語の中で，閾値を超える悪口度をもつ単語のみを素性とする．重みは一律とし，単語の存在のみによって判別する分類方法とする．この方法にした理由は，悪口文は悪口単語の影響によって悪口文と判断される事がほとんどであり，悪口単語の有無だけでも十分に判断できると考えたからである．

悪口文分類において，悪口単語の有無でほとんどが悪口文だと判断できるが，悪口単語が否定されている場合は悪口単語を含んでいても悪口文とならない．そのため，否定を表す語 (否定語) を考慮しなければならない．

本手法では，否定語と文節内共起している悪口単語は，悪口単語として扱わないこととする．否定語は「ない」のみを取り扱う．文節の切り出しには係り受け解析器「Cabocha(3)」を使用した．

5 評価実験

2 ちゃんねるの書き込みを入力文として，提案手法を用いて悪口文と非悪口文に分類する実験を行った．入力文中の単語から悪口単語のみを素性とする提案手法に対して，入力文中の単語を全てを素性とする手法をベースラインとした．実験には，2 ちゃんねるより収集した悪口文 1400 文と非悪口文 1400 文を使用した．これらの文を使って 5 分割交差検定を行い，最終的な精度は平均値を提示する．評価指標には適合率，再現率，F 値を使用した．評価結果を図 1 に示す．

図 1 より，最も F 値が高くなったのは閾値が-0.2 の時であり F 値は 89.97 である．これはベースラインの F 値よりも約 5 ポイント高い結果となっている．このことから，悪口度をもとにした素性選択手法は有効であることが分かる．

また，F 値が最も高かった閾値が-0.2 であったことから悪口度が高い単語のみでの判別は難しく，悪口度が著しく低い単語のみを素性から削除するだけで精度は十分に向上することが分かった．

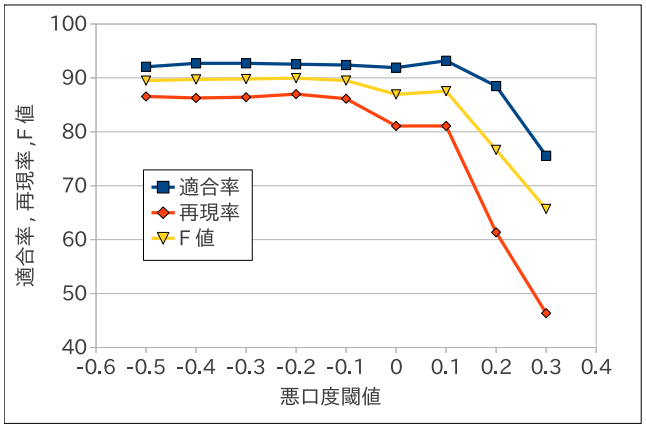

図 1: 悪口文の分類精度

6 考察

誤り解析をした結果，造語の認識誤りが原因であるものが主だった．例文として以下の提示する．

(例文) 意味がわからんスレたてるな競馬鹿

この例文の場合，「競馬鹿」を含むことが悪口文と判別される原因である．しかし，単語分割をする際に「競馬」と「鹿」という悪口度の低い単語に分割されていた．素性選択の際に排除され，悪口の原因となる要素がなくなったため非悪口文に分類された．

本稿では MeCab を使用したことで「キモヲタ」などのある程度の造語に対しては認識することができたが，全ての造語に対して対応はできていない．

主な分類誤りの原因は単語の切り分けに関する問題であったため，悪口度を用いた素性の選択手法は十分に有効であると考える．今後は 2 ちゃんねるの書き込みに限らず多くの種類の文を入力として実験を行い，問題を分析して改善に努めたい．

7 おわりに

本稿では 2 ちゃんねるの書き込みを対象に誹謗中傷を表す文 (悪口文) を検出する手法を提案した．文を構成する単語に対して悪口文に含まれるか否かを示す悪口度を与えた．悪口度の算出には 2 つの基本単語との共起頻度の偏りを算出する SO-PMI を用いた．悪口極性の基本単語には MI が高く，多くの悪口単語と共起している語が適切であることが分かった．また，非悪口極性には悪口単語と共起せず，単独での出現頻度が高い語が適切であることが分かった．この悪口度をもとに SVM の素性の単語を選択して文分類を行った．

その結果，F 値においてベースラインよりも約 5 ポイント向上した．

この手法が 2 ちゃんねる以外の文に対しても同様の結果が得られるか今後実験していきたい．

使用した言語資源及びツール

(1) 電子掲示板サイト “2 ちゃんねる”, http://www.2ch.net/

(2) 形態素解析器「MeCab」, Ver.0.98, http://mecab.sourceforge.net/

(3) 係り受け解析器「CaboCha」, Ver.0.52, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/˜taku/software/cabocha/

(4) Web 日本語 N グラム第 1 版, http://www.gsk.or.jp/catalog/GSK2007-C/

(5) SVM 学習ツール 「Tiny SVM」, Ver.0.09, http://chasen.org/˜taku/software/TinySVM/

参考文献

[1] José María Gómez Hidalgo, Ignacio Giráldez, and Manuel de Buenaga. Text Categorization for Internet Content Filtering. *Revista Iberoamericana de Inteligencia Artificial*, pp. 34–52, 2003.

[2] 池田和史, 柳原正, 松本一則, 滝嶋康弘. 係り受け関係に基づく違法・有害情報の高精度検出方式の提案. 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM2010), C9-5, 2010.

[3] 松葉達朗, 桝井文人, 井須尚紀. 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 5D-4, pp. 383–386, 2010.

[4] Perter D. Turney and Michael L. Littman. Unsupervised. learning of semantic orientation from a hundred-billion-word corpus. *National Research Council, Institute for Information Technology, Technical Report ERB-1094(NRC-44929)*, 2002.

[5] Guangwei Wang and Kenji Araki. Modifying SO-PMI for Japanese weblog opinion mining by using a balancing factor and detecting neutral expressions. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pp. 189–192. Association for Computational Linguistics, 2007.