2 ちゃんねるを対象とした悪口表現の抽出

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

1 はじめに

Web 上には，他者への誹謗中傷 (悪口) が多く存在する．最近では小中学生などの低年齢層も他者への悪口を書き込み，ネットいじめとして社会的問題となっている [1]．主なネットいじめの現場となるのが電子掲示板 (BBS) である．電子掲示板は情報交換や相談・議論の場として主に利用され，連日多くの書き込みが行われている．しかし，犯罪予告や，他人の個人情報の暴露，誹謗中傷を目的とした書き込みも多く存在する．巨大電子掲示板サイト “2 ちゃんねる (1) ” は匿名性が強く，悪口を書き込みやすいため，特にその傾向が強い．

悪口は人権問題に関わるため，管理されなければならない．悪口を管理するために，人手で Web ページをパトロールをしている企業や自治体はある．しかし，人手での監視は大きな負担であるため，可能な限り自動にするべきである．携帯コミュニティサイトを運営している (株) 魔法 i らんどの調査結果[1]では，ネットいじめを減らす方法として「システムで中傷する言葉を書けなくする」ことを挙げており，悪口を特定する技術は必要とされている．

自動で悪口を管理する堅実な方法としては悪口表現辞書の使用したフィルタリングがある．悪口表現辞書があれば，悪口書き込みを自動で発見でき，対処が容易になるはずである．

しかし，現時点では公開されている悪口表現辞書はない．そのため，我々は多くの悪口が存在する 2 ちゃんねるから悪口表現を抽出し，悪口表現辞書の構築を目的とする．2 ちゃんねるに限定した理由は Web 全体の悪口表現の種類数と 2 ちゃんねるの悪口表現の種類数には大きな差はないと仮定したからである．本稿では悪口表現辞書の構築の第一歩として，悪口表現の使われ方に着目し，悪口表現の抽出手法を提案する．使われ方とは，ある単語列と悪口表現の連接確率を考慮したもので n-gram を使用した抽出方法である．

1.1 悪口表現の定義

一概に悪口表現といっても曖昧である．本稿での悪口表現は，特定の他者に対して直接的に侮辱や誹謗中傷している単語，句とする．皮肉のような文脈や他の情報に依存する中傷は対象としない．以下に対象とする悪口表現を含む文の例を示す．

(a) あの政治家死ね
(b) 奴らはバカな暇人野郎

(a) では「死ね」が悪口表現となる.「死ね」のような表現は周辺単語に依存せず，単語で悪口表現になるため単語で抽出する．(b) では「バカ」が悪口表現になり得る．しかし，本稿では悪口表現を単語に限定していないため,「バカな暇人野郎」のような句を収集する．単語に限定しない理由は，悪口表現は単語で悪口と判断出来ない場合があるからである．例えば,「バカ」と単語で他者に向けると悪口表現として認識される．しかし,「バカうまい」など他の単語との組み合わせにより悪口表現とならない場合がある．そのため，本稿では悪口表現を単語に限定せず，句も対象とする．表 1 に本稿で対象とする悪口表現の例を示す．

表 1: 悪口表現の具体例

悪口表現
みんなまとめて逝け
死んでくれって思う
バカな暇人野郎
マジうざい
キモイ！
ヲタは地獄に落ちろ
死ね

2 関連研究

悪口表現の抽出に類似した研究には松葉ら [2] の研究がある．松葉らは学校非公式サイトの悪口表現を有害情報として扱い，抽出を行っている．抽出のために，レーベンシュタイン距離の利用で同義・異表記の単語を統一している．また，掲示板の書き込みの有害・無害判断や書き込みの悪質度を測定している．我々は n-gram 確率を使用し，周辺単語から悪口表現の抽出する．そして，本稿では悪口表現抽出のみを行う．

悪口表現の抽出は評価表現の抽出と類似している．評価表現の抽出の研究に Turney and Littman[3] の研究がある．Turney and Littman は “excellent” や “poor” などの評価表現を種辞書にして，共起頻度を求めることで評価極性の強さを求めている．種辞書を使用する点では本研究と同じである．しかし，2 ちゃんねるのようなくずれた日本語を扱う場合に，共起では形態素解析誤りの影響が大きい．我々は n-gram を使用することで，過分割された形態素をまとめて扱い，形態素解析誤りの影響を軽減させている．

n-gram を使って未知語や専門用語を抽出する研究に森ら

[1] http://ipolice.jp/4.pdf

[4] の研究がある．森らは大規模なコーパスからの単語抽出とその単語の品詞の推定を行っている．文字列の前後に隣接する n-gram に着目する点は本稿と同じである．森らは品詞情報の連接確率を使用しているが，我々は品詞情報を全く使用していない．2 ちゃんねるの文では正確に品詞を付与することは困難であるため，品詞情報は無視した．また，我々は抽出対象を単語に限定していない．

3 手法

本手法は以下の 4 つの構成要素から成る．図 1 はシステムの概略図である．

1. 悪口表現種辞書の構築
2. 悪口文の収集
3. 悪口単語 n-gram モデルの作成
4. 悪口表現抽出

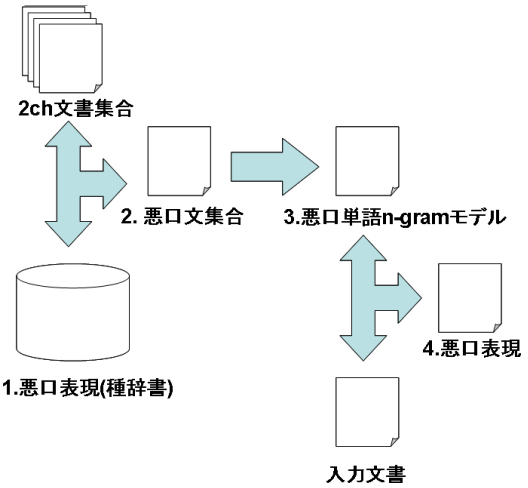

図 1: 提案手法の処理の流れ

3.1 悪口表現種辞書の構築

n-gram モデルの作成のためには多くの悪口表現を含んだ文 (悪口文) が必要である．しかし，大量の文の中から悪口文のみの抽出を人手で行うのは大きな負担である．そこで，我々は人手で 2 ちゃんねるから悪口表現を抽出し，小規模な悪口表現辞書を作成した．この辞書を種辞書として悪口文を自動で収集する．種辞書に登録されている悪口表現の数は 103 件であり，表 1 のような悪口表現が登録されている．

3.2 悪口文の収集

悪口表現種辞書を使って 2 ちゃんねるから記事を読み込み，悪口文を収集する．2 ちゃんねるの文で種辞書の登録語を含む文を悪口文とした．現在の登録語では毎日，約 2000 記事を解析し約 6500 文の悪口文を収集できている．以下は収集した悪口文の例である．

- つか，官僚死ねや
- 泥棒ゴミクズ団体はさっさと吊ってこい！
- いちいちひねくれた言い方すんな キモイ奴だな
- こんなんでイチイチ騒ぐなボケカス。

3.3 悪口 n-gram モデルの作成

我々は単語 n-gram を使用して，悪口表現と連接しやすい単語列を求める．今回は 1〜5-gram を考慮する．ここで問題となるのが悪口表現が 1 語とは限らないということである．複数の単語から構成される悪口表現の周辺単語を n-gram で求めようとした時，5-gram 全てが悪口表現となる場合がある．この場合，周辺単語は獲得出来ないため，我々は前処理として，悪口表現を 1 語に汎化して周辺単語を獲得できる状態にした．

n-gram モデルを作成するために SRILM[(2)] を使用した．SRILM は低頻度問題の対策として自動でバックオフスムージングを行う．n-gram モデルを作成する際には，文のわかち書きが必要である．そのために形態素解析器茶筌 [(3)] を使用し，茶筌の辞書には IPA 品詞体系 [(4)] を使用した．わかち書きの際に，活用形による確率の分散を避けるために，文の形態素は全て原形にして扱った．

悪口表現の直前に連接する単語列 (1〜4-gram) を左連接属性と呼ぶ．また，悪口表現の直後に連接する単語列 (1〜4-gram) を右連接属性と呼ぶ．悪口表現の直前は前向き単語 n-gram，直後は後ろ向き単語 n-gram でモデルを作成する．

種辞書をもとに 2 ちゃんねるの悪口文を収集し続けた結果，悪口文は約 20 万文となった．また，悪口表現と連接しやすさを求めるために非悪口文を 2 ちゃんねるから約 50 万文を収集した．総数約 70 万文を使用して前向きと後ろ向きの単語 n-gram モデルを 2 つ作成した．さらに，各モデルの悪口表現を含む単語 n-gram だけを抽出し，合わせて 1 つにしたモデルを悪口単語 n-gram モデルとした．モデルの一部を表 2 に示す．<悪口> に該当する部分が悪口表現候補となり，抽出対象となる．ただし，実際に使用する場合は<悪口> を削除する．これにより，モデルは 1〜4-gram となる．左側の数値は連接確率である．本稿ではこの数値が高いほど，悪口表現と連接しやすいとしている．

表 2: 悪口単語 n-gram の例

悪口単語 n-gram モデル	
0.94	<悪口> だ な 日 本
0.67	<悪口> は さっさと 日本 から
0.62	<悪口> は 何でも 他人 の
0.94	ない 化学 の 専門 <悪口>
0.22	顔 見る と 大体 <悪口>
0.41	相当 身勝手 だ 単なる <悪口>

3.4 悪口表現の獲得

作成した悪口単語 n-gram モデルを用いて悪口表現を抽出する．表 2 のように単語 n-gram には悪口表現との連接確率が付与されている．入力文が悪口単語 n-gram モデルの表現を含むなら左連接属性か右連接属性かを考慮し，一致した部分の前方か後方を抽出する．さらに，抽出した部分が他の悪口 n-gram を含むなら再度抽出する．抽出の順番は連接確率の高いものから優先的に使用する．以下に抽出までの流れの例を示す．

入力文　マスゴミのクズどもって，何でこうなる事が選挙前から解りきってたのに 捏造や印象操作してまで○○[2]の宣伝しまくったわけ？

原形化　マスゴミ の クズ どもって , 何 で こうなる 事 が 選挙 前 から 解る きる てる た のに 捏造 や 印象 操作 する て まで ○○ の宣伝 する まくる た わけ ？

適用される n-gram　どもって，

属性　右連接属性

抽出される悪口表現　マスゴミ の クズ

4 評価実験

提案した手法を用いて悪口文と非悪口文を入力し，正確に悪口表現を抽出できるかを実験した．さらに，予備実験として，新しい悪口表現をいくつ獲得できるかを検証した．新しい悪口表現を多く獲得出来れば，辞書の悪口表現の登録数が増える．そうなれば，多くの悪口文を収集できる．

今回は種辞書に新しい悪口表現の追加を視野に入れたため，非悪口表現の抽出する数を抑止したかった．そのため，再現率よりも適合率を優先した．悪口表現の抽出実験の評価は抽出された文字列を人が見て，悪口表現を含むかどうかを判断した．

4.1 評価セットの作成

評価セットとして学習データとは別に 2 ちゃんねるから悪口文，非悪口文の文書セットを人手で用意した．内訳は悪口文が 378 文，非悪口文は 382 文である．

4.2 悪口表現抽出実験

悪口表現の n-gram モデルの左連接属性と右連接属性のみを別々に考慮した時を適合率の結果を図 2 に示す．再現率の結果を図 3 に示す．適合率はシステムが悪口表現と判断した部分が本当に悪口表現だった場合の割合であり，再現率は全悪口表現の中でシステムが出力した悪口の割合である．閾値は使用する n-gram を制限するための連接確率の閾値である．閾値が 0.9 や 0.8 などの場合は適合率が 1 であったが，再現率は最高でも 0.3 程度だった．

[2]実際には個有名詞だが，本原稿上では “○○” と表記する．

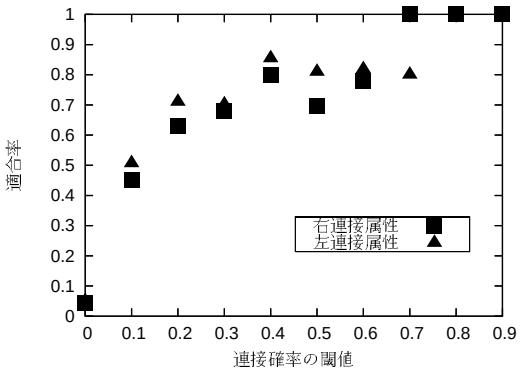

図 2: 閾値による適合率の推移

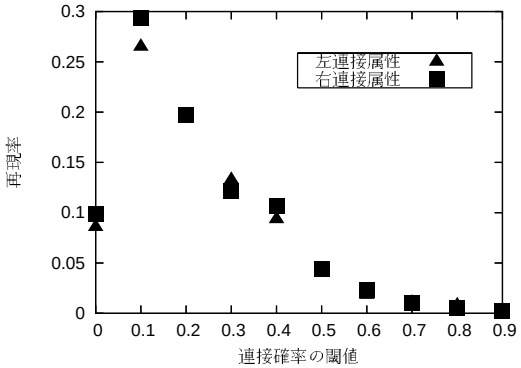

図 3: 閾値による再現率の推移

4.3 新しい悪口表現の獲得

本節ではシステムが出力した悪口表現の中で種辞書に登録されていない語を獲得できた数の評価を行う．抽出した悪口表現が種辞書に登録されているか否かを人手で判断した．

図 4 に閾値による新しい悪口表現の獲得数の推移を示す．閾値が低いほど，新しい悪口表現の獲得数は多くなった．しかし，閾値を設定しない (閾値が 0) 場合，獲得数が少ない．これは抽出した悪口表現が他の悪口表現を含む場合に再度抽出することが原因である．悪口単語 n-gram は 1〜4-gram で構成されているが，閾値を設定した場合は 1-gram を使用することはほとんどない．しかし，閾値を設定しない場合は 1-gram も使用するため，繰り返し悪口表現の抽出が行われる．その結果，出力される悪口表現は短くなり，意味を保持せず悪口表現が抽出できてない．抽出された悪口表現の数が少ないため，新しい悪口表現の数も極端に少なくなった．

5 考察

5.1 適合率と再現率について

閾値が 0.9 や 0.8 など高ければ確実に悪口表現を抽出できたが，この時に抽出できた悪口表現の数は 3 件程度だった．また，再現率は最も高いもので 0.3 程度だった．この

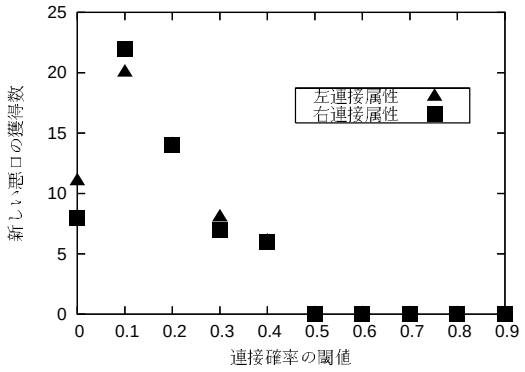

図 4: 新しい悪口表現の獲得数

ことから，悪口表現にのみに連接しやすい単語列の数は少ないことが分かる．悪口表現は定型的に存在するわけではない．より悪口表現の特徴を見極め，悪口表現に適した手法の考案が必要である．

今回は種辞書への登録を考慮したため，適合率を優先させたが，悪口表現の獲得数が少ないため，辞書の拡張は見込めない．再現率の向上にも努めなければならない．

5.2 抽出範囲について

本稿では連接確率を用いて悪口表現を抽出できるかを検証した．今回の手法では悪口表現が含まれるか否かの判定と，悪口表現の位置の特定が可能であるが，悪口表現の範囲の特定は出来ない．悪口表現は句や単語であるため，今回の実験では悪口表現候補を広い範囲で抽出した．そのため，悪口表現以外の語も悪口表現として抽出される．これからの課題として，どこまでが悪口表現かを正確に特定する方法も検討しなければならない．

5.3 新しい悪口表現の獲得について

獲得した新しい悪口表現を表 3 に示す．

表 3: 新しく獲得した悪口表現の例

悪口表現
キモオタロリコン
消えてしまえ，馬鹿
デブ婆ァ
スタイル悪い
カス芸人
馬鹿男女
負け豚の遠吠え
クズねらー

表 3 に示すように，今回の手法で閾値を操作すれば新しい悪口表現は獲得できる．しかし，新しい悪口表現を獲得するためには閾値を低くしなければならない．閾値を低くすれば，非悪口表現を多く獲得してしまう．そのため，閾値を低くすることが解決策にはならない．我々は，別の方法で非悪口表現の抽出数を減少させなければならない．

表 3 の例の他に,「糞○○」という新しい悪口表現を多く獲得できている.「糞」は 1 語では別の意味の可能性があるため，悪口表現として登録はできない．しかし，糞ガキ，糞ゲー (ゲーはゲームの意)，のような名詞と連接することで悪口表現として確定的になり，辞書に登録できるようになる．このような同じ単語の造語は同じ周辺文字列を持つため，今回の手法で獲得できた．

6 まとめと今後の課題

本稿では 2 ちゃんねるの書き込みから悪口表現を抽出する手法を提案した．提案手法は n-gram 確率を用いた単純な手法である．評価実験の結果，閾値が高ければ高確率で悪口表現を抽出できるが，抽出できる数は極端に少ない．今後は悪口表現が定型的に存在するわけではないことを考慮し手法の改良を行っていく．

閾値を変化させて悪口表現を抽出した時に，新しい悪口表現を獲得できた．今後も新しい悪口表現を種辞書に登録し続けていき，辞書の大規模化を目指す．

本稿の結果をベースラインとして，悪口表現の抽出の精度向上を目指す．

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

(2) 言語モデル作成ツールキット「SRILM」, http://www.speech.sri.com/projects/srilm/.

(3) 形態素解析器「茶筌」,Ver.2.3.3, 奈良先端科学技術大学院大学 松本研究室, http://chasen.naist.jp/hiki/ChaSen/

(4) IPA 品詞体系日本語辞書「IPADIC」, Ver.2.7.0, 奈良先端科学技術大学院大学 松本研究室, http://chasen.naist.jp/stable/ipadic/

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, http://www.mext.go.jp/bmenu/houdou/20/11/08111701/001.pdf, 2008.

[2] 松葉達朗, 里見尚宏, 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. 情報処理学会研究報告, NL192-15, 2009.

[3] 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.

[4] 森信介, 長尾眞. n グラム統計によるコーパスからの未知語抽出. 情報処理学会論文誌, Vol. 39, No. 7, pp. 2093–2100, 1998.