2009-IOT-4 (38)
2009 / 3 / 6

掲示板の発言に対する自動判別を用いたユーザ教育支援手法の改良

一藤　裕† 今野　将†† 曽根　秀昭†††

† 東北大学大学院情報科学研究科　〒 980–8578 宮城県仙台市青葉区荒巻字青葉 6–3
†† 千葉工業大学工学部　〒 275–0016 千葉県習志野市津田沼 2–17–1
††† 東北大学サイバーサイエンスセンター　〒 980–8578 宮城県仙台市青葉区荒巻字青葉 6–3
E-mail: †ichifuji@mail.tains.tohoku.ac.jp, ††konno.susumu@it-chiba.ac.jp, †††sone@isc.tohoku.ac.jp

あらまし　インターネット上で学校または地域ごとに集まりコミュニケーションをとる学校裏サイトのような電子掲示板では，匿名性などの特徴や誤解から，陰湿ないじめに発展し社会問題となっている．これは，電子掲示板に書き込んだ発言が他者にどう影響を与えるかを知らないことが原因の一つとして考えられる．そこで我々は，個々の発言が他者にどう影響を与えるかの指標を確立し，その影響度を発言者へ示すことにより発言者が問題となる発言をしないように教育を支援する手法を提案した．本稿では，以前提案した手法では対応できなかった発言に対応するため，学習データベースに登録する品詞を選択する．また，発言の長さに応じた発言の判別を行うのではなく，2 つの学習データを利用して発言の判別を機械的に行う手法を提案する．

キーワード　電子掲示板, 社会問題，倫理教育

Improving the educational assistant tool for BBS users using comments evaluation system

Yu ICHIFUJI†, Susumu KONNO††, and Hideaki SONE†††

† Graduate School of Information Sciences, Tohoku University　Aramaki Aza Aoba 6–3 Aoba-ku, Sendai, Miyagi 980–8578 Japan
†† Faculty of Engineering, Chiba Institute of Technology　2–17–1 Tsudanuma Narashino, Chiba 275–0016 Japan
††† Cyber Science Center, Tohoku University　Aramaki Aza Aoba 6–3 Aoba-ku, Sendai, Miyagi 980–8578 Japan
E-mail: †ichifuji@mail.tains.tohoku.ac.jp, ††konno.susumu@it-chiba.ac.jp, †††sone@isc.tohoku.ac.jp

Abstract　Electronic bulletin board systems (BBS) have become a common communication tool on the Internet. Some people who use BBS anonymously sometimes write opinions which make other people on the BBS feel bad. This is especially true as users become younger and younger and the use of BBS on the Internet and cellular phones becomes more widespread. Users don't understand how such comments can affect readers. As a result, such negative statements sometimes lead younger users to behave inappropriately on BBS. The authors have developed an assistance tool which shows users what kind of influence their statements might have on other people on the BBS. Hopefully this will make all BBS users more aware of the effects of their words on others. This should lead to a more positive atmosphere on the BBS. To improve the accuracy of evaluation for each comment, we introduce new word class to our assistant tool. Also, we propose a new automatic evaluation method using two types of samples.

Key words　BBS, Social issues, Ethics education

1. はじめに

現在，学校裏サイトと呼ばれる学校のいじめの温床となっているサイトがインターネット上に多数存在し，社会問題となっている．これらのサイトでは，各学校ごと，もしくは，その地域の学生によって，生徒や先生への不満や誹謗・中傷が書き込まれており，文部科学省の調べによると，書き込まれた発言の約 27%に「死ね」「殺す」といった暴力的な言葉が存在すると

報告されている [1]．この問題に対処するためには，学校裏サイトを把握し上記のような問題となる発言（以下，“問題発言”）をしないよう教育することが必要である．

問題発言をさせない・減少させるために，現在，NG ワードフィルタの設置や掲示板へのアクセスブロック，掲示板の監視パトロールなどの対策が取られている．また，学校裏サイトを把握するために，学校裏サイトチェッカー [2] というサイトが作成され，数多くの学校裏サイトが登録されるようにもなっている．しかし，このような対策のみであると，だれでも気軽に利用できる掲示板の特性を失わせることになる．また，本音の討論においては，多少口汚い言葉が混じるものであり，このような言い争い（以下，“フレーミング” と呼ぶ）を全て規制してしまうと，掲示板の存在意義すら失わせるという観点から，すべてのフレーミングを規制する必要はないという意見も存在する [4] [5] [6]．著者らもこの意見に賛同し，規制だけでなく，ユーザへのインターネットの倫理教育が必要不可欠であると考える．なぜなら，問題発言はインターネットの匿名性に対する誤解や自分の発言が相手にどう受け取られるかを正しく理解していないことが原因の一つとして考えられるからである．

そこで著者らは，インターネットの個々のユーザに対し，各発言が他者にどう影響を与えるかを既出発言を利用しベイズ理論を用いて影響力を算出し，問題発言かそうでないかを自動的に判断し，問題発言の場合，投稿前にユーザへ注意を促すシステムを提案した. [3] このシステムは，掲示板の発言を判断する基準がないことから，既出発言を学習データとして用い判断基準とし，新たな発言をベイズ理論を用いて問題発言かどうか判別するものである．学習データには，発言を単語に分解し助詞・助動詞を除いたものと単語のペアを作成したものの 2 つを利用しそれぞれを発言の長さに応じて使い分けていた．しかし，これでは判別できないものや誤判定をしてしまうことが問題が存在する．また，発言の長さによってどの学習データを利用するかを決定していたが，どの程度の長さの時にどのデータを使うかの明確な基準がない問題もある．そこで本稿は，誤判定してしまう発言に対応するため，除外していた助詞・助動詞のなかから，判断に有用な言葉を追加することを提案する．また，機械的に判別を行うために，発言の長さに対応させるのではなく，2 つのデータを利用し，2 つの結果を考慮した判別システムを提案する．これにより，未対応の発言に対応させ，また，機械的な判別の精度を上昇させることを実現する．

2. 既存手法と問題点

はじめに既存手法による発言判別の方法を述べる．その後，既存手法の問題点を述べ改良方法を提案する．

2.1 既存手法による発言判別

既存手法では，発言の影響力を算出するために，掲示板の既出発言を誹謗中傷発言のような他者に不快感のみを与える発言（以後，“問題発言” と呼ぶ）とそれ以外の発言の 2 つのクラスに分類し，それを基準とした．この手法は，図 1 のように学習フェーズと判別フェーズの 2 つで構成されている．

学習フェーズでは，発言を判別するための基準を学習させる．

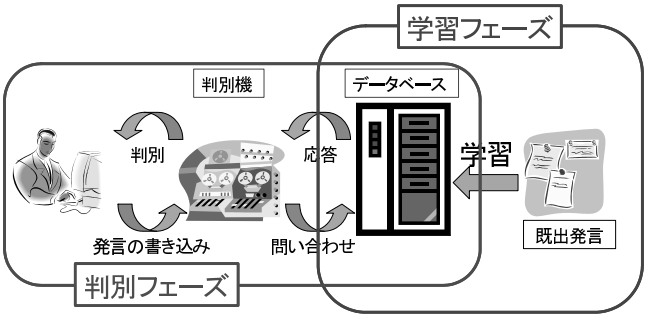

図 1　提案手法の概要

具体的には，既出発言を人間の判断により分類する．その後，単語に分解し助詞・助動詞を取り除く．残った単語から単語のペアを図 2 のように作成し，学習データベースにそれぞれを登録する．

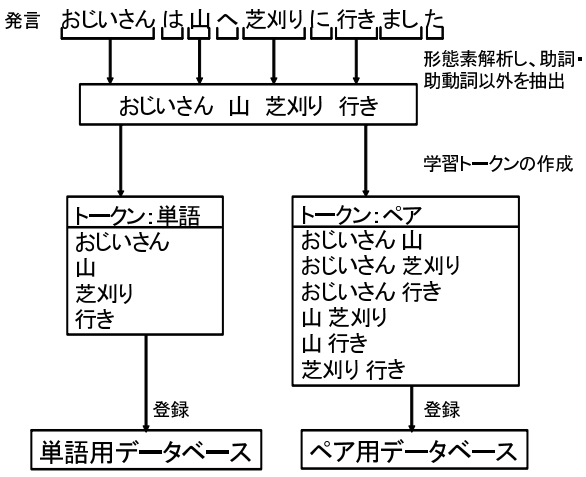

図 2　データベースへ登録するための分解図

判別フェーズでは，判別対象の発言の長さに応じて判別を行う．短い発言は発言中の単語がその発言を象徴していると考えられるため，単語を利用して発言の判別を行う．また，発言が長い場合，引用文を用いて発言する場合や，最後に否定する場合に対応させるために，単語のペアを利用して発言の判別を行う．判別の流れは 3 のようになる．

新たな発言があるとき，まず，その発言を形態素解析し，必要な品詞の単語を抜き出す．抜き出した単語数から，発言の長さを決定する．その後，長さにあった学習データを用い，発言の影響度を算出し，その結果を発言者へ示す．

単語の場合，発言の影響度を算出するために，まず，出現単語の影響度をそれぞれ算出する．表 1 の問題発言クラスにおける x^i の総数を $S_b(x^i)$，通常発言クラスにおける x^i の総数を $S_g(x^i)$，学習させた問題発言の総数を BN，通常発言の総数を GN とすると，ある単語 x の影響度 $\pi(x^i)$ は式 3.2 で算出される．

$$\pi(x) = \frac{\frac{S_b(x^i)}{BN}}{(\frac{S_b(x^i)}{BN}) + 2 \times \frac{S_g(x^i)}{GN}} \tag{1}$$

発言 y の影響度を $p(y)$ とすると，$pi(x^i)$，GN，BN を用い

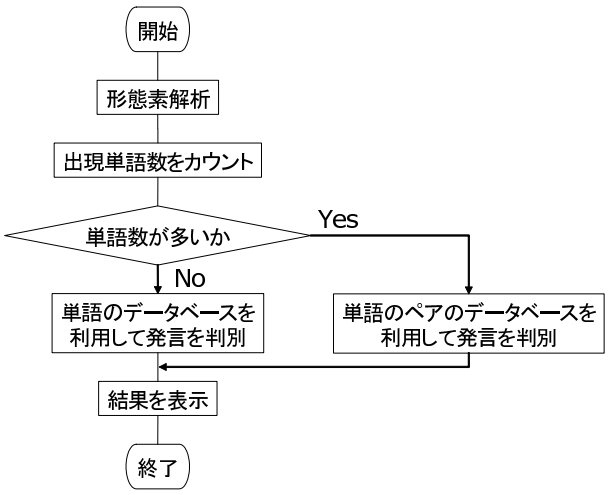

図 3　発言の判別の流れ

て，式 3.2 で算出される．

$$PB = \frac{BN}{BN + GN} \tag{2}$$

$$p(y) = \frac{(PB)^{1-m}\prod_{i=1}^{m}\pi(x^i)}{(PB)^{1-m}\prod_{i=1}^{m}\pi(x^i) + (1-PB)^{1-m}\prod_{i=1}^{m}(1-\pi(x^i))} \tag{3}$$

このようにして，発言者へ，自身が発した発言が他者にどう受け取られるかを式で示す．単語のペアの場合も同様に算出する．算出された値が 1 に近いほど，誹謗中傷発言と判別され，0 に近いほど問題のない発言クラスに分類される．通常発言と問題発言の閾値を決め，その閾値を超えた場合，発言者へ問題発言となりうることを示す．

2.2　既存手法の問題点と提案

既存手法では，発言の判別基準の学習時において，助詞・助動詞をすべて取り除いているが，助詞のなかにもいくつか重要な単語が含まれている．「ない」などの否定語である．つまり，「○○しない」「○○はよくない」などの否定部分「ない」「ず」「ねえ」などが判別基準に含まれていないため，「この問題発言はよくない」という発言が問題発言そのものとして判断されてしまうことがある．そこで本稿では，助詞のうち否定語を学習データとして利用することを提案する．

また，既存手法では発言の長さ（出現単語数）に応じて判別を行っていたが，どの程度長さの発言を長い発言として捉えるか難しい．そのため，発言の長さに応じて単語単体と単語のペアの場合を 2 つに分けるのではなく，両方の結果を踏まえ判別を行うことを考える．つまり，発言の長さによらず，単語の場合と単語のペアの場合の 2 通りの判別を行い，2 つの結果を総合して，発言の判別を行うことを提案する．

3.　既存手法の改良

既存手法と同様に，改良手法も発言を判別するための学習を行う学習フェーズと学習データを用いて判別を行う判別フェーズの 2 つによって構成されている．以下で説明を行う．

3.1　学習フェーズの改良

人手を用いて発言を問題発言と通常発言に分類後，発言ごとに形態素解析し単語に分解する．従来手法では，品詞のうち助詞・助動詞をデータとして用いないこととしていた．改良後は，除外していた助詞のうち，否定を意味する語（「ない」「ねぇ」など）を学習データとして採用する．その後，単語のデータベースには分解した単語を，単語のペアのデータベースには分解した単語から単語のペアを作成し，データベースに登録し学習させる．データベースには，登録単語または登録単語のペアと問題発言に出現した回数と通常発言に出現した回数が登録されており，表 1 のようになっている．

表 1　単語のデータベースの登録例

登録単語	問題発言クラス	通常発言クラス
お前	66	130
ありがとう	0	14
うざい	20	3
・	・	・

データベースへの登録について詳しく述べる．まず，既出発言を問題発言と通常発言の 2 クラスに人手を用いて分類し，それぞれのクラスごとに処理を行う．1 発言ごとに形態素解析を行い，必要とする品詞の単語を抜き出す．単語の学習データベースには，抜き出した単語とそれぞれのクラスでの出現数が登録される．単語のペアの学習データベースには，抜き出した単語ですべてのペアを作成し，その後，単語のペアのデータベースに単語のペアとそれぞれのクラスでの出現数を登録する．登録までの流れは図 4 のようになる．

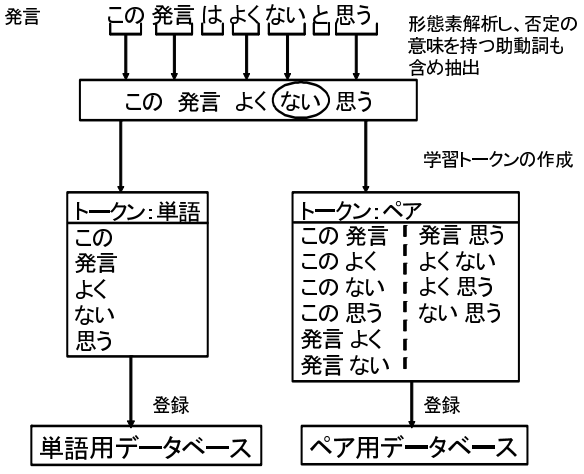

図 4　データベースへ登録するための分解図

このように 2 種類の学習データを準備することにより，反対意見を述べることやフレーミングなどを容認しつつ，誹謗中傷の発言を自動判別することを目指す．

3.2　判別フェーズの改良

投稿される発言が問題発言と通常発言のどちらに分類されるか数値で算出し判別を行う．新たな発言があるとき，まず，その発言を形態素解析し，必要な品詞の単語を抜き出す．

単語の場合，発言の影響度を算出するために，まず，出現単

語の影響度を式を用いてそれぞれ算出する．その後，新たに投稿された発言が他者にどう受け取られるかを式を用いて算出する．単語のペアの場合も同様に算出する．算出された値が 1 に近いほど，問題発言と判別され，0 に近いほど通常発言と判別される．分類された結果，問題発言だった場合，発言者へ問題発言の可能性があることを示す．既存手法では，発言の長さ（出現単語の数）に応じてどちらのデータを使って判別するかを選択していた．改良手法では，発言の長さによらず，2 つのデータを使って発言の判別を行う．そのため，以下の 4 つのパターンの判別結果が得られることになる．

（1） 単語と単語のペアによる判断がともに問題発言

（2） 単語と単語のペアによる判断がともに通常発言

（3） 単語によって通常発言と判断され単語のペアによって問題発言と判断される場合

（4） 単語によって問題発言と判断され単語のペアによって通常発言と判断される場合

パターン 1 は，どちらも問題発言と判断しているので問題発言と判断する．パターン 2 も同様にどちらも通常発言と判断しているので通常発言と判断する．パターン 3 については，問題発言に出現する単語がない，または，少ないが，ペアを作る場合問題発言に出現するものが多いという結果である．これは，笑いの意味を示す「w」や強調する「！」などを使った発言を判別するときに出現する可能性がある．つまり,「w」単体などはよく出現するため，目立った特徴量を持たないが,「w-w」や「w-！」などのように組み合わさった場合挑発発言に見られる形となり，パターン 3 のような判定となる可能性がある．しかし，単語単体では問題発言ではないという結果が出ているため，パターン 3 は通常発言と判別する．パターン 4 は条件付き問題発言と判別することにする．本手法では，フレーミングや多少の口汚い言葉はある程度許容する立場をとっている．このような発言には，ある程度問題発言に出現するような単語が使われている．その結果，単語のみに着目し発言の判別を行うと，このような発言はすべて問題発言として判断され，批判的な発言や単語が使えなくなる可能性がある．したがって，単語による発言判定では問題発言となる発言もある程度は許容する必要がある．そこで，単語のペアの結果を利用する．パターン 4 では，単語のペアによる判別では通常発言と判別されることから，批判的な内容かもしれないが，フレーミングのような発言の可能性が高いとみなすことができる．

4. 検 証 実 験

本稿では，従来手法の改良が効果的であることを示すために検証実験を行う．学習フェーズでは取り扱う品詞を変え，判別フェーズでは単語による判別とペアによる判別を両方利用して判別を行う．従来手法と改良手法の判別結果の比較を行う．

4.1 実 験 準 備

今回，いわゆる学校裏サイトと呼称される掲示板の発言を検証実験に利用している．2004 年 11 月から 2008 年 4 月までに書き込まれた発言を対象とし，人手を用いて問題発言を 520 発言，通常発言を 1276 発言抽出し，学習させている．形態素解析は，“mecab” という形態素解析エンジンを使用し，ネットで多様される造語などをあらかじめ登録している [7]．また，従来手法も改良手法も同じ学習データを用いている．

4.2 実 験 内 容

改良手法の有効性を示すために，従来手法と改良手法の判別結果の比較と，判別に使われる単語・ペアの比較を行う．また，単語と単語のペアを利用した場合の組み合わせ方法を調べるために，次の 4 パターンと人間による判断との比較を行う．

（1） 単語と単語のペアによる判断がともに問題発言

（2） 単語と単語のペアによる判断がともに通常発言

（3） 単語によって問題発言と判断され単語のペアによって通常発言と判断される場合

（4） 単語によって通常発言と判断され単語のペアによって問題発言と判断される場合

パターン 1，2 は人間の評価と一致するかを比較する．パターン 3，4 は人間の評価がどのような分布になるかを調べ，パターン 3，4 が出現した場合，その発言が通常発言と問題発言のどちらに評価するべきかを明らかにする．判別は 0.75 を基準とし，0.75 以上の場合，問題発言と判断し，0.75 未満の場合，通常発言と判断する．

4.3 実験サンプル

検証実験に使用した掲示板の発言は 2007 年 5 月から 2008 年 11 月までの 895 個である．実験に使用したデータのサンプルの 1 部を以下に示す．データは，書いた人物を特定されないため，一部改変している．実験には，オリジナルデータを利用している．

（1） わるいけど俺オタクじゃないからオタク用語使ってもわかんないからｗ

（2） 君，もっと頭使おうか．１０秒たってからじゃないと連続投稿できないんだよ．バカ

（3） お前らについていけない．名前出されたぐらいでビビッてるオタクどもめ

1，2，3 の発言はどれも，複数の人間による判断は通常発言と問題発言で割れている．ただし，1 は通常発言と判断する人が多く，2，3 の発言は問題発言と判断する人が多い．

4.4 実 験 結 果

実験サンプルに対する既存手法と改良手法の判断結果を表 2 に示す．

表 2 実験サンプルに対する判別結果

	既存手法		改良手法		人手
発言番号	単語	ペア	単語	ペア	人による判断
1	0.828	判別不可	0.886	4.46E-05	通常発言側
2	0.99	0.99	0.998	0.0095	問題発言側
3	0.865	0.999	0.967	0.43	問題発言側

発言 1 では,「ない」という単語が既存手法では取り除かれていたため，ペアが作成できず判別できなかった．また,「オタク」という言葉はあまりよい意味で使われないが，改良手法では,「オタク＋ない」というペアが作成できるため，単語のペアでも判別が可能となっている．また，単語ではかなり高い値で

あるがペアでは低い値であるため，この発言は条件付問題発言となる．実際に人による判断も，人により問題発言と判断されることがあった．よって，両方を使って正しい判断ができていると言える．

実験サンプル掲示板に対し既存手法と改良手法による 4 パターンの判別と人手による判断の比較検証を表 3 表 4 に示す．

表 3　パターン 1，2 の判別結果

	パターン 1		パターン 2	
手法	一致	不一致	一致	不一致
既存	36	56	497	57
改良	37	80	425	43

表 4　パターン 3，4 の判別結果

	パターン 1		パターン 2	
人手による評価	通常	問題	通常	問題
既存手法	35	3	137	74
改良手法	37	4	183	86

表 3 より，既存手法と改良手法でそれほど差異が見られなかった．今回は，助動詞に含まれる否定の意を持つものを加えたので，そのような助動詞を含む発言に対しては実験の通り有効になる結果が得られた．ただし，そのような発言が今回の実験対象にさほど含まれていなかったため全体的に差異が見られなかったと考えられる．

表 4 より，単語による評価が通常発言で単語のペアによる評価が問題発言となる場合，既存・改良手法ともに 9 割以上が通常発言と判断される結果となった．したがって，パターン 3 の判別結果が得られたときは，その発言は通常発言と判別することとする．また，逆に単語による評価が問題発言で単語のペアによる発言が通常発言と判別する場合，6 割が問題発言，4 割が通常発言と人間によって判別されていることから，今後このパターン 4 の場合を詳しく分析することが必要となる．

5. おわりに

本稿では，学校裏サイトにおける誹謗中傷発言を減らすために，発言が投稿されるまえにその発言が他のユーザにどう影響を与えるかを数値化し問題発言かどうかを自動的に判別することを目的とした．そのために，既存手法で含まれていない否定の意味をもつ助動詞を追加した．また，機械的に判断するために，単語と単語のペアの両方のデータを用いて，判別パターンの組み合わせによって発言を通常発言と問題発言に分類する手法を提案し実装を行い検証実験を行った．その結果，否定の意を持つ助動詞の追加により，既存手法では判別できなかった発言を判別可能とすることができた．また，判別パターンを組み合わせることにより，機械的に判別することが可能となった．

文　献

[1] 文部科学省, http://www.mext.go.jp/

[2] 学校裏サイトチェッカー，http://schecker.jp/

[3] 一藤 裕, 今野 将, 曽根 秀昭, “発言の長さに応じた電子掲示板における発言の評価方法に関する研究”, IEICE Technical report SITE, vol.108, No.75, pp.31-34, 2008.

[4] 大澤幸生, 松村真宏, 中村洋, “フレーミングは議論を阻害するか -2 ちゃんねるは何故面白い？”， IEICE technical report. IA, Vol.102, No.143, pp.55-60, 2002.

[5] 柴内康文, “言い争う—「フレーミング論争の検証」”， 現代のエスプリ, 川浦康至 (編), vol.370, 至文堂, 1998.

[6] 松村真宏，三浦麻子，柴内康文，大澤幸生，石塚満, “2 ちゃんねるが盛り上がるダイナミズム”，情報処理学会誌, vol.45, no.3, pp.1053-1061, 2004.

[7] 形態素解析エンジン “mecab”, http://mecab.sourceforge.net/