［論文］

ID交換掲示板における書きこみの隠語表記揺れを考慮した有害性評価

Method for Estimation of Harmfulness of ID-Exchange BBS Based on Lexical Jargonizations

安彦 智史†，長谷川 大‡，プタシンスキ ミハウ§，中村 健二ㇵ，佐久田 博司*
Satoshi ABIKO, Dai HASEGAWA, Michal PTASZYNSKI, Kenji NAKAMURA, and Hiroshi SAKUTA

†仁愛大学 人間学部コミュニケーション学科
‡東京工科大学 メディア学部
§北見工業大学 情報システム工学科
ㇵ 大阪経済大学 情報社会学部
＊青山学院大学 理工学部情報テクノロジー学科

† Department of Communication, Faculty of Human Studies, Jin-ai University
‡ School of Media Science, Tokyo University of Technology
§ Department of Computer Science, Kitami Institute of Technology
ㇵFaculty of Information Technology and Social Sciences, Osaka University of Economics
＊Department of Integrated Information Technology, College of Science and Engineering, Aoyama Gakuin

要旨

プライベートチャットアプリケーションの ID 交換を目的とした掲示板(ID 交換掲示板)において違法・有害な情報を含む書き込みが増加傾向にある．ID 交換掲示板では，多様な隠語表現を用いたやり取りが行われており意図的に崩された日本語が多く含まれるため，従来の手法では有害性評価を行うことが困難である．そこで本研究では，ID 交換掲示板における隠語表現を分類し，特に表層的な表記揺れが生じる環境下でも有害性判定を行える手法を検討する．

Abstract

Recently generic forum boards, such as Bulletin Board Systems (BBS) have experienced an increase of illegal and harmful activities, especially on BBS, which purpose is to exchange user contact IDs for further private chat applications (so called “ID-exchange BBS”). On such BBS, lexical transcription is often jargonized, or modified intentionally, making it difficult to extract information using standard tools. In this study, we first study the typology of harmful jargonized expressions on ID-exchange BBS. Based on this typology we propose a method dealing with the intentional transcription modifications on ID-exchange BBS in a robust way. In the evaluation, we developed a system applying the proposed method and verified the performance in estimating harmfulness of BBS entries. We also further improved the system by applying an automatic sentence pattern extraction method to separate harmful from non-harmful entries. The experiment confirmed that it was possible to eliminate most of the erroneous transcription modifications with the proposed method.

1. はじめに

ネットワークサービスの利用環境は著しい進展を見せている．その中で，出会い系サイトや誹謗・中傷などの書き込みを含む掲示板など，違法・有害な情報を含むサイトは増加傾向にあり，青少年の安全なインターネット利用は重要な課題[1][2][3]となっている．2013 年に警察庁が発表したサイバー犯罪の検挙状況によれば，検挙数の大半をネットワーク利用犯罪が占めており，2009 年（3,961 人）から 2013 年（6,655 人）にかけて約 1.7 倍へと増加を続けている．その内訳を見ると児童(18 歳未満)の関係する「児童買春・児童ポルノ法違反(児童ポルノ)」，「青少年保護育成条例違反」はいずれも増加の傾向にあり，2009 年（1,249 人）から 2013 年（2,306 人）までで約 1.8 倍となった．児童のネットワーク利用犯罪被害において，かつては悪意ある大人による誘い出しや児童買春の取引場は出会い系サイトが主となっていたが，2008 年の出会い系サイト規制法の法改正により事業者の実態把握促進や被害防止措置の義務化が行われた．これにより被害が SNS サイトな

どのコミュニティサイトへと移行している．特に最近では無料通話アプリや出会い系スマホアプリと呼ばれるスマートフォン用アプリケーションに起因する被害も数多く報告されている．スマートフォン用アプリケーション内で不健全な交流がされる場合には，これまでの出会い系サイトやコミュニティサイトとは異なり外部からアプリケーション内のやりとりを知ることが困難である．また児童にとってもより身近なデバイスを利用して直接的なやりとりが行われてしまうため犯罪の温床となっている．スマートフォン用アプリケーション内で交流が行われるにあたって事前に各ユーザが持つアカウントとそれぞれ紐付けされたIDを交換する必要があり，その交換の場は主に掲示板（以下，ID交換掲示板と表記）で行われている．

従来からあるWebサイトを対象に有害性評価を行う取り組みとしては，人手によるサイバーパトロールの他，既定のキーワードとの一致に基づく手法[4]と統計的アプローチに基づく手法[5]-[11]の2種類の研究が行われている．前者の既定のキーワードとの一致に基づく手法では，解析対象となる情報の中に，事前に作成した有害情報に関連するキーワードが出現するかどうかを判定し，その出現する割合に基づき情報の有害性を評価する．後者の有害情報の特徴に基づく手法では，SVM(Support Vector Machine)[12]やNative Bayes Classifier[13]などの識別機を用いて教師データから学習した有害情報の特徴に基づき，情報が有害かどうかを評価している．

しかし，ID交換掲示板には，従来掲示板やWebサイトとは異なる二つの特徴がある．一つ目は，記事がほぼすべて短文であり掲示板の中での会話の流れや文脈が無い点である．例えば，「条件有りで会える方　東京10代」など，ID交換掲示板では，具体的な交流を閉じた環境のアプリケーションで行うことを前提としているため，ほとんどの書き込みがアプリケーションの登録ID，目的，場所等の最小限の情報で構成されている．そのため，書き込みの相関関係は無く，書き込みの時系列を用いた有害性評価[14]を行うことは困難である．また，要件となる単語のみで構成された書き込みも多く，係り受け関係を考慮した手法では精度が低下すると考えられる．二つ目は，多様な隠語表現が存在する点である．ID交換掲示板では，従来から出会い系掲示版などで利用されていた「神待ち」や「割り切り」といった援助交際を示唆する隠語などの他に，伏字や当て字，意図的に単語を崩した表現を用いた様々な隠語表現が存在している．そのため，有害情報の特徴に基づく手法では形態素に分解することが難しく，正しい教師データの構築を行うことが困難である．また，キーワード一致ベースの手法では，上記した多様な隠語表現により，膨大な単語の登録と更新作業が必要となる問題がある．

現在行われているサイバーパトロールやこれらの研究により違法性が認められた場合は，違法情報として警察機関やIHC（インターネットホットラインセンター）等を通じてプロバイダに連絡が届き，該当サイトへアクセスできなくなるよう処理される．しかし，ID交換掲示板の場合は具体的な記載では無く，多様な隠語表記によるやり取りが行われているため，違法性評価が困難である．例えば「高校生の方，川崎で9月9日に援助交際できる方いませんか．3万払います．私は20代です」という書き込みがあれば，その書き込みがあったID交換掲示板自体が出会い系サイト禁止法違反となり，IHCに連絡してサイトの削除を要請することが可能である．また，書き込みを行った者も売春防止法により摘発される可能性がある．しかし，ID交換掲示板では，近い意味で用いられていると考えられる書き込みでも「JKさん、えん　川崎　20代」のように記載がされる．これらの隠語が含まれる書き込みは，限り無く違法に近いように解釈できるが厳密に違法情報と言いきれないため，違法情報と見做されず，「違法とは見做されないが青少年にとって有害である」という意味とされる有害情報と区分される場合がある．サイバーパトロールでは，パトロール実施者がID交換掲示板の書き込みを閲覧し，違法情報や有害情報，無害な情報を判断する必要があるが書き込みに多様な隠語表現が含まれている場合，パトロール実施者の知識や経験により判定精度が異なってくるといった問題が生じる．

そこで，本研究では，サイバーパトロールにおける有害性判定支援を目的として，ID交換掲示版のような多様な隠語表現が含まれており短文で文脈が無視された環境においても有害性を評価できる仕組みを検討する．本論文の構成として，まず2章では，ID交換掲示板の記事に含まれる隠語表現の調査を行う．そして，隠語表現を特徴ごとに分類し，それぞれの隠語表現の特徴について考察する．次に3章では，ID交換掲示板に含まれる記事の有害性評価手法を検討する．ID交換掲示板に含まれる記事の中には，隠語が含まれている

ために違法とは判定できないが，限りなく違法に近い書き込みが多数存在する．そこで本章では，ウェブサイトの違法有害判定に利用される各種法令を基に隠語が含まれる ID 交換掲示板の書き込みにおける有害性について定義し，隠語が含まれ，違法性が高いと思われる書き込みについても評価できる有害性の評価項目を策定する．これにより，隠語が含まれる書き込みであっても，書き込みに含まれる有害性を分類し，評価できる仕組みを検討する．4 章では，2 章と 3 章に示した ID 交換掲示板に含まれる隠語の特性と有害性評価項目を基に，隠語の表記揺れを解消し，有害性評価を行う手法を提案する．そして，本手法の有効性を評価するために実証実験を行い，適合率，再現率，F 値，正解率から結果を考察する．さらに，5 章と 6 章では，有害性評価の精度向上を目的として，4 章の表記揺れ解消手法により表記揺れが確認できた単語を辞書登録し，Ptaszynski ら[15]が提案している教師あり学習 SPEC を用いた有害判定を行い, SVM との比較結果を考察する．最後に，7 章では本研究の総括と今後の研究課題についてまとめる．

2. ID 交換掲示板の特性

2.1. ID 交換掲示板における隠語表現の調査・分類

本研究では，ID 交換掲示板に含まれる多様な隠語表現を分類するために，3 つの異なる ID 交換掲示板から 900 件の書き込みを収集し，記事に含まれる隠語の傾向と隠語が含まれる割合について調査した．尚，本研究における隠語の定義は，「特定の範囲内でのみ通用するような，一般とはかけ離れた意味をもたされた言葉」あるいは「伏字や当て字等を用いて表現した一般的な言葉」とする．本研究で分類した隠語の種類を表 1，書き込みに含まれる隠語の割合を図 1，実際の書き込み例を図 2 に示す．900 件の書き込みの内，隠語が含まれた書き込みは 38%であった．表 1 より，本研究では，ID 交換掲示板に使用されている隠語をさらに「短縮・意味」，「表層的言い換え」，「マスク」，「当て字」，「言外の情報」の 5 つに分類を行った．ID 交換掲示板に含まれる隠語はこの 5 つの内のいずれか，もしくは複数が含まれており，900 件の書き込みの中に「短縮・意味」が 235 件，「表層的言い換え」が 247 件，「マスク」が 54 件，「当て字」が 53 件，「言外の情報」が 77 件含まれていた．また，図 1 より，ID 交換掲示板に使用される隠語表現の割合は短縮・意味や表層的言い換えに分類されるものが多く，次いで言外の情報，マスク，当て字となった．

2.2. 各隠語表現の特性

本研究では，ID 交換掲示板の有害性評価を実現するために，2.1 節の分類結果より，それぞれの隠語の特性について考察する．

- 短縮・意味

短縮・意味に分類される隠語には，援助交際を「えん」のように短縮して表現した隠語や女子高生女子中学生を「jkjc」と表現するように単語の内容がわかる意味を持たせて省略させたものなどが含まれる．ID 交換掲示板では違法・有害性が含まれる書き込みの殆どが出会い系に関する書き込みであり，掲示板で使用されている短縮・意味に分類される隠語についても，出会い的に関連する単語が多い．例えば，2.1 節による調査において，短縮・意味に該当する書き込みは 235 件あったが，そのうちの五分の一にあたる 54 件が「さぽ，援，円，〇」などの援助交際を示す単語であった．

- 表層的言い換え

表層的言い換えは，「えっ，，，，ち，，い，，」や「大!人の関!係きぼう～」のように単語の間に記号や別の文字，空白等を挟んで表現した隠語である．また，「エ，，，，，ち，，，い」のように平仮名やカタカナ，小文字，全角半角を複合的に使って表現されるものある．表層的言い換えは，人手で確認すると意味がわかるが形態素解析のような機械的に単語を分割するような仕組みを適応することは困難である．表層的言い換えでは，書き込みにより間に含まれる記号の数を自由に変えられるため，隠語辞書等の固定の単語を登録する仕組みでは対応する難しい．ID 交換掲示板は，プライベートチャットアプリケーションの ID を公開し交換することを目的とした掲示板であるが，アプリケーション ID をそのまま公開すると，業者により自動的にアプリケーション ID が収集され，個人宛てにスパムが配信されることがある．そのため，このような表層的に文言を変更した表現が掲示板内に普及したと考えられる．

表 1　隠語表現の種類

	説明	具体例
短縮・意味	名詞や動詞を省略や言い換えをして表現	ぷちなど OK なんで相談下さい。
表層的言い換え	単語の間に記号などの文字列を含めて表現	えっ、、、、、、、ち、、、い
マスク	単語の一部を記号などに置き換えて表現	え。ち　え、、ち
当て字	単語の一部を類似する他の文字に置き換えて表現	女の子ばっかり <わえててずるい！
言外の情報	直接的な単語を用いずに特定の意味を表す表現	意味わかる人 条件あり

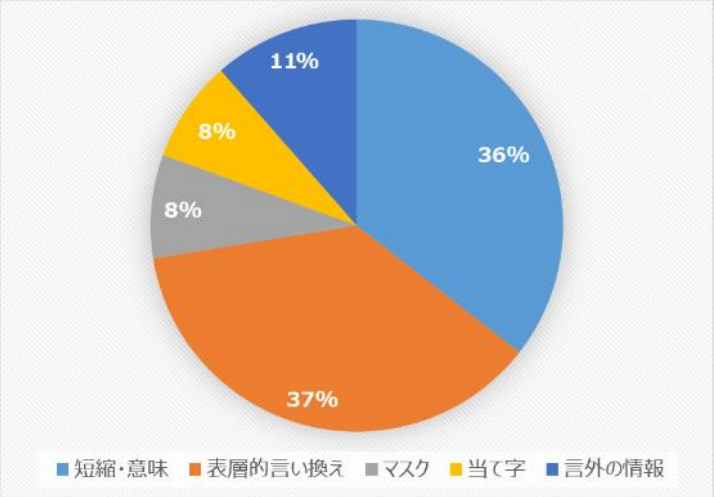

図 1　隠語表現の種類別割合

図 2　隠語表現が含まれる書き込み例

- マスク

マスクは，「動画〇影」のように単語の一部をマスキングした表現を含んだ隠語である．短縮・意味に分類されるものと同様に従来の掲示板から多く利用されている．ID 交換掲示板では，表層的言い換えと複合的に利用されることもある．

- 当て字

当て字では，カタカナの「チ」を漢字の「千（せん）」，平仮名の「く」を記号の「<(小なり)」，アルファベットの「H(エイチ)」を「Ｉ－Ｉ（アイ　ハイフン　アイ）」のように異なる文字で表現した隠語である．短縮・意味に分類される隠語と同様に，ID 交換掲示板では違法・有害性が含まれる書き込みの殆どが出会い

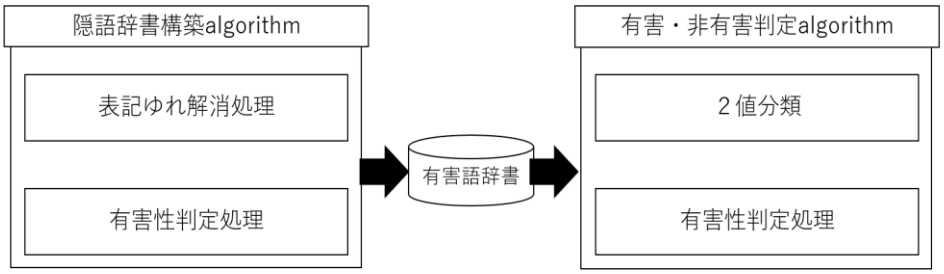

図 3　本研究の構成

表 2　本データセットの種類

No	データセット	使用箇所	節
1	900 件	隠語表現の調査	2.1 節
2	600 件	提案手法の辞書構築	4.3 節
3	300 件	正解データの作成, 実証実験	4.4 節,4.5 節,5.4 節

系に関する書き込みであり，使われる単語も出会い系やセクスティング（※性的なコンテンツを含むテキストメッセージ，sex（セックス）+texting（テキスト書き込み）から由来される．）に関連する単語がほとんどである．

- 言外の情報

言外の情報とは，「条件ありで会えるよ」，「意味分かる人きて」のように特別な隠語や援助交際を示唆する単語を使用していない表現である．これらの表現は通常の掲示板では特別な意味をもたない書き込みであるが，ID 交換掲示板に記載されている場合は援助交際を示唆していることが多い．

3.　研究の概要

3.1.　研究の目的と内容

本研究では，ID 交換掲示板を対象に，書き込みごとの有害性評価を行うことを目的とする．2 章の調査により ID 交換掲示板には様々な隠語表記が含まれていることを確認した．ID 交換掲示板の書き込みの中には，特に「表層的言い換え」や「短縮・意味」に該当する隠語表現の出現頻度が高い．しかし，1 章に挙げたように従来からある Web サイトを対象に有害性評価を行う取り組みの中では，これらの隠語表現に対する対応方法についての対策が示されているものは少ない．そこで，本研究では，ID 交換掲示板において発生しやすいと考えられる法令違反の抑止・取り締まりを目的として有害性を定義し，これらの隠語表現が含まれる環境下においても書き込みごとの有害性について評価する仕組みを検討する．

3.2.　研究の構成

本研究の構成を図 3 に示す．4 章では，ID 交換掲示板に含まれる隠語の表記揺れを解消する手法を提案し，本研究手法を用いたルールベースの辞書を構築することで有害性評価を行う．そして，5 章では，有害性評価の精度向上を目的として，4 章で構築した辞書を活用した教師あり分類の手法を提案し，6 章で手法の実証実験を行う．また，本研究では，異なる 3 つの ID 交換掲示板からランダムに収集したデータセットを複数用いる．本研究のデータセットを表 2 に示す．

3.3.　有害性の定義

電子掲示板に書き込みが行われた際に，プロバイダや掲示板オーナの対応ガイドラインとして利用されているものは，ホットライン運用ガイドライン検討協議会から「ホットライン運用ガイドライン」や電気電信事業者協会らから「インターネット上の違法な情報への対応に関するガイドライン」が策定されている．サ

イバーパトロールでは，これらのガイドラインに記載された情報をもとに違法，有害，無害の判定を行う．これらのガイドラインでは主に，a)売春防止法，b)出会い系サイト規制法，c)薬物関連法規，d)賃金業法関連法規，e)その他の法規に関連する違法性情報への対応へのガイドラインが示されている．本研究では，この中でも特に未成年者の被害が多数報告されている，a)，b)に関する有害性判定を対象とする．a)，b)に関連するガイドラインでは，売春防止法（第 5，6 条），出会い系サイト規制法（第 6 条）の違反の抑止・取締りを目的としており，これらの違反の構成要件として以下の 6 つの表現をあげている．

- 売春をうかがわせる表現（e.g.援交（えん、円、ぷち、さぽ）、条件あり、意味わかる人、など）
- 料金を示唆する表現（e.g.5 千、20k、苺（15）、2 で募集、など）対償を供与するまたは受けることを意味する表現（e.g.お小遣い、お礼、困ってる子、など）
- 児童を意味する表現（e.g.JS、JC、JK、J◯、15〜18、など）
- 性交または性交類似行為をもとめる表現（e.g.エッチしたい、セフレ募集、など）
- 交際を求める表現（e.g.会える人、絡みましょう、お茶しよう、一緒に遊んでくれませんか、など）

これらの表現のうち単一または複数の表現が含まれている場合には違法性が高いと判断される．また，現在インターネット上の違法・有害情報の通報受付窓口であるインターネットホットラインセンターでは，違法情報のほかにも青少年の健全育成を害する情報として，セクスティング等を含む書き込みも有害情報としてフィルタリング事業者等に提供している．そこで本研究では，セクスティングが含まれる書き込みを青少年に有害な表現と定義し，これら 7 つの表現が含まれていることを有害性の定義として用いる．

4. 隠語の表記揺れ解消

本章では，まず，2 章に示した隠語による表記揺れを解消することを目的に隠語を概念化し，表現を統一する手法を提案する．そして，表現を統一した書き込みに対して有害性評価を行うことで，隠語の表記揺れ解消処理の有効性を考察する．

4.1. 隠語の概念化

本研究で提案する隠語の表記揺れ解消手法の概要を図 4 に示す．本手法では，ID 交換掲示板に記載された書き込みに対して，5 つの処理を行うことにより発見した隠語を概念化し，有害性を評価する．まず，隠語の概念化処理を行うにあたり，ID 交換掲示板の書き込みに含まれる隠語表現から，当て字として使われる「<」や「苺」，「千」などを登録した当て字辞書，表層的言い換えに利用されている「,」や「・」，「/」などの記号や感嘆符，文字を登録した表層的言い換え辞書を構築した．当て字変換処理では，入力データに当て字が含まれていた際に当て字辞書を用いて当て字をひらがなに変換する．また，数字については全角と半角が混在しているため，全て半角に変換する．次に，表層的言い換え変換処理では，入力データに表層的言い換え辞書に該当する記号が連続して使用された場合に記号を除去した書き込みデータと記号を 1 つのみ残した複数のデータを作成する．複数のデータを作成する理由については，表層的言い換えだけでなくマスクや短縮に対応するためである．「エッチ」という単語を例に挙げると，表層的言い換えでは「えっ、、、、、、、ち」，マスクでは「エ◯チ」などのケースがある．この場合，単純に記号を除去するのみではマスクに対応することができない．そこで，書き込みの中に表層的言い換えが含まれていた場合，記号を全て除去した書き込みデータと記号を 1 つのみ残したデータを作成することで表層的言い換えとマスクのどちらの隠語表現であっても対応することが可能となる．そして，ひらがな化処理では入力された書き込みデータに含まれるすべての漢字と全角半角を含むカタカナをひらがな化する．ID 交換掲示板の記事では，漢字，ひらがなや半角カタカナ，全角カタカナや大文字小文字が入り混じって使用されることが多い．ここでは，MeCab のひらがな化機能を利用することで記事をひらがなに変換する．小文字変換機能ではひらがな化されたデータセットに含まれる小文字を全て大文字に統一を行う．最後に，エラーリストによる文字除去処理では，誤検出防止のためのエラーリストを作成し，特定の文字を除去した．援助交際を示唆する単語の中には，「えん」のように短い文字で構成されものが多い．そのため，「あまえんぼう」や「えんきょり」などの単語で誤検出となる可能性がある．エラーリストによる文字除去リストでは，このような誤検出となる単語を予め登録して除外した．尚，

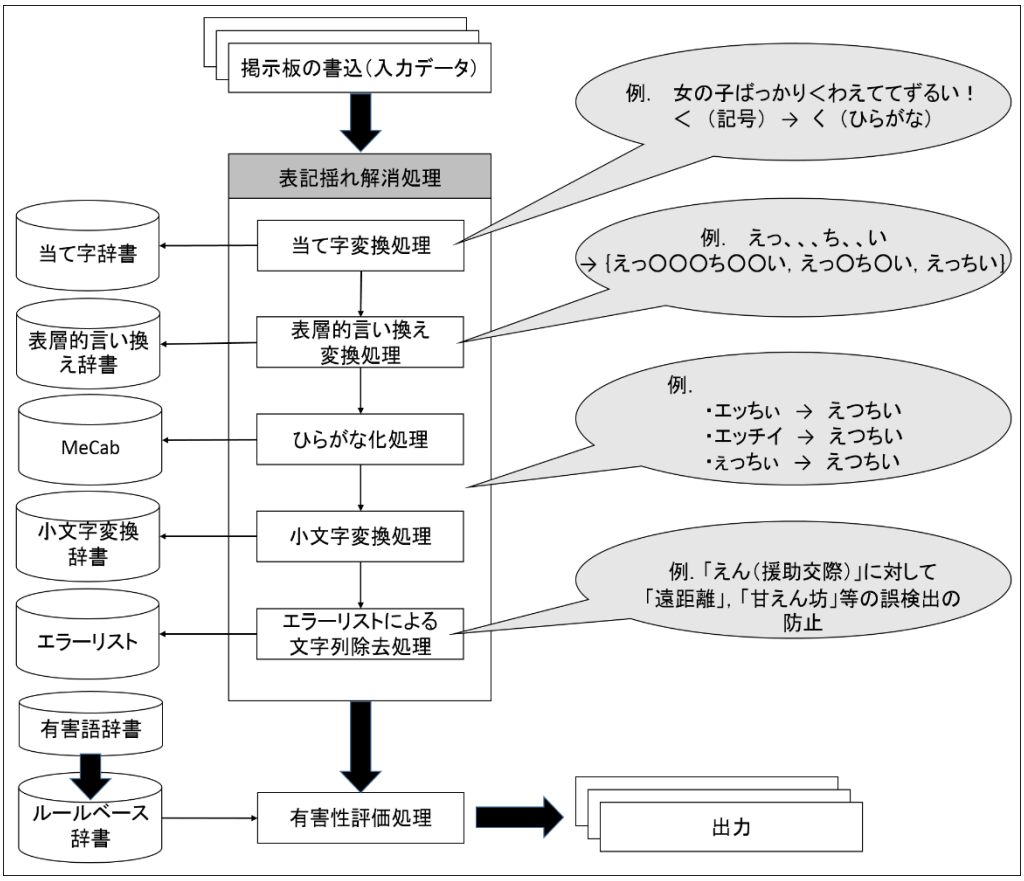

図 4　隠語の概念化

本提案手法における処理の順序について，記事内に当て字や表層的言い換えに該当する隠語表現が含まれていたると，正しくひらがな化処理を行うことができないため，当て字変換処理と表層的言い換え処理を最初に行う必要がある．そして，ひらがな化処理を行った後に小文字変換処理を行うことで，全ての漢字に含まれる小文字も大文字に変換することが可能である．本提案手法では，この 5 つの一連の処理を行うことによって，ID 交換掲示板上における表記揺れを含む書き込みの概念化を行った．

4.2. 有害性評価処理

有害な表現を有害語辞書に登録する．そして，有害な表現を全てひらがな化し，辞書に含まれる表現を一文字ずつマスク化することでルールベースの辞書を自動構築した．ルールベース辞書のイメージを表 3 に示す．最後に，隠語表記揺れ解消処理により出力されたデータに対して，ルールベース辞書を用いた分類を行う．これにより，ID 交換掲示板の記事ごとに，どのような有害性が含まれているか評価する．

4.3. 有害性評価

4 章で提案した隠語表記揺れ解消手法の有効性を検証するために，提案手法を実装したベースラインシステムを構築する．ベースラインシステムでは，まず，表 2 に記載した 600 件の書き込みに含まれる隠語表現から隠語表記揺れ解消処理に使う各辞書と有害性評価を行うためのルールベースの辞書を構築した．また，エラーリストに登録する文字列については，辞書構築時に収集した書き込みに対して全てひらがな化処理を行い，有害語性評価処理時に誤判定と判断した単語を全て登録した．それぞれの辞書に登録された件数を表 4 に示す．そして，隠語表記揺れ解消手法の処理を行ったデータセットと原文のデータセットを用意し，それぞれに対して，有害性項目の分類を行うことで，隠語表記揺れ解消手法の有効性を評価する．

表 3　ルールベース辞書のイメージ

有害性評価項目	有害な表現例	ルールベースの辞書例
売春をうかがわせる表現	援助	えんじょ
		○んじょ
		え○じょ
		えん○よ
		えんじ○

表 4　各辞書の登録件数

当て字辞書	表層的言い換え辞書	小文字変換辞書	エラーリスト	ルールベースの辞書
30 件	46 件	23 件	45 件	352 件

表 5　著者 3 名による有害表現の有無評価の一致率

	売春	料金	対償	児童	性交	交際	青少年に有害
正例の一致数 (3 名)	10	1	10	6	60	113	5
負例の一致数 (3 名)	272	299	279	293	199	97	219
一致率 (3 名)	94.0%	100.0%	96.3%	99.7%	86.3%	70.0%	74.7%
正例の一致数 (2 名)	15	1	18	7	84	161	28
負例の一致数 (2 名)	285	299	282	293	216	139	272
一致率 (2 名)	100%	100%	100%	100%	100%	100%	100%

4.4. 正解データの作成

表 2 に記載した 300 件の書き込みを対象に，著者 3 名が 7 種の各有害表現について，それが含まれているか否かの 2 値で注釈付けを行った（表 5）．さらに，3 名が注釈付けを行った各書き込みに対して，2 章で上述した隠語表現が含まれている否か 4 値で注釈付けを行った（表 6）．注釈付けの一貫性を調査するために，各有害性項目の 3 名および 2 名の評価結果が一致した書き込みの正例数，負例数および注釈が一致した割合を算出した．表 5 より，全体として 3 名の一致率および 2 名の一致率はともに高いことから，注釈付けにおいて個人間の解釈による差異は小さいといえる．ここでは，2 名が一致した注釈付け結果を正解データに採用した．また，表 6 より，有害・無害に関わらず隠語表現が混在している環境下で正解データを作成できていることを確認した．

4.5. 実験内容

表 2 に記載した書き込み 300 件を用いて，隠語表記揺れ解消手法の処理を行ったデータセットと原文のデータセットを用意し，それぞれに対して有害性項目の分類を行う．隠語表記揺れ解消手法の処理を行ったデータセットにはルールベースの辞書を用いる．そして，原文のデータセットにはルールベースの辞書を構築する課程で作成した概念化していない辞書を用いた．そして，書き込みに含まれていた単語が辞書に登録されていてればその単語の有害性評価項目を参照し，その書き込みの有害性項目とすることで分類する．最後に，4.4 節で作成した正解データと照会することで，適合率(Precision)と再現率(Recall)，F 値(F-score)，正解率(Accuracy)を算出し，これを比較した．

4.6. 結果および考察

表記揺れ解消処理前データセットの実験結果を表 7 に表記揺れ解消処理後データセットの実験結果を表 8 に示す．料金の項目において F 値は 0.003 低下し，他の項目においてはいずれも処理後データセットが処理前データセットを上回った．これは，ID 交換掲示板に利用される隠語表現は出会い系や猥褻な表現に起因する同義語が多く，本提案手法のように辞書の登録件数が少なくても表記揺れを解消することで精度があがった

表 6　有害表現と隠語表現を含む書き込みの分類

有害で隠語を含む	有害で隠語を含まない	無害で隠語を含む	無害で隠語を含まない
110 件	86 件	44 件	60 件
37%	29%	15%	20%

表 7　表記揺れ解消処理前データセットの実験結果

	売春	料金	対償	児童	性交	交際	青少年に有害	平均
Precision	0.600	0.022	0.875	1.000	0.875	0.807	0.227	0.629
Recall	0.400	1.000	0.389	0.714	0.167	0.673	0.179	0.503
F-score	0.480	0.043	0.538	0.833	0.280	0.734	0.200	0.444
Accuracy	0.957	0.850	0.960	0.993	0.760	0.737	0.867	0.875

表 8　表記揺れ解消処理後データセットの実験結果

	売春	料金	対償	児童	性交	交際	青少年に有害	平均
Precision	0.722	0.020	0.867	1.000	0.771	0.772	0.247	0.628
Recall	0.867	1.000	0.722	1.000	0.440	0.820	0.643	0.785
F-score	0.788	0.040	0.788	1.000	0.561	0.795	0.356	0.618
Accuracy	0.977	0.840	0.977	1.000	0.807	0.773	0.783	0.880

ためと考えられる．また，処理前データセットと処理後データセットにおける各項目の F 値の差については平均が+0.17 であり，これを上回った売春，対象，性交の 3 項目に分類される書き込みは表記揺れが起こりやすい傾向があると考えられる．本研究では，タスクの特性上，有害な書きこみを網羅的に発見できる方がよいため再現率が最も重要となる．再現率については，どの有害性項目においても本提案手法処理前に比べ，良好な精度を得られた．また，性交と青少年に有害の有害性項目において，再現率が他の項目よりも低い理由は 2 つ考えられる．1 つ目は他の項目と比較して隠語表現が多様である点である．2 つ目は有害表現に用いられる単語の品詞が多様である点である．1 つ目に関しては，今後辞書データを増やすことで対応可能だと考えられる．しかし，2 つ目に対応するためには，登録された隠語表現からルールベースの辞書を構築する際に隠語表現の活用語も登録するなどルールの追加が必要だと考えられる．次に，料金項目の適合率が低い理由として，処理前データセット，処理後データセットと共に数字を含んでいた場合に料金項目を付加する処理を行っていたためだと考えられる．料金と誤判定された書き込みには「20 代」などの年齢の表記が多く含まれており，これらの表記により適合率が低下したと考えられる．また，今回実験対象とした 300 件のデータには正解となるデータが 1 件しか含まれていなかったため，料金の項目に関してはデータ件数を増やして実験する必要がある．本実験から，ID 交換掲示板の書き込みには隠語の表記揺れが含まれており，これらの表記揺れが有害性評価に影響を与えていることが明らかになった．本実験は，有害な表現を登録した辞書とルールベースの辞書との比較を行っているが 1 章に挙げていた教師あり学習を用いて有害性評価を行っている手法についても，有害記事を学習する段階で記事に表記揺れが生じた場合，正しく形態素に分解することができない．そのため，隠語の表記揺れを考慮しなければ精度が低下すると考えられる．

5.　有害・非有害の判定手法による性能向上手法の提案

本章では，有害性項目の分類精度の向上を目的に，教師あり分類を用いた手法の提案を行う．まず，教師あり分類に用いる手法とその設定方法について説明する．そして，上述した表記揺れ解消手法を用いた前処理手法について説明する．

5.1.　SPEC 文パターン抽出手法の説明

書き込みを有害・非有害に分類するためには，まず Ptaszynski ら[15]によって提案された Language Combinatorics（言語の組み合わせ論）のアイデアを応用した．このアイデアでは，文章のような言語的実体は

要素（単語，句読点など）の繰り返し無しの順序付け組み合わせの集まりとして知覚できることを前提とされている．さらに，多くの異なった文章に現れる最も頻度の高い組み合わせは，文パターンとして定義されている．Ptaszynski らはこのアイデアを文中パターン自動抽出システム SPEC に応用した．SPEC は，SVM などをベースにした一般的な文書分類手法が高い結果を得られないタスクにて幅広く応用されている[16]．そのタスクとは，例えば，ネット上のいじめの検出など隠語や未知語など表記が変更されているタスク，つまり分類対象となる言語現象の複雑さ及び不規則性が極めて高いタスクである．SPEC が他の分類器と異なる点は，単語（BoW）や N グラムのみならず，文内のすべてのパターンを素性として使っていることであり，そのため，より広い範囲で分類を行うことができる．なお，本手法は，⑴ 素性（＝パターン）抽出，⑵ 重み計算，⑶ 分類，⑷ ヒューリスティックルールの適用と閾値調整という 4 つの部分から構成される．

5.1.1. 素性抽出

本手法では分類用の素性を以下の通りに抽出する．まず，訓練データとして要素（単語，句読点等）に分けられた文より繰り返し無しの順序付け組み合わせがすべて抽出される．各 n-要素の文より，k-組み合わせクラスタを抽出することができ，さらに 1≦k≦n となっており，各 k-要素の組み合わせクラスタは数式 1 のように 2 項係数となっている．この段階では，SPEC システムでは，{1,...,n}の間のすべての値の k クラスタの組み合わせを生成する．すべてのクラスタにおいて生成されるすべての組み合わせの数は，数式 2 にあたる．

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1}$$

$$\sum_{k=1}^{n} \binom{n}{k} = \frac{n!}{1!(n-1)!} + \frac{n!}{2!(n-2)!} + \ldots + \frac{n!}{n!(n-n)!} = 2^n - 1 \tag{2}$$

次に，すべての組み合わせでは，非後続の要素の間にはアスタリスク（“*”）が置かれ，頻出のパターン（出現＝2 回以上）のみが残される．このように生成されたパターンに対して重みが算出され，テストデータの分類用の素性として利用される．

5.1.2. 重みの計算

パターンの出現頻度をもとに，各パターンの正規化された重み wj を数式 3 の通り算出する．訓練データのポジティヴの重み Opos を訓練データ内総合重み Opos+Oneg で割ったうえ，算出結果を+1 と−1 の範囲に合わせるために結果から 0.5 を引きそして 2 でかけるというカスタムなシグモイド関数を利用する．さらに重みは次のように変更することができる．まず，重み計算には，パターンの長さ k（要素の数）とパターンの出現頻度 O という各パターンにある 2 つの特徴が重要となる．あるパターンはあるデータセットをどの程度代表しているかはその 2 つの特徴によって定義付ける．なお，重みは，以下の通り変更できる．

・パターンの重みを長さ k でかける（以降 length awarded，略 LA）
・パターンの重みを長さ k と出現頻度でかける（以降 length and occurrence awarded，略 LOA）

5.1.3. 分類

分類では，テストデータの各文に対して，その文でマッチングされたすべてのパターンの重みの合計から分類スコアを数式 4 の通り算出する．

$$w_j = \left(\frac{O_{pos}}{O_{pos} + O_{neg}} - 0.5\right) * 2 \tag{3}$$

$$score = \sum w_j, (1 \geq w_j \geq -1) \tag{4}$$

5.1.4. ヒューリスティックスルールの適用と閾値調整

上述の 5.1.2 項のように生成されたパターンのリストを次のようなヒューリスティックルールによって変更することができる．ある反対特徴（ポジティブ・ネガティブ，または有害・非有害）を持った文のコレクションからパターンを抽出すると，そのパターンリストには必ず 3 種類のパターンが確認できる．

・片方にしか現れないユニークなパターンと，
・両方に同じ頻度で現れ重みが 0 となるパターン（以降“0 パターン”）
・頻度が同じではないが両方に現れるパターン（以降“曖昧パターン”）
なお，分類用のパターンリストはこのように変更できる．
・すべてのパターンを利用する（以降 all patterns，略 ALL）
・すべての曖昧パターンを削除する（以降 ambiguous patterns deleted，略 AMB）
・0 パターンのみ削除する（以降 zero-patterns deleted，略 0P）

また，上記のように生成されたパターンには，要素分割パターンと要素連続パターン（N グラム）が必ず出現する．文書分類の課題では，N グラムの用いることは一般的なため，実験は，要素分割パターン（patterns，略 PAT）の他に N グラムのみ（ngrams，略 NGR）という設定でも行っている．最後に，訓練データにおける偏りを考慮しなければならない．両方の訓練データは完全に同じではない限り，どの訓練データにも必ず片方への偏りが起こる．その偏りのもとには，片方に文数が多かったことと，文そのものが平均的に長かったことなどがある．偏りが起きるため，分類では，スコアの閾値を，例えば，0 で固定させると最適な結果より低い結果が出ることが予想される．なお，スコアの偏りを解消するためには，閾値最適化を行っている．結果は，閾値の各段階において，適合率（Precision），再現率（Recall），標準 F 値（balanced F-score），正解率（Accuracy）で算出する．

5.1.5. SPEC の応用における従来研究

近年，SPEC は様々なテキスト分類課題に応用されている[18]．以下にその一部を要約する．まず，Ptaszynski ら[19]は，感情文と非感情文の分類に SPEC を用いている．SPEC による文から感情的パターンを抽出する手法は，SVM などの一般的な分類器より大幅に結果を改良できた．
また，Ptaszynski ら[17]はネットいじめを検出するためのテキストパターンを抽出する問題にも SPEC を適用している．ここでは，掲示板への書き込みを収集したデータセットにおいていじめに該当する書き込みの判別を試みており，SPEC が既存のネットいじめ検出手法と比較して精度が向上したことを確認した．
また別の研究では，Nakajima ら[20]が，トレンド予測のために未来に関する表現の分析において SPEC を応用した．この実験では，未来を表す文は少数のパターンが頻出し，現在や過去を表すパターンは数が多く，出現頻度が分散する傾向にあることを確認した．

5.2. データセットの前処理

SPEC を用いた実験を行う前に，書き込みの前処理を行う必要があった．SPEC は要素（単語・形態素など）に分割された入力文をもとに訓練・分類を行っているため，訓練データ・テストデータを要素に分割する必要があった．まず，有害情報と無害情報の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．形態素とは，それ以上分解したら意味をなさなくなるところまで分割して抽出された最小の文字列を指す．掲示板の書込みに対し形態素解析を行うため，形態素解析に MeCab(日本語辞書 ipadic)を利用した．形態素解析の解析例を以下の図 5 に示す．

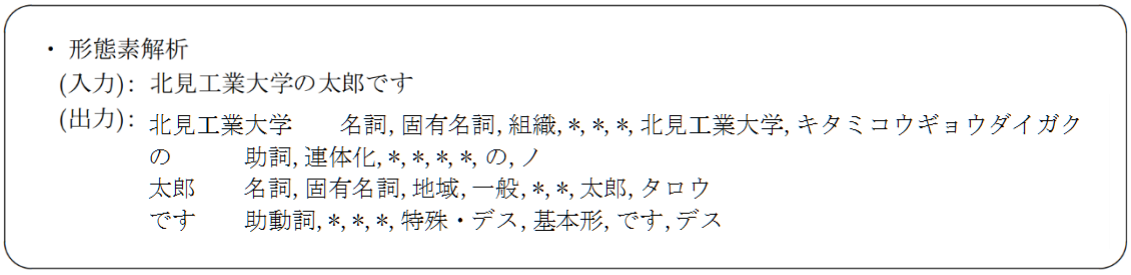

図 5　形態素解析の解析例

MeCab 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，表記の揺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

保つために上述した表記揺れ解消手法を用い，データセット内の掲示板に頻出する有害語のリストを MeCab 用のカスタムユーザ辞書への登録を行った．次節で処理について詳述する．

5.3. 有害単語の辞書登録

有害情報を構成する 1 つの要素として「絵ッ血」や「え・ち」などの特有の単語(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おらず，未知語判定，もしくは解析失敗をしてしまう可能性が高い．しかし，有害単語は直接誹謗中傷単語も多く，書込みに含まれるだけで有害情報となりうる重要な要素である．これらの単語を解析失敗してしまうと，有害情報と無害情報の分類は著しく困難となる．そこで，有害単語を表記揺れ解消手法を用いて収集して整理し，あらかじめ形態素辞書の登録した．以下に作業の手順を示す．

（1） (データセットを表記揺れ解消手法にて処理し，表記揺れが確認された単語を抽出する．
（2） 抽出した単語の表記揺れの形を見出し語として用いて，もととなった形態素の形態素情報を追加し，ユーザ辞書に登録する．
（3） ユーザ辞書を用いて形態素解析されたデータセットの再解析を行い，形態素に分割する．

上記の手順に従って，有害単語の辞書登録及びデータセットの前処理を行った．登録した有害単語は，解析器の辞書登録定義に従って，見出し語，品詞情報，優先コスト，読み，発音，活用語の情報を付与している．優先コストは，値が低ければ低いほど，その単語から優先して形態素解析を行う．有害単語は，重要な有害情報構成要素なので取りこぼしは避けたい．よって，最優先で解析されるように，他の全ての単語より低い 300 という値から，文字列のバイト数を引くことによって，最長一致法による最優先解析をするようにした．

6. 評価実験

6.1. 有害・非有害判定結果

実験結果としては，まず，10 分割交差検定では有害・非有害の判定手法について性能を確認した．その後，一番高い性能を確認した手法のバージョンを用いて，ID 交換掲示板の詳細分類手法の前段階にて改良する方法として応用し，最終結果を確認した．

有害・非有害の判定は手法のどのバージョンが一番高い性能を得られたかを確認するのには，いくつかの評価基準を用いた．まず，適合率，再現率，F 値，正解度の一番高い結果を得られたバージョン．そしてそれが複数の場合, 閾値の最適化を用いて算出した最適なバージョンを oversampling 無しの条件と oversampling 有の条件で確認した．最高適合率は，oversampling 無しの場合，P=0.934(NGR-LA，NGR-LA-0P，R=0.813,F=0.870,A=0.776)，oversampling 有りの場合，P=1.000(NGR-LA-0P,R=0.99,F=1.00,A=1.00)となった．最高の再現率は，OS 有の場合はどの条件でも R=1.000(P=0.917,F=0.957,A=0.917)となり，OS 無の場合は NGR と NGR-LA のみの場合 R=1.000(P=0.95,F=0.97,A=0.97)となったが，そのほかの条件もほとんど 1.000 に近い結果を得られた. 最も高かった F 値は, OS 無の場合は F=0.958(NGR,NGR-LA,NGR-LA-0P,P=0.919,R=1.000,A=0.920)となり，OS 有の場合は F=1.000(NGR-LA-0P,P=1.000,R=0.99,A=1.000)となった．最も高かった正解率は，OS 無の場合はすべての条件では Acc=0.917(P=0.917,R=1.000,F=0.957)となり，OS 無の場合は，Acc=1.000(NGR-LA-0P,P=1.000,R=0.99,F=1.000)となった．比較のため，同じデータセットを文書分類のベースラインと考えられるサポートベクターマシン（SVM）を用い，Bag-of-Words（BOW）モデルで訓練を行った．モデルの重み設定には，訓練データ内の単語の出現回数（occurrence），出現頻度（tf）及び出現頻度×逆文書頻度（tf-idf）と 3 種類の重みを用いた．SVM による分類結果は表 9 に示す．重み計算として単語の出現回数を用いた場合，oversampling の有無に限らず結果は比較的に低かった．特に oversampling 有の場合，Precision が 74.7%と一番高かった反面 Recall が 23.4%と極めて低く F 値は 35.6%となった．tf 及び tf-idf の間については変更が見られず，どれも同じ結果となった．特に Recall においての改善が見られ，どれも 100%となった．

表 9　SVM を用いた分類結果(Bag-of-Words, 10 分割交差検定の平均)

	oversampling 無し				oversampling 有			
重み設定	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score	Accuracy
occurrence	0.679	0.984	0.804	0.672	0.747	0.234	0.356	0.556
tf	0.682	1.000	0.811	0.682	0.526	1.000	0.689	0.526
tf-idf	0.682	1.000	0.811	0.682	0.526	1.000	0.689	0.526

表 10　OS 無の場合の SPEC の結果と SVM の結果

	oversampling 無し												
	Highest F-score w/ threshold				Highest Precision w/ threshold				Highest Accuracy w/ threshold				BEP
	Pr	Re	F1	Acc	Pr	Re	F1	Acc	Pr	Re	F1	Acc	(P=R=F)
PAT-ALL	0.917	1.000	0.957	0.917	0.921	0.965	0.943	0.892	0.917	1.000	0.957	0.917	0.921
PAT-0P	0.917	1.000	0.957	0.917	0.922	0.977	0.949	0.903	0.917	1.000	0.957	0.917	0.922
PAT-AMB	0.917	1.000	0.957	0.917	0.923	0.986	0.953	0.911	0.917	1.000	0.957	0.917	0.921
PAT-LA	0.917	1.000	0.957	0.917	0.926	0.979	0.952	0.909	0.917	1.000	0.957	0.917	-
PAT-LA-0P	0.917	1.000	0.957	0.917	0.926	0.979	0.952	0.909	0.917	1.000	0.957	0.917	-
PAT-LA-AMB	0.917	1.000	0.957	0.917	0.917	1.000	0.957	0.917	0.917	1.000	0.957	0.917	-
NGR-ALL	0.917	1.000	0.957	0.917	0.923	0.790	0.851	0.745	0.917	1.000	0.957	0.917	0.922
NGR-0P	0.917	1.000	0.957	0.917	0.922	0.984	0.952	0.909	0.917	1.000	0.957	0.917	0.922
NGR-AMB	0.917	1.000	0.957	0.917	0.920	0.875	0.897	0.816	0.917	1.000	0.957	0.917	0.919
NGR-LA	**0.917**	**1.000**	**0.957**	**0.917**	**0.934**	**0.813**	**0.870**	**0.776**	**0.919**	**1.000**	**0.958**	**0.920**	**0.924**
NGR-LA-0P	0.917	1.000	0.957	0.917	**0.934**	**0.813**	**0.870**	**0.776**	**0.919**	**1.000**	**0.958**	**0.920**	0.923
NGR-LA-AMB	0.917	1.000	0.957	0.917	0.921	0.965	0.943	0.892	0.917	1.000	0.957	0.917	-

表 11　OS 有の場合の SPEC の結果と SVM の結果

	oversampling 有り												
	Highest F-score w/threshold				Highest Precision w/threshold				Highest Accuracy w/ threshold				BEP
	Pr	Re	F1	Acc	Pr	Re	F1	Acc	Pr	Re	F1	Acc	(P=R=F)
PAT-ALL	1.000	0.970	0.980	0.980	1.000	0.970	0.980	0.980	1.000	0.970	0.980	0.980	0.970
PAT-0P	1.000	0.970	0.980	0.980	1.000	0.970	0.980	0.980	1.000	0.970	0.980	0.980	0.970
PAT-AMB	1.000	0.970	0.980	0.980	1.000	0.970	0.980	0.980	1.000	0.970	0.980	0.980	0.973
PAT-LA	1.000	0.970	0.980	0.980	1.000	0.970	0.980	0.980	1.000	0.970	0.980	0.980	0.974
PAT-LA-0P	1.000	0.970	0.990	0.990	1.000	0.970	0.990	0.990	1.000	0.970	0.990	0.990	0.972
PAT-LA-AMB	1.000	0.970	0.980	0.980	1.000	0.970	0.980	0.980	1.000	0.970	0.980	0.980	0.970
NGR-ALL	1.000	0.970	0.990	0.990	1.000	0.970	0.990	0.990	1.000	0.970	0.990	0.990	0.981
NGR-0P	1.000	0.970	0.990	0.990	1.000	0.970	0.990	0.990	1.000	0.970	0.990	0.990	0.982
NGR-AMB	0.990	0.980	0.980	0.980	1.000	0.950	0.980	0.980	0.990	0.980	0.980	0.980	0.979
NGR-LA	**1.000**	**0.990**	**0.990**	**0.990**	**1.000**	**0.990**	**0.990**	**0.990**	**1.000**	**0.990**	**0.990**	**0.990**	**0.993**
NGR-LA-0P	**1.000**	**0.990**	**1.000**	**1.000**	1.000	0.990	0.990	0.990	**1.000**	**0.990**	**1.000**	**1.000**	**0.991**
NGR-LA-AMB	1.000	0.980	0.990	0.990	1.000	0.980	0.990	0.990	1.000	0.980	0.990	0.990	0.981

表 12　F 値が最も高かった 2 つの手法の比較

	NGR-LA			
	OS 無し		OS 有り	
	F-score	Accuracy	F-score	Accuracy
NGR-LA-0P	0.3293	0.3293	0.0565	0.0829
	p>0.05	p>0.05	p>0.05	p>0.05

しかし，SPEC の結果と違って SVM は oversampling 無しの場合の方が結果が高く，F 値が 81.1%，Precision と Accuracy が同じく 68.2%となった．次に，ブレイクイブンポイント（BEP）を算出した．ブレイブレイクイブンポイント(BEP)とは，適合率と再現率を同じ値とした場合の精度である．BEP を用いることでバランスを取った分類条件を算出することが可能である．SPEC（OS 無）の場合は，NGR-LA の条件は BEP＝0.924 となり，SPEC（OS 有）は，NGR-LA が 0.993 その次に高い結果は NGR-LA-0P が 0.991 となった．結果は表 10 と表 11 に示している．すべての結果から以下の考察を行うことができる．SVM よりも SPEC を用いた手法の方が結果は高かった. 特に NGR-LA 及び NGR-LA-0P の 2 つの条件がどの評価基準でも高い結果が得られた. そこで，2 つの間の相違は統計的に有意差を持っているかについて検証するためにＴ検定を行った．Ｔ検定の結果は表 12 に示す．結果から 2 つの間には有意な差は確認されなかった.

その意味は，NGR-LA の方が若干高い結果となったが，どれも信頼度の高い結果であると解釈できる．最後に，OS 有場の合は結果が低くなると予想していたが，実際には高かったことが確認できた．その理由として

表 13　一般分類器による分類結果．交差検定は 10 分割及び 3 分割，
最高点は太字．結果は有害／非有害を個別に分類した平均値．

10 分割交差検定	OS 無し				OS 有			
	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score	Accuracy
Naïve Bayes	0.845	0.827	0.835	0.827	0.851	0.849	0.848	0.849
kNN (k=1)	0.826	0.713	0.754	0.713	0.857	0.799	0.790	0.799
kNN (k=2)	0.837	0.743	0.777	0.743	0.773	0.663	0.625	0.663
RIPPER	0.757	0.87	0.810	0.870	0.886	0.852	0.849	0.852
C4.5	0.757	0.867	0.808	0.867	0.942	0.935	0.935	0.935
Random Forest	0.892	0.877	0.825	0.877	**0.968**	**0.966**	**0.965**	**0.966**
3 分割交差検定	OS 無し				OS 有			
	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score	Accuracy
Naïve Bayes	0.844	0.833	0.838	0.833	0.872	0.868	0.867	0.868
kNN (k=1)	0.814	0.713	0.752	0.713	0.829	0.739	0.720	0.739
kNN (k=2)	0.817	0.777	0.794	0.777	0.789	0.634	0.578	0.634
RIPPER	0.757	0.867	0.808	0.867	0.866	0.816	0.810	0.816
C4.5	0.757	0.867	0.808	0.867	0.931	0.92	0.919	0.920
Random Forest	0.892	0.877	0.825	0.877	**0.961**	**0.958**	**0.958**	**0.958**

表 14　提案手法(ベースライン)と提案手法を SPEC で改善させた分類の最終結果

	提案手法（ベースライン）								提案手法（SPEC による改良）							
	売春	料金	対償	児童	性交	交際	青少年に有害	平均	売春	料金	対償	児童	性交	交際	青少年に有害	平均
Precision	0.722	0.020	0.867	1.000	0.771	0.772	0.247	0.628	0.722	0.022	0.929	1.000	0.783	0.783	0.277	0.645
Recall	0.867	1.000	0.722	1.000	0.440	0.820	0.643	0.785	0.867	1.000	0.813	1.000	0.456	0.890	0.643	0.810
F-score	0.788	0.040	0.788	1.000	0.561	0.795	0.356	0.618	0.788	0.043	0.867	1.000	0.576	0.833	0.387	0.642
Accuracy	0.977	0.840	0.977	1.000	0.807	0.773	0.783	0.880	0.977	0.853	0.987	1.000	0.823	0.827	0.810	0.897

表 15　Paired Student T-test の結果(SPEC 無し　VS　SPEC 有り)

Precision	p=0.0980
Recall	p=0.1354
F-score	p=0.0698
Accuracy	**p=0.497**

は，強制的に増えたとしても，ネガティブな訓練データ（非有害）が増え，そこから抽出されたパターンに重みがかかり，OS 無の場合で間違って有害と判断されていたものは正しく非有害と判断された．

6.2. 一般分類との比較

最後に，分類結果の傾向をさらに確認するために SVM 以外の一般分類器との比較を行った．比較に用いたのは，従来研究[21]にてネットいじめの自動検出のために用いられた分類器のすべてを用いた．それらの分類器を以下の通りに説明する．分類結果は表 13 に示す．NaïveBayes は素性間の強い（単純な）独立性を前提にありベイズの定理に基づいた確率的分類器である．kNN（k 近傍法）は，特徴空間内に入力事項に最も近い例に基づいた分類器. 本研究では, 比較のために k=1 と k=2 のパラメータを用いた. RIPPER（Repeated Incremental Pruning to Produce Error Reduction）アルゴリズムは段階的に分類ルールを学習し最適化する．ノイズの多いテキストの分類において効果的である．C4.5 は決定木生成アルゴリズムであり，最初はラベル付きのデータセットから決定木を生成し，各木のノードにおいて決定を行うために最適な分割を行う．RandomForest は集団学習アルゴリズムであり，ランダムサンプリングされた訓練データから学習した多数の決定木を使用する．

多くの分類器（特に決定木ベース）が SVM のベースラインを超え，本研究で扱っているようなノイズの多

いデータに対して SVM が一般分類器と同様あるいはそれ以下の性能を示した．また，元のデータは比較的に少量であったため，10 分割交差検定の他に 3 分割交差検定を行い，分類器間の結果の傾向を調査した．3 分割交差検定の場合の結果は全体的に低かったが，結果の間には類似した傾向がみられ，決定木ベースの分類器の結果及び NaiveBayes が高く，kNN は低くなった．また，Oversampling はほとんどの場合プラスに働き，RandomForest が一番高い結果（F=0.965）得られたが，提案手法には及ばなかった．

6.3. ID 交換掲示板詳細分類への応用

評価実験で一番結果が高かったのは，N グラムのパターンリストでは重み計算にパターンの長さ導入した設定になった（NGR-LA）であった．場合によって 0 パターンを削除した設定（NGR-LA-0P）の方が高い結果を得られていたが，その二つの設定の間には有意差は得られなかったため，有害性の詳細分類への応用実験では前者を利用することにした．さらに，結果は閾値が-0.5 の場合は一番高かったため，詳細分類実験のときもこの閾値を用いた．実験結果は表 14 に示している．結果が 100%になっている場合以外，SPEC による前段階有害性判定を用いたほうが，良好な結果を得られた．さらに改良前と改良後の結果の有意性を求めた結果，適合率，再現率，F 値は統計的有意に近いが有意ではない（not quite statistically significant）という結果が出たが，正解率の場合は 5%程度での有意性は得られた．有意性検定の結果は表 15 に示している．結果の詳細分析をすると，SPEC の方では非有害を有害に誤って分類したケースはなかった．一方，有害性を持った書き込みを非有害と判断されたケースは3つあった. 3つとも以下のように得られたスコアとともに表している．

スコア 例文
-0.501 彼氏いない女の子いないかないたら絡もう通話しよ
-0.761 あと、若い子でし○ぎくれる子いないかな？
-1.520 クリスマス★コスどうかなあ？ちょっとハデかなあ～？色々写メあるよん♪

この中でスコアが一番高い例はほとんど閾値と重なっておりボーダーラインケースとなった．特に「絡もう」とか「通話」という表現がスコアに貢献している．しかし，「彼氏」，「女の子」，「いないかな」などのような表現は非有害の書き込みにも表れやすいのでボーダーラインケースになったことが考えられる．また，有害なのに，非有害と判断された残りの２つのケースについては，前述と同じく「いないかな」の他，「若い子」，「クリスマス」などのような投稿者のプライバシーにかかわりやすいものは，犯罪や法律違反にかかわる書き込みではなく，一般的な出会い（友達，話し相手探しなど），で使われやすいため，非有害という判断になった．また，非有害の書き込みの方が特殊文字（★，♪）が出現しやすいためスコアをさらに下げる効果があった．最後に，辞書に登録されていなかった「し○ぎ」という単語も有害語として正しく検出されなかったこともスコアに影響を与えたことが考えられる．さらに，頻出する有害パターン・非有害パターンの詳細分析を行った．非有害の場合は，例えば「ちょっと」，「よろ」（よろしくの省略），「ーー」（音の延長）や「泣」というのがよく見られた．とくに「泣」は，有害ではなく正直な個人的な悩みの告白や，軽い会話交換において顔文字の働きを果たしている．また，顔文字も頻繁に検出されていた．有害の頻出パターンには，以下のようなパターンが頻繁に見られた．一番よく見られたのは，「LINE」やその工夫して変更した表記（「ライン」，「L1NE」など）があったが，辞書登録がされており正確に検出することができた．また，書き込みが投稿された法律違反らしき目的を直接表す表現も，トピックとして検出されたケースも頻繁であった．その一番目立つグループとして以下のようなものがあげられる．

・【「女」，「募集」】
・【「気軽に」，「連絡」，「絡んで」】
・【「ID」，「追加」】
・【「彼女」，「探し」】
・【「相手」，「募集中」】
・【「通話」・「電話」，「しよ」】

更に，文字通りにセクスティングの書き込みもあり，「エッチ」，「オナニー」，「セフレ」などのような表現が頻繁に見られた．しかし，直接意味を表す言葉の他に，「話し」，「てくれる」，「[で]ください」，「たり[～たり]」，「(笑)」など文法のレベルで特徴を表す表現も見られた．上記のような直接な書き込みもあるが，特に隠語で書かれコンテンツがマスクされている有害書き込みは一番検出しにくいと想定されている．このような文法的な特徴は，変化する隠語表現を含む有害な書き込みを検出する方法として利用可能である．

7. おわりに

本研究では，ID 交換掲示版のような多様な隠語表現が含まれ，短文で文脈が無視された環境においても有害性を評価できる仕組みを検討した．まず，ID 交換掲示板を対象に多様な隠語表現を解析し，隠語の分類を行った．そして，分類結果を基に隠語表記揺れ解消手法を考案し，手法を実装したベースラインシステムの構築を行った．ベースラインシステムの実験では，本システムを実装したデータセットと原文のデータセットに対して有害性評価を行い，正解データと比較を行うことで本手法の有効性を示した．このことから，サイバーパトロールの実施者が隠語の意味を正しく解釈できないような場面において，本システムを利用することで，有害情報の検出精度向上に寄与できると考えられる．さらに，本システムの有用性について確認するため，福井県警に本成果の利用用途や目的についてヒアリングを行った．その結果，有用であるとの評価を頂いたため，福井県警からの委嘱を受け，引き続きサイバーパトロールでの実施検証を行う予定である．次に，ベースラインシステムの精度を向上することを目的に，文パターン自動抽出手法 SPEC を用いた２項分類と有害性項目の詳細分類を行った．２項分類では，書き込み文章における有害・非有害の判定を行い，十分割交差検定により精度を確認した．そして，詳細分類では，SPEC を用いた提案手法と SPEC を用いないベースラインシステムを比較し，SPEC を用いた提案手法の有用性を示した．全体的には，今回は十分な結果を得られたが，データセットが比較的に小規模（300 件）であったため，評価実験では正解率（Accuracy）の統計的有意差は得られたが, F 値や本研究で重要とする再現率の有意差は得られなかった. 将来の課題としては，まずデータセットの拡張を行う予定である．例えば，ネットいじめを追及している研究[17]では，BBS 書き込みを 3,000 件のデータセットを応用することが妥当であると強調されており，今後このようにデータセットの拡張を行う．拡張方法としては，本研究で作成したシステムを利用し，ID 交換掲示板をクローリングしブートストラッピング手法を用いてデータセットの自動的，または半自動的な拡張を試みたい．

謝辞

本研究は JSPS 科研費 15K16543 の助成を受けたものです.

参考文献

[1] 小針誠 (2009). 学校裏サイトにおける「ネットいじめ」の構造と対策. 児童心理, 金子書房, 63(13), pp. 94-99.
[2] 桑子博行 (2011). 我が国における児童ポルノのブロッキングの仕組みと今後の展望 (特集インターネット上に氾濫する違法・有害情報の現状と対策). 警察学論集, 64 (8), pp. 67-93.
[3] 苗村憲司 (2011). サイバー防犯ボランティアの意義と育成支援方策について (特集インターネット上に氾濫する違法・有害情報の現状と対策). 警察学論集, 64 (8), pp. 51-66.
[4] Lee, W., Lee, S. S., Chung, S., and An, D. (2007). “Harmful Contents Classification Using the Harmful Word Filtering and SVM.” Proceedings of the 7th International Conference on Computational Science, 4489, pp. 18–25.
[5] Guermazi, R., Hammami, M., and Hamadou, A. (2007). “Combining Classifiers for Web Violent Content Detection and Filtering.” Proceedings of the 7th International Conference on Computational Science, pp. 773–780.
[6] Lee, P. Y., Hui, S. C., and Fong, A. C. M. (2002). “Neural Networks for Web Content Filtering.”IEEE Intelligent Systems, 17 (5), pp. 48–57.
[7] Du, R., S.-N. R. and Susilo, W. (2003). “Web Filtering Using Text Classification.” The 11th IEEE International Conference on Networks, 11, pp. 325–330.
[8] Chandrinos, K. V., Androutsopoulos, I., Paliouras, G., and Spyropoulos, C. D. (2000). “Research and Advanced Technology for Digital Libraries: 4th European Conference, ECDL2000 Lisbon, Portugal, September 18–20, 2000 Proceedings.”, pp. 403–406.

[9] 菊池琢弥, 内海彰 (2010). 語の共起情報に基づく有害サイトフィルタリング手法. 第 9 回情報科学技術フォーラム講演論文集, FIT2010 (2), pp. 1‐6.
[10] 池田和史，柳原正，松本一則，滝嶋康弘 (2010). HTML 要素に着目した違法・有害サイト検出手法の提案と評価. 第 9 回情報科学技術フォーラム講演論文集, FIT2010 (2), pp. 7‐12.
[11] 松葉達明, 里見尚宏, 桝井文人, 河合敦夫, 井須尚紀 (2009). 学校非公式サイトにおける有害情報検出. 言語理解とコミュニケーション研究会技術研究報告, 109 (142), pp. 93‐98.
[12] 中村健二，田中成典，大谷和史，山本雄平 (2009). セキュアライフの創出を目指した安全知の獲得に関する研究―電子掲示板からの犯行予告の抽出―. 土木情報利用技術論文集, 18,pp. 269‐280.
[13] 吉村卓也，藤井雄太郎，伊藤孝行 (2011). Robinson 型判定手法を用いた単語共起フィルタの検証. 第 10 回情報科学技術フォーラム講演論文集, FIT2011 (2), pp. 85‐90.
[14] 中村健二，田中成典，北野光一，寺口敏生，大谷和史 (2012). マルチエージェントクローラを用いた有害ユーザの効率的発見手法. 情報処理学会論文誌, 53 (1), pp. 90‐104.
[15] Ptaszynski, M., Rzepka, R., Araki, K., and Momouchi, Y. (2011). “Language combinatorics: A sentence pattern extraction architecture based on combinatorial explosion.” International Journal of Computational Linguistics (IJCL), 2 (1), pp. 24–36.
[16] Nakajima, Y., Ptaszynski, M., Honma, H., and Masui, F. (2016). “A Method for Extraction of Future Reference Sentences Based on Semantic Role Labeling.” IEICE Transactions on Information and Systems, 99-D (2), pp. 514–524.
[17] Ptaszynski, M., Masui, F., Kimura, Y., Rzepka, R., and Araki, K. (2015). “Extracting Patterns of Harmful Expressions for Cyberbullying Detection.” Proceedings of 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC’15), pp. 370–375.
[18] Ptaszynski, M., Lempa, P., Masui, F. (2015). A Modular System for Support of Experiments in Text Classification. Technical Transactions, pp. 229-243, publ. by Cracow University of Technology.
[19] Ptaszynski, M., Masui, F., Rzepka, R., Araki, K. (2017). Subjective? Emotional? Emotive?: Language Combinatorics based Automatic Detection of Emotionally Loaded Sentences, Linguistics and Literature Studies, Vol. 5, No. 1, pp. 36-50.
[20] Nakajima, Y., Ptaszynski, M., Honma, H., Masui, F. (2014). Investigation of Future Reference Expressions in Trend Information. In Proceedings of the 2014 AAAI Spring Symposium Series, “Big data becomes personal: knowledge into meaning”, pp. 31-38.
[21] Michal Ptaszynski, Juuso Kalevi Kristian Eronen and Fumito Masui. (2017). "Learning Deep on Cyberbullying is Always Better Than Brute Force", IJCAI 2017 3rd Workshop on Linguistic and Cognitive Approaches to Dialogue Agents (LaCATODA 2017), Melbourne, Australia, August 19-25.

著者略歴

安彦 智史（あびこ さとし）

2013 年関西大学大学院総合情報学研究科総合情報学専攻博士課程後期課程修了．同年，青山学院大学附置情報メディアセンター 助手．2016 年仁愛大学人間学部コミュニケーション学科講師に着任し，現在に至る．

長谷川 大（はせがわ だい）

2012 年北海道大学大学院情報科学研究科メディアネットワーク専攻博士課程修了．同年，青山学院大学理工学部情報テクノロジー学科助手．2013 年同助教．2017 年に東京工科大学メディア学部助教に着任し，現在に至る．

プタシンスキ ミハウ（ぷたしんすき みはう）

2010 年北海道大学大学院情報科学研究科博士後期課程修了，博士（情報科学）学位取得．同年日本学術振興財団研究員．2012 年北見工業大学非常勤研究員．2013 年同大学工学部助教に着任し，現在に至る．

中村 健二（なかむら けんじ）

2009 年関西大学大学院総合情報学研究科総合情報学専攻博士課程後期課程修了．同年関西大学ポスト・ドクトラル・フェロー．2010 年立命館大学情報理工学部助手．2012 年大阪経済大学情報社会学部准教授に着任し，現在に至る．

佐久田 博司（さくた ひろし）

1974 年東京大学工学部卒業．1979 年同大学大学院工学博士課程修了．1981 年同大学工学部金属工学科助手．1981 年(株)日立製作所日立工場．1983 年同技師．1984 年長岡技術科学大学工学部助手．1992 年青山学院大学理工学部助教授．1997 年マサチューセッツ工科大学客員助教授．2004 年青山学院大学理工学部教授．2007 年マサチューセッツ工科大学客員教授．現在に至る．