3P1-6in2

有害表現抽出に対する種単語の影響に関する一考察

A study on the effect of seed word selection on harmful expression extraction

畠山 鈴生[*1] 桝井 文人[*1] プタシンスキ・ミハウ[*1] 山本 和英[*2]
Suzuha Hatakeyama Fumito Masui Michal Ptaszynski Kazuhide Yamamoto

[*1]北見工業大学
Department of Computer Science,Kitami Institute of Technology

[*2]長岡技術科学大学
Department of Electrical, Electronics, and Information Engineering, Nagaoka University of Technology

We study a social problem called cyberbullying, which is a new form of bullying. In this study, we improre pronely propose method for detecting harmful entries in order to help mitigate the problem. We discuss the effect of seed word selection in order to refine the existing method. We collect several kinds of word sets using different approaches. By evaluating the precision of each approach, we were able to compare and analyse how change of the seed word sets influences the performance of the exiting methods.

1. はじめに

「ネット上のいじめ」が新たないじめの形として社会問題化している.「ネット上のいじめ」とは，携帯電話やパソコンを通じて Web 上のいわゆる学校非公式サイトの掲示板などに特定の人物 (子ども等) への誹謗中傷を書込んだり，嫌がらせメールを送ったりする行為である [1]．このようないじめに対処するために，学校関係者や保護者などが主体となってネットパトロール活動を行っている．しかしながら，ネットパトロールは主に人手による作業が中心であり，多数の掲示板に記述された膨大な書込みの中から有害情報を探し出すには計り知れない労力と時間を要する．そのため，これらの作業にかかる人的コストや作業従事者の身体的・精神的影響も懸念される．加えて，最近では許可された人物しか書込みを閲覧できない掲示板が増えつつあるため，ネットパトロール自体が困難な状況になっている．

上記の問題に対して，情報科学的アプローチを用いて対処しようとする研究が報告されている．新田ら [2] はカテゴリ別関連度最大化手法を提案している．彼らは，松葉ら [3] の有害極性判定手法を拡張し，少数の種単語を複数のカテゴリに分類，各カテゴリとの関連度の最大値を有害極性値とすることで，有害書込みを 90 %の精度で判定できると報告している．一方で，有害書込みにも非有害書込みにも現れやすい曖昧な表現や，個人情報で構成される有害書込みに対する誤判定が発生することについても言及している．

本研究では，新田らのカテゴリ別関連度最大化手法の性能向上を目的として，種単語の規模や組み合わせを変えることが処理結果に与える影響を検証する．

2. 従来手法

松葉ら [3] は，Turney[4] の関連度判定手法 ($PMI-IR$) を拡張して有害書込みとの関連度である有害極性値を算出し，少数の種単語を用意することで大量の有害書込み候補を効率よく発見できる有害極性判定手法手法を提案した．さらに，新田ら [2] は松葉ら [3] の有害極性判定手法を拡張して，種単語のカテゴリ化と関連度の最大値を取得する考えを導入したカテゴリ別関連度最大化手法を提案し，精度及び再現率を向上させた．

カテゴリ別関連度最大化手法は，有害書込みとの関連度である有害極性値を算出することで，有害書込みを 90%の精度で判定できると報告している．ところが，本研究で再評価実験を行ったところ，精度は 64%まで下がっていた．これは，新たな有害語や隠語が増えていることが原因の一つであると考えられる．また，1 章で述べたように有害書込みにも無害書込みにも現れやすい曖昧な表現や，個人情報で構成される有害書込みに対する誤判定が発生するという問題が残されており，対策として種単語を増やすということを述べている．

石坂らの手法 [5] は，コーパスから悪口文を自動検出するものであり，書込みに含まれる単語に対して悪口の度合いを示す SO 値 (*Semantic Orientation using Pointwise Mutual Information*) を計算する．SO 値とは，対象の単語が事前に用意した 2 つの基本単語のどちらと文書内共起しやすいかを，相互情報量を用いて定量化した値のことである．SO 値が 0 以上の場合は悪口単語，それ以外の場合は非悪口単語と判定する．石坂らの手法を用いて発見した悪口単語を新田らの手法の種単語とし，種単語の組み合わせや規模を考慮することにより性能向上が期待できる．また，この手法は Web 上の電子掲示板「2 ちゃんねる」を対象とした実験を通して，一定の有効性が確認されている．石坂らは，SO 値を算出するテストデータとして 2,735 単語の中から 3 人の評価者が悪口単語か否かを判断し，3 名が悪口と判断した 76 単語を使用している．悪口単語 (76 単語) の例としては「DQN[*1]」や「キチガイ」,「無能」等が挙げられる．

石坂らの手法を用いて種単語の規模を拡大し，新田らの手法 [2] を改良する．そして，種単語の規模や組み合わせが処理結果にどのような影響を及ぼすかについて検証する．

3. 種単語による効果の検証

3.1 種単語候補の選別

前章で述べた改良手法の効果を検証するために，石坂らの手法を用いて，種単語候補の選別を行った．

以下に選別の手順を示す．

1. テストデータは悪口単語 76 単語，初期データは石坂らが

[*1] ヤンキー (不良) や非常識人を意味するネットスラング

用意した悪口基本単語 14 単語と新田らの種単語 9 単語を合わせた 23 単語，非悪口基本単語 17 単語を用いてテストデータに対して SO 値を算出した．新田らの種単語 9 単語は，元々人手で判断された単語であり，SO 値を算出するにあたり，大きな影響を与える単語と判断し，加えた．

2. 上記 1 の結果，テストデータに対して 0 以上の SO 値を算出した基本単語の組み合わせが 44 パターンあった．44 パターン中，悪口基本単語は 23 単語中の 17 単語だった．

3. 上記 2 で得られた悪口基本単語 17 単語と非悪口基本単語 17 単語のセットを $seed17$，悪口基本単語を新田らの種単語 9 単語に変えた場合を $seed9$ として実験を行った．テストデータは悪口単語 76 単語に石坂らの非悪口単語 17 単語を加えた合計 93 単語とした．$seed17$ と $seed9$ の 2 セットを石坂らの手法に応用し，テストデータに対して SO 値を算出する．

4. 上記 3 で算出した SO 値を用いて悪口単語候補と非悪口単語候補を選別するための式 (1) の通り閾値 α_1 を設定する．

$$\alpha_1 = \mu \pm 2\sigma \qquad (1)$$

ここで μ は平均値，σ は標準偏差とする．また，悪口単語候補を w，非悪口単語候補 h とし，閾値を超える単語 ($h \leqq \alpha_1 \leqq w$) を種単語候補として選別する．

3.2 種単語候補選別結果と実験

選別した種単語候補を以下の表 1 に示す．

表 1: 選別した悪口単語候補と非悪口単語候補

	悪口単語候補	非悪口単語候補
seed17	ダセー，クズマスゴミ，バカサヨ，マジキモ，イボヲタ，ゴキヲタ，糞尿	絞り込み，買い上げ，振替，降順，素敵，美しい，可愛い，机，引換，赤い，太陽，チューリップ，夏
seed9	ビッチ，イボヲタ，目糞，脱糞，糞虫，バカサヨ，ダセー，糞尿，ゴキヲタ，マジキモ，クズマスゴミ，愚民	絞りこみ，チューリップ，買い上げ，素敵，太陽，美しい，机，夏，四角い，降順，赤い，可愛い，引換，寄生虫，振替

表 2 に，実験で用いた種単語の組み合わせを示す．表 2 に示す 6 セットの種単語の組み合わせを用意し，それらを新田らの手法 [2] に対して，これらの種単語を設定して性能の変化を調べた．ベースラインとして，石坂らが人手で用意した悪口単語 110 単語のうち相互情報量 (MI)[5] が高かった上位 5 単語 ($case5$) と，新田ら [2] の 9 単語 ($case6$) を種単語を使用した結果を用いた．テストデータは，新田らが用いている有害 1,508 文，非有害 1,490 文の計 2,998 文とした．

表 2: 実験に用いた種単語の組み合わせ

$case1$	$seed17$ の 7 単語
$case2$	$seed9$ の 12 単語
$case3$	$seed17$ に新田らの種単語 9 単語を加えた 16 単語
$case4$	$seed9$ に新田らの種単語 9 単語を加えた 21 単語
$case5$	石坂ら [5] による悪口基本単語候補 5 単語
$case6$	新田ら [2] による種単語 9 単語

3.3 実験の結果

前節で述べた実験結果，F 値の最大値を図 1 に示す．

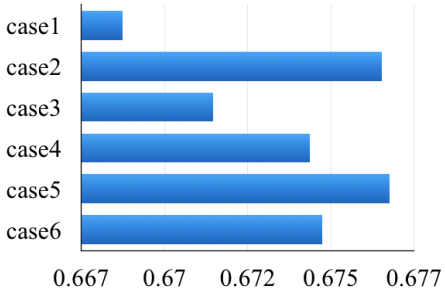

図 1: $case1$〜$case6$ の F 値の最大値

F 値については，$case5$ が最も高い値を示したが，精度の最大値が最も高かったのは $case4$ であった．

$case1$ から 6 の統計的有意差を検証するために，マクネマー検定 [6] を行った．マクネマー検定とは，対応のある 2 組の標本の比率の差を検定する．データ 1 とデータ 2 で誤って分類されたサンプル数を表 3 に従って，a から d のそれぞれの合計を求める．そして，統計値を計算する．a はデータ 1 とデータ 2 のどちらも有害と判断した合計である．b はデータ 1 では非有害，データ 2 では有害と判断した合計である．c はデータ 1 では有害，データ 2 では非有害と判断した合計である．d はデータ 1 とデータ 2 のどちらも非有害と判断した合計である．

表 3: 2 つのデータの性能対応表

		データ 1	
		正	誤
データ 2	正	a	b
	誤	c	d

今回は実験対象データ 2,998 文中，$phrase$ がないと判断した文を排除した 1,975 文を対象として，マクネマー検定を行った．

各実験結果に有意な差があるかどうかを検証した．$case5$ とその他の $case$，$case6$ とその他の $case$ を比較したマクネマー検定の結果を表 4 に示す．

表 4: 実験結果における有意差

データ	*case*1	*case*2	*case*3	*case*4	*case*5	*case*6
*case*5	***	***			–	***
*case*6	***	*	**	**	***	–

*p≤0.5, **p≤0.1, ***p≤0.01

表 4 の *case*5 との比較では，*case*1，*case*2，*case*6 で有意差が得られた．*case*3 と *case*4 では，どの有意水準においても有意差が得られなかった．また，*case*6 との比較では，有意水準 0.1%で *case*1，*case*5 で有意差が得られ，有意水準 5%では，全ての結果で有意差が得られた．これらの結果は種単語の規模が性能に必ずしも寄与しないことを示している．

4. 人間解析による精緻化

前章で作成した種単語抽出と，人間判断との違いを比較検証するために，まず単語の印象調査を Web アンケートを用いて実施した．これは，人間判断した単語を新田らの手法の種単語として用いると，石坂らの手法によって判断された単語よりも精度及び再現率が向上するかどうかを検証するために行った．アンケートの対象単語は，石坂らの手法で用いられている悪口単語 76 単語，非悪口単語 17 単語，新田らの手法の種単語の 9 単語のうち重なっている単語を除き，合計 101 単語とした．

評価方法としては,「無害」,「少し無害」,「わからない」,「少し有害」,「有害」の 5 段階評価とした．アンケートは，15 名 (男性 10 名，女性 5 名) から回答を得られた．実際に使用したアンケート画面の例を図 2 に示す．

図 2: アンケート画面の一部

今回は，アンケート結果の半数以上を基準として単語を絞った．アンケート結果として，半数以上が「有害」,「有害」＋「少し有害」,「無害」＋「少し無害」,「無害」と判断された単語数を以下の表 4 に示す．

表 5: アンケート結果 (半数以上が判断した単語数)

「有害」	「有害」＋「少し有害」	「無害」＋「少し無害」	「無害」
5	44	30	17

*case*1〜*case*6 の各単語とアンケートの「有害」5 単語は，全て一致する場合はなかった．また，*case*1〜*case*6 の単語には，「セックス」,「フェラ」,「ビッチ」といった卑猥語が含まれていたが，アンケート「有害」5 単語の中には含まれていなかった．このことから，システムが判断した単語と人間解析との単語では，違いがあることがわかった．アンケート結果を新田らの手法の種単語とし，石坂らの手法を用いて選別した単語を種単語とした場合との性能を比較していく．

アンケート結果で半数以上が「有害」と判断した 5 単語 (*case*7) を新田らの手法の種単語として実験を行った．また，*case*1〜6 の単語とアンケート結果で半数以上が「有害」＋「少し有害」であった単語と同じであった単語が 18 単語あったので，これらも種単語 (*case*8) として実験を行った．

*case*1〜6 の種単語の詳細と追加実験の種単語として用いた *case*7 と *case*8 の詳細を以下の表 5 に示す．

表 6: 種単語の組み合わせ *case*1〜*case*6 と実験に用いた種単語 *case*7 と *case*8 の詳細

*case*1	糞尿，バカサヨ，マジキモ，クズマスゴミ，イボヲタ，ゴキヲタ，ダセー
*case*2	糞尿，バカサヨ，マジキモ，クズマスゴミ，イボヲタ，ゴキヲタ，ダセー，ビッチ，目糞，脱糞，糞虫，愚民
*case*3	糞尿，バカサヨ，マジキモ，クズマスゴミ，イボヲタ，ゴキヲタ，ダセー，セックス，ヤリマン，フェラ，死ね，殺す，殴る，きもい，うざい，不細工
*case*4	糞尿，バカサヨ，マジキモ，クズマスゴミ，イボヲタ，ゴキヲタ，ダセー，ビッチ，目糞，脱糞，糞虫，愚民，セックス，ヤリマン，フェラ，死ね，殺す，殴る，きもい，うざい，不細工
*case*5	死ね，消えろ，蛆虫，カス，死ねよ
*case*6	セックス，ヤリマン，フェラ，死ね，殺す，殴る，きもい，うざい，不細工
*case*7	死ね，死ねよ，殺せ，殺す，クズマスゴミ
*case*8	死ね，消えろ，蛆虫，カス，ヤリマン，フェラ，殺す，きもい，うざい，不細工，ビッチ，クズマスゴミ，脱糞，糞虫，ダセー，ゴキヲタ，マジキモ，死ねよ

5. 結果と考察

前章で述べた *case*7 と *case*8 を種単語として用いた実験結果を以下の図 3，図 4 に示す．また，比較対象として，*case*1〜6 のうち精度の最高値が最も高かった *case*4 を図 5 に示す．

各図の y 軸は精度及び再現率，x 軸は実験で用いた文数 (件数) を有害かどうかを判定する閾値 (件数) として示している．閾値 (件数) を 50 件毎に設定し，それぞれの精度と再現率を表している．

全体的には閾値が小さな値の際，高い精度を示し，閾値が大きくなるにつれて精度は除々に低下していることがわかる．しかし，閾値 800 付近から，精度が再び向上しているので，文脈によって有害にも非有害にもなる文が混在していると思われる．

一方，再現率は逆の傾向を示している．閾値 850 付近から再び再現率が上昇しているのは，今回設定した種単語と関連の低い有害語を示しており，文脈によって有害にも無害にも変化する単語が存在することを示唆している．

*case*4 と他の *case* を比較すると，精度，再現率ともに大きな違いがみられた．精度に関しては，*case*7 と *case*8 の閾値 350 付近まではほぼ一定の値を保っているのに対し，*case*4 は

急激に減少している．再現率に関しては，*case*7 と *case*8 の閾値 450 付近までは一定に上昇し，そこから閾値 850 付近までほぼ一定でまた上昇しているのに対し *case*4 はほぼ一定に上昇し続けている．

このことから，人手で判断した種単語による有害語は高い極性値を持っており，精度に影響を及ぼすことがわかった．

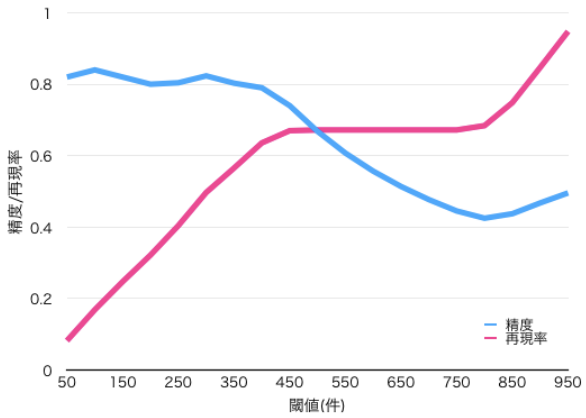

図 3: *case*7 の結果

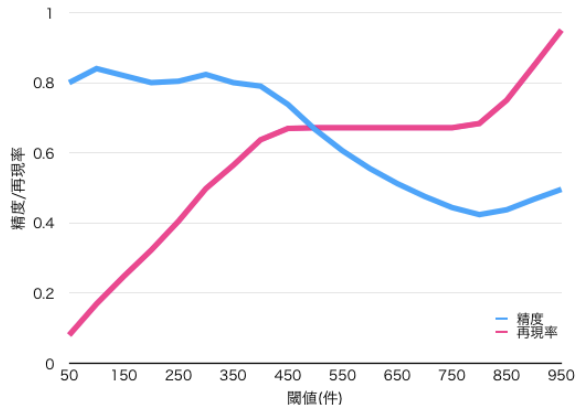

図 4: *case*8 の結果

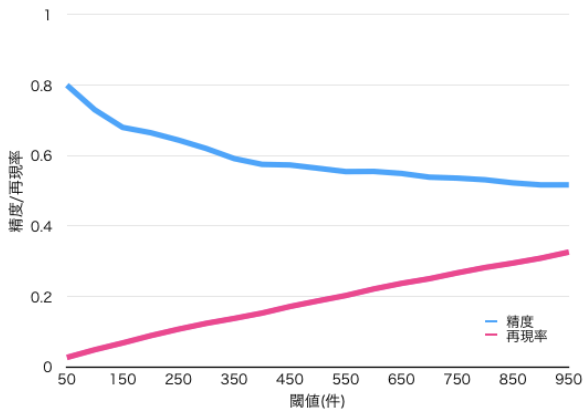

図 5: *case*4 の結果

6. おわりに

本研究では，新田らのカテゴリ別関連度最大化手法の性能向上を目的として，種単語の規模や組み合わせを変えることにより，処理結果にどのような影響があるかの検証を行った．石坂らの手法を用いて得られた悪口単語候補を新田らの種単語と組み合わせることで種単語と拡張し，比較実験を行った結果，種単語の組み合わせにより平均精度は向上することがわかった．このことから，種単語の選別，組み合わせにより性能向上が期待できる．また，Web アンケートを実施し，システムが判断した有害単語と人手で有害と判断された単語にはどのような違いがあるか検証を行った．その結果，人手で判断した単語を種単語として用いると多くの有害語が高い極性値を持っており，精度に影響を及ぼすことがわかった．

今後の課題として，文脈によって有害にも非有害にもなる文に対応する種単語を考慮することが挙げられる．

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] 新田大征，桝井文人，プタシンスキ・ミハウ，木村泰知，ジェプカ・ラファウ，荒木健治 :“カテゴリ別関連度最大化手法に基づく学校非公式サイトの有害書込み検出”，第 27 回人工知能学会全国大会発表論文集，(2013.6).

[3] 松葉達明，桝井文人，河合敦夫，井須尚紀 :“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”，言語処理学会第 17 回年次大会発表論文集，P2-26(2011.3).

[4] 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).

[5] 石坂達也，山本和英:“Web 上の誹謗中傷を表す文の自動検出”，言語処理学会第 17 会年次大会発表論文集，pp.131-134(2011.3).

[6] *McNemar*, Quinn : “Note on the sampling error of the difference between correlated proportions or percentages”. Psychometrika Vol. 12, No. 2, pp. 1532013157, (1947.6).