2M2-OS-34a-1

ドメインにより意味が変化する単語に着目した猥褻な表現のフィルタリング

Study of Harmful Expressions Filtering Focusing on Word's Meaning Depending on Document's Domain and Context

近江 龍一 *1　西原 陽子 *2　山西 良典 *2
Ryuichi Omi　Yoko Nishihara　Ryosuke Yamanishi

*1立命館大学大学院情報理工学研究科
Graduate School of Information Science and Engineering, Ritsumeikan University

*2立命館大学情報理工学部
College of Information Science and Engineering, Ritsumeikan University

In the context of harmful information for young people is frequently shared on the Internet, information filtering methods have been studied for their mental growth. Though the direct harmful expressions have been filtered by the existing methods, the indirect harmful expressions have been not covered yet; the indirect harmful expression denotes a metaphor and some words with unprintable characters. Such indirect expressions are not always harmful. Whether such expressions are harmful or not depends on the domain and the context of the information. The goal of our research is to develop a novel filtering method for harmful sentences which is especially harmful to young people. In this paper, the varieties of obscene sentences were analyzed. Through the analysis, we found that the obscene sentences could be generally classified into four classes. The features of each class have been suggested from the discussions.

1. はじめに

インターネット上には未成年に対して有害な情報が溢れている．有害な情報とは法に触れてしまうものや，差別に関するもの，アダルト系，暴力表現，ギャンブルに関するもの，出会い系，グロテスクな表現と様々なものがある．有害な情報は，特にインターネット上の電子掲示板やブログ，同人サイトなど，一般の人が自由に書込み，アップロードができる所で確認できる．例えば，電子掲示板の 2 ちゃんねる *1 やイラストコミュニケーションサービスの pixiv*2 がある．これらのサイトには有害な情報が掲載されることも少なくはない．

有害な情報を未成年に見せないように，年齢制限や有害な表現でフィルタリングする対策が行われている．フィルタリング技術の多くは，情報中に有害な表現が含まれるかどうかを判定し，有害な表現を含むならばその情報をフィルタリングする．一方で，有害な表現を含むかどうかを判定する場合に，間接的な表現があると判定が困難になることも多い．有害な情報の中でも特に性表現に関するものは，その情報の性質から，直接的なものもあれば間接的なものまで多種多様な表現が用いられてきた．直接的な性表現とは例えば，性器や性行為を表す単語である．間接的な性表現とは例えば，情報が掲載されているドメインや文脈に即した暗喩や伏せ字を用いたものである．直接的な性表現を用いるとフィルタリングされやすくなるため，敢えて間接的な性表現を用いることもある．例えば，掲示板での売春に関する書込みの一部を伏せ字にして，客の勧誘を助けた事件があった *3．この事件のサイトの情報では伏せ字として,「＊交」などが使われていた．伏せ字が用いられたとしても，多くの人は掲示板に書かれているこの情報が，援助交際の略語を暗示した「援交」だと判断できる．人間であれば，単語自体の表層的な文字が伏せられていたとしても，前後の文脈や情報が掲載されているウェブサイトの特性から総合的に，この伏せ字が暗喩する対象を判断することができる．他にも，「バナナ」で男性器を暗喩するなど，単語や表現は使われるドメインや文脈によって意味が変化する．このような多種多様に間接的に記述された有害な表現を含む情報をフィルタリングするためには，単語やその表現が使われているドメインや文脈に応じて意味が変化することを捉える必要がある．

本研究では，ドメインや文脈により単語の意味が変化することに着目し，有害な表現を含む情報をフィルタリングする手法を新しく提案する．研究の第一段階として，本論文では，猥褻な表現の種類と特徴について分析を行う．分析結果を考察し，有害な情報の新しいフィルタリング手法について検討する．

2. 関連研究

有害文書のフィルタリングの研究は多くなされている [3]．菊池ら [4] の研究では 2 つの単語の共起確率を用いることにより，有害サイトを高精度に検知する手法を提案している．この手法では，迷惑メールのフィルタリングに多く用いられているベイジアンフィルタを応用している．ベイジアンフィルタでは過去の有害サイト，非有害サイトから得られた情報から 1 つの単語の有害確率を計算する．そして，1 つの単語ごとの有害確率を結合して対象が有害かどうかの判断をしている．菊池らはこのベイジアンフィルタの考え方に基づき，2 語の共起による有害確率を計算している．Deyue ら [1] も同様に複数の単語の共起を用いてフィルタリングしている．中村ら [7] は Web ページをブロック単位で分割してから共起関係を用いる手法を提案している．これらの手法はサイトや Web ページの単位でフィルタリングを行うことができる．一方で 1 文単位のフィルタリングは困難である可能性が高い．本論文では文書中の 1 文ごとのフィルタリングを目指す．これにより，真に有害な表現を含む部分的な情報だけを除き，それ以外の情報は閲覧できるようにする．

連絡先: 近江龍一，立命館大学情報理工学部，滋賀県草津市野路東 1-1-1

∗1 http://www.2ch.net/
∗2 http://www.pixiv.net/
∗3 http://happymail.co.jp/age/?enhmpc01

表 1: 分析に用いた小説データ．論文での仮 ID と小説 URL，作者名，小ジャンルを示す．

ID	小説の URL	作者名	小ジャンル
1	http://www.pixiv.net/novel/show.php?id=7291993	みみどん	BL
2	http://www.pixiv.net/novel/show.php?id=7294856	水青	BL
3	http://www.pixiv.net/novel/show.php?id=7294158	帯	BL
4	http://www.pixiv.net/novel/show.php?id=7295465	木月	BL
5	http://www.pixiv.net/novel/show.php?id=7298387	松永 奏多	BL
6	http://www.pixiv.net/novel/show.php?id=7294216	なんこ	BL
7	http://www.pixiv.net/novel/show.php?id=7295301	花火☆ gcat	BL
8	http://www.pixiv.net/novel/show.php?id=7289312	ネギ助	BL
9	http://www.pixiv.net/novel/show.php?id=7297755	緑@ついったー	NL
10	http://www.pixiv.net/novel/show.php?id=7297605	兎綺	NL

池田ら [2] は係り受けを用いた有害情報のフィルタリングの手法を提案している．この手法では，有害な文からは有害な係り受け文節組を，無害な文からは無害な係り受け文節組を検出することで有害情報のフィルタリングの精度を向上させている．さらに概念辞書を用いて係り受け文節組を抽象化し，より多くの表現を検出している．係り受けは一般的な単語の組合せ，使われ方を前提としており，暗喩された有害表現に対応することは困難である可能性が高い．本研究では，暗喩された有害表現の検出を目指す．

栗原ら [5] は暗喩されて表現されている名詞句の意味解析の提案をしている．この手法では，暗喩の名詞句である ”A の B ”の意味解析を行うために，暗喩の名詞句 ”A の B ”に特化した辞書を作る．そして，その辞書を検索することで意味解析をしている．暗喩の表現に，特に有害表現の暗喩は様々なものがあることが確認されている [6]．本研究では名詞句に加えて，他の形式の暗喩についても扱えるようにする．

3. 文の種類を分析する方法

猥褻な表現が含まれる小説をテキストデータとして用意し，テキストデータに含まれる猥褻な表現に関する文の種類を分析した．本論文では，pixiv に投稿されていた小説を分析に用いた．

3.1 分析に用いた小説のデータ

本研究では猥褻な表現に関する文を集めるために，pixiv に投稿されている R-18 小説を用いた．R-18 小説に分類される小説の中から，2016 年 10 月のウィークリーランキング Top10 の小説を選択し，分析に用いた．

10 個の小説の URL と小ジャンルを表 1 に示す．小ジャンルはノーマルラブ (NL と略す）とボーイズラブ（BL と略す）があり，それぞれ 2 個，8 個であった．

3.2 データからの文の切り出し

小説のテキストデータから，文を切り出した．文は必ず句点で区切られているわけではなく，記号で区切られていることもあった．句点や記号が複数個使われ，文が区切られていることもあった．そこで，文字と文字の間に記号または句点が含まれている場合に，最後の記号または句点を文の区切りと見なし，文章を 2 つに切り分けた．

3.3 4 種類の猥褻な表現の分類クラス

切り出された文を猥褻な表現とそれ以外に分類し，猥褻な表現の特徴を得たい．本論文では猥褻な表現として，性表現を扱うこととする．猥褻な表現の特徴を得るために，人手で予備的な調査を行い，以下の 4 種類の猥褻な表現の分類クラスを設定した．

- クラス 1: 文が猥褻な文脈に含まれ，かつ文に直接的な性表現が含まれるクラス
- クラス 2: 文が猥褻な文脈に含まれ，かつ文に間接的な性表現が含まれるクラス
- クラス 3: 文が猥褻な文脈に含まれないが，文に直接的な性表現が含まれるクラス
- クラス 4: 文が猥褻な文脈に含まれるが，文に直接的，間接的いずれの性表現も含まれないクラス

1 つの文の中にクラス 1 と，クラス 2 の特徴を持つものも存在した．このような場合は，1 つの文が 2 つのクラスに同時に属するものとして分類した．4 種類のクラスについて，3.3.1 から 3.3.4 で，例をあげて説明する．

3.3.1 クラス 1: 猥褻な文脈に含まれ，直接的な性表現を含む文

一つ目のクラス 1 は，文が猥褻な文脈に含まれ，かつ直接的な性表現を含む文のクラスとする．このクラスに含まれる文の例としては「一松は急いでカラ松から目をそらし、ギンギンになった自分の（男性器を示すカタカナ 3 文字）を扱くふりをした。」がある

3.3.2 クラス 2: 猥褻な文脈に含まれ，間接的な性表現を含む文

二つ目のクラス 2 は，文が猥褻な文脈に含まれ，かつ間接的な性表現を含む文のクラスとする．このクラスに含まれる文の例としては「びゅびゅっと一松の口の中に暖かい液体が広がる。」がある．

3.3.3 クラス 3: 猥褻な文脈に含まれないが，直接的な性表現を含む文

三つ目のクラス 3 は，文は猥褻な文脈に含まれないが，直接的な性表現を含む文のクラスとする．このクラスに含まれる文の例としては「前戯とかは大丈夫なんだ。」がある．

3.3.4 クラス 4: 猥褻な文脈に含まれるが，いずれの性表現も含まない文

四つ目のクラス 4 は，文は猥褻な文脈に含まれるが，直接的，間接的，いずれの性表現も含まない文のクラスとする．このクラスに含まれる文の例としては「先ほど丹念に舐め解したため、すぐに舌の侵入を許してしまう。」がある．

これらの 4 つのクラスを用いて，猥褻な表現に関する文を分類した結果を 4. にて示す．

表 2: 小説中の猥褻な表現に関する文を 4 種類のクラスに分類した結果. () 内に猥褻な表現に関する文を 100%としたときの割合を示す.

ID	全文数	猥褻な表現に関する文	クラス 1	クラス 2	クラス 3	クラス 4
1	1,626	1,425	377 (26.4%)	170 (11.9%)	5 (0.4%)	942 (66.0%)
2	324	125	66 (52.8%)	40 (32.0%)	0 (0%)	55 (44.0%)
3	480	59	5 (8.5%)	34 (57.6%)	0 (0%)	24 (40.7%)
4	1,568	336	42 (12.5%)	60 (17.9%)	0 (0%)	240 (71.4%)
5	317	306	14 (4.6%)	64 (20.9%)	0 (0%)	234 (76.4%)
6	570	67	2 (3.0%)	4 (6.0%)	0 (0%)	61 (91.0%)
7	252	75	23 (29.9%)	19 (24.7%)	3 (6.5%)	35 (45.5%)
8	469	150	32 (21.3%)	45 (30.0%)	0 (0%)	55 (36.7%)
9	720	346	46 (13.3%)	77 (22.3%)	0 (0%)	243 (70.2%)
10	683	309	8 (2.6%)	62 (20.1%)	0 (0%)	241 (78.0%)
平均	700.9	319.9	61.5 (19.2%)	57.5 (18.0%)	0.8 (0.2%)	213 (66.5%)

4. 文の分類結果と考察

小説の各文を 3.3.1 から 3.3.4 で説明したクラスに分類した結果を説明する．文を猥褻な表現に関するものとそれ以外に分類し，その後，猥褻な表現に関する文をクラス 1 からクラス 4 のいずれかに分類した．分類は全て人手で行い，第一著者が行った．複数人にアンケートをとるなどは，本論文では行わなかった．

各小説の分類結果を表 2 に示す．10 個の小説の中に文は全部で 7,009 文あった．そのうち猥褻な表現に関する文は 3,199 文あった．クラス 1 に分類された文は 615 文，クラス 2 に分類された文は 575 文，クラス 3 に分類された文は 8 文，クラス 4 に分類された文は 2,130 文であった．いずれにも分類されないものはなかった．なお，同時に 2 クラス以上に属する文もあるため，1 ドキュメントに対する%が 100 を越えることがある．

5. 文の特徴の考察

本論文では猥褻な表現に関する文のクラスは 4 種類あると定義した．本節では，この 4 種類のクラスに含まれる文の特徴を考察し，フィルタリングに使える特徴を検討する．

クラス 1 は，猥褻な文脈に含まれ，かつ直接的な性表現が含まれる文のクラスであった．クラス 1 の文は直接的な性表現の単語が使用されているため，猥褻な単語の辞書を作ることにより，猥褻な表現に関する文をフィルタリングできる．

クラス 2 は，猥褻な文脈に含まれ，かつ間接的な性表現が含まれる文のクラスであった．クラス 2 の文は間接的な性表現の単語を含むため，辞書を用いてのフィルタリングは困難である．間接的な性表現の単語は，単語や文が含まれるドメインや文脈により意味が変化する．ドメインや文脈を特定した上で，間接的な性表現の単語を判定する方法を考える必要がある．1 章で触れたバナナという単語を例に挙げると，官能小説ではバナナという単語は男性器という意味で使われることがあり，フィルタリングの対象となる．しかし，普通の会話で使われる場合は問題ない．この場合，例えば文中や前後の文に出現する有害な単語の含有率，あるいはバナナと有害な単語の共起に注目することにより有害な情報にたどり着く可能性などを考慮することで対応可能と考える．

クラス 3 は，猥褻な文脈に含まれないが，直接的な性表現が含まれる文のクラスであった．クラス 1 と違う点は，猥褻な文脈には含まれないが，猥褻な表現に関する文という点である．一つの小説の中で文脈は連続しているため，猥褻でない文脈と猥褻な文脈が存在した際に，突然に猥褻な文脈へと切り替わることは少ないと考えられる．これらの文は，猥褻でない文脈から猥褻な文脈へと切り替わる境界で出現することが多かった．したがって，文脈の変化の兆しを評価し，フィルタリングする必要があると考えられる．さらに，クラス 3 に分類された文については，猥褻かどうかの評価が割れる可能性も高いため，今後，複数人にアンケートをとり，分類の見直しをする必要がある．

クラス 4 は，猥褻な文脈に含まれるが，文に直接的，間接的いずれの性表現も含まれない文のクラスであった．クラス 4 の文は性表現の単語を含まないが猥褻な表現に関する文である．このクラスの文は表 2 において，2,130 文あり，4 つのクラスの中で最も多かった．最も数が多かったことから，猥褻な表現の多くは文脈と暗喩を用いて表現されることが多いと示された．したがって，猥褻な表現を高い精度でフィルタリングするためには，文脈や暗喩を評価する必要があることがデータから確認できた．クラス 4 の文をフィルタリングするためには，まず文が含まれる文脈を特定する必要がある．その後，単語や単語の組合せの特徴を調査し，フィルタリングする方法を考える必要がある．

6. まとめ

本論文では，ドメインにより意味が変化する単語に着目した有害な情報のフィルタリングのために，猥褻な表現に関する文の種類を分類した結果を報告した．本論文では，猥褻な表現に関する文のクラスとして 4 つのクラスを定義した．4 つのクラスとは，クラス 1: 文が猥褻な文脈に含まれ，かつ文に直接的な性表現が含まれるクラス，クラス 2: 文が猥褻な文脈に含まれ，かつ文に間接的な性表現が含まれるクラス，クラス 3: 文が猥褻な文脈に含まれないが，文に直接的な性表現が含まれるクラス，クラス 4: 文が猥褻な文脈に含まれるが，文に直接的，間接的いずれの性表現も含まれないクラスであった．pixiv の R-18 指定の小説を 10 件用意し，小説に含まれる猥褻な表現に関する文を人手で 4 つのクラスに分類した．分類した結果，最も文が多かったクラスはクラス 4 で，最も文が少なかったクラスはクラス 3 であった．各クラスの特徴を考察した結果，クラス 1 は単語辞書によりフィルタリング可能，クラス 2 からクラス 4 は文が含まれるドメインや文脈を判定する必要があり，その上でフィルタリングに利用可能な特徴を調査する必

要があることが明らかになった.

参考文献

[1] Deyue Deng，大塚 孝信，伊藤 孝行：複数単語共起フィルタリングにより大規模化するデータを処理する有害文分類手法の提案，研究報告知能システム，Vol.2013，No.2，pp.1-8，(2013).

[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式の提案，DEIM Forum 2010 C9-5, (2010).

[3] 石坂 達也， 山本 和英：２ちゃんねるを対象とした悪口表現の抽出，言語処理学会第 16 回年次大会発表論文集，pp.178-181，(2010).

[4] 菊池 琢弥，内海 彰：語の共起情報に基づく有害サイトフィルタリング手法，情報科学技術フォーラム講演論文集，Vol.9，No.6，pp.1-6，(2013).

[5] 栗原 健，松本 和幸，土屋 誠司，任 福継：意味素に基づく隠喩の名詞句 “A の B” の意味解析，研究報告情報学基礎（FI），Vol.2010，No.1，pp.1-6，(2010).

[6] 永田 守弘: 官能小説用語表現辞典, ちくま文庫, (2006).

[7] 中村 健二，田中 成典，山本 雄平，安彦 智史：共起関係の抽出範囲を考慮した有害情報フィルタリング手法，情報処理学会論文誌，Vol.54，No.2，pp.571-584，(2013).