テキストに基づく違法有害記事の削除作業支援方式

笠原要，藤野昭典，永田昌明

NTT コミュニケーション科学基礎研究所
{kasahara.kaname,　fujino.akinori, nagata.masaaki}@lab.ntt.co.jp

1. はじめに

消費者生成メディア(Consumer Generated Media, CGM)の記事には、犯罪に関わる違法情報や青少年に有害な内容等が含まれている場合があり社会問題となっている。CGMを提供する事業者は、記事の違法有害性を適宜判定して削除・通報等することが求められているが、多数の記事を目視確認する場合、人的稼動を要する点が問題である。そこで本稿では、その作業を効率化する支援方式を検討した。

2. 違法有害記事の削除作業支援

2.1 CGMコンテンツの違法有害記事削除作業

ブログやSNS、レンタル掲示板、口コミサイト等の利用者自身がコンテンツを投稿するCGMが急速に社会に普及している。総務省調査では、国内の主要ブログとSNSのユーザー数は延べ9,800万と推測している[1]。記事を継続的に投稿する人はこの一部であるが、記事へのコメントや評価まで含めれば、多数のユーザーが関わってコンテンツが生成されている。

CGMコンテンツの特徴としては、件数の多さと共に、多種多様な情報が含まれていることである。最新の口コミ情報や専門/ニッチ/最新情報が集積しており、インターネット利用者にとって有益な知識源となっている。しかしながら、事業者が直接調達する従来のインターネット上のコンテンツと異なり実質的には誰でも記事投稿できるので、事業者が内容を直接コントロールできない。そのため、犯罪を予告/誘導/告白する違法な内容や青少年にとって有害な内容がCGMコンテンツに含まれる場合があり、社会的な問題となっている。CGM事業者は、違法有害な投稿記事に対する外部の通報に基づいて削除を行っているが、より主体的な対応が求められている。例えば、青少年の利用に適した健全なモバイルコンテンツの認定を行う第三者機関であるEMA(モバイルコンテンツ審査・運用監視機構)のサイト運用管理体制認定基準[2]では、サイト事業者が毎日のサイトパトロールを行うことが求められている。

CGMサイトの一部では、独自に禁止語辞書を作成し、それを含む投稿を自動削除しているが、禁止語を予想して別の表現で書き込みされると対処できない。一方、禁止語を増やしすぎると、問題がない投稿まで削除してしまう。そのため、青少年向けのCGMサイトを中心として、事業者や委託された監視事業者が直接投稿記事を目視確認して削除することが行われている。違法な記事やサイトに適合しない有害な記事に確実に対応できるが、要する稼働コストが高くなってしまうので、EMA 等のコンテンツ認定機関の基準に叶うサイトは依然少数である。

上記状況を鑑み、CGM 事業者が行う記事チェック作業コストを軽減する支援方法を検討する。CGM コンテンツの種類には画像や音声、動画もあるが、本稿ではテキストのみを対象とする。

2.2 違法有害記事の特徴

CGM コンテンツに現れる違法有害記事の特徴について説明する。第1に、投稿記事の違法有害性は、画一的ではなくCGMサイトによって異なることが挙げられる。犯罪の予告や違法な物品の取引のような内容は常に違法な記事といえる。一方、犯罪に間接的に結びつくような内容や、有害性に関する内容についてはサイト毎の対応は様々である。例えば出会いを仲介するサイトでは、サイト上でのコミュニケーション以外は禁止する場合があれば、メールアドレスの交換まで許容するサイト、実際の出会いの希望まで記事に含めることも許容するサイトまで様々存在する。そのため、同一の投稿記事に対してもサイトによって削除対応が異なる。

そのため、違法有害性の判定作業の支援方式では、固定的な判定基準を緻密にルール化して利用することでは汎用性に欠けてしまう。CGMサイトの違法有害性の基準に柔軟に対応できる支援が求められる。

第2の違法有害記事の特徴としては、その件数が無害な記事に比して極めて少ないことが挙げられる。一部レンタル掲示板サービスのように削除基準を利用者に決定させるポリシーを取るCGMサービスでは、スレッド内に常識的には有害と言える記事が多数含まれる場合もある。しかし、コンテンツの監視作業を専門の事業者に委託するCGMサイトの場合には、含まれる違法有害記事は数%以下であることが多い(監視事業者聞き取りに基づく)。そのため、圧倒的に多数の無害な記事をできるだけ目視チェックさせないような支援方式が適当である。

そして、違法有害記事を含むCGMコンテンツでは、新語や新しい言語表現が多数現れることが重要な特徴である。そのため、新聞記事等の一般的なテキスト

の処理で検証されている形態素解析や構文解析ツールをそのまま使った場合、想定される結果が得られない場合がある。辞書を対応させることも考えられるが、新語の表現は月単位で変化する場合もあるので、解析結果の誤りを想定した支援方式が望まれる。

2.3 提案方式

これまでに説明した違法有害記事の特徴及び削除作業の特徴を考慮した支援方法について説明する。

図1は、その概要である。投稿された CGM 記事のテキストに着目し、違法有害性判定済みの教師データに基づき、無害な記事を分類し取り除き、残った記事を CGM 事業者の目視確認対象とする。また、対象記事について違法有害な表現を自動抽出し、それを強調して記事とともに確認作業者の端末画面に表示する。これにより作業者の確認作業対象とする投稿記事の件数を削減させ、目視作業においても有害な表現を探しやすくすることで作業効率を向上させる。

無害な記事の除去では、できるだけ多くの無害な記事を分類することが目的であるが、それ以上に、有害な記事を誤って無害と分類することを避けることが重要である。そこで、以下の2つの尺度を用いる(図2)。

UUR (Unblocked-contents Under-block Rate)[3]は、システムが無害と判定した記事中に含まれる実際には有害な記事が含まれる割合である。また、目視チェック削減率とは、当初の記事の件数に対するシステムが出力した無害記事の割合を表す。CGM 事業者が期待する UUR の値(例えば 0.01%)の元に、目視チェック削減率をできるだけ削減できる分類方法を選択することになる。

		システム出力	
		有害 p	無害 n
正解	有害 p	pp	np
	無害 n	pn	nn

Unblocked-contents **U**nder-block **R**ate

UUR = np ／(np + nn)

目視チェック削減率

= (np + nn) ／ (pp + pn + np + nn)

図2 無害記事除去の評価尺度

また、CGM 事業者によって違法有害基準が異なるので、事業者が過去に行った判定結果を教師データとして、その傾向に基づく無害記事の分類を行う。分類学習の方法としては、SVM、naive Bayes や、ベイジアンスパムフィルタ方式を利用して実験的に比較検討した。

記事に対する違法有害な記述の抽出については、

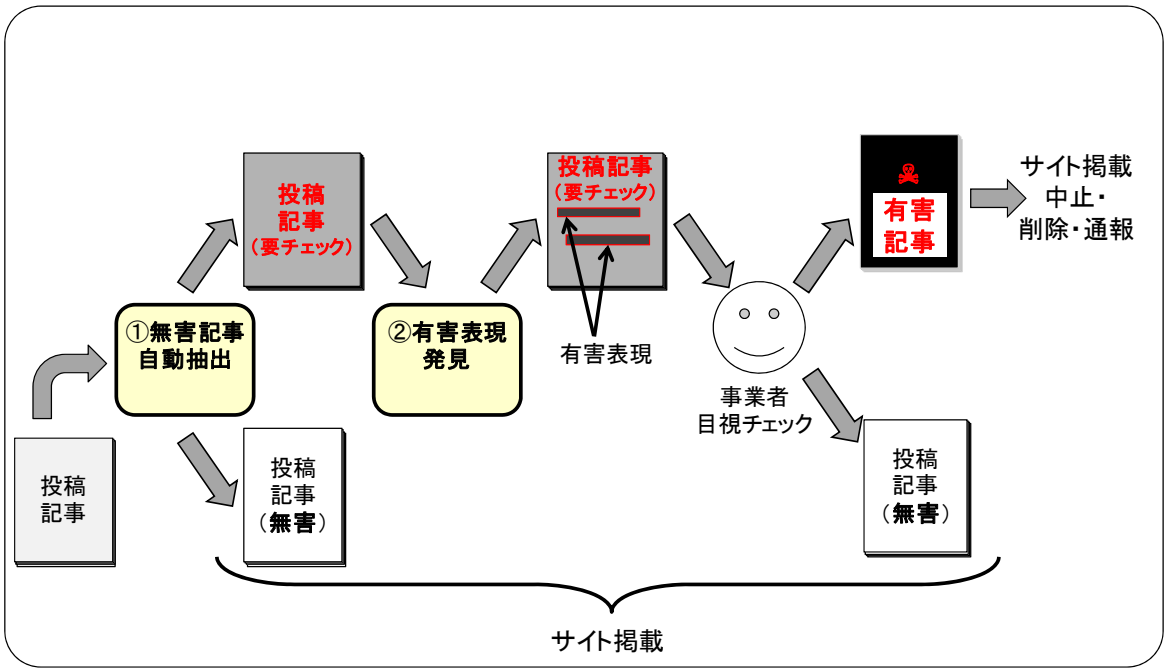

図1 提案方式概要

れよりも短い節等のみに現れる場合も考えられる。一方、個々の単語の違法有害性を判定し、その結果を CGM 事業者の目視作業の参考情報することも考えられる。しかし、その語に曖昧性がある場合に作業者は結局、単語の前後を確認する必要がある。そのため、単語が連続する一定の表現を抽出する方法を検討した。

この場合、記事の各単語や文字が有害な表現であるかを同定するチャンク同定問題に帰着できると考え、アプローチが類似するテキストからの人物説明記述の抽出方法[4]を参考として、3種類のラベル[5]を設定した。

I：現在位置のトークン（単語、文字）が違法有害表現の先頭以外の一部である
O：現在位置のトークンは違法有害な表現ではない
B：現在の位置は違法有害な表現の先頭である

違法有害であるかの分類と同様に表現のラベリングも、CGM サイトの基準によって結果が異なると考えられるので、訓練データに基づく機械学習方法を用いて比較検討した。

3. 実験

3.1 実験方法

10-20 代女性を主要ターゲットとして提供されているモバイル系 CGM サイトのブログ記事及び記事コメント 72 万件を実験データとした。CGM 事業者の違法有害判定基準に従い、専門家によって目視で違法有害であるか無害であるか確認した。981 件（約 0.13%）の違法有害な記事が発見された。有害記事については、特徴付ける違法有害な表現を抽出した。違法有害基準としては、違法行為や個人情報の開示や卑猥表現、サイト外の商行為に関する記述であった。例えば個人情報に関しては住所の市区町村名までは可としているように、それぞれについて詳細な分類から基準が選択されている。違法有害の判定結果が付与された記事データについて、64.8 万件を教師データ、7,200 件を評価用のテストデータとして用いた。テストデータ中に違法有害記事は 108 件含まれていた。また、素性抽出のための形態素解析には、MeCab を用いた[6]。

提案方式における無害記事を除去するための分類プログラムとして、naive Bayes は CPAN の Algorithm::NaiveBayes(http://search.cpan.org/~kwilliams/Algorithm-NaiveBayes-0.04/lib/Algorithm/NaiveBayes.pm)、SVM は、TinySVM [7]を用いた。訓練データは4分割し、分類の UUR を 0.0001(0.01%)以下の時に無害記事削減率を最大化するように分類の閾値を最適化した。

違法有害表現のラベリングには、SVM の実装の YamCha[8]、CRF[9]の実装である CRF++ (http://crfpp.sourceforge.net/)を用いた。抽出された違法有害表現と専門家が抽出した表現箇所の IO ラベル部分についての適合率、再現率、F 値で評価した。

3.2 無害記事の除去

72,000 件のテストデータについて、種々の分類手法を適用した。評価結果を表1に示す。

素性としては、自立語の原形と品詞の組を用い、素性の出現頻度を重みとしている。naive Bayes と SVM では、自立語以外の形態素の出現頻度も含めて重みを正規化している。また、分類閾値の最適化の際に、UUR が 0.01%を満たさない場合はそれに近い条件で最適化した。Robinson、Robinson-Fisher、Graham はそれぞれ、ベイジアンスパムフィルタリングの方法である。

無害記事の削除率については、Graham の方法が約 7 割と高い値となっているが、UUR は 0.5%と条件を満たしていない。そのため、条件を満たし削除率が最も高い(58%)Robinson の方法を用い、続く方式の評価を行った。

表1 無害記事除去方法の比較

分類器	UUR (%)	削減率	分類精度	有害記事再現率
Robinson	0.0048	0.580	0.581	0.981
Robinson -Fisher	0.0052	0.538	0.539	0.981
Graham	0.0473	0.704	0.705	0.778
naive Bayes	0.0285	0.245	0.245	0.945
naive Bayes (normalized)	0.0000	0.082	0.083	1.000
SVM	0.0000	0.001	0.001	1.000
SVM (normalized)	0.0000	0.000	0.001	1.000

3.3 有害表現ラベリング

ラベリング手法の比較を行った（表2）素性としては、形態素の原形と品詞を用いた。また、ウィンドウ幅±5 として CRF と SVM を適用した。その結果、適合率、再現率共に CRF を用いた結果が優れていた。

表2 ラベリング方法の比較

Method	Precision	Recall	F 値
CRF	0.1899	0.1733	0.1813
SVM	0.0714	0.1481	0.0964

素性を獲得するための形態素解析が適切であるかを確認するため、種々の素性を用い CRF でラベリングを行い評価した(表3)。表層表現及び原形については、ウインドウ幅を±5とし、バイグラムまで考慮した。また、文字の場合には同等のウィンドウ幅で n=4 まで考慮した。品詞については、ウィンドウ幅±1、ユニグラムまで考慮した。

文字を素性とした場合には、他の素性に比べて評価値が下がっており、形態素解析が有効であることが推測される。一方品詞も利用した場合には、大きな評価値の向上は見られず、品詞情報が活用されていない。

表3 素性の比較

素性	Precision	Recall	F 値
原形+品詞	0.1899	0.1733	0.1813
表層	0.1832	0.1753	0.1792
原形	0.1833	0.1709	0.1768
文字	0.1505	0.1686	0.1590

4. おわりに

本稿では、CGM 事業者が行う投稿記事のチェック作業を支援する方法として、違法有害であるかを判定したデータを教師データとして無害な記事を高精度に除去し、残りの記事について、違法有害な表現を抽出、付加情報として与える支援方法を提案し、既存の分類、ラベリング手法を適用して比較を行った。

今後は、既存の禁止語辞書を用いる方法と比較を行い作業効率の向上の評価を行う予定である。また、違法有害表現のラベリングにおいて、品詞情報が活用出来ない問題については、形態素解析結果を分析し、方式の改善を計る予定である。

謝辞

本研究において、NTT コミュニケーション科学基礎研での実務実習の一環で実験に協力して下さった長岡技術科学大学の一野瀬翔吾さんに感謝いたします。

参考文献

[1] 総務省: ブログ・SNSの経済効果の推計, http://www.soumu.go.jp/main_content/000030547.pdf , 2010.

[2] EMA: コミュニティサイト運用管理体制認定基準, http://www.ema.or.jp/dl/communitykijun.pdf , 2010

[3] Resnick, P. J., Hansen, D. L., & Richardson, C. R.: Calculating error rates for filtering software. Communications of the ACM, 47(9), pp.67-71, 2004.

[4] 西田成臣, 森辰則: 機械学習を用いた二段階洗練化手法による人物説明記述の抽出, 情報研究報, NL-67, pp.79-84 , 2008.

[5] Erik F. Tjong Kim Sang and Jorn Veenstra, Representing Text Chunks. In: ″Proceedings of EACL'99″,1999.

[6] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, EMNLP-2004, pp.230-237 (2004).

[7] Kudo, T. (2002). TinySVM: Support Vector Machines. http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/

[8] T. Kudo, Y. Matsumoto, "Chunking with support vector machines", Proc. of the 2nd Meeting North American Chapter of the Association for Computational Linguistics, 2001.

[9] Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and labeling Sequence Data, In Proc. of 18th International Conference on Machine Learning, pp.282-289, 2001.

[10] G. Robinson. Spam Detection, http://radio.weblogs.com/0101454/stories/2002/09/16/spamDetection.html, 2002.

[11] G. Robinson. A Statistical Approach to the Spam Problem. Linux Journal, No. 107, 2003.

[12] P. Graham. A Plan For Spam, http://www.paulgraham.com/spam.html, 2002.