3O1-3in

ツイート炎上抑制のための包括的システムの構築

Building a Comprehensive System for Preventing Flaming on Twitter

大西 真輝 Masaki Onishi　澤井 裕一郎 Yuichiro Sawai　駒井 雅之 Masayuki Komai　酒井 一樹 Kazuki Sakai　進藤 裕之 Hiroyuki Shindo

奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Insutitute of Science and Technology

It has been a social problem that a user who has made an inappropriate remark on the Internet receives a harsh reaction from other users, which is called “flaming”. We built a comprehensive system that aims to prevent flaming especially focused on Twitter. The system predicts whether the user's input will cause flaming. presents the user expected replies to the user's tweet, and suggests a corrected version of the tweet. The user is expected to be informed of the danger of flaming and gain knowledge about how to correct the tweet.

1 はじめに

近年，Twitter や Facebook といったソーシャル · ネットワーキング · サービス (SNS) の利用者数は増加の一途を辿りづつけている．必然的に，インターネット上のコミュニケーションは巨大なものとなり，今や SNS は，現代のインターネットにおいて必要不可欠なコミュニケーションツールとなった．しかし，個人の失言に対し誹謗中傷や非難が殺到してしまうといった負の側面が SNS には存在する．それらは一般に“炎上”と呼称され，現代のインターネットが抱える大きな社会問題の一つである．

そこで我々は，Twitter における炎上事象の抑制を目的とする包括的なシステムを構築した．先行研究として岩崎ら [岩崎 13] によって，人間の介入なく機械的に炎上防止を行うシステムが構想されているが，システムの実現には至っていない．本システムでは，炎上防止のみを目的とするのではなく，炎上の原因となる言語表現が存在するとし，その言語表現を改めるよう情報発信者に促すことで，炎上リスクを抑制する．システムが持つ機能は以下の 3 つに大別される．

(1) 入力テキストに対する炎上可能性の判定

(2) 炎上する表現を訂正したテキストの提案

(3) 予想される返信の提示

収集した炎上事例に対し，交差検定を行ったところ，本システムによって F 値 0.74 で炎上を検知することが可能であった．

連絡先: 大西 真輝，奈良先端科学技術大学情報科学研究科，
〒 630-0192 奈良県生駒市高山町 8916-5,
onishi.masaki.oe3@is.naist.jp

2 炎上事例の分析

岩崎ら [岩崎 13] は，炎上事例を犯罪自慢と価値観の押し付けに分類できるとしている．そのうち犯罪自慢についてはベイジアンフィルタを用いることで 81%の精度で検知が可能であるが，価値観の押し付けは同様な手法では検知が困難であるとしている．

我々は，“炎上事例には分類に関わらず他者を不快にさせる言語表現が含まれている”という仮定をし，過去に Twitter 上で炎上した事例を 100 件収集し分析を行った．このとき，

- そのツイート一つで炎上したと判断できる．
- 画像や動画による炎上ではなく，テキストのみで炎上している．
- 引用（リツイート）数が 20 以上ある．
- まとめサイトなどで取り上げられている．

という 4 つの条件を満たす事例を炎上事例として取り扱った．結果を表 1 に示す．比較のため，ランダムに抽出した一般のツイート 100 件についても同様の観点から分析を行い，表 1 に併記している．表中の「直接的な表現」とは，“バカ，アホ，死ね”といった他者を直接的に侮蔑する表現を指し，「補助的な表現」とは，“(笑)，ｗｗｗ”などの前後の文脈によっては，他者を煽ったり貶しているように捉えられる表現のことを指す．

分析結果から，直接的な表現と補助的な表現を合わせた不快な表現の割合は，一般のツイートの 15%に対して，炎上ツイートは 88%と，炎上ツイートにはより多くの不快な表現が含まれていることが判明した．このことから，単語の表層情報という単純な特徴量から炎上検知が可能であると考えられる．

表 1: 事例中に不快な表現が含まれる割合

事例	直接的な表現	補助的な表現	含まれない
炎上	68%	20%	12%
一般	9%	6%	85%

これは岩崎ら [岩崎 13] の分類における価値観の押し付けについても，同様であることを示唆している．

3 システムの概要

本章ではシステムの主要機能について説明する．システムは，次の 3 つの機能によって，多様な側面からユーザに炎上を抑制する機会を与えることを目的としている．

(1) 炎上検知と原因の指摘

(2) 表現を訂正したテキストの提案

(3) 予想される返信の提示

3.1 SVM による炎上可能性の判定

(1) の機能では，入力テキストに対し，炎上可能性の判定と，炎上の原因となる単語の指摘を行う．炎上検知の手法として，機械学習の一つである SVM を採用した．具体的には，以下の手順に沿って行われる．

手順 1.1 形態素解析機 MeCab [工藤 04] を用いて入力テキストを単語単位に分かち書きする．

手順 1.2 機能語以外の単語の原形と，小林ら [小林 05]，東山ら [東山 08] の評価極性辞書による極性を特徴量として抽出する．単語の原形は unigram から trigram まで用いた．

手順 1.3 手順 1.2 で抽出した特徴を用いて，一般のツイートと炎上ツイートを分類するように学習した SVM によって，入力テキストの炎上の成否を判定する．炎上と判定された場合，分離平面からの距離を用いて炎上可能性を視覚的に表示する．

手順 1.4 手順 1.3 で炎上と判定された場合，入力テキスト中の炎上単語リスト内に該当する単語を指摘する．炎上単語リストは，SVM のモデルデータ内の重みを利用することによって，原因となる単語を収集し作成した．

上記の手順によって，単純に炎上するか否かの 2 値分類を行うのみでなく，炎上可能性の度合いや言語表現を指摘することによって，炎上リスクの明確化や原因を学ぶ事が可能となる．

3.2 word2vec を用いた訂正テキストへの変換

(2) の機能では，炎上の原因であると判断された単語を適切な表現に訂正したテキストに変換する．訂正方法として，炎上の原因と判断された単語を, 意味的に類似する適切な単語へと置換する操作を行う．このとき類似した単語を機械的に獲得するために，単語同士の類似度を word2vec [Mikolov 13] によって計算した．

word2vec の学習には，Twitter で収集した 50,000,000 件の日本語ツイートを利用した．各ツイートに対し，URL やハッシュタグの削除などの前処理を行い，MeCab によって形態素解析を行った．形態素解析の結果から，単語の原型に品詞情報を付与し分かち書きしたものを，word2vec の学習データとして使用した．

具体的な訂正手順は，以下の通りとなる．

手順 2.1 3.1 節の炎上単語リスト内の各単語に対し，word2vec を用いて類似度の高い単語を算出する．

手順 2.2 手順 2.1 で算出した単語の中から炎上単語リストに含まれず，単語の品詞が同一であるという条件の基に，類似度の一番高い単語を採用する．

手順 2.3 手順 2.2 で採用した単語が活用語であった場合は，置換前の単語と同様の活用を行った後に置換し，活用語でない場合は，そのまま置換を行う．

3.3 ニューラルネット言語モデルを用いた返信の生成

(3) の機能では，ユーザが入力したツイートに対して予想される返信を生成し，ユーザにツイートの訂正を促す．返信の自動生成には，ニューラルネット言語モデル [Yoshua 03] を使用した．

n-gram ニューラルネット言語モデルは，式 (1) で表されるように，直前の $n-1$ 単語 $w_{i-n+1}, .., w_{i-1}$ が与えられたときの次の単語 w_i の出現確率を，ニューラルネットによりモデル化する．

$$P(w_i|w_{i-n+1}, .., w_{i-1}) \quad (1)$$

ニューラルネット言語モデルは，文脈長 n が大きい場合に，n-gram の頻度に基づく言語モデルに比べて頑健に単語の出現確率を推定できる．

返信を生成する際は，式 (1) により計算される確率分布から，1 単語ずつサンプリングする．そして，終端記号が生成された時点で文の生成を停止する．

表 2: ニューラルネットのパラメータ

パラメータ	値
ユーザのツイートの語彙数	10000
生成する単語の語彙数	30000
単語ベクトルの次元数	100
隠れ層 h の次元数	500
ユーザのツイートの特徴語数 m	3
文脈長 $n-1$	10

3.3.1 ニューラルネットの構成

本機能では，ユーザが入力したツイートに対する返信の生成を目的とする．そのため，式 (2) のように，次の単語の予測を，直前に生成した単語列 $w_{i-n+1},..,w_{i-1}$ に加えて，ユーザが入力したツイートが含む単語のうち特徴的な m 個の単語 $t_1,..,t_m$ によって条件づける．

$$P(w_i|w_{i-n+1},..,w_{i-1},t_1,..,t_m) \tag{2}$$

まず，ユーザのツイートに含まれる特徴的な単語 $t_1,..,t_m$，直前に生成した単語 $w_{i-n+1},..,w_{i-1}$ を入力とし，それぞれの単語を，ルックアップテーブル C^{tw}，C で単語ベクトルに変換する．次に，式 (3)，(4) で重み行列 W_h から隠れ層 h の値を計算する．

$$C^{tw}_{avg} = \frac{1}{m}\sum_{i=1}^{m} C^{tw}(t_i) \tag{3}$$

$$h = \tanh(W_h(C^{tw}_{avg}, C(w_{i-n+1}),.., C(w_{i-1}))) \tag{4}$$

そして，式 (5) のように，重み行列 W_o と softmax 関数により次の単語の確率分布に変換する．

$$P(w_i|w_{i-n+1},..,w_{i-1},t_1,..,t_m) = \mathrm{softmax}(W_o h) \tag{5}$$

ユーザのツイートに含まれる特徴的な単語を抽出する方法として，単語の頻度に基づくサンプリングを行った．語彙中の各単語は，コーパス中の頻度順に単語 ID が割り振られている．コーパス中で低頻度である語が，ユーザのツイートを特徴付ける単語であるという考えに基づき，単語 ID に比例する確率で，m 個の単語をユーザのツイートから復元抽出する．例えば，ユーザのツイートが“飲酒 運転 なう”であった場合，“飲酒 (ID=2896)”，“運転 (ID=1288)”，“なう (ID=100)”をそれぞれ 2896 : 1288 : 100 の割合で抽出する．

ニューラルネットのパラメータを表 2 に示す．訓練データとして，インターネット上の掲示板データから取得したデータを使用した書き込み-返信対 3,722,927 文対を使用した．訓練において学習されるパラメータは，C，C^{tw}，W_h，W_o である．パラメータの更新には ADADELTA [Zeiler 12] を使用した．

表 3: 炎上検知精度

使用した特徴	*Recall*	*Precision*	*F*1
表層情報のみ	0.68	0.73	0.71
+評価極性辞書	0.74	0.75	0.74

4 評価実験

炎上検知の精度については，収集したツイートデータを用いて交差検定を行うことで評価を行った．訂正したテキストの提案と，予想される返信の提示については機械的な評価が困難であるため，本システムを被験者に使用してもらい，アンケートをとることで評価した．

4.1 炎上検知の精度評価

炎上検知の評価を行うため，人手によって炎上ツイートをさらに 500 件収集した．この時，リツイート数 15 件以上，テキストのみで炎上すると思われるという条件に基づくツイートに対象を限定した．2 章で収集した炎上ツイート 100 件と合わせた計 600 件の炎上ツイート，並びにランダムに抽出した 1,800 件の一般のツイートを実験データとして使用し，5 分割交差検定を行った．結果を表 3 に示す．

結果として，F 値で 0.74 という高い値を得た．表層情報のみの場合に対し，評価極性辞書を用いた場合では，Recall が大幅に改善されていることがわかる．このことから評価極性の情報が，他者を不快にさせる表現の取得漏れを防いでいると考えられる．

4.2 アンケート結果によるシステム評価

成人男性 9 人に実際にシステムを使用してもらい，システムの主観評価を目的としたアンケートを行った．その結果を抜粋して表 4 に示す．

アンケートでは，訂正 · 返信の両機能において，あまり良い評価を得ることができなかった．特に有効性に対する評価が，我々の期待に反する結果となった．我々は各機能の性能向上が，各機能の有効性の評価につながると考えており，より適切な表現訂正と，より尤もらしい返信生成が今後の課題であると考えている．

5 システムの出力とエラー対策

実際のシステムが行う検知や出力の例を表 5 に示す．

表 5 中の S1 では“情弱”が真逆の意味である“情強”に誤って訂正されてしまっている．これは，word2vec では，対義語

表 4: システムのアンケート結果

質問内容	はい	いいえ
訂正テキストに関するアンケート		
訂正された表現は炎上の防止に役立ったと思うか？	44%	56%
訂正されたテキストは、日本語として自然な文章だったか？	56%	44%
訂正後の意味合いは訂正前と変わっていなかったと思うか？	33%	67%
提示された返信に関するアンケート		
返信を見ることによって炎上を抑制する効果があったと思うか？	33%	67%
提示された返信は、日本語として自然な文章だったか？	67%	33%
入力文の内容に則した返信が返ってきたか？	33%	67%

表 5: システムの検出・出力例

	入力テキスト	炎上表現	訂正後の文	生成された返信例
$S1$	情弱とキモオタは死ね！	情弱，キモオタ，死ね	情強とオタクは苦しめ！	笑わす中もらうわ寝るわ
$S2$	飲酒運転なう	飲酒運転	飲酒なう	何かのことね ww

でも文脈中での使用法が類似していれば，類似度が高いとされてしまうためである．対策として，変換候補を複数用意し，変換前と変換後のテキスト間で，含意関係認識を行うなどが考えられる．また，返信の提示では，日本語として意味を成さないテキストが出力されてしまっている．この現象はテキスト生成時に文法規則を取り入れることで，抑制することができるのではないかと考えられる．

また，S2 の“飲酒運転”のような犯罪自慢に関しては，表現を訂正するのではなく，犯罪であると警告する方が好ましい．そのため，犯罪か否かを別の枠組みで判定する必要があると考えられる．

6 おわりに

本研究では，Twitter 上での炎上を抑制するために，3 つの機能からなるシステムを構築した．炎上事例の分析から，炎上事例には他者を不快にさせる特徴的な言語表現が多く存在することがわかった．単語の表層とその評価極性判定のみを特徴量として学習した SVM による実験では，F 値 0.74 で炎上検知が可能であった．訂正したテキストと予測される返信の提示については，アンケート結果での評価が悪く，日本語として尤もらしいテキストを生成することが，今後の課題であると考えられる．

謝辞

本研究は奈良先端科学技術大学院大学情報科学研究科の 2014 年度 Creative and International Competitiveness Project の助成を受けて実施しました．心から感謝の意を表します．

参考文献

[岩崎 13] 岩崎祐貴, 折原良平, 清雄一, 中川博之, 田原康之, 大須賀昭彦: CGM における炎上の同定とその応用, in *The 27th Annual Conference of the Japanese Society for Artificial Intelligence* (2013)

[工藤 04] 工藤拓, 山本薫, 松本裕治: Conditional Random Fields を用いた日本語形態素解析, 情報処理学会自然言語処理研究会 SIGNL-161, pp. 89–96 (2004)

[小林 05] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一: 意見抽出のための評価表現の収集, 自然言語処理, pp. 203–222 (2005)

[東山 08] 東山昌彦, 乾健太郎, 松本裕治: 述語の選択選好性に着目した名詞評価極性の獲得, 言語処理学会第 14 回年次大会論文集, pp. 584–587 (2008)

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space., in *In Proceedings of Workshop at ICLR* (2013)

[Yoshua 03] Yoshua, B., Réjean, D., Pascal, V., and Janvin, C.: A neural probabilistic language model, *The Journal of Machine Learning Research*, Vol. 3, pp. 1137–1155 (2003)

[Zeiler 12] Zeiler, M. D.: ADADELTA: an adaptive learning rate method, *arXiv preprint arXiv:1212.5701* (2012)