情報処理学会論文誌　データベース　Vol.12　No.1　1–5 (Jan. 2019)

テクニカルノート

叫喚ツイート抽出手法の拡張

熊本 忠彦[1,a)]

受付日 2018年6月4日, 採録日 2018年7月31日

概要：マイクロブログの 1 つである Twitter では，突発的な感情の発露を表すために，「日本が勝ったぁああ」や「それはやめて～～ーー」のような叫喚表現化したツイートを用いることがある．これまでの先行研究では，こういった叫喚ツイートの叫喚表現化された部分を検出し，元の表現（「勝った」や「やめて」）に変換することで，既存の様々な辞書を利用できるようにするための手法やツイートから叫喚ツイートを抽出し，投稿者の感情の大きな変化を検知するという手法が提案されている．しかしながら，抽出される叫喚ツイートの種類についてはあまり深く検討されておらず，比較的単純な正規表現により検索可能な叫喚ツイートのみが抽出されていた．そこで本論文では，先行研究で提案された正規表現を拡張し，より多くの叫喚表現に対応できるようにするともに，提案手法による叫喚ツイートの抽出割合や抽出精度を評価することで，その有効性を検証する．

キーワード：ツイッター，叫喚，中村明の基本感情

On a Method for Extracting More Shouting Tweets from Twitter

Tadahiko Kumamoto[1,a)]

Received: June 4, 2018, Accepted: July 31, 2018

Abstract: On Twitter, one of microblog services, shouting tweets like “Our national soccer team wooooon!” and “Stoooop it!!!” are often used in order to express a sudden emotional effusion. In some previous studies, methods for identifying word lengthening in tweets and transforming them into their original expressions have been proposed so that a variety of existing lexicons and dictionaries can be available. In other previous studies, methods for extracting shouting tweets from Twitter and detecting a huge change of emotions of the users who posted the tweets have been proposed. In these studies, however, what kinds of word lengthening should be detected is not considered very deeply. They have used comparatively simple regular expressions to extract shouting tweets from Twitter. This paper, therefore, proposes six regular expressions to extract more shouting tweets from Twitter, and shows effectiveness of the regular expressions by calculating extraction ratio of shouting tweets and accuracy on extraction of the shouting tweets.

Keywords: Twitter, word lengthening, Akira Nakamura’s basic emotions

1. はじめに

マイクロブログの 1 つである Twitter では，突発的な感情の発露を表すために，ツイートの一部（文末や文中の母音，長音記号など）を繰り返すことで，「日本が勝ったぁああ」や「それはやめて～～ーー」のような叫喚を表現することがある．このような叫喚表現は，ほとんどの場合，辞書に登録されていないため，形態素解析や単語出現頻度の算出といった各種テキスト処理の妨げとなる．そのため，先行研究 [1], [2] では，ツイートの叫喚表現化された部分を正規形（叫喚表現を含まない表現）に変換することで，既存の様々な辞書を利用できるようにするための手法が提案されている．一方，叫喚ツイート（叫喚表現を含むツイート）を抽出し，投稿者の感情の大きな変化を検知するという研究 [3], [4] や叫喚ツイートとそうでないツイートでは受ける印象がどのように異なるかを分析した研究 [5] も行われており，叫喚表現に関して様々な研究が行われているこ

[1] 千葉工業大学
Chiba Institute of Technology, Narashino, Chiba 275–0016, Japan
[a)] kumamoto@net.it-chiba.ac.jp

とが分かる．

しかしながら，いずれの研究においても，対象となる叫喚表現の種類についてはあまり深く検討されておらず，比較的単純な正規表現により検索可能な叫喚ツイートのみが抽出されていた．そこで本論文では，浅井らの手法 [2] で用いられていた正規表現を拡張し，より多くの種類の叫喚表現に対応できるようにするとともに，提案手法による叫喚ツイートの抽出割合や抽出精度を評価することで，その有効性を検証する．

以下に本論文の構成を示す．まず，2 章で関連研究について述べ，本論文の新規性を示す．次に，3 章で Twitter からツイートを収集し，浅井らの手法 [2] を用いて叫喚ツイートを抽出するとともに，4 章で叫喚表現ではないと判定されたツイートを分析し，新たな叫喚ツイートを抽出するための正規表現を提案する．さらに，5 章で提案手法による叫喚ツイートの抽出割合と抽出精度を評価し，浅井らの手法 [2] と比べることで，その有効性を検証する．最後に，6 章で本論文のまとめと今後の課題について述べる．

2. 関連研究

先行研究 [1], [2] では，ツイートの叫喚表現化された部分を正規形に変換するための手法が提案されている．たとえば，Brody らは，英語のツイートを対象に，単語の一部が 3 回以上繰り返された「niiiice」や「reallly」のような叫喚表現を検出し，正規形（「nice」や「really」）に変換する手法を提案している [1]．浅井らは，日本語のツイートを対象に，同じ母音が 3 回以上繰り返された「うわあああああああ」や「ねむいいぃぃぃぃぃ」のような叫喚表現を検出し，正規形（「うわあ」や「ねむい」）に変換する手法を提案している [2]．

一方，叫喚ツイートを抽出することで，投稿者の感情の大きな変化を検知するという研究 [3], [4] もある．たとえば，高橋らは，日本語のツイートからひらがなとカタカナの 5 回以上の繰返しもしくは記号「！」の 7 回以上の繰返しを抽出することで，投稿者の感情が大きく表れたツイートを高い精度（F 値 0.786）で抽出できることを示している [3]．山本らは，ニコニコ動画に登録されている楽曲動画に対し投稿されたコメントから形容詞（形容動詞を含む）や日本語文字の 3 回以上の繰返し，サビ区間中に投稿されたコメントを抽出し，サポートベクタマシン（SVM）に対する素性として利用することで，その楽曲動画の印象（cute, sorrow, cheerful, fresh, cool, aggressive, darkness）を比較的高い精度（印象別 F 値 0.535〜0.758）で推定できることを示している [4]．

また，ツイートの一部を叫喚表現化することにより，ツイートから受ける印象がどのように変化するかを分析した研究 [5] もある．たとえば，熊本は，叫喚表現（文末や文中の母音や長音，記号などの繰返し）を含むツイートの印象と叫喚表現を正規化したツイートの印象を 267 人の Twitter ユーザが参加するアンケート調査に基づいて調べ，その結果，叫喚表現化により，(1) 20 代では「嫌い」が強くなる，(2) 30 代では印象は変わらない，(3) 40 代では「嫌い」と「恥ずかしい」が強くなるといったことや (4) 男性では「嫌い」と「恥ずかしい」が強くなる，(5) 女性では印象は変わらないといったことを明らかにしている．

しかしながら，いずれの研究においても，比較的単純な正規表現によって検索可能な叫喚ツイートのみが対象となっていた．本論文では，より多くの叫喚表現を検索できるよう正規表現を拡張するとともに，その抽出割合と抽出精度を評価し，既存手法 [2] と比較している点が新しい．

3. ツイートの収集と叫喚ツイートの抽出

本章では，Twitter からツイートを収集し，浅井らの手法 [2] を用いて叫喚ツイートを抽出する．

表 1 に示したように，まず，2017 年 6 月 2 日〜4 日の 3 日間，Twitter Streaming API を用いて Twitter からツイートを収集した．その結果，約 200 万個のツイートを収集することができたが，この中にはいわゆるリツイートも含まれている．リツイートに叫喚表現が含まれている場合，そのオリジナルのツイートも収集されてしまう可能性があり，2 重にカウントされる恐れがある．そこで，リツイートは公式リツイートか非公式リツイートかに関係なく両方とも除外することにした．結果，ツイートの数は約 150 万個になった．この約 150 万個のツイートから 20,000 個のツイートをランダムにサンプリングし，浅井らの手法 [2] を用いて叫喚ツイートを抽出したところ，212 個の叫喚ツイートが抽出された．抽出割合は 1.06%ということになる．なお，浅井らは同じ母音の 3 回以上の繰返しを叫喚表現と定義しており，**表 2** に示した正規表現を採用している．

表 1　叫喚ツイート収集に関する予備実験
Table 1　Preparatory experiment for collecting shouting tweets.

総収集ツイート数	2,040,844
RT 除外ツイート数	1,476,899
ランダムサンプリングされた 20,000 ツイートから浅井らの手法 [2] により抽出された叫喚ツイート数（抽出割合）	212 （1.06%）

（注）ツイート収集期間：2017 年 6 月 2 日〜4 日

表 2　浅井らの手法 [2] で用いられた正規表現
Table 2　Regular expression used in Asai et al.'s method [2].

```
あ{3,}|い{3,}|う{3,}|え{3,}|お{3,}|ぁ{3,}|ぃ{3,}
|ぅ{3,}|ぇ{3,}|ぉ{3,}|ア{3,}|イ{3,}|ウ{3,}|エ{3,}
|オ{3,}|ァ{3,}|ィ{3,}|ゥ{3,}|ェ{3,}|ォ{3,}
```

表 3　提案手法で用いる正規表現
Table 3　Our proposed regular expressions.

叫喚表現タイプ	正規表現
母音の繰返し	[あぁアァアァ]{3,}｜[いぃイィイィ]{3,}｜[うぅウゥウゥ]{3,}｜[えぇエェエェ]{3,}｜[おぉオォオォ]{3,}
長音記号の繰返し	(ー｜～｜—｜-｜‐){3,}
「ん」/「ん”」の繰返し	[んンンン]{2,}｜([んンンン][”゛゙_]){2,}
小さい「っ」の繰返し	[っッッ]{2,}[^ぁ-ケヲ-ン]｜[っッッ]{2,}$
濁点付母音の繰返し	([ぁ-おァ-オァ-ォア-オ][”゛゙_]){3,}
^＋長音記号	([ぁ-ケヲ-ン](\^｜\ ^)(ー｜～｜—｜-｜‐))$｜[ぁ-ケヲ-ン](\^｜\ ^)(ー｜～｜—｜-｜‐)[^\^\ ^]

（注）下線を引いてあるカタカナは半角文字であることを表している．

4.　叫喚ツイート抽出用正規表現の提案

前章で述べたように，浅井らの手法 [2] では，ランダムサンプリングにより得た 20,000 個のツイートから 212 個の叫喚ツイートを抽出することができた．そこで本章では，浅井らの手法により叫喚ツイートではないと判定された残り 19,788 個のツイートから叫喚しているツイートを抽出し，場合分けすることで，主なものを抽出するための正規表現を設計することにした．結果，**表 3** に示した 6 個の正規表現を得た．それぞれの正規表現を「|」でつなげることにより，一括での正規表現検索も可能である．

表 3 において，「母音の繰返し」は浅井らが用いた正規表現を拡張したものとなっている．浅井らの正規表現では「あああ」や「オオオ」のように同じ母音が 3 回以上繰り返されている表現が抽出されるが，著者が提案する正規表現では「ああア」や「ォォオ」のように読みが同じなら文字種（ひらがな，カタカナ，大文字/小文字，全角文字/半角文字）が異なる母音でも 3 回以上繰り返されている場合には叫喚表現として抽出することができる．「長音記号の繰返し」は 5 種類の長音記号（ー，～，—，-，‐）の任意の組合せ（3 文字以上）を対象としている．『「ん」/「ん”」の繰返し』は，「ん」もしくは濁点付きの「ん」が 2 回以上繰り返された場合，その表現を叫喚表現として抽出するというものである．『小さい「っ」の繰返し』は，「っ」や「ッ」もしくは半角カタカナの「ッ」が 2 回以上繰り返された場合を対象としているが，ひらがなやカタカナが続く場合は除外される．「ん」や「ん”」，「っ」の繰返しは，文法的な日本語文章ではほとんど見当たらないため，本論文では 2 回以上の繰返しを叫喚表現とした．「濁点付母音の繰返し」は，一部の漫画やアニメで使われるようになった「あ”」や「ウ”」など濁点を付与された母音が任意の組合せで 3 回以上繰り返された場合に，その表現を叫喚表現として抽出するというものである．「あ”」などの繰返しも文法的な日本語文章ではほとんど見当たらないが，「あ”あ”」のような 2 回の繰返しは間投詞としてとらえる方が好ましい場合もあると考え，本論文では 3 回以上の繰返しを叫喚表現とした．また，濁点付母音の特殊性に鑑み，「ウ”あ”あ”」や「ぅ”お”あ”」のように異なる母音の組合せでも叫喚表現として抽出することにした．「^＋長音記号」は，ひらがなもしくはカタカナの後に「^」と長音記号（ー，～，—，-，‐）が続く場合を対象としているが，さらに「^」が続く場合は除外される．これは，「(^ー^)」のような顔文字を抽出しないためである．なお，表 3 中の下線が引かれたカタカナは半角文字であることを表している．

5.　性能評価

本章では，提案手法の性能を評価するために，浅井らの手法（表 2 参照）と提案手法（表 3 参照）による叫喚ツイートの抽出割合と抽出精度を評価し，比較する．

まず，叫喚ツイートの抽出対象となるツイートを取得した．具体的には，表 1 に示した RT 除外ツイート（1,476,899 個）の中から 2017 年 6 月 2 日に収集されたものを抽出し，464,563 個のツイートを取得した．

次に，この 464,563 個のツイートを対象にそれぞれの手法を用いて叫喚ツイートを抽出した．その結果を**表 4** に示す．表 4 によれば，浅井らの手法では 5,499 個（1.18%）の叫喚ツイートを抽出できたが，提案手法では 12,894 個（2.78%）と 2 倍以上の叫喚ツイートを抽出できており，抽出数で 7,395 個，抽出割合で 1.59 ポイントの増となっていることが分かる．ここで，提案手法における各正規表現の貢献を示すために，それぞれの正規表現による叫喚ツイートの抽出数と抽出割合を調べた．結果を**表 5** に示す．一番抽出数が多かったのは「母音の繰返し」であり，同じ母音の繰返しである浅井らの手法（表 4 参照）と比べても，抽出数で 1,118 個，抽出割合で 0.24 ポイントの増となっている．次に抽出数が多かったのは「長音記号の繰返し」であり，「母音の繰返し」に匹敵する 5,589 個（1.20%）の叫喚ツイートを抽出できている．他の 4 つの正規表現については，抽出数が 103～617，抽出割合が 0.02%～0.13%と，あまり多くなかった．なお，表 5 の計が表 4 の値と合わないのは，複数の正規表現にヒットするツイートが一定数あったためである．

次に，各正規表現の抽出精度を評価した．具体的には，各正規表現により抽出された叫喚ツイート（表 5 参照）か

表 4　各手法により抽出された叫喚ツイート数
Table 4　Number of shouting tweets extracted by each method.

	抽出数	抽出割合（%）
浅井らの手法 [2]	5,499	1.18%
提案手法	12,894	2.78%
（増分）	+7,395	+1.59 ポイント

表 5　提案手法の各正規表現により抽出される叫喚ツイート数
Table 5　Number of shouting tweets extracted by each of our proposed regular expressions.

	抽出数	抽出割合（%）
母音の繰返し	6,617	1.42%
長音記号の繰返し	5,589	1.20%
「ん」/「ん”」の繰返し	617	0.13%
小さい「っ」の繰返し	554	0.12%
濁点付母音の繰返し	188	0.04%
^+長音記号	103	0.02%
計	13,668	

表 6　各正規表現における叫喚ツイートの抽出精度
Table 6　Accuracy on extraction of shouting tweets by each regular expression.

	抽出精度（%）
浅井らの手法 [2]	97%
母音の繰返し	96%
長音記号の繰返し	97%
「ん」/「ん”」の繰返し	97%
小さい「っ」の繰返し	96%
濁点付母音の繰返し	100%
^+長音記号	100%

らランダムサンプリングによりツイートを 100 個ずつ抽出し，叫喚表現を含んでいるか否かを判定した．その結果を**表 6** に示す．

浅井らの手法と「母音の繰返し」では，「おおんおおおん」や「メェェェ」のような擬音語を叫喚表現ではないと判定した．「長音記号の繰返し」では長音記号を単なる直線/波線として用いている場合，叫喚表現ではないと判定した．『「ん」/「ん”」の繰返し』では，「（略）カモンカモ〜ンんーフフフフバイバーイ！」[*1] や「（略）今まではオスメス交互に運んんできてるものと（略）」のように偶然「ん」が繰り返された場合やタイポにより「ん」が繰り返された場合を叫喚表現ではないと判定した．なお，「んん」には「んん」を間投詞（感嘆詞）として用いている場合があり，もし間投詞として用いられた「んん」を叫喚表現ではないと判定するならば，抽出精度は 77%になる．ただし，「ん」が 3 回以上繰り返された場合や「んん〜」のように長音記号と組み合わされた場合は，叫喚表現と判定する．『小さい「っ」の繰返し』では「ヴァックショ————ッッ」などの擬音語を叫喚表現ではないと判定した．「濁点付母音の繰返し」と「^+長音記号」では抽出精度は 100%であった．抽出割合の低い特殊な言語表現であり，そのため，抽出精度が高くなったものと考えられる．ツイート例としては「あ゛あ゛あ゛！！！布団あったけえ！！！！！」や「はぇ^〜すっごい安い…」などがあげられる．

[*1] 実際のツイートでは「フフフフ」は半角文字であった．

6. おわりに

本論文では，Twitter から「日本が勝ったぁああ」や「それはやめて〜〜ーー」のような叫喚ツイートを抽出するために，6 個の正規表現を提案した．それぞれ，「母音の 3 回以上の繰返し」，「長音記号の 3 回以上の繰返し」，『「ん」/「ん”」の 2 回以上の繰返し』，『小さい「っ」の 2 回以上の繰返し』，「濁点付母音の 3 回以上の繰返し」，「^+長音記号」を検索するための正規表現であり，それぞれの正規表現を「|」でつなげることにより，一括での正規表現検索も可能となる．

今後の課題として，抽出した叫喚表現を正規形に変換するための手法の開発があげられる．これまでの先行研究 [1], [2] でもそのような手法は提案されているが，実際のツイートにおける叫喚表現化はより複雑に行われており，十分にカバーしているとはいえない．今回提案した各正規表現に対し，より高精度な変換手法を開発する必要があるといえる．

謝辞　本研究の一部は JSPS 科研費 17K00430 による研究助成の成果であり，ここに記して感謝の意を表す．

参考文献

[1] Brody, S. and Diakopoulos, N.: Cooooooooooooooollllllllllll!!!!!!!!!!!!!!: Using Word Lengthening to Detect Sentiment in Microblogs, *Proc. Conference on Empirical Methods in Natural Language Processing*, Edinburgh, United Kingdom, pp.562–570 (2011).
[2] 浅井洋樹，秋岡明香，山名早人：きたあああああああああああああああ！！！！！ 11：マイクロブログを用いた教師なし叫喚フレーズ抽出，第 5 回データ工学と情報マネジメントに関するフォーラム（DEIM2013），A4-4 (2013).
[3] 高橋雄太，片岡義雅，浅井洋樹，山本祐輔，秋岡明香，山名早人：繰り返し表現を含んだ感情的なツイートの抽出，第 4 回データ工学と情報マネジメントに関するフォーラム（DEIM2012），C3-6 (2012).
[4] 山本岳洋，中村聡史：視聴者の時刻同期コメントを用いた楽曲動画の印象分類，情報処理学会論文誌データベース，Vol.6, No.3, pp.61–72 (2013).
[5] 熊本忠彦：叫喚表現化によるツイート印象の変化の分析，ARG Web インテリジェンスとインタラクション研究会，No.8, WI2-2016-1, pp.1–6 (2016).

熊本 忠彦　（正会員）

千葉工業大学情報科学部情報ネットワーク学科教授．1988 年筑波大学第三学群情報学類卒業．1990 年同大学大学院修士課程修了．同年郵政省通信総合研究所（現，国立研究開発法人情報通信研究機構）入所．2007 年千葉工業大学准教授を経て，2010 年より同大学教授．感性情報処理とその応用に関する研究に従事．1996 年博士（工学）（筑波大学）．FIT2004 論文賞，IMECS 2010/2014 Best Paper Award 各受賞．電子情報通信学会，人工知能学会，日本データベース学会，日本知能情報ファジィ学会，日本感性工学会各会員．

（担当編集委員　榊 剛史）