6P-07

ツイートに含まれるイベント名称の自動認識

佐久間 盛貴* 町田 翔* 延澤 志保*
*東京都市大学知識工学部 *東京都市大学大学院工学研究科

1 はじめに

近年，Social Networking Service(以下 SNS) の普及により多くの人々は手軽につながりを作りコミュニケーションを取ることが可能となっている．現在，日本国内で利用者が多い代表的な SNS といえば Twitter や Facebook, Instagram，LINE 等が挙げられる．Twitter はツイートと呼ばれる短文の形で簡単に利用者の受け取りたい情報が閲覧できる．

Twitter の利用者数は年々増加しており [1]，得られる情報も膨大なものになっている．Twitter の利点のひとつに，イベントなどの参加者の生の声をリアルタイムに閲覧できる点が挙げられる．これを活用すれば，イベントの名称や会場などの属性的な情報だけでなく，その場で内容や感想などを参考にしながらイベントに参加したり，参加者の感想を参考にして参加を検討したりすることが可能となる．そこで本研究では，ツイートに含まれるイベント情報の自動抽出を行うため，ツイート中のイベント名称の自動認識を行う手法を検討する．

2 イベント名称認識の課題

イベント名称は一般に,「東京コミックコンベンション 2017」など，イベントの内容 (「イルミネーション」や「フェスティバル」)，会場や地域，時期などで構成される複合語である．イベント名称の構成要素および出現順序に確固とした規則性はなく,「第 21 回 2017 神宮外苑いちょう祭り」のようにイベント名称と認識しやすいものから「視線で花咲くアート展」のように自動での認識が困難と考えられるものまで趣向の凝らされたものも多い (表 1) ．表 1 に挙げたイベント名称すべてに対応できるようなパターンの作成は困難であり，またイベント名称の構成に規則がないことからも，イベント名称すべてに対応可能なパターンの作成は難しいものと考えられる．

表 1: イベント名称の例

山梨ぶどうスイーツ収穫祭
視線で花咲くアート展
進撃の巨人 体感型ゲーム“ 解 ”～技術都市奪還作戦～
TOKYO TOWER HIGHBALL GARDEN
HAPPY！ハロウィン・スイーツ 2017
東京コミックコンベンション 2017
六本木開館 10 周年記念展 天下を治めた絵師 狩野元信
miyazaki イルミネーション in 2017
しな水ハッピーハロウィン 2017
第 21 回 2017 神宮外苑いちょう祭り
いけばな龍生展 2017 — IKE-BUZZ —
THE ドラえもん展 TOKYO 2017
2017 年赤坂ル・アンジェ教会 クリスマスイルミネーション

また，ツイート中ではイベント名称を略す事例も多い．例えば「東京コミックコンベンション 2017」についてのツイート 339 件を調べたところ，その中で出現した 368 個のイベント名称のうち 27.2%が「東京コミコン 2017」，24.7%が「東京コミコン」，22.8%が「コミコン」と略されていた反面，名称をそのまま記述していたのは 11.4%に留まった．また，英字表記である「TokyoComicCon」も 13.6%存在していた (図 1)．このように，ツイートでは

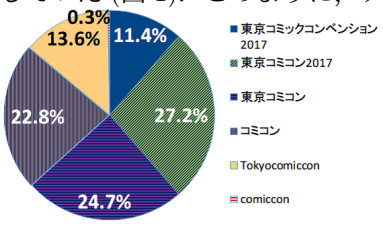

図 1: イベント名称のツイッターでの出現例

複数の略称や通称が混在することがある．これらのイベント名称の認識はパターンマッチングでは不十分であり，イベント名称か否かを推定する手法が必要である．

略称や通称を正しく扱うためには，それらの名寄せの処理が必要となる．山田らは最長共通部分列比 (Longest Common Subsequence Ratio: LCSR) を用いてイベント名称の名寄せを行う手法を提案している [2]．共通部分列とは，2 つの列において連続または非連続にかかわらず同じ要素が同じ順序で出現した部分列のことを指し，共通部分列のうち最も長いものを最長共通部分列と呼ぶ．山田らは 2 つのイベント名称の最長共通部分列長を長い方のイベント名称の文字数で割った最長共通部分列比を予め設定した閾値と比較し，名寄せ処理を行っている [2]．

3 出現語に着目したイベント名称の推定

3.1 提案手法の概要

イベント名称の構成要素は自由度が高く，名寄せより先に，まずイベント名称部分の推定を行う必要がある．

本稿では，イベント名称中に出現する語とその周辺に出現する語とに着目し，イベント名称を含むツイートとそれ以外のツイートを自動判別する手法を提案する．本手法により，イベント名称やその略称，通称などを含むイベントを認識することが可能となれば，ここからイベント名称部分を抽出することで，さまざまな形態のイベント名称に対応することが期待できる．

Automatic Recognition of Event Names Embedded in Tweets
Moritaka Sakuma*, Sho Machida*, and Shiho H.Nobesawa*.
* Faculty of Knowledge Engineering, Tokyo City University
* Graduate School of Engineering, Tokyo City University

3.2 イベント名称を含むツイートの特徴

イベント名称を含むツイートは,「イベント会場に行ったら混んでいた」や「明日のイベントの待ち合わせ」など，イベントに絡む行動が記述されていることが多い．また，イベント名称自体も，自由度は高いとはいえ，地名 (「東京」など) や年 (「2017」など)，イベントの内容を示す語 (「展」など) のように特徴的な語の並びであることが多い．表 2 に，ツイートの例を挙げる．正例とラベルの振られているツイートはイベント名称を含むもので，イベント名称を含まないツイートを負例とする．表 2 中のイベント名称を下線部で示す．表 2 の「イルミリオン」はイルミネーションイベントの名称だが，一般的な語とは言えない．また「ムーミンマーケット 2018」も，構成要素である「ムーミン」「マーケット」「2018」はそれぞれイベントと決められるものではなく，これらの並びが複合語としてひとつのイベント名称を成している．それに対して表 2 の負例に含まれる「イルミネーション」は，イベント名称に頻繁に含まれる語だが，これらの例ではイベント名称ではなく一般的な語として扱う必要がある．

表 2: ツイートの正例と負例

ラベル	分類されたツイート
正	さがみ湖イルミリオン すごかったー
正	横浜そごうの ムーミンマーケット 2018 にて。ムーミンハウス！
負	イルミネーション見に行きたい (n 回目)
負	明日せっかく女の子やるからイルミネーション撮影がんばるぞ

3.3 イベント名称を含むツイートの自動推定

本研究では，イベント名称を含むツイートの自動推定を行うため，Support Vector Machine(SVM) を用いる．SVM は高次元特徴空間において線形関数の仮説空間を用いる学習システムである [3]．あらかじめ用意した正例，負例データから特徴を学習し，各データ点から距離が最大となる分離平面を決定する．この学習データから未学習データに対しても高い認識能力を得ることが出来る．

本研究では，イベント名称を含むツイートとそれ以外のツイートそれぞれについて，構成要素である形態素を調べて各語の出現頻度を値とするベクトルを作成し，これを素性として機械学習を行う．

4 実験結果と考察

本研究では，Twitter から得たツイートを本文中のイベント名称の有無によって正例，負例に分類したものをコーパスとして，イベント名称を含むツイートの認識実験を行った．本研究では，Walker+[4]，レッツエンジョイ東京 [5] の 2 つのイベント情報サイトに記載のあるイベントを認識対象とする．上記のイベントサイトからイベント名称の抽出を行い，抽出されたイベント名称を用いてツイートを検索し取得する．このようにして得たイベント名称を含むツイートを正例，ランダムに得たツイートのうちイベント名称を含まないものを選んで負例とし，イベント名称の認識実験を行った．

ツイート 1,224 件を対象としたイベント名称認識実験の結果，本システムの出力は 88.2%の正解率を得た (表 3)．表 3 で，システム出力が「正」であるとは，入力ツイートがイベント名称を含むと判定されたことを,「負」とはイベント名称を含まないと判定されたことを示す．すなわち，正例を正とした 298 ツイート，負例を負と出力した 823 ツイートがそれぞれ正解であり，これらの和の全体に対する割合が 88.2%である．表 3 中の割合表記はそれぞれ正例全体，負例全体でのそれぞれの出力の割合を指す．正例の正解は 87.9%，負例の正解は 88.4%とほぼ同等の正解率を示しており，本手法は正例，負例の双方について高い正解率を実現したと言える．

表 3: イベント名称認識結果

	システム出力				
	正		負		計
正例	298	(87.9%)	41	(12.1%)	339
負例	103	(11.6%)	782	(88.4%)	885
計	401		823		1,224

表 4 にシステムの出力例を挙げる．

表 4: システム出力例

正解	出力	ツイート
正例	正	キモい展 に行ったら本当にキモかった。#キモい展
正例	負	今日から 5 日まで象の鼻パーク周辺にて開催される スマートイルミネーション横浜 2017 のアワード部門に参加しています。
負例	正	今日は久々に病棟のお仕事があったので病棟行ったらイルミネーション凄かった (/ ´△ ‘＼)

5 まとめ

本論文では，Twitter など SNS の膨大な数のユーザー投稿から興味のあるイベント情報を検索する際の補助として，投稿内容からイベント名称の有無を認識する手法を提案した．

イベント名称は構成要素の自由度が高く，認識は難しい．さらに，ツイート中ではイベント名称をそのまま記述せず略称や通称とする事例も多く，イベント名称をパターンマッチングで検索するのは無理がある．本研究では，さまざまなイベント名称に高頻度で出現する形態素に着目することで，88.2%の正解率を得た．

参考文献

[1] Social Media Lab, ”2017 年 11 月更新！ 11 のソーシャルメディア最新動向データまとめ”, https://gaiax-socialmedialab.jp/post-30833/.

[2] 山田 渉, 菊池 悠, 落合 桂一, 鳥居 大祐, 稲村 浩, 太田 賢, ”マイクロブログを用いたイベント情報抽出技術 ”, 情報処理学会論文誌, Vol.57, No.1, p.123-132, 2016.

[3] Nello Cristianini, John Shawe-Tylor 著, 大北 剛 訳, ”サポートベクターマシン入門 ”, 共立出版, 2005.

[4] Walker+: https://www.walkerplus.com/.

[5] レッツエンジョイ東京: https://www.enjoytokyo.jp/.