DEIM2021 J14-4

会話 tweet を用いた観光口コミ抽出手法

伊東　駿哉†　　小林　亜樹††

† 工学院大学情報学部情報通信工学科　〒 163–8677 東京都新宿区西新宿 1–24–2
†† 工学院大学情報学部情報通信工学科　〒 163–8677 東京都新宿区西新宿 1–24–2
E-mail: †j017022@ns.kogakuin.ac.jp, ††aki@cc.kogakuin.ac.jp

あらまし　観光口コミ収集目的などで，Twitter から観光地に関する tweet を抽出する手法が求められている．地名，位置情報を用いる古典的な手法，それらを核に時系列，フォロー関係，共起語により拡張する手法などが研究されてきた．これに対し本論文では，古典的な手法による tweet 集合を核として，主に tweet の会話関係を取り込んで対象 tweet を拡張する手法を提案する．会話（thread tree）が直接的な言及を続けている可能性が高いのに対して，共起語ベースの拡張では関係のない tweet を多数拡張することが予想されるため，これらを比較し，フィルタリング手法についても検討する．最後に，事例ベースで提案手法の有効性を評価する．

キーワード　Twitter, 情報抽出, 観光情報

1 はじめに

2019 年にインターネットの普及率が 90% を超え，情報検索の為であったり，SNS を利用したりと全世代に幅広くインターネットが利用されている [1]．JTB 総合研究所によるとインターネットの普及率に伴って旅行中にガイドブックではなく WEB ページやマイクロブログを用いて，観光地について検索している人が増えている [2]．マイクロブログの一種である Twitter では，tweet と呼ばれる 140 文字以内の短いテキストを投稿することができるサービスで，手軽にユーザ自身が体験したことや身の回りで起きた出来事や話題である情報やニュースがリアルタイムに近い形で投稿されていたり，他のユーザへ返信を行うことが出来る Reply 機能が実装されている．2019 年第 1 四半期の時点で，3 億 3,000 万人の月間アクティブユーザー数を記録している [3]．

Twitter にはキーワード検索機能があるが，ユーザが入力する必要があり，tweet 内に検索語が含まれている tweet しか見ることが出来ない．例えば，観光地へ行くとなり，周辺情報の評判や情報を調べようと思った際に，公式の検索機能から観光地名で検索すると，検索単語である観光地名が含まれている tweet は得ることは出来るが，観光地に対する評判や情報，感想がある tweet 以外も得てしまい，公式検索機能には限界がある．そこで，特定の観光地名が含まれている tweet から欲しい情報が含まれている tweet のみを抽出出来ればと考えた．そこで，特定の話題に対して検出する手法については地名や位置情報を用いる古典的な手法，それらを核に時系列，フォロー関係，共起語により拡張する手法などが研究されてきたが，それらにも限界がある．

本論文は古典的な手法による tweet 集合を核として，主に tweet の会話関係を取り込んで対象 tweet を拡張する．拡張された tweet 集合から観光地名が含まれていて，かつ否定的，肯定的な人の感情語 (楽しいやつまらない，綺麗など) が共起している tweet の抽出を行う．それにより，観光地に対する評判や情報，感想がある tweet が抽出できるのではないかと考える．本論文では会話が直接的な言及を続けている可能性が高いのに対して，単 tweet からとの比較を行い，提案手法の有効性を評価する．

2 関連研究

Twitter から，文書中から情報の検出を試みる研究は数多くある [4] [5] [6] [7]．これらは観光地毎に人々がどのような感想，評価を行っているのかを知ることが出来るような tweet を取得することを目的とする本研究とは異なり，観光地名が含まれていない tweet から観光地に言及している tweet を抽出しようとする研究 [4] [5]，本研究と似た目的を掲げるものの抽出手法の異なる研究 [6] また，観光地とは無関係に tweet のリプライ関係に着目した情報抽出を提案しているもの [7] に分類される．

渡邉ら [4] は，観光地名が含まれていない tweet から観光地に関する感想を抽出したいという目的から 2 種類の実験を行った．1 つ目は Twitter Streaming API で取得した tweet から，観光地名を含む tweet の特徴語を抽出し，その特徴語を含む抽出まで，2 つ目は Twitter 社が提供している公式のキーワード検索を用いて観光地名が含まれている tweet を検索し，tweet した前後の tweet と，観光地名が含まれている tweet をしたアカウントのフォロワの tweet，観光地名が含まれている tweet の前後の tweet において観光地名を含まない画像付き tweet に関する感想があるかを人手で評価する．という 2 種類の実験を試みた．2 種類の実験から通常の検索よりも多くの感想が抽出する事が出来，観光地名を含んでいない tweet にも観光地に対する感想が含まれていることが分かった．本論文とは観光地に対する tweet を抽出したいという点で関連しているが，特徴語の抽出や公式のキーワード検索から tweet を抽出したという点が本研究とは異なる部分である．

小原ら [5] は，tweet から地域連想語，パターンマッチング

を用いて観光情報の抽出，ユーザの居住地の把握を目的としている．ここでの地域連想語とは，地名や特産品，施設名のように特定の都道府県を連想することができる単語のことを示す．TwitterAPI で取得した tweet を用いており，その際に retweet やアプリを用いて tweet されたものは除去している．得られた tweet に対して形態素解析を行い表記や品詞情報を取得．地域連想語辞書を用いて tweet に含まれる地域連想語を取得し形態素解析した結果と地域連想語を用いてパターンマッチングをおこない，観光地に関する tweet を取得．結果として，502 件の Tweet を抽出することができた．tweet の中には，他人の行動に関する Tweet や天気・災害に関する Tweet を誤って取得してしまっていた．本論文とは観光地に対する tweet を抽出する手法が関連しているが，tweet から地域連想語を用いていなかったり，目的が異なる部分である．

免田ら [6] は，Twitter に投稿された観光地に関する tweet を利用し，観光地の情報の推薦を行うシステムの開発を行うことを目的としている．tweet の抽出方法としては Java のライブラリにある Twitter4j [8] を用いて文章データを取得したそこから Sen という形態素解析機を使用し，その中であらかじめ京都の地名・建造物名を入れた固有名詞辞書というのを用いてマッチングを行い，キーワードである金閣寺という固有名詞を含む tweet を抽出，データベースに格納．金閣寺という固有名詞が含まれている tweet から無造作に 40 件選び，どれくらいポジティブかネガティブかという評価をしていた．本論文とは同じ観光地に関する tweet を利用するという点で関連しているが，tweet の抽出方法と固有名詞辞書というのを用いてマッチングを行っているのが異なる部分である．

また藤田ら [7] は，一般的なテキスト処理手法である LDA を用いてトピック分析を行い，K-means によるクラスタリングを行う．このとき，tweet 群を reply 関係で結ぶことで，特定の話題について言及していると考えられる tweet を，会話単位でまとめることができると考え，一つの話題に留まることが期待される会話に注目した話題検出を試みた．本手法では，準実時間での話題抽出と時間経過に伴う話題の推移を捉えることを企図して，tweet 集合を一定の時間間隔に区切って処理する．この処理単位を slot と呼び，1slot の時間幅を slot 幅と呼ぶ．Tweet の収集については，Streaming API を用いて新たな 1slot まで毎に元 tweet 集合を得，bot アカウントによる tweet を除外後に reply 先 tweet を lookup API で収集して（これを会話追跡と呼ぶ）当該 slot の処理 tweet 集合を得る．結論としては単 tweet モデルよりも，Reply 関係を用いたモデルはよりクラスタの精度が分かれる傾向にあった．本論文では藤田ら [7] の拡張した tweet 集合を本研究にて用いている．異なる点は，目的が違うという部分である．

3 提案手法

3.1 概　　要

本研究では，既存のマイクロブログにおける目的投稿の抽出手法と異なり，会話と呼ぶ一連のスレッドツリーをひとまとめとして抽出を行う．藤田らは先行研究 [6] において，tweet の in_reply_to プロパティを用いて reply 先 tweet を追跡することで，一連の reply によるスレッドツリーを取得する手法を開発しており，本研究ではこの手法を用いる．会話の取得は，Streaming API で取得した tweet 集合を元にして，逐次 lookup API を用いて reply 先 tweet を取得していくという手順である．

次に，ここで得られた各会話を 1 文書と見做して，目的の口コミを含む文書を判別していく．口コミの判別は，各文書を形態素解析した上で解析結果で出力される原形を Bag of words として取り扱い，この BoW に対して観光地名と日本語極性辞書の語が共起する文書を全て口コミと判定する手法である．

3.2 用語の定義

3.2.1 会　　話

先行研究 [6] にて reply 機能を用いたやりとりを会話と呼び，その関係を根つき有向木でモデル化している．本論文ではこの会話関係を用いた tweet から抽出手法を提案する．まず有向木でモデル化した会話の構造例を図 1 に示す．一つの tweet から連なる reply のやりとりは，一種の会話のように見ることができると考え，木全体を 1 会話とする．本手法ではこの 1 会話を 1 文書として扱い，会話 tweet と呼ぶ．

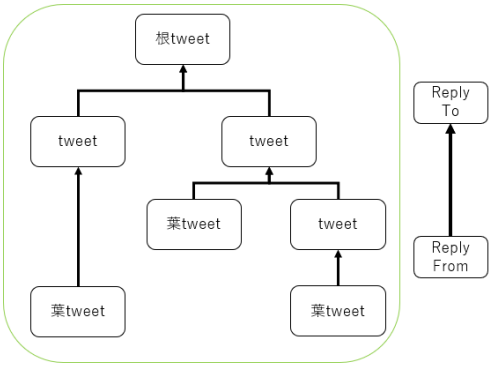

図 1: 会話の構造例

3.2.2 口　コ　ミ

本研究では，観光地毎に人々がどのような感想，評価を行っているのかを知ることが出来るような tweet を取得することを目的としている．特に，このように得られた tweet を整理して表示することで，観光地に関する人々の感想を容易に把握することが出来るようになるため，旅行計画等の立案に寄与できることを目指している．

そこで，観光地について言及しており，かつ評価や感想を述べているような tweet 文のことを口コミと呼ぶこととする．この口コミにどのような文が該当するかの例を図 2 に示す．

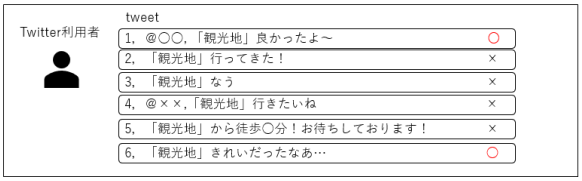

図 2: 口コミについて

ここでは，1 から 6 まで付番した tweet を枠囲いした各行に示しており，末尾には，本研究における口コミであるか否かを〇×で示している．具体的には，

（1） 「良かった」という評価を述べているため，口コミである．

（2） 単に事実のみを述べているため，口コミではない．

（3） 単に事実のみを述べているため，口コミではない．

（4） 願望を述べているのみであるため，口コミではない．

（5） 勧誘であり，感想を含まないため，口コミではない．

（6） 「きれい」という感想を含むため，口コミである．

のように判定される．

本論文では，感想や意見である多くの場合，肯定的または否定的な意味を示す語を含むと考えられるため，日本語評価極性辞書 [10] 収載の語を含む tweet を口コミと見做す判定手法を提案する．この辞書内の語の一部を図 3 に示す．

深謝	p	～する（行為）自分
深甚	p	～である・になる（評価・感情）主観
深憂	n	～する（行為）
申し分	e	～がある・高まる（存在・性質）
真	p	～である・になる（評価・感情）主観
真っ向勝負	e	～である・になる（状態）客観
真価	p	～である・になる（評価・感情）主観
真剣	p	～である・になる（評価・感情）主観
真剣勝負	e	～する（行為）
遅延	n	～する（出来事）
遅刻	n	～する（出来事）
撤回	e	～する（行為）
撤廃	e	～する（行為）
上達	p	～する（出来事）
上品	p	～である・になる（評価・感情）主観
上品さ	p	～がある・高まる（存在・性質）

図 3: 日本語評価極性辞書の一部

ここに見るように,「楽しい」という語は例えば，広告，勧誘等にも用いられるため，この判別手法では十分な分類が行えないことは予想される．しかし，その定量評価は不明であるため，本稿ではこの定義による判定を試み，結果を分析する．

3.3 tweet 集合の取得

会話 tweet を得る為に用いる tweet 集合の取得について説明する．

tweet 集合の取得には Twitter 社が提供する 2 種類の API を使用する．まず，Streaming API を利用して tweet 集合 T_s を取得する．tweet 集合 T_s は全 tweet の 1%が含まれている．次に lookup API に対するクエリとして，tweet 集合 T_s から取得した in_reply_to_status_id を指定する．得られた tweet から再帰的に Reply tweet を取得し，T_l を得る．

図 4 のように，Streaming API で得た tweet 集合の T_s と lookup API で取得した tweet 集合の T_l の 2 つの和となる tweet 集合を T_u で表わし，これをこの後の会話抽出処理における処理対象 tweet 集合とする．

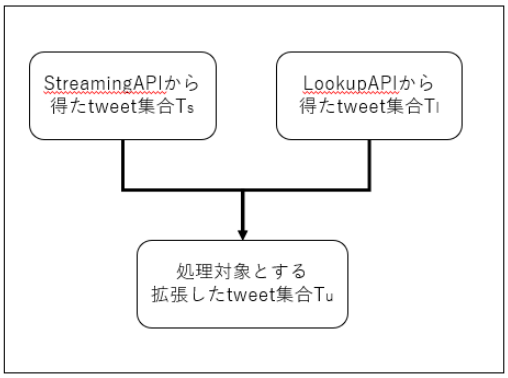

図 4: tweet 集合

3.4 会話 tweet にする処理

本手法では，日本語評価極性辞書の単語を用いて観光地に関する口コミを収集することを目的としているため，処理対象とする tweet は次の条件を満たすものとする．

- 日本語の tweet のみ
- リツイートや引用リツイートを除く tweet
- 顔文字で使われる補助記号や絵文字，URL を除去した tweet

これら 3 点を満たしている tweet から会話 tweet にしていく．会話 tweet にするには，3.3 節で集めた tweet 集合 T_u を取得した際に得られた属性の id と in_reply_to_status_id に着目して会話 tweet にしていく．in_reply_to_status_id に書かれている数値にはツイートされたときに振り分けられる id の数値が書かれているので，同じ数値がある tweet 同士を結び付けて 1 会話を 1 文書とする．

3.5 口コミの抽出

本手法では，1 会話 1 文書から観光地に関する口コミを収集することを目的として挙げている．図を用いて共起している会話 tweet の取得方法について説明する．

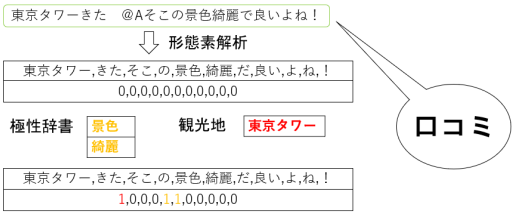

図 5: 口コミの判別手順

まず，3.4 節で 1 会話 1 文書 (図 5 の緑で囲っている部分) を形態素解析器である MeCab を用いて，形態素解析する．本手法では SNS で用いられる最近の用語にも追従するために，辞書には mecab-ipadic-NEologd を用いる．例えば「国立新美術館」という文字列があり，形態素解析を行うと「国立」「新」「美術館」の 3 つに分割されてしまい，一つの単語としてみなされず，そのまま判別を行うと誤判定されてしまう可能性があるので，軽減するために用いている．また本手法での口コミであるか否かの判定には，別に用意した辞書内の語が含まれているか

否かの情報を必要と宇する．このとき，投稿された tweet 内では一般に用言は活用形であるため，語の一致を行うために形態素解析器によって得られた語の原形情報を用いており，この目的でも形態素解析器を利用している．あらかじめ形態素に分割しておくことで，辞書に含まれる語の tweet 内の単語内文字列への部分一致が防げるほか，原形での一致判定で，活用を無視して語の存否を判定できるといった処理上の利点がある．

観光地名かつ極性辞書内の語が含まれているかを判定する．口コミの判別には，形態素解析によって得られた語の活用を無視して語の一致を見るため原形を取り扱う．形態素解析結果の原形を黒く囲っているように Bag of words として扱い，この BoW に対して観光地名かつ日本語評価極性辞書に載っている語 (図 3) が共起している文書を，すべて口コミ文書と判定する．

4 実　　験

本研究では，従来 1 tweet 単位での情報抽出を行っていたのに対して，抽出する情報単位を会話単位とすることによって，複数の tweet にまたがって抽出対象としての判定を行う要素が分散している場合にも，当該一連の tweet を抽出できるようになることを目指している．そこで，提案手法がこの目的を達成しているかどうかを検証するため，観光口コミ情報の抽出を，提案手法と比較手法とでそれぞれ行い，その結果を比較し分析する．比較手法には，従来の 1 tweet を 1 文書として扱う，1 tweet 1 文書モデルを充てる．なお，この文脈で，提案手法は 1 会話 1 文書モデルと呼ぶ．

4.1 対象 tweet

本実験で使用する tweet は Streaming API で 2019 年 4 月 27 日から 10 日分収集し，lookupAPI から収集した tweet を含めた tweet 集合 T_u を実験で用いた．この期間を選んだ理由は，本研究の目的が観光地に関する口コミ tweet の抽出であるため，旅行者が増えるであろう盛んなゴールデンウィーク期間がデータ収集に適していると考えたためである．対象とする観光地としては人気観光スポット TOP50 [11] の上位 5 位までに入っていった観光地を対象として実験を行った．下記に 5 つの観光地を示す．

- 京都府：清水寺
- 東京都：浅草寺
- 京都府：金閣寺
- 東京都：東京タワー
- 東京都：東京スカイツリー

本実験の処理対象とする tweet 集合は，提案手法による tweet 取得対象となった tweet 集合 T_u とし，これを比較手法，提案手法双方の共通の処理対象 tweet 集合とする．提案手法の対象となる 1 会話 1 文書モデルの総数は約 49 万件．比較手法では，tweet 集合 T_u の各 tweet それぞれを独立した文書として扱う．したがって，比較手法の対象となる 1 tweet 1 文書モデルの総数 は約 800 万件を用いる．

表 1: 用いる tweet 集合 T_u

1 会話 1 文書数	494,899 件
1 tweet 1 文書数	8,278,156 件

4.2 結　　果

3.3 で述べた T_u から提案手法で定義した会話 tweet から口コミであると判定された会話数，観光地名が含まれる会話数の結果を表 2 に，観光地名は含まれていないが極性辞書の語が含まれている会話数，どちらも含まれていない会話数を表 3 に示す．

表 2: 1 会話 1 文書から口コミを収集した結果

観光地	口コミ判定された会話数	観光地名入り会話数
清水寺	37 件	41 件
浅草寺	19 件	19 件
金閣寺	24 件	25 件
東京タワー	68 件	71 件
東京スカイツリー	8 件	10 件

提案手法である会話 tweet を見ると，根 tweet には，観光地名のみしか出現しなかった tweet が後々の Reply に評価や感想が述べられているのが出現しているの確認し，提案手法の有効性を確認することが出来た．観光地名を含むのに極性語を含まない会話 tweet を見ると,「楽しかった」や「きれい」という感想や評価が含まれているが口コミに判定されなかった会話 tweet があった．このことに関しては，極性辞書には「楽しさ」,「楽しみ」,「綺麗」という単語での登録はあるが,「楽しい」や「きれい」と言った単語の登録がなく，判定する事が出来なかった．

次に比較手法である 1 tweet1 文書から口コミであると判定された tweet 数，観光地名が含まれる tweet 数の結果を表 3 に示す．

表 3: 1tweet1 文書から口コミを収集した結果

観光地	口コミ判定された tweet 数	観光地名入り tweet 数
清水寺	90 件	126 件
浅草寺	59 件	81 件
金閣寺	64 件	84 件
東京タワー	282 件	429 件
東京スカイツリー	35 件	62 件

単 tweet でも，本手法を適用して口コミの抽出を行ってみたが，口コミだと判定された tweet の半分以上が評価や感想が述べておらず，たまたま極性辞書の語と観光地名が共起していた為に口コミであると判定された．

5 おわりに

StreamingAPI から得た tweet に加え，リプライ関係を含めた tweet 集合から会話 tweet にし，観光地名かつ日本語極性辞書が共起している tweet を収集した．1tweet1 文書からでは観光地名と評価や感想が述べられている tweet が少なく，事実しか述べていない tweet が多かったが，会話 tweet に拡張したこ

とで，事実のみの tweet から，感想や評価が述べられている Reply tweet も加わったので，本手法の会話 tweet は有効であることを確認出来た．しかし，問題もあり，口コミである tweet の抽出の際に使用していた日本語評価極性辞書内の語に「綺麗」という語があるが，tweet する際にキレイやきれいと平仮名や片仮名で表記されている可能性もあり，考慮していなかった点が挙げられる．これらは日本語評価極性辞書を基に自らで評価や感想述べている際によく出現する語を追加することで，幅広く対応できるのではないかと考え，明らかにしていきたい．また，マイナーな観光地を対象とした時にどのような tweet があり，どういう特性があるのか見ることが出来れば面白いと思うので，こちらについても明らかに出来ればと思う．

文　　献

[1] 総務省｜令和 2 年版 情報通信白書｜インターネットの利用状況 https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r02/html/nd252120.html.

[2] 【データ】デジタル化による旅行スタイルの変化　ＪＴＢ総合研究所調べ https://www.kankokeizai.com/%E3%80%90%E3%83%87%E3%83%BC%E3%82%BF%E3%80%91%E3%83%87%E3%82%B8%E3%82%BF%E3%83%AB%E5%8C%96%E3%81%AB%E3%82%88%E3%82%8B%E6%97%85%E8%A1%8C%E3%82%B9%E3%82%BF%E3%82%A4%E3%83%AB%E3%81%AE%E5%A4%89%E5%8C%96/.

[3] Q1 2019 Earnings Report https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Slide-Presentation.pdf.

[4] 渡邉 百合子，吉野 考, “観光地名なしツイートからの観光地に関する感想の抽出手法”, 情報処理学会論文誌 59(1), pp. 43-51, 2018.

[5] 小原 基季， 森田 和宏， 泓田 正雄， 青江 順一，“Twitter 本文を用いた観光情報抽出及び分析システムの構築”， 人工知能学会全国大会論文集 29, pp. 1-3, 2015.

[6] 免田 哲矢， VictorKryssanov， 林 勇吾， 小川 均，“Twitter を用いたリアルタイム情報収集による観光地情報推薦システム”, 第 73 回全国大会講演論文集 2011(1), pp. 647-648, 2011.

[7] 藤田 俊之， 小林 亜樹，“災害情報抽出のための Reply 関係を用いた話題抽出”， DEIM Forum, 2020.

[8] Twitter4j，http://twitter4j.org/ja/index.html.

[9] mecab-ipadic-NEologd : Neologism dictionary for MeCab, https://github.com/neologd/mecab-ipadic-neologd/blob\\/master/README.ja.md.

[10] FrontPage / Open Resources / Japanese Sentiment Polarity Dictionary，http://www.cl.ecei.tohoku.ac.jp/index.php?Open\\\%20Resources\%2FJapanese\%20Sentiment\%20Polarity\%20Dictionary.

[11] 【2019 年版】みんなが行った国内の人気観光スポットランキング TOP50！，https://tripnote.jp/japan/popular-spot-ranking2019,