DBSJ Japanese Journal
Vol. 16-J, Article No. 22, March 2018

マイクロブログからの関連実世界観測情報の抽出

Extracting Related Real-World Observations from Microblog

新田 直子♡ 吉武 真人♢
中村 和晃♠ 馬場口 登♣

Naoko NITTA Masato YOSHITAKE
Kazuaki NAKAMURA Noboru BABAGUCHI

近年，**Twitter** を代表とするマイクロブログに，多くのユーザから実世界における観測情報がリアルタイムに投稿される．これらの実世界観測情報は，地名など各地の観測対象を表すローカル語を含む場合が多く，ローカル語を用いた抽出方法が提案されている．しかし実際には，これらのローカル語を含まない観測情報も存在する．そこで本研究では，ローカル語を含む観測情報に，各観測対象の特徴を表す単語が含まれるという仮定に基づき，ローカル語を含む観測情報集合から学習した単語の意味表現を用いて，共通したローカル語を含まないが，関連する観測情報の抽出を目指す．

A large number of users post what they observe around themselves to microblogs. Since the observations at specific locations are often described with words representing the observed locations or events, these observations can be discovered by finding the posts with local words which are uniquely used at specific locations. Based on the assumptions that the discovered observations also contain words defining the semantics of the observed locations or events, this paper proposes a method for extracting the observations without the local words based on their semantic relevancy to the local words.

1. はじめに

全世界に 3 億人以上のアクティブユーザを抱える Twitter [1] には，多様な情報が世界各地から投稿される．Twitter の主な投稿形式は，ツイートと呼ばれる 140 文字以下の短文である．スマートフォンなどの普及に伴い，時間や場所を問わず手軽に投稿することができるため，多くのユーザが実世界において観測した情報をその場で投稿しており，新聞やテレビなどのメディアに比べリアルタイム性の高い情報が多い．このような特性から，Twitter のユーザ一人一人を，地震を観測した時に震度情報を出力する震度計などの物理センサに対し，ソーシャルセンサと捉え，Twitter からセンサが観測した実世界情報の獲得を試みる研究が進められている [2], [3].

特定の対象に関する観測情報を獲得したい場合，観測対象の関連語を用いる手法が提案されている．例えば Sakaki ら [4] は，地震の震源地を推定するため，予め地震に関する単語を関連語として人手で設定することにより，投稿位置の緯度・経度情報であるジオタグが付与されたツイートから地震の観測情報を抽出した．また土屋ら [5] は，予め準備した鉄道の運行トラブルに関するツイート集合から関連語を学習し，さらなる鉄道運行トラブルの観測情報の抽出に利用した．これらの研究では，各観測対象に対して，関連語や過去の観測情報の事例を人手で与えなければならない．

そこで，ユーザからクエリにより指定される対象に対し，自動的な関連語の決定により観測情報を抽出する手法も提案されている．Massoudi ら [6] や藤木ら [7] は，観測対象に特徴的な事象が発生した際，その事象を表す単語と観測対象を表すクエリの同一ツイート内での共起頻度が一時的に高くなると考え，クエリと短期的に共起する単語を関連語とした．この手法では，多様な観測対象の関連語を現在までのツイートから自動的に決定できる．しかし，ユーザからクエリが与えられてから関連語を決定し観測情報を抽出するため，ユーザがクエリを与えてから観測情報を取得するまでにタイムラグが生じる．

一方，観測対象を限定せず，様々な実世界観測情報を抽出する手法も提案されている．これらの手法では，実世界観測情報が地名など各地の観測対象を表すローカル語を含む場合が多いと考え，ローカル語を用いて実世界観測情報を抽出する．例えば Watanabe ら [8] は，位置情報サービスである Foursquare の投稿から地名をローカル語として収集し，収集したローカル語を含むツイートを各地の観測情報とした．また，地名だけでなく，地名を表す略語や特産品など，様々な地域に特有な語をローカル語として Twitter から収集する研究も行われており，これらの研究を組み合わせれば，より多様な観測情報を獲得できると考えられる．ローカル語抽出の研究では，ジオタグ付きツイートから地理空間的に局所性の高い単語を抽出するアプローチが中心となっている [9], [10], [11]. 各単語に対して適切な時区間において空間的局所性を調べることにより，地名のような常に同じ場所を示す単語だけではなく，各地のイベントなどを表す一時的なローカル語も合わせて，リアルタイムに抽出できる．よって，抽出された各地の観測対象を表すローカル語を含むツイートを，多様な対象の観測情報として抽出できるが，実際には，ユーザは観測情報を投稿する際，地名や他人と同じ表現を用いるとは限らないため，既存手法では抽出できない実世界観測情報が存在する．

そこで本研究では，Twitter に投稿されるツイートから，ローカル語で表される多様な対象に対し，ローカル語を用いず投稿された観測情報を抽出することを目的とする．提案手法ではまず，ジオタグ付きツイートに含まれる各単語の空間的局所性に基づき，各地の観測対象を表すローカル語を抽出する [11]．ここで，各ローカル語により表される対象の観測情報は，ローカル語の有無に依らず，その対象を特徴付ける単語を多く含むと考えられる．そこで次に，意味的に近い単語ほど類似度が高くなるような単語のベクトル表現 [12], [13] を，ローカル語を含む実世界観測情報集合から学習する．最後に，学習された単語のベクトル表現に基づき，各ツイートとローカル語の関連度を算出し，各ローカル語に対し，ローカル語を含まないが関連度の高い単語を多く含むツイートを，関連する実世界観測情報として抽出する．

2. 提案手法

Twitter には，多くのユーザが様々な場所から多様な実世界観測情報を投稿するが，その投稿の多くには，各地の観測対象を表すローカル語が含まれると考えられる．ローカル語は，自らが表す観測対象が存在する位置でのみ使用されることが多いと推定されるため，まずジオタグ付きツイートから，各単語の空間的局所

♡ 正会員 大阪大学大学院工学研究科
naoko@comm.eng.osaka-u.ac.jp
♢ 非会員 大阪大学大学院工学研究科
yoshitake@nanase.comm.eng.osaka-u.ac.jp
♠ 非会員 大阪大学大学院工学研究科
k-nakamura@comm.eng.osaka-u.ac.jp
♣ 非会員 大阪大学大学院工学研究科
babaguchi@comm.eng.osaka-u.ac.jp

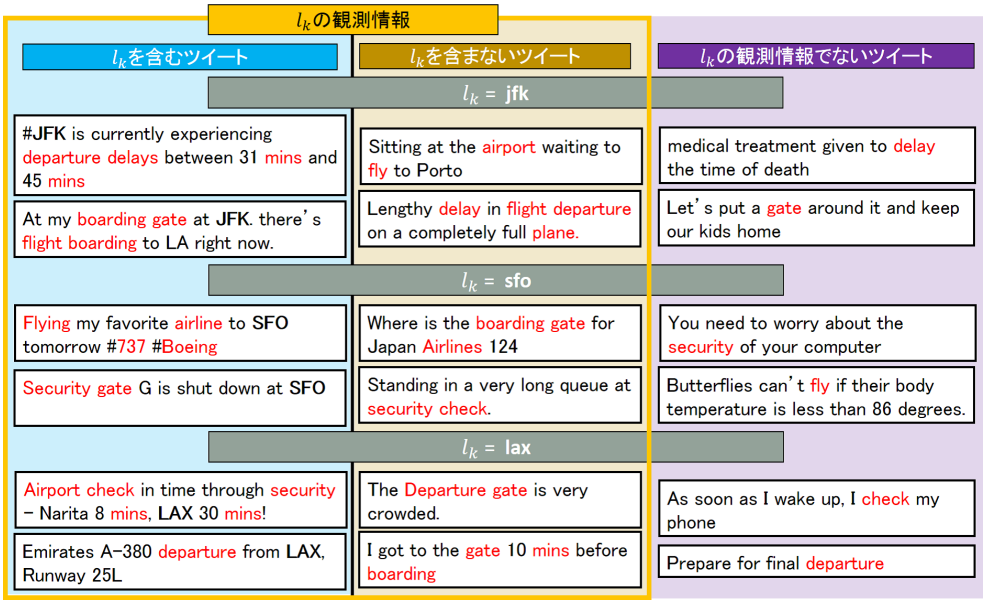

図 1: ローカル語を各地の空港とした時の観測情報の例

性に基づき，ローカル語集合 $L = \{l_k | k \in \mathbb{N}\}$ を抽出し，各ローカル語 l_k を含むツイートを l_k に関する実世界観測情報として抽出する．しかし，実世界観測情報が必ずしもローカル語を含むとは限らない．例えば，l_k がサンフランシスコ国際空港を表す略語である “sfo” である場合，l_k に関連する観測情報には，“Standing in a very long queue at security check.” のような，l_k を含んでいないツイートも存在する．よって，全てのツイートを，l_k を含むツイート，l_k を含まないが l_k の観測情報であるツイート，l_k の観測情報でないツイートに分類できると考えられる．

ローカル語を各地の空港とした時の実世界観測情報の例を図 1 に示す．各ローカル語 l_k に関するツイートは，ローカル語の有無によらず “departure” や “boarding”，“gate” などの空港から連想される単語を多く含む．一方，l_k に関連しないツイートは，ローカル語から連想される単語を含む場合もあるが，ローカル語からは連想されない単語も多く含むと考えられる．このように，ローカル語から連想される単語同士は，関連した対象に関する観測情報において共に使われることが多い．そこで，各単語が用いられるコンテキストの類似性から単語間の意味的な関係性を学習する word2vec により得られた単語ベクトルに基づき，ローカル語に関する実世界観測情報を抽出する．

以上より，提案手法は図 2 に示すように以下のステップにより構成される．

Step1) ローカル語の抽出
ジオタグ付きツイート中に用いられる各単語の空間的局所性に基づき，各地の観測対象を表すローカル語をその位置と共に抽出する [11]．さらに，抽出された位置において投稿されたローカル語を含むジオタグ付きツイートを実世界観測情報の一部として抽出する．

Step2) 単語の意味表現の学習
word2vec [12], [13] により学習される単語のベクトル表現は，各単語を意味的に表現できることが知られているため，以降では単語の意味表現と呼ぶ．Step1) において収集したローカル語を含むツイート集合を用いて，word2vec により，単語の意味表現を学習する．

Step3) 実世界観測情報の抽出
各ローカル語 l_k が抽出された位置において投稿された全てのツイートに対し，Step2) で学習した単語の意味表現に基づき l_k との関連度を算出し，関連度の高いものを l_k が表す対象に関連する実世界観測情報として抽出する．

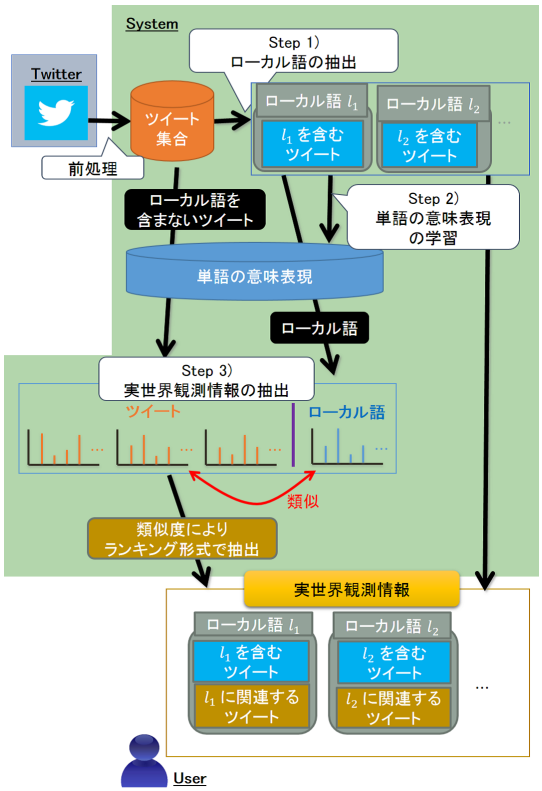

図 2: 提案手法の概要

次節以降で，各ステップの詳細について述べる．

2.1 ローカル語の抽出

実世界観測情報は，観測対象を表すローカル語を含む場合が多いため，まず，ローカル語を抽出することにより，実世界観測情報の一部を獲得する．ローカル語はその地域に固有のものを表してい

ると考えられるため，空間的に狭い範囲で繰り返し用いられる単語をローカル語として抽出する．この時，Twitter に投稿される各単語に対して適切な時区間を設定した上で空間的局所性を調べることにより，各地の場所や特産品を表す語などの恒常的なローカル語と共に，イベントを表す語などの一時的なローカル語を抽出できる [11].

まず，ツイートを収集している地理空間全体を，各エリアの投稿数がほぼ均等となるよう J 個のエリア $A = \{a_j | j = 1, \cdots, J\}$ に分割する．ローカル語は地名や特産品，イベントなど各地で観測される対象を表す名称が多く，主に名詞から構成されると考えられるため，ジオタグ付きツイートが投稿される度，形態素解析により品詞のタグ付けを行い，名詞のみを抽出する．ここで，例えば “Michigan” という単語が州の名であるのに対し，“Michigan Stadium” は施設名であるように，ローカル語を複合名詞として抽出することにより示す場所や意味がより限定されると考えられるため，複合名詞を抽出する．

ジオタグ付きツイートが投稿されるごとに，ツイートに含まれる Z 個の名詞 $u_z (z = 1, \cdots, Z)$ の出現履歴を更新する．u_z の局所性は，TFIDF 法を応用し，以下の式により算出される．

$$tfidf^{max}_{u_z} = f^{max}_{u_z} \cdot idf_{u_z} \quad (1)$$

$$idf_{u_z} = \log \frac{J}{|A_{u_z}|}, \ where \ A_{u_z} = \{a_j | f_{u_z,j} \neq 0\} \quad (2)$$

ただし，$f_{u_z,j} (j = 1, \cdots, J)$ はエリア j における単語 u_z の出現頻度，$f^{max}_{u_z} = \max_j f_{u_z,j}$ は全エリアの中の u_z の最大出現頻度，$|A_{u_z}|$ は u_z が出現したエリア数，J は全エリア数を表す．u_z が特定のエリアで頻繁に出現したときに $tfidf^{max}_{u_z}$ は高くなるため，$tfidf^{max}_{u_z} \geq R$ を満たしたとき u_z をローカル語 l_k として抽出する．さらに，イベントを表す語のような一時的に空間的局所性をもつローカル語は，それ以外の時区間ではどのエリアでも使用され得る．よって，長期間の出現履歴を保持していた場合，一時的な空間的局所性を検知することが困難となる．そこで $tfidf^{max}_{u_z} < r$ を満たしたとき u_z の出現履歴を一度削除し，新たに収集し始めることにより，一時的な空間的局所性を正しく検知できるような時区間での出現履歴を保持するよう対処する．また，$tfidf^{max}_{u_z} \geq R$ を満たして u_z をローカル語 l_k として抽出するときも，u_z のこれまでの出現履歴を削除し，新たに収集を開始する．これにより，一度ローカル語として抽出された u_z が新たな出現履歴において $tfidf^{max}_{u_z} < r$ を満たしたとき，過去の一時的なローカル語であると判断し，ローカル語集合から削除する．

抽出された各ローカル語 l_k に対し，対応するエリア集合 A_{l_k} から投稿された l_k を含むツイートを，l_k が表す対象の実世界観測情報の一部として抽出する．

2.2 単語の意味表現の学習

実世界観測情報には，ローカル語の有無に依らず，観測対象ごとに用いられる単語に特徴があると考えられる．そこで，ローカル語を含まないが，ローカル語が表す観測対象に関するツイートを抽出するため，前節で抽出したローカル語を含むツイート集合から単語間の意味的な関係性を学習する．このとき，前処理として，URL である “http(s)://...” やユーザ名を表す “@” で始まる単語は除去する．さらに，ユーザ名を表すものとは別に “@ yankee stadium” のように，“@” の後ろにスペースを入れた上で地名などが記述される場合も多く見られる．この場合における “@” の出現位置は，ツイートの後方であることが多い．ここでは一般的な語の関係性を学習するため，単語 “@” で始まる単語列に加え，ローカル語も除去した上で，word2vec [12], [13] を適用する．

word2vec は，学習用コーパスとして大量の文書集合が与えられたとき，各単語の周辺単語を推定するような 2 層からなるニューラルネットワークを用いて，同じコンテキストで用いられる単語対ほど近くなるような各単語のベクトルを学習する．ここでは，各ローカル語と共に用いられる単語対が近くなるよう，前節で抽出した各ローカル語に対する実世界観測情報集合を一つの文書とみなし，全てのローカル語に対する文書集合を学習用コーパスとして word2vec に与える．

表 1: ローカル語の例

ローカル語	ツイート数	出現日数
ohio	1,688	30
texas	1,540	30
arizona	636	30
santa barbara	505	30
north dakota	116	28
central park	1722	30
yankee stadium	976	30
sfo	193	23
desert trip	280	5
boston red sox vs new york yankees	11	2
akron marathon	10	1
heavy rain	10	1

2.2 実世界観測情報の抽出

l_k が表す観測対象に関する実世界観測情報は，エリア集合 A_{l_k} から投稿されると考えられるため，まず，各ローカル語 l_k に対し，A_{l_k} から投稿された，l_k を含まないジオタグ付きツイートを l_k が表す対象に関連する観測情報の候補として抽出する．

l_k を含むツイートに含まれる単語集合は，l_k を特徴付けると考えられるため，これらの単語のベクトルの平均を l_k の意味表現として算出する．また，l_k を特徴付ける単語が経時的に変わる場合もあると考えられるため，l_k の意味表現は 1 日毎に，その日の l_k を含むツイートから，新たに算出する．l_k に関連する実世界観測情報を含むツイートも，l_k を特徴付ける単語が多く含まれていると考えられるため，各ツイートに含まれる単語のベクトルの平均を，各ツイートの意味表現として算出し，ローカル語とツイートのベクトル間のコサイン類似度をローカル語とツイートの関連度として算出する．l_k との関連度の高いツイートを抽出することにより，ローカル語に関連する実世界観測情報の抽出を実現する．

3. 実験

Twitter の Streaming API を用いて，アメリカ本土を緯度が 24 度から 49 度，経度が-125 度から-66 度の範囲と設定し，2016 年 9 月 8 日から 2016 年 10 月 9 日に投稿された 6,655,763 件のジオタグ付きツイートを収集した．まず，収集したツイートから，多様な観測対象を表す語としてローカル語を抽出し，初めの 25 日間における，全てのローカル語に対して抽出された，ローカル語を含む実世界観測情報を学習用コーパスとして word2vec により単語の意味表現を学習した．学習時に適切なパラメータを決定するため，評価用ツイート集合における実世界観測情報の抽出実験を行った．最後に，決定したパラメータを用いて学習した単語の意味表現に基づき，実際にローカル語に対応するエリア集合から投稿されたローカル語を含まないツイートから，関連する実世界観測情報を抽出し，その結果について考察した．

3.1 ローカル語の抽出

アメリカ本土を各エリアの投稿数が均等となるよう $J = 256$ エリアに分割し，各ツイートに対して，Brill’s Tagger [14] により品詞のタグ付け，TermExtract [15] により複合語を抽出した．得られた全ての名詞に対して，$R = 16.63$，$r = 6.97$ [11] としてローカル語を抽出した結果，期間中に削除されたローカル語も含め，51,166 語のローカル語が抽出された．得られたローカル語の例と，30 日間における各ローカル語を含むツイート数及び出現日数を表 1 に示す．“texas” や “arizona” のような空間的に広い範囲に存在する州や都市を表す地名，“yankee stadium” や “sfo” のような特

表 2: “texas” を含む実世界観測情報例

日付	ツイート本文
10/7	Want to work in #Coppell. Texas? View our latest opening: (URL) #Job #NowHiring #GetHired #IT #Jobs #Hiring
10/7	Interested in a #job in #Irving. Texas? This could be a great fit: (URL) #CustServ #CustomerCare #CustomerService
10/7	Want to work at Kindred At Home? We’re #hiring in #Grapevine. Texas! Click for details: (URL) #Job #Sales #Jobs
10/9	Interested in a #job in #Dallas. Texas? This could be a great fit: (URL) #driver #cdl (URL)
10/9	Oklahoma may have won the red river shootout. but Texas won when (URL)
10/9	Interested in a #job in #DALLAS. Texas? This could be a great fit: (URL) #Retail #Hiring #CareerArc

表 3: “sfo” を含む実世界観測情報例

日付	ツイート本文
10/7	Sunrise at SFO as I wait for a standby. Will I get lucky today? @flysfo @united #travel #sunrise (URL)
10/7	Our flight attendant Karen on flight 5456 from Redmond to SFO was first class all the way! Thanks @united for hiring great crew.
10/7	Self service buffet at the airport lounge.... jaaaaa @ American Express Centurion Lounge At SFO (URL)
10/9	Excited for @United family day at SFO (@ United Technical Operations in San Francisco. CA) (URL)

表 4: “desert trip” を含む実世界観測情報例

日付	ツイート本文
10/7	Desert Trip! Tonite Bob Dylan and the Stones. Tomorrow Neil Young (URL)
10/7	The Desert Trip begins. @ Desert Trip 2016 (URL)
10/7	Concert in the desert! Bob Dylan and The Rolling Stones tonight! @ Desert Trip 2016 Platinum (URL)
10/9	Neil Young deserttripindio amazing!! @ Desert Trip 2016 (URL)
10/9	Neil Young’s set was absolutely mind blowing. #musicfamily #deserttrip @ Desert Trip 2016 (URL)
10/9	Perfect way to escape reality. Music and fireworks. Desert Trip: Day 2. Thank you @neilyoung (URL)

定の位置に存在する施設を表す地名だけでなく，“desert trip” や “boston red sox vs new york yankees” のような一時的なイベントを表す表現がローカル語として抽出された．ローカル語を含むツイート数は様々であるが，州名や都市名，施設名を含むツイートはほぼ毎日のように出現している一方，イベントを表すローカル語を含むツイートは短期的にしか出現しないことが分かる．

各ローカル語が表す対象の観測情報の一部として，各ローカル語に対応するエリア集合から投稿されたローカル語を含むツイートを抽出したところ，30 日間で 836,574 件のツイートが得られた．例として，空間的に広い範囲を表すと考えられるローカル語 “texas”，空間的に狭い範囲を表すと考えられるローカル語 “sfo”，一時的なイベントを表すと考えられるローカル語 “desert trip” に対して得られたツイートの一部を，それぞれ表 2，3，4 に示す．“texas” の観測情報には，フットボールの試合に関する情報もあるが，多くの求人情報が抽出されている．“sfo”，“desert trip” のような，より場所が限定されるローカル語に対しては，得られた観測情報もより詳細なものとなっている．

求人情報のような同一の形式で大量に投稿されるツイートが word2vec の学習用コーパスに含まれる場合，これらのツイートに含まれる単語同士の関連性が非常に高くなり，意味表現に基づきローカル語を含まない観測情報を抽出する際，同じ形式のツイートが大量に抽出されると考えられる．よってこのようなツイートは，実世界観測情報としても，意味表現の学習時においてもノイズとなり得る．このような同一の形式で大量に投稿されているツイートを，以降ではノイズツイートと呼ぶ．

3.2 評価用ツイート集合を用いた提案手法の評価

実世界観測情報の抽出精度を定量的に評価するため，まず，評価用ツイート集合を抽出する．本研究で抽出対象とする，ローカル語を含まないが，ローカル語で表される観測対象に関するツイートの正解データを人手で抽出することは非常に困難であるため，これを疑似的に作成するため，主にツイートの検索に利用されるハッシュタグに注目した．ハッシュタグとは，ツイート中の単語の前に語 “#” をつけることにより，その単語を，ツイートに付与されたタグとみなす機能である．観測対象を表す単語がハッシュタグとして用いられる場合もあり，ツイートの本文中の単語をハッシュタグにするユーザも存在するが，多くの場合，本文とは独立して記述される．そこで，ローカル語がハッシュタグで記述されているツイートから，ローカル語のハッシュタグを除去し，ローカル語を含まない観測情報の正解ツイートとする．ただし，ハッシュタグではスペースを使うことができないため，多くのユーザが，複合語はスペースを除去して記述する．よって複合語であるローカル語の正解ツイートは，これを考慮して抽出した．

次に，ローカル語に対応するエリア集合とは異なるエリアから投稿されたツイートは，ローカル語が表す対象の観測情報ではないと考えられるため，これらを不正解ツイートとしてランダムに抽出した．このとき，実環境での動作を想定し，正解ツイートと不正解ツイートの合計が，ローカル語に対応するエリア集合から投稿されたツイート数と同数になるように，不正解ツイートの数を決定した．

以上のようにして，収集したツイートの最後の 5 日間における各ローカル語に対し，評価用ツイート集合を抽出した．ただし，各ローカル語に対し，ローカル語を含むツイートが少ない場合，適切な意味表現を算出できない可能性がある．そこでローカル語を含むツイートが 1 日で 10 件以上存在するローカル語を対象とした．これにより，のべ 817 語のローカル語に対し，35,923 件の正解ツイート，999,439 件の不正解ツイートを抽出した．

これらの評価用ツイート集合を用い，各ローカル語に対して提案手法により正解ツイートが正しく抽出できるか評価する．ローカル語とツイートの意味表現のコサイン類似度に閾値を設定する

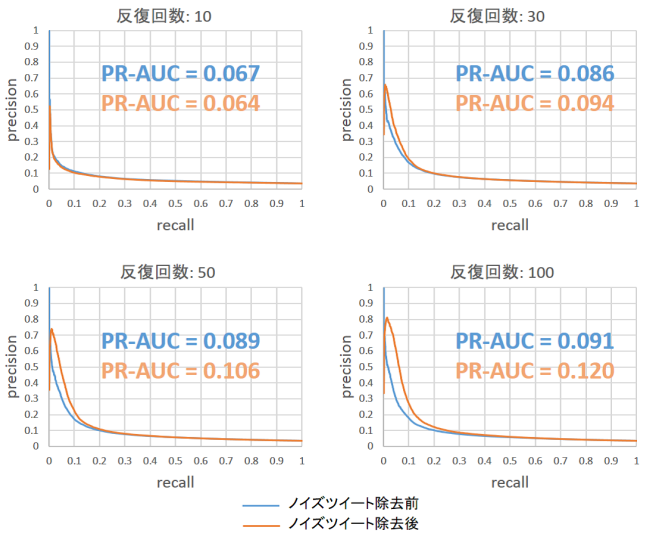

図 3: ノイズツイート除去前／後のコーパスによる学習における反復回数 *iter* による抽出精度の変化

ことにより，抽出された観測情報における誤検出の割合を示す適合率（precision）と，正解ツイートにおける未検出の割合を示す再現率（recall）が算出できる．以降では，コサイン類似度に対する閾値を変化させたときの適合率と再現率の関係を示す PR 曲線の下面積（PR-AUC）を評価指標とする．

3.2.1 ノイズツイートの影響の評価

まず，word2vec の学習用コーパスに含まれるノイズツイートの影響を検証するため，初めの 25 日間のローカル語を含む全てのツイートをコーパスとして学習したモデルと，ノイズツイートを除いた全てのツイートをコーパスとして学習したモデルを用意し，評価用ツイート集合を用いて比較実験を行った．今回は頻出するノイズツイートに用いられる形式を人手で 13 パターン設定し，これらのパターンに合致するツイートを除去した．コーパス内のツイート数，単語数，語彙数は，ノイズツイート除去前のコーパスで，659,615 件，4,552,955 語，146,691 語，ノイズツイート除去後のコーパスで，392,142 件，2,488,157 語，136,324 語となった．

単語の意味表現の学習には，gensim[16] の word2vec ライブラリを利用した．学習に関するパラメータは，Mikolov ら [12], [13] の実験を参考に，単語ベクトルの次元数 *dim* を 300，学習モデルを skip-gram，ウインドウサイズを 10，学習する単語の最低出現回数を 5，ネガティブサンプリングに用いる単語数を 15 とした．意味表現を学習する単語をコーパス内の出現回数で限定した結果，ノイズツイート除去前のコーパス，除去後のコーパスの語彙数は，それぞれ 25,829 語，21,516 語となった．Mikolov らは，2400 万語から 60 億語のコーパスを用いて学習の反復回数 *iter* を 1 や 3 に設定していたが，本研究で扱うコーパスはこれに比べて非常に小さいため，*iter* を 10 から 100 の間で変化させ，結果を比較した．

各反復回数における PR 曲線を図 3 に示す． *iter* = 10 の時には，結果にほぼ差は現れなかったが，反復回数を多くするにつれ，ノイズツイートを除去したコーパスで学習したモデルの方がより多く正解ツイートのみを上位に抽出できた．反復的に学習する目的は，コーパスのサイズが十分でないときに学習の不足を補うことであるが，ノイズツイートを含むコーパスでは，少ない反復回数でも，同じ形式で大量に投稿されたノイズツイートによって反復的に学習されており，反復回数によらず，ノイズツイートと類似したツイートが多く上位に抽出された．

iter = 100 とし，ノイズツイート除去前のコーパス，除去後のコーパスでそれぞれ学習した単語の意味表現を用いて，上位に抽出

表 8: パラメータによる PR-AUC の変化

*dim**iter*	10	30	50	100	200	300
100	0.051	0.060	0.066	0.074	0.083	0.085
300	0.064	0.094	0.106	0.120	0.129	0.133
600	0.066	0.115	0.122	0.131	**0.141**	**0.143**
1000	0.065	0.119	0.126	0.136	**0.142**	**0.144**

されたツイート例をそれぞれ表 5，6 に示す．ただし，表中の *cos* は，コサイン類似度を表す．まず，10 月 9 日の “santa barbara” に対する抽出結果に注目すると，ノイズツイート除去後のコーパスで学習した場合には，santa barbara のレストランなどの情報が抽出されているのに対し，ノイズツイート除去前のコーパスで学習した場合は，天気に関する情報を上位に誤抽出している．この結果より，“santa barbara” を含む天気に関するノイズツイートが多く存在し，ノイズツイートを除去しない場合，ノイズツイートの影響により，天気に関するツイートが上位に多数抽出されたと言える．10 月 4 日の “texas” に対する抽出結果についても，コーパスの違いにより異なる結果が得られたが，どちらの場合も不正解ツイートを上位に誤抽出したことがわかる．ノイズツイート除去後のコーパスにより学習した場合においても非常に形式が類似したノイズツイートが上位 2 件に抽出されており，これは学習に用いたコーパスにおいてノイズツイートが除去しきれなかったためと考えられる．このようなノイズツイートの形式を人手で完全に設定することは困難であるため，ローカル語を含む観測情報を抽出した後，ノイズツイートを自動的に判定する方法の考案が必要である．一方で，10 月 9 日の “central park” に対する抽出結果のように，いずれのコーパスで学習しても，結果に違いが生じないローカル語も存在した．

学習を繰り返しても再現率は低かったが，下位に含まれた正解ツイートの多くは，表 7 に示すような除去されなかったノイズツイートや，ハッシュタグは含まれるが，観測対象の特徴を表すような単語を含まず，観測情報として重要性の低いツイートであった．これらの結果より，ローカル語をハッシュタグとして含むツイートが必ずしも正解ツイートとして適切とは言えないため，より適切に抽出結果を評価する方策が必要と考えられる．しかしノイズツイート除去後のコーパスにより学習した場合の方が，適切な観測情報を抽出できるローカル語が多いと言えるため，以降の実験では，ノイズツイート除去後のコーパスを word2vec の学習用コーパスとする．

3.2.2 パラメータによる影響の評価

次に，word2vec の学習におけるパラメータの影響を検証する．本研究と Mikolov らの実験における大きな違いは，コーパスのサイズである．そこで，コーパスのサイズにより最適値が大きく変化すると考えられる，単語ベクトルの次元数と学習の反復回数の影響を評価する．まず単語ベクトルの次元数 *dim* に対し，Mikolov らは $dim = 50, 100, 300, 600$ の 4 通りで精度の変化を検証しており，次元数が大きいほど精度が向上する結果を示した．ここではこの 4 通りから，*dim* = 50 を除外し，*dim* = 1000 を加えた 4 通りについて検証する．また学習の反復回数に対しては，7 億 8300 万語のコーパスを用いて 3 回の反復による精度向上が示されている．本研究ではこれと比較し非常に小さなコーパスで学習するため，反復回数は重要なパラメータと考えられる．よって，反復回数 *iter* を前節で設定した $10, 30, 50, 100$ に $200, 300$ を加えた 6 通りについて検証する．

単語ベクトルの次元数と学習の反復回数を変更し word2vec を学習した際の PR-AUC を表 8 に示す．単語ベクトルの次元数，学習の反復回数のいずれも大きい方が PR-AUC は向上したが，*dim* = 600，*iter* = 200 程度でほぼ収束した．

3.3 実世界観測情報の抽出結果の考察

最後にノイズツイート除去後のコーパスを用いて *dim* = 600，*iter* = 200 として word2vec を学習した後，提案手法により，実際に各

表 5: ノイズツイート除去後のコーパスにより学習した場合の上位抽出例

ローカル語	日付	*cos*	ツイート本文	正解ラベル
santa barbara	10/9	0.748	Til next time CA. Toronto is beautiful. Your people are the nicest. Can't wait to come back!··· (URL)	negative
santa barbara	10/9	0.741	Little, tasty Devils. #finchandfork #santabarbara #california #devileddeggs #eggs @ Finch and··· (URL)	positive
santa barbara	10/9	0.738	Hot sunny #santabarbara day requirements: ice cold white #wine & comfy A51 #hat @ Area 5.1 Winery (URL)	positive
texas	10/4	0.935	@ebbtideapp Tide in Queensboro Bridge, New York 10/04/2016 High 12:27pm 4.8 Low 6:25pm 0.6 High 12:47am 4.2 Low 6:25am 0.8	negative
texas	10/4	0.919	@ebbtideapp Tide in Quonset Point, Rhode Island 10/04/2016 Low 4:10pm 0.2 High 10:58pm 3.4 Low 4:12am 0.3 High 11:18am 3.7	negative
texas	10/4	0.669	Back feeble at the seaport spot in lower east side! : grim_stagram. #skateboarding #skateboard··· (URL)	negative
central park	10/9	0.844	Great view coming home tonight #nyc #concretejungle #newyork #skyline #centralpark #manhattan··· (URL)	positive
central park	10/9	0.825	Birthday celebration of a legend #johnlennon #centralpark #groupies #imagine #nyc #birthday··· (URL)	positive
central park	10/9	0.806	#newyorkcity #centralpark @ Top Of The Rock NYC (URL)	positive

表 6: ノイズツイート除去前のコーパスにより学習した場合の上位抽出例

ローカル語	日付	*cos*	ツイート本文	正解ラベル
santa barbara	10/9	0.759	Weather now: clear sky, 71 ° F, 10 mph north wind. (URL)	negative
santa barbara	10/9	0.748	WEATHER: Environment Canada calling for mainly sunny skies today and a high of 13 C in Toronto. @CP24 (URL)	negative
santa barbara	10/9	0.744	It might be too late to evacuate but all that good sunshine was calling my name. Clear skies and··· (URL)	negative
texas	10/4	0.626	Work (@ Texas Workforce Commission - Main Building in Austin, TX) (URL)	negative
texas	10/4	0.593	See a virtual tour of my listing on 13 Harbourside LANE 7138 #HiltonHeadIsland #SC #rea... (URL) (URL)	negative
texas	10/4	0.590	The Cage ops director shuffling the cards at employee open house at Rock and Brews··· (URL)	negative
central park	10/9	0.843	Great view coming home tonight #nyc #concretejungle #newyork #skyline #centralpark #manhattan··· (URL)	positive
central park	10/9	0.815	Birthday celebration of a legend #johnlennon #centralpark #groupies #imagine #nyc #birthday··· (URL)	positive
central park	10/9	0.805	#newyorkcity #centralpark @ Top Of The Rock NYC (URL)	positive

表 7: ノイズツイート除去後のコーパスにより学習した場合の下位抽出例

ローカル語	日付	*cos*	ツイート本文	正解ラベル
oakville	10/7	0.135	#Oakville 13:30 ENE8.3kts G10.2kts 1017.92hPa Falling	positive
oakville	10/7	0.135	#Oakville 13:00 ENE6.5kts G8.8kts 1018.13hPa Falling	positive
new york city	10/9	0.192	This is #NewYorkCity @ DUMBO, Brooklyn (URL)	positive

ローカル語に対応するエリア集合から投稿されたローカル語を含まないツイートから，ローカル語が表す対象に関する観測情報を抽出した．多様な観測対象に対する抽出結果を確認するため，空間的に広い範囲を示すローカル語，特定の位置を示すローカル語，一時的なローカル語に対する抽出結果についてそれぞれ考察する．

まず，空間的に広い範囲を表すローカル語に対しては，同じローカル語でも日によって抽出結果に違いが見られた．表 9，10 にそれぞれ，10 月 7 日，9 日におけるローカル語 “texas” に対して抽出されたツイート例を示す．10 月 7 日にはテキサス州でメジャーリーグの試合が行われたため，これに関するツイートが多く投稿された結果，“texas” の意味表現が野球に関する単語により特徴付けられ，野球の観測情報が抽出されたと考えられる．一方 10 月 9 日は多様なツイートが抽出された．“texas” のような州名を表すローカル語の場合，ローカル語を含むツイートには多様なものが存在するため，特徴的な事象が発生しなければ意味を限定できず，適切に関連情報を抽出できないと考えられる．

次に特定の位置を示すローカル語に対する抽出結果として，10 月 7 日におけるローカル語 “sfo”，“alumni stadium” に対して抽出されたツイート例をそれぞれ表 11，12 に示す．“sfo” に対する抽出結果の上位 5 件のうち，4 件がサンフランシスコ国際空港に関する投稿であった．これらの投稿は互いに共通する単語も少なく，空港内の独立した観測情報であるが，“sfo” の意味表現が空港に関する単語により特徴付けられ，これらの単語に基づき正しく抽出できたと考えられる．また，“alumni stadium” はボストン大学にあるフットボール場であり，この日開催されたフットボールの試合に関するツイートが上位に抽出された．このように，ローカル語が空港やフットボール場である場合，州名などに比べると投稿される観測情報の内容が限定されるため，観測対象を表す単

表 9: 10/7 におけるローカル語 “texas”に対する抽出結果例

順位	*cos*	ツイート本文
1	0.671	Playoff baseball, great friends, and one mother of a hot dog. @ Globe Life Park in Arlington (URL)
2	0.651	win or lose we still know how to have a good time @ Globe Life Park in Arlington (URL)
3	0.634	October baseball #nevereverquit #becausebaseball #latergram @ Globe Life Park in Arlington (URL)
4	0.630	Game time!!! Let’s go Rangers! #NeverEverQuit @ Globe Life Park in Arlington (URL)
6	0.629	Bucketlist item...A big league playoff game. CHECK!!! #social #ALDS #mlb (@ Globe Life Park in Arlington - @mlb) (URL)

表 10: 10/9 におけるローカル語 “texas”に対する抽出結果例

順位	*cos*	ツイート本文
1	0.682	Just gonna sit down wind. I ran home dropped off Leia & put on an IOTA shirt so I don’t get fined. #meeting (URL)
2	0.636	Gets rid of the background singer. Goes full blues boogies. Less Nashville. More Rock n roll.… (URL)
3	0.629	In case you missed it, I’m giving away the last Jenn Saddle Bag in the ”sold out” tan color!… (URL)
4	0.627	Another great OU /Texas wknd n the books. Always great to see friends down here n Big D. And… (URL)
5	0.626	So much pain behind these smiles but we gotta keep PUSHin! Going harder than ever. Love you… (URL)

表 11: 10/7 におけるローカル語 “sfo”に対する抽出結果例

順位	*cos*	ツイート本文
1	0.644	Arrived, at what I still think is America’s nicest airport terminal. #flying #travel #SFOairport… (URL)
2	0.609	First flight on Alaska, let’s see what the future of Virgin America looks like. @ San Francisco… (URL)
3	0.591	And the ”Mother of the Year” award goes to the new mom in seat 2A on @jetblue flight 1435… (URL)
4	0.591	Fun House mirror in the women’s bathroom gives me the long legs I always wanted. (URL) (URL)
5	0.589	Omg there’s a Yoga room in this terminal and my flight is delayed. Namast! (@ San Francisco International Airport) (URL)

表 12: 10/7 におけるローカル語 “alumni stadium”に対する抽出結果例

順位	*cos*	ツイート本文
1	0.741	Just outside of #Boston. #3 Clemson in town to face #BC. #RedBandana Game tonight. ESPN right… (URL)
2	0.733	So many #Clemson fans! (@ Boston College in Boston, MA) (URL)
3	0.710	Attending Clemson vs Boston College tonight!!!! #bcfootball #clemson #lovecollegefootball @… (URL)
4	0.701	#fbf to Clemson @ BC 2012! Go Tigers! @ Boston College (URL)
5	0.678	Football game shenanigans. #bostoncollege #clemson #footballgame #bcvsclemson… (URL)

表 13: 10/7 におけるローカル語 “desert trip”に対する抽出結果例

順位	*cos*	ツイート本文
1	0.652	Watching the Rolling Stones sound check from the light house hill behind our campsite. Gonna be… (URL)
2	0.652	Dylan played more hits in that set than the previous three times I’ve seen him combined AND he still left out Like a Rolling Stone. Bad ass
3	0.640	Like a Rolling Stone is a great song but I mean, it’s SO Dylan to play a hits filled set and leave out his biggest one. Love it.
4	0.626	Just got a new beer to try. I’ll let you know what I think of it later.... (Night Owl Pumpkin Ale) (URL)
5	0.622	When I’m already missing my husband, but my fam I ain’t seen in a min ALL LOVE HIM, asking for him, now I’m messin him more eh haha

語によってローカル語が適切に特徴付けられ，ローカル語を含まなくても関連する観測情報を上位に抽出できたと考えられる．

最後に，一時的なイベントを表すローカル語に対する抽出結果として，10 月 7 日におけるローカル語 “desert trip” に対して抽出されたツイート例を表 13 に示す．この日はライブイベントである “desert trip” の初日であり，このライブに出演するミュージシャンに関するツイートが上位に抽出された．イベントなども投稿される観測情報の内容が限定されるため，ローカル語の意味表現を算出するのに十分な数のローカル語を含む観測情報が投稿されれば，ローカル語を含まないが関連する観測情報も抽出可能と考えられる．

4. おわりに

本研究では，Twitter に投稿されるツイートから多様な実世界観測情報を抽出する手法を提案した．提案手法ではまず，各地に存

在する多様な観測対象を表すローカル語を，単語の空間的局所性に基づき抽出し，抽出したローカル語を含むツイートを，各対象に関する観測情報とする．これらの観測情報から学習した単語間の意味表現に基づき，さらに，ローカル語を含まないが関連した観測情報を抽出する．2016 年の 30 日間にアメリカから投稿された，投稿位置を示すジオタグが付与されたツイートから，実世界観測情報を抽出したところ，空港やライブイベントのように観測情報の内容がある程度限定される対象に対しては，ローカル語の意味が適切な単語により特徴付けられ，これらの単語を含む関連する観測情報が正しく上位に抽出された．今後の課題として，ノイズツイートの自動的な判定方法，及び抽出結果の定量的評価方法の検討が挙げられる．

[謝辞]

本研究の一部は，科学研究費補助金（基盤（C）26330137，基盤（S）16H06302）の助成を受けたものである．

[文献]

[1] “Twitter,” https://twitter.com

[2] 榊 剛史, 松尾 豊, “ソーシャルセンサとしての Twitter −ソーシャルセンサは物理センサを凌駕するか？−”, 人工知能学会誌, 27(1), pp.67–74, 2012.

[3] A. Sheth, “Citizen Sensing, Social Signals, and Enriching Human Experience,” IEEE Internet Computing, 13(4), pp.87–92, 2009.

[4] T. Sakaki, M. Okazaki, and Y. Matsuo, “Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development,” IEEE Trans. on Knowledge and Data Engineering, 25(4), pp.919–931, 2013.

[5] 土屋 圭, 豊田 正史, 喜連川 優, “マイクロブログを用いた鉄道の運行トラブル発生期間および付帯情報の抽出”, データ工学と情報マネジメントに関するフォーラム, B3–2, 2014.

[6] K. Massoudi, M. Tsagkias, M. D. Rijke, and W. Weerkamp, “Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts,” Proc. European Conf. on Advances in Information Retrieval, pp.362–367, 2011.

[7] 藤木 紫乃, 上田 高徳, 山名 早人, “経時的な関連語句の変化を考慮したクエリ拡張による Twitter からの情報抽出手法”, データ工学と情報マネジメントに関するフォーラム, C9–5, 2013.

[8] K. Watanabe, M. Ochi, M. Okabe, and R. Onai, “Jasmine: A Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs,” Proc. Int’l Conf. on Information and Knowledge Management, pp.2541–2544, 2011.

[9] Z. Cheng, J. Caverlee, and K.Lee, “You are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users,” Proc. Int’l Conf. on Information and Knowledge Management, pp.759–768, 2010.

[10] H. -W. Chang, D. Lee, M. Eltaher, and J. Lee, “@Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage,” Proc. Int’l Conf. on Advances in Social Networks Analysis and Mining, pp.111–118, 2012.

[11] 上村 卓也, 新田 直子, 中村 和晃, 馬場口 登, “マイクロブログからのリアルタイム地域情報抽出”, データ工学と情報マネジメントに関するフォーラム, C7–1, 2017.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Proc. Int’l Conf. on Learning Representations, 12 pages, 2013.

[13] T. Mikolov, l. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” Proc. Int’l Conf. on Neural Information Processing Systems, pp.3111–3119, 2013.

[14] E. Brill, “A Simple Rule-based Part of Speech Tagger,” Proc. Workshop on Speech and Natural Language, pp.112–116, 1991.

[15] “TermExtract,” http://gensen.dl.itc.u-tokyo.ac.jp.

[16] “gensim,” https://radimrehurek.com/gensim/index.html

新田 直子 **Naoko NITTA**

1998 年，大阪大学基礎工学部情報工学科卒業．2003 年，同大大学院博士後期課程修了．2002〜2004 年，日本学術振興会特別研究員．2003〜2004 年，コロンビア大客員研究員．現在，大阪大学大学院工学研究科准教授．博（工）．メディア理解に関する研究に従事．

吉武 真人 **Masato YOSHITAKE**

2014 年，大阪大学工学部電子情報工学科卒業．2016 年，同大大学院博士前期課程修了．現在，株式会社本田技術研究所勤務．

中村 和晃 **Kazuaki NAKAMURA**

2005 年，京都大学工学部情報学科卒業．2010 年，同大大学院情報学研究科博士後期課程研究指導認定退学．2010 年，同大大学院法学研究科助手．現在，大阪大学大学院工学研究科助教．博士（情報学）．画像・映像認識，人物行動理解に関する研究に従事．

馬場口 登 **Noboru BABAGUCHI**

1979 年，大阪大学工学部通信工学科卒業．1981 年，同大大学院博士前期課程修了．1982 年，愛媛大学工学部助手．大阪大学工学部助手，講師，産業科学研究所助教授を経て，現在，大学院工学研究科教授．1996〜1997 年，UCSD・文部省在学研究員．工博．メディア処理，プライバシー保護画像処理に関する研究に従事．電子情報通信学会フェロー．