DEIM Forum 2017 D3-5

# マイクロブログからの関連実世界観測情報の抽出

吉武　真人†　　新田　直子†　　中村　和晃†　　馬場口　登†

† 大阪大学大学院工学研究科　〒 565–0871 大阪府吹田市山田丘 2–1
E-mail: †yoshitake@nanase.comm.eng.osaka-u.ac.jp,
††{naoko,k-nakamura,babaguchi}@comm.eng.osaka-u.ac.jp

**あらまし**　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，地名など各観測対象の位置を特定する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各単語を意味的に表現するベクトルを用いて，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．

**キーワード**　マイクロブログ，情報抽出，実世界観測情報

## 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観測対象の位置を示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]．各単語の時間的局所性も考慮すれば，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，各地の観測対象を表すと考えられるローカル語を含むツイートを抽出することにより，多様な対象の観測情報が抽出できるが，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時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，関連度の高いローカル

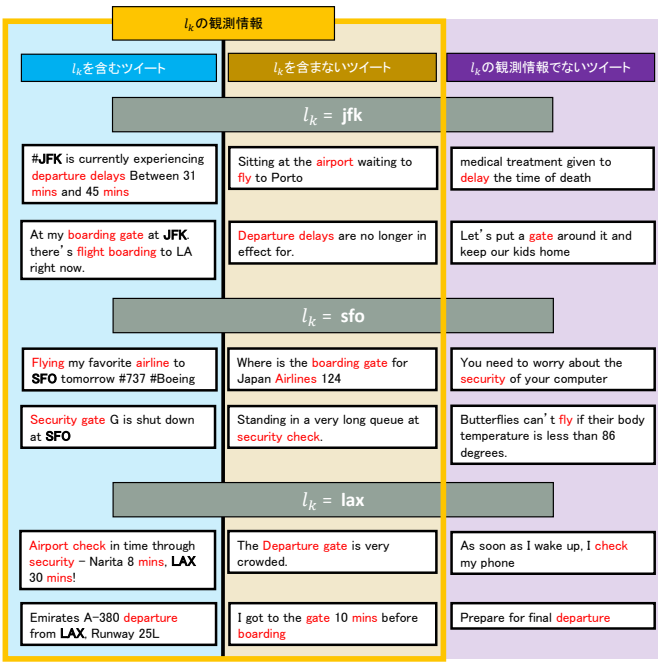


図 1　ローカル語を各地の空港とした時の観測情報の例

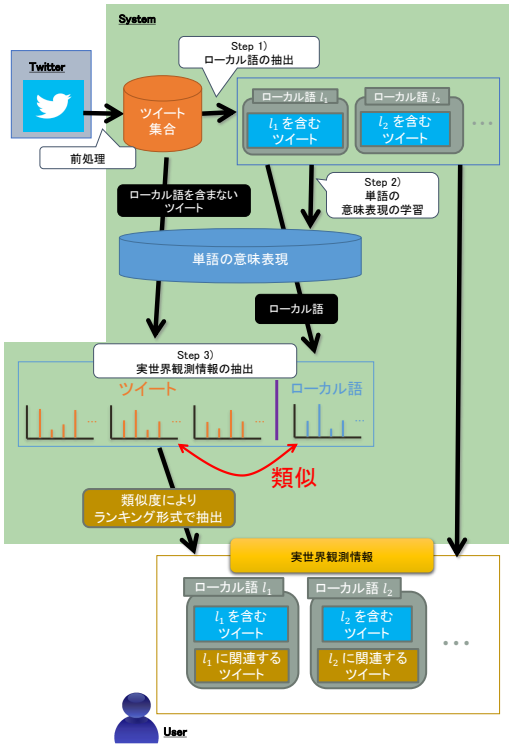


図 2　提案手法の概要

語を含まないツイートを，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，観測対象の位置を示すローカル語が含まれると考えられる．そこで，まず，ジオタグ付きツイートから，各単語の空間的局所性に基づき，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 $l_k$ に関するツイートには，ローカル語の有無に依らず “departure” や “boarding”，“gate” などの空港から連想しうる単語を多く含む．一方，$l_k$ に関連しないツイートには，ローカル語から連想しう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しうる単語同士は，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．

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，観測対象の位置を示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の時空間的局所性に基づいて抽出することにより，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，名詞を抽出する．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．

$$tfidf_{u_z}^{max} = f_{u_z}^{max} \cdot idf_{u_z} \quad (1)$$

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

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，$f_{u_z}^{max} = \arg\max_j f_{u_z,j}$ は同一エリアにおける $u_z$ 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 $tfidf_{u_z}^{max}$ は高くなるため，$tfidf_{u_z}^{max} \geqq R$ を満たす $u_z$ をローカル語 $l_k$ として抽出し，$u_z$ の出現履歴を削除する．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を適切に扱うため，$tfidf_{u_z}^{max} < r$ を満たす $u_z$ の出現履歴を削除し，$u_z$ がすでにローカル語として抽出されていた場合には，ローカル語集合からも削除する．

各ローカル語 $l_k$ に対し，$a_k$ 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，ローカル語に関するツイートを抽出するため，単語間の意味的な関係性を，前節で抽出したローカル語を含むツイート集合から学習する．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，このような表現に加え，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．

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 |

### 2.3 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 $a_k$ 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，エリア $a_k$ 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

2016/9/8〜2016/10/9 に，アメリカから投稿されたジオタグ付きツイートを収集し，30 日間のツイート 6,655,763 件を実験に用いた．　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 日間のツイートにおける，ローカル語を含むツイートをコーパスとして word2vec により単語の意味表現を学習した．次に，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，得られ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，実際にローカル語が示す場所から投稿されたローカル語を含まないツイートに対し，ローカル語に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

各パラメータを，$J = 256$，$R = 16.63$，$r = 6.97$ [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

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) |

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を示す州や都市を表す地名，“yankee stadium” や “sfo” のような特定の位置を示す施設を表す地名だけでなく，“desert trip” や “boston red sox vs new york yanees” のような一時的なイベントを表すイベント名がローカル語として抽出された．いずれのローカル語においても，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，イベント名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．

各ローカル語に対する実世界観測情報の一部として，各ローカル語が示す場所から投稿されたローカル語を含むツイートを実世界観測情報の一部として抽出したところ，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”，空間的に狭い範囲の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．ここに示した，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データセット内に存在する場合，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，このような単語のいずれかと関連性をもつローカル語の関連情報を抽出する際には，同じ形式のツイートが大量に抽出されると考えられる．ローカル語が表す観測対象に関する多様な実世界観測情報の抽出結果を評価する際，このようなツイートはノイズとなるため，テストデータから予め除去しておくことが必要であると考えられる．また，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

提案手法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．提案手法で抽出を目指す，ローカル語を含まないが，ローカル語に関連するツイートのテストデータを正確に作成することは困難であるが，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”#” 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．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，そのツイートに関連する観測対象を表す．本研究で抽出したいツイートは，ローカル語に関連するが，ローカル語を含まないツイートであるため，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たものを，正解ツイートとして抽出した．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については，これを考慮して抽出した．

次に，ローカル語が示す場所とは異なる場所から投稿されたツイートは，ローカル語には関連しないと仮定し，ローカル語

が示すエリア以外の場所から投稿されたツイートを，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，ローカル語が示す場所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，データセットの最後の 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，ローカル語の意味表現は各ローカル語を含むツイートに含まれる単語により算出されるため，ローカル語を含むツイートが少ない場合には，適切な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．

### 3.2.1 ノイズツイートの影響の評価

word2vec により，単語の意味表現を学習する際，コーパスに含まれるノイズツイートの影響を評価するため，初めの 25 日間のローカル語を含むツイートをコーパスとして学習したモデルと，そこからさらにノイズツイートを除去した上で，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．このとき，コーパス内のツイート数，単語数，語彙数は，ノイズツイートを除去していないコーパスは，659,615 件，4,552,955 語，146,691 語であり，ノイズツイートを除去したコーパスは，392,142 件，2,488,157 語，136,324 語であった．

実装には，gensim [14] 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，最適化をサンプル数 15 のネガティブサンプリングとした．このとき，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，ノイズツイートを除去していないコーパス，除去したコーパスの語彙数は，それぞれ 25,829 語，21,516 語となった．また，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．

コーパスにノイズを含む場合と含まない場合の比較結果を図 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，ノイズを含まないコーパスで学習したモデルの方が優れた結果となった．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，ノイズツイートは，同じ形式で大量に投稿されるため，反復回数が少ない場合においても，類似したツイートによって反復的に学習されるため，反復回数による影響が小さくなったと考えられる．

$iter = 100$ とし，ノイズツイートを除去したコーパス，除去していないコーパスでそれぞれ学習したとき，上位に抽出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，10/9 の “santa barbara” に関

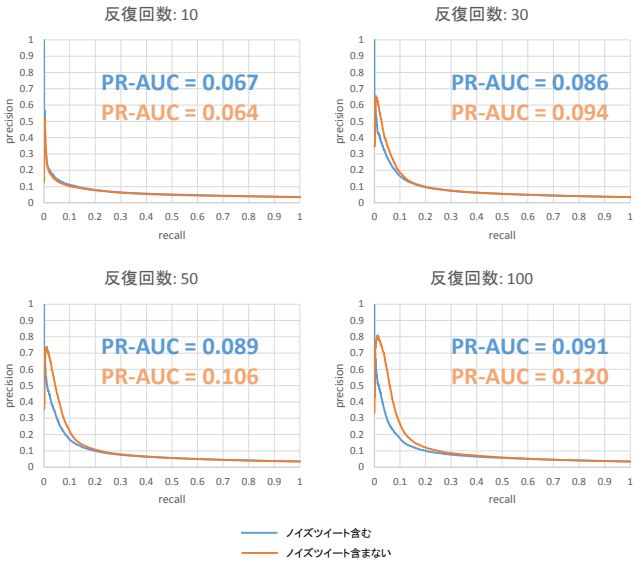


図 3 $iter$ を変化させた場合のコーパスにおけるノイズの有無による比較

する抽出結果に注目すると，ノイズを除去した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，ノイズを除去していない場合には，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には，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考えられる．10/4 の “texas” に関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データを上位に誤抽出したことがわかる．このとき，ノイズツイートを除去したコーパスの場合において抽出されている上位 2 件のツイートは非常に形式が似ており，これについてもノイズツイートとして除去すべきであった可能性がある．しかし，このようなノイズツイートを人手で完全に除去するのは困難であるため，これを機械的に除去するシステムを組み込むことが必要だと考えられる．

一方で，10/9 の “central park” に関する抽出結果のように，どちらのコーパスで学習しても，結果に違いが現れないローカル語も存在した．また，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．この結果より，どちらのコーパスで学習しても適切にツイートを抽出できるローカル語，誤抽出が多いローカル語の他に，ノイズツイートを除去したコーパスの方が，適切な観測情報を抽出できるローカル語も存在する．よって，以降の実験では，コーパスにノイズツイートを除去したものを用いる．

### 3.2.2 パラメータによる影響の評価

次に，word2vec による単語の意味表現の学習における適切なパラメータを得るため，パラメータを変化させ，評価用ツイート集合を用いて評価する．本研究と Mikolov らの実験環境におけ

表 5　ノイズツイート除去コーパスの場合の上位抽出例

| 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 #deviledeggs #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　ノイズツイートを除去していないコーパスの場合の上位抽出例

| 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　ノイズツイートを除去したコーパスの場合の下位抽出例

| 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 |

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 $dim$ は，Mikolov らは $dim = 50, 100, 300, 600$ の 4 通りに設定し，精度の変化を検証しているが，次元数を多くすればするほど精度が向上する結果となっている．よって，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，学習の反復回数については，7 億 8300 万語のコーパスにおいて，3 回することにより，精度の向上が示されている．Mikolov らの他の実験では，特に記述

表 8　パラメータによる抽出精度

| $dim \backslash iter$ | 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** |

されていないため，反復は行っていないと思われるが，本研究では，非常に小さなコーパスで学習するため，前節でも示した通り，反復回数は重要なパラメータであると考えられる．よって，反復回数 $iter$ を前節で設定した $10, 30, 50, 100$ に $200, 300$

表 9　10/7 におけるローカル語”texas”の場合の抽出結果例

| 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”の場合の抽出結果例

| 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”の場合の抽出結果例

| 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”の場合の抽出結果例

| 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) |

を加えた 6 通りについて検証する．

以上で設定した単語ベクトルの次元数と学習の反復回数について，PR-AUC を評価指標として網羅的にシミュレーションを行った結果を表 8 に示す．検証した中では，単語ベクトルの次元数と学習の反復回数のどちらに対しても，PR-AUC の値は単調増加であったが，$dim = 600, iter = 200$ 程度で精度はほぼ収束した．よって，以降の実験では，$dim = 600, iter = 200$ とする．

### 3.3　実世界観測情報の抽出結果の考察

最後に，提案手法により，実際に各ローカル語が示す場所から投稿された，ローカル語に関連するが，ローカル語を含まないツイートを抽出した．多様な実世界観測情報を抽出できるかを評価するため，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示すローカル語に関する関連情報の抽出結果を示す．この場合，同じローカル語に対し異なる日付において抽出した結果に違いが見られた．この結果の例として，10/7，10/9 におけるローカル語 “texas” に関する抽出ツイート例をそれぞれ表 9，10 に示す．10/7 には，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って特徴付けられ，野球の観測情報が抽出されたと考えられる．一方，10/9 には，texas に関連しない多様なツイートが抽出された．これは，“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は多様なものが存在する．そのため，特徴的な事象が発生しなければ，特定の観測対象に関する単語により特徴付けられないため，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関する関連情報の抽出結果として，10/7 におけるローカル語 “sfo”，“alumni stadium” に関する抽出ツイート例を表 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．特に，3 位に抽出されたツイートは，幼児と飛行機に搭乗する母親の気遣いを賞賛する内容であり，少数の人間しか観測していないと思われる情報であっても抽出で

表 13　10/7 におけるローカル語”desert trip”の場合の抽出結果例

| 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 |

きた．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は限定されると考えられるため，観測対象に関連する単語によってローカル語が適切に特徴付けられ，ローカル語に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関する関連情報の抽出結果として，10/7 におけるローカル語 “desert trip” に関する抽出ツイート例を表 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．この場合も，“sfo” の例と同様に，ローカル語を含む投稿がライブに関する内容に限定されるため，これに関する単語によってローカル語が適切に特徴付けられ，ローカル語に関連する観測情報を上位に抽出できたと考えられる．

## 4. ま と め

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ため，地域に特有な語であるローカル語を含むツイートから単語間の意味的な関係性を学習し，各ローカル語を観測対象として，関連する実世界観測情報を抽出する手法を提案した．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など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，それぞれの観測対象に関連する単語によりローカル語が特徴付けられたため，観測情報が上位に抽出された．今後の課題として，ノイズツイートの自動的な除去方法，及び，抽出結果の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)2633037）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？―”, 人工知能学会論文誌, Volume 27, Issue 1, pp. 67–74, 2012.
[3] A. Sheth, “Citizen Sensing, Social Signals, and Enriching Human Experience,” IEEE Internet Computing, Volume 13, Issue 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 Transactions on Knowledge and Data Engineering, Volume 25, Issue 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”, 第 6 回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,” Proceedings of European Conference 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”, 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, 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,” Proceedings of the 20th ACM International Conference 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,” Proceedings of ACM International Conference 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,” Proceedings of International Conference 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”, 第 9 回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,” Proceedings of International Conference 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,” Proceedings of International Conference on Neural Information Processing Systems, pp. 3111–3119, 2013.
[14] “gensim,” https://radimrehurek.com/gensim/index.html