DEIM2020 J4-3(day1 d18)

SNSを用いた地域の『やわネタ』検出

山城　拓郎†　　牛尼　剛聡††

† 九州大学芸術工学部　〒 815–8540 福岡県福岡市南区塩原 4–9–1
†† 九州大学芸術工学研究院　〒 815–8540 福岡県福岡市南区塩原 4–9–1
E-mail: †yamashiro.takurou.784@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　本研究では、Twitter に投稿された情報を用いて、それぞれの地域における観光やグルメなどのカジュアルな話題である『やわネタ』を抽出する手法を提案する。本手法では、まず、Twitter に投稿されたジオタグ付きツイートを訓練データとして、ツイートのテキストのみからそのツイートが対応する地域を予測する分類機を機械学習を利用して学習させ、その学習器を用いたブートストラップ法によって学習器の訓練を繰り返し、地域の特徴を幅広く抽出する。次に、対象とする地域に関して他のユーザが興味を持つ話題を頻繁に投稿するユーザを抽出し、そのユーザらが投稿する情報から、対象とする地域の『やわネタ』を抽出する。

キーワード　地理情報と SNS，ソーシャルビッグデータ，機械学習

1 はじめに

近年，スマートフォンをはじめとする個人用情報デバイスの普及に伴い，ソーシャル・ネットワーキング・サービス (SNS) が発展し、利用者数が増大している。また、スマートフォンに搭載された GPS を利用した位置情報を用いたサービスが拡大している。

Twitter[1]は、ツイートと呼ばれる 140 文字以下の短文を投稿する SNS である。Twitter では、ユーザのそのときどきの自分の状況や感じたことをリアルタイムに投稿するが、投稿には投稿した場所の位置情報を付加できる。位置情報が付加されたツイートを、位置情報付きツイートと呼ぶ。位置情報付きツイートを利用すれば、そこに含まれるテキストと位置情報によって、特定の位置でどんなことが起こっているのかを把握できる可能性がある。これまでにも、位置情報付きツイートを利用して Twitter を様々な実世界でのイベント検出のためのソーシャルセンサとして利用する研究が数多くなされている [1]。

ソーシャルセンサとは、Twitter やブログなどのソーシャルメディアに投稿された内容を利用して実世界を観測する時に、ユーザを物理センサと同様の機能を持つセンサとして捉える考え方で、人間が観測できる事象すべてを感知することができる。近年は Twitter などの即時性の高い SNS が拡大したことで、事象が発生してからそれを発信するまでのタイムラグが小さくなり、短期間での時系列変化を対象にした研究が可能になってきている。

Twitter をソーシャルセンサとして用いる際に問題になる点として、位置情報が付加されているツイートが少なく [2]、日本語のツイートに対しては約 0.18%しか位置情報が付与されていない [3] ことが挙げられる。この理由から、 Twitter をイベント検出のためのソーシャルセンサとしてそのまま利用することは困難である。

1 : https://twitter.com

一方『やわネタ』とは、マスメディア業界で使われる用語で、様々なニュースの中でも政治や社会問題などの扱うのに慎重さを必要とする重たい話題である『かたネタ』と対照的に、地域のグルメ情報やイベント情報などのように肩の力を抜いて楽しむことができる話題である。

『かたネタ』は公共の情報ソースやテレビや新聞などのメディアで比較的簡単に知ることができるが、『やわネタ』を集めるには口コミを集めたり取材をする等の人的コストが必要になる。

本研究では、位置情報が付加されていないツイートに関連付く位置を予測し、地域のカジュアルな話題である『やわネタ』を検出する手法について述べる。

2 関連研究

本章では、本研究に関連してツイートの位置推定に関する研究について述べる。

2.1 TF-IDF の応用

森國ら [4] [5] は、実世界の空間をいくつかのエリアに分割し、ツイートの投稿が行われたエリアを式 (1) を用いた最尤推定により推定している。A は推定対象エリアの集合、a は推定対象エリア、W_t はツイート t に含まれる単語の集合、$p(w)$ はデータ中 で単語 w が出現する確率、$p(a|w)$ は単語 w がエリア a で出現する確率である。

$$\arg\max_{a \in A} p(a;t) = \sum_{w \in W_t} p(a|w)p(w) \tag{1}$$

また、単語の位置を推定する際にノイズとなるツイートを排除するため、TF-IAF フィルタを利用している。TF-IAF は式 (2) で示される。$C(w)$ は単語 w の総出現回数、A は推定対象エリアの 集合、A_w は単語 w の出現したエリアの集合である。

$$
\begin{aligned}
TFIAF(w) &= TF(w) \cdot IAF(w) \\
TF(w) &= \log(C(w)) \\
IAF(w) &= \log \frac{|A|}{|A_w|}
\end{aligned}
\tag{2}
$$

TF-IAF フィルタは、情報検索などに使われる TF-IDF [6] の概念を用いたフィルタで、地域ごとに偏りのない単語をノイズとして除去できる。これを利用することによって地理的な分布に影響を及ぼさない単語を除去し、推定精度を向上させる。本来の TF-IDF の概念をそのまま適用すると、TF の計算式はそれぞれのエリアごとの単語 w の出現回数になるはずであるが、この研究ではツイートに位置情報を関連付けることが目的となっているため、単語の出現頻度の総数を利用している。

2.2 ジオコーディング

宮内ら [7] は、ジオコーディングと呼ばれる技術を利用してツイートの位置を推定している。つつとは、地名やスポット名を含むツイートに対してその場所の位置情報を付加する手法である。

この研究では、ジオタグ付きツイートの中に含まれる固有名詞をスポット名の候補とし、あらかじめ作成しておいたスポット情報辞書の情報をもとにジオコーディングによって位置情報を付加する。ジオタグが持つ位置情報とジオコーディングによって付加された位置情報に 10 km 以上の誤差が存在すれば不正解, 1 km 以下の誤差であれば正解のラベルをツイートに付加する。誤差が 1 km 以上 10 km 以内であるツイートに関しては判定が困難であるため教師データとして利用しない。このようにしてラベルが付加されたツイートの様々な特徴を分析し位置推定のための学習に適したツイートの持つ特徴について分析している。

本研究では、従来のエリアごとの分類タスクに Bi-directional LSTM を用いたブートストラップ法によってラベル付きデータを増やすことで、より有用な地域に関する情報を抽出する。

3 提 案 手 法

3.1 概　　要

本研究では、ツイートの投稿の位置の推定のために実空間をメッシュで分割し、ツイートに関連づくメッシュを推定する。メッシュの分割方法は JIS として規定された地域メッシュ [8] を参考にする。地域メッシュは、それぞれの区画を緯度と経度の区間によって分割したものであり、区間の長さごとに 6 段階で規定されている。例えば第 3 次地域メッシュでは、緯度差は 30 秒、経度差は 45 秒で、1 辺の長さは約 1km である。本稿では第 3 次地域メッシュを南北に 3 つ、東西に 3 つ並べた 9 つを 1 つの約 3 km 四方のメッシュとして利用する。

本研究で提案する手法では、Bi-directional LSTM (Long short-term memory) [9] によってジオタグ付きツイートのテキストデータから、ツイートが投稿されたメッシュを予測するモデルを構築する。次に、ブートストラップ法 [10] によってそれぞれのメッシュと関連付けられるツイートの量を増やしメッシュごとの特徴を幅広く抽出していく。

3.2 機械学習を用いた地域メッシュごとの特徴量の学習

ツイートの位置推定のために Bi-directional LSTM [9] を用いたモデルを利用する。LSTM [11] は代表的な再帰的ニューラルネットワーク (RNN) の一種であり、時系列情報を学習する際に勾配が消滅する問題を解決している。Bi-directional LSTM は、時系列順方向に学習した LSTM のモデルと時系列逆方向に学習した LSTM のモデルの出力を統合したものである。Bi-directional LSTM は、LSTM よりデータ全体のコンテクストを適切に学習することができると言われている。

本手法で利用するモデルに対して、ツイートのテキストデータをに対して形態素解析を行ったものを入力データとし、ツイートの位置を表すメッシュの 1 次元 One-hot ベクトルを正解ラベルとする。

3.3 AF フィルタ

森國ら [4] [5] の研究では、ツイート地域分類に利用するデータを選別するのに TF-IDF の概念を利用したいくつかのフィルタを利用している。中でも該当の単語が出現する頻度をフィルタリングの対象にした AF フィルタは最も高い効果を発揮している。本研究では、地域の情報を表す特徴量を抽出する際にノイズとなる単語をフィルタリングするために、AF の大きい単語を除去する手法を利用する。AF は式 3 で表される。A_w は単語 w が出現したエリアの集合である。

$$
AF(w) = |A_w| \tag{3}
$$

3.4 ブートストラップ法

ブートストラップ法 [10] は、半教師あり学習の一種で少量のラベル付きデータを利用したサンプリングを繰り返すことによって、ラベル付きデータの量を増やすフレームワークである。本研究では、少量の位置情報付きツイートを利用してメッシュごとのラベル付きデータを増やすためにこの手法を利用する。

今回提案する手法では、図 1 、図 2 のようにしてサンプリングを行う。まず、図 1 のように構築した学習済みモデルを利用してジオタグのついていないツイートがどのメッシュでツイートされたものを判別させ、それぞれのメッシュに対しての確率がパラメータとなったベクトルを得る。ある特定のメッシュの確率が閾値より高かったものをその地域に関連するツイートとしてサンプリングし、そのツイートにメッシュごとのラベルを付加する。このときラベルを付加したものを 1 次地域ツイートと呼ぶ。

次に同じように、ラベルを付加したツイートのみを集めて、再度 Bi-directional LSTM によってどのメッシュのラベルが付加されているのかを予測するモデルを構築し、その学習済みモデルを利用してジオタグの付いていないツイートがどのメッシュでツイートされたものを判別させ、メッシュごとのラベルを付加する。このときラベルを付加したものを 2 次地域ツイートと呼ぶ。

このようにして、ジオタグの付いていないツイートを判別・

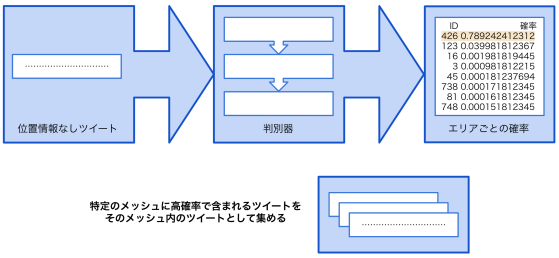

図 1　ジオタグの付いていないツイートを分類

学習を繰り返すことでそれぞれのメッシュと紐づくツイートの量を増やしていく。

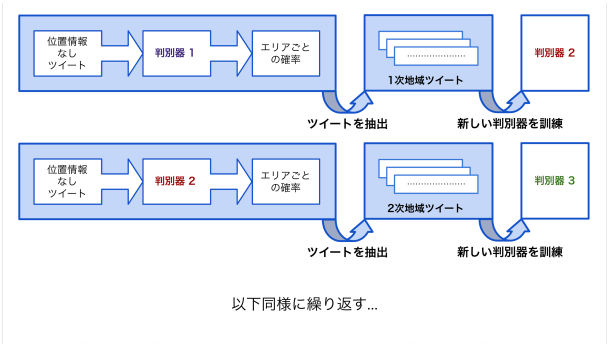

図 2　判別・学習を繰り返しサンプリング

3.5　トピックを示すキーワードの利用

任意のトピックを示すキーワードを設定する。トピックを示すキーワードは BTM (Biterm topic model) [12] によってツイートから抽出したものの中から選定する。BTM とは、biterm と呼ばれる同じ文書に出現する単語対の集合を分析することで共起しやすい単語の集合としてトピックを抽出する手法であり、従来のトピックモデルである LDA と比べて、短いテキストでも適切にトピックを抽出できるのが特徴である。

抽出した n 次地域ツイートの中から設定したキーワードを含むツイートを検索し、ヒットしたものを『やわネタ』とする。

4　実　　験

4.1　データセット

Twitter Streaming API[2] を用いて、2019 年 8 月 10 日から 2019 年 12 月 10 日の期間で福岡県を含む矩形領域内で投稿されたジオタグ付きツイートを 158,437 件取得した。ツイートの特徴を確認するために、メッシュごとに TF-IAF を計算し、TF-IAF 値が大きい単語を抽出しマップ上に表示させた。なお、本章で用いる TF-IAF は 2 章で述べた式 (2) で表されるものではなく、本来の TF-IDF の概念をそのまま用いており式 (4), (5), (6) で示される。$C(w_a)$ は単語 w のエリア a での出現回数、A は推定対象エリアの 集合、A_w は単語 w の出現したエリアの集合である。

2 : https://dev.twitter.com/streaming/overview

$$TFIAF(w,a) = TF(w,a) \cdot IAF(w) \tag{4}$$

$$TF(w,a) = \log(C(w_a)) \tag{5}$$

$$IAF(w) = \log \frac{|A|}{|A_w|} \tag{6}$$

図 3 のように 1 件以上のジオタグ付きツイートが投稿されたメッシュの北西の角にピンを立てた。山間部や海上にはジオタグ付きツイート 1 件も投稿されていないメッシュが存在することがわかる。

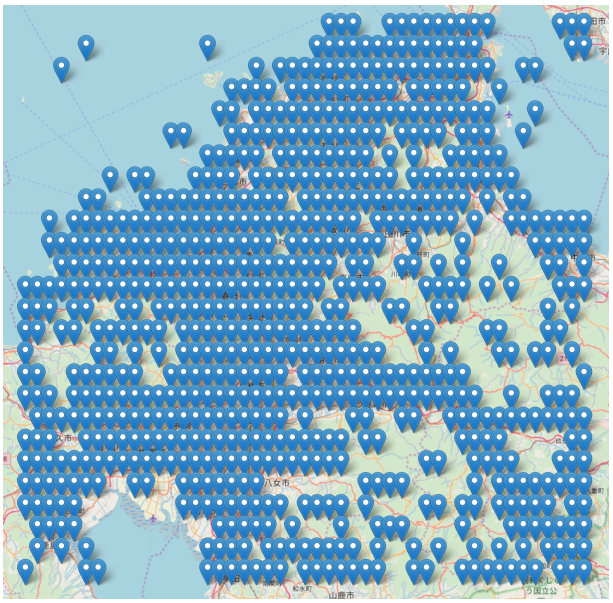

図 3　福岡県内のジオタグ付きツイートが存在したメッシュ

図 4 、図 5 は、それぞれ博多駅周辺、九州大学伊都キャンパス周辺の特徴語をマップ上に表示させている様子である。地域のランドマークや店などが特徴語として抽出できている。

ジオタグ付きツイートのうち 20% の 31,687 件をテストデータとし残りの 128,630 件を訓練に使用した。また、前処理としてツイートの中に含まれる URL を除去した。

4.2　判別・学習に使うモデルの構築

ツイートを学習する前に、いくつかの前処理を行った。まず文章を単語ごとに分かち書きにするために mecab-ipadic-NEologd[3] を使用した。モデルの埋め込み層として学習済み fastText[4][5] を利用した。学習済み fastText は 2017 年 01 月 01 日時点の日本語版 Wikipedia[6] のデータをコーパスとして学習している。

モデルの詳細なパラメータは表 1 の通りである。

3 : https://github.com/neologd/mecab-ipadic-neologd
4 : https://fasttext.cc/
5 : https://drive.google.com/file/d/0ByFQ96A4DgSPUm9wVWRLdm5qbmc/view/
6 : https://ja.wikipedia.org

図 4　博多駅周辺の特徴語

表 1　Bi-directional LSTM のパラメータ

損失関数	categorical cross entropy
最適化手法	Adam
バッチサイズ	100

Layer	Output Shape
Embedding	100, 300
Bi-directional LSTM	64
Dense	1290

4.3　ジオタグ付きツイートを用いた学習

図 6 はジオタグ付きツイートを用いた学習におけるエポックごとの正解率の推移である。20 エポックの時点で、正解率は 83.1% であった。この結果から、ジオタグ付きツイートにはそれぞれの地域の特徴が十分に表れていると考えられる。

さらに、地域の情報を表す特徴量を抽出する際にノイズとなる単語を除去するため AF フィルタを利用する。表 7 はツイート中に出てくる単語を全て分かち書きにし AF の大きい方から降順に並べた時の全体の単語数に対する累積比の値を示したものである。極端に出現回数の大きな単語がツイートに含まれる単語数の大部分を占めていることがわかる。

AF の大きな単語から順にフィルタリングの対象とした。フィルタリングの対象となった単語はベクトルとして埋め込む際に全てのパラメータが 0 のベクトルとして処理する。図 8 は全単語数に対するフィルタリング対象の単語数の割合を 10%ずつ変

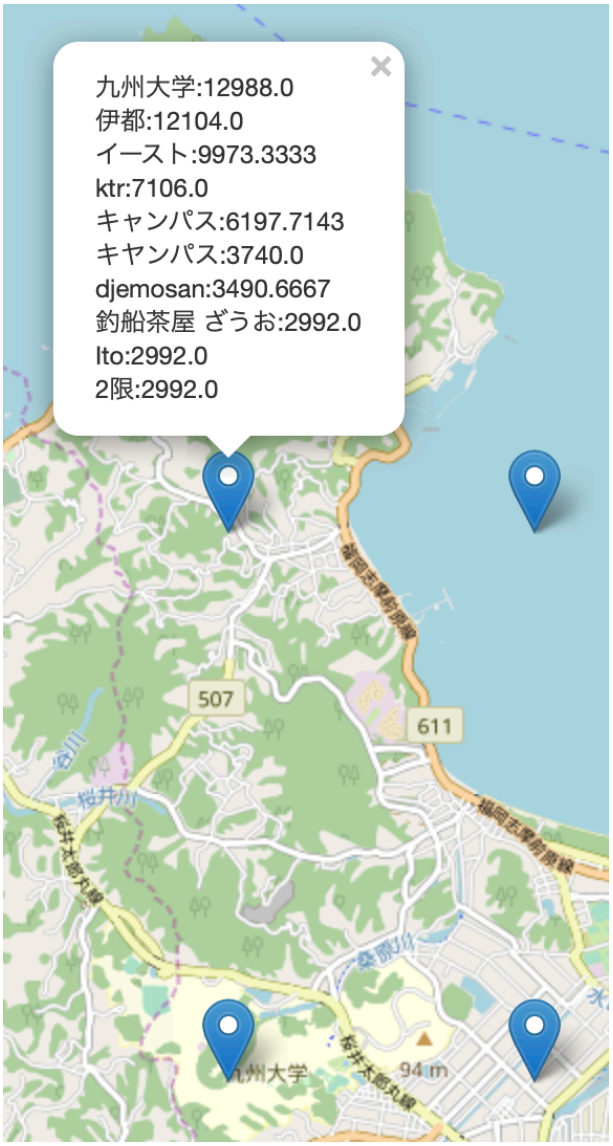

図 5　九州大学伊都キャンパス周辺の特徴語

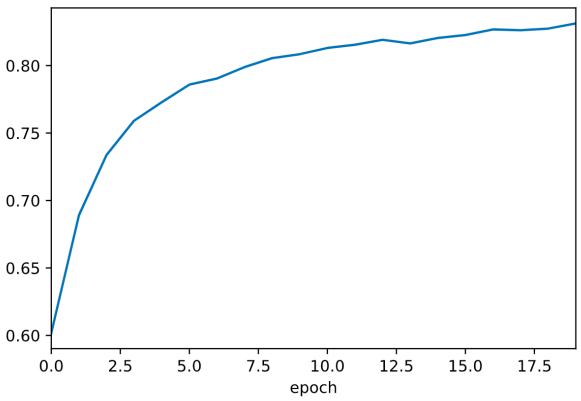

図 6　正解率の推移

化させた時のジオタグ付きツイートを用いた学習における 20 エポックの時点での正解率の推移である。全体の 2 割の単語をフィルタリングした時がもっとも高い正解率 (83.4%) を記録した。全単語数の 6 割まではフィルタリングしても大きく正解率

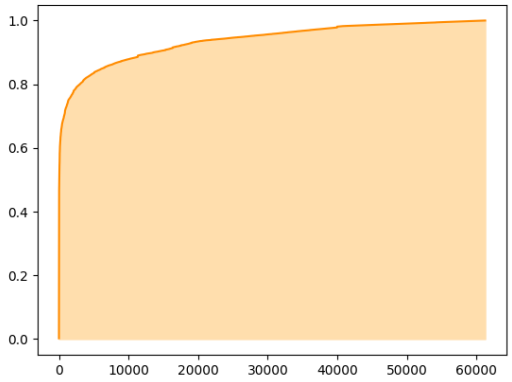

図 7　単語を AF の大きいものから並べた時の累積比

に影響を与えないことがわかる。

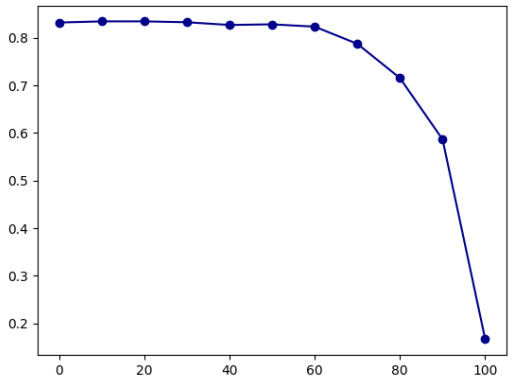

図 8　フィルタリング対象の単語数の割合を 10%ずつ変化させた時の正解率の推移

4.4　ブートストラップ法

ジオタグ付きツイートを用いて構築した学習済みモデルを用いてブートストラップ法を行った。入力するジオタグの付いていないツイートとして、Twitter のプロフィールの場所を “福岡” に設定しているユーザが 2019 年 12 月 20 日に行ったツイート 573,442 件を利用した。それぞれのメッシュに対しての確率がパラメータとなったベクトルの最大値が 0.5 を超えたツイート 66,921 件 を 1 次地域ツイートとして収集した。収集したメッシュごとのツイートの件数は表 2 のようになった。

メッシュ番号 485 の示すメッシュは、博多駅および周辺の繁華街地域を含む。収集した 1 次地域ツイートの 85.3% がメッシュ番号 485 の示すメッシュと紐づけられている。

表 3 はそれぞれのメッシュと紐づけられたツイートの一例である。表 3 から地域の情報とは全く関連のない、本質的に位置を特定することが困難なツイートも抽出してしまっていることがわかる。特にメッシュ番号 485 の示すメッシュでは、表 4 に示されるような地域の特徴ではないノイズを多分に学習してしまった可能性が考えられる。

表 2　メッシュごとのツイートの件数

メッシュ番号	件数
485	57084
128	1436
438	1302
488	1241
725	1129
486	846
483	757
533	329
436	315
383	303
…	…

表 3　収集した 1 次地域ツイート

メッシュ番号	内容	確率
485	びゃあああああ	0.6938424
486	博多駅前 一等地 の ビル 建て替え は 大きい 。	0.95899934
484	中央区 の 福岡ドーム です 。	0.63920003
…	…	…

表 4　学習に用いた位置情報付きツイート

メッシュ番号	内容
485	ぴえん (@東京純豆腐 in 福岡市, 福岡県)
486	ただいまあああああああああああああああああああああ (@博多駅 in 福岡市, 福岡県) c
487	楽しかったンゴ帰るンゴまたくるンゴ (@福岡空港 in 福岡市, 福岡県) c
…	…

4.4.1　他の地域のツイートを用いた学習

ノイズによる影響を複数のメッシュに分散させるために、ノイズを含むツイートが大量に存在すると考えられる他の地域のツイートを用いて学習を進める。Twitter Streaming API[7] を用いて、2019 年 12 月 1 日から 2019 年 12 月 10 日の期間で東京 23 区を含む矩形領域内で投稿されたジオタグ付きツイートを 348,053 件取得した。福岡県内で投稿されたジオタグ付きツイートと東京 23 区内で投稿されたジオタグ付きツイートを連結し一つの位置情報付きツイートデータとした。この位置情報付きツイートデータ 543,889 件を 4.2 で用いた学習器と同じモデルでツイートを学習させた。学習器の埋め込み層では、AF の大きな単語から順に全単語数の 20%は全てのパラメータが 0 のベクトルとして処理した。20 エポックの時点で、正解率は 79.2%であった。

4.4 と同様にジオタグ付きツイートを用いて構築した学習済みモデルを用いてブートストラップ法を行った。入力するジオタグの付いていないツイートとして、Twitter のプロフィールの場所を “福岡” に設定しているユーザが 2019 年 12 月 20 日に行ったツイート 573,442 件を利用した。それぞれのメッシュ

7 : https://dev.twitter.com/streaming/overview

に対しての確率がパラメータとなったベクトルの最大値が 0.5 を超えており、該当のメッシュが福岡県内のメッシュであるツイート 12,934 件 を 1 次地域ツイートとして収集した。収集したメッシュごとのツイートの件数は表 5 のようになった。全体として抽出されたツイートの数が減っているが、特にメッシュ番号 485 の示すメッシュと紐づくツイートの数が減っており、地域と関連のないツイートの抽出を抑えることができたと考えられる。

表 5　メッシュごとのツイートの件数

メッシュ番号	件数
485	9021
488	1074
128	489
486	471
483	255
436	181
440	158
857	137
484	119
438	105
…	…

4.5　トピックを示すキーワードの利用

1 次地域ツイートを BTM によって分析し、トピックを示すキーワードを抽出していく。また BTM によって分析するツイートの前処理として、固有名詞・一般名詞・サ変接続名詞以外の品詞を持つ単語、ひらがな・カタカナ・アルファベット・数字のいずれかで構成され単語長が 1 のものを除去した。B 抽出したキーワードを含むツイートを検索した。例としていくつかのキーワードで検索した結果を示す。表 6, 7, 8 はそれぞれ “フェス”, “うどん”, “学校” を含む 1 次地域ツイートの一部である。結果により、トピックに関連するそれぞれの地域の話題を抽出できていることがわかる。

表 6　“フェス” を含む 1 次地域ツイート

メッシュ番号	ツイート内容
485	【新しい記事です】HASAMI・ MINO 器フェス詳細はこちら #fear #イベント #event #福岡
485	赤い衣装を新調しましたので 赤い 4 人組であの曲歌います 特別なこと満載だよー よかもんフェス 2019Xmas 直前 SP
128	【日曜日は「北九州 2019」へ】、 消防車をはじめ防災関係機関の特殊車両の展示や、 人気キャラクターショー、 テレビでおなじみの天達 気象予報士による防災講演など、 楽しみながら防災について学べるイベントです。
…	…

表 7　“うどん” を含む 1 次地域ツイート

メッシュ番号	ツイート内容
485	丼ぶりと京風うどんの「なか卯福岡半道橋店」 グランドオープン
485	福岡市役所の皿うどん 480 円。 笑ってしまうほど野菜たっぷり。 リンガーハットなら 680 円くらいでしょうか。 儲からないのか近々食堂閉鎖。惜しい。 #福岡ランチ
485	「イチカバチカ」のごぼ天うどん「やまみ」の博多天ぷら 「かわ屋」のとり皮「おにまる」 の海鮮全般「田中田」のぜいたく丼 「梟」の豚骨担々麺「想夫恋」の日田焼きそば 福岡出身者りな助のガチおすすめ
…	…

表 8　“学校” を含む 1 次地域ツイート

メッシュ番号	ツイート内容
299	古賀市千鳥小学校にて、 ドラムと昼休みミニライブを行ってきました。 講師バンドは私を含め学習支援に 熱意のある福岡のミュージシャン達が集まり、 千鳥小に夢を与える為に結成した 『CHIDREAMS』です。小学校で授業…
535	小学校で預かっている迷子のインコちゃん。 飼い主さんは、さぞかし心配されてる事でしょう 。。。 #春日市 #セキセイインコ
484	#明日の福岡コンサート 2019 年 12 月 21 日 土 14:00 黒崎ひびしんホール・中ホール 福岡県立八幡工業高等学校吹奏楽部 第 53 回定期演奏会
…	…

5　ま　と　め

本稿では、位置情報が付加されていないツイートに位置情報を付加し、地域の特徴を幅広く学習させつつ、ユーザベースの情報を利用することで地域のカジュアルな話題である『やわネタ』を検出する手法を提案した。今後は、ユーザベースの情報や時系列情報を利用してジオタグのついていないツイートから地域に関する特徴を抽出できるようにする予定である。

文　　献

[1] 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?(¡特集¿Twitter とソーシャルメディア). 人工知能学会誌, Vol. 27, No. 1, pp. 67–74, January 2012.

[2] 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 (CIKM), pp. 759–768, 2010.

[3] 橋本康弘, 岡瑞起. 都市におけるジオタグ付きツイートの統計.

人工知能学会誌, Vol. 27, No. 4, pp. 424–431, July 2012.
[4] 森國泰平, 吉田光男, 岡部正幸, 梅村恭司. ツイート投稿位置推定のためのノイズとなる単語の除去手法. 第 7 回データ工学と情報マネジメントに関するフォーラム, DEIM '15
[5] 森國泰平, 吉田光男, 岡部正幸, 梅村恭司. ツイート投稿位置推定のための単語フィルタリング手法. 情報処理学会論文誌 データベース, Vol. 8 No. 4, pp. 16–26, December 2015
[6] 北研二, 津田和彦, 獅々堀正幹. 索引語の抽出と重み付け. 情報検索アルゴリズム, pp. 33–40. 2002.
[7] 宮内天士, 白川 真澄, 原 隆浩, 西尾章治郎. 高精度なツイート投稿位置の推定に有効なルールの抽出. 第 8 回データ工学と情 報マネジメントに関するフォーラム, DEIM '16
[8] https://www.jisc.go.jp/app/jis/general/GnrJIS NumberNameSearchList?showjisStdNo=X0410
[9] Graves, A., Fernández, S., Schmidhuber, J. (2005). Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. He 15th International Conference on Artificial Neural Networks: Formal Models and Their Applications - Volume Part II, 799–804.
[10] 小町 守，工藤 拓，新保 仁，松本裕治: Espresso 型ブートストラッピング法における意味ドリフトのグラフ理論に基づく分析，人工知能学会論文誌，Vol.25, No.2, pp.233–242 (2010).
[11] Sepp Hochreiter and Jurgen Schmidhuber, " Long short-term memory", Neural computation, Vol.9 No.8 pp.1735–1780, 1997.
[12] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng, "A Biterm Topic Model for Short Texts", WWW '13 Proceedings of the 22nd international conference on World Wide Web, pp.1445-1456, 2013