1L-05

# Twitterで発信される病気症状の可視化に向けたTwitterユーザの居住地推定手法の検討

松本真拓† 松原香太‡ 安藤一秋†‡

香川大学

## 1. はじめに

Twitterをソーシャルセンサとして利用することで，人間社会における物理的・社会的イベントをリアルタイムに検出する研究が行われている[1].

本研究では，Twitter上から様々な病気症状を含むツイートを抽出し，地域別・時系列別に可視化することを目的とする．病気情報をリアルタイムに取得して地図上に表示することで，病気・症状の流行を把握できるだけでなく，原因や特徴の分析，未知の病気の発生検知なども可能になると考えられる．

本稿では，Twitter上で発信される様々な病気症状の地域別での可視化に向け，ツイートしたユーザの居住地を都道府県別に推定する手法について検討する．我々の先行研究[2]では，地名に属する固有表現の出現回数および固有表現と共起する語を素性として居住地を推定する手法を提案した．本稿では，推定性能の向上を目指して，場所を示す固有表現を含むツイート文をクラスタリングし，ツイートに出現する固有表現の居住を示す尤度に基づいて居住地を推定する手法について検討する．

## 2. 関連研究

日本に在住しているTwitterユーザの居住地を推定する手法としては，投稿内容を用いるコンテンツベースの手法やユーザのフォロー関係を用いるグラフベースの手法などが存在する．

コンテンツベースの手法として，森國らは，tf-idfの概念を用いてノイズとなる単語をフィルタリングすることで，都道府県別のツイート投稿位置を約22.0%の正解度で推定する手法[3]を提案している．また，坂本らは，「バカ」や「アホ」などの類語の使用頻度の違いにより，7地方における居住地を49.0%のF値で推定する手法[4]を提案している．

グラフベースの手法として，廣中らは，位置情報付きツイートを用いてソーシャルグラフを作成し，ユーザの居住地を推定する手法[5]を提案している．市区町村別の推定で29.2%，都道府県別で53.8%のF値を得ている．

我々の先行研究[2]では，ツイートから抽出した地名に属する固有表現の出現回数および固有表現と共起する語を素性として，LightGBM[6]で推定する手法を提案し，都道府県別の推定で72.0%の適合率を得ている．

本稿では，固有表現の出現するツイート文の周辺語をより考慮したモデルを検討することで，推定性能の向上を目指す．

---

Twitter User's Residence Prediction for Visualization of Disease Symptom in Tweets

†Masahiro Matsumoto・Kagawa University

‡Kyota Matsubara・Kagawa University

†‡Kazuaki Ando・Kagawa University

## 3. ベースライン手法と提案手法

本稿では，我々の先行研究[2]と同様，場所を示す固有表現辞書を基に，Twitterユーザの居住地を推定する．

先行研究では，場所を示す同じ固有表現が含まれていても，出現するツイートの内容によって，固有表現が居住地を示す場合と旅行先や地元などの居住地以外の場所を示す場合を区別していなかった．そこで，固有表現を含むツイート文をクラスタリングし，ツイートに出現する固有表現の居住を示す尤度に基づいてユーザの居住地を推定する手法を検討する．

### 3.1 固有表現辞書の構築

場所を示す固有表現辞書の構築法について説明する．固有表現辞書の作成には，goo固有表現抽出API[7]を用いる．goo固有表現抽出APIは，日本語文からORG（組織名），PSN（人名），LOC（地名）などのクラスに属する固有表現を抽出できる．本稿では，goo固有表現抽出APIによってLOC（地名）クラスの固有表現と判定された単語を，地名に属する固有表現として用いる．そして，抽出した固有表現をGoogle Geocoding API[8]によって都道府県と結びつけ，場所を表す固有表現と都道府県が対となる辞書を構築する．

また，辞書補完を目的として，Web上から収集した駅名，市町村名，ショッピングモール名，公園名などを固有表現辞書に追加する．

### 3.2 ベースライン手法

まず，比較手法であるベースライン手法について述べる．ベースライン手法は，固有表現辞書と以下の式(1)を用いて，ユーザ $u$ が居住している都道府県$a^*$を推定する．

$$a^* = argmax\, p(a;u)$$

$$p(a;u) = \sum_{w \in d_a} p(a|w) * count(w,u) \qquad (1)$$

$$p(a|w) = \frac{count(w,a)}{count(w)}$$

ここで，$a$ は推定対象の都道府県，$d_a$ は都道府県 $a$ に該当する固有表現の集合， *count*(*w*)を固有表現 $w$ の出現回数, *count*(*w*, *a*)を都道府県 $a$ 在住のユーザが固有表現 $w$ をツイートした回数，*count*(*w*, *u*)をユーザ $u$ が固有表現 $w$ をツイートした回数である．ツイート文のようなくだけた文に対する形態素解析の精度は低いため，パターンマッチングにより，固有表現の出現回数をカウントする．

ベースライン手法は，我々の先行研究で提案した手法[2]と同様，固有表現が居住地を示す尤度を考慮しない手法である．提案手法は，この手法をベースに改良する．

### 3.3 提案手法

「香川の方に小旅行してみた」や「香川在住の学生ラ

イダーです!」のように，同じ“香川”という固有表現が出現するツイート文において，“香川”が居住地を示す尤度は異なる．そこで，場所を示す固有表現を含むツイートを抽出し，固有表現ごとにツイート文集合を構築する．そして，各集合をクラスタリングして，ツイートに出現する固有表現の居住を示す尤度を求めることで，上記の問題に対応する手法を検討する．

ツイート文中の名詞，動詞，形容詞の加重平均をツイート文の分散表現とし，X-means 法[9]を使用して，各集合をクラスタリングする．単語分散表現には，株式会社ホットリンクが公開している「日本語大規模 SNS+Web コーパスによる単語分散表現モデル[10]」を用いる．

提案手法では，式(2)を用いて，クラスタ内での尤度に基づいて居住地を推定する．

$$a^* = argmax\, p(a;u)$$
$$p(a;u) = \sum_{w \in d_a} \sum_{i \in m_w} p_i(a|w) * count_i(\mathrm{w},u) \qquad (2)$$
$$p_i(a|w) = \frac{count_i(w,a)}{count_i(w)}$$

ここで，$count_i(w)$をクラスタ $i$ 内での固有表現 $w$ の出現回数，$count_i(w,a)$をクラスタ $i$ 内で都道府県 $a$ のユーザが固有表現 $w$ をツイートした回数，$count_i(\mathrm{w},u)$をユーザ $u$ がクラスタ $i$ 内に含まれる固有表現 $w$ をツイートした回数とする．また，$m_w$は固有表現 $w$ を含むツイート文集合のクラスタ数である．

## 4. 評価実験

### 4.1 実験設定

評価実験に用いるデータセットは，松原らの研究[11]で構築されたものを利用する．2019 年 6 月 26 日～30 日に 23,538 ユーザから収集した 12,993,817 件のツイートから goo 固有表現 API によって得られた地名に属する固有表現 12,635 語のうち Google Geocoding API により各都道府県に写像できた 9,893 語および辞書拡張によって得た 13,786 語の合計 23,443 語を固有表現辞書として用いる．

データセットのうち，情報量の少ないユーザを省くためツイート数が 100 件未満または固有表現の出現回数が 10 回未満のユーザを排除し，データセットの 8 割を学習データ，2 割をテストデータに利用して評価する．評価指標には，47 都道府県の適合率，再現率，F 値の平均値を用いる．

提案手法の比較手法として，先行研究[2]で提案した手法（先行手法）と，3.1 節で示したベースライン手法（クラスタリングを用いない手法）を用いて，都道府県別で居住地推定した結果を比較する．

### 4.2 評価結果と考察

先行手法およびベースライン手法，提案手法の結果を表 1 に示す．先行手法と比較して，提案手法は，再現率と F 値に関しては大きな違いは見られなかったが，適合率は 3.0 ポイントと大幅に向上した．また，ベースライン手法と提案手法を比較した場合，提案手法が適合率，再現率，F 値の全てで上回っている．

以上の結果より，ツイート文集合のクラスタリングにより，固有表現のツイート内での位置づけが考慮されていると考えられる．

エラーの傾向としては，全ての手法において，都市圏のように都道府県をまたぐ移動が多い場所に対して，適合率が平均より約 8.5 ポイント低い結果となった．また，提案手法では，他の手法と比べて，北海道や沖縄に対する適合率が低かった．その原因としては，北海道や沖縄は観光地であり，居住していないユーザであっても言及しやすいことから，居住者とそれ以外のユーザにおける言及差が少ないことが影響していると考えられる．

表 1　都道府県別の居住地の推定結果

| | 適合率 | 再現率 | F 値 |
|---|---|---|---|
| 先行研究 | 0.721 | 0.709 | 0.712 |
| ベースライン手法 | 0.737 | 0.692 | 0.702 |
| 提案手法 | **0.751** | **0.710** | **0.718** |

## 5. おわりに

本稿では，Twitter で発信される病気症状の可視化に向け，ユーザの居住地を都道府県別に推定する手法について検討した．固有表現を含むツイートを抽出し，文単位の集合をクラスタリングすることで，推定性能の向上を実現した．今後は，言及されやすい地名に対して，固有表現の出現回数に依存しない手法を検討する．また，提案手法では，同じユーザが複数のツイートにまたがって発信する内容を考慮できていない．そのため，複数のツイート間の関係を考慮した手法を検討する．

## 参考文献

[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?”，人工知能学会誌，Vol.27, No.1, pp.67–74, 2012.
[2] 松本他, “Twitter で発信される病気症状の可視化に向けた Tweet からのユーザの居住地推定の検討”, 第 19 回情報科学技術フォーラム公演論文集, 2020
[3] 森國他，“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, pp16–26, 2015
[4] 坂本他，“類語の出現頻度に着目した居住地の推定に関する調査研究”, 第 34 回ファジィシステムシンポジウム 講演論文集, pp857-858, 2018
[5] 廣中他，“日本における居住地推定に利用するためのフォロー関係の調査”，人工知能学会論文誌，Vol.32, No.1，pp.WII-M_1-11，2017.
[6] G. Ke, et al., “LightGBM: a highly efficient gradient boosting decision tree”, in Proc. of the 31st International Conference on Neural Information Processing Systems, pp.3149-3157, 2017.
[7] goo ラボ｜API｜固有表現抽出 API, https://labs.goo.ne.jp/api/jp/named-entity-extraction/
[8] Google Geocoding API, https://developers.google.com/maps/documentation/geocoding/
[9] D. Pelleg, et al., “X-means: Extending K-means with Efficient Estimation of the Number of Clusters”, Proc. of the 17th International Conference on Machine Learning, pp.727-734, 2000.
[10] 松野他，“日本語大規模 SNS+Web コーパスによる単語分散表現のモデル構築”，人工知能学会全国大会論文集 2019，pp.1-3，2019.
[11] 松原他，“Twitter で発信される病気症状の可視化に向けた Tweet 内容を用いたユーザの居住地推定”，情報処理学会 第 82 回全国大会講演論文集，pp.393-394, 2020.