D-007

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
Twitter User's Residence Estimation Using Tweet Content for Visualization of Disease Symptom in Tweets

松本 真拓† 松原 香太† 安藤 一秋†
Masahiro Matsumoto Kyota Matsubafra Kazuaki Ando

1. はじめに

病気の流行を迅速に察知することは，流行拡大への対処という観点から重要な課題となっている．近年，Twitter から Tweet を収集・分析することで，インフルエンザや新型コロナウイルスなどの感染症の流行具合を推定し，地域別に可視化する研究が注目されている[1]．本研究では，Twitter 上の Tweet から様々な病気症状を抽出し，時系列・地域別等で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，地理的遷移の把握や原因分析が可能になる．

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する．先行研究として松原らは，Tweet から抽出した地名に属する固有表現を素性に利用することで，都道府県別に居住地を推定する手法[1]を提案した．しかし，出身地や活動地など，居住地とは異なる都道府県に特有の固有表現への対応や，利用する固有表現数の増加に伴うベクトルの次元数の増加などの問題があった．そこで本稿では，地名に属する固有表現を都道府県へ写像することでベクトルを圧縮する手法と，地名に属する固有表現の異なりを考慮するための重み付けについて提案する．また，ユーザの移動に関する素性を追加することで，推定性能に与える影響を確認する．

2. 関連研究

廣中らは，フォロー関係と位置情報付き Tweet を用いてソーシャルグラフを作成し，居住地を推定する手法[2]を提案し，市区町村別の推定で 29.2%，都道府県別の推定で 53.8%の F 値を得ている．

松本らは，市区町村名やスポット名などの地名を含む Tweet に対して，居住を示す Tweet かどうかのラベル付けしたものを Support Vector Machine（SVM）で学習し，事実性解析を行う手法[3]を提案し，78.0%の正解率を得ている．前後 Tweet との関係から居住を示す Tweet かどうかを人手で判断するため，ラベル付与のコストが高くなり，教師データの拡充は難しい．

松原らは，プロフィールに「都道府県市区町村名」＋“在住”の記載があるユーザの Tweet を収集し，記載された都道府県を正解ラベルとしてデータセットを構築した．Tweet から抽出した地名に属する固有表現を素性に利用し，XGBoost（eXtreme Gradient Boosting）で学習する手法[1]を提案し，構築したデータセットに対して都道府県レベルの推定で 70.6%の F 値を得ている．しかし，都市圏ユーザに対する推定では，活動地域と居住地域が異なる場合が多く，推定性能が低い結果となっている．また，出身地や活動地など，居住地に関係しない都道府県特有の固有表現を考慮できていない点や，利用する固有表現数の増加に伴うベクトルの次元数の増加などの問題があった．

† 香川大学 Kagawa University

3. 提案手法

本稿では，松原らの手法[1]と同様に，都道府県レベルの居住地推定を多クラス分類問題と捉え，Tweet 文中の地名に属する固有表現を基に，機械学習を用いて，Twitter ユーザの居住地を推定する．本稿では，松原らの手法におけるベクトルの次元数問題の改善と，居住地と居住地以外の固有表現の違いを考慮するため，地名に属する固有表現を都道府県へ写像することでベクトルを圧縮する手法と，地名に属する固有表現の重み付けについて提案する．また，ユーザの移動に関する素性についても提案する．

3.1. 固有表現の抽出

固有表現の抽出には，goo 固有表現抽出 API[4]を用いる．goo 固有表現抽出 API は，日本語文から，ART（人工物名），ORG（組織名），PSN（人名），LOC（地名），DAT（日付表現），TIM（時刻表現），MNY（金額表現），PCT（割合表現）のいずれかのクラスに属する固有表現を抽出する．本稿では，goo 固有表現抽出 API によって LOC クラスに属すると判定された語を，地名に属する固有表現として用いる．

3.2. 地名に属する固有表現の都道府県への写像

抽出した地名に属する固有表現をすべて素性に利用する場合，抽出された語数により，ベクトルの次元数が激増する問題がある．そこで，抽出した地名に属する固有表現を各都道府県に写像することで，次元数を圧縮する方法について提案する．

抽出した固有表現の各都道府県への写像には，Google Geocoding API[5]を用いる．Google Geocoding API では，地名に属する固有表現を入力することで，固有表現が示す場所の住所を得ることができる．獲得した住所に含まれる都道府県名により，固有表現を 47 都道府県に対応付けることでベクトルを 47 次元に圧縮し，これをベース素性に利用する．また，住所に都道府県名が含まれない固有表現は，海外の地名を示すものや地名を示さないものであると考えられるため，除去する．

3.3. 地名の重み付け

地名に属する固有表現には，居住地を示す Tweet に含まれやすい表現と居住地以外を示す Tweet に含まれやすい表現が存在するといえる．例えば，有名な観光地については居住者以外が言及する可能性が高く，日常生活で使われる地名は居住者が言及する可能性が高いと考えられる．そこで，地名に属する固有表現に対する重み付けを検討する．安田らの手法[6]を参考に，地名 t の重み $w(t)$を以下のように定める．

$$w(t) = \log(\frac{P(t)+1}{N(t)+1} + 1) \quad (1)$$

ここで，$P(t)$は地名を示す固有表現 t をその固有表現が示す都道府県の居住者が Tweet した頻度，$N(t)$は非居住者が Tweet した頻度を示す．

本稿では，各ユーザの Tweet 文中の地名に属する固有表現の出現頻度とその固有表現の重みの積の合計を各都道府県の素性値とし，47 次元のベクトルを作成する．

3.4. ユーザの移動に関する素性

Tweet 文中には，活動地や出身地など，居住地と関連のない固有表現も出現すると考えられる．そこで，Tweet 文中に表出するユーザの移動に関する情報を素性に利用することで，推定性能の向上を目指す．

本稿では，ユーザの移動に関する情報として，地名に属する固有表現と共起する「移動を示す語」に着目する．移動を示す語として，Weblio 類語辞典[7]を参考に人手で表 1 に示す 10 クラスの語群を選定した．地名に属する固有表現を含む Tweet 文を Mecab で形態素に分割し，10 クラスの各語と共起する各都道府県の固有表現を抽出し，470 次元の素性をベクトルに追加する．

表 1　移動素性として選択した語群

行く・向かう	住む・在住（する）
来る	乗る
戻る・帰る	帰郷（する）・帰省（する） 地元（入り）・里帰り（する）
着く・到着（する）	移る・引っ越す
旅行（する）・観光（する）	居る・滞在（する）

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[1]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 Tweet 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 語を素性とする．BaseLine 手法には，抽出した 9,893 語の固有表現の出現回数を素性値とした 9,893 次元のベクトルによる都道府県別の居住地推定法を用いる．提案手法には，地名に属する固有表現を 47 都道県に写像した手法（手法 1）と，手法 1 にユーザの移動素性を追加した手法（手法 2）を用いて，それぞれの手法の推定性能を比較する．

分類器には，LightGBM[8]を用いる．プロフィールに記載されている都道府県レベルの居住地と推定居住地を比較し，完全一致した場合を正解とする．5 分割交差検証により評価し，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．

4.2. 評価結果と考察

評価結果を表 2 に示す．BaseLine 手法と手法 1 を比較した場合，BaseLine 手法の推定性能の方が高い結果となった．これは，複数の都道府県に特有である固有表現を Google Geocoding API により 1 つの都道府県に写像したことが影響していると考えられる．例えば，都道府県を跨いで展開しているスーパーマーケットや複数の都道府県に存在する市区町村名が 1 つの都道府県に集約されることで情報量が低下し，推定性能が低下したと考えられる．そこで，地名に属する固有表現を複数の都道府県に写像する手法を改良する必要があると考えられる．

次に，BaseLine 手法と手法 2 を比較した場合，推定性能に大きな違いは見られなかった．しかし，ベクトルの次元数を比較した場合，BaseLine 手法は 9,893 次元であるのに対し，手法 2 では 517 次元と大幅に削減できていることから，手法 2 に優位性はあるといえる．

最後に，手法 1 と手法 2 を比較した場合，手法 2 の推定性能が高い．Tweet 文中の地名に属する固有表現のみでなく，移動に関する周辺語も考慮することで，推定性能が向上することが確認できた．本稿では，移動を表す語のみに着目したが，ユーザの日常的な行動や過去の行動を表す素性を追加することで，居住地とは異なる地名に属する固有表現による誤分類を防ぐことができると考えられる．

表 2　居住地の推定結果

	適合率	再現率	F 値
BaseLine	0.719	0.708	0.711
手法 1	0.707	0.696	0.700
手法 2	0.720	0.707	0.712

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．都道府県別の居住地推定を多クラス分類問題と捉え，地名に属する固有表現と，固有表現に共起する移動を示す語を考慮することで，推定性能が向上することを確認した．

今後は，地名に属する固有表現を複数の都道府県に写像する手法の改良と，ユーザの日常的な行動や過去の行動を表す素性などについて検討し，推定性能の向上を目指す．また，病気・症状を含む Tweet を発信し，居住地がプロフィールに明記されていないユーザを対象に，提案手法の有効性を評価する．

参考文献

[1] 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，2020.
[2] 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.
[3] 松本他，“Twitter を用いた感染症発生動向の可視化”，人工知能学会情報アクセスと可視化マイニング研究会（第 15 回），SIG-AM-15-08，pp.48-53，2017.
[4] goo ラボ｜API｜固有表現抽出 API，https://labs.goo.ne.jp/api/jp/named-entity-extraction/
[5] Google Geocoding API，https://developers.google.com/maps/documentation/geocooding/start
[6] 安田他，“ブログ作者の居住地推定”，言語処理学会 第 12 回年次大会 発表論文集，2006.
[7] Weblio 類語辞典，https://thesaurus.weblio.jp/
[8] G. Ke, et al., “LightGBM: a highly efficient gradient boosting decision tree”, in Proc. of NIPS’17, pp.3149-3157, 2017.