5N-07

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

松原 香太† 安藤 一秋‡
香川大学工学部† 香川大学創造工学部‡

1. はじめに

病気の流行を迅速に察知することは，病気の予防や流行への対処などの観点から重要である．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を提案する．提案手法では，Tweet中の出現単語や固有表現などを素性に利用して居住地を推定する．プロフィールに居住情報を記載しているユーザと記載していないユーザのTweetから構築したデータセットを用いて，提案手法の有効性を検証する．

2. 関連研究

廣中らは，位置情報付き Tweet とフォロー関係を用いて居住地を推定する手法[1]を提案し，市区町村レベルの推定で 29%，道府県レベルの推定で 54%の F 値を得ている．しかし，位置情報を含んだ Tweet は全体の 1%に満たないため，Tweetに位置情報を付与しない大多数のユーザを対象にできない問題がある．

松本らは，県名，市名，駅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かどうかのラベルを付与したものを Support Vector Machine（SVM）で学習し，都道府県レベルで居住地を推定する手法[2]を提案し，78%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必要があるため，ラベリングコストが高くなり，教師データの拡充は難しい．

3. データセットの構築

実験に利用するデータセットの構築法について説明する．ユーザのプロフィール情報を検索できるサービス Twpro API を用いて，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を収集する．その後，TwitterAPI を用いて，1 ユーザにつき，Reply と Retweet を除く，最大 3,200 件の Tweet を収集する．Bot Tweet を除外するため，クライアント名が “Twitter for iPhone”, “Twitter for Android”, “Twitter Web Client”, “Twitter for iPad” である Tweet のみを対象とする．また，極端に情報量が少ないユーザを除外するため，100Tweet 未満のユーザ，全Tweetが100語彙未満で構成されているユーザをフィルタリングする．

本稿では，2019 年 6 月 26 日～30 日に 23,538 ユーザから 12,993,817 件の Tweet を収集し，データセットとして利用する．

4. 提案手法

Twitter ユーザは，意識，無意識に関わらず，特定地域の情報を発信していると考えられる．そこで，Tweetに含まれる地域特有の情報を持つ単語を素性に利用することで，都道府県レベルで居住地を推定する手法を提案する．

都道府県レベルの居住地推定は，多クラス分類問題と捉えられる．そこで本稿では，XGBoost （eXtreme Gradient Boosting）を用いて居住地を推定する．XGBoost の入力は，各ユーザの Tweet 集合であり，パラメータ tree_method は gpu_hist，objective は multi:softprob とする．

Tweet から素性を抽出するため，形態素解析（MeCab）と goo 固有表現抽出 API [5]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，次元数の影響も確認するために，単語数に応じた 4 つの素性について検討する．なお，素性値には，出現回数を利用する．以下，4つの素性について説明する．

4.1. Noun 素性

Tweet内の名詞以外の品詞は，地域特有の情報を持つ単語になりにくいと考え，名詞のみを素性に利用する．また，各都道府県において，ある名詞を含むTweetが5人未満の場合，利用されにくい名詞と考え，素性から除外する．また，数字，アルファベット，記号，ストップワードも除外し，最終的に 38,592 語を素性として利用する．

4.2. Entity 素性

地域特有の情報を持つ単語になる可能性がある ART, ORG，LOC クラスの 23,374 語を素性として利用する．

4.3. LOC 素性

「居住情報を明記しているユーザは地名をつぶやきやすい」という事前調査を基に，LOC クラスの 12,635 語を素性として利用する．

4.4. LOC'素性

LOC 素性には，海外の国名，都市名も含まれている．目的とする居住地推定の粒度は，都道府県レベルであるため，LOC 素性からカタカナのみで構成される単語をフィルタリングした，11,270 語を素性として利用する．

5. 評価実験

5.1. 実験設定

提案手法の有効性を評価するため，データセットを用いて，パターンマッチングによる手法（BaseLine 手法）と，提案手法の 4 素性別の推定性能を比較する．BaseLine 手法では，まず，各ユーザの Tweet から都道府県市区町

A Method for Estimating Twitter User's Residence Using Tweet Content for Visualization of Disease Symptoms in Tweets
† Kyota Matsubara, Faculty of Engineering, Kagawa University
‡ Kazuaki Ando, Faculty of Engineering and Design, Kagawa University

村名を抽出し，都道府県別に集計する．そして，集計値が最大となる都道府県をそのユーザの居住地として推定する．なお，集計値が同値になった場合は，より人口の多い都道府県を居住地として推定する．都道府県市区町村名には，総務省の全国地方公共団体コード[3]に記載されている1,788件の内，ノイズになる可能性が高い，ひらがなのみの語，一文字の語を排除した1,682件を利用する．人口は，総務省統計局の都道府県別人口と人口増加率[4]に記載されている平成27年度版を利用する．

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．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，提案手法は5分割交差検証で評価する．

5.2. 評価結果と考察

評価結果を表1に示す．BaseLine手法とその他を比較した場合，BaseLine手法の再現率とF値が大きく減少している．これは，居住地以外の都道府県について言及したTweetの割合が高いことを示している．つまり，都道府県について言及しているTweetの割合が小さいユーザが存在していると考えられる．

Entity素性を用いた場合，推定性能が最高の結果を得た．固有表現を用いた他の素性と比較した場合，最も単語数が多く，推定に必要な情報が担保されていることが最高性能を得た理由として考えられる．Entity素性とNoun素性を比較すると，すべての評価指標において，Entity素性を用いた結果が上回っている．次元数（単語数）もNoun素性よりEntity素性の方が小さいことから，Entity素性の方が居住地推定に適しているといえる．Entity素性とLOC素性を比較した場合，推定性能に大きな差はない．人手で居住地を推定する際は，都道府県市区町村名が大きな手掛かりとなるが，機械学習による推定においても，それらが地域特有の情報を持つ単語になることがわかる．LOC素性とLOC'素性では，すべての評価指標でLOC'素性の結果が上回っている．海外の国名，都市名については，地域による出現数の差はなく，地域特有の情報を持つ単語にならないため，居住地推定には不必要な情報であるといえる．また次元数を減らすことで，計算時間の短縮だけでなく，推定性能も向上することを確認した．

表1　居住地の推定結果

	適合率	再現率	F値
BaseLine	0.630	0.341	0.416
XGB(Noun)	0.709	0.700	0.702
XGB(Entity)	**0.725**	**0.719**	**0.720**
XGB(LOC)	0.715	0.703	0.706
XGB(LOC')	0.720	0.712	0.714

6. 居住地不明記ユーザに対する評価

プロフィールに居住情報を記載したユーザと記載していないユーザにおいて，Tweet本文における地域特有の情報を持つ語の扱い方が異なる可能性がある．そこで，居住情報をプロフィールに記載していないユーザに対して，提案手法により居住地推定を試みる．

病気症状の可視化に向け，gooヘルスケア家庭の医学[6]に記載されている病気症状のうち，年齢や性別に関係しない14症状を含む内容をTweetし，かつ居住情報を記載していないユーザに対して提案手法による推定性能を確認する．対象となるユーザの内，Tweet内容から居住地を明確に判断できる82ユーザを人手で抽出し，正解ラベルを付与し，これらをテストデータに利用する．データセットを教師データに利用して構築した分類器を用いて，テストデータに対する居住地を推定する．正解ラベルと推定居住地を比較し，完全一致した場合を正解とする．

評価結果を表2に示す．表1ではEntity素性を用いた場合に最高の推定性能を得たが，表2ではEntity素性よりNoun素性の方がすべての評価指標において結果が上回っている．この結果から，Noun素性にはEntity素性（ART, ORG，LOCクラス）以外にも地域特有の情報となる素性が埋もれているといえる．今後は，ART，ORG，LOCクラス以外の素性について検討する必要がある．

固有表現を用いた素性であるEntity素性とLOC素性，LOC'素性を比較した場合，LOC'素性で推定性能が最高の結果を得た．つまり，LOC'素性は固有表現の中で地域特有の情報を持つ単語になり得ることが確認できた．一方で「居住情報を明記しているユーザは地名を呟きやすい」という調査結果の下，LOC系の素性を検討したが，テストデータを人手で抽出する際，地名を含むTweetに着目し過ぎた可能性もある．今後は，無作為に抽出したユーザに対して評価を行う必要がある．

表2　居住地不明記ユーザに対する居住地の推定結果

	適合率	再現率	F値
XGB(Noun)	**0.629**	**0.684**	**0.639**
XGB(Entity)	0.609	0.651	0.600
XGB(LOC)	0.617	0.651	0.605
XGB(LOC')	0.623	0.653	0.612

7. 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を提案した．Entity素性を利用した手法により，居住地を明記するユーザに対して72.5%，Noun素性を利用した手法により，不明記のユーザに対して62.9%の平均適合率を得た．今後は，ART，ORG，LOCクラス以外の素性や単語の重み付けなどについて検討する．また，無作為に抽出したユーザに対して大規模な評価を行う．

参考文献

[1] 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.
[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.1，2017.
[3] "総務省｜電子自治体｜全国地方公共団体コード"，http://www.soumu.go.jp/denshijiti/code.html
[4] "総務省統計局｜人口・世帯｜都道府県別人口と人口増加率"，https://www.stat.go.jp/data/nihon/02.html
[5] "gooラボ｜API｜固有表現抽出API"，http://www.soumu.go.jp/denshijiti/code.html
[6] "gooヘルスケア｜家庭の医学"，http://health.goo.ne.jp/