4P-06

ツイートの内容と天気情報を用いたTwitterユーザの居住地推定

松本真拓† 安藤一秋‡

†香川大学大学院 ‡香川大学

1. はじめに

近年，流行している新型コロナウイルスは，発熱や咳などの一般的な症状のみでなく，味覚障害や嗅覚障害など，様々な症状が報告されている．したがって，様々な病気・症状の発生状況の地域性や時系列性を特定することで，未知の病気の発生・流行を察知することができると考えられる．本研究では，ソーシャルメディアの1つであるTwitterを用いて，感染症の流行状況だけでなく，腹痛や頭痛といった各種症状の発症状況を調査し，都道府県別や時系列別などで可視化するシステム[1, 2]の構築を目的とする．

Twitterから病気の流行状況と病気症状の発症状況の地域性を察知するためには，病気・症状情報を発信しているユーザの位置情報を特定する必要がある．Twitterで顕在的に得られる位置情報としては，ユーザの発信するツイートに付与されている位置情報やユーザが自由に記述できるプロフィールの Location 項目の情報などがある．しかし，ツイートに位置情報を付与して発信しているユーザやプロフィールの Location 項目に自身の正確な居住地域を示しているユーザは少ない[3]．そのため，顕在的な位置情報以外を用いてユーザの居住地を推定する必要がある．また，既存のツイート内の単語分布を用いてユーザの居住地を推定する手法においては，地域的な単語を含まないユーザに対して，ユーザ数の多い大都市圏へと居住地を誤推定するという問題がある[4]．そのため，ユーザ数の少ない地方のユーザに対しては，ツイート内の単語分布以外の情報を追加して予測する必要がある．

そこで本稿では，ツイート内に地域特徴語を含まない地方ユーザに対して居住地を正確に推定することを目的に，我々の先行研究[5]で提案した天気情報を用いたモデルによる地方ユーザの居住地予測に対する有効性について考察する．

2. 関連研究

インフルエンザのみを対象とし，その流行状況をTwitterから抽出して可視化する既存システムとして「インフルくん[6]」がある．「インフルくん」では，ツイートに含まれるインフルエンザに関する情報を収集し，都道府県別にマッピングしている．しかし，ツイート発信者の居住地の特定に位置情報付きツイートやユーザプロフィールの Location 項目を利用しているため，およそ75%のツイートに対しては位置情報を付与できておらず地域不明となっている．したがって，ユーザ数の少ない地方のユーザに対しては，位置情報付きツイートやプロフィールの Location 項目を用いずにユーザの居住地を推定する手法が特に必要となる．

我々の先行研究[5]では，英語圏ユーザの居住地推定において高性能を得ているツイート内容を用いた既存の深層学習モデルに対して，天気情報を含むツイートとツイート発信時間の天気情報を照らし合わせるモデルを追加することで，推定性能が向上することを報告している．

本稿では，我々の先行研究において提案した深層学習モデルを用いて，天気情報を追加することでツイートのみを用いる場合と比較して，地方別でどのように推定性能が向上するのかについて考察する．

3. 提案手法

我々の先行研究[5]で提案したユーザのツイート内容とアメダスの観測データを用いて居住地を推定する深層学習モデルについて述べる．提案モデルのアーキテクチャを図1に示す．

提案モデルは，ツイート内容モデルと天気情報モデルで構成される．ツイート内容モデルでは，ツイート内容からユーザの居住地を都道府県別に推定するのに対し，天気情報モデルでは，天気情報を含むツイートおよびツイート発信時間の天気情報を参照することによりユーザの居住地を地方別に推定する．天気情報モデルによる地方別予測結果をツイート内容モデルによる都道府県別予測に反映することで，居住地推定性能の向上させている．

4. 実験設定

本稿では，ツイート内容モデルのみを用いて推定する手法と，我々の先行研究で提案したツイート内容モデルに天気情報モデルを組み込んだ手法の性能を比較することで，天気情報を深層学習モデルに取り入れることによる地方ユーザの居住地推定に対する有効性を検証する．実験には，プロフィールに自身の居住地を明記している 204,965 ユーザを収集し，184,965 ユーザを学習データとし，検証データとテストデータには，それぞれ10,000ユーザを利用する．

5. 実験結果

天気情報モデルによる地方ユーザに対する居住地推定への影響を見るために，都道府県別予測結果が正解ラベルと同地方への予測となっている場合に

Prediction of Twitter User's Location Using Tweets and Weather Information

† Matsumoto Masahiro・Kagawa University

‡ Ando Kazuaki・Kagawa University

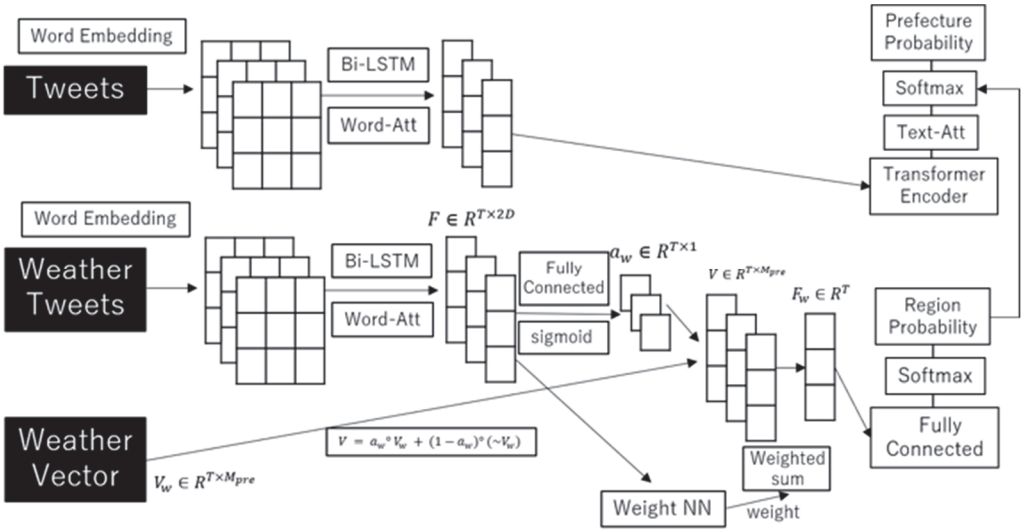

図 1　提案モデルのアーキテクチャ

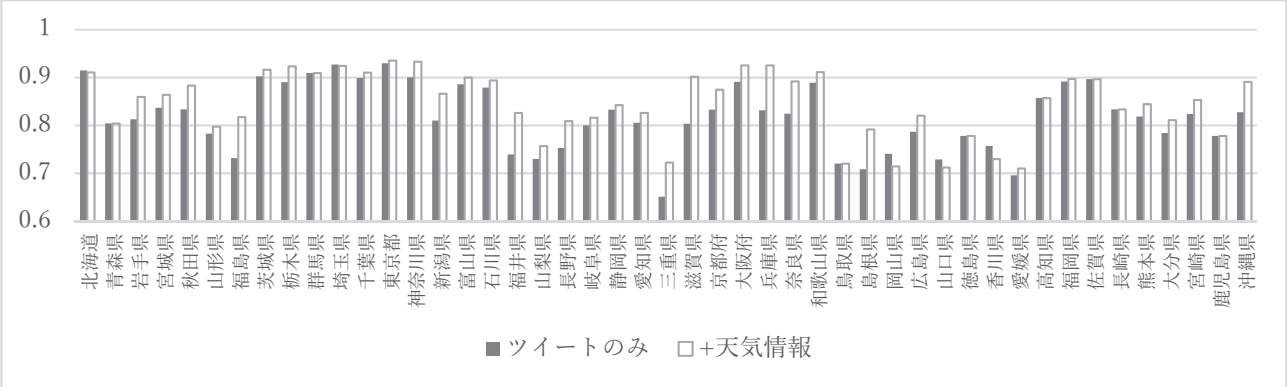

図 2　都道府県別の再現率

正解としたときの都道府県別の再現率を図 2 に示す．

図 2 より，天気情報を用いることで，首都圏においては再現率の向上はあまり見られないが，東北地方や中部地方，関西地方のユーザに対しては，再現率が大きく向上していることがわかる．単語分布のみで予測した場合，地域特徴語を含まないユーザについては居住地を首都圏に予測してしまう傾向があるが，地域特徴語を含まないユーザであっても，天気情報を利用することで，正しい地方へ居住地を推定できることを確認した．しかし，中国地方や四国地方のユーザに対しては，再現率があまり向上していない．原因として，中国地方や四国地方では降雨が少なく，特徴的な降水が他地方と比較して少ない可能性が考えられる．

6. まとめ

本稿では，我々の先行研究で提案した天気情報を用いたモデルによる地方ユーザの居住地予測に対する有効性について考察した．単語分布のみでは，ツイート内に地域特徴語を含まないユーザの居住地を首都圏に予測する問題があったが，天気情報を追加することで地方ユーザに対する再現率が向上することを確認した．しかし，中国・四国地方のユーザに対しては，特徴的な降水が少ないため，天気情報では予測が難しいことも確認した．今後の課題として，天気情報のみでなく地震等の発生状況などを予測に用いることについて検討する．

参考文献

[1] 安藤他，“ツイートされる病気・症状の可視化に向けた症状の事実性解析”，言語処理学会第 27 回年次大会発表論文集，3 pages，2021.

[2] 松本他，“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”，IPSJ 第 83 回全国大会講演論文集，1-393-1-394，2021.

[3] 橋本他，“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，Vol27，No4，pp.424-431，2012.

[4] 近藤他，“アメダスの観測データを用いた Twitter ユーザの居住地推定の試み”，第 10 回 Web インテリジェンスとインタラクション研究会オンライン・プロシーディングス，pp.31-36，2017.

[5] 松本他，“Twitter で発信される病気・症状の可視化に向けたツイート内容と天気情報に基づく Twitter ユーザの居住地推定”，ARG Web インテリジェンスとインタラクション第 17 回研究会予稿集，pp33-38，2021.

[6] インフルくん，http://mednlp.jp/influ_map/