

ツイートの内容と天気・地震情報を用いた 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

新型コロナウイルスに感染すると発熱や咳、味覚障害などの様々な症状が現れることが報告されている。様々な病気・症状の動向を時系列・地域別に捉えることで、既存の病気の流行や未知の病気の発生を察知することができる。本研究では Twitter においてリアルタイムに発信される病気・症状に関するツイートの動向を観察することで、様々な病気・症状の発生・流行状況を時系列・都道府県別に可視化するシステム[1,2]の構築を目的とする。

病気・症状に関するツイートを地域別に察知するためには、ツイート発信者の位置情報を特定する必要がある。しかし、ツイートを位置情報を付与して投稿するユーザやプロフィールの location 項目に居住地を明記しているユーザは非常に少ない[3]。そのため、位置情報付きツイートやユーザプロフィールの location 項目以外の情報を用いてツイート発信者の位置情報を推定する必要がある。

ツイート発信者の居住地を推定する手法は数多く提案されている。しかし、ユーザのツイートに含まれる単語分布を用いて居住地を推定する既存手法においては、地域的な特徴を表す単語（地域特徴語）を含まないユーザの居住地を人口の多い大都市圏へと誤推定する問題がある[4]。ツイートを地域特徴語を含まないツイートしか発信しないユーザであっても、自身の居住地に関する現在の天気情報といった自然現象をツイートする場合がある。

我々の先行研究[5]において、地域特徴語を含まないユーザに対する居住地推定性能の向上を目的に、天気に関するツイートとツイート発信時間の天気情報をもとに Twitter ユーザの居住地を推定する手法を提案した。しかし、天気情報のみでは、特徴的な降水が少ない中国地方や四国地方のユーザの居住地を推定することが難しい問題があった。

本稿では、地震発生を知らせるツイートの存在に注目し、我々の先行研究[5]で提案した深層学習モデルに、地震情報を入力することで、Twitter ユーザの居住地推定に対する有効性を確認する。

2. 関連研究

Twitter ユーザの居住地を推定する手法については、これまでに様々な手法が提案されているが、近年では、英語圏のユーザに対して深層学習を用いたモデルが注目されている。深層学習を用いた Twitter ユーザの居住地推定手法としては、Miura ら[6]や Huang ら[7]のモデルが挙げられる。これらの深層学習モデルは、ユーザのツイート内容やユーザ同士の関係性など Twitter 内部の情報のみを入力に利用している。そのため、地域特徴語をツイートしないユーザや幅広い地域のユーザと関わりのあるユー

ザに対しては、居住地の推定が難しい課題がある。

我々の先行研究[5]では、英語圏のユーザに対する居住地推定において高性能な結果を得ているツイート内容を用いたモデルに対して、ユーザの発信する天気に関する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て活用するモデルを追加することで、居住地の推定性能が向上することを確認した。

本稿では、天気情報以外の情報として地震情報を追加することで、天気情報のみでは推定が難しかったユーザの居住地推定に与える影響を確認する。

3. 提案手法

本稿では、我々の先行研究で提案したモデルに地震情報を追加する手法を提案する。近年の日本において、様々な規模の地震が発生している。そして、地震発生後には、Twitter のトレンドワードになるなど、ユーザの反応数が多い。

図 1 に提案モデルを示す。提案モデルは、ツイート内容モデルと外部情報モデルで構成される。ツイート内容モデルはユーザのツイートを入力とする Huang らのモデル[7]を簡略化したモデルである。外部情報モデルは、外部情報ツイートと外部情報ベクトルを入力とする。外部情報ツイートとは、天気情報を対象とした場合、「雨が降っている」などの対象とする事象に関するツイートのことである。外部情報ベクトルとは、外部情報ツイート発信時の対象とする事象の情報を含むベクトルである。天気情報を対象とした場合は、ツイート発信時に雨が降っている場合は 1、雨が降っていない場合は 0 を要素にもつ 47 都道府県ごとの 47 次元ベクトルである。最終的には、外部情報モデルが地方別に居住地を予測した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、ユーザの居住地を都道府県別に推定する。

地震情報を対象とする場合、外部情報ツイートは地震情報に関するツイートである。地震情報に関するツイートの取り扱い法を決めるため、「地震の発生」と「ツイート内の単語有無」の 2 つの事象の自己相互情報量 (PMI) の大きい単語上位 20 件を求め、そのいずれかの単語が含まれるツイートを外部情報ツイートに利用する方法を検討したが、地震と関係のないノイズツイートが多く含まれることを確認した。そこで本稿では、「地震」、「揺れ」のどちらかの単語を含むツイートのみを外部情報ツイートに利用する。外部情報ベクトルは、Web スクレイピングした地震情報を用いて、ツイート発信時の 6 時間前以内に震度 2 以上の地震が発生している都道府県は 1、地震が発生していない都道府県は 0 を要素にもつ 47 次元ベクトルである。

4. 実験設定

データセットは、先行研究[5]と同じものを利用し、病気・症状情報をツイートしているユーザのうち、プロフィールに自身の居住地を明記している 204,965 ユーザのう

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

[†] Matsumoto Masahiro · Kagawa University

[‡] Ando Kazuaki · Kagawa University

ち、学習データを 184,965 ユーザ、検証データ、テストデータを、それぞれ 10,000 ユーザとする。ユーザのプロフィールに記載されている居住地を正解ラベルとし、推定居住地と比較することで推定性能を評価する。

5. 実験結果

47 都道府県別予測結果の Precision, Recall, F 値を表 1 に示す。表 1 より、ツイート内容のみを用いる場合と比べ、ツイート内容に加えて天気情報および地震情報を追加することで都道府県別予測結果の F 値が約 2.3 ポイント向上していることがわかる。しかし、ツイート内容と天気情報のみを用いた場合とツイート内容と天気及び地震情報を用いた場合を比較した場合、F 値の差は約 0.3 ポイントしか向上していない。

提案モデルでは、外部情報モデルがユーザの居住地を地方別に予測し、その結果をツイート内容モデルにおいて都道府県別に予測した結果に反映する。そこで、予測結果が正解ラベルと同地方になっていた場合も正解とした場合の推定性能を表 2 に示す。表 2 より、地方別に予測する結果においては、ツイート内容モデルに天気・地震情報の両方を追加した場合、F 値が約 3.2 ポイント高くなることからわかる。次に、地方別予測結果の再現率を図 2 に示す。図 2 から、天気情報のみでは特徴的な降雨が少なく再現率の向上があまり見られなかった中国地方などにおいて地震情報を追加することで推定性能が向上している。最後に、表 3 に外部情報モデルのみによる地方別予測の結果のマクロ平均を示す。表 3 より、外部情報モデルによる地方別予測の結果は、Precision, Recall, F 値の全てにおいて天気情報のみを用いるよりも、天気情報と地震情報を併用した場合の推定性能が高い。

6. 考察

ツイート内容に天気情報のみでなく地震情報を追加することで、地方別予測の推定性能が向上することがわかった。また、天気情報のみでは特徴的な降雨が少なく推定性能が向上しなかった中国地方などのユーザに対して、地震情報を追加することで推定性能を向上できることを確認した。しかし、天気情報のみでなく地震情報も用いることで外部情報モデルによる地方別予測結果は向上するものの、提案モデル全体で見た場合、外部情報として天気情報のみを用いた場合と天気・地震情報を用いた場合の都道府県別の推定性能の差は微増であった。また、同エポック数で学習した場合、天気・地震情報を用いた提案モデルの都道府県別予測結果が、天気情報のみを用いた場合よりも低いことも確認した。

提案モデルは、外部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した後、都道府県別予測結果の誤差を最小化して学習しているため、外部情報モデルによる地方別予測結果の向上により、ツイート内容モデルが未学習になるなど、ツイート内容モデルの都道府県別学習に悪影響を及ぼしている可能性もある。よって、今後は、外部情報モデルの予測結果を活用する方法について検討を継続する。

7. おわりに

本稿では、ツイートコンテンツに天気情報と地震情報を併用する深層学習モデルを提案し、ユーザの居住地推定に対する有効性を確認した。今後は、都道府県別の予測結果を向上するため、外部情報モデルの予測結果を活用する方法について検討を継続する。

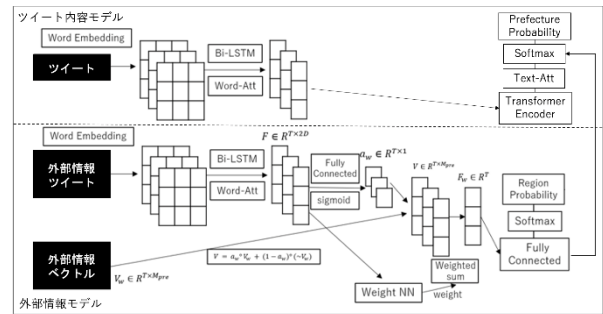


図1 提案モデル

表1 都道府県別予測結果

	Precision	Recall	F 値
ツイート内容のみ	0.7245	0.6929	0.7046
+天気情報	0.7730	0.6899	0.7244
+天気・地震情報	0.7835	0.6873	0.7277

表2 同地方予測も正解とした場合の予測結果

	Precision	Recall	F 値
ツイート内容のみ	0.8369	0.8245	0.8305
+天気情報	0.8619	0.8389	0.8489
+天気・地震情報	0.8817	0.8462	0.8629

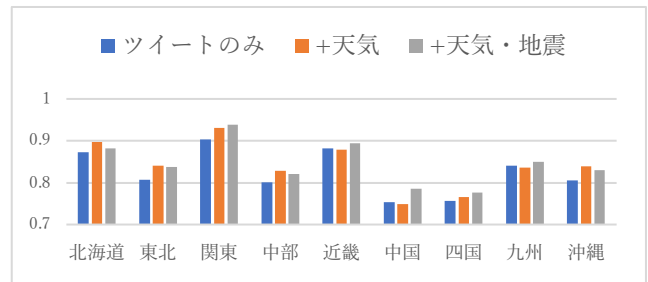


図2 地方別予測結果の再現率

表3 外部情報モデルによる地方別予測結果

	Precision	Recall	F 値
天気情報のみ	0.5483	0.3877	0.4354
天気・地震情報	0.6646	0.4850	0.5412

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”, NLP 2021, 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 2021, 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, Vol.27, No.4, 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 W12 予稿集, 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”, 第 17 回 W12 予稿集, pp.33-38, 2021.
- [6] Y. Miura, et al., “Unifying text, metadata, and user network representations with a neural network for geolocation prediction,” in Proc. of ACL2017, pp. 1260-1272, 2017.
- [7] B. Huang, et al., “A hierarchical location prediction neural network for twitter user geolocation,” in Proc. of EMNLP-IJCNLP2019, pp.4732-4742, 2019.