

## BERT を利用した SNS ユーザの居住地推定 Estimation of SNS User's Place of Residence using BERT Language Model

平松 洪輝<sup>†</sup>  
Hiroki Hiramatsu

安藤 一秋<sup>‡</sup>  
Kazuaki Ando

### 1. はじめに

本研究では、SNS に日本語で発信された病気・症状を含む投稿を収集・分析することで、感染症の流行状況や各種症状の発症状況を地域別や時系列別に可視化するシステムの構築を目的とする。このシステムを利用することで、未知の感染症を含めた様々な病気・症状の動向を捉えることが可能になる。

病気・症状に関する投稿を地域別に検知するためには、投稿者の位置情報を特定する必要がある。しかし、自らの居住地をプロフィールに明記しているユーザは少ない[1]ため、発信者の居住地をプロフィール以外の情報を用いて推定する技術が必要になる。

我々の先行研究[2, 3]では、X (旧 Twitter) ユーザを対象に、ツイート発信者の居住地を推定する深層学習モデルを提案した。これまでの研究では、ツイート内容に加えて天気・地震情報を活用したモデル[2]や、ツイート内容に加えてフォロー情報を活用したモデル[3]を提案した。これらのモデルは、いずれもツイート内容からユーザの居住地を推定するモデル (ツイート内容モデル) を基盤としているが、ツイートテキストからの特徴抽出を単語分散表現によるベクトル化と Bi-LSTM 層を用いて実現していた。

本稿では、このツイートテキストからの特徴抽出部を BERT[4]に置き換え、文脈を考慮した、より高性能なテキスト埋め込み表現を得ることで、ツイート内容モデルの推定性能の向上を図る。

### 2. 先行研究

我々の先行研究では、X (旧 Twitter) ユーザの居住地を推定する複数の手法を提案してきた。

松本ら[2]は、ユーザのツイート内容に加えて、ツイートに含まれる天気・地震情報を用いてユーザの居住地を推定する深層学習モデルを提案した。実験の結果、47 都道府県別予測の F 値で 0.728 を達成した。天気・地震情報は特に Precision の改善に大きく寄与し、両者を同時に利用した場合に Precision の最良値を得ている。一方で、Recall はツイート内容のみを用いた場合に最良となり、天気・地震情報の追加は Recall の改善には寄与しなかった。松本らの手法で Recall が低くなる理由は、データセットにおけるユーザ数が都道府県ごとに偏っているためである。その結果、モデルが大都市に対して過学習してしまい、ユーザ数の少ない地方部の都道府県での予測が難しくなっている。また、ユーザ数が比較的多い都道府県でも、大都市に隣接する都道府県 (例: 神奈川県, 京都府) では、大都市へ予測が集中してしまい、Recall が低くなる傾向が見られた。

ツイート内容のみを用いる場合、大都市とその周辺地域の地理的および社会的境界が曖昧であるため、SNS 上の活動が混在し、モデルによる正確な居住地推定が難しくなる。そこで、平松ら[3]は、SNS のフォロー機能に注目し、ツイート内容とフォロー情報を組み合わせた居住地推定モデルを提案した。地理的に近いユーザ同士は SNS 上でフォロー関係にある可能性に基づき、これを居住地推定の手がかりとした。実験の結果、47 都道府県別予測の Recall で 0.761 を達成し、特にデータセット中のユーザ数が少ない地方部の都道府県において、ツイート内容のみを用いた場合と比較して、平均で約 8.75 ポイント向上することに成功した。

### 3. 提案手法

本研究では、ツイート内容からユーザの居住地を推定するモデルを「ツイート内容モデル」とよぶ。先行研究[2]で提案したツイート内容モデルは、Huang ら[5]によって提案されたモデルを簡略化したものである。具体的には、文字レベルの CNN 層を省略し、ツイート内容のみから特徴を抽出するという点で簡略化されている。従来のツイート内容モデルでは、ツイートテキストは単語分散表現によりベクトル化され、Bi-LSTM 層と注意機構を用いて特徴を抽出していた。本稿では、ツイートテキストからの特徴抽出を BERT に置き換えたモデルを提案する。これにより、文脈を考慮した、より高性能なテキスト埋め込み表現が得られることから、ツイート内容モデルの推定性能の向上が期待できる。

ツイートテキストの特徴抽出から居住地推定までの処理の詳細を説明する。まず、ユーザ  $u_i$  のツイート列  $T_i = \{t_1, t_2, \dots, t_n\}$  について、各ツイートを個別に BERT に入力し、[CLS]トークンの最後の隠れ状態をもとに、ツイート埋め込み表現  $H_i = \{h_{1,CLS}, h_{2,CLS}, \dots, h_{n,CLS}\}$  を得る。ツイートを個別に BERT に入力することで、BERT の系列長の制限に影響されることなく、利用可能な情報を最大限に活用できる。一方で、ツイートを個別に処理することで、ツイート間の関連性が無視されてしまう。そこで、ツイートの埋め込み表現  $H_i$  を、Transformer Encoder 層に入力し、ツイート間の相互関係を考慮した埋め込み表現  $H'_i$  を得る。その後、field-level multi-head attention 層[5]を利用して  $H'_i$  を集約することで、最終的にユーザの居住地を 47 都道府県別に予測するための特徴量  $f_i$  を得る。

### 4. 評価実験

#### 4.1 実験設定

データセットは、先行研究[2]と同じものを利用する。病気・症状情報をツイートしているユーザで、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、推定居住地と完全

<sup>†</sup> 香川大学大学院創発科学研究科 Graduate School of Science for Creative Emergence, Kagawa University

<sup>‡</sup> 香川大学創造工学部 Faculty of Engineering and Design, Kagawa University

一致した場合を正解とする。評価指標には、47 都道府県別予測の Precision, Recall, F 値のマクロ平均を用いる。

提案手法で使用した BERT は、東北大学が公開している日本語事前学習済みモデル bert-base-japanese-v3<sup>1</sup>を用いる。ツイートテキストの最大系列長を 120, エポック数を 5 とし、検証データを用いて最もロスの少ないエポックを採用する。また、先行研究[2]のツイート内容モデルを同じデータセット分割で訓練し、これを従来手法 (ベースライン) として性能を比較する。

#### 4.2 結果と考察

47 都道府県別予測結果の Precision, Recall, F 値のマクロ平均を表 1 に示す。また、従来手法による都道府県別 Precision を図 1 に、提案手法による都道府県別 Precision を図 2 に示す。

表 1 より、BERT を利用した提案手法によって、Precision が 0.8000 を達成し、従来手法と比較して約 2 ポイント向上していることがわかる。ツイート内容のみを用いて Precision で 8 割台を達成したのは本研究では初めてであり、BERT の有効性が示された。また、図 1 および図 2 より、都道府県ごとの性能には顕著な地域差が確認できる。具体的には、提案手法により東京都で約 5 ポイント、大阪府で約 8 ポイントの Precision の低下を確認している。一方、地方部の多くの都道府県では Precision が向上しており、特に東北地方、中国地方、四国地方では平均で約 5 ポイントの改善が確認できた。

ユーザ数が少ない地方部の都道府県で Precision が向上したことは、BERT が従来手法では捉えきれなかったユーザを正確に推定できるようになったことを示しており、BERT による文脈を考慮したテキスト埋め込みが効果的に機能したといえる。一方で、東京都および大阪府で Precision が低下した原因としては、投稿内容の多様性がノイズとなり、BERT がそのノイズに過学習した可能性がある。大都市では、ユーザのツイート内容が多様であり、この点が居住地推定の性能低下に影響した可能性がある。つまり、BERT の高度な表現力は、データの少ない地方部では有利に作用したが、データの豊富な都市部では、細部のノイズにまで過学習してしまい、その結果として予測性能が低下したと考えられる。この問題を解決するためには、居住地推定の対象ツイートやユーザを選別し、ノイズを除去するプロセスが必要である。具体的には、地域特徴語や方言的な要素などを含まないユーザ、すなわち居住地推定が事実上不可能なユーザを除外するための対策が必要であると考える。このようなノイズの影響を減少させることで、BERT は東京や大阪においても高いパフォーマンスを発揮する可能性がある。また、提案手法とフォロー情報を組み合わせることで、F 値の最高性能を達成できることが期待できる。

表 1 都道府県別予測結果

	Precision	Recall	F 値
従来手法 (ベースライン)	0.7801	<b>0.6943</b>	0.7305
提案手法	<b>0.8000</b>	0.6855	<b>0.7319</b>

<sup>1</sup> <https://huggingface.co/tohoku-nlp/bert-base-japanese-v3>

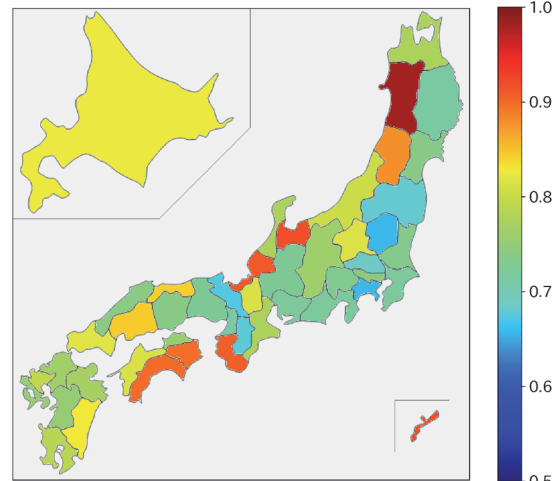


図 1 従来手法による都道府県別 Precision

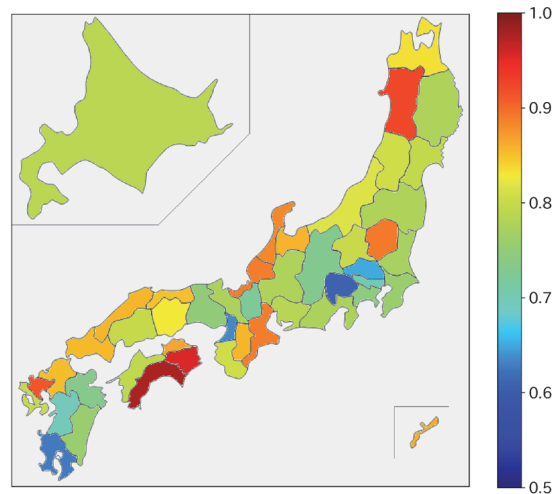


図 2 提案手法による都道府県別 Precision

#### 5. おわりに

本稿では、先行研究で提案したツイート内容モデルの性能向上を目的として、ツイートテキストからの特徴抽出モジュールを BERT に置き換えたモデルを提案した。実験の結果、特に地方部の Precision が大幅に向上し、全体の Precision で初めて 8 割台を達成した。

今後は、フォロー情報と組み合わせた実験や、居住地推定が不可能なユーザの特定と除去により、より頑健な居住地推定手法の開発に取り組みたい。

#### 参考文献

- [1] 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).
- [2] 松本他, “ツイート内容と天気・地震情報を用いた Twitter ユーザの居住地推定”, 情報処理学会第 85 回全国大会講演論文集, pp.1-499-1-500 (2023).
- [3] 平松他, “投稿内容とフォロー情報を用いた SNS ユーザの居住地推定”, 情報処理学会第 86 回全国大会講演論文集, pp.1-463-1-464 (2024).
- [4] J.Devlin, et al., "BERT: Pre-training of deep bidirectional transformers for language understanding.", in Proc. of NAACL-HLT2019, pp. 4171-4186, (2019).
- [5] B.Huang, et al., "A hierarchical location prediction neural network for twitter user geolocation," in Proc. of EMNLP-IJCNLP2019, pp.4732-4742 (2019).