Vol.2021-CSEC-93 No.27
Vol.2021-IOT-53 No.27
2021/5/14

深層学習を用いたTwitterユーザの位置推定の試み

笠井 遥輝[1]　利光 能直[2]　菊田 翼[2]　嶋田 里聖[1]　川越 響[1]　田畑 唯斗[1]　齋藤 孝道[1]

概要：Twitter におけるユーザ（アカウント）の位置推定について様々な手法が提案されている．しかし，多くの手法は言語依存であり，英語を用いるユーザを対象としている．そのため，日本国内のユーザの位置推定にそのまま適用することはできない．また，ツイートに位置情報を添付できるジオタグという機能は，正解データとして用いられていたが，2019 年に廃止された．その結果，正解データを用意することが難しくなった．そこで，本論文では，ジオタグの代わりに，位置情報共有機能を持つソーシャルメディア，Swarm から収集した情報を用いて正解データを作成し，その上で，国内の Twitter ユーザの情報を用いて，深層学習による都道府県単位でのユーザの位置推定，推定結果の評価，および改善点の考察を行った．

キーワード：位置推定，深層学習，Twitter，Swarm

User Geolocation Estimation on Twitter using with Deep Neural Networks

Abstract: Various methods have been proposed for estimating the geolocation of users (accounts) in Twitter. However, most of the methods are language-dependent and target English-speaking users, so they cannot be applied to the geolocation estimation of users in Japan. In addition, geotagging, a feature that allows users to attach geolocation information to tweets, was used as the correct answer data, but was discontinued in 2019. As a result, it has become difficult to prepare correct answer data. In this paper, instead of geotagging, we used information collected from Swarm, a social media platform with geolocation information sharing feature, to create the correct answer data. Then, using the information of domestic Twitter users, we estimated the geolocation of users by prefecture using deep neural networks, evaluated the estimation results, and discussed the points that need to be improved and the points that can be improved.

Keywords: Geolocation Estimation，Deep Neural Networks，Twitter，Swarm

1. はじめに

本論文では，国内の Twitter ユーザを都道府県単位で位置推定した．

スマートフォンの普及とともに，Twitter や Facebook，Instagram などのソーシャルメディアも急速に普及している．特に，Twitter はユーザの世界的な広がりとリアルタイム性の高さから，ユーザの居住地などを推定する研究に利用されている．

Twitter には，ジオタグというスマートフォンからのツイートに自身の現在位置を添付できる機能や，ユーザがプロフィールに自宅や職場などの位置情報を記載できる機能がある．しかし，Jurgens ら [1] によると，ツイートへのジオタグの添付は投稿全体の 3 %未満であり，利用率が極めて低い．実際に，Twitter 社は利用率が低いことを理由に，2019 年にジオタグを廃止しており，現在は利用できない．また，プロフィールに記載できる位置情報は，容易に書き換え可能であり，実際に自宅や職場の位置情報を記載しているユーザは極めて少ない．そこで，ユーザが添付している位置情報に頼らず，位置を推定する様々な手法を提案されている．たとえば，Huang ら [2] の研究のように，ツイートの文脈，プロフィールの記載事項，フォローフォロワーの関係などの情報を用いた位置推定の研究が行われている．これらの位置推定の研究の目的は，ユーザの自宅や職場，居場所などを国や地域の粒度で把握することである．

[1] 明治大学
Meiji University
[2] 明治大学大学院
Graduate School of Meiji University

しかし，提案されている Twitter における位置推定の研究の多くは言語依存であり，ある特定の言語や特徴にしか対応しない．そのため，特定の国や地域で，その手法を使うことができない場合，その国や地域の言語や特徴に合わせた位置推定の手法を構築する必要がある．特に，多くの位置推定の研究は，英語圏のユーザを対象としており，日本のユーザを対象とした位置推定の研究は少ない．また，ジオタグは位置推定の研究において正解データとして用いられていた．ジオタグが廃止されたことにより，正解データを用意することが難しくなった．

そこで，本論文では，ジオタグの代わりに，位置情報共有機能，および Twitter との連携機能を持つソーシャルメディア，Swarm から収集した位置情報を用いて正解データを作成し，その上で，国内の Twitter ユーザの情報を用いて，深層学習による Twitter ユーザの位置推定，推定結果の評価，および改善点の考察を行った．

2. 関連研究

Twitter における位置推定に関連する研究を示す．

Zheng ら [3] によれば，Twitter における位置推定は三つの種類に分けることができる．

- **Home Location Prediction(HLP)**
  Home Location は Twitter ユーザの長期的な居住地を指す．居住地を推定することで，ローカルコンテンツの推薦や位置情報に基づく広告の提示などに応用することができる．HLP では，正解データとなる居住地はユーザが記載しているプロフィールから収集することができる．しかし，プライバシー保護の観点から，ユーザのプロフィールには偽の位置情報や，空の情報，位置情報とは無関係な情報などが含まれることが多い．いくつかの研究では，プロフィールの位置情報ではなく，ユーザの投稿に添付されたジオタグを居住地の正解データとして集計している．集計方法としては，ジオタグの中で含まれる頻度が最も高い都市や，最初に投稿された有効なジオタグ，ジオタグの幾何学的中央値を求めることなどが挙げられる．
- **Tweet Location Prediction(TLP)**
  Tweet Location はツイートが投稿された実際の場所を指す．ツイートが投稿された位置を推定することで，Twitter ユーザの行動履歴を詳細に把握することができる．HLP とは異なり，一般的にツイートに添付されたジオタグに基づいて推定が行われる．
- **Mentioned Location Prediction(MLP)**
  Mentioned Location は，ツイートで言及されている場所を指す．地名を指す言葉の抽出や，抽出した言葉の指す位置の識別を行うことで，ツイートの内容理解が容易になる．MLP は，災害や疾病管理関連のアプリケーションに応用できる可能性がある．

また，位置情報の推定には，基本的に以下の評価指標が用いられている．

- **Accuracy(精度)**：正しく予測された位置の割合
- **Acc@161**：正解の位置から半径 161km(100 マイル) の誤差を許容した際に，正しく予測された位置の割合
- **Median(中央値)** ：距離誤差の中央値
- **Mean(平均値)** ：距離誤差の平均値

Lourentzou ら [4] は，ツイートのみを情報として，深層学習による位置推定を行った．その際，以下の三種類のデータセットを用いて米国の州や地域の分類を行った．

- **GeoText**
  Eisenstein ら [5] によって作成され，9,500 ユーザによる 38 万件のツイートを集めたデータセットである．また，各ユーザの地理的座標が記載されている．
- **TwUS**
  Roller ら [6] によって作成され，北米の 45 万ユーザによる 3,800 万件のツイートを集めたデータセットである．また，このデータセットはそれぞれのユーザが行った全てのツイートを集めたもので，一番最初に行ったジオタグ添付のツイートを正解データとして用いている．
- **TwWORLD**
  Han ら [7] によって作成され，140 万ユーザによるツイートを集めたデータセットである．このデータセットは，米国だけでなく全世界のデータを含んでいる．

これら三つの異なるデータセットを用いた推定精度を表 1 に示す．

表 1 Lourentzou らによる推定精度

データセット	Acc @ 161(%)
GeoText	29
TwUS	43
TwWORLD	21

三浦ら [8] は深層学習を用いて Twitter における位置推定を行った．609,691 ユーザによる 7,653,295 件のツイートを学習データとし，6,221 ユーザによる 78,173 件のツイートを検証用データとし，メタデータとツイートを統合してモデルの学習を行った．この結果，ツイートのみでの推定精度は約 40.9%で，Median は 69.5km であった．また，メタデータを含めた場合の推定精度は約 47.5%で，Median は 16.1km となった．この結果より，メタデータを含めた方が予測精度が高くなることが示された．

Huang ら [2] は，階層構造のニューラルネットワークを用いてツイートやメタデータを利用した位置情報の推定方法を提案した．この方法では，プロフィールやツイート，ユーザのフォローフォロワーの関係から構築されたネットワークを推定に利用している．用いたデータセットは，Twitter-US（Lourentzou らが用いた TwUS と同様），

Twitter-World（Lourentzou らが用いた TwWORLD と同様），WNUT の 3 種類である．WNUT は，Han ら [9] によって発表され，100 万ユーザの学習データと，それぞれ 1 万ユーザのテストデータ及び検証データを集めたデータセットである．

Huang らは用いたデータの組み合わせごとに精度を算出した．その結果を表 2 に示す．なお，精度評価は全て Acc@161 である．また，表中では Twitter-US と Twitter-World をそれぞれ TwUS と TwWorld，フォローフォロワーの関係を FF 関係と表現している．

表 2 Huang らによる推定精度

用いた情報	TwUS	TwWORLD	WNUT
ツイートのみ	57.1	40.1	52.9
ツイートとメタデータ	61.1	56.4	73.1
ツイートと FF 関係	72.7	68.4	73.1

Singh ら [10] は Twitter における位置推定の利用方法として，洪水のような災害において，Twitter を用いて助けを求める被災者を特定するアルゴリズムと，その位置を推定するシステムを考案した．Singh らは，インドで実際に発生した洪水関連のツイートを 3 万件収集し，まずそれぞれのツイートを投稿したユーザの優先度の高さを分類する機械学習ベースのシステムを構築した．この研究における優先度の高さとは救助の必要性の度合いを示す．それにより，より優先度の高いツイートを抜き出し，それらのツイートに対して位置の推定を行うことを可能にした．位置推定を行うにあたって，いくつかのステップを踏んでいる．まず，助けを求める該当のツイートにジオタグが付けられているかどうかを確認する．この段階で，ジオタグが添付していることが確認できればジオタグの情報がそのままユーザの位置となる．ジオタグが付いていない場合，該当のツイートを行ったユーザの過去 7 日間のツイートを取得し，ユーザ位置に関するマルコフモデルを生成し，モデルから位置推定を行う．最終的な位置推定の精度は，総推定値に対する推定成功率で測定され， 推定精度は 87%とされている．

3. 実験データ

本節では，本論文での位置推定の実験に用いるデータについて説明する．

3.1 Swarm での投稿を共有したツイートとユーザ情報の収集

本論文では，Swarm の投稿を Twitter で共有しているユーザの情報を集める．そのために，Twitter API[11] を用いて 2020 年 11 月 7 日〜11 月 13 日および 2021 年 3 月 1 日〜3 月 7 日に投稿された，”swarmapp.com”というキーワードを含む国内のツイートおよびそのツイートを行ったユーザ情報を収集した．

3.2 ツイートの収集

3.1 節で収集したユーザが 2020 年 11 月 14 日〜11 月 24 日および 2021 年 3 月 8 日〜3 月 15 日に投稿したツイートを，Twitter API を用いて収集した．

本論文で収集し，位置推定に用いた情報について表 3 に示す．なお，各情報の文字数の上限は Twitter 社が設定している上限である．

表 3 位置推定に用いた情報

情報	情報の説明	文字数の上限
Name	アカウントの表示名	50
Location	プロフィール欄の位置情報	30
Description	プロフィール文	160
Tweet	ツイート	140

本論文では，あるユーザの Name，Location，Description のいずれかが英語や絵文字，数字など，日本語でない文字のみの情報が含まれていた場合，ツイートを含めたそのユーザの情報を位置推定に用いないこととした．

今回，本論文用に集めたデータ数はユーザ情報が 7,106 件，ツイートが 11,317,845 件であった．

3.3 情報の前処理

本実験で用いる情報はいずれも自然言語である．したがって，深層学習のために前処理として，自然言語処理を行う必要がある．

まず，データに対して形態素解析を行った．形態素解析ツールとして，Mecab を用いるとともに，分かち書き辞書として mecab-ipadic-NEologd を用いた．ここで分かち書き辞書とは，形態素解析を行う際に参照される辞書である．分かち書き辞書内の単語を基に文章を単語に分割する．Twitter から収集されるデータには，新語や固有表現が多く含まれる．mecab-ipadic-NEologd は，オープンソース・ソフトウェアであり，頻繁に更新されているため，新語や固有表現に強く，単語数が多い．

また，Twitter から収集したデータには，位置推定する上でノイズとなり得る単語が多く含まれる．それらの単語を除去するため，地名や地域の固有表現抽出を行った．固有表現抽出には，形態素解析と同様のツール，辞書を用いた．

3.4 正解ラベルの作成

以下に正解ラベルの作成手順を示す．

3.4.1 Swarm の投稿が示す位置の緯度経度を収集する

3.2 節で収集したツイートから，Swarm での投稿を共有したツイート（以下，Swarm ツイート）を抽出した．抽出したツイート内の URL にアクセスし，スクレイピングを行うことで，各 Swarm ツイートの緯度経度を取得した．3.2 節で収集したツイートのうち，Swarm ツイートは 1,324,878 件であった．

3.4.2 ユーザごとに緯度経度を定める

3.4.1 節で収集した緯度経度を基に，以下の手順で全ユーザについて，ユーザごとの緯度経度を決定した．本論文では，ユーザごとの緯度経度とは，あるユーザが最も多く Swarm ツイートを投稿した地域の緯度経度と定義した．

(1) あるユーザの Swarm ツイートとその緯度経度を抽出する
(2) あるユーザの 1 つの Swarm ツイートについて，その緯度経度を中心として半径 30km 以内にある，同一ユーザの他の Swarm ツイートの個数をカウントする．上記の操作をそのユーザの全 Swarm ツイートについて行う
(3) 上記の個数が最も多い Swarm ツイートの緯度経度をユーザごとの緯度経度とする

3.4.3 正解ラベルの作成

本論文では，都道府県単位の位置推定を行った．そのため，3.4.2 節で定めたユーザごとの緯度経度は正解ラベルとしてそのまま用いない．まず，国土交通省の位置参照情報ダウンロードサービス [12] を利用し，市区町村単位の緯度経度とその市区町村が属する都道府県を取得した．次に，取得した市区町村の緯度経度のうち，ユーザごとの緯度経度に最も近い市区町村の緯度経度を求めた．最後に，その市区町村が属する都道府県を位置推定における正解ラベルとした．

4. 実験

4.1 実験概要

本論文では，Twitter と Swarm の情報を用いて，Twitter ユーザの位置推定を行った．実験では以下の点を検証した．

- 表 3 に挙げた各情報を一つずつ用いて位置推定した場合の精度評価
- 表 3 に挙げた全ての情報を用いて位置推定した場合の精度評価
- 都道府県ごとの推定結果の調査
- 位置推定に影響を与える情報の調査

4.2 実験の詳細

本論文で行う 3 つの実験の詳細について説明する．なお，評価指標は 5.1 節で説明する表 4 に，データセットを学習用とテスト用にユーザ単位で分割した件数およびツイートの件数を示す．

表 4 学習用，テスト用データの件数

データの種類	ユーザ件数	ツイート件数
学習用データ	4,974	8,375,205
テスト用データ	2,132	2,942,640

4.2.1 実験 1 の詳細

実験 1 では，表 3 に挙げた情報のうち，一つを用いた深層学習を 4 パターン行い，その上で，位置推定および精度評価を行った．そのために，各情報を学習するためのニューラルネットワークを作成した．作成したニューラルネットワークの構造は 4.3 節で説明する．

4.2.2 実験 2 の詳細

実験 2 では，表 3 に挙げた全ての情報を用いた深層学習を行い，その上で位置推定および精度評価を行った．そのために，Keras の concatenate を用いて，各情報について作成したニューラルネットワークを連結した．

本実験では比較のため，実験 1 で学習し，パラメータ調整済みのニューラルネットワークを結合した場合と，パラメータを調整していないニューラルネットワークを結合した場合の 2 パターンの深層学習を行った．どちらの場合も，推定モデルは同様の構造である．なお，パラメータ調整済みのニューラルネットワークに含まれる層についてはパラメータを更新しないようにした．

また，都道府県ごとの推定結果を調査するため，パラメータ調整済みのニューラルネットワークを連結した推定モデルについて，都道府県ごとの推定結果も求めた．

4.2.3 実験 3 の詳細

実験 3 では，実験 1，2 の結果を踏まえて，位置推定に影響を与える情報を調査するため，情報のうち 1 つを除外して，3 つの情報を用いた深層学習を 4 パターン行い，その上で位置推定および精度評価を行った実験 2 と同様，Keras の concatenate を用いて，各情報について作成したニューラルネットワークを連結した．なお，本実験ではパラメータ調整済みのニューラルネットワークを連結した．

4.3 ニューラルネットワークの構造

ニューラルネットワークの構造は，文章中の単語をベクトル化する一層の Embedding 層，文章中の単語の長期的な依存関係を学習する一層の Long Short Term Memory(LSTM) 層である．Tweet 学習用のニューラルネットワークには，学習高速化と過学習防止の目的で，BatchNormalization を用いた．損失関数は交差エントロピー関数，最適化関数には Adam を使用した．

また，Embedding 層で文章中の単語をベクトル化する際，次元数を指定する．各情報を学習する際に指定した次元数を表 5 に示す．

表 5 単語ベクトル化の次元数

情報	ベクトル化の次元数
Name	20
Location	30
Description	50
Tweet	50

5. 実験結果

5.1 位置推定精度算出に用いた指標

位置推定の評価指標として，Precision のマクロ平均, Recall のマクロ平均, Accuracy，F1 値のマクロ平均，F1 値を使用した．F1 値は，都道府県ごとの位置推定の評価指標として使用した．以下で，本論文の位置推定に合わせてそれぞれの評価指標について説明する．

- Precision のマクロ平均：特定の都道府県と位置推定したデータのうち，実際にその都道府県のデータである割合を全都道府県について算出し，平均した値
- Recall のマクロ平均：特定の都道府県のデータのうち，実際にその都道府県と位置推定したデータの割合を全都道府県について算出し，平均した値
- Accuracy：位置推定の正解率
- F1 値のマクロ平均：Precision のマクロ平均と Recall のマクロ平均の調和平均
- F1 値：都道府県ごとの Precision と Recall の調和平均

以下に，Precision のマクロ平均, Recall のマクロ平均, Accuracy， F1 値のマクロ平均，F1 値を求める式を示す．

$$MacroPrecision = \frac{1}{47}\sum_{i=1}^{47}\frac{TP_i}{TP_i + FP_i}$$

$$MacroRecall = \frac{1}{47}\sum_{i=1}^{47}\frac{TP_i}{TP_i + FN_i}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$MacroF1 = \frac{2 \cdot MacroPrecision \cdot MacroRecall}{MacroPrecision + MacroRecall}$$

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

上記の位置推定の精度評価指標の算出の際には，表 6 に基づき，各都道府県ごとに，TP, TN, FP, FN を算出した．

表 6 ある都道府県における位置推定の TP,TN,FP,FN の分類

		実際の位置	
		対象の都道府県	その他
位置推定	対象の都道府県	TP	FP
	その他	FN	TN

5.2 実験 1 の結果

表 3 に挙げた情報を一つずつ用いて位置推定した結果を表 7 に示す．

この結果から，Tweet のデータを用いた場合の位置推定が全評価指標において最高精度となることが分かった．また，Name と Description のデータを用いた場合の位置推定は，Precision，Recall，F1 値のマクロ平均が 0.1 より低くく，ほとんど位置推定できないことが分かった．

表 7 実験 1 の結果

推定モデル	Precision	Recall	Accuracy	F1
Name	0.03	0.02	0.17	0.02
Location	0.18	0.12	0.28	0.14
Description	0.04	0.02	0.15	0.02
Tweet	0.55	0.35	0.60	0.40

5.3 実験 2 の結果

表 3 に挙げた情報を全て用いて位置推定した結果を表 8 に示す．なお，表 8 の Trained はパラメータ調整済みの二ューラルネットワークを連結した推定モデルであり，Untrained はパラメータを調整していない二ューラルネットワークを連結した推定モデルである．

この結果から，パラメータ調整済みのニューラルネットワークを連結した推定モデルより，パラメータを調整していないモデルを連結した推定モデルの方が，全ての評価指標において高精度となることが分かった．また，いずれの推定モデルも実験 1 の Tweet を用いた場合より精度が低下することが分かった．

表 8 実験 2 の結果

推定モデル	Precision	Recall	Accuracy	F1
Merged(Trained)	0.18	0.12	0.33	0.15
Merged(Untrained)	0.25	0.20	0.37	0.21

また，データ数と位置推定の精度の関係性を調査するため，データ数が多い上位 10 都道府県，F1 値が高い上位 10 都道府県を表 9，表 10 にまとめた．なお，表 9，表 10 の F1 値は都道府県ごとの F1 値であり，F1 値のマクロ平均ではない．

表 9, 10 から東京都, 愛知県, 大阪府, 兵庫県がデータ数と F1 値ともに上位 10 都道府県に入っており，特に東京都はデータ数が最も多く，F1 値が最も高いことが分かった．

表 9 データ数上位 10 都道府県

rank	都道府県	データ数	F1
1	東京都	1933	0.44
2	神奈川県	702	0.17
3	愛知県	504	0.37
4	大阪府	497	0.27
5	埼玉県	370	0.12
6	千葉県	301	0.08
7	静岡県	219	0.18
8	兵庫県	200	0.22
9	京都府	186	0.10
10	北海道	184	0.17

表 10 F1 値上位 10 都道府県

rank	都道府県	データ数	F1
1	東京都	1933	0.44
2	愛知県	504	0.37
3	広島県	93	0.33
4	富山県	46	0.31
5	熊本県	39	0.29
6	大阪府	497	0.27
7	福岡県	170	0.27
8	新潟県	118	0.24
9	兵庫県	200	0.22
10	岡山県	69	0.22

5.4 実験 3 の結果

表 3 に挙げた情報のうち，1 つを除外して位置推定した結果を表 11 に示す．なお，本節の比較は全て実験 2 のパラメータ調整済みのニューラルネットワークを連結した場合との比較である．

この結果から，Location と Tweet を除外した場合，全評価指標において全ての情報を用いた場合より精度が低下することが分かった．特に，Tweet を除外した場合，全度評価指標において最低精度となることが分かった．また，Name と Description を除外した場合，Accuracy を除く評価指標において全ての情報を用いた場合より精度が向上することが分かった．

表 11 実験 3 の結果

除外した情報	Precision	Recall	Accuracy	F1
Name	0.19	0.18	0.31	0.17
Location	0.13	0.12	0.26	0.12
Description	0.23	0.18	0.33	0.19
Tweet	0.07	0.06	0.17	0.06

6. 考察

6.1 実験 1 について

実験 1 の結果について，各情報に地名を含めているユーザの割合という観点から考察した．

表 12 を見ると，Name に地名を含めているユーザの割合は，0.04 であった．この結果から，Name を用いた位置推定の精度が低い原因は，Name に地名を含めているユーザが極めて少ないことであると考えられる．

Description に地名を含めているユーザの割合は 0.43 であり，Name の約 10 倍であるにも関わらず，Description を用いた位置推定の精度は，Name を用いた位置推定の精度と同程度になった．これは，Description に含まれている地名の多くが，正解データと関連がないことが原因であると考えられる．

Location に地名を含めているユーザの割合は 0.62 であった．Location はユーザの位置情報であるはずだが，約 4 割のユーザは Location に地名などの位置情報と関連がない情報を記載していることが分かった．また，位置推定の精度を見ると，Accuracy が 0.28 となっているため，Location に地名が含まれている場合でも，正解データと関連がない地名である可能性がある．

ユーザごとに地名を含む Tweet が必ず存在していることが分かった．また，ユーザごとの Tweet に含まれる地名数の平均は約 640 個であった．含まれる地名数が多いことが，Tweet を用いた位置推定の精度が最も高い原因であると考えられる．

表 12 地名を含むユーザの割合

情報	地名を含むユーザの割合
Name	0.04
Location	0.62
Description	0.43
Tweet	1.00

6.2 実験 2 について

都道府県ごとの推定結果について考察を行う．

データ数，F1 値ともに上位 10 都道府県に入っているのは，都市部の中心となっている都府県が多い．このことから，本実験で用いたデータセットに関しては，都市部の中心となっている都道府県は地方の都道府県より位置推定しやすいと考えられる．しかし，都道府県ごとのデータ数に偏りがあることが，この傾向の原因となっている可能性もある．

6.3 実験 3 について

各情報を除外した場合の精度と実験 2 の精度を比較し，各情報が位置推定に与える影響について考察する．

Name または Description を除外した場合の Accuracy を除く精度評価指標が，実験 2 より高くなったことにより，Name と Description が位置推定に影響を与えない，あるいは精度低下の影響があると考えられる．実験 1 でも，Name または Description を用いた場合，他の 2 つの情報を用いた場合と比べて，精度が著しく低くなっていた．

Locaiton または Tweet を除外した場合の全ての精度評価指標が，実験 2 より高くなったことにより，Location と Tweet が位置推定に対して精度向上の影響があると考えられる．特に，Tweet を除外した場合の精度が著しく低下したことより，Tweet は本論文で用いた情報の中で最も位置推定に影響を与えると考えられる．

7. 今後の課題

7.1 実験データについて

本論文で収集したデータは，都道府県ごとでデータ数の偏りが大きい．東京都をはじめとする首都圏に属する都道府県のデータ数は多い傾向にあるが，地方にある都道府県

のデータ数は少ない傾向にある．都道府県ごとのデータ数の偏りが小さくなれば，位置推定の精度が向上する可能性がある．また，深層学習を行うには総データ数が少ないため，さらにデータを収集することで位置推定の精度が向上する可能性がある．

7.2 位置推定に用いた情報について

本論文ではアカウントの表示名，プロフィール欄の位置情報，プロフィール文，ツイートを用いた位置推定の実験を行った．

2 節で挙げた Huang らの研究では，フォローフォロワーの関係のネットワークを用いることで高い精度での位置推定を実現している．フォローフォロワーの関係のネットワークを用いることにより，位置推定の精度が向上する可能性がある．

また，Twitter から取得できるメタデータは本論文で用いた情報やフォローフォロワーの関係以外にも他にもいくつかある．たとえば，アカウントの作成日時や使用端末などが挙げられる．これらの情報も含めた位置推定を行うことで，精度が向上する可能性がある．

7.3 日本語の自然言語処理について

本論文では，日本語の文章に対して形態素解析を行っているが，一つの単語でも漢字表記や平仮名表記などの違いがある．たとえば,「東京都」と「とうきょうと」のように同じ意味の単語でも，分かち書きした結果，誤って解釈し異なる意味として処理されてしまうことがある．表記揺れによる誤った解釈を減らすことにより，精度が向上する可能性がある．

7.4 推定モデルについて

2 節で挙げた三浦らの研究では，ツイート以外のメタデータを同時に用いることで精度が高くなることが示されているが，本論文では複数の情報を用いた位置推定の精度の方が低くなった．また，同研究でパラメータ調整済みのニューラルネットワークを連結した場合の精度が，パラメータ未調整のニューラルネットワークを連結した場合より高くなることが示されているが，本論文ではパラメータ調整済みのニューラルネットワークを連結した場合の精度の方が低くなった．いずれの原因も解明できていない．今後の研究では，この原因を実験データの偏りおよび不足と仮定し，実験データの問題点の解決に取り組む．

8. 研究倫理

我々は，Menlo report[13] の精神に則り，倫理的配慮をして実験を行った．実験を行う際，個人識別はせずプライバシーを遵守した．本論文で使用したデータセットは学術的な目的にのみ使用し，我々の研究室にて厳重に保管されており，他者への売却および提供をしない．

9. まとめ

本論文では，Twitter から取得したプロフィール情報やツイートと，Swarm から取得した情報を用いて深層学習を行い，その上で Twitter ユーザの位置推定の実験を行った．

まず，1 つの情報を用いた実験を行い，その後複数の情報を用いた実験を行った．結果として，プロフィールに記載された位置情報とツイートはユーザの位置推定に有用な可能性があることがわかった．

参考文献

[1] David Jurgens, Tyler Finnethy, James Mccorriston, Yi Tian Xu, and Derek Ruths. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice, 2015.

[2] Binxuan Huang and Kathleen M. Carley. A hierarchical location prediction neural network for twitter user geolocation, 2019.

[3] X. Zheng, J. Han, and A. Sun. A survey of location prediction on twitter. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 30, No. 9, pp. 1652–1671, 2018.

[4] Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. Text-based geolocation prediction of social media users with neural networks. In *2017 IEEE International Conference on Big Data (Big Data)*, pp. 696–705. IEEE, 2017.

[5] Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1277–1287, Cambridge, MA, October 2010. Association for Computational Linguistics.

[6] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1500–1510, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[7] B. Han, P. Cook, and T. Baldwin. Text-based twitter user geolocation prediction, 2014.

[8] Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. A simple scalable neural networks based model for geolocation prediction in twitter. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pp. 235–239, 2016.

[9] Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pp. 213–217, 2016.

[10] Jyoti Prakash Singh, Yogesh K Dwivedi, Nripendra P Rana, Abhinav Kumar, and Kawaljeet Kaur Kapoor. Event classification and location prediction from tweets during disasters. *Annals of Operations Research*, Vol. 283, No. 1, p. 737–757, 2019.

[11] Twitter API Documentation — Docs — Twitter Developer. https://developer.twitter.com/en/docs.

[12] 国土交通省 — 位置参照情報ダウンロードサービス. https://nlftp.mlit.go.jp/isj/.
[13] D. Dittrich and E. Kenneally. The menlo report: Ethical principles guiding information and communication technology research. u.s. department of homeland security, 2012.