# ツイート投稿位置推定のためのノイズとなる単語の除去手法

森國泰平 ‡† 吉田光男 ‡‡† 岡部正幸 ‡‡‡†† 梅村恭司 ‡‡‡‡†

† 豊橋技術科学大学 情報・知能工学系 〒 441–8580 愛知県豊橋市天伯町雲雀ヶ丘 1–1
†† 豊橋技術科学大学 情報メディア基盤センター 〒 441–8580 愛知県豊橋市天伯町雲雀ヶ丘 1–1
E-mail: ‡ morikuni14@ss.cs.tut.ac.jp, ‡‡ yoshida@cs.tut.ac.jp, ‡‡‡ okabe@imc.tut.ac.jp,
‡‡‡‡ umemura@tut.jp

**あらまし** ツイートから特徴を抽出し、投稿位置と対応をさせることで、Twitter を実世界を観測するリアルタイムセンサと考えることができる。しかし多くのツイートには位置情報が付加されていないために、Twitter をリアルタイムセンサとして活用することは困難である。そこで本研究では、ツイートの投稿位置を推定することで、より多くのツイートに位置情報を付加する問題に取り組む。提案する手法では、1 件のツイート内容のみを使用して推定を行う。また、ツイート中の単語から地理的な分布に偏りのない単語を除く手法を提案し、推定精度が向上することを示す。

**キーワード** Twitter, 地域語, 位置情報推定

## 1. はじめに

Twitter(注1) は、ツイートと呼ばれる短文を投稿することができるソーシャル・ネットワーキング・サービス (SNS) である。Twitter ユーザは、ツイートに自身の近況などをリアルタイムに記述し、同時に記述時の位置情報（ジオタグ）を付加して投稿する場合もある。ツイートから特徴を抽出し、付加された位置情報と組み合わせることで、その瞬間における地域の特徴を知ることができる。このことから、Twitter は実世界でのイベントを観測するリアルタイムセンサ（ソーシャルセンサ）として活用することが可能である。ソーシャルメディアをセンサとして活用する研究は、これまでに数多く行われている [1]。

Twitter をリアルタイムセンサとして活用する利点の 1 つとして、一般的なセンサでは観測できない事象を観測できることが挙げられる。例えば、一般的なセンサでも地震や台風などの災害自体を観測することはできるが、人的被害の状況や支援物資の不足を観測することは困難である。2011 年 3 月 11 日に発生した東日本大震災では、実際に Twitter がこのような情報伝達の手段として利用されており [2]、Twitter をリアルタイムセンサとして活用することには十分な意義があるといえる。

Twitter をリアルタイムセンサとして活用する場合に問題となるのは、位置情報が付加されたツイートが少なく [3]、日本語のツイートにおいても約 0.18%しか存在しないことである [4]。そのため観測できる情報が少なく、実際に Twitter をリアルタイムセンサとして活用することは困難である。投稿位置の推定によって、できるだけ多くのツイートに位置情報を付加することができれば、この問題を克服できると考えられる。

本研究では、各地域における単語の出現頻度を学習し、位置情報が付加されていないツイートの投稿位置を推定する。ツイート (投稿内容) には地域を特定することのできる単語が含まれていると考えられるため、本研究では 1 件のツイートのみを使用して投稿位置の推定を試みる。また、ツイートに含まれる単語から地理的な分布に偏りのない単語 (ノイズ) を取り除くことで、推定精度が向上することを示す。

## 2. 関連研究

Twitter をリアルタイムセンサとして活用するための研究は、これまでにも多く行われている [1]。Aramaki ら [5] は、Twitter からインフルエンザの流行を検出する方法を提案し、都道府県単位でインフルエンザの流行を可視化している (注2)。Sakaki ら [6] は Twitter を用いて日本で発生した震度 3 以上の地震の震源地を高精度で特定している。

SNS を用いて位置情報推定を行う研究は多く存在する。推定手法としては、投稿内容などを用いて推定を行うコンテンツベースの手法とユーザの友人関係などを用いて推定を行うグラフベースの手法という 2 つの手法に大きくに分けられる。

コンテンツベースの手法では、推定対象となるエリアごとに単語の出現頻度を学習し、投稿内容に含まれる単語を用いて位置情報を最尤推定することが多い。SNS に投稿される文章には、地理的な分布に偏りのない単語が多く含まれている。これらの単語をノイズとして除き、地理的な分布に偏りのある単語 (地域語) を抽出する手法として地域語フィルタが用いられる場合がある。Cheng ら [3] がユーザの居住地を推定するために提案した地域語フィルタでは、地理的に狭い範囲にのみ出現する単語を地域語とする。三木ら [7] らは、ツイートの投稿位置を推定するために TF–IDF [8] の概念を用いた地域語フィルタを提案している。三木らの提案した地域語フィルタでは、地域語の時間的局所性を考慮しており、一定期間ごとに地域語を更新することで、推定精度が向上したと報告している。地域語フィルタの他に、ツイートが投稿される位置がまばらであることから、頻度情報をスムージングする手法も提案されている [3]。単語の出現頻度の学習については、これまでに述べたようにツ

(注1) : https://twitter.com (accessed 2015–03–20)

(注2) : http://mednlp.jp/influ (accessed 2015–03–20)

イートのジオタグに含まれる位置情報を用いて学習を行うものが多いが、伊川ら [9] は、位置情報サービス (Foursquare(注3) など) から投稿されたツイートを用いて学習を行っている。

グラフベースの手法では、友人の居住地の位置情報を用いてユーザの居住地を推定する場合がある。Backstrom ら [10] は Facebook(注4) で居住地が記述されたユーザを用いて学習を行い、友人は近くに住んでいるという仮説から、居住地の推定を推定している。

本研究では、三木らと同様にコンテンツベースの手法によってツイートの投稿位置を推定する。三木らによる投稿位置推定は、推定を行ったツイートについての推定精度 (適合率) 向上と地域語の特定精度向上を目的にしていたが、本研究では、全てのツイートを対象とした推定精度 (再現率) 向上を目的とする。多くのツイートに位置情報を付加することができれば、Twitter をリアルタイムセンサとして活用することが可能になる。

## 3. 提 案 手 法

### 3.1 概　　要

本研究では、位置情報付きツイートに含まれる単語から、各エリアの単語の出現頻度を学習し、位置情報が付加されていないツイートの投稿位置を推定する。ツイート (投稿内容) には、方言やランドマーク名など、地理的局所性のある単語が含まれていると考えられる。これらの単語の分布を用いることで、位置情報が付加されていないツイートの投稿位置を推定できる可能性がある。そのため、本研究ではツイート内容のみを使用して投稿位置を推定する。また、Twitter をリアルタイムセンサとして活用するためにはリアルタイム性が求められることや、ツイート数の少ないユーザのツイートに対しても投稿位置を推定できることが望ましいため、推定に使用するのは 1 件のツイート内容のみとする。

投稿位置の推定は、式 (1) を用いた最尤推定によって行う。$A$ は推定対象エリアの集合 (4.1 節で詳述)、$a$ は推定対象エリア、$W_t$ はツイート $t$ に含まれる単語の集合、$p(w)$ はデータ中で単語 $w$ が出現する確率、$p(a|w)$ は単語 $w$ がエリア $a$ で出現する確率である。

$$\arg\max_{a \in A} p(a;t) = \sum_{w \in W_t} p(a|w)p(w) \tag{1}$$

これは単純にツイート内容のみを用いた推定であり、このままでは以下の 2 つの問題が存在する。

- ツイートに含まれるほとんどの単語は、地域を特定するための重みが小さい
- ツイートの分布がまばらであるため、観測誤差による影響が大きい

これらの問題を克服するために、Cheng ら [3] は地域語フィルタと頻度情報のスムージング手法を提案した。地域語フィルタとは、ツイートに含まれる単語から地域語とノイズを分類する手法である。頻度情報のスムージングとは、エリアごとの単語の出現頻度を平滑化することによって、観測誤差を小さくするための手法である。Cheng らは、地域語フィルタが推定精度を大きく向上させ、頻度情報のスムージングによる推定精度向上は大きくなかったと報告している。

本研究では 1 件のツイート内容のみを用いて推定を行うため、ツイートに含まれる単語の不足も問題となる。ツイートには最大でも 140 文字しか記述することができないため、推定に使用できる単語は少ない。そのため、以下の 2 つの要因によってツイートの投稿位置を推定できない場合がある。

- ツイートに含まれる単語が全て未知語である
- 地域語フィルタによって、ツイートに含まれる全ての単語がノイズとして除かれた

ここで問題とするのは、地域語フィルタで多くの単語を除くことによって、位置の推定が行えなくなることである。位置を推定できないツイートが増えると、推定精度 (再現率) の低下に繋がる。そのため、多くの単語を除かず、推定精度を向上させることのできる地域語フィルタが必要となる。

### 3.2 単語の出現頻度の学習

提案手法はエリアごとの単語の出現頻度を用いてツイートが投稿されたエリアを推定する最尤推定であり、全てのエリアについて各単語の出現頻度を学習する必要がある。そのため、位置情報が付加されたツイートに対して形態素解析を行い、単語の集合に分割する。さらに、エリアデータの境界情報から、ツイートが投稿されたエリアを特定し、エリアと各単語の出現頻度を対応づける。

### 3.3 ノイズとなる単語の除去

本研究では、ツイートの位置が推定できなくなるという問題に対処し、推定精度を向上させる地域語フィルタとして TF–IAF フィルタを提案する。TF–IAF フィルタは、情報検索などで用いられる TF–IDF [8] の概念を地域語フィルタに適用したものである。$C(w)$ を単語 $w$ の総出現回数、$A$ を推定対象エリアの集合、$A_w$ を単語 $w$ の出現したエリアの集合として式 (2) で表される。各単語について TF–IAF の値を求め、特定の閾値よりも値が小さいものをノイズとして分類する。

$$\begin{aligned} TF\text{–}IAF(w) &= TF(w) * IAF(w) \\ TF(w) &= \log(C(w)) \\ IAF(w) &= \log \frac{|A|}{|A_w|} \end{aligned} \tag{2}$$

TF–IAF フィルタは、ツイートにおける単語の出現頻度を TF の値で評価するため、ツイートに含まれる可能性の高い単語が地域語として判定されやすいという特徴を持っている。そのため、位置を推定できくなるツイートが少なくなると考えられる。また、投稿位置を推定したいツイートからは地理的な分布に偏りのない単語 (ノイズ) が除かれるため、推定精度の向上も期待ができる。

TF–IAF フィルタのもう 1 つの特徴として、三木ら [7] とは異なり、TF–IDF の概念をそのまま地域語フィルタに適用したものではない点が挙げられる。TF–IDF の概念をそのまま地域語フィルタに適用すると TF の計算式が、単語 $w$ の総出現回

(注3) : https://ja.foursquare.com (accessed 2015–03–20)
(注4) : https://www.facebook.com (accessed 2015–03–20)

数ではなく、あるエリアにおける単語 $w$ の出現回数となる。この場合に計算される TF–IAF の値は、特定のエリアで多く使用されるという直感的な地域語の判定基準になっているといえる。しかし本研究では、地域語の特定精度向上を目的とするのではなく、多くのツイートに位置情報を付加することを目的としているため、単語の出現頻度を全体で評価し、ツイートに多く含まれる単語を地域語とすることが重要となる。

### 3.4 頻度情報のスムージング

Cheng らはスムージングによって推定精度が向上した報告している。本研究では、Cheng らが提案した 4 つのスムージングうち、最も推定精度が向上した Lattice-based Neighborhood Smoothing を使用する。これは、日本を緯度経度 1 度単位の格子状に区切り、それぞれの格子における単語の出現確率を各エリアの出現確率に分配する手法である。Lattice-based Neighborhood Smoothing にはスムージングの強さを示すパラメータ $\mu$ と $\lambda$ が存在する。$\mu$ は格子間のスムージングの強さ、$\lambda$ は格子から都市へのスムージングの強さを調整するパラメータである。本稿ではこれらのパラメータの値を 0.5 で固定し、最適なパラメータの探索は今後の課題とする。

## 4. 実験設定

### 4.1 エリアデータ

エリアデータとは、学習と推定における領域を区別するための境界データと中心座標データである。山や川で隔てられたエリアでは、生活の様子や文化、方言などが異なる可能性がある。地理的な意味を考慮せずに緯度経度でエリアの分割を行うと、エリアの中で複数の方言が混在してしまう恐れがある。そこで、本研究では地理的に意味のある境界で分けられていると考えられる、都道府県や市区町村の領域データの利用を検討する。エリアとして扱う領域は、いくつかの分割の粒度を考えることができ、関西や関東などの地方単位や、都道府県単位、市区町村単位、町丁（丁目や字など）単位などが挙げられる。これらのうち、本研究では市区町村単位の領域をエリアとして使用する。

エリアデータは総務省統計局の「平成２２年国勢調査（小地域） 2010/10/01」(注5) から、境界データ (世界測地系緯度経度・G-XML 形式) を取得した。このデータは、町丁単位で分けられたデータである。町丁は県番号 (KEN) と県内通し番号 (SEQ_NO2) の 2 つのメタデータで一意に求まる。それぞれのデータについて、Boundary タグで囲まれた境界全体を含む矩形座標データ、Geometry タグで囲まれた境界を示す詳細な座標データ、Property タグで囲まれた都道府県名や中心座標などを示すメタデータの 3 つの情報が含まれている。

本研究では Geometry タグに含まれる詳細な境界データをエリアの境界として使用する。また、取得したデータは町丁単位のデータであるため、本研究でエリアとして扱う市区町村単位の境界データに変換する必要がある。県番号 (KEN) と県内都市番号 (CITY) の 2 つのメタデータで一意に市区町村が求められることから、以下のアルゴリズムによって、市区町村単位の境界データに変換を行った。

（1） (KEN, CITY) の 2 つのメタデータをキーとして、町丁単位のデータをグループ化する

（2） 各グループについて、境界データを結合する (境界データ完成)

（3） 各グループについて、結合された境界データの重心を求める (中心座標データ完成)

境界データの結合と重心の計算には、MySQL 5.6 の Multi-Polygon 型と Centroid 関数を使用した。

### 4.2 ツイートデータ

2013 年 1 月 1 日から 2013 年 2 月 1 日に投稿された位置情報付きのツイートデータを Twitter Streaming API(注6) を用いて収集した。さらに、日本国内からの投稿に限定するため、ツイートの投稿位置を示すメタデータ coordinates(注7) に指定された座標が、エリアデータのいずれかのエリアに含まれるツイートのみを抽出した。また、Bot による自動投稿の影響を少なくするために、以下の条件にあてはまるツイートを Bot による投稿として除いた。

（1） Twitter クライアント名に「NightFoxDuo」を含む

（2） ツイート内容に「きつねかわいい !!! 」を含む

（3） 緯度経度が (34.967096, 135.772691) である

（4） Twitter クライアント, アカウント名, 表示名, プロフィールのうち、1 つ以上に「BOT」「Bot」「bot」「人工無能」のいずれかを含む

先の 4 つの条件の中で (1)(2)(3) は「夜狐八重奏＋(注8)」の影響を除くことを目的としている。「夜狐八重奏＋」は iOS 向けの Twitter クライアントであり、ツイートのジオタグに京都府の伏見稲荷大社の座標を自動で追加する機能や、「きつねかわいい !!! (X 回目)」(X には投稿回数が入る) というツイートを自動投稿する機能がある。これらの機能によって多くのツイートが本来の位置情報と関係なく京都府伏見区で観測されるため、本研究ではこのクライアントのユーザを Bot として扱う。収集した 4,650,177 件のツイートから Bot による投稿を除いた結果、4,353,276 件のツイートが抽出できた。

Twitter をリアルタイムセンサとして活用するためには、ツイートが投稿されたときに位置を推定できなくてはならない。そのため、学習に利用するツイートデータは評価に利用するデータよりも過去に投稿されたものである必要がある。よって、2013 年 1 月 1 日から 2013 年 1 月 31 日までの 4,226,554 件のツイートを学習データとし、2013 年 2 月 1 日の 126,722 件のツイートの評価データとする。

### 4.3 評価方法

推定結果の評価には、正解度 (Accuracy)、適合率 (Preci-

(注5)：http://e-stat.go.jp/SG2/eStatGIS/page/download.html (accessed 2015–03–20)

(注6)：https://dev.twitter.com/streaming/overview (accessed 2015–03–20)

(注7)：https://dev.twitter.com/overview/api/tweets (accessed 2015–03–20)

(注8)：http://www58.atwiki.jp/nightfox (accessed 2015–03–20)

sion)、F 値 (F-measure) を用いる。F 値は一般的に再現率と適合率によって求められるが、今回は正解度と適合率を用いて計算を行う。ただし、正解度は推定対象ツイートのうち推定に成功したツイートの割合であるため、実質的には再現率と同様の指標となっている。また、地域語フィルタを評価するための指標として推定可能割合 (Estimatability) を用いる。各評価指標は以下の式で表される。

$$
\begin{aligned}
ErrDist(t) &= d(l_{act}, l_{est}(t)) \\
Accuracy(T,x) &= \frac{|\{t|t \in T \wedge ErrDist(t) \leqq x\}|}{|T|} \\
Precision(T,x) &= \frac{|\{t|t \in T \wedge ErrDist(t) \leqq x\}|}{|T_{est}|} \\
F\text{–}measure(T,x) &= \frac{2 * Accuracy(T,x) * Precision(T,x)}{Accuracy(T,x) + Precision(T,x)} \\
Estimatability(T) &= \frac{|T_{est}|}{|T|}
\end{aligned}
$$

$ErrDist(t)$ はツイート $t$ の実際の位置 $l_{act}$ と推定結果の位置 $l_{est}$ との距離、$T$ は推定対象ツイートの集合、$T_{est}$ は位置を推定できたツイートの集合、$x$ は推定成功とする最大の距離 (正解距離) である。

### 4.4 Cheng らが提案した地域語フィルタとの比較

提案手法の有効性を確認するために、Cheng ら [3] が提案した地域語フィルタとの比較を行う。式 (3) は Cheng らが提案した地域語フィルタに使用されるモデルである。$a$ は中心となるエリア、$l$ は単語 $w$ の地域性の強さを示す指標、$C$ は単語 $w$ がエリア $a$ で出現する確率、$A_w$ は単語 $w$ が出現するエリアの集合、$d_i$ はエリア $a$ とエリア $i$ との距離である。全ての単語について $a$ と $l$ の値を求め、$l$ について閾値を設けることで地域語を特定する。

$$
\begin{aligned}
&\arg\max_{a,l} f(a,l;w) \\
&= \sum_{i \in A_w, i \neq a} \log C d_i^{-l} + \sum_{i \notin A_w, i \neq a} \log\left(1 - C d_i^{-l}\right) \qquad (3)
\end{aligned}
$$

Cheng らが提案した地域語フィルタと提案手法である TF–IAF フィルタとをフィルタした単語の特徴について比較する。単語の特徴による影響と単語数による影響とを分離するためには、フィルタした単語の数が同程度であることが望ましい。そのため、式 (3) で $l$ が $1.0 \times 10^{-3}$ の精度で求まらない単語、および $l$ が限りなく大きくなる単語 ($|A_w| = 1$) をフィルタし、TF–IAF フィルタにおいても同程度の単語数をフィルタするように閾値を設定した。

## 5. 実験結果

### 5.1 推定可能割合と性能の関係

Cheng らが提案した地域語フィルタと提案手法である TF–IAF フィルタとを用いて推定を行った結果を比較する。

推定可能割合を比較した結果が表 1 である。Cheng らが提案した地域語フィルタ、TF–IAF フィルタ共に 80%程度の単語をフィルタしている。フィルタなしのときの推定可能割合約 100%と比べ、Cheng らが提案した地域語フィルタでは推定可能割合が約 47%と大幅に低下している。この原因として Cheng らが提案した地域語フィルタでは、単語の地域性の強さを示す指標 $l$ に、単語の出現頻度が考慮されていないことが挙げられる。そのため、ツイートに多く含まれる単語を除き、推定を行えないツイートが増えたと考えられる。一方、TF-IAF フィルタでは推定可能割合が約 96%となり、位置を推定可能なツイートはほとんど減少しないという結果になった。

表 1　地域語フィルタの比較 – 推定可能割合

| フィルタ | フィルタした単語の割合 | 推定可能割合 |
|---|---|---|
| フィルタなし | 0% | 99.96% |
| Cheng | 81.39% | 47.16% |
| TF–IAF | 83.04% | 95.51% |

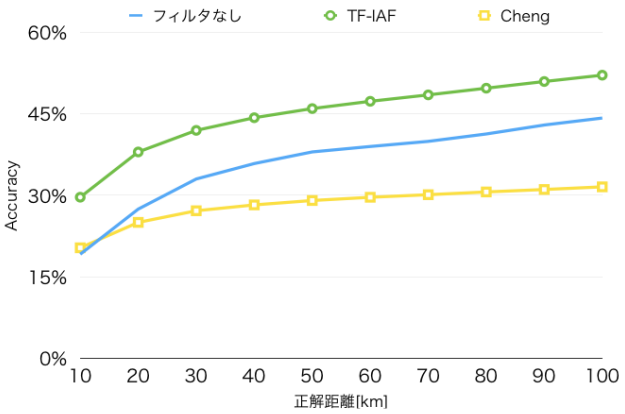


図 1　地域語フィルタの比較 – 正解度

正解距離を 10km から 100km まで変化させた場合の正解度、適合率、F 値を比較した結果が図 1、図 2、図 3 である。正解度と F 値については TF–IAF フィルタが最高性能を示し、正解距離 30km において正解度約 43%、F 値約 44%となった。適合率については Cheng らが提案した地域語フィルタが最高性能を示し、正解距離 30km において約 58%となった。Cheng らが提案した地域語フィルタの正解度が低くなった原因は、推定可能割合が大きく低下したためであると考えられる。このことから、正解度 (再現率) の向上を目的としている本研究においては、TF–IAF フィルタがより適した地域語フィルタであるといえる。そのため以降の実験では、TF–IAF フィルタの性能のみを評価する。

### 5.2 単語数と性能の関係

フィルタした単語数の変化によって推定性能がどのように変化するのかを調べる。フィルタする単語数を約 1%単位で変化させるために、TF–IAF フィルタの閾値を調整した。なお、TF–IAF の値には同値が多く存在するため、1%単位で単語数を変化させられなかった区間が存在する。

フィルタした単語数を変化させた場合の推定可能割合を図 4 に示す。フィルタした単語が多くなるにつれ、推定可能なツイートが減少することが分かる。約 99%の単語をフィルタしたときの推定可能割合は約 27%となった。

フィルタした単語数を変化させた場合の正解度、適合率、F 値の変化を、図 5、図 6、図 7 に示す。正解距離は 10km、30km、

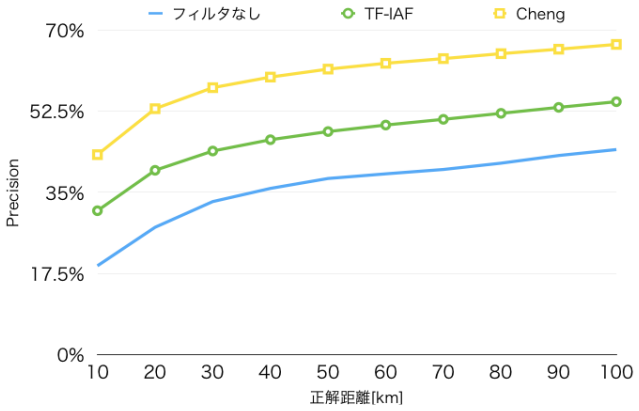


図 2　地域語フィルタの比較 – 適合率

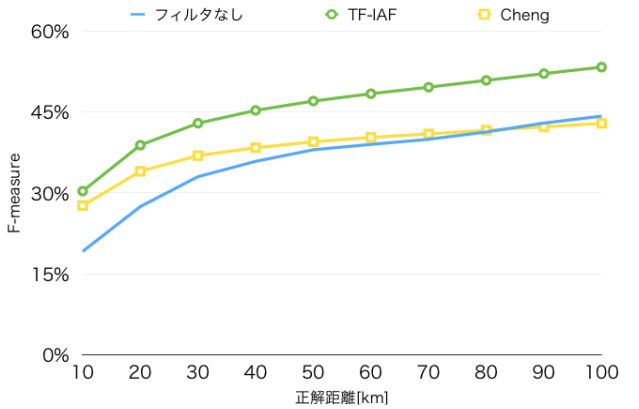


図 3　地域語フィルタの比較 – F 値

50km、100km と変化させた。図 5 と図 7 から、フィルタした単語数が約 83%の時に再現率と F 値が最大になり、フィルタした単語数が多すぎた場合と少ない場合との双方で精度が低下していることがわかる。一方、図 6 から、フィルタした単語が多ければ多いほど適合率が向上することが分かり、正解距離 30km において約 99%の単語をフィルタしたときの適合率は約 83%となった。このことから、地域語フィルタが地域語を特定したために正解度が向上したのではなく、ツイート中からノイズとなる単語を取り除くことができたために推定精度が向上したと考えられる。

### 5.3　都道府県別の性能比較

都道府県別の正解度を比較する。正解距離を 10km とし、日本全体での推定結果を都道府県別に評価した結果が図 8 である。人口の多い都道府県では高精度で推定を行えていることが分かる。特に東京都では約 61%のツイートを 10km 以内に推定できるという結果になった。また、推定結果のマイクロ平均は正解度、適合率、F 値が約 30%、約 32%、約 31%であり、都道府県別の推定結果のマクロ平均は正解度、適合率、F 値ともに約 19%となった。

### 5.4　時刻別の性能比較

時刻別の正解度を比較する。2 月 1 日のツイートを 1 時間単位で分割し、時刻ごとに評価した結果が図 9 である。正解距離は 10km、30km、50km、100km と変化させた。夜間に比べ昼

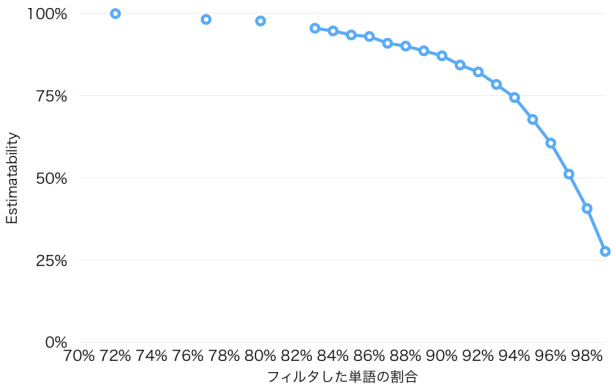


図 4　単語数による影響 – 推定可能割合

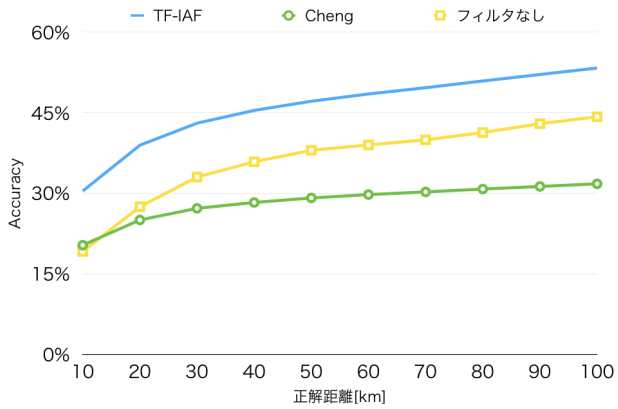


図 5　単語数による影響 – 正解度

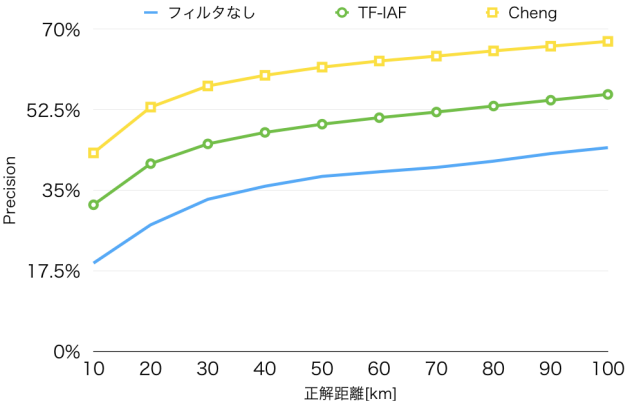


図 6　単語数による影響 – 適合率

間の正解度が高い傾向が見られる。また、8 時と 18 時の正解度が極大値をとっていることから、出勤や退勤の時刻に正解度が高くなることがわかる。特に 8 時は正解度の最大値をとっており、正解距離 30km においては約 52%となった。これは、ツイートに駅名などが含まれる可能性が高くなるためであると考えられる。

### 5.5　推定結果の例

投稿位置の推定に成功したツイートと失敗したツイートの例を表 2 と表 3 に示す。地名や駅名などが含まれているツイートでは推定に成功していることがわかる。推定に失敗したツイー

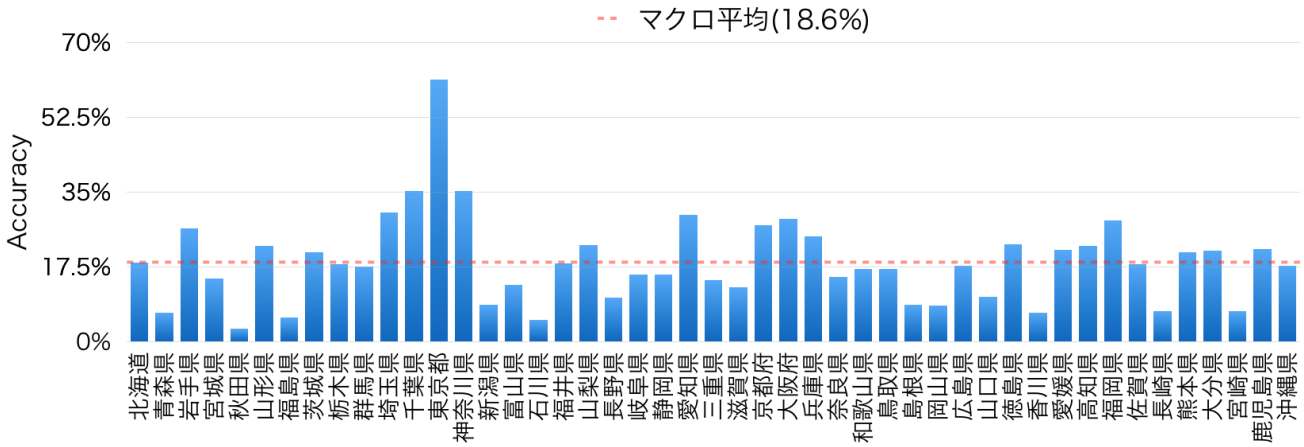


図 8　都道府県別の推定精度 – 正解度

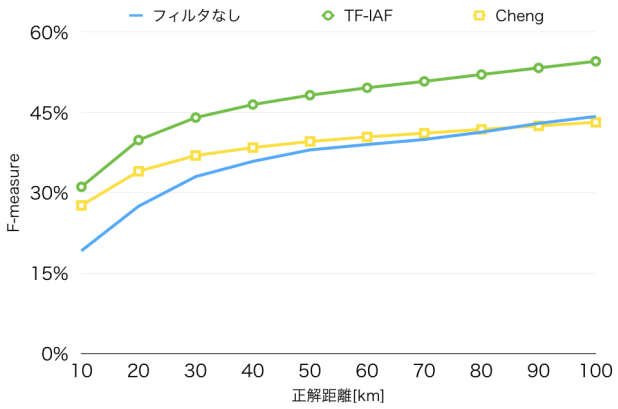


図 7　単語数による影響 – F 値

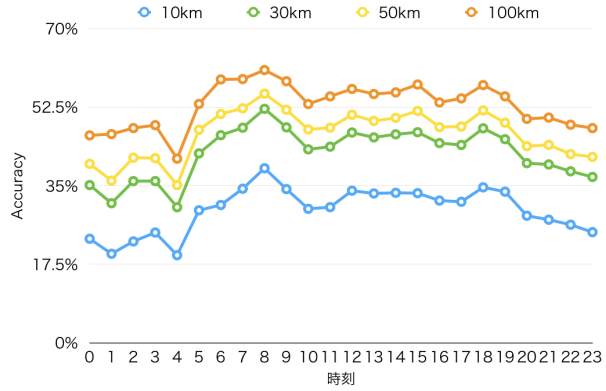


図 9　時刻別の推定精度 – 正解度

トでは、地域を特定できないような単語しか含まれないもの、現在位置と関係のない地名を含んでいるもの、全ての単語が除かれて推定を行えなかったツイートなどがある。このことから、本質的に位置を特定できないツイートが存在することが分かる。

推定に成功したツイートの例 3 件について、推定値上位 10 件の結果を図 10、図 11、図 12 に示す。赤色で中に黒点のあるマーカーが推定結果 (推定値 1 位) の座標である。図 12 では東京周辺に推定候補が集中しているが、図 10、図 11 ではいくつもの地域にばらけて分布している。これは「高畑」が地名だけでなく人名にも用いられることや、地名ではない単語が地域語

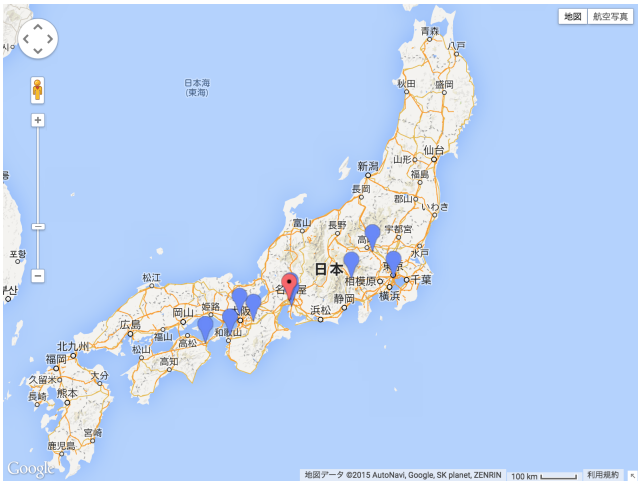


図 10　「高畑駅にタッチ！ http://t.co/fcw2uTIa」の推定結果

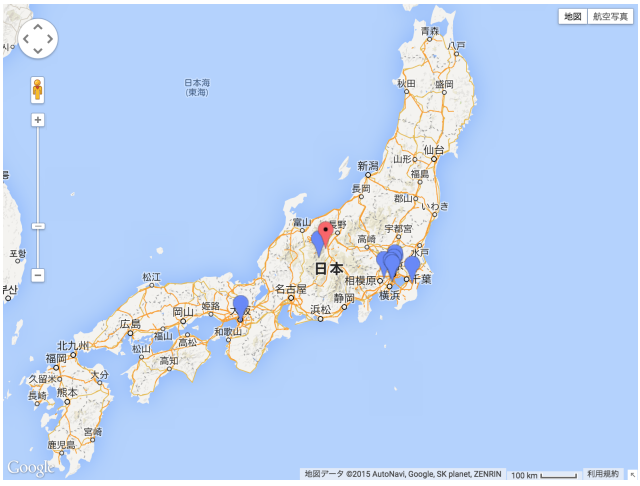


図 11　「安曇野とうちゃーく！雪、すくなっ(--;」の推定結果

として分類されることにより、特定の狭い範囲に関連付けられなかったためであると考えられる。

### 5.6　スムージングと性能の関係

スムージングによって推定精度がどのように変化するのかを調べる。本研究では Cheng ら [3] が提案した Lattice-based Neighborhood Smoothing を使用しているが、スムージングを

表 2　推定に成功したツイートの例

| ツイート内容 | 正解エリア | 正解座標 (緯度, 経度) | 推定エリア | 推定座標 (緯度, 経度) |
|---|---|---|---|---|
| 高畑駅にタッチ！ http://t.co/fcw2uTIa | 愛知県名古屋市中川区 | (35.140, 136.853) | 愛知県名古屋市中川区 | (35.121, 136.853) |
| 安曇野とうちゃーく！雪、すくなっ(--; | 長野県安曇野市 | (36.301, 137.924) | 長野県安曇野市 | (36.281, 137.868) |
| 代々木からなかなか動かないなー... | 東京都渋谷区 | (35.684, 139.705) | 東京都渋谷区 | (35.644, 139.715) |

表 3　推定に失敗したツイートの例

| ツイート内容 | 正解エリア | 正解座標 (緯度, 経度) | 推定エリア | 推定座標 (緯度, 経度) |
|---|---|---|---|---|
| あーねむ、ねよよー | 兵庫県宝塚市 | (34.823, 135.344) | 東京都港区 | (35.628, 139.749) |
| 滋賀に帰ったら一人暮らしでもするか (^^ゞ | 三重県四日市市 | (34.942, 136.609) | 滋賀県大津市 | (35.189, 135.906) |
| 帰宅。 | 埼玉県富士見市 | (35.838, 139.553) | 推定不可能 | - |

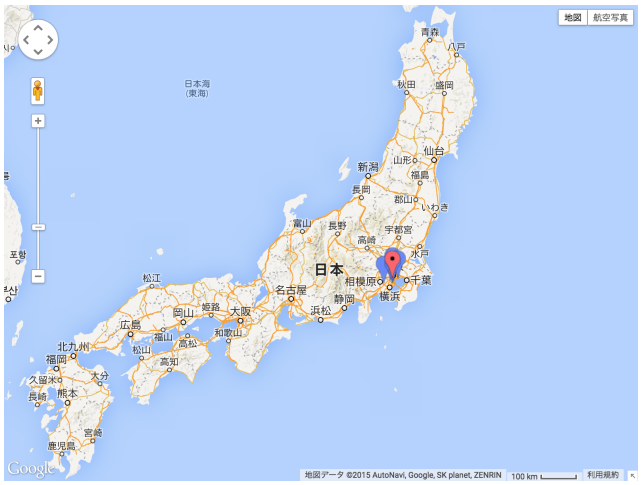


図 12　「代々木からなかなか動かないなー...」の推定結果

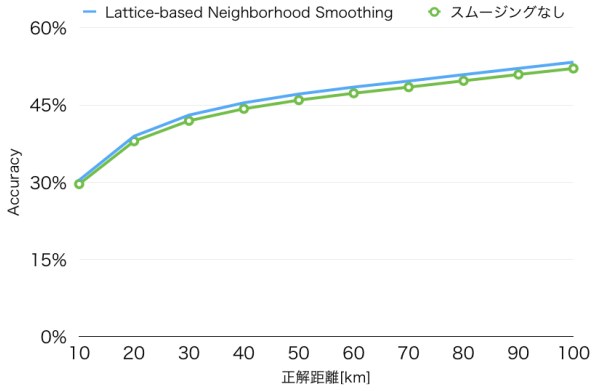


図 13　スムージングによる影響 – 正解度

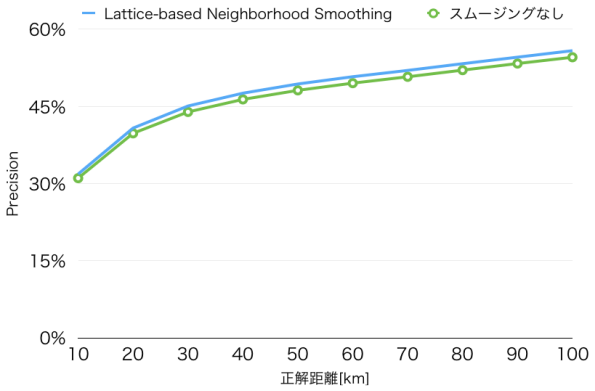


図 14　スムージングによる影響 – 適合率

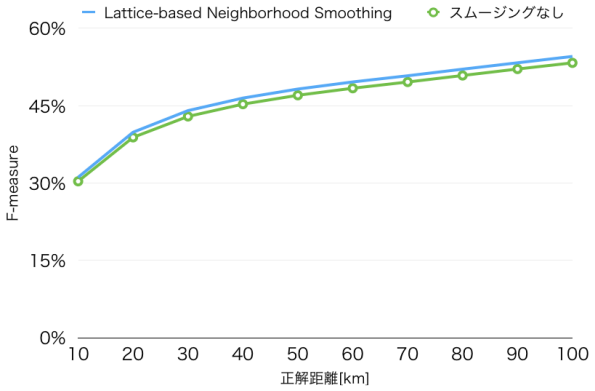


図 15　スムージングによる影響 – F 値

行わなかった場合の結果と比較することで、その有効性を検証する。

正解度、適合率、F 値を比較した結果が図 13、図 14、図 15 である。正解度、適合率、F 値のいずれについても、すべての距離において、スムージングを行った場合の推定精度が約 1%向上するという結果になったことから、居住地推定のためのスムージング手法が投稿位置推定にも有効であることが示唆される。

### 5.7　エリアの粒度と性能の関係

これまでの実験では、市区町村レベルのエリアデータを使って学習と推定を行った。取得したエリアデータは町丁レベルでエリアを表すデータであったため、町丁レベルのエリアデータにおいても学習と推定を行い、地区町村レベルの結果と比較する。

推定可能割合を比較した結果が表 4 である。フィルタした単語の割合は約 81%とほとんど同じであるが、推定可能割合では町丁レベルが市区町村レベルより高いという結果になった。これは式 2 において、TF の値がエリア数に依存しないことが理由として考えられる。町丁レベルではエリア数が増加し、IAF の重みが小さくなったことにより、TF の値を大きく評価した結果になったといえる。

正解距離を 10km から 100km まで変化させた場合の正解度、適合率、F 値を比較した結果が図 16、図 17、図 18 である。全ての結果において、市区町村レベルのほうが高精度で位置を推定できるという結果になった。これはエリア数が増加したことによって、各エリアにおけるツイートの分布がよりまばらに

表 4　エリアの粒度による影響 – 推定可能割合

| エリアレベル | フィルタした単語の割合 | 推定可能割合 |
|---|---|---|
| 町丁 (丁目、字など) | 80.70% | 99.90% |
| 市区町村 | 80.73% | 97.71% |

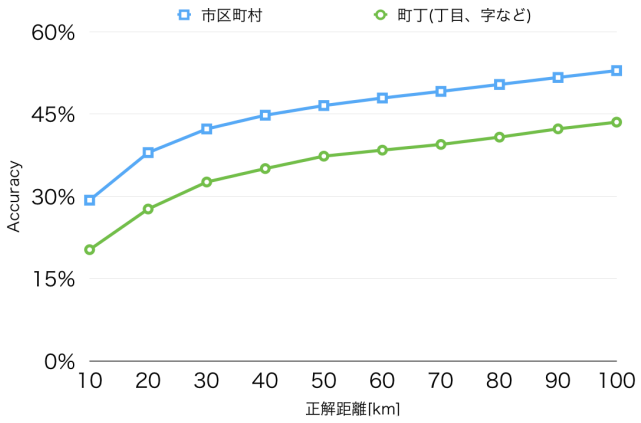


図 16　エリアの粒度による影響 – 正解度

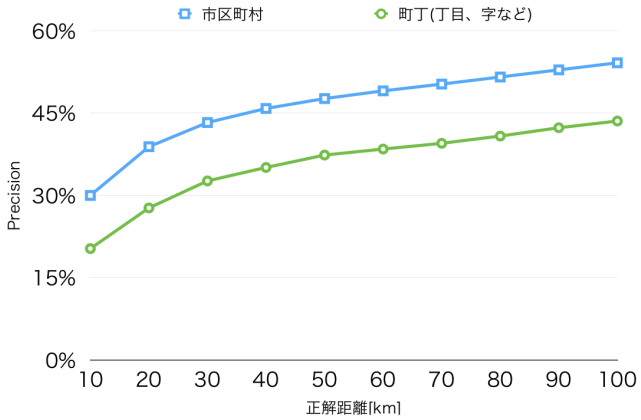


図 17　エリアの粒度による影響 – 適合率

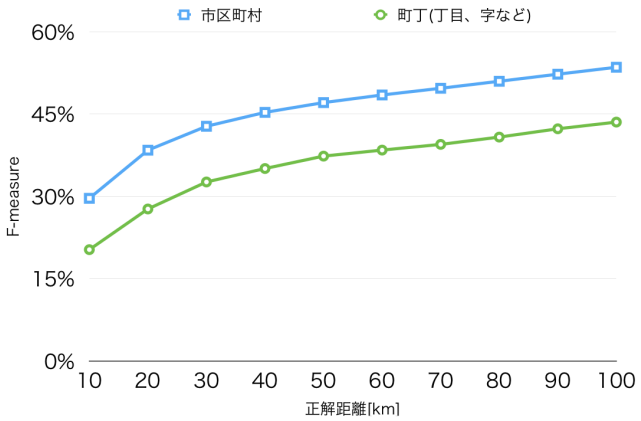


図 18　エリアの粒度による影響 – F 値

なってしまったためであると考えられる。学習データを増やし観測誤差による影響を少なくすることができれば、町丁レベルでの推定精度向上に繋がる可能性がある。

## 6. おわりに

本研究では、単語の出現頻度を学習し、1 件のツイート内容のみから、投稿位置の推定を行った。さらに、ツイート中からノイズとなる単語を取り除く TF–IAF フィルタを提案した。TF–IAF フィルタは、出現頻度の高い単語を大きく評価することで、多くの単語を除きすぎるという問題を克服する地域語フィルタである。TF–IAF フィルタを用いて推定を行った結果として、位置を推定できないツイートが減少し、推定精度 (再現率) が向上した。

今回の実験では、地域語フィルタが地域語を特定したために推定精度が向上したのではなく、ツイート中からノイズとなる単語を取り除くことができたために推定精度が向上したと考えられる結果となった。そのため、地域語の特定を行うという観点ではなく、ノイズを除くという観点で単語のフィルタを行うことが推定精度を向上させる可能性がある。

今後の課題として、1 件のツイート内容のみではなく、ユーザの情報やツイート間の時間的関係を考慮した推定を行うことが挙げられる。使用できる情報が増えることで、本質的に推定を行うことが困難なツイートや単語数が少ないツイートに対しても、より正確な投稿位置推定を行えると考えられる。

## 文　献


[1] 榊剛史, 松尾豊. ソーシャルセンサとしての Twitter –ソーシャルセンサは物理センサを凌駕するか?–. 人工知能学会誌, Vol. 27, No. 1, pp. 67–74, January 2012.

[2] 吉次由美. 東日本大震災に見る大災害時のソーシャルメディアの役割–ツイッターを中心に–. 放送研究と調査, Vol. 61, No. 7, pp. 16–23, July 2011.

[3] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pp. 759–768, Toronto, ON, Canada, October 2010.

[4] 橋本康弘, 岡瑞起. 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, July 2012.

[5] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter. In *Proceedings of the 2011 Conference on Emperical Methods in Natural Language Processing*, EMNLP '11, pp. 1568–1576, Edinburgh, Scotland, UK., July 2011.

[6] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pp. 851–860, Raleigh, North Carolina, USA, April 2010.

[7] 三木翔平, 新田直子, 馬場口登. 単語の地理的局所性の経時変化を考慮したツイートの発信位置推定. 第 6 回データ工学と情報マネジメントに関するフォーラム, DEIM '14, 兵庫県淡路市, March 2014.

[8] 北研二, 津田和彦, 獅々堀正幹. 索引語の抽出と重み付け. 情報検索アルゴリズム, pp. 33–40. 2002.

[9] 伊川洋平, 榎美紀, 立堀道昭. マイクロブログのメッセージを用いた発信場所推定. 第 4 回データ工学と情報マネジメントに関するフォーラム, DEIM '12, 兵庫県神戸市, March 2012.

[10] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pp. 61–70, Raleigh, North Carolina, USA, April 2010.