

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

三木 翔平[†] 新田 直子[†] 馬場口 登[†]

[†] 大阪大学大学院工学研究科 〒565-0871 大阪府吹田市山田丘 2-1

E-mail: [†]miki@nanase.comm.eng.osaka-u.ac.jp, ^{††}{naoko,babaguchi}@comm.eng.osaka-u.ac.jp

あらまし 本研究では、マイクロブログの代表である Twitter に投稿される各ツイートに対し、ツイートに含まれる単語の地理的局所性に基づき発信位置を推定する手法を提案する。提案手法ではまず、発信位置の緯度・経度を表すジオタグが付与された少数のツイートから単語の地理的局所性を解析し、特定の地域から発信される地理的局所性の高い単語であるローカル語と、ローカル語が示す位置情報を対にして学習する。提案手法は特に、地理的局所性が経時的に変化する単語が存在することに着目し、短期間に投稿されたジオタグ付きツイートからの学習を繰り返す逐次学習により、ローカル語の追加およびローカル語が示す位置情報の更新を行い、発信位置推定精度の向上を目指す。キーワード ローカル語, マイクロブログ, 位置推定, 経時変化, 逐次学習

1. はじめに

近年、人々が各地で観測した情報を観測時間、場所情報と共にソーシャルメディアに投稿することから、実世界に関する様々な情報をソーシャルメディアから獲得する研究が進められている。なかでも Twitter [1] はリアルタイム性の高い情報源として注目されている。Sakaki ら [2] は「地震」や「揺れ」などの特定のキーワードを含む、緯度・経度で表されるジオタグと呼ばれる位置情報付きの投稿（ツイート）の追跡により地震の発生をリアルタイムに検出し、震源地を推定した。Lee ら [3] は注目した領域のツイート数やユーザ数の急激な変化に基づき局所的に人が集中するイベントを検出した。しかし、ツイートにジオタグを付与するユーザはまだ非常に少なく、全体のツイートの 1% にも満たない。そのため、検出可能なイベントも地震など観測情報が豊富なものに限定される。

この問題に対し、ジオタグを付与していないユーザが発信したツイートの内容から位置を推定する手法が提案されている。Paradesi [4] は Gazetteer と呼ばれる地名辞書を用いて、ツイート中に含まれる地名をもとにツイートと地理空間上の位置を対応付けた。Cheng ら [5] は、ツイート中には地名の略称や、地名以外にも特定の場所と潜在的に関連のあるローカル語が含まれるという考えに基づき、ジオタグ付きのツイートからローカル語を学習した。Cheng らの手法ではまず、長期間にわたって収集した大量のジオタグ付きツイートから、ツイートに含まれる各単語の地理的分布を推定する。単語のうち Gazetteer に記載された地名に関する語は、ローカル語を判定する学習器を構築するために用いられ、Gazetteer に記載されていない単語の中から学習器によりローカル語と判定された語のみを抽出する。これにより、例えばアメリカの Houston 周辺と関連のローカル語として “rockets” が得られたと述べている。そしてジオタグを付与していないユーザが発信した一連のツイートにローカル語が含まれる場合、そのローカル語の地理的分布をもとにユーザの居住地を都市レベルで推定する。Cheng らは、1 ユーザあ

たり 100 ツイートもあれば比較的高い精度での推定が可能であると主張している。Chang ら [6] は、Cheng らと同様に単語の地理的分布を推定することによりユーザの位置推定を行っているが、推定した各単語の地理的分布と一般的な単語であるストップワードの地理的分布とを比較することにより、Gazetteer を用いることなくローカル語を決定した。これらの研究では、ローカル語は地名など常に同じ位置を表すことを前提としている。

ツイートには文字数の制限があり、位置推定の手がかりが少ないことから、単一のツイートに対して位置推定をする場合にはより多くのローカル語が必要となる。ローカル語として利用可能なものには、施設や建物名など常に同じ場所を表す語の他に、一時的に特定の位置を表したり、表す位置が経時的に変化するイベントに関する語などが挙げられる。このようなローカル語は長期間では地理的局所性を示さないが、短期間のツイート集合において局所性を示すと考えられる。よって短い時区間ごとにローカル語を学習する必要があるが、ツイート中の全単語について期間ごとに何度も地理的分布を推定することは現実的ではない。そこで本研究では、短期間に収集したジオタグ付きツイートから、全単語の地理的分布を推定しない簡便な方法によるローカル語の抽出を繰り返すことにより、ローカル語の経時変化を考慮した単一のツイートに対する位置推定を実現する。提案手法では実世界の地理空間を複数のエリアに分割し、エリアごとに投稿されたツイートの内容を比較することにより、あるエリアで頻出し、かつ他のエリアでは稀にしか出現しない単語をローカル語とし、ローカル語が発信された場所のジオタグとともに抽出する。このとき、ローカル語が出現したエリアを各期間で比較し、その変化に応じて各ジオタグに対する重み付けを行い、ローカル語の地理的分布の経時的変化に対応する。位置推定の際には、ツイートに含まれるローカル語の重み付きジオタグ集合から、カーネル密度推定を用いてツイートの地理的分布を推定する。

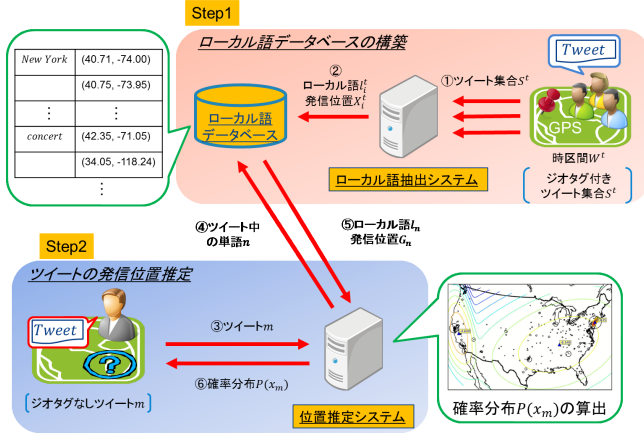


図 1 提案手法の概要

2. 提案手法

提案手法は、ジオタグが付与されていないツイート t が発信された位置の緯度・経度 $x = (lat, lon)$ の確率分布 $P_t(x)$ の推定を目的とする。 $P_t(x)$ を推定するには、一定の期間 W の間に投稿されたジオタグ付きのツイート集合 S から抽出したローカル語 l_k ($k = 1, \dots$) およびローカル語が発信された位置の重み付きジオタグ集合 G_k からなるデータベースを用いる。提案手法は図 1 に示すように、Step1) ローカル語データベースの構築および Step2) ツイートの発信位置推定の 2 つのステップにより構成される。

Step1) では、Twitter に投稿されたジオタグ付きツイートの解析によりローカル語のデータベースを構築する。ジオタグ付きのツイート集合 S から、ローカル語 l_k および各ローカル語を含むツイートが発信された位置を示すジオタグ $x_{k,n}$ ($n = 1, \dots, N_k$) をそれぞれ抽出する。各ローカル語が示す位置の経時変化を考慮するため、ジオタグ $x_{k,n}$ に対してローカル語が表す位置の変化率に応じた重み $\omega_{k,n}$ ($0 \leq \omega \leq 1$) を付与した重み付きジオタグ集合 $G_k = (\omega_{k,n}, x_{k,n})$ をローカル語 l_k とともにデータベースへと格納する。この処理を W ごとに逐次的に繰り返すことによりデータベースを更新する。

Step2) では、ジオタグが付与されていないツイートの発信された位置を推定する。ツイート t にローカル語 l_m ($m = 1, \dots, M$) が含まれるとき、 l_m の発信位置の重み付きジオタグ集合 $\bigcup_{m=1}^M G_m$ から、カーネル密度推定を用いてツイートの発信源の確率分布 $P_t(x)$ を推定する。次節より各ステップの詳細を述べる。

2.1 ローカル語データベースの構築

ローカル語は地名をはじめとする位置に関する語であり、主に名詞から構成されると考えられる。また、任意のツイートはユーザによりさまざまな場所から発信されるのに対し、ローカル語を含むツイートは局所的に特定の場所で集中して発信され、なおかつ他の場所では稀にしか出現しないはずである。以上を踏まえて提案手法ではまず、実世界の地理空間を複数のエリアに分割し、各エリアから発信されたツイートに含まれる名詞を抽出する。次に、抽出した名詞をエリア間で比較することによ

り、単語ごとに地理的分布を算出することなくローカル語を決定し、各ローカル語が発信された位置のジオタグ集合とともに抽出する。この処理を時区間ごとに繰り返し、新しく得られたローカル語を追加すると同時に、ジオタグ集合に対して逐次重み付けをすることによりデータベースを更新する。

以降でその各処理について詳しく述べる。

2.1.1 前処理

ツイートは人間の実世界観測に基づいて発信されており、人の多い場所ほど発信されるツイートの数も多い。提案手法では地理空間上をエリアに分割するが、エリアごとのツイート数が異なると各エリアの名詞の出現回数にばらつきが生じ、正確にローカル語の抽出ができない。そこで、四分木アルゴリズムを用いて地理空間上を各エリアのツイート数になるべく均等になるように分割する。四分木アルゴリズムとは 2 次元の空間を同じ大きさの 4 つの象限に再帰的に分割していくものであり、発信されるツイート数が多い地域ほど細かいエリアに分割される。こうして得られたエリアを a_j ($j = 1, \dots, J$) とする。

次に、ローカル語の候補となる名詞のみをツイートから抽出するために、ツイートに対して形態素解析を行う。また、TermExtract [7] と呼ばれるキーワード自動抽出システムを用いることにより、複数の単語からなる複合名詞を生成する。これは、例えば“*Brooklyn*”という単語が都市名であるのに対し、“*Brooklyn Museum*”は施設名であるように、ローカル語を複合名詞として抽出することにより示す場所や意味が限定されたと考えられるためである。以下では名詞・複合名詞を区別せず、名詞と呼ぶものとする。

2.1.2 ローカル語の抽出

前処理により得られた J 個のエリア a_j 内で発信された名詞集合を比較することにより、各エリアに固有な名詞をローカル語として抽出する。本研究では、出現頻度の低い施設や建物名に関する語や、複数の場所と関連のある語などもローカル語の対象とするため、以下の式により名詞 u_i がエリア a_j にどの程度固有であるかを表すスコア $Score_{i,j}$ を算出する。

$$Score_{i,j} = n_{i,j} * idf_i \quad (1)$$

$$idf_i = \log \frac{J}{d_i} \quad (2)$$

ただし、 $n_{i,j}$ は a_j における名詞 u_i の出現回数、 d_i は u_i を含むエリア数である。 $n_{i,j}$ は a_j 内での u_i の出現回数が多ければ大きくなるのに対し、 idf_i は逆文書頻度と呼ばれ、 u_i が多くのエリアで出現する場合には低く、特定のエリアにしか出現しない場合には高くなるような、一般語フィルタとしての役割を持つ。 $Score_{i,j} \geq Th$ となる u_i をローカル語 l_k と決定し、 l_k を含むツイートが発信された位置を示すジオタグ $x_{k,n}$ とともに抽出する。ここで、 $Score_{i,j}$ は a_j 内の他の名詞 $u_{i'} (\neq u_i)$ の出現回数 $n_{i',j}$ に依らず決定されるため、ローカル語となるための閾値は $d_i = 1$ の時の出現回数 $n_{i,j}$ の下限値 θ を設定することにより決定される。また、複数のエリアと関連のある名詞もエリア数に応じた出現回数の下限値を超えればローカル語として抽出される。

2.1.3 データベースの更新

ローカル語データベースは時区間 W において抽出されたローカル語およびそのジオタグ集合を繰り返し格納することにより更新する．新たに抽出されたローカル語の追加により、ツイート中に頻出する地名だけでなく、施設や建物名のように出現頻度の低いローカル語も学習される．また、抽出したローカル語の中にはイベントに関する語のように時区間 W ごとに地理的分布が変化する場合がある．このようなローカル語を位置推定に用いる際には、経時変化を考慮して最適な地理的分布を推定する必要がある．以上を踏まえ提案手法では、時区間ごとに新たに抽出されたローカル語およびそのジオタグ集合をデータベースに追加すると同時に、過去の時区間のジオタグ集合に対して重み付けを行うことによりデータベースを更新する．

まず、過去の時区間のジオタグ集合に対する重み付けについて述べる．時刻 τ において抽出されたローカル語集合を L' 、時刻 τ においてデータベースに格納されているローカル語集合を L とする． $l_k \in L' (k = 1, \dots)$ が L に含まれる場合には、データベース内のジオタグ集合に対する重みが増加される．このとき、ローカル語が常に同じエリアから抽出される限り、どの時区間に抽出されたジオタグ集合も位置推定の際には同等に信頼できるが、ローカル語が時区間ごとに異なるエリアから抽出される場合、ジオタグ集合は過去のものであるほど信頼度は低いと考えられる．そのため、更新前の重みを $\omega_{k,n}^\tau$ 、更新後の重みを $\omega_{k,n}^{\tau+1}$ としたとき、 $\omega_{k,n}$ をエリアの変化率 γ_k を用いて以下の式で更新する．

$$\omega_{k,n}^{\tau+1} = \omega_{k,n}^\tau - \gamma_k \quad (3)$$

ただし、 $\omega_{k,n}^{\tau+1} < 0$ のとき $\omega_{k,n}^{\tau+1} = 0$ とする．また、 l_k が常に同じエリアから抽出されるとき $\gamma_k = 0$ 、 l_k が全く異なるエリアから抽出されるとき $\gamma_k = 1$ となればよい．そのため、 γ_k を l_k が異なる時区間において抽出されたエリアの類似度 s_k に基づき以下の式で決定する．

$$\gamma_k = 1 - \sqrt{2s_k - s_k^2} \quad (4)$$

$$s_k = \frac{A_k \cap A'_k}{A_k \cup A'_k} \quad (5)$$

ただし、 A_k は l_k が更新前に抽出されたエリア集合であり、 A'_k は l_k が S から抽出された際のエリア集合である．

A_k は以下の式で更新する．

$$A_k^{\tau+1} = \begin{cases} A_k^\tau \cup A'_k & (\text{if } A_k^\tau \cap A'_k \neq \emptyset) \\ A'_k & (\text{otherwise}) \end{cases} \quad (6)$$

さらに、突発的に出現するローカル語はデータベースに残すと位置推定精度を低下させる要因となるため、更新時に除去する必要がある． L に格納されているローカル語のうち、 S において λ 個以上のエリアで出現し、ローカル語と判定されなかった l_k については、更新前の重みに依らず $\omega_{k,n}^{\tau+1} = 0$ とする．

最後に、 L に含まれていない新たに抽出されたローカル語を以下の式の通りデータベースに追加する．

$$L^{\tau+1} = L^\tau \cup L' \quad (7)$$

このとき、 $l_k \in L' (k = 1, \dots)$ を含むツイートが発信された位置を示すジオタグ $x_{k,n} (n = 1, \dots, N_k)$ は、重みの初期値を $\omega_{k,n} = 1$ とする．こうして更新したデータベースを用いてツイートの発信位置を推定する．次節でツイートの推定手法について述べる．

2.2 ツイートの発信位置推定

構築したローカル語のデータベースをもとにジオタグが付与されていないツイートの発信位置を推定する．まず、位置推定の対象となるツイート t が M 個のローカル語 $l_m (m = 1, \dots, M)$ を含むとき、全ローカル語の発信位置の重み付きジオタグ集合 $\bigcup_{m=1}^M G_m$ から、カーネル密度推定を用いてツイートの発信源の確率分布 $P_t(x)$ を推定する．カーネル密度推定とは標本データから確率密度関数を連続的に推定するための手法であり、データの密度が高い点を複数点求めることができる．地点 x におけるカーネル密度推定の推定値は以下のように求められる．

$$P_t(x) = \frac{1}{h\Omega} \sum_{m=1}^M \sum_{n=1}^{N_m} \omega_{m,n} * K\left(\frac{x - x_{m,n}}{h}\right) \quad (8)$$

$$\Omega = \sum_{m=1}^M \sum_{n=1}^{N_m} \omega_{m,n} \quad (9)$$

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad (10)$$

ただし、 N_m は l_m のジオタグ数、 h はバンド幅、 $K(\cdot)$ はカーネル関数であり、実験ではガウス関数を用いた．また、バンド幅は Silverman [8] が提案した以下の式で求める．

$$h = 1.06\hat{\sigma}(M * N_m)^{-1/5} \quad (11)$$

ただし $\hat{\sigma}$ はジオタグ集合の標準偏差である．

3. 評価実験

提案手法の有効性を以下の2つの観点から検証する．

- (1) 抽出されたローカル語の妥当性
- (2) データベースの更新の有用性

以下ではまず実験に用いたデータセットについて説明し、次にそれぞれの実験結果について考察する．

3.1 データセット

Twitter の Streaming API を用いて 2013 年に、緯度が 24 度から 49 度、経度が -125 度から -66 度、すなわちアメリカ本土から発信されたジオタグ付きツイートを収集し、そのうちのべ 30 日間のツイート 62,779,944 件を実験に用いた．各時区間の間隔 W を 24 時間とし、各日に発信されたツイートのうち、ランダムに選択した 0.2% を位置推定のテストデータとし、残りをローカル語のデータベースを構築するための学習データとした．また、前処理において四分木アルゴリズムを適用した結果、エリア数 $J = 279$ となった．

提案手法の有効性は位置推定の推定効率、推定精度および平均誤差に基づき評価する．推定効率とはテストデータのうちローカル語を含むツイート T の数を表し、推定精度および平均誤差は以下に示すように Accuracy (ACC) および Average Error Distance (AED) で評価する．

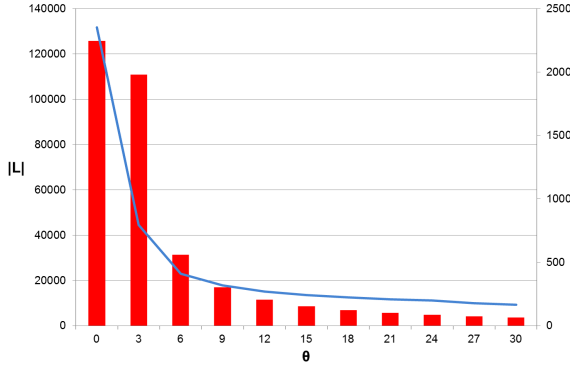


図 2 θ と $|L|$ および $|T|$ の関係 (DB_1)

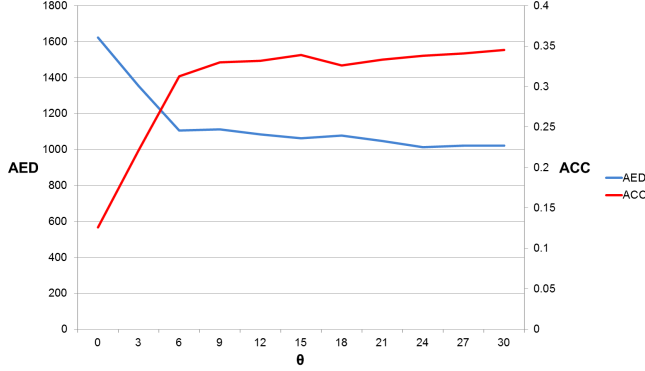


図 3 θ と ACC および AED の関係 (DB_1)

$$ACC = \frac{|\{t|t \in T \wedge ErrDist(t) \leq 160[km]\}|}{|T|} \quad (12)$$

$$AED = \frac{\sum_{t \in T} ErrDist(t)}{|T|} \quad (13)$$

$$ErrDist(t) = d(x_{act}(t), x_{est}(t)) \quad (14)$$

ただし, $x_{act}(t)$ は実際にツイート t が発信された位置, $x_{est}(t)$ は t が発信された推定位置であり, 本実験では $x_{est}(t) = \arg \max_x P_t(x)$ とした. また, $ErrDist(t)$ は推定結果の誤差 [km] を表し, $ErrDist(t) \leq 160[km]$ のときツイート t は正しく位置推定が出来たものとする.

3.2 ローカル語抽出の妥当性

提案手法により短期間のツイート集合から抽出されたローカル語を用いて, 位置推定の効果を検証する. 長期間のツイート集合から抽出した場合のローカル語と位置推定結果を比較するため, 本実験では 10 日目までの全 21,494,710 件のツイートから構築したデータベース DB_1 および, 10 日目の 2,280,012 件のツイートから構築したデータベース DB_2 を用意し, 10 日目のテストデータ 4,569 件に対して位置推定を行った.

図 2 は, θ を変化させた際の DB_1 に含まれるローカル語数 $|L|$ および推定効率 $|T|$ を示す. $\theta = 0$ はツイート集合に含まれるすべての名詞をローカル語と判定することを意味する. $\theta = 3$ のとき, $|L|$ がわずかに減少したのに対して $|T|$ が大幅に減少した. これは, 多くのツイートが表 3 に示すような少数の一般語のみを含むためであり, 一般語が低い閾値により除去されたと考えられる. 図 3 は θ を変化させた際の DB_1 の ACC および AED を示す. $\theta = 6$ のとき ACC , AED はともに大幅に改

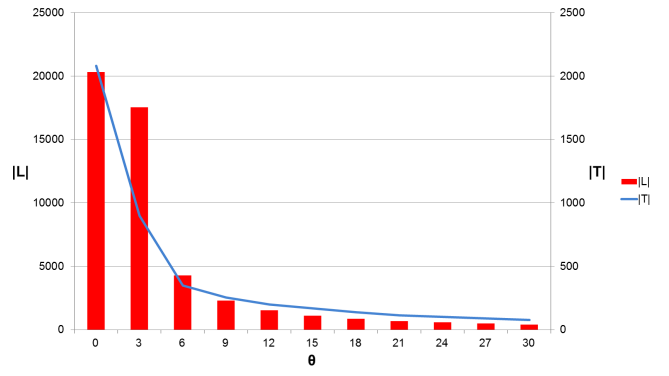


図 4 θ と $|L|$ および $|T|$ の関係 (DB_2)

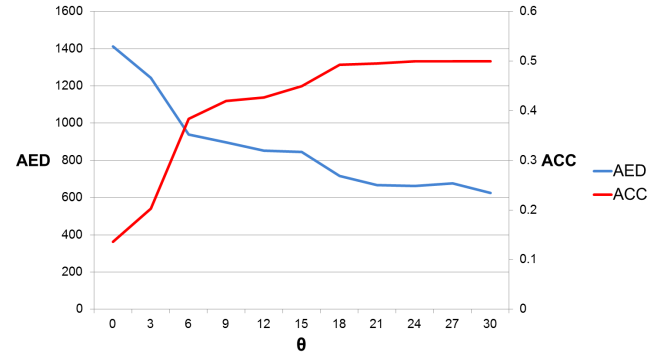


図 5 θ と ACC および AED の関係 (DB_2)

表 3 ローカル語を含まないツイート例

That was a pretty damn good movie
Missing my love bag!
Well time to sleep.

善された. さらに閾値を上げることにより ACC や AED はわずかに向上したが, それにともなって $|T|$ も減少しており, 両者はトレードオフの関係にある.

図 4 は θ を変化させた際の DB_2 に含まれるローカル語数 $|L|$ および $|T|$ の関係を示す. DB_1 と比較すると, ツイート数がおよそ 10 分の 1 になったことにより $|L|$ は全体的に大幅に減少した. 短期間で学習可能なローカル語の語量は限られるが, 長期間にわたり逐次的にツイートを収集することにより多くのローカル語を学習することが可能であると考えられる. 一方, $|L|$ が大幅に減少したにも関わらず, DB_1 と DB_2 では $|T|$ はほとんど変化はなかった. これより, 位置推定に必要なローカル語の多くは 1 日分のツイートからも抽出可能であると言える. 図 5 は 1 日分のツイート集合からローカル語を抽出した場合の θ と ACC および AED の関係を示す. DB_1 と同様, $\theta = 6$ のときに ACC および AED はともに大幅に改善された. また, 同じ θ の値でも DB_1 と比較すると全体的に ACC や AED は向上した. これは, 同じ閾値でもツイート集合が小さければ小さいほどローカル語と決定する際の出現回数の条件が厳しくなるためであり, より地理的局所性の高いローカル語が抽出されたと考えられる. 特に, $\theta = 24$ のとき $ACC = 0.5$, $AED = 662[km]$ となり, 推定したツイートのうち半数にあ

表 1 各 DB で正しく位置推定できたツイート例およびその推定誤差 (太字はローカル語を示す)

DB_1 でのみ正しく推定できたツイート例	$ErrDist$ [km]
Pit stop before our walk on the seven mile bridge . w/ @wanderthemap (@ Mrs. Mac's Kitchen)	84.56
Delaware is kinda cool	4.79
@YasonoJ yaso please come to banana bay tonight	60.07
DB_2 でのみ正しく推定できたツイート例	
I-75 North Bound! Go Cats! #CSW	72.05
Got tickets to see Luke Bryan with the best roommates!	89.72
Duke game with my favorite little boy! @ Wallace Wade Stadium	3.95
DB_1, DB_2 どちらも正しく推定できたツイート例	(DB_1/DB_2)
Happy Halloween! @dnbnikki # disney # disneyland	4.47/1.04
Manhattan here we go	36.08/39.57
One thing I love about living in Arizona , I can tan until December.	38.60/44.35

表 2 ローカル語を含むが正しく位置推定できなかったツイート例およびその推定誤差 (太字はローカル語を示す)

DB_1 でのみ推定可能だが, 正しく推定できなかったツイート例	$ErrDist$ [km]
Waited 13 minutes at Dunkin for my donut.	239.68
I talked about Harry Potter for a solid 2 hours last night.	1211.94
@HippieeeeLoveee: I would be ok with moving into a teepee in the mountains .	2033.51
DB_2 でのみ推定可能だが, 正しく推定できなかったツイート例	
Breaking the boots out for the first time this September .	287.14
Menzel told the fans in the stadium to be quiet and they did...	690.64
Start to Oktoberfest!! (@ Ore House Restaurant w/ @johnsonianb)	2544.03
DB_1, DB_2 どちらも推定可能だが, どちらも正しく推定できなかったツイート例	(DB_1/DB_2)
All packed up and on our way to Dublin , VA to see the @oakridgeboys :-)	496.26/523.20
@heidimo6 yay!! That is excellent news!! Where in Ohio do u live?	1086.49/920.24
Maybe I'll go to LAX when I go to Wisconsin . @BrannanHudson14	3318.70/3196.31

る 51 件のツイートの発信位置が正しく推定された。この結果は、先行研究である Cheng ら [5] がユーザに対して行った位置推定の推定精度 ($ACC = 0.51$) とほぼ同じであり、平均誤差 ($AED = 857[km]$) に関しては Cheng らを上回った。Cheng らの手法では 5ヶ月という長期間にわたるツイート集合から学習した 3,183 語のローカル語を用いて、1 ユーザあたり 1,000 件分のツイートからユーザの位置を推定するが、提案手法では 1 日という短期間で学習したローカル語を用いて単一のツイートに対して高い精度で位置推定が可能である。

表 1 は $\theta = 6$ のときにそれぞれの DB で正しく位置推定ができたツイート例を推定誤差とともに示す。上の 3 件は出現頻度の低い地名に関するローカル語を含むため、 DB_2 ではローカル語として抽出されずに DB_1 でのみ正しく位置推定ができた。次の 3 件は地理的分布が経時的に変化するイベントに関するローカル語と考えられ、 DB_1 では抽出されずに DB_2 でのみ正しく位置推定ができた。下の 3 件は常に同じ位置を示し、なおかつ出現頻度の高い地名に関するローカル語を含むため、どちらの DB でも正しく位置推定ができた。

しかし、テストデータの中にはローカル語を含むが正しく位置推定が出来なかったツイートも存在する。表 2 は DB_1, DB_2 それぞれにおいて正しく位置推定が出来なかったツイートの例である。上の 3 件は DB_1 でのみ位置推定が可能だが、推定結果が正しくなかったツイート例である。これらのツイートが含

むローカル語は、地理的局所性はそれほど高くないが出現回数が極めて多いためにローカル語と判断されたことや、同じ地名でも複数の位置を示す語であることが推定位置を誤った要因と考えられる。次の 3 件は DB_2 でのみ位置推定が可能だが、推定結果が正しくなかったツイート例である。短期間のツイート集合から抽出したローカル語は DB_1 と比較して高い地理的局所性を示すはずだが、依然として出現回数が極めて多い語や、複数の位置を示すイベントに関する語も含まれた。下の 3 件はどちらの DB でも位置推定が可能だったが、どちらも推定結果が正しくなかったツイート例を示す。これらは、ツイートの発信位置と異なる場所について言及しており、ツイートの文脈を考慮した高度な自然言語処理を必要とする。

3.3 データベースの更新の有効性

データベースの更新による推定効率、推定精度および推定誤差への影響を考察する。本実験では、以下の 3 つのローカル語データベース DB_1, DB_2, DB_3 を用意し、30 日分のテストデータに対して位置推定を行う。

DB_1 : 最初の 10 日間のツイートを 1 つの集合とみなして構築したデータベース

DB_2 : 常に最新の 1 日分のツイートのみから構築するデータベース

DB_3 : 1 日ごとに更新を繰り返しながら構築するデータベース
 DB_1 は長期間で一度だけローカル語を学習したものであり、前

表 4 各 DB を用いた位置推定結果 (30 日間平均)

DB	DB ₁	DB ₂	DB ₃
T	340.733	246.766	260.700
ACC	0.359	0.381	0.397
正解ツイート数	122.733	94.433	103.366
AED	993	951	928

節と同じデータベースを用いる．このデータベースの構築手法は先行研究の考え方に基き、データベースが更新されることはない．DB₂ および DB₃ は短期間でのローカル語の学習を繰り返す点は共通するが、DB₂ は時区間ごとに新たにデータベースを再構築するのに対し、DB₃ はデータベースに新たなローカル語を追加しながら重み付きのジオタグ集合を更新する．ただし、前節の実験結果より推定効率と ACC、AED のトレードオフを考慮して $\theta = 6$ とし、DB₃ を更新する際の閾値は $\lambda = 5$ とした．

図 6 に各 DB に含まれるローカル語数 $|L|$ の推移を示す．DB₁ に含まれるローカル語は長期間のツイート集合から抽出されるため、地名や施設名などの常に同じ位置を表す単語から構成されると考えられる．DB₁ は更新しないため $|L|$ は一定であるが、10 日間のツイート集合から一括でローカル語を抽出するため他の DB と比較して $|L|$ は大きい．DB₂ は日別のツイート数や名詞数のばらつきにより多少の変動はあるが、常に 3,000 語前後のローカル語が抽出された．DB₃ は過去の不要なローカル語を除去しつつ、新たに抽出されたローカル語を追加するため、結果として $|L|$ は徐々に増加した．30 日経過した時点で DB₁ には及ばないが、さらに更新を繰り返すことにより DB₁ を上回る可能性がある．図 7 は各 DB の $|T|$ の推移を示す．DB₁ は $|L|$ が最も大きいことから、すべての日において DB₂ および DB₃ を上回った．また、DB₃ は常に DB₂ をわずかに上回る結果となったが、更新を繰り返しても $|L|$ ほど大きな差は見られなかった．図 8 および図 9 は各 DB の ACC および AED の推移を示す．DB₁ は日別のばらつきが大きく安定した推定精度が得られなかったが、データベースを更新しないため日数の経過とともに徐々に推定精度は低下した．DB₃ は DB₂ と比較すると ACC はわずかに上回り、AED も同程度もしくはわずかに改善された．

表 4 は各 DB を用いた位置推定結果の 30 日間の平均をまとめたものである．DB₁ は豊富なローカル語を利用して他の DB よりも多くのツイートに対して位置推定が可能であるため $|T|$ は最も高いが、ACC や AED は最も低い結果となった．DB₂ は短期間でローカル語を学習するためローカル語数 $|L|$ は最も少なく、それに伴い $|T|$ も最も低い．しかし、ローカル語の経時変化を考慮するため ACC および AED は DB₁ よりも良い結果が得られた．DB₃ は ACC および AED の値は最も良い結果を示し、DB₂ よりも多くのツイートの位置を正しく推定することができた．DB₁ と比較すると正解ツイート数では及ばないものの、DB₁ で正しく位置推定ができなかったツイートが含むローカル語の多くを除去したことにより、結果として ACC および AED が改善された．これは、ローカル語の示す

位置の経時変化を考慮してデータベースを更新した提案手法が推定効率および推定精度、平均誤差すべての観点から位置推定に有効であることを示す．

表 5 は DB₃ によって正しく位置推定されたツイート及びその推定誤差を示す．上の 5 件は DB₁ では正しく推定されなかったツイートであり、DB₃ では “thunder” のように地理的分布が経時的に変化するイベントに関すると思われるローカル語を抽出し、正しく地理的分布を推定した．下の 5 件は DB₂ では正しく推定されなかったツイートであり、DB₃ では “Pewaukee” のように小さな地名など出現頻度の低いローカル語の抽出が可能である．

最後に、データベースを 30 日間で更新したことによるローカル語の地理的分布の変化を考察する．ここではまず、データセットに用いた 30 日間のうち 14 日間でローカル語として抽出された “Florida Georgia Line” を例に挙げる．Florida Georgia Line とはアメリカで活動するミュージシャンであり、実験期間中のうち 19 日間はアメリカ国内でライブツアーを敢行したことが確認できた．図 10 は日別に “Florida Georgia Line” の発信源の確率分布が最も高かった点を実際にライブが行われた場所とともに地図上に表したものである．Florida や Georgia などの地名を含んでいるにも関わらず、毎日異なるエリアでローカル語として抽出された．ライブが行われた 19 日間のうち 14 日間はライブ会場付近のローカル語として抽出され、ライブが行われていない日にはローカル語として抽出されなかった．これより、提案手法では示す位置が経時的に変化するローカル語を抽出し、更新されたジオタグ集合から適切な地理的分布を推定できたとと言える．

もうひとつの例として、“rain” は 30 日間のうち 26 日間でローカル語として抽出された．図 11 はローカル語 “rain” をもとに正しく位置推定が出来たツイートを推定位置に配置したものである．雨が観測される場所は経時的に変化するが、提案手法ではその経時変化に対応してツイートの位置を正しく推定できた．気象に関するローカル語の他にも、交通・ライブ・スポーツチームの試合などに関するローカル語が各地で抽出されており、地名だけでなくイベントに関するローカル語を用いることにより多くのツイートに対して位置推定が可能である．

4. ま と め

本稿では、Twitter に投稿される各ツイートに対し、ツイートに含まれる位置を表す語の経時変化を考慮した発信位置推定手法を提案した．短期間に投稿されたジオタグ付きツイートからローカル語の学習を繰り返す逐次学習により、ローカル語の地理的分布の経時変化に対応すると同時にローカル語数を増加させることを試みた．抽出したローカル語の妥当性を検証する実験では、地理的局所性が特に高いローカル語を用いることにより単一のツイートに対しても 50% の高い精度で発信位置を正しく推定できることを確認した．また、データベース更新の有効性を検証した実験では、高い推定効率を維持しつつ、推定精度 39.7% および平均誤差 928km とともにデータベースを更新しない手法よりも良好な結果が得られた．これより、推定効率、

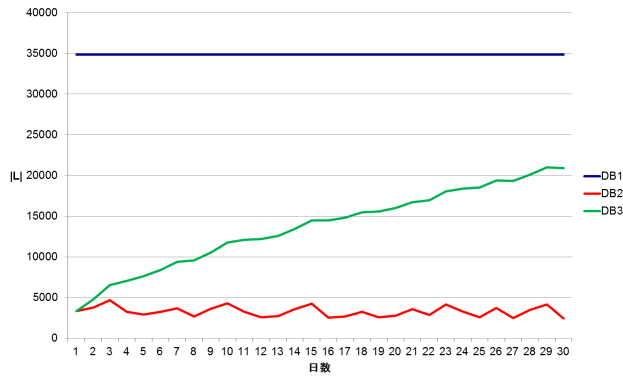


図 6 各 DB の日別の $|L|$ の推移

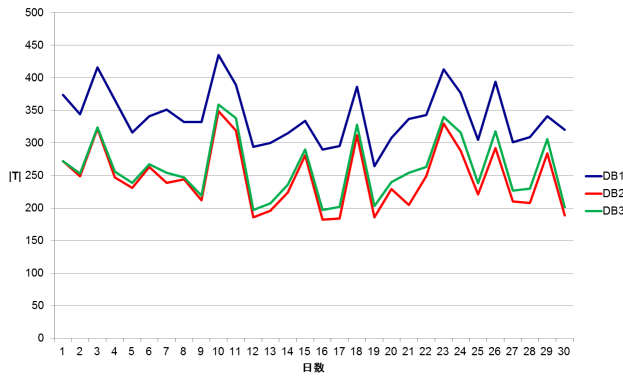


図 7 各 DB を用いた日別の $|T|$ の推移

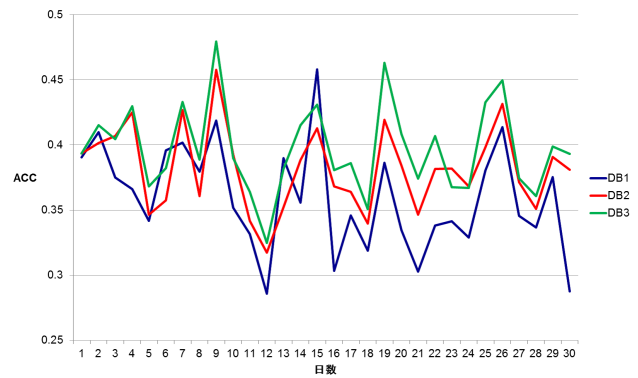


図 8 各 DB を用いた日別の ACC の推移

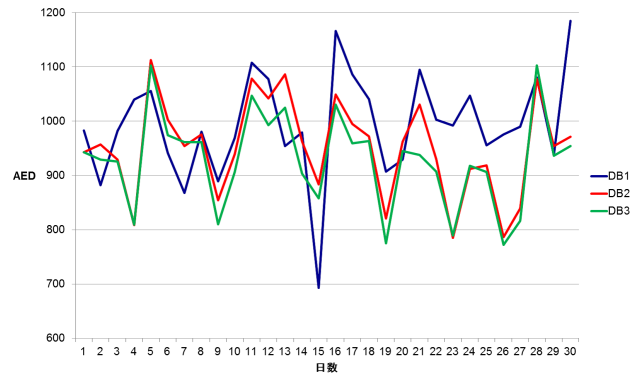


図 9 各 DB を用いた日別の AED の推移

表 5 DB_3 で正しく位置推定できたツイート及びその推定誤差 (太字はローカル語を示す)

DB_1 では推定不可	$ErrDist$ [km]
He likes the redskins	28.12
C'mon Cowboys! Lets take care of bidness	137.46
thunder while driving is a little scary	69.27
I just want to sit here with the rain and Justin Vernon all day #obligations	7.57
@WilliamSandman @muhfucka_jones Will is butthurt after the Eagles lost	159.47
DB_2 では推定不可	
Vacation @ Wellfleet , Cape Cod	6.19
@angelbabyy no ill be at ETSU . I told u yesterday when we were at the movies.	34.13
Oktoberfest with the fam #family #sunday #fun #beer @ Snowbird Ski	0
@aprilmaey_ I'm sorry cuh. I left all my basedness at sun city :/	7.87
@AbbyCB not in Pewaukee anymore!? Congrats, where did you find a place!?	3.14

推定精度および平均誤差のすべての観点からローカル語データベース更新の有効性を示した。今後の課題として、さらに長期間にわたる大規模なデータを用いた実験、および同一ユーザーの一連のツイートをもとにローカル語を含まないツイートの位置推定手法の検討が挙げられる。

文 献

- [1] "Twitter," <https://twitter.com/>.
- [2] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," *Proceedings of International World Wide Web Conference (WWW)*, pp. 851–860, 2010.
- [3] R. Lee, S. Wakamiya, and K. Sumiya, "Discovery of Unusual Regional Social Activities Using Geo-tagged Microblogs," *World Wide Web Spacial Issue on Mobile Services on the Web*, 14(4) pp. 321–349, 2011.
- [4] S. Paradesi, "Geotagging Tweets using Their Content," *Proceedings of International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pp. 355–356, 2011.
- [5] Z. Cheng, J. Caverlee, and K. Lee, "You are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users," *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 759–768, 2010.
- [6] H.-W. Chang, D. Lee, M. Eltaher, and J. Lee, "@Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage," *Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 111–118, 2012.
- [7] "TermExtract," <http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb.html>.
- [8] B. W. Silverman, "Density Estimation for Statistics and Data Analysis," *Chapman and Hall*, 1986.

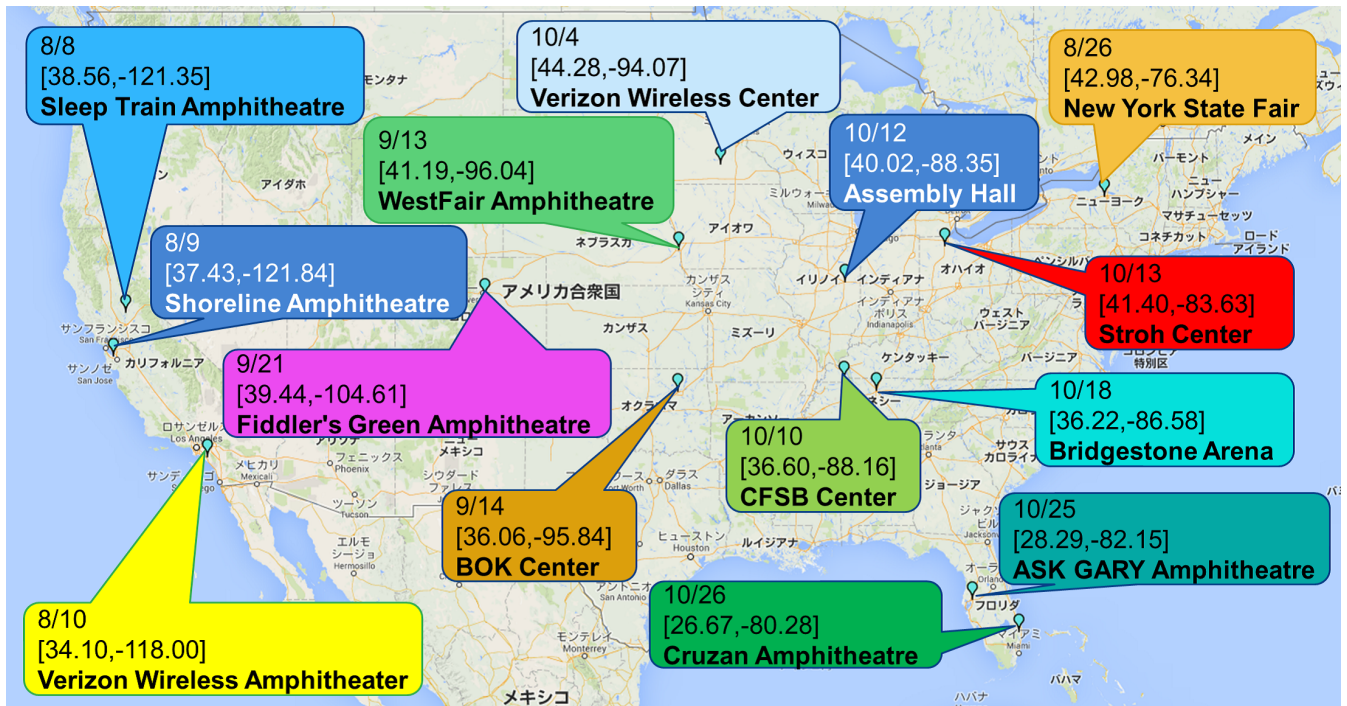


図 10 ローカル語 “Florida Georgia Line” が日別に示す位置と実際にライブが行われた場所

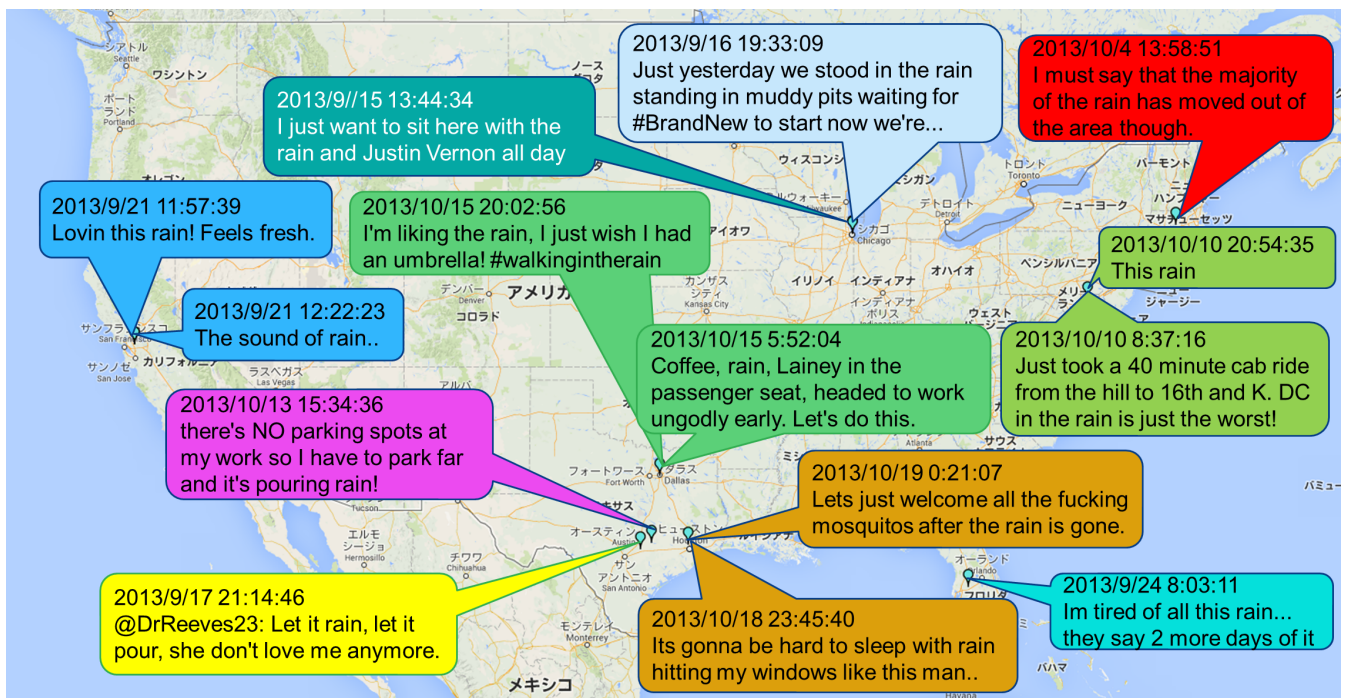


図 11 ローカル語 “rain” を用いて正しく位置推定が出来たツイートの内容と推定位置