Vol.2015-NL-223 No.6
2015/9/28

英文マイクロブログにおける 地域固有単語共起にもとづくユーザ位置推定

石田和成[†1]

英文マイクロブログにおける地域固有の単語共起にもとづき，情報発信者の位置推定を行った．単語共起は，単語の意味の多義性を表現するデータ構造として効果的であり，知識の体系化やコミュニティの抽出，スパムの同定などに用いられる．本研究に先立ち，日本国内で発信された日本語の位置情報付きマイクロブログを用いた情報発信者の位置推定を行い，単語共起による位置推定は地域固有の単語と比べ精度が高いことを確認した．本研究では，全世界で発信された英語の位置情報付きマイクロブログを用いた位置推定を行い，単語共起による位置推定の有効性を確認した．

User Location Estimation on English Microblogs with Area Specific Term Co-occurrence

KAZUNARI ISHIDA[†1]

This paper discusses user location estimation on English microblogs with area specific term co-occurrence. Term co-occurrence is a convenient data representation form to handle multiple meaning terms for various purposes such as organizing information, extracting communities, detecting spams, and so on. In previous paper, we employed geo-tagged Japanese microblogs published in Japan to apply user location estimation methods, and then we confirmed that user location estimation with term co-occurrence is more effective than that with area specific term. In this paper, we employed geo-tagged English microblogs published globally to apply the user location estimation methods, and then we also confirmed the effectiveness of user location estimation with term co-occurrence.

1. はじめに

英文マイクロブログにおける地域特有の単語共起にもとづき，情報発信者の位置推定を行い，地域トピックの考察を行う．スマートフォンの普及は，マイクロブログのデータ量急増をもたらした．しかし，ソーシャルメディアにおいて，位置情報の付加されたデータが占める割合は非常に少ない．地方都市は大都市と比較し，その傾向が顕著である．潜在的に存在する地方都市の情報を収集するには，所在地の不明なユーザの大まかな位置推定が必要である．位置推定を行うために，本研究に先立ち，日本国内で発信された日本語データに関して，地域特定スコアにもとづく情報発信者の位置推定手法を提案した．地域特定スコアは，位置情報付きデータに含まれる単語共起の出現頻度，平均緯度経度，標準偏差にもとづき定義した．単語共起は，多義的な単語の意味を表現するデータ構造として効果的であり，知識の体系化[1]やコミュニティの抽出[2]，スパムの同定[3]などに用いられる．本研究では，単語共起を用いた位置推定手法を，全世界の英文マイクロブログデータに適用し，その有効性について考察する．

2. 関連研究

マイクロブログにおける位置推定手法として以下のような研究がある．Dalvi ら[4]は，空間的なモデルを用いたオブジェクトとツイートのマッチングを行うために，ユーザとオブジェクトの距離のモデル, 言語モデル(ユニグラム，バイグラム）を定義し，EM アルゴリズムによる学習を行った．位置の定まったオブジェクトとしてレストランを選び, Yahoo ローカルの 2009 年 12 月から 2011 年 1 月までのデータから，750,000 のレストランを抽出し，ツイートの位置推定を行った．これに対し，本研究では，オブジェクト（話題）の位置は地域に固定されないものとして取り扱う．

Bo ら[5]は，地域特定語を用いてテキスト分類問題にもとづくツイートの位置予測を行った．地域区分の方法として行政区分を用い，人口の少ない地域は，隣接する人口の多い地域と統合することにより，地域間の情報格差に対処する．この地域区分にもとづき，地域特定語の決定するため，語を，(1)ローカルワード（1 地域に属する），(2)セミローカルワード（n 地域に属する），(3)コモンワードの 3 種類に分類した．語の特徴量として，単語頻度と地域頻度に加え，情報利得を用いた．この研究では，地域の範囲について，モデル構築時に地域統合の制約を加えているが，本研究では，分析時に地域の粒度を選択できる手法を提案する．

Cheng ら[6]は，地域特定のキーワードにもとづくユーザ位置の推定アルゴリズムを提案した．地域特定キーワードを選定するために，Backstrom ら[7]が提案した，語の地理

†1 広島工業大学
Hiroshima Institute of Technology

ⓒ2015 Information Processing Society of Japan

的な集中と散らばりのモデルを用い，ユーザ位置推定を行った．この研究では，単語の地域性を用いているが，本研究では,地域特有の単語共起も合わせて用いることにより，位置推定精度を改善する．

Ishida[8]は総務省統計局の定める地域メッシュにもとづき位置推定を行った．Roller ら[9]は，言語モデルと地域区分として適応的グリッドを用いた位置推定を行った．それに対し，本研究では，得られた結果の解釈が容易な，行政地域ごとの地域区分を用いる．

3. 位置推定

位置情報付きマイクロブログから，単語毎に緯度経度を集計し，地域特有の単語を特定する．そのため，位置情報付きツイートから，ツイートにおける名詞を単語として抽出する．また，単語毎に緯度経度の平均，標準偏差を求める．これら統計量にもとづき，単語の地域固有スコアを定義する．以下の手順で位置推定と推定精度の評価を行う．

1. 位置情報付きツイートをデータセットから抽出
2. 位置情報付きツイートを発信したユーザ（ジオユーザ）を抽出
3. ツイートから名詞を単語として抽出
4. 単語の平均緯度経度，標準偏差の計算
5. 単語（単語共起）の地域固有得点データベースを構築
6. ジオユーザの全ツイートをデータセットから抽出
7. ジオユーザの全ツイートと，単語（単語共起）の地域固有スコアにもとづく，ジオユーザの位置推定
8. 推定位置と実際位置の比較にもとづく精度の評価

用いるデータセットは，Twitter public stream のサンプルaを用いて，2015 年 4 月 1 日から 2015 年 8 月 17 日まで収集した全世界の英文ツイートである．ここではユーザの言語設定が英語(en)の場合，そのユーザのツイートは英文ツイートとして扱う．データセットにおけるツイート数は 195,901,490，単語の種類は 46,822,206，ユーザ数は 29,655,251 である.また,位置情報付きツイートは 5,061,719 と全ツイートの約 2.58%，位置情報付きツイートを発信したユーザ数は 2,072,867 と全ユーザの約 7.0%である．

3.1 手法 1：単語を用いた位置推定

位置情報付きツイートから単語を抽出し，単語の出現した緯度経度の平均と標準偏差を求める．ここで，予備的な実験にもとづき，扱う単語は名詞のみとした．単語の品詞決定には GPoSTTL(Enhanced Brill's Tagger)bを用いた．表 1 に抽出された位置情報付き単語の頻度と位置情報の例を示す．ここで得られた単語の平均緯度経度について，全世界の都市データcにもとづき作成した,緯度経度と都市のデータベースを用いて，単語と都市との対応関係を抽出する．さらに，単語の地域特定スコアを定義する（式１）．

表 1　単語頻度と出現緯度経度

単語	頻度	緯度	経度	都市(位置)
selangor	3399	3.14	101.74	Kuala Lumpur, Malaysia
pathum	576	13.80	100.48	Bangkok, Thailand
asimah	236	29.37	47.98	Kuwait City, Kuwait
makati	304	14.53	121.05	Makati, Philippines
cheras	206	3.13	101.72	Kuala Lumpur, Malaysia
prakan	165	13.75	100.47	Bangkok, Thailand
tangerang	237	−6.23	106.85	Jakarta, Indonesia
subang	530	2.84	101.94	Kuala Lumpur, Malaysia
setar	172	6.08	100.38	Alor Setar, Malaysia
東京都	149	35.67	139.70	Tokyo, Japan
binghamton	251	43.16	−77.57	Rochester, United States
nonthaburi	143	13.75	100.47	Bangkok, Thailand
cebu	483	10.37	123.72	Cebu City, Philippines
#unagi	1	1.48	103.73	Johor Bahru, Malaysia
mentai	2	3.15	101.70	Kuala Lumpur, Malaysia

$$Score = tf \times \exp\left(-\sqrt{sx^2 + sy^2}\right) \quad \cdots (1)$$

ここで，各単語についての位置情報付き単語の頻度(***tf***)，経度の標準偏差(***sx***)，緯度の標準偏差(***sy***)を用いている．この定義により，地理的分散が小さく出現頻度の単語は，都市（位置）を特定する単語として高いスコアを得る．このスコアにもとづき，全ツイートに含まれる単語を用いて，ジオユーザの位置推定を行う．各ジオユーザについて，ツイートから抽出した単語に対応する都市のスコアを加算する．これをこのユーザの全単語について行うことにより，ユーザの推定位置（都市）のランキングが得られる．このランキングでトップの都市をユーザの推定位置とする．

3.2 手法 2：地理的散らばり，頻度を制限した位置推定

手法 1 では，地域特定スコアを定義し，単語と都市との関連の強さを計算することにより，単語を用いたユーザの位置推定手法を定義した．ただし，この手法では，出現頻度が非常に高い単語の場合，緯度経度の散らばりが大きい場合でも，比較的高いスコアが得られる可能性がある．
そのため第２の方法では，地理的散らばりと単語頻度に閾値を設け，位置推定に用いる単語を制限し，地域特定スコアを用いる．これらの閾値の設定により，出現頻度が高く緯度経度の散らばりの大きい単語による，位置推定精度の

a https://stream.twitter.com/1.1/statuses/sample.json
b http://gposttl.sourceforge.net/
c https://en.wikipedia.org/wiki/Lists_of_cities (2015 年 8 月 31 日参照)

ⓒ2015 Information Processing Society of Japan

低下を防ぐ．先に行った研究(Ishida [10])と同様に，単語出現頻度の上限については 50000，緯度経度の標準偏差の上限については 2.0 を用いることとした．

3.3 手法 3：単語共起を用いた位置推定

手法 2 では，地理的散らばりや出現頻度の閾値にもとづく地域特定スコアを用いた，ユーザの位置推定手法を定義した．しかし，通常，単語は多義的で，複数の意味を持つものが多いため，異なる意味で用いられている同一表記の単語が，位置推定精度を低下させる可能性がある．そのため，第 3 の手法では，単語共起を用いた位置推定手法を定義する．単語共起における 2 つの単語のうち，一方の単語のみ，地理的散らばりや出現頻度の閾値を用い，位置推定に用いる単語共起を制限する．

双方の単語に制限を課す場合，有効な単語共起が得られる確率が非常に低く，位置推定に利用できる十分な，単語共起と住所の対応関係が得られない．また，双方の単語とも制限無しとした場合，位置推定にとって有用な情報を持たない単語共起が多数含まれるため，位置推定精度の低下や，計算量の爆発といった問題が生じる．

そこで一方の単語のみに閾値を設定した単語共起について，方法 1 でのべた単語のスコアと同様に，単語共起にもとづく地域特定スコアを定義する．この単語共起は，両方の単語が閾値の制約を満たす場合もある．ここで 1 つのツイートにおける単語共起は，含まれる単語すべてのペアである．ツイートの文字列の最大は１４０文字と非常に短いため，同一ツイート内にある単語共起には有意な意味があると考える．表 2 に単語共起の頻度と位置情報の例を示す．

表 2　単語共起頻度と出現緯度経度

単語 1	単語 2	頻度	緯度	経度	場所
alam	selangor	1946	3.15	101.70	Kuala Lumpur, Malaysia
bangkok	pathum	1177	13.76	100.47	Bangkok, Thailand
city	makati	485	14.55	121.03	Makati, Philippines
kedah	setar	416	6.11	100.37	Alor Setar, Malaysia
makati	manila	385	14.55	121.03	Makati, Philippines
cebu	city	497	10.31	123.88	Cebu City, Philippines
ampang	selangor	388	3.15	101.70	Kuala Lumpur, Malaysia
bangkok	chatuchak	490	13.71	100.48	Bangkok, Thailand
makati	metro	371	14.55	121.03	Makati, Philippines
#earthquake	#healdsburg	3	37.78	−122.42	San Francisco,United States
#earthquake	clearlake	4	37.98	−122.18	San Francisco,United States
cagsawa	volcano	2	12.63	121.80	Makati, Philippines
mentai	sushi	14	3.06	101.90	Kuala Lumpur, Malaysia
#unagi	sushi	2	1.38	103.78	Singapore, Singapore

3.4 位置推定精度の評価

単語や単語共起にもとづく位置推定結果について，ツイートに付与された実際の位置情報にもとづき評価を行う．ユーザの位置推定結果においては，地域特定単語や単語共起による位置スコアの合計にもとづき，推定された都市が順位付けされる．そのうち，一番得点の高い都市をユーザの推定位置とする．この推定位置と，実際にユーザが滞在した位置との距離にもとづき，位置推定結果を評価する．この評価方法にもとづき，3.1，3.2，3.3 でそれぞれ定義した，単語による位置推定（手法 1），制限付き単語による位置推定（手法 2），制限付き単語共起による位置推定（手法 3）を比較する．

図 1，2 は，推定された位置と実際の位置との誤差に関するユーザの度数分布の推移を表す． 3 つの手法すべてにおいて，誤差 1000km 以下のユーザ頻度が高い．そのうち手法 3（Method 3，単語共起を用いた位置推定）が 582420 と最も高く，次いで手法 1 （Method 1，単語を用いた位置推定）が 532157，手法 2（Method 2，単語の地理的分散，頻度を制限した位置推定）は 527806 であった．このように，誤差 1000km 以下のユーザ頻度の観点からすると，地域特定単語の代わりに，地域特定単語共起を用いる手法 3 を用いることにより，位置推定の精度が向上することがわかる．

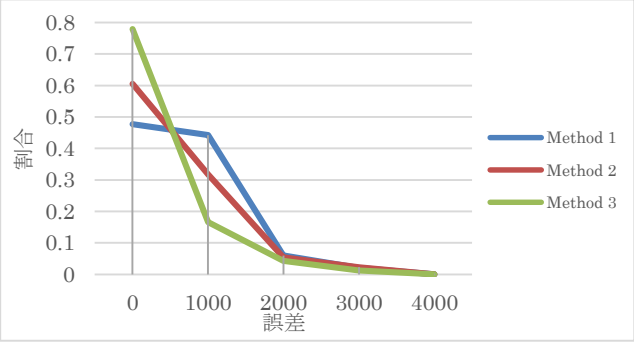

図 1　推定誤差とユーザ割合

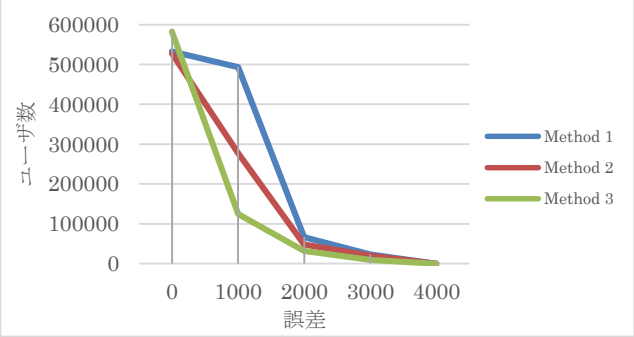

図 2　推定誤差とユーザ数

4. 英語と日本語のデータ比較

前節において全世界の英文マイクロブログデータを用いて，単語共起にもとづく利用者位置推定の有効性について確認した．4.1 ではこのデータと，以前位置推定に用いた日本国内の日本語マイクロブログデータとを比較する．4.2

では，位置推定に用いた日本国内および全世界の都市データの比較を行う．加えて英文マイクロブログにおける都市の位置推定誤差の考察を行う．

4.1 英語および日本語のマイクロブログデータ

英文マイクロブログデータの収集期間は 2015 年 4 月から 8 月の約 5 か月間である．日本語データの収集期間は 2011 年 3 月から 2014 年 5 月の約 38 か月である．2 つのデータセットの収集時期は期間，長さともに異なる．また，英語と日本語の区別は，各ツイートに付与された発信者の言語情報(en または ja)にもとづく．各ツイート個別に言語判定は行っていない．そのため，ここで扱う英語データセットにおける日本語発信，日本語データセットにおける英語発信も含まれる可能性がある．

位置推定においては，名詞の単語や単語共起の出現頻度を用いる．英語データセットにおけるユーザ数とツイート内に出現した名詞の種類，頻度を表 3 に示す．また，日本語データセットにおけるユーザ数とツイート内に出現した名詞の種類，頻度を表 4 に示す．ジオタグデータの割合について英語と日本語を比較すると，ユーザ数では約 7.0%と 2%，単語種類では約 7.3%と約 1.7%，頻度では約 2.0%と 0.4%と，英語データセットの方が全体に占めるジオタグデータの割合が高いことがわかる．

表 3 英語データセットの利用者数と単語種類，単語頻度

	全データ	ジオタグデータ	割合
ユーザ数	29,655,251	2,072,867	0.0699
単語種類	46,822,206	3,395,740	0.0725
単語頻度	1,208,441,397	24,709,867	0.0204

表 4 日本語データセットの利用者数と単語種類，単語頻度

	全データ	ジオタグデータ	割合
ユーザ数	15,886,866	350,415	0.0221
単語種類	48,961,892	827,869	0.0169
単語頻度	3,214,110,936	13,425,539	0.0042

また，全ツイートにおけるジオタグ付きツイートの割合は，日本語では 1%未満であるのに対し，英語では約 2.6%と，英語データにおいて位置情報付きデータが多いことがわかる（表 5）．

表 5 全ツイートにおけるジオタグツイートの割合

	全ツイート数	ジオタグ付きツイート数	割合
日本語	263,581,826	1,132,580	0.0043
英語	195,901,490	5,061,719	0.0258

ここで，一度以上ジオタグ付きツイートを発信した利用者（ジオユーザ）におけるジオタグ付きツイートの平均発信数は，日本語利用者では約 3.23 であるのに対し，英語利用者では約 2.44 と，日本語ジオユーザは英語ジオユーザと比べ頻繁に位置情報を発信していることがわかる（表 6）．つまり，日本語利用者は英語利用者と比べ，比較的限られた利用者が位置情報発信を行うジオユーザであるが，日本語ジオユーザの位置情報発信頻度は，英語ジオユーザと比べ高いことがわかる．

表 6 ジオユーザにおけるジオタグツイートの発信割合

	ジオタグツイート数	ジオユーザ数	平均発信数
日本語	1,132,580	350,415	3.23
英語	5,061,719	2,072,867	2.44

4.2 都市データの比較

英文マイクロブログの位置推定は全世界の都市データ[d]を用いた．それに対し日本語マイクロブログの位置推定は全日本国内の都市データ[e]を用いた．全世界の都市分布を図 3 に示す．

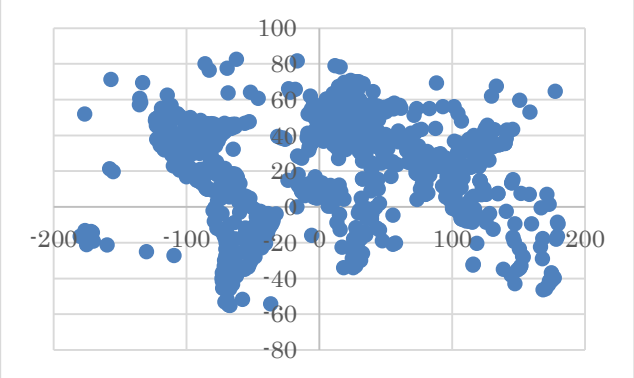

図 3 全世界の都市分布

図 4 に国別の都市数と都市間平均距離の分布を示す．図においてフランス（14 都市，平均距離 7895.4km）とアメリカ合衆国（94 都市，平均距離 3810.5km）の都市間平均距離や都市数が突出している．

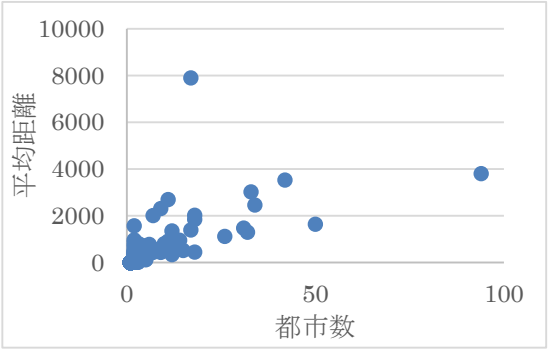

図 4 国別都市数と都市間平均距離

d https://en.wikipedia.org/wiki/Lists_of_cities (2015 年 8 月 31 日参照)
e http://nlftp.mlit.go.jp/isj/ (2015 年 8 月 31 日参照)

ⓒ2015 Information Processing Society of Japan

フランスは，マタウトゥ，バス＝テール，フォール・ド・フランス，サンピエール島およびミクロン島，フランス領ギアナ，マムズ，サン＝ドニ，ヌメアといった，本国から離れた領土を持つ（図５）．また，アメリカ合衆国は，アラスカ州，ハワイ州，アメリカ領サモア，プエルトリコ島，アメリカ領ヴァージン諸島，グアム，サイパン島といった本国から離れた領土を持つ（図６）．これら遠隔地にある多数の領土が都市間平均距離や都市数の突出の要因である．

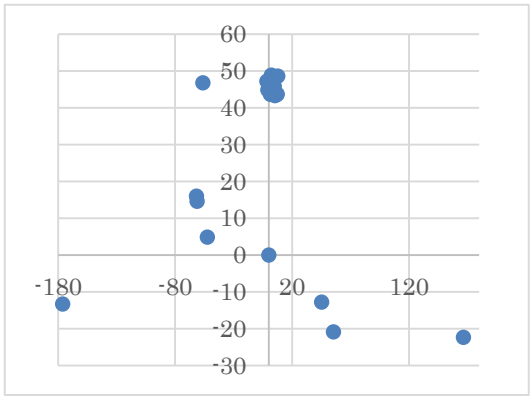

図５　フランスの都市分布

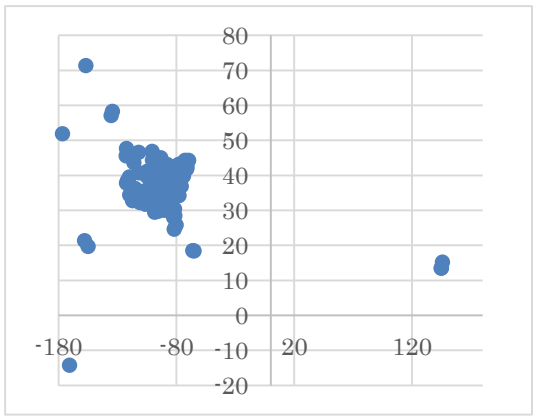

図６　アメリカ合衆国の都市分布

次いで，日本国内の都市分布を図７に示す．

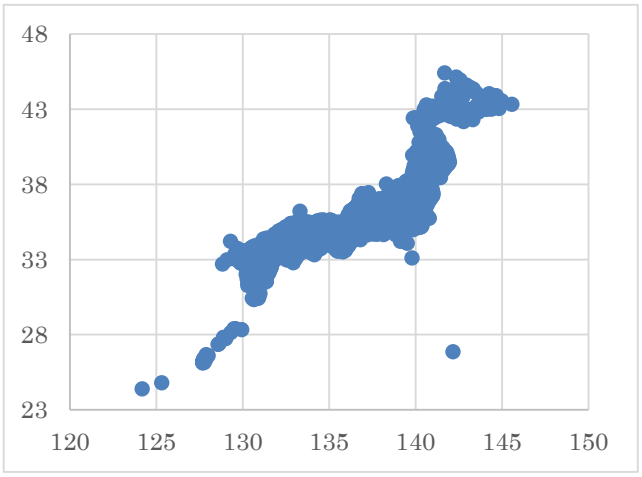

図７　日本国内の都市分布

図８に都道府県別の都市数と都市間平均距離の分布を示す．図において北海道（108 市町村，平均距離 159.5km），東京都（57 市町村，平均距離 81.2km），鹿児島県（42 市町村，平均距離 179.6km），沖縄県（21 市町村，平均距離 81.7km），島根県（15 市町村，平均距離 83.6km）の都市間平均距離や市町村数が突出している．北海道は面積が広いため，市町村数が多く，平均距離が長い．その他の都道府県について，東京都は小笠原諸島，鹿児島県は奄美群島，沖縄県は石垣島や宮古島といった離島が平均距離の長さの要因である．

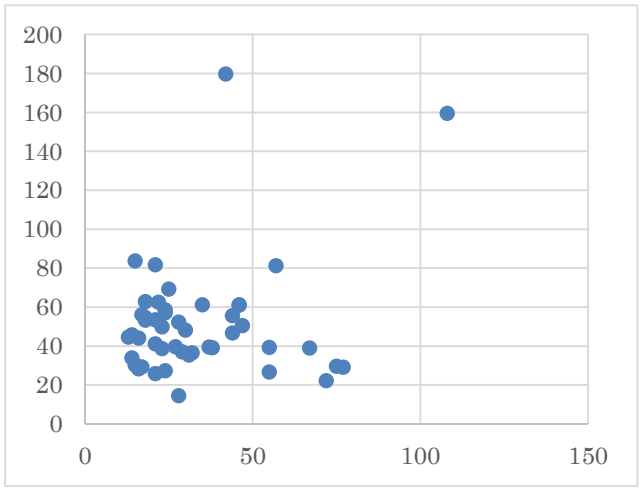

図８　都道府県別都市数と都市間平均距離

4.3 地域別位置推定誤差

図９において方法１による位置推定結果で得られた位置推定件数と実際の位置との誤差の関係を示す．件数が多く誤差が少ない都市として，ジャカルタ(件数 28,186，誤差 258.3km)，クアラルンプール(件数 26,711，誤差 266.3km)，マカティ(件数 26,092，誤差 584.9km)，ロスアンゼルス(件数 25,935，誤差 1578.1)がある．この４都市に着目し，位置推定方法と誤差の関係を考察する．

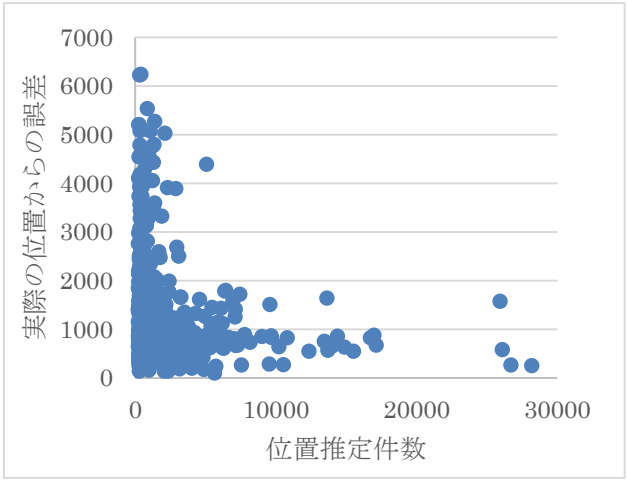

図９　位置推定件数と誤差

方法 2 で得られた結果では，ジャカルタ(件数 24,884，誤差 345.7km)，クアラルンプール(件数 24,884，誤差 284.5km)，マカティ(件数 23,358，誤差 468.4km)，ロスアンゼルス(件数 18,940，誤差 1348.1)である．マカティ，ロスアンゼルスは誤差が減少しているが，ジャカルタ，クアラルンプールは誤差が増大している．

方法 3 で得られた結果では，ジャカルタ(件数 21,832，誤差 222.8km)，クアラルンプール(件数 23,535，誤差 298.4km)，マカティ(件数 21,509，誤差 305.6km)，ロスアンゼルス(件数 15,060，誤差 1174.9)である．ジャカルタ，マカティ，ロスアンゼルスでは誤差が減少しているが，クアラルンプールでは誤差が増大している．

方法 1 と比べ方法 3 は誤差が減少する傾向にあるが，ですべての都市についての誤差が減少するわけではないことがわかる．

次いで，図 10 に方法 1 による位置推定結果で得られた位置推定件数と推定された位置との誤差の関係を示す．件数と誤差が突出した都市として，プライア(件数 41,265，誤差 6523.7km)がある．それに対して，方法 2，3 で得られた結果においては，プライアは除外されている．このように，用いる単語に制約を置き，単語共起を用いることにより，誤った位置推定が低減されていることがわかる．

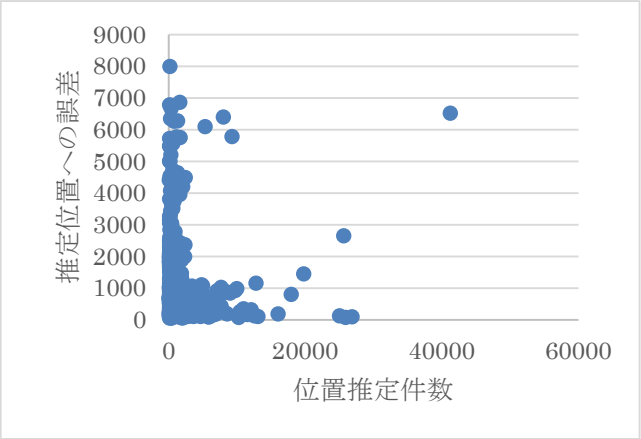

図 10　位置推定の件数と誤差

5. おわりに

英文マイクロブログにおける地域固有表現にもとづき，情報発信者の位置を推定する手法の有効性を確認するために，位置情報付きデータにおける単語出現頻度，平均緯度経度，標準偏差，地域特定スコアを求め，情報発信者の位置推定を行った．また，本研究で位置推定に用いた英文マイクロブログのデータと，以前位置推定に用いた日本国内の日本語マイクロブログデータとの比較を行った．さらに，日本国内および全世界の都市データの比較と，全世界の都市に関する位置推定方法と位置推定誤差の関係性を考察した．

参考文献

1)　Ishida, K. and Ohta T., “An approach for organizing knowledge according to terminology and representing it visually,” IEEE Transactions on Systems, Man, and Cybernetics, Part C, Vol. 32, No. 4, pp. 366-373, 2002.
2) Ishida, K., “Extracting Latent Weblog Communities: A Partitioning Algorithm for Bipartite Graphs,” Proceedings of the 2nd Annual Workshop on the Weblogging Ecosystem - Aggregation, Analysis and Dynamics in the 14th International World Wide Web Conference (WWW2005), Makuhari Messe, Chiba, Japan, May 10 - 14, 2005.
3) Ishida, K, "Extracting Spam Blogs with Co-citation Clusters," Proc. Of the 17th International World Wide Web Conference (WWW2008), April 21 - 25, 2008.
4) Dalvi N., Kumar R., and Pang B., "Object Matching in Tweets with Spatial Models," WSDM’12, February 8-12, 2012, Seattle, Washington, USA.
5) Bo H., Cook P., and Baldwin T., "Geolocation Prediction in Social Media Data by Finding Location Indicative Words," Proceedings of COLING 2012: Technical Papers, pages 1045-1062, COLING 2012, Mumbai, December 2012
6) Cheng Z., Caverlee J., and Lee K., "A Content-Driven Framework for Geolocating Microblog Users," ACM Transactions on Intelligent Systems and Technology, Vol. 4, No. 1, Article 2, Publication date: January 2013.
7) Backstrom, L., Kleinberg, J., Kumar, R., and Novak, J. (2008). Spatial variation in search engine queries. In Proceeding of the 17th international conference on World Wide Web, WWW ’08, pages 357-366, Beijing, China. ACM.
8) Ishida K., "Extracting Geo-Social Information based on Geo-Tagged Social Media," 4th World Congress on Social Simulation (WCSS 2012), National Chengchi University, Taipei, Taiwan, September 4-7, 2012 .
9) Roller S., Speriosu M., Rallapalli S., and Wing R., Jason Baldridge, "Supervised Text-based Geolocation Using Language Models on an Adaptive Grid," Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1500-1510, Jeju Island, Korea, 12-14 July 2012.
10)　Ishida, K., “Estimation of User Location and Local Topics Based on Geo-tagged Text Data on Social Media,” the 4th IIAI International Congress on Advanced Applied Informatics, Okayama Convention Center, Okayama, Japan, July 12 – 14, 2015.

ⓒ2015 Information Processing Society of Japan