Twitterを用いた地域性の強いスポットの抽出

石川　彰夫† 　松本　一則† 　高井　公一† 　安田　圭志† 　服部　元†

† 株式会社 KDDI 総合研究所　〒 102–8460 東京都千代田区飯田橋 3–10–10
E-mail: †{ao-ishikawa,matsu,ko-takai,ke-yasuda,ge-hattori}@kddi-research.jp

あらまし　近年，音声認識の利用が広がっており，その認識の精度も向上している．しかし，地名などの固有名詞は，同音でも地域によって意味が異なる場合もあり，認識に失敗する原因となることもあった．この問題を解決するために，地域依存単語（特定の地域において他地域よりも頻繁に用いられる単語）を抽出して辞書に反映することにより，認識精度を向上させる研究が行われている．一方，Twitter に投稿されたツイートの中のジオタグ付きツイートを収集することにより，地域依存単語を抽出することが可能となる．本研究では，地域依存単語の抽出と，それを適用する地理的範囲（スポット）を決定することを目的として，ジオタグ付きツイートから地域依存単語を抽出するとともに，Echelon 解析を適用し，その単語が使用されている地域を特定することを試みた．その上で，スポットの適切さに関する評価を行い，手法の有効性を確認した．

キーワード　Twitter，Echelon 解析，スポット

1 はじめに

近年，スマートスピーカーの普及に代表されるように，音声認識のアプリケーションとしての利用や，さらには翻訳技術と組み合わせた音声翻訳の利用も進んでいる．その認識の精度も向上しており，一般的な単語を用いた会話であれば実用上は問題ない精度を達成している．しかしながら，そうでない単語，特に地名や施設名，商品名などの固有名詞においては，同音でも地域によって意味が異なる場合もあり，認識に失敗する原因となることもあった．この問題を解決するために，特定の地域において他地域よりも頻繁に用いられる単語（以下，地域依存単語）を抽出する研究が行われている．抽出した単語を，音声認識のための辞書に反映することにより，認識の精度の向上が期待できる．

一方，今日では，ソーシャルメディアを通じて多くの人々が多様な情報発信を行うようになった．Twitter は主要なサービスの一つであり，ユーザはツイートと呼ばれる 140 文字以内の短文を投稿することができる．Twitter には携帯端末からも時間や場所を問わず手軽に投稿することができるため，多くのユーザが情報を観測したその場所でツイートしており，他メディアと比較して実時間性も高い．これらのツイートの中には，投稿した時点のユーザの位置を緯度・経度で表したジオタグ と呼ばれる位置情報付きのツイートも含まれている．そのため，ジオタグ付きツイートを収集することにより，特産品・方言・地名や地名を表す略語などの様々な特定の地域において他地域よりも頻繁に用いられる単語を地域依存単語として抽出することが可能となる．

本研究では，ジオタグ付きツイートから地域依存単語が使用されている地理的範囲（スポット）を特定するとともに，その単語が使用されているスポットを特定することで，音声認識辞書を利用者の場所に応じて切り替える応用を想定している．

第 2 章では，地域依存単語の抽出に関連する従来の研究を紹介する．第 3 章では，本研究の提案手法について説明する．第 4 章で提案手法を評価するための実験の概要と結果およびその評価を述べ，第 5 章でまとめる．

2 関連研究

従来，ジオタグ付きツイートから地域依存単語を抽出する様々な研究が行われている [1] [2] [3] [4]．ツイートではなく，Flickr に投稿されるジオタグ付き画像に付与されるタグを用いて地域情報を抽出したアプローチもある [5]．しかしながら，いずれの研究でも，地域依存単語が使用されたスポットを求めてはいない．

一方，ツイート以外の分野では Echelon 解析が使われている．空間的自己相関の観点からスポットの有無を検定する手法 [6] [7] や，全領域の中を一定の規則に基づいた小領域で走査することで，スポットを抽出する手法 [8] [9] や，疾病の地域集積性を検討する手法 [10] [11] などが提唱されている．

しかし，これらの先行研究による手法は，スポットが不必要に大きくなり辞書の切り替えなどの用途に使用できない等の課題があった．

3 提案手法

国土地理院においては，約 80km 四方の大きさの不連続の四角形で全国を覆った「一次メッシュ」を用いている．本研究では，この一次メッシュが定義された北緯 20 度から北緯 46 度まで，東経 122 度から東経 154 度までの領域を緯線方向と経線方向にそれぞれ 512 等分した約 4km 四方のメッシュ（以下,「基底メッシュ」）を取り扱う．これらのメッシュは，緯度・経度に基づきすき間なく網の目に区画されたほぼ正方形の形状であることから，位置の表示が明確で簡便にできるため，ジオタグ中の緯度・経度のペアと相互対応が容易である．

本研究では，スポットが不必要に大きくなるという課題を解決するため，対象となる地域の基底メッシュに対して，隣接基底メッシュ間の相関に基づいて基底メッシュを併合する手法を採る．即ち，隣接する基底メッシュの集合がスポットである．それによって階層構造に基づく尤度比の高い，適切な大きさのスポットを得られる．

併合の方式として Echelon 解析 [12] [13] を用いる．Echelon 解析とは，空間的な位置を表面上のデータの高低に基づき分割し，空間データの位相的な構造を系統的かつ客観的に見つけるために開発された手法である．また，分割により得られた空間データの構造を表すグラフを，Echelon デンドログラムと呼ぶ．

このように，本研究では，ジオタグ付きツイートを収集して単語の出現頻度の偏りを利用した Echelon 解析によりスポットを抽出することを試みる．

提案手法の手順を説明する．

（1） 一定期間のツイートの中でジオタグ付きツイートを収集する．BOT による投稿など意味を持たないツイートをフィルタリングする．

（2） 基底メッシュごとに 10 ツイート以上で出現した単語の中から，AIC [14] を用いて地域依存単語を抽出する．

（3） Echelon 解析を適用し，Echelon デンドログラムを作成する．

（4） 特徴づけが強い基底メッシュから優先的に基底メッシュをスキャンすることにより，基底メッシュを併合する．

以上の手法により，スポットを高精度に抽出することが可能となる．

以降では，提案手法の各手順について詳細を述べる．

3.1 ジオタグ付きツイートの収集

Twitter を対象とし，緯度・経度で表されるジオタグが付与されたツイートを収集する．その際，単語の出現頻度の違いに対応するため，ツイートに含まれる単語毎に出現履歴を記録する．Twitter Search API によるツイート検索結果 (レスポンス) は JSON 形式のテキストで得られる．ツイートに位置情報が付加されている時，レスポンスの項目”coordinates”に緯度・経度情報が記述されている．次に，ツイートを形態素で分析して単語を取り出している．なお，不要なツイートを排除するためのフィルタの条件については，4 章の実験条件で述べる

3.2 赤池情報量基準（AIC）に基づく地域依存単語の抽出

収集したジオタグ付きツイートに出現する単語それぞれについて，表 1 に示すように，AIC に用いる 2 × 2 表を作成する [15].

表 1 単語の分類

	当該地域内に	当該地域外に
出現する	n_{11}	n_{12}
出現しない	n_{21}	n_{22}

- n_{11} : 当該地域内で発せられたある単語を含むツイート数
- n_{12} : 当該地域外で発せられたある単語を含むツイート数
- n_{21} : 当該地域内で発せられたある単語を含まないツイート数
- n_{22} : 当該地域外で発せられたある単語を含まないツイート数

本研究では，以下の手順により，AIC に基づいて判定値 aic を求めた．なお，$L_{independent}$ は自由パラメータが 2 個の場合の対数誤差を，$L_{dependent}$ は自由パラメータが 3 個の場合の対数誤差を，$aic_{independent}$ は自由パラメータが 2 個の場合の AIC を，$aic_{dependent}$ は自由パラメータが 3 個の場合の AIC をそれぞれ表す．

$$
\begin{aligned}
h &= n_{11} + n_{12} \\
k &= n_{11} + n_{21} \\
N &= n_{11} + n_{12} + n_{21} + n_{22} \\
L_{independent} &= h \log h + k \log k + (N-h) \log (N-h) \\
&\quad + (N-k) \log (N-k) - 2N \log N \\
L_{dependent} &= n_{11} \log n_{11} + n_{12} \log n_{12} \\
&\quad + n_{21} \log n_{21} + n_{22} \log n_{22} \\
aic_{independent} &= -2L_{independent} + 4 \\
aic_{dependent} &= -2L_{dependent} + 6 \\
aic &= aic_{independent} - aic_{dependent} \qquad (1)
\end{aligned}
$$

3.3 Echelon デンドログラムの生成

抽出した地域依存単語に対し，基底メッシュ毎の単語の AIC の判定値に基づいた Echelon 解析を行い，Echelon デンドログラムを生成する．Echelon 解析は，市区町村や州などに分けられた地域上の 1 変量値に対して，空間的な位置を何らかの判定値の高低に基づき分割し，空間データの位相的な構造を系統的かつ客観的に見つけるために開発された解析法である．Echelon 解析で使われる Echelon デンドログラムは，それら空間データの構造を表現したグラフである．

下記の変数を用いる．

- メッシュID : $m(i,j)$
- $tw(i,j)$: 基底メッシュ$m(i,j)$ からの投稿の数
- 単語 $w(k)$: $1 \leqq k \leqq K$
- 標準辞書の単語（標準単語）: $w(1), w(2), \ldots, w(L)$ $L \leqq K$
- $n(i,j,k)$: 基底メッシュ$m(i,j)$ において，単語 $w(k)$ が出現する投稿の数
- $P(i,j,k) = n(i,j,k)/tw(i,j)$: 基底メッシュ$m(i,j)$ での単語 $w(k)$ 出現率
- $\tilde{Pw}(k) = \sum_{i,j} n(i,j,k) / \sum_{i,j} tw(i,j)$: 単語 $w(k)$ の出現率
- $\tilde{Nw}(i,j,k) = \tilde{Pw}(k) \times tw(i,j)$: 基底メッシュ$m(i,j)$ における単語 $w(k)$ の期待出現数
- $\sigma(k) = [\sum_{i,j} \{n(i,j,k) - \tilde{Nw}(i,j,k)\}^2]^{\frac{1}{2}}$: 単語 $w(k)$ の出現数の分散
- $ll(i,j,k)$: 基底メッシュ$m(i,j)$ での単語 $w(k)$ の有用性

ここで，基底メッシュ$m(i,j)$ に出現する標準単語 $w(k)$ の中で $ll(i,j,k)$ が最小 (または 少ない方から N 番目) の単語 w' の $ll(i,j,w')$ の値を θ とし，$ll(i,j,k') \geqq \theta$ となる k' を不足単語とみなす．

基底メッシュ$m(i,j)$ からの $tw(i,j)$ 件の投稿に単語 $w(k)$ を含む投稿が $n(i,j,k)$ 件観測される対数尤度を使用する．

$$ll(i,j,k) = n(i,j,k)\log\frac{n(i,j,k)}{tw(i,j)} + (tw(i,j)-n(i,j,k))\log\frac{tw(i,j)-n(i,j,k)}{tw(i,j)}$$

3.4 基底メッシュの併合

本研究の利用する Echelon 解析では，基底メッシュの併合の判定値に式 (1) の値を用いる．以下に Echelon デンドログラムによる基底メッシュ併合の過程を示す．まず，基底メッシュに対し，上下左右の 4 近傍の連結した基底メッシュの値よりも高い値からなる基底メッシュの集団を 1 つ選択し，第 1 ピークとする．続いて，第 1 ピークに属さない基底メッシュの中から，上下左右の 4 近傍の連結した基底メッシュの値よりも高い値からなる基底メッシュの集団を 1 つ選択し，第 2 ピークとする．以降，第 1 ピーク～第 N ピークのいずれにも属さない基底メッシュの中から第 $N+1$ ピークを作成する処理を，ピークが作成できなくなるまで反復する．以上の方法により，スポットが生成される．

4 実　　験

第 3 章で述べた提案手法に基づいて，スポットを抽出する実験を行った．

4.1 ツイートの収集

実験では，2012 年 10 月 1 日から 2018 年 9 月 30 日までの 6 年間のジオタグ付きツイートを対象とした．スポットを抽出する地域として，東京都，愛知県，長崎県を対象として実験を行った．

BOT による投稿など意味を持たないツイートを排除するために，本実験では以下に示すフィルタを適用した．

- ツイート文から短縮 URL を除き完全一致したものを 1 つだけ残して削除
- ツイート文からリツイート文字列（RT@xxxxx:）を除き完全一致したものを 1 つだけ残して削除
- source(ユーザエージェント) または sourceURL に特定文字列 (BOT ツール名等) を含むものを削除
- ツイート文に特定文字列 (BOT 特有発言等) を含むものを削除
- 特定ユーザ (BOT 等) から発言されたものを削除

フィルタを適用した後の各スポットのツイート件数は，表 2 の通りである．

出現頻度の低い単語を対象とすると精度に悪影響があると思われるため，10 ツイート以上で出現した単語のみを対象とすることとした．10 ツイート以上で出現した単語数と，ツイートに含まれる単語数を，表 3 に示す．

表 2　ツイート件数

地域	件数
全国	125,824,804 件
東京都	22,384,160 件
愛知県	7,951,838 件
長崎県	947,483 件

表 3　10 ツイート以上に出現する単語の数

地域	単語数（総単語数）
全国	953,235 件（4,430,303 件）
東京都	348,610 件（2,042,490 件）
愛知県	127,531 件（827,819 件）
長崎県	28,363 件（224,851 件）

4.2 Echelon 解析の結果

Echelon 解析による併合の結果を以下に示す．それぞれ，図 1 が東京都，図 2 が愛知県，図 3 が長崎県の解析結果である．図中の橙色の部分がスポットを表しており，見やすくするためにスポット内のツイート件数が 1000 件以上のスポットに絞って表示している．

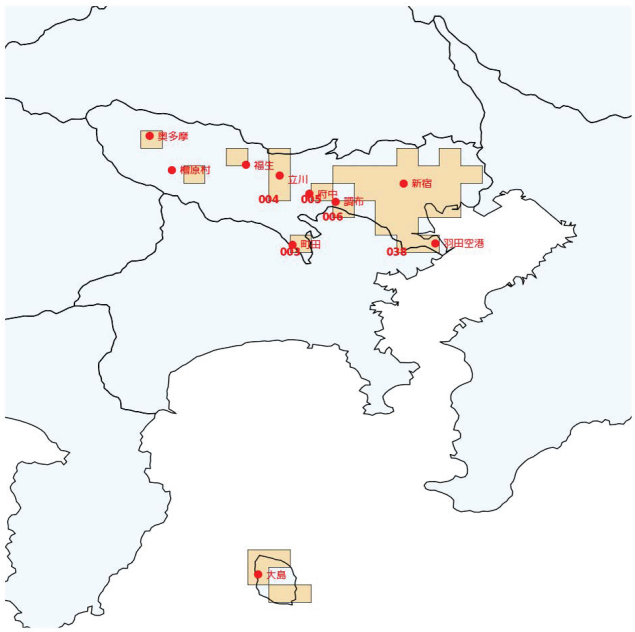

図 1　Echelon 解析による解析結果（東京都）

図中の赤文字は実際の地名を表している．いずれの地域においても，都市部には周辺と比較して大きなスポットが存在していることが分かる．これは仮説として，人が集まる都市部では移動手段が充実しており人の移動が容易であるため，広い範囲で同じ言葉が使われているからと考えられる．

4.3 スポットの範囲の適切さの評価

ここでは，抽出したスポットの適切さを，音声認識性能改善の観点に立ち評価を行なう．具体的には，スポットごとに含まれるツイートから言語モデルを構築し，各スポットの評価データにおける未知語率とパープレキシティーにより評価を行なう．

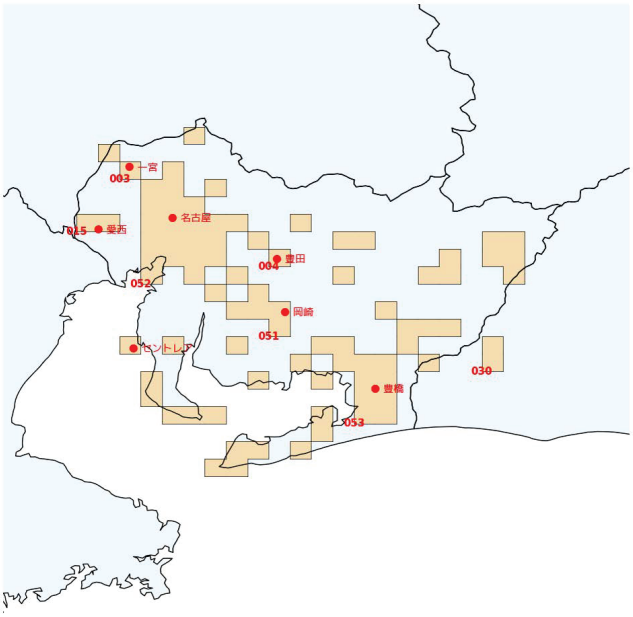

図 2　Echelon 解析による解析結果（愛知県）

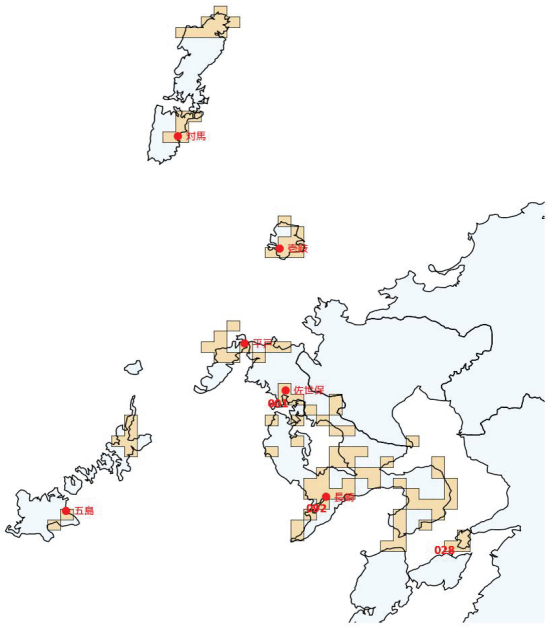

図 3　Echelon 解析による解析結果（長崎県）

一般に，言語モデルにおける未知語率やパープレキシティーは，その値が低いほど音声認識精度が良いとされている．評価対象のスポットとして，図 1 より新宿と町田，図 2 より名古屋と岡崎，図 3 より佐世保と長崎の，計 6 スポットを選んだ．

ツイートから抽出したスポットとその併合スポットのツイートの件数を，表 4 に示す．併合スポットで得られたツイートの実際の件数はもっと多いが，東京都と愛知県に比べて長崎県のデータ量が少なかったため，学習データ量を長崎県のそれに揃えた．ここで，単語数は学習文から出現する異なり語数を記しており，評価文にのみ出現する単語は件数に含めていない．

表 5 に，本実験における未知語率を示す．この表において，

表 4　データ件数

スポット	学習データ	評価データ	単語数
新宿	76595	961	30373
町田	77480	985	27548
名古屋	77149	984	27855
岡崎	76404	966	26200
佐世保	78131	984	26737
長崎	77634	979	27094
全国	77689	-	28252

行と列はそれぞれ，評価データのスポットとモデルのスポットを表す．

また，表 6 は，各スポットの学習データから 3-gram 言語モデルを構築し，各スポットの評価データを元にパープレキシティを算出した結果である．言語モデル学習のツールキットとしては SRILM [16] を使用した．この表においても，行と列はそれぞれ，評価データのスポットとモデルのスポットを表す．いずれのスポットにおいても，言語モデルの学習データとテストデータのスポットが一致した場合，異なるスポットや全国の言語モデルの結果と比較して，未知語率やパープレキシティが最も低くなっていることが確認できた．該当するパープレキシティの値をイタリック体で示す．以上のことから，提案手法を用いて生成されたスポットが有効であると考えられる．

次に，地図情報として得られる行政地域と Echelon 解析によるスポット抽出とを比較するため，一例として，名古屋市の境界と，今回の実験で抽出した名古屋スポットを対象とし，未知語率とパープレキシティの比較を行なった．結果を，それぞれ表 7 と表 8 に示す．「Echelon」とあるのが Echelon 解析で得られたスポットであり，「地図」とあるのが地図上の行政地域である．ここでも，各言語モデルの学習データ量は，同一となるように揃えている．

表 7 の未知語率と表 8 のパープレキシティがともに小さくなっており，同じ傾向であることが判る．また，評価データが「Echelon」の場合に Echelon の値が地図の値よりも大きくなっているのに対し，評価データが「地図」の場合は Echelon の値と地図の値で同等となっていることから，Echelon 解析によるスポット抽出結果が，言語的特徴を捉えつつ，地域性も網羅できていると言える．

5　おわりに

本研究では，Twitter に投稿されるツイートから，Echelon 解析を用いて単語の空間的局所性に基づいて地域依存単語を抽出し，スポットを抽出する手法を提案した．その上で，スポット毎にパープレキシティを算出し，抽出結果の定量的評価を行った．

実験の結果，いずれのスポットでも異なるスポットや全国と比較してパープレキシティが低くなった．これにより，提案手法により生成したスポットが，言語的特徴を反映した地理的範囲を表現しうることが分かった．

今後の課題として，空間的局所性だけではなく時間的局所性

表 5　未 知 語 率

モデル＼評価データ	新宿	町田	名古屋	岡崎	佐世保	長崎
新宿	*13.89%*	17.23%	19.88%	18.17%	21.74%	19.95%
町田	20.08%	*12.69%*	21.39%	19.16%	22.31%	20.81%
名古屋	19.65%	19.61%	*14.54%*	17.27%	22.04%	20.31%
岡崎	23.32%	21.12%	21.17%	*11.29%*	22.67%	20.83%
佐世保	23.66%	22.37%	23.32%	19.08%	*10.37%*	18.18%
長崎	23.23%	20.89%	22.17%	18.20%	19.13%	*10.42%*
全国	18.32%	18.33%	19.93%	16.41%	18.70%	17.92%

表 6　パープレキシティ

モデル＼評価データ	新宿	町田	名古屋	岡崎	佐世保	長崎
新宿	*189.18*	185.37	193.25	216.41	224.24	195.82
町田	202.34	*149.66*	192.37	196.95	207.05	191.19
名古屋	200.67	177.45	*166.85*	193.95	202.88	182.61
岡崎	211.01	191.00	192.29	*166.78*	204.45	183.08
佐世保	219.90	192.47	205.38	205.52	*147.33*	187.70
長崎	214.74	196.80	205.64	200.52	207.41	*159.26*
全国	210.69	194.23	202.96	204.18	219.74	187.64

表 7　名古屋の未知語率

モデル＼評価データ	Echelon	地図
Echelon	*14.54%*	*13.42%*
地図	15.95%	13.60%

表 8　名古屋のパープレキシティ

モデル＼評価データ	Echelon	地図
Echelon	*166.85*	174.06
地図	*172.43*	174.78

も考慮した時空間的局所性に基づく単語の抽出方法の検討が挙げられる．また，Echelon 解析で得られたスポットと地図上の行政地域との比較を，名古屋以外の地域においても確認したい．

謝　　辞

本研究は，総務省「グローバルコミュニケーション計画の推進 -多言語音声翻訳技術の研究開発及び社会実証- I. 多言語音声翻訳技術の研究開発」の一環として実施したものです．

文　　献

[1] 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.759768, 2010.

[2] 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.111118, 2012.

[3] 上村卓也, 新田直子, 中村和晃, 馬場口登, “マイクロブログからのリアルタイム地域情報抽出”, データ工学と情報マネジメントに関するフォーラム, C71, 2017.

[4] 新田直子, 吉武真人, 中村和晃, 馬場口登: ”マイクロブログからの関連実世界観測情報の抽出”, 日本データベース学会和文論文誌, Vol.16-J, Article, No.22, 8 pages, March 2018.

[5] T. Rattenbury and M. Naaman, “Methods for Extracting Place Semantics from Flickr Tags,” ACM Transaction on the Web, 3(1), 30 pages, 2009.

[6] P. Moran, ”The interpretation of statistical maps”, Journal of the Royal Statistical Society B, Vol.10, pp.243251, 1948.

[7] L. Anselin, ”Local indicators of spatial association-LISA”, Geographic Analysis, Vol.27, pp.93115, 1995.

[8] S. Openshaw, M. Charlton, C. Wymer, and A. W. Craft, ”A mark 1 geographical analysis machine for the automated analysis of point data sets”, International Journal of Geographical Information Systems, Vol.1, pp.335358, 1987.

[9] J. Besag, and J. Newell, ”The detection of clusters in rate diseases”, Journal of the Royal Statistical Society, Series A, Vol.154, pp.143155, 1991.

[10] T. Tango, ”A class of tests for detecting ’general’ and ’focuses’ clustering of rate diseases”, Statistics in Medicine, Vol.14, pp.23232334, 1995.

[11] T. Tango, ”A test for spatial disease clustering adjusted for multiple testing”, Statistics in Medicine, Vol.19, pp.191204, 2000.

[12] W. L. Myers, G. P. Patil, and K. Joly, ”Echelon approach to areas of concern in synoptic regional monitoring”, Environmental and Ecological Statistics, Vol.4, pp.131152, 1997.

[13] K. Kurihara, ”Classification of geospatial lattice data and their graphical representation”, ClasEchelon sification, Clustering and Data Mining Applications, pp.251258, 2004.

[14] H. Akaike, ”Information theory and an extension of the maximum likelihood principle”, Proceedings of the 2nd International Symposium on Information Theory, pp,267-281, 1973.

[15] Keiji Yasuda, Panikos Heracleous, Akio Ishikawa, Masayuki Hashimoto, Kazunori Matsumoto, and Fumiaki Sugaya, ”Building a Location Dependent Dictionary for Speech

Translation Systems", International Conference on Computational Linguistics and Intelligent Text Processing 2017, pp.482-491, 2017.

[16] The SRI Language Modeling Toolkit. http://www.speech.sri.com/projects/srilm/