2H3-5

オープンデータに基づく地域オントロジを用いたソーシャル分析

Social Analytics using Region Ontology Created by Open Data

村上 明子 *1　伊川 洋平
Akiko Murakami　Yohei Ikawa

*1日本アイ・ビー・エム株式会社 東京基礎研究所
IBM Research - Tokyo

Recently, Twitter data can be thought that one of the most important information source at the disaster time, such as earthquake, heavy rain and flood, typhoon, etc. Twitter data contains time and textual information, and some of the data also contains location information. Due to the privacy issue, recently most users do not add location information for their post obviously, however location names and landmark buildings are found frequently in the posted messages. In this paper we use location-related Open Data for identify the location area of the post. We map each name of location and landmark buildings to a certain area for identify which area is most urgent for recovery from the disaster. We also prototyped a visualization system which can view residents thoughts and sentiments by time and areas at the disaster time.

1. はじめに

災害時に災害の現場のことを最も把握しているのは現場にいる人である．現場にいる人からの情報は災害情報の把握として非常に重要であり，現場の情報をさまざまな方法で取得する方法が近年多く試みられている．ソーシャルメディアの情報は，このような災害の現地での情報を取得する手段の一つとして多くの期待が寄せられており，2014 年 2 月の関東甲信越地方の豪雪の際には長野県佐久市市長が Twitter で住民に情報の提供を呼びかけ情報を収集し迅速な対応に役立てたなどの実例も出てきている．また，災害時にハッシュタグ等を用いて情報を発信するように住民に呼びかけるなどの活動も行われている．さらに報道の現場では，ソーシャルメディアから事件や事故などの初期情報を取得し，取材活動に役立てようといった動きも見られている．

災害に限らず，このようなソーシャルでのリアルタイム情報では空間情報も重要な情報のひとつとなる．ソーシャルメディアの中には緯度経度のような地理空間情報を付与できる機能を持ったものも多くあるが，すべての発言に発信位置の情報がついているとは限らない．特に，コンテキストを共有した仲間同士の発言では，施設名や道路名など正確な地名ではなく共有した知識に基づいた地理情報で情報が交換される．この地理情報は広い範囲を示していたり，その土地の固有の表現であったりするため，この情報を元に位置情報を推測したとしても，高い精度は多く期待できない．

さらに，ソーシャルメディアには情報の信頼性という問題がある．悪質なデマであったり，悪意はなかったとしても，伝聞による不確かな情報が多くソーシャルメディアには投稿されている．

地理情報の不正確さと発言の信頼性のなさを鑑みても，災害発生時のような情報の少ない中ではソーシャルの情報は重要な情報のひとつであることには変わりはない．そのため，筆者らはソーシャル発言の地理情報を地点ではなく地域で把握し，多く言及されている地域はどこかを地域間の差で把握したり，時間による変分を把握することで，ノイズや信頼度の低い発言を排除する方法を提案する．本研究ではそのために必要な，発言内の地理的な情報から地域情報へ変換するための言語リソースについて考察する．また，その言語リソースに基づいて分析されたソーシャルメディアのデータが，どのように利用されるかについても議論する．

連絡先: 村上 明子，akikom@jp.ibm.com

2. ソーシャル発言に見られる地理手掛かりランドマーク

災害時，ソーシャルメディアには多くの発言が投稿されるが，その中には建造物や場所の名前といったことから特定の地域に関する発言だとわかるものがある．下記にその例を示す．

1. 「明日は**福知山球場** 行く予定だったが福知山球場も水没して大変な事になってるので流石に練習試合はなさそう.」
2. 「**神田川**来た．超濁流ですけど (´Д｀) 氾濫しそうで怖い.」

この発言の中で，**福知山球場**は京都府福知山市にある野球場，**神田川**は東京都に流れる河川の名前である．このように，直接的な住所の表記はなどはなくとも発言内容に関する地域を知ることができる．このような場所の手掛かりになる語を「地理手掛かりランドマーク」と呼ぶことにする．

ソーシャルの分析で，発言内容と空間情報を関連付けるためには，この発言中の地理手掛かりランドマークから空間情報へマップするための言語リソースが必要となる．

3. 地域オントロジー

前章で示したように，Twitter や Facebook といった SNS の発言中では場所を示す際に住所そのものを書き下すことはまれである．例えば、「東京スカイツリー」に行った事を表現するときに「スカイツリーに行った」とは記述するが,「墨田区押上一丁目に行った」と記述することはほぼないといってよい．そのため，ソーシャルの発言内から場所を推定するためには，地名だけではなく地理手掛かりランドマークの表現を捉える必要がある．

地理手掛かりランドマークには，建造物や施設名を示すものと，河川名や山の名前のように地域全体を指すものなど，い

くつかの種類が考えられる．また～地域や～流域といった表現も，ひとつの地理手掛かりランドマークであると考えられる．

地理手掛かりランドマークから発言の内容がどこの地域に関するものなのかを関連付けることで，現象や住民の感情がどの地域で多く起こっているのかなどの多くの情報を集められると考えられる．例えば図 1 に示すようなある川の流域での発言を考える．地域として「AA 区」,「BB 区」,「CC 区」,「DD 区」の 4 つ，ランドマークとしてソーシャルの発言中に「XX 川」「YY 橋」「ZZ 小学校」が現れたと考える．

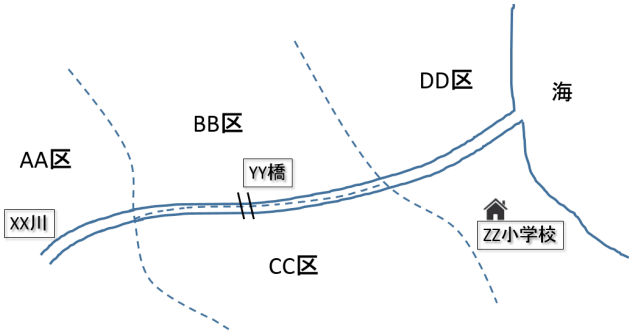

図 1: 地理手掛かりランドマークと地域の例

「ZZ 小学校」は「DD 区」に位置している小学校であり，この地理手掛かりランドマークを含んだ発言は DD 区に関するものであると考えることができる．一方で「YY 橋」は「BB 区」と「CC 区」にまたがって架かっており，この 2 つの地域のどちらか，あるいは両方の地域について関するものであると考えることができる．さらに,「XX 川」は広くこの川の流域を指すため、特定の地域を表したものではなく，この 4 つの地域すべての地域に関する発言と捉えられる．

この例のように、地理手掛かりランドマークを地域にマッピングするという作業は，そのランドマークが示す範囲と，ランドマークの場所と地域との関係，この 2 つに依存する．そのため，ランドマークと地域をマップするためには，まず地域を特定し，その後に各ランドマークが示す範囲について検討する必要がある．このような関連付けを行うことに必要な言語リソースを「地域オントロジー」と呼ぶこととする．

4. オープンデータによる地域オントロジーの作成

地名などの各種情報から緯度経度などの地理座標への変換はジオコーディングと呼ばれ，一般的なウェブ上での地図サービスで実装されている．ジオコーディングは文字列で与えられた各種情報と，地理座標の対となるデータによって実現できる．

今回作成したいものはこのような各種情報と地理情報との一対一対応ではない．それは，前述の橋の例のように 2 つの地域にわたるものであったり，川の例のように複数の地域にわたるものであったりするからである．そのため，必要となる地域オントロジーは，地理手掛かりランドマークと地域への一対多対応のものとなる．この章では，その地域オントロジーを作成する方法について述べる．

4.1 分析対象地域の決定

結び付けられる地理手掛かりランドマークの文字列の情報は，地名，建造物，河川や山の名前などの自然物の名前などさまざまである．前節で述べたように，これらの示す範囲はその種別によって異なる．また，たとえそれが狭い範囲を示すランドマークであったとしても，前にあげた橋の例のように地域の境界線上にある場合は 2 つの地域に属することもありえる．したがって、地域オントロジーを作成する際にはまず分析地域の定義が必要となる．

まずは分析する必要がある対象の領域を決定し，その中での地域を決定する．例えば，ある河川の氾濫に対する住民の不安を捉えたいのであれば，その河川の流域にあたる地域を対象領域とする．分析する地域の区分は分析の目的によって決定する。例えば行政がその行政区ごとの住民の感情を分析したいのであれば，区や市といった行政区の単位で設定することもできる．各地域の範囲は，緯度経度のポリゴンで設定しておく．

4.2 地理手掛かりランドマークと地域へのマップの作成

次に，地理手掛かりランドマークの文字列と地域へのマップを作成する．これは，地理手掛かりランドマークの種別ごとに行う．

地図上でほぼ位置が確定する地理手掛かりランドマークである建造物や施設名の場合は，その地理座標と各地域の範囲のポリゴンを比較し，どの地域に属するかを判定する．建造物や施設名の地理座標は，自治体や国土地理協会（http://www.kokudo.or.jp）などで公開されているデータを用いることができる．また国立情報学研究所の GeoNLP プロジェクトでは，施設名等とそれに対応する地理座標のデータをオープンデータとして公開している (https://geonlp.ex.nii.ac.jp/)．

注意すべきは，これらの地理手掛かりランドマークが分析する地域の境界線上に位置している場合である．この場合は，どちらの地域に明確に属するかは判定できないため，2 つ（または 3 つ以上の場合もある）の地域に属していると考える．

河川，山，公園などある程度の領域を表す地理手掛かりランドマークは，分析する地域にそのランドマークがどのように広がっているのかを調査し，マップを作成する必要がある．例えば，東京で「神田川」は 23 区のうち杉並区、中野区、新宿区、豊島区、文京区、千代田区、台東区、中央区を流れる河川であるため，区ごとの分析をする際にはこの 8 つの区すべてのことに言及しているとする．

地名などのように県や市、町名といった階層構造を持っている地理手掛かりランドマークは，その階層構造を利用して地域へのマッピングを行う．その際，熊本県と熊本市のように，単に「熊本」とあったときにどちらの階層なのか曖昧性がある場合がある．その場合は，より抽象度の高い県レベルでの階層であるとする．これらの階層つき地名オントロジーは，Wikipedia（http://www.wikipedia.org/）などから得ることができる．

5. 地域オントロジーを用いたソーシャル分析の例

地域オントロジーを用いて、発言中に含まれる地理手掛かりランドマークを地域にマッピングさせることで，地域ごとの発言の量や，それに含まれる感情を比較することが可能となる．また，地域間の差だけでなく時間的変化を見ることで，ある地域に関してある感情が多く起こったり，落ち着いたりといった現象を可視化することができる．

筆者らは発言中の「怖い」や「不安だ」といった表現を捉える分析とこれらの地理情報を組み合わせることで，どの地域に不安な人が多いか，といった分析を行った [1]．これは熊本県熊本市の白川流域において大雨による河川の氾濫に関する住民

の不安があがったと想定し,, 2012 年の実際の水害時のデータを元に作成したサンプルのツイートを用いて，避難所のある地域ごとに時刻情報とともに可視化したものである．その分析イメージを 2 に示す．

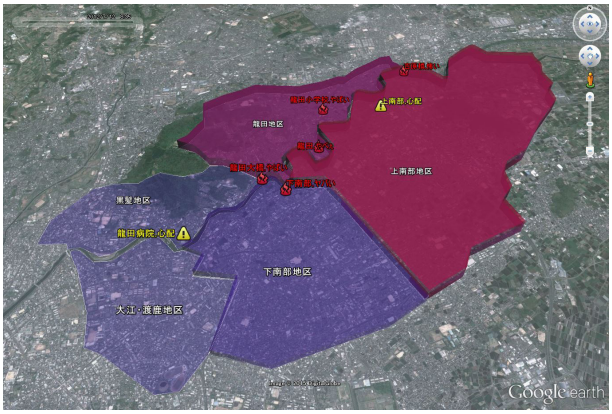

図 2: 熊本白河流域における地域ごとの Twitter 内の不安度の可視化

地域ごとの不安の多さを大小を比較することにより，どこの地域の住民が他の地域に比べ不安を多く感じているのか，といったことを知ることができる．また，ソーシャルの発言時刻を情報として加えることで，時刻の経過による地域ごとの不安の増加，あるいは解消していく様子を知ることも可能となる．

6. 関連研究

外部知識からオントロジーを作成する研究は，人手で作成された分類基準のある日本語語彙大系を用いて Wikipedia の情報から大規模オントロジーを作成する研究 [2] など数多くあげられる．この研究では Wilipedia の中の is-a 関係に着目し，地名に限らず組織などの階層のあるオントロジーを大規模に作成可能とするものである．

ソーシャルメディアの発言の位置情報推定の研究も多く，ソーシャルのつぶやきから地震の震源地を知る研究 [3] や，位置に関連する地理的局所性のある用語を取得しその用語に基づいて推定する研究 [4]，単語の地理的局所性の時系列変化を利用した発信位置推定の研究 [5] などある．

本研究では，上記のような一般的な地理情報やオントロジーではなく，分析のための地域に基づいたオントロジーを作成することが必要になる．これらの一般的なオントロジーからの変換による地域オントロジーの自動作成は今後の課題である．

また，ソーシャルの発言を災害の早期発見に役立てようという研究も多くある．ツイッターなどのリアルタイムのつぶやきから土砂災害の予兆などを検知する試みや [6]，火災などの災害の初動を検知するといった研究 [7] などがあげられる．

7. まとめ

本論文では，ソーシャルの発言等、テキストに含まれている地理手掛かりランドマークを地域にマッピングするための地域オントロジーの提案と，それの作成方法について述べた．また，その地域オントロジーを用いて，地域間の発言の分析による災害時の情報可視化の可能性についても述べた．多くの発言を地域に集約することによって，ソーシャルメディアでの地理情報の表記の曖昧性と信頼性の欠如を補完か可能になるのではと考えている．

今後の課題としては，ソーシャルにおける発言者の地域間の偏りをどう扱うのかという問題が挙げられる．これは，地域間の居住者の数，あるいはソーシャルツールを使っている人の数など，偏りにさまざまな要因がある．また，悪意あるいは無作為のノイズが混入したときのノイズ耐性などについても議論をする必要があると考えている．

参考文献

[1] 村上 明子, 伊川 洋平, 「Twitter を用いた災害時の住民感情の分析」, DEIM2015, 2015

[2] 柴木優美, 永田昌明, 「山本和英:日本語語彙大系を用いた Wikipedia からの汎用オントロジー構築」, 情報処理学会研究報告, 自然言語処理研究会報告 2009-NL-194-4, 2009

[3] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. ”Earthquake shakes Twitter users: real-time event detection by social sensors.” Proceedings of the 19th international conference on World wide web. ACM, 2010

[4] S. Paradesi, “ Geotagging Tweets using Their Content, ”Proceedings of International Florida Artificial Intelligence Research Society Conference (FLAIRS), pp. 355?356, 2011.

[5] 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」, DEIM2014, 2014

[6] http://www.nilim.go.jp/lab/bcg/kisya/journal/kisya20140714.pdf

[7] 斎藤翔太, 伊川洋平, 鈴木秀幸, 村上明子, 「Twitter を用いた災害情報の早期発見」, 言語理解とコミュニケーション研究会（NLC）, 2014