2B4-NFC-02a-2

字・町名をキーとした災害時 Twitter 情報の抽出と地図への展開

-「どこ」で「何」が起きているのかを知る-

Twitter data mining for Crisis Mapping

原 久美子[*1] Kumiko Hara
木野 泰伸[*2] Yasunobu Kino
鳥海 不二夫[*3] Fujio Toriumi

[*1] 筑波大学 University of Tsukuba
[*2] 筑波大学 University of Tsukuba
[*3] 東京大学 The University of Tokyo

After 2005 Hurricane Katrina, the significance of citizen-based crisis management has been increasing so as to cope with natural disasters. One of the most effective tools for the crisis management must be crisis mapping. Crisis mapping strongly supports the situation awareness, which is the essential stage of the crisis management. However, the hurdle of utilizing the text data in SMS or SNS for crisis mapping is quite high. This paper attempts to contrive the technique of getting the coordinates from tweets that have no geo-tag. Based on the analysis of Twitter data which was recorded in 2011 Tohoku earthquake, the authors propose a filtering technique. The result of the test indicates that lower layer of Japanese address format can be used as the key.

1. はじめに

2011 年 3 月 11 日,宮城県沖を震源地とするマグニチュード 9.0 の東日本大震災が発生した.この震災において日本は,一国家の対応能力を遥かに超える問題を次々と突きつけられる事態に陥った.この大震災を契機に,巨大災害時の危機管理のあり方が官民問わず広く議論されている.

災害時を含む危機管理は,長らく国家や公共機関の責務とされて来た.しかし,2005 年のカトリーナハリケーン災害を緒として,近年では一般市民による自発的危機管理の重要性が増して来ている.2010 年のハイチ地震では,紙のベースマップが壊滅状況に陥ったハイチのために,世界中のアマチュア地図作製家がインターネット上においてデジタル地図を作製し,その地図上に SMS(携帯電話のメッセージング機能)の情報が集約された[Laituri, 08].また,市民による自発的情報収集の手段の一つとして Twitter が有効であることが,マルセイユ森林火災の例等で明らかになっている[Longueville, 09].災害時の地図による情報の可視化は,危機管理の重要なステージである SA (Situation Awareness)を強力に支援する.

地図を用いた危機管理には直接位置情報が必要であるが,2010 年の Sysomos 社の調査によれば,発信場所の緯度経度情報とともに発信される tweet は全 tweet のうちのわずか 0.23%にとどまる.自然言語に含まれる日本の地名を手がかりとする直接位置情報の特定については,テキスト中の経路情報を用いる手法[野秋, 04],観光情報に特化した手法[石野, 10]などの先行研究がある.しかし,災害時の状況把握に有益な tweet を抽出し,「どこで(座標値)何が(トピック)起きているのか」というデータへと効率的に変換する手法はいまだ確立されていない.本研究では,東日本大震災時の Twitter ログの分析を行い,危機管理に有益な tweet を字・町名(注)をキーとして抽出する手法の提案と検証を行った.これは,地名階層上位からの検索では抽出漏れとなる情報を,可能な限り抽出しようとする試みである.具体的には次の分析を行い,抽出手法の検討をし,一都道府県単位を対象とする情報の抽出手法を提案,検証した.なお、本研究での位置特定とは、字・町域代表点レベル精度とする.

1-被災地の地名と災害キーワードを同時に含む tweet の一般的特徴を把握する
2-字・町名で抽出した tweet のノイズの種類と原因を調査する.
3-住所表記階層上位の地名が出現しない tweet からのハッシュタグによる位置特定の可否を調査する
4-住所表記階層上位の地名が出現しない tweet から,必要な tweet を抽出する手がかりとなる"ユーザープロフィールの住所"の記載率と記述パターンを調査する
5-字・町名のみが含まれる災害関連 tweet の重要性を分析する

(注)住居表示法施行済み地域については"丁目"部分を含まないものとする.以下,字・町名という用語についてこの説明を省く

2. 東日本大震災時の tweet 分析

東日本大震災時に記録された Twitter のログを用い,宮城県についての tweet 分析を行った.

2.1 データと分析ツール

図 1 に示すように分析用データの作成を行い,6 つのデータセットを作成した.データセット E を除くそれぞれのデータセットには 3/13, 3/14, 3/21 の日別で複数の種類がある.

分析用災害キーワードは「情報、電気、安否、状況、水道、ガス、状態、食料、ガソリン」とした.

データ作成および分析用ツールは Quantum GIS, MeCab, MeCab ユーザー辞書, Python コード,および UNIX コマンドである.MeCab ユーザー辞書は『位置参照情報ダウンロードサービス (国土交通省)』データから作成した.

連絡先:原久美子, 筑波大学ビジネス科学研究科,
s1245009@u.tsukuba.ac.jp

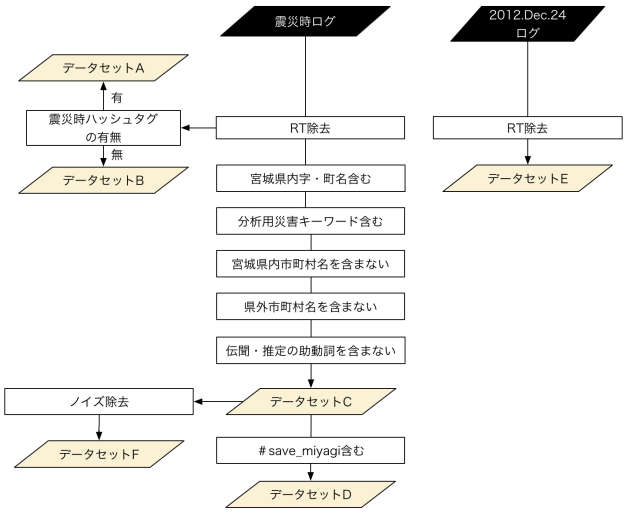

図 1: 分析用データの概要

2.2 災害関連 tweet の特徴

被災地の地名と同じ単語を含む多数の tweet の中から,災害関連 tweet を抽出し得る特徴の有無を確認するためにこの分析を行った.分析に用いたデータはデータセット A および B である.データセット A と B について,次の項目の比較を行った結果を表 1 に示す.

- 1tweet あたりのバイト数平均(1tweet の長さ)
- ポジティブ顔文字(113 種)とポジティブ感情記号(♪,☆)数/tweet 数のスコア
- 助動詞のうち,“です/ます”の数/tweet 数のスコア

表 1:災害時 tweet の一般的特徴

	3/13		3/14		3/21	
	災害	通常	災害	通常	災害	通常
総数(単位 1,000)	36	11,160	33	13,827	11	15,008
1tweet の平均バイト数	202	110	204	108	204	94
Positive 顔文字・記号	0.53	4.27	0.68	4.70	1.64	3.17
ですます	68.38	30.40	63.61	27.90	44.67	22.10

2.3 字・町名のノイズ分析

字・町名を検索キーとして抽出を行った場合のノイズの混入状況と原因を分析した.分析に用いたデータは 3/13 のデータセット C (63,497 件)である.

この分析では,まずデータセット C に出現する字・町名の出現頻度をカウントした.次に出現頻度上位 100 の字・町名を含む tweet の数を宮城県内の家屋被災状況マップに重ね合わせ(資料出所:総務省統計局・政策統括官・統計研修所ホームページ,東京大学空間情報科学研究センターCSV アドレスマッチングサービス),異常値を示すものをノイズと判定した.これらのノイズは 4 種類に分別可能であることがわかった.結果を次に示す.

- 1 文字の字・町名であるため,他地域の地名の一部,一般名詞の一部が抽出されてしまっているもの(例:関「東」,「関」西,関「西」,青「森」,「大」震災,「原」発)
- その日のニュースや検索ランキング上位に存在する人名・ニュースや検索ランキング上位に存在する宮城県外の地名の混入(例:「枝野」官房長官,「宮崎」新燃岳噴火,「岡田」幹事長)
- 全国に重複が多数あるか,一般的に良く知られた地名の混入 (例:五反田,福岡,沼津)
- 一般名詞が抽出されているもの(例:桜,平成)

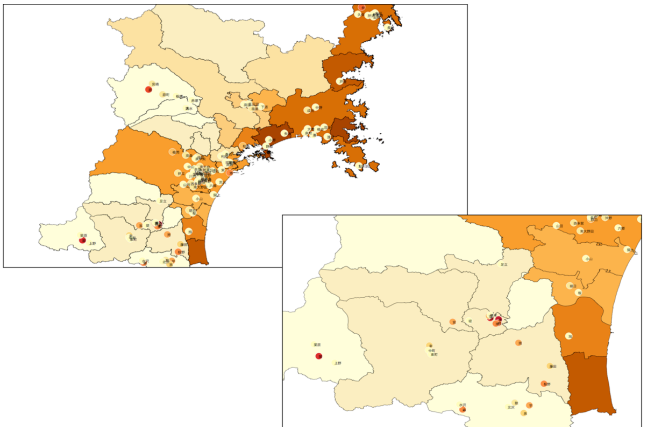

図 2:字・町名出現頻度と被災状況の重ね合わせ

2.4 ハッシュタグを用いた字・町名の位置特定可否分析

地名階層上位の地名が tweet 中に出現しない場合,ハッシュタグでどの程度位置を特定することが可能かを分析した.

#save_miyagi を含む 3/13 のデータセット D は 428 件で同日データセット C(63,497 件)のうち 0.6%である.この 428 件のうちハッシュタグで位置特定可能な tweet は 302 件 70.5%である.

2.5 ユーザープロフィール住所の記述分析

twitter のユーザープロフィール住所の記述パターンを分析した.この分析では次の 2 点を確認した.分析に用いたデータはデータセット E(1,297,962 件)である.

1- ユーザープロフィールの記述率(架空地名等を含む)
2- ユーザープロフィール住所の記述パターン

ユーザープロフィール住所の記載率 は 55%(1,297,962 件中の 715,874 件)であった.

ユーザープロフィール記載のあるもののうち,宮城県のユーザープロフィール住所記述パターンは,“宮城県”(784 件),”宮城“(474 件),”宮城(県)+郡・市町村“(1,030 件),みやぎ”(211 件),“ミヤギ”(10 件),“M(m)iyagi”(194 件),“杜の都”(226 件),“牛タン(県,王国)”(66 件),“笹かま(県,王国)(4 件)”,“S(s)endai”(364 件)等の表記が使われている.市名についてのカウントは宮城県内人口第 2 位の自治体”石巻市”単独(14 件),第 3 位の自治体”大崎市“単独(3 件)となっている.

2.6 字・町名のみが含まれる災害関連 tweet の重要性

字・町名をキーとする手法の妥当性を確認するため、字・町名のみが含まれる tweet について、内容の重要性分析を行った.分析に用いたデータは 3/13 のデータセット F(2,874 件)である.重要性の判断基準を「報道では得られない情報が含まれているか否か」と定義し,目視により確認を行った。

3/13 データセット F のうち、重要性がある tweet は 1,283 件 44.6%である.目視で把握した特徴を次に記す.

- 情報の内容が具体的である(例:道路の損壊状況,商店の在庫状況,行列の長さ等)
- 被災者から被災者に向けた情報交換メッセージが多い

- 友人・知人同士の情報交換 tweet の文体は「ですます体」の率が低い
- 字・町名に付随して道路,公共施設名,病院名,商業施設名などが多く抽出される

3. 分析結果の考察

分析結果の考察を次にまとめる.

3.1 災害関連 tweet の特徴

災害時関連 tweet は,次の 3 つの特徴を持つ.

1- バイト数が 200 以上である

2- ですます体で記述される

3- ポジティブな感情表現の顔文字および記号を含まない

災害関連の tweet については,1 つの tweet でより多くの情報を伝えたいという心理,またポジティブな感情の表現を自粛するべきという心理が働くものと考えられる.外部への発信を意識するため「ですます体」で記述されることが多くなり,過去の tweet を無意識に模倣して文章を組み立てるために,文体がパターン化されるのではないかとも考えられる.処理すべき tweet 数が膨大な場合は,これらの特徴を用いた抽出も検討すべきである.

3.2 字・町名ノイズの分析

字・町名のノイズの原因は明確である.一般名詞と人名の誤抽出については,当日のニュースランキングに出現する単語の共起語,検索ランキング上位に出現する単語の共起語を NG ワードとして設定すればよい.これらの NG ワードによって誤抽出を大きく減じることが可能である.1 文字の地名/全国で重複する地名/被災地外にある有名な地名については,市町村行政区名を添加して抽出用辞書を作成する,辞書コストの調整を行う,NG ワード設定で誤抽出を防ぐといった方法が考えられる.ただし,いずれの方法でも字・町名が単独で出現した場合の抽出漏れの懸念があることは否めない.

3.3 ハッシュタグの利用可否

“#地名“,“save_地名”というパターンのハッシュタグは市町村名を含まない tweet の位置特定の手がかりとして使用可能である.東日本大震災時には災害発生から数時間でエリア名(”tohoku”等)・県名のハッシュタグが設定され,1 日程度経過して各市町村名のハッシュタグが設定された.位置特定用ハッシュタグとしては地域名のアルファベット,日本語の 2 種類を考慮に入れておくべきである.

3.4 ユーザープロフィール住所の利用

ユーザープロフィールの住所を大別すると,a.都道府県名(漢字,ひらがな,アルファベット) b.政令都市名(漢字,ひらがな,アルファベット) c.地域の特産物・歴史上の有名人 の3つのパターンが存在する.県名に続けて市町村名が出現するケースは上述 3 つのうち a に含まれると考えてよい.いずれのケースでも,位置特定の手がかりとして有効である.

3.5 字・町名のみが含まれる災害関連 tweet の割合と重要性

市町村名を含まず,字・町名のみを含む tweet には,報道では得られない生の情報が含まれている率が高い.Tweet 量全体に対する比率は小さいが,数千件単位と数が多いため,市民発・現地発の 1 次あるいは疑似 1 次情報として重要性が高いと考えられる.また,字・町名をキーとすることにより医療機関名,公共施設名,座標データの入手が難しい商業施設名を抽出可能であるメリットは無視できない.道路に関する情報についても,地点名が明瞭であるため有益であると考える.

4. tweet の抽出手法提案

前章までの分析および考察から,字・町名をキーとして,危機管理に有益な tweet を抽出する手法を次のように提案する.

1) 字・町名を基本とし,次に該当するものに市町村行政区名を添加して MeCab ユーザー辞書を作成する
   - 文字数が 1 である字・町名
   - 一般名詞と重複する字・町名
   - 県内,全国で重複する字・町名
   - エリア外の政令都市,政令都市の駅名と同じ字・町名

2)NG ワードを設定する.
   -当日のニュース記事における,出現頻度の高い人名と接尾詞の組み合わせ,出現頻度の高い人名と共起する単語
   -検索ランキング上位の人名・抽出対象地区外の地名と共起する単語
   -抽出対象地区外の地名と共起する単語

3) 抽出処理の流れを図 3 に示す.抽出された A, B, C, D のデータは,座標付与により地図上で扱うことが可能となる.

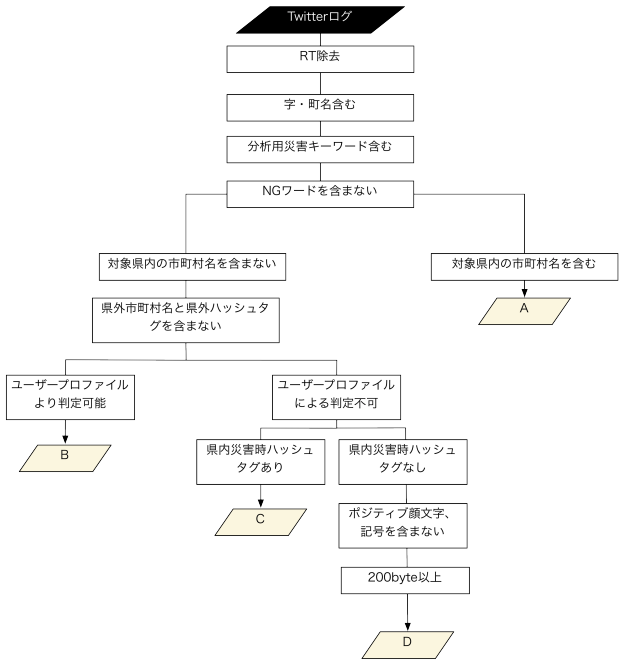

図 3 提案手法概念図

5. 検証

RT を含めない 3/15 の Twitter ログのうち,「ガソリン」「コンビニ」「診察」「復旧」のいずれかのキーワードと字・町名を同時に含む tweet を用い検証を行った.このデータには,正解フラグを人の手によって付与している.

5.1 各種設定

- NG ワード:長官,輪番,グループ,幹事長,噴火,枝野さん,寝ろ,枝野氏
- Positive 顔文字,記号:24 種類
- ハッシュタグ: “#miyagi”

5.2 検証結果

抽出を行った結果を表 2 に示す.なお,このデータにはユーザープロフィールが含まれていないため,ユーザープロフィールを用いた処理プロセスは実行していない.

表 2:提案手法による抽出状況

	tweet数	フラグ
RTなし	15,384,190	
字・町名あり	136,289	348
キーワード抽出	1,723	348
県内市町村名あり	678	259
県内市町村名なし	1,045	89
NGワードなし		
県内市町村名あり	655	255
県内市町村名なし	793	89
県内外とも市町村名なし	698	81
#miyagiタグあり	13	1
#miyagiタグなし	685	80
顔文字・記号なし	677	79
顔文字・記号あり	8	1
200byte以上	367	43
200byte以下	317	37

5.3 検証結果考察

提案手法を用い,正解フラグ付きのデータ 348 件のうち,311 件(89.3%)を正しく抽出し,42 件が失われた.
各ステップにて失われるデータの割合は次のようになる.

- NG ワードの有無 4 件 1%
- Positive 顔文字と記号 1 件 0.2%
- 1tweet のサイズによる判定 37 件 10%

NG ワードおよび positive 顔文字・記号を用いた tweet 除外による消失は,誤差として許容範囲と考える.しかし, tweet サイズによる除外については抽出すべきデータの 10%が失われてしまうため,抽出元となるデータの量,また抽出の目的によって使い分けることが必要となる.

6. 抽出 tweet の地図上展開テスト

提案手法により、前章の検証よりも多くの災害関連キーワードを用いて抽出した 3/12 の tweet に対し、MeCab を用いて座標を与え,地図上に展開した(代表点の座標取得は東京大学空間情報科学研究センターCSV アドレスマッチングサービスを利用).さらに、プロットした tweet を総務省統計局の字・町名ポリゴンでカウントし,tweet 数の大小によって塗り分け図に表現したものを図 4 に示す.

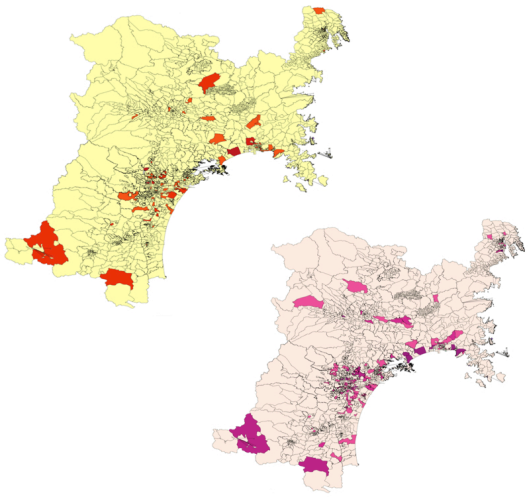

図 4:「安否」に関する 3/12 の tweet(上),「状況」に関する 3/12 の tweet(下) Quantum GIS 使用.

7. まとめ

字・町名をキーとして危機管理に有益な tweet を抽出する手法については,本研究の目的を原始的なレベルにおいて果たしたものと考える.しかし,抽出判定に及ぼす各要素の影響のモデル化には至らず,更なるデータの精査が必要である.また,字・町名単独出現と現地発1次情報との関連性を定量的に証明する必要性がある.本研究で作成した MeCab のユーザー辞書についてはコスト設定等の整備が充分でなく,高い精度を得るために辞書の調整を行わねばならない.

地図上での個人発テキスト情報可視化については,リアルタイムで状況把握が可能であるという地図の強みが災害時に生かされるはずである.その反面,同一人物の繰り返し投稿が高い値として表現されてしまう,投稿のなかった地域では何も起きていないという誤解を与えてしまうという問題も発見した.Twitter のデータのみでは発信者の属性が偏りがちであるため,携帯電話のテキストメッセージングサービスである SMS を災害時に集積できるシステムの実現が待たれる.

参考文献

[Laituri, 08] Laituri, M., & Kodrich, K. : On line disaster response community: People as sensors of high magnitude disasters using Internet GIS, Sensors, 2008

[Longueville, 09] De Longueville, B., Smith, R. S., & Luraschi, G.: OMG, from here, I can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires, In Proceedings of the 2009 International Workshop on Location Based Social Networks, 2009

[野秋, 04] 野秋浩三, 相良毅, 有川正俊:歩道ネットワークと地名辞書を基本とした日常的な場所表現を対象としたジオコーディング手法, 電子情報通信学会第 15 回データ工学ワークショップ論文集, 2004

[石野, 10] 石野亜耶, 難波英嗣,竹澤寿幸: 旅行ブログエントリからの観光情報の自動抽出, Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, 2010