1H2-NFC-02a-4

災害情報基盤構築に向けた テキストデータからの地理情報抽出システム

Anomaly Detection Local Areas Using Social Media

榊 剛史[*1] Takeshi Sakaki
原 久美子[*2] Kumiko Hara
吉田 光男[*2] Mitsuo Yoshida
鳥海 不二夫[*1] Fujio Toriumi
篠田 孝祐[*3] Kosuke Shinoda
栗原 聡[*3] Satoshi Kurihara
風間 一洋[*4] Kazuhiro Kazama
野田 五十樹[*5] Itsuki Noda

[*1]東京大学 The University of Tokyo
[*2]筑波大学 The University of Tsukuba
[*3]電気通信大学 The University of Electro-Communications
[*4]和歌山大学 Wakayama University
[*5]産業技術総合研究所 National Institute of Advanced Industrial Science and Technology

We are planning to develop geo-location information extraction system from text data. This system makes it possible to convert gps information and location name extracted from social media posts into administrative district information. Extraction of administrative district information is important for disaster management because support activities and rescue operations are performed by the administrative section in the case of disaster situation. In this paper, we propose a method to estimate the location of sending form tweets. From next year, we will try to develop applications for disaster management based on this mission-critical system.

1. はじめに

災害発生時に，いち早く正確な情報を収集し，また収集した情報を迅速に伝えるべき人，組織に伝達することは，災害支援活動において重要なタスクとなる．

2011 年 3 月 11 日に発生した東日本大地震においても，様々な情報が伝わらないことにより，災害支援活動に支障が生じた事例が多数挙げられる．例えば，各地での不足物資に関する情報が正しく伝達されなかったために需要を上回る支援物資が現地に送られ，その整理や補完に人や場所のリソースが割かれてしまうという事態が発生した[*1]．逆に，災害発生時に正しく情報が伝達されることで，災害支援活動のサポートや避難行動の一助となった事例もある．Google が Person Finder という安否情報の交換サービスを立ち上げたことで．被災者による安否情報の発信，家族や知人の安否確認がスムーズに行われた[*2]．また，ホンダとパイオニアによる被災地の通行実績情報が Google MAP 上で提供されることで，通行可能な道をいち早く知ることができるようになり，被災者の避難行動や災害支援者が現地到着が効率化されたとされている[*3]．

ソーシャルメディアの普及に伴い，広域災害発生時に様々な情報がソーシャルメディア上に発信されるようになってきた．例えば，東日本大地震においては，被災地の現状や安否情報，不足物資，通行可能ルートなど災害支援活動に役立つ様々な情報が Twitter 上に投稿されていた．また，2014 年 2 月中旬の記録的降雪に伴う山梨県，群馬県等での孤立世帯の情報について，現地入りできないためにマスコミによる報道が行われない中，個人により多数の情報が Twitter に投稿されていた[*4]．つまり，災害時にソーシャルメディアに投稿された情報を収集し，適切に提示することで避難行動や災害支援活動を補助することが可能になると考えられる．

災害時にソーシャルメディアに投稿される情報において，どこから投稿されたか，すなわち情報の発信場所はソーシャルメディアの情報を役立てる上で重要な手掛かりとなる．例えば，物資不足情報，通行可能経路情報，安否情報いずれの情報にしても，適切に活用するためには位置情報が不可欠である．また，被災地にいるユーザに情報を伝達する際も，そのユーザがどこにいるかによって，伝えるべき情報が異なる．このようにソーシャルメディアの投稿を収集する際は，その投稿から位置情報を取得する手法が必要となる．そこで，本近未来チャレンジでは、入力された投稿を分析し，入力された投稿の発信場所，その緯度経度情報，その行政区域を出力するようなシステム（Web API）を提供することを目指す．災害支援活動や避難行動は行政区域単位で行われるため，より情報を有効に活用するために行政区域を出力する．

本稿では，そのようなシステムを実現するために，ソーシャルメディアの投稿から発信場所の位置情報を推定する手法を提案する．情報の収集しやすさ，リアルタイム性の高さから Twitter 上の投稿を対象とする．提案手法では，複数の手法を用いてツイートの発信場所を推定し，それらの結果を統合し，発信場所を絞り込んで行くアプローチを採用する。位置情報の推定手法としては，1. GPS 情報を用いる手法，2. 投稿者のプロフィールを用いる手法，3. 投稿者のソーシャルグラフを用いる手法，4. ツイート内の地名を用いる手法の 4 手法を用いる．

2. 関連研究

いくつかの研究において，災害時にどのように Twitter が活用されるかの分析が行われている．Miyabe らは東日本大地震における Twitter の利用傾向を分析している [Miyabe 12]. Mendoza らは 2010 年のチリ地震において，地震後数時間後か

連絡先: 榊 剛史，東京大学大学院工学系研究科，東京都文京区本郷 7-3-1 工学部 3 号館

∗1 http://communityarts.jp/wp-content/uploads/Asahi-Shinbun-11.06.11.pdf
∗2 http://google.org/personfinder/japan
∗3 http://response.jp/article/2011/03/15/153269.html
∗4 http://matome.naver.jp/odai/2139264049251471001

ら数日後まで，Twitter ユーザがどのように振舞ったかについて調査を行ない，その特徴を明らかにしている [Mendoza 10]．これらの研究では，災害発生時にはソーシャルメディア，特に Twitter がよく活用されることが言及されている．

ソーシャルメディアのユーザプロフィールに記載されている地名を位置情報として用いている研究や [Hecht 11]，ユーザのリンク情報や, ユーザとリンク関係にあるユーザの位置情報を用いてユーザの位置情報を推定する手法も提案されている [Backstrom 10]．

ソーシャルセンサを用いる研究のうち、位置情報が必要となる代表的な研究として局所的なイベントを検出する手法が提案されてきた [Lee 11a, Walther 13, Lee 11b, Li 12]．これらの研究においては，位置情報として，ツイートに付与された GPS 情報 [Lee 11b] やユーザの居住地情報 [Li 12]，特定地域の地名リストを用いている [Walther 13]．その中で，Watanabe らの研究は機械学習の手法を用いてツイートへの地名付与を実現している．[Watanabe 11]．

また固有表現抽出の手法を用いて，ソーシャルメディアの投稿から地名を抽出する手法も提案されている [Ji 09, Lin 04]．さらに Ritter らは CRF を適用することでツイートへの品詞付与エラーを減らすと共に LabeledLDA を適用することで，既存手法と比べ固有表現抽出の精度が 25% 改善することを示している [Ritter 11]．しかし，日本語を解析する際には品詞を付与する前に単語分割を行う必要があるが，既存の形態素解析ツールでは，口語的な文書において単語分割に失敗することが多い．そのため，この手法を日本語ツイートにそのまま適用することは難しい．

3. データセット

本研究では，震災前後に投稿された日本語のツイートを収集した．収集手順は，以下の通りである．

1. 日本語 Twitter ユーザのリストを作成
2. 1. のリストのユーザが震災前後に投稿したツイートを収集

収集したデータセットの詳細は以下の通り．

- ユーザリストに含まれるユーザ数：130 万ユーザ
- ツイート数：356,118,522 ツイート
- 収集期間：2011 年 3 月 7 日〜3 月 24 日

4. 提案手法

本節では，ツイートの発信場所を推定するための提案手法について述べる．まずは，本研究で用いる代表的な 4 つの位置情報推定手法の概要とその特徴について述べる．その後，それらを相補的に組み合わせて，ツイートの発信場所を精度良く推定する手法を提案する．

4.1 個別の位置情報推定手法

ここでは，本稿で用いる代表的な 4 つの位置情報推定手法を挙げる．

4.1.1 GPS 情報の利用

ここでは，投稿に付与された GPS 情報を用いて投稿者の位置情報推定を行う．

モバイル機器でツイートを投稿する場合，その GPS 情報をツイートに付与することができる．この GPS 情報はジオタグと呼ばれる．一般的に，GPS による情報は人間の認知機能よりもはるかに正確に現在位置を表現できるため，このジオタグを位置情報として用いる．この手法は，正確かつ詳細な位置情報が得られる反面，まれに GPS の値を偽装するユーザがいるため，誤った位置情報が得られる可能性がある．一般的に GPS が付与されたツイートは全体の 0.2% 程度であるため，本アプローチで位置情報が取得できる可能性は低い [Lee 11a]．

4.1.2 ツイート内地名の利用

ここでは，投稿に含まれる地名を用いて投稿者の位置情報推定を行う．

地名は含むツイートは数多く投稿されている．そこで，その地名を抽出し，地名→緯度経度変換を行い，位置情報として取得する．ただし．投稿者がツイート内で言及している地名は，1．投稿者がいる場所，2．投稿者が興味を持っている場所のいずれかの場合である．そのため，ツイート中の地名を投稿者の地名とする場合、低くない確率で誤った位置情報が得られる可能性がある．

4.1.3 ユーザプロフィールの利用

ここでは，投稿者のユーザプロフィールから投稿者の位置情報推定を行う．

Twitter のユーザプロフィールにはユーザの居住地を入力する項目がある．またユーザプロフィール自体に居住地を書き込むユーザもいる．ただし，ユーザ毎に正確性や具体性が異なる．例えば,「東京都北区王子」のように大字まで入力しているユーザもいれば,「東京都」や「日本」のように大ざっぱな記入をしているユーザ，さらには「夢の中」「この世のどこか」など実際の居住地とは無関係の情報を入力しているユーザもいる．そこで，本研究では推定可能なユーザのみ居住地を推定し，それを場所情報として利用した．ユーザの居住地は，市町村名及び都道府県名（漢字，ひらがな，カタカナ，アルファベット表記いずれか）がユーザプロフィールに含まれているか否かにより判定した．本手法は多くのユーザに適用できるものの，都道府県単位という粗い単位でしか位置情報を推定することができない．

4.1.4 ソーシャルグラフの利用

ここでは，投稿者のソーシャルグラフから投稿者の位置情報推定を行う．

既存研究より，投稿者とリンク関係のあるユーザは，投稿者と同じ地域に住んでいる割合が高いことが知られている [Backstrom 10]．そこで，投稿者位置情報が未知であった場合，投稿者とリンク関係のあるユーザのプロフィールを用いて，投稿者の位置情報を推定を行う．

具体的には，投稿者とリンクある各ユーザについて，都道府県単位で居住地判定を行う．その後、最も多かった都道府県の割合が全体の 0.3 を超えている場合に，その都道府県を投稿者の居住地と見なす．

本手法は多くのユーザに適用できるものの，都道府県単位という粗い単位でしか位置情報を推定することができない．

4.2 複数手法を組み合わせた位置情報推定手法

前述した 4 つの位置情報推定手法を組み合わせて，ツイートの発信場所を推定する手法を提案する．

まず，上記 4 つの手法は得られる位置情報の粒度から 2 つに分けることができる．1．GPS 情報を用いる手法と 2．ツイート内地名を用いる手法については，詳細な位置情報が得られる可能性がある．その反面，2．については位置情報の偽装や単に地名について言及しているだけの場合，実際のツイート発信地から大きくずれてしまう可能性がある．1．についてはそも

表 1: 提案手法によるツイート発信地推定の評価

	地域判定	地域一致	地域不一致
可能	0.728 (182/250)	0.212 (53/250)	0.516 (129 /250)
不可能	0.272 (68/250)		

そも GPS が付与されているツイートの割合が低いため，位置情報が抽出できる可能性が低い．一方，3．ユーザプロフィールを用いる手法，4．ソーシャルグラフを用いる手法については，都道府県単位という粗い粒度でしか位置情報が推定できない．その反面，これらの位置情報は投稿者の居住地域を表していると考えられるため，2．ツイート内地名を用いる手法と比較して，ツイート発信地から大きくずれる可能性は低いと考えられる．また 1．と比較してユーザプロフィールやソーシャルグラフは全てのユーザが持っているものであるため，位置情報を抽出できる可能性が高い．

そこで，提案手法では，下記の様に位置情報推定を行う．まず，3 と 4 の手法を用いてツイート発信地の大体の地域を絞った後, 1 と 2 の手法を用いて詳細なツイート発信場所を絞り込む．仮に後者で得られる詳細な位置情報が前者で絞り込んだ地域から外れている場合, 位置情報としては用いないものとする．

このように，

- 粒度は粗いが，確度の高い位置情報
- 粒度は詳細であるが，確度の低い位置情報

という 2 種類の推定位置情報を組み合わせることで，より確実に位置情報を推定することを目指す．

5. 評価実験

提案手法の評価実験を行う．本提案手法においては，東日本大地震における被災地 5 箇所を選定し，その地名を含むツイートをデータセットより抽出する．それらに対し提案手法を適用し，それらの情報発信元について特定すると共に，被災地域からの情報が抽出できているかを検証する．

具体的な手順は下記の通り．

1. データセットより，東日本大地震において被災地域 5 箇所の地名を含むツイートをそれぞれ 50 ツイートずつ無作為に抽出する．今回用いた地名は，被害の大きかった地域として「石巻」「大船渡」，中程度の被害だった地域として「筑波」，原子力発電所事故が発生した場所として「双葉町」，被害の少なかった地域として「本郷（文京区）」を選定した．
2. 各ツイートに提案手法を適用し，投稿者の居住地推定及びツイート内地名，GPS データによるツイート発信地の絞り込みを行う．
3. 各ツイートで判定された発信地について人手で評価を行う．

結果は表 1 の通り．表 1 より，全体としてユーザプロフィール，ソーシャルグラフから地域（都道府県）判定ができたのが全体の 0.728 である．また，判定された地域と含まれる地名の地域が一致したものは，0.212 であった．つまり，地名を含むツイートのうち，実際にその地名から発信したと推定されるツイートは全体の五分の一程度であった．なお，今回の抽出したデータには GPS が付与されていたツイートが存在しなかったため，今回は評価対象外とした．

表 2: 発信場所が推定できたツイートの地域別割合

石巻	大船渡	双葉町	筑波	本郷
0.208 (11/53)	0.094 (5/53)	0.00 (0/53)	0.226 (12/53)	0.471 (25/53)

また発信場所が推定できたツイートのうち，各地域ごとの割合は表 2 の通りである．表 2 より，発信場所が推定できたツイートの多くは，被害の少なかった本郷，筑波から発信されていることが分かる．また，双葉町という地名を含むツイートで双葉町から発信されていると推定されたものは，0 件であった．

表 3 に，実際に発信場所が推定できたツイートの例を示す．表 3 より，確かにこれらのツイートは当該地域から発信されていると考えられる．ただし，これらのツイートについて，本当に当該地域から発信されているかを厳密に判定することは困難である．そのため，今回は正解／不正解の判定は行わず，ツイートを定性的に分析することで，あくまで「正解らしい」と判定するだけにとどめる．

6. 終わりに

本稿では，広域災害時に災害支援活動や避難行動に役立つツイートを収集するために，ツイートの発信場所を推定する手法を提案した．

まず，既存の位置推定手法の特徴を明らかにした後，互いの手法の長所が短所を補うように，各手法を相補的に組み合わせる事によって，詳細な発信場所の推定を実現した．評価実験を通じて，実際に提案手法により被災地から投稿されたと推測されるツイートを収集することができた．

今後は，提案手法をツイートの投稿場所推定 API として実装する．すなわち，ツイートを入力することで，その投稿地域，緯度経度情報およびその行政区域を出力するような API である．この API を本近未来チャレンジの成果として，誰もが利用可能な形で提供していく予定である．

7. 謝辞

本研究を行なうにあたり，ツイートデータの収集に協力していただいたクックパッド株式会社の兼山元太氏及び株式会社ホットリンクに感謝する．また，本研究は Microsoft Research Asia University Relations の助成を受けた．

参考文献

[Backstrom 10] Backstrom, L., Sun, E., and Marlow, C.: Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity, in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 61–70, ACM Press (2010)

[Hecht 11] Hecht, B., Hong, L., Suh, B., and Chi, E. H.: Tweets from Justin Bieber's Heart: the Dynamics of the Location Field in User Profiles., in *Proceedings of the 2011 Annual Conference on Human factors in Computing Systems*, CHI '11, pp. 237–246, ACM Press (2011)

表 3: 発信場所が推定できたツイートの例

まみたすのお家も海の近くだから心配 (; _;) 石巻悪いで有名になりすぎだぁ
再度。宮城県石巻市蛇田字西境谷地の川沿い、三菱自動車跡地後ろの二階だて住宅に取り残されています。二階まで浸水しそうです。子供が 2 人いるため屋根に登ることもできません。救助要請おねがいします。
大船渡のおともだち連絡つかないよどうしよづｙｓ
茨城大学、筑波大学、茨城県立医療大学ともに後期日程の学力検査は行わないことになりました。当初は延期が考えられていたようですが、この状況での実施は困難とのことです。

[Ji 09] Ji, R., Xie, X., Yao, H., and Ma, W.-Y.: Mining City Landmarks from Blogs by Graph Modeling, in *Proceedings of the Seventeen ACM International Conference on Multimedia*, MM '09, p. 105, ACM Press (2009)

[Lee 11a] Lee, C.-H., Yang, H.-C., Chien, T.-F., and Wen, W.-S.: A Novel Approach for Event Detection by Mining Spatio-Temporal Information on Microblogs, in *Proceedings of International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '11, pp. 254–259, IEEE (2011)

[Lee 11b] Lee, R., Wakamiya, S., and Sumiya, K.: Discovery of Unusual Regional Social Activities using Geotagged Microblogs, *World Wide Web*, Vol. 14, No. 4, pp. 321–349 (2011)

[Li 12] Li, R., Lei, K. H., Khadiwala, R., and Chang, K.-C.: TEDAS: A Twitter-based Event Detection and Analysis System, in *IEEE 28th International Conference on Data Engineering*, ICDE '12, pp. 1273–1276, IEEE (2012)

[Lin 04] Lin, J. and Halavais, A.: Mapping the Blogosphere in America, in *Workshop on the Weblogging Ecosystem at the 13th International World Wide Web Conference*, Vol. 18 (2004)

[Mendoza 10] Mendoza, M., Poblete, B., and Castillo, C.: Twitter under crisis: can we trust what we RT?, in *Proceedings of the SOMA 2010*, pp. 71–79, New York, New York, USA (2010), ACM Press

[Miyabe 12] Miyabe, M., Miura, A., and Aramaki, E.: Use Trend Analysis of Twitter after the Great East Japan Earthquake, in *Proceedings of the 2012 ACM conference on Computer Supported Cooperative Work*, CSCW'12, pp. 175–178 (2012)

[Ritter 11] Ritter, A., Clark, S., Mausam, , and Etzioni, O.: Named Entity Recognition in Tweets: An Experimental Study, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 1524–1534, ACL (2011)

[Walther 13] Walther, M. and Kaisser, M.: Geo-spatial Event Eetection in the Twitter Stream, in *Proceedings of the 35th European conference on Advances in Information Retrieval*, ECIR'13, pp. 356–367, Springer-Verlag (2013)

[Watanabe 11] Watanabe, K., Ochi, M., Okabe, M., and Onai, R.: Jasmine: a Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs, in *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pp. 2541–2544, ACM (2011)