

Twitter からの被災時の行動経路の自動抽出および可視化

石野亜耶

小田原周平

難波英嗣

竹澤寿幸

広島市立大学大学院 情報科学研究科

{ishino, odawara, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

1. はじめに

2011年3月11日14時46分に発生した東日本大震災では、日本で観測史上最大の Mw (モーメントマグニチュード) 9.0 を記録し、岩手県から千葉県までの8県におよぶ広域で震度6弱以上の強い地震動が発生した。東北日本の太平洋側で10メートルを超える大津波が襲来、東京電力福島第一原子力発電所の事故にともなう大規模な避難が行われるなど、我々が経験したことのない大規模かつ複合的な大災害となった¹。このような災害時には、被災者の避難経路や救援物資の配送に利用可能な経路の情報は非常に重要な情報である。

また、震災発生時以降、手軽に情報を発信できるツールとして注目を集めている Twitter 上では、安否確認を求めるツイートや、被災地の交通情報を発信するツイートなど、震災に関連する様々な情報が発信された。自然言語処理研究者を中心としたプロジェクト ANPI NLP[1]では、Twitter などから、個々に述べられている人の安否情報に注目し、Google社の“Google Person Finder”上のデータと照合しながら、最新の安否情報を整理する活動が行われた。本研究では、ANPI NLPで提供されている東日本大震災情報に関連する Twitter のデータから、機械学習を使用して、ユーザの行動経路を抽出する手法を提案する。震災時の行動経路のデータベースを作成することで、被災者や、輸送物資を配送する援助者へ、利用可能な経路情報などの有益な情報を提供できることが期待できる。また、得られた行動経路を地図上にマッピングする事により、集約してユーザに提示し、閲覧できるシステムを開発する。

本論文の構成は以下の通りである。2節ではシステム動作例、3節では関連研究、4節では提案手法、5節では実験結果と考察について述べ、6節で本稿をまとめる。

2. システム動作例

本研究で構築した震災時の行動経路提示システムについて、その動作例を紹介する。図1は行動経路提示システムの画面である。図1には、約50件の行動経路をマッピングしてある。

行動経路として、ユーザが移動元(図中①)から移動先(図中②)に移動したことがわかるようになっている。また避難経路には、移動手段に応じたアイコンが表示されている。図中③の場合は、移動手段としてバスが用いられているため、バスのアイコンが表示されている。また、アイコンをクリックすると、詳細な情報(図中④)が表示できるようになっている。

本研究では、ANPI NLPで提供されている Twitter データから、ユーザの行動経路として、移動元、移動先、移動手段を抽出する手法を提案する。



図1 行動経路提示システムの動作例

¹ <http://www.mlit.go.jp/saigai/bosaijoho/>

3. 関連研究

被災地域内での移動の参考となる情報を提供するサイトに、「自動車通行実績情報マップ²」、「通れた道マップ³」がある。これらのサイトでは、GPS を搭載した車両から収集した通行実績情報が公開されている。

Twitter から震災関連の情報を抽出する研究として、Sakaki ら[2]の研究がある。Sakaki らは、Twitter のツイートを観測することにより、即座に地震発生地を検出を行う手法を提案している。Sakaki らはこの研究の有効性を示し、地震発生地検出アプリケーションとして、地震リポートシステムを開発している。Sakaki らは、地震の発生地抽出を行うシステムの開発を行っており、本研究での行動経路の抽出とは異なる。

Web から行動経路を抽出するための研究がいくつかある。郡ら[3]は、ブログからユーザの行動時の代表的な経路とその文脈を抽出し、それらを地図上にマッピングすることにより、集約して提示するシステムを提案している。システムとしてはまず、ブログ内に現れる各地名が、実際にビジターがその場所を訪れたという文脈で使用されているかを判定し、訪れていると判定された場合はその地名を破棄するという地名フィルタを作成する。次に、作成した地名フィルタにより取得した各ブログエントリの「訪れた地名」に順序づけを行い、地名によるシーケンシャルパターンを生成し経路抽出を行っている。郡らは、研究対象を京都に関するブログに限定としているが、本研究では震災関連のツイートを対象としている点で異なる。Davidov ら[4]は、Web から交通手段や経路の地理的なネットワークを見つける手法を提案している。Davidov らは、種となる特定の地域を表す項の小さな集合が与えられると、その地域の場所の名前とともに、連結性と交通に基づいたグラフを発見するアルゴリズムを提案している。このアルゴリズムでは、“[Transport] from A to B”のような“AからBへ”という表層パターンを使用して経路抽出を行っている。Davidov らは“from A to B”という限られたパターンで抽

2

http://www.google.co.jp/intl/ja/crisisresponse/japanquake2011_traffic.html

³ <http://g-book.com/pc/spot/Tohoku-Jishin.asp>

出をおこなっているが、本研究では機械学習を利用して行動経路抽出を行うため、より多くのパターンの行動経路抽出が可能であると考えられる。

本研究と同様に、機械学習を使用して行動経路を抽出する研究がある。Ishino ら[5]は、機械学習を利用することで、旅行ブログから、旅行者の行動経路を抽出する手法を提案している。Ishino らは、旅行者の行動経路を抽出するため、以下の5種類のタグを使用している。

- FROM：移動元
- TO：移動先
- VIA：経由地、経由道路など
- METHOD：移動手段
- TIME：移動に要した時間

Ishino らは、機械学習にCRFを使用している。機械学習には、以下の一般的な行動経路の抽出に使用できる11種類の素性と、旅行者の行動経路の抽出に特化した4種類の素性を使用している。

■ 一般的な行動経路の抽出に利用できる素性

- 単語
- 品詞
- 括弧（「」、「』」など）
- from：“から”、“を出發”などFROMの手掛かり語かどうか（40語）
- to：“まで”、“に到着”などTOの手掛かり語かどうか（271語）
- via：“経由”、“を通過”などVIAの手掛かり語かどうか（43語）
- via2：高速道路（101件）
- method：“飛行機”、“自動車”などMETHODの手掛かり語かどうか（148語）
- method2：乗り物リスト（128件）
- method3：列車、高速バスの愛称・路線名（2,033件）
- time：“分”、“時間”などTIMEの手掛かり語かどうか（77語）

■ 旅行者の行動経路の抽出に特化した素性

- from_to1：“博物館”、“温泉”など観光名所の後に付きそうな単語（45語）
- from_to2：“観光”、“駅”など行き先の後に付きそうな単語（11語）
- from_to3：観光名所リスト（13,779件）
- from_to4：駅名・空港名（9,437件）

Ishino らは、上記の手法により、精度 80.3%、再現率 46.8%を得ている。Ishino らは、旅行ブログを対象としているが、本研究では Twitter を利用する点で異なる。また、Ishino らは、旅行者の行動経路の抽出に特化した素性を利用している。しかし、本研究では、震災時の行動経路の抽出を目的としているため、震災時の行動経路の抽出に特化した素性を機械学習に使用する。

4. Twitter からの被災時の行動経路の自動抽出

本研究では、ANPI NLP で提供されている震災情報に関連する Twitter データから、行動経路を抽出する手法を提案する。4.1 節では、行動経路の自動抽出手法、4.2 節では、機械学習を使用した行動経路の自動抽出について説明を行う。

4.1. 行動経路の自動抽出手法

本研究では、図 2 に示した行動経路情報を含んだツイートから、機械学習を用いて行動経路を抽出する手法を提案する。

(ツイート例 1)
東松島市野蒜字下沼の親戚は、鳴瀬浅井公民館に避難しているとのことです。
(ツイート例 2)
弟が車で実家の矢本に行き 情報を得てきましたので、よろしければ私の TL 読んでください。

図 2 行動経路を含んだツイート例

Ishino らの旅行ブログから旅行者の行動経路を抽出する手法では、3 節で示した 5 種類のタグを使用している。しかし、震災時には、目的地である移動先以外の場所を経由する 경우가少ないといった特徴がある。そのため本研究では、VIA タグは使用せず、経由地や経由道路にも TO タグを付与する。また、ツイートは 140 文字という文字制限があるため、時間情報まで記述されているツイートは少数であった。よって本研究では、TIME タグは使用しない。以上の特徴を考慮し、Twitter から行動経路を抽出するために、以下の 3 種類のタグを定義する。また、図 2 のツイートにタグを付与した例を図 3 に示す。本研究では、ツイート内の各単語に、以下で定義するタグを付与するという系列ラベリング問題として考え、機械学習を

用いてタグの自動付与を行う。

- FROM：移動元
- TO：移動先、経由地、経由道路
- METHOD：移動手段

(ツイート例 1)
<FROM>東松島市野蒜字下沼</FROM>の親戚は、<TO>鳴瀬浅井公民館</TO>に避難しているとのことです。
(ツイート例 2)
弟が<METHOD>車</METHOD>で実家の<TO>矢本</TO>に行き 情報を得てきましたので、よろしければ私の TL 読んでください。

図 3 図 2 のツイートにタグを付与した例

4.2. 機械学習を使用した行動経路の自動抽出

本研究では、機械学習として CRF を使用した。CRF 基本手法は与えられた文に含まれる語を分類するのに使用した。素性とタグは以下のように CRF に与える。

- (1) ターゲットとなる単語から、CRF に与える前後の単語数 k
- (2) ターゲットとなる単語の前に存在する、ターゲットからの距離が k 以内に現れる単語
- (3) ターゲットとなる単語の後に存在する、ターゲットからの距離が k 以内に現れる単語

我々は予備実験の結果から、 $k=4$ と定めた。また、Ishino らが使用した、一般的な行動経路の抽出に使用できる 11 種類の素性から、TIME に関する素性を除いた 10 種類の素性に加え、以下に示す震災時の行動経路に特化した 4 種類の素性を使用する。

- hinan_from：“から避難”、“で被災”など FROM の手掛かり語かどうか (10 語)
- hinan_to：“小学校”、“公民館”など避難所に使用される施設名 (22 語)
- hinan_to2：“へ避難”、“へ逃げ”など TO の手掛かり語かどうか (5 語)
- hinan_method：“で救助”など METHOD の手掛かり語かどうか (3 語)

5. 実験

◆データセット

ANPI NLP で提供されている震災情報に関連する Twitter のデータ 49,263 件を使用した。Twitter のデータには、行動経路が含まれるツイート以外にも、被災地で活動を行う自衛隊への応援メッセージや、募金活動に関する情報、デマ情報に関する注意喚起などのツイートが大量に含まれる。そのため、本研究では、ANPI NLP で提供されている Twitter データから、“移動”、“避難”など行動経路を記述される際に使われやすい単語が含まれるツイート 1,314 件を実験対象とした。このツイートの FROM、TO、METHOD タグを手手で付与したデータを、機械学習に用いた。手手で付与したタグの数を表 1 に示す。

また、比較手法として、3 節に示した、Ishino が旅行ブログから旅行者の行動経路を抽出する際に使用した素性を用いる実験を行った。

表 1 手手で付与したタグの数

	訓練用 (件)	テスト用 (件)
FROM	235	57
TO	399	102
METHOD	61	18

◆機械学習と評価尺度

機械学習には CRF を用いた。また、精度と再現率を用いて評価を行った。

◆実験結果

実験の結果を以下の表 2 に示す。表 1 より、比較手法に比べ、提案手法が精度で 2.5 ポイント、再現率で 0.5 ポイント向上した。

表 2 行動経路の自動抽出の実験結果

	提案手法		比較手法	
	精度 (%)	再現率 (%)	精度 (%)	再現率 (%)
FROM	38.2	22.8	47.1	28.1
TO	48.1	36.3	46.8	35.3
METHOD	81.8	50.0	66.7	44.4
平均	56.0	36.4	53.5	35.9

FROM、TO の精度低下の原因について考察する。震災時の行動経路については、図 3 のツイート例 1 のように、地名や施設名が詳細に記述される傾向がある。そのため、形態素解析の失敗などの原因により、正しくタグが付与できない場合が多くあった。この問題は、Graham ら [1] が使用した、東北地方に特化した地名辞書を手掛かり語として使用することで解決できると考えられる。

6. おわりに

本研究では、ANPI NLP で提供されている震災関連の Twitter データから、ユーザの行動経路を抽出する手法を提案した。素性には、震災時の行動経路の抽出に特化した手掛かり語を収集し、素性として使用した。その結果、精度 56.0%、再現率 36.4% を得た。震災時の行動経路の抽出に特化した素性を使用しなかった場合に比べ、精度 2.5 ポイント、再現率 0.5 ポイント改善することができた。また、提案手法により得た経路情報を利用し、マッピングし提示する行動経路提示システムの構築を行った。

参考文献

- [1] Graham Neubig, Yuichiro Matsubayashi, Masato Hagiwara and Koji Murakami. Safety Information Mining — What can NLP do in a disaster —. Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP), pp. 965-973, 2011.
- [2] Takeshi Sakaki, Makoto Okazaki and Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. 19th International World Wide Web Conference (WWW2010), 2010.
- [3] 郡 宏志, 服部 峻, 手塚 太郎, 田島 敬史, 田中 克己. ブログからのビジターの代表的な経路とそのコンテキスト抽出. 情報処理学会研究報告データベースシステム研究会, Vol.2006, No.78, pp.35-42, 2006.
- [4] Dmitry Davidov and Ari Rappoport. Geomining : Discovery of Road and Transport Networks Using Directional Patterns. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp.267-175, 2009.
- [5] Aya Ishino, Hidetsugu Nanba and Toshiyuki Takezawa. Automatic Compilation of an Online Travel Portal from Automatically Extracted Travel Blog Entries. Proceedings of the 18th international Conference on Information Technology and Travel & Tourism (ENTER2011), pp.113-124, 2011.