F-6

災害関連ツイート要望・対応策マッチングコーパスの作成

Constructing a Corpus of Requests and Treatments in Tweets during a Crisis

佐野 大樹†　　イシュトバン・ヴァルガ†　　風間 淳一†
Motoki Sano　　István Varga　　Jun'ichi Kazama
橋本 力†　　鳥澤 健太郎†
Chikara Hashimoto　　Kentaro Torisawa

1．はじめに

東日本大震災において、Twitter は安否確認や孤立した被災者の救助に活用され（徳田 2011）、災害時に有効なコミュニケーション手段、また、言語資源として着目されている（Neubig and Matsubayashi, 2011, Okazaki and Matsuo, 2011, Sakaki, Toriumi and Matuso, 2011, Sano, Varga, Kazama and Torisawa, 2012）。震災発生後インターネットサービスが利用できなかった地域もあるが、情報支援プロボノ・プラットフォーム（2012）による面談調査（*N*=186）では、「役に立ったインターネットサービス（発生後数時間、複数回答）」として、Twitter と回答とした人の割合が 10.8%であり、Yahoo! (4.8%)、mixi（4.8%）、Google（3.8%）と比べて高い割合を占めている。

このような高い評価を受けた理由の 1 つとして、Twitter は新聞やテレビなどの他のメディアに比べ、地理的に限定された地域や特定の個人に関する情報を発信・入手しやすいという特徴が挙げられる。先述した情報支援プロボノ・プラットフォーム（2012）の調査では、Twitter に対して「災害情報やインフラ関係など、さまざまな情報を得ることができた。友人の安否確認もできた。ボランティアの情報を得られた」（岩手県 男性）「ガソリンや食料などの生活物資に関する情報収集にはラジオ福島の HP やツイッターがとても役に立ちました」（福島県 女性）などといった意見があり、要望を満たす情報を含むツイートを効率的に抽出し発信者に届けられるのであれば、被災者が抱える問題の解決に繋がる身近で実用的な対応策を Twitter の情報を活用することで提供できるようになると考えられる。

そこで我々は、要望を発信したユーザには解決策を、解決策を発信したユーザには要望を自動伝達するシステムの構築を進めている。このシステムは、例えば「*○○市に住んでいるものです。粉ミルクを避難所に送ってください。*」という要望に対して「*○○市では、粉ミルクの配布を行っています。お困りの方は、○○市△△小学校まで*」といった、要望に含まれる表現（*粉ミルク*）と同じ、もしくは、類似した表現を含むツイートをマッチさせるだけでなく、「*腎臓病を患っている父のため、○○避難所から離れられません*」といった問題提起を介する間接的な要望に対して「*○○病院にて、24 時間透析治療を受けられます。連絡先×××-×××*」といったツイートをマッチさせるなど、問題（「腎臓病」）から予想される対処方法（「透析」）を Web データから得た知識を用いて推測し、対応策とマッチさせることを可能とする。

本稿では、この計画の一部として行われている「要望-対応策マッチングコーパス」構築の概要と方法について説明する。なお、問題解決を求める要望を含むツイートを「要望ツイート」、要望を満たす対応策に関する情報を含むツイートを「対応策ツイート」、「*粉ミルクを避難所に送ってください。*」のように明示的にどのような対策をとってほしいか表す要望を「直接要望」(direct request)、「*腎臓病を患っている父のため、○○避難所から離れられません*」のようにどのような対策をとってほしいか明示されておらず、問題提起によって対応策の必要性を示唆する要望を間接要望　(indirect request)とよぶ（Sano et al., 2012）。

2．要望-対応策マッチングコーパスの概要

「要望-対応策マッチングコーパス」は構築中のシステムの評価用データとして作成しているもので、1 つの要望ツイートに対して人手で特定した対応策ツイートを複数収録している。例えば、間接要望「*避難所に入るのを遠慮して車で生活している人に、エコノミー症候群になる人が増えてきています*」という間接要望に対して、以下のような対応策ツイートがマッチされる。

> **対応策ツイート 1**: *【低体温症•エコノミー症候群を防ぐ】避難所の皆様、少しでも体を動かしましょう☆タオルつかみ☆足下にタオルや毛布を敷く。足指先で布をつかみ手前にひく。左右交互にいっちに、いっちに。末端を動かすとポカポカしてきます。足指ジャンケンも効果的！トイレが気になると思いますが水分補給もね。*
>
> **対応策ツイート 2**: *東北地方太平洋沖地震の被災者の皆様へ−静脈血栓塞栓症、ロコモティブシンドロームの対策について−　－　Ｉｎｆｏｓｅｅｋ　ニュース　https://×××　日本整形外科学会　https://×××*
>
> **対応策ツイート 3**: *ミカンの中の筋にはヘスペリジンが含まれ血管の血栓溶解作用もあります。是非、食べてください　(https://×××)*

対応策ツイート 1 は、要望ツイートと同じ「エコノミー症候群」という表現が、対応策ツイート 2 ではエコノミー症候群の別名「静脈血栓塞栓症」が、対応策ツイート 3 には、「エコノミー症候群」という表現もその別名も含まれていないが、エコノミー症候群に対する対応方法の「血栓溶解」という表現が含まれる。このように、要望-対応策マッチングコーパスでは、災害時に対応が必要とされた一次的・二次的な問題に関する要望ツイートとマッチする対応策ツイートを、別名、同義語、予防・対処方法を考慮し収録している。

-
†情報通信研究機構,
National Institute of Information and Communications Technology

3. コーパス構築の方法と手順

3.1 データ

要望ツイートと対応策ツイートの抽出には、2011 年 3 月 10 日から 2011 年 4 月 4 日までに投稿された東日本大震災関連の約 2 億 2 千万ツイート（提供元:（株）ホットリンク 120GB）を使用している。データには、#jishin、#jisin、#hinan 、 #earthquake 、 #tsunami 、 #anpi 、 #jishin_e 、 #save_ibaraki、などのハッシュタグをもつツイートや、「デマ」「募金」「義援金」などをキーワードとして収集されたツイートが含まれ、収集対象ユーザ数は約 140 万件である。

3.2 データの検索

コーパス構築にあたって、要望ツイートと対応策ツイートを検索するためにオープンソースの Apache Solr を利用して作成した全文検索エンジンを用いた。ただし、Solr のランキング機能は使用しておらず、検索キーワードを含む（AND 検索）ツイートを無作為に表示するようにした。ラインキング機能を使用しなかったのは、デフォルトのスコアリング方法では、文字数が少ないツイートのほうが多いツイートに比べて高いスコアを得るように設計されているためである。例えば、「水」というキーワードで検索した場合、ツイート本文が「*水不足*」のものと「*○○市の△△小学校の避難所では、水が不足*」のものとでは、前者のほうが上位にランキングされてしまい、意味のある要望や対応策がコーパスにほとんど含まれないということになってしまう。このような偏りを防ぐため、スコアリングの結果、抽出される要望ツイート・対応策ツイートに偏りがでないように、無作為に結果を表示するようにした。

3.3 手順 1: 地震関連問題リストの作成

要望-対応策マッチングコーパスを構築するため、まず、地震災害で直面する問題のリストを作成した（以下、「地震関連問題リスト」）。地震関連問題リストを作成したのは、地震災害に関連する問題キーワードのうち構築中のシステムがどの程度カバーできるのかを評価できるようにするためである。3 名のアノテータに以下の指示を与え、作成することにした。

- ●自分が地震災害にあったと仮定して、具体的で（「病気」「物資不足」など抽象的なものは避ける）緊急性が高い問題（停電、断水など）、及び、問題の対象となるもの（水、ガソリンなど）をリストアップする。
- ●一次的な被害だけでなく、二次的な被害もリストアップする（例：エコノミー症候群）
- ●問題、及び、問題の対象を連想した場合は、連想したことが、実際、災害時に問題となっていたというような発言があるかないか、Web で調べる。裏付けとなる URL を作業ファイルに記載する。
- ●Web サイトを調べて、問題、及び、問題の対象となるものをリストアップしてもよい。
- ●各自 10 時間かけて、 連想できる、調べられる限りのものを、リストアップする。（時間は厳密に測る）
- ●問題は、名詞、もしくは、名詞句で表す。

アノテータ 3 名がそれぞれ作成したリストをマージした結果、359 件（異なり）の問題キーワードリストを作成できた。例を以下に示す。なお、問題キーワードには、Hashimoto et al. (2012)の excitation の概念に基づき「+」（活性）か「-」（不活性）というラベルが付与されている。「+」は、問題キーワードに該当する対象が発生・存在・効果（及びその準備）を発揮することで問題となるもの（停電、断水など）に、「-」は、問題キーワードに該当する対象が回避される、消滅、不足、弱体化することが問題となるもの（水、ガソリンなど）に付与されている。

> *ヘドロ +, 下着 -, 肺血栓塞栓症 +, 放射能 +, 米 -, インシュリン -, 血液 -, 看護士 -, アルコール依存症 +, 洗濯 -, 電池 -, 粉ミルク -, 寝具 -, 注射針 -*

3.4 手順 2: 要望リストの作成

地震関連問題リストを作成した後に、「要望リスト」の作成を行った。要望リストは、地震関連災害リストにあげられた問題について、直接要望と間接要望となるツイートをそれぞれ 1 つ人手で集めたものである。例えば「風邪薬 -」という問題キーワードに対しては、以下の要望ツイートが特定された。

> **直接要望**: *【○○町情報】心臓病、高血圧、糖尿病、風邪薬――薬が全般的に欠乏している。手配求む。*
> **間接要望**: *○○市の友人より。乾麺、粉もの小麦粉や、ホットケーキミックスなど、日持ちのできるもの。子供用、大人用の風邪薬や生理用品、赤ちゃんのオムツ、カセットコンロのガス。これらが不足しいてます。*

作業はアノテータ 4 名が担当し、3.2 で述べた検索エンジンを用いて要望ツイートを検索した。アノテータが使用したキーワードは、問題キーワードのみである。「～ください」「～ほしい」などの要求表現を検索キーワードに追加することで、要望ツイートの検索は容易になるが、そのような方法を用いた場合、抽出される要望ツイートには類似した要求表現が含まれ、多様な表現形式をカバーできない可能性がある。そこで、無作為に表示される検索結果を上位から確認していき、一番最初に特定した直接要望・間接要望を要望ツイートとして特定してもらうことにした。

抽出された要望ツイートは、問題キーワードに関連しない要望を含むか否かによって「single」（単数）か「multi」（複数）に分類されている。例えば、先述した間接要望では「風邪薬」以外にも「乾麺」「生理用品」などが要望されているため「multi」と分類される。なお、要望される行為によって、直接要望が好まれるか、間接要望が好まれるかは異なる（Sano et al., 2012）。このため、直接要望か間接要望のいずれかしか要望ツイートが見つからない問題キーワードもある。また、ツイートデータからは要望ツイートが見つからないものもある。

3.5 手順 3: 要望ツイートと対応策ツイートのマッチング

要望リストの構築後、要望リストに収録された要望ツイートと対応策ツイートをマッチさせる作業を開始した。他と同様に、この作業も人手で行われている。マッチングの例については、2節にて説明した通りである。

要望リストの作成では直接要望・間接要望を1つずつ収集したが、本作業では特定できた全ての対応策ツイートを収録することにした。検索は、以下に示す方法で行われる。

●各要望ツイートについて、10回検索をする。
●10回のうち、3回は指定された問題キーワードを含む検索を行う。問題キーワード以外にも、場所情報や対応策を探すのに各アノテータが妥当だと考える動詞を利用する。
●2回は問題キーワードの同義語（別名）を含む検索を行う。
●残りの5回は、問題キーワード（異表記・同義）を含まない検索を行う。

例えば、問題キーワードが「エコノミー症候群」の場合、「エコノミー症候群」を検索キーワードに含む検索を 3 回し、「エコノミー症候群」の同義語を含む検索（例えば、「静脈血栓塞栓症」）を 2 回、残りの 5 回は「エコノミー症候群」や「静脈血栓塞栓症」を含まない検索をする。

アノテータは、それぞれの検索で、検索結果上位 30 件に含まれる対応策ツイートを全て特定し収録する。基本的に、要望ツイート 1 つについて 300 件（10 回×上位 30 件）のツイートを、問題キーワード別では 600 件（直接要望 300 件・間接要望 300 件）を確認することになる。

なお、同義語や対処方法など、検索で利用するキーワードを考えるために、Web 上の情報を利用することを許可した。例えば、「エコノミー症候群」に関する要望の場合、エコノミー症候群の対処方法を「エコノミー症候群」「対処」などのキーワードを用いて Web 上で検索すると「血栓溶解」といったような対処方法を見つけることができることを作業マニュアルに記載した。

特定された対応策ツイートは、要望リストの場合と同様に、問題キーワードに関する要望以外の要望の対応策として機能できるか否かによって「single」（単数）か「multi」（複数）に分類される。例えば、2 節の対応策ツイート 1 は、「低体温症」の対応策にもなるため「multi」と分類される。また、このような分類以外にも、対応策ツイートは以下に示す 3 つのうちいずれかに分類されている。各分類の定義と例を示す。

1.どのような対応策で解決されたかは記載されていないが、問題が解決したことを示すツイート
- **要望ツイート:** *○○県○○市○○からの情報です。やっと電気が通ったとの連絡が入りました。ガスは三週間位無理そうです お水も出ないそうです*（問題キーワード「水」）
- **対応策ツイート:** *水が出たＴＴ　@○○県○○市*

2. どのような対応策で解決されるかは記載されていないが、問題が解決することを示すツイート
- **要望ツイート:** *【○○県】○○町はペットボトルの水やお茶、紙コップ、紙皿、使い捨てカイロ、紙オムツ、毛布、缶詰・レトルト食品等を募集*（問題キーワード「お茶」）
- **対応策ツイート:** *○○市からの救援物資、紙オムツやお茶などが自衛隊の車に詰まれて出発してました。明日には到着するんじゃないかな。*

3.どのような対応策で解決された・されるか記載されているツイート
- **要望ツイート:** *家を失って、家族が亡くなって、食料が届かず、寒くて凍死‥‥。信じたくないけど、○○でそれが起こってる。*（問題キーワード「凍死」）
- **対応策ツイート:** *東日本大震災で、体育館や教室は床がフローリングで石や木ですから熱を下から奪われます アルミの保温シートやアルミホイルをセロテープで張り合わせて段ボールの底に貼っても効果的です*

4. まとめと今後の展望

本稿では、要望-対応策マッチングコーパスの概要と構築方法について述べた。要望ツイートと対応策ツイートをマッチングできるシステムを実現し、災害時に発生する問題の解決に貢献できるよう計画を進めていく予定である。

謝辞

本研究で利用しているデータは（株）ホットリンク様よりご提供頂いた。ここに記して、感謝致します。

参考文献

Hashimoto, C., K. Torisawa, S.De Saeger, J. Oh and J. Kazama. (2012). Excitatory or Inhibitory: A New Semantic Orientation Extracts Contradiction and Causality from the Web, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 619–630.

情報支援プロボノ・プラットフォーム (2012). 『３.１１被災地の証言 東日本大震災 情報行動調査で検証するデジタル大国・日本の盲点』インプレスジャパン.

Neubig, G. and Y. Matsubayashi (2011). Safety information mining-What can NLP do in a disaster-. *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pp.965-973.

Okazaki, M. and Matsuo, Y. (2011). Semantic twitter: Analyzing tweets for real-time event notification. in J.G. Breslin et al (Eds.) *BlogTalk* 2008/2009, LNCS 6045. pp.63-74, Berlin: Springer-Verlag.

Sakaki, T., F. Toriumi and Y. Matuso (2011). Tweet trend analysis in an emergency situation. *Proceedings of the Special Workshop on Internet and Disaster*.

Sano, M., I. Varga, J. Kazama, and K. Torisawa (2012). Requests in tweets *during* a crisis: A systemic functional analysis of tweets on the Great East Japan Earthquake and the Fukushima Daiichi nuclear disaster. *Papers from the 39th International Systemic Functional Congress*, pp.135-140.

徳田雄洋 (2011). 『震災と情報−あのとき何が伝わったか』岩波書店.