2B4-NFC-02b-3

東日本大震災時のツイートデータからのイベントの因果関係の抽出

Causality Analysis of Events Using Tweets in Great East Japan Earthquake

風間 一洋[*1] Kazuhiro Kazama　高橋 公海[*2] Masami Takahashi　鳥海 不二夫[*3] Fujio Toriumi　榊 剛史[*3] Takeshi Sakaki　栗原 聡[*4] Satoshi Kurihara　篠田 孝祐[*5] Kosuke Shinoda

野田 五十樹[*6] Itsuki Noda

[*1]和歌山大学 Wakayama University　[*2]日本電信電話 (株) NTT　[*3]東京大学 The University of Tokyo　[*4]電気通信大学 The University of Electro-Communications

[*5]慶應義塾大学 / 理化学研究所 Keio University / Riken　[*6]産業技術総合研究所 AIST

This paper presents a method to extract causal relationships of events from Twitter. We extracted event-specific words, which are frequently used in a specific period, from tweet archives. Next, we make a series of event-specific words for each user and make a transition relationship matrix by counting their anteroposterior relationships between event-specific words. Existence or nonexistence of causality, its direction, and its strength are determined by analyzing a transition relationship matrix. In fact, we make a causal realationship network from tweet archive in the Great East Japan Earthquake and visualize it.

1. はじめに

近年のソーシャルメディアの普及により，単に自分の考えや生活に関するメッセージの投稿だけでなく，ソーシャルグラフを活用した情報収集やメッセージ交換などもおこなわれるようになってきた．このようなソーシャルメディア上の行動は日常生活のかなりの割合を占めることから，ソーシャルメディア上の情報を整理・再構成することで，人間の行動パターンや実世界で発生しているイベントを推測することができる．

本稿では，大規模な事件に関連して発生する多数のイベント群とその間の因果関係を，Twitter のツイートから抽出する手法を提案する．まず，Twitter のツイートアーカイブから，東日本大震災の後に特に頻繁につぶやかれるようになった単語を，地震によって発生したイベントを表すイベント関連語として抽出する．次に，各ユーザごとにそのイベント関連語の前後関係をカウントし，その出現頻度から単語出現の因果関係の有無と方向を決定する．実際に，2011 年 3 月 11 日に発生した東日本大震災の前後のツイートアーカイブを用いて因果関係を抽出し，その可視化結果を示す．

2. イベントの因果関係の分析

2.1 インターネットデータからの因果関係の抽出

インターネットから入手できるデータからの因果関係抽出に関しては，さまざまな研究が存在する．例えば，佐藤らは Web 上の膨大なデータの複文や重文を分解して単一の事象を表す単文を抽出し，その間の因果関係の強さを調べて因果ネットワークを抽出する手法を提案した [佐藤 06]．また，中島らは Web から収集した時系列データから，接続標識などの手がかりと，各季節のイベントの出現情報や共起情報を機械学習し，時期依存性を持つイベント連鎖を抽出する手法を提案した [中島 13]．

これに対して，Twitter のツイートデータは，一般的な Web ページよりも口語的であり，最大 140 文字の制限と即時性の強さから，ある一連のストーリーを構築する文章が分割して投稿される傾向が強いことから，新たな手法の開発が必要である．. そこで，本稿では，イベントを文章として取り出す代わりに，イベント関連語を抽出し，二つのイベント関連語間の遷移頻度から因果関係を推定する．

2.2 データセット

3 月 5 日から 24 日の間に，Twitter API[*1]を用いて 200 件以上日本語でツイートしたアクティブなユーザのツイートを収集し，さらに収集漏れを減らすために，後日各ユーザに対して再収集し，それをデータセットとして使用した．200 件は，Twitter API の呼び出し 1 回で取得できる最大ツイート数である．データセットの規模は，ツイート数が 362,435,649 件，ユーザ数が 2,711,473 人である．データセットには，ツイート ID（64 ビット整数），ツイートしたユーザのスクリーン名，本文，ツイート元．ツイート時間，リプライ先のツイート ID，リプライ先のスクリーン名が含まれる．

2.3 ツイートからの単語の抽出

まず最初に，ツイート関連語の候補となる単語を抽出するために，ツイートのテキストから URL，ハッシュタグ，スクリーン名を除去してから，Mecab [*2]で日本語形態素解析し，非自立，数，接尾，ナイ形容詞語幹を除く名詞を抽出した．ただし，発言後も別のユーザのツイート中に繰り返し出現する公式リツイート及び非公式リツイートの元メッセージ部分は削除した．なお，新語や流行語，専門用語なども複合語として抽出されるように，標準の IPA 辞書に加えて，はてなキーワード[*3]や原子力百科事典 ATOMICA[*4]の用語を辞書に追加した．

連絡先: 風間 一洋，NTT 未来ねっと研究所，〒 180–8585 東京都武蔵野市緑町 3–9–11，kazama@ingrid.org

*1 http://apiwiki.twitter.com/
*2 http://mecab.sourceforge.net/
*3 http://d.hatena.ne.jp/keyword/
*4 http://www.rist.or.jp/atomica/

2.4 イベント関連語の抽出

東日本大震災と関連が深い単語だけを対象にするために，次の 3 つの条件を満たす名詞を抽出し，イベント関連語とした．

1. 地震発生から一週間以内の出現ツイート数が 1,000 件以上
2. 一日の出現確率がピークの日が地震発生から一週間以内
3. ピークの日の出現確率が，地震発生前の 10 倍以上

2.5 イベント関連語の遷移頻度行列の作成

即時的・逐次的に発言される Twitter は，起こった出来事を後からまとめて書く場合と異なり，発言の完全性は期待できない．つまり，実世界で順番に発生した三つのイベントのイベント関連語を w_0, w_1, w_2 とした場合に，Twitter 上では必ずしも $w_0 \to w_1 \to w_2$ のように発言されると限らない．例えば，あるユーザは $w_0 \to w_2$ だけを，別のユーザは $w_1 \to w_2$ だけを発言するかもしれない．

まず，各ユーザごとに $w_i \to w_j \to w_k \to w_l$ のようなイベント関連語の発言順の系列を作成する．なお，イベント関連語はイベント発生後は何度も繰り返し発言される傾向があるので，初出の場合のみ記録する．

次に，得られたイベント関連語系列を $w_i \to w_j$ のような二つの単語間の遷移関係に分解して，その遷移頻度 f_{ij} を計算し，遷移頻度行列 F を作成する．なお，単語の初出だけしか考慮しないので同じ単語同士の遷移はなく，対角成分は 0 となる．

$$F = \begin{bmatrix} f_{0,0} & \cdots & f_{0,n-1} \\ \vdots & \ddots & \vdots \\ f_{n-1,0} & \cdots & f_{n-1,n-1} \end{bmatrix} \quad (1)$$

2.6 因果関係の決定

遷移頻度行列 F の各要素の頻度の総和を N とすると，この行列で単語 w_i から単語 w_j に遷移する確率は $p(w_i \to w_j) = f_{i,j}/N$ となるが，これはツイート中に出現したすべての単語 w_i と単語 w_j に対する確率である．つまり，単語の出現頻度 f_i, f_j の影響を受けており，特に $f_i \ll f_j$ のように大きく異なる場合は，現実とは逆に $p(w_i \to w_j)$ より $p(w_j \to w_i)$ が大きくなるという問題がある．

そこで，正規化された単語 w_i から単語 w_j への移行確率 $\tilde{p}(w_i \to w_j)$ を，単語出現確率 $p(w_i)$ と $p(w_j)$ を用いて，次のように定義する．

$$\tilde{p}(w_i \to w_j) = p(w_i) \times p(w_i \to w_j) \times \frac{1}{p(w_j)} = \frac{f_{i,j} \times f_i}{N \times f_j} \quad (2)$$

ここで，$p(w_i) = f_i/M$（M は単語出現頻度の総和）である．

単語 w_i と単語 w_j の間の因果関係の有無は，$\tilde{p}(w_i \to w_j)$ と $\tilde{p}(w_i \to w_j)$ を比較して決定する．まず，$\tilde{p}(w_i \to w_j)$ と $\tilde{p}(w_i \to w_j)$ の値が近い場合には共起関係，値が大きく異なる場合には因果関係が存在すると見なす．現実には，抽出時のノイズの影響で，因果関係が存在しない方向の頻度が 0 になることはほとんどない．

さらに，因果関係の方向は，$\tilde{p}(w_i \to w_j) \gg \tilde{p}(w_i \to w_j)$ の場合は $w_i \to w_j$，$\tilde{p}(w_i \to w_j) \ll \tilde{p}(w_i \to w_j)$ の場合は $w_i \leftarrow w_j$ であると見なす．これは，$\tilde{p}(w_i \to w_j)/\tilde{p}(w_i \to w_j) \leq 1/T$ または $\tilde{p}(w_i \to w_j)/\tilde{p}(w_i \to w_j) \geq T$ の成立により判定する．閾値 T は正の整数である．

3. 因果関係ネットワークの可視化

実際に東日本大震災のツイートデータから抽出した因果関係ネットワークを図 1 に示す．閾値 $T = 10$ とし，$\tilde{p}(w_i \to w_j)$ が大きい順に 150 個の因果関係を抽出して，Cytoscape の Force-Directed Layout を用いて可視化した．なお，一部の単語に存在した表記の揺れは統一している．

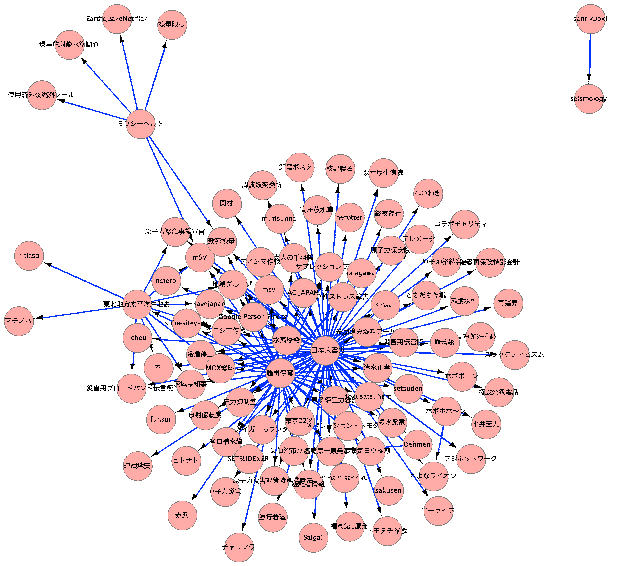

図 1: 因果関係ネットワークの可視化

この図から,「東日本大震災」,「東北地方太平洋地震」,「輪番停電」,「水素爆発」から，多くの震災関連語が導きだされていることがわかる．ただし，地震関係，原発関係，AC ジャパンの CM 関係など，東日本大震災には関連していても，互いの関連が少ないような複数のトピックが混在している問題がある．

4. おわりに

本稿では，同時多発的に発生したイベント群とそれらの因果関係を Twitter のツイートから抽出する手法について延べ，実際に東日本大震災時のツイートアーカイブから抽出した因果関係ネットワークの可視化結果を示した．

今後の課題としては，まず Latent Dirichlet Allocations（LDA）などの手法を用いたトピックを考慮したイベント関連語の抽出により，因果関係の分離と精度向上を計ることである．次に，イベント関連語と，それと同時に使われる補足語をまとめて単語グループとして表示することで，イベントの詳細を把握できるようにすることである．最後に，例えば $w_0 \to w_1 \to w_2$ と $w_0 \to w_2$ という因果関係を $w_0 \to w_1 \to w_2$ に統一するような，因果関係ネットワークの簡略化をおこなうことである．

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に感謝する．また，本研究は NTT との共同研究，および科研費（24300064）の助成を受けた．

参考文献

[中島 13] 中島 直哉, 吉永 直樹, 鍛治 伸裕, 豊田 正史, 喜連川 優：時期依存性を有するイベント連鎖の獲得, in *DEIM Forum 2013* (2013)

[佐藤 06] 佐藤 岳文, 堀田 昌英：Web マイニングを用いた因果ネットワークの自動構築手法の開発, 社会技術研究論文集, Vol. 4, pp. 66–74 (2006), Takefumi Sato and Masahide Horita