ツイートデータからのイベントの因果関係の抽出法

Event Causality Extraction of Events Using Tweets

風間 一洋[1*]　鳥海 不二夫[2]　榊 剛史[2]　栗原 聡[3]　篠田 孝祐[4,5]　野田 五十樹[6]
Kazuhiro Kazama[1], Fujio Toriumi[2], Takeshi Sakaki[2], Satoshi Kurihara[3],
Kosuke Shinoda[4,5], and Itsuki Noda[6]

[1] 和歌山大学 / Wakayama University,
[2] 東京大学 / The University of Tokyo,
[3] 電気通信大学 / The University of Electro-Communications,
[4] 慶應義塾大学 / Keio University,
[5] 理化学研究所 / Riken,
[6] 産業技術総合研究所 / AIST

Abstract: 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. Furthermore, we simplify an extracted graph structure by removing redundant causal edges. In fact, we make a causal relationship network from tweet archive in the Great East Japan Earthquake. We analyze the network structure and show that proposed method is suitable for extracting causal relationships.

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・再構成すれば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・把握できると考えられる．そこで，Twitter で毎日つぶやかれる膨大なツイートから，実世界で発生した因果関係のあるイベントの連鎖を抽出することができれば，実世界の状況を体系づけたり，概観することが容易になるはずである．

本稿では，実世界の事象に関連して発生するイベント群に関する時系列的な情報を，Twitter のツイート群から因果関係ネットワークとして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，注目したいイベントの関連語を抽出し，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て得られるイベント関連語の遷移頻度行列を用いて単語出現の因果関係の有無と方向を決定し，さらにネットワーク簡略化を適用して因果関係ネットワークを抽出する．実際に 2011 年 3 月 11 日に発生した東日本大震災に関するイベント関連語に関して抽出した因果関係ネットワークを評価して，その有効性を示す．

*連絡先：和歌山大学システム工学部
和歌山県和歌山市栄谷 930
kazama@sys.wakayama-u.ac.jp

2　ツイートからの因果関係抽出

Twitter は，社会的な要素を備えたコミュニケーションネットワークであり，誰もが参加できるソーシャルメディアである．そのリアルタイム性の高さから，例えば実世界で発生した出来事を観測するために用いられ，ソーシャルセンサ（Social Sensor）とも呼ばれている [Sakaki 10]．そこで，実世界で発生したイベントに関する情報を収集するだけでなく，それらの因果関係を分析できれば，Twitter で毎日つぶやかれる膨大なツイートから得られる情報を体系づけたり，俯瞰することが容易になるはずである．

このような因果関係の抽出には，因果関係の原因と結果の節を繋ぐ接続詞に着目する方法，構文パターンから抽出する方法，モダリティから因果関係の強さを

決定する方法など，文単位で分析する手法を使うことが多い．ただし，このような既存研究のほとんどが，新聞記事やブログ，Web ページなどの，比較的長い文章を扱うことを前提としている．

これに対して，Twitter のツイートは，そのリアルタイム性の高さと最大 140 文字の制限により，口語的な短い文章をほとんど校正せずに素早く投稿する傾向が強く，従来のアプローチでは因果関係を推定することは難しい．

そこで本稿では，文ではなく，単語に着目し，因果関係の決定に手がかり表現や構文パターンを使う代わりに，多数のユーザのツイートストリームにおける単語の出現順序を集計する集合知的なアプローチにより，イベントの因果関係を抽出する手法を提案する．

3 関連研究

テキスト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，それらの文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]．石井らは,「ため」や「を受けて」のような因果関係を示す手がかり表現を含む文節から抽出した事象 SVO 構造をマージする過程を繰り返すことで，ネットワークを増分的に構築する手法を提案した [石井 10]．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]．Sakaji らは，日経新聞の過去記事から手がかり表現と構文パターンを用いて因果関係を抽出する手法を提案した [Sakaji 08]．乾らは,「ため」を接続標識として用いて抽出した因果関係知識を抽出し,「事態」と「行為」の組み合わせにより因果関係を cause 関係，effect 関係，precond 関係，means 関係の 4 種類に分類した [乾 04]．青野らは，Web 文書から把握したい辞書を表す検索語と手がかり表現を用いて要因として抽出した事象をさらに要因検索することを繰り返し，階層的に獲得した因果関係を因果関係ネットワークとして可視化する方法を提案し，抽出された因果関係を分析した [青野 10]．澤村らは，東日本大震災に関する新聞記事から，10 種類の手がかり標識を用いて因果関係を抽出し，その結果と原因の語彙の一致を Jaccard 係数で調べて接続した後に．さらに HDP-LDA を用いて同じ潜在的トピックを持つ因果関係を接続し，因果関係連鎖を抽出する方法を提案した [澤村 13]．

既存研究は文の接続関係や因果関係を表す手がかり表現を用いることが多いが，本研究は文章長が短く，口語的で構文解析もうまくできない Twitter のツイートのような実データを想定して単語単位で扱う点，手がかり表現を用いない点，単語の出現の前後関係の統計的解析で因果性を求める点で異なる．

4 イベントの因果関係ネットワークの抽出

イベント関連語の出現系列を分解して得られるイベント関連語の組の出現頻度を元に，確率的に因果関係を推定すると共に，そのグラフ構造から冗長な因果関係を除去することで簡素化し，イベントの因果関係ネットワークを抽出する方法について述べる．

4.1 単語系列の抽出

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．

そこで，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．これはツイート間に限らず，同一ツイートまたは同一文内の場合でも，単語の順序を考慮する．この理由は，日本語では同一ツイートまたは同一文中で先に出現する単語が原因を，後で出現する単語が結果を示すことが多いからである．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こととする．

ここで，3 人のユーザが $w_0 \to w_1 \to w_2$ という因果関係を持つ単語系列をツイートする例を，図 1 に示す．w_0, w_1, w_2 はイベント関連語であり，矩形はツイートを示し，時間は左から右に流れるものとする．図 1 では，ユーザ 0 からは $w_0 \to w_1 \to w_2$，ユーザ 1 からは $w_1 \to w_2$ という単語系列が抽出される．これに対して，ユーザ 2 では $w_0 \to w_0 \to w_1 \to w_1$ という順序で単語が出現するが，初出だけを記録するので抽出される単語系列は $w_0 \to w_1$ となる．

4.2 単語の出現頻度行列の作成

n 個のイベント関連語 $w_i (0 \leq w_i \leq n-1)$ があるとする．まず，単語 w_i の出現頻度 f_i を求めて，単語頻

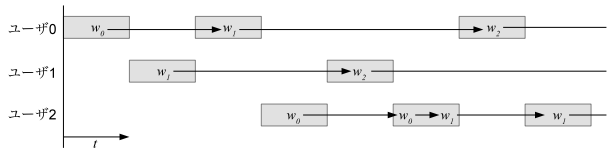

図 1: 単語系列の例

度行列 W を作成する．

$$W = [f_0, f_1, \cdots, f_{n-1}] \tag{1}$$

ここで，単語頻度行列 W の各要素の総和は M とする

4.3 単語の遷移頻度行列の作成

得られた単語系列を $w_i \rightarrow w_j$ 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 $f_{i,j}$ を計算し，遷移頻度行列 F を作成する．

$$F = \begin{bmatrix} f_{0,0} & \cdots & f_{0,n-1} \\ \vdots & \ddots & \vdots \\ f_{n-1,0} & \cdots & f_{n-1,n-1} \end{bmatrix} \tag{2}$$

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．ここで，遷移頻度行列 F の各要素の頻度の総和を N とする．

4.4 単語間の遷移確率の計算

遷移頻度行列中で単語 w_i から単語 w_j への遷移が存在する確率 $p(w_i \rightarrow w_j)$ は，次のように表すことができる．

$$p(w_i \rightarrow w_j) = \frac{f_{i,j}}{N} \tag{3}$$

ただし，$p(w_i \rightarrow w_j)$ は単語 w_i から単語 w_j への遷移確率ではないので，単語の出現頻度 f_i, f_j の影響を受ける．例えば，単語の出現頻度 f_i が大きいほど大きく，単語出現頻度 f_j が小さいほど小さくなる傾向があり，特に $f_i \ll f_j$ のように出現頻度が大きく異なる場合は，実際の遷移確率とは逆に，$p(w_i \rightarrow w_j)$ より $p(w_j \rightarrow w_i)$ が大きくこともある．

そこで，単語 w_i から単語 w_j への遷移確率 $p(w_j|w_i)$ を，単語出現確率 $p(w_i)$ と $p(w_j)$ を用いて，次のように求める．

$$\begin{aligned} p(w_j|w_i) &= p(w_i) \times p(w_i \rightarrow w_j) \times \frac{1}{p(w_j)} \\ &= \frac{f_i}{M} \times \frac{f_{i,j}}{N} \times \frac{M}{f_i} \\ &= \frac{f_{i,j} \times f_i}{N \times f_j} \end{aligned} \tag{4}$$

ここで，単語 w_i が出現する確率 $p(w_i)$ は次の通りである．

$$p(w_i) = \frac{f_i}{M} \tag{5}$$

4.5 単語の因果関係の決定

単語 w_i と単語 w_j の間の因果関係の有無は，$p(w_j|w_i)$ と $p(w_i|w_j)$ を比較して決定する．まず，$p(w_j|w_i)$ と $p(w_i|w_j)$ の値が大きく異なる場合に限り，因果関係が存在すると見なす．

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

ただし，現実にはさまざまなノイズが存在するために，因果関係が存在しない方向の頻度が 0 にならないことも多い．因果関係決定は，単語の出現頻度が低いほどノイズの影響を受けやすいので，$f_{i,j} \geq C \vee f_{j,i} \geq C$ の条件を満たす場合に限りおこなうこととする．

4.6 因果関係ネットワークの簡略化

すでに述べたように，実世界で因果関係があるイベント関連語が，Twitter 上でもすべて観測できるとは限らない．つまり，実世界で因果関係があるイベント関連語の系列は，Twitter 上では必ずしも完全な系列のまま観測できないので，例えば $w_0 \rightarrow w_1 \rightarrow w_2$ という順序関係があった場合に，$w_0 \rightarrow w_2$ のように途中が欠落した系列としても観測されることがあり，それが必要以上に因果関係を複雑化してしまう．

そこで，このような冗長な因果関係を除去することで，因果関係ネットワークの構造を簡略化する．具体的には，ある単語 w_i が w_j と w_k の間に $w_i \rightarrow w_j$，$w_i \rightarrow w_k$ のような因果関係があり，w_j と w_k の間にも因果関係がある場合には，次のように簡略化する．

1. w_j と w_k の間に $w_j \rightarrow w_k$ という因果関係が成り立っている場合は，$w_i \rightarrow w_k$ の因果関係を除去する（図 2a）．
2. w_j と w_k の間に $w_j \leftarrow w_k$ という因果関係が成り立っている場合は，$w_i \rightarrow w_j$ の因果関係を除去する（図 2b）．

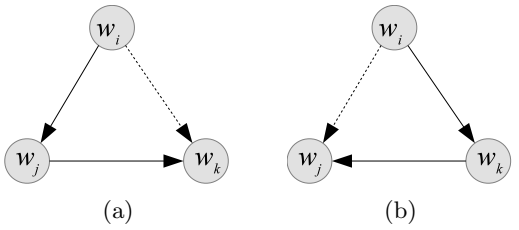

図 2: 因果関係ネットワークの簡略化

5 評価

5.1 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．

5.2 イベント関連語の抽出

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文章以外の URL，ハッシュタグ，スクリーン名などの文字列を除去してから，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，ユーザ自身の発言ではないことから削除した．

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．

さらに，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て用いた．

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

この結果得られたイベント関連語は 180 語である．

なお，例えば「東日本大震災」という名称は 2011 年 4 月 1 日の持ち回り閣議で決定されたために，それまでは「東北・関東大震災」,「東北地方太平洋沖地震」などの多くの名称が使われていた．このような表記の揺れに関しては，人手で作成した辞書を用いて正規化処理をおこなった．

5.3 因果関係ネットワークの可視化

$T = 10, C = 150$ として，提案手法を用いて抽出した因果関係ネットワークを Cytoscape 3.0.2 の Prefuse Force Directed Layout を用いて可視化した．その可視化結果を，図 3 に示す．

可視化結果中の次数が高い単語に着目すると,「東日本大震災」,「輪番停電」,「ミリシーベルト」,「ウエシマ作戦」などが相当し，入次数と出次数に注目すると，これらを次の 3 種類に分類できる．

1. 出次数の方がかなり多い単語（「東日本大震災」,「ミリシーベルト」）
2. 入次数の方がかなり多い単語（「ウエシマ作戦」）
3. 入次数と出次数がほぼ同じ単語（「水素爆発」,「輪番停電」）

1 のタイプは，原因となったイベントを示す単語だと考えられる．つまり,「東日本大震災」は地震,「ミリシーベルト」は原発事故による放射線を示すイベント関連語であり，それらの単語から多くのイベントが引き起こされている様子が表されていると推測できる．特に「東日本大震災」は多くのイベント関連語の原因となっていることから，出次数は与えた影響の強さや範囲に関連していると推測できる．

3 のタイプは，因果関係の結果が，新たな因果関係の原因になるような，因果関係の連鎖を示す単語だと考えられる．図 3 の右下を見ると,「東日本大震災」から「原子炉格納容器」を経て「水素爆発」に到達しているが，さらに「原子力保安院」や「福島第 1 原発」を経て「輪番停電」に到達しており，震災発生から原発事故を引き起こし，そのため輪番停電をおこなわざるをえなくなったという震災直後の事情を窺い知ることができる．

因果関係ネットワークは基本的には無閉路有向グラフになると予測されるが，2 のタイプはそれに反する．さらに,「アオシマ作戦」から「ウエシマ作戦」が引き起こされたように可視化されているが，実際の時系列的な順序は逆である．これらは実世界のイベントでは

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

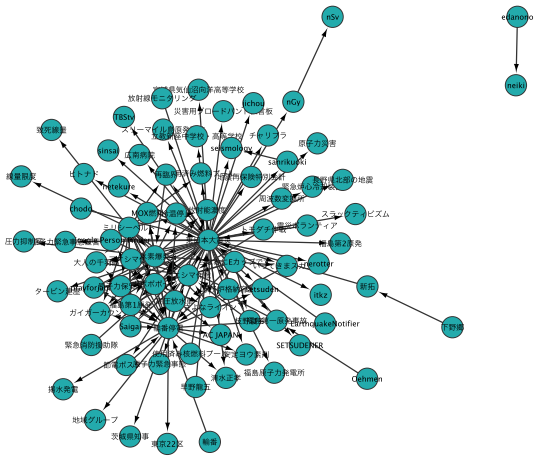

図 4: 簡略化しない場合の可視化結果

なく，口コミで広がった自発的な支援活動であることから因果関係が正しく推測しにくかった可能性がある．しかし,「大人の子守唄」,「Google Person Finder」,「トモダチ作戦」,「アオシマ作戦」などの，震災後のさまざまな支援活動と結びついている点は興味深い．

5.4 ネットワーク簡略化の有効性

提案手法では [風間 13] の手法にネットワークの簡略化を追加し，さらに細部に改良を施している．提案手法と対比させるために，簡略化を用いない場合の可視化結果を図 4 に示す．

ネットワークの基本統計量は，簡略化を施す前はノード数 81，エッジ数 148，平均直径 5，クラスタ係数 0.223 であり，簡略化を施した後はノード数 75，エッジ数 83，平均直径 8，クラスタ係数 0 であった．

なお，簡略化を用いないと，特に出次数が多い原因を示すイベント関連語が密に結びつく傾向があり，因果関係を読み取るのが難しい．これに対して，提案手法は冗長なエッジが削除されたことにより，可視化結果において次数が高いイベント関連語が分離して配置されるようになり，より多くの情報をよりわかりやすく可視化できるようになったと言える．

6 おわりに

本稿では，Twitter のツイートアーカイブから，イベント群の因果関係を抽出し，因果関係ネットワーク 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，有効性を示した．

今後の課題は以下の通りである．まず，今回用いたイベント関連語の集合は震災直後の一週間以内にバーストした単語を選択したが，イベントの因果関係を調べるためには，必ずしも充分とは言えない．そこで，注目しているイベントに関するイベント関連語を自動的により適切に抽出する必要がある．

次に，現在はイベント関連語を用いているが，既存研究のほとんどが文として扱っていることからわかるように，本稿のように単語だけで扱うと発生したイベントを理解するために必要な情報が不足してしまう．そこで，イベント関連語だけでなく，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扱うなどの方法で，ユーザに提示する情報量を増やすことを考えている．

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 JSPS 科研費 24300064 の助成を受けた．

参考文献

[青野 10] 青野 壮志, 太田 学：要因検索による因果関係ネットワークの構築と因果知識の獲得, in *DEIM Forum 2010* (2010)

[乾 04] 乾 孝司, 乾 健太郎, 松本 裕治：接続標識「ため」に基づく文章集合からの因果関係知識の自動獲得, 情報処理学会論文誌, Vol. 45, No. 3, pp. 919–933 (2004)

[石井 10] 石井 裕志, 馬 強, 吉川 正俊：因果関係ネットワークの増分的な構築について, 第 72 回情報処理学会創立 50 周年記念全国大会, 第 5 巻, pp. 239–240 (2010)

[風間 13] 風間 一洋, 高橋 公海, 鳥海 不二夫, 榊 剛史, 栗原 聡, 篠田 孝祐, 野田 五十樹：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, 第 28 回人工知能学会全国大会 (2013)

[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)

[Sakaji 08] Sakaji, H., Sekine, S., , and Masuyama, S.: Extracting Causal Knowledge Using Clue Phrases and Syntactic Patterns, in *7th International Conference on Practical Aspects of Knowledge Management (PAKM 2008)*, pp. 111–122 (2008)

[Sakaki 10] Sakaki, T., Okazaki, M., and Matsuo, Y.: Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors, in *Proceedings of the 19th International Conference on World Wide Web*, pp. 851–860 (2010)

[澤村 13] 澤村 瞳, 小林 一郎：文書内の事象間の関係抽出への取り組み, 第 28 回人工知能学会全国大会 (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

図 3: 因果関係ネットワークの可視化結果