1H3-NFC-02b-3

Twitterのイベントの因果関係の分析

Causality Analysis of Events Using Tweets

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

*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産業技術総合研究所 The National Institute of Advanced Industrial Science and Technology

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

ソーシャルメディアの普及に伴い，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ようになったことから，その情報を整理・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 の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できれば，実世界の状況を概観できるはずである．

本稿では，Twitter のツイート群から実世界の事象に関連して発生するイベント群の因果関係ネットワークを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する因果関係ネットワークを評価して，その有効性を示す．

2. ツイートからの因果関係抽出

Twitter は誰もが参加できるソーシャルメディアであり，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としての役割を持つ．情報伝播・交換の即時性が高いことから，Sakaki らは実世界で発生した出来事を観測するためのソーシャルセンサ（Social Sensor）としての利用を提案した [Sakaki 10]．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．これは既存研究が，新聞記事やブログ，Web ページなどの比較的長くしっかり書かれた文章を前提としているからであり，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い Twitter では，従来のアプローチで因果関係を推定することは難しい．

そこで，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ことで因果性の有無と方向を判定する集合知的な手法を用いる．

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

連絡先: 風間 一洋 (kazama@ingrid.org)
和歌山大学システム工学部情報通信システム学科
〒 640–8510 和歌山県和歌山市栄谷 930

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 イベント関連語の抽出

ツイート中でイベントについて語るために使われる名詞をイベント関連語と呼ぶ．イベント関連語の抽出方法は種々考えられるが，本稿では暫定的に東日本大震災という大きなイベントの後に明確なバースト性を持つ単語とした．

まず，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 [*1] 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 [*2] や原子力百科事典 ATOMICA[*3] の用語を辞書に追加した．

さらに，次の 3 つの条件を満たす名詞を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

4.2 イベント関連語系列の作成

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わけではなく，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，まずユーザごとに $w_i \to w_j \to w_k \to w_l$ のような発言順のイベント関連語系列を作成する．イベント関連語の順序は，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も考慮する．この理由は，日本語の文章では，同一ツイートまたは同一文内で，先に出現するイベント関連語が原因を，後で出現するイベント関連語が結果を示すことが多いからである．なお，イベント関連語は何度も繰り返し発言される傾向があるので，初出のみ記録する．

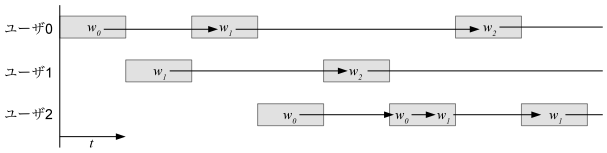

図 1: イベント関連語系列の例

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イベント関連語を含むツイートを投稿する例を，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．ユーザ 0 から $w_0 \to w_1 \to w_2$，ユーザ 1 から $w_1 \to w_2$ というイベント関連語系列が抽出される．ユーザ 2 では $w_0 \to w_0 \to w_1 \to w_1$ という順序でイベント関連語が出現するが，初出だけを記録するのでイベント関連語系列は $w_0 \to w_1$ となる．

4.3 イベント関連語の出現頻度行列の作成

次に，n 個のイベント関連語 $w_i (0 \leq w_i \leq n-1)$ の出現頻度 f_i から，出現頻度行列 W を作成する．

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

ここで，出現頻度行列 W の各要素の総和は M とする

4.4 イベント関連語の遷移頻度行列の作成

イベント関連語系列を二つのイベント関連語間の遷移関係 $w_i \to w_j$ の遷移頻度 $f_{i,j}$ から，遷移頻度行列 F を作成する．

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

ここで，遷移頻度行列 F の各要素の総和を N とする．イベント関連語の初出しか考慮しないので，対角成分は 0 である．

4.5 イベント関連語間の遷移確率の計算

遷移頻度行列中でイベント関連語 w_i から w_j への遷移が存在する確率 $p(w_i \to w_j)$ は，次のように計算できる．

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

ただし，$p(w_i \to w_j)$ は w_i から w_j への真の遷移確率ではなく，出現頻度 f_i と f_j の影響を受ける．例えば，イベント関連語の出現頻度 f_i が大きいほど大きく，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 $p(w_i \to w_j)$ より $p(w_j \to w_i)$ が大きくなることがある．

イベント関連語 w_i の出現確率 $p(w_i)$ を以下の通りとする．

$$p(w_i) = \frac{f_i}{M} \tag{4}$$

この時，イベント関連語 w_i から w_j への遷移確率 $p(w_j|w_i)$ を，出現確率 $p(w_i)$ と $p(w_j)$ を用いて，次のように求める．

$$p(w_j|w_i) \quad = \quad p(w_i) \times p(w_i \to w_j) \times \frac{1}{p(w_j)}$$

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

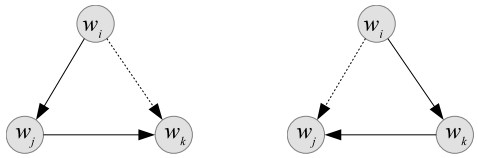

図 2: 因果関係ネットワークの簡略化

$$= \frac{f_i}{M} \times \frac{f_{i,j}}{N} \times \frac{M}{f_i} = \frac{f_{i,j} \times f_i}{N \times f_j} \quad (5)$$

4.6 イベント関連語の因果関係の決定

イベント関連語 w_i と 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．しかし，イベント関連語 w_i と w_j の間に成立している関係が因果性を持たない共起関係だったり，現実データのノイズのために因果関係が存在しない方向の遷移確率が 0 にならないことも多い．

そこで，$p(w_j|w_i)$ と $p(w_i|w_j)$ の確率の値が大きく異なる，つまり $p(w_j|w_i) \gg p(w_i|w_j)$ または $p(w_j|w_i) \ll p(w_i|w_j)$ の場合に，イベント関連語 w_i と w_j の間に因果関係が存在するとみなす．ただし，確率が低い場合には真の因果関係なのかどうかが疑わしいので，ある閾値を下回る場合は除外する．

すなわち，因果関係 $w_i \to w_j$ は，以下の条件を満たす時に存在するとする．

$$p(w_j|w_i) \geq p(w_i|w_j) \times T \quad (6)$$

$$p(w_j|w_i) \geq P \quad (7)$$

また，因果関係 $w_j \to w_i$ は，以下の条件を満たす時に存在するとする．

$$p(w_i|w_j) \geq p(w_j|w_i) \times T \quad (8)$$

$$p(w_i|w_j) \geq P \quad (9)$$

ここで，T は $T \geq 1$，P は $0 \leq P \leq 1$ である．

4.7 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として観測できず，例えば $w_0 \to w_1 \to w_2$ という順序関係があった場合に，$w_0 \to w_2$ のように途中が欠落した系列として観測されることが因果関係を複雑化すると考えられる．そこで，このような冗長な因果関係を除去して，因果関係ネットワークを簡略化する．

簡略化の対象は，3 個のイベント関連語の間に因果関係が成り立っている場合である．例えば，イベント関連語 w_i と w_j，w_k の間に $w_i \to w_j$，$w_i \to w_k$ のような因果関係があるとする．さらに w_j と w_k の間に $w_j \to w_k$ という因果関係が成り立っている場合は，w_i から w_k に到達する経路が 2 つあることになるので，ショートカットである $w_i \to w_k$ の因果関係を除去する（図 2a）．この因果関係を削除しても，w_i と w_k の間の因果性が保存されることに注意する．同様に，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，$w_i \to w_j$ の因果関係を除去する（図 2b）．

実際には，次の手順でネットワーク全体を簡略化する．

1. 入次数が 0，出次数が 1 以上のノードを探す．見つからない場合は終了する．
2. 指定されたノードから幅優先探索でエッジの組を見つけて，図 2 のどちらかの条件に合致する場合に簡略化する．
3. 1 に戻る．

5. 評価

5.1 データセット

3 月 5 日から 24 日の間に，Twitter API[*4] 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て，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 因果関係ネットワークの分析

データセットから抽出さ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ので，このような表記の揺れは人手で作成した辞書を用いて正規化した．

このイベント関連語集合を用いて，$T = 10, P = 0.01$ として因果関係ネットワークを抽出した．ノード数は 121 個，エッジ数は 169 本，平均次数は 2.793，クラスタ係数は 0.0 であった．

なお，ネットワーク簡略化をおこなわない場合は，ノード数は 121 個，エッジ数は 182 本，平均次数は 3.008，クラスタ係数は 0.061 となる．つまり，ノード数は同じでも，構造は簡略化されていることがわかる．

5.3 因果関係ネットワークの可視化

抽出した因果関係ネットワークを Cytoscape 3.0.2 の Force Directed Layout を用いて可視化した結果を図 3 に示す [Shannon 03]．

この可視化結果を見ると,「東日本大震災」(98) ,「輪番停電」(43) ,「ミリシーベルト」(13) ,「水素爆発」(12) など，次数が高い単語がいくつか存在した．これらは，他の事象を引き起こす原因となった単語だと考えられる．つまり,「東日本大震災」は地震,「輪番停電」は停電,「ミリシーベルト」は原発事故による放射線の影響,「水素爆発」は原発事故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．

なお，今回は因果関係決定に関するノイズを除去するために，以前のような頻度ではなく確率を使った．これは全体的なバランスが良くなる反面，次数が高いノードに近い部分ほど密になり，そこからのホップ数が少なくなる傾向があり，因果関係の連鎖が抽出されにくくなることがわかった．

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する手法について延べ，実際に東日本大震災時

*4 http://apiwiki.twitter.com/

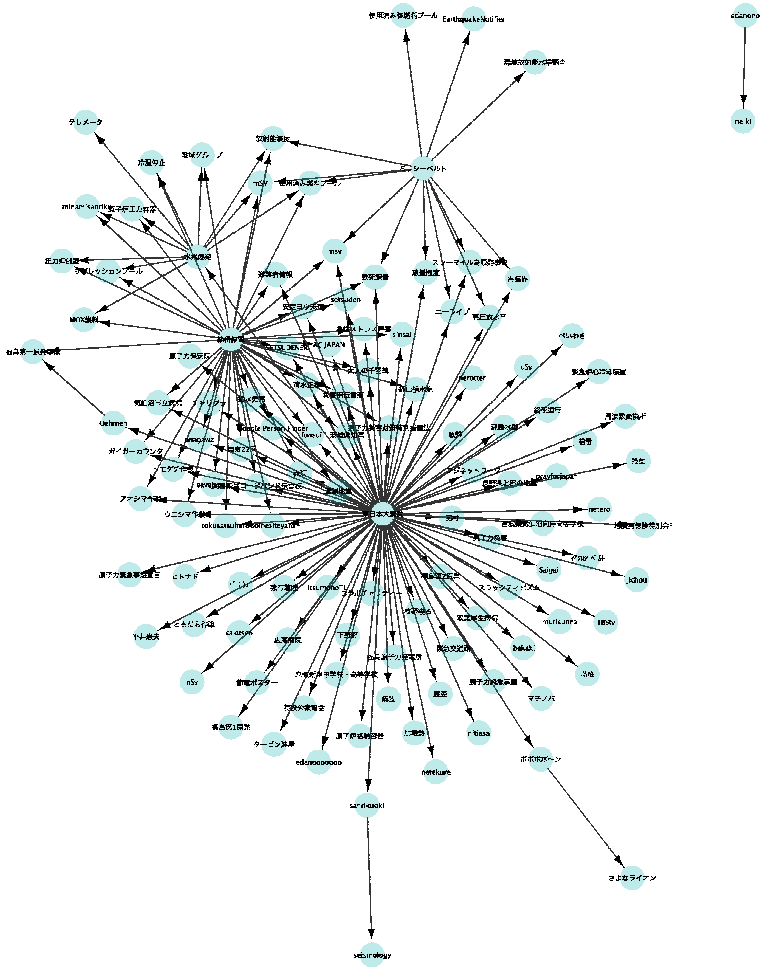

図 3: 因果関係ネットワークの可視化結果

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を可視化すると共に，それを分析した．

なお，本手法の結果は与えるイベント関連語群に大きく影響される．今回イベント関連語としてバーストした単語を用いたが，イベントの因果関係を把握するためには必ずしも充分ではなく，また不適切と思われる因果関係も見受けられた．そこで，LDA などのトピック分類手法を使ってトピックごとのイベント関連語群を抽出する予定である．

また，既存研究のほとんどが文を扱っていることからわかるように，イベント関連語単独では発生したイベントを理解するためには不十分である．そこで，イベント関連語と同時に使われる補足語をまとめて扱うなどの方法を検討中である．

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 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)

[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)

[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)

[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)

[Shannon 03] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T.: Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks, *Genome Research*, Vol. 13, pp. 2498–2504 (2003)