1F3-6

文書内の事象間の関係抽出への取り組み

A Study on Extracting Relations among Events in Documents

澤村瞳　　小林一郎
Hitomi Sawamura　Ichiro Kobayashi

お茶の水女子大学大学院人間文化創成科学研究科
Advanced Sciences, Faculty of Sciences, Graduate School of Humanities and Sciences Ochanomizu University

In this paper, we aim to extract causal relations in multiple documents and then overlook the whole causal chains among the events described in the documents. In order to extract causal relations from documents, we focus on the words recognized as clue expressions for causal relations. By connecting multiple causal relations to make a network of causal chains, we have introduced matching methods based on word similarity and topic similarity. We have verified our proposed method is useful for extracting causal chains from actual newspaper articles.

1. はじめに

本研究では，複数文書内に現れる事象間の因果関係を抽出し，事象に対する全体像を俯瞰することを目的とする．事象間の因果関係を捉えるために，文中に現れる節間関係や手がかり表現に着目し因果関係を抽出する．さらにそれぞれの文書から抽出された断片的な原因と結果の因果関係をつなぐため，原因および結果の表現間の柔軟なマッチングとして，Jaccard 係数による語彙集合の類似度の算出および，Hierachical Dirichlet Process Latent Dirichlet Allocation(HDP-LDA) を使いトピックを抽出し，文書中の表層的な情報および潜在的な情報から事象間の関係の抽出を試みる．

2. 因果関係抽出

先行研究では多くの手法で因果関係抽出を行なっている．乾ら [1] の研究では文書の特徴から因果関係を抽出している．坂地 [3] らは構文パターンと手がかり表現を用い，正確に因果関係を抽出している．佐藤 [4] の研究では Web 中の文書を用い因果関係ネットワークを視覚的に表す研究を行なっている．

2.1 抽出対象因果関係

因果関係を抽出するためには，節を繋ぐ際の接続詞に表現される因果関係を抽出することや構文パターンを捉えて因果関係を抽出する手法，また因果関係の強さをモダリティを考慮して決める手法なども存在する．本研究では，因果関係を抽出する方法として先行研究を参考にし，節間関係を示す表層表現を手がかり標識に基いて因果関係を抽出する．右上の表 1 には，節間関係に現れる因果関係を示す接続詞を示す．これを元に因果関係を抽出する．

また，助詞と名詞においても，以下の 10 表現について因果関係を抽出する．

「結果」,「場合」,「理由で」,「目的で」,「れば」,「影響で」,「より」,「に伴う」,「たら」,「受け」

表 1: 因果関係抽出に利用する節間関係

節間関係	表層表現
理由	〜ので，〜せいで
条件	〜ならば
目的	〜ために，〜のに，〜べく
逆説	〜けれど

2.2 因果関係の連鎖

2.2.1 語彙の一致による因果関係抽出

因果関係の連鎖を抽出するためには, ある因果関係の結果が他の因果関係の原因であることを判別する必要がある．本研究では，ある文中に示される因果関係の結果と他の文中に示される因果関係の原因において表れている語彙を対象に Jaccard 係数に基づき類似性を取ることにより，その繋がりを捉えることにする．今，文 1 の単語の集合を A，文 2 の単語の集合を B とすると，Jaccard 係数は式 (1) で表される．

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

式 (1) では，比較する文における単語数（文長）の影響を考慮していないため，文長を考慮した文の類似度として本研究では式 (2) を採用する．

$$Jaccard'(A,B) = Jaccard(A,B) \times \sqrt{|A \cup B|} \quad (2)$$

2.2.2 トピックの一致による因果関係抽出

イベント間の隠れた因果関係を発見するために，本研究では文書の潜在的意味を推定する階層ディリクレ課程を用いた配分法 Hierarchical Dirichlet Process Latent Dirichlet Allocation(HDP-LDA) を使って潜在的トピックを抽出した．HDP-LDA は予めトピックの数を与えなくとも，自動でトピックを抽出する言語モデルであるため，与えるトピック数が推定結果に大きな影響を及ぼすことはない．

2.2.3 因果関係抽出の流れ

図 1 に文書から因果関係を抽出し，因果関係の連鎖を構築する概要を示す．

連絡先: 澤村瞳，お茶の水女子大学大学院人間文化創成科学研究科理学専攻情報科学コース小林研究室，東京都文京区大塚 2-1-1，sawamura.hitomi@is.ocha.ac.jp

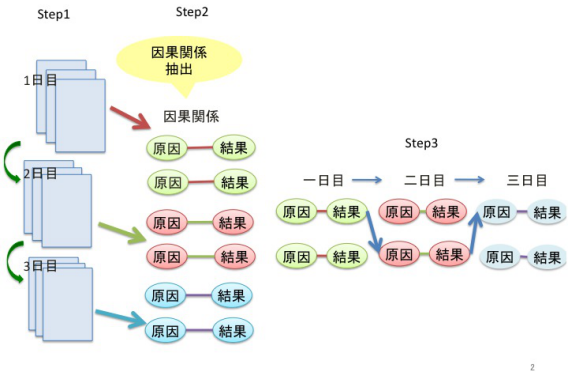

図 1: 因果関係の抽出と連鎖構築の概要

図 1 に示す概要を処理の流れに沿って説明する．

step 1. 因果関係抽出
1 文から上記に示した因果関係を表現する接続詞（表 1）または手がかかり標識があるものを集める．

step 2. 原因と結果のペア生成
step1 で抽出した文を因果関係表現を前後に，前の部分を「原因」，後ろの部分を「結果」として，2 つのペアに分ける．

step 3. 対象期間における因果関係連鎖の生成
1 文の「結果」と他の文の「原因」に含まれる単語の一致 Jaccard 係数によって測り，閾値を越えるものを因果関係連鎖として採用する，さらに，HDP-LDA によって対象の文書から抽出した潜在的トピックにより，1 文の「結果」と他の文の「原因」が同じトピックで値が高い時に，因果関係連鎖として採用する．トピックの一致についての因果関係連鎖の詳細を下記に示す．

1) 潜在的トピックを因果関係のある全ての文を対象に抽出する．

2) ある文の「結果」と他の文の「原因」のトピックが同じであるかを確認し，確率分布に基づいて二つの部分のトピックと単語から調べていく．上記二つの部分の単語の一致や，同じトピックであるとみなした時にそこに因果関係が存在すると見なし繋いでいく．

3) そのペアは事前に定められた閾値を用いて同じトピックとみなされ，かつ類似度の高いペアに因果関係があるとみなし繋げていく．

3. 実験仕様

3.1 実験仕様

使用したデータは，朝日新聞，読売新聞，河北新聞の東日本大震災に関する記事の 3 月 11 日から 3 月 13 日までの新聞記事 621 記事において，因果関係を抽出する事ができた 1894 文を用いた．連続する 3 日間において，因果関係の連鎖があるかを Jaccard 係数の値と HDP-LDA によるトピックを元に調査した．HDP-LDA の設定として，サンプリング手法に Gibbs サンプリング，イテレーション 10 回，トピック分布のハイパーパラメータ α は，α=2.8 と設定する．γ はガンマ分布 (1,1) より得られ，ここでは Teh et al.[7] を参考にし、γ=0.04 と設定する．

3.2 実験結果

Jaccard 係数による一致で得られた因果関係のうち上位 25 件を図 2 に，その中の番号ついている文を表 2 に示す．

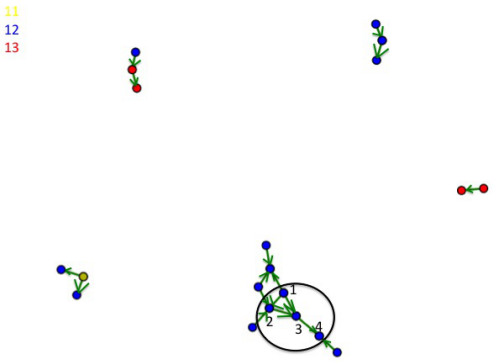

図 2: Jaccard 係数の一致による因果関係連鎖

表 2: 図 2 中の文 1，2，3，4

1	爆発で格納容器が破壊され れば、大量の放射性物質が環境に放出されることになる。
2	微量の放射性物質を含む水蒸気が外部に放出される程度深刻ではないが、燃料棒が損傷して露出し、水蒸気と反応して爆発するような事態にな れば、大量の放射性物質が外部に放出されることになる。
3	放射性物質の環境放出が不可避となったことを 受け、政府は午前５時４４分、周辺住民の避難指示範囲を半径３キロから１０キロに拡大した。
4	政府が１２日早朝、福島第一原発のある福島県大熊、双葉両町の住民に対する避難指示を半径３キロから１０キロ以内に拡大したことを 受け、両町の住民移動が始まった。

次に潜在的トピック抽出により得られた 8 個のトピックにおける上位 10 個の単語を表 4 に示す．HDP-LAD によるトピックの一致で得られた因果関係のうち上位 25 件を図 3 に，その中の番号ついている文を表 3 に示す．

4. 考察

4.1 語彙の一致による因果関係連鎖抽出

図 2 において 12 日のイベントが多く抽出されている理由は震災翌日のため多くの重要な情報とそれらの理由などの説明が多く記述されていたためと考える．まるで囲まれている部分をみてみると 1 → 3 では爆発が起こり放射性物質が放出したという因果連鎖がとれていることがわかる．また 1 → 2 → 3 と 2 を入れることによって，燃料棒が露出することにより放射物質が放出されるという詳しい因果連鎖をとることができた．3 → 4 では 3 での「住民の避難指示」を受け 4 で「住民移動が始まった」という因果連鎖が抽出できた．

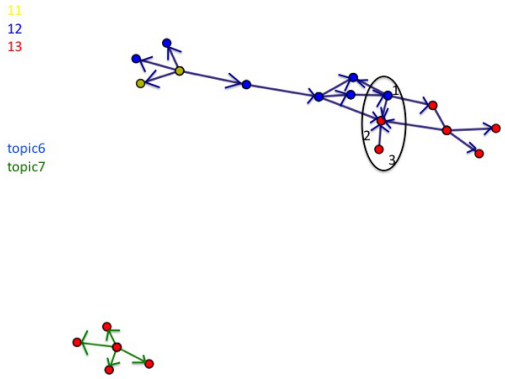

図 3: 潜在的トピックの一致による因果関係連鎖

表 3: 図 3 中の文 1，2，3

1	政府は１１日夜、東北・関東大地震の影響で自動停止した福島県の東京電力福島第１原発の１、２号機で、外部からの電力供給が失われるなど緊急に対策を講じる必要があるとして、原子力災害対策特別措置法に基づく「原子力緊急事態宣言」を発令した。
2	東日本巨大地震では、東京電力福島第一原発１～３号機、同第二原発の全４基、東北電力女川原発の全３基、日本原子力発電東海第二原発の計１１基が、強い揺れにより自動停止した。
3	保安院によると、地震による停電で外部からの電力供給が失われたことや、冷却水をさらに冷やす海水を取り込み、動かすポンプが津波で被害を受けたことなどにより、福島第一原発２号機や、同第二原発１、２、４号機などでは、冷温停止までに時間がかかっているという。

4.2 トピックの一致による因果関係連鎖抽出

図 3 において抽出されたトピックに対してそれぞれ見てみるとトピック 1 は「宮城」,「福島」,「岩手」などの単語が出てきているので東北のトピックとなっている. トピック 2 は「電話」,「携帯」,「メール」などから災害時の連絡手段のトピックである. トピック 3 では「東京」,「新宿」,「横浜」,「川崎」,「江東」から首都圏のトピックである. トピック 4 では「予定」,「運転」,「中止」,「全線」などからイベントの開催事情についてである. トピック 5 では「避難」,「津波」,「がれき」などから津波や地震の被害に関してのトピックである. トピック 6 では「原発」,「爆発」,「放射」,「物質」などから原発爆発に関してのトピックである. トピック 7 では「地震」,「津波」,「観測」,「気象庁」などから地震や津波の観測についてのトピックである. トピック 8 では「空港」,「キャンセル」,「タクシー」,「満席」などから交通事情に関してのトピックであることがわかる. トピックだけの図に対しての考察 1 → 2，3 → 2 はともに原発爆発に関するトピックでまとまっているが，因果関係がとりにくい.

4.3 語彙およびトピックの一致による因果関係連鎖抽出

図 4 においてトピックのみからでは事象間の因果関係を明確に俯瞰できないこともある一方，Jaccard 係数のみによる因果関係抽出では語の一致に依存してしまうため潜在的な因果関係を抽出することができないと考えられる．実験結果からそれらのことを確認し，Jaccard 係数に基づいて得られた因果関係にトピックに基づいて得られた因果関係を追加することが妥当と判断した．図 4 と表 5 その結果を示す．Jaccard 係数によって得られた上位 20 のイベントに対しては，原発爆発と地震や津波の観測の二つのトピックが関与していることがわかり，Jaccar 係数に基づく因果関係抽出では得られなかった緑の丸とオレンジの丸の部分の因果連鎖が抽出ができていることがわかる．ここで A の因果連鎖について具体的にその内容を見てみると 1 → 2 では原発爆発のトピックで対応という理由でつながっているが，二つの文の内容において表層的に明示される因果関係はあまり確認されない．B の方を見てみると 3 → 2 では 3 での「原子力緊急事態宣言」が 4 での「住民に対する避難指示」であることがわかる．また 3 → 4 では 3 で

表 4: HDP-LDA により抽出された各トピックの上位 10 単語

トピック 1	トピック 2	トピック 3	トピック 4
宮城	電話	東京	予定
午後	営業	取引	運転
被災	携帯	午後	中止
福島	被災	地震	全線
被害	場合	帰宅	再開
地震	メール	新宿	延期
岩手	サービス	横浜	試合
現在	支店	川崎	東日本
日本	対応	施設	東京

トピック 5	トピック 6	トピック 7	トピック 8
避難	福島	地震	空港
津波	原発	津波	キャンセル
男性	爆発	観測	タクシー
車	避難	気象庁	レンタカー
家	東京電力	震度	庄内
近く	発電	震源	自宅
自宅	物質	沖	満席
がれき	放射	発生	列
地震	会見	マグニチュード	客

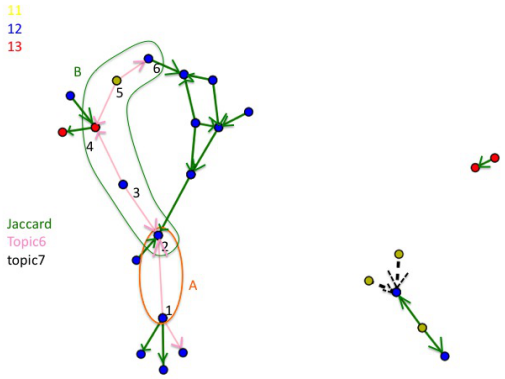

図 4: Jaccard+HDP-LDA により抽出された因果関係

表 5: 図 4 中の文 1，2，3，4，5

1	総理からどのような指示が　総理ご自身が専門家のみなさんの話、経済産業大臣の話を聞きながらやっている ので、指示を出すというよりも、しっかりと住民のみなさまの健康の観点からつねに最悪のケースを想定して、万全の措置をとるということを事実上指示しながら経産大臣、保安院、原子力安全委員会、東電などと対応させていただいているところ。
2	政府が１２日早朝、福島第一原発のある福島県大熊、双葉両町の住民に対する避難指示を半径３キロから１０キロ以内に拡大したことを 受け、両町の住民移動が始まった。
3	政府は１１日夜、東北・関東大地震の 影響 で自動停止した福島県の東京電力福島第１原発の１、２号機で、外部からの電力供給が失われるなど緊急に対策を講じる必要があるとして、原子力災害対策特別措置法に基づく「原子力緊急事態宣言」を発令した。
4	東京電力福島第一原子力発電所の爆発の 影響、新潟県は１３日、原発周辺の放射線の監視業務で派遣した職員２人が被曝（ひばく）したと発表した。
5	東京電力は１１日、宮城沖地震の 影響、福島県の福島第一原発の１号機と２号機が自動制止して高温になっている原子炉の炉心を、水を循環させて冷やせない状態になっている可能性がある、と発表した。
6	東京電力は地震で自動停止している福島第二原発１～４号についても、原子炉を覆っている格納容器内部の圧力を下げる ため、弁を開けて放射性物質を含んだ空気を外部に放出させることを検討する、と発表した。

「福島第一原発の自動停止」と 4 での「福島第一原子力発電所」に関係があることがわかり，2 と 4 の文は 3 を介して原発が爆発後のことがわかる．一方, 5 → 4 では 5 での「冷やせていない炉心」が原因で 4 で「爆発」したことがわかる．また 5 → 6 では 5 の「原子炉の炉心」が 6 で「原子炉を覆っている格納容器内部の圧力を下げる」と関連していることがわかり，5 の「炉心」を介して 4 と 5 がつながっていることがわかった．このことにより，語彙が完全に一致していなくても潜在的トピックを用いることで，別の言葉で言い換えられて因果関係が生起しているということもわかった．

5. おわりに

本研究では，新聞記事の文中に存在する因果関係を抽出し，抽出された因果関係の結果と原因を Jaccard 係数を使い表現の一致に基づいて繋げることにより，因果関係の連鎖を構築した．一方，トピックの一致のみに基いて繋げると，同じ話題で繋がってはいるが，因果関係連鎖の抽出は確認できなかった．そこで Jaccard 係数による語彙の一致に基づき繋いだ後に，同じトピックの物を繋げることにより，より柔軟な因果関係の連鎖を抽出できた．今後は因果関係の連鎖だけでなく，様々な関係性を視野に入れ抽出する事を試みる．また，因果関係抽出手法を災害の被害を軽減する対策を考えるためのツールとして役立たせたいと考えている．

参考文献

[1] 乾考司，奥村学文書内に現れる因果関係の出現特性調査情報処理学会研究報告，2005-NL-167，2005

[2] 大友謙一，柴田知秀，黒橋禎夫，述語項構造の共起情報と節間関係の分布を用いた事態間関係知識の獲得，言語処理学会第 17 回年次大会，2011.

[3] 坂地泰紀，竹内康介，関根聡，増山繁，構文パターンを用いた因果関係抽出，言語処理学会第 14 回年次大会，E5-5, (2008).

[4] 佐藤岳文，掘田昌英，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.

[5] 青野荘志，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，DEIM Forum2010，2010.

[6] 柴田知秀，黒橋禎夫，述語項構造の共起情報と格フレームを用いた事態間知識の自動獲得，IJCNLP2011，2011.

[7] Yee Whye Teh，Michael I. Jordan，Matthew　J. Beal，David M. Blei，Hierarchical Dirichlet Processes，Journal of the American Statistical Association，Vol.101，2004.