文章内の事象間関係の抽出と俯瞰技術の開発

理学専攻・情報科学コース　澤村 瞳 (指導教員：小林 一郎)

1　はじめに

近年，Web などの文書データの増加に伴い，大量の情報から迅速に有益な情報を抽出する手法が必要とされている．本研究では，複数文書内に現れる事象間の関係を抽出し，事象に関する全体像を俯瞰する 2 つの手法を提案する．2 章では，因果関係が存在している文を抽出し，それらの因果関係の連鎖を構築することにより事象に関しての理解を深める事を目的とする，因果関係抽出に基づく事象の俯瞰分析手法を提案する．3 章では，文書全体の潜在的意味を解析し，トピックの変遷に伴う事象の生起や人物の行動から潜在する事象の関係抽出に基づく俯瞰分析手法を提案する．

2　因果関係抽出に基づく事象の俯瞰分析

2.1　概要

文書内に表れ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 [1] や手がかり表現 [2] に着目し因果関係を抽出する．さらに因果関係を持つ文から原因と結果の対を抽出し，他の文が持つ因果関係と繋ぐため，ある文の結果と他の文の原因に含まれる語彙の一致による表層的な情報下での関連性，及び，HDP-LDA [3] を用いて抽出されたトピックによる潜在的な情報下での関連性から事象間の関係を抽出する．

2.2　提案手法

因果関係連鎖判定には，文 A 中に示される因果関係の結果部と文 B 中に示される因果関係の原因部において現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．さらに，文長を考慮し式 (1) を定義する．

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

また，事象間の隠れた因果関係を発見するために，階層ディレクレ過程を用いて文書の潜在的意味を推定する HDP-LDA を使用し潜在的トピックを抽出した．

2.3　実験

2.3.1　実験仕様

対象データは，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 621 件のニュース記事である．

2.3.2　実験結果および考察

HDP-LDA により推定されたトピックを表 1 に示す．また，Jaccard 係数による語彙の一致で得られた因果関係連鎖に，潜在的トピックに基づいて得られた因果関係を追加した上位 25 件を図 1 に，その中の番号が付いている文を表 2 に示す．

実験結果から，Jaccard 係数に基づいた因果関係連鎖の判定において，明確な因果連鎖を発見できたが，潜在的トピックに基づいた因果関係連鎖では同じトピックでまとまってはいるが，因果関係としては判定が難しいことが分かった．そこで，Jaccard 係数に基づき得られた因果関係連鎖に潜在的トピックにより得られた因果関係連鎖を追加することにより，明確な因果関係連鎖を抽出するとした．

表 1: 抽出された各トピックの上位単語

トピック	上位単語	トピック名
topic1	宮城 午後 福島 被災 被害 自信 岩手	東北
topic2	電話 営業 携帯 被災 場合 メール サービス	連絡手段
topic3	東京 取引 午後 地震 帰宅 新宿 横浜	首都圏
topic4	予定 運転 中止 全線 再開 延期 試合	イベントの開催事象
topic5	避難 津波 男性 家 車 近く 自宅	津波や地震の被害
topic6	福島 原発 爆弾 避難 東京電力 発電 物質	原発爆発
topic7	地震 津波 観測 気象庁 震度 震源 沖	津波や地震の観測
topic8	空港 キャンセル タクシー レンタカー 庄内 自宅	交通事情

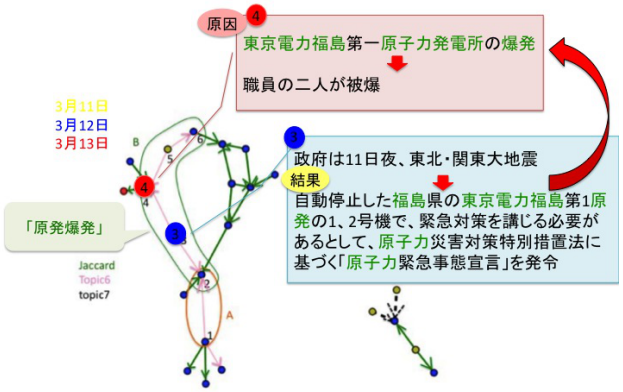

図 1: Jaccard+HDP-LDA による因果関係連鎖[2]

3　潜在事象の関係抽出に基づく俯瞰分析

3.1　概要

ニュース記事などの文書データから自動で話題性のある事象（以下，イベント）を抽出することを目的とする．HDP-LDA を用い潜在的トピックを推定し，そのトピックを元にキーワードや人物を抽出することで，イベントの発見を行う．

3.2　提案手法

まず全文書を HDP-LDA を用い，トピックを推定する．各文に対して推定された各トピックの中で，一番重みの多いトピックをその文とトピックとみなす．トピック推定した文集合から，ある時におけるキーワードを抽出する．文献 [4] を用いて以下の式を用いて重みが大きい単語をキーワードとして抽出することで，その時に重要なイベントを認識する．式 (2) は前の時間を考慮し，なお他のトピックの単語出現度を考慮した式である．式 (3) はあるトピックのみに着目して，重要単語を抽出した式である．t は時間，k はトピック，TF は文中の単語出現度，λ は重みパラメータ，ISF は総文数/単語が出現する文を表す．

$$Weight(w)_k^t = \frac{TF(w)_k^t}{\sum_k TF_k^t} \times \exp(-\lambda \times Weight(w)_k^{t-1}) \qquad (2)$$

$$Weight(w)_k^t = TF(w)_k^t \times ISF \qquad (3)$$

人物や，国家，組織などをイベントに関連する情報

表 2: 図 4 中の文 1，2，3，4，5，6

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。

とみなし，出現頻度からトピックやイベント等と，どのように関与しているかを抽出する．また，人物同士の Simpson 係数に閾値を加えた (式 (4)) を用いて共起を測ることにより，人物同士の関係の強さを示した．

$$R(X,Y) = \begin{cases} \dfrac{|X \cap Y|}{min(|X|,|Y|)} (if |X| > k \wedge |Y| > k) \\ 0 \quad (otherwise) \end{cases} \tag{4}$$

3.3　実験

3.3.1　実験仕様

対象データは，毎日新聞の 911 テロに関する記事の 2001 年 9 月 12 日から 10 月 10 日までの 1438 件のニュース記事である．

3.3.2　実験結果および考察

HDP-LDA により推定されたトピックを表 3 に示す．全文に対して表 3 で抽出されたトピックで重み付けをし，トピックの個数をグラフで表し，式 (4) を用いて，全文書中の人物同士の関係の強さを抽出したものを表 4 に示す．

表 3: 抽出された各トピックの上位単語

トピック	上位単語	トピック名
topic0	米国 テロ 米 支援 攻撃 多発 同時 日本	政治
topic1	ニューヨーク テロ 米国 ビル 世界 貿易 米	ニューヨーク
topic2	米国 テロ イスラム 攻撃 戦争 米 報復 世界	アフガニスタン
topic3	基地 米 出港 同時 予定 キティ テロ ホーク	米軍 基地
topic4	活動 実施 措置 自衛隊 規定 武器 十 支援	自衛隊 支援
topic5	預金 外貨 運用 月末 証券 残高 ドル 増加	為替 金融
topic6	スイス 航空 経営 エア 株式 赤軍 保有 灯油	航空 会社
topic7	判事 スコットランド 発効 リビア 法廷 虐殺	裁判

図 2 から政治とアフガニスタンに関するトピックが多くを占めていることがわかった．この二つのトピックの文数のグラフが連動しているところも多いことから，政治の動きとアフガニスタンの動きは大きく関わっ

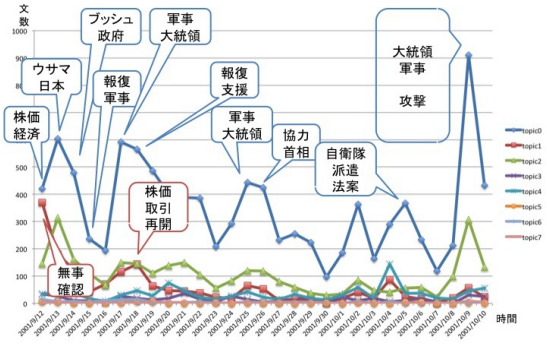

図 2: 各トピックの文書数とキーワード

表 4: 人物同士の関係が強い上位 10 組

順位	人物 1	人物 2	Simpson 係数
1	サミュル・ハク党首	タリバン	1.0
1	ファズド・ラフマン	タリバン	1.0
1	ブッシュ	ライス大統領補佐官	1.0
1	マスード将軍	タリバン	1.0
5	ウサマ・ビンラディン	サミュル・ハク党首	0.83
6	ブッシュ	パウエル国務長官	0.73
7	サッタル外相	タリバン	0.72
8	パウエル国務長官	タリバン	0.71
9	ブッシュ	シラク	0.70
10	ウサマ・ビンラディン	ブッシュ	0.69

ていることがわかった．topic0 の政治のトピックに関して，キーワードを追っていくと，米大統領や日本政府の動きがわかった．また topic1，ニューヨークのトピックに関して，文書数が多くなっている日には，テロの発生や，株取引の再開など，イベントが発生していることがわかった．また，人物同士の関連性を抽出したところ,「タリバン」に着目するとイスラム組織の人物らが上位に出現し．続いて「パウエル国務長官」や「サッタル外相」が出現した．これは,「タリバン」の動きに関する対策に関する共起だと考えられる．次に,「ブッシュ」に着目すると「ライス大統領補佐官」や「パウエル国務長官」との共起はアメリカの政治に関することであり,「シラク」は外交を示していることがわかる．これらのことから，同じ組織や国に所属している組み合わせが上位に出てくることがわかった．

4　おわりに

本研究では，事象の俯瞰分析技術の開発として，因果関係を用いた事象を詳しく理解する手法と，潜在的意味に基づいた手法を提案した．今後の課題としては，リアルタイムでの俯瞰分析を行い，提案手法をより有用性の高いものにしたいと考えている．

参考文献

[1] 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 回年次回，2011.

[2] 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

[3] Yee Whye Teh，Michael I. Jordan，Matthew j. Beal，David M. Blei，Hierachical Dirichlet Processes，Journal of American Statistical Association，Vol.101，2004.

[4] Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai J. Gao, Xin Tong, and Huamin Qu TextFlow: Towards Better Understanding of Evolving Topics in Text IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 17, NO. 12, 2011