複数文書内の事象間の因果関係抽出に基づく俯瞰分析への取り組み

澤村 瞳　(指導教員：小林 一郎)

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，日本語 WordNet で語彙間の類義関係を加えて算出する $Jaccard+Wordnet$ を導入した．

2　因果関係抽出

2.1　抽出対象因果関係

因果関係を抽出するためには，節をつなぐ際の接続詞に表現される因果関係を抽出すること [2] や構文パターンを捉えて因果関係を抽出する手法 [3]，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 [4] など様々な手法が存在する本研究では，因果関係を抽出する方法として，節間関係を示す以下の表層表現（表 1 参照）と手がかり標識に基づいて因果関係を抽出する．

表 1: 節間関係を示す表層表現

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
同時	～ならば

具体的には，以下の 10 表現を手がかり標識とし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,」

2.2　因果関係の連鎖

因果関係の連鎖を発見ためには，ある因果関係の結論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．このことから，抽出した因果関係の結論部と原因部が同じものを示しているかを判別しなくてはならない．本研究では結論部および原因部において表現されている語彙を対象に ‘語彙の一致’ と，‘語彙の一致’ に ‘同義語 ’ を組み合わせた二つの観点から因果関係を抽出することを試みる．

2.2.1　Jaccard 係数による語彙の一致

類似度を判定する文の中から名詞，動詞を抽出し語彙の集合をそれぞれ A，B とする時，集合 A，B，の一致度を示す指標となる Jaccard 係数を以下に示す．

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

語彙集合 A，B をそれぞれ因果関係の結論部と他の因果関係の原因部とみなして，時間順序を考慮した上で因果関係の繋がりを Jaccard 係数の値によって判定する．

2.2.2　同義語の判定

同義語の判定には日本語 WordNet[5] を用いる．WordNet は単語を類義関係のセット（synset）でグループ化をしており，1 つの synset が 1 つの同義語集合に対応している．また synset は上位語と下位語の関係でリンクしており，類似度の尺度を用いて上位語階層の概念への最短経路に基づいた 0 から 1 の範囲のスコアを返す．例として「停止」と「中止」の語彙の類似度を取ると 1.0 であり,「状況」と「周囲」の類似度は 0.5 となる．

2.2.3　Jaccard ＋ 日本語 Wordnet

日本語 WordNet を利用して得られた語彙の類似度を $sim(A,B)$ とすると，Jaccard ＋ 日本語 WordNet により求められる，単語 $a(a\in A)$ と単語 $b(b\in B)$ の類似度を式 (2) の様に定義する．

$$Jaccard+WordNet=\frac{\sum_{a\in A,b\in B} sim(a,b)}{|A\cup B|} \quad (2)$$

式 (2) より，Jaccard 係数では語彙の表層的な一致しか取れなかったものが，語彙の類義関係が取れるようになり，より柔軟なマッチングが実現されると期待される．ここで，式 (2) の値は [0,1] ではないため，類似の判定は値の相対的な大きさによって行う．

2.3　因果関係抽出のながれ

図 1 に文中から因果関係を抽出する流れを示す．

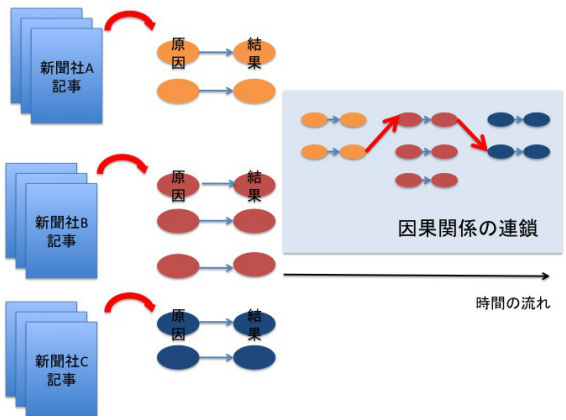

図 1: 因果関係抽出のながれ

step 1.　因果関係抽出
1 文中に 2.1 節にに示した因果関係を表現する接続詞または手がかり標識があるものを集める．

step 2.　原因部と結論部のペア生成
因果関係表現を前後に，前の部分を原因部，後ろ

の部分を結論部として，それぞれから名詞と動詞を取り出し，原因部と結論部の 1 文を 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

ある文の原因部と他の文の結論部におけるそれぞれの名詞，動詞の Jaccard 係数による語彙の一致と，$Jaccard + WordNet$ にる一致をそれぞれ測り，予め設定した閾値を越えるものを因果関係の連鎖として採用する．

3 実験

3.1 実験仕様

使用したデータは，朝日新聞，読売新聞，河北新聞の東関東大震災に関する記事の 3 月 11 日からの 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ことができた 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 係数と，Jaccard 係数 + 日本 WordNet の値を基に調査した．それぞれの指標から得られる因果関係の連鎖を可視化ツールの一つである NetworkX[6] を用いて表す．

3.2 実験結果

一文から抽出された因果関係に対して，結論と原因の総当たりの組を 3 日間において求め，その結果，Jaccard 係数の値が 0.3 以上ものと，WordNet の類似度が 0.5 以上のペアを単語の意味が一致しているとし，$Jaccard + WordNet$ が 0.34 以上の結論と原因の組に対して，因果関係の連鎖があると判断した．

図 2，3 に各々取得された因果関係上位 25 件の連鎖を NetworkX で示し，その内の一部の因果関係の連鎖を詳細に示す．

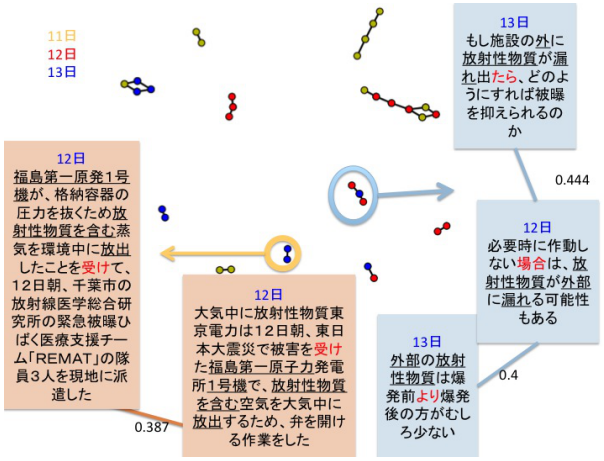

図 2: Jaccard 係数による因果関係の連鎖

3.3 考察

図 2 の 12 日から 13 日の因果関係の連鎖を見ると，Jaccard 係数が 0.4 以上の値で「放射性物質の拡散」についての因果関係がとれていることがわかる．Jaccard 係数による語彙の一致では，文の中で重要な語彙が一致していたため因果関係の話題が一貫しているものが連鎖として取れていた．一方，図 3 の $Jaccard+Wordnet$ を使用した場合は，上位 25 件の因果関係の連鎖の中には，「放射性物質の拡散」についての連鎖は入ってお

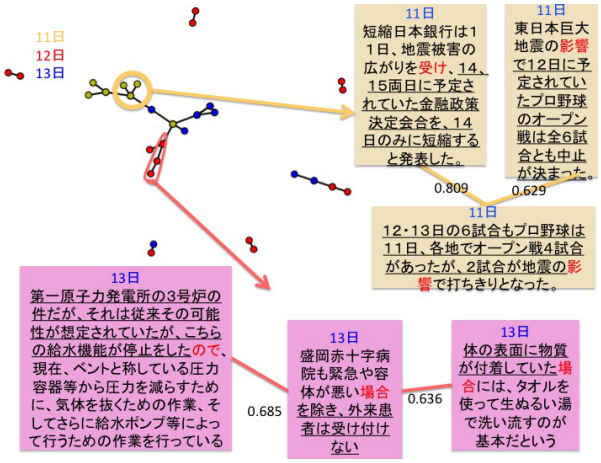

図 3: $Jaccard + WordNet$ による因果関係の連鎖

らず，値が高いとされる因果関係の連鎖も「金融」と「プロ野球の開幕」の話が繋がってしまっており，内容が不明瞭なものが多くなってると判断する．このことは 3 月 11 日から 13 日の 3 日という短期間において 1 つの話題に対して統一された語彙が使われていたことを示している．これにより，Jaccard 係数を用いた語彙の一致の方が良い結果がでた．$Jaccard+WordNet$ では因果関係に対して，不要な繋がりを取ってしまったと考えられる．

4 おわりに

東関東大震災に関する 3 月 11 日から 13 日までの 3 日間の記事において因果関係の連鎖を抽出することを試みた．本研究では 2 つの指標を導入し因果の連鎖を抽出することを試みたが，結果として語彙間の類義関係を考慮しない Jaccard 係数基づく因果関係連鎖の判定が優れているという結果になった．このことは短い期間での新聞記事は統一された語彙を使用する傾向が強く，語彙の類義関係まで捉えた因果関係連鎖の抽出は却って不正確なものになってしまったためと考える．今回使用した Jaccard 係数は異なる語彙集合 2 つの類似度を測ったが，他にも 1 つの語彙集合からもう一方の語彙集合の被覆度を考慮した指標などもあり，今後様々な指標の導入を検討したい．

参考文献

[1] 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.

[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 年.

[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] 日本語 WordNet http://http://nlpwww.nict.go.jp/wn-ja/

[6] NetworkX http://http://networkx.lanl.gov/