

# 複数文書内の事象間の因果関係抽出への取り組み

澤村 瞳<sup>†</sup>      小林一郎<sup>‡</sup>

<sup>†</sup>お茶の水女子大学 理学部 情報科学科

<sup>‡</sup>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

## 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、文章中に現れる手がかり表現に着目し、それらの表現をきっかけに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繋ぐため、原因および結果の表現間の柔軟なマッチングを導入する。

## 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、節を繋ぐ際の接続詞に表現される因果関係を抽出すること [1] や構文パターンを捉えて因果関係を抽出する手法 [2]、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 [3]。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、節間関係を示す以下の表層表現と手がかり標識に基づいて因果関係を抽出する。

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, ～から*, ～せい
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」

### 2.2 因果関係の連鎖

因果関係を連鎖するためには、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現されている語彙を対象に Jaccard 係数に基づき柔軟に類似性を取ることににより、その繋がりを捉えることにする。その指標を以下に示す。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) で表わされる。

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

### 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。

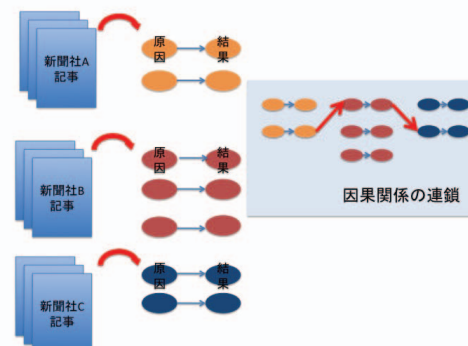


図 1: 因果関係の抽出と連鎖構築の概要

#### 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) または手がかり標識があるものを集める。

#### step 2. 原因と結果のペア生成

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、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

1 文の「結果」と他の文の「原因」に含まれる単語の一致を Jaccard 係数によって測り、閾値を越えるものを因果関係の連鎖として採用する。

An Approach to Extraction of Causal Relation among Events in Multiple Documents

<sup>†</sup>Hitomi SAWAMURA(g0820518@is.ocha.ac.jp),

<sup>‡</sup>Ichiro KOBAYASHI(koba@is.ocha.ac.jp)

<sup>†</sup>Dept. of Information Sciences, Faculty of Science, Ochanomizu University, 2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610

<sup>‡</sup>Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University, 2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610

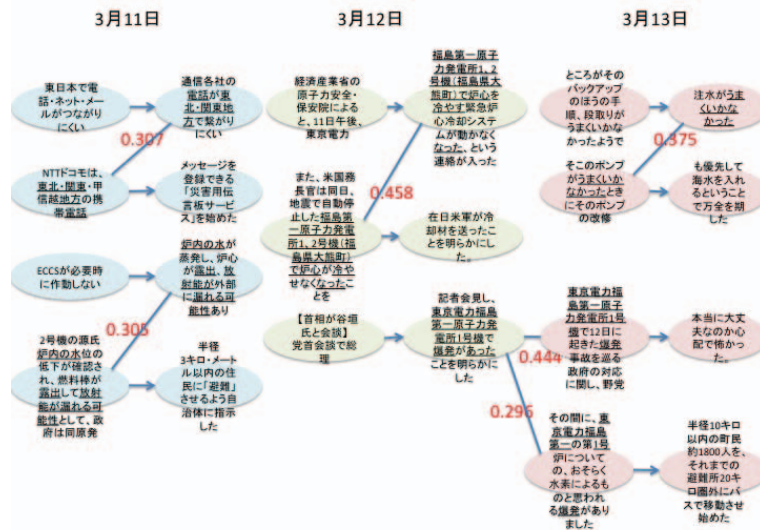


図 2: 抽出された因果関係連鎖の一部

### 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係数の値を基に調査した。

#### 3.2 実験結果

1文から抽出された因果関係に対して、結果と原因の総当たりの組を3日間において求め、Jaccard係数が高い上位100件に対して、実際の因果関係があるかを人手で判断した。その結果、Jaccard係数の値が0.29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。これにより求められた因果関係の数を集計したものを表2に示す。

表 2: 3日間の抽出された因果関係の数

	11日	12日	13日
11日	11	0	0
12日		29	32
13日			34

図2に取得された因果関係の連鎖の一部を示す。

#### 3.3 考察

表2において、3月11日の記事において、因果関係が見られるのは抽出した上位100個中11個確認されたが、3月11日の新聞記事の中にある因果関係が12日、13日より少ない結果になっていることがわかる。これは、11日の新聞記事は大災害の内容を簡潔に印象強く短い文で表現された記事が多かったためであることが当日の新聞記事を調べることでより分かった。

また、11日から翌日にかけて、さらに翌々日にかけては、因果関係は一つも抽出されていないが、これは、11日の新聞記事に因果関係を示す文そのものが少なかったことと、因果関係の連鎖の発見に語彙の完全一致のみを採用しているために、本来、抽出されるべき因果関係の連鎖が抽出できなかった可能性などが考えられる。12日は巨大地震が津波を起し、様々な災害を引き起こしたことが報道され、11日より多くの因果関係が抽出されている。13日はさらにその数が増えていること、また、12日から13日に亘る因果関係の連鎖の数が12日の数より増えていることから災害の影響が拡大しながら連鎖していることがわかる。

### 4 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。現時点では、結果と原因を単純に表現の一致だけに基づいて因果関係を繋げることをしているが、今後は、単語集合の被覆関係や単語の類義関係なども考慮して、より柔軟に因果関係の連鎖を抽出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回年次大会, 2011年.
- [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] 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.
- [4] 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.