2021 年 8 月 26 日

# 言語処理を用いたテキスト分析

≪因果関係抽出手法≫

- 因果関係が文に含まれているかどうかを判定.
- 因果関係が含まれている文から，**原因表現**と**結果表現**を抽出.

## 1. 金融分野での因果情報の重要性

- 膨大な金融情報を分析し，因果関係を抽出することで投資家の投資判断を手助けする手法が注目されている.
- 本章では，因果関係を抽出する上で重要な手掛かりとなる表現（**手がかり表現**）を利用して，因果関係を自動で抽出する手法を紹介.


因果関係イメージ

サブプライムローンの危機 **により** ，世界不況が起こった.

原因表現　　手がかり表現　　結果表現

## 2. 手がかり表現

- 手がかり表現から因果関係を抽出する.
- しかしながら，手がかり表現は因果関係以外を表すことがあることに注意する（日本語の難しいところ…）.

【例文①】　**因果関係を表す**

- 日本市場では消費者などの抵抗感**から**，遺伝子非組み換え品に限定していない一般大豆のニーズが減退している.

【例文②】　**因果関係は表さない**

- 同社は全国の延べ約十万人の会員**から**，約九十億を会員費として集めている.

- 一方で，手がかり表現を持たないが，因果関係を持つ文章も存在する．今回は，手がかり表現によって示されれている因果関係を扱う.

【例文③】　原因表現　　結果表現

- 高松市でも送電に落雷があり，市内の約 50%の世帯で 12 分停電した.

## 3. 因果関係の判定

- 手がかり表現が因果関係を持つか持たないかを判定するために，**因果関係判定手法**を用いる．ここでは，機械学習手法（Support Vecter Machine：SVM）を用いた.
- ここで扱う因果関係は，原因もしくは理由と結果を示し，手がかり表現を伴って出現するものに限定.

**因果関係を含む文**

- こうした経済指標やユーロ現金導入に**伴う**消費動向の微妙な変化を**背景に**，米同時テロ以降に広がった景気悲観論は急速に後退している．
- サリドマイドやスモン被害**を受け**，同省は一九八〇年に医薬品の副作用被害を救済する制度を創設した．

**因果関係を含まない文**

- 長野県信用組合(長野市，丸山彰一理事長)は五日**から**，法人向けにインターネットバンキングの取り扱いを始める．
- 大きさは幅が約三十二・八センチ，奥行き約三十・六センチ，高さ約四十八・一センチ**で**，重さは約三・七キロ．

- フィルタリング手法で用いる素性[1]として採用したもの．

**構文的な素性**　（日本語文において因果関係を表すためによく用いられる表現を利用するため）

- 助詞のペア

**意味的な素性**　（因果関係を示す語彙の関係を利用する）

- 拡張言語オントロジー

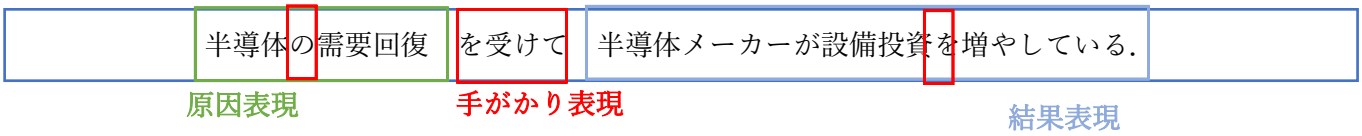
**それ以外の素性**

- 手がかり表現の直前形態素の品詞
- 文に含まれる手がかり表現
- 形態素ユニグラム
- 形態素バイグラム

≪構文的な素性の例≫

【例文④】

➡構文解析を用いて，手がかり表現に関係のある助詞（~の，~を）だけを素性として獲得．



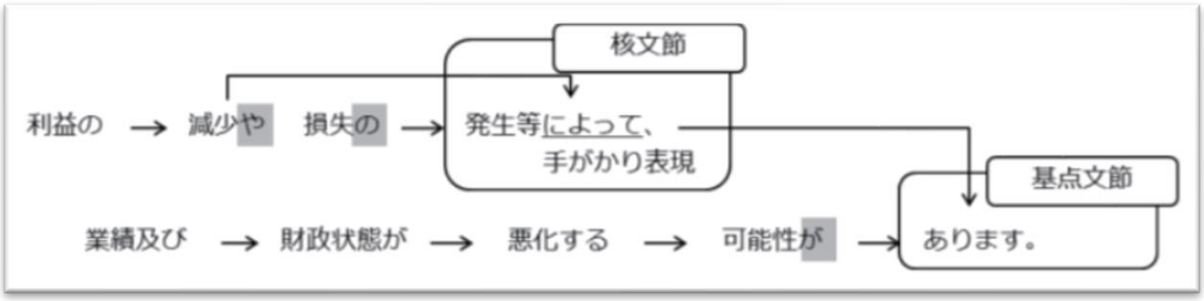

## 3.1 構文的な素性（助詞のペア）

- 特徴を表している助詞のみを獲得する必要がある（素性として役に立たないため）．
- 構文情報を用いて助詞を獲得することで，手がかり表現に依存している助詞を素性として獲得する．
- 助詞ペアの抽出は，Extraction of pairs of particles に従う（P89 参照）．

≪用語の定義≫

- 核文節　　：手がかり表現を含む最後尾の文節
- 基点文節　　：核文節の係り先となる文節

[1] 機械学習の際，データを分類する手がかりとなる情報のこと．「学習素性」ともいう．例えば，単語の品詞を推定するモデルを機械学習する際には，その単語の前後に出現する語や品詞が学習素性として使われる．言い換えれば，前後に出現する語や品詞を手がかりとして対象語の品詞を推定するモデルを学習する．学習素性として何を使うかは，機械学習に基づく自然言語処理の成否を決める重要な要因である．

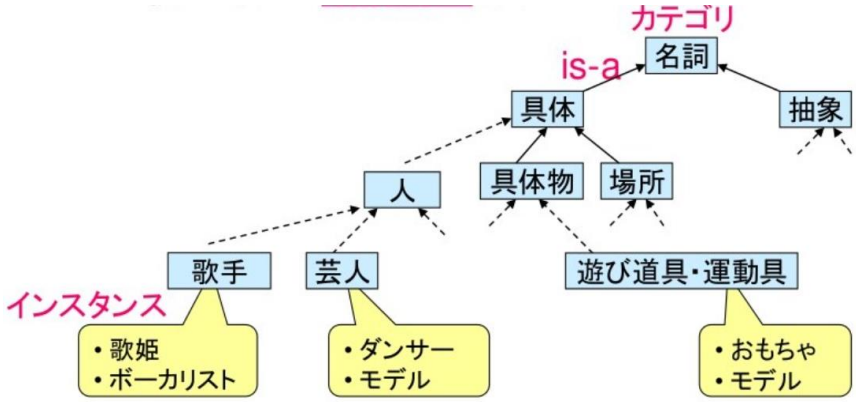



核文節と基点文節の例

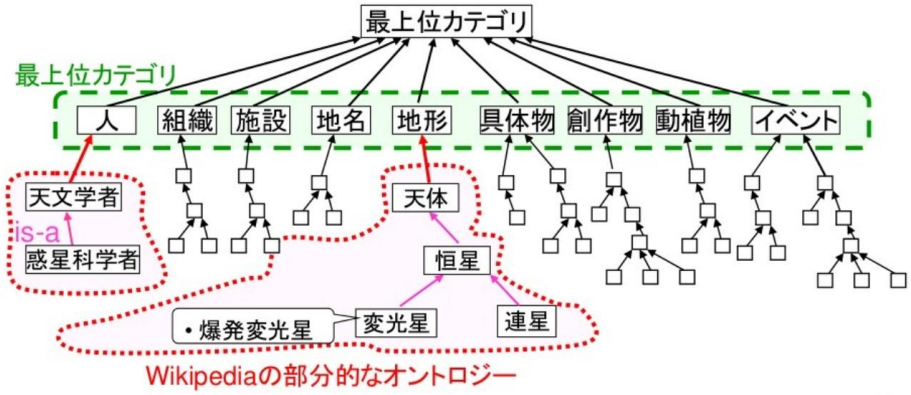

助詞ペアの例

## 3.2 意味的な素性（拡張言語オントロジー）

- 小林らが作成した言語オントロジーを拡張言語オントロジーと定義して，利用する．
- 今回は，日本語語彙体系から作成された拡張言語オントロジーを用いる．
  - 日本語語彙体系[2]とは，人手で作成された$is-a$関係からなる大規模なオントロジーであり，1 つに統一された階層構造を持つ．カテゴリは 3000 件で，インスタンスは 30 万件．最上位のカテゴリはジャンルを分類するためのカテゴリになる．




日本語語彙体系のイメージ



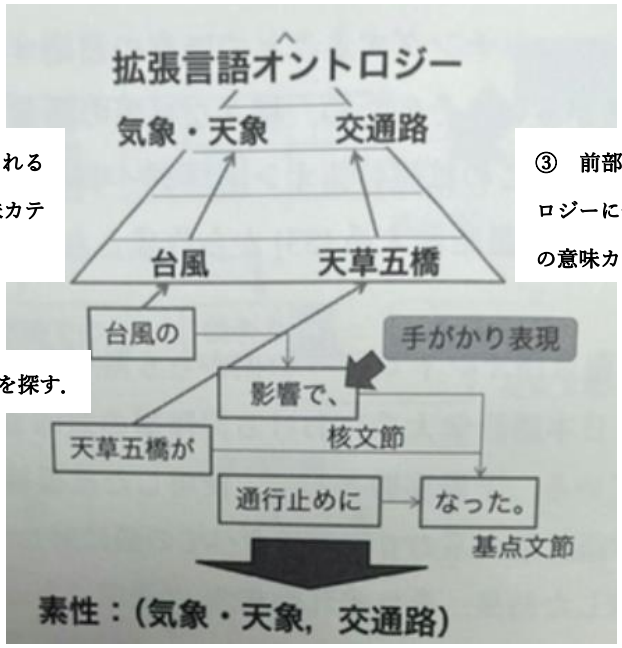

最上位カテゴリのイメージ

[2] 参考資料：https://www.slideshare.net/jnlp/11-nlp-shibaki

- 本資料では，拡張言語オントロジーの上から 6 階層目の意味カテゴリを素性として使用（理由については，P91 の 11 行目から参照）.

≪拡張言語オントロジー素性の取得例≫

② 文節に拡張言語オントロジーに含まれる語（台風）があれば，上から 6 層目の意味カテゴリを前部語彙とする.




③ 前部語彙同様に，文節に拡張言語オントロジーに含まれる語があれば，上から 6 層目の意味カテゴリを後部語彙として獲得.

① 核文節に係っている文節（台風の）を探す.

## 3.3 タグなしデータからの追加学習データの獲得

- タグが付けられた学習データを用いて，タグなしデータから追加学習データを獲得するにあたり，手がかり表現が持つ意味に着目.
- 以下の文を手がかり表現「のため」に置換すると，因果関係が見える．この性質を利用して追加学習データを獲得する.

円高により，日本経済が悪化した.

円高 | のため | 日本経済が悪化した.

- 因果関係を持たない文では，手がかり表現とその前後に因果関係がないことを示す必要がある.
- 追加データを獲得する手続き（Extracting additional learning data）は P93 参照.

## 4. 因果関係の抽出

因果関係を抽出するにあたり，原因結果をそれぞれ「原因表現」,「結果表現」と定義する.
このとき，手がかり表現と原因表現・結果表現の出現位置を 5 通りに分類した（P95 参照）.
本研究では，この 5 通りのパターンから因果関係を獲得するアルゴリズムを用いて因果関係を抽出.

## 5. 因果関係抽出結果

業績発表記事~因果関係を抽出．学習データは，1995 年～2005 年の日経新聞記事からランダムに抽出した「手

がかり表現」を含む 1000 文を使用.
テキストに実際の結果が記載（P96～97）.

表 6.3　手がかり表現の一覧.

| | | | | |
|---|---|---|---|---|
| を背景に | を背景に、 | を受け、 | ため、 | に伴う |
| を反映して | をきっかけに | により、 | に支えられて | を反映し、 |
| が響き、 | ためで、 | を受けて | から、 | が響いた。 |
| ため」 | が影響した。 | による。 | を受けて、 | に伴い |
| ため。 | が響く | が響いている | が響いている。 | で、 |
| このため、 | その結果、 | この結果、 | に伴い、 | ためだ。 |
| によって | により | ためで | このため | |

表 6.4　抽出した因果関係の例.

| | |
|---|---|
| 原因 | 主要納入先の自動車や半導体などが設備投資を削減した |
| 結果 | 売上高、経常利益の予想を下方修正。 |
| 原因 | 日産自動車など国内メーカーが減産した |
| 結果 | 純正カーステレオの売り上げが一割弱減少した。 |
| 原因 | 利益率の高い土木製品が一五%減った |
| 結果 | 会社設立来初の経常、最終赤字。 |
| 原因 | 個人消費の低迷と冷夏で、子供服とベビー服の販売が不振で八%の減収になる |
| 結果 | キムラタンは十五日、九四年三月期の経常損益が五億円の赤字になる見通しだ、と発表した。 |

## 6. 意外な因果関係の抽出

**原因**　猛暑　➡　飲料水の売り上げ好調　（なんとなくわかる）
**結果**　猛暑　➡　飲料水を運ぶ**段ボール**の売り上げ好調　（思いつきにくい…）

因果関係の抽出の為のスコアリングについて紹介.

≪会社に関連するキーワードのスコアリング≫

$$Score(w, cp) = \frac{W\big(w, S(cp)\big)}{max_{w'} W\big(w', S(cp)\big)} \tag{6.1}$$

$$W\big(w, S(cp)\big) = tf\big(w, S(cp)\big) H\big(w, S(cp)\big) \log_2 \frac{N}{df(w)} \tag{6.2}$$

$$H\big(w, S(cp)\big) = -\sum_{d \in S(cp)} P(w, d) \log_2 P(w, d) \tag{6.3}$$

$$P(w, d) = \frac{f(w, d)}{\sum_{d' \in S(cp)} f(w, d')} \tag{6.4}$$

$S(cp)$　：会社$cp$の決算短信 PDF の集合

$tf(w, S(cp))$　：$S(cp)$に含まれる単語$w$の頻度

$N$　：決算短信 PDF の数

$H(w, S(cp))$　：PDF $d$ に出現する単語$e$の出現確率$P(w, d)$に基づくエントロピー