

# 構文パターンを用いた因果関係の抽出

坂地 泰紀 (豊橋技術科学大学 sakaji@smlab.tutkie.tut.ac.jp)

竹内 康介 (ランゲージ・クラフト研究所)

関根 聡 (ニューヨーク大学 sekine@cs.nyu.edu)

増山 繁 (豊橋技術科学大学 masuyama@tutkie.tut.ac.jp)

## 1 はじめに

因果関係に関する知識は、自動要約や質問応答システムなど、幅広い自然言語アプリケーションにとって重要な知識のひとつである。しかしながら、大規模な文章集合から因果関係に関する知識を手で獲得するのはコストと時間がかかり、かつ世の中の因果関係を全て書き出すのは不可能である。そこで、我々は 4 つの構文パターンと手がかりとなる表現 (以下、手がかり表現) を用いて因果関係に関する知識を自動的に抽出する手法を提案する。データには因果関係に関する知識が多く含まれている景気動向に関する記述がある記事 (以下、景気動向記事) を用いた。今回、我々は因果関係を抽出する上で重要となる手がかり表現の調査を行った後、因果関係の抽出を試みた。

### 1.1 因果関係に関する知識

今回、我々が扱う因果関係に関する知識は、テキスト上に直接現れる表層的な因果知識である。本稿では、それらは「原因・理由」を示す表現と、それに対応する「出来事」、もしくは「判断」を示す表現の組で構成されるものを指す。**原因・理由**は出来事の原因・理由を表す場合と、判断の理由 (根拠) の場合がある [3]。

「円高が進行したため、輸出が減少した」という文の場合、「円高が進行した」という表現は「輸出が減少した」という表現の根拠となる。根拠を表す表現 (以下、根拠表現) 「円高が進行した」、結果を表す表現 (以下、結果表現) 「輸出が減少した」の組が因果関係に関する知識となる。ここで、「ため、」は手がかり表現となる。

なお、「円安になったため、日本の景気が上昇した」という文の場合、本来は「円安になった」と「日本の景気が上昇した」の間に、「輸出が増加する」、「日本は加工貿易国である」などの背景知識がある。しかしながら、背景知識を抽出することは難しい。そのため、本論文ではテキスト上に直接現れる因果知識だけを扱い、背景知識は扱わない。

## 2 関連研究

乾らは接続標識「ため」を含む複文から因果関係に関する知識を獲得し、それらを 4 つの因果関係 (*cause*, *effect*, *precond*, *means*) に自動的に分類する手法を提案している [1]。乾らが因果関係に関する知識を獲得するために対象とした文は、**タメ複文**と名付けられた、接続標識「ため」を含む複文のみである。この方法では、「ため」以外の手がかりとなる語や複文ではない場合に、因果関係に関する知識を取得できないため、我々は様々な手がかり表現を使用し、かつ手がかり表現を含む全ての文を因果知識獲得の対象にしている。我々が今回扱う「原因・理由」を

表す因果知識は、乾らの *cause* 関係に対応する。

佐藤らは重文、複文のみを対象として、格フレームを用いて因果知識を獲得している [5]。それに対して、我々は重文、複文に限らず、手がかり表現を含む全ての文を対象にして因果関係に関する知識獲得を行っている。また、我々は格フレームではなく、構文パターンを用いて因果関係に関する知識を獲得している。

## 3 手がかり表現の調査

因果関係に関する知識を獲得するにあたり、手がかり表現の数と種類について調査を行う。データは因果関係に関する知識が多く含まれていると考えられる景気動向記事 300 記事を用いる。景気動向記事は坂地ら [6] の手法を用いて、1990 年から 2005 年の日経新聞から獲得した。

### 3.1 タグ付けルール

手がかり表現の調査をするため、景気動向記事 300 記事に以下のタグを付与する。

**CLUE タグ** 手がかり表現を示すタグ

**BASIS\_VP タグ** 根拠表現で、かつ動詞句を示すタグ

**BASIS\_NP タグ** 根拠表現で、かつ名詞句を示すタグ

**RESULT タグ** 結果表現を示すタグ

**RESULT\_SBJ タグ** 結果表現の主部を示すタグ

**RESULT\_PRED タグ** 結果表現の述部を示すタグ

**RESULT\_BF タグ** 結果表現でかつ、根拠表現が根拠表現の前に出現する場合を示すタグ

**INV タグ** 名詞句と動詞句を反転させる語を示すタグ

**IGNORE タグ** 無視する語を示すタグ

手がかり表現には、助詞や読点を含める。助詞の種類や読点の有無によって、手がかり表現が因果関係に使われる割合が異なる可能性があると考えたためである。

結果表現の主部と述部が連続して出現する場合は、結果表現の主部と述部を合わせて **RESULT** タグを付与する。結果表現の主部と述部が離れて出現する場合は、主部には **RESULT\_SBJ** タグを、述部には **RESULT\_PRED** タグを付与する。

調査を進めると、手がかり表現によって、根拠表現となる表現が名詞句か動詞句かが決まっているという傾向が見られた。ただ、形式名詞の「の」、「こと」があると、その主要部分は動詞句であるのに、名詞句に対応する手がかり表現が使われることになる。この点を明確にするために形式名詞などに対して、**INV** というタグを付与した。動詞句と名詞句については、野呂ら [2] を参照した。

### 3.2 タグ付け結果

乾らは手がかり表現 (乾らの論文では手がかり標識と呼んでいる) について調査した結果を報告している [4]. 乾らは調査の対象テキストとして, 毎日新聞 1995 年度版の社会面の記事を用いた. 乾らが調査した手がかり表現を出現頻度と共に表 2 に示す. 我々が調査した結果と, 乾らが調査した結果を表 1 に示す.

表 1 調査した記事数, 手がかり表現の総数, 手がかり表現の異なり数

|         | 記事数 | 総数  | 異なり数 |
|---------|-----|-----|------|
| 乾らの調査結果 | 750 | 219 | 34   |
| 我々の調査結果 | 300 | 695 | 154  |

表 1 より, 記事あたりの手がかり表現数が乾らの調査結果に比べ, 我々の方が多くなっている. これは, 読点や助詞の扱いは異なるが, 我々が用いた景気動向記事が因果関係を多く含むことが主因であると考えられる. そのため, 因果関係に関する知識を抽出する目的で使用するには, 景気動向記事は良いコーパスだと言える.

我々が調査した手がかり表現について, それが手がかり表現として使用された頻度, コーパス中の頻度 (総数), 手がかり表現として使用された割合を表 3 に示す. 表 3 には, 頻度が高い 10 個の手がかり表現を示してある.

表 2 乾らが調査した手がかり表現

|                                                                           | 頻度  |
|---------------------------------------------------------------------------|-----|
| ため                                                                        | 120 |
| で                                                                         | 35  |
| 結果 ので と                                                                   | 5   |
| 場合 ば ことから                                                                 | 4   |
| から                                                                        | 3   |
| 理由で 目的で 影響で より ように<br>よう として ところ が                                        | 2   |
| 背景には 続いて 事故で 事件で 取り入れようと<br>際に 際 限り れると み による ており<br>せようと ことで うとして ことによって | 1   |

表 3 手がかり表現として使用された頻度 (頻度), コーパス中の頻度 (総数), 手がかり表現として使用された割合 (割合)

|      | 頻度  | 総数   | 割合    |
|------|-----|------|-------|
| で    | 155 | 2385 | 0.065 |
| による  | 48  | 244  | 0.197 |
| で、   | 48  | 340  | 0.141 |
| から   | 46  | 646  | 0.071 |
| ため、  | 30  | 55   | 0.545 |
| を背景に | 24  | 26   | 0.923 |
| から、  | 17  | 46   | 0.367 |
| の影響が | 14  | 18   | 0.778 |
| に伴う  | 14  | 23   | 0.609 |
| を受け、 | 12  | 14   | 0.857 |

表 3 より, 我々が手がかり表現を調査した結果, 「で」や「から」, 「による」などの手がかり表現が数多く出現することが分かった. しかしながら, これらの表現は因果関係に用いられる割合は小さいため, 因果関係に関する知識を自動的に取得するために用いるのは適切ではないと考えられる. そこで, 因果関係に

関する知識をより多く正確に取得するために, 適切な手がかり表現を獲得する. 手がかり表現に以下に示すスコアを付与し, 順位付けを行った.

$$Score = \log(TF(t_i)) \times R(t_i) \quad (1)$$

ここで,  $TF(t_i)$  は表現  $t_i$  が手がかり表現として使用された頻度である.  $R(t_i)$  は表現  $t_i$  が手がかり表現として使用された割合である. スコアの高い手がかり表現 10 個を表 4 に示す.

表 4 降順に並べたスコアが高い表現

|       | 頻度 | 総数 | 割合    | スコア   |
|-------|----|----|-------|-------|
| を背景に  | 24 | 26 | 0.923 | 1.274 |
| を背景に、 | 10 | 10 | 1.000 | 1.000 |
| を受け、  | 12 | 14 | 0.857 | 0.925 |
| を挙げる  | 10 | 12 | 0.833 | 0.833 |
| ため、   | 30 | 55 | 0.545 | 0.806 |
| に伴う   | 14 | 23 | 0.609 | 0.698 |
| に伴い、  | 6  | 7  | 0.857 | 0.667 |
| を反映して | 6  | 7  | 0.857 | 0.667 |
| に加え、  | 9  | 13 | 0.692 | 0.661 |

表 2, 表 3 より, 乾らの手がかり表現と, 本稿における手がかり表現の定義は異なるが, 頻度の高い順に並べた手がかり表現は乾らが調査した手がかり表現と似ているものが多い. しかしながら, 乾らの調査では「ため」が最も頻度が高くなっているのに対して, 我々の調査では「で」が最も頻度が高い手がかり表現になっている. これは, 本稿では手がかり表現に読点を含めたため, 「ため」と「ため、」が別の手がかり表現として扱われたことと, 分野による違いに起因すると考えられる.

## 4 因果関係に関する知識抽出パターン

根拠表現の位置は手がかり表現の直前に出現する. それに対して, 結果表現は様々な場所に出現する. そのことから, 結果表現の位置によって, 抽出方法が異なってくるため, 以下の 4 つの構文パターンを用いた.

Pattern A: 結果表現の主部と述部が存在する場合

Pattern B: 手がかり表現の直後に結果表現が存在する場合

Pattern C: 結果表現が前の文である場合

Pattern D: 結果表現が根拠表現より前に出現する場合

ここで, 手がかり表現が含まれる最後尾の文節を手がかり表現の**核文節**, 核文節の係り先の文節を**基点文節**と定義する. 因果関係に関する知識の抽出手法の概要を以下に示す.

Step 1: 手がかり表現が含まれる文を探す.

Step 2: 手がかり表現に「。」が含まれている, もしくは手がかり表現の後に「。」があるなら, Step 4 を実行する. そうでなければ, Step 3 を実行する.

Step 3: 基点文節が動詞句であり, かつ基点文節が係り先である文節の中に係り助詞, もしくは格助詞を含むものがあれば, Pattern A を適用して終了する. そうでなければ, Pattern B を適用して終了する.

Step 4: 核文節に係っている文節に係り助詞が含まれている場合, Pattern D を適用して終了する. そうでなければ,

Pattern C を適用して終了する。

Pattern A と B, Pattern C と D の根拠表現の獲得方法は同じである。Pattern A, B は核文節の前に存在する文節に、核文節の後続に存在する文節に係る文節が出現するまで、次々と繋ぎ合わせていったものを根拠表現として獲得する。Pattern C, D は核文節の前にある文節に、核文節に係り、かつ接続助詞、もしくは係り助詞を含む文節が出現するまで、次々と繋ぎ合わせていったものを根拠表現として獲得する。

以下の節ではパターンごとの結果表現の獲得方法について述べる。

#### 4.1 Pattern A

基点文節の前に存在する文節を、核文節が現れるまで、次々と基点文節へ繋ぎ合わせていったものを結果表現の述部として獲得する。基点文節に係り先である文節に係り助詞、もしくは格助詞が含まれている場合、その文節に係り先である文節を次々と繋ぎ合わせたものを結果表現の主部として獲得する。

例えば、図 1 の場合、基点文節「下落気味となろう。」に、文節「総じて」を結合することで結果表現の述部「総じて下落気味となろう。」を獲得している。また、基点文節「下落気味となろう。」に係る文節「国際価格は、」には係り助詞が含まれているため、それに係る文節「農産物の」を結合し、結果表現の主部「農産物の国際価格は、」を獲得している。

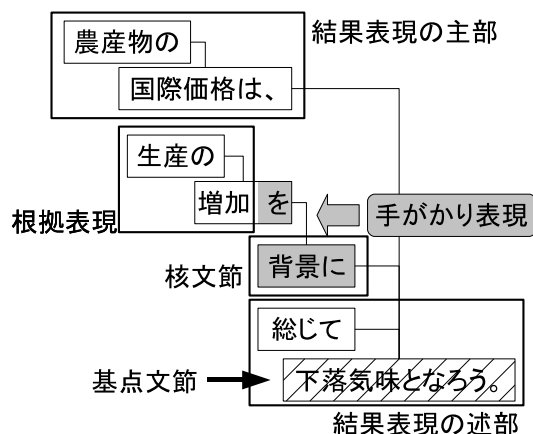


図 1 Pattern A の例

#### 4.2 Pattern B

基点文節が動詞句なら、Pattern A の結果表現の述部と同様の獲得方法で結果表現を獲得する。基点文節が名詞句なら、基点文節から順に次の文節を辿っていく。もし、現在注目している文節の係り先が、核文節の前にある文節から係られている場合、核文節から現在注目している文節までの文節と注目している文節から助詞を取り除いたものを連結させ、それを結果表現として獲得する。文節に格助詞が含まれる場合、文節に「ほうが」、「方が」が含まれている場合、文節に「など、」が含まれている場合も同様の処理を行うことで、結果表現を獲得する。

例えば、図 2 の場合、基点文節「復興需要が」は名詞句であるため、そこから次の文節へと辿っていく。基点文節「復興需要が」に係る文節「一息ついた。」は核文節より前にある文節から係られているため、基点文節「復興需要が」から助詞を取り除い

たものを結果表現「復興需要」として獲得している。

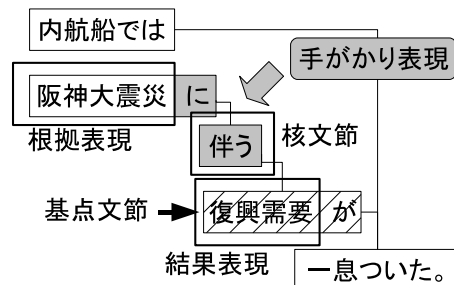


図 2 Pattern B の例

#### 4.3 Pattern C

手がかり表現が含まれる文の直前に存在する文を、結果表現として獲得する。

#### 4.4 Pattern D

核文節に係る文節に係り助詞が含まれる場合、その文節から助詞や読点、形式名詞「の」を取り除いたものに、その文節に係り先である文節を、次々と繋ぎ合わせたものを結果表現として獲得する。

例えば、図 3 の場合、手がかり表現の核文節「ためだ。」に係り先の文節「あるのは、」には係り助詞が含まれているため、助詞や読点、名詞「の」を取り除き、それに係る文節を結合し、結果表現「生産に下げどまりの動きがある」を獲得している。

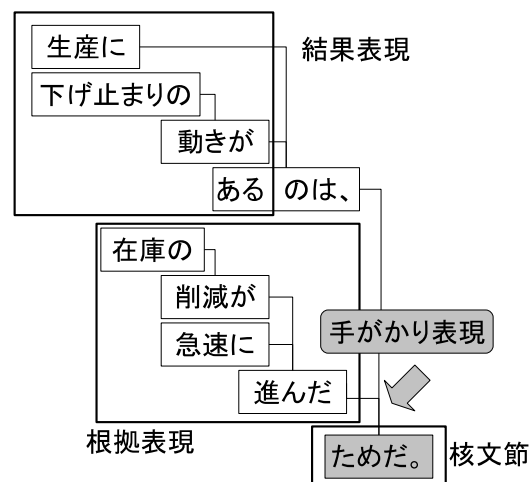


図 3 Pattern D の例

## 5 実験と考察

実験には新たにタグを付けた景気動向記事 200 記事と、2.2 で手がかり表現に付与したスコアの高い順 40 個の手がかり表現のうち、「で、」「ほか、」「ため」「による」を除いた 36 個の手がかり表現を用いた。「で、」「ほか、」「ため」「による」は 40 個の手がかり表現のうち、手がかり表現に使用された割合を降順に並べたときの下位 4 件である。係り受け解析には Cabocha\*1 を適用した。

本手法を用いて抽出してきた因果関係に関する知識を、すでにタグ付けされているデータに対して機械的に評価すると、タ

\*1 <http://chasen.org/~taku/software/cabocha/>

表5 根拠表現, 結果表現, 両方共正解の場合, それぞれの適合率

| 手がかり表現   | 抽出数 | 適合率   |       |       |
|----------|-----|-------|-------|-------|
|          |     | 根拠表現  | 結果表現  | 両方    |
| を背景に (A) | 5   | 0.800 | 0.600 | 0.600 |
| を背景に (B) | 14  | 0.923 | 0.923 | 0.786 |
| を挙げる (A) | 5   | 0.400 | 0     | 0     |
| ため、(A)   | 6   | 0.500 | 0.200 | 0.333 |
| ため、(B)   | 24  | 0.708 | 0.664 | 0.667 |
| に伴う (B)  | 20  | 0.750 | 0.500 | 0.450 |
| に加え、(B)  | 10  | 0.400 | 0     | 0     |
| ためだ。(C)  | 9   | 0.889 | 0.889 | 0.889 |
| ためだ。(D)  | 6   | 1.000 | 0.833 | 0.833 |
| により (A)  | 9   | 0.778 | 0.667 | 0.667 |
| により (B)  | 11  | 0.455 | 0.273 | 0.091 |
| を受けて、(B) | 5   | 0.800 | 0.400 | 0.400 |
| の効果が (B) | 8   | 0.750 | 0     | 0     |
| の影響も (B) | 5   | 0.800 | 0     | 0     |
| の影響が (A) | 5   | 1.000 | 0     | 0     |
| の影響が (B) | 15  | 0.733 | 0     | 0     |
| により、(B)  | 14  | 0.929 | 0.786 | 0.786 |
| によって (A) | 5   | 0.600 | 0.600 | 0.600 |
| によって (B) | 10  | 0.500 | 0.500 | 0.500 |
| から、(A)   | 12  | 0.833 | 0.667 | 0.583 |
| から、(B)   | 30  | 0.700 | 0.567 | 0.533 |
| Total    | 292 | 0.757 | 0.526 | 0.449 |

表6 根拠表現と結果表現の失敗の原因

|      | 総失敗数 | 抽出失敗 | 過抽出 |
|------|------|------|-----|
| 根拠表現 | 71   | 20   | 51  |
| 結果表現 | 139  | 36   | 103 |

グ付けの際に誤って付けたタグが存在したり, 表現の範囲がほんの少しだけ異なるが意味的に同じ場合, 正しくないと判定されてしまう。そこで, 本手法を正確に評価するために, 実験結果をタグ付け者が判断した。

手がかり表現と4つの構文パターンを用いて因果関係に関する知識を抽出し, 根拠表現, 結果表現, 両方共正しかった場合, それぞれの適合率を求めた。抽出数が5以上であった手がかり表現と, 全体の抽出数と適合率を表5に示す。36個の手がかり表現で表されている因果関係全体の再現率は, 根拠表現抽出0.917, 結果表現抽出0.809であった。

表5より, 「を背景に」, 「により、」, 「ためだ。」などの手がかり表現を用いた場合, 高い適合率で因果関係に関する知識を抽出できている。しかしながら, 手がかり表現「を挙げる」, 「に加え、」, 「の影響が」などを用いた場合, 結果表現の取得は全て失敗している。これは, 手がかり表現「を挙げる」, 「に加え、」, 「の影響が」などを用いた場合, 本来, 結果表現でないものを取り得てしまうためである。

## 6 エラー解析

エラーの原因を調べたところ, 構文パターンによる抽出の失敗(抽出失敗)と, 因果関係を含んでいないものから抽出したものの(過抽出)の2種類のエラーがあった。抽出失敗の数と過抽出の数を調べた結果を表6に示す。

抽出失敗の原因を調査するために, 過抽出を除いたデータ

表7 結果表現, 根拠表現の両方が含まれているデータから抽出したときの抽出数, パターンごとの適合率

|           | 抽出数 | 適合率   |       |       |
|-----------|-----|-------|-------|-------|
|           |     | 根拠表現  | 結果表現  | 両方    |
| Pattern A | 38  | 0.978 | 0.816 | 0.747 |
| Pattern B | 113 | 0.901 | 0.761 | 0.761 |
| Pattern C | 13  | 0.867 | 0.917 | 0.846 |
| Pattern D | 6   | 1.000 | 1.000 | 1.000 |

を用いて, パターンごとの適合率と, 根拠表現と結果表現の両方が含まれているデータから抽出したときの抽出数を求めた。結果を表7に示す。例えば, 「**第一は<BASIS\_NP>バブル崩壊</BASIS\_NP><CLUE>による</CLUE><RESULT>株価、地価など資産価格の下落</RESULT>だ。**」という文から, 抽出を試みたとき, 本来, 結果表現として「株価、地価など資産価値の下落」を抽出すべきところで, 「株価」を抽出してしまっていた。これは, 構文パターンが対応できていないことにより, 失敗したと考えられる。この結果から, 構文パターンに改良の余地があると考えている。

過抽出の例を以下に示す。例えば, 「意見は業種によって割れた。」という文は, 手がかり表現「によって」を含んでいるため, 抽出対象となる。しかしながら, 因果関係を含んでいないため, 過抽出となってしまう。過抽出を減らすことは今後の課題の一つであるが, 例を分析したところ, 根拠表現, 結果表現に含まれる動詞の特徴を利用することで, 過抽出を減らすことができると考えている。

## 7 まとめ

我々は因果関係に関する知識を取得するため, 文章中に手がかりとなる表現の総数と因果関係に用いられている数を調査した。取得した手がかり表現と4つの構文パターンを用いて, 因果関係に関する知識の抽出を試みた。「を背景に」, 「により、」, 「ためだ。」などの手がかり表現を用いた場合は, 高い適合率で抽出できることを確認した。

他の様々な分野のコーパスから手がかり表現を獲得し, 手がかり表現の分布を調査することを考えている。

## 参考文献

- [1] 乾孝司, 乾健太郎, 松本裕治. 接続標識「ため」に基づく文章集合からの因果関係知識の自動獲得. 情報処理学会論文誌, Vol. 45, No. 3, pp. 919–933, 2004.
- [2] 野呂智哉, 橋本泰一, 徳永健伸, 田中穂積. 大規模日本語文法の開発. 自然言語処理, Vol. 12, No. 1, pp. 3–32, 2005.
- [3] 庵功雄. 新しい日本語学入門. スリーエーネットワーク, 2001.
- [4] 乾孝司, 奥村学. 文章内に現れる因果関係の出現特性調査. 計量国語学, 第25巻, pp. 123–144, 2005.
- [5] 佐藤浩史, 笠原要, 松澤和光. テキスト上の表層的因果知識の獲得とその応用. 信学技報 (TL98-23), pp. 27–34, 1999.
- [6] 坂地泰紀, 酒井浩之, 増山繁. 景気動向を示す根拠表現の抽出と分析. 電子情報通信学会技術研究報告「言語理解とコミュニケーション」, pp. 151–156, 2007.