論　文

決算短信 PDF からの原因・結果表現の抽出

坂地　泰紀[†a)]　　酒井　浩之[†]　　増山　繁[††]

Extracting Causal Expressions from PDF Files of Summary of Financial Statements

Hiroki SAKAJI[†a)], Hiroyuki SAKAI[†], and Shigeru MASUYAMA[††]

あらまし　本論文では，決算短信 PDF から原因・結果表現を抽出する手法を提案する．近年，証券市場における個人投資家の比重が増大しており，個人投資家に対して投資判断の支援を行う技術の必要性が高まっている．個人投資家にとって，企業に関する情報は投資判断において重要である．その情報の一つに，企業業績に関する原因・結果がある．例えば，原因「猛暑」，結果「冷房需要の盛り上がり」といった情報を投資家に提示することで，「猛暑」の場合には，「冷房需要」が高まる可能性があることを個人投資家が知ることができるというメリットがある．そこで，我々は，企業が Web ページに掲載する決算短信 PDF から原因・結果を自動的に抽出する手法の開発を行った．

キーワード　テキストマイニング，情報抽出，知識発見

1. まえがき

近年，人工知能分野の手法や技術を，金融市場における様々な場面に応用することが期待されており，例えば，膨大な金融情報を分析して投資判断を支援する技術が注目されている．更に，最近では証券市場における個人投資家の比重が増大しており，個人投資家に対して投資判断の支援を行う技術の必要性が高まっている．

投資家にとって，企業の業績に関する情報は，投資判断を行ううえで重要であるが，企業の業績だけでなく，その業績要因に含まれる原因と結果が重要である．例えば，原因「猛暑」，結果「冷房需要の盛り上がり」といった情報を投資家に提示することで，「猛暑」の場合には，「冷房需要」が高まる可能性があることを個人投資家が知ることができるというメリットがある．そして，その原因「猛暑」に対する結果「冷房需要の盛り上がり」から，猛暑の年には，冷房に関する事業を行っている企業の業績が好調に推移することが期待できる．しかしながら，証券市場の上場企業数は約 3,500 社と多いうえに，近年では年に 4 回，決算発表がある．更に，大幅な業績の修正を行う場合にも業績修正発表を行う必要があるため，人手によって多くの企業の業績要因に含まれる原因と結果を取得するのには多大な労力を要する．

そこで，本研究では，企業が Web ページに掲載する決算短信 PDF に着目し，決算短信 PDF から原因・結果を自動的に抽出する手法を提案する．本手法は大まかに分けて二つのステップからなる．一つ目のステップでは，素性に言語オントロジーや構文情報を用いた機械学習手法により，決算短信 PDF から原因・結果を含む文を抽出する．二つ目のステップでは，抽出した原因・結果を含む文から原因・結果表現 (次章で後述) を抽出する．原因・結果表現抽出には，構文情報を用いた Pattern を用い，更に原因・結果表現を抽出するために用いる手がかりとなる表現をブートストラップ法を用いて自動的に追加獲得する．

2. 本手法の概要

本章では，本手法全体の概要を説明する．ここで，原因・結果を，それぞれ，原因表現，及び，結果表現と本論文では定義する．以下に，本手法の概要を示す．

† 成蹊大学，武蔵野市
Seikei University, Musashino-shi, 180–8633 Japan
†† 豊橋技術科学大学，豊橋市
Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan
a) E-mail: hiroki_sakaji@st.seikei.ac.jp
DOI:10.14923/transinfj.2014JDP7119

Step 1: 各企業サイトから決算短信 PDF を収集し，収集した PDF をテキスト(注1)に変換する．
Step 2: 得られたテキストデータから，原因・結果判定手法 [1] を用いて，原因・結果を含む文を抽出する．
Step 3: 原因・結果を含んでいると判定された文から原因・結果表現抽出手法 (**4.** で後述) を用いて，原因を示す原因表現と結果を示す結果表現の対を抽出する．
□

前章で説明した二つのステップは，Step 2 と Step 3 に該当する．また，Step 1 にて，企業 3,821 社の企業 Web ページから 106,885 個の決算短信 PDF ファイルを取得した．

3. 原因・結果を含む文の抽出

本手法では，原因・結果を抽出するうえで重要な手がかりとなる表現 (手がかり表現と定義する) を利用して，原因・結果を抽出する．例えば，「ため」は，原因・結果を抽出するうえで重要な手がかり表現となる．しかしながら，手がかり表現には，原因・結果以外の意味をもつものがある．例えば，「あなたの**ため**に、花を買った。」という文中の手がかり表現「ため」は，原因ではなく，目的の意味を表している．このような場合に対応するために，まず，半教師有り学習を用いたフィルタリング手法 [1] を適用し，原因・結果を含む文を決算短信 PDF から抽出する．原因・結果を含む文を抽出する手法は機械学習手法 (SVM) を用いているため，次節で説明する素性を獲得する．

3.1 素　　性

原因・結果を含む文を抽出するために，文から特徴 (素性) を獲得する．本研究では，表 1 の素性を獲得する．

我々は，原因・結果を含むか否かの判定のため，構文的な素性，意味的な素性を用いる．構文的な素性を用いることにより，日本語文において原因・結果表現を表すためによく用いられる表現を利用するという狙いがある．例えば，「半導体**の**需要回復**を受けて**半導体メーカーが設備投資**を**増やしている。」という文に含まれる助詞と手がかり表現の並び「～の～を受けて～を～」が原因・結果を表している可能性が高い．そこで，構文解析を用いて手がかり表現に関係のある助詞だけを素性として獲得する．また，意味的な素性として拡張言語オントロジーを用いることにより，原因・結果を示す語彙の関係を利用するという狙いがある．各素性の具体的な獲得方法については，[1] を参照されたい．

表 1 素性の一覧
Table 1 List of features.

構文的な素性
- 助詞のペア

意味的な素性
- 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

(注1)：PDF をテキストに変換するツールとして pdftotext を用いた．

4. 原因・結果表現の抽出

本章では，決算短信 PDF からの原因・結果表現の抽出方法について述べる．本手法では，原因・結果表現を抽出するうえで重要な手がかりとなる表現 (手がかり表現と定義する) を利用して，原因・結果を含む文群から原因・結果表現を自動的に抽出する．文献 [2] に準拠し，原因・結果表現は，出来事 (結果) とその理由 (原因) の組から構成されるとするが，本論文では，1 文中，または，隣り合う 2 文中に直接表現されている表層的なものに限定する．例えば，「サブプライムローンの危機により，世界不況が起こった」という文の場合，「世界不況が起こった」は結果表現，「サブプライムローンの危機」は原因表現，「により」は手がかり表現となる．これらの結果と原因は，手がかり表現「により」によって明確に示されている．

我々は，経済新聞記事を調査することにより，手がかり表現と原因・結果表現の出現位置を 4 通りに分類するとともに，原因・結果表現を抽出するための手がかり表現を取得した [3]．その 4 通りの分類を Pattern A から D とした．本研究では，この四つの Pattern に新たな Pattern E を追加した五つの Pattern を用いる．五つの Pattern と各 Pattern の関連図を図 1 に示す．また，各 Pattern の例を図 2 に示す．本手法は，この 5 通りの Pattern から原因・結果表現を獲得するアルゴリズム [3] を用いて，原因・結果表現を抽出する．「その結果」や「そのため」といった手がかり表現を伴い，2 文にまたがっている原因・結果表現を抽出するために追加した Pattern E を含めた五つの Pattern を用いた原因・結果表現抽出手法は，我々の知る限り新規手法である．

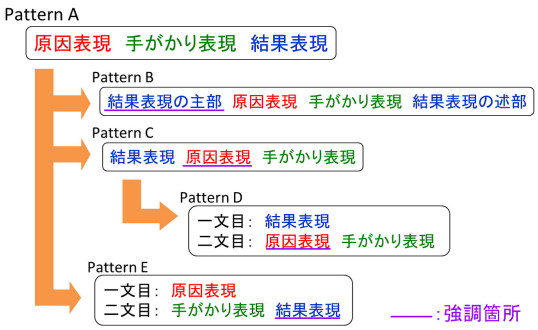

図 1　各 Pattern の関連図
Fig. 1　An association chart of patterns.

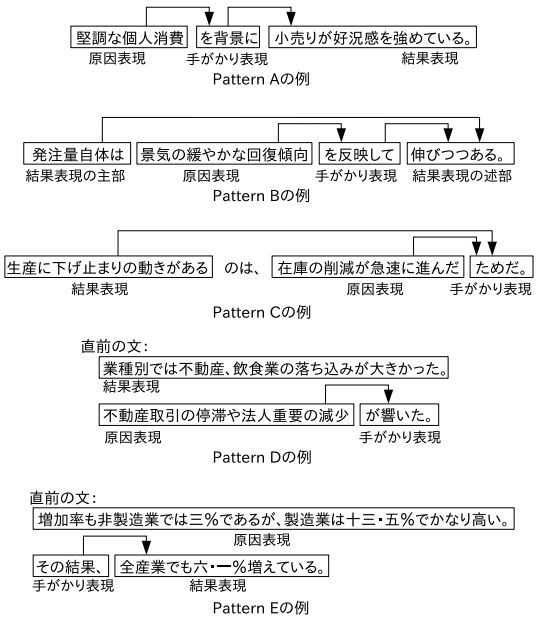

図 2　各 Pattern の例
Fig. 2　Examples of patterns.

図 1 において，出現頻度の最も高い Pattern A (原因表現，手がかり表現，結果表現の順に並ぶ) [3] が基本型であると考え，Pattern A と他の Pattern との関連性を矢印を用いて表している．Pattern B は，基本型から，強調のため結果の主部が文頭へ移動したものである．Pattern C は，結果を強調するため基本型を倒置したものである．Pattern D と E は一文にすると長くなるので，原因と結果を 2 文に分割したものである．Pattern A を分割したものが Pattern E であり，Pattern C を分割したものが Pattern D となっている．また，Pattern D と E では，それぞれ，手がかり表現を含む文が強調されるようになっている．

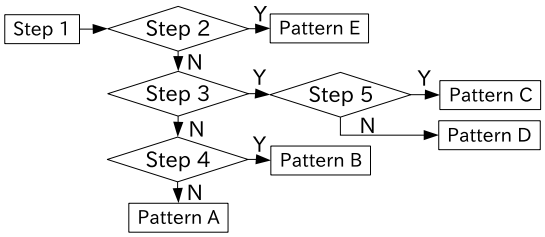

図 3　*Identification of patterns* の概要
Fig. 3　An outline of *Identification of patterns*.

4.1　適切な表現形式の識別

本節では，対象文が与えられたときに，上記に示した Pattern のうち，どの Pattern を適用するかを識別する手続き (*Identification of patterns*) について説明を行う．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．[3] にて提案した手続きに，Pattern E を追加した *Identification of patterns* の概要を図 3 に示す．

[*Identification of patterns*]
Step 1:　手がかり表現を含む文を取得する．
Step 2:　手がかり表現が文頭に出現する場合，Pattern E を適用した後，処理を終了する．そうでなければ，Step 3 を実行する．
Step 3:　手がかり表現に「。」が含まれている，若しくは，手がかり表現の後に「。」があるなら，Step 5 を実行する．そうでなければ，Step 4 を実行する．
Step 4:　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中に係り助詞，若しくは，格助詞を含むものがあれば，Pattern B を適用する．そうでなければ，Pattern A を適用し，処理を終了する．
Step 5:　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 C を適用する．そうでなければ，Pattern D を適用する．　□

例えば，対象文として「暖冬により暖房用燃料の販売が低調だった。」という文が与えられた場合，まず，Step 1 において手がかり表現「により」で，この文を取得することができる．次に，Step 2 で手がかり表現が文頭に存在しないため，Step 3 へ行く．Step 3 では，手がかり表現に句点が含まれていないので，Step 4 へ行く．最後に，この文の基点文節は，「低調だった。」という動詞句であるが，基点文節に係っている文節の中に係り助詞，若しくは，格助詞を含む文節が存在しないため，Pattern A が適用される．

4.2 手がかり表現の優先度

手がかり表現「を背景に、」と手がかり表現「を背景に」のように，部分一致している手がかり表現が幾つか存在する．そのため，*Identification of patterns* の Step 1 において，上記のような場合は左端から部分一致している表現のうち，最長文字数の手がかり表現を適用することとする．例えば，「円高を背景に、売り上げが減少している。」という文があった場合，Step 1 において適用する手がかり表現は「を背景に、」となる．

5. 原因・結果表現抽出手法の改良

4. で示した原因・結果表現抽出手法を決算短信 PDF に適用した場合，以下のような文において正しく原因・結果表現を抽出できなかった．

> 主な要因といたしましては、利益剰余金が四半期純損失と剰余金の配当により 2 億 5 千 8 百万円減少したことによります。

例えば，上記の決算短信 PDF に含まれる文に対して，手がかり表現「により」で原因・結果表現を抽出しようとした場合，原因表現として「利益剰余金が四半期純損失と剰余金の配当により 2 億 5 千 8 百万円減少した」，結果表現として「ます。」を抽出してしまう．そこで，既存の手がかり表現「により」に「ます。」を加えた新たな手がかり表現「によります。」等を抽出し，既存の手がかり表現に加えることで上記のような問題に対応する．また，上記文に含まれる「主な要因といたしましては」は，「利益剰余金が四半期純損失と剰余金の配当により 2 億 5 千 8 百万円減少した」が原因表現を示し，前文が結果表現であることを示す文字列である．このような文字列が決算短信 PDF に数多く散見されたため，この文字列を獲得し，原因・結果表現抽出に用いる新たな手法を開発する．本研究では，決算短信 PDF の文頭に特徴的に出現する文字列を **Prefix Pattern** と定義する．

5.1 新しい手がかり表現の獲得

本節では，「によります。」などの新しい手がかり表現を獲得する手法について述べる．Prefix Pattern は「によります。」などの新しい手がかり表現を伴って出現するため，新しい手がかり表現を獲得する必要がある．まず，「ます。」などの接尾辞を既存の手がかり表現の末尾に加えたものが，決算短信 PDF 中に存在するか否かを調べ，もし，存在すれば新たな手がかり表現として獲得する．新たな手がかり表現獲得に用いた接尾辞一覧を以下に示す．

> ます。 あります。 います。 おります。 です。

本手法を適用し，新しい手がかり表現を獲得した結果，以下に示す手がかり表現を獲得することができた．

> を受けております。 によります。
> によっています。
> によっております。 ためであります。

5.2 Prefix Pattern の獲得

本節では，Prefix Pattern を獲得する手法について述べる．「主な要因といたしましては」などの Prefix Pattern の末尾には係助詞「は」が存在することに着目する．図 4 に示す正規表現を作成し，これを用いて Prefix Pattern の候補を獲得する．ここで，「.」はワイルドカード，「*」は 0 回以上の繰り返し，「?」は最短一致を，それぞれ意味する．図 4 では，例として手がかり表現「によります。」を用いている．実際には，末尾に現れる手がかり表現全てを用いて上記のような正規表現を作成し，Prefix Pattern の候補を獲得する．

抽出した Prefix Pattern の候補の中には，不適切なものも存在する．そこで，「因」，「増加」，「減少」のいずれかの語を含むものを Prefix Pattern として獲得する．ここで獲得した Prefix Pattern を **6.** で用いるため，精度良く Prefix Pattern を獲得する必要がある．そのため，**7.** で後述する学習データ (決算短信 PDF のみ) において，調査したところ，上記の語を含むものの全てが Prefix Pattern であったため，これらの語を用いた．ただし，例外として「これは」は上記の語を含んでいないが，Prefix Pattern とした．上記の処理を Algorithm 1 に示す．

Algorithm 1 において，*Regular Expression* は図 4 に示した正規表現に一致した文字列を返す関数である．

図 4 Prefix Pattern 候補の獲得の例
Fig. 4 Examples of extracting candidate prefix patterns.

Algorithm 1 Extracting Prefix Pattern

Input: Sentence Set $S = (s_0, s_1, \ldots, s_n)$ and Clue Set $C = (c_0, c_1, \ldots, c_m)$
Output: Prefix Pattern Set P

```
P ← ∅
for each s ∈ S do
    for each c ∈ C do
        p ← RegularExpression(s, c)
        if p includes (これは or 因 or 増加 or 減少) then
            P ← P + p
        end if
    end for
end for
return P
```

表 2　Prefix Pattern の例
Table 2　An example of prefix patterns.

また主な減少要因としましては　要因は
これは　増加の理由は
この主な要因といたしましては
この要因は　売上高の減少をカバーしたのは
その主な要因といたしましては　主因は
増加要因は　経常利益の増加要因は

もし，一致しなければ，空文字を返す．

Prefix Pattern 獲得手法を適用した結果，285 個の Prefix Pattern を獲得することができた．獲得できた Prefix Pattern の例を表 2 に示す．

5.3　Prefix Pattern と新しい手がかり表現を用いた原因・結果表現抽出手法

Prefix Pattern を用いた原因・結果表現抽出手法について述べる．Prefix Pattern の末尾は係助詞であるため，Pattern を識別する手続き *Identification of patterns* の Step 5 における Pattern C の判別と重複してしまう．そこで，改良手法では Pattern C を適用しないようにする．具体的には，Pattern を識別する手続き *Identification of patterns* の Step 5 を以下のように変更する．

Step 5:　末尾に出現する手がかり表現「によります。」などが文に含まれていた場合，文頭が Prefix Pattern であれば，Pattern D を適用する．

これにより，決算短信 PDF に特徴的に数多く現れる Prefix Pattern を伴った原因・結果表現を抽出できるようになる．ただし，Pattern C での原因・結果表現を抽出できなくなる．**7.** で後述する学習データを調べたところ，Pattern C に当てはまる原因・結果表現は存在せず，Prefix Pattern に当てはまる原因・結果表現は 28 個存在した．この結果から，Pattern C に当てはまる原因・結果表現を抽出できなくなったとしても，Prefix Pattern による原因・結果表現の方が数多く獲得できると考えられる．また，本章で説明した Prefix Pattern の獲得と，それを用いた原因・結果表現抽出に関しては，決算短信 PDF に特徴的に出現する原因・結果表現を抽出するための手法となっており，我々が初めて提案する新規手法となっている．

Algorithm 2 Bootstrapping

Input: Clue Set $C = (c_0, c_1, \ldots, c_n)$ and Prefix Pattern Set $P = (p_0, p_1, \ldots, p_m)$ and Iteration Number T
Output: Clue and Prefix Pattern Set $\left(\acute{C}, \acute{P}\right)$

```
Ć ← C
Ṕ ← P
for t ← 1...T do
    Ć ← Ć + getClues(Ṕ)
    Ṕ ← Ṕ + getPatterns(Ć)
end for
return (Ć, Ṕ)
```

6.　ブートストラップ手法による手がかり表現と Prefix Pattern の追加獲得

5. の手法を用いて，手がかり表現の追加獲得と Prefix Pattern を獲得したが，十分な量の手がかり表現と Prefix Pattern を獲得できなかった．そこで，ブートストラップ手法を用いて，手がかり表現と Prefix Pattern の追加獲得を行う．ブートスラップ手法の概要を Algorithm 2 に示す．

Algorithm 2 において，**5.** で獲得した手がかり表現を既存の手がかり表現に加えたものを初期に与えられる手がかり表現とし，獲得した Prefix Pattern を Prefix Pattern 集合とする．また，T は反復回数を表す．*getClues* と *getPatterns* については，それぞれ，**6.1** と **6.2** で後述する処理である．

6.1　手がかり表現の獲得

本節では，ブートストラップ手法の *getClues* に当たる Prefix Pattern を用いての手がかり表現の獲得について述べる．そのため，まず，図 5 に示す正規表現を作成し，手がかり表現候補を獲得する．Prefix Pattern 集合に存在する Prefix Pattern，それぞれに対して正規表現を作成し，決算短信 PDF から手がかり表現候補を獲得する．図 5 で示すように，格助詞から文末までを手がかり表現候補として獲得する．ただし，**7.** で後述する手がかり表現から句読点を除いたものを含む場合に，手がかり表現候補として獲得する．

手がかり表現候補の中には手がかり表現として適切

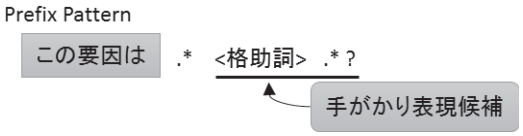

図 5　手がかり表現候補の獲得の例
Fig. 5　An example of extracting candidate clues.

でないものも存在するため，手がかり表現の選別を行う．ここで，様々な Prefix Pattern と共起する手がかり表現候補は適切であるという仮定に基づき，適切な手がかり表現を獲得するためのスコアにエントロピーを用いる．2 回以上出現した手がかり表現候補に対して，以下の式 (1) を用いてスコアを計算する．なお，スコアは 0 から 1 の値を取るように正規化している．

$$Score(c) = \frac{H(c)}{max_c H(c)} \tag{1}$$

$$H(c) = -\sum_{e \in E(c)} P(e,c) \log_2 P(e,c) \tag{2}$$

$$P(e,c) = \frac{f(e,c)}{\sum_{e' \in E(c)} f(e',c)} \tag{3}$$

ただし，
$E(c)$: 手がかり表現候補 c と共起する Prefix Pattern の集合
$max_c H(c)$:　全てのエントロピー $H(c)$ の中で最大のもの
$P(e,c)$:　Prefix Pattern e と手がかり表現候補 c が共起する確率
$f(e,c)$:　Prefix Pattern e と手がかり表現候補 c の共起数

$Score(c)$ がしきい値 α 以上の手がかり表現候補を手がかり表現として獲得する．

6.2　Prefix Pattern の獲得

本節では，ブートストラップ手法の *getPatterns* に当たる手がかり表現を用いての Prefix Pattern の獲得について述べる．**5.2** で示した図 4 の正規表現を用いて Prefix Pattern 候補を獲得する．**5.2** では，特定の語が含まれている Prefix Pattern 候補を Prefix Pattern として獲得していたが，ここでは様々な手がかり表現と共起する Prefix Pattern 候補は適切であるという仮定に基づき，エントロピーを用いて選別を行う．2 回以上出現した Prefix Pattern 候補に対して，以下の式 (4) を用いてスコアを計算する．なお，スコアは 0 から 1 の値を取るように正規化している．

$$Score(e) = \frac{H(e)}{max_e H(e)} \tag{4}$$

$$H(e) = -\sum_{c \in C(e)} P(c,e) \log_2 P(c,e) \tag{5}$$

$$P(c,e) = \frac{f(c,e)}{\sum_{c' \in C(e)} f(c',e)} \tag{6}$$

ただし，
$C(e)$: Prefix Pattern 候補 e と共起する手がかり表現の集合
$max_e H(e)$:　全てのエントロピー $H(e)$ の中で最大のもの

$Score(e)$ がしきい値 α 以上の Prefix Pattern 表現候補を Prefix Pattern として獲得する．

7. 評価実験

評価データには，決算短信 PDF からランダムに 30 ファイルを用いた．評価者は，投資歴 15 年の個人投資家に依頼し，人手でタグを付与した．評価者が 30 ファイルに対して，原因・結果を含む文を抽出し，抽出された文に対してタグ（「原因」,「結果」）を付与(注2)したところ，478 個の原因・結果表現対が存在した．決算短信 PDF から原因・結果を含む文を抽出するための学習データとして，経済新聞記事において手がかり表現を含む文 2,064 と，決算短信 PDF において手がかり表現を含む文 1,296 を用いた．ただし，学習データにおける決算短信 PDF 1,296 文は，評価データ以外のデータを用いた．学習データは，原因・結果表現を含む 1,454 文と，原因・結果表現を含まない 1,906 文から構成される．形態素解析器としては Mecab(注3) を用い，係り受け解析器としては Cabocha [4] を用いた．学習器には SVM$^{\text{Light}}$(注4) を用い，カーネルは線形を用いた．初期手がかり表現には，表 3 に示す 34 個を用いた．

表 4 に我々の既提案手法 [3] と，新たな Pattern E を追加した本手法 1，**5.** で述べた改良手法（本手法 2），本手法 2 に **6.** で述べた手法を追加した本手法 3 の評価結果を示す．本手法 3 は，しきい値 α が 0.7，反復回数が 3 のときの結果を用いている．これは，学習データ (決算短信 PDF のみ) において，本手法 3 を適用した場合，上記パラメータの場合が最も良い結果

(注2)：手がかり表現を伴わない原因・結果表現にもタグを付与した．
(注3)：http://code.google.com/p/mecab/
(注4)：http://svmlight.joachims.org/

表 3　初期手がかり表現
Table 3　Initial clues.

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 4　評 価 結 果
Table 4　Evaluation results.

	精度	再現率	F 値	抽出数	手がかり表現数
既提案手法	0.81	0.56	0.66	330	30
本手法 1	0.82	0.62	0.71	360	34
本手法 2	0.84	0.65	0.74	371	39
本手法 3	0.84	0.68	0.75	389	77

表 5　反復回数 3 回の場合の評価結果
Table 5　Evaluation results of loop 3.

しきい値 α	精度	再現率	F 値	手がかり表現数	Prefix Pattern 数
0.3	0.80	0.66	0.72	1152	1381
0.4	0.81	0.67	0.73	250	1174
0.5	0.82	0.68	0.74	127	779
0.6	0.84	0.68	0.75	91	677
0.7	0.84	0.68	0.75	77	337
0.8	0.84	0.68	0.75	72	295

表 6　しきい値 α が 0.7 の場合の評価結果
Table 6　Evaluation results of α 0.7.

反復回数	精度	再現率	F 値	手がかり表現数	Prefix Pattern 数
1	0.84	0.66	0.74	52	288
2	0.84	0.68	0.75	64	296
3	0.84	0.68	0.75	77	347
4	0.84	0.68	0.75	86	417
5	0.83	0.68	0.75	130	427

であったためである．精度は，手法で抽出した原因・結果表現対のうち，正解だった割合を示す．再現率は，評価データに含まれる 478 個の原因・結果表現対をどのくらい網羅できたかを示す．F 値は，精度と再現率の調和平均である．この評価では，抽出した表現が評価者によってタグ付けられた表現を含む場合に正解とした．逆に，タグ付けられた表現の一部のみを抽出していた場合は，不正解とした．

表 5 に反復回数 3 回に固定し，しきい値を変化させた場合の精度，再現率，F 値を示す．更に，表 6 にしきい値 α を 0.7 に固定し，反復回数を変化させた場合の精度，再現率，F 値を示す．表 7 に，本手法により抽出した原因・結果表現の例を幾つか示す．

7.1　汎用性の検証

提案した本手法 1 が他の領域のテキストに適用可能か否かを検証するために実験を行った．本手法 2 と本手法 3 に関しては，決算短信 PDF に特徴的に出現する Prefix Pattern を用い，かつ，手がかり表現を「ですます調」に対応させた手法であるため，汎用性がないと判断し，実験は既提案手法と本手法 1 を用いて行う．決算短信 PDF と似ている文書として業績発表記事を汎用性の検証のために用いる．業績発表記事とは，経済新聞記事において企業の業績発表に関する記述が存在する記事 [5] である．**7.** の評価者に依頼し，業績発表記事集合からランダムに取得した業績発表記事 30 記事に対して，タグを付与した．その結果，51 の原因・結果表現対が存在した．業績発表記事は，決算短信 PDF と比べ短い文書であるため，同じ 30 記事でも含まれている原因・結果表現の数が少なくなった．業績発表記事での評価結果を表 8 に示す．

7.2　Pantel らの手法との比較

6. で説明したブートストラップ手法と，Pantel et al. が提案している手法との比較実験を行った．Pantel et al. の提案している Espresso は，パターン p を用いて二つの名詞対 $i = x, y$ を獲得する手法である [6]．我々は，Espresso を本タスクに適応させるため，Prefix Pattern と手がかり表現をそれぞれ Espresso におけるインスタンスとパターンとした．Prefix Pattern 候補の信頼度を式 (7) で計算し，手がかり表現候補の信頼度を式 (8) で計算する．

$$r_l(i) = \frac{\sum_{p \in P} \frac{pmi(i,p)}{max_{pmi}} r_\pi(p)}{|P|} \tag{7}$$

$$r_\pi(p) = \frac{\sum_{i \in I} \frac{pmi(i,p)}{max_{pmi}} r_l(i)}{|I|} \tag{8}$$

ここで，P は手がかり表現候補の集合，I は Prefix Pattern 候補の集合である．また，次の式 (9) で $pmi(i,p)$ を計算する．

$$pmi(i,p) = \log \frac{P(i,p)}{P(i)P(p)} \times \frac{C_{ip}}{C_{ip}+1} \times \frac{min(\sum_{k=1}^{|P|} C_{ik},\ \sum_{j=1}^{|I|} C_{jp})}{min(\sum_{k=1}^{|P|} C_{ik},\ \sum_{j=1}^{|I|} C_{jp}) + 1} \tag{9}$$

ここで，C_{ip} は，p と i が共起する頻度である．そして，まず，本手法と同様の初期手がかり表現を与え信頼度 r_l を求め，その上位 N 個を Prefix Pattern として獲得した．次に，信頼度 r_π を求め，その上位 N 個

表 7　抽出した原因・結果表現の例
Table 7　Examples of extracted causal expressions.

企業名	原因表現	結果表現
日本製紙グループ	実需低迷および夏場の天候不順	板紙は、段ボール原紙などの国内販売数量が前期を大幅に下回りました。
神栄	東日本大震災やタイにおける大洪水の影響による自動車メーカーの生産調整	車載空調センサは、取扱いは減少となりました。
パウダーテック	暖冬予想によるカイロメーカーの減産や製品価格の低下	鉄粉製品は、売上が大幅に減少しました。
TOA	昨今の防犯への関心の高まりとニーズの拡大	当社のセキュリティ関連商品の売上も好調に推移しました。

表 8　業績発表記事における評価結果
Table 8　Evaluation results of financial articles on business performance of companies.

	精度	再現率	F 値	抽出数	手がかり表現数
既提案手法	0.95	0.69	0.80	37	30
本手法 1	0.93	0.75	0.83	41	34

表 9　拡張 Espresso との比較実験結果
Table 9　Evaluation results of our method and Espresso.

	精度	再現率	F 値	手がかり表現数	Prefix Pattern 数
本手法 2	0.84	0.65	0.74	39	285
本手法 3	0.84	0.68	0.75	77	347
拡張 Espresso	0.82	0.69	0.75	311	585

を手がかり表現として獲得した．これを定められた回数，繰り返し，Prefix Pattern と手がかり表現を獲得した．Espresso では，しきい値を設けて獲得していないため，信頼度の高い順の上位 N 個を獲得する．実験では，N の値をそれぞれ 100 とした．原因・結果表現を含む文か否かを判定する手法の学習データ (決算短信 PDF のみ) において，拡張 Espresso を実行したところ，反復回数 3 回が最も良い F 値となったため，これを採用した．表 9 に，実験結果を示す．本手法 3 の α の値は 0.7，反復回数は 3 回であり，拡張 Espresso の N の値は 100，反復回数は 3 回となっている．

8. 考　　察

表 4 より，本手法 1 の精度と再現率が既提案手法を上回った．これは，Pattern E による原因・結果表現の網羅率の向上に加えて，Pattern E の抽出精度が高かったためである．

また，本手法 2 と本手法 3 が精度，再現率，F 値の全てにおいて既提案手法と本手法 1 を上回った．これは，決算短信 PDF に特徴的に出現する Prefix Pattern を伴う原因・結果表現対を抽出できるようになったことに起因する．例えば，既提案手法では抽出できなかった以下の例を本手法では抽出できていた．

> <r>（投資活動によるキャッシュ・フロー）投資活動の結果、４６４百万円のキャッシュ・フローの減少（前期比３４．７％減）となりました。</r>これは主に、店舗の新規出店による有形固定資産取得のために３９３百万円の支出と保証金差入９５百万円を行ったためであります。

ここで，タグで囲まれた部分は原因表現を示し，<r>タグで囲まれた部分は結果表現を示す．上記例では，新しい手がかり表現「ためであります。」が存在したため，抽出することができた．

更に，本手法 3 は本手法 2 より再現率が向上し，結果的に F 値が向上している．これは，手がかり表現と Prefix Pattern をブートストラップ手法により追加獲得できた結果による．追加で獲得できた手がかり表現と Prefix Pattern を表 10 に示す．

通常，ブートストラップ法(注5)を用いると，しきい値 α が高ければ精度が高くなり，再現率が低くなる．逆に，しきい値 α が低くなれば精度が低くなり，再現率が高くなる．それに対して，表 5 より，しきい値 α が高くなるにつれて，精度，再現率共に向上し，しきい値 α が低くなるにつれて，精度，再現率共に低下している．上記現象の理由は，以下のとおりである．

本手法 3 はブートストラップ法であるため，しきい値を下げたり，反復回数を多くすると，より多くの Prefix Pattern や手がかり表現を獲得する．しきい値が高いうちは，適切な手がかり表現や Prefix Pattern を獲得するが，しきい値がある一定値を超えると，不適切な手がかり表現や Prefix Pattern を獲得してし

(注5)：ブートストラップ法とは，言語処理におけるブートストラップ手法全般を指している．それに対し，ブートストラップ手法とは，本手法 3 のブートスラップ手法を指す．

表 10　ブートストラップ手法により追加獲得した手がかり表現と Prefix Pattern の例
Table 10　Examples of additional clues and prefix patterns.

手がかり表現	によるものであります。　による収入であります。　によるものです。 により減少したものです。　による収入の増加であります。　により増加いたしました。
Prefix Pattern	その主な原因は　その主な内容は　これらは　主なものは　主な増減は 増加した主な要因は　流動資産の増加の主な要因は　この主因は

まう．上記のような状態になると，正しい手がかり表現「により」を獲得していたとしても，より長い「により堅調に推移いたしました。」という不適切な手がかり表現を獲得してしまう．その結果，**4.2** より，原因・結果表現抽出において，最も文字数が多い手がかり表現を採用するため，適切でなく，かつ，文字数が多い手がかり表現が存在した場合，これを採用してしまう．つまり，適切な手がかり表現を獲得していたとしても，それより長い不適切な手がかり表現を獲得してしまうと，適切な原因・結果表現を抽出できない．このことより，多くの手がかり表現を獲得すればするほど，精度だけでなく，元々正しく抽出できていた原因・結果表現も抽出できなくなり，再現率も下がってしまう．これが，しきい値が低くなれば精度が低くなり，再現率も下がってしまった理由である．

また，本手法 3 は，本手法 2 で獲得した Prefix Pattern と手がかり表現を種 (初期 Prefix Pattern 集合と初期手がかり表現集合) として利用している．通常のブートストラップ法では，我々の実験よりも少数の種から始め，しきい値を下げて数多くの手がかり表現を獲得し再現率が向上していく．それに対して本手法 3 では，既に多くの Prefix Pattern と手がかり表現を網羅している状態からスタートしているため，F 値のピークがしきい値が高いときになってしまったと考える．これが，しきい値が高くなるにつれて，精度，再現率共に向上している理由である．

表 8 より，決算短信 PDF に対して適用した結果に比べ，業績発表記事に対して適用した結果の方が高い F 値を示した．これは，業績発表記事は新聞記事であり，既提案手法は経済新聞記事を適用対象として開発した手法であることから，高い性能を示したと考える．また，決算短信 PDF における本手法 1 と，業績発表記事における本手法 1 の両方とも，F 値 0.7 以上を達成することができた．この結果より，本手法 1 に汎用性がある可能性を示すことができた．しかしながら，十分な数の様々な領域のテキストにおいて実験したわけではないため，今後の課題として，更なる汎用性検証のため，他の様々な領域のテキストでの実験を行うことを検討している．

また，業績発表記事における本手法 1 において，以下のような例は正しく抽出できていなかった．

> <r>ダイオーズが三十日に発表した二〇〇二年三月期連結決算は、最終損益が七億三千万円の黒字（前の期は十三億円の赤字）と、過去最高益となった。</r> 米国部門の売り上げが八％伸びたのに加え、清掃用品レンタル事業で利益率が改善したことが寄与。

上記の例では，本手法 1 で用いた手がかり表現に「が寄与。」がなかったため，原因・結果表現を抽出できなかった．今後の課題として，「が寄与。」のような手がかり表現も自動的に獲得する手法が必要となってくる．

表 9 より，F 値に関しては本手法 3，拡張 Espresso 共に 0.75 と同じ値になった．ただし，精度に関しては本手法 3 の方が高く，再現率に関しては拡張 Espresso の方が高くなった．結果に大きな差がでなかったのは，我々が提案している Prefix Pattern や手がかり表現の定義を拡張 Espresso が用いているためであると考えている．また，本論文で提案しているブートストラップ手法を用いても，拡張 Espresso を用いても性能が向上したことから，本タスクにおけるブートストラップ法の有効性が確認された．しかしながら，本手法 3 のブートストラップ手法では α の値を決めるだけでよいが，Espresso ではインスタンスとパターンのそれぞれにおいて，上位何個を獲得するかを決めるパラメータが必要となる．このことから，本手法 3 のブートストラップ手法の方が，Espresso に比べて，パラメータ数が少ないという利点がある．

8.1　エラー解析

各手法の False Positive と False Negative を表 11 に示す．ただし，本手法 3 はしきい値 α が 0.7，反復回数が 3 のときである．表 11 より，既提案手法，本手法 1，本手法 2，本手法 3 の順番で False Negative が少なくなっていることが分かる．これは，**5.** の手法や，**6.** の手法の改善による結果であると考えられる．

表 11　各手法の False Positive と False Negative の数
Table 11　Number of false positives and Number of false negatives.

	False Positive	False Negative	合計
既提案手法	64	212	276
本手法 1	64	182	246
本手法 2	58	165	223
本手法 3	62	151	213

8.1.1　False Positive

以下に，False Positive の例を示す．

> 今後の景況につきましては、輸出の下支えにより<r>製造業を中心とした緩やかな回復が予測される</r>

上記の文片から原因表現として「輸出の下支え」，結果表現として「製造業を中心とした」を抽出した．文片に正解のタグが付与されているように，正しい結果表現は「製造業を中心とした緩やかな回復が予測される」である．これは，係り受け解析の誤りによって，誤った結果表現を抽出してしまっていた．この問題を解決するためには，五つの Pattern を改良し，係り受け解析の誤りにも対応できるようにする必要がある．

8.1.2　False Negative

以下に，False Negative の例を示す．

> 百貨店業界におきましては、同業・他業態との競争激化に加え、消費マインドの冷え込みから、<r>売上高が 3 年連続で前年割れとなっております。</r>

上記の文には，原因表現として「同業・他業態との競争激化に加え、消費マインドの冷え込み」，結果表現として「売上高が 3 年連続で前年割れとなっております。」が含まれているが，本手法では抽出されていなかった．これは，判定手法により，この文に原因・結果が含まれていないと判定されてしまったためである．この問題を解決するためには，判定手法の再現率を向上させる必要がある．

9.　原因・結果表現検索システム

本章では，提案手法を用いて抽出した原因・結果表現を検索するシステムについて説明する．我々は，ユーザの投資判断支援を行うため，抽出した原因・結果表現を検索できる試作的なシステムを構築した．まず，

図 6　検索結果画面の例
Fig. 6　A causal expressions search system.

3,821 企業の Web ページから決算短信 PDF の収集を試みた結果，106,885 個の決算短信 PDF を収集することができた．収集した決算短信 PDF に対して本手法 3 (しきい値 α が 0.7，反復回数が 3) を適用し，原因・結果表現を抽出した結果，1,218,526 個の原因・結果表現を抽出した．抽出した原因・結果表現を用いて LAMP (Linux, Apache, MySQL, PHP) 環境においてシステムを構築した(注6)．本システムでは，各企業に紐づいた原因・結果表現を検索することができる．図 6 に，検索結果の例を示す．

図 6 では，原因表現に「天候不順」を含む原因・結果表現対を検索している．図 6 より，日本製紙グループの 2010 年 4 月に発表された決算短信 PDF に含まれる原因表現「実需低迷および夏場の天候不順」，結果表現「板紙は、段ボール原紙などの国内販売数量が前期を大幅に下回りました。」を検索することができた．それにより，個人投資家は，「天候不順」となった場合に，日本製紙グループの売り上げが下がる可能性があることを知ることができる．

10.　関 連 研 究

Khoo らは人手で作成したパターンを用いて，新聞記事や医療データベースから原因・結果を抽出する手法を提案している [7], [8] が，結果表現と原因表現が同じ文に含まれている必要がある．これらの研究は，初期の研究として非常に重要であるが，原因・結果を抽出する対象が限定されているため，抽出結果も限定的

(注6)：http://hawk.ci.seikei.ac.jp/CS/

となる．それに対して，本手法では重文，複文や文をまたがった対象からも原因・結果を抽出することができるため，数多くの抽出結果が得られることが期待できる．

Chang らは手がかり表現と語の組の出現確率を用いて，二つの名詞句間の原因・結果を抽出する手法を提案している [9]．また，Girju は手がかり表現に基づいて自動的に WordNet に含まれる名詞句間の原因・結果の検出と抽出を行う手法を提案している [10]．彼らの研究は名詞句の組を原因・結果の対象としているため，他の表現間の原因・結果を抽出することができないが，本手法では名詞句だけでなく動詞句や文をも対象としている．

大森らは不具合事例文書から製品・部品に関する因果関係抽出手法を提案している [11]．不具合事例文においてモノ（物体）を示す実体語を素性に用いて機械学習を行い，因果関係を表す文を抽出している．Ishii らは新聞記事から因果関係を抽出し，因果関係ネットワークを構築する手法を提案している [12]．それに対して，本手法では不具合事例文書や新聞記事ではなく，決算短信 PDF に特化した手法となっている点が異なっている．

酒井らは，企業 Web ページから取得したキーワードを使用して，企業の業績発表記事から重要な業績要因を抽出し，それを検索するシステムを作成した [5]．業績要因は本研究の原因表現に相当する．藏本らは，新聞記事に含まれる単語を抽出し，共起解析，主成分分析，回帰分析からなる CPR 法を用いて長期市場動向の分析を行っている [13]．これらの研究に対して，本研究では，業績要因や単語だけではなく，原因・結果表現対を抽出している．

11. むすび

本研究では，決算短信 PDF から原因・結果表現を自動的に抽出する手法の提案を行った．本手法は大まかに分けて二つのステップから構成され，一つ目のステップでは，決算短信 PDF から原因・結果を含む文を抽出した．二つ目のステップでは，抽出した原因・結果を含む文から原因・結果表現を抽出した．原因・結果表現抽出には，構文情報を用いた Pattern を用いた．更に，決算短信 PDF に特徴的に出現する Prefix Pattern と手がかり表現をブートストラップ手法を用いて自動的に獲得し，これを用いて原因・結果表現を抽出する手法を新たに開発した．決算短信 PDF に合わせて手法を改良することで，精度 0.85，再現率 0.68，F 値 0.75 を達成した．

今後の課題として，Pattern C の決算短信 PDF への適用が考えられる．そのためには，Prefix Pattern を適用するか，それとも，Pattern C を適用するかを適切に判断するアルゴリズムが必要となる．

また，手がかり表現の優先度を文字数の多さではなく，他の尺度を用いることが今後の課題に挙げられる．手がかり表現の文字数だけではなく，手がかり表現の出現頻度などを用いて優先度を求めることができると考えられる．

謝辞　ニューヨーク大学関根聡研究准教授には，本研究に関して有益なご助言を頂いた．本研究の一部は日本学術振興会科研費 (C)26330359 の援助で行われた．

文　献

[1] 坂地泰紀，増山　繁，“新聞記事からの因果関係を含む文の抽出手法，” 信学論（D），vol.J94-D, no.8, pp.1496–1506, Aug. 2011.

[2] 庵　功雄，新しい日本語学入門（第 2 版），スリーエーネットワーク，2012.

[3] H. Sakaji, S. Sekine, and S. Masuyama, “Extracting causal knowledge using clue phrases and syntactic patterns,” 7th International Conference on Practical Aspects of Knowledge Management (PAKM), pp.111–122, 2008.

[4] 工藤　拓，松本裕治，“チャンキングの段階適用による日本語係り受け解析，” 情処学論，vol.43, no.6, pp.1834–1842, 2002.

[5] 酒井浩之，増山　繁，“企業の業績発表記事からの重要業績要因の抽出，” 信学論（D），vol.J96-D, no.11, pp.2866–2870, Nov. 2013.

[6] P. Pantel and M. Pennacchiotti, “Espresso: Leveraging generic patterns for automatically harvesting semantic relations,” Proc. 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL, pp.113–120, 2006.

[7] C.S. Khoo, J. Kornfilt, R.N. Oddy, and S.H. Myaeng, “Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing,” Literary and Linguistic Computing, vol.13, no.4, pp.177–186, 1998.

[8] C.S. Khoo, S. Chan, and Y. Niu, “Extracting causal knowledge from a medical database using graphical patterns,” Proc. 38th ACL, pp.336–343, 2000.

[9] D.-S. Chang and K.-S. Choi, “Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities,” Information Processing and Management, vol.42, no.3, pp.662–678, 2006.

[10] R. Girju, “Automatic detection of causal relations for

question answering," ACL Workshop on Multilingual Summarization and Question Answering, pp.76–83, 2003.
[11] 大森信行，森　辰則，"不具合事例文書からの製品・部品に関する因果関係抽出手法の検討，" 言語処理学会第 18 回年次大会，pp.1192–1195, 2012.
[12] H. Ishii, Q. Ma, and M. Yoshikawa, "Incremental construction of causal network from news articles," J. Information Processing, vol.20, no.1, pp.207–215, 2012.
[13] 藏本貴久，和泉　潔，吉村　忍，石田智也，中嶋啓浩，松井藤五郎，吉田　稔，中川裕志，"新聞記事のテキストマイニングによる長期市場動向の分析，" 人工知能学会論文誌，vol.28, no.3, pp.291–296, 2013.

（平成 26 年 9 月 4 日受付，12 月 29 日再受付，
27 年 2 月 4 日早期公開）

坂地　泰紀　（正員）

2009 年豊橋技術科学大学大学院工学研究科修士課程知識情報工学専攻修了．2012 年同大学院工学研究科博士後期課程電子・情報工学専攻修了．博士（工学）．2012 年株式会社ドワンゴ入社．2013 年成蹊大学理工学部情報科学科助教．自然言語処理，特に，テキストマイニングの研究に従事．人工知能学会，言語処理学会等各会員．

酒井　浩之　（正員）

2002 年豊橋技術科学大学大学院工学研究科修士課程知識情報工学専攻修了．2005 年同大学院工学研究科博士後期課程電子･情報工学専攻修了．博士（工学）．2005 年豊橋技術科学大学知識情報工学系助手．2007 年豊橋技術科学大学知識情報工学系助教．2010 年豊橋技術科学大学情報・知能工学系助教．2012 年成蹊大学工学部情報科学科講師．2014 年成蹊大学理工学部情報科学科准教授．自然言語処理，特に，テキストマイニング，テキスト自動要約の研究に従事．人工知能学会，言語処理学会，情報処理学会等各会員．

増山　　繁　（正員）

1977 年京都大学工学部数理工学科卒業．1982 年同大学大学院博士後期課程単位取得退学．1983 年同修了 (工学博士)．1982 年日本学術振興会奨励研究員．1984 年京都大学工学部数理工学科助手．1989 年豊橋技術科学大学知識情報工学系講師．1990 年同助教授．1997 年同教授．2010 年同大学院工学研究科情報・知能工学専攻教授．2005 年より 2012 年まで，同大学インテリジェントセンシングシステムリサーチセンター教授兼務．2010 年同大学人間・ロボット共生リサーチセンター教授兼務．アルゴリズム工学，特に，並列アルゴリズム等，及び，自然言語処理，特に，テキスト自動要約等の研究に従事．言語処理学会，情報処理学会等各会員．