# 企業業績要因文における因果関係の有無判定手法の提案

坂本 大祐 (豊橋技術科学大学　知識情報工学系　sakamoto@smlab.tutkie.tut.ac.jp)
坂地 泰紀 (豊橋技術科学大学　知識情報工学系　sakaji@smlab.tutkie.tut.ac.jp)
酒井 浩之 (豊橋技術科学大学　知識情報工学系　sakai@smlab.tutkie.tut.ac.jp)
増山 繁 (豊橋技術科学大学　知識情報工学系　masuyama@tutkie.tut.ac.jp)

## 1　はじめに

現在，ウェブページや新聞記事を含む大規模な機械可読文書が入手可能となっている．多くの機械可読な文書中には，実アプリケーションに役立つ様々な情報があり，テキストマイニング技術を用いることで，それらを獲得することが可能である．そのような情報の一つとして因果関係がある．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にとって重要な情報源となる．しかし，大規模文書からそのような知識を人手で獲得するには高いコストと時間を要する．そのため，文書から因果関係に関する知識を自動的に抽出することを試みた研究がいくつか存在する [1][2][5].

因果関係に関する知識を抽出する研究のほとんどは，因果関係であると推定される文字列の組を自動的に抽出するだけで，それが本当に因果関係を持っているか否かを判定していない．そのため，抽出された原因と結果の組の間に因果関係が存在するか否かは，結局，人間が判断しなければならない．そこで，本研究では特定の手がかり表現を含む文に含まれる原因表現と結果表現の間に因果関係があるか否かを自動的に判定する手法を提案する[*1]．本研究では，因果関係有無の判定手法を開発するにあたり，手がかり表現の直前の形態素と，文に含まれる助詞に着目した．因果関係有無の判定手法により，既存研究の精度向上が見込まれ，より実用性の高い情報の獲得が期待できる．

## 2　関連研究

Sakaji ら [5] は手がかり表現と構文情報を使用し，景気動向に関する記述を含む記事から結果と原因の組を抽出する手法を提案している．文献 [5] では，特定の表現やパターンに合致する文から原因・結果表現の抽出を行うが，抽出された組に因果関係があるかどうかは判定していない．

乾ら [2] は接続標識「ため」を含む複文から因果関係についての知識を獲得し，それらを 4 種類の因果関係（cause, effect, precond, means）に自動分類する手法を提案している．本研究では，接続標識「ため」に比べて広い多義性を持つ手がかり表現 “で、” を含む文を，因果関係の判定対象とする．

Girju [1] は手がかり表現に基づき，英文から因果関係を表現する名詞句の対を抽出する手法を提案している．この手法は，名詞句の対を選択する際に，英語の概念辞書である WordNet[*2]を利用している．しかし，本研究では Sakaji ら [5] の手法に基づき原因表現と結果表現を抽出するため，名詞句だけでなく動詞句も原因・結果表現に含むことができる．

## 3　判定対象とする文

### 3.1　業績要因文について

本研究では，Sakai ら [4] の手法を用いて日経新聞（1990 年から 2005 年までの 16 年間）記事から取得された，企業の業績発表に関する記事における業績要因を含む文（以下，業績要因文）を判定対象とする．対象をこのように限定したのは，経済新聞記事の業績要因文が因果関係に関する知識を多く含むことと，対象の限定により文が因果関係を持つ場合と持たない場合とを見分けやすくするためである．

### 3.2　業績要因文を構成する表現について

本研究では Sakaji ら [5] の手法を参考とし，業績要因文が以下の表現から構成されるものと考える．

- 原因表現（場合により主部・述部に分かれる）
- 結果表現
- 手がかり表現

手がかり表現とは，文書から因果関係を抽出する際の手がかりとなる接続標識を指す．日本語における手がかり表現の例として，“ため”，“による” などがある．本研究では業績要因文のうち，手がかり表現である “で、” を含むものを判定対象とする．これは，手がかり表現 “で、” が多様な意味を持つため，その他の手がかり表現を含む文に比べて因果関係を含む割合が低いためである．手がかり表現 “で、” が文中においてどのような意味を表す場合にその文が因果関係を含まないかについては，4.2 節で後述する．

Sakaji らは，手がかり表現を含む文を，文の構成によって以下の 4 パターンに分類している [5].

**パターン A**　結果表現が原因表現の後に出現し，手がかり表現は原因表現と結果表現を連結する．

**パターン B**　結果表現の主部と述部が一つの文中に分かれて存在し，手がかり表現は結果表現の主部と述部を連結する．

**パターン C**　結果表現が手がかり表現を含む文の直前の文であり，手がかり表現は原因表現となる部分を示す．

**パターン D**　結果表現が原因表現の前に出現し，手がかり表現は原因表現，および，結果表現に係られる．

[*1] Sakaji らの手法を用いることで，原因表現と結果表現を求めることができる．

[*2] http://wordnet.princeton.edu/

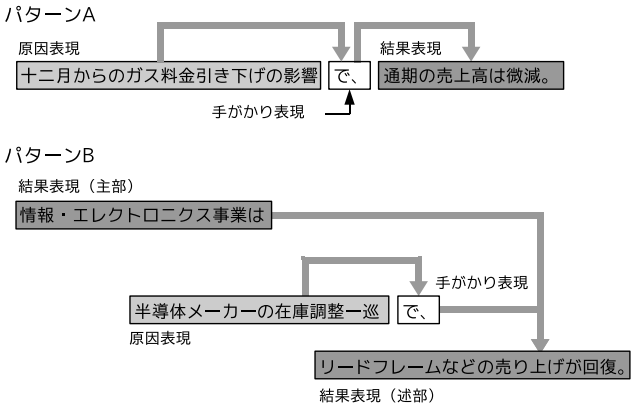


図 1　パターン A，B に該当する文の例

上記パターンのうち，手がかり表現 “で、” を含む文は，構文上，パターン D には該当しない．また，パターン C は連続する二つの文から原因表現と結果表現を選定するものであるが，Sakai らの手法により抽出される業績要因文はそれぞれ一文ずつであるため，除外する．従って，本研究ではパターン A，または，B に該当するものを判定対象とする．パターン A，B に該当する文の例を図 1 に示す．

## 4　判定対象とする文の調査

### 4.1　業績要因文のラベル付け

調査のために，手がかり表現 “で、” を含む文について，原因表現が結果表現の原因になっていると考えられるものに「因果関係あり」，それ以外のものに「因果関係なし」のラベルをそれぞれ割り当てた．「因果関係あり」とラベル付けされた文の例を図 2 ，「因果関係なし」とラベル付けされた文の例を図 3 にそれぞれ示す．

1. 菓子パンもあんパンなどが堅調で、販売増となったようだ。
2. 個人消費の低迷や猛暑の影響で、収益性の高い洋菓子やチョコレートなどの需要が落ち込んだ。
3. 十二月からのガス料金引き下げの影響で、通期の売上高は微減。

図 2　「因果関係あり」とラベル付けされた業績要因文の例

1. 電子化成品部門で、利益貢献度が高い電子材料の「高純度金属ヒ素」の需要が減少している。
2. 二〇〇〇年三月期の連結決算で、退職給与引当金を中心に過去最大の八千二百億円の特損を計上。
3. 売上高見込みは三二%増の二千四百億円で、従来予想を二百億円上方修正した。
4. 通信機器が同二〇%増と高い伸びで、コンピューターも好調だった。

図 3　「因果関係なし」とラベル付けされた業績要因文の例

手がかり表現 “で、” を含む業績要因文 2,000 件のうち，1,408 件が「因果関係あり」，592 件が「因果関係なし」とラベル付けされた．

### 4.2　因果関係を含まない文のパターン

ラベル付けを行った結果，手がかり表現 “で、” を含む業績要因文が因果関係を含まないと判定される特定のパターンが確認できた．図 4 にそれらを示す．業績要因文が図 4 に示すパターンのいずれかに該当する場合，因果関係を持たない可能性が高いと考えられる．なお，各項目のパーセンテージは「因果関係なし」とラベル付けされた文のうち，そのパターンに該当するものがどの程度の割合を占めるかを表す．

**手がかり表現の前に特定の語が現れる場合**

- 場所や地域などを表す語 (図 3 の 1)．約 10%．
- 期間や時期などを表す語 (図 3 の 2)．約 12%．
- 金額やパーセンテージなどを表す語 (図 3 の 3)．約 24%．
- その他特定の語（「一方」など）．約 27%．

**並列**

原因表現と結果表現の事象が並列関係にある場合 (図 3 の 4)．約 27%．

図 4　因果関係を持たないパターン

## 5　提案手法

業績要因文が因果関係を持つか否かの判定には Support Vector Machines(SVM) [6] を使用する．SVM に用いる素性として，以下に示すものを使用する．

- 形態素のユニグラム
- 形態素のバイグラム
- 助詞のペア
- 手がかり表現の直前形態素の品詞

### 5.1　前処理

素性の取得を行う前に，各業績要因文に対して以下の処理を行い，文を整形する．

**括弧とその中身の除去**

半角括弧 “()”，全角括弧 “（）” とその中身は素性取得の際のノイズとなる可能性があるため，形態素解析を行う前の段階で除去する．

**数字の置換**

連続して出現する数字を一つの “〇（レイ）” に置き換える．

### 5.2　助詞のペア

業績要因文において，一方が原因表現，他方が結果表現に出現する助詞の組合せを素性とする．助詞のペアを素性として使用する理由は，原因表現と結果表現が並列関係にある文の判定に助詞が有効であると考えられるためである．

助詞のペアは，以下のアルゴリズムに従って取得する．係り受け解析には CaboCha [3] を使用した．なお，核文節，および，基点文節は Sakaji らの手法に基づき，以下のように定義する．

**核文節**　手がかり表現を含む文節
**基点文節**　核文節の係り先となる文節

Step 1:　業績要因文を文節ごとに区切る．
Step 2:　文節を先頭から走査する．
Step 3:　文節が核文節に係る場合
- その文節に含まれる助詞を前部助詞リストに追加する．

Step 4:　文節が基点文節に係る場合
- その文節が核文節以前に位置する場合はスキップする．
- その文節に含まれる助詞を後部助詞リストに追加する．

Step 5:　前部助詞が取得できなかった場合
- 原因表現のうち，核文節に一番近い助詞を前部助詞リストに追加する．
- 原因表現に助詞が存在しない場合は前部助詞リストに null を追加する．

Step 6:　後部助詞が取得できなかった場合
- 結果表現のうち，基点文節に一番近い助詞を後部助詞リストに追加する．
- 結果表現に助詞が存在しない場合は後部助詞リストに null を追加する．

Step 7:　前部助詞と後部助詞の全ての組合せ (助詞のペア) を素性とする．

上記アルゴリズムによる処理の例を図 5 に示す．

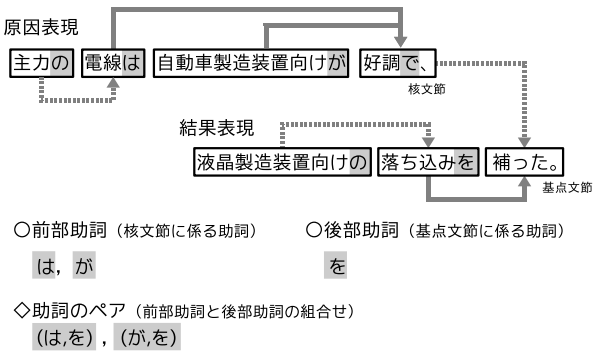


図 5　助詞ペア取得の例

## 6　実験と考察

日経新聞における企業の業績発表記事 302,837 件（1990 年から 2005 年までの 16 年分）から，Sakai らの手法を用いて業績要因文 139,666 件を抽出した．抽出された業績要因文のうち手がかり表現 “で、” を含む 7,546 件の一部に対して人手で「因果関係あり」,「因果関係なし」のラベル付けを行った．

ラベル付けした文のうち，1,200 件（因果関係あり・なし，それぞれ，600 件ずつ）を学習データとして使用し，507 件（因果関係あり 303 件，なし 204 件）をテストデータとして判定を行った．

判定には $SVM^{Light}$[*3]を用いた．またカーネルは線形を使用した．判定結果を表 1 に示す．なお，表中において，Uni は形態素のユニグラム，Bi は形態素のバイグラム，Pair は助詞のペア，Class は手がかり表現の直前形態素の品詞情報を表す．

表 1 に示す通り，F 値が最も高かったのは形態素のバイグラムと手がかり表現の直前形態素の品詞情報を使用した場合であるが，他の素性を使用した場合との差は僅かである．特に，形態素のユニグラム・バイグラムは他の素性と併用せずとも良好な結果が得られていることがわかる．これは，本研究が因果関係の有無判定の対象を企業の業績要因文のうち，特定の手がかり表現 “で、” を持つもののみに限定したためだと考えられる．判定の対象を限定することで文中に現れる語彙，特に，因果関係の判定に有効となる語彙が限られるため，特徴語や重要語といった語の選択の必要性がなかったと考えられる．

表 1　テストデータの分類結果

| 使用した素性 | 精度 (%) | 再現率 (%) | F 値 |
|---|---|---|---|
| Uni | 85.6 | 82.2 | 83.9 |
| Bi | 89.2 | 84.5 | 86.8 |
| Uni, Bi | 89.6 | 85.2 | 87.3 |
| Uni, Pair | 89.1 | 83.2 | 86.0 |
| Uni, Class | 89.4 | 85.8 | 87.6 |
| Bi, Pair | 91.2 | 82.5 | 86.6 |
| Bi, Class | 90.7 | 86.8 | 88.7 |
| Uni, Pair, Class | 91.3 | 86.1 | 88.6 |
| Bi, Pair, Class | 92.4 | 84.5 | 88.3 |
| All | 91.5 | 85.5 | 88.4 |

$SVM^{Light}$ により生成されたモデルファイルを解析し，各素性のスコアを算出した．そのうち，因果関係の判定に有効であったもの（スコアの絶対値が大きかったもの）の上位 10 件をそれぞれ表 2，3 に示す．

表 2　素性のスコア 上位 10 件

| スコア | 素性 | 素性グループ |
|---|---|---|
| 0.322820 | 名詞-形容動詞語幹 | Class |
| 0.267719 | 名詞-サ変接続 | Class |
| 0.264473 | (の, が) | Pair |
| 0.247823 | こと | Uni(Cause) |
| 0.228851 | の | Uni(Cause) |
| 0.202575 | ことで | Bi(Cause) |
| 0.197765 | 効果 | Uni(Cause) |
| 0.197167 | たこと | Bi(Cause) |
| 0.182212 | 効果で | Bi(Cause) |
| 0.159440 | 投資 | Uni(Cause) |

表 3　素性のスコア 下位 10 件

| スコア | 素性 | 素性グループ |
|---|---|---|
| -0.395677 | も | Uni(Effect) |
| -0.346811 | (が, も) | Pair |
| -0.324537 | ○ | Uni(Cause) |
| -0.219866 | 一方で | Bi(Cause) |
| -0.219866 | 一方 | Uni(Cause) |
| -0.166597 | 中 | Uni(Cause) |
| -0.165695 | 中で | Bi(Cause) |
| -0.159959 | のは | Bi(Cause) |
| -0.148275 | が | Uni(Cause) |
| -0.127912 | (null, は) | Pair |

表 2, 3 の項目である「素性グループ」は，その素性が 5 節で示すどの素性グループに属しているかを表し，(Cause)，(Effect) の表記はその属性が原因表現，結果表現のどちらから取得されたものであるかを表す．

[*3] http://svmlight.joachims.org/

表 2 に示す素性のうち上位 2 つが手がかり表現の直前形態素の品詞情報である．品詞が「名詞-形容動詞語幹」である形態素には，「好調」「不振」などがあり，同様に，品詞が「名詞-サ変接続」である形態素には「影響」「低迷」などがある．これらの形態素は，因果関係を含む文において手がかり表現の直前にたびたび出現することから，高いスコアが与えられたと考えられる．

表 3 に示す素性のうち最もスコアが小さいものが結果表現から取得されたユニグラム「も」で，次にスコアが小さいものが助詞のペア (が, も) となっている．手がかり表現 “で、” を含む業績要因文が並列構文となる場合，3.2 節で示した例文のように，結果表現に助詞「も」が含まれる．並列構文である業績要因文は因果関係を含まないため，前述した素性に小さなスコアが与えられたと考えられる．

また，スコアの絶対値の大きな素性のうちの多くが，原因表現から取得されたものであることがわかる．さらに，それらの素性のほとんどが手がかり表現 “で、” の直前に出現するものである．これより，今回判定対象とした業績要因文においては，原因表現中の形態素，特に，手がかり表現の直前に出現する形態素が因果関係の有無に大きく影響していると考えられる．

## 7 エラー解析

テストデータ 507 件について手がかり表現 “で、” の意味に応じて表 4 のように人手で分類した．なお，分類されたデータの数がテストデータ数と異なるのは，複数の意味を持つと解釈できる文が確認されたためである．表 4 において，「原因・理由・動機を表す」に分類された文のみが因果関係を含む．

各素性を組み合わせて判定実験を行い，誤判定した文の意味とその数を調査した．各素性の組合せごとの結果を表 5 に示す．

表 4　意味によるテストデータの分類

| 意味 | 数 |
|---|---|
| 原因・理由・動機を表す | 303 |
| 動作・作用を行うときの状況・状態を表す | 123 |
| 並列を表す | 30 |
| 動作・作用の行われる場所を表す | 25 |
| 動作・作用の行われる時を表す | 11 |
| 逆接を表す | 11 |
| その他 | 10 |

表 5　誤判定した文の意味による内訳

| | 原因 | 状況 | 並列 | 場所 | 時 | 逆接 | 他 | 計 |
|---|---|---|---|---|---|---|---|---|
| Uni | 54 | 16 | 8 | 12 | 3 | 1 | 2 | 96 |
| Uni,Pair | 51 | 12 | 4 | 8 | 3 | 2 | 2 | 82 |
| Uni,Class | 43 | 10 | 5 | 10 | 2 | 1 | 3 | 74 |
| Uni,Pair,Class | 42 | 9 | 4 | 7 | 1 | 1 | 3 | 67 |
| Bi | 47 | 15 | 9 | 6 | 0 | 0 | 1 | 78 |
| Bi,Pair | 53 | 12 | 4 | 5 | 0 | 0 | 3 | 77 |
| Bi,Class | 40 | 9 | 9 | 5 | 2 | 0 | 2 | 67 |
| Bi,Pair,Class | 47 | 8 | 4 | 5 | 2 | 0 | 2 | 68 |
| Uni,Bi | 45 | 11 | 7 | 7 | 2 | 1 | 2 | 75 |
| All | 44 | 9 | 5 | 5 | 1 | 1 | 3 | 68 |

表 5 より，素性の組合せによって誤判定した文の意味の内訳が異なることがわかる．なお，前述のとおり，文が因果関係を含むと判定されるのは，手がかり表現が文中において「原因・理由・動機を表す」場合のみである．そのため，表 5 における誤判定とは，手がかり表現が「原因・理由・動機を表す」文については「因果関係を含む文が “因果関係なし” と判定された」ことを意味する．反対に，それ以外の文については「因果関係を含まない文が “因果関係あり” と判定された」ことを意味する．形態素のユニグラム・バイグラムの双方において，手がかり表現の直前形態素の品詞情報を素性として併用した場合，因果関係を含む文と状況を表す文の誤判定は減少している．また，同様に助詞ペアを素性として併用した場合は，双方ともに並列構文の誤判定は減少している．しかし，バイグラムと助詞ペアの組合せについては因果関係を含む文の誤判定が増加しているため，結果的にバイグラムのみを使用した場合とほとんど差がない．ユニグラムに助詞のペアと品詞情報を組み合わせた場合は全体的な誤判定の減少が見られるが，バイグラムの場合は前述のとおり助詞ペアによる誤判定の増加が起こるため，ユニグラムほど大きな効果は見られない．

## 8 まとめ

我々は手がかり表現 “で、” を含む業績要因文がどのような場合に因果関係を持つのか調査し，判定のため素性の選択を行った．判定対象が限定されていることから，素性として形態素のユニグラムまたはバイグラムのみを使用した場合でも良好な結果が得られることがわかった．また，助詞ペアの使用は並列構文の判定には効果が見られたが，バイグラムと組み合わせた場合には誤判定の原因となることがわかった．

今後の課題として，因果関係有無の判定対象を，手がかり表現 “で、” を含む，より一般的な文へと拡張することが考えられる．

## 参考文献


[1] Roxana Girju. Automatic detection of causal relations for question answering. In *In Proceedings of Annual Meeting of the Association for Computational Linguistics, Workshop on Multilingual Summarization and Question Answering - Machine Learning and Beyond*, 2003.

[2] **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.

[3] **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**. **情報処理学会論文誌**, Vol. 43, No. 6, pp. 1834–1842, 2002.

[4] H. Sakai and S. Masuyama. Cause information extraction from nancial articles concerning business performance. *IEICE Trans. on Information and Systems*, Vol. E91-D, No. 4, pp. 959–968, 2008.

[5] Hiroki Sakaji, Satoshi Sekine, and Shigeru Masuyama. Extracting causal knowledge using clue phrases and syntactic patterns. *7th International Conference on Practical Aspects of Knowledge Management (PAKM)*, pp. 111–122, 2008.

[6] V.N. Vapnik. *Statistical Learning Theory*. Wiley, 1999.