3L4-OS-26b-2

決算短信PDFからの因果関係抽出に基づく過去事象間の関連表示システム

A Causal Expressions Search System for PDF Files of Summary of Financial Statements

坂地泰紀 *1　酒井浩之 *1　増山繁 *2
Hiroki Sakaji　Hiroyuki Sakai　Shigeru Masuyama

*1成蹊大学　*2豊橋技術科学大学
Seikei University　Toyohashi University of Technology

This paper proposes a system that searches causal expressions from pdf files of a summary of financial statements. First, our method extracts causal expressions from the pdf files. Then, we construct a search system for the extracted causal expressions. Finally, we evaluate our system.

1. はじめに

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．さらに，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因果関係が重要である．例えば，原因「猛暑」，結果「冷房需要の盛り上がり」といった因果関係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」の因果関係から，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．さらに，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因果関係を取得するのには多大な労力を要する．そのため，我々は，企業の業績発表に関する記事から因果関係を抽出する手法を提案した [坂地 13]．しかしながら，[坂地 13] の手法では，日本経済新聞記事を対象としているため，これを用いた因果関係検索システムを作成したとしても，著作権の関係で一般には公開できない．さらに，大企業の業績発表は，経済新聞に業績発表記事として掲載される可能性が高いが，証券市場に上場している企業数約 3,500 社の全ての業績発表が記事になるとは限らず，そのため，業績発表記事のみを対象としている [坂地 13] の手法では，全ての企業を網羅できない．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 に着目した．企業の Web ページに掲載されている決算短信 PDF を使用できれば，業績発表記事を対象とするより多くの企業を対象にすることができる．そこで，本研究では，決算短信 PDF から因果関係を抽出し，抽出した因果関係を検索するシステムの開発を行う．

連絡先: 坂地泰紀，成蹊大学，東京都武蔵野市吉祥寺北町 3-3-1，hiroki_sakaji@st.seikei.ac.jp

2. システム構築手法

因果関係判定手法と因果関係抽出手法を用いて因果関係を抽出した後に，抽出した因果関係を検索することが可能なシステムを作成する．以下に，システム構築手法を示す．

Step 1: 各企業サイトから決算短信 PDF を収集する．収集した PDF をテキスト *1 に変換する．

Step 2: 収集したテキストデータから，因果関係判定手法 [坂地 11] を用いて，因果関係を含む文を抽出する．

Step 3: 因果関係を含んでいると判定された文から因果関係抽出手法 (節 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因果関係として抽出する．

Step 4: 抽出した因果関係を保存した因果データベースを作成し，因果関係を検索できるシステムを構築する．

3. 因果関係を含む文の抽出

本手法では，因果関係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，因果関係を抽出する．例えば,「ため」は，因果関係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，因果関係以外の意味を持つものがある．例えば,「あなたのために、花を買った。」という文中の「ため」は，原因・結果ではなく，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，半教師在り学習を用いたフィルタリング手法 [坂地 11] を適用し，因果関係を含む文を決算短信 PDF から抽出する．そして，抽出された文に対して，次節で述べる因果関係抽出手法を適用し，因果関係を抽出する．

4. 因果関係の抽出

本節では，決算短信 PDF からの因果関係を表す表現の抽出方法について述べる．ここで，原因・結果を，それぞれ，原因表現と結果表現と本論文では定義する．本手法では，因果関係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，決算短信 PDF から因果関係を自動的に抽出する．文献 [庵 12] に準拠し，因果関係は，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

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

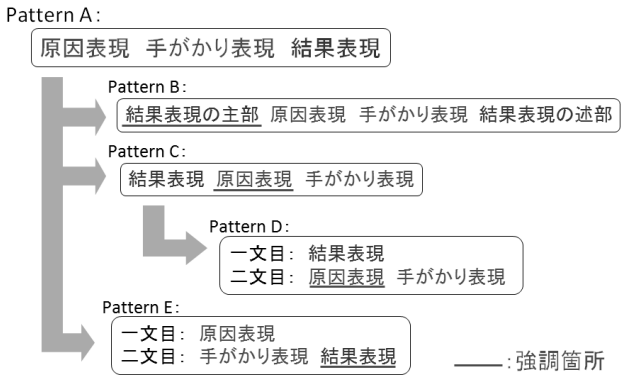

図 1: 各 Pattern の関連図

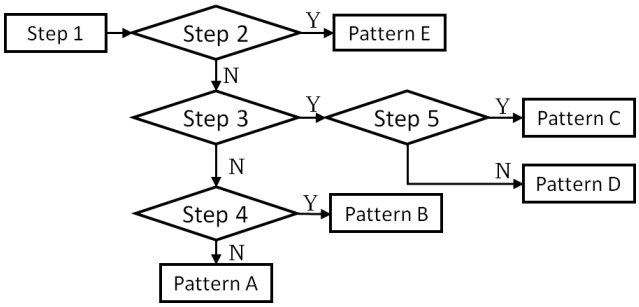

図 2: Pattern 識別の概要

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，以前，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 5 通りに分類し，因果関係を抽出するための手がかり表現を取得した [Sakaji 08]．その 5 通りを Pattern A から D とし，図 1 に示す．本手法は，この 5 通りの Pattern から因果関係を獲得するアルゴリズムを用いて，因果関係を抽出する．

図 1 において，我々は Pattern A は基本型であると考えた．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 文に分割したのものである．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．

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．*Identification of patterns* の概要を図 2 に示す．

[*Identification of patterns*]

Step 1: 手がかり表現を人手で与え，それ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, Step 6 を実行する．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，もしくは，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，もしくは，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 を適用する．Step 6 を実行する．

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．

Step 6: 手続きを終了する． □

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，もしくは，格助詞を含む文節が存在しないため，Pattern A が適用される．

5. 因果関係抽出手法の改良

節 4. で示した因果関係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因果関係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」で因果関係を抽出しようとした場合，原因表現として「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」，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パタンである．このようなパタンが決算短信 PDF に数多く散見されたため，このパタンを獲得し，因果関係抽出に用いる新たな手法を開発する．本研究では，このような文頭に出現するパタンを **Prefix Pattern** と定義する．

表 1: Prefix Pattern の例

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

5.1 新しい手がかり表現獲得

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.「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。

上記の結果より，数多くの手がかり表現を獲得できると予想したが，実際には各接尾辞に対応した手がかり表現が一つずつしか存在しなかった.

5.2 Prefix Pattern 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」が存在することから，これを用いて獲得する．図 3 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 回以上の繰り返しを意味する．図 3 では，例として手がかり表現

図 3: Prefix Pattern 候補の獲得の例

「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．そのため,「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．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.

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 1 に示す.

5.3 Prefix Pattern を用いた因果関係抽出手法

Prefix Pattern を用いた因果関係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 を伴った因果関係を抽出できるようになる．ただし，Pattern C での因果関係を抽出できなくなる.

6. 評価実験

決算短信 PDF から因果関係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．形態素解析器としては Mecab[*2] を用い，係り受け解析器としては Cabocha[工藤 02] を用いた．学習器には SVM^{Light}[*3] を用い，カーネルは線形を用いた．手がかり表現には，[Sakaji 08] で獲得された手がかり表現から抽出精度の低い手がかり表現を除いた 34 個を用いた.

評価データには，学習データに用いたものを除いた決算短信 PDF からランダムに 20 ファイルを用いた．20 ファイルに対して人手で因果関係を表すタグ（「原因」,「結果」）を付与したところ，277 個の因果関係が存在した.

表 2 に経済新聞記事を対象とした既提案手法と，節 4. で述べた改良手法（本手法）の評価結果を示す．精度は，手法で抽出した因果関係のうち，正解だった割合を示す．再現率は，評価データに含まれる 277 個の因果関係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.

表 2: 評価結果

	精度	再現率	F 値	抽出数	手がかり表現数
既提案手法	0.82	0.60	0.69	201	34
本手法	0.85	0.65	0.73	211	39

7. 考察

表 2 より，本手法の方が精度，再現率，F 値の全てにおいて既存手法を上回った．これは，決算短信 PDF に特徴的に出現する Prefix Pattern を伴う因果関係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.

*2 http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
*3 http://svmlight.joachims.org/

8. システム構築

本節では，決算短信 PDF から抽出した因果関係を検索できるシステム構築について述べる．まず，各企業 Web ページから決算短信 PDF を収集した結果，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 に対して本手法を適用し，因果関係を抽出した結果，1,181,945 個の因果関係を抽出した．今回は，LAMP(Linux, Apache, MySQL, PHP) 環境においてシステムを構築した *4．本システムでは，各企業に紐づいた因果関係 (原因表現と結果表現の組) を検索することができる．図 4 に，検索結果の例を示している．

図 4: 検索結果画面の例

図 4 では，原因表現に「天候不順」を含む因果関係を検索している．図 4 より，神栄株式会社の 2011 年 5 月に発表された決算短信 PDF に含まれる原因表現「年度はじめの天候不順や猛暑の影響」，結果表現「当社グループのアパレル分野でも低調に推移しました。」を検索することができた．また，企業名「ソニー」，原因表現「プレイステーション」で検索したところ，2013 年 10 月に発表された決算短信 PDF から以下の因果関係を抽出することができた．

原因表現 「プレイステーション 4」(以下「PS4TM」) の導入に向けた研究開発費の増加及び PSVita の戦略的価格改定の影響

結果表現 ゲーム分野は前年同期に比べ損益が大幅に悪化しました。

9. 関連研究

Chang らは手がかり表現と語の組の出現確率を用いて，2 つの名詞句間の因果関係を抽出する手法を提案している [Chang 06]．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因果関係の検出と抽出を行う手法を提案している [Girju 03]．彼らの研究は名詞句の組を因果関係の対象としているため，他の表現間の因果関係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．

Bethard らは Syntactic 素性と Semantic 素性を用いて，動詞対に対して因果関係があるか否かの判定を行う手法を提案している [Bthard 08]．上記で述べた研究では，名詞句や動詞句に限って因果関係を抽出している．しかしながら，因果関係を構成する原因・結果表現は名詞句や動詞句のみとは限らない．本手法では，手がかり表現を対象に因果関係を表す意味かどうかを判定しているため，動詞句でも名詞句でも判定することが可能である．

10. まとめ

本研究では，決算短信 PDF から因果関係を抽出し，それを検索するシステムの構築を行った．決算短信 PDF に特徴的に出現する Prefix Pattern を自動的に獲得し，これを用いて因果関係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 0.65，F 値 0.73 を達成した．今後の課題として，Pattern C の場合も因果関係を抽出できるように手法を改良することが挙げられる．

参考文献

[Bthard 08] Bthard, S. and H.Martin, J.: Learning Semantic Links from a Corpus of Parallel Temporal and Causal Relations, in *in Proceedings of ACL-08*, pp. 177–180 (2008)

[工藤 02] 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)

[Chang 06] Chang, D.-S. and Choi, K.-S.: 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)

[Girju 03] Girju, R.: Automatic detection of causal relations for Question Answering, in *In ACL Workshop on Multilingual Summarization and Question Answering*, pp. 76–83 (2003)

[Sakaji 08] Sakaji, H., Sekine, S., and Masuyama, S.: Extracting Causal Knowledge Using Clue Phrases and Syntactic Patterns, in *7th International Conference on Practical Aspects of Knowledge Management (PAKM)*, pp. 111–122 (2008)

[庵 12] 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)

[坂地 11] 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, 電子情報通信学会論文誌 D, Vol. J94-D, No. 8, pp. 1496–1506, (2011)

[坂地 13] 坂地 泰紀, 酒井 浩, 増山 繁：企業業績発表記事からの因果関係抽出, 人工知能学会第 11 回金融情報学研究会, pp. 37–43 (2013)

*4 http://hawk.ci.seikei.ac.jp/CS/