# アナリストレポートからのアナリスト予想根拠情報の抽出

# Extraction of Basis Information on Analyst's Forecasts from Analyst Reports

酒井浩之 [1*]　柴田 宏樹 [1]　平松 賢士 [2]　坂地泰紀 [1]
Hiroyuki Sakai [1]　Hiroki Shibata[1]　Kenji Hiramatu[2]　Hiroki Sakaji [1]

[1] 成蹊大学 理工学部 情報科学科
[1] Department of Computer and Information Science, Faculty of Science and Technology, Seikei University
[2] 株式会社アイフィスジャパン
[2] IFIS JAPAN LTD

**Abstract:** We propose a method of extracting basis information on analyst's forecasts from analyst reports. Our method extracts basis information on analyst's forecasts from analyst reports by using frequent expressions (e.g., "earning capacity") and clue expressions (e.g., "is expected"). The frequent expressions and clue expressions are extracted from the analyst reports automatically. We evaluated our method and it attained 75.0% precision and 61.7% recall, respectively.

## 1　はじめに

近年，投資家に対して投資判断の支援をおこなう技術の必要性が高まっている．そのため，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の支援を行う技術が注目されている．その一例として，日本銀行が毎月発行している「金融経済月報」，内閣府が調査・公表している景気ウォッチャー調査，TDNET 適時開示情報閲覧サービス [1] や企業 Web ページなどで公開される決算短信をテキストマイニングの技術を用いて分析する研究などが盛んに行われている [1][3][6][5].

アナリストレポートとは，証券アナリストが企業の経営状態や収益力などを調査し，その結果をまとめたレポートのことである．アナリストレポートには，証券アナリストの調査や分析に基づき，その企業の業績予測や，株価や事業の今後の展望などが記述され，それらのアナリスト予想をもとにレーティング [2] が付与されている点が特徴である．高度な専門知識をもつ証券アナリストによる詳細なレポートは投資判断のための最も重要な情報源のひとつであり，アナリストレポートの内容が株価の変動要因にもなりうる．例えば，アスクルの 2016 年 7 月 6 日の株価は急反発したが，その要因は国内大手証券がレーティングを最上位に格上げしたからである [3]．毎日のように多くのアナリストレポートが発表され，特に決算発表が近くなると本数は増えるが，多い日は一日 1200 本ものアナリストレポートが発表されることもあり，その全てを熟読することは難しい．そのため、アナリストレポートの内容を十分把握できないケースも想定される。

アナリストレポートのレーティングは証券アナリストの調査，分析によるアナリスト予想に基づいて付与されるが，レーティングの変化だけでなく，そのアナリスト予想の根拠となる情報が投資判断では重要である．そこで本研究では，投資判断やアナリストレポートの内容を把握するうえで重要な，アナリスト予想の根拠情報を含む文を自動的に抽出する手法を提案する．例えば,「世界経済の回復に加え、米ドル安や中国の供給削減期待などから市況は急回復し、同社の輸出も今後大きく改善する可能性が高い。」といった文を自動的に抽出する．以降，アナリスト予想の根拠情報を含む文をアナリスト予想根拠文と定義する．本研究により，アナリストレポートから抽出されたアナリスト予想根拠文のみを提示することで，アナリストレポートの内容把握に必要な時間を削減できたり，アナリスト予想根拠文の中からその企業の事業に関連する根拠情報の

*連絡先：成蹊大学 理工学部 情報科学科
〒 180-8633 東京都武蔵野市吉祥寺北町 3-3-1
E-mail: h-sakai@st.seikei.ac.jp

[1] https://www.release.tdnet.info/inbs/main.html

[2] アナリストレポートにより多少の差異があるが，例として Strong Buy，Buy，Hold，Sell，Strong Sell の 5 段階のレーティングが付与されている．

[3] http://news.finance.yahoo.co.jp/detail/20160707-17267814-ifis-stocks

みを選別することで，その企業への投資判断やそのアナリストレポートを熟読するかどうかを判断するための情報となることが期待できる．

## 2　アナリスト予想根拠情報の抽出

### 2.1　アナリストレポートからの手がかり表現の自動獲得

アナリスト予想根拠文の抽出は，まずは，業績発表記事から業績要因表現を抽出した酒井らの手法 [4] を適用し，アナリストレポートからアナリスト予想根拠文を抽出するのに有効な手がかりとなる表現（例えば「予想する」「寄与しよう」等．以降,「手がかり表現」と定義）を獲得する．そして，獲得された手がかり表現等を使用することで，アナリスト予想根拠文の抽出を行う．手がかり表現を獲得する手法の概要を以下に示す．

**Step 1:** 少数の手がかり表現（具体的には,「予想する」,「考える」,「高い」の 3 つの表現を用いる）を人手で与え，それに係る節を取得する．

**Step 2:** 取得した節の集合から，その中で共通して頻繁に出現する表現を共通頻出表現（例えば「可能性」「急回復」等）として，後述の基準を用いて抽出する．

**Step 3:** 共通頻出表現が係る節を取得し，その中から新たな手がかり表現を，後述の基準を用いて抽出する．

**Step 4:** 獲得した手がかり表現から，それに係る節を取得する．

**Step 5:** Step 2 から Step 4 を，新たな手がかり表現と共通頻出表現が獲得されなくなる，もしくは，予め定めた回数まで繰り返す（図 1 を参照）．□

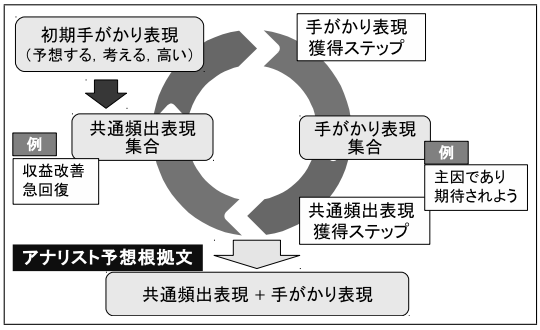


図 1: 共通頻出表現・手がかり表現自動獲得手法の概要

Step 2 では，手がかり表現に係る節の集合から，適切な共通頻出表現を選別する．具体的には，様々な手がかり表現に係っている共通頻出表現は適切であるという仮定に基づき，共通頻出表現が手がかり表現に係る確率に基づくエントロピーを式 1 で求め，その値が，ある閾値 $T_e$ 以上の共通頻出表現を選別する．

$$H(e) = -\sum_{s \in S(e)} P(e,s) \log_2 P(e,s) \tag{1}$$

ここで，アナリストレポートの集合において，

$S(e)$**:** 共通頻出表現 $e$ が係る手がかり表現の集合．

$P(e,s)$**:** 共通頻出表現 $e$ が手がかり表現 $s$ に係る確率．式 2 で求める．

$$P(e,s) = \frac{f(e,s)}{\sum_{s' \in S(e)} f(e,s')} \tag{2}$$

ここで，$f(e,s)$ は，共通頻出表現 $e$ が手がかり表現 $s$ に係る回数である．

エントロピー $H(e)$ は，共通頻出表現 $e$ がアナリストレポート集合において様々な手がかり表現に均一の確率で係っている場合に高い値をとる．

閾値 $T_e$ は，以下の式 3 によって設定する．

$$T_e = \alpha \log_2 |Ns| \tag{3}$$

ただし，$Ns$ は共通頻出表現を取得するのに使用した手がかり表現の集合，$\alpha$ は定数 $(0 < \alpha < 1)$ である．以下に，定数 $\alpha$ を 0.50 とした場合に抽出された共通頻出表現をいくつか示す．

収益改善，収益力，ピークアウト，高成長持続，急回復

Step 2 で共通頻出表現の抽出を行った後，Step 3 にて，その選別した共通頻出表現から新たな手がかり表現を獲得する．ここでも，様々な共通頻出表現が係っている手がかり表現は適切であるという仮定に基づき，手がかり表現の選別を行う．具体的には，手がかり表現候補に対して共通頻出表現が係る確率に基づくエントロピーを求める．

$$H(s) = -\sum_{e \in E(s)} P(s,e) \log_2 P(s,e) \tag{4}$$

ただし，アナリストレポート集合において，

$E(s)$**:** 手がかり表現 $s$ に係る共通頻出表現の集合．

$P(s,e)$**:** 手がかり表現 $s$ に対して共通頻出表現 $e$ が係る確率．式 5 で求める．

$$P(s,e) = \frac{f(s,e)}{\sum_{e' \in E(s)} f(s,e')} \tag{5}$$

ここで，$f(s,e)$ は，手がかり表現 $s$ に対して共通頻出表現 $e$ が係る回数である．

そして，ある閾値以上の候補を手がかり表現として抽出する．閾値は，共通頻出表現と同様に式 3 によって設定するが，$Ns$ は新たな手がかり表現を獲得するのに使用した共通頻出表現の集合である．

以下に，定数 $\alpha$ を 0.50 とした場合に獲得された手がかり表現をいくつか示す．

寄与しよう，主因であり，奏功し，牽引する，織り込んでいる，注目点である，顕在化する，期待されよう

なお，手がかり表現を獲得するのに使用したアナリストレポートの総数は 22178 であり，獲得された共通頻出表現の数は 966 個，手がかり表現の数は 674 個であった．

## 2.2 手がかり表現と共通頻出表現を使用したアナリスト予想根拠文の抽出

アナリストレポートから獲得された「手がかり表現」と「共通頻出表現」を使用して，アナリストレポートからアナリスト予想根拠文を抽出する．以下に手法を示す．

**Step 1:** アナリストレポートから，手がかり表現を含む文を抽出する．

**Step 2:** Step 1 で取得された文を係り受け解析し，手がかり表現を含む文節を取得する．

**Step 3:** Step 2 で取得した文節に係っている文節を取得し，連結する．

**Step 4:** Step 3 の処理を，係り元の文節がなくなるまで繰り返す．

**Step 5:** 連結された文節に共通頻出表現が 3 個以上，含まれていた場合，その文をアナリスト予想根拠文として抽出する（図 2）． □

表 1 に，アナリストレポートから抽出されたアナリスト予想根拠文をいくつか示す [4]．

## 2.3 文末手がかり表現の自動獲得

2.1 節で取得された手がかり表現は 1 文節で構成される．しかし，実際は 2 文節以上の文節で構成される手がかり表現も多く存在する．例えば「可能性があると

[4]アナリストレポートの著作権の関係上，実際に抽出された文に変更を加えてある．

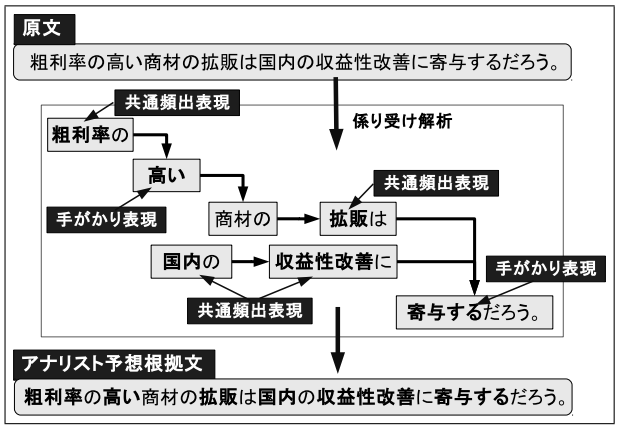


図 2: 手がかり表現と共通頻出表現を使用したアナリスト予想根拠文の抽出

みる。」という表現は手がかり表現として有効であるが，「可能性が」，「あると」，「みる。」の 3 文節で構成される．しかし，2.1 節の手がかり表現獲得手法では，このような多文節で構成される手がかり表現は獲得されず，また 1 文節の「みる。」だけでは有効な手がかり表現ではない．そこで，2.2 節で抽出されたアナリスト予測根拠文の文末に着目し，文末に出現しており，かつ，多文節で構成される手がかり表現を取得する（以降，文末に出現しており，かつ，多文節で構成される手がかり表現を文末手がかり表現と定義する．）．文末に着目した理由は，アナリストレポートには文末に特徴的な表現が多く出現している傾向があるからである．

文末手がかり表現を抽出するにあたり，まず，2.2 節で抽出されたアナリスト予想根拠文の文末に出現する 1 文節（以降，文末文節と定義）を取得する．その結果，例えば以下のような文末文節が取得された．

ある。，みている。，みる。，みられる。，なった。，なる。

これらの文末文節は手がかり表現として有効ではない．しかし，文末文節に係っている文節列を取得し，文末文節と組み合わせることで，有効な文末手がかり表現となる可能性がある．そこで，文末手がかり表現を獲得するにあたり，文末文節に係っている文節列を取得する．以下に，「みる。」に係る文節列の例を示す．

可能性が高いと，可能性があると，限定的と，会社計画を上回ると

このように，「みる。」だけでも非常に多くの文節列が取得され，文末文節とそれに係る文節列との組み合わせ数は膨大なものになる．そこで，文末文節とそれに係る文節列との組み合わせを絞り込む．具体的には，文末文節 $c$ に係る文節列 $p$ に対して以下の式 6 でスコア

表 1: アナリスト予想根拠文の例

| | |
|---|---|
| 文例 1 | 粗利率の高い機器の拡販は国内の収益性改善に寄与するだろう。 |
| 共通頻出表現<br>手がかり表現 | 国内，粗利率，収益性改善，拡販<br>高い，寄与する |
| 文例 2 | 当社保有の独自技術を用いた再生医療は事業として十分成立すると考えており、その中で当社がオンリーワンであることの価値は高い。 |
| 共通頻出表現<br>手がかり表現 | 価値 事業 当社<br>高い |

を求め，このスコアがある閾値を上回る文節列のみを抽出する．

$$Score(p,c) = -f(p,c)\sqrt{fp(p)}\log_2 P(p,c) \quad (6)$$

$$P(p,c) = \frac{f(p,c)}{N(c)} \quad (7)$$

ただし，アナリストレポートから取得したアナリスト予想根拠文の集合において，

$P(p,c)$: 文末文節 $c$ から取得される文節列 $p$ の出現確率，

$f(p,c)$: 文末文節 $c$ から取得される文節列 $p$ の取得回数．

$N(c)$: 文末文節 $c$ から取得される文節列の総数

$fp(p)$: 文節列 $p$ に含まれる文節の数，

ここで，文末文節 $c$ から取得される文節列の中で，3 回以上出現し，かつ，$Score(p,c)$ が平均値以上の文節列を抽出する．そして，抽出された文節列と文末文節 $c$ を連結した表現を，文末手がかり表現として獲得する。獲得された文末手がかり表現の数は，10,339 個であった．以下に獲得された文末手がかり表現の例をいくつか示す．

割高感はないとみられる。，上回る成長を見込む。，
増益基調が続くとみる。，改善することになる。，
押し上げる要因となろう。，下回る恐れがある。

文末手がかり表現を使用したアナリスト予想根拠文の抽出は，2.2 節で示した手法と同様である．

## 3 アナリスト予想根拠文からの重要文抽出

アナリストレポートによっては多くの文で構成される長い文章も多く存在する．そのような長いアナリストレポートからは，多くのアナリスト予想根拠文が抽出されるため，それを全て提示しては投資家に対する負担が大きいことに変わりはない．そこで，抽出されたアナリスト予想根拠文に対してスコアを付与し，そのスコアに基づいて重要なアナリスト予想根拠文を抽出する．

アナリスト予想根拠文へのスコア付与は，アナリスト予想根拠文に含まれるキーワードにスコアを付与し，そのキーワードのスコアの和とする．ここで，アナリストレポートが対象としている企業におけるキーワードのスコアを求める必要があるが，これは，その企業の決算短信 PDF から求める．企業ごとのキーワードのスコア計算に決算短信 PDF を使用した理由は以下のとおりである．対象とする企業に関する情報源として，決算短信 PDF の場合は，どのような上場企業であっても年に 4 回の決算短信が発表されるのに対し，企業によって発表されるアナリストレポートの数はまちまちである．株式市場における注目度が高い大企業であれば，それを対象としたアナリストレポートは多く発表されるのに対し，そうではない企業を対象としたアナリストレポートの発表数は少ない．そのため，アナリストレポートをキーワードのスコア付与の情報源とした場合，企業によってはアナリストレポートの数が不足する可能性がある．

決算短信 PDF を使用したキーワードのスコア付与は，企業 $t$ の決算短信 PDF における名詞 $n$ に対して，以下の式 8 で重み $W(n,S(t))$ を計算することで行う．

$$W(n_i,S(t)) = (0.5+0.5\frac{TF(n_i,S(t)}{\max_{j=1,\ldots,m} TF(n_j,S(t))}) \times H(n_i,S(t))\log_2\frac{N}{df(n_i)} \quad (8)$$

ここで，

$S(t)$: ある企業 $t$ の決算短信 PDF の集合．

$TF(n,S(t))$: $S(t)$ において，名詞 $n$ が出現する頻度．

$H(n,S(t))$: $S(t)$ の各決算短信 PDF である $d$ に名詞 $n$ が出現する確率 $P(n,d)$ に基づくエントロピー．以下の式 9 によって求める．

$$H(n,S(t)) = -\sum_{d\in S(t)} P(n,d)\log_2 P(n,d) \quad (9)$$

$df(n)$: 名詞 $n$ を含む決算短信 PDF をもつ企業の数．

表 2: 決算短信 PDF から抽出された企業キーワードの例

| 企業名称 | 企業キーワード |
|---|---|
| 東芝 | 電子デバイス部門, インフラ部門, デバイス部門, ストレージ, 医用システム事業 |
| 大日本印刷 | エレクトロニクス, 印刷事業, 液晶カラーフィルター, 情報コミュニケーション部門 |
| カゴメ | 野菜飲料, 野菜生活, 果美食品, 生鮮トマト, 飲料事業, お茶飲料, 植物性乳酸菌ラブレ |
| エーザイ | 医薬品, アリセプト, パリエット, 医薬品事業, 抗がん剤, 製薬用機械 |
| 三菱商事 | 資源関連, 金融事業, エネルギー事業, ＬＮＧ, 石油, インフラ事業 |
| セイコーエプソン | デバイス精密機器事業, センサー産業機器事業, インクジェットプリンター |

$N$: 決算短信 PDF を収集した企業の数.

$W(n, S(t))$ は，情報検索で一般的な $tf \cdot idf$ 値を，1 つの企業の決算短信 PDF の集合を 1 つの文書とみなして求め，さらに，その企業の決算短信 PDF 集合においてまんべんなく出現している場合に高い値をとる尺度を組み合わせたものである．従って，$W(n, S(t))$ は，企業 $t$ の決算短信 PDF の集合中に多く，かつ，まんべんなく出現し，他の企業の決算短信 PDF には出現していない名詞 $n$ に対して大きな値が割り当てられる．表 2 に，上記の手法によって，企業ごとの決算短信 PDF から高い重みが付与されたキーワード（以降，企業キーワードと定義）をいくつか示す.

企業 $t$ のアナリストレポートから抽出したアナリスト予想根拠文 $cs(t)$ のスコア $W(cs(t))$ は，アナリスト予想根拠文に含まれる企業キーワードのスコアの和とする.

$$W(cs(t)) = \sum_{n_i \in T(cs(t))} W(n_i, S(t)) \qquad (10)$$

ここで，$W(n_i, S(t))$ は企業 $t$ における企業キーワード $n_i$ の重み，$T(cs(t))$ は，アナリスト予想根拠文 $cs(t)$ に含まれる企業 $t$ の企業キーワードの集合である.

以下に，1 つのアナリストレポートから抽出されたアナリスト予想根拠文に対して高いスコアが付与された文の例をいつくか示す [5].

- 健康食品事業の売上高は前年同期比２％の減収にとどまったが、営業利益は同８０％の大幅な増益を達成、弊社の事前想定の同５５％増を上回った。
- 主力の乳製品事業において、高採算の商品の予想増収率を前回予想の前期比１０％から同２０％増に引き上げた。

# 4 評価

本手法を実装し，22,178 個のアナリストレポートから手がかり表現と共通頻出表現を獲得した．企業キーワード抽出のための決算短信 PDF は，123,042 個の決算短信 PDF ファイルを使用した．実装にあたり，形態素解析器として MeCab[6]，係り受け解析器として CaboCha[2] を使用した.

評価において，手がかり表現と共通頻出表現を獲得したアナリストレポートではないアナリストレポートからアナリスト予想根拠文を抽出した．正解データとして，12 個のアナリストレポートを，手がかり表現と共通頻出表現を獲得したアナリストレポートを含まないアナリストレポート集合から無作為に選択し，その中の合計 468 文から人手にてアナリスト予想根拠文を抽出して作成した．次に，選択したアナリストレポートから本手法にてアナリスト予想根拠文を抽出し，そのアナリスト予想根拠文と正解データの文が一致すれば正解とし，精度，再現率，F 値を算出した．比較手法として，以下の手法と本手法とを比較する.

**本手法１**： 手がかり表現と文末手がかり表現を使用する手法

**本手法２**： 文末手がかり表現を使用せず，2.1 節の手法で獲得された手がかり表現のみを使用する手法

**本手法３**: 手がかり表現として，文末手がかり表現のみを使用する手法

**Lead(X) 手法**： アナリストレポートの最初から $X$ 文までを，アナリスト予想根拠文とするベースライン手法

評価結果を表 3 に示す.

# 5 考察

本手法 1 の精度は 75.0% であり，良好な精度を達成した．ベースライン手法である Lead 法と比較しても，精度，再現率，ともに高くなっている．Lead 法はアナリストレポートの最初から $X$ 文までをアナリスト予想根拠文とする手法であるが，Lead(10) の精度は 56.36%

[5] アナリストレポートの著作権の関係上，実際に抽出された文に変更を加えてある.

[6] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

表 3: 評価結果

| 手法 | 精度 (%) | 再現率 (%) | F 値 |
|---|---|---|---|
| 本手法 1 | 75.00 | 61.76 | 67.73 |
| 本手法 2 | 77.02 | 55.88 | 64.76 |
| 本手法 3 | 77.94 | 25.98 | 38.97 |
| Lead(10) | 56.36 | 30.39 | 39.48 |
| Lead(20) | 48.63 | 52.45 | 50.46 |

であり，アナリストレポートの最初にアナリスト予想根拠文が出現しているわけではないことが分かる．

文末手がかり表現を使用していない本手法 2 と比べ，本手法 1 は精度を大きく低下させることなく，再現率が向上している．酒井らの手法 [4] で獲得できる手がかり表現は 1 文節で構成されているため,「上回る成長を見込む。」のような複数文節で構成される手がかり表現を獲得できないが，文末に出現する文節に係る文節列と組み合わせることで，複数文節で構成される手がかり表現を獲得することができた．しかし，本手法 3 の結果より，複数文節で構成される文末手がかり表現のみではアナリスト予想根拠文の抽出数が多くはなく，再現率が低い結果となった．

本手法 1 の再現率は 61.76% であり，精度と比べると低く，まだ抽出できないアナリスト予想根拠文が存在している．以下に本手法にて抽出できなかったアナリスト予想根拠文をいつくか示す．

- インバウンド需要は年数億円の寄与と推定する。
- 「ＩＴソリューション」は、流通業案件が伸長し同４％増収となった。

いずれも，手がかり表現として「推定する」や「増収となった。」が獲得されていれば，抽出することができる．手がかり表現を獲得するために使用したアナリストレポートには「推定する」や「増収となった」といった表現は出現しているため，より多くの手がかり表現を獲得できるようにする必要がある．例えば，本手法では手がかり表現を獲得するための初期手がかり表現として「予想する」「考える」「高い」の 3 つを使用したが，この初期手がかり表現の調整や，式 3 における閾値を決定するための定数 $\alpha$ の調整を行い，より多くの手がかり表現を獲得できると考える．

## 6 まとめ

本研究では，投資判断やアナリストレポートの内容を把握するうえで重要な，アナリスト予想の根拠情報を含む文（アナリスト予想根拠文）を自動的に抽出する手法を提案した．本研究により，アナリストレポートから抽出されたアナリスト予想根拠文のみを提示することで，アナリストレポートの内容把握に必要な時間を削減できたり，アナリスト予想根拠文の中からその企業の事業に関連する根拠情報のみを選別することで，その企業への投資判断やそのアナリストレポートを熟読するかどうかを判断するための情報となることが期待できる．本手法は，酒井らの手法 [4] を適用し，アナリストレポートからアナリスト予想根拠文を抽出するのに有効な手がかり表現を獲得した．そして，獲得された手がかり表現等を使用することで，アナリスト予想根拠文の抽出を行い，抽出されたアナリスト予想根拠文の文末から新たな手がかり表現（文末手がかり表現）を獲得することで，より多くのアナリスト予想根拠情報を抽出した．評価の結果，本手法は精度 75.0%，再現率 61.7% であり，良好な精度，再現率を達成した．

## 参考文献


[1] 和泉潔, 後藤卓, 松井藤五郎: 経済テキスト情報を用いた長期的な市場動向推定, 情報処理学会論文誌, Vol. 52, No. 12, pp. 3309–3315 (2011).

[2] 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).

[3] 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).

[4] Sakai, H. and Masuyama, S.: Cause Information Extraction from Financial Articles Concerning Business Performance, *IEICE Trans. Information and Systems*, Vol. E91-D, No. 4, pp. 959–968 (2008).

[5] 酒井浩之, 西沢裕子, 松並祥吾, 坂地泰紀: 企業の決算短信 PDF からの業績要因の抽出, 人工知能学会論文誌, Vol. J98-D, No. 5, pp. 172–182 (2015).

[6] 山本裕樹, 松尾豊: 景気ウォッチャー調査を学習データに用いた金融レポートの指数化, 2016 年度人工知能学会全国大会 (2016).