# 業種別企業業績要因を含む新聞記事の抽出

# Extraction of Newspaper Articles including Cause Informations Concerning Business Performances for each Industry

丸澤 英将 [1*]　和泉 潔 [1]　坂地 泰紀 [1]　田村 浩道 [2]
Hidemasa Maruzawa[1]　Kiyoshi Izumi[1]　Hiroki Sakaji[1]　Hiromichi Tamura[2]

[1] 東京大学大学院工学系研究科
[1] Graduate School of Engineering, The University of Tokyo
[2] 野村證券株式会社
[2] Nomura Securities Co.,Ltd.

**Abstract:** These days, a growing number of individual investors is attracting public attentions even in Japan, and securities companies are actively providing them with investment informations. Especially, analyst reports written by professional security analysts are important investment judgment materials, but their timing of publication varies by brands. In this paper, using the structures of causal relationships of sentences in analyst reports, the ways of security analysts paying attention to cause informations concerning business performances were learned, and newspaper articles including similar causal relationships were extracted. We aim to realize a real-time investment supporting system.

## 1　はじめに

近年，わが国でも個人投資家の増加が注目されており，証券会社も積極的に投資情報の提供を行っている．中でも，市場分析の専門家である証券アナリストが企業業績を予想するアナリストレポートは，重要な投資判断材料だが，発行時期は銘柄ごとにまちまちである．この間欠性を補うシステムとして，日々発行される新聞の記事などのデータから企業業績を変動させる要因になりうる経済イベント（業績要因）を即時に獲得してデータベースに蓄積し，顧客からの問い合わせに自然言語で回答する対話型投資支援システムが考えられる．例えば，ある銘柄の株価の変動要因を尋ねる質問に回答するなどの使い方が想定できる．

上記の目的のため，アナリストレポート中でどのような経済イベントが注目され，企業業績の予想の根拠として用いられているかという因果関係の特徴を学習することで，新聞など別の媒体で報じられている経済イベントから，アナリストの行う企業業績予想を推測することを考える．特定の文の特徴を学習して，別の文章から類似の文を獲得する手法として，単に文全体に含まれる単語の組で一致度を測る bag-of-words 法よるものがあるが，その手法では背後にある因果関係を把握できていないという問題がある．

## 2　提案手法

対話型投資支援システムのための業績要因データベース構築に至るまでの流れを概説する．まず，アナリストレポートの文中で頻出する因果関係の構造を抽出する．次に，因果関係のうちの原因を示す文の部分が指す内容を，企業業績予想の根拠情報として獲得する．その根拠情報と類似する内容を指す文を新聞記事中から探し出し，根拠となりうる経済イベントを取得する．同様の経済イベントから，過去にアナリストがどのような予想を導いたかを参照し，新聞記事中の経済イベントによる企業業績の変化を予想する．このようにして生成した根拠情報，業績予想のリストを，企業業績要因データベースとする．

本稿では，この流れのうち，新聞記事中からアナリストレポートの根拠となりうる経済イベントを取得する段階までを論じる．

*連絡先：東京大学大学院工学系研究科システム創成学専攻
和泉・坂地研究室
〒 113-8654 東京都文京区本郷 7-3-1
E-mail: m2016hmaruzawa@socsim.org

### 2.1 アナリストレポートからの根拠部，予想部の抽出

初めに，アナリストレポート中の因果関係を抽出するために，酒井らのブートストラップ法による手法 [?] を用いた．この手法では，アナリストの予想根拠文を特徴付ける手がかり表現と，手がかり表現に係る節の中で共通して頻繁に出現する共通頻出表現を定義する．最初に少数の手がかり表現と共通頻出表現を与えることで，互いに係り受け関係にある新たな共通頻出表現と手がかり表現が連鎖的に獲得される．

この手法を用いるに当たって，本研究では，特にアナリストの予想を示す文の部分と，その予想の根拠を示す文の部分を分離して抽出する工夫をした．前者を予想部，後者を根拠部と呼ぶこととする．また，各根拠部が指す経済イベントを，根拠情報と呼ぶこととする．アナリストレポート中の文の例を示す．

> 原油安及び探鉱費の増加を主因に，YY.M 期の純利益予想を下方修正した．

この場合,「(を) 主因に,」を根拠部手がかり表現として，それに係る文の部分「原油安及び探鉱費の増加」を根拠部とする．一方,「(を) 下方修正した.」を予想部手がかり表現として，それに係る文の部分を予想部「YY.M 期の純利益予想」として，根拠部とは完全に分離して抽出する．なお，根拠情報は,「原油価格が下がった一方，探鉱にかかるコストが上がった」という経済イベントを指す．

この工夫により，以降に記す根拠部の特徴量抽出の際に，アナリストの予想を示す表現を排除し，予想の根拠として用いられた文の部分のみを対象とすることができる．また，ある根拠からどのような予想が導かれたという因果関係の対応が得られるため，新聞記事中の経済イベントによる企業業績の変化を予想するための学習データとすることができる．なお，企業業績に関する記述を原因・結果表現それぞれに分けて抽出する手法には，坂地ら [?] によるものがあるが，文章パターンの認識が決算短信に特化されている．アナリストレポートではより多様な表現が用いられていたため，本研究では，基礎手法として汎用性の高い酒井ら [?] の手法をもとにした．

### 2.2 根拠情報の業種別特徴の学習

次に，得られた根拠部手がかり表現から，根拠部の特徴を学習する．まず，先に獲得した予想部の手がかり表現と係り受け関係にある文の部分を，根拠部として抽出する．この根拠部を形態素解析し，英単語を除く名詞に分類されるもののうち,「数，接尾，非自立」の下位分類を除いた形態素の組を取得する．この名詞の組を，全根拠部の名詞の組中での tf-idf 値を用いてベクトル化したものを，根拠部の特徴量とする．

すなわち，各組中の名詞について以下の値を計算し，その組の特徴ベクトルの長さが 1 となるように正規化したものを特徴量とする．

$$\frac{n_{i,j}}{\sum_k n_{k,j}} \cdot \left( \log \frac{|\mathrm{D}|}{|d : t_i \in d|} + 1 \right) \tag{1}$$

ここで，$n_{i,j}$ はアナリストレポート中の根拠部 $d_j$ の名詞の組における名詞 $t_i$ の個数，D はアナリストレポート中の根拠部の名詞の組全ての集合である．

ここで，根拠部とその根拠部を抽出したアナリストレポートが言及している銘柄が属する業種の関係に注目する．業種とは，事業内容の銘柄をまとめた分類で，業種別に根拠部の特徴を学習することで，意味のある特徴を獲得できると期待される．

同じ業種に属する銘柄についての根拠部の集合には，似た根拠情報を指す集合が存在すると考えられる．例えば,「鉄鋼・非鉄」業種に属する銘柄の根拠部には，鉄や銅の資源価格に関する根拠情報を指すものが多く含まれると推測される．逆に，似た根拠情報を指す根拠部の集合でも，特定の業種に偏って存在するものと，様々な業種に満遍なく存在するものがあると考えられる．例えば，ベンゼンの資源価格に関する根拠情報を指す根拠部は,「化学」業種に偏って存在するであろうが，為替に関する根拠情報を指す根拠部は，幅広い業種に満遍なく存在すると推測される．

そこで，先に得た根拠部の特徴量を用いて根拠部を多クラス分類し，各クラスの根拠部がどの業種についての根拠部であるかの頻度分布を計算する．根拠部の特徴量には名詞の組を用いているため，この多クラス分類は，根拠部が指す根拠情報をおおよそ多クラス分類していると見なすこともできる．

$$f_{n,m} = |v : v \in (\mathrm{C_n} \cap \mathrm{I_m})| \tag{2}$$

ここで，$f_{n,m}$ はクラス $n$ の根拠部が業種 $m$ についての根拠部である頻度，$v$ は根拠部の特徴ベクトル，C は根拠部の特徴ベクトルを分類したクラスを表す集合 ($n = 1, 2, ...N_\mathrm{C}, N_\mathrm{C}$ : クラスの総数)，I は根拠部の属する業種を表す集合 ($m = 1, 2, ...N_\mathrm{I}, N_\mathrm{I}$ : 業種の総数) である．

さらに，この頻度分布のクラスごとの偏りを，平均情報量を用いて定量化する．

$$-\sum_m f_{n,m} \log_2 f_{n,m} \tag{3}$$

この平均情報量が小さいほど，特定の業種に偏って存在する根拠部が属するクラスであり，平均情報量が

大きいほど，様々な業種に満遍なく存在する根拠部が属するクラスであると言える．

## 2.3 新聞記事からの業種別根拠情報の獲得

前節までで，アナリストレポートから抽出される根拠情報を多クラス分類し，各クラスの根拠情報がどの業種の銘柄の業績予想の根拠となりうるかの頻度分布を得た．以降，これらの各クラスの代表点である重心ベクトルと頻度分布を用いて，新聞記事から新たな根拠情報を獲得する．

まず，新聞記事の文章を表す特徴量を設計する．新聞記事の文章から，アナリストレポートでの根拠部の特徴量を得るために使用した名詞を抽出する．ただし，単に特徴量に使用した名詞に一致する名詞のみを抽出した場合，抽出される根拠情報が限られてしまう問題が生じる．同じ根拠情報を指す場合でも，アナリストレポートと新聞記事ではわずかに異なる表現を用いていることや，学習に用いたアナリストレポートの発行された期間では特定の経済イベントのみが注目され，逆の経済イベントを指す表現が獲得できないことが想定されるからである．前者の例として，株価が上がり続けていたという根拠情報を指すために，アナリストレポートでは「持続的な株価の上昇」，新聞記事では「株価続騰」と表現されることがある．後者の例としては，学習に用いたアナリストレポートの発行された期間では「原油価格の下落」という経済イベントのみが注目されていた場合，同じような銘柄の業績に影響を与えるであろう「原油価格の高騰」という表現が無視されてしまうことがある．そこで，新聞記事の文章中の名詞を，構文上の出現位置の特徴を用いて分散表現を生成する Word2Vec 法 [?] を使用することで，文脈上の類似度の高い名詞まで抽出できるよう拡張する．

こうして抽出した新聞記事の文章中の名詞の組を，アナリストレポートでの根拠部の特徴量を得るために使用した tf-idf 値を用いてベクトル化することで，新聞記事の文章の特徴量とする．

すなわち，各組中の名詞について以下の値を計算し，その組の特徴ベクトルの長さが 1 となるように正規化したものを特徴量とする．

$$\frac{n_{i,l}}{\sum_k n_{k,l}} \cdot \left( \log \frac{|\mathrm{D}|}{|d : t_i \in d|} + 1 \right) \tag{4}$$

ここで，$n_{i,l}$ は新聞記事の文章 $a_l$ の名詞の組における名詞 $t_i$ の個数，D はアナリストレポート中の名詞の組全ての集合である．

ここで，新聞記事のうち根拠情報として獲得するのにふさわしくない一部の記事を除外する．経済記事の中には，新聞社が独自に企業の決算の内容を予想する「観測記事」と呼ばれる記事がある．観測記事は過去の経済イベントを事実として報じるものではなく，将来の予想を伝えるものであるため，マーケットレポートに似た性質の文章である．そのため，観測記事を抽出する正規表現を用いて，対象から除外する．また，実際の企業の決算内容を報じる「決算記事」も，マーケットレポートでの予想部にあたる文章となるため，同じく正規表現を用いて，対象から除外する．

観測記事，決算記事を除いた新聞記事の文章の特徴ベクトルと，根拠情報を分類した各クラスの重心ベクトルとのコサイン類似度を求め，新聞記事の文章と各クラスとの類似度とする．

$$s_{l,n} = v_l \cdot g_n \tag{5}$$

ここで，$s_{l,n}$ は新聞記事の文章 $a_l$ とクラス $\mathrm{C_n}$ との類似度，$v_l$ は新聞記事の文章 $a_l$ の名詞の組の正規化した特徴ベクトル，$g_n$ はクラス $\mathrm{C_n}$ の重心ベクトルを長さ 1 に正規化したベクトルである．

さらに，各クラスとの類似度と，そのクラスの根拠情報がどの業種の銘柄の業績予想の根拠となりうるかの頻度分布との加重平均をとることで，新聞記事の文章がどの業種に属する銘柄の業績予想の根拠となりうるかの指標とする．以下，この指標を新聞記事の文章の各業種への業績寄与度と呼ぶ．

$$c_{l,m} = \sum_n s_{l,n} \ f_{n,m} \tag{6}$$

ここで，$c_{l,n}$ は新聞記事の文章 $a_l$ の業種 $m$ への業績寄与度である．

新聞記事をその文の各業種への業績寄与度を用いることで，特定の期間中の新聞記事のうち，各業種に属する銘柄の業績予想の根拠となりうる重要記事を一覧できることが期待される．ただし，単に業績寄与度の降順で並べると，様々な業種で満遍なく業績寄与度が高い根拠情報が混在してしまうことがある．例えば，日経平均株価の動向などがこれに該当する．この影響を取り除くため，各新聞記事の文章の全業績寄与度中，各業種への業績寄与度の値の偏差値を求める．

$$\begin{aligned} \mathrm{dev(c_{l,m})} = &\quad \frac{c_{l,m} - \mu_l}{\sigma_l} \cdot 10 + 50 \\ \mu_l = &\quad \tfrac{1}{N_\mathrm{I}} \textstyle\sum_m c_{l,m} \\ \sigma_l = &\quad \tfrac{1}{N_\mathrm{I}} \textstyle\sum_m (c_{l,m} - \mu_l)^2 \end{aligned} \tag{7}$$

ここで，$\mathrm{dev(c_{l,m})}$ は新聞記事の文章 $a_l$ の全業績寄与度中，業種 $m$ への業績寄与度の値の偏差値である．

特定の業種についてこの偏差値が高い新聞記事の文章は，その業種への業績寄与度が特徴的に高いことを意味する．一方，この偏差値のみの降順で新聞記事を並び替えると，業績寄与度のみで並べた場合と逆に，い

ずれの業績寄与度もわずかしかないが，その業種への業績寄与度だけが少しだけ高いという新聞記事の文章が混在してしまうことがある．例えば，企業業績とは関わりの薄いスポーツ記事などがこれに該当する．したがって，各業種への業績寄与度とその値の偏差値の調和平均をとったものを，新たに業績関連度指数と定義する．ただし，調和平均を求める際には，各業種への業績寄与度の値は平均 50，標準偏差 10 に正規化する．

$$r_{l,m} = \frac{2\ c'_{l,m}\ \mathrm{dev}(\mathrm{c'_{l,m}})}{c'_{l,m} + \mathrm{dev}(\mathrm{c'_{l,m}})} \tag{8}$$

ここで，$r_{l,m}$ は新聞記事の文章 $a_l$ の業種 $m$ への業績関連度指数，$c'_{l,n}$ は新聞記事の文章 $a_l$ の業種 $m$ への業績寄与度を平均 50，標準偏差 10 に正規化した値である．

新聞記事をその文の各業種への業績関連度指数の降順で並べることで，特定の期間中の新聞記事のうち，各業種に属する銘柄の業績予想の根拠となりうる重要記事を一覧できる．

# 3 実験設定

## 3.1 アナリストレポートからの根拠部，予想部の抽出

アナリストレポートからの根拠部，予想部の抽出に当たって，初期表現を以下のように選定した．初期の手がかり表現には，名詞の後に動詞「する」が続く熟語動詞の出現頻度上位 20 位の中から，根拠・予想を示す箇所に高確率で用いられる表現を選んだ．また，初期の共通頻出表現には，熟語の出現頻度上位 20 位の中から，根拠・予想を示す箇所に高確率で用いられる表現を選んだ．特に，根拠部の初期の共通頻出表現には，ポジティブ・ネガティブの判断を含む表現を用いるようにした．

選定の結果，根拠部の抽出では，初期の手がかり表現，共通頻出表現にそれぞれ「考慮し，反映し，評価し」,「増益，改善，成長」を用いた．予想部の抽出では，初期の手がかり表現，共通頻出表現にそれぞれ「継続する，予想する」,「利益，業績，売上」を用いた．学習データには，野村證券株式会社の Global Markets Research レポート（2013 年下半期発行分，日本株 216 銘柄の表紙部分）を用いた．

## 3.2 根拠情報の業種別特徴の学習

係り受け解析器として CaboCha [?]，形態素解析器に MeCab [1]を使用した．多クラス分類には，k-means 法を用い，k=100 とした．銘柄を分類する業種には,「野村 19 業種分類」（化学，鉄鋼・非鉄，機械，自動車，電機・精密，医薬・ヘルスケア，食品，家庭用品，商社，小売り，サービス，ソフトウェア，メディア，通信，建設，住宅・不動産，運輸，公益，金融）を用いた．

[1]http://taku910.github.io/mecab/

## 3.3 新聞記事からの業種別根拠情報の獲得

新聞記事には，日経新聞の 2014 年の記事（スポーツ記事など，経済記事以外も含む）119,767 件を用いた．Word2Vec 法のモデルには，ロイター社の 2003 年から 2013 年の経済記事の文章をコーパスとし，200 次元で分散表現を生成するよう学習したものを用いた．文脈上近い意味の名詞とみなす類似度の閾値には，0.7 を使用した．観測記事を抽出する正規表現には，見出しと最初の 1 文について，1ヶ月分を人手で仕分けることで得た以下の表現を用いた．

「.*決算予想.* 」または
「 (利益 | 損益).*(る見通しだ | たようだ | たもようだ | になりそうだ | になった公算が大きい). 」

同じく，決算記事を抽出する正規表現には，人手で得た以下の表現を用いた．

「決算 (を | で).*発表.*た.」　または
「(利益 | 損益).*(円 | ドル).*(発表 | だった).*.」または
「(利益 | 損益).*(% | 倍に)(増えた | 減った | 増加した | 減少した | 増となった | 減となった | 上回った | 下回った). 」または
「(利益 | 損益).*(最高 | 最低 | 黒字 | 赤字).*た.」または
「(円 | ドル)(から | に).*(上げた | 下げた | 修正した). 」

抽出した重要記事について，トップ 30 における精度を求めた．比較対象には，因果関係の構造に注目して根拠部を分離することをせず，単にアナリストレポートの各文全体から特徴量となる名詞を抜き出して特徴量に用いた bag-of-words 法による，トップ 30 における精度を用いた．精度の算出に当たっては，重要記事抽出対象の時期のアナリストレポート中で，5 つ以上の銘柄においてアナリストによる業績予想の根拠とされていた概念を正解とし，人手で評価した．

# 4 結果・考察

## 4.1 アナリストレポートからの根拠部，予想部の抽出

根拠部，予想部それぞれで新たに抽出された共通頻出表現，手がかり表現の一部を示す．

根拠部共通頻出表現
　下落，上振れ，好調さ，減速，悪化
根拠部手がかり表現
　織り込んで，踏まえ，主因に
予想部共通頻出表現
　成長，増収，増益，採算改善，コスト削減
予想部手がかり表現
　見込まれる，期待される，続こう

根拠部，予想部ともに，それぞれの特徴を捉えた共通頻出表現，手がかり表現が新たに抽出できている．特に，根拠部の共通頻出表現は，初期表現にはポジティブな意味を持つ単語のみを与えたにも関わらず，対応するネガティブな表現も抽出できている．一方，予想部の共通頻出表現は，ポジティブな表現が大半を占める結果となった．これは，学習の対象としたアナリストレポートの発行時期が，いわゆるアベノミクスによる市場全体の回復期にあたるため，アナリストの予想がポジティブな方向に偏っていたためと考えられる．

## 4.2 根拠情報の業種別特徴の学習

根拠部の手がかり表現をもとに，6,655 文の根拠部を得，この根拠部から，特徴量に用いる名詞を 2,968 個得た．根拠部を多クラス分類した結果，1 つのクラスに属する根拠部の数は最低 12 個，最大 1,492 個となった．平均情報量上位 10 クラス，すなわち，様々な業種に満遍なく存在する根拠部が属するクラスについて，その重心ベクトルを組成する代表的な名詞の組を記す．

海外の事業環境の好調さ，人件費の増加など，幅広い銘柄において業績に影響を与える根拠情報を示唆しているクラスが並んでいる．

平均情報量下位 10 クラス，すなわち，特定の業種に偏って存在する根拠部が属するクラスについて，その重心ベクトルを組成する代表的な名詞の組と，そのクラスが偏在する業種を記す．

資源価格や為替前提の変更などは，特定の業種の銘柄の業績にのみ影響を与えやすい根拠情報といえる．一方，決算に関する根拠情報などは，後述するように，学習データに偏りがあるために不適切な特徴が獲得されてしまっている．また，金融業種が特徴的なクラスを構成する傾向があることが分かる．

表 1: 平均情報量上位 10 クラス．

| クラスの重心ベクトルの組成 |
|---|
| 販売 法人税 低下 月 減収 |
| 事業 環境 航空 海外 好調 |
| 業績 好調 予想 足元 月 |
| 増加 費用 人件費 経費 研究開発費 |
| 継続 判断 投資 レーティング 決算 |
| 効果 これら 買収 税 連結 |
| 可能 業績 開発 達成 良 |
| 影響 ジェネリック 円高 為替 薬価改定 |
| 成長 中期 利益 期待 事業 |
| 計画 会社 中期 営業利益 営業増益 |

表 2: 平均情報量下位 10 クラス．

| クラスの重心ベクトルの組成 | 偏在する業種 |
|---|---|
| 発表 取材 決算 材料 期待 | 金融 |
| 前提 為替 変更 下期 判断 | 機械 |
| 効率 系列 資本 高水準 比較 | 小売り |
| シナジー 統合 経費削減 システム | 金融 |
| 銅 鉱山 金属 価格 下落 | 鉄鋼・非鉄 |
| 上昇 低下 基準 予想 不具合 | 公益 |
| セクターバリュエーション 低下 上昇 | 機械 |
| ヒアリング 決算 継続 月 急激 | 化学 |
| 投信 解約 外債 販売 回り | 金融 |
| 自動車保険 収支 改善 等級 制度 | 金融 |

## 4.3 新聞記事からの業種別根拠情報の獲得

新聞記事をその文の各業種への業績関連度指数の降順で並べた重要記事のうち，化学業種についての 2014 年 1 月〜3 月の重要記事の見出しを記す．

大王紙――高値圏，海外事業が拡大
ナイロン原料が下落，アジア価格，中国で荷余り感.
苦戦の中国，１５年黒字へ，花王，現地大手の販売網活用，設備投資，アジア中心
ベンゼン，３月４％安，アジア向け，中国で需給緩和.
ベンゼン上昇，１０ヵ月ぶり高値，1月アジア向け.
ベンゼン価格，３ヵ月ぶり下落，アジア向け.
配合飼料３期ぶり上げ，４～６月価格，農家向け，原料高で.
古紙，輸出価格が下落，４月積み，需要鈍り９～１０％安.
鉄スクラップ値下がり，市中価格，アジア需要伸びず，鋼材の値上げに弱材料.
価格差を読む（1）発電燃料と電力―安い石炭利幅大きく，重油は採算割れ水準.

繊維，化学原料の資源価格の動向など，化学業種に属する銘柄の業績に関係があると考えられる根拠情報を含む記事が並んでいる．本手法によるトップ 30 における精度は 70 %，bag-of-words 法によるトップ 30 における精度は 57 %だった．この業種は，因果関係の構造に注目して根拠部を分離したことで，特徴量の学習が適切に行われたといえる．

次に，自動車業種についての 2014 年 1 月～3 月の重要記事の見出しを記す.

円高，一時１００円台，ＮＹ市場.
為替市場，円高に備え，円買いオプション，半年ぶり高水準.
円相場の膠着続く，米景気期待・中国不安が綱引き.
円売り持ち高解消一服，投機筋，円高圧力和らぐ.
円高，じわり天井感，海外投機筋に実需の壁
円高１０３円台，東京市場.
円横ばい，１０２円３３～３４銭
投機筋，円売り膨らむ，６年半ぶり水準，日米の金利差拡大.
ＮＹ市場，円一時１０３円台.
為替――円安基調が一服か

主に為替に関する根拠情報が並んでいる．これは，自動車業種に属する銘柄のアナリストレポートでは，為替が主要な根拠情報として用らいれているためと考えられる．特に，今回学習の対象としたアナリストレポートの発行時期が，いわゆるアベノミクスによる急激な円安の進行が注目されていた時期にあたるため，為替要因に偏って言及されており，その特徴を強くしたものと推測される．このように，新聞記事中に直接自動車に関係する名詞が無い場合でも根拠情報として獲得できることが，因果関係を考慮した本手法の強みと言える．本手法によるトップ 30 における精度は 100 %，bag-of-words 法によるトップ 30 における精度は 83 %だった．一方，為替のみで業績予想が行われるのはこの時期に特異な現象であることも想定される．すなわち，精度が高くても再現率が低い可能性がある．時期によらない根拠情報を獲得するためには，学習対象の期間を拡大する，アナリストレポート以外の時期によらない根拠情報を含む学習データと組み合わせるなどの工夫が必要である.

最後に，家庭用品業種についての 2014 年 1 月～3 月の重要記事の見出しを記す.

デル，新興国で積極投資，中国など，ソフト販売を拡大.
ピジョン，ブラジル進出，育児用品，新興国投資を倍増，品質・ブランド生かす.
香港・味千中国――高値圏，和風ラーメン回復（アジア新興国ＮＯＷ）
新興国――中国の景気動向に注目
中国，世界から投資１月１６％増.
新興国――中国の製造業景況感に注目
欧州企業，鈍い収益回復，１０～１２月，新興国景気の減速響く.
中国蒙牛乳業――成長期待で戻り歩調に
Ｇ２０閉幕，成長持続へ危機感共有――新興国の不満，抑え込む，先進国，中国は改革
香港株――改革にらみ神経質な展開

新興国に関する雑多な記事が混在しており，十分に適切な根拠情報が取得できていない．本手法によるトップ 30 における精度は 53 %，bag-of-words 法によるトップ 30 における精度は 63 %だった．(例に挙げた 3 業種についてのトップ 30 における精度を表**にまとめる.) これは，学習の対象としたアナリストレポートにおいて，家庭用品業種に属する銘柄の業績予想には，主に新興国における当該銘柄の商品の販売動向が言及されており，このうち新興国を表す名詞がこの業種の特徴として過学習されてしまったためと推測される．また，一部の業種では，そもそも学習データが少ないために意味上のまとまりのある特徴を学習できていない場合もあった．このような学習の偏りや不足を防ぐためには，根拠情報として用いられなかった過去の新聞記事を負例として学習に用いる，決算短信など根拠情報を得られる他のデータを学習の対象に組み入れるなどの工夫が必要である.

また，本稿の段階では，膨大な数の新聞記事が各業種の根拠情報のとして適切であるかという正解データが未整備のため，得られた結果に対して十分定量的な評価ができていない．専門家の監修による正解データ

表 3: 業種別重要記事のトップ 30 における精度 (%).

| 手法 | 化学 | 自動車 | 家庭用品 |
|---|---|---|---|
| 本手法 | 70 | 100 | 53 |
| bag-of-words 法 | 57 | 83 | 63 |

の作成とそれを用いた結果の再現率，F 値などによるさらなる定量評価は，今後の課題とする．

## 5 まとめ

アナリストレポートの文章の因果関係の構造から，証券アナリストが企業業績要因として注目する情報を業種別に学習し，新聞から同様の因果関係を含む記事を抽出することで，即時性の高い投資支援の実現を目指す構想を提示した．その初段として，業種別に抽出した重要記事に対する簡単な評価・考察を行った．

結果，一部の業種では因果関係を学習していないと抽出できない記事を適切に獲得できたが，学習データの期間的な偏りや内容面での偏り，量の不足により，不適切な結果となった業種もあった．学習データの質・量の向上，結果のさらなる定量評価が今後の主な課題である．

## 参考文献

[1] 酒井浩之, 柴田宏樹, 平松賢士, 坂地泰紀 : 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, 人工知能学会第 *17*回金融情報学研究会, pp. 25–30 (2016)

[2] 坂地泰紀, 酒井浩之, 増山繁 : 決算短信 PDF からの原因・結果表現の抽出, 電子情報通信学会論文誌 *D*, Vol J98-D, No. 5, pp. 811–822 (2015)

[3] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean : Distributed Representations of Words and Phrases and their Compositionality, *NIPS 2013*, pp. 3111–3119 (2013)

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