テキストマイニングによる金融レポートの自動生成支援

Generation Support of Financial Reports by Textmining

丸澤 英将[1]　和泉 潔[1]　坂地 泰紀[1*]　田村 浩道[2]
本廣 守[2]
Hidemasa Maruzawa[1]　Kiyoshi Izumi[1]　Hiroki Sakaji[1]　Hiromichi Tamura[2]　Mamoru Motohiro[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: Recently, with the increase of individual investors, the necessity of investment support technologies is increasing. Although analyst reports on which professional securities analysts forecast business performances or stock prices of companies are regarded as important investment decision materials, writing an analyst report is heavy burden. In this research, we summarize newspaper articles and support the generation of analyst reports by using knowledge of information features which are referred to as reasons for analysts' forecasts of business performances or stock prices in analyst reports.

1　はじめに

近年，我が国でも証券市場における個人投資家の比重が増大しており，個人投資家の投資判断を支援する技術の必要性が高まっている．個人投資家が重視する投資判断材料の一つに，証券会社が発行するアナリストレポートがある．アナリストレポートとは，証券市場調査・分析の専門家である証券アナリストが，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，企業の業績や株価に対する証券アナリストの予想が示され，その根拠として，その企業の取り組む事業の近況・財務状況（企業のファンダメンタルズ）や事業に影響を与える経済・政治・社会状況（マクロ経済のファンダメンタルズ）などが言及される．

これら根拠として言及される情報は，証券アナリストの独自の調査によるものも含まれるが，規模の大きい企業のファンダメンタルズやマクロ経済のファンダメンタルズは，新聞などの媒体でも報じられるものである．ただし，媒体で報じられる様々な経済情報の中で，どの情報が企業の業績や株価に影響を与えるものであるかを見極めるには，証券アナリストの高度な専門知識を必要とする．アナリストレポートを参考にする個人投資家にとっては，企業の業績や株価に対する証券アナリストの予想だけでなく，証券アナリストがどのような情報を根拠として重視することで，その予想を導き出したのかという見極めの観点が重要である．例えば，酒井らは株式市場における次のような現象に注目している [1]．2012 年度上期のパナソニックの連結業績の発表では，前年同期比で売上高は減少したが，営業利益は増加したという内容であった．しかし，社長は「今回の大幅な業績の下ぶれの根本的な原因は，本業の不振にある」と語った．この発言が嫌気され，パナソニックの 2012 年 11 月 1 日の株価はストップ安となった．このように，株価を予想するためには，場合によっては決算発表中の営業利益の値ではなく，本業に関する社長発言を重視すべきという判断は，専門的な知見を必要とするものといえる．

そのため，アナリストレポート中で業績・株価予想の根拠として言及される情報の特徴を捉え，新聞記事などの媒体から証券アナリストが注目するであろう情報に絞って自動的に要約する技術が重要である．この技術は，特に次のような点で有用である．アナリストレポートを作成してきた証券アナリストにとっては，レポート作成の支援に活用できる．執筆経験者によると，決算発表の時期には多くのレポート発行が集中し，膨大な経済情報の中から，企業の業績や株価の変動を引き起こす要因について人手で整理する作業は負担が大きいという．本技術により有用な経済情報を要約し，情報整理の作業時間が短縮できれば，より独自性の高い調査・分析に注力したり，執筆するレポートの数を増

*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: staff@socsim.org

やしたりすることができる．アナリストレポートを参照してきた個人投資家にとっても，投資判断材料の充実に繋がる．アナリストレポートが発行される頻度は銘柄[1]によって大きく異なり，注目度の高い銘柄は四半期ごとの決算発表に合わせて度々発行されるが，発行されることが少ない銘柄もある．証券市場の上場企業数が東京証券取引所だけでも 3500 社近くに上る中で，特に個人投資家は保有する銘柄の選定理由が個々人で異なっており，各人が注目している銘柄のアナリストレポートが必ずしも頻繁に発行されているとは限らない．そこで，本技術により証券アナリストの執筆作業が効率化され，発行されるレポートの銘柄数・頻度が増えたり，個々のレポートの質がより高まったりすれば，個人投資家の投資判断材料を充実させることができる．

この目的に応用可能な技術として，近年，自然言語処理やテキストマイニング技術の進展により，テキストデータから自動的に重要な情報を抽出する技術が発達してきている [2, 3]．しかし，これらの要約技術は，そのままでは事象の背景にある因果関係を考慮できない．一方，文の因果関係の構造に注目し，原因表現を取り出す手法も提案され始めている [4]．丸澤らは，この文の原因表現を取り出す技術を応用し，アナリストレポート中で業績・株価予想の根拠として言及される情報の特徴を学習した知識を，重要情報抽出技術に組み込むことで，業績変動要因文を新聞記事から抽出している [5]．ただし，抽出した文がアナリストレポートの執筆に有用であるかの実務家による評価は行われていない．本研究では，丸澤らの手法を用いて，新聞記事からアナリストレポート執筆に有用な要約文を自動生成するシステムを構築し，実務家の評価により実用性の検証を行う．

2　アナリストレポート執筆支援文自動生成の全体の流れ

アナリストレポート執筆支援文の自動生成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，因果関係のうち原因表現を，証券アナリストによる企業業績・株価予想の根拠情報[2]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，根拠情報となり得る企業業績要因を取得する．このようにして取得した企業業績要因をまとめ，アナリストレポート執筆支援文を自動生成する．全体の流れを図 1 に示す．

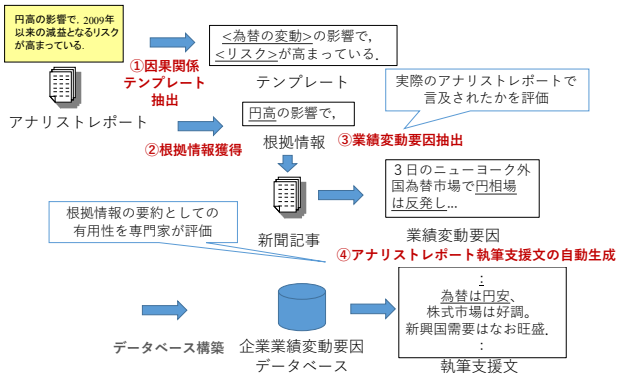

図 1: アナリストレポート執筆支援文自動生成の全体の流れ

本システムの実装例として，Web サーバー上のシステムとして実装したものの動作画面を図 2，3 に示す．

システムには，一定期間のアナリストレポートのテキストデータを与え，予め根拠情報の特徴を学習させておく．そして，データベース中に要約対象となる新聞記事のテキストデータを格納し，図 2 のように，要約対象として用いる新聞記事の期間，注目する銘柄の業種，自動生成文の並べ方を入力として指定し，Search ボタンを押す．システムは指定された期間（図の例では 2014 年 1 月〜3 月）の新聞記事のテキストデータを参照し，指定された業種についての学習した根拠情報と類似する内容を指す文を探し出す．それらの文のうち重要度の高いものを，図 3 のように，指定された並べ方によりいくつか並べて提示する．各文には，抽出元の記事の発行日と見出しを併記する．並べ方には，記事の発行日の時系列順，後述する業績関連度指数順のいずれかを用いる．提示された文を参考にして，証券アナリストがレポートを執筆する，という用途を想定する．

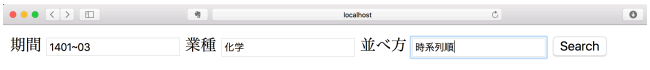

図 2: 提案手法によるシステムの入力画面

[1]証券会社に上場されている株式の企業名

[2]本稿では，酒井ら [6] に倣い，アナリストレポートの中で企業の業績や株価の変動を引き起こす要因として言及されている情報を，「（アナリスト予想の）根拠情報」と呼ぶ．また，一般に，企業の業績に影響を与える要因を「業績変動要因」と呼ぶ．

図 3: 提案手法によるシステムの出力画面

*原油安及び探鉱費の増加*__を主因に,__
（根拠部）　**（根拠部手がかり表現）**

*YY.M 期の純利益予想*__を下方修正した.__
（予想部）　**（予想部手がかり表現）**

図 4: アナリストレポート中の文の例

3　新聞記事からの根拠情報となり得る業績変動要因の取得

本システムに用いる，根拠情報となり得る業績変動要因の新聞記事からの取得には，丸澤らの手法 [5] を用いる．本節に，その手法を簡単に述べる．

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中の因果関係の抽出には，酒井らのブートストラップ法による手法 [4] を用いる．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，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．前者を予想部，後者を根拠部と呼ぶ．アナリストレポート中の文の例を図 4 に示す．

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上に，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．

3.2　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，根拠部がどのような業績変動要因を指し示しているかという意味的な特徴を学習する．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次の値を計算し，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．

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

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．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．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．

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次の式のように計算する．

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

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，次の式のように平均情報量 e を用いて定量化する．

$$e = -\sum_m f_{n,m} \log_2 f_{n,m} \qquad (3)$$

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．

3.3 新聞記事からの業種別根拠情報の獲得

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．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 [7] 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次の式の値を計算し，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．

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

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．

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「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次の式で求め，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．また，$\cdot$ は内積演算子である．

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得るかの頻度分布との加重平均を次の式のように計算することで，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得るかの指標とする．

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

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

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を防ぐため，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次の式で求める．

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

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，N_I は業種の総数である．

最後に，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を防ぐため，各業種への業績寄与度とその値の偏差値の調和平均を次の式のようにとったものを，業績関連度指数と定義する．

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

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，$c'_{l,m}$ 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．

4 アナリストレポート執筆支援文の自動生成

前節の手法により，各業種に属する銘柄の業績予想の根拠となり得る重要記事（業績関連度指数の高い記事）を取得できる．これらを，特定の時期・業種について抽出し，発行日・見出しを付して任意の数並べることで，アナリストレポートの執筆支援文として提示する．

手法全体の適用手順の具体例を示す．学習対象のアナリストレポートにおいて，化学業種に属する銘柄では,「原油安を主因に，純利益予想を下方修正した.」や「原油価格の持続的な下落により，純利益の減少が見込まれる.」のように，原油安が主な根拠情報として言及されることが特異的に多かったとする．この場合，まず 3.2 節の処理により,「原油安」や「原油価格の持続的な下落」というテキストが根拠部として抽出される．

そして，3.2 節の処理により，これらの根拠部を特徴ベクトルで表現して比較することで，これらが同じ原

油安について言及している類似するテキストであり，ある業種に偏って存在するテキスト群（クラス）であることが定量化される．

このように学習した情報を用いて，3.3 節の手法で新聞記事中の文を検索する．新聞記事の中に「原油価格続落」を報じる文があった場合，この文が「原油安」や「原油価格の持続的な下落」と類似したテキストであり，化学業種の根拠情報と関連が高い業績変動要因であることが計算される．この計算を特定期間の新聞記事の全文に渡って行うことで，その期間中の特定業種の根拠情報となり得る業績変動要因を含む記事を一覧できる．

さらに，本節のように，発行日・見出しを付して任意の数並べることで，化学業種について，表 1 のような文の列を提示することができる（表中の記事は架空のものである）．

表 1: アナリストレポート執筆支援文の自動生成例

文 (発行日 見出し)
会合後，原油先物は最安値を更新した．(2015/4/xx OPEC 会合，減産合意ならず)
国際原油相場の下落は xヶ月連続．(2015/4/xx 国際原油相場，xヶ月連続の下落)
原油価格の暴落が止まらない．(2015/5/xx 商品先物市場，総じて低調)

このような文の列を直近の新聞記事から自動生成することで，化学業種に属する銘柄についてのアナリストレポートを執筆する証券アナリストに対して，化学業種に属する銘柄の主要な業績変動要因である原油安についての要約文を提示し，執筆の支援を行うことができる．

5 評価

本手法で抽出した文の列について，表 2 の条件で，証券会社の実務家 3 名に評価を依頼した．

学習データには，野村證券株式会社の Global Markets Research レポート（2014 年下半期発行分，日本株 216 銘柄の表紙部分）を用いた．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．新聞記事には，日本経済新聞本紙の朝・夕刊の地方面を除く 2015 年度の記事（スポーツ記事など，経済記事以外も含む）を用いた．

アナリストレポートからの根拠部の抽出では，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．形態素解析器に MeCab[3]を，係り受け解析器に CaboCha[8] を使用した．多クラス分類には，k-means 法を用い，k=100 とし，実装に Python ライブラリ scikit-learn 0.19.1[4]を利用した．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・決算記事を抽出する正規表現には，丸澤ら [5] と同じものを用いた．

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名詞を抜き出してそれらの tf-idf 値を特徴量に用いた単純 bag-of-words（BOW）法を用いた．

評価基準は表 3 のように指定した（評価には専門的な判断が求められるため，負担を考慮し，全業種ではなく無作為に選んだ 5 業種のみの評価とした）．

表 2: 有用性評価実験の条件

期間	2015 年 4 月〜2016 年 3 月中の各四半期の計 4 期間
業種	無作為に選んだ 5 業種（化学，自動車，小売り，食品，住宅・不動産）
並べ方	時系列
抽出文の提示方法	各期間・業種の抽出文 5 つずつを，どちらが提案手法によるものかを伏せて比較手法によるものと並べて提示
参考情報の提示方法	当該期間・業種の実際のアナリストレポートの概要を事前に提示し，根拠情報の参考としてもらう

表 3: 有用性評価実験の評価基準

評価	当該期間における当該セクターの企業の業績に影響を与えそうな情報の抽出について，以下のいずれに当てはまるか
○	よく抽出できている．
△	一部含むが、要点にずれがある．
×	ほとんど含まない．

[3]http://taku910.github.io/mecab/
[4]http://scikit-learn.org/

6　結果

まず，3 名の実務家による評価の一致を測るため，Cohen の κ 値を 2 名の組み合わせごとに求めた．結果を表 4 に示す．

表 4: 有用性評価結果の Cohen の κ 値

評価者 A & 評価者 B	0.54
評価者 B & 評価者 C	0.72
評価者 C & 評価者 A	0.38
平均値	0.55

今回の評価結果は一致度が中程度と言える．評価がややばらついた原因は，何をアナリストレポートの根拠情報とするべきかは実務家でも判断が分かれることがある専門的な事項であり，一部の文で評価が 1 名は○，1 名は△，1 名は×と分かれたケースが見られたことが挙げられる（実際のアナリストレポートの執筆は，業種ごとに専門の担当者が行うが，今回は 3 名の評価者に 5 業種全てについてそれぞれ評価してもらった）．そのため，評価の多数決ではなく，評価○を 1 点，△を 0.5 点，×を 0 点と点数化して 3 名の評価の平均点を算出し，すべての文が○と評価された場合の点数を 100%として集計した．

業種ごとの集計結果を図 5 に示す．(図中の業種は，比較手法である単純 BOW 法での精度の降順に並べた．)

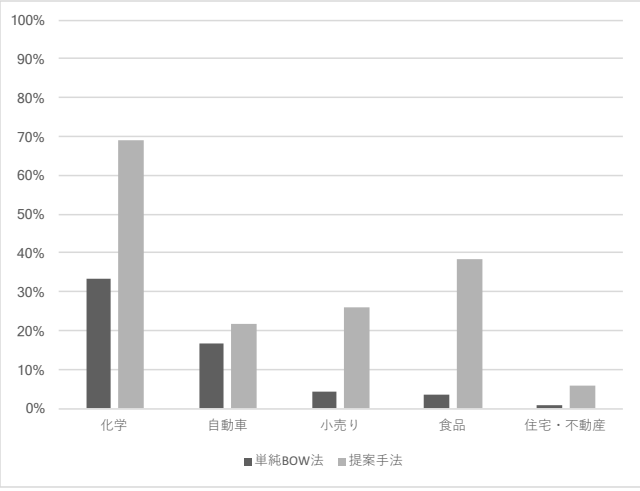

図 5: アナリストレポート執筆支援文の有用性評価による精度

比較手法では，化学業種での精度が 30%を超えたのみで，自動車業種では 20%，小売り，食品，住宅・不動産業種では 10%を下回る精度だった．一方，提案手法では，化学業種で 70%近い精度を達成したほか，食品業種で大幅に精度が改善した．自動車，小売り，住宅・不動産業種での精度も，いずれも比較手法のものを上回った．

比較手法での精度の低さは，アナリスト予想の根拠情報になり得る業績変動要因を新聞記事から抽出するという今回の問題設定に対しては，単なるキーワードマッチングでは，実務家による有用性評価に耐え得る精度に至らないことを示していると言える．

7　考察

提案手法によって良好な精度を達成した化学業種での抽出文例の比較を表 5 に示す．

表 5: 自動抽出文例（化学業種）

提案手法	石油製品の取引価格がアジアで軒並み下落している．
比較手法	各地の寺社などに油のような液体がまかれた事件は，警察庁によると 14 府県 43ヶ所（1 日現在）に被害が広がった．

いずれの文も「油」という，化学業種に属する銘柄のアナリストレポートの文中に出現するキーワードを含んでいる．原油・石油・灯油などの原料・加工品の需給動向は，化学業種に属する銘柄についての代表的な根拠情報である．しかし，比較手法で抽出された文中での油は，これらの需給動向とは関係が薄い．それに対し，提案手法では「取引価格」の「下落」という需給動向を表す名詞や,「アジア」という当該時期に主要な消費地としてアナリストレポートの文中で度々言及されていた名詞を含んだ文を抽出できている．これは，提案手法において，アナリストレポートの文中から原因表現の部分だけを取り出して特徴量を設計しているため，これらの根拠情報のキーワードを重視した新聞記事からの文抽出ができることの効果が現れているものと考えられる．

次に，提案手法によって大幅な精度の改善が見られた食品業種での抽出文例の比較を表 6 に示す．

表 6: 自動抽出文例（食品業種）

提案手法	伊藤忠商事はチョコレート原料の加工・販売事業に参画する．
比較手法	ショットピーニングは微小な金属粒を材料の表面にぶつけて耐久性を高める加工技術．

いずれの文も「原料・材料」や「加工」という，食品業種に属する銘柄のアナリストレポートの文中に出現するキーワードを含んでいる．食品の原材料の加工

事業や技術の動向は，食品業種に属する銘柄についての代表的な根拠情報である．しかし，比較手法で抽出された文中での材料加工は，食品ではなく金属についてである．それに対し，提案手法では食品の加工・販売事業についての文を抽出できている．これは，化学業種での考察と同様に,「販売」や「事業」への「参画」という根拠情報として同時に出現することが多いキーワードを重視できているためと考えられる．なお，食品業種では化学業種に比べて，提案手法・比較手法いずれの精度も低い．いずれも原料について言及されることが多い業種だが，化学業種では原油など限られた種類の原料が繰り返し言及されるのに対して，食品業種では多様な原料が言及され，キーワードとして学習することがより困難であることが一因と考えられる．

また，提案手法で抽出された文が，企業自体は食品業種に分類されていない「伊藤忠商事」の事業動向についてであることに注目されたい．新聞記事中に出現する企業名や企業との関連性が高いキーワードを基に，抽出文が関連する業種を特定することも可能だが，その場合，この文は「伊藤忠商事」という企業名を基に商社業種に分類されることになる．アナリストレポートの根拠情報として言及されるのは，当該業種の動向に限られず，原料の採掘・輸送・加工，製品の輸送・販売・宣伝，サービスの提供など一連のサプライチェーンで関わる業種全般の動向に及ぶ．この点を考慮した情報抽出が可能な点も，因果関係に注目する提案手法の特徴と言える．さらに，提案手法で算出する業績関連度指数は，業種ごとに関連度を連続値で評価するため，文が関連する先を単一の業種に限定せず，業績へ影響が及ぶ全業種について関連度を見ることが出来る．

最後に，提案手法・比較手法いずれも精度が低かった住宅・不動産，自動車業種について，抽出文例の比較をそれぞれ表 7，表 8 に示す．

表 7: 自動抽出文例（住宅・不動産業種）

提案手法	安倍晋三内閣は法人税や消費税の見直しを進めてきたが，所得税の抜本改革は放置してきた．
比較手法	財務省が２８日に発表した２０１４年～４月，１５年３月の税収実績は前年同期比１２・３％増と高い伸びになった．

住宅・不動産，自動車業種はそれぞれ税，外国為替の話題が多く抽出され，そのほとんどが実務家評価で根拠情報とは直接は関係ない文と判定された．表に掲載した提案手法による抽出文は，その中でも，実務家による評価が１名は○，１名は△，１名は×と分かれたもので，これらの業種の中では比較的関係ありと評価された文である．

表 8: 自動抽出文例（自動車業種）

提案手法	シティグループ証券は２０１７年の円ドル相場を年平均１ドル＝１２９円と予想していますが，２０年には１１６円まで円高になると予想しています．
比較手法	ドル／円１ドル＝１１２．１７～１１２．２０円（７０銭の円高）【中略（ユーロ／円，ユーロ／ドル相場の値)】（東京市場１２時時点）．

住宅・不動産業種では，この時期には消費税や相続税法改正の影響が根拠情報として盛んに議論されていた．不動産は高額な消費・相続の対象であるため，税の影響は大きい．そのため，税制・税収の話題が多く抽出されたと考えられる．政府の税収は関係が薄いが，税制改革の動向は根拠情報になり得るということである．

一方，自動車業種では，前章で述べたのと同様にこの時期も外国為替動向が根拠情報として大きく注目されていた．日々の具体的な相場の値を報じる文は関係が薄いが，長期の外国為替動向の予想は根拠情報になり得るということである．今回の提案手法では名詞のみを特徴量に用いているため，外国為替についての文を抽出することはできても，その中で，言及期間の長短や過去の事実と将来の予想の区別を付けることは困難である．このことから，新聞記事の文自体に，今回アナリストレポートの文に用いたような文構造解析を行えば，これらを区別し，より精度を向上できる可能性がある．

全般的に，小売り，食品，住宅・不動産のように消費動向が主な根拠情報となる業種では，キーワードに一般的な名詞が混在してしまうため，社会面の記事など業績変動要因の少ない文を誤って抽出してしまうことが多かった．この問題は，今回の提案手法では根拠情報を含む文から作った特徴量という正例のみを学習していることに一因があると考えられ，根拠情報を含まない文から作った特徴量を負例として与えることで改善する可能性がある．

8　まとめ

本研究では，証券アナリストの業務支援やそれによる個人投資家の投資判断材料の充実のため，アナリストレポートの中で企業の業績変動要因として言及されている根拠情報を学習し，根拠情報となり得る企業の業績変動要因を新聞記事から抽出することで，アナリストレポートの執筆を支援する要約文を自動生成する手法を提案した．アナリストレポートからの根拠情報

の学習では，文の係り受け解析を行い，背後にある因果関係に注目した処理を行った．また，業種別に根拠情報を学習・分類することで，業種ごとの企業業績への寄与度を定量化した．これにより，企業の業績変動要因を新聞記事から期間・業種別に抽出することが出来た．

本提案手法は，証券会社の実務家による評価で，実験を行った全業種について比較手法の単純 BOW 法より高い有用性を示した．これは，提案手法が文の因果関係の構造を反映した学習を行っており，因果関係の把握が重要な経済分野に適したテキストマイニング手法であるためと考えられる．一方，一部の業種での取得精度には課題が残り，文抽出の有用性においても改善の余地がある．根拠情報を含まない文を負例として与えて学習することなどで精度を改善し，新聞記事の即時的な取得や実際のアナリストレポートに近い形式での出力を行うことで，より実用的なアナリストレポート執筆支援文の自動生成システムを構築できることが期待される．

参考文献

[1] 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. 30, No. 1, pp. 172–182 (2015)

[2] Otterbacher, J., Erkan, G., and Radev, D. R.: Using random walks for question-focused sentence retrieval, *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 915–922 (2005)

[3] Filippova, K., Surdeanu, M., Ciaramita, M., and Zaragoza, H.: Company-oriented extractive summarization of financial news, *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 246–254 (2009)

[4] Sakai, H. and Masuyama, S.: Cause Information Extraction from Financial Articles Concerning Business Performance, *IEICE transactions on information and systems*, Vol. 91, No. 4, pp. 959–968 (2008)

[5] 丸澤英将, 和泉潔, 坂地泰紀, 田村浩道：業種別企業業績要因を含む新聞記事の抽出, 第 *19* 回金融情報学研究会, pp. 71–77 (2016)

[6] 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, 第 *17* 回金融情報学研究会, pp. 25–30 (2016)

[7] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.:Distributed representations of words and phrases and their compositionality, *Advances in neural information processing systems*, pp. 3111–3119 (2013)

[8] 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)