決算短信から抽出した業績要因文の事業セグメントに基づく分類と業績文の抽出

Classification based on business segments of causal information and extraction of performance sentences from summary of financial statements

村野壮人 [1]　酒井浩之 [1]　坂地泰紀 [2]　江口潤一 [3]

Taketo Murano[1], Hiroyuki Sakai[1], Hiroki Sakaji[2], and Junichi Eguchi[3]

[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] 東京大学
[2] The University of Tokyo
[3] 大和証券投資信託委託株式会社
[3] Daiwa Asset Management

Abstract: In this research, we propose a method to automatically classify sentences including causal information concerning business performance (e.g. “Orders of semiconductor manufacturing equipment were good.”) extracted from summary of financial statements of companies based on business segments of the companies. Moreover, we propose a method to extract performance sentences from summary of financial statements. For example, the sentences including causal information extracted form summaries of financial statements of SUBARU Co., Ltd. are classified to either “automobile” segment or “aerospace” segment. In addition, our method extracts performance sentences, e.g. “Sales were ¥3,262.0 billion, an increase of ¥93.7 billion (2.9%) compared with the previous fiscal year.”, by deep learning and automatically generates training data.

１．はじめに

近年，投資家に対して投資判断の支援を行う技術が求められており，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いる．例えば，企業の発行している「決算短信」をテキストマイニングの技術を用いて解析し，経済市場を分析する研究などが行われている[1][2][3][4][5].

投資家が投資活動を行うにあたり，上場企業の業績情報の収集は必要不可欠である．また，業績情報の中でも特に業績要因が投資判断において重要である．なぜなら，業績回復の要因が，その企業の主力事業が好調であることであったならば株価への影響は大きいが，株式売却益の計上などの特別利益の計上が要因であるならば株価への影響は軽微であるからである．

関連研究として，酒井らは企業の決算短信 PDF から業績要因文を自動抽出する研究を行っている[1]が，抽出された業績要因がその会社の中でどの程度の重要性かを判断するのは難しい．例えば，株式会社 SUBARU では事業セグメントが「自動車」と「航空宇宙」の２つあり，それぞれのセグメントごとの 2016 年 3 月 31 日から 2017 年 4 月 1 日までの利益は自動車セグメントで 397,657 百万円，航空宇宙セグメントで 9,102 百万円となっており，大きな差がある．そのため，SUBARU にとって自動車セグメントの業績要因は航空宇宙セグメントの業績要因よりも重要であると考えられる．そのため，酒井らの研究によって抽出された業績要因がどの事業セグメントに属するかを分類し，さらに，事業セグメントごと業績によって重要度を付与できれば，投資判断を行ううえで重要な情報源となると考える．

そこで本研究では，例えば，SUBARU の決算短信から抽出された業績要因文である「国内の登録車につきましては、全面改良を行った「インプレッサ」

に加え、「レヴォーグ」および「フォレスター」の販売が好調に推移したことにより、売上台数は 12.6 万台と前期に比べ 1.5 万台(13.3%)の増加となりました。」に「自動車」等の事業セグメントに自動で分類する手法を提案する．また，それに付随し，分類されたセグメント毎の重要度の業績による付与を目標とし，業績文の自動抽出についても行う．例えば，「売上高は、自動車売上台数の増加などにより、為替変動に伴う売上高の減少を吸収し、過去最高となる 3 兆 3,260 億円と前期に比べ 937 億円(2.9%)の増収となりました。」といった業績文を深層学習によって抽出する．

2．業績要因文の事業セグメントに基づく分類

業績要因文の事業セグメントに基づく分類をするにあたって，各企業ごとの事業セグメント名が必要になる．その事業セグメント名を自動で収集できることが望ましいが，今回の研究では人手で事業セグメント名を収集し，その事業セグメントに基づいた分類を行うものとする．また，業績要因文を分類するにあたって，学習データを人手にて作成し、その学習データによる機械学習手法による方法が考えられる．しかし、本タスクにおいては企業ごとに分類に必要な学習データが必要になるため，全ての企業に対応できる学習データの作成には多大な労力を要する．そのため，実際に業務に利用する際のことを考慮し，機械学習に用いる学習データをもなるべく自動的に生成し，業績要因文を事業セグメントに分類をすることを目標とする．以上の説明をふまえ，本手法の概要を以下に示す．

Step 1: 決算短信 PDF や企業 Web ページから各企業の事業セグメント名を人手で収集する．

Step 2: 決算短信 PDF から抽出された業績要因文に直前に出現した事業セグメント名を付与したものを教師データとして K 近傍法で分類する．

Step 3: 自動で直前に出現した事業セグメント名を付与した業績要因文を，業績要因文に含まれる企業キーワード（後述）の合計スコアを元に閾値で絞り，それを教師データとして K 近傍法によって分類する．

Step 2 のみでは精度が低すぎるため，Step 3 の処理を実行した．

2.1. 各企業の事業セグメント名の収集

企業の多くは，事業セグメントという，企業の構成単位によって，業務内容を分類している．本研究では事業セグメント名の自動抽出はせず，人手によって企業ごとに収集したセグメント名を業績要因文に付与していく．表 1 に人手で収集した企業と事業セグメントの例を記す．

表 1：企業毎の事業セグメント

企業名	事業セグメント名
日本電信電話	地域通信事業，長距離・国際通信事業，移動通信事業，データ通信事業
SUBARU	自動車，航空宇宙
三菱地所	ビル事業，生活産業不動産事業，住宅事業，海外事業，投資マネジメント事業，設計監理事業，ホテル事業，不動産サービス事業
花王	ビューティケア事業，ヒューマンヘルスケア事業，ファブリック&ホームケア事業，ケミカル事業
日立製作所	情報・通信システム，社会・産業システム，電子装置・システム，建設機械，高機能材料，オートモティブシステム，生活・エコシステム

2.2. 業績要因文への事業セグメント名の付与

業績要因文へ事業セグメント名を付与するにあたって，業績要因文の直前，または文中に出現していた事業セグメント名を付与する．この時，同じ文中に複数の事業セグメント名が出現していた場合，その文へ付与するセグメント名は「無し」とした．

例えば，SUBARU で「一方、軽自動車につきましては、4 月に「ルクラ」、「プレオ」、「プレオバン」を投入したことや、「サンバー」シリーズが前年同期を上回る台数で推移したことにより、売上台数は 51 千台と前年同期比 5 千台(10.6%)の増加となりました。」といった業績要因文が抽出された場合，文中に事業セグメントの一つである「自動車」が含まれているため，「自動車」セグメントに分類される．

2.3. k 近傍法による業績要因文の分類

2.2 節の手法によって生成されたデータを学習データとして，セグメント名の分類を，K 近傍法を用いて行う．また，文中に複数の事業セグメント名が出

現していた場合，その業績要因は特定の事業セグメントに属さないと判断し，セグメント名を「無し」とした．ただし，2.2 節の手法によって作成した学習データによる分類は精度が低いため，学習データの絞り込みを行った．具体的には，学習データとする業績要因文に含まれる企業キーワードのスコアの合計値を算出し，企業毎のその平均値に 0.8 をかけたものを閾値とし，閾値未満となった業績要因文を教師データから除外した．ここで，企業キーワードとは酒井らの手法[1]によって決算短信から抽出された，企業ごとの重要なキーワードである．企業キーワードは，企業 t の決算短信ＰＤＦ集合 S(t)に含まれる名詞 n に対して、以下の式でスコア W(n ,S(t)) を計算し，スコアが大きい名詞を企業キーワードとして抽出する．

$$W(n_i,S(t)) = \left(0.5 + 0.5\frac{tf(n_i,S(t)}{max_{j=1,\dots,m}TF\left(n_j,S(t)\right)}\right) \times H(n_i,S(t)) \times \log_2\frac{N}{df(n_i)}$$

ここで，
S(t): ある企業 t の決算短信の集合．
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) の各決算短信である d に名詞 n が出現する確率に基づくエントロピー．
df(n): 名詞 n を含む決算短信をもつ企業の数．
N: 決算短信を収集した企業の数．

企業キーワードとスコア W(n ,S(t)) の例を表 2 に示す．

表 2：企業キーワードの例

企業名	企業キーワード	W(n ,S(t))
花王	コンシューマープロダクツ事業	7.65
	コンシューマープロダクツ	7.15
	ヒューマンヘルスケア事業	6.26
	ヒューマンヘルスケア	5.74
SUBARU	レガシィ	1.50
	インプレッサ	1.48
	フォレスター	1.44
	宇宙事業部門	1.05
	航空宇宙事業部門	1.05

生成した学習データ D と，学習データ以外の決算短信から新たに抽出した業績要因文 T の文書間類似度を(sim)を以下の式によって求め，学習データにおける事業セグメントが付与された文集合の近傍となる業績要因文を求める．

$$sim(V_d,V_t) = \frac{V_d \cdot V_t}{\|V_d\| \cdot \|V_t\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

$$V_d = \{x_1, x_2, \dots, x_n\}, V_t = \{y_1, y_2, \dots, y_n\}$$

ここで，
V_d: 学習データに含まれる業績要因文 D の企業キーワードを要素，そのスコアを要素値としたベクトル
V_t: テストデータとなる業績要因文Ｔの企業キーワードを要素，そのスコアを要素値としたベクトル

求められた近傍となる文に付与された事業セグメント名による投票で，テストデータの業績要因文に付与する事業セグメント名を決定する．

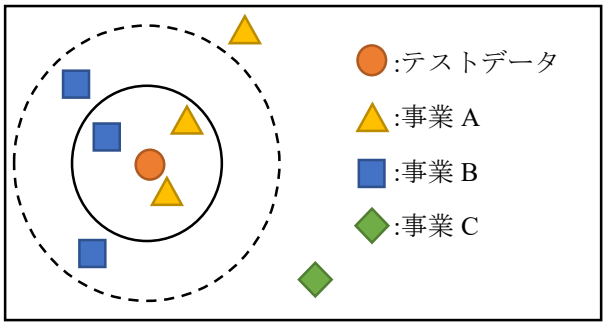

図 1: k 近傍法による分類

図 1 の例では最近傍法ならテストデータに最も近い事業 A に分類され，k=3 でやった場合実線の円の中に入っている 3 つのセグメントで投票が行われる．この場合，事業 A が 2 つ，事業 B が 1 つなので事業 A に分類される．k=5 の場合，点線の円の中に含まれる 5 つのセグメントで投票が行われ，事業 A が 2 つ，事業 B が 3 つとなるので，テストデータは事業 B に分類される．

３．業績文の自動抽出

事業セグメントごとの業績を業績要因文に付与し事業セグメントと結びつけ，重要度とするため業績文を抽出する．なお，事業セグメントごとの業績は１つの表にまとまっていることが多いが，その表は企業ごとにフォーマットが異なっている．そのため，表から事業セグメントごとの業績を直接，抽出することは困難であり，決算短信に含まれる文から業績文を抽出する必要があった．

業績文の抽出には深層学習を用い，そのための学習データも自動生成する．学習データはランダムに

選んだ 1,000 社の過去の決算短信 PDF から以下の条件に合致する文をそれぞれ正例，負例とする．

正例は，「売上」「億円となりました」のどちらも含まれている文，例えば，「当社の当第 3 四半期累計売上高は、主に前年同期比出荷ビットの増加により、32.2%増の 4,222 億円となりました。」や，「HE&S 分野の売上高は，液晶テレビの販売台数が減少しましたが、主に為替の好影響により、前年同期比 11.8%増加し、2,638 億円となりました。」である．

負例は，文中に算用数字，漢数字，共に含まない文，例えば「品質管理及びコンプライアンスに関する教育の強化につきましても継続的に推進しております。」や，「以下、前年同期比については、当該変更を反映した前年同期の数値を用いております。」のようなものである．

以上の条件において抽出された，それぞれ正例 1273 文，負例 328911 文を学習データとし，学習データの正例，負例のどちらにも出現している内容語素性 395 個とした深層学習を用いて分類する．

深層学習のモデルの入力層のノード数は学習データから抽出された素性（395 語）と同じとし，隠れ層はノード数 1,000 が 3 層，ノード数 500 が 3 層，ノード数 200 が 3 層，ノード数 100 が 3 層の計 12 層とする．出力層は 1 要素 である．エポック数は 50，活性化関数として，ReLU を使用した．

本モデルによって業績文（正例）として識別された文の誤分類を調べたところ，「これらの結果、売上台数の合計は、90.6 万台と前期に比べ 9.3 万台(11.4%)の増加となりました。」のような，車の台数に対する文などで，「円」という言葉が入っていないことがわかった．そのため，「円」を含んでいない文を除去した．本手法により抽出できた業績文の例を以下に示す．

- 当第 1 四半期連結累計期間の営業収益は、海外ビジネスにおける為替影響があったものの、国内ビジネスの規模拡大などにより、3,735 億円(前年同期比 3.7%増)となりました。
- 一方、負債は前期と比較して 4 億 11 百万円(6.0%)減少し、65 億 8 百万円となりました。
- また、自営店と加盟店の売上を合計したチェーン全店売上は 4 兆 5,156 億 5 百万円(前年同期比 5.2%増)となりました。

4． 業績文と事業セグメント名の出現位置による分類

業績要因文への事業セグメントの分類を，３章で抽出した業績文の出現位置を利用して行う手法を試みた．事業セグメント名が出現する文と業績文の間を同一セグメントに対する文集合と仮定し，その中に業績要因文が出現したときのみ，業績要因文に事業セグメント付与を行う．しかし，決算短信には企業毎に書き方の特徴があり，事業セグメント名が出現した直後に業績文が出現し，そのあとに業績要因文が出現する場合があったため，他の事業セグメント名が出現するまで，業績文が複数回出現した時，最後の業績文までをひとまとまりとして，そのまとまりの中に出現した業績要因文を分類した．2 節と 4 節での，業績要因文の事業セグメント分類した結果を表 3 に示す．

表 3: 業績要因文のセグメント分類

企業名	事業セグメント名	業績要因文
花王	ビューティケア事業	日本では、「ビオレ」の洗顔料や日焼け止め、乾燥性敏感肌ケア「キュレル」の売り上げが伸長し、前期を上回りました。
	ファブリック&ホームケア事業	アジアでは、インドネシア、タイで衣料用洗剤「アタック」が好調に推移し、台湾、香港で、抗菌機能を高めた衣料用液体洗剤を発売して市場を活性化し、売り上げが伸長しました。
SUBARU	自動車	また、軽自動車につきましては、新型車「シフォン」が販売に寄与したものの、その他車種が減少したことにより、売上台数は 3.3 万台と前期に比べ 0.1 万台(3.4%)の減少となりました。
	航空宇宙	(航空宇宙事業部門)防衛省向け製品では、新多用途ヘリコプター「UH-X」の契約に基づく開発本格化などにより、売上高は前期を上回りました。

5. 評価

5.1 業績要因文の分類の評価

業績要因文のセグメント分類の評価は，事業セグメント名を収集した上場企業の中から選んだ 11 社の企業の決算短信において，自動生成したままの学習データを用いた最近傍法（自動生成），閾値によって学習データの再生成を行ったものを用いた最近傍法（最近傍法），k=3 の k 近傍法（k=3），業績文の位置を利用して分類したもの（業績文利用）の場合の 4 パターンで自動的に業績要因文への事業セグメント名の付与をし，精度を求めた．評価結果を表 4 に示す．また，このとき，どのセグメントにも属さないと分類された業績要因文は全体の数から除外して計算をする．表中の分子は正しく分類された業績要因文の数，分母は抽出されたのち，セグメント名が付与された業績要因文の数である．

また，近傍法での分類が有効かを調べるため，学習データを自動生成したのち，手動で修正を加えた．「花王」，「SUBARU」の２社で最近傍法を実行した場合の精度を表 5 に示す．

表 5: 手動学習データでの最近傍法

企業名	精度
花王	0.9 (9/10)
SUBARU	1 (10/10)

5.2 業績文の抽出の評価

業績文の抽出は，業績要因文の事業セグメントの分類に使用したものと同じ 11 社をで行った．
11 社の最新の決算短信 PDF から抽出した．深層学習による分類と，分類の後に，「円」を含まないものを除去した二通りの精度を表 6 に示す．

表 6: 業績文の抽出精度

深層学習による抽出	0.928 (65/70)
「円」を含まないものを除去	0.984 (62/63)

6. 考察

業績要因文のセグメント分類は，企業ごとの決算短信の書き方に大きく左右されてしまうため，各企業での精度の差が大きくなってしまっている．例として，「花王」，「NTT ドコモ」の 2 社はどの分類法を用いたとしても，多くの業績要因文に対し，セグメント分類が行われ，その精度も良いが，「トヨタ自動車」，「三菱地所」は業績要因文の抽出の時点で抽出結果に誤ったものも多いため，分類ができなくなっている．「ヤフー株式会社」では，明確にセグメント分類できる業績要因文が 2 文しかないにもかかわらず，無理に分類しようとしてしまうため，精度が下がってしまっていたが，業績文を利用した分類の場合は２文のみを適切に分類することができている．しかし，「日本たばこ産業」の場合，他の手法ではうまく分類できていたが，業績文の抽出の再現度が低いことと，業績文の出現する位置の関係上，業績文を利用した場合，全てが未分類となってしまった．

このように，企業によって決算短信の書き方の特徴に多く作用されてしまう．そのため，複数の分類法を企業ごとに使い分けることが良いと考える．しかし，実際に運用するにあたって同じシステムで多くの企業に対応できることが求められるため，2.3 節で行った企業キーワードのスコアを用いた学習データの絞り込みによる誤分類の除去や，k 近傍法と業績文利用の 2 つの和集合を取ることにより，精度・再現率の向上をしたい．

また，「NTT ドコモ」では「通信事業」という事業名に対する文を「通信関連」などと表記されるなど

表 4: 業績要因文のセグメント分類数と精度

企業名	自動生成	最近傍法	k=3	業績文利用
日本たばこ産業	6/8	6/7	6/7	0/0
セブン&アイ・ホールディングス	3/6	2/4	1/3	1/1
花王	7/8	3/3	5/5	6/6
ヤフー株式会社	2/8	2/5	1/5	2/2
中国工業	5/9	0/1	0/0	6/7
村田製作所	0/4	0/1	1/1	1/1
トヨタ自動車	5/10	0/0	0/0	2/9
SUBARU	8/10	6/8	6/8	8/9
三菱地所	2/9	0/0	0/0	0/0
日本電信電話	4/6	1/1	0/0	0/0
NTT ドコモ	6/6	2/2	8/9	5/7
合計（精度）	48/84 (0.571)	22/32 (0.688)	28/38 (0.737)	31/42 (0.738)

の表記揺れが起きていることが確認できた．このような表記揺れは他の企業の決算資料でも見られるので，人手で収集した事業セグメント名を元に表記揺れに対応できるようにするか，自動で事業セグメント名を収集できるようにする技術が求められる．

7. まとめ

本研究では，企業の決算短信 PDF から抽出した業績要因文に対する事業セグメントへの分類手法を提案した．業績要因文の抽出では，人手で収集した事業セグメント名を元に学習データを自動生成し，生成された学習データを用いた k 近傍法によって，業績要因文を分類した．評価の結果，k=3 での k 近傍法で精度 73%，また，業績文の位置を利用した分類でも精度 73%となり，比較的，良好な精度を得ることができた．今後は再現率の向上と共に，業績文中に出現する業績情報による業績要因文への重み付けも行う予定である．

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] 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] 北森詩織，酒井浩之，坂地泰紀, “決算短信 PDF からの業績予測文の抽出”，電子情報通信学会論文誌 D, vol.J100-D, no.2, pp.150‐161, 2017.

[4] 酒井浩之，松下和暉, “決算短信からの業績要因文の抽出”，第 11 回テキストアナリティクス・シンポジウム, pp.87-91, 2017.

[5] 室野莉沙，酒井浩之，坂地泰紀，ベネット ジェイスン, “決算短信から抽出した原因・結果表現の意外性の判定”，第 11 回テキストアナリティクス・シンポジウム, pp.93-98, 2017.