文間の結束性に基づく決算短信における業績要因文の抽出

中山 祐輝　　津々見 誠　　村上 浩司
楽天株式会社 楽天技術研究所
{yuki.b.nakayama, makoto.tsutsumi, koji.murakami}@rakuten.com

1 はじめに

ウェブ上には，経済新聞や決算情報などの市場動向について書かれたテキストが膨大に蓄積されている．それらを有効利用するために，金融テキストマイニングの研究が盛んである [3]．金融テキストマイニングとは，テキスト集合から投資に有用な知識を抽出する処理の総称であり，基盤技術として「かつおやさばなどの漁獲が好調だった。」のような業績につながった要因を抽出し，要因の重要度を推定することが重要である．本論文では，このような要因を含む文を業績要因文と呼ぶ．業績要因文の重要度が推定されると，機関投資家と個人投資家双方において，業績に対する株価への影響度合いを知る手がかりとなるため，投資判断の支援に役立つ．また，業績要因文の抽出結果を要約に応用することで，証券アナリストが行うレポート作成の効率化につながる．

本論文では，テキスト情報の中で決算短信に着目し，第一ステップとして決算短信から業績要因文を抽出する手法を提案する．決算短信は全ての上場企業で一般公開されており，かつ最も早く提供される決算情報であることから，可用性と速報性の観点で優れており，市場変動を知るためにより有益な情報源となる．

本研究は，業績要因文を内的要因と外的要因に大別する．内的な業績要因文とは,「対象企業全体またはその属性」において，財政上ではないかつ業績結果につながった「取り組みまたは状態変化」を含む文と定義する．以下の例文を用いて具体的に説明する．

(1) 世界経済は、海外政治情勢が不安定となり地政学的リスクが増した一方で、国際金融市場は安定を保ち、世界各国の景気は緩やかな回復基調で推移しました。

(2) これらの結果、(a) 日本綜合飲料事業では、キリンビール (株) におけるビール類の販売数量の減少や、/(b) キリンビバレッジ（株）で前第 1 四半期連結会計期間に費用として計上していた一部販売費を売上高から控除した影響で減収となりましたが、/(c) 各事業会社でコスト削減等による収益性改善の取り組みが進み、/(d) 増益となりました。

例文 (1) は，対象企業ではなく世界経済の動向に関する記述であり，業績要因文ではない．例文 (2) の (a) と (c) は,「ビール類の販売数量の減少」と「収益性改善の取り組みが進み、」という商品の属性に関する状態変化と取り組みがそれぞれ記述されており，かつこれらは業績結果に結びついたため，業績要因である．よって，例文 (2) は業績要因文である．ここで，業績結果とは定量・定性に関わらず，利益情報が記述された部分であり，(d) の「増益」や「営業利益は 1 億円」が業績結果である．一方，(b) は費用の計上という財政の状態変化に関する記述なので，業績要因ではない．このような財政上の要因に起因する業績結果の場合，株価への影響は軽微である [4] と判断できるため，本研究では財政上の要因を業績要因の対象とはしない．

業績要因の候補が業績結果とは異なる文に含まれる場合は，文脈で判断する．例えば，同じ決算短信において例文 (2) の前に以下のような文があるとしよう．

(3) なお、缶・小型 PET 容器を中心とした販売目標管理や SCM コスト削減の取り組みを継続し、収益力の向上を図りました。

例文 (3) 中の取り組みは，例文 (2) 中の「コスト削減等による収益性改善の取り組み」につながり,「増益」という結果につながったことから，例文 (3) は業績要因文とする．

外的な業績要因文とは，為替変動に関する要因を含む文と定義する．

(4) プロフェッショナルプリンティング事業の売上収益は為替影響もあり増加となりました。

円高が進むと輸出関連株が安くなるなど為替変動が株価に影響するため，業績要因文の対象とする．

2 関連研究

Jacobs [2] らは，ニュース記事内の経済イベントと株価変動の関係性を明らかにするために，経済イベントを 10 種類のカテゴリに分類する手法を提案した．彼らは，企業によって発行される決算情報の業績結果をイベントのカテゴリとしており，決算情報を用いたという点で本研究と類似している．しかし，本研究は業績結果に対する要因の特定が目的である．

Ding [1] らは，ニュース記事内から (行為者，行動，対象，日時)=(マイクロソフト，買収する，ノキアの携帯事業，2013 年 9 月) のような四つ組でイベントを抽出し，それらを株価予測に応用した．本研究の業績要因もイベントとみなせる．しかし，行動だけでなく (a) のような企業の状態変化も対象とする点で彼らとは異なる．

決算短信から業績要因文を抽出する研究に酒井ら [4] の先行研究がある．彼らは， 業績要因文の明確な定義がされていないことから、本論文ではまずその定義を試

みる．また，彼らの手法は，決算短信を文の集合ととらえ，文間のつながりを考慮しない．我々の提案手法は談話構造を意識した手法であり，結束性を用いて更なる改善を図る．

3 既存手法

酒井らの手法は，図 1 のように企業キーワードを含む文節が，係り受けをたどって手掛り表現に到達する文を業績要因文として抽出する．企業キーワードとは，商品名や部門名等，その企業にとって重要なキー ワードである．例えば，図 1 中の「かつお」「さば」「漁獲」という語が「日本水産」の企業キーワードにふさわしい．手掛り表現とは,「好調だった。」のような業績要因となる状況や変化を表す用言的な表現である．手掛り表現は，業績発表記事からまず収集され，その後業績要因文の抽出結果を用いて，決算短信から追加で収集される．企業キーワードの収集は，まず企業 t の決算短信に出現する名詞 n に対して，$W(n, F(t))$ を計算する．

$$W(n, F(t)) = H(n, F(t))TF(n, F(t)) \log_2 \frac{N}{df(n)} \quad (1)$$

$F(t)$ は企業 t の決算短信集合である．そして，$W(n, F(t))$ が企業 t における $W(n, F(t))$ の平均より大きく，かつ $\log_2(N/df(n))$ が 1 より大きい名詞を企業キーワードとして抽出する．式 (1) は，$f \in F(t)$ 中で名詞 n が決算短信 f で出現する確率のエントロピーと，$F(t)$ 中の文書全てを結合して一つの文書とみなした TF-IDF からなる．つまり，$F(t)$ 中の決算短信に多くかつ，まんべんなく出現し，他の企業の決算短信には出現しない名詞に対して高い値をとる．

4 提案手法

提案手法は，決算短信が文の集合であるとらえるのではなく，談話構造をなしているとみなし，文間の結束性に着目する．結束性はテキストの一貫性を実現するための道具であり，テキストの表層による文法的なつながりである．提案手法では，結束性を持つ文同士において，一方の抽出結果が片方の結果に依存すると仮定する．例えば，同じ決算短信に以下の文同士が隣接する場合を考える．

(5) 米ドルの平均為替レートは 108.15 円と前年同期に比べ、11 %の円高で推移しました。

(6) また、新興国通貨の為替レートも円高で推移し、特に中南米国通貨は米ドルやユーロを超える円高で推移しました。

例文 (6) の「また」など，文頭に位置する接続詞は，例文 (5) の話題に対して並列構造，対比，詳細化などを実現するための結束性を担う．このような結束性を持つ場合，片方が業績要因文の可能性が高ければ（低ければ），もう一方も業績要因文でありそう（なさそう）である．図 2 に提案手法の概要を示す．以下で，各ステップを説明する．

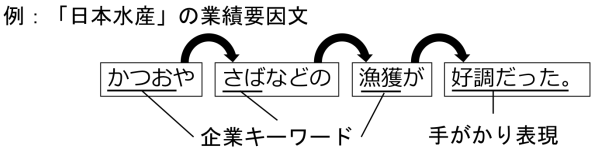

図 1: 先行研究における業績要因文の抽出手法

信頼性が最大と最低の業績要因文を選択 3 節で述べた先行研究の手法を適用し，抽出された業績要因文それぞれに対して，信頼度を計算する．企業 t の決算短信 $f \in F(t)$ における i 行目の業績要因文 s_i の信頼度 $conf(t, f, i)$ に基づいて，信頼度が最大の行番号 $M(t, f)$ と最低の行番号 $L(t, f)$ を取得する．

$$M(t, f) = \arg\max_i conf(t, f, i) \quad (2)$$

$$L(t, f) = \arg\min_i conf(t, f, i) \quad (3)$$

$$conf(t, f, i) = \frac{1}{|K(t, f, i)|} \sum_{n \in K(t,f,i)} W(n, F(t)) \quad (4)$$

ここで，$K(t, f, i)$ は i 行目の業績要因文中で，係り受けをたどって手掛り表現に到達する企業キーワードの集合である．先行研究では，そのような企業キーワードの有無によって業績要因文かどうかを判定しているため，信頼性を推定する上で重要な企業キーワードである．$conf(t, f, i)$ は，$n \in K$ を満たす全ての n で W(n,F(t)) が高いほど，高い値をとる．以降，説明の都合上 M と L の引数は省略する．

信頼度が最低の文における結果を反転 業績要因文として抽出された中で，信頼度が最小の文は業績要因文の可能性は低いと考え，s_L を業績要因文でないとする．

セグメントごとに結束性を特定 本節の冒頭で述べた結束性を持つ文同士の依存関係は，同じ話題に対して存在すると考え，セグメントごとに s_L と s_M を中心として結束性を特定する．セグメントとは，ある見出しについて記述された本文の一部分である．セグメントの特定は，句点のない行が見出しであるという傾向を用いることで行う．提案手法は，文同士が業績要因文の判定において依存関係をもつ可能性のある結束性として以下の三つを用いる．

- 接続：文頭の接続詞（e.g., また，そして，一方）、文頭の副詞（e.g., 特に）
- 参照：文頭の代名詞（これ，それ）
- 語彙的結束性

接続と参照については，以下の手順で隣接した文において結束性の有無を特定する．以下は，s_L の例であるが，s_M についても同じ処理を行う．

1. $i = L - 1, j = L + 1$
   (前向きステップ)
2. s_i が見出しであれば 5. へ

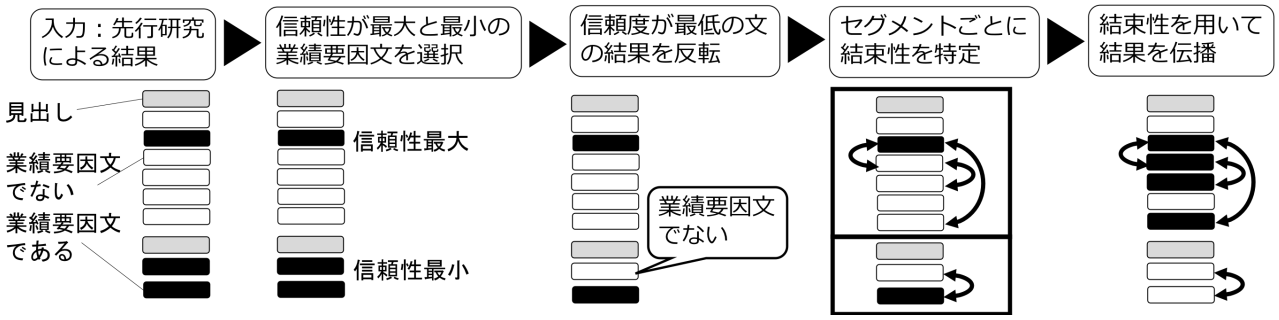

図 2: 提案手法の概要

3. s_i の文頭に接続詞もしくは副詞があれば，s_i と s_{i+1} の間に結束性があるとし，なければ 5. へ
4. i を一つ減らし，2. へ

(後向きステップ)

5. s_j が見出しであれば，処理を終了
6. s_j の文頭に接続詞もしくは副詞があれば，s_j と s_{j-1} の間に結束性があるとし，なければ処理を終了
7. j を一つ増やし，5 へ

語彙的結束性については，文間の語の重複に着目する．一般的に文は，主題を示す部分（主題部）とそれ以外の部分（非主題部）に分けられる [5]．本研究では「において, におきまして, は, につきまして, について」のいずれかにマッチする文節を主題部，それ以外を非主題部とする．二文間で主題に語の重複がある場合は，類似した話題の可能性があるため，一方の抽出結果が片方の結果に依存すると考えられる．また，二文間で主題部と非主題部の間に語の重複がある場合も，例文 (7) と例文 (8) のように，同じ話題に対して詳細化される可能性があるため，同じことが言える．以上の考えに基づいて，s_M と s_L それぞれの名詞の形態素集合を用いて次の処理を行う．同じセグメント内の文について，主題部内の名詞の形態素集合と上記の集合に一つでも重複する形態素があれば，両者の間に結束性があると判定する．

(7) 国内卸売事業は、花種子の売上高が減少しましたが、野菜種子と資材の売上高が増加しました．
(8) 品目別では、野菜種子は、ブロッコリーが増加となりましたが、ダイコンが減少となりました。

結束性を用いて結果を伝播 s_L と s_M から全ての結束関係をたどり，結果を伝播させる．図 2 では，まず s_M と結束性をもつ 4 行目と 7 行目の文が業績要因文であると判定され，さらに 4 行目の文から結束性をたどることで 5 行目の文も業績要因文として抽出される．一方，s_L から結束性をたどることで，10 文目は業績要因文でないと判定される．

5 評価実験

提案手法の有効性を検証するために，先行研究を比較対象として実験を行なった．酒井らが選択した 10 社の企業における決算短信から 1 文書ずつ抜き出し，計 10 文書を評価用のデータとした．次に，上記 10 件の決算短信に対して，2 名の判定者が独立に業績要因文を人手で特定した．判定者は第一著者と第二著者であり，前者は 6 件，後者は 4 件に対してアノテーションを行なった．判定者には予備的な試行を通して判定基準を周知徹底するよう努めた．その後，アノテーションの一致率を測定するために，アノテーション済みの決算短信 1 件を互いに交換し独立にアノテーションを行い，2 文書におけるアノテーションの一致率を測定した．その結果，2 文書それぞれの一致率は，κ 値で 0.76 と 0.86 であった．

表 1: 手掛り表現と企業キーワードの獲得に用いたデータ

		先行研究	本研究
日経経済新聞	記事数	71,070	64,121
	文字数/記事	325.2	336.3
決算短信	企業数	3,821	3,635
	文書数	106,885	31,251

先行研究の手法を再現するための手掛り表現と企業キーワードの獲得は，表 1 に示すデータを用いた．また，手掛り表現の収集については，公開プログラム[1]を用いた．決算短信は，EDINET[2]から収集した．

表 2 に各企業における評価データの基礎情報と実験結果を示す．各企業で優れた手法の F 値は太字にした．提案手法は，10 企業中 6 企業で先行研究より優れた F 値を達成した．表 3 の (a) は，提案手法により業績要因文として正しく抽出できた例である．(a) では 26 行目の文が信頼性の最も高い業績要因文として抽出され，かつ主題部の間で「飲料」という形態素が重複したという語彙的結束性を用いたことで，33 行目の文を業績要因文として抽出できた．しかし，(b) の 62 行目の文は「冷蔵庫」という語が企業キーワードとして抽出されなかった一方で,「コスト」という企業キーワードとしてふさわしくない語が抽出された結果，信頼度が低くなり，信頼性が最低の業績要因文として特定されたことによって，業績要因文として抽出されなかった．提案手法は，先行研究による企業キーワードの抽出手法に基づいて信頼度を計算するため，その手法を改良することが提案手法の精度向上につながる．

[1] http://www.ci.seikei.ac.jp/sakai/clupes.html
[2] http://disclosure.edinet-fsa.go.jp

表 2: 評価用データの詳細と実験結果

証券番号	企業名	決算短信文数	業績要因文数	先行研究			提案手法		
				精度	再現率	F 値	精度	再現率	F 値
1332	日本水産	98	14	0.80	0.57	**0.67**	0.78	0.50	0.61
1377	サカタのタネ	66	20	0.72	0.90	0.80	0.72	0.90	0.80
2503	キリン HD	74	35	0.85	0.80	0.82	0.88	0.83	**0.85**
2730	エディオン	43	6	0.40	0.33	0.36	0.50	0.50	**0.50**
4674	クレスコ	149	9	0.36	0.56	**0.43**	0.27	0.33	0.30
4911	資生堂	70	19	0.64	0.74	0.68	0.67	0.74	**0.70**
6502	東芝	472	14	0.23	0.86	0.36	0.24	0.86	**0.37**
6724	セイコーエプソン	98	25	0.50	0.96	0.66	0.52	0.96	**0.68**
6758	ソニー	401	69	0.45	0.57	0.50	0.45	0.57	0.50
9468	KADOKAWA	70	25	0.45	0.43	0.44	0.50	0.48	**0.49**
	平均	154.1	23.6	0.53	0.671	0.57	0.55	0.666	**0.58**

表 3: 抽出結果の例

(a) キリン HD

行番号	文	業績要因文？		
		正解	先行研究	提案
26 (s_M)	ノンアルコール・ビールテイスト飲料の販売数量は、「キリン零 ICHI（ゼロイチ）」の販売数量が引き続き伸長したことにより、前年から約 6 割増加しました。	Yes	Yes	Yes
33	健康を基軸にした価値創造に挑戦した健康・スポーツ飲料カテゴリーでは、機能性表示食品「キリンサプリ」ブランドから新商品を発売し、カテゴリー全体の販売が伸長しました。	Yes	No	Yes

(b) 日本水産

行数	文	業績要因文？		
		正解	先行研究	提案
54	＜医薬原料、機能性原料、機能性食品＞			
55 (s_M)	・医薬原料は、後発品使用促進策の影響があり苦戦したが、乳児用粉ミルクに添加する DHA などの機能性原料の販売が国内外とも堅調に推移したことに加え、特定保健用食品「イマークＳ」など通信販売の広告宣伝費削減もあり増収・増益となった。	Yes	Yes	Yes
56	＜臨床診断薬、産業検査薬、医薬品＞			
62 (s_L)	・営業再開した冷蔵庫の効果もあり売上は増加したものの、労務費や電力料などのコストが増加し、前年同期並みの利益となった。	Yes	Yes	No

提案手法は，セグメントが細かい場合に，結果の伝播が限定される．表 3(b) では，55 行目の文が最も信頼性の高い業績要因文と認定されたものの，粒度の細かい見出しが頻発することで，セグメントが小さくなり，他の文に結果を伝播させることができかなった．このような事例に対処するためには，セグメント間の関係を特定する必要がある．また,「サカタのタネ」と「ソニー」は，先行研究と全く同じ結果となり，結果の伝播が行えなかった．これは，s_M と s_L の二文のみを結果の伝播に利用していることが一因としてあり，どの文を対象としてどのように結果を伝播させるかが今後の課題である．

6 おわりに

本論文では，決算短信から業績要因文を抽出する手法を提案した．本論文の貢献は，業績要因文の定義を行ったことと，文間の結束性を利用して抽出結果を伝播させる手法を考案し，実験を通してその有効性を検証したことである．結果の伝播方法と企業キーワードの抽出手法を改良することが今後の課題である．

参考文献

[1] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of EMNLP*, pp. 1415–1425, 2014.

[2] Gilles Jacobs, Els Lefever, and Véronique Hoste. Economic event detection in company-specific news text. In *Proceedings of the First Workshop on Economics and Natural Language Processing*, pp. 1–10, 2018.

[3] 和泉潔, 松井藤五郎. 金融市場における最新情報技術：8. 金融テキストマイニング研究の紹介. 情報処理, Vol. 53, No. 9, pp. 932–937, 2012.

[4] 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.

[5] 柴田知秀, 黒橋禎夫. 談話構造解析に基づくスライドの自動生成. 自然言語処理, Vol. 13, No. 3, pp. 91–111, 2006.