意外性のある原因・結果表現の決算短信からの抽出

Extraction of rare cause-result expressions from summary of financial statements

酒井 浩之 * ¹	坂地 泰紀 * ²	室野 莉沙 *1	北島 良三 * ¹	ベネット ジェイスン *3
Hiroyuki Sakai	Hiroki Sakaji	Risa Murono	Ryozo Kitajima	Jason Bennett
*1成蹊大争	老 *2東京	大学 *3三	井住友アセット	マネジメント株式会社
Seikei University The University of T		y of Tokyo Sun	Sumitomo Mitsui Asset Management Company, Limited	

In this research, we propose a method to extract rare cause-result expressions from summary of financial statements. For example, our method extracts effect expression that "Sales of drinking paper containers has increased" with respect to cause expression that "due to hot summer", as a rare cause-result expression. It is difficult to imagine that the sales of "drinking paper containers" may increase in the case of "hot summer". Our method calculates the conditional probability with words contained in the result expression co-occurring with words contained in the cause expressions. Moreover, our method determines that cause information is contained in the result expression or not by using deep learning. Our method extracts rare cause-effect expressions by using the conditional probability and the cause information.

1. はじめに

近年,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を支援する技術が注目されている. さら に,最近では証券市場における個人投資家の比重が増大してお り,個人投資家に対して投資判断の支援を行う技術の必要性が 高まっている [和泉 11][藏本 13].

投資家にとって企業の業績に関する情報は投資判断を行う うえで重要であるが、企業の業績だけでなく、その業績要因に 含まれる原因とそれに対する結果が重要である。例えば、原因 が「猛暑の影響により」において、結果が「冷房需要の盛り上 がり」といった情報を投資家に提示することで、「猛暑」の場 合には「冷房需要」が高まる可能性があることを個人投資家が 知ることができるというメリットがある。そして、その原因に 対する結果の情報から、猛暑の年には冷房機器に関する事業を 行っている企業(空調機器メーカーや家電量販店など)の業績 が好調に推移することが期待できる。

企業の業績に関する情報は、企業が決算ごとに発行する決 算短信に記載されており,業績要因に含まれる原因と結果に関 する情報も記述されている.しかしながら,証券市場の上場企 業数は約3,600社と多いうえに、近年では年に4回、決算発 表がある*1. そのため,人手によって多くの決算短信を読み, 原因と結果の情報を取得するのには多大な労力を要する. そこ で、坂地らは、決算短信から例えば「夏場の猛暑の影響により 冷房需要が盛り上がりました。」のような原因・結果表現を自 動的に抽出する手法を提案している [坂地 15]. しかし,坂地 らの手法によって抽出された原因・結果表現のままでは投資に 活用することは難しい. なぜなら, 原因「猛暑の影響により」 に対する結果「冷房需要の盛り上がり」の情報は誰にでも予 想できることである.従って,猛暑のときに冷房のような空調 機器を製造している企業の業績が良くなることは自明であり, 投資に有用な情報とは言い難い.しかし,例えば原因「猛暑の 影響により」に対して結果「除草関連商品の売り上げが好調

連絡先: 酒井浩之, 成蹊大学, 東京都武蔵野市吉祥寺北町3-3-1, h-sakai@st.seikei.ac.jp

*1 1年を4期に分け、3か月ごとに企業が公表する. 平成15年より全上場企業に義務づけられている

になる」や「飲料用紙容器の販売が増加した」という情報は、 個人投資家にとっては簡単には思いつきにくく意外性がある. すなわち、猛暑のときは雑草の成長が早く、結果として除草関 連商品の売り上げが好調になることを示唆している.従って、 猛暑のときに除草に関連する事業を行っている企業に対して投 資をしておけばよいことがわかる.そのため、室野らは決算短 信から抽出した原因・結果表現のなかから*2、意外性のある 原因・結果表現を判定する手法を提案した [室野 17].

しかし,室野らの手法では,入力として「猛暑」「冷夏」の ような「原因となる語」と定義された語を必要としている.す なわち,原因となる語として「猛暑」を入力し,「猛暑」を含む 原因表現に対応する結果表現のなかから意外性のある結果表現 を判定する.そのため,「猛暑」「冷夏」「暖冬」「厳冬」のよう な季節に関する語や天災に関する語であれば分かりやすいが, それ以外の意外性のある原因・結果表現を得るためには適切な 「原因となる語」が必要となるため,投資に対する専門知識や 企業の事業に関する専門知識がないと活用が難しいという問題 があった.さらに,意外性のある原因・結果表現を多く獲得し 投資のための知識ベースとして運用するためには,特定の「原 因となる語」に限定することは望ましくない.そのため,本研 究では,「猛暑」のような「原因となる語」を必要とせず,決算 短信のみから意外性のある原因・結果表現を抽出することを目 的とする.

関連研究として,酒井らは決算短信から業績要因を含む文 (例えば「半導体製造装置の受注が好調でした。」)を抽出する 手法を提案している [酒井 15][酒井 17]. 北森らは決算短信か ら業績予測文 (今後の業績予測に関する情報が記述されている 文)を抽出する手法を提案している [北森 17a][Kitamori 17b]. 上記の研究では,業績に対しての要因,予測に関する情報を抽 出しているため、「~が好調」のため業績が良くなるのような 原因・結果情報ともいえる.しかし,好調であれば業績が良く なるのは自明なので,意外性のある情報は抽出できず,本研究 とは目的が異なる.

^{*2} 原因・結果表現の抽出には坂地らの手法 [坂地 15] を使用した.

2. 原因・結果表現の意外性の判定

まずは,室野らの手法 [室野 17] について簡単に述べる.以下に手法の概要を示す.

Step 1: 決算短信から原因・結果表現を抽出する.

- Step 2: 原因となる語(「猛暑」など)を指定し,原因となる 語を含む原因表現に対する結果表現から名詞を抽出する.
- Step 3: 原因表現に Step 2 で指定した原因となる語が出現す るときに、対する結果表現に Step 2 で抽出された名詞が 出現する条件付き確率を求める
- Step 4: Step 3 で求めた条件付き確率と,企業キーワード(後述)を用いて,意外性のある原因・結果表現であるかを 判定するための意外性スコアを求め,意外性スコアが高いものを意外性のある原因・結果表現と判定する.

2.1 条件付き確率の算出

Step 3 における条件付き確率を用いる理由として,例えば, 原因表現に「猛暑」を含むときは,結果表現に「冷房」という 名詞は多く出現し,逆に,原因表現が「猛暑」を含むときには, 結果表現に「除草剤」は多く出現しないという仮定に基づく. したがって,原因に「猛暑」が出現するときに,結果に「冷房」 が出現する条件付き確率は大きくなり,「除草剤」が出現する 条件付き確率は小さくなることが予想できる.意外性があるの は「除草剤」であり,条件付確率が小さい名詞対を含む原因・ 結果表現は意外性があると判定できる.原因となる語 k が原 因表現に出現したときに,結果表現に名詞 w が出現する条件 付き確率 P(w|kw) を以下の式 1 で算出する.

$$P(w|k) = \frac{f(w,k)}{f(k)} \tag{1}$$

ここで, f(w,k)は, 全ての原因・結果表現の集合において k が含まれている原因表現に対応する結果表現に名詞 w が出現 する頻度を表し, f(k)は, 全ての原因・結果表現の集合にお いて k が含まれている原因表現が出現する頻度を表す.

2.2 企業キーワードに基づくスコア

酒井らは決算短信からその企業にとって重要なキーワードを 抽出する手法を提案しており [酒井 15], そのようなキーワー ドを企業キーワードと定義する.決算短信からの企業キーワー ドの抽出は,企業 t の決算短信における名詞 n に対して以下 の式 2 で重み W(n, S(t)) を計算することで行う.

$$W(n, S(t)) = (0.5 + 0.5 \times \frac{tf(n, S(t))}{\max tf(n, S(t))})$$
$$\times H(n, S(t)) \times \log_2 \frac{N}{df(n)}$$
(2)

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. begineqnarray

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.

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.

ここで,企業*t*における重要な原因・結果表現を判定するために,企業*t*の決算短信において,原因となる語*k*を含む原因・結果表現に対して以下の式で求めるスコア*KS*(*k*,*t*)を求め,意外性の判定に使用する.

$$KS(k,t) = \frac{2S_1(k,t)S_2(k,t)}{S_1(k,t) + S_2(k,t)}$$
(3)

$$S_1(k,t) = c_1 \sum_{n \in T_1(k)} \frac{W(n, S(t))}{\max_{n'} W(n', S(t))}$$
(4)

$$S_2(k,t) = c_2 \sum_{n \in T_2(k)} \frac{W(n,S(t))}{max_{n'}W(n',S(t))}$$
(5)

ここで, c_1 , c_2 はそれぞれ原因表現,結果表現に含まれる企 業キーワードの個数, $T_1(k)$ はkを含む原因表現に含まれてい る企業キーワードの集合, $T_2(k)$ はkを含む原因表現に対する 結果表現の中に含まれている企業キーワードの集合である.

2.3 意外性スコア

企業 t の決算短信において,原因となる語 k を含む原因・結 果表現の意外性を判定するための指標である意外性スコアを以 下の式で求める.

$$SS(k,t) = \frac{KS(k,t)}{\sum_{w \in r(k)} P(w|k)}$$
(6)

ここで, r(k) は, k を含んでいる原因表現に対する結果表現の 中に含まれている名詞の集合である. 意外性スコアが大きい原 因・結果表現を意外性があると判定するため,条件付き確率に 基づくスコアを分母に,企業キーワードに基づくスコアを分子 とする式を意外性スコアとした.

3. 意外性のある原因・結果表現の抽出

前節の手法では「猛暑」のような原因となる語を入力として 必要とするが、本節では、原因となる語を指定しなくても決算 短信から意外性のある原因・結果表現を抽出する手法について 述べる.具体的には、「原因となる語」として原因表現に含ま れる全ての名詞 N-gram を採用し、意外性スコアの大きいも のを出力する.しかし、上記の方法では原因・結果表現として 不適切な表現が多く抽出されてしまう.多くの原因・結果表現 を分析したところ、例えば以下に示すような、結果表現に業績 要因に関する情報が含まれている原因・結果表現が適切であっ た*³.

・原因:多雨によるブラジル出し鉄鉱石の荷動き鈍化 ・結果:ケープサイズ船市況は低迷しました

従って,意外性スコアが高い原因・結果表現の中から結果表現 に業績要因に関する情報が含まれているかどうかを判定し,そ のような原因・結果表現のみを抽出する.

3.1 結果表現における業績要因の判定

結果表現に業績要因に関する情報が含まれているかどうか の判定に深層学習を用いる.具体的には,決算短信から業績要

*3 ケープサイズ船とはパナマ運河が通航できない大型船を表す.

因を含む文を抽出する手法 [酒井 17] を,結果表現にのみ適用 する.以下に深層学習を用いた業績要因の判定手法について述 べる.

- Step 1: 酒井らの手法 [酒井 15] を用いて決算短信から抽出し た業績要因文から,手がかり表現の"拡張手がかり表現" を獲得する
- Step 2: 業績要因文に対して前述の企業キーワードを用いて スコアを付与する.
- Step 3: 拡張手がかり表現を含み,かつ,スコアが高い業績 要因文を正例,手がかり表現,企業キーワードをともに 含まない文を負例として学習データを自動生成する.
- Step 4: 自動生成された学習データを使用し,深層学習にて 結果表現の業績要因を判定する.

酒井らの手法 [酒井 15] では、「好調でした」「不振でした」と いった業績要因文の抽出に有効な手がかり表現をブートスト ラップ的に獲得し、そのような手がかり表現と企業キーワード を使用して,業績要因文の抽出を行っている.本手法では,手 がかり表現にいくつかの文節を追加することで、"拡張手がか り表現"と定義する文節列(例えば「受注が好調でした」)を 獲得し,例えば拡張手がかり表現「極めて好調でした」を含む 業績要因文のみを抽出する. さらに,業績要因文に含まれる企 業キーワードを使用して業績要因文にスコアを付与し、スコア の高い業績要因文のみを抽出する. 上記の処理を行うことで, 精度 80% 程度である酒井らの手法による抽出精度を 100% 近 くまで高め*4,抽出された業績要因文を学習データの正例と する. 負例は、決算短信の企業キーワードと手がかり表現、ど ちらも含まない文とする.そして、997 企業の決算短信から 30,839 文の業績要因文を正例として生成した.また、負例は 正例と同数とし、従って 61,678 文の学習データを生成した.

自動生成された学習データに基づき,深層学習により業績要 因を判定する.まず,入力層の要素となる語(素性)を選択す る.具体的には,自動生成された学習データにおいて正例の業 績要因文に含まれる内容語(名詞,動詞,形容詞)に対して, 以下の式7にて重みを計算する.

$$W_p(t, S_p) = TF(t, S_p) \times H(t, S_p)$$
(7)

ここで,

S_p: 学習データにおいて正例に属する業績要因文の集合

- $TF(t, S_p)$: 文集合 S_p において, 語 t が出現する頻度
- H(t,S_p): 文集合 S_p における各業績要因文に含まれる語 t の 出現確率に基づくエントロピー

同様に,負例の文に含まれる内容語に対しても同様に重み $W_n(t,S_n)$ を計算する.ただし, S_n は学習データにおいて負例に属する文の集合となる.

ここで、ある語 t の正例における重み $W_p(t, S_p)$ が負例にお ける重み $W_n(t, S_n)$ の 2 倍より大きければ、その語 t を素性と して選択する. もしくは、語 t の負例における重み $W_n(t, S_n)$ が正例における重み $W_p(t, S_p)$ の 2 倍より大きければ、その語 t を素性として選択する. 上記の条件を課すことで、正例、負 例における特徴的な語のみを素性として選択し、正例、負例、 ともによく出現するような一般的な語を素性から除去する.上 記の手法により、61,678 文の学習データから 4,549 語が素性 として選択された.

深層学習のモデルについて以下に述べる.入力は 61,678 文 の学習データから抽出された 4,549 語を要素,語 t における $W_p(t, S_p)$,もしくは, $W_n(t, S_n)$ の大きいほうを要素値とし たベクトルとする.モデルの入力層のノード数を入力ベクト ルの次元数と同じ 4,549 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 250 が 3 層、ノード数 125 が 3 層の計 12 層とする.出力層は 1 要素である.また、 活性化関数として、ReLUを使用した.

学習されたモデルにより,結果表現の業績要因を判定する.

4. 評価

本手法の評価を行うため、本手法を実装した.実装にあたり、 形態素解析器として MeCab*5 を使用した.決算短信*6 から 抽出した原因・結果表現、1,446,247 文において、原因表現に 含まれる 662,372 個の名詞 N-gram を「原因となる語」とし て意外性スコアを求める.さらに、意外性スコアが高い原因・ 結果表現の結果表現に対して業績要因の判定を行い、結果表現 に業績要因を含む原因・結果表現を抽出する.表??に、本手法 にて抽出された意外性のある原因・結果表現をいくつか示す.

評価は本手法にて抽出された原因・結果表現を評価者が判 定し,意外性の有無を評価した.意外性の有無を評価すること は評価者の知識にも依存するため困難であるが,ここでは,個 人投資家(すなわち,その分野の専門知識は持たない)を想定 し,投資歴が10年以上ある個人投資家に依頼して評価した. 例えば、原因表現が「多雨によるブラジル出し鉄鉱石の荷動き 鈍化」、結果表現が「ケープサイズ船市況は低迷しました」の 場合,大型船による運搬や鉱物の採掘に携わっている人であれ ば意外性がないかもしれないが, その分野に対する専門知識を 持たない個人投資家にとっては,鉄鉱石の荷動き鈍化→大量の 鉄鉱石を運搬するには大型船が必要→ケープサイズ船の受注 減という関係を思いつくのは困難である.従って、意外性が有 ると判定できる.また,原因が結果に間接的に影響がある場合 は意外性有りと判定した. 例えば, 原因が「猛暑の影響」, 結 果が「飲料用紙容器の販売が増加した」の場合、猛暑→飲料が 売れる→飲料用紙容器の販売が増加のように, 猛暑と飲料用紙 容器の間にもう一つの原因がある.このような場合は機械的に 意外性有りと判定した.逆に、原因が「猛暑の影響」、結果が 「飲料が好調」であれば、原因が結果に直接的に影響があるた め,意外性無しと判定される.なお,本手法でも原因・結果表 現として不適切な表現が抽出されることがあるが、それは意外 性無しと判定した.図1に意外性スコアの上位 N における精 度の変化を示す. ここで,図中の Baseline は結果表現に業績 要因の判定を行わない場合である.

5. 考察

本手法による意外性スコアの上位 70 における精度は 0.64 で あり、ベースラインである結果表現に業績要因の判定を行わな い場合と比べて、精度が大幅に向上した.室野らの手法では、 「猛暑」のような原因となる語を指定し、適切な原因となる語 を指定できれば 80% 程度の精度を達成できたが、原因表現に

^{*4} 再現率は大幅に低下する.

^{*5} http://taku910.github.io/mecab/

^{*6 3,821} 社の企業 Web ページから取得した 106,885 個の決算短信 PDF ファイル

原因表現	結果表現	
患者さんや医療現場に対するドライアイの疾患啓発	角膜疾患治療剤の「ヒアレイン点眼液」は、順調な	
活動	伸びを示しました。	
オンライン請求義務化推進に伴うレセプトコンピュー	医療機関の電子カルテの導入が停滞した	
タ導入に対する補助金の影響		
計量法改正に伴う需要増	メディカル計量器が好調だった	
原発事故に伴う放射能汚染問題、政治情勢の混乱に	建設コンサルタント業界においても、引き続き厳し	
伴う公債発行特例法案の決議遅れによる予算の執行	い状況で推移いたしました。	
抑制などの影響		

表 1: 抽出された原因・結果表現



図 1: 意外性スコア上位 N における精度の推移

含まれる全ての名詞 N-gram を原因となる語として指定すると、不適切な表現が多く抽出されてしまうことが分かる.

しかし,例えば以下のように,ベースライン手法では抽出さ れたが,本手法では結果表現に業績要因が含まれていると判定 されず,除去された原因・結果表現も存在した.

・原因: EUの砂糖制度改革による輸出補助金削減 ・結果:原糖市況は急伸し

本手法では,業績要因の判定に深層学習を用いており,その学 習データを自動生成しているため,多くの語で構成される文が 学習データの正例となる傾向にある.そのため,上記の例のよ うな短い結果表現は,業績要因を含むと判定できないことが 多い.

本手法による誤抽出では,例えば以下のような,原因と結果 に意外性が乏しいものがあった.

・原因:新製品の手術画像用レコーダーの発売が第3四 半期にずれこんだ

・結果:医用画像記録再生機器は減収となりました。

本手法では,結果に「医用画像記録再生機器」のような低頻度 の語が出現すると条件付き確率が低くなり,原因と結果に意外 性が乏しいにもかかわらず意外性スコアが大きくなる.

6. まとめ

本稿では,決算短信から意外性のある原因・結果表現を自動 的に抽出する手法について述べた.本手法は,原因表現にある

名詞が出現するときに結果表現に名詞が出現する条件付き確率と,決算短信から抽出された企業キーワードのスコアを用いて意外性スコアを算出することで意外性の判定を行った. さらに,原因・結果表現の結果表現に対して業績要因の判定を行うことで,不適切な表現を除去した.本手法を評価した結果,本手法の精度は 0.64 となり,ベースライン手法よりも高い精度を達成した.

参考文献

- [和泉 11] 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)
- [北森 17a] 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)
- [Kitamori 17b] Kitamori, S., Sakai, H., and Sakaji, H.: Extraction of sentences concerning business performance forecast and economic forecast from summaries of financial statements by deep learning, in *IEEE Symposium on Computational Intelligence for Financial Engineering & Economics*, pp. 67–73 (2017)
- [藏本 13] 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)
- [室野 17] 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)
- [酒井 15] 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)
- [酒井 17] 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)
- [坂地 15] 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)