決算短信における業績要因・業績結果の因果関係の抽出

Extraction of causal relation between performance factors and performance results from summaries of financial statements

加藤悠太[1]　酒井浩之[1]　坂地泰紀[2]　北島良三[1]　江口潤一[3]

Yuta Kato[1], Hiroyuki Sakai[1], Hiroki Sakaji[2], Ryozo Kitajima[1], Junichi Eguchi[3]

[1] 成蹊大学 理工学研究科 理工学専攻
[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] 大和証券投資信託株式会社
[3] Daiwa Asset Management

Abstract: In this paper, we propose a method for extracting causal relation between performance factors and performance results from summaries of financial statements by deep learning. For example, our method extracts “Sales of copper tubes for air conditioners declined due to high demand due to the hot summer heat caused by the hot summer of the year” as a performance factor, and “Consolidated operating profit was 4,859 million yen” as a performance result corresponding to the performance factor. By extracting such causal relation, it is possible to analyze what kind of factors have changed the performance.

１．はじめに

近年，証券市場における個人投資家の比重が増大しており，個人投資判断を支援する技術の必要性が高まっている．そのため，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[1][4][5].

例えば，決算短信から原因・結果表現を抽出する技術が提案されている[1]. 既存手法[1]では, 例えば, 原因「猛暑」, 結果「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．しかし，既存手法[1]は原因・結果表現を抽出するための手がかり表現と Pattern を定義し，その Pattern に当てはまる原因・結果表現をすべて抽出する手法であるため，原因・結果表現ではあるが投資判断の支援とは無関係な情報も多く抽出されてしまう．

そこで，本研究では，決算短信から業績要因とその業績結果の因果関係の情報を含む文を抽出することを目的とする．例えば，業績要因「一昨年の猛暑の影響による高需要の反動によりエアコン用銅管の売り上げが減少した。」，その業績結果「連結営業利益は４，８５９百万円（前連結会計年度比１９．５％減）に留まりました。」という情報が含まれている文（業績要因文・業績結果文）を抽出する．この情報を大量の決算短信から自動的に抽出することで，ある要因の場合は業績が好調と言った企業を検索することが可能になる．

関連研究として，酒井らは手がかり表現と企業ごとの重要なキーワードを用いて決算短信から業績要因を抽出する手法を提案している[6]．さらに，文献[6]で抽出された業績要因文を用いて学習データを自動生成し，自動生成された学習データからより多くの業績要因文を抽出する手法を提案している[4]. しかし, 酒井らの手法では業績要因を抽出できるが, それに対応する業績結果を取得しておらず，業績要因と業績結果の因果関係を得ることはできない．また，室野らは，既存手法[1]の結果から，意外性のある原因・結果表現を判定する手法を提案している[7][8]. それに対して本研究では, 既存手法[1]の結果を業績要因・業績結果でさらに絞り込み，より高い精度を高めることを目的としている．

本手法では, 既存手法を用いて決算短信から原因・

結果表現を抽出する．その中から，原因表現が業績要因であるもの，結果表現が業績結果であるものを，深層学習を用いて判別する．

２．決算短信からの原因・結果表現の抽出

本手法では既存手法[1]の決算短信 PDF から抽出された原因・結果表現に対し，業績要因であるか，業績結果であるかを判定する．その既存手法[1]である原因・結果表現の抽出の概要を以下に示す．

２.１．原因・結果表現の抽出の概要

Step １：各企業サイトから決算短信 PDF を収集し，収集した PDF をテキストに変換する．
Step ２：得られたテキストデータから，原因・結果判定手法[2]を用い，原因・結果を含む文を抽出する．
Step ３：原因・結果を含んでいると判定された文から，手がかり表現と構文情報を用いた Pattern を使用して原因を示す原因表現と結果を示す結果表現の対を抽出する．

２.２．原因・結果を含む文の抽出

原因・結果を抽出するうえで重要な手がかりとなる表現を利用し，原因・結果を抽出する．

まず，半教師あり学習を用いたフィルタリング手法[2]を適用し，原因・結果を含む文を決算短信 PDF から抽出する．原因・結果を含む文を抽出する手法は SVM を用いるため，表１で説明する素性を使用している．

表 1 素性の一覧

構文的な素性
・助詞のペア
意味的な素性
・拡張言語オントロジー
上記以外の素性
・手がかり表現の直前形態素の品詞
・文に含まれる手がかり表現
・形態素ユニグラム
・形態素バイグラム

２.３．原因・結果表現の抽出

手がかり表現を利用し，原因・結果を含む文集合から原因・結果表現を自動的に抽出する．文献[3]に準拠し，原因・結果表現は出来事とその理由の組みから構成されるとするが，既存研究[1]では 1 文中，または隣り合う 2 文中に直接表現されている表層的なものに限定する．原因・結果表現を抽出するための構文 Pattern や手がかり表現など，具体的な抽出手法については文献[1]を参照されたい．

３．業績要因・業績結果判定のための学習データの生成

本手法では，既存手法[1]によって決算短信から抽出された原因・結果表現，それぞれから，業績要因，業績結果を深層学習にて判定する．深層学習による業績要因，業績結果それぞれの判定モデルを生成するため，学習データが必要となる．その学習データは自動生成するが，その業績要因文の学習データの生成手法[4]を以下に示す．

３.１．業績要因抽出のための手がかり表現

酒井らの手法[6]では，業績要因抽出のための手がかり表現を決算短信から自動的に獲得する．手がかり表現は以下の手法で獲得される．

Step１:少数の手がかり表現(「が好調」,「が不振」)を人手で与え，それに係る節を取得する．
Step２：取得した節の集合から，その中で共通して頻繁に出現する表現（「売り上げ」等）を共通頻出表現として抽出する．
Step３：共通頻出表現が係る節を取得し，その中から新たな手がかり表現を抽出する．
Step４：獲得した手がかり表現から，それに係る共通頻出表現を取得する．
Step５：Step２から Step４を，新たな手がかり表現と共通頻出表現が獲得されなくなる．もしくは，予め定めた回数まで繰り返す．

具体的な手がかり表現の獲得手法については，文献[6]を参照されたい．

３.２．業績要因の学習データの自動生成

獲得された手がかり表現を拡張した手がかり表現（拡張手がかり表現と定義）と，企業ごとの重要なキーワード（企業キーワードと定義）を用い，業績要因文の学習データの自動生成する手法について述べる．企業キーワードの抽出については文献[6]を，拡張手がかり表現については，文献[4]を参照されたい．

Step１：企業の決算短信から既存手法[4]を用いて業績要因文を抽出する．
Step２：抽出した業績要因文に対し，企業キーワードを含みかつ拡張手がかり表現を含む文を抽出し，学習データの正例とする．
Step３：企業の決算短信から，企業キーワードと手がかり表現，どちらも含まない文を抽出する．
Step４：Step３で抽出した文のうち文字数が一定以上の文を学習データの負例とする．

上記の処理を行うことにより，業績要因文の学習データを生成した．学習データは正例 12,2447 文，負例 12,2447 文の合計 244,894 文が生成された．

３.３．業績結果の学習データの自動生成

業績結果の学習データの生成手法[5]を以下に示す．ランダムに選んだ,000 社の決算短信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

「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

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。」である．

以上の条件により業績結果文の学習データを生成する．学習データは正例 13,158 文、負例 339,198 文の合計 352,356 文が生成された．

４．業績要因・業績結果の抽出

本手法の概要を以下に示す．

Step１：決算短信から文を抽出する．
Step２:抽出された文から既存手法[1]により原因・結果表現を抽出する．
Step３：抽出された原因表現を業績要因文の学習データによって構築された深層学習モデルにより，業績要因の判定を行う．
Step４：抽出された結果表現を業績結果文の学習データによって構築された深層学習モデルにより，業績結果の判定を行う．
Step５：原因表現が業績要因と判定され，かつ，結果表現が業績結果であると判定されたもののみを業績要因・業績結果として抽出する．

業績要因・業績結果の抽出手法の概要を図 1 に示す．

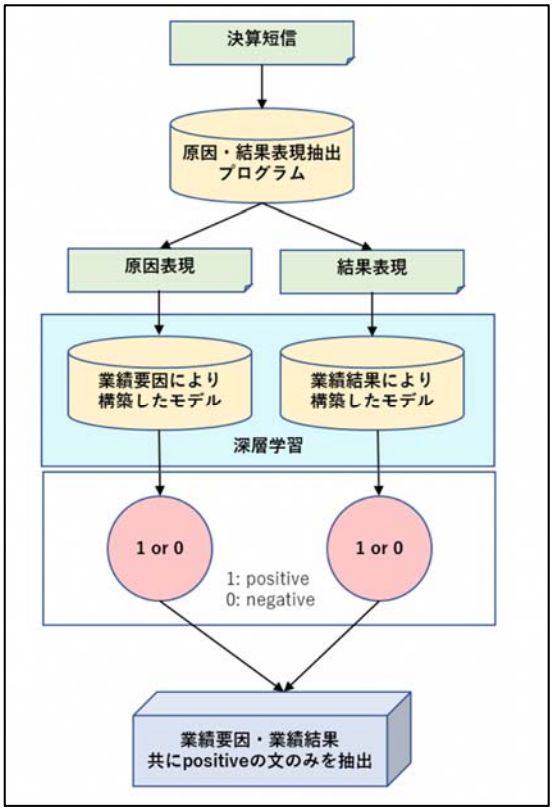

図 1　業績要因文・業績結果文の抽出の概要

４.１．業績要因・業績結果の判定

本手法では，業績要因，業績結果の判定を深層学習モデルにより行った．深層学習モデルには多層パーセプトロンを採用した．学習データから文献[4]の手法により素性を抽出し，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の数と同じとした．中間層のノード数は，ノード数 1,000 の隠れ層が 3 層，ノード数 500 の隠れ層が 3 層，ノード数 200 の隠れ層が 3 層，ノード数 100 の隠れ層が 3 層の計 12 層とする．出力層はノード数 1 とし，エポック数は 50 として学習を行った．

本手法により業績要因・業績結果として抽出された因果関係の例を表２に示す．

表 2 業績要因・業績結果の因果関係の例

企業名	
業績要因	業績結果
明治海運	
消費増税後の駆け込み需要の反動減からの回復が当初の予想よりも遅れていることや、一部の季節商品の販売不振が続く	日曜雑貨事業の売上高が 17 億 7 千 6 百万円減少した
昭和電工	
リチウムイオン電池材料はスマートフォン向けに加え車載向けの出荷が増加し増収となりましたが、昭光通商株式会社は減収となりました	当セグメントの売上高は６９９億６６百万円となりましたが、営業利益は主にリチウムイオン電池材料の数量増により１０億４６百万円となりました
コマツ	
【産業機械・車両他】産業機械・車両他部門では、産業機械事業およびフォークリフト事業が堅調	売上高は１，４７８億円となり、当部門のセグメント利益は１３９億円、売上高セグメント利益率は９．４％となりました

５．評価

本手法により抽出された業績要因・業績結果の因果関係の評価は, 上場企業 2663 社の決算短信を対象にして行った.

比較手法とし，深層学習で判定された業績原因に対し，企業キーワードでフィルタリングをかけての判定，結果表現に対し,「円」が含まれている場合を業績結果として判定を行った場合を比較した. また，ベースラインとして，本手法による業績要因・業績結果の判定を行う前（すなわち，既存研究[1]による抽出のみ）の場合の結果を評価した.

評価用のデータは上場企業 2663 社の決算短信から各企業の原因･結果表現 1 組ずつ抽出し, 合計 2663 の評価セットにおいて，人手にて業績要因・業績結果を判定し，正解データを作成した．それぞれの手法での精度，再現率を表３に示す.

表３　業績要因・業績結果判定の精度・再現率

手法 （業績要因 / 業績結果）	精度 (%)	再現率 (%)
ベースライン	3.81	100
深層学習モデル / 深層学習モデル	83.33	38.46
深層学習モデル / 「円」で判定	86.04	71.15
深層学習モデル + 企業キーワードでフィルタリング / 深層学習モデル	85.71	34.62
深層学習モデル + 企業キーワードでフィルタリング / 「円」で判定	91.55	62.5

６．考察

深層学習モデルにより判定した結果，既存手法[1]と比べ，はるかに精度が上がった。最も精度が高かったものは，業績要因を深層学習モデルにより判定後，企業キーワードでフィルタリングをかけ，業績結果は「円」で判定をした場合であった．しかし，企業キーワードでフィルタリングをかけないものに比べるとやや再現率が落ちてしまった．また，企業キーワードで絞り込みをかけてしまうと新規事業などの現状の企業キーワードにない単語が出てきてしまう場合に対応できない．そのため，多層パーセプトロン以外の手法で，文中の出現単語の前後関係を考慮するような時系列データを処理することが可能な Recurrent Neural Networks の一種である LSTM(Long Short Term Memory)ネットワークを用い精度の向上を試みた. しかし, 精度が最大で 84.85%, 再現率が 32.69%と落ち込む結果となってしまった. これは素性の選択やモデルの構築が決算短信の文と合っていない可能性が考えられる．よって様々な素性での学習やモデルの構築をし，精度，再現率の向上を図る予定である.

本研究では業績要因・業績結果を決算短信から抽出したが，業績要因・業績要因である原因・結果表現や，数は少ないが業績結果・業績要因である因果関係も存在し，そのような情報も投資判断にとって有用であると考える．今後は，業績要因・業績要因である原因・結果表現や，業績結果・業績要因である原因・結果表現の抽出も行う予定である.

７．まとめ

本稿では，決算短信の原因・結果表現から業績要因・業績結果を抽出する手法を提案した．具体的には，業績要因の判定用深層学習モデルと業績結果の判定用深層学習モデルをそれぞれ作成し，両モデルが業績要因・業績結果と判定された原因・結果表現のみを業績要因・業績結果の因果関係であると判別した．評価の結果，精度は 83.33%，再現率は 71.15%と，高い精度で抽出することができた．

参考文献

[１] 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.

[２] 坂地泰紀，増山繁: “新聞記事からの因果関係を含む文の抽出手法”，電子情報通信学会論文誌 D, vol. J94-D, No. 8, pp.1496–1506, (2011)

[３] 庵功雄: “新しい日本語学入門(第 2 版)”，スリーエーネットワーク, (2012)

[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] 村野壮人，酒井浩之，坂地泰紀，江口潤一, “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”，第 19 回金融情報学研究会, pp.59-64, 2017.

[６] 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.

[７] 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.87-91, 2017.

[８] Hiroki Sakaji, Risa Murono, Hiroyuki Sakai, Jason Bennett, Kiyoshi Izumi, “Discovery of Rare Causal Knowledge from Financial Statement Summaries”, IEEE Symposium on Computational Intelligence for Financial Engineering & Economics (IEEE CIFEr'17), Hawaii, November, 2017.