金融機関のテキストデータに基づく企業の業績要因の抽出

Extraction of information on performance factors from financial institution texts

近藤 浩史[1]　大沼 俊輔[1]　中込 祐平[2]　遠藤 公志郎[2]

三橋 尚文[2]　佐藤 雪子[2]　酒井 浩之[3]

Hirofumi Kondo[1], Shunsuke Onuma[1], Yuhei Nakagome[2], Koshiro Endo[2]

Naofumi Mitsuhashi[2], Yukiko Sato[2] and Hiroyuki Sakai[3]

[1]株式会社日本総合研究所

[1]The Japan Research Institute, Limited

[2]株式会社三井住友銀行

[2] Sumitomo Mitsui Banking Corporation

[3]成蹊大学

[3]Seikei University

Abstract: For financial institutions, it is important to monitor the performance and performance factors of corporate customers. Financial institutions accumulate large amounts of data to monitor customer information including their performance. It is relatively easy to extract information on performance from structured data, while it is difficult for non-structured data like text data. In this research, we tried to extract sentences that represent customer's performance factors from text data created by financial institutions. Our proposed method using deep learning extracts sentences similar to sentences prepared as correct examples in advance. Due to the difference in properties of text data, our method showed better performance than the prior research.

１．はじめに

金融機関にとって取引先企業の業績および業績要因を把握することは重要である．そのため，企業の公開情報（決算書など）の確認や，企業へのヒアリングを通して，業況の把握に努めている．その過程において，様々な資料が作成され，大量のデータが保管されている．

大量のデータから業績に関する情報を抽出する場合，構造化されたデータに対しては比較的容易である一方，テキストデータのような非構造データに対しては難しいケースがある．一般に，テキストデータの検索ではキーワード検索が用いられるが，キーワードに一致した文が必ずしも業績を述べた文ではないこともある．そのため，予め大量のデータから業績文や業績要因文のみを抽出しておく必要がある．

そこで，本研究では金融機関の社員が企業との面談を通して作成した大量のテキストデータ（以下，企業内テキストと記載）から，業績文および業績要因文の抽出を試みる．

近年，テキストアナリティクス技術の進展により，テキストデータから企業の業績要因文等の抽出手法が研究されてきた．先行研究[１]では，決算短信 PDF から業績要因文を抽出する手法を提案し，良好な精度を得ている．その他，先行研究[２][３][４]のように，決算短信 PDF から原因・結果表現の抽出や，業績予測文の抽出も研究されている．また，先行研究[５]では有価証券報告書から因果関係文を抽出している．

先行研究では企業の公開情報を扱っているが，本研究では金融機関内にあるテキストデータを対象とする点が異なる．決算短信等は文書の性質上，記載すべき内容がある程度固定されており，文書内に業績を述べた文が存在する．一方，企業内テキストは内部文書につき，記載内容は多種多様である．そのため，先行研究の手法を企業内テキストに適用すると，業績要因文が高精度に取得できない恐れがある．

そこで，本研究では深層学習により業績要因文を抽出する手法（以下，深層学習による手法と記載）を提案する．本手法は予め決めた業績要因文の正解

データと類似する文を抽出する手法である．また，本研究では，比較のため先行研究[１]を再現した手法（以降，先行研究手法と記載することがある）も実装し，企業内テキストに対して適用する．

なお，深層学習による手法では，正解データの準備が必要であり，人手で準備するには手間がかかる．本研究では内閣府が公表する景気ウォッチャー調査[６]の景気判断理由集を活用し，正解データ作成の省力化にも取り組んだ．

結果として，文章の性質の違いから，先行研究[１]を再現して適用するよりも，深層学習による手法が良い性能で業績要因文を抽出できることが分かった．

２．業績要因文の抽出手法

まず決算短信PDFから業績要因文を抽出する先行研究[１]の概略について述べ，次に深層学習による手法について述べる．

２．１．先行研究の概略

先行研究[１]は，文中の「手がかり表現」と「企業キーワード」に着目して，業績要因文を抽出する手法を提案している．

手がかり表現は，[１]によると「業績要因となる状況，状態，変化を表す用言的な表現」と定義される．「堅調だった」「低迷した」等が具体例であり，業績要因文によく含まれる表現と解釈できる．

企業キーワードは，ある企業の決算短信 PDF に含まれる名詞のうち，企業にとって重要な名詞のことである．[１]から具体例を挙げると，ソニーの企業キーワードとしては「液晶テレビ」が挙げられる．

企業キーワードは名詞の TF-IDF およびエントロピーを基準にして抽出される．抽出基準を直感的に説明すると「他企業にはあまり出現しないが，ある企業にはよく出現し，かつ，当該企業の決算短信のそれぞれの文書に偏りなく出現する単語」となる．

業績要因文は，手がかり表現と企業キーワードの２つを用いて抽出される．まず手がかり表現を含む文を検索し，手がかり表現に係る節に企業キーワードを含む文を業績要因文として抽出する．[１]によると，ソニーの例では，企業キーワード「液晶テレビ」，手がかり表現「減少した」を元に，「この大幅な減収は，主に液晶テレビの販売台数が大幅に減少したことによるものです」という業績要因文が取得される．

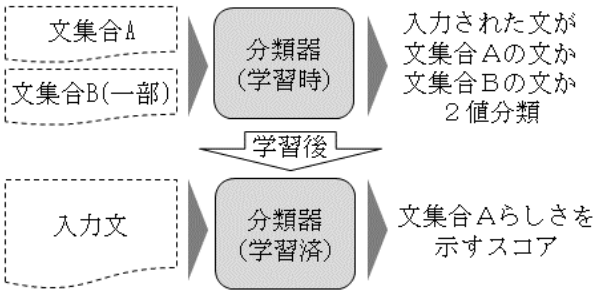

図１：分類器の構築

２．２．深層学習による手法

２．２．１　対象データの絞り込み

先行研究[１]を参考に，深層学習による手法においても「手がかり表現」を含む文を対象に業績要因文を抽出する．分析対象となる文書の性質が変わったとしても，業績要因文には「堅調だった」といった表現がよく含まれると考えられるためである．

手がかり表現の抽出手法は先行研究[１]と同様とし，手がかり表現を文中に含む文を抽出することで対象データを絞り込む．

２．２．２　業績要因文の抽出器の構築

抽出器は文集合 A と文集合 B が与えられたときに，文集合 A の類似文を文集合 B から抽出するモデルである．抽出器は以下に示す 4 つのテキスト分類器をアンサンブルして構成する．

① TF-IDF/LR モデル

学習で使用した文に含まれる単語の TF-IDF を特徴量として，ロジスティック回帰で文章を分類するモデル．

② CNN/NN モデル

先行研究[７]をベースとしたモデル．

③ 双方向 LSTM/NN モデル

双方向 LSTM（以降，BiLSTM と記載）とニューラルネットワークを組み合わせたモデル．

④ SWEM/LR モデル

先行研究[８]に SWEM-concat と記載されたモデル．SWEM-concat の特徴量を文の特徴量として，ロジスティック回帰で文章を分類するモデル．

以上 4 つの分類器を使用し，抽出器を以下①②の手順で構築する．

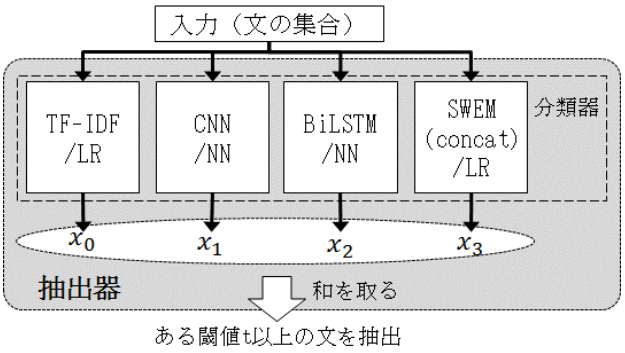

図２：抽出器の概略図

①　各分類器の学習（図１）

各分類器は入力文の特徴量を生成し，入力文が文集合 A の文か，文集合 B の文かを 2 値分類する．各分類器は同じ学習データから，独立に学習させる．学習済み分類器は入力文の文集合 A の文らしさを示すスコアを出力する．

なお，各分類器の学習時には，文集合 B の全体を学習データとするのではなく，文集合 B からサンプリングした文集合 B の一部を用いる．これは，各文集合に含まれる文数の偏りを小さくするためである．

②　抽出器の構築（図２）

学習済みの分類器をアンサンブルする．抽出器は入力された文に対して，各分類器が出力する文集合 A の文らしさを示すスコアの和を集計し，予め決めた閾値よりも大きな値の文を抽出して出力する．

このように構築した抽出器に対して，文集合 B の全文を入力し，抽出器から出力された文は，文集合 A に似た文章と言うことができる．

本研究では文集合 A が抽出すべき業績要因文の正解データで，文集合 B が企業内テキスト（手がかり表現含む）となる．

２．２．３　正解データの作成

企業内テキストから正解データを多数作成するためには，大量の文を読む必要があり，手間がかかる．

そこで，内閣府が公表する景気ウォッチャー調査の景気判断理由集を活用し，以下①②の手順で正解データを作成した．景気判断理由集を使用した理由は，景気判断理由集の文に，企業の業績要因文と似た文が存在するからである．例えば「観光客が増えているが、お盆時期の天候不順の影響で見込みほど売上が伸びなかった。」といった文が含まれている．

なお，企業内テキストのうち，景気ウォッチャーの景気判断理由集の文と類似する文を景気ウォッチャー類似文と記載する．

①　景気ウォッチャー類似文の抽出

前節で説明した抽出器の仕組みを用いて，景気ウォッチャー類似文を抽出した．具体的には，前節で説明した文集合 A を景気ウォッチャー調査の景気理由判断集の文，文集合 B を企業内テキストとして抽出器を構築した．

②　景気ウォッチャー類似文に対するラベル付け

景気ウォッチャー類似文は厳密には業績要因文ではない．そのため，抽出済みの景気ウォッチャー類似文の各文に対して，業績要因文の正解データとするかどうかをラベル付けした．

景気ウォッチャー類似文（業績要因文に類似する文）に対してのみラベル付けを実施することで，正解データの作成が省力化された．

３．実装

３．１．テキストの前処理

企業内テキストは 2010 年 1 月～2018 年 8 月に作成されている．企業内テキストは複数の文書から成り，文書をさらに文に分割して使用する．文を単語に分割する際には，MeCab を使用し，辞書として mecab-ipadic-NEologd[9]と独自の金融用語辞書を組み合わせて使用した．また，意味ある文を構成しないと想定される短文（動詞，名詞，形容詞の合計が 5 単語以下の文）は事前に除去した．このようにして全体で約 5,000 万件の文を得た．以降，前処理済みの企業内テキストを抽出対象文と呼ぶ．

本研究で使用した，景気ウォッチャーの景気判断理由集は 2013 年 1 月～2018 年 11 月に公開されたデータである．企業内テキストと同様に前処理を実施するが，短文は除去していない．以降，前処理済みの景気ウォッチャー調査の景気判断理由集の文を調査文と呼ぶ．

３．２．先行研究の手法

手がかり表現は，[１０]を Python に移植したプログラムを元にして，調査文から取得した．調査文を使用した理由は，前述のとおり業績要因文に近い表現で記載されているからである．

景気判断理由集から手がかり表現を抽出すると，不適切と考えられる手がかり表現が抽出された．例えば「多い」「少ない」等の一般的な形容詞が挙げられる．人手で不適切な手がかり表現を除去し，結果として 226 件の手がかり表現を得た．

キーワード抽出および業績要因文の抽出は[1]を元に実装した．本実装を用いて抽出対象文から業績要因文を得た．

３．３．深層学習による手法

３．３．１　対象データの絞り込み

抽出対象文から手がかり表現を含む文を絞り込む際には, 3.2 節で述べた先行研究と同じ手がかり表現を用いた．抽出対象文から手がかり表現を含む文を抽出したところ，約 170 万件の文を得た．

３．３．２　正解データの作成

2.2.3 節で述べた正解データ作成の際には，調査文および抽出対象文から, それぞれ 3 万文（計 6 万文）をランダムにサンプリングして，景気ウォッチャー類似文の抽出器を構築した．

取得した景気ウォッチャー類似文のうち，24,057 文に対して人手でラベル付けを行い，業績要因文 16,568 文と非業績要因文 7,489 文を得た．

３．３．３　抽出器の構築

業績要因文の抽出器の構築では，表 1 に示すデータを使用した．

人手でラベル付けしたデータに加えて，調査文および抽出対象文からランダムにサンプリングした文を追加した. 調査文と抽出対象文を追加する理由は, 業績要因文と非業績要因文の特徴を幅広にモデルが学習できるようにするためである．

文集合 A には抽出すべき業績要因文の正解データを設定するため，人手でラベル付けした業績要因文と調査文を加えた．また，文集合 B には非業績要因文を設定するため，人手でラベル付けした非業績要因文と抽出対象文を加えた．なお，文集合 B に加える抽出対象文はランダムにサンプリングするため，業績要因文を含む可能性があるが，件数が少ないと想定されるため問題にはならない．

構築した抽出器を用いて，抽出対象文（手がかり表現含む）から業績要因文を抽出した．なお，抽出器の閾値は 1.0 とした．

表１：抽出器の構築に使用したデータ件数

文集合	人手でラベル付け	ランダムに追加	合計
A	16,568	15,875（※１）	32,443
B	7,488	24,390（※２）	31,878

※１：調査文から取得

※２：抽出対象文から取得

４．評価

４．１．抽出精度の評価

抽出した業績要因文から 1,000 文をランダムにサンプリングし，人手で業績要因文であるかどうかをラベル付けして抽出精度を評価する．ラベル付けは「業績要因文」「業績関連文」「間違い」の 3 値分類とした．なお，業績関連文とは「猛暑の影響で客足が遠のいている」といった遠回しに業績を述べた文を示す．

表２に深層学習による手法と先行研究手法の抽出精度を記載した．深層学習による手法の精度は 65.3%（業績関連文と合わせると 85.6%）であり，比較的高い精度で業績要因文を抽出できた．

表２：ラベル付け結果の件数と割合

	深層学習による手法	先行研究手法
業績要因文	653 (65.3%)	284 (28.4%)
業績関連文	203 (20.3%)	108 (10.8%)
間違い	144 (14.4%)	608 (60.8%)

４．２．再現率の評価

再現率は, ランダムにサンプリングした 10 社を元に評価する. 本来は全データを元に評価すべきだが, 業績要因文の全量を把握することが困難なためサンプリング評価する. サンプリングした 10 社の抽出対象文（手がかり表現含む）および抽出した業績要因文を，人手でラベル付けして評価した．ラベル付けは精度評価と同様に 3 値分類とした．

表３はサンプリングした 10 社において，再現率・精度・F 値を算出した結果である．深層学習による手法では高い再現率を達成している．

表３：サンプリングした 10 社に対する再現率・精度・F 値の評価

	業績要因文のみ		業績要因文 + 業績関連文	
	先行研究手法	深層学習による手法	先行研究手法	深層学習による手法
再現率	0.72	0.92	0.68	0.86
精度	0.33	0.55	0.48	0.80
F 値	0.45	0.69	0.56	0.83

４．３．考察

再現率，精度ともに深層学習による手法が先行研究よりも高い性能を発揮している．ただし，表３によると，先行研究手法も，精度は 0.33 と低いが，再現率は 0.72 と比較的良好な結果となっている．

この原因は企業キーワードの取得にあると考える．企業内テキストは，決算短信 PDF とは異なり，企業の業績を述べた文だけでなく，様々なトピックを扱っている．そのため，企業の商品名や部門名が企業キーワードとして抽出されるだけでなく，企業内テキストに良く記載されている金融商品に関する事項・書類名・地名なども抽出されやすい．

したがって，抽出した企業キーワードの中には業績要因文に入るべきキーワードも含んでいるために，再現率は比較的高くなるが，全く関係ないキーワードも含むため精度が低下すると考えられる．

一方，深層学習による手法は，企業内テキストの文の特徴が，予め準備した正解データの文の特徴と類似する文を抽出するため，企業キーワードに依存せずに業績要因文を抽出できたと考える．

５．まとめ

本研究では企業内テキストデータを対象に，深層学習により業績要因文を抽出する手法を提案した．企業キーワードに着目して業績要因文を抽出する先行研究よりも，提案手法が良い性能を示すことが分かった．これは企業内テキストが業績に関する文だけでなく，様々なトピックを使った文が多く含まれるためである．

本研究が提案した深層学習による手法は企業キーワードによらず業績要因文を抽出できるが，正解となる業績要因文の作成に手間がかかることが欠点と考えられる．本研究では，景気ウォッチャー調査の景気理由判断集を活用することで，正解データの作成も大幅に省力化することができた．

今後の課題としては更なる性能向上のために抽出モデルを工夫することなどが挙げられる．

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] 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，坂地 泰紀，室野 莉沙，北島 良三，ベネット ジェイスン：意外性のある原因･結果表現の決算短信からの抽出，第 32 回人工知能学会全国大会，(2018)

[５] 佐藤 史仁，佐久間 洋明，小寺 俊哉，田中 良典，坂地 泰紀，和泉 潔：有価証券報告書からの因果関係文の抽出，第 32 回人工知能学会全国大会，(2018)

[６] 内閣府 景気ウォッチャー調査 https://www5.cao.go.jp/keizai3/watcher/watcher_menu.html （2019/1/28 アクセス）

[７] Yoon Kim : Convolutional Neural Networks for Sentence Classification，EMNLP2014，(2014)

[８] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao and Lawrence Carin : Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms, ACL2018, (2018)

[９] 佐藤敏紀，橋本泰一，奥村学：単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討，言語処理学会第 23 回年次大会，(2017)

[１０] 手がかり表現自動獲得プログラム (CLue Phrases Extraction Software), https://www.ci.seikei.ac.jp/sakai/clupes.html (2019/01/28 アクセス)