極性と重要度を考慮した決算短信からの業績要因文の抽出

大村 和正 [1]　白井 穂乃 [2]　石原 祥太郎 [2]　澤 紀彦 [2]
[1] 京都大学大学院情報学研究科　[2] 株式会社日本経済新聞社
omura@nlp.ist.i.kyoto-u.ac.jp
{hono.shirai,shotaro.ishihara,norihiko.sawa}@nex.nikkei.com

概要

本稿では，極性と重要度を考慮した決算短信からの業績要因文の抽出手法を提案する．提案手法は 2 段階の学習データの自動生成から成り，重要度を考慮した要因分類器の学習データと極性分類器の学習データを決算短信から生成する．これらの自動生成データにより，業績要因文の抽出精度の改善と高精度な極性付与ができることを示す．また，中規模な評価データを人手で構築し，この再現ができるように必要な情報を公開する．

1　はじめに

ウェブ上には日々膨大なテキストデータが蓄積されており，それらを活用するための技術の需要が高まっている．その中でも，金融ドメインのテキストを対象としたマイニング手法は，投資支援や経済分析への応用が期待されることから近年盛んに研究されている [1]．本研究では，このような金融テキストマイニングの 1 タスクである「決算短信からの業績要因文の抽出」に取り組む．

決算短信とは，上場企業が決算発表を行う際に開示する，当期の経営成績等をまとめた書類である (表 1 上段)．決算短信は企業動向をいち早く報じるものであるため，投資判断に欠かせない情報源である一方，文章量が多く要点の把握には労力を要する．このため，投資判断の参考になる「業績変化の要因が記述された文 (**業績要因文**)」を自動抽出することができれば，投資支援として有用であると考えられる．このような背景のもと，業績要因文の抽出に向けた手法は複数提案されてきた [2, 3, 4, 5, 6]．

決算短信は原則再配布が認められていないためにオープンデータがなく，注釈付けも多大な労力を要することから，いかに機械学習モデルの学習データを自動生成するかがひとつの争点となっている．例えば，酒井らは業績要因の手がかりとなる表現と各企業にとって重要なキーワードを自動獲得し，それらをベースに学習データを生成している [5]．この手法は工程が若干複雑である点に難がある．これに対し，大村らは決算短信の要約記事である業績発表記事 (表 1 下段) を利用した，簡素なデータ生成手法を提案している [6]．この手法も決算短信に含まれる業績と無関係な記述[1)]が負例であることを十分に学習できないため，精度に改善の余地がある．

本研究では大村らの手法 [6] を拡張し，学習データの生成を 2 段階にすることで，前述の問題に対処する．具体的には，業績発表記事から生成した学習データで分類器を訓練し，それを用いて決算短信に擬似ラベルを付与することで，よりタスクに適応した学習データを生成する．これにより，業績要因文の抽出精度が改善することを示す．

また，この拡張に乗じて，実応用での需要が高い極性[2)]や重要度の付与が可能となるようにデータ生成を工夫する．具体的には，疑似ラベルを付与する際に重要度を考慮したスコアを付与し，これと並行

表 1　ある上場企業の決算短信と，それに対応する業績発表記事の例（抜粋）．

決算短信	… 百貨店業での営業収益は 398,338 百万円 (前年同期比 31.4%減)、営業損失は 16,863 百万円 (前年同期は営業利益 6,563 百万円) となりました。 国内百貨店におきましては、新型コロナウイルス感染症の拡大に伴う緊急事態宣言の発出を受け、全店で食料品フロアを除く臨時休業を実施しました。5 月末には全店で全館営業を再開いたしましたが、多くのお客様の来店を見込んだ営業施策や販売促進策の中止や開催方法の見直しをしたことに加え、外出を控える動きも依然強く、売上高は大きく減少いたしました。また、渡航制限で …
業績発表記事	… が 25 日発表した 2020 年 3〜11 月期の連結決算で、最終損益は 243 億円の赤字 (前年同期は 164 億円の黒字) だった。新型コロナウイルスの感染拡大に伴う外出自粛やインバウンド (訪日外国人) 需要の大幅な落ち込みによる販売減少が響いた。 …

1)　「〜は以下のとおりとなりました。」といった注記など．
2)　その要因による売上高や利益の増減の向きを指す．

して極性分類器の学習データを生成する．これにより，重要度の予測は改善の余地があるものの，高精度な極性付与ができることを示す．極性や重要度を考慮する手法はこれまでにも個別に提案されてきた [7, 8, 9] が，総合的に扱う点が差分である．

本研究のもう一つの貢献として，中規模な評価データを人手で構築し，この再現に必要な情報を公開する[3]．評価データはそのまま公開できないため，注釈と決算短信の取得元の情報から評価データが再現できるように整備を進める．

2 タスク設定

本研究の対象を明確にするため，決算短信に含まれる文を以下のように分類定義する．

業績文 当期の業績変化のみを述べた文

暗黙的な業績要因文 文内で当期の業績変化と紐付いていないが，要因であると判断される文

明示的な業績要因文 文内で当期の業績変化とその要因が述べられた文

その他 上記のいずれにも該当しない文

例えば，表 1 上段において，1 文目は業績文である．また，2 文目は「全店で食料品フロアを除く臨時休業を実施」という要因が述べられているものの，業績変化と紐付いていないため，暗黙的な業績要因文である．さらに，3 文目は「売上高は大きく減少」という業績変化とその要因が述べられているため，明示的な業績要因文である．

本研究では，決算短信に含まれる各文が「明示的または暗黙的な業績要因文であるか否か」を判定する 2 値分類タスクとして定式化する．なお，財政状態や将来予測に関する記述 [10, 11] は対象外とする．以降，この 2 値分類タスクを「要因分類」，明示的または暗黙的な業績要因文を「正例」，そうでない文を「負例」と呼ぶ．

3 提案手法

提案手法は図 1 のように 2 段階の学習データの自動生成から成る．

3.1 第 1 段階

大村らの手法 [6] に従って業績発表記事から学習データを生成する．具体的には，業績発表記事から数字を含む文を業績文 (負例)，含まない文を暗黙的

3) https://github.com/omukazu/ecifr

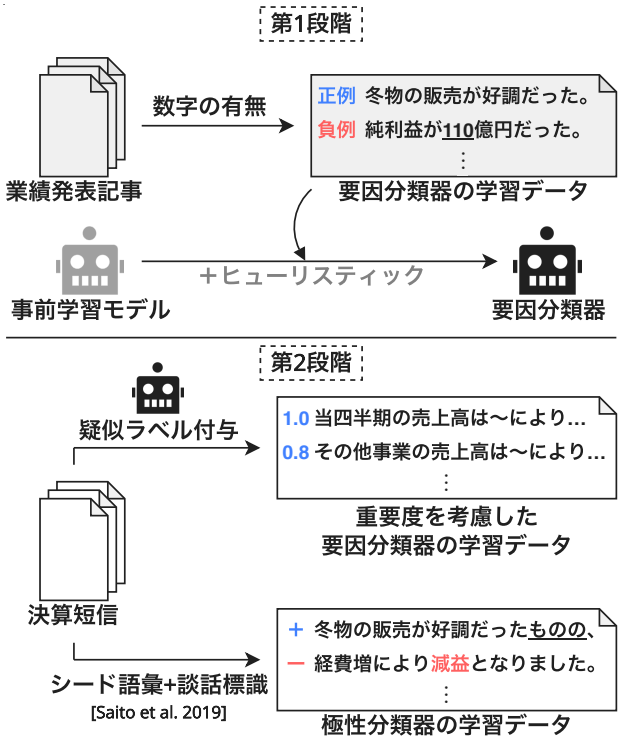

図 1 提案手法の概要図．

な業績要因文 (正例) として抽出する．また，記事の第 1 段落から抽出した暗黙的な業績要因文と業績文を適当な接続詞でつなぐことで明示的な業績要因文を生成し，これを正例として加える．

ヒューリスティックの導入 決算短信には業績と無関係な記述[1]が多く含まれる．そのような記述によく見られる文体の文が負例であると学習させるため，以下のいずれかの条件を満たす文のラベルを「負例」に上書きする．

- 末尾の文節が過去形または体言止めでない
- 末尾の節の意志性が強い

1 点目は将来予測に関する記述等を，2 点目は「〜に努めました。」のような定性的だが業績変化との因果関係が曖昧な文を想定したものである．

最後に，生成したデータを用いて，入力文が正例である確率を予測する要因分類器を訓練する．

3.2 第 2 段階

重要度を考慮した要因分類器の学習データと極性分類器の学習データを決算短信から生成する．

重要度を考慮した要因分類器の学習データ 第 1 段階で訓練した要因分類器を使用し，決算短信から抽出した各文に重要度を考慮したスコア ($\in [0, 1]$) を付与することで生成する．手順を以下に示す．

1. 要因分類器の予測確率が 0.8 以上である文を正例，0.2 以下である文を負例として抽出し，それ以外の文を除外する．
2. 負例の文はスコア 0 をラベルとする．
3. 正例の文は，それが抽出元の決算短信において冒頭から i 番目の正例の文であるというメタ情報をもとに，以下の計算式で算出されるスコアをラベルとする．

$$\text{score}(i) = \max(1.0 - 0.02 \times i, 0.8)$$

決算短信には「企業全体の概況 → 主力事業の概況 → その他事業の概況」という談話構造がよく見られ，冒頭に近いほど重要な要因である可能性が高いことを 3 の計算式に反映している．

極性分類器の学習データ　Saito らの手法 [12] を参考に以下の手順で生成する．

1. 「増収」や「減益」など，業績変化の極性を強く表すシード語彙を人手で策定する．
2. シード語彙を含み，それと「原因・理由」または「逆接」の談話関係を表す談話標識で接続される節の組を取得する．
3. シード語彙を含む節はそれにもとづく極性を，それと接続される節は談話関係にもとづいて伝播される極性をラベルとする．

最後に，生成した 2 つのデータを用いて，入力文のスコア (=重要度) を予測する要因分類器と極性分類器を訓練する．極性と重要度を同時に予測する分類器の構築は今後の課題である．

4　実験

提案手法の有効性を検証するために実験を行う．オープンデータは存在しないため，評価データは人手で構築し，学習データは自動生成する．

4.1　評価データの構築

まず，株式会社日本経済新聞社が提供するニュース配信サービス「日経電子版」[4] における業界分類 [5] に従って，各業界 10 企業ずつ計 150 件の決算短信を人手で収集した．対象期間は 2021 年 4 月から 2022 年 4 月までの 1 年間とした．

続いて，各決算短信 PDF から四半期決算に関する定性的情報が記述されたテキストを抽出し [6]，正規表現ベースの文分割処理 [7] を適用した．こうして得られた各文に著者らが人手で注釈付けを行い，3 人以上の合意があったラベルに集約した．また，正例と評価された文はその極性と重要度も付与した．極性はポジティブ・ネガティブ・判別不能のいずれかを，重要度は決算短信 1 件につき最大 3 文まで重要文を主観で決め，重要文であるか否かを付与した．

最後に，2:8 に分割して開発データとテストデータを構築した．評価データの統計を表 2 に示す．この再現に必要な情報は公開する．

表 2　評価データの統計．

ラベル	極性	開発	テスト
正例	ポジティブ	188	908
	ネガティブ	49	351
	判別不能	15	70
負例		380	2,076

表 3　学習データの統計．要因分類器の学習データは負例に偏っていたため，5 万文ずつに絞っている．また，「正例」は 0.8 以上のスコアが付与された文を指す．

データ名	ラベル	数
重要度を考慮した要因分類器の学習データ	正例	50,000
	負例	50,000
極性分類器の学習データ	ポジティブ	8,488
	ネガティブ	6,732

4.2　学習データの生成

業績発表記事の取得　提案手法の第 1 段階で使用する業績発表記事は「日経電子版」[4] から取得した．具体的には，2018 年 1 月から 2021 年 1 月までの 3 年間を対象とし，メタデータのトピック情報に「企業決算」ラベルが付与されているものを取得した．この結果，上場企業 1,023 社の業績発表記事を計 3,322 件取得した．また，総文数は 43,893 文であった．

決算短信の取得　ベースラインおよび提案手法の第 2 段階で使用する決算短信はウェブから自動収集した．業績発表記事の取得時と同様に，2018 年 1 月から 2021 年 1 月までの 3 年間を対象とした．この結果，上場企業 3,653 社の決算短信を計 31,771 件取得した．取得された PDF に対し，4.1 節と同様に文を抽出した結果，総文数は 574,000 文であった．

提案手法　3 節の手法に従って学習データを生成した．過去形または体言止めであるか否かの判別と節間の談話関係の認識は，構文解析器 KNP [8] [13] が

4）https://www.nikkei.com
5）https://www.nikkei.com/help/markets/helpindex.html#gyoshu-needs
6）詳細は付録を参照されたい．
7）https://github.com/ku-nlp/python-textformatting
8）https://github.com/ku-nlp/knp

表 4　テストデータに対する実験結果．異なる 3 つのシード値で fine-tuning した結果の平均と標準偏差を記載している．「ヒューリスティックのみ」は，3.1 節の条件を満たさない文を全て「正例」とした時の精度である．

設定		適合率	再現率	F 値	正解率 (重要文)
ヒューリスティックのみ		60.2	83.6	70.0	—
ベースライン [5]		62.8 ± 6.0	**86.6 ± 3.7**	72.6 ± 3.4	27.7 ± 5.1
	+ ヒューリスティック	81.6 ± 1.6	78.7 ± 4.4	80.0 ± 1.8	31.6 ± 1.8
提案手法	第 1 段階	47.6 ± 2.5	84.0 ± 6.4	60.6 ± 0.9	16.4 ± 1.8
	第 1 段階 + ヒューリスティック	79.4 ± 4.9	78.9 ± 4.3	79.0 ± 0.5	29.6 ± 2.0
	第 2 段階 (要因分類)	**84.0 ± 2.3**	85.3 ± 1.7	**84.6 ± 1.7**	**40.0 ± 2.9**
	第 2 段階 (極性分類 − ポジティブ)	95.5 ± 1.1	96.3 ± 0.4	95.9 ± 0.5	—
	第 2 段階 (極性分類 − ネガティブ)	90.2 ± 0.8	88.2 ± 3.0	89.2 ± 1.5	—

付与する言語素性をもとに自動で行った．また，意志性の判別は，Kiyomaru & Kurohashi の手法 [14] に従って構築された意志性分類器を使用し，スコア ($\in [0, 1]$) が 0.3 以上のものを意志性が強いとみなした．生成したデータの統計を表 3 に示す．

ベースライン　手がかり表現と企業キーワードによる業績要因文の抽出手法 [5] をベースラインとした．手法に従って正例 85,241 文，負例 75,175 文から成る学習データを生成した．また，3.1 節のヒューリスティックの有無による精度の変化も調査した．

4.3　モデルの訓練・評価

分類器のモデルはいずれも RoBERTa [15] を採用した．事前学習モデルは早稲田大学が公開している日本語 RoBERTa base モデル[9] を使用した．

要因分類　分類器の性能を F 値で評価した．ベースライン・提案手法ともにモデルが過学習する傾向が見られたため，開発データに対する F 値をステップごとに測り，F 値が最大となるステップのパラメータでテストデータに対する性能を評価した．

重要度の予測　精度は重要文の正解率として評価し，以下の手順で算出した．

1. 決算短信ごとに各文の予測値を出す．
2. 各決算短信の重要文の数 N_i を既知として，予測値上位 N_i 文を取得する．
3. N_i 文のうち，実際に重要文であった数 M_i を求め，$\sum_i M_i / \sum_i N_i$ を正解率として計算する．

極性分類　表 2 の評価データのうち，正例かつポジティブまたはネガティブのラベルが付与されたものに対して，分類器の性能を F 値で評価した．開発データに対する F 値が最大となるエポックのパラメータでテストデータに対する性能を評価した．

9)　https://huggingface.co/nlp-waseda/roberta-base-japanese

4.4　実験結果・定性的分析

テストデータに対する実験結果を表 4 に示す．誤分類の実例は付録を参照されたい．

要因分類　提案手法について，疑似ラベルによる全体的な性能の改善が見られる．また，ベースライン・提案手法ともにヒューリスティックを導入することで適合率が大きく向上している．決算短信に含まれる業績と無関係な記述が負例であることを十分に学習させることが重要だと考えられる．

定性的分析では，経済や景気について述べた文を正例だと誤分類する傾向が見られた．主語の大きさに着目するような工夫が必要だと考えられる．

重要度の予測　重要文の正解率は提案手法による改善が見られるが，依然として精度が低い．文脈を見ないと企業全体・主力事業・その他事業のいずれについて述べているか判別できない文が多いことが原因として挙げられる．文章単位での解析に拡張すること，疑似ラベルを付与する際に各文が対象とする事業セグメントを解析し [16]，それをもとにスコアを付与することを検討する．

極性分類　F 値は 95.9，89.2 と比較的高い精度に達している．定性的分析では「赤字幅が減少」といった極性が反転する表現も正しく解析できていた一方，数量推論やドメイン知識が必要な文を誤分類する傾向が見られた．シード語彙に「前四半期比〜%」といった定量表現を加えることを検討する．

5　おわりに

本稿では極性と重要度を考慮した決算短信からの業績要因文の抽出手法を提案した．今後は極性と重要度の同時学習や重要度の予測精度の改善，そして構築した分類器とテキスト生成モデルを組み合わせた決算短信の要約記事の自動生成などを検討する．

謝辞

再現実験にご協力いただき，精度改善についてご助言を頂いた中間康文さんに感謝いたします．

参考文献

[1] 坂地泰紀, 和泉潔, 酒井浩之. 金融・経済ドメインを対象とした言語処理. 自然言語処理, Vol. 27, No. 4, pp. 951–955, 2020.

[2] 西沢裕子, 酒井浩之. 企業の決算短信 pdf からの業績要因の自動抽出. 電子情報通信学会技術研究報告 = IEICE technical report : 信学技報, 2013.

[3] 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.

[4] 中山祐輝, 津々見誠, 村上浩司. 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. 言語処理学会 第 25 回年次大会, 2019.

[5] 酒井浩之, 松下和暉, 北島良三. 学習データの自動生成による決算短信からの業績要因文の抽出. 知能と情報, Vol. 31, No. 2, pp. 653–661, 2019.

[6] 大村和正, 白井穂乃, 石原祥太郎, 澤紀彦. 決算短信からの業績要因文の抽出に向けた業績発表記事からの訓練データの生成. 言語処理学会 第 28 回年次大会, 2022.

[7] 酒井浩之, 増山繁. 企業の業績発表記事からの重要業績要因の抽出. 電子情報通信学会論文誌 D, Vol. 96, No. 11, pp. 2866–2870, 2013.

[8] 磯沼大, 藤野暢, 浮田純平, 村上遥, 浅谷公威, 森純一郎, 坂田一郎. 業績変動を考慮した決算短信からの重要文抽出. 情報処理学会 第 227 回 自然言語処理研究会, 2016.

[9] 酒井浩之, 坂地泰紀, 山内浩嗣, 町田亮介, 阿部一也. 深層学習と拡張手がかり表現による業績要因文への極性付与. 第 18 回 人工知能学会 金融情報学研究会, 2017.

[10] 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. 100, No. 2, pp. 150–161, 2017.

[11] 河村康平, 高野海斗, 酒井浩之. 決算短信からの業績要因を含む業績予測文の抽出. 2021 年度 人工知能学会全国大会 (第 35 回), 2021.

[12] Jun Saito, Yugo Murawaki, and Sadao Kurohashi. Minimally Supervised Learning of Affective Events Using Discourse Relations. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**, 2019.

[13] Sadao Kurohashi and Makoto Nagao. KN Parser: Japanese Dependency/Case Structure Analyzer. In **Proceedings of the International Workshop on Sharable Natural Language Resources**, pp. 48–55, 1994.

[14] Hirokazu Kiyomaru and Sadao Kurohashi. Minimally-Supervised Joint Learning of Event Volitionality and Subject Animacy Classification. In **Proceedings of the AAAI Conference on Artificial Intelligence**, 2022.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. **CoRR**, Vol. abs/1907.11692, , 2019.

[16] 高野海斗, 酒井浩之, 北島良三. 有価証券報告書からの事業セグメント付与された業績要因文・業績結果文の抽出. 人工知能学会論文誌, Vol. 34, No. 5, 2019.

表 5　提案手法に従って構築された要因分類器および極性分類器の誤分類例.

正解	予測	文
正例	負例	主な要因は、持分法投資利益の増加であります。
負例	正例	当第 1 四半期連結累計期間におけるわが国経済は、新型コロナウイルス感染症の影響により、依然として厳しい状況にありました。
ポジティブ	ネガティブ	当社グループはバッテリーセパレーター事業を第 2 四半期に譲渡しました。
ネガティブ	ポジティブ	販売費及び一般管理費は営業所の統廃合により人件費や賃借料が減少した一方、広告宣伝費用が増加した結果、前年同四半期より 13 百万円増加し、6 億 28 百万円となりました。

A　付録

A.1　決算短信 PDF からのテキスト抽出

PDF からのテキスト抽出には pdfminer[10] を使用した．日本語のテキスト抽出に対応しており，ページごとにテキストを抽出できる点などが理由である．

本研究が対象とする業績要因文は通常，本文の冒頭に「当四半期決算に関する定性的情報」などと題したセクションを設けて，その中で記述される．この定性的情報を高精度で抽出する (将来予測に関する記述などのノイズを減らす) ために，以下の手順でテキスト抽出を行った．

1. 文字列「目次」が含まれるページを先頭から順に探索し，最初に見つかったページより後のページを取得する．
2. 定性的情報の次に述べられやすい項目を表すフレーズ[11]が含まれるページを先頭から順に探索し，最初に見つかったページのフレーズ以前のテキストを抽出する．

要は，目次と定性的情報の次に述べられやすい項目を表すフレーズに挟まれる部分を定性的情報として抽出している．評価データの生成元の決算短信 150 件は全て，上記の手法で定性的情報が記述されたテキストを抽出できることを確認した．

A.2　誤分類の実例

提案手法に従って構築された要因分類器および極性分類器の誤分類例を表 5 に示す．これらを少し考察すると，例えば 1 番目の誤答例は「末尾の文節が過去形または体言止めでないならば負例」というヒューリスティックをモデルが重視していることが原因だと考えられる．改善案として，確信度の高い正例のデータから文末を現在形にした疑似データを生成し，これを混ぜて訓練することが挙げられる．

また，4 番目は本文で言及した数量推論などを要する文の誤答例である．定量表現をもとに費用が全体として増えたことを理解する必要があるため，定性的な極性表現を中心に学習させる現状のアプローチでは難しいと考えられる．シード語彙の拡張で対応できるか否かを今後調査する．

10）https://github.com/pdfminer/pdfminer.six

11）本研究では，「財政状態に関する」，「財政状態の」，「将来予測情報に関する」，「業績予想に関する」，「今後の見通し」の 5 つとした．

表 6　シード語彙の一覧.

ポジティブ	ネガティブ
増収, 増益, 好調, 堅調 順調, 回復, 向上, 改善	減収, 減益, 低調, 低迷 不振, 悪化, 鈍化, 停滞

表 7　訓練時のパイパーパラメータ.

パラメータ名	パラメータ値	
	要因分類器	極性分類器
バッチサイズ		32
エポック数	1	3
学習率	5e-5	1e-1
最大トークン長		128
Optimizer		AdamW
Betas		(0.9, 0.98)
重み減衰		1e-2
Scheduler	Linear decay with linear warmup	
Warmup	100 (steps)	0.1 (ratio)
シード値		{0, 1, 2}

A.3　シード語彙の一覧

極性分類器の学習データの生成時に策定したシード語彙の一覧を表 6 に示す．

A.4　ハイパーパラメータ

要因分類器および極性分類器の訓練時のハイパーパラメータを表 7 に示す．