日興リサーチセンター
SMBC日興証券グループ

テキストマイニングを用いた
有価証券報告書からの未来志向文の抽出

Research Report
2019 年 3 月

投資工学研究所
小寺　俊哉
佐藤　史仁
佐久間　洋明
田中　良典

要　　約

有価証券報告書や決算短信、新聞記事などのテキストデータには、企業の業績や業績予想の根拠、経営戦略、進行中の施策、今後の取り組み、認識しているリスクなど、投資判断において重要な情報が含まれている。近年、これら数値情報以外の情報を取得する技術として、テキストマイニングの金融分野への応用が進み、テキストデータから投資判断に有用と思われる非財務情報の効率的な抽出を試みる研究が行われている。しかしながら、これらの研究の多くは、業績やその根拠など主に企業の過去の活動に焦点をあてており、将来における取り組みなどの企業の未来の活動に関する情報について抽出を試みた例はそれほど多くない。投資判断が企業の将来における活動やその成果の確度に大きく依拠するとすれば、企業が今後何をするのか、どんな活動に重点をおくのかを把握することは、過去の情報を把握すること以上に重要であると言えよう。

そこで本稿では、企業の今後の活動に関して記述している文（未来志向文）を抽出することを目的とし、モデルの開発を行った。未来志向文は、その企業の取り組みに関する記述かつ未来に関する記述であるという前提の下、2 つの機械学習による判定モデルを組み合わせ、企業の今後の取り組みに関する記載の多い有価証券報告書の「対処すべき課題」から、未来志向文の抽出を試みた。その結果、本研究で構築したモデルは適合率、再現率、F 値いずれの評価指標においても平均で 0.9 を超える非常に高い性能を示した。また、本モデルを用いて有価証券報告書あたりの未来志向文の文数や「対処すべき課題」全文に対する割合の推移を見ると、「対処すべき課題」全文の数の増減にかかわらず未来志向文の文数は増加傾向にあり、全文に対する割合も 0.55 程度から 0.65 程度に単調に増加していた。

本稿で提案した未来志向文抽出モデルを用いることで、有価証券報告書等のテキストデータから今後の企業の取り組みに関する情報を、網羅的かつ効率的に抽出することができ、これらの情報を用いた新たな投資判断材料の獲得が期待できるだろう。

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

目次

1. はじめに
2. 未来志向文の抽出方法の全体像
2.1 利用データ
3. 未来表現文の抽出方法
3.1 未来表現文のラベル付与
3.2 未来表現文の抽出
3.2.1 語形によるマッチング
3.2.2 文末手がかり表現の自動獲得
3.2.3 機械学習による判定
4. 目的手段文の抽出方法
4.1 目的手段文のラベル付与
4.2 目的手段文の抽出
5. 結果
6. 考察
7. おわりに

1. はじめに

近年、金融市場の分析においてテキストマイニングの注目度が増している。例えば、日本銀行が発行していた金融経済月報や新聞記事に対して、テキストマイニングを用いて金融市場を分析する研究等が行われている。和泉ら[和泉他 2011]は金融経済月報のテキストデータを用いて日本国債市場での運用テストを行い、テキストマイニングが長期的な市場分析に有効であることを示した。また、藏本ら[藏本他 2013]は日本経済新聞のテキスト情報を用いて、和泉ら[和泉他 2011]が提案したCPR法を応用し、より広範で長期的な分析を行った結果、日経平均株価と TOPIX の 1 ヶ月後の騰落について、6 割以上の精度で予測できることを示した。これらの研究では、日本国債市場や日本株式市場などのマクロ指標を対象とした分析において、テキストマイニングの有効性を示した。一方で、ミクロな観点から個別企業を対象に分析する場合には、財務情報等の数値情報だけでなく、ニュースや新聞記事、SNS、決算短信、有価証券報告書等のテキスト情報等を用いて、各企業の特性や業績の理由及び根拠、経営戦略に関する情報、進行中又は今後予定されている施策、抱えているリスク等の多岐にわたる定性的な情報も加味し総合的に評価することが効果的だと考えられる。ただし、これらの評価を行うには専門的な知識や経験が必要なうえ、多数の企業間の相対的な評価を行う場合には、それらの企業を同時に一定の基準を持って評価しなければならないという難しい問題がある。

このような問題を解消するため、テキストマイニングの技術を使い、テキストデータから重要部分を

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

自動的に判断し抽出する研究や、定性的なテキストデータを定量的に評価する研究が盛んに行われている。例えば、坂地ら[坂地他 2015]は業績等に関する記述のある決算短信に着目し、機械学習を用いて自動的にそれらのテキストデータから投資判断に有用な情報として因果関係文を抽出するモデルを開発した。また、佐藤ら[佐藤他 2017][佐藤他 2018]は、このモデルを有価証券報告書に応用して因果関係文の抽出を試みた。因果関係文とは、出来事（結果）とその理由（原因）の組から構成される文と定義される。例えば、原因「猛暑」による結果「冷房需要の盛り上がり」等の因果関係を提示することで、「猛暑」の際には「冷房需要」が高まる可能性があるという情報を得ることが出来る。このように因果関係文は、業績要因の分析などに用いることができ、主に過去に起きた事象に関する非財務情報として重要な情報であると考えられる。

しかしながら、過去に関する分析だけでなく、その企業が今後どのような目的で、どのような手段を用いて、認識している課題に取り組んでいくのかという情報もまた、投資判断を行う上で必要不可欠な情報であろう。このような情報が記載されている開示書類として有価証券報告書が挙げられる。その中の「対処すべき課題」の項目には、主に各企業の事業上、財務上の対処すべき課題が述べられており、さらに投資家との対話促進の観点から、経営方針や経営環境に関しても記載することが求められている[1]。

将来における取り組みに関する情報を含む文を取得する際に、単にその企業の未来に関する情報を含む文を抽出するだけでは、その企業の取り組みなどの目的や手段に関係のない文も含まれ、一方で、その企業の取り組みなどの目的や手段に関する情報を含む文を抽出するだけでは、過去に行った取り組みを取得してしまう恐れがある。つまり、その企業の未来に関する情報と、その企業の取り組みなどの目的や手段に関する情報について同時に含む文を取得する必要がある。もし、このような企業の将来の取り組みに関する情報を含む文を自動的に取得することが可能となれば、その企業が認識している課題、それらの課題に対する具体的な取り組みやその目的や手段及び進捗を効率的に把握することが出来るだろう。さらに複数の企業のこれらの情報を網羅的に集約することが出来れば、財務指標の変化の予測や今後の業績の安定性、収益性の高低、ガバナンスの良し悪し等について、企業間や業種間における客観性の高い比較分析等へ応用することができる。

そこで、本稿では、有価証券報告書の「対処すべき課題」の項目から、未来に関する情報を含む特徴文（なんらかの特徴を含む文を、以降特徴文と呼ぶ）を「未来表現文」、企業の目的や手段に関する情報を含む特徴文を「目的手段文」、これら２つの特徴を同時に持つ特徴文を「未来志向文」と定義し、これらの特徴文を自動的に抽出するモデルの開発とその評価を行った。また、追加分析として、「対処すべき課題」にこれらの特徴文がどの程度含まれるかを確認するため、開発したモデルを用いて実際に

[1] 2017 年 3 月 31 日以後に終了する事業年度に係る有価証券報告書では、「対処すべき課題」の項目が、「経営方針、経営環境及び対処すべき課題等」と変更されており、その内容は内閣府令により「最近日現在において連結会社が経営方針・経営戦略等を定めている場合には、当該経営方針・経営戦略等の内容を記載すること。また、経営上の目標の達成状況を判断するための客観的な指標等がある場合には、その内容について記載すること。」、「最近日現在における連結会社の経営環境並びに事業上及び財務上の対処すべき課題について、その内容、対処方針等を具体的に記載すること。」と定められている。
https://www.fsa.go.jp/news/28/syouken/20170214-1/02.pdf

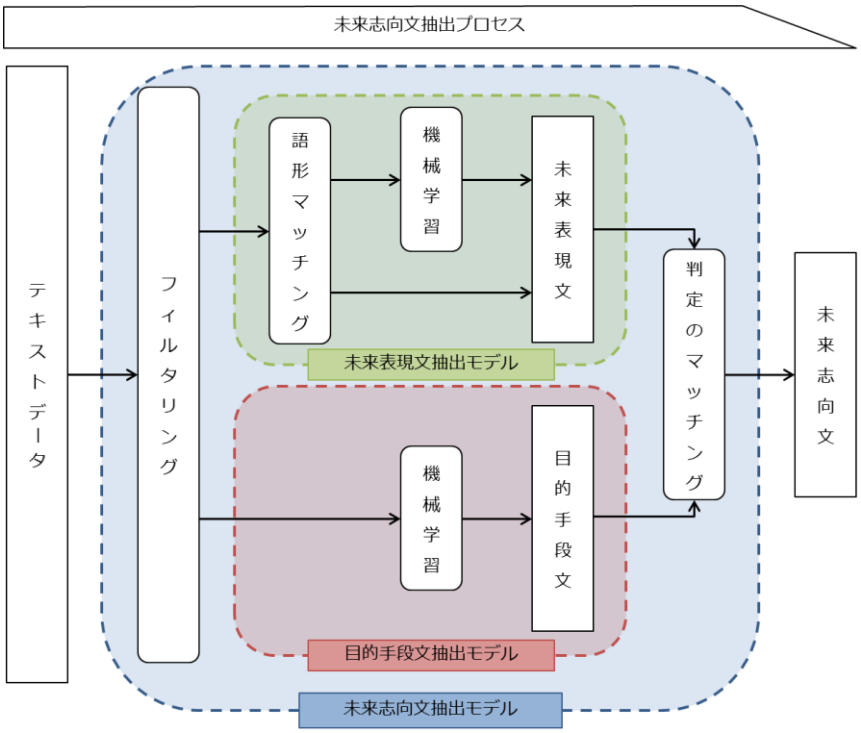
本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

抽出を行った。そして、抽出した各特徴文を年ごとに集計しその推移の全体的な傾向を確認した。

2. 未来志向文の抽出方法の全体像

本稿で提案する未来志向文の抽出モデルは、未来表現文抽出モデルと目的手段文抽出モデルの 2 つのモデルの組み合わせによって構築される。図表 1 にテキストデータから各々のモデルを通して未来志向文が抽出される過程を示す。

図表 1. 未来志向文抽出プロセス



（出所）日興リサーチセンター作成

まず、テキストデータから不要な部分を取り除くフィルタリング処理が実行される。ここでは、文字数が 2 や 3 の極端に短い文や「【対処すべき課題】」の項目そのものを表すタイトル部分などが除外される。次に、テキストデータは未来表現文抽出モデルと目的手段文抽出モデルにより各々の判定結果が付与される。それぞれのモデルの具体的な処理は後述する。最後に未来表現文と目的手段文の両方のモデルにおいて正例と判定されたものが未来志向文として抽出されることとなる。本稿では、未来表現文、目的手段文、未来志向文それぞれにおいて当てはまるものを正例、当てはまらないものを負例と呼ぶこととする。ここで、未来表現文とは未来に関する情報を含む特徴文であり、目的手段文とは企業の目的や手段に関する情報を含む特徴文のことである。また、未来志向文とは上記 2 つの特徴を同時に有する特徴文のことである。各特徴文の例文を図表 2 に示す。文 1 は豊田自動織機の 2017 年 3 月期、文 2 はデンソーの 2010 年 3 月期、文 3 は住石ホールディングスの 2009 年 3 月期、文 4 はヤマハ発動機の 2016 年 12 月期の有価証券報告書から取得した文を示している。

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

図表 2. 各特徴文の例文

No	未来表現文	目的手段文	未来志向文	例文
文 1	正例	正例	正例	・企業活動を通じて住みよい地球と豊かな社会づくりに取り組むとともに、クリーンで安全な優れた品質の商品を提供する
文 2	正例	負例	負例	今後の世界経済については、引き続き内需が好調な中国を中心とするアジアが世界を牽引し、緩やかに回復が続き、先進国・新興国ともにプラス成長するものと見ています。
文 3	負例	正例	負例	事業環境の急激な変化にあっても、安定的な成長を続けるべく昨年 10 月より持株会社体制に移行しました。
文 4	負例	負例	負例	また、これらに限らず株式の買付行為の中には、その態様によっては、企業価値及び株主共同の利益を害するものも存在します。

（出所）各社有価証券報告書より日興リサーチセンター作成

2.1　利用データ

本稿の分析に用いたテキストデータの基本情報を図表 3 に示す。対象テキストとして有価証券報告書を用いた。対象期間は 2008 年～2018 年とし、各年において取得することの出来た各企業の本決算の有価証券報告書の PDF ファイルをテキスト化した。そして、それらの中から「対処すべき課題」または「経営方針、経営環境及び対処すべき課題等」（以下、2 つを併せて「対処すべき課題」とする）の項目を抽出した後、文単位に分割し、テキストデータとした。

また、本稿における自動抽出モデルでは、教師あり学習と呼ばれる機械学習手法を用いた。教師あり学習では、入力されるデータと対応する教師信号（以下、ラベルと呼ぶ）、つまり正解符号が必要となる。この入力データとラベルのセットを学習データと呼び、これを用いてモデルがデータの特徴を学習することとなる。また、完成したモデルが適切に判定出来ているか評価するために用いられる入力データとラベルのセットを評価データと呼ぶ。

本稿では、学習データ及び評価データを作成するために用いる文の抽出においては、より広範な表現に対応できるように、有価証券報告書の時期や業種[2]を分散させ、それぞれ 935 文（5 社×17 業種×11 年）の文を抽出した。評価データは、後述する未来表現文、目的手段文、未来志向文において同じデータを用いてラベルを付与し、評価を行った。

[2] 本稿では、 東京証券取引所が定めた東証 17 業種分類を用いた。

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

図表 3. テキストデータの基本情報

対象	内容
期間	2008 年～2018 年
企業	有価証券報告書が取得することが出来た上場企業
有価証券報告書	発表日が対象期間に含まれている本決算の有価証券報告書
項目	「対処すべき課題」、「経営方針、経営環境及び対処すべき課題等」

（出所）日興リサーチセンター作成

3. 未来表現文の抽出方法

3.1 未来表現文のラベル付与

英語では助動詞や文法的特徴から比較的容易に時制を判断できる一方で、日本語では未来の時制を判断するのは容易ではないため、時制に関する定義が難しく様々な議論がなされている。例えば、龍城[龍城 2013]は「特に日本語の場合は、良く知られているように、時間的概念の過去、現在、未来は存在するが、時制としては過去と非過去という 2 つの時制しかない。」と述べており、文中における節単位での時制転移について日英語を比較して議論している。このように日本語における時制の定義は、英語の様にある程度明確なルールに基づいて決定することは難しく、様々な要素から複合的に考える必要がある。その中で庵[庵 2001]は、図表 4 の様に、「シテイル」と「シタ」という表現はそれぞれ「―てい―」と「―φ―」という形式でアスペクト[3]的に対立し、「―る―」と「―た―」はテンス[4]的に対立することにより分類できるとしている。

図表 4. 日本語のテンス及びアスペクトによる対立

アスペクト / テンス	完成相	継続相
非過去	スル	シテイル
過去	シタ	シテイタ

（出所）庵[庵 2001]を参考に日興リサーチセンター作成

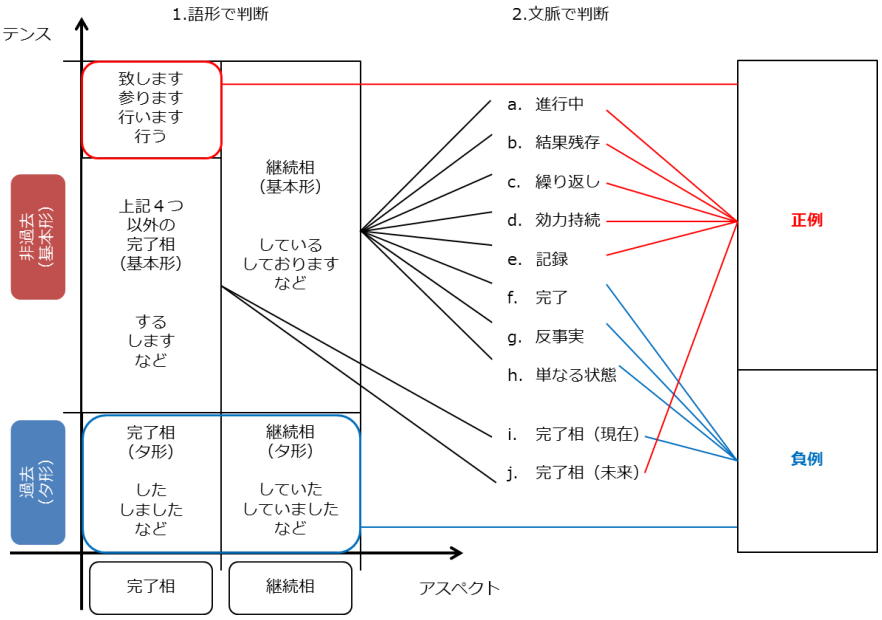
これを基に本稿で考案した未来表現文のラベル付与に関するルールを図表 5 に示す。本稿では文末が「”た。”」で終わるものをタ形、それ以外のものを基本形と呼ぶこととし、タ形のものであれば文脈に関わらず過去を示す文と判断し、未来表現文における負例とした。また、基本形の中で「”致します。”、”

[3] ある動作の局面や状態（開始直前、開始直後、継続、反復等）を表す言語学・文法学の用語。相とも呼ばれる。
[4] 発話時を基準とした述語の時間軸での位置付け（過去、現在、未来）。時制。英語では、動詞の活用によってその動詞の表す動作や状態がいつ行われたかが示されるが、日本語では活用形が無いため、動詞のみでテンスを表すことが出来ない。「リンゴが落ちていた。」を例に取ると、テンスは過去、アスペクトは（落ちている状態の）継続となる。

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

参ります。”、”行います。”、”行う。”」及びそれらの平仮名表現のものは、常に未来に行う内容を含む文であると判断し、文脈に関わらず正例であるとした。例えば、「新製品の開発により成長機会を確保することに注力致します。」といった文がこれに該当する。

図表 5. 人手による未来表現文のラベル付与の判断基準



（出所）日興リサーチセンター作成

次に、上記 4 つ以外の完了相基本形について、厳密なルールに基づいて分類する方法が筆者らの知る限りでは存在しなかったため、文脈から人手によって正例負例を判断し、正例となる部分を完了相（未来）、負例となる部分を完了相（現在）と呼ぶこととした。

最後に、庵[庵 2001]では「―てい―」で示される継続相に着目し、その文の意味により以下の a~h に分類し、a~e は「継続」、f,g は「以前」のカテゴリとし、h は形容詞適用法の別物として扱っている。そこで本稿ではこれらを参考に、継続相基本形の文について、a~e に含まれる文を未来表現文の正例として、f~h に含まれる文を負例として人手によって判定し、ラベル付与することとした。

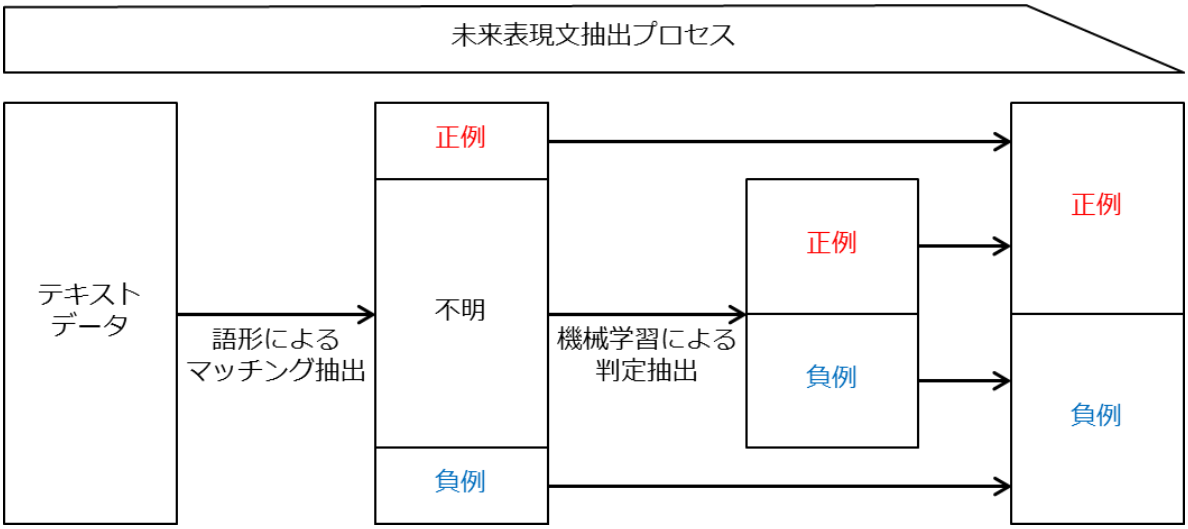
a. 進行中 （例：田中さんは部屋で本を読んでいる。）
b. 結果残存 （例：窓ガラスが割れている。）
c. 繰り返し （例：田中さんは毎朝ジョギングをしている。）
d. 効力持続 （例：この橋は 5 年前に壊れている。）
e. 記録 （例：犯人は 3 日前にこの店でうどんを食べている。）
f. 完了 （例：彼からの手紙を受け取ったとき、彼は既に死んでいた。）
g. 反事実 （例：彼が助けてくれなかったら、私は死んでいた。）
h. 単なる事実 （例：学校の北側に高い山がそびえている。）

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

3.2　未来表現文の抽出

本稿では未来表現文の抽出を、「語形によるマッチング抽出」と「機械学習による判定抽出」の 2 つの手法を複合的に用いて行った。この抽出手法の全体像は図表 6 に示す。

図表 6. 未来表現文抽出プロセス



（出所）日興リサーチセンター作成

まずは、文末の語形のみによって、タ形にマッチングする文を負例に、「”致します。”、”参ります。”、”行います。”、”行う。”」及びそれらの平仮名表現を含む文を正例として抽出した。次に、これら以外の文に対して機械学習を用いて未来表現文の判定を行った。ところで、正例と判定した上記４つの文末表現に類似した表現は、同様に正例として判別されることが想定された。そのため、本稿では、酒井ら[酒井他 2016]の手法を参考に、文末手がかり表現を自動的に獲得し、その情報を後述の機械学習による判定の素性として用いた。

3.2.1　語形によるマッチング

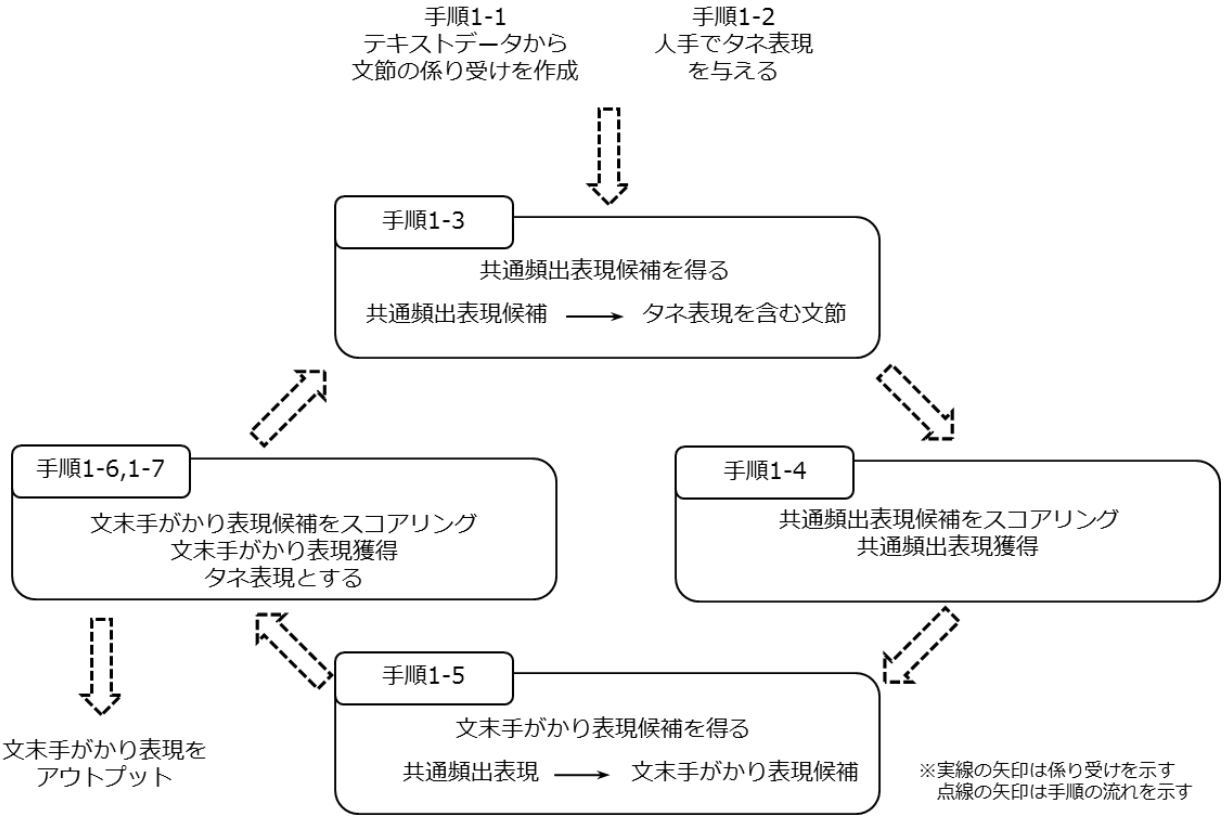
未来表現文の判定には文末の表現の特徴を利用した。例えば「”参ります。”」という表現であれば、その文はほぼ将来の取り組みについて記述されているため、この表現がある場合は正例と判断した。また、文末が「”た。”」のタ形の文は過去や現在までに完了した内容を示していると考えられるため負例と判断した。この文末の特徴を利用し、全てのテキストデータに対してまずは文末の表現を確認し「”致します。”、”参ります。”、”行います。”、”行う。”」及びそれらの平仮名表現のいずれかに合致するものは正例、文末の表現が「”た。”」に合致するタ形のものは負例と機械的に判定した。この文末表現の語形マッチングで判定出来なかったものに対して後述の機械学習の判定を行った。

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

3.2.2 文末手がかり表現の自動獲得

文末表現が「”致します。”、”参ります。”、”行います。”、”行う。”」及びそれらの平仮名表現に合致するものは正例と判定できるという特性から、それらの類似表現は機械学習の判定において有効な素性情報になると考えられる。本稿では、そのような文末表現を「文末手がかり表現」と呼ぶこととした。それら文末手がかり表現を獲得するために、拙稿[小寺他 2018]と同様に酒井ら[酒井他 2016]の手法を参考に文末手がかり表現の自動獲得手法を試みた。本稿では、文末手がかり表現に対して係る文節の内、多くの文末手がかり表現に共通して頻繁に係る文節は、既存の文末手がかり表現と類似した意味の文末表現に対して係る可能性が高いと考えた。そのような文節を共通頻出表現として獲得し、次に共通頻出表現が係る文末表現の内、同様に共通して頻出するものを新たな文末手がかり表現として獲得した。例えば、既存の文末手がかり表現として「”行います。”」を与え、共通頻出表現として「”施策を”」を獲得する。その獲得した共通頻出表現により文末表現の「”進めていきます。”」を獲得し、これを新たな文末手がかり表現とする。この手順を繰り返し行うことで文末手がかり表現を自動で獲得した。文末手がかり表現の自動獲得手法のイメージ図を図表 7 に、具体的な手順を手順 1-1～手順 1-7 に、獲得した文末手がかり表現の一部を図表 8 に、共通頻出表現の一部を図表 9 に示す。文末手がかり表現の自動獲得手法によって 74 個の文末手がかり表現と 215 個の共通頻出表現を獲得した。

図表 7. 文末手がかり表現自動獲得手法イメージ図



（出所）日興リサーチセンター作成

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

図表 8．文末手がかり表現例

獲得した文末手がかり表現の一部		
努める。	推進しております。	目指しております。
図っております。	計画しております。	所存です。
図っていく。	実施しております。	拡大します。

（出所）日興リサーチセンター作成

図表 9．共通頻出表現例

獲得した共通頻出表現の一部		
事業展開を	再構築を	改善を
推進し、	向け、	よう、
目指し、	向けて	収益性を

（出所）日興リサーチセンター作成

手順1-1．対象となるテキストデータに係り受け解析器 CaboCha[5][工藤，松本 2002]を用いて構文解析[6]し、文節に区切り係り受けを作成する。

手順1-2．少数の文末手がかり表現（以下「タネ表現」とする）を人手で与える。本稿ではタネ表現として「”致します。”、”参ります。”、”行います。”、”行う。”」の 4 つの表現を与えた。

手順1-3．式 1 によりタネ表現の数から閾値T_eを算出する。また、タネ表現を含む文末文節を取得し、それら文末文節に係る文節を共通頻出表現候補として集合を作成する。

$$T_e = \alpha \log_2 |N_s| \tag{1}$$

α：定数（$0 < \alpha < 1$）、本稿では 0.5 とする

$|N_s|$：タネ表現の数

手順1-4．各共通頻出表現候補eに対して式 2,3 により$H(e)$を算出し、手順 1-3 で算出した閾値に対し、$H(e) \geq T_e$のものを共通頻出表現e'として獲得する。

$$H(e) = -\sum_{s \in S(e)} P(e,s) \log_2 P(e,s) \tag{2}$$

$$P(e,s) = \frac{f(e,s)}{\sum_{s' \in S(e)} f(e,s')} \tag{3}$$

[5] CaboCha とは、機械学習に基づく日本語係り受け解析器で、文を入力すると文節の係り受け関係が出力として得られる。https://taku910.github.io/cabocha/

[6] 構文解析とは、文の構造を求めることであり、日本語の場合は、文節間の係り受けを求める場合が多い[石田，金 2012]。

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

$S(e)$：共通頻出表現候補eが係るタネ表現の集合

$P(e,s)$：共通頻出表現候補eがタネ表現sに係る確率

$f(e,s)$：共通頻出表現候補eがタネ表現sに係る回数

手順1-5. 式 4 により共通頻出表現の数から閾値T_sを算出する。また、それら共通頻出表現が係る先の文節を文末手がかり表現候補として集合を作成する。ただし、本稿では文末表現のみを文末手がかり表現候補とした。

$$T_s = \alpha \log_2 |N_{e'}| \tag{4}$$

$|N_{e'}|$：共通頻出表現の数

手順1-6. 各文末手がかり表現候補s''に対して式 5,6 により$H(s'')$を算出し、手順 1-5 で算出した閾値に対し、$H(s'') \geq T_s$のものを文末手がかり表現として獲得する。

$$H(s'') = -\sum_{e' \in E(s'')} P(s'', e') \log_2 P(s'', e') \tag{5}$$

$$P(s'', e') = \frac{f(s'', e')}{\sum_{e'' \in E(s'')} f(s'', e'')} \tag{6}$$

$E(s'')$：文末手がかり表現候補s''に係る共通頻出表現の集合

$P(s'', e')$：文末手がかり表現候補s''に共通頻出表現e'が係る確率

$f(s'', e')$：文末手がかり表現候補s''に共通頻出表現e'が係る回数

手順1-7. 獲得した新たな文末手がかり表現をタネ表現とし、手順 1-3 に戻る。ただし、新たな文末手がかり表現が得られなくなる、または手順 1-3~手順 1-7 を指定の回数繰り返した場合は終了する。

3.2.3 機械学習による判定

機械学習による判別モデルを用いる際に、学習、評価及び学習後の判別処理において、文の特徴を表す素性が必要となる。本稿では図表 10 に示す 9 種類の素性を利用した。素性の作成において、形態素解析[7]では形態素解析器 MeCab[8] [工藤他 2004]を、構文解析では係り受け解析器 CaboCha を用いた。機械学習の手法としてサポートベクターマシン（SVM）[9]を利用した。

[7] 形態素とは、言語学で使われる専門用語で、“意味の最小の単位”と説明され、テキストを形態素に分割することを形態素解析という[石田 2008]。

[8] MeCab は、京都大学情報学研究科-日本電信電話株式会社コミュニケーション科学基礎研究所 共同研究ユニットプロジェクトを通じて開発されたオープンソース形態素解析エンジン。文を入力すると、その文の形態素と各形態素の品詞が出力される。http://taku910.github.io/mecab/

[9] python の機械学習のオープンソースライブラリ scikit-learn（http://scikit-learn.org/stable/index.html）を用いた。

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

図表 10. 未来表現文抽出機械学習モデルの素性概要

素性の名前	概要
形態素ユニグラム	学習データに含まれる頻度 2 以上の形態素。
形態素バイグラム	学習データに含まれる頻度 2 以上の隣り合った形態素のペア。
文末手がかり表現	文末手がかり表現の自動獲得手法により得られた文末手がかり表現の内、学習データに含まれるもの。
共通頻出表現	文末手がかり表現の自動獲得手法により得られた共通頻出表現の内、学習データに含まれるもの。
副詞	学習データに含まれる品詞が副詞の形態素。
文末 2 文節のペア	学習データに含まれる文末 2 分節のペア。
文末の品詞	学習データに含まれる文末の形態素の品詞。
句点の有無	文末に句点が存在するか否か。
文の属性	文が「タイトル」「箇条書き」「注記」「本文」等のいずれの属性に属するか。

(出所) 日興リサーチセンター作成

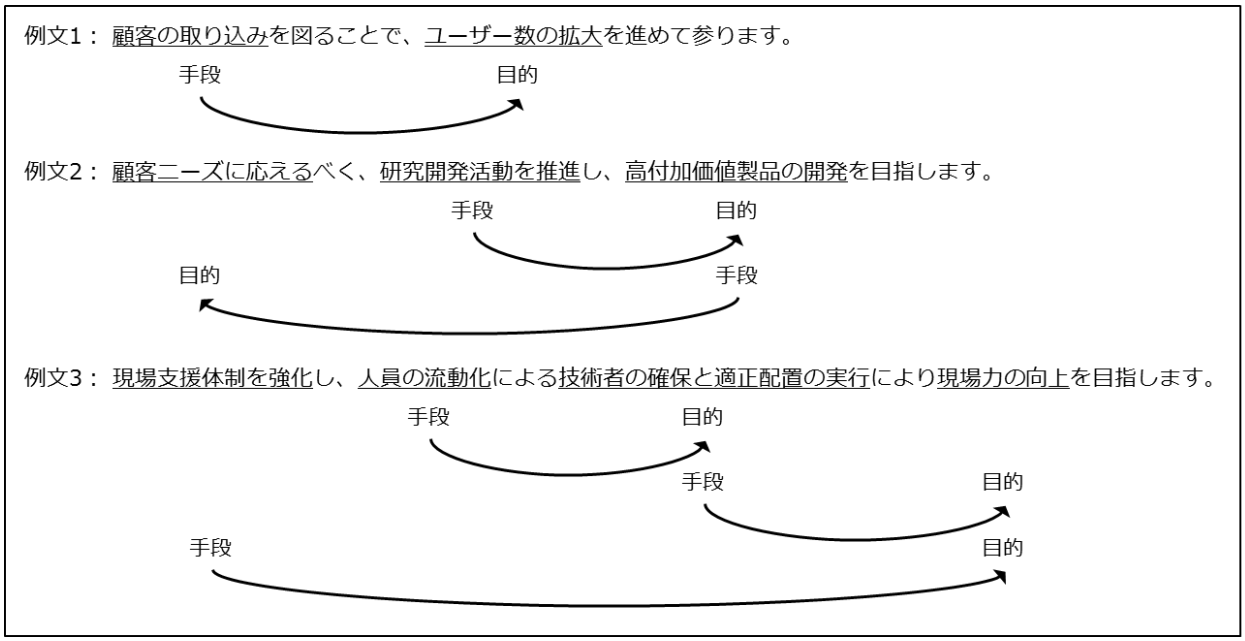
4. 目的手段文の抽出方法

4.1 目的手段文のラベル付与

日本語において、手段部分と目的部分をルールで分けることが難しいため、文脈から判断し分類を試みた。図表 11 は文中における目的と手段の関係の構造について示している。図表 11 の例文 1 に示すように、基本的には手段部分が目的部分に係るという係り受け構造が見られる。この例では「顧客の取り込み」が手段部分となり、「ユーザー数の拡大」が目的部分となる。しかしながら、例文 2 や例文 3 のように、例えば例文 3 では「人員の流動化」が手段部分となり、目的部分「技術者の確保と適正配置の実行」に係る。ところが、この「技術者の確保と適正配置の実行」は手段部分として、「現場力の向上」を目的部分として係る。つまり、着目する語群は、係り受けの組合せ次第で、手段にも目的にも成り得てしまい、またこれらの事例は多数見られた。加えて、文頭の「現場支援体制を強化」が手段部分として同じく「現場力の向上」に係るといった入れ子の構造になる場合も多々存在した。このように、文中の手段部分や目的部分を厳密に分けることは難しいため、本稿では文中に目的もしくは手段の少なくともいずれかを含むものを正例として判定した。

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

図表 11．文中における目的と手段の関係性の構造



（出所）日興リサーチセンター作成

4.2　目的手段文の抽出

上述のように目的手段文の判定はルールベースで行えないため、機械学習のみを用いて判定を行った。本稿では図表 12 に示す 7 種類の素性を利用した。入力する素性データは、未来表現文と同様の処理を行い作成した。ただし、ここでは未来表現文の判定結果が目的手段文の判定にも関係があるものとして素性に加えた。未来表現文と同様に機械学習手法としてサポートベクターマシン（SVM）を利用した。

図表 12．目的手段文抽出機械学習モデルの素性概要

素性の名前	概要
形態素ユニグラム	図表 10 の形態素ユニグラムと同様。
形態素バイグラム	図表 10 の形態素バイグラムと同様。
読点の前の品詞	学習データに含まれる読点の前の形態素の品詞（読点が無い場合は読点無しという素性を加える）。
文末の品詞	図表 10 の句点の前の品詞と同様。
句点の有無	図表 10 の句点の前の品詞と同様。
文の属性	図表 10 の句点の前の品詞と同様。
未来表現文判定結果	未来表現文抽出モデルでの判定結果（正例か負例）。

(出所) 日興リサーチセンター作成

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

5. 結果

図表 13 は左上、右上、左下の順に未来表現文、目的手段文、未来志向文の抽出モデルの評価結果を示している。平均/合計の欄は、適合率、再現率、F 値については正例と負例の数で加重平均した値を、データ数、判定数については合計を示す。未来表現文抽出モデルと目的手段文抽出モデルのいずれにおいても正例の判定で適合率、再現率、F 値それぞれで 0.9 を超える非常に高い評価結果であった。また、負例においても最低でも 0.765 と、ある程度高い水準を示した。平均を見ると全ての評価指標で 0.9 前後であった。同様に、未来志向文抽出モデルでは、正例はいずれの評価指標においても 0.9 を超える高い性能を示した。一方で、負例においても全ての指標で 0.8 を上回る良好な性能を示した。また、平均を見ると全ての指標が 0.9 を上回った。

図表 13. 各抽出モデルの評価結果

	適合率	再現率	F 値	データ数	判定数
正例	0.936	0.993	0.964	726	770
負例	0.970	0.766	0.856	209	165
平均/合計	0.944	0.942	0.940	935	935

	適合率	再現率	F 値	データ数	判定数
正例	0.936	0.924	0.930	714	705
負例	0.765	0.796	0.780	221	230
平均/合計	0.896	0.894	0.895	935	935

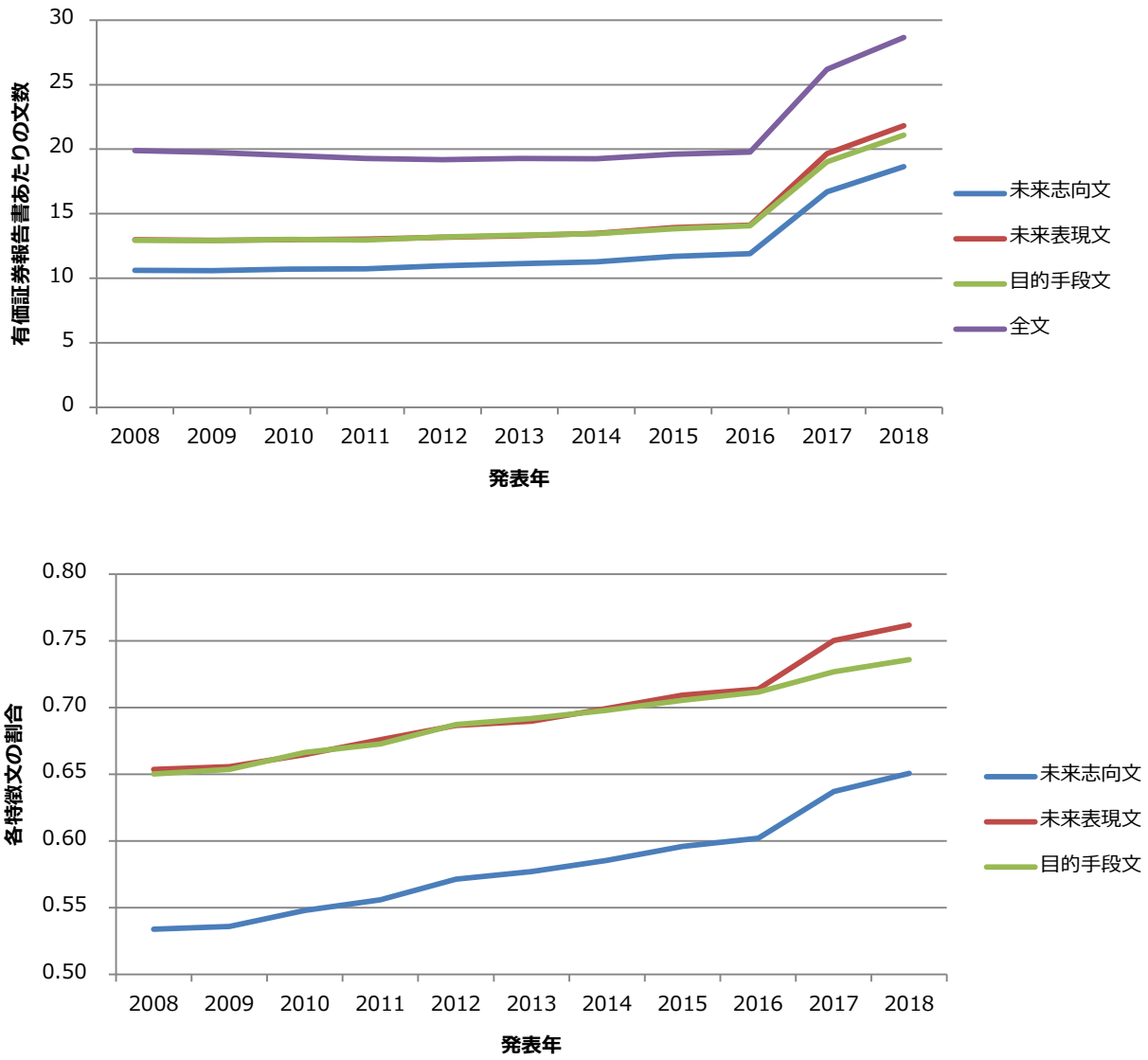
	適合率	再現率	F 値	データ数	判定数
正例	0.918	0.964	0.940	631	662
負例	0.916	0.822	0.867	304	273
平均/合計	0.918	0.918	0.916	935	935

（注）左上、右上、左下の順に未来表現文、目的手段文、未来志向文抽出モデルの評価結果を示す。
表中数値は、例えば正例の適合率はモデルによって正例と判定されたデータの内、実際に人によるラベルが正例であったものの割合を示し、正例の再現率は人によるラベルが正例であったデータの内、モデルによって正例と判定されたものの割合を示す。F 値は適合率と再現率の調和平均を示す。正例のデータ数は評価データ中に実際にラベルが正例であったデータの数を、正例の判定数は評価データ中のモデルにより正例と判定されたデータの数を示す。平均/合計の欄は、適合率、再現率、F 値について正例と負例の数で加重平均した値を、データ数、判定数については合計を示す。
(出所) 各社有価証券報告書より日興リサーチセンター作成

有価証券報告書あたりの「対処すべき課題」中の全文数、及び未来志向文、未来表現文、目的手段文それぞれの文数、並びに各特徴文数の全文数に対する割合の発表年ごとの推移を図表 14 に示す。有価証券報告書あたりの全文の数は 2012 年まで僅かに減少していたが、その後は増加傾向となり 2017 年から大きく増加した。一方で、各特徴文の数は僅かではあるが増加傾向にあり、2017 年以降大きく増加した。また、分析期間中においてどの特徴文も全文に占める割合が年々増加しており、「対処すべき課題」全文に占める未来志向文の割合は 0.55 程度から 0.65 程度に拡大した。

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

図表 14. 有価証券報告書あたりの各特徴文、全文の数の推移（上）とその割合（下）



（注）上図は発表年ごとの有価証券報告書あたりの各特徴文と「対処すべき課題」全文の数、下図は発表年ごとの各特徴文の「対処すべき課題」全文に対する割合の推移を示す。
（出所）各社有価証券報告書より日興リサーチセンター作成

6. 考察

本稿で提案した未来表現文抽出モデル、目的手段文抽出モデル、未来志向文抽出モデルの判定性能は、いずれも良好なものであった。しかしながら、これらのモデルに用いたデータの特性として、どのタイプの抽出文においても正例が多く負例が少ないという、データ数の偏りがあった。つまり、全て正例と判定されればそれだけで正解しやすくなるため、モデルが高性能に見えてしまう懸念があった。しかしながら、全てのモデルについて判定数とデータ数はほぼ一致しており、むやみに正例と判定したことによって高い性能を示している訳ではなく的確に判定できていると言えるだろう。さらに、興味深い点として、未来志向文における負例の再現率とＦ値の評価指標が、未来表現文及び目的手段文のそれよりも

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

高い値を示した。これは未来志向文抽出モデルでは、未来表現文抽出モデルと目的手段文抽出モデルの両方で正例とならない限り正例と判定されないために、いずれかのモデルにおいて実際に負例のものを正例と誤判定したとしても、もう片方のモデルで適切に判定することで、誤判定された文を相補的に負例にしたと考えられる。本稿で提案したモデルは、有価証券報告書の「対処すべき課題」にある固有の特徴を用いることなく、日本語の文章の特徴を利用している。その中で、目的や手段を含み、かつ未来に関する記述であることを独立に評価し、最後にこれらを総合的に評価することで未来志向文の抽出を試みたが、これらのアプローチが有効であったことの証左であり、同時に二つの機械学習を組み合わせて判別するアプローチの有効性も示すことが出来た。

本稿では、抽出した各特徴文について、有価証券報告書あたりの文数及び全文に対する割合の推移を発表年ごとに見た。2016 年までは僅かではあるが、全文数は 2012 年まで減少しそれ以降増加したが、各特徴文の数は増加傾向にあった。しかしながら、「対処すべき課題」全文に占める各特徴文の割合は単調に増加する傾向にあった。このことは、「対処すべき課題」全文を減らした際には、今後の取り組みに関する記述を削除することはせず、「対処すべき課題」の文数を増やす際には、今後の取り組みに関する記述を優先して加筆したことを示唆している。観察期間中は、リーマンショックやギリシャショック、東日本大震災、アベノミクス等、経済環境や経営環境にとってインパクトの大きい出来事が多々発生した期間であったが、どのような環境下においても、企業にとって今後の取り組みを示している文は、投資家に対する重要なメッセージとして位置付けられていることの表れかもしれない。一方で、2017 年以降は全文、各特徴文ともに大きく増加し、全文に対する割合も大きく増加した。これは、内閣府令により「対処すべき課題」が「経営方針、経営環境及び対処すべき課題等」へと変更され、それまでの対処すべき課題に加え、経営方針や経営環境についての記述が追加されたことに起因すると思われる。政府は、アベノミクスにおける成長戦略において、日本再興戦略及び未来投資戦略を実施してきた[10]。その中で、政府は中長期的投資の促進の施策として、「企業の戦略的投資や、ESG 要素も念頭に置いた中長期的な企業価値向上に資する開示」を図る[11]など、企業と投資家との対話を促進している。未来志向文の文数や「対処すべき課題」全体における割合の増加は、このような環境下、企業の情報開示への姿勢が変化してきていることを表しているのかもしれない。

7. おわりに

本稿では、目的表現や手段表現を含みかつ未来を表現している文を未来志向文と定義して、自動で抽出するモデルを構築し、有価証券報告書の「対処すべき課題」の項目から未来志向文の抽出を行った。本稿の提案モデルは、適合率、再現率、F 値の平均でいずれの評価指標でも 0.9 を超える非常に高い性能を示した。また、提案モデルを用いて「対処すべき課題」全文から抽出した各特徴文の数や割合の推移を見たところ、近年における、各社の法定開示資料における将来の取り組みに関する開示意識の高ま

[10] 首相官邸 HP「これまでの成長戦略について」（https://www.kantei.go.jp/jp/singi/keizaisaisei/kettei.html）
[11] 例えば、未来投資戦略 2018（全体版）p132
（https://www.kantei.go.jp/jp/singi/keizaisaisei/pdf/miraitousi2018_zentai.pdf）

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

りを示唆していた。ただし、2017 年以降に関しては、法令による開示方法変更の影響を強く受けており、将来に関する取り組みの記述が、経済・経営環境よりも寧ろ、法令対応により大きく変化しやすいことを示唆している。つまり、法令・規則による記載内容の変更や記載箇所の変更は、その企業の業績や方針等の変化とは、無関係な変化をもたらすことになる。この点、適時開示資料や法定開示資料を用いて時系列分析等を行う際には、テキストデータの連続性を担保するために、対象とするテキストデータを選択的に限定するような技術が必要になるだろう[田中 2018]。

今後の課題として、抽出した未来志向文の活用方法の検討や適用範囲の拡張を考えたい。例えば、未来志向文を用いて各社の取り組みを定量的に分析することや、同業他社間での類似する未来志向文を抽出し、各社の取り組みの機微な違いを比較することにも役立てることが出来るだろう。また、本稿で提案した手法は、その抽出条件に文書固有の特徴を含んでいないため柔軟性に富んでおり、他の文書へ応用することも容易であると考えられる。例えば、各企業の社長メッセージや経営計画資料といった幅広い文書からの未来志向文の抽出が期待できる。そのような様々な情報媒体から各企業の将来の取り組みに関する情報を集約し分析することでより多面的な企業分析が可能になるだろう。

＊ 本稿の作成にあたり、東京大学大学院工学系研究科教授和泉潔先生と特任講師坂地泰紀先生に、多大なるご指導とご助言を頂いた。心から感謝の意を表したい。なお、本稿の内容・意見は全て当社に属する。

＊ 本稿における提案手法の未来志向文抽出技術は現在特許出願中（出願番号「特願 2019-001952」）である。

以上

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。

参考文献

庵功雄[2001], 「テイル形、テイタ形の意味の捉え方に関する一試案」, 一橋大学留学生センター紀要, (4), pp.75-94, 2001

石田基広[2008], 「R によるテキストマイニング入門」, 森北出版, 2008.

石田基広・金明哲編著[2012], 「コーパスとテキストマイニング」, 共立出版, 2012.

和泉潔・後藤卓・松井藤五郎[2011], 「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, Dec 2011.

工藤拓・松本裕治[2002], 「チャンキングの段階適用による日本語係り受け解析」, 情報処理学会論文誌, Vol.43, No.6, pp.1834-1842, Jun 2002.

工藤拓・山本薫・松本裕治[2004], 「Conditional Random Fields を用いた日本語形態素解析」, 情報処理学会研究報告自然言語処理 (NL), Vol.2004, No.47, pp.89-96, May 2004.

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[2013], 「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.

小寺俊哉・佐藤史仁・田中良典[2018], 「テキストマイニングによる有価証券報告書の因果関係文以外の特徴文の抽出」, 日興リサーチレビュー, 2018 年, 1 月号.

酒井浩之・柴田宏樹・平松賢士・坂地泰紀[2016], 「アナリストレポートからのアナリスト予想根拠情報の抽出」, 第 17 回金融情報学研究会, pp.25-30, 2016

坂地泰紀・酒井浩之・増山繁[2015],「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, May 2015.

佐藤史仁・佐久間洋明・小寺俊哉[2017], 「テキストマイニングによる有価証券報告書の因果関係文の抽出」, 日興リサーチレビュー, 2017 年, 10 月号.

佐藤史仁・佐久間洋明・小寺俊哉・田中良典・坂地泰紀・和泉潔[2018], 「有価証券報告書からの因果関係文の抽出」, 人工知能学会全国大会論文集 JSAI2018, 2O404, 2018.

龍城正明[2013], 「日英語節複合における時制表現に関する考察：話者の状況の視点という概念を通して」, 同志社大学英語英文学研究, 90, pp.190-216, 2013.

田中良典[2018], 「テキストマイニングによる有価証券報告書の開示変更への対応に関する分析」, 日興リサーチレビュー, 2018 年, 9 月号.

本資料は、信頼性の高いデータから作成されておりますが、当社はその正確性・確実性に関し、いかなる保証をするものではございません。本資料は、情報提供を目的としており、投資勧誘を目的としたものではございません。証券投資に関する最終判断は、投資家ご自身の判断でなさるようにお願いいたします。本資料の著作権は当社に帰属し、本資料の転用および販売は固く禁じられております。