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による有価証券報告書の因果関係文の抽出

Research Report
2017 年 10 月

投資工学研究所
佐藤　史仁
佐久間　洋明
小寺　俊哉

要　約

近年、テキストマイニングなどの人工知能分野の技術を、金融市場における分析に導入する研究が盛んに行われている。例えば、決算短信等のテキストデータから投資判断に有益な情報として業績等の因果関係文を抽出する手法の研究がされている。同様の手法で、有価証券報告書の「業績等の概要」の他、決算短信にはない、「対処すべき課題」や「事業等のリスク」などの項目から因果関係文を抽出することで、業績のみならずリスク対策や企業の施策などを把握するための有力情報が取得できると考えられる。しかし、有価証券報告書から因果関係文を抽出した報告はない。そこで、本稿では、坂地ら[坂地, 増山 2011]の手法を参考に、有価証券報告書から因果関係文を抽出する判別モデルを機械学習により作成し、因果関係文を抽出した。また、有価証券報告書の項目ごとに含まれる因果関係文の数や、坂地ら[坂地他 2015]を参考に原因表現と結果表現の抽出を行い、それぞれの表現に含まれる単語に違いがあるかを確認した。その結果、精度の高い判別モデルが作成できた。また、リスクに関する因果関係文が年々増えていること、「対処すべき課題」では企業の存続に関わる課題に関して企業ごとに認識している様々な要因が記載される傾向などが示唆された。これらは、有価証券報告書独自の投資判断に有益な情報の効率的な抽出や、テキストデータの定量評価手法を構築する上で有力な情報として役立つだろう。

目次

1. はじめに

投資判断に利用できる基本的な情報の１つとして、財務データやマーケットデータなどの数値データがある。これら数値データの特徴は、構造化しやすく統計解析なども比較的容易かつ大量に実施可能である点が挙げられる。実際にいくつものデータベンダーによってデータベース形式などの構造化データとして提供されており、合わせて分析ツールも提供される場合が多い。一方、数値以外のデータとして、ニュース、新聞記事、決算短信や有価証券報告書などに含まれるテキストデータがある。これらテキストデータの特徴は、過去や将来の業績に対する理由及び根拠、経営戦略に関する情報、進行中の施策、新商品発表情報、抱えているリスクや企業の不祥事に関する情報など、投資判断において数値データにはない重要な情報を含んでいる点が挙げられる。しかしながら、構造化することが難しく、そのほとんどは投資家が直接読まなければ投資判断に利用できない場合が多かった。近年、この問題に対し、テキストマイニングなどの人工知能分野の技術を、金融市場における分析に導入して解決を試みる研究が盛んに行われている。例えば、ある企業に関連するニュースがその株価にとってポジティブに働くかネガティブに働くか（極性）でニュースを定量化し、株式リターンとの関係を分析した研究[沖本, 平澤 2014] [五島他 2015]が挙げられる。これらの研究は、投資家が直接そのニュースを読まなくとも、ニュースに極性を付与し定量化することで、投資戦略に活用できることを示唆している。

テキストデータの定量化以外にも、テキストデータから投資判断等に関する重要文を抽出する手法の研究が行われている。磯沼ら[磯沼他 2016]は、決算短信から決算記事を自動生成することを目的として、決算記事などから推定した企業の事業セグメントの重要度と極性及び重要単語を用い、決算短信から決算記事を構成する文を抽出する手法を提案している。このような技術は公表されたテキストデータから投資家が注目すべき情報のみを要約し、即座に提示するツール等に一部応用され始めている[1]。坂地ら[坂地, 増山 2011]は、過去の業績や製品の売れ行きなどに対する原因と結果を含む文を因果関係文とし、その特徴的な表現と機械学習によって、因果関係文を経済新聞の記事から抽出する手法を提案している。さらに、坂地ら[坂地他 2015]は、決算短信の因果関係文から原因表現と結果表現を抽出する手法を用いて、企業の事業活動に付随する様々な因果関係を含む文や、ある原因(例えば、猛暑）などに対する結果(例えば、エアコン)とそれに結びつく企業を検索できるシステムを開発し公開している[2]。因果関係文の抽出以外にも、決算短信から業績の要因を含む文を抽出した研究[酒井他 2015b][3]や、業績の予測を含む文を抽出した研究[北森他 2017]があり、さらに抽出された文に極性を付与する研究[酒井他 2015a]も行われている。このように、テキストデータからの重要文の抽出手法は、投資家が投資判断に有益な情報を効率良く把握することを可能にし、また、ある特定のテーマを持った重要文の定量化データを既存の分析や投資戦略等へ導入することで、新しい投資戦略や手法の開発に役立つだろう。しかしながら、多様なテキストデータからある特定のテーマを持った重要文についての抽出手法やその定

[1] 決算短信から自動で決算記事を作成する決算サマリーというサービス（β版）が（株）日本経済新聞社から提供されており、このサービスに磯沼ら[磯沼他 2016]の技術も導入される予定となっている。
[2] CS（Causal expressions Search system）として Web に公開されている。http://hawk.ci.seikei.ac.jp/CS/
[3] CEES（Causal Expressions Extraction System）として Web に公開されている。http://hawk.ci.seikei.ac.jp/cees/

量化が多岐にわたり研究されているが、現在のところ、それらの統一的な抽出手法や数量化モデルは存在しない。言い換えれば、抽出対象となるテキストデータや重要文によって、抽出手法や定量化手法に工夫が必要であると言える。

有価証券報告書は、上場会社が証券取引所から求められている適時開示資料である決算短信に対し、金融商品取引法により提出が定められている開示資料である。有価証券報告書は、決算短信と比較して速報性はないものの[4]、「業績等の概要」等の業績に関する情報だけでなく、「対処すべき課題」や「事業等のリスク」など投資判断に有益と考えられる情報をより多く含む。また、ニュースや新聞記事は限られた企業に関する情報が多いが、有価証券報告書は全ての上場企業から公表されているという利点もある[5]。投資判断に有益なテキスト情報としては、まずは業績に関する情報が考えられる。どんな事象が原因で、そのような業績結果となったのかを知るには、業績に関する項目の因果関係文を抽出すれば良い。また、何をリスク要因や経営課題として捉えて、それに対しどんな対策を講じているのかを知るには、リスクやその企業がもつ課題に関する項目の因果関係文を抽出することで把握できると考えられる。つまり、有価証券報告書の各項目から因果関係文を抽出することは、単純に業績に関する原因と結果だけでなく、その企業のリスク対策や目指している方向性などを把握するための有力情報を取得することになると言えるだろう。にもかかわらず、因果関係文の抽出に関する研究の中で、新聞記事や決算短信を対象とした研究はあるものの、有価証券報告書を対象にした報告はない。そこで、本稿では、有価証券報告書から「業績等の概要」、「対処すべき課題」、「事業等のリスク」を対象に因果関係文の抽出を行った。具体的には、坂地ら[坂地, 増山 2011]の手法を参考に、有価証券報告書から因果関係文を抽出する判別モデルを機械学習により作成し、有価証券報告書の因果関係文の抽出を行った。さらに、抽出した文における原因部分と結果部分の関係を調べるため、坂地ら[坂地他 2015]を参考に、得られた因果関係文から原因表現と結果表現を抽出した。最後に、有価証券報告書の項目ごとに含まれる因果関係文の数及び原因表現と結果表現に含まれる単語に違いがあるかを確認した。

2. 利用したデータの定義

本稿においてテキストマイニングの対象としたデータの定義を図表 1 に示す。対象企業は TOPIX1000 の構成企業とした。対象テキストは、有価証券報告書の項目のうち、「業績等の概要」、「対処すべき課題」、「事業等のリスク」のテキストとした。なお、有価証券報告書のテキストデータは、PDF ファイルからテキストデータを抽出後、クリーニングを行った上で文を抽出したデータを用いた。

[4] 決算短信は、決算期末後 45 日以内の開示が適当とされ、30 日以内の開示がより望ましいとされている。一方、有価証券報告書は、やむを得ない場合を除いて事業年度末から 3 ヵ月以内の公表が求められている。

[5] 例えば金融庁の EDINET から入手が可能。http://disclosure.edinet-fsa.go.jp/

図表 1　テキストマイニングの対象としたデータ

対象	内容
期間	2008 年～2016 年
企業	TOPIX1000 の構成企業
有価証券報告書	発表日ベースで対象となる期間に含まれている本決算の有価証券報告書
項目	「業績等の概要」、「対処すべき課題」、「事業等のリスク」

(出所) 日興リサーチセンター作成

今回テキストマイニングの対象となった文の数について年ごと項目ごとの内訳を図表 2 に示す。項目ごとに確認すると、各年通して「業績等の概要」の文の数が相対的に多く、「対処すべき課題」の文の数が相対的に少ない。また、年による項目ごとの文の数の推移を見ると、「事業等のリスク」が徐々に増加している。その他の項目は多少の変動があるものの、比較的安定的な推移を示している。全期間、全項目の合計は 874,858 文となった。

図表 2　テキストマイニングの対象とした文の数の項目別内訳

発表年	「業績等の概要」	「対処すべき課題」	「事業等のリスク」	合計
2008	39,377	25,996	29,725	95,098
2009	40,401	26,914	31,286	98,601
2010	39,184	25,839	31,412	96,435
2011	39,028	25,466	32,939	97,433
2012	38,490	24,824	33,044	96,358
2013	37,651	25,184	33,666	96,501
2014	37,467	24,561	34,666	96,694
2015	36,925	25,282	35,630	97,837
2016	37,665	24,923	37,313	99,901
全期間	346,188	228,989	299,681	874,858

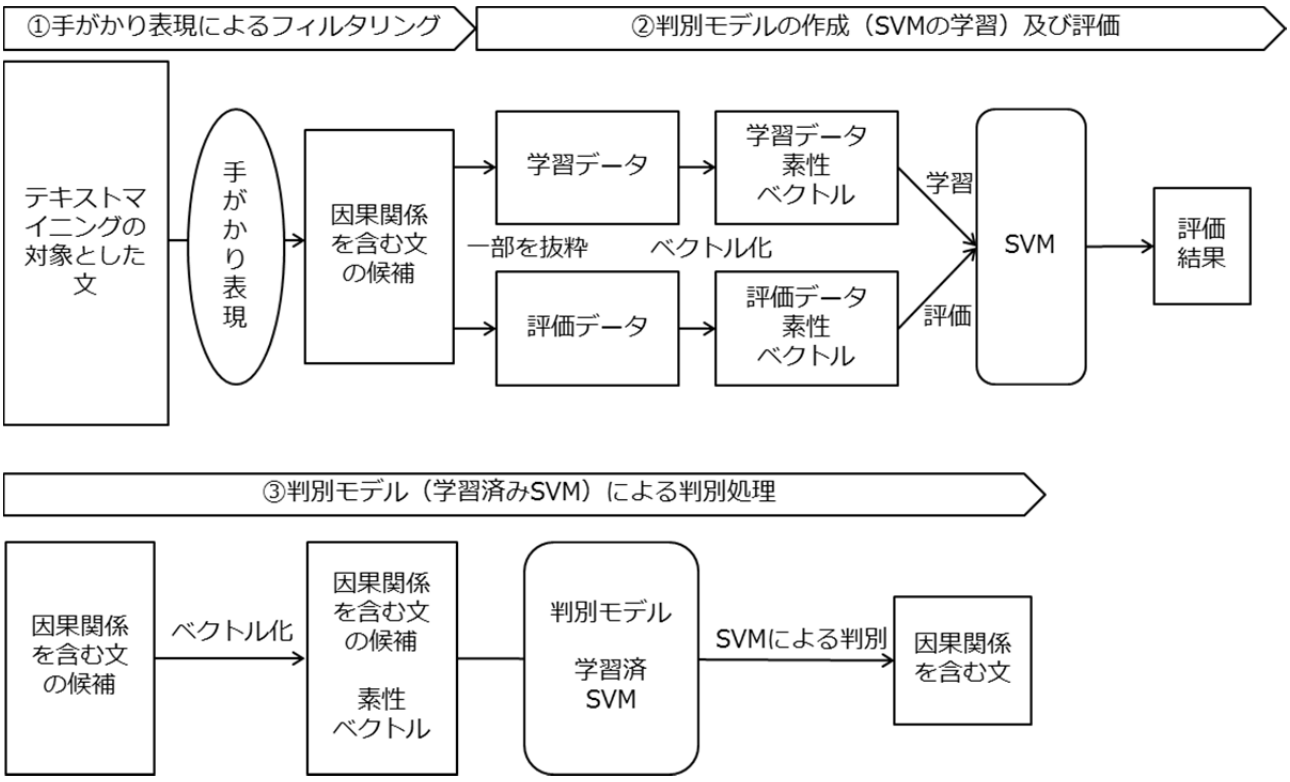
(出所) 日興リサーチセンター作成

3. 各種テキストマイニング手法

3.1 因果関係文の抽出手法

本稿では、新聞記事から因果関係を含む文を抽出する研究[坂地, 増山 2011]で用いられた手法を適用した。この手法は、より広範の因果関係文を対象とできることや高い抽出性能となったことが報告されている。本稿で行った因果関係を含む文の抽出の実際の処理として、まず、テキストマイニングの対象とした文から手がかり表現で因果関係を含む文の候補を抽出した。次に、この因果関係を含む文の候補の一部から学習データ及び評価データを作成した。そして、素性でベクトル化した学習データでサポートベクターマシン（以下、SVM）[6]を学習させ判別モデルを作成した。判別モデルの評価は、素性でベクトル化した評価データで行った。最後に、因果関係を含む文の候補を素性でベクトル化し、判別モデルで因果関係を含む文を抽出した。なお、SVM のカーネルは線形を用いた。手がかり表現と素性及び学習データと評価データについては後述する。図表 3 は因果関係文の抽出処理の概要を示す。また、抽出対象となる因果関係を含む文と抽出対象外である因果関係を含まない文の具体例を図表 4 に示す。

図表 3　因果関係文の抽出処理の概要



(出所) 日興リサーチセンター作成

[6] サポートベクター―マシン（SVM）とは、教師あり学習と呼ばれるパターン認識モデルの一つで、いくつかのラベルを付与されたデータを読み込み、それらの特徴（素性）を学習する事で、未学習のデータを分類する事の出来る機械学習の手法である。2値分類を解く上では、現在知られている多くの学習モデルの中でも優れた識別能力があると言われている。本稿では、python の機械学習のオープンソースライブラリである scikit-learn（http://scikit-learn.org/stable/index.html）を用いた。

図表４　因果関係を含む文と含まない文の例

因果関係を含む文	シューズ部門では、ランニングブームの継続と、フィッティングの取組みを強化したこと**により**、ランニングシューズの販売が堅調に推移いたしました。
	また、新興国を中心とする旺盛な需要や新しいエネルギー資源の開発など**を背景に、**当社グループの事業環境は好転しております。
	当連結会計年度におけるわが国経済は、米国の金融危機に端を発する世界経済の減速による影響**を受け、**年度後半から企業収益が大幅に減少し、雇用情勢が悪化するなど、景気は急速に悪化しました。
因果関係を含まない文	また、通商、独占禁止、特許、消費者、租税、為替管制、環境・リサイクル関連の法規制**を受けております。**
	現下の危機及び変動の大きい市場環境の中**で、**当社グループは、リカバリー・プランを遂行していく。
	平成 17 年 7 月 1 日**から、**製造たばこの販売に際しては、これらの規定に従っております。

注：太字は手がかり表現を示す。
(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

3.1.1 手がかり表現によるフィルタリング

因果関係を含む文の抽出の第一ステップとして、手がかり表現を含む文の抽出を行う。手がかり表現とは因果関係文を判定する上で重要な手がかりとなる表現を示す。例えば、「猛暑日が連続したため、飲料水の売上が伸びた。」という文の「ため、」が手がかり表現となる。本稿では、決算短信と有価証券報告書は記載される文が類似していることから、決算短信の手がかり表現[坂地他 2015]を参考に 37 個の手がかり表現を選定した。選定した手がかり表現を含む文が、因果関係を含む文の候補となる。ただし、2 文にまたがる因果関係や、手がかり表現が含まれていない文は対象外とした。選定した手がかり表現を図表 5 に示す。

図表５　手がかり表現一覧

を背景に	を背景に、	を受け、	ため、	に伴う
に伴い、	を反映して	をきっかけに	により、	に支えられて
によって	を反映し、	が響き、	ためで、	を受けて
から、	により	が響いた。	ため」	が影響した。
による。	ためで	ためだ。	を受けて、	に伴い
ため。	が響く	が響いている	が響いている。	で、
を受けております。	によります。	によっております。	ためであります。	によっています。
響いております	響いています			

(出所) 日興リサーチセンター作成

3.1.2 サポートベクターマシンによる因果関係文の抽出

（１）学習データ及び評価データの作成方法

SVM の学習データ及び評価データは因果関係を含む文の候補から下記の手順 1-1～手順 1-4 の手順で抽出した後、因果関係を含む場合に正例、含まない場合に負例とするラベル付与を人手で行った。文の抽出においては、精度の高い判別モデルを作成するため、学習データについてより広範の表現を抽出できるような工夫を行った。具体的には、有価証券報告書に記載される文が時期や業種、項目によって特徴が異なることが考えられるため、これらが均一に抽出されるようにした。

手順1-1　図表 1 に示した対象有価証券報告書を、有価証券報告書の発表日ベースで年ごとに振り分ける。

手順1-2　手順 1-1 の各年の有価証券報告書から業種ごとにランダムに 3 社を抽出。計 459（3 社×17 業種×9 年）の有価証券報告書が抽出される。業種は東京証券取引所が定めた東証 17 業種分類を利用する。

手順1-3　手順 1-2 で抽出した 459 の有価証券報告書から手がかり表現によるフィルタリングで因果関係を含む文の候補を抽出する。

手順1-4　手順 1-3 で抽出した各有価証券報告書の文から、「業績等の概要」、「対処すべき課題」、「事業等のリスク」の各項目からランダムに 1 文ずつ抽出。合計 1377（459×3）文が抽出される。なお、1 文も抽出されなかった有価証券報告書があった場合は、既に抽出済みの有価証券報告書以外で、同年、同業種内でランダムに取得された有価証券報告書を使用して同じ項目の 1 文をランダムに抽出する。

学習データ及び評価データをそれぞれ 1377 文抽出した後、人手で正例と負例のラベルを付与した。正例と負例の誤判定の発生を抑えるために、少なくとも 3 人の判定が一致するような判定手順で作業を行った。判定員は金融業務に従事する実務者が担当した。正例と負例のラベル付与の手順を手順 2-1～手順 2-4 に示した。学習データ及び評価データとも共通の手順となる。得られたラベル付きデータの内訳は、学習データが正例 782 文、負例 595 文、評価データが正例 733 文、負例 644 文となった。

手順 2 -1　抽出した学習データ（または評価データ）1377 文を 3 グループ（各 459 文）に分割する。
手順 2 -2　判定員計 9 人を各グループに 3 人ずつ割り当てる。
手順 2 -3　グループ内のそれぞれの文に対して、判定員 3 人が同じ判定だった場合はその判定を採用し正例または負例のラベルを付与する。
手順 2 -4　手順 2-3 でラベルが付与されなかった文に対しては、別の判定員 2 人が判定を行う。5 つの判定のうち、同じ判定が 3 つ以上となった判定を採用しラベルを付与する。

（2）素性の作成

判別モデルとして SVM を用いる際、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。本稿では坂地ら[坂地, 増山 2011]を参考に 4 つの素性を利用した。各素性の概要を図表 6 に示す。因果関係を含む文の特徴を適切に捉えるために、文に含まれる手がかり表現と構文的な素性である助詞のペアが素性として含まれている。SVM に入力される最終的なデータは、取得した全ての素性を並べて、文に含まれている素性を 1、含まれていない素性を 0 としたベクトルとなる。素性の作成においては、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[9]では係り受け解析器 CaboCha[10][工藤, 松本 2002]を用いた。

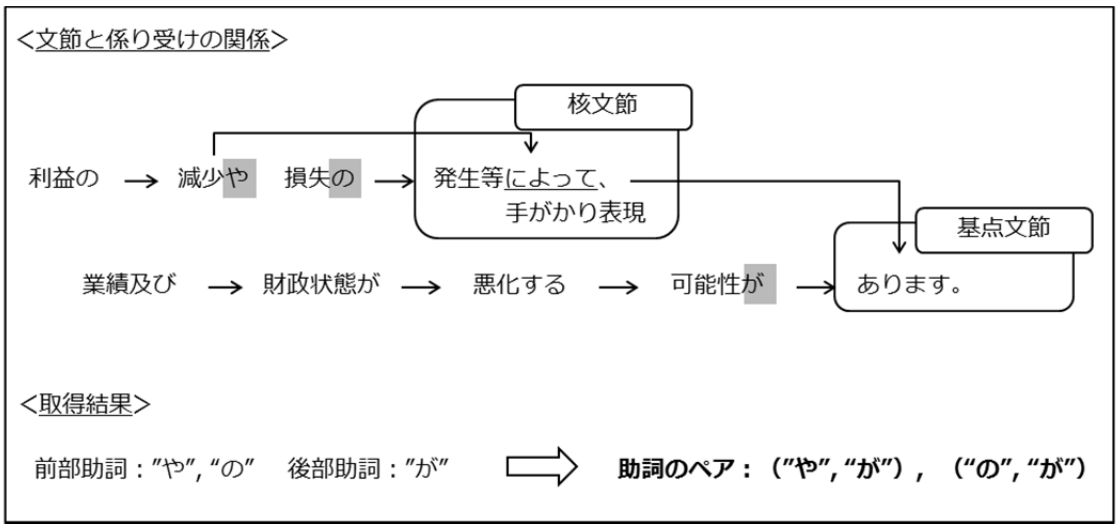
[7] 形態素とは言語学で使われる専門用語で、“意味の最小の単位”と説明され、テキストを形態素に分割することを形態素解析という[石田　2008]。
[8] MeCab は 京都大学情報学研究科–日本電信電話株式会社コミュニケーション科学基礎研究所 共同研究ユニットプロジェクトを通じて開発されたオープンソース 形態素解析エンジン。https://taku910.github.io/mecab/
文を入力すると、その文の形態素と各形態素の品詞が出力として得られる。
[9] 構文解析とは、文の構造を求めることであり、日本語の場合は、文節間の係り受けを求める場合が多い[石田 2012]。
[10] サポートベクターマシンに基づく日本語係り受け解析器。https://taku910.github.io/cabocha/
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。

図表 6　素性の概要

素性の名前	概要
助詞のペア	核文節に係る文節に含まれる助詞を前部助詞、基点文節に係る文節の助詞を後部助詞とし、前部助詞と後部助詞を合わせた全ての助詞のペア（重複を除く）。ただし、前部助詞が取得できない場合は、核文節より前の最も近い文節の助詞を前部助詞とする。存在しない場合は欠損値。また、後部助詞が取得できない場合は、核文節より後で基点文節に最も近い助詞を後部助詞とする。存在しない場合は欠損値。 ※核文節は手がかり表現を含む文節。基点文節は核文節の係り先の文節。 ※助詞のペアの取得イメージは図表 7 を参照。
文に含まれる手がかり表現	図表 5 を参照。
形態素ユニグラム	因果関係を含む文の候補を形態素解析器で分解した形態素のうち、頻度が 2 以上のものを抽出し重複を除いたもの。
形態素バイグラム	因果関係を含む文の候補を形態素解析器で分解し、隣り合った全ての形態素ペア（重複を除く）。

(出所) 日興リサーチセンター作成

図表 7　助詞のペアの取得イメージ



注：図表上部の <文節と係り受けの関係> は、係り受け解析を行った文節と係り受けの関係を示している。一塊の文字列が文節を示し、矢印は係り先を示す。
(出所) 日興リサーチセンター作成

3.2　原因・結果表現の抽出

抽出した因果関係を含む文から、構文パターンによって原因・結果表現を抽出する手法[坂地他 2015][Sakaji et al. 2008]を用い、原因表現と結果表現を抽出した。この手法は、原因表現と結果表現及び手がかり表現の出現パターンを調査し、5 つの構文パターンに分類したものを用いて、構文マッチングによって原因表現と結果表現を抽出するものである。ただし、本稿の手法では、前述したとおり同一文内に原因・結果表現が含まれている場合に限られるため、構文マッチングの対象となるパターンは、図表 8 に示した 3 パターンとなる。抽出処理はこの手法を実装した原因・結果表現抽出プログラム[11]を用いた。さらに、抽出された原因表現と結果表現を形態素解析し、各表現、各項目で頻度が高かった単語を確認した。ただし、品詞が形容詞、副詞、名詞のものだけを対象とした。頻度の数え方は、各有価証券報告書、各項目の原因表現、結果表現ごとに単語が含まれているかどうかで数え、各原因表現・結果表現に複数回出現した場合は 1 としてカウントしている。

図表 8　原因表現と結果表現及び手がかり表現の出現パターン

パターン名	構成イメージ
パターン A	原因表現＋手がかり表現＋結果表現
パターン B	結果表現の主部＋原因表現＋手がかり表現＋結果表現の述部
パターン C	結果表現＋原因表現＋手がかり表現

(出所) 日興リサーチセンター作成

4. 結果

4.1　判別モデルの評価結果

SVM を学習させ作成した判別モデルの評価結果を図表 9 に示す。因果関係を含む文の場合も、含まない文の場合も、精度、再現率、F 値[12]とも 0.8 を超える結果となった。この結果は、同様の手法で重要文を用いて行った評価と比較して遜色ない結果と言える。

[11] https://github.com/tetsuwaka/CausalExtraction

[12] 精度とは、どれだけ正確に正しい判定ができたかを示す指標。再現率とは、対象となるデータをどれだけ正しく判別できているかの網羅性を示す指標。F 値とは、一般的にトレードオフの関係にある精度と再現率のバランスを示し、精度と再現率の調和平均で計算される指標（F 値=2×精度×再現率/(精度＋再現率)）。

図表 9　判別モデルの評価結果

	精度	再現率	F 値	データ数
因果関係を含む文	0.85	0.89	0.87	733
因果関係を含まない文	0.87	0.82	0.84	644
平均／合計	0.86	0.86	0.86	1377

注：因果関係を含む文を正とすると、精度は正と判別したデータのうち、実際に正であるものの割合を示す。再現率は実際に正であるデータのうち、正しく判別されたものの割合を示す。F 値は精度と再現率の調和平均を示す。平均／合計の欄は、精度、再現率、F 値について、因果関係を含む文と含まない文の数で加重平均した値を、データ数については合計を示す。
(出所) 日興リサーチセンター作成

4.2　因果関係文抽出結果

図表 10 は抽出された因果関係文の結果であり、年ごと項目ごとに集計した因果関係文の数と抽出率（各年各項目の因果関係文の数を、同年同項目のテキストマイニングの対象とした文（図表 2 参照）の数で割って 100 を掛けた値）を示す。まず、抽出された因果関係文の合計は 176,886 文であり、抽出率は 20.2%となった。次に項目別に確認すると、「業績等の概要」では、因果関係文の数と抽出率の割合が、2009 年に増大し、それをピークに減少傾向がみられた。「対処すべき課題」では、因果関係文の数と抽出率ともに 3 項目中最も少なかった。「事業等のリスク」については、因果関係文の数が年々増加する傾向がみられた。

図表 10　抽出された因果関係文と抽出率の項目別内訳

発表年	「業績等の概要」	「対処すべき課題」	「事業等のリスク」	合計
2008	8,820（22.4%）	1,370（5.3%）	7,332（24.7%）	17,522（18.4%）
2009	10,455（25.9%）	1,652（6.1%）	8,478（27.1%）	20,585（20.9%）
2010	9,693（24.7%）	1,490（5.8%）	8,552（27.2%）	19,735（20.5%）
2011	9,365（24.0%）	1,598（6.3%）	9,032（27.4%）	19,995（20.5%）
2012	9,221（24.0%）	1,427（5.7%）	9,116（27.6%）	19,764（20.5%）
2013	8,700（23.1%）	1,402（5.6%）	9,318（27.7%）	19,420（20.1%）
2014	8,389（22.4%）	1,406（5.7%）	9,698（28.0%）	19,493（20.2%）
2015	8,692（23.5%）	1,440（5.7%）	10,048（28.2%）	20,180（20.6%）
2016	8,378（22.2%）	1,357（5.4%）	10,457（28.0%）	20,192（20.2%）
全期間	81,713（23.6%）	13,142（5.7%）	82,031（27.4%）	176,886（20.2%）

注：括弧内は抽出率。抽出率は、各年各項目の因果関係文の数/同年同項目のテキストマイニングの対象とした文の数×100。
テキストマイニングの対象とした文の数は図表 2 を参照。
(出所) 日興リサーチセンター作成

4.3　原因・結果表現の抽出結果

図表 11 は「業績等の概要」における原因表現及び結果表現のそれぞれについて、高頻度で出現した上位 30 の単語を示す。原因表現については、「価格、投資、経済、市場、円高、海外、震災」などが結果表現にはない単語として出現している。結果表現のみに出てきた単語は、「売上高、利益、損失、営業利益、増収、減収、増益、減益、前期」などであった。共通して出現している単語として、「増加、減少」などの数量の変化に関する単語があった。

図表 11　「業績等の概要」の原因表現及び結果表現に高頻度で出現した単語一覧

原因表現			結果表現		
1~10 位	11~20 位	21~30 位	1~10 位	11~20 位	21~30 位
増加	取得	固定資産	連結	販売	当期
減少	支出	低迷	増加	支出	損失
影響	会計	震災	会計	事業	利益
需要	計上	月	年度	需要	影響
事業	年度	効果	減少	減収	改善
販売	投資	市場	売上高	大幅	増益
連結	関連	円高	前期	増収	収益
価格	経済	拡大	推移	収入	景気
売上	景気	海外	営業利益	回復	減益
金	推移	改善	売上	資金	状況

(出所) 日興リサーチセンター作成

図表 12 は「対処すべき課題」における原因表現及び結果表現のそれぞれについて、高頻度で出現した上位 30 の単語を示す。原因表現については、「取締役会、株主総会、世界、震災、市場、景気、増加、減少、変動」などが結果表現にはない単語として出現している。結果表現のみに出てきた単語は、「企業価値、地位、共同、方針、対応、目的、維持、判断、買収防衛」などであった。共通して出現している単語として、「株主、皆様、新株予約権、利益、需要、規模、経済、影響」などの単語が抽出された。

図表 12　「対処すべき課題」の原因表現及び結果表現に高頻度で出現した単語一覧

原因表現			結果表現		
1~10 位	11~20 位	21~30 位	1~10 位	11~20 位	21~30 位
当社	月	プラン	当社	目的	地位
株主	世界	景気	株主	企業価値	平成
影響	需要	利益	利益	影響	状況
経済	規模	拡大	方針	環境	ない
事業	理由	株主総会	共同	基本	損害
新株予約権	取締役会	無償	皆様	対応	回復
平成	市場	任期	事業	可能	経済
環境	震災	減少	プラン	規模	判断
皆様	行為	変動	株式	需要	買収防衛
株式	増加	対抗	維持	新株予約権	ハンド

(出所) 日興リサーチセンター作成

図表 13 は「事業等のリスク」における原因表現及び結果表現のそれぞれについて、高頻度で出現した上位 30 の単語を示す。原因表現については、「環境、経済、規制、変化、動向、要因、自然災害、信用」などが結果表現にはない単語として出現している。結果表現のみに出てきた単語は、「財政、財務、成績、資産、費用、損失、悪影響」などであった。共通して出現している単語として、「発生、リスク、為替、製品、業績、経営」などの単語が抽出された。

図表 13　「事業等のリスク」の原因表現及び結果表現に高頻度で出現した単語一覧

原因表現			結果表現		
1~10 位	11~20 位	21~30 位	1~10 位	11~20 位	21~30 位
当社	状況	低下	影響	悪影響	減少
グループ	経済	増加	当社	事業	製品
変動	可能	販売	グループ	成績	費用
事業	悪化	自然災害	可能	状況	当行
影響	為替	動向	業績	リスク	為替
発生	製品	信用	発生	財務	市場
市場	価格	取引	財政	増加	販売
環境	要因	業績	状態	価格	資産
変化	規制	経営	変動	損失	活動
リスク	変更	予期	経営	低下	悪化

(出所) 日興リサーチセンター作成

5. 考察

因果関係文の抽出結果に関しては、テキストマイニングの対象とした文の約 20%が抽出された。つまり、有価証券報告書の「業績等の概要」、「対処すべき課題」、「事業等のリスク」の文全体に対して、約 1/5 が何らかの原因結果を記しており、これらの文を抽出することで、投資判断に有益な情報の効率的な把握に役立てられるだろう。抽出結果について項目別に推移を確認すると、まず、「業績等の概要」については、因果関係文の数と抽出率が 2009 年に増大し、それをピークに減少傾向が見られた。2008 年 9 月のリーマン・ショックによる影響で、翌年の 2009 年に発表された有価証券報告書の「業績等の概要」の記載文が増加し、その後、徐々に元の水準に戻っていったものと推察される。次に、「対処すべき課題」については、抽出率が「業績等の概要」や「事業等のリスク」と比較して低かった。この理由の 1 つとして、因果関係文は過去形の表現に対して多く含まれる傾向にあるが、「対処すべき課題」の文は現在形や未来形の表現が多い傾向にあるためだと考えられる。最後に、「事業等のリスク」では因果関係文の数が年々増加する傾向が確認できた。これは、リスクが多様化していることや、リスク情報は 1 度開示されると削除されにくく、その結果リスク情報の開示の数が増加を続けている[野田 2016]と考えられる。

原因表現と結果表現の抽出結果に関しては、項目ごとに原因表現と結果表現で含まれる出現頻度が上位 30 の単語に以下の特徴があった。まず、「業績等の概要」について、原因表現では、「経済、市場、円高、海外」など経済やマーケット環境と関連する単語が結果表現にはない単語として抽出された。また、「震災」も原因表現のみで抽出された。結果表現のみに出てきた単語を確認すると、「売上高、営業利益、増収、減益」などの業績に関わる具体的な個別の損益項目の単語が多かった。これらのことから、経済やマーケット環境、自然災害を業績の原因に記載する企業が多いと推察される。また、「前期」という単語が結果表現に出ていることから、前期と比較して売り上げや利益がどうだったかを示す結果表現が多いことも推察される。次に、「対処すべき課題」について、原因表現では、「取締役会、株主総会、世界、震災、市場、景気、増加、減少、変動」などの単語が結果表現にはない単語として抽出された。これらを一括りにする解釈は難しく、対処すべき課題として、その原因を何と認識するかは、企業によって様々であるのかもしれない。結果表現のみに出てきた単語を確認すると、「企業価値、地位、買収防衛、方針、目的、維持、判断」など会社及び組織運営に関連するような単語が抽出された。「対処すべき課題」の結果表現は企業としての存続に関する内容が記載されている場合が多いことが推察される。共通して出現した単語に、「株主、皆様」が抽出された。これはステークホルダーに関する記載が多いことが推察される。最後に、「事業等のリスク」について、原因表現では、「環境、経済、規制、自然災害、動向、変化、要因」など主に企業を取り巻く環境に関する単語が結果表現にはない単語として抽出された。リスクの要因は規制、災害、経済及びそれらの変化などに分類できるだろう。結果表現のみに出てきた単語を確認すると、「財政、財務、成績、資産、費用、損失」などの業績に関する個別の項目を包括する単語であるものが多かった。企業は規制、災害、経済とそれらの変化をリスク要因とし、その影響を「業績等の概要」のような個々の項目ではなく、それらを包括するより大きな項目に対して記

載していることが多いと推察される。

6. まとめ

本稿では、有価証券報告書から因果関係を含む文を抽出する判別モデルを作成した。また、作成した判別モデルを用いて抽出した因果関係文を年別、項目別に集計し、その特徴を確認した。さらに、因果関係文から原因表現と結果表現を抽出し、各表現に出現する単語から項目別の特徴を確認した。

判別モデルに関しては、精度、再現率、F 値とも 0.8 を超える結果となり、有価証券報告書から因果関係文を抽出する判別モデルとして高い性能を示すことができた。この判別モデルを用いることで、他のテキストデータにはない有価証券報告書独自の投資判断に有益な情報を効率よく抽出し、把握することが可能になったと言えるだろう。また、抽出された因果関係文や原因表現、結果表現は、項目ごとに異なる特徴があった。因果関係文の抽出数をみると、「業績等の概要」では、業績に多大な影響を及ぼすイベントによって因果関係文が増加する可能性があること、「対処すべき課題」では、他の項目と比較して因果関係文の含まれる割合が少なく、その他のテーマを持つ文の割合が多いこと、「事業等のリスク」では、リスクの多様化や過去のリスク情報が記載され続ける傾向を反映して因果関係文が年々増加していることが示唆された。また、原因表現と結果表現の頻出単語をみると、因果関係文の記載の傾向として、「業績等の概要」では、個別の損益項目に関する業績結果の原因として経済やマーケット環境が示され、「対処すべき課題」では、企業の存続に関わる課題に関して企業ごとに認識している様々な要因が示され、「事業等のリスク」では、企業を取り巻く環境を原因とする業績の包括的項目への影響が示されていることが示唆された。これらの特徴の違いは、企業の業績評価や業績予測、経営戦略や施策、認識するリスクの変化の把握等の評価を目的とした場合、結果だけでなく原因も考慮したテキストデータの定量評価手法を構築する上で有力な情報となり得るだろう。

今後の課題として、本稿で明らかになった有価証券報告書の 3 つの項目に対する因果関係文の原因表現と結果表現の特徴を踏まえた定量評価を行いたい。また、これら定量化したデータと企業の業績や株価との関係性を分析し、投資戦略への応用を検討していきたい。

＊ 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助教坂地泰紀先生に、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。

以上

参考文献

五島圭一・高橋大志・寺野隆雄[2015], 「ニュースのテキスト情報から株価を予測する」, 第 29 回人工知能学会全国大会 大会論文集, Vol.29, pp.1-3, May 2015.

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

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.

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 [2016], 「業績変動を考慮した決算短信からの重要文抽出」, 研究報告自然言語処理 (NL), Vol.2016-NL-227, No.6, pp.1-6, Jul. 2016.

北森詩織・酒井浩之・坂地泰紀[2017], 「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, pp150-161, Feb. 2017.

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, pp1834-1842, 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, pp89-96, May 2004.

野田健太郎 [2016],「有価証券報告書における定性情報の分析と活用」, 経済経営研究 , Vol.37, No.1, May 2016.

沖本竜義・平澤英司 [2014], 「ニュース指標による株式市場の予測可能性」, 証券アナリストジャーナル, Vol.52, No.4, pp.67-75, Apr. 2014.

酒井浩之・小林義和・坂地泰紀[2015a],「企業の決算短信ＰＤＦから抽出した業績要因への極性付与」, 第 15 回 金融情報学研究会, pp.7-12, 2015.

酒井浩之・西沢裕子・松並祥吾・坂地泰紀[2015b], 「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.

坂地泰紀・竹内康介・関根聡・増山繁 [2008],「構文パターンを用いた因果関係の抽出」, 言語処理学会 第 14 回年次大会 発表論文集, pp.1144-1147, Mar. 2008.

坂地泰紀・増山繁 [2011],「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, pp1496-1506, Aug. 2011.

Hiroki Sakaji, Satoshi Sekine, Shigeru Masuyama[2008], “Extracting Causal Knowledge Using Clue Phrases and Syntactic Patterns”, 7th International Conference on Practical Aspects of Knowledge Management (PAKM), pp.111-122, Yokohama, Japan, 2008.

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, pp811-822, May 2015.