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以外の特徴文の抽出

Research Report
2018年1月

投資工学研究所
小寺　俊哉
佐藤　史仁
田中　良典

要　約

近年、金融市場においてテキストマイニングの注目度が増し、金融経済月報や企業の決算短信等様々な媒体に対してテキストマイニングを用いた研究が盛んに行われている。坂地ら[坂地他, 2015]は機械学習を用いて、決算短信から原因・結果を含む因果関係文を特徴文として抽出する手法を開発した。佐藤ら[佐藤他, 2017]は坂地らの手法を参考に、有価証券報告書から因果関係文の抽出を試みたところ、「対処すべき課題」の項目では因果関係文以外の特徴文の存在が示唆された。そこで、本稿では有価証券報告書の「対処すべき課題」の項目における企業の課題に関する記述を「課題文」と定義し、課題文の文末表現の特徴から、手がかりとなる表現を自動獲得手法により増幅し、それらを用いて課題文の抽出を試みた。その結果、抽出した課題文を確認したところ判別モデルの精度は良好であった。また、抽出した課題文の数は「対処すべき課題」全文の数を上回るペースで年々増加する傾向であり、近年の有価証券報告書における課題に関する開示意識の高まり等が示唆された。さらに、抽出した課題文を用いて、単語ごとに各業種における特有度合いをスコア化し、業種特有の課題に関する単語の抽出を試みた結果、それぞれの業種の課題に則した単語を抽出することが出来た。このように本稿での提案手法は、有価証券報告書を用いた分析手段として有用であり、投資判断情報の効率的な獲得に役立つだろう。

目次

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導入の注目度が増している。例えば、日本銀行が毎月発行している金融経済月報や、新聞記事に機械学習を用いたテキストマイニングの技術により、経済市場を分析する研究等が盛んに行われている。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手法が長期的な市場分析に有効であると示した。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、日経平均と TOPIX の 1 ヶ月後の騰落予測において 6 割以上の精度であった。これらの研究では、指標等のマクロ的な対象の分析におけるテキストマイニングの有効性を示した。しかし、ミクロな観点から個別企業を対象に分析する場合は、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を総合的に評価することが効果的だと考えられる。ただし、それらの評価には専門的な知識や経験が必要なうえ、企業間の相対的な評価を行う場合、大量の企業を同時に一定の基準を持って評価するのは困難であるという問題がある。

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、自動でテキストデータから重要部分を判断し抜き出す研究や定性的なテキストデータを定量化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、業績等の記述のある決算短信に着目し、機械学習を用いて自動で決算短信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ことの出来るモデルを開発した。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とした。ところで、決算短信は上場企業が証券取引所から求められている適時開示資料であることに対し、有価証券報告書は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[1]。その特性として、決算短信と比較して速報性はないもの[2]、「業績等の概要」等の業績に関する項目だけでなく、「対処すべき課題」や「事業等のリスク」等の決算短信にはない項目があり、投資判断に有用と思われる情報が多く含まれている。そこで佐藤ら[佐藤他, 2017]は、坂地らの手法を有価証券報告書に応用し、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 3 項目から原因と結果の因果関係文を特徴文として抽出を行った。その結果の 1 つとして、「対処すべき課題」の項目において他 2 項目と比較して抽出された因果関係文の数の割合が少なかった。その要因の 1 つは当該項目の特徴に因るものと考えられる。企業内容等の開示に関する内閣府令では「対処すべき課題」の項目を「最近日現在における連結会社（連結財務諸表を作成していない場合には提出会社）の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

[1] 有価証券報告書は、金融商品取引法第二十四条により提出が定められている。金融庁の EDINET から入手可能。http://disclosure.edinet-fsa.go.jp/

[2] 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。

等を具体的に記載すること」としている。さらに、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から「対処すべき課題」を「経営方針、経営環境及び対処すべき課題等」に変更することとなっている。このように「対処すべき課題」での因果関係文の数の割合が少なかったことと、企業が取り組むべき課題を記述しているという項目の特徴を踏まえると、当該項目には因果関係とは異なる特性の企業経営に係る重要な情報が含まれている可能性があると考えられる。

本稿では、有価証券報告書の「対処すべき課題」の項目における因果関係文以外の特徴文を企業の課題に関する記述の文であるとした。この「対処すべき課題」における特徴文を「課題文」と定義し、課題文の特徴として文末の表現に注目して抽出を行った。しかしながら、これらの課題文から課題を抽出できたとしてもそれぞれの企業における課題は多種多様であると推測され、多くの企業について一律で単純な比較・評価を行うのは難しいだろう。そこで、多種多様な課題を整理するための一助として、抽出した課題文から、単語ごとに各業種における特有度合いをスコア化し、業種特有の課題に関する単語の抽出を試みた。

2. 課題文の抽出方法

2.1 利用データ

本稿で用いるテキストデータの基本情報は図表 1 に示す。対象期間は 2008 年～2016 年とし、対象企業は各年における本決算の有価証券報告書発表時に東京証券取引所の市場第一部に上場している企業とした。対象企業の本決算の有価証券報告書 PDF ファイルをテキスト化し、文単位に分割した。そして、それらの中から「対処すべき課題」の項目の文をテキストデータとした。

図表 1　対象テキストデータの基本情報

対象	内容
期間	2008 年～2016 年
企業	有価証券報告書発表時、東京証券取引所の市場第一部に上場している企業
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」

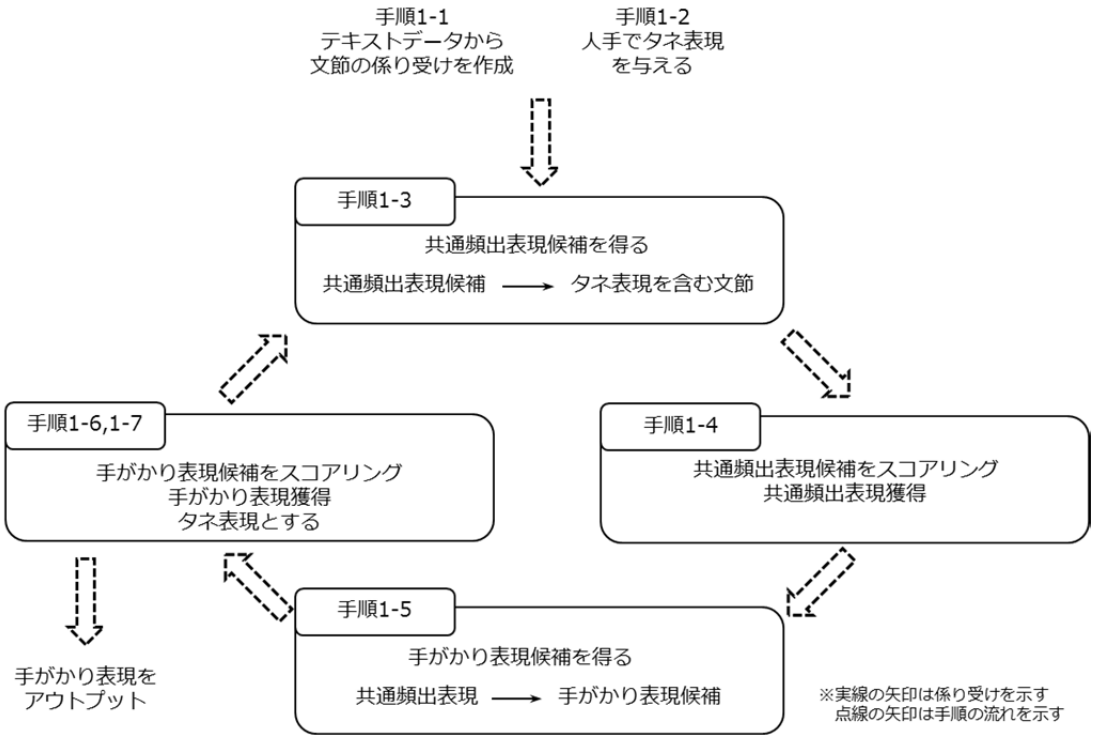
（出所）日興リサーチセンター作成

2.2 手がかり表現の自動獲得

課題文を抽出するにあたり、まずは課題文の特徴を捉える必要がある。「対処すべき課題」の項目における課題に関する記述の文は、「取り組みます。」や「参ります。」のように文末に特徴が見られた。それらの文末表現を手がかりとすることで、課題文を抽出した。本稿では、そのような文末表現を「手がかり表現」と呼ぶこととした。

それら手がかり表現を獲得するために、酒井ら[酒井他, 2016]の手法を参考に手がかり表現の自動獲得手法を試みた。本稿では、手がかり表現に対して係る文節の内、多くの手がかり表現に共通して頻繁に係る文節は、既存の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手がかり表現として獲得した。例えば、既存の手がかり表現として「取り組みます。」を与え、共通頻出表現として「改善に」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「努めます。」を獲得し、これを新たな手がかり表現とする。ただし、本稿では獲得する手がかり表現を文末表現に絞るため、末尾が句点のもののみを対象とした。この手順を繰り返し行うことで手がかり表現を自動で獲得した。手がかり表現の自動獲得手法のイメージ図を図表 2 に、具体的な手順を手順 1-1～手順 1-7 に、獲得した手がかり表現の一部を図表 3 に、共通頻出表現の一部を図表 4 に示す。手がかり表現の自動獲得手法によって 107 個の手がかり表現と 390 個の共通頻出表現を獲得した。この手がかり表現を用いて課題文を抽出することとする。

図表 2 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

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[3][工藤, 松本 2002]を用いて構文解析[4]し、文節に区切り係り受けを作成する。

手順1-2. 少数の手がかり表現（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「取り組みます。」、「参ります。」、「致します。」、「行います。」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 0.4 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。

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

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

$S(e)$：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。

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

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

[3] 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/

[4] 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]。

$E(s'')$：手がかり表現候補s''に係る共通頻出表現の集合

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

$f(s'', e')$：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。

図表 3　手がかり表現例

獲得した手がかり表現の一部		
図ります。	実行します。	目指します。
展開します。	注力する。	推進します。
強化いたします。	予定しております。	行っております。

(出所) 日興リサーチセンター作成

図表 4　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	構造改革を	収益拡大を
グローバルに	改善に	研究開発を
徹底して	今後も	体制作りを

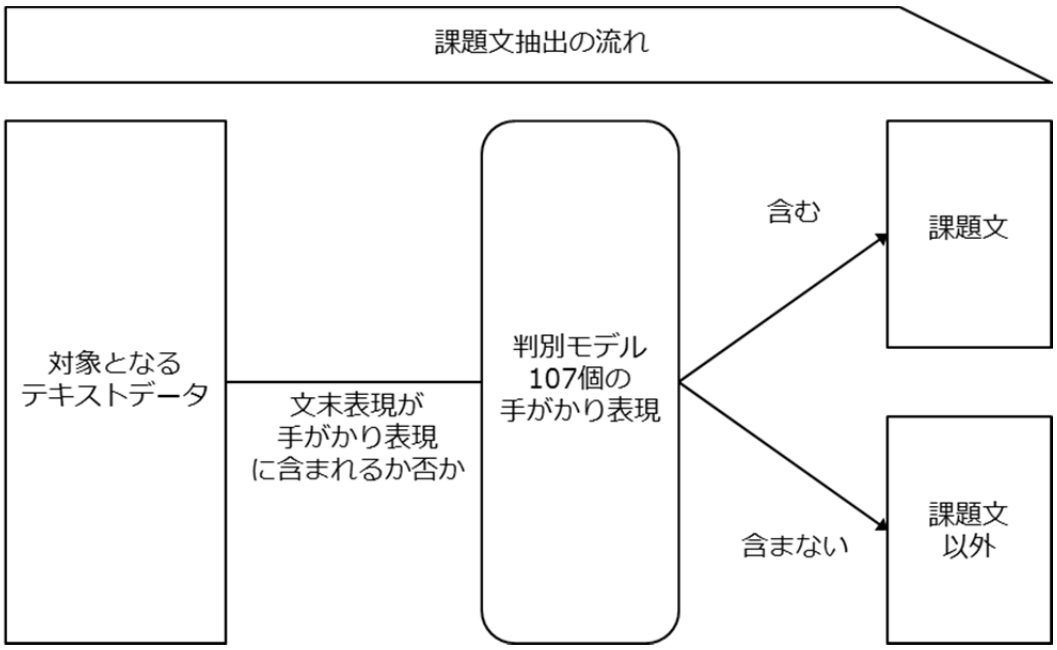
(出所) 日興リサーチセンター作成

2.3　課題文の抽出

本稿では、課題文の特徴は文末にあるとし、文末表現を手がかり表現として獲得した。この手がかり表現を用いて、各文に対して、手がかり表現を含む場合は課題文、含まない場合はそれ以外の文と判別するモデルを作成した。判別モデルのイメージを図表 5 に示す。

また、判別モデルの性能評価を行うために、抽出した課題文の判定を行った。抽出した課題文から無作為に 100 文を抽出し、複数人で正誤を判定した。その際、図表 6 のような文を課題文とし、判定基準として、企業が今後取り組むべき課題を含んでいる文を課題文、完了形や、具体的な課題内容を含んでいない文等をそれ以外の文とした。その結果により判別モデルの性能を評価した。

図表 5　課題文判別モデルのイメージ図



(出所) 日興リサーチセンター作成

図表 6　課題文の例

課題文の例
売上拡大を課題とし、グローバル展開に取り組みます。
収益力の改善に向けて、店舗運営の効率化を推進致します。

(出所) 日興リサーチセンター作成

2.4　業種特有の課題に関する単語の抽出

次に抽出した課題文から業種特有の課題に関する単語を抽出することを試みた。本稿では業種区分として、図表 7 に示す東京証券取引所が定めた東証 17 業種分類を用いて各業種特有の課題に関する単語を抽出することとした。

図表 7　本稿で用いる業種分類

東証 17 業種分類			
食品	自動車・輸送機	電力・ガス	金融（除く銀行）
エネルギー資源	鉄鋼・非鉄	運輸・物流	不動産
建設・資材	機械	商社・卸売	
素材・化学	電機・精密	小売	
医薬品	情報通信・サービス その他	銀行	

(出所) 日興リサーチセンター作成

まず、抽出した課題文を MeCab[5]によって形態素[6]に分割した。その形態素を単語と表記することとした。ここで、分析対象は名詞と動詞とし、また単独では意味をなさないものや、課題文全体での頻度が 10 未満の単語は分析対象から除くこととした。各単語の全企業での出現頻度、各業種での出現頻度及び各企業での出現頻度を算出した。各単語に対して出現確率を計算し、全企業内での業種間のエントロピー[7]と、各業種内での企業間のエントロピーを求め、それらエントロピーの比を$Score$とした。この$Score$が大きい単語、つまり全体では特定の業種に偏って出現するが、業種内では万遍なく様々な企業に出現するような単語を業種特有の課題に関する単語とした。計算の手順を以下の手順 2-1～手順 2-5 に示す。

手順2-1　単語tに対し、式 7 を用いて業種sec_iでの頻度を全企業での頻度で除し、単語tの業種sec_iごとの出現確率$P_{all}(t, sec_i)$を算出する。

$$P_{all}(t, sec_i) = \frac{\text{業種}sec_i\text{での単語}t\text{の頻度}}{\text{全企業での単語}t\text{の頻度}} \tag{7}$$

手順2-2　単語tに対し、式 8 により業種間のエントロピー$H_{all}(t)$を算出する。ただし、ここでエントロピーが 0 のものは、1 つの業種でのみ出現する単語で、一般性の無いものといえるため除くものとする。

$$H_{all}(t) = -\sum_{i=1}^{17} P_{all}(t, sec_i) \log_2 P_{all}(t, sec_i) \tag{8}$$

[5] 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/

[6] 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]。

[7] エントロピーとは、乱雑さや平均情報量と訳され、ばらつき具合を表している。つまり、エントロピーが高ければ、全体で万遍なく出現することを示し、低ければある特定の所で出現していることが分かる。

手順2-3　次に単語tに対し、式 9 により業種sec_iにおける企業com_jの頻度を業種sec_i全体の頻度で除し、単語tの業種sec_iにおける企業com_jの出現確率$P_{sec_i}(t, com_j)$を算出する。

$$P_{sec_i}(t, com_j) = \frac{\text{企業}com_j\text{での単語}t\text{の頻度}}{\text{業種}sec_i\text{での単語}t\text{の頻度}} \tag{9}$$

手順2-4　単語tに対し、式 10 により業種sec_iにおける企業間のエントロピーを算出する。

$$H_{sec_i}(t) = -\sum_{com_j \in COM(sec_i)} P_{sec_i}(t, com_j) \log_2 P_{sec_i}(t, com_j) \tag{10}$$

$COM(sec_i)$：業種sec_iに含まれる企業の集合

手順2-5　最後に単語tに対し、式 11 のように手順 2-4 で算出した業種sec_iにおける企業間のエントロピーを手順 2-2 で算出した業種間のエントロピーで除することで、単語tの業種sec_i内での$Score$を算出する。

$$Score(t, sec_i) = \frac{H_{sec_i}(t)}{H_{all}(t)} \tag{11}$$

3. 結果

3.1 課題文の抽出

判別モデルの性能評価に関して、抽出した課題文から無作為に抽出した 100 文を複数人で確認したところ、87 文が課題文と判定された。

また、図表 8 は、各年の「対処すべき課題」全文の数と抽出した課題文の数及びその割合を表している。「対処すべき課題」全文の数は、2010 年、2011 年、2012 年で減っているものの、それ以降 2016 年までは増えていた。また、課題文の割合は、16%～20%程度で増加傾向であった。

図表 9 は抽出した課題文の発表年、企業名、課題文の内容の例を表している。ただし、下線部分は課題文抽出に用いた手がかり表現を示している。

図表 8 課題文の抽出割合

発表年	「対処すべき課題」全文の数	課題文の数	割合（%）
2008	38,760	6,371	16.4
2009	38,898	6,556	16.8
2010	37,567	6,375	16.9
2011	36,867	6,238	16.9
2012	36,712	6,519	17.7
2013	37,912	6,838	18.0
2014	39,082	7,297	18.6
2015	40,855	7,685	18.8
2016	41,910	7,975	19.0
計	348,563	61,854	17.7

（出所）日興リサーチセンター作成

図表 9　抽出された課題文

発表年	企業名	課題文
2009	トヨタ自動車	また、需要拡大の見込まれる資源国・新興国については、商用車や低価格車の商品力強化を着実に**進めていきます。**
2015	日本電信電話	さらに、平成 24 年度末で約 700 名であった課長相当職以上の女性管理者について、管理者全体に占める割合を平成 32 年度までに倍増させるという目標を掲げており、多様性の尊重と機会均等に向けても**取り組んでまいります。**

（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

3.2　業種特有の課題に関する単語の抽出

各業種における業種特有の課題に関する単語として*Score*が 1 位～10 位の単語を図表 10～図表 13 に示す。ただし、特定の企業、製品等を表すような単語は除いた。

食品業種では「加工食品、茶、ビール」、エネルギー資源業種では「メタンハイドレート」、自動車・輸送機業種では「軽自動車、燃費、小型車」、電力・ガス業種では「火力発電所、原子力、石炭火力発電所、原子力発電」、小売業種では「大型店、開店、陳列、フォーマット、新店」、金融（除く銀行）業種では「投資信託」、不動産業種では「空室率、テナント、マンション」等の単語が抽出された。

図表 10　各業種の*Score*が 1 位～10 位の単語

食品	エネルギー資源	建設・資材	素材・化学
加工食品	探鉱	積算	ポリマー
砂糖	埋蔵量	施工	高分子
酒類	メタンハイドレート	ガラス	無機
チョコレート	ＳＳ	官庁	コンパウンド
茶	輸送	臨海	酸化チタン
冷凍食品	潤滑油	EPC	塩素
飲料	石油	土木	衣料
ビール	石油製品	ゼネコン	エアバッグ
自動販売機	化学製品	耐火	顔料
乳酸菌	石油精製	受注競争	企業化

(出所) 日興リサーチセンター作成

図表 11　各業種の*Score*が 1 位～10 位の単語

医薬品	自動車・輸送機	鉄鋼・非鉄	機械
ＯＴＣ	軽自動車	特殊鋼	油圧ショベル
癌	ゴム	亜鉛	パチンコ
学術	燃費	電線	パチスロ
創製	小型車	圧延	工作機械
腎	製品	鋳鉄	原子力
疾患	生産	めっき	遊技
新薬	技術	鋼管	建設機械
疼痛	部品	製鋼	EPC
錠	取り組む	ケーブル	熱処理
神経	強化	鋼板	機種

(出所) 日興リサーチセンター作成

図表 12　各業種の*Score*が 1 位～10 位の単語

電機・精密	情報通信・サービス その他	電力・ガス	運輸・物流
高周波	興行	経年	沿線
インバータ	視聴	火力発電所	船隊
カーエレクトロニクス	アニメ	沿線	コンテナ
真空	BPO	原子力	観光資源
計測	com	送配	ダイヤ
AV	番組	石炭火力発電所	列車
ＭＥＭＳ	ゲームソフト	混焼	バリアフリー化
加熱	制作	エネファーム	海運
プリンティング	通信サービス	原子力発電	不動産業
ワイヤレス	オンラインゲーム	ヒートポンプ	運航

(出所) 日興リサーチセンター作成

図表 13　各業種の*Score*が 1 位～10 位の単語

商社・卸売	小売	銀行	金融（除く銀行）	不動産
専門商社	全店	バンク	預り	空室率
卸売	大型店	住宅ローン	債券	分譲
商社	開店	リスクテイク	引受	不動産業
酒類	客数	中小企業金融	アドバイザリー	分譲マンション
水産物	陳列	銀行	証券	用地
安全保障	必需	融資	融資	物件
繊維	フォーマット	地域金融機関	投資信託	テナント
商	新店	貸出	銀行	マンション
仕る	ドミナント	金融サービス	貸付	店舗
メディア	接客	全店	金融サービス	オフィスビル

(出所) 日興リサーチセンター作成

4.　考察

4.1　課題文の抽出

本稿における判別モデルの性能評価に関して、100 文中 87 文について企業が今後取り組むべき課題を含んでいると判定されたことから、良好な精度と言えるだろう。

課題文の抽出結果において、「対処すべき課題」全文の数は 2012 年から 2016 年にかけて増加傾向であった。これは、対象を東京証券取引所の市場第一部に上場している企業としているため、対象企業数

が増加していることが１つの要因として考えられる。しかしながら、抽出された課題文の数の割合は対象期間において増加傾向であった。「対処すべき課題」全文の数が増えているため、課題文の数が増えるのは容易に想定できるが、その上で課題文の数の割合が増えているのは、近年の有価証券報告書における課題に関する開示意識の高まりや、企業の事業規模や事業領域の拡大に伴う課題の多様化を示唆しているかもしれない。

抽出された例として、2009 年のトヨタ自動車と 2015 年の日本電信電話の課題文を挙げた。2009 年のトヨタ自動車の課題文では、課題と捉えている対象地域や、対象商品についての情報が記されており、事業戦略における課題に関する情報を得ることが出来た。2015 年の日本電信電話の課題文では、人事面での課題が記されており、抽出された課題文から事業戦略以外の課題に関する情報も取得することが出来た。このように本稿における提案手法によって抽出した課題文は、企業の課題に関する情報を含んでおり、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。

4.2　業種特有の課題に関する単語の抽出

業種特有の課題に関する単語として、食品業種と金融（除く銀行）業種において抽出された単語は、それぞれ「加工食品、茶、ビール」、「投資信託」といった企業が取り扱う商品そのものに関すると推察される単語が抽出され、企業が今後注力すべき商品として課題に挙げている可能性が考えられる。エネルギー資源業種では、「メタンハイドレート」という新たなエネルギー資源として期待される単語が抽出され、それに対する開発技術に関する課題を認識していることが推察される。自動車・輸送機業種では「軽自動車、燃費、小型車」といった単語が抽出され、国内市場で販売が好調な小型車及び軽自動車に関する事業戦略を課題として注力していることや、これらに必要な低燃費技術の研究開発が企業にとって重要な課題になっていることが連想される。電力・ガス業種では、「火力発電所、原子力、石炭火力発電所、原子力発電」といった単語が抽出されており、2011 年 3 月 11 日に発生した東日本大震災の影響による事業環境の変化に対応するための事業構造の見直しが注力すべき課題として挙げられているのかもしれない。小売業種では「大型店、開店、陳列、フォーマット、新店」といった単語が抽出され、新店や大型店舗及び陳列方法に関する店舗戦略に課題を抱えているということが考えられる。一方で、商品に関する単語は抽出されず、このことは企業が課題として商品戦略よりも店舗戦略を重視していることが推察される興味深い結果であった。

このように、業種によって事業環境や事業戦略、商品戦略、技術開発等の多岐にわたる課題に関する様々な単語が抽出されたが、それらは各業種における課題を表していた。意外なことに経営戦略に関する単語はあまり抽出されなかった。このことは経営戦略に関する単語は業種を問わず課題とされるため、業種特有の課題としては抽出されなかったことが考えられる。

5. おわりに

本稿では、有価証券報告書の「対処すべき課題」における因果関係文以外の特徴文として企業の課題に関する記述の文の抽出を行った。その結果、企業の課題に関する情報を含み、投資判断に有益な情報となり得る課題文を抽出することが出来た。本稿の提案手法により情報収集の効率化を図ることが出来る可能性を示唆した。ただし、本稿で用いた抽出手法では、手がかり表現の有無の判定のみで判別を行ったため、文末が体言止め等のものは抽出対象から除外した。また、出現頻度が著しく低い手がかり表現は手がかり表現の自動獲得手法で獲得することが難しく、そのような手がかり表現を含む課題文は対象外となっている。今後は、文末だけでなく、文頭、文中の表現や、係り受け、あるいは前後の文の内容などを考慮した判別モデルの作成が必要になるかもしれない。

さらに、抽出した課題文を用いて、業種特有の課題に関する単語を抽出することが出来た。今後の発展として、時系列での課題の変化の抽出や、課題のクラスタリング、他項目と対応させた分析等が挙げられるだろう。

＊ 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

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

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.

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, pp1834-1842, Jun. 2002.

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, Vol28, No. 3, pp. 291-296, 2013.

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, pp811-822, 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.