# 有価証券報告書からの因果関係文の抽出 Extraction of Causal Knowledge from Annual Securities Report

左藤 史仁 <sup>*1</sup> 佐久間 洋明 <sup>*1</sup>	小寺 俊哉 <sup>*1</sup>	田中 良典 <sup>*1</sup>	坂地 泰紀 <sup>*2</sup>	和泉 潔 <sup>*2</sup>
Fumihito Sato Hiroaki Sakuma	Shunya Kodera	Yoshinori Tanaka	Hiroki Sakaji	Kiyoshi Izumi
*1 日興リサーチセンター株: Nikka Research Contar Ins	式会社	*2 東京大学	之 大学院工学系研	F究科

In annual securities report, various information such as corporate policy, risk management, R&D, and so on, is included other than business performance. Previous researches proposed the extraction methods of important sentences containing causal information from financial articles and texts but not annual financial reports. In this paper, we applied these extracting methods based on SVM discriminant model to annual securities reports in our original way. Our method indicated high performance and all evaluations, that were precision, recall and F-score, showed more than 0.8. By using our model, useful information from annual securities reports would be collected effectively, which allow us to make unique investment decisions.

## 1. はじめに

投資判断に利用できる情報として, 財務データやマーケット データなどの数値データの他,ニュースや新聞記事,決算短信, 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、過去又は将来の業績に対する理由及 び根拠や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,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[五島 2015] [沖本 2014]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. 例えば、坂地ら「坂地 2011] 「坂地 2015] は、文の表現と機械学 習によって、過去の業績や製品の売れ行きなどに対する因果 関係文を経済新聞の記事や決算短信から抽出する手法を提案 している.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 得られるとした.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 2015b]や,業績の予 測を示す文を抽出した研究[北森 2017]がある. さらに, 抽出さ れた重要文に極性を付与する研究[酒井 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\*1,「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\*2. 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を抽出する判別モデル を提案する.具体的には、坂地ら[坂地 2011]の手法を応用し、

連絡先: 佐藤 史仁, 日興リサーチセンター株式会社, 〒103-0026 東京都中央区日本橋兜町 8-1

<sup>\*1</sup> 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.

<sup>\*2</sup> 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.企業の HP の他、金融庁の EDINET な どから入手可能. http://disclosure.edinet-fsa.go.jp/

有価証券報告書専用の因果関係文を抽出する機械学習を用いた判別モデルを作成した.データは TOPIX1000 を構成する 企業の 2008 年から 2016 年までの有価証券報告書(本決算)に 含まれるテキストデータを用いた.

## 2. 因果関係文の判別モデル

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 2011]で用いられた手法を応用した.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\*3)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、ソフトマージン SVM でのペナ ルティの重要度を表すコストパラメータ C は 1.0 とした. 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1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 2 に示す. ただし,太字は手がかり表現 を示している.

図表1 因果関係文の抽出処理の概要





因果関係を	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、新興国を甲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。
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。
	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。

図表2 因果関係を含む文と含まない文の例

#### 2.1 手がかり表現によるフィルタリング

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 2015]を参考に 37 個の手がかり表現を選定した.これらの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の例を図表 3 に示す.

図表	3	手がかり表現の	例
			~ -

から、	を背景に、	を受けております。
を反映し、	を反映して	に支えられて
によって	により	ためであります。
に伴う	に伴い、	を受け、

## 2.2 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:** 分析対象の有価証券報告書を,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.

<sup>\*3</sup> 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を用いた.

http://scikit-learn.org/stable/index.html

- **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 2011] を参考に 4 つの素性を利用した.各素性の概要を図表 4 に示す.ただし,図表 5<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表しており,一塊の文字列が文節を,矢印が係り先を示している.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,形態素解析では形態素解析器 MeCab\*<sup>4</sup>[工藤 2004]を,構文解析では係り受け解析器 CaboCha\*<sup>5</sup>[工藤 2002]を用いた.

図表4 素性の概要

素性の	概要			
名前				
助詞の	核文節に係る文節に含まれる助詞を前部助			
ペア	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5を参照.			
文に含	図表3を参照.			
まれる				
手がかり				
表現				
形態素	因果関係を含む文の候補を形態素解析器で分			
ユニ	解した形態素のうち, 頻度が 2 以上のものを抽			
グラム	出し重複を除いたもの.			
形態素	因果関係を含む文の候補を形態素解析器で分			
バイ	解し, 隣り合った全ての形態素ペア(重複を除			
グラム	<).			

<sup>\*4</sup> http://taku910.github.io/mecab/

図表5 助詞のペアの取得イメージ



## 3. 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 6 に示す.ただし、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している.

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値とも0.8を超える結果となった.また、図表7に抽出 された因果関係を含む文の例を示す.

図表6 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	

図表7 判別モデルの因果関係文抽出結果例

項目名	因果関係文として抽出された文
業績等の 概要	投資活動によるキャッシュ・フローは、船舶の 取得による支出などにより、当連結会計年度 は1,455億40百万円のマイナスとなりまし た。
事業等の リスク	一方、9月以降には取引先からの返品が発 生するため、第4四半期の収益が低下いた します。
対処すべき 課題	ロール事業では、事務機の構造変化や高耐 久化の加速による補修品市場の縮小といっ た環境変化により販売の低迷が想定されま す。

#### 4. 考察

判別モデルの評価は、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値とも0.8を超え、良好な判別 結果となったが、この要因としていくつかの理由が考えられる.1 つは、上手くモデルの学習が機能するような学習データが準備 できたことが考えられる.実際に、年、業種、項目ごとに万遍なく 文を取得していることや、手がかり表現でフィルタリングしたこと で正例候補が絞られ、結果的に正例と負例の割合がおおよそ 6対4となり、偏りなく広範の表現の学習データが準備できた. その他、本稿で対象とした有価証券報告書の因果関係を含む

<sup>\*5</sup> https://taku910.github.io/cabocha/

文の構文や単語は似たものが多かったという可能性や,本稿の 素性で上手く捉えられるようなシンプルな特徴を持つ文が多か った可能性も考えられる.ここで、シンプルな特徴を持つ文は、 因果関係を含むか否かの人による判別が容易であると考え, SVM の性能評価の追加確認を行った. 評価データとしては, 手順 2-3 において, 最初の判定員 3 人の判定が一致した文を 判別の容易な文としたデータ、反対に、3人の判定が一致しな かった文を判別の難しい文としたデータのそれぞれを用いた. その結果を図表 8 に示す. 判別の容易な文については, 判別 性能の評価が高く,一方で,判別の難しい文は判別性能の評 価が低い結果となった. 判別の難しい文は 5 人の判定員で判 別されており、単純にルール化できるような因果関係を含むか 否かの基準だけではなく、判定員それぞれの異なる経験と様々 な観点から判別されている.他方, SVM では,限られた学習デ ータと決められた素性で学習しているため,ルール化できるよう な基準は反映できても、個々人の経験や様々な観点までは反 映しきれなかったのかもしれない. このようなニュアンスと呼べる 要素を加味するには、まず、判別が難しい文の学習データを増 やすことや、判別の難しい文に対する因果関係が含まれるか否 かの判別基準となる特徴などを,素性として新たに加える等の 工夫が必要になると考えられる.

判別 難易	因果 関係 有無	精度	再現率	F値	データ数
E	有	0.93	0.93	0.93	590
勿	無	0.91	0.91	0.91	454
甜作	有	0.58	0.74	0.65	143
美田	無	0.75	0.60	0.67	190

図表8 判別モデルの追加の評価結果

### 5. 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 る判別モデルを提案した.作成したモデルの評価では、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 値とも0.8を超え、良好な判別結果となった.この判別モデル により有価証券報告書独自の投資判断に有益な情報の効率的 な抽出が期待できるだろう.

また本稿における SVM による判別モデルについて、判別が 難しい文の学習データを増やすことや、判別の難しい文の因果 関係が含まれるか否かの判別基準となる特徴などを素性として 新たに加える等の工夫をすることで、判別モデルのさらなる性 能の向上が見込めるだろう.

#### 参考文献

- [五島 2015] 五島圭一, 高橋大志, 寺野隆雄: 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, (2015)
- [北森 2017] 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, pp150-161, (2017)
- [工藤 2002] 工藤拓, 松本裕治: 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, (2002)

- [工藤 2004] 工藤拓,山本薫,松本裕治: 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, (2004)
- [沖本 2014] 沖本竜義, 平澤英司: 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, (2014)
- [酒井 2015a] 酒井浩之,小林義和,坂地泰紀: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 2015b]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)
- [坂地 2011] 坂地泰紀, 増山繁: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, (2011)
- [坂地 2015] 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, pp811-822, (2015)