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コロナ禍における有価証券報告書の記載の変化に関する分析

Short Review
2020 年 10 月

投資工学研究所
川﨑　正勝
佐藤　史仁

1. はじめに

2019 年 12 月に中国で発生が確認された新型コロナウイルス感染症（COVID-19）が、世界中に影響をもたらしている。発生確認以来、感染が拡大していった新型コロナウイルス感染症に対し各国はその対応に追われ、日本でも、政府や地方自治体による異例の対応が相次いだ。政府は、2020 年 2 月 26 日に大規模イベントの中止や縮小、同月 27 日には全国すべての小学校、中学校、高校、特別支援学校について臨時休校を要請した。さらに、4 月 7 日に 7 都府県を対象に緊急事態宣言が発出され、9 日後の 16 日には対象地域が全都道府県に拡大された。地方自治体でも、東京都が協力金を支給することを条件に特定の業種・施設に対して休業や時短営業の要請を行うなど様々な対策が行われた。一方、個人では新型コロナウイルス感染症の発生・拡大によって、テレワークや在宅勤務の拡大、不要不急の外出や大人数での飲食に対する自粛意識の高まりなどが生じ、生活や消費行動に影響が及んだ。これらは企業の経営環境に波及しており、例えば、観光業界では、県外への移動自粛や入国制限などによって旅行客は激減し経営環境が悪化した。一方で、ビデオ会議などのリモートオフィスソフトを手掛ける企業などの経営環境に対しては、この状況が追い風となっているだろう。このように、多くの企業の経営環境は新型コロナウイルス感染症の影響で大きく変わっている。

経営環境の変化に対して、経営者の認識する経営課題やその対策も変化しており、これらの認識の変化は、例えば、有価証券報告書の記載の変化として表れていることが考えられる。特に、「経営方針、経営環境及び対処すべき課題等」[1]の項目では、新型コロナウイルス感染症の発生・拡大以前の「コロナ前」とその後の「コロナ後」とで、経営者の経営環境に対する認識の変化に対応して、記載の内容も大きく異なっていることが想定される。逆に言えば、この記載の変化を調べることで、「コロナ前」と「コロナ後」とでの経営者の認識する課題と対策の変化が把握でき、企業の今後の展望やリスクを推察するための有力な情報を得られる可能性がある。

しかし、多くの企業の有価証券報告書を確認することは、人手と労力の負担が大きいという問題がある。また、上記の項目には、「経営課題やその対策の記述がされている重要な文」以外の文、例えば、「マクロ経済環境に関する事実のみを述べた文」や、「過去の課題への対処に対する単なる結果を述べた文」

[1] 2019 年 1 月 31 日付で公布・施行された「企業内容等の開示に関する内閣府令の一部を改正する内閣府令」で、「経営方針・経営戦略等について、市場の状況、競争優位性、主要製品・サービス、顧客基盤等に関する経営者の認識の説明を含めた記載」が求められている。

なども多い。このような問題に対しテキストマイニングの分野では、目的に合った重要な文を効率的に取得する手法が提案されている。例えば、坂地ら[坂地・増山 2011]は新聞記事から、質問応答システムなどに役立つ情報である原因や結果を含む文である因果関係文を、自動で抽出する技術を開発した。また、小寺ら[小寺他 2019]と田中ら[Tanaka 他 2019]は、将来に関する文で、かつ、目的や手段に関する内容を有する文を未来志向文と定義し自動で抽出する技術を開発[2]した。さらに、文書の特徴を把握することを目的に、重要語を抽出する手法も提案されている。酒井ら[酒井他 2015]は、決算短信から業績要因を含む文を抽出する試みにおいて、単語が出現する多さと単語が出現する文書の偏りを考慮した term frequency–inverse document frequency 法（以下 TF・IDF 法）、及び文書集合での単語出現の「ばらつき」度合いを考慮したエントロピーとを組み合わせて、企業の重要なキーワードを取得し、これと手がかり表現を用いて決算短信から業績要因を含む文を定量的に抽出した。

このように、必要な文や文書の特徴を表すキーワードを自動で抽出し集計することで、対象とする文書の全体的なテーマや記載内容の傾向を効率的に把握できることが期待される。そこで、「コロナ前」と「コロナ後」で経営者がどのような課題や対策を考えているのか、またその変化を把握するために、「コロナ前」として 2019 年 3 月期決算、「コロナ後」として 2020 年 3 月期決算の有価証券報告書における「経営方針、経営環境及び対処すべき課題等」の項目を対象に、テキストマイニングを用いて全体的な記載の変化の確認を行った。ここで、確認の対象とする文は未来志向文のみとした。経営者が認識している経営課題や対策を含む文は、過去ではなく未来の内容を含み、課題や対策が目的や手段として表現されている文であると仮定し、未来志向文のみを分析対象とすることが妥当だと考えた。

具体的にはまず、2020 年 3 月期（以下、当年度）及び 2019 年 3 月期（以下、前年度）の特徴語を抽出し、各年度のテーマや傾向の解釈を行った。特徴語は、その年度の課題や対策のテーマを表す単語として、未来志向文に含まれる各単語の各年度に対する特有度と重要度からなる特徴語スコアを計算し、スコアの高い単語とした。次に、「コロナ前」の課題や対策のテーマが、「コロナ後」でも継続しているのかを確認するため、「コロナ前」の特徴語の「コロナ前」と「コロナ後」における出現数の増減を確認した。

2. 未来志向文抽出モデルと対象テキストデータ

未来志向文の抽出に用いる機械学習モデルには小寺ら[小寺他 2019]のモデル[3]を利用し、入力データは図表 1 に示す対象テキストデータを用いた。対象テキストデータは、2020 年 8 月 19 日時点で東京証券取引所の市場第一部に上場している企業における、3 月期決算の有価証券報告書の「経営方針、経営環境及び対処すべき課題等」から抽出した文とした。なお、2020 年 3 月期決算の有価証券報告書については 2020 年 7 月末までに EDINET[4]に提出されたものを用いた。未来志向文の抽出について、

[2] 日興リサーチセンター株式会社、国立大学法人東京大学が 2019 年に特許を取得（特許番号：特許第 6615392 号）。
[3] モデルにおけるパラメータは、小寺ら[小寺他 2019]の機械学習モデルの学習用データセットを使った学習によって調整されたパラメータを利用。
[4] 金融商品取引法に基づく有価証券報告書等の開示資料に関する電子開示システム（https://disclosure.edinet-fsa.go.jp/）。

小寺ら[小寺他 2019]と田中ら[Tanaka 他 2019]は PDF ファイルから抽出していたが、本稿では XBRL[5]ファイルから抽出を行った。また、対象テキストデータの作成過程で、余分なスペースの削除やアルファベットの半角化などテキストマイニングにおける一般的なクリーニング処理を行った。なお、分析対象の有価証券報告書は 2020 年、2019 年 3 月期決算のものであるが、特徴語スコアは年度特有度合いを考慮しており、スコア算出対象年度のほか、それ以前の過去データも必要なため、2016 年～2018 年 3 月期決算の未来志向文も抽出した。

図表 1　対象テキストデータの基本情報

対象	内容
企業	2020 年 8 月 19 日時点で東証一部に上場している企業
文書	対象企業の 2016 年～2020 年の 3 月期決算の有価証券報告書
項目	「経営方針、経営環境及び対処すべき課題等」

（注）2020 年 3 月期決算の有価証券報告書は 2020 年 7 月末までに提出されたもの

（出所）日興リサーチセンター作成

3. 特徴語スコアの算出方法

特徴語スコアの算出方法は、酒井ら[酒井他 2015]の決算短信の文書からの企業の重要なキーワードを抽出する手法と同様に以下の（1）式を利用した。ただし、TF 及び IDF の計算では、キーワードの抽出対象として企業ごとの決算短信の集合を 1 つの文書と見なすのではなく、年度ごとの全ての対象テキストデータを 1 つの文書と見なした。エントロピーの計算では、企業の決算短信の集合でなく、年度の有価証券報告書の集合を対象とした。また、IDF について、単語がどれだけスコア算出対象年度特有かを考慮するため、前年度の特徴語スコアは 2016~2019 年、当年度では 2016~2020 年の 3 月期決算の有価証券報告書の集合を対象に計算を行った。

$$\text{スコア} = TF \times IDF \times \text{エントロピー} \qquad (1)$$

（1）式の、それぞれの構成要素の計算式及び意味については図表 2 で示した。なお、特徴語スコアの算出で用いる単語は、抽出した未来志向文に対し MeCab[6]を用いた形態素解析によって得られる形態

[5] XBRL（eXtensible Business Reporting Language）は、各種事業報告用の情報（財務・経営・投資などの様々な情報）を作成・流通・利用できるように標準化された XML ベースのコンピュータ言語。
（https://www.xbrl.or.jp/modules/pico1/index.php?content_id=9）

[6] 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/）

素[7]とした。また、分析対象とする単語の品詞は名詞とし、単独では意味をなさないものは分析対象から除いた。

図表２　構成要素の計算式及び意味

項目	計算式	意味
TF	年度tの全対象テキストデータに対する、 $\frac{\text{単語}w\text{の出現数}}{\text{全ての単語の出現数の総和}}$	年度tにおける単語wの 出現のしやすさ
IDF	$\log_2 \frac{\text{年度の数}}{\text{単語}w\text{が出現する年度の数}}$	年度単位でみた 単語wの珍しさ
エントロピー	$-\sum_{d_t} P(w,d_t)\log_2 P(w,d_t)$ ただし、$d_t \in \{\text{年度}t\text{の対象有価証券報告書の集合}\}$ としたとき、d_tの対象テキストデータに対して、 $P(w,d_t) = \frac{\text{単語}w\text{の出現数}}{\text{全ての単語の出現数の総和}}$	年度tにおける 単語wが出現する 有価証券報告書の多さ

（出所）日興リサーチセンター作成

[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]。

4. 前年度と当年度の特徴語によるテーマの解釈

前年度と当年度の特徴語スコアが上位 20 の単語（以下ではこれらの単語を各年度の特徴語とする）を図表 3 に示した。

図表 3　前年度と当年度の特徴語

順位	前年度	当年度
1	SDGs	新型コロナウイルス
2	持続可能な開発目標	感染拡大
3	令和	SDGs
4	DX	収束
5	RPA	自粛
6	デジタルトランスフォーメーション	DX
7	万博	禍
8	RoboticProcessAutomation	外出
9	マネー・ローンダリング	COVID
10	CASE	持続可能な開発目標
11	MaaS	消毒
12	スタートアップ	時差出勤
13	気候	着用
14	洋上風力発電	資金繰り
15	資金供与	新型コロナウィルス
16	tainableDevelopmentGoals	デジタルトランスフォーメーション
17	RE	手洗い
18	失敗	出勤
19	TCFD	CASE
20	Society	検温

（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

両年度の共通のテーマとして、「SDGs 関連（SDGs、持続可能な開発目標、tainableDevelopmentGoals）」と「DX 関連（DX、デジタルトランスフォーメーション）」「自動車業界のトレンド関連（CASE[8]）」が挙げられる。「SDGs 関連」と「DX 関連」の特徴語は、同じ意味であ

[8] Connected、Autonomous、Shared & Services、Electric、コネクテッド、自動運転、シェア＆サービス、電動化の頭文字。

り、英語、日本語、又は略語の表記違いとなっている。「tainableDevelopmentGoals」は「SustainableDevelopmentGoals」がクリーニングや形態素解析の際に単語が分割されたと考えられる。いずれも、近年の注目度の高い話題であり、それが課題として認識されていることが窺える結果となった。「SDGs 関連」については、前年度に、世界規模で抗議活動が行われている気候対策に関する「TCFD[9]」「気候」「洋上風力発電」も抽出された。「SDGs 関連」は国際的な社会課題であり、日本でも取り組みが広がっていることが示唆される。「DX 関連」については、前年度に、「RPA」「RoboticProcessAutomation」の DX に関わる具体的な単語が抽出された。また、「自動車業界のトレンド関連」については、「CASE」が両年度の、「MaaS[10]」が前年度の特徴語として抽出された。

当年度のみで出現している特徴語に注目すると、「新型コロナウイルス」「COVID」「新型コロナウィルス」、さらに「感染拡大」「収束」「自粛」「禍」と新型コロナウイルス感染症に関連する単語が多く抽出された。また、新型コロナウイルス感染症の拡大に伴い、連日ニュースでもよく耳にした企業や個人の対策や課題に関する単語である「時差出勤」「出勤」の勤務体制や「消毒」「着用」「手洗い」「検温」の衛生管理、「資金繰り」や「外出」が抽出された。当年度の特徴語は総じて新型コロナウイルス感染症関連であり、優先すべき経営課題のテーマが前年度から大きく変化したことが推察される。

一方、前年度のみで出現している特徴語を確認すると、「マネー・ローンダリング」「資金供与」が抽出された。これは、2019 年 10 月から 11 月にかけて、日本で「第 4 次 FATF 対日相互審査」の現地調査が審査団により行われたが、経営者が審査への対応を重要な経営課題ととらえていたことを示唆する。また、「万博」は、2025 年に大阪で開催されることが 2018 年 11 月に決まったことを受けた、前年度ならではの特徴だと言える。

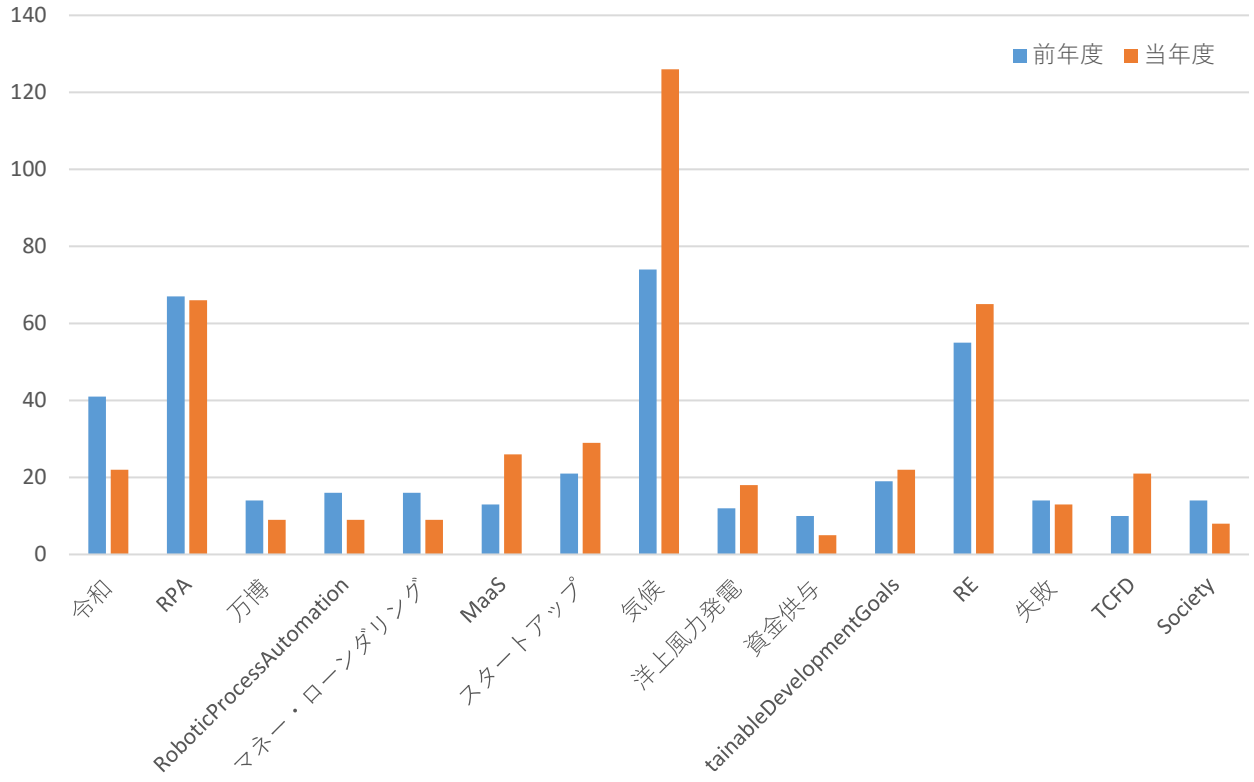
当年度は新型コロナウイルス感染症が課題・対策の大きなテーマとして新たに出現したものと考えられるが、前年度にテーマとなっていた「マネー・ローンダリング」「資金供与」「万博」などが、当年度の課題・対策として継続して認識されているのかを確認するため、対象の特徴語の出現数を比較した。図表 4 に前年度のみに抽出された特徴語について、前年度、当年度各対象単語の出現数を示す。なお、ここでの対象単語の出現数は、未来志向文中の単語を数えて算出した。

特徴語として前年度のみに抽出された単語の中で、「令和」「万博」「RPA」「RoboticProcessAutomation」「マネー・ローンダリング」「資金供与」「失敗」「Society」の出現数は減少となった。「RPA」と「RoboticProcessAutomation」は同義語であるため併せて考えれば出現数の減少率は小幅にとどまっていた。「失敗」についても減少率は小幅となっていた。これら以外の単語の内、「令和」「万博」「RPA」「RoboticProcessAutomation」「マネー・ローンダリング」「資金供与」の出現数が減少した単語については、経営環境に影響を与えたと考えられる前年度特有のイベントが関係していることが推察される。具体的には「令和」は元号改正、「万博」は大阪での開催決定、「マネー・ローンダリング」「資金供与」は「第 4 次 FATF 対日相互審査」の現地調査があった。このことは、イベント通過後、これらの課題や

[9] Task Force on Climate-related Financial Disclosures、気候関連財務情報開示タスクフォース。
[10] Mobility as a Service、モビリティーのサービス化。

対策を挙げる企業の減少やその優先度の低下を表しているのかもしれないが、0 にはなっておらず、当年度の課題や対策のテーマとしての特徴語にはならなかったものの、課題や対策は継続している企業があることを示す。一方、前年度の特徴語となった単語の出現数が当年度に増加した単語については、それら単語が表す課題や対策を挙げる企業の増加や記載が充実したことが考えられるものの、当年度の特徴語とはならなかった。新型コロナウイルス感染症関連の単語が当年度特有であったことや多くの企業で課題や対策として記載があったため、相対的に当該単語の特徴語スコアの順位が低下し、特徴語として抽出されなかったことが考えられよう。これらの結果は、年度の特徴語となった単語の出現数が減少、増加のいずれの場合であっても、特徴語の表す課題や対策は次年度も継続する傾向にあることを示唆している。

図表４　前年度のみに出現した特徴語の前年度と当年度の単語の出現数



（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

参考として、今回抽出された特徴語が記載されている未来志向文の例を図表 5 に示す。

図表 5 特徴語が記載されている未来志向文の例

文例	未来志向文
SDGs 関連	また、SDGs に関する社会的な関心の高まりや、地球温暖化や気候変動によって発生する自然災害等が地域経済及び当行グループにとっての大きなリスクとなっていることを踏まえ、本業を通じた SDGs への取組強化を進めてまいります。
SDGs 関連	近年、お客様から SDGs および気候変動問題の取り組みに対する情報提供の要請が増加していることもあり、今後、更に SDGs および気候変動問題への対応を積極的に進めていきます。
DX 関連	取引量の増加や業務の複雑化が進む中、業務プロセスの見直しや RPA(ロボティック・プロセス・オートメーション)等のデジタル技術の活用を積極的に推進することにより、効率的な業務運営体制を構築し、競争力の基盤強化を図る。
自動車業界のトレンド関連	・自動車の素材については、燃料電池車(FCV)、電気自動車(EV)、ハイブリッドカー等の更なる開発や CASE の浸透に向けた各種素材の取扱いを拡大していきます。
新型コロナ関連	新型コロナウイルス感染症が世界的な流行となる中、当社グループでは全事業所の従業員を対象にテレワーク・時差出勤・直行直帰等を実行し、従業員の安全・健康の確保と感染の防止に努めております。
新型コロナ関連	また、新型コロナウイルス感染拡大防止及び従業員の安全を考慮し、始業前及び実務開始前の検温、出退勤時のマスク着用、手洗い等を義務づけております。
新型コロナ関連	資金繰りについてもコミットメントライン借入枠 800 億を維持しつつ、特別融資などの有利な条件の借入を行いながら、キャッシュ・フロー重視の経営をバランスよく行っている状況であります。
マネー・ローンダリング	マネー・ローンダリング対策等の金融犯罪未然防止を含むコンプライアンス・リスクへ厳格に対応するなど、グループ会社一体となったリスクガバナンスの高度化を進めてまいります。
万博	また、万博開催が決定し統合型リゾート(IR)の誘致が期待される大阪エリアでは、新しい事業の創出を目指します。

（注）有価証券報告書の「経営方針、経営環境及び対処すべき課題等」から特徴語を含む未来志向文を抜粋。特徴語を示すため、特徴語については当社が赤字にし下線を追加した。

（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

5. おわりに

本稿では、「コロナ前」と「コロナ後」で経営者がどのような課題や対策を考え、それがどう変化したかを把握するため、2019 年 3 月期決算と 2020 年 3 月期決算の有価証券報告書からその年度特有の課題や対策を表すと考えられる特徴語を抽出し分析した。「コロナ後」の特徴語として、新型コロナウイルス感染症関連の単語が多数抽出され、経営者の認識する課題や対策に対して、新型コロナウイルス感染症による影響が大きかったことが確認された。また、「コロナ前」の課題・対策は、「コロナ後」でも継続していることが分かった。

特徴語スコアの算出には、TF 及び IDF、エントロピーを用いた。TF とエントロピーでは、年度における単語や出現している文書数が多いほどスコアが高く、業種などの一部の企業のみで出現する単語は出現回数が相対的に少ないためスコアが低くなりやすい。一方、年度による特有度合いの高い単語を特徴語とするため組み入れた IDF によって、単年でしか出現しないような単語はスコアが高く、逆に毎年出現している単語の IDF の値は 0 となり、特徴語スコアも 0 となる。これにより、「当社」や「事業」などの有価証券報告書で一般的に用いられるような課題や対策に関係のない頻出用語は特徴語として候補から外れることとなる。しかし、株主還元など、全ての株式会社で常に取り組むべき課題もある。このような一般的な課題・対策に関する単語は出現回数が高いものの、IDF の値が 0 となり特徴語とはならない。このような課題を表す単語を特徴語として抽出する場合は、別な手法が必要になるだろう。また、本稿で抽出した特徴語の出現回数を確認すると、前年度で出現回数が 100 を超えたのは「SDGs」のみであった。このことは、本稿の年度特有の課題としての特徴語は、業種としての課題や、一部の企業で取り組みが始まり徐々に広がっていく課題を表していることを示唆する。しかしながら、当年度の特徴語であった新型コロナウイルス感染症関連の単語は、ほぼ全企業で記載されており、新型コロナウイルス感染症の影響の大きさを示唆するとともに、例年にないケースだったと言える。このようにテキストマイニングによって経営者の認識する課題を捉えることは、各企業の課題と対策に対する大きなテーマや、業種のテーマ、これから拡大していくテーマの把握につながり、投資判断の一助となるだろう。

＊ 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。

以上

参考文献

Tanaka, Y.・Kodera, S.・Sato, F.・Sakaji, H.・Izumi, K. [2019] , 「The Extraction of the Future-Oriented Sentences from Annual Reports」, 2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI), pp.679-684, 2019.

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

小寺俊哉・佐藤史仁・佐久間洋明・田中良典 [2019], 「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」, 日興リサーチレビュー, 2019 年, 3 月号.

酒井浩之・西沢裕子・松並祥吾・坂地泰紀[2015], 「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・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.