Vol.2019-DBS-169 No.19
Vol.2019-IFAT-136 No.19
2019/9/10

有価証券報告書の分析に基づく重要な新着ニュースの発見

米田 宏生[1,a)]　湯本 高行[1,b)]　礒川 悌次郎[1]　上浦 尚武[1]

概要：近年，大企業の買収や公的機関からの経済指標の発表などニュースで報道される出来事が経済に大きな影響を与えることがある．そして，これらのニュース情報はインターネット上でリアルタイムに取得できる．インターネット上の新着ニュース記事から重要な出来事を抽出できれば，即時的に企業の経営戦略に対する判断や投資家の投資判断などに活用できる．しかし，常に配信されているニュース記事すべてを人手で確認することは労力がかかる上に，リアルタイム性が失われる．また，ニュース記事のすべてが重要な情報というわけではなく，経済的な分析をする上で不要な情報も含まれている．そこで，本研究では有価証券報告書を分析することで，経済に関連があると考えられる事象を名詞のみで表現する事象パターンを作成する．そして，事象パターンを用いて，ニュース記事から経済的影響のある事象を自動で抽出することを目的とする．さらに，抽出事象を時系列データと紐付けることで応用できる手法であることを確認する．

1. はじめに

近年，大企業の買収や公的機関からの経済指標の発表などニュースで報道される出来事が経済に大きな影響を与えることがある．そして，インターネットが浸透した現代社会では，ニュースサイトなどでリアルタイム性の高い情報を得ることができる．このようなインターネット上の新着ニュース記事から重要な出来事を抽出できれば，即時的に企業の経営戦略に対する判断や投資家の投資判断などに活用できると考えられる．しかし，ニュース記事は膨大な数が配信されており，すべてのニュース記事を人手で確認するのは非常に労力がかかり，その上時間がかかることからリアルタイム性が失われる．また，ニュース記事のすべてが重要な情報というわけではなく，経済的な分析をする上で不要な情報も含まれている．

このような文中に含まれる要因と結果という因果関係を自動抽出する際に，手がかり表現を用いた手法がとられることがある．しかし，ニュース記事で手がかり表現が少なく要因が明示的に書かれていないためにこの手法で抽出することは困難である．

そこで，本研究では因果関係が手がかり表現とともに書かれている有価証券報告書を分析することで，要因が明示的に書かれていないニュース記事から経済的影響のある事象を自動で抽出する手法の開発を行う．有価証券報告書の経済的影響のある事象を名詞の組み合わせで示す，事象のパターンを作成することで，パターンマッチングを用いてニュース記事からの事象抽出を考える．また，手がかり表現による抽出は手がかり表現さえあれば抽出対象になるので，要因の内容に着目せず，重要さは考えていない．一方で，事象のパターンは選出された名詞によって表現されるので，抽出される文の内容が重要な事象に関することになっていると考えられる．

[1] 兵庫県立大学
[a)] ei19y030@steng.u-hyogo.ac.jp
[b)] yumoto@eng.u-hyogo.ac.jp

2. 関連研究

文中から因果関係を自動で抽出する研究として，手がかり表現と手がかり表現が文中で現れる場所を考慮して因果関係を抽出する坂地らの研究 [1] がある．また，手がかり表現「ため」に焦点を当てた乾らの研究 [2] がある．これは因果関係が行為か事態か判断し，cause, effect, precond, means の 4 つの因果関係を定義している．応用として手がかり表現を用いて評価表現を取得する高野らの研究 [3] がある．文中の手がかり表現と文の構造を考慮することで，評価表現をブートストラップ的に抽出することに成功している．しかし，どれも手がかり表現を用いた抽出手法であり，手がかり表現が少ないニュース記事には適用しにくい．そこで，本研究では要因を表す事象をパターンにすることで，手がかり表現が使えないテキストに対する抽出を行う．

3. 予備調査

自然言語処理において，手がかり表現を用いることで因果関係を抽出できることがある．これは手がかり表現が文

中にあれば，その前後に要因と結果が存在する，といったものである．有価証券報告書では手がかり表現を用いた因果関係の抽出が行われている [4]．予備調査として，有価証券報告書内で出現頻度の高い上位 5 個の手がかり表現「により」,「があり」,「もあり」,「ことから」,「こと等から」,「ことなどから」,「を背景に」,「影響」を用いて毎日新聞の経済面のニュース記事 40 件の 401 文と chABSA-dataset[*1]に収録されている有価証券報告書 200 件の 5077 文に対して因果関係の抽出を行った．全文に対する抽出文の割合を表 1 に示す．chABSA-dataset は感情解析を行うデータセットで，収録されている有価証券報告書の一部に専門家が付加した極性情報がある．極性は有価証券報告書の文が業績面からポジティブかネガティブかに分類した情報であり，この 2 つのどちらかに分類された文には経済的影響のある事象が記載されていると考えられる．ただし，全ての文に対して極性が付けられるわけではない．また，経済的影響のある事象というのは「円高の進行」や「政府による新政策の発表」といった日本経済における生産，流通，消費に影響を及ぼすと考えられる出来事のことである．ニュース記事の極性については著者のうちの 1 名が独自に判断して情報を付加した．表 1 より有価証券報告書は全体の約 56%が極性を持っており，極性が付加されているもののうち 58%を手がかり表現で抽出できる．一方でニュース記事では全体の約 54%が極性を持っており，手がかり表現で抽出できるのは，極性が付加されているもののうち約 7.8%である．全体での極性を持っている文の割合がほぼ同じなのにも関わらず，手がかり表現で抽出できる割合は大きく違う．このことから，ニュース記事では手がかり表現の出現頻度が低く，この方法では要因の抽出が難しいといえる．

表 1 予備調査

	極性情報あり	極性情報あり& 手がかり表現あり
有価証券報告書	56%	32%
ニュース記事	54%	4%

4. 事象パターンを用いた重要ニュースの発見

本節ではニュース記事の経済的影響のある事象を抽出するために，事象パターンを作成し，実際にニュース記事を対象に抽出を行う手法を提案する．本手法で事象パターンとは経済的影響をある事象に含まれる名詞を用いて表現される二つ以上の名詞のセットである．本手法では有価証券報告書を基に，経済的影響のある事象を表現する事象パターンを作成する．たとえば，「日本経済は，円高の進行により景気の減速懸念が高まった」という文において，手がかり表現やルールを用いることで経済的影響のある事象として「円高の進行」という出来事を有価証券報告書から抽出する (4.1 節)．次に抽出した経済的影響のある事象をより抽象化して別の事象の抽出に利用するために，名詞の組合せで事象を表現することを考える．まず，事象が記載されている文と記載されていない文を比較することで，事象が記載されている文に特徴的に現れる名詞の選出を行う (4.2 節)．ここで「円高」や「進行」が経済的影響のある事象に特徴的に現れる名詞であれば，アソシエーション分析を行い，＜円高，進行＞という名詞の集合，つまり事象パターンを作成する (4.3 節)．本稿では事象パターンを＜＞で囲み，上記のように表現する．最後に事象パターンを用いてニュース記事の経済的影響のある事象を含む文の抽出を行う (4.4 節)．

4.1 有価証券報告書内の経済的影響のある事象の抽出

3 章で示したようにニュース記事には手がかり表現の出現頻度が低く，一方で有価証券報告書では出現頻度が高くなっている．そこで，手がかり表現で有価証券報告書から要因を抽出することにより，それらを利用して事象パターンを作成することを考える．利用した手がかり表現一覧を表 2 に示す．また，抽出範囲としては，「手がかり表現よりも前に存在し，最も近い距離にある読点まで」という独自に定義したルールに従う．もし手がかり表現よりも前に読点が存在しない場合は，文頭までを抽出範囲とする．「ＬＰガス業界においては，世帯人員の減少や省エネ機器の普及により，家庭用の需要は減少傾向が続いた」では手がかり表現「により」を基準に，前にある読点までを抽出すると，「世帯人員の減少や省エネ機器の普及」が抽出できる．「霜降りひらたけの収穫・出荷を開始したことよりきのこの生産量は増加しました」では読点がないため，手がかり表現「により」から文頭までの「霜降りひらたけの収穫・出荷を開始したこと」を抽出する．

表 2 手がかり表現一覧

により	ことから	こと等から	ことなどから
もあり	があり	を背景	に支えられて
を受け，	を受けて，	を受けて	影響
に伴い	に伴う	に伴い，	を反映して
を反映し，	によって	が響き，	ため，

4.2 経済的影響のある事象に現れる特徴的な名詞の選出

4.1 節で抽出した経済的影響のある事象の含まれる文を形態素解析することで名詞のみを取得する．経済的影響のある事象に現れる特徴的な名詞の選出することで，簡潔で重要な名詞のみで構成される事象パターンの作成を考える．特徴的な名詞の選出するために，カイ二乗検定 [5] とリスク比 [6] を利用する．カイ二乗検定とは 2 つの変数に関連が言えるのか否かを判断するために用いられる検定である．また，リスク比とは 2 つのデータ内である事象が起こる確率をもとにして算出される，起こりやすさを数値化したものである．表 3 を用いると次の式で表される．

[*1] https://github.com/chakki-works/chABSA-dataset

$$\text{リスク比} = \frac{\frac{A}{A+B}}{\frac{C}{C+D}} \quad (1)$$

表 3 クロス集計表

	対象名詞を	
	含む文数	含まない文数
抽出できた文の集合	A	B
抽出できなかった文の集合	C	D

カイ二乗検定で特徴的な名詞を選出した後に，リスク比を用いて更に名詞の絞り込みを行う．カイ二乗検定もリスク比も 4.1 節で抽出した経済的影響のある事象を含む文の集合とそれ以外の抽出されなかった文の集合を比較データとする．まず，ストップワードを用いて必要のない名詞の削除を行う．本研究では slothlib[*2]のストップワードを用いた．次にカイ二乗検定においては有意水準 5%で特徴的な名詞を選出した．ここで有意水準を 1%にしなかったのは，経済的影響が考えられる事象に関係してくる名詞が多く排除される可能性があるためである．しかし，有意水準 5%では過剰に名詞を取得しており，また，1%では過剰に排除されていることから，カイ二乗検定により選出された名詞を対象に，リスク比を用いてさらに特徴的な名詞を絞り込む．複数のリスク比をしきい値として実験した結果，リスク比 2 以上の名詞を経済的影響のある事象に現れる特徴的な名詞として選出する．次に複合名詞の作成を行う．複合名詞というのは 2 つ以上の名詞が結合することで生成される 1 つの名詞のことである．特徴的な名詞を選出した段階では「熊本地震」というような複合名詞は，JUMAN[7] の解析により「熊本」と「地震」に分けて出力されている．JUMAN とは形態素解析という，文を意味を持つ最小の単語である形態素に分けることができるシステムである．このように複合名詞が分かれて出力されていたら，事象パターン作成に悪影響を及ぼす可能性がある．そこで文中で連接する名詞は複合名詞として扱う．

4.3 事象パターン作成

有価証券報告書に記載されている経済的影響のある事象に現れる特徴的な名詞に対して，事象パターンを次の 3 ステップで作成する．

4.3.1 固有名詞の抽象化

文中では出現頻度が高い固有名詞と低い固有名詞が考えられる．例えば，「米国の政策」というフレーズからは「米国」と「政策」からなる事象パターンが，「バチカン市国の政策」というフレーズからは「バチカン市国」と「政策」からなる事象パターンが作成されることが考えられる．しかし，アソシエーション分析では支持度にしきい値を設けて，事象パターンを作成するため，「バチカン市国」といった出現頻度の低い固有名詞を含む事象パターンの作成は不可能であり，ニュース記事における抽出もできない．ここで全トランザクション T と相関ルール X → Y がある時，支持度 [8] と確信度 [8] は以下の式で表される．

$$\text{支持度} = |X \cap Y|/|T| \quad (2)$$

$$\text{確信度} = |X \cap Y|/|X| \quad (3)$$

ニュース記事で出現頻度の低い固有名詞に関する経済的影響のある事象も抽出できるようにするために，固有名詞の抽象化を行う．JUMAN で解析して得られた情報のうち，「地名」，「組織名」，「人名」の情報が付加される固有名詞をそれぞれ，[地名]，[組織名]，[人名] というタグに置き換える．こうすることで，＜ [地名]，政策＞といった事象パターンが作成され，出現頻度の低い固有名詞に関する事象の抽出も可能になる．また，すべての固有名詞を抽象化するのではなく，抽象化せずに扱うべき固有名詞もある．そのような固有名詞の判断に確信度を用いる．アソシエーション分析で相関ルールを取得した際に，確信度が高いということは，対象の相関ルールのアイテムが同時に出現する可能性が高いと考えられる．よって，相関ルールの確信度がしきい値以上であれば固有名詞が存在しても抽象化を行わない．本手法では確信度のしきい値を 70%と定めた．

4.3.2 相関ルールからの事象パターン作成

アソシエーション分析から得た相関ルールのアイテムのセットを事象パターンとして取得する．まず，4.3.1 に示す抽象化を行わず，支持度が 0.01%，確信度が 50%以上で相関ルールを作成する．作成された相関ルールのうち確信度が 70%以上であれば，抽象化を行わず，アイテムのセットを事象パターンとして取得する．次に確信度が 70%以下の相関ルールの固有名詞を抽象化し，再びアソシエーション分析を行って，事象パターンを作成する．

4.3.3 類似事象パターンの統一

43.2 で得られる事象パターンには＜米国, 政権, 政策＞と＜米国, 政権, 動向＞というような類似したものが見られる．このような似ている事象パターンを統一するために，Dice 係数を用いる．2 つの事象パターン間の Dice 係数が 0.5 以上であれば，類似している事象パターンとする．Dice 係数の式 [9] を次に示す．

$$\text{Dice} = 2|A \cap B|/(|A| + |B|) \quad (4)$$

ここで 2 つの事象パターンが類似しているとなった場合の共通している名詞を共通名詞，それぞれの事象パターンにしか存在しない名詞を付属名詞として取得する．上の例の場合は，共通名詞が「米国」と「政権」，付属名詞が「政策」と「動向」となっている．統一後は，共通名詞に加え，付属名詞が 1 つ以上含まれる名詞のセットを事象パターンとして扱う．この場合，＜米国, 政権, 政策＞，＜米国, 政権, 動向＞，＜米国, 政権, 政策, 動向＞の 3 つの事象パターンが考えられる．

[*2] http://svn.osdn.net/svnroot/slothlib/

4.4 ニュース記事内の経済的影響のある事象の発見

ここでは，ニュース記事から事象パターンを用いて経済的影響のある事象を抽出する手法を示す．重要なニュースというのはニュース記事本文に原因などの重要な事象が記載されていると考えられる．そこで，ニュース記事本文を抽出対象とし，本文中に事象パターンが当てはまる文があれば，そのニュース自体が経済的影響のある出来事を報道していると考えることができる．以上より重要なニュースを発見するということは，ニュース記事本文に経済的影響のある事象が記載されている文を発見することと同義と捉えることができる．

4.4.1 名詞間の係り受け関係を考慮した事象の抽出

事象パターンを基に，ニュース記事の経済的影響のある事象の抽出を行う際に，事象パターンを構成する名詞間に係り受けの関係をチェックする．「米国の政策が日本の経済に影響を及ぼした」という文に対して＜米国，政策＞という事象パターンを用いれば「米国の政策」という部分に係り受けの関係があるため抽出対象にする．一方で「米国は『日本の政策は修正すべきだ』と声明を出した」という文では「米国は」という文節が「声明を」という文節に係っているため，事象パターンの「米国」と「政策」に文脈的関係がないと判断し，抽出されない．このように事象パターンを構成する名詞間の関係を考慮することで，文脈を考慮した抽出が可能になっている．

4.4.2 複合名詞のみで表現される事象の抽出

事象パターンは 2 つ以上の名詞のセットとして作成している．＜競争，激化＞という事象パターンを用いた場合，「競争が激化して」というようなフレーズは抽出できるが，「競争激化による」という複合名詞のみで表現される事象の抽出はできない．そこで，事象パターンを構成する名詞を組み合わせて作成される複合名詞も抽出対象にする．＜競争，激化＞であれば，2 つの名詞を組み合わせることで，「競争激化」と「激化競争」という複合名詞を作成でき，これら複合名詞を含む文を抽出対象とする．よって，「競争激化による」というフレーズを含む文も抽出可能となる．

5. 評価実験

本章では作成した事象パターンを用いてニュース記事における経済的影響のある事象の抽出を行った結果の評価と抽出事象を実際に利用されている時系列データと関連付けることで事象の有用性の評価を行う．

5.1 事象パターンを用いた抽出手法の評価

5.1.1 実験方法

2014 年の 1 月から 3 月の間に毎日新聞の経済面に掲載された記事 2356 件の 21049 文を対象に，作成した事象パターンを用いて経済的影響のある事象が書いてある文を抽出する．また，事象パターンで抽出できた文数と同じ数の文をランダムでニュース記事から抽出する．そして抽出した文の評価を行うためにクラウドソーシングを用いる．抽出した文に対して，クラウドソーシングで経済的影響がある事象が記載されているかを表 3 の回答のように 4 段階で評価してもらう．その結果に対して表 3 の点数に示すような点数付けを行う．この時，各文 10 人分の回答を取得し，その平均点数を文の評価とする．評価基準として，平均点数が 3.0 以上の抽出文を経済的影響のある事象が記載されている文とし，3.0 未満の抽出文を記載されていない文とする．そして，点数付けした結果をもとに，適合率と再現率を算出する．表 4 の混同行列を用いて，適合率と再現率の定義を以下に示す．

$$適合率 = TP/(TP + FP) \tag{5}$$

$$再現率 = TP/(TP + FN) \tag{6}$$

表 4 回答の数値化

回答	点数
記載されている	4
おそらく記載されている	3
おそらく記載されていない	2
記載されていない	1

表 5 混同行列

		正解データ	
		正	負
予測データ	正	TP	FP
	負	FN	TN

5.1.2 実験結果

ニュース記事 21049 文に対して事象パターンを用いた結果，182 文の文が抽出できた．また事象パターンで抽出できた文以外からもランダムに 182 文を抽出した．これらに対して点数付けを行った結果を表 5 に示す．このとき，事象パターンを用いて抽出できなかった文を経済的影響のある事象が記載されていない文として扱っている．事象パターンを用いた抽出文に対する適合率は 90.7%，また再現率は 65.7%となった．つまり，事象パターンを用いて抽出できた文には高い可能性で経済的影響のある事象が記載されているが，再現率の低さより，経済的影響の事象を見逃していることがわかる．

表 6 抽出結果と評価

		クラウドソーシングによる評価	
		影響あり	影響なし
事象パターンによる抽出結果	影響あり	165	17
	影響なし	86	96

5.2 抽出事象の有用性に関する実験

事象パターンで抽出した事象は経済に影響を与える大きな出来事であると考えることができる．つまり，そのような出来事を表現する文章を構成する名詞の特徴を分析することで，抽出できている内容の有用性を評価することができ，重要なニュースの発見に繋がると考える．そこで，事象パターンを用いて抽出した事象と経済指標の変化を紐付けることで抽出事象の有用性を評価する．

5.2.1 実験方法

本実験では時系列データを 2014 年の全ての抽出事象に紐付けることで，変動が大きい日の事象を求める．次に変動が大きい日の抽出事象に特徴的な名詞を特定し，それらを分類する．

まず，2014 年の毎日新聞の経済面，社会面と国際面を対象に事象パターンを用いて抽出を行った結果と 2014 年の TOPIX-17 データの紐付けを 1 日単位で行う．今回，TOPIX-17 データの「商社・卸売」を対象に分析を行う．ここで，ニュース記事は新聞を対象としているため，事象は発生した日の次の日の新聞に掲載されることが考えられる．一方で TOPIX-17 データは事象が発生した時にすぐに変動すると考えられる．そこで紐付ける際は，抽出事象を報道日の前日の TOPIX-17 データと紐付ける．また，事象パターンを用いて抽出できるのは事象を含む文であるが，抽出できた文を含むニュースタイトルも重要な情報を含んでいると考えられるため，TOPIX-17 データと紐付ける．

次に紐付けたデータに対して，TOPIX-17 データの 5 日間の移動平均をとり，移動平均と実際の TOPIX-17 データとの差の絶対値が 2000 以上の日を変動の大きい日として，紐付けたデータを取得する．そして名詞を分類するために，変動が大きい日の事象を含む文とタイトルの名詞を抽出していく．同時に変動が大きい日以外のデータに対しても名詞を抽出し，カイ二乗検定を行うことで変動が大きい日に特徴的に現れる名詞の特定する．

次に変動が大きい日に特徴的に現れていた名詞を分類する．分類するために，名詞にタグを付けることを考える．タグの種類として固有名詞を意味する「組織名」,「地名」,「人名」を考える．また，経済的影響のある事象を分析しているため,「経済」タグも必要になると考える．他には，経済に影響を与える出来事として株価や業績の変動が考えられるので「変化」タグ，物の売れ行きや生産状況などが経済に関連してくると考えられるため，変化の対象になる「物品」タグが必要になる．さらに，JUMAN の解析結果のドメイン情報は名詞がよく使われている分野の分類であるため，名詞の特徴を捉えるのに適切であると考え，ドメイン名をタグをして利用することを考える．「組織名」,「地名」,「人名」タグは JUMAN の解析結果で固有名詞と判断された名詞に付与する．例えば「三菱重工」であれば「組織名」というタグが付く．「経済」タグは Wikipedia の記事に付いているカテゴリを参考に付与していく．例えば「金利」という特徴的な名詞にタグ付けを行う時,「金利」という文字列を記事タイトルに含む Wikipedia 記事「政策金利」を見つける．次に,「政策金利」のカテゴリ「経済指標」よりカテゴリの内容が経済に関係する内容であるかの有無を人手で確認する．経済に関係する内容であると判断したら,「金利」に「経済」のタグを付ける．「変化」タグや「物品」タグは人手によって判断して付与する．「変化」タグは「衰退」や「減益」,「物品」タグは「小麦」や「新薬」が対象になる．JUMAN のドメイン名を基したタグ付けはドメイン情報を持つ名詞を対象に行う．ここで，1 つの名詞が複数のドメイン情報を持つ場合は，複数のタグを付ける．ここまででタグが付与されなかった名詞は「抽象」というタグを付与する．以上のように名詞の分類を行い，その結果より分析を行う．

5.2.2 実験結果

移動平均をしきい値として TOPIX-17 データの変動が大きい日を調査すると，2014 年は 38 日あった．事象パターンを用いた抽出事象が報道された日付けと変動が大きい日が一致したのは 38 日中 30 日あった．また，30 日間で抽出できた事象数は 171 件であった．一方で，それ以外の日，つまり変動が小さい日の合計事象抽出数は 774 件であった．この 2 つの事象の集合からカイ二乗検定 [5] を用いて抽出した変動が大きい日の特徴的な名詞にタグ付けを行った結果を表 6 に示す．この時の有意水準は 5%とした．1 日当たりの抽出数を比較すると，変動が大きい日では 5.7 件で，変動が小さい日は 2.4 件で 2 倍以上の差がついていることから変動が大きい日には重要なニュースが集中していると考えられる．「政治」,「ビジネス」や「経済」といった経済に影響を与えると考えられる分野の名詞が多く出てきている．

表 7 名詞についたタグとその個数

タグの種類	個数	タグの種類	個数
政治	31	科学・技術	8
ビジネス	28	スポーツ	7
経済	22	健康・医学	6
変化 *	22	文化・芸術	5
人名	21	家庭・暮らし	5
組織名	18	メディア	4
地名	16	交通	2
物品 *	16	レクリエーション	1
教育・学習	16		

*人手によるタグ

5.3 考察

5.1 節の結果から，適合率が高いことより，抽出文における経済的影響が記載されているという信頼性は高いといえる．しかし，17 文は経済的影響がないと評価された文であった．経済的影響のないと評価された文の例として，事象パターン＜競争, 激化＞による「今後の世界競争の激化

は必至で油断できないのが現実だ」や＜金融市場＞による「デリバティブ」は『派生した』という意味の英語で，金融市場では株式や債券など元になる資産に関連して価格が決まる『金融派生商品』を指す」があった．これらはニュース記事の筆者の意見や言葉の説明の内容になっている文である．このような内容の文を抽出対象外にするには文の種類を分類 [11] できれば改善すると考えられる．

また，再現率が低いことより，経済的影響が記載されいるにも関わらず抽出できていない文があることがわかる．抽出できていなかった文として，「FRB は昨年 12 月，リーマン・ショック後の金融危機対策として導入した量的緩和策の縮小開始を決定」や「株高で株式や投資信託の取引が活発化し，手数料収入が拡大，各社とも大幅利益を達成した」などがあった．作成した事象パターンの中に金融に関するパターンや株価に関するパターンは作成されていたが，事象パターンの名詞と完全一致しなければ抽出対象とならないため，抽出ができなかった．これを改善するには事象パターンを増やすことや抽象化の対象名詞の拡大が考えられる．また，同じような意味を持つ名詞も抽出対象とすることができれば，より多くの抽出ができると考えられる．

5.2 節の結果より，「政治」，「ビジネス」，「経済」といった明らかに経済に影響を与える分野の名詞が変動が大きな日の特徴として現れている．これらのタグがついた名詞を含む抽出文の例としては，「経済」タグのついた「金融」を含む，「イスラム教徒が多いインドネシアや中東などの新興国の経済成長で，『イスラム金融』は急拡大している」などがあり，重要な事象だといえる．また，「変化」のタグがついた名詞では「減益」を含む，「ただ，中国勢の台頭で競争は激化し，サムスンの 13 年 10〜12 月期連結決算は営業減益だった」などがあり，企業の業績の変化を示していることから重要な事象であるといえる．「物品」タグがついた名詞の抽出の例は，「キャメル」や「ケント」といったたばこの銘柄を含む，「米たばこ大手 :『キャメル』が『ケント』買収か」などあり，企業の買収といった重要な出来事になっている．今回の名詞の分類を行っていく中で「教育・学習」のタグが思っていたよりも多かった．これらの内容としては，若者の人口減少に対する抽出が多く，少子高齢化という重要な社会問題に繋がる抽出結果となっていた．また，今回は Wikipedia のカテゴリ情報を参考にして「経済」のタグをつけたが，もっと Wikipedia のカテゴリ情報を利用すれば，より細かくタグ付けを行うことができ，詳細な分析が行えると考えられる．

6. おわりに

本研究では，有価証券報告書を分析し，事象パターンを作成・利用することで，ニュース記事から経済的影響のある事象を抽出する手法を提案した．事象パターンを作成するために，有価証券報告書の経済的影響の事象に現れる特徴的な名詞を対象に絞ったアソシエーション分析を行った．また，ニュース記事内の経済的影響のある事象を抽出するために，文脈を考慮した事象の抽出や複合名詞のみで表現される事象の抽出を行った．

評価実験はクラウドソーシングで事象パターンで抽出した文や抽出できなかった文に経済的影響が記載されているかを評価してもらうことで行った．抽出した文の適合率は 90.7%と高い数値になったことより，抽出文に経済的影響のある事象が記載されているという信頼性は高いと言える．一方で，再現率は 65.7%と低く，事象パターンを用いずにランダムに抽出した文の中にも経済的影響が記載されていると評価された文が含まれていた．また，時系列データと抽出結果を関連付けることで抽出事象の有用性を評価した．その結果，時系列データに大きな変動がある日の抽出事象には経済に関連する名詞が含まれていることが確認でき，事象パターンを用いた抽出事象は現実のデータと組合せて応用が可能だとわかった．今後の課題としては経済的影響のある事象を取りこぼしなく抽出することが重要となってくる．そのために事象パターンの網羅性を向上させる必要がある．

謝辞 本研究は JSPS 科研費 JP19H04116 の助成を受けたものです．

参考文献

[1] 坂地 他, “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 回年次大会論文集, pp.1144-1147, 2008.
[2] 乾 他, “接続標識「ため」に基づく文書集合からの因果関係知識の自動獲得”, 情報処理学会論文誌 Vol.45 No.3, pp.919-933, 2004.
[3] 高野 他, “因果関係に着目した口コミ Web サイトからの評価表現抽出”, 人工知能学会論文誌 24 巻 3 号 C, pp.322-332, 2009.
[4] 佐藤 他, “有価証券報告書からの因果関係文の抽出”, 2018 年度人工知能学会全国大会（第 32 回）,2018.
[5] Oyama et al., “Query Modification by Discovering Topics from Web Page Structures”, In Proceedings of the 6th Asia Pacific Web Conference (APWEB 2004), Lecture Notes in Computer Science Vol.3007, pp.553-564, 2004.
[6] 藤田, “アウトブレイクデータ解析の一般的な手法と注意点”, 環境感染誌 Vol.29 no.2, pp.80-92, 2014.
[7] 松本 他,“日本語形態素解析システム JUMAN 使用説明書 version2.0”, NAIST Technical Report, 1994.
[8] 岡田 他, “相関ルールとその周辺”, オペレーションズ・リサーチ : 経営の科学 Vol9 No.9, pp.565-571, 2002.
[9] 松尾 他, “Web 上の情報からの人間関係ネットワークの抽出”, 人工知能学会論文誌 20 巻 1 号 E ,pp.46-56,2005.
[10] 笹野 他,“構文・述語項構造解析システム KNP の解析の流れと特徴”, 言語処理学会 第 19 回年次大会 発表論文集,pp.110-113,2013.
[11] 嶋田 他, “SVM を用いた株価短報における意見文と事実文の抽出”, 言語処理学会 第 19 回年次大会 発表論文集, pp.15-17, 2013.