Vol.2016-NL-227 No.6
2016/7/29

業績変動を考慮した決算短信からの重要文抽出

磯沼 大[†1]　藤野 暢[†2]　浮田 純平[†3]　村上 遥[†4]
浅谷公威[†1]　森 純一郎[†1]　坂田 一郎[†1]

概要： 近年，記事生成などへの自動要約技術の適用が注目されている．本研究で対象とする決算記事は，どの事業や事象が企業全体の業績変動に大きな影響を及ぼすかといった記者の知見をもとに作成される．したがって自動要約においてはこうした記者の知見を抽出し，情報抽出･要約に適用する技術が必要である．本研究では過去の決算短信と決算記事から業績変動と短信文の掲載パターンを学習し，記事に掲載されるべき文を決算短信から抽出する手法を提案する．提案手法は 2 パートに分かれ，第 1 パートでは各事業セグメントの業績変動と，記事掲載パターンを学習することにより，各事業セグメントの重要度を判定する．第 2 パートでは，判定した各事業セグメントの重要度と極性判定を用いることで各文の重要度を評価し，抽出を行う．極性判定では，決算記事中の各表現に関する極性を自動で獲得し，非負値行列因子分解（NMF）による極性値推定を行うことで，決算記事に未出現の表現も含めた多様な表現に関する極性の獲得を可能にした．提案手法を適用して決算短信から抽出された文と実際の決算記事を比較した実験において，重要文抽出精度の評価を行い，事業セグメントの重要度判定と NMF による極性推定の有用性を確認した．

キーワード：テキストマイニング，情報抽出，決算短信，極性判定，NMF

1. はじめに

近年，個人株主数はオンライン株取引の普及も伴い増加傾向にある．決算発表の内容を要約した新聞記事（以下決算記事）は個人株主の投資判断において頻用される情報媒体の一つである．近年電子版新聞の普及により，決算記事には迅速性（決算発表後速やかに決算記事が配信されること），網羅性（より多くの企業に関する決算記事が配信されること）が求められている．一方でこれらを人手によって実現するには多大な人件コストを要する．したがって，自動要約技術による記事作成支援ニーズが高まっている．

次に決算記事に必要な要件について述べる．本稿では決算記事として各企業の売上高・純利益などの業績と，その業績要因文が記述されている記事を想定する．具体的には「A 社の売上高は X 億円だった．S 事業では訪日客需要の増加に伴い売上が増加した.」といった記述を想定する．一般的に業績要因文は，当該企業の業績変動に大きな影響を及ぼす重要な事業セグメントはどれか，注目するべき重要な商品や商材はどれかといった記者の知見をもとに記述される．したがって自動要約においてはこうした記者の記事作成における知見を抽出し,情報抽出･要約に適用する技術が必要である．

本稿で対象とする決算短信は，当期の業績とその業績の要因や当期の取り組みが数十の文によって記述されている．多事業展開している企業では，多くの場合各事業セグメントについてセグメント別の売上高・利益と，その要因についてセクションで分けて記述されている．したがって決算短信から重要な業績要因文を抽出する場合，まず重要な事業セグメントを特定し，その事業セグメントのセクションから重要文を抽出することが妥当であると考えられる．

以上の背景から，本稿では以下 2 つの手法を提案する．

- 各事業セグメントの業績変動と決算記事への掲載パターンを学習し，記事掲載の観点から事業セグメントの重要度を判定する手法
- 判定された事業セグメントの重要度に加え，各文の極性判定を行い，業績変動を最もよく説明する事象に関する文(以下重要文)を抽出する手法

極性判定では，決算記事中の各表現に関する極性を自動で獲得し，非負値行列因子分解（Non-Negative Matrix Factorization; NMF）[1]による極性値推定を行うことで，決算記事に未出現の表現も含めた多様な表現に関する極性の獲得と，精緻な極性判定を可能にした．

本稿では，2 節で重要文抽出や極性判定に関する関連研究について述べる．3 節では業績変動に基づいてセグメントの重要度を判定する手法について説明し，4 説ではセグメントの重要度と極性判定を用いた重要文の抽出方法について述べる．5 節の評価実験では，提案手法の有用性を確認するために，判定された重要セグメントと重要文を実際の決算記事と比較することで精度評価・分析を行う．

2. 関連研究

要約のための重要文抽出の手法として，PageRank[2]を応用した手法が提案されている[3][4]．Erkan らは文をグラフ構造化し，語彙の類似度の観点から各文のグラフにおける固有中心性 LexRank を算出し，重要度の指標として採用している．グラフの各ノードは各文を示し，エッジは各文を Bag-of-Words と tf-idf 法によってベクトル化し，そのコサイン類似度を採用している．それらを用いて隣接行列を作成し，隣接行列の固有ベクトルを計算することで各文の重要度を算出している．Otterbacher らは Erkan らの LexRank に各文のトピックとの関連性を加味した topic-sensitive

†1 東京大学大学院 工学系研究科 技術経営戦略学専攻
†2 東京大学大学院 新領域創成科学研究科 人間環境学専攻
†3 東京大学 医学部 医学科
†4 東京大学大学院 医学系研究科 生体物理医学専攻

LexRank を提案している[5]．これは抽出したいトピックに関する単語群をクエリとして定義し，隣接行列の値に各文間の類似度に加え，クエリと文の関連性を加える事で，クエリと関連性の高く，かつ各文と関連している文に高い重要度が与えられる．Filippova らは topic-sensitive LexRank を各企業の複数の経済ニュースの要約抽出に応用している[6]．クエリとして Yahoo! Finance の企業情報に含まれる各単語を用いることにより，企業固有の単語を含みながら，全体と関連度の高い文の抽出を行っている．また企業に関する説明文など，経済ニュースの要約文として採用されるべき情報を含まないにも関わらず複数のニュースで同様の表現が用いられることにより他の文との関連度が高くなってしまう．これらの文の重要度を補正するために，各文間の類似度の計算の際にクエリに頻繁に含まれている単語の重み付けを小さくするバイアス項を用いている．

文の極性判定に関して，Xiaowen らは意見を示す単語だけでは文の極性を判定できず，意見の側面も極性判定には必要であることを論じている[7][8]．例えば”long”という意見を示す単語を含む文について，デジタルカメラに関して”battery life is long”という文は肯定的であるのに対し，”it takes a long time to focus”という文は否定的である．したがって文の極性判定には意見を示す単語のみではなく，(バッテリーの耐久時間, 長い)といった意見の側面も含めた単語対の極性を用いるべきだと提案している．極性辞書の作成には文脈一貫性を利用し極性を獲得する研究がある[9][10]．Nasukawa らは，文書中では極性が連続する文脈が形成されることが多く，同文内や近接文間の極性は一致しやすいと述べており，これを文脈一貫性と呼んでいる．文脈一貫性の性質に基づき，一部の既知の極性表現を用いて， 未知の極性表現の獲得に成功している．

企業の決算短信から業績要因文を抽出する研究として酒井らの研究がある[11]．最初に「が好調」,「が不振」という手がかり表現を与え，それに係る節中に頻繁に出現する表現を共通頻出表現として獲得する．そして共通頻出表現が係る節中に頻繁に出現する表現を手がかり表現として新たに獲得するというプロセスを繰り返し，獲得した共通頻出表現と手がかり表現を用いて業績要因文を抽出している．また酒井らは業績要因文の重要度評価や極性付与を行っており．重要度評価に関して酒井らは各企業の Web ページから重要キーワードを抽出し，文中に含まれる各キーワードの重要度を用いて決算記事中の業績要因文の重要度を評価している[12]．極性付与に関しては「推移」,「増加」,「減少」を含んだ表現に着目し，係り受け解析を用いて（堅調に，推移）や（コスト，増加）といった極性表現を獲得し，人手で極性値を付与している[13]．

以上をふまえて，本稿の貢献について述べる．文章のグラフ構造化と固有ベクトル中心性算出による文の重要度評価[3]は「他の文に関連する文は重要である」という前提に基づいている．これは一般的な新聞記事予約や複数文書要約等各文が依存し合っている文書の要約では効果的であるが，決算短信のように各文が比較的独立である文書の要約には適さない．また業績との関連性という観点からの重要度を考慮できず，方向性が異なる．本稿では各文が関連する事業セグメントについて，業績変動と掲載パターンを抽出することで事業セグメントの重要度を判定し，重要文抽出に用いる手法を提案する．

企業固有の単語との関連性を加味した文の重要度評価[5] [6][11] [12]に関しては本稿でも 4.3 節の各文の重要度評価の際に用いている．また，極性値の取得においても，Xiaowen らの単語対の概念[7]や Nasukawa らの文脈一貫性の概念[9]を参照して取得を行っている．一方単語対の数は非常に多く，それらの網羅的な取得や，極性推定を行う一般的な手法が存在しないことから，本稿では NMF による未出現単語対の極性値推定を提案し，多様な単語対の極性値取得を行なっている．

3. 事業セグメントの重要度判定

提案手法の概要を図 1 に示す．提案手法は 2 パートに分かれており，第 1 パートでは各事業セグメントの業績変動と，記事掲載パターンを学習することにより，各事業セグメントの重要度を判定する．第 2 パートでは判定された事業セグメントの重要度と極性判定により各文の重要度を評価し，重要文抽出を行う．本節では第 1 パートについて説明する．第 1 パートの概要を図 2 に示す．

3.1 各事業セグメントの業績取得

多事業展開している企業の多くでは，決算短信で各事業セグメントにおけるセグメント売上高・利益を公表している．企業全体の売上高や利益は構造化データとして決算短信に添付されている xbrl ファイルに記述されているが，これらは構造化データとして公表されておらず文章中に記述されているため，本稿では後述する表現によるパターンマッチングによって取得した．

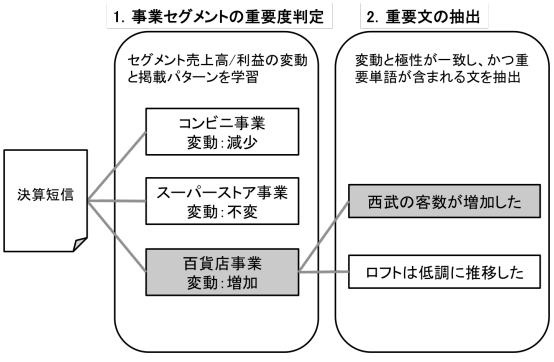

図 1　提案手法の概要

Figure 1　The outline of proposed method

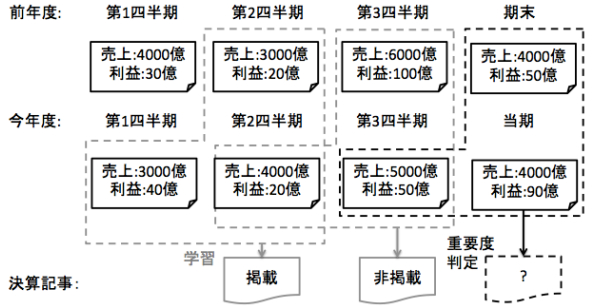

図 2　重要セグメントの判定手法概要

Figure 2　The outline of important segment identification

セグメント売上高・利益は各企業によって表現方法がそれぞれ異なる．それぞれについて概ね以下の表現が存在する．これらの表現を用いてパターンマッチングを行い，セグメント売上高･利益それぞれの値を取得した．

- セグメント売上高：セグメント売上高，セグメント収益，売上高，営業収益，営業総収入
- セグメント利益：セグメント利益，セグメント損失，営業利益，営業損失

3.2 特徴量の設計と学習

取得したセグメント売上高・利益，ならびに企業全体の売上高・利益を用いて，学習のための特徴量を設計する．本稿では掲載される事業セグメントについて，セグメント売上高・利益が企業全体の売上高・利益に占める割合が大きい，セグメント売上高・利益が前年同期，前期に比べ大きく変化しているという仮説をもとに式(1)のように説明変数を設計した．表 1 に説明変数で使用する変数の定義を示した．今期のセグメント売上高･利益それぞれについて，前年同期・前期のそれぞれとの変化量を算出し，それの企業全体の売上高・利益の変化量に対する割合を算出している．こうして，前述した仮説を反映する特徴量を設計する．

$$\boldsymbol{x} = \left(\frac{s_c - s_p}{S_c - S_p}, \frac{s_c - s_f}{S_c - S_f}, \frac{r_c - r_p}{R_c - R_p}, \frac{r_c - r_f}{R_c - R_f}\right) \tag{1}$$

一方目的変数yは事業セグメントの記事掲載における重要度であり，学習データでは当該事業セグメントが記事掲載されている場合に1を，されていない場合は0を付与している．本稿では学習データを自動で取得するために事業セグメント名とその略称が決算記事内に含まれているかどうか dice 係数を用いて判定し，目的変数を算出している．ここで dice 係数とは集合間の類似度である[14]．単語を文字集合と見立て，事業セグメント名w_sの文字集合$S(w_s)$と記事内の各単語w_tの文字集合$S(w_t)$間の dice 係数を算出し，その最大値が閾値以上であった場合にセグメント名またはその略称が含まれていると判定している（式(2)，式(3)）．本稿では，閾値として$threshold = 0.7$を用いている．

表 1　説明変数で使用する変数の定義

Table 1　The definition of variables

	セグメント売上高	全体の売上高	セグメント利益	全体の利益
今期	s_c	S_c	r_c	R_c
前年同期	s_p	S_p	r_p	R_p
前期	s_f	S_f	r_f	R_f

$$y = \begin{cases} 1 \; if \; \max_{w_t} dice(S(w_s), S(w_t)) > threshold \\ 0 \end{cases} \tag{2}$$

$$dice(S(w_s), S(w_t)) = \frac{2 \cdot |S(w_s) \cap S(w_t)|}{|S(w_s)| + |S(w_t)|} \tag{3}$$

得られた学習データ集合をもとに，サポートベクター回帰により事業セグメントsの重要度$y(s)$を推定する．

4. 極性判定による重要文抽出

本説では，3 節で判定されたセグメントの重要度，極性判定，及び重要単語の含有判定によりセグメントの業績変化を最もよく説明する事象に関する文を抽出する手法について述べる．

4.1 極性辞書の作成

極性とはある文がどのくらい肯定的か，あるいは否定的かを示す指標である．本稿ではある事象が売上高や利益の増加と関連する場合正の極性を，減少と関連する場合負の極性を持つと定める．

ある事象に関する文の極性を判定する際に参照する事象と極性値の集合を極性辞書という．本稿では事象の表現方法として単語対を用いる．「好調」など単語単体で極性が定まるものも存在するが，一方で「増加」については「客数が増加した」という文脈においては正の極性を持ち，「販管費が増加した」という文脈では負の極性を持つ．したがって単語単体ではなく，（客数，増加）などの単語対で事象を表現する．

次に既存決算記事から各単語対と極性値の取得方法について説明する．本稿で対象とする決算記事では前述したように売上高・利益とその業績要因文が記述される．したがって売上高・利益が増加した場合，その業績要因文は正の極性を持つ場合が多く，減少した場合業績要因文は負の極性を持つ場合が多いと考えられる．そこで業績要因文を係り受け解析し，係り受け元と係り受け先の単語対(i, j)について，売上高･利益増加時の出現回数$p_{i,j}$を総出現回数$n_{i,j}$で除すことでその単語対の極性値$v_{i,j}$を算出している(式(4))．極性値が1に近いほどその単語対は正の極性の傾向を持ち，0に近いほど負の極性の傾向を持つ．

$$v_{i,j} = \frac{p_{i,j}}{n_{i,j}} \tag{4}$$

係り受け先（単語m個）

係り受け元（単語n個）	増える v_1	増加 v_2	減る v_3	減少 v_4
販売台数 u_1	0.8	0.7	0.2	0.1
販管費 u_2	0.2	0.1	0.8	0.9
出店数 u_3	0.9	0.8	0.1	?

図 3　極性辞書の行列化の例
Figure 3　The example of polarity matrix

4.2 NMF による極性推定

極性値が得られた単語対は決算記事に出現するものに限られるため，未出現の単語対についても極性値を取得するために極性値の推定を行う. 極性辞書の各単語対について，係り受け元のn個の単語を行成分, 係り受け先のm個の単語を列成分として行列$D \in \{x, ? \mid 0 \leq x \leq 1, x \in R\}^{n \times m}$として行列化し, NMF を用いて未出現の単語対の極性値を推定する．ただし?の要素は未出現の単語対の極性値である．図 3 に示した例では，決算記事に未出現の（出店数，減少）という単語対を他の単語対の極性値を用いて補完することを考える．NMF では各単語は任意のk次元の潜在変数ベクトルで表わされると考え，単語対の極性値は単語対を構成する 2 単語の潜在変数ベクトルの内積で表されると捉える．その内積と実際の極性値との誤差をもとに潜在変数ベクトルを学習することで，極性値が未知の単語対(i,j)の極性を$\widehat{D}_{ij}$によって推定する（式(5)，式(6)）．本稿では潜在変数ベクトルの次元として$k = 30$を用いている．

$$\widehat{D} = U \cdot V^T \approx D \tag{5}$$

$$U \in R^{n \times k}, V \in R^{m \times k} \tag{6}$$

4.3 重要文の抽出

文が含まれている事業セグメントの重要度と極性判定，及び重要単語の含有判定により重要文を抽出する．重要単語とは各企業において売上高・利益規模の大きい事業名や商品名のことを指す．重要単語は当該企業の過去の決算記事において頻出し，かつ他企業の決算記事には出現しにくい単語であるという仮定から，単語tの企業cにおける重要度$e_{t,c}$を式(7)，式(8)，式(9)に示す tf-idf 値で表現する．ただし，$n_{t,c}$は単語tが企業cの決算短信に出現した回数，N_tは単語tが決算短信に出現した企業数，Nは総企業数である．

$$e_{t,c} = tf_{t,c} \cdot idf_t \tag{7}$$

$$tf_{t,c} = \frac{n_{t,c}}{\sum_k n_{k,c}} \tag{8}$$

$$idf_t = \log \frac{N}{N_t} \tag{9}$$

次に企業cの各業績要因文lについて，単語群W_lに含まれている各単語tの重要度$e_{t,c}$の総和と，lが含まれている事業セグメントs_lの重要度$y(s_l)$を 3 節の手法に基づき取得する．また文lの係り受け解析により取得した単語対集合P_l中の各単語対(i,j)の極性値$v_{i,j}$の平均を算出する．ここで，当該セグメントの前年同期比の売上変化量dが0未満である場合，極性値として反転した$1 - v_{i,j}$を用いている．これは業績の変化と合致した極性の文を抽出するためである．算出した重要度の総和と極性値の平均を乗じた値を文lの企業cにおける重要度$w_{l,c}$とし，重要度が上位の文を重要文として抽出する（式(10)）．

$$w_{l,c} = \begin{cases} y(s_l) \cdot \dfrac{1}{|P_l|} \displaystyle\sum_{(i,j) \in P_l} v_{i,j} \cdot \sum_{t \in W_l} e_{t,c} \ (if\ d \geq 0) \\ y(s_l) \cdot \dfrac{1}{|P_l|} \displaystyle\sum_{(i,j) \in P_l} (1 - v_{i,j}) \cdot \sum_{t \in W_l} e_{t,c} \ (if\ d < 0) \end{cases} \tag{10}$$

5. 実験

本実験では 3 節セグメントの重要度判定と 4 節業績要因文抽出のそれぞれについて実装し，精度評価を行った．セグメントの重要度判定では，各セグメントの重要度を連続値ではなくサポートベクターマシンによって 2 値分類し，対応する決算記事に当該セグメントの内容が記述されているか人手で確認を行い，精度評価を行った．重要文抽出では，重要度上位 5 つの文を重要文として抽出し，対応する決算記事に抽出文と同等の内容が記述されているか人手で確認し，精度評価を行った．実装に際し，形態素解析では MeCab[15]，係り受け解析では CaboCha[16]を用いた．

5.1 データセット

本稿では，日本国内金融商品取引所の上場会社が開示している決算短信を対象とする．決算短信は多くの企業で四半期毎に公表され，同時に売上高・利益などの業績が xml 形式で記述された xbrl ファイルも公表される．それらは各企業のホームページ並びに日本取引所グループの適時開示情報閲覧サービスにて取得が可能になっている．

データセットは 2013 年 1 月から 2016 年 3 月に開示された決算短信と，その決算発表について記述した日本経済新聞社の決算記事を用いた．学習データセットは 40 社 312 文書 923 セグメント，テストデータセットは 20 社 20 文書 62 セグメントの決算短信・決算記事を用いた．真の重要セグメント数は 62 セグメント中 42 セグメント，真の重要文数は 572 文中 93 文である．

表 2　重要セグメントの判定結果
Table 2　The result of important segment identification

	適合率	再現率	F 値
提案手法	0.92	0.65	0.76

5.2 実験結果・分析

まず重要セグメント判定の結果を表 2 に示す．提案手法では適合率が高く，掲載と判定されたセグメントの多くが実際の決算記事において掲載されていたことが分かる．

各セグメントの特徴量と掲載・非掲載との関係を考察するために，特徴量の一部である前期比売上高，前年同期比利益の変動値をプロットした散布図を図 4 に示す．右上，右下にプロットされたセグメントは業績変動が大きく，実際に記事掲載されており正しく重要セグメントと判別できていることが分かる．一方，左真中辺りにプロットされた業績変動が小さいが記事掲載されているセグメントもあり，これらを重要セグメントと判定できず再現率を下げていると考えられる．判定に失敗したセグメントは，資産売却や減損処理による特別利益や特別損失が発生し決算記事で取り上げられたものの，セグメント売上・利益には計上されず業績変動が小さいものが多数を占めた．したがって再現率を向上させるには特別利益や特別損失が発生したセグメントについて記事掲載を判定する手法が別途必要になる．

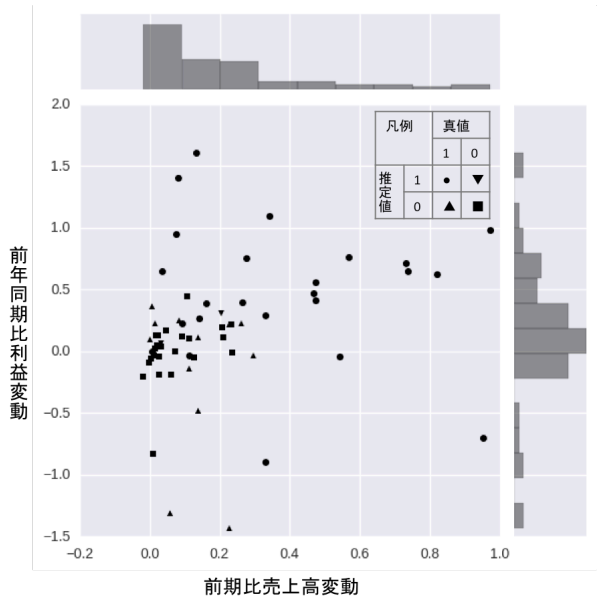

図 4　特徴量と掲載・非掲載の関係
Figure 4　The relation between features and segment appearance in the article

表 3　重要文抽出の結果
Table 3　The result of important sentence extraction

	適合率	再現率	F 値
(A) 単語の重要度判定 +極性辞書による極性判定	0.53	0.57	0.55
(B) (A)+セグメントの重要度判定	0.57	0.61	0.59
(C) (A)+NMF による極性推定	0.58	0.62	0.60
(D) (A)+セグメントの重要度判定 +NMF による極性推定	0.66	0.71	0.68

次に重要文抽出について比較実験を行った結果を表 3 に示す．比較実験では(A)セグメントの重要度判定と NMF による極性推定を用いない場合，即ち既存研究で提案されている単語の重要度判定と極性辞書を用いた極性判定による重要文抽出をベースラインに，(B)セグメントの重要度判定のみを加えた場合，(C)NMF による極性推定のみを加えた場合，(D)セグメントの重要度判定と NMF による極性推定の両方を加えた場合（提案手法）の 4 パターンについて精度比較を行った．比較の結果，本稿で提案するセグメントの重要度判定, NMF による極性推定の両方を用いた場合，両方を用いない場合よりも F 値が向上し，提案手法の有用性が確認された．

比較実験の結果について考察する．セグメントの重要度判定を加えない場合では，売上高・利益規模が小さく，実際の記事では注目される対象ではないが,「好調に推移した」などの極性表現に誘引されて誤抽出された例が目立った．決算短信では事業セグメントの規模に関する記述がほぼ無く,また業績変動の多寡に関する表現も限られているため，こうした定量的な重要度判定は有効性をもたらすと考えられる．また，数十の文から少数の重要文を判定する本実験のようなタスクでは，注目するべき箇所を特定できるという点で有効に働くと考えられる．

NMF による極性推定を行った場合では, 取得できなかった単語対の極性値が推定され文の極性判定がよりに精緻になったことで．精度が向上したと考えられる．例えば「飲食分野においては、クライアント接点の強化等に引き続き取り組んだ結果、取引店舗数が拡大しました（中略）」という文は，極性辞書のみでは抽出できなかったが，極性推定を行うことで正しく抽出された．これは極性辞書のみでは取得できなかった（強化，取り組む）という単語対について極性値を 0.7 と推定できたためである．また（店舗数，拡大）という単語対については極性辞書では極性値が 0.5 だったが，極性推定により極性値が 0.9 になり肯定的な極性を持つと補正された．極性辞書には（店舗数，増加），（受注，拡大）や（受注，増加）といった単語対に正の極性値が付与されており，これらによって極性値が推定されたと考えられる．

提案手法により決算短信から抽出した文と実際の決算記事を株式会社サンドラッグの2015年第3四半期決算発表を例に比較する．決算短信からの抽出文は「ドラッグストア事業は、消費増税後の反動減の回復により化粧品等を中心に販売が増加いたしました。なお、ドラッグストア事業の出店などの状況は、34 店舗を新規出店し、3 店舗のスクラップ&ビルドと 42 店舗を改装したほか、5 店舗の閉店とフランチャイズ 2 店舗を解約し活性化を図りました。第 3 四半期に入り暖冬の影響で使い捨てカイロやハンドクリーム等の季節商材は苦戦いたしましたが、引き続き食品等の販売強化や都市部を中心に拡大するインバウンド需要への対応に注力したことにより、既存店売上高は前年同期を大きく上回りました。ディスカウントストア事業は、消費増税後の反動減の回復により日用品、雑貨の販売が増加いたしました。第 3 四半期に入り暖冬の影響で灯油や季節家電・衣料品等は前年を下回りましたが、食品等の販売が好調に推移したことにより、既存店売上高は前年同期を大きく上回りました。」であった．一方，対応する実際の決算記事は「ドラッグストア大手のサンドラッグが 10 日発表した 2015 年 4~12 月期の連結決算は、純利益が前年同期比 38%増の 165 億円だった。4~12 月期として最高益を更新した。都市部の店舗で訪日外国人客向けに化粧品の販売が伸びた。売上高は 14%増の 3789 億円だった。訪日外国人需要が堅調だったほか、取り扱いを強化した食品の販売も 1 割程度増えた。45 店の新規出店にくわえ、59 店で化粧品や食品を強化する改装を実行し、既存店売上高が 8.5%増えた。暖冬の影響で使い捨てカイロやハンドクリームなど冬物商品の売れ行きは鈍った。」だった．「食品販売の強化」，「訪日外国人需要が堅調」や「季節商材の苦戦」といった内容が抽出できていることが確認された．

6. おわりに

本稿では，各事業セグメントの業績変動と決算記事への掲載パターンを学習し，事業セグメントの重要度を判定する手法と，判定された事業セグメントの重要度に加え，各文の極性判定を行い，重要文を抽出する手法を提案した．極性判定では，NMF による極性値推定を行うことで，多様な表現に関する極性の獲得を可能にした．提案手法を適用し，決算短信から抽出された文と実際の決算記事を比較した実験では，事業セグメントの重要度判定と NMF による極性推定を加えた場合重要文抽出精度が向上し，提案手法の有用性を確認した．

今後の課題として，資産売却や減損処理による特別利益や特別損失が発生した場合に，それらが記事掲載の観点から重要な事象かどうかを判定する手法が精度向上には必要である．また，決算短信から抽出した文の要約にも取り組むことを考えている．

謝辞

本研究は，東京大学大学院工学系研究科技術経営戦略学専攻松尾豊特任准教授と，同学術支援専門職員である椎橋徹夫氏から多くのご助言を頂きました．この場を借りて厚く御礼を申し上げます．

参考文献

[1] Lee, D. D. and Seung, H. S.: Learning the parts of objects by non-negative matrix factorization, Nature, Vol. 401, No. 6755, pp. 788-791 (1999).
[2] Lawrence, P., Sergey, B., Rajeev, M. and Terry, W.: The PageRank citation ranking: bringing order to the web, Tech. Rep. SIDL-WP-1999-0120, Stanford University (1999).
[3] Günes, E. and Radev, D. R.: LexRank: Graph-based lexical centrality as salience in text summarization, Journal of Artificial Intelligence Research, Vol. 22, No. 1, pp. 457-479 (2004).
[4] Mihalcea, R. and Tarau, P.: Textrank: Bringing order into texts, In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 404–411 (2004).
[5] Jahna, O., Erkan, G. and Radev, D. R.: Using random walks for question-focused sentence retrieval, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 915-922 (2005).
[6] Katja, F., Surdeanu, M., Ciaramita, M. and Zaragoza, H.: Company-oriented extractive summarization of financial news, Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 246-254 (2009).
[7] Xiaowen, D., Liu, B. and Yu, P. S.: A holistic lexicon-based approach to opinion mining, Proceedings of the 2008 international conference on web search and data mining, ACM, pp. 231-240 (2008).
[8] Murthy, G. and Liu, B.: Mining opinions in comparative sentences, Proceedings of the 22nd International Conference on Computational Linguistics, Association for Computational Linguistics, Vol. 1 (2008).
[9] Kanayama, H. and Nasukawa, T.: Fully automatic lexicon expansion for domain-oriented sentiment analysis, Proceedings of the 2006 conference on empirical methods in natural language processing, Association for Computational Linguistics, pp. 355-363 (2006).
[10] 乾孝司，梅澤佑介，山本幹雄：評価表現と文脈一貫性を利用した教師データ自動生成によるクレーム検出，自然言語処理，Vol.20, No.5, pp. 683-705 (2013).
[11] 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).
[12] 酒井浩之，増山繁：企業の業績発表記事からの重要業績要因の抽出，電子情報通信学会論文誌 D, Vol. 96, No. 11, pp. 2866-2870 (2013).
[13] 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 人工知能学会 金融情報学研究会，pp. 7-12 (2015).
[14] Dice, L. R.: Measures of the amount of ecologic association between species, Ecology, Vol. 26, No. 3, pp. 297-302 (1945).
[15] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, EMNLP, Vol. 4, pp. 230-237 (2004).
[16] Kudo, T. and Matsumoto, Y.: Japanese dependency analysis using cascaded chunking, proceedings of the 6th conference on Natural language learning, Association for Computational Linguistics, Vol. 20, pp. 1-7 (2002).