文脈を考慮した決算短信からの業績要因抽出

Extraction of corporate performance factors from summaries of financial statements considering context

加藤悠太 [1]　　酒井浩之 [1*]
Yuta Kato[1]　　Hiroyuki Sakai[1]

[1] 成蹊大学 理工学研究科 理工学専攻
[1] Department of Computer and Information Science, Faculty of Science and Technology, Seikei University

Abstract: In this paper, we proposed a method for automatically extracting sentences containing corporate performance factors from summaries of financial statements at high accuracy. More specifically, only sentences that can be determined to be corporate performance factors in the summaries of financial statements with high probability are extracted, and those sentences are used as training data. We trained the neural network by using the word representation of the training data. Furthermore, by using the appearance tendency of the corporate performance factor sentence specific to summaries of financial statements as a bias of model output, it became possible to extract with a higher f-measure than related work that performs filtering processing using corporate keywords .

1　はじめに

近年，証券市場における個人投資家の比重が増加しており個人投資家への投資判断を支援する技術の必要性が高まっている．そのため，人工知能分野における手法を金融市場の様々な場面に応用することが期待されている．例として，日本銀行が毎月発行している「金融経済月報」や経済新聞記事を解析し，経済市場を分析する研究などが盛んに行われている [1][2].

投資家にとって，企業の業績情報を収集することは重要であり，その中でも業績の要因となる情報が重要である [3][4]．しかしながら，証券市場における上場企業は膨大な数存在し，その企業が四半期に 1 回，決算発表を行う．そのため，人手により多くの企業の業績要因を把握することは困難である．決算短信の文において，この業績の要因が含まれている文を，本研究では「業績要因文」と定義する．

決算短信から業績要因文を抽出する手法とし，既存手法では特定の文を見つける上で手がかりとなる表現である「手がかり表現」を拡張した「拡張手がかり表現」と，企業の重要なキーワード「企業キーワード」による重みを用いて業績要因文を抽出する．その手法は適合率のみは非常に高い結果で抽出することが可能なため，その抽出した業績要因文を学習データの正例として深層学習を行う．この学習モデルによる業績要因文の抽出を行い，その抽出された文に対し企業キーワードによるフィルタリングをする手法が提案されている [5].

しかしながら，既提案手法は深層学習による文の抽出後，企業キーワードによるフィルタリング処理を行うことによって適合率を向上させているため，新規事業のような過去の決算短信で出現していなかったキーワードが含まれる業績要因文を抽出することができない．

そのため，本手法では企業の重要なキーワードによるフィルタリング処理を行わず，業績要因文の文脈や前後関係を考慮した手法によって既提案手法と同程度の適合率，再現率で業績要因文の抽出をすることを目的とする．

2　関連研究

既存手法 [6] を用い，決算短信から業績要因文，手がかり表現，企業キーワードを抽出する．そしてさらに手法 [5] により手がかり表現から拡張手がかり表現を獲得し，抽出された業績要因文に対し企業キーワードを用いてスコアを付与する．そのスコアと拡張手がかり表現により学習データの自動生成を行う．生成された学習データを用い，深層学習（多層パーセプトロン）

*連絡先：成蹊大学
〒 180-8633 東京都武蔵野市吉祥寺北町 3-3-1
E-mail: h-sakai@st.seikei.ac.jp

にて決算短信から業績要因文を抽出する．既存手法の概要を図 1 に示す．

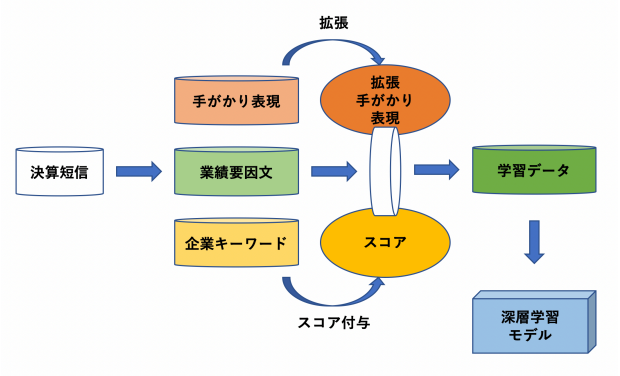

図 1: 既存手法 [5] の概要図

2.1 学習データの自動生成

獲得された拡張手がかり表現と企業キーワードを用い，以下の手順により学習データの自動生成を行う．

Step1: 決算短信から手法 [6] により業績要因文を抽出する．

Step2: 抽出した業績要因文に対し，企業キーワードを含みかつ拡張手がかり表現を含む文を抽出する．

Step3: 業績要因文に含まれている企業キーワードのスコアの合計をその業績抽出文のスコアとして付与する．

Step4: 業績要因文に付与したスコアの企業ごとの平均値を算出し，平均値より大きいスコアをもつ業績要因文を学習データの正例データとする．

Step5: 企業の決算短信から企業キーワードと手がかり表現のどちらも含まない文を抽出する．

Step6: Step5 で抽出した文のうち文字数が一定以上の文を学習データの負例データとする．

2.2 特徴量選択

自動生成した学習データを用い，業績要因文を抽出する．学習データにおいて正例の業績要因文に含まれる内容語（名詞，動詞，形容詞）に対し，以下の式にて重みを計算する．

$$W_p(t, S_p) = TF(t, S_p)H(t, S_p) \quad (1)$$

S_p: 正例に属する業績要因文の集合．

$TF(t, S_p)$: 文集合 S_p において語 t が出現する頻度．

$H(t, S_p)$: 文集合 S_p における各業績要因文に含まれる語 t の出現確率に基づくエントロピー．

$$H(t, S_p) = -\sum_{s \in S_p} P(t, S_p) \log_2 P(t, S_p) \quad (2)$$

$$P(t, S_p) = \frac{tf(t, s)}{\sum_{s \in S_p} tf(t, s)} \quad (3)$$

ここで，$P(t, s)$ は業績要因文 s における語 t の出現確率を表し，$tf(t, s)$ は文 s において語 t が出現する頻度を表す．

次に，負例の文に含まれる内容後（名詞，動詞，形容詞）に対しても同様に重みを計算する．

$$W_n(t, S_n) = TF(t, S_n)H(t, S_n) \quad (4)$$

ここで，S_n は学習データにおいて負例に属する文の集合を表す．

以下の条件どちらかが成り立つ語 t を特徴量として選択する．

$$W_p(t, S_p) > 2W_n(t, S_n) or W_n(t, S_n) > 2W_p(t, S_p) \quad (5)$$

2.3 業績要因文の抽出

多層パーセプトロンで学習されたモデルにより決算短信から業績要因文を抽出する．その後，精度を高めるために，抽出された文に含まれている企業キーワードのスコアの合計をその文のスコアとして付与し，スコアが高いかつ文末が「た」文のみを業績要因文として抽出する..

3 分散表現

3.1 単語の分散表現

既存手法 [5] では単語の重みを計算し特徴量選択を行っていたが，本手法では新規事業のような未知の単語が出現する文や業績要因文の中でも判定が曖昧な文に対し分類を行うことを可能にするため，単語の分散表現を用いることにより精度向上を図る．具体的には，分散表現獲得の基礎的な手法として，word2vec[7][8] と，より単語の文脈的な解釈を表すことが可能な ELMo[9] を使用した． 本手法では，word2vec の学習は Skip-gram により行った．

ELMo のモデル構造は双方向の LSTM から構成されており，単語の文脈情報を獲得することができる．例えば「増加」といった単語のベクトルを得たい場合，word2vec では「増加」という単語の周辺単語の出現確率から学習を行い，その結果「増加」という単語の分

散表現を一意で持つことになる．しかし ELMo では単語のコンテキストを考慮することにより，その単語が文脈的にどう扱われるのかという情報を分散表現として表すことが可能となる．そのため,「増加」という単語の文中での意味を表現することが可能となる．

3.2 文の分散表現

単語以外では文に対する分散表現の研究もされている．比較対象としてトピックを考慮した SCDV[10] を用いた．SCDV は単語がトピッククラスタに属する確率を GMM による分類から算出し，その確率と単語の *idf* 値から文の分散表現を生成し，その要素をスパースする手法である．

4 ELMo を用いた LSTM による業績要因文の抽出

ELMo は双方向 LSTM によって単語のコンテキストを学習した分散表現を獲得することが可能であり，それを word2vec のようなベクトルに連結することによって精度を上げることが可能とされている．また，一般的には ELMo 自体がコンテキストを学習しているため，獲得した分散表現に対して時系列性を考慮したモデルによって学習する必要はないとされている．しかしながら ELMo の学習は非常にコストが高いため，短期間でモデルの学習をすることは難しい．そのため，予め wikipedia の一般的なコーパスによって学習されたモデル [11][12] を用い，そのモデルの出力する分散表現を一般的な日本語におけるコンテキストと考え，さらにそのモデルの出力である決算短信の分散表現を LSTM によって学習を行う．これにより ELMo により決算短信を学習することに近づくと考えた．

本来 ELMo は双方向 LSTM によって学習を行うが，ELMo における決算短信の分散表現をさらに学習する際には，ある程度日本語としてのコンテキストは学習されていると判断し，図 2 のような順方向 LSTM のみで学習を行い，その出力を多層パーセプトロン（MLP）の入力として用いる．

5 業績要因文の出現傾向におけるバイアス値

決算短信の文集合において，業績要因が出現する頻度や傾向を分析した．その結果，図 3 のように業績要因文は連続して出現することが多いと分かった．この傾向は一度業績の要因について言及し始めると，それ

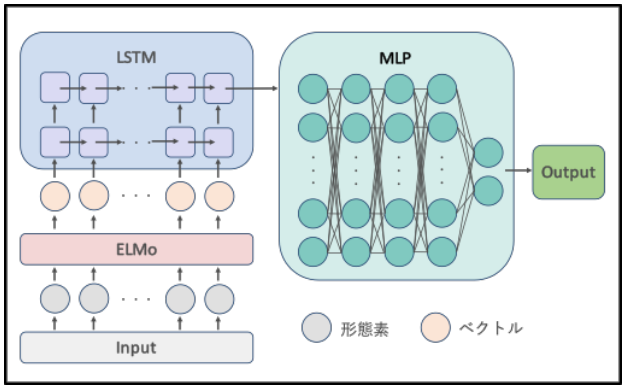

図 2: ELMo の分散表現を入力としたモデルの概要図

に関連する業績の要因をまとまった形で記すからであると考えられる．これを利用し業績要因が含まれてい

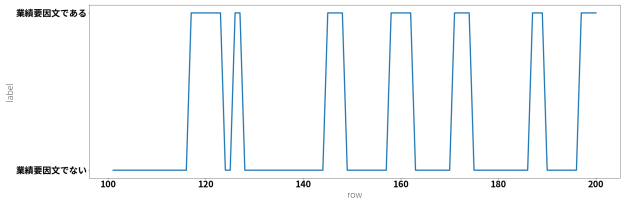

図 3: 業績要因文の出現傾向

ると判断された文の前後には業績要因が含まれている可能性が高いという仮説のもと，ある文の判定に対しその前後のモデルの出力をバイアスとしてかけることにより，決算短信における業績要因文の出現傾向の特徴を考慮した判定結果を得ることとした．これにより，決算短信に含まれる業績要因文の位置関係に基づく特徴をバイアス値として表現することができ，より決算短信に特化した分類結果を得ることが可能になる．このため，決算短信から業績要因文の抽出における精度向上が期待できる．

連続した決算短信の文 $s_1, s_2, \cdots, s_i, s_I$ があったとき，以下の式にて前後の文のバイアスをかけた判定結果 o'_i を得る．

$$o'_i = \frac{o_i + b_i}{2} \tag{6}$$

$$b_i = \frac{\sum_{-c \le j \le c, j \neq 0} 2^{|c|-|j|} o_{i+j}}{\sum_{-c \le j \le c, j \neq 0} 2^{|c|-|j|}} \tag{7}$$

このとき，

o_i: 文 s_i をモデルの入力とした時の出力結果．

b_i: 文 s_i に対し，前後の出力結果を加重平均したバイアス値．

c: コンテキストウィンドウ．決算短信の前後何文までをバイアスとして含めるか表すパラメータ．

これにより，業績要因文が集合している場合には業績要因文である可能性が高いというバイアスがかかり，また，業績要因が含まれていない文が集合している場合には業績要因が含まれていない可能性が高いというバイアスがかかることになり，決算短信における業績要因文の出現傾向の特徴を考慮した判定が可能となる．

6 実験

既存手法 [5] と本手法で結果を比較するため，自動生成された学習データを用い，word2vec，ELMo による単語の分散表現を特徴量とした MLP，LSTM[13][14] による業績要因文の抽出を行った．また，文の分散表現として SCDV を用いた MLP で業績要因文の抽出を行った．

word2vec，SCDV による分散表現を用いたモデルではハイパーパラメータ探索として k-fold 法を用いた 3 分割交差検証を行い，ベイズ最適化 [15] により探索を行った．ELMo に関しては計算コストが高いため，少量の汎用的なパラメータによってグリッドサーチを行った．

7 評価

7.1 分散表現を用いた業績要因抽出の評価

分散表現を用いた場合の業績要因の抽出結果について評価を行った．その評価結果を表 1 に示す．既存手法と比較し，ELMo を特徴量として LSTM と MLP により判定を行なった結果は，企業キーワードによるフィルタリング処理を行わなくとも，既存手法と同等の F 値を得ることができた．また，word2vec を特徴量として LSTM と MLP により判定を行なった結果に関しても既存手法には劣るが比較的良好な結果が得られた．

7.2 業績要因文の出現傾向をバイアスとした業績要因文の抽出

分散表現を用いた場合の業績要因抽出のモデルの出力に対し，決算短信の業績要因文出現傾向をバイアスとした実験について評価を行った．その評価結果を表 2 に示す．ELMo の分散表現を用いたモデルの出力にバイアスをかけた結果，コンテキストウィンドウ c を 1，2 に設定した場合，適合率，再現率共に向上する結果となった．

この結果から判定する前後の文との関係を考慮し，モデルの出力に対しバイアスとしてかけることにより精度が向上することがわかった．

表 1: 分散表現を用いた業績要因抽出の評価結果

	適合率	再現率	F 値
既存手法 MLP 企業キーワードフィルタ	0.827	0.724	0.772
特徴量: SCDV モデル: MLP	0.358	0.724	0.523
特徴量: word2vec モデル: MLP	0.462	0.859	0.601
特徴量: word2vec モデル: LSTM,MLP	0.609	0.834	0.704
特徴量: ELMo モデル: MLP	0.376	0.945	0.538
特徴量: ELMo モデル: LSTM,MLP	0.723	0.847	0.780

表 2: 出現傾向をバイアスとした抽出結果

	c	適合率	再現率	F 値
特徴量: ELMo モデル: MLP	0	0.376	0.945	0.538
	1	0.406	0.945	0.568
	2	0.413	0.949	0.575
特徴量: ELMo モデル: LSTM,MLP	0	0.723	0.847	0.780
	1	0.787	0.859	0.821
	2	0.796	0.876	0.835

8 考察

実験結果から，単語の文脈を考慮した分散表現を用いることにより，既存手法を上回る F 値を得ることができた．これは決算短信の 1 文における語彙数の分散が大きく，文の長さが 2 倍以上になる文も多く存在する．このため，tf 値やエントロピーなどの単語数による影響を受けやすい特徴量から分散表現にすることで，文の長さの影響を軽くすることが可能になったと考えられる．また，文脈による単語の扱われ方を考慮した表現に変えることで，業績の要因について書かれている文の単語のコンテキストを特徴として学習することが可能になったことも精度向上につながったと考える．

決算短信から業績要因文の抽出後，企業キーワードによるフィルタリング処理を行わなくすることで，学習データに存在しない企業キーワードになりうる単語が出現する文を決算短信から抽出することが可能になった．このため，企業が新規事業に対する業績要因として発表している文についてもフィルタリングによって除外されることはなくなった．また，新規事業の場合は新規で出現する単語が多く，重みの算出に影響がある．それを単語のコンテキストを考慮した分散表現に変えることや，新規事業に関する単語であっても，単

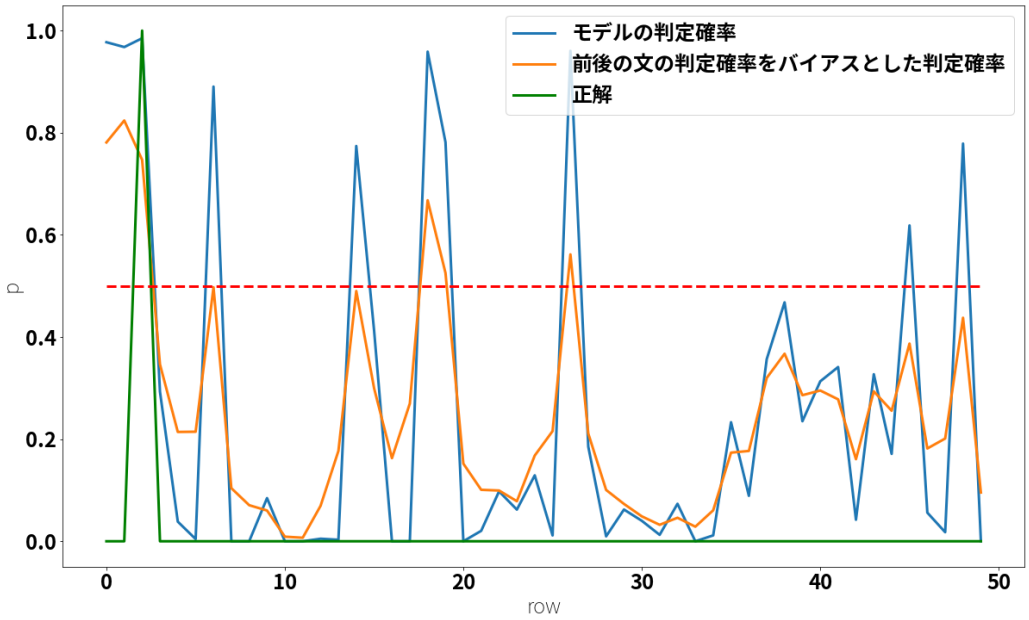

図 4: バイアスを用いた場合の判定結果の比較

語の分散表現を得るために使用した wikipedia には含まれていることは多いので，より新規事業に関する文を抽出することが可能になった一因と考えられる．

今回，文脈的解釈を含んだ分散表現が有用であると判断できたため，ELMo に SCDV のトピックのような観点を加えることでより精度が上がる可能性もある．そのため，トピックとコンテキストの双方を組み込むことが可能な分散表現についても決算短信の分析において有用であるか判断が必要である．

ELMo の分散表現については 1024 次元で行なったが，ベクトルの圧縮や学習データのサンプリングを行い，サンプリングしたデータで最適な学習パラメータを探索することも必要であると考えられる．

業績要因文であるかどうか判定する文の前後の判定結果を利用し，モデルの出力にバイアスをかける手法では有用な適合率，再現率が得られた．これは業績要因文であると誤判定してしまうことを前後の出力をバイアスとして用いた場合には抑制することが可能になったためである．図 4 からも前後の文の出力結果に偏り，誤判定が抑制していることが読み取れる．

前後の文が判定結果に依存するのであれば，単語や文字で使われる n-gram を文単位での文 n-gram として扱うことで判定結果が変わることになり，この結果を見ることで前後の文が判定文に対しどのように作用しているかを分析することが可能になると考えられる．近年では，bert[15] のような隣接文判定をタスクの一つとして解いているモデルもあるため，この検証も必要である．

9 むすび

本研究では，決算短信から業績要因を抽出する手法を企業の重要なキーワードや手がかり表現によるフィルタリングを行わず，文脈を考慮した手法によって既提案手法と同等の適合率，再現率で業績要因文の抽出を行った．具体的には，単語の頻度に基づく特徴量から単語の周辺確率やコンテキストを分散表現として表したものを利用し，ニューラルネットにより学習，判定を行うことで F 値 0.78 と既提案手法と同等の精度で抽出することが可能となった．さらに，決算短信における業績要因文の出現傾向の特徴として，業績要因が含まれる文は文章内で固まった状態で出現する性質を利用し，モデルの出力に対し，さらに前後のモデルの出力をバイアスとしてかけることで高い精度で抽出することができた．評価の結果，F 値が 0.83 と良好な結果を得ることができた．

今後の課題として，学習データと評価データの乖離を減らすことや，企業や決算短信ごとでの業績要因文の書き方を特徴として捉えることでバイアス値の調整や，学習するときの特徴量自体により前後の文の特徴を含ませることがあげられる．

参考文献

[1] 和泉潔, 後藤卓, 松井藤五郎: 経済テキスト情報を用いた長期的な市場動向の推定, 情報処理学会誌, Vol. 52, No. 12, pp. 3309–3315 (20011)

[2] 蔵本貴久, 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: 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, Vol. 28, No. 3, pp. 291–296 (2013)

[3] Hiroyuki, Sakai., Shigeru, Masuyama.: Assigning Polarity to Causal Information in Financial Articles on Business Performance of Companies, *IEICE Transactions on Information and Systems*, Vol. E92-D, No. 12, pp. 2341–2350 (2009)

[4] Hiroyuki, Sakai., Shigeru, Masuyama.: Cause Information Extraction from Financial Articles Concerning Business Performance, *IEICE Transactions on Information and Systems*, Vol. ED, No. 4, pp. 959–968 (2008)

[5] 酒井浩之, 松下和暉, 北島良三: 学習データの自動生成による決算短信からの業績要因文の抽出, 日本知能情報ファジィ学会誌, Vol. 31, No. 2, pp. 653–661 (2019)

[6] 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. J98-D, No. 5, pp. 172–182 (2015)

[7] Tomas, Mikolov., Kai, Chen., Greg, Corrado., Jeffrey, Dean.: Efficient Estimation of Word Representations in Vector Space, *ICLR Workshop* (2013)

[8] Xin, Rong.: word2vec Parameter Learning Explained, *arXiv preprint arXiv:1411.2738*, (2014)

[9] Matthew, E, Peters †., Mark, Neumann., Mohit, Iyyer., Matt, Gardner., Christopher, Clark., Kenton, Lee., Luke, Zettlemoyer.: Deep contextualized word representations, *Proc. of NAACL*, (2018)

[10] Mekala, Dheeraj., Gupta, Vivek., Paranjape, Bhargavi., Karnick, Harish.: SCDV: Sparse Composite Document Vectors using soft clustering over distributional representations, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 659–669, (2017)

[11] Che, Wanxiang., Liu, Yijia., Wang, Yuxuan., Zheng, Bo., Liu, Ting.: Towards Better UD Parsing: Deep Contextualized Word Embeddings, Ensemble, and Treebank Concatenation, *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 55–64, (2018)

[12] Fares, Murhaf., Kutuzov, Andrey., Oepen, Stephan., Velldal, Erik.: Word vectors, reuse, and replicability: Towards a community repository of large-text resources, *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pp. 271–276, (2017)

[13] S, Hochreiter., J, Schmidhuber.: Long short-term memory, *Neural Computation*, 1735-1780, (1997)

[14] Felix, A, Gers., Jurgen, Schmidhuber., Fred, Cummins.: Learning to forget: Continual prediction with LSTM *Neural Computation*, pp. 2451–2471, (2000)

[15] Devlin, Jacob., Chang, Ming-Wei., Lee, Kenton., Toutanova, Kristina.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv preprint arXiv:1810.04805*, (2018)