3H1-OS12a-4

製品・部門情報の企業業績要因表現からの抽出

Extraction of Product and Section Information from Causal Expressions on Business Performance of Companies

西崎 海人 Kaito Nishizaki　酒井 浩之 Hiroyuki Sakai　増山 繁 Shigeru Masuyama

豊橋技術科学大学 知識情報工学系
Department of Knowledge-based Information Engineering, Toyohashi University of Technology

We propose a method for the extraction of product and section information included in causal expressions extracted from financial articles on business performance of companies. Our method makes the pattern by using the rule obtained from the investigation of the causal expressions. Then, product and section information is extracted by using the pattern.

1. はじめに

近年，個人投資家が増加しており，投資判断を支援するための技術が必要になってきている．適切な投資判断を行うためには，企業の業績情報を収集することは重要である．投資判断だけでなく，企業経営においても同業他社の業績を分析することは重要である．業績分析のための知識源のひとつとして新聞記事が挙げられ，新聞記事から業績要因[*1](例えば,「新型の自動車の売り上げが好調だった」) を取得する方法 [Sakai 08] が提案されている．さらに，取得された業績要因に対して経常的な業績要因と一時的な業績要因に分類する方法 [藤村 09] が提案されており，一時的な業績要因よりも経常的な業績要因の方が株式市場に与える影響が大きいことが明らかにされている．さらに，経常的な業績要因であれば事業と関連する業績要因として高い重要度を付与し，一時的な業績要因であれば低い重要度を付与することで，個人投資家の投資判断を支援するのに有効な情報源となることが期待される．

ここで，経常的な業績要因に含まれる製品・部門情報は，業績要因に重要度を付与する際に活用できる情報であると考えられる．例えば，液晶テレビを主力事業としているS社にとって「冷蔵庫は堅調だった」よりも,「液晶テレビの売り上げが増加した」の方がインパクトが大きい．そのため，S社にとっては，前者よりも後者に対して高い重要度を付与すべきであり，重要度を付与する処理のためには,「冷蔵庫」や「液晶テレビ」という製品の情報が必要となる．また，製品の情報を含まず部門の情報のみを含む業績要因文[*2]が存在するため,「化成品部門」というような部門の情報も必要となる．

本研究では，藤村らの手法 [藤村 09] を用いて取得された経常的な業績要因文から，製品・部門情報の抽出を行う手法を提案する．

2. 関連研究

新聞記事から業績要因を抽出する研究として，Sakai ら [Sakai 08] は手がかり表現と共通頻出表現の二つを定義し，これらの表現をブートストラップ的に獲得し，取得された表現を用いて業績発表記事中から業績要因の抽出を行っている．そして，それらに対して極性 (positive，negative) の付与を行っている [Sakai 09]．また，藤村ら [藤村 09] は Sakai らの手法 [Sakai 08] を用いて取得された業績要因に対して，Support Vector Machine(SVM) を使用して経常的な業績要因文と一時的な業績要因文に分類している．さらに，イベントスタディ法に基づく分析により，一時的な業績要因よりも経常的な業績要因の方が株式市場に与える影響が大きいことを明らかにしている．

一方，固有表現抽出の研究として，Web やオンラインの新聞記事などの大規模なコーパスから固有表現に使われる語を自動的に学習する試みがいくつかなされている．代表的なものとしては文献 [Riloff 99],[Collins 99],[Yangarber 02] などがあげられる．これらの研究では，ブートストラッピングなどの手法を用いて与えられた少数の単語 (seed) から自動的な固有表現抽出を試み，よく使われる固有表現が比較的たやすく抽出できると期待されている．また，新山ら [新山 04] はコンパラブルな新聞記事 (同一の出来事を記述した異なる新聞社の記事) に着目した教師なし学習における固有表現例の収集について提案をしている．

それらに対して，本研究では経常的な業績要因文から製品・部門情報を抽出することに特化しており，それに特化した人手で作成したパタンを用いて抽出を行う．

3. 業績要因文とは

本研究における業績要因文は Sakai らの手法 [Sakai 08] によって新聞[*3]中の企業業績発表記事から取得されたものである．業績要因文には経常的な業績要因文と，一時的な業績要因文が存在し，藤村らの手法 [藤村 09] によっていずれかに分類される．分類された経常的な業績要因文を対象として，製品・部門情報の抽出を行う．経常的な業績要因文の一例を以下に示す．

> 六月から投入した「牛焼肉定食」や「豚生姜（しょうが）焼定食」の販売が好調だったほかファストフード業態の拡大も寄与した。

また，一時的な業績要因文の一例を以下に示す．

> 社宅跡地の売却益八十三億円などを特別利益に計上した。

連絡先: 豊橋技術科学大学，豊橋市天伯町雲雀ヶ丘 1-1，0532-44-6867，0532-44-6873，nishizaki@smlab.tutkie.tut.ac.jp

∗1　企業の業績に影響を与える要因

∗2　業績要因を含む文

∗3　日本経済新聞（1990 年から 2005 年）

4. 抽出対象とする情報

本研究における抽出対象となる情報の例を以下に示す．ここで，太字は抽出対象となる語を指す．上の三件は製品情報，下の二件は部門情報の例である．なお，例に示した“ スポーツ部門 ”や“ 半導体事業 ”のように部門名，事業名を部門情報として定義する．

- **レーザービームプリンター**や**複写機**など**事務機器**の好調が続き、**トナー**など**消耗品**も好調なため。
- **トイレタリー用品**や**化粧品向けの香料（フレグランス）**は価格低下が響き、売上高は横ばいにとどまりそう。
- 主力製品の**コーヒー飲料「W」**は六%増と伸びたものの、「**J**」は前期並みと伸び悩んだ。
- ただ、販売価格の低下が響いて**スポーツ部門**全体では前の期を下回った模様。
- **半導体事業**は上期に七百二十億円の営業利益を稼いだが、下期は上期に比べて大幅な減益が避けられない。

5. 製品・部門情報抽出手法

本研究では，人手で作成したパタンを用いて経常的な業績要因文から製品・部門情報を抽出する手法を提案する．

5.1 経常的な業績要因文の分類

抽出対象である経常的な業績要因文に対する調査の結果，藤村らの手法 [藤村 09] で取得された経常的な業績要因文中には，抽出対象を含まない文が存在した．抽出対象を含まない文に対して製品・部門情報の抽出を行うとノイズが多く含まれる可能性が高いため，前処理として抽出対象を含む文と含まない文に Support Vector Machine(SVM) で分類を行った．抽出対象を含まない経常的な業績要因文の一例を以下に示す．

> 「U 社」などの競合店に対抗して値下げやポイントカード制度の導入を進めたが、客足が遠ざかったうえ客単価も低下している。

5.1.1 使用する素性とデータセット

藤村らの手法 [藤村 09] で取得された経常的な業績要因文を形態素解析[*4]して形態素に分割し，形態素列を取得する．形態素列から形態素のユニグラム，バイグラム，トライグラムを生成し，これらの業績要因文中での頻度を求め，それらを素性とする．

データセットには，学習用データセット (データ数 1,195 件，正例 543 件，負例 652 件) と評価用データセット (データ数 1,000 件，正例 444 件，負例 556 件) を用いる．これらは，藤村らの手法 [藤村 09] で取得された経常的な業績要因文 2,195 件より作成した．ここで，正例は抽出対象を含む経常的な業績要因文，負例は抽出対象を含まない経常的な業績要因文とする．

5.1.2 評価実験

評価実験では，素性にユニグラムを用いたケース (以下，ケース 1)，ユニグラムとバイグラムの二つを用いたケース (以下，ケース 2)，ユニグラム，バイグラム，トライグラムの三つを用いたケース (以下，ケース 3) の三種類について，SVM^{light}[*5]を使用して実験を行った．

5.1.3 評価結果

評価結果を表 1 に示す．分類結果は正解率 86.4%となり，良好であると言える．素性の種類を変えながら実験を行った結果，正解率はケース 3 が最も高い値を示した．なお，ケース 1 およびケース 2 もほぼ同等の値を示した．

表 1: 評価結果（単位：%）

実験名	精度	再現率	正解率
ケース 1	84.7	84.5	86.3
ケース 2	84.3	84.9	86.3
ケース 3	84.5	84.9	86.4

5.2 パタン作成

5.1.1 章で述べた学習用データセットの正例 543 件に対する調査の結果，抽出対象となる語が出現する特定の型が存在することを確認した．以下に例を示す．

- 空気清浄機 **など**環境機器 **の売り上げ**が**伸び**た。
- デジカメ **や**ステッパー の**好調**が続くため。
- 液晶 **や**半導体 **など**の 部品**部門** が**伸び**る。
- 顧客の設備投資抑制を受けて**主力の**制御システム販売 が**伸び**悩む。
- **利益率が高い**通信機器**向け**フェライト・コア **の販売**が引き続き増える。
- 携帯電話**向けの**発振器 **など**が増加し、今期の連結純利益が前期比二・七倍の六十八億円と期初予想を達成できる見通しのため。
- 管球**部門** は 半導体 **や**液晶製造**用の**露光ランプ が**大幅**に**増加**し、一〇%程度の増収になる。

太字部分の周辺に抽出したい製品・部門情報が現れる傾向があったため，これらの特徴を用いてパタンを作成した．作成したパタンの総数は 50 である．以下に，作成したパタンの一部を示す．ここで，［　　］は抽出対象となる形態素列 (5.3 章で述べる) を指す．

- ［　　］など［　　］
- ［　　］や［　　］
- ［　　］ 助詞 + 好調
- ［　　］ 助詞 + 不振
- 利益率 が 高い ［　　］
- ［　　］[向け][の]［　　］
- ［　　］[用][の]［　　］
- ［　　部門］
- ［　　事業］

*4 形態素解析器として MeCab（http://mecab.sourceforge.net/）を用いた．

*5 SVM^{light}，http://svmlight.joachims.org/

5.3 提案手法

提案手法について述べる．

Step 1. 形態素列の取得

抽出対象を含む経常的な業績要因文を形態素解析して形態素に分割し，形態素列を取得する．ただし，複合名詞はそれ以上分割を行わない．

Step 2. 語の重み付け

Step 1 で取得した形態素列に対して，パタンをそのまま適用すると多くのノイズも取得されるため，Step 1 で取得した形態素列に含まれる各々の語に対して重み付けを行う．語の重み付けには IDF(文書頻度の逆数) を使用する．IDF の定義は式 1 のとおりである．

$$g_i = \log \frac{n}{n_i} \tag{1}$$

ただし，n は文書数 (業績要因文の総数)，n_i は文書頻度 (Step 1 で取得した形態素列に含まれる語 w_i を含む業績要因文の数) である．IDF を用いることにより，多数の業績要因文で使用されている語の重みは低くなり，語の重みが低い場合は抽出対象から除外する．試行錯誤の結果，本提案手法では閾値を 5 と設定し，語の重みが閾値以下であるものを除外した．また，地域を指す語のみである場合も必要な情報ではないため，人手で抽出対象から除外した．

Step 3. パタンの適用

パタンを適用して，製品・部門情報を取得する．

なお，製品情報と部門情報の区別はしておらず，どちらも同じ処理段階で取得される．

6. 評価実験

今までに述べた製品・部門情報抽出手法の評価を三つのデータの結果に基づいて行う．

6.1 評価データ

5.1.1 章で述べた評価用データセットから，SVM^{light} による分類 (ケース 3) で正例として得られた経常的な業績要因文をデータ 1 とする．データ 1 には 446 件の経常的な業績要因文が正例として分類されていた．分類の正解率は表 1 に示したように，86.4%である．また，抽出手法自体を評価するために，評価用データセットから，人手による分類で正例として得られた経常的な業績要因文 444 件をデータ 2 とする．参考として，学習用データセットから，人手による分類で正例として得られた経常的な業績要因文 543 件 (パタンを作成した際に調査に用いたもの) をデータ 3 とする．

6.2 評価結果

評価結果を表 2 に示す．SVM による分類で得られたデータ 1 では，精度 55.3%，再現率 54.6%とあまり良好な結果とは言えない．人手による分類で得られたデータ 2 では，データ 1 と比較して，精度，再現率ともにそれぞれ 9.30%，10.1%増加している．

表 2: 評価結果（単位：%）

対象データ	精度	再現率
データ 1	55.3	54.6
データ 2	64.6	64.7
データ 3	70.5	70.2

6.3 抽出結果

提案手法による抽出結果の例を表 3 および表 4 に示す．表 3 は正しく抽出された例，表 4 は誤って抽出された例である．ここで，下線部分は抽出された情報を指す．表 4 に示した誤りについてエラー解析を行ったところ，次のような問題を確認した．なお，各問題点への対応策については 7 章の考察で述べる．

問題点 1. 抽出対象を一つの情報として扱うか否か

“ トイレタリー用品や化粧品向けの香料（フレグランス）” (表 4，1 行目) は一つの情報として抽出されているが，本来は“ トイレタリー用品 ”と“ 化粧品向けの香料（フレグランス）” の二つの情報として抽出すべきものである．一方で，“ プリンターや複写機向けのローラー ”(表 3，5 行目) のように一つの情報として抽出すべきものも存在する．提案手法では，後者を優先したパタン「［　　］［や］［　　］［向け］［の］［　　］」を作成した．両者ともこのパタンにより抽出されているため，このような誤抽出が起こった．

問題点 2. 因果関係を含む業績要因文

“ S 社の家庭用ゲーム機に新型が登場した影響で旧世代向けのメモリーカードが急減するものの、増収効果で吸収する。” (表 4，2 行目) のように，因果関係の原因に該当する製品・部門情報が含まれる業績要因文が存在する．表 4 の 2 行目では因果関係の原因に該当する“ S 社の家庭用ゲーム機に新型が登場 ”から“ S 社の家庭用ゲーム機 ”が抽出されているが，因果関係の原因に該当する製品・部門情報は本来抽出すべきではない．また，表 4 の 3 行目については“ 半導体製造装置 ”と“ 液晶製造装置 ”の二つのみを抽出すべきである．なお，“ 半導体製造装置 ”が抽出されていない点については次で述べる．因果関係の原因に該当する製品・部門情報であるか否かを認識せずにパタンを適用しているため，このような誤抽出が起こった．

問題点 3. 語の重み付けによる抽出漏れ

単にパタンが当てはまらず抽出漏れが起こった場合以外に，語の重み付けによる抽出漏れが確認できた．提案手法では，IDF による語の重み付けを行い，語の重みが閾値以下であるものを抽出対象外と設定した．しかし，抽出対象外の語に，“ パソコン ”,“ 携帯電話 ”,“ メモリー ”のような抽出対象とすべき語が含まれていた．また，“ 半導体製造装置 ”という語も抽出対象外となっていたため，表 4 の 3 行目のような抽出漏れが起こった．

問題点 4. 抽出対象外とすべき語

“ 一二%増”(表 4，4 行目) は，製品・部門情報ではないので本来抽出すべきではない．しかし，「［　　］ 助詞 + 好調」というパタンによって，このような誤抽出が起こった．この誤抽出に対応するには，“ 一二%増”のような表記を抽出対象から除外する必要があると考える．

表 3: 正しく抽出された例

・レーザービームプリンター や 複写機 など 事務機器 の好調が続き、トナー など 消耗品 も好調なため。
・主力製品「C」や 濃縮タイプ製品 の販売が好調。
・半導体部門 が伸びたことに加え、パソコン用の記憶装置 など 電子機器の受託生産（ＥＭＳ）事業 が好調だった。
・パチンコ機向けの画像処理ＬＳＩ（大規模集積回路）が好調だった。
・バンドーは ベルト製品 に加え、プリンターや複写機向けのローラー などが好調。

表 4: 誤って抽出された例

・トイレタリー用品や化粧品向けの香料（フレグランス）は価格低下が響き、売上高は横ばいにとどまりそう。
・S 社の家庭用ゲーム機 に新型が登場した影響で 旧世代向けのメモリーカード が急減するものの、増収効果で吸収する。
・パソコンや携帯電話、携帯音楽プレーヤー など デジタル関連機器 の市場拡大を背景に、メモリーや 液晶ディスプレー の需要が増加し、半導体製造装置や 液晶製造装置 の販売が回復する。
・セラミックス部門 で フロッピーディスク用磁気ヘッド事業 から撤退したため伸び悩んだが、磁石 など マグネット部門 は 一二%増 と好調に推移した。

7. 考察

まず，6.2 章で述べた評価結果について考察する．SVM による分類で得られたデータ 1 と人手による分類で得られたデータ 2 では，精度，再現率ともにそれぞれ 9.30%，10.1%の差が表れたが，これは SVM による分類の正解率が 86.4%であることが影響している．また，人手による分類で得られたデータ 2 とデータ 3 では，パタンを作成した際に調査に用いたデータ 3 に対して，調査に用いていないデータ 2 は精度，再現率ともにそれぞれ 5.90%，5.50%低下している．データ 3 には存在しないパタンがデータ 2 に存在するとも考えられるが，データ 3 の再現率が 70.2%であることから，パタン作成を行った段階で既にパタンが網羅できていない．人手によるパタン作成では，パタンが網羅できていない場合，抽出できない情報がでてきてしまうため，半自動的なパタン作成が必要と考える．

次に，6.3 章で述べた抽出結果について考察する．問題点 1 については，抽出対象が並列関係か否かを判断する必要があると考える．並列関係か否かが判断できれば，並列関係である場合のパタンと並列関係でない場合のパタンを作成し，それにより誤抽出に対応できると考える．問題点 2 に対応するには，因果関係の原因に該当する製品・部門情報である場合は抽出しないパタンを作成する必要があると考える．問題点 3 については，閾値など語の重み付けについて再検討する必要がある．問題点 4 については，地域を指す語のみである場合を人手で抽出対象から除外したように，%が含まれる場合は抽出対象から除外することで対応できると考える．

8. まとめ

本研究では，人手で作成したパタンを用いて経常的な業績要因文から製品・部門情報を抽出する手法を提案した．評価結果は，SVM による分類で得られたデータ 1 で精度 55.3%，再現率 54.6%とあまり良好とは言えない．また，パタン作成を行った段階で既にパタンが網羅できていないことを確認した．さらに，エラー解析を行ったところ，6.3 章で述べたような問題点を確認した．

今後の課題として，7 章の考察で述べたように，半自動的なパタン作成の実現や誤抽出への対応策を試みるなどして，精度，再現率を改善したい．

謝辞

本研究の一部は，日本学術振興会科研費基盤研究 C(22500129)，若手研究 B(21700158)，電気通信普及財団，および人工知能財団の支援に基づいて行われた．

参考文献

[Sakai 08] H. Sakai, S. Masuyama,“ Cause Information Extraction from Financial Articles Concerning Business Performance”, IEICE Transactions on Information and Systems, Vol.E91-D, No.4, pp.959-968, April 2008.

[Sakai 09] H. Sakai, S. 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.

[藤村 09] 藤村真太郎, 酒井 浩之, 増山 繁,“ 企業業績要因文の経常的か否かに基づく分類とイベントスタディ法に基づく分析”, 第 23 回人工知能学会全国大会, 2009.

[Riloff 99] Ellen Riloff and Rosie Jones, 1999,“ Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping”, In Proceedings of the AAAI 1999.

[Collins 99] Michael Collins and Yoram Singer, 1999,“ Unsupervised Models for Named Entity Classification”, In Proceedings of the EMNLP 1999.

[Yangarber 02] Roman Yangarber, Winston Lin and Ralph Grishman, 2002,“ Unsupervised Learning of Generalized Names”, In Proceedings of the COLING 2002.

[新山 04] 新山祐介, 関根聡,“ コンパラブルな新聞記事からの固有表現の自動抽出”, 固有表現と専門用語ワークショップ, 2004.