3H1-3

企業の業績発表記事からの業績要因抽出と極性付与

Causal Information Extraction from Articles Concerning Business Performance of Companies and Polarity Assignment

酒井浩之[*1] 増山繁[*1]
Hiroyuki Sakai Shigeru Masuyama

[*1]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 of assigning polarity to causal information extracted from Japanese financial articles concerning business performance of companies. Our method assigns polarity, "positive" or "negative", according to business performance to causal information, e.g. "*zidousya no uriage ga koutyou*: (Sales of cars are good)" (The polarity "positive" is assigned in this example.). First, our method classifies articles concerning business performance into positive articles and negative articles. Using this classified sets of articles, our method assigns polarity, "positive" or "negative", to causal information extracted from the set of articles concerning business performance. Since our method does not need training dataset for assigning polarity to causal information, our method is able to assign polarity to causal information that consists of any sequence of words. We evaluated our method and it attained 73.8% precision and 46.1% recall of assigning positive, and 71.8% precision and 53.1% recall of assigning negative, respectively.

1. はじめに

現在，日本経済の成長を促す方策として，日本政府は「貯蓄から投資へ」を推奨している[*1]．投資家にとって，企業の業績に関する情報を収集することは重要であるが，実際の業績に関する情報だけでなく，その業績要因が重要である．なぜなら，業績拡大の要因が，その企業の主力事業が好調であることであったならば株価への影響は大きいが，株式売却益の計上などの特別利益の計上が要因であるならば株価への影響は軽微であるからである．しかし，証券市場の上場企業数は約 2500 社と多いうえに，近年では年に 4 回の決算発表がある．さらに，大幅な業績の修正を行う場合にも業績修正発表を行う必要があるため，人手によって全ての企業の業績要因を取得するのには多大な労力を要する．そのため，我々は，経済新聞記事から企業の業績発表記事を抽出し，その中から，業績要因（例えば，「新型の自動車の売り上げが好調だった」）を抽出する手法を提案した [Sakai 08]．抽出された業績要因は，例えば，証券アナリストへの支援材料として利用できる．しかし，より有効な情報として利用するためには，抽出した業績要因に対して業績に対する極性（「ポジティブ」,「ネガティブ」）を付与する必要がある．例えば，業績要因「ソフト販売の収益が寄与する」に対しては「ポジティブ」,「繊維部門の不振が響く」に対しては「ネガティブ」のラベルを付与する．業績要因に対して極性を付与することで，業績要因を使用した景気動向予測[*2]，および，業績要因に基づいて株取り引きを行うコンピュータトレーディングにも応用できることが期待できる．そのため，本稿では，業績要因に対して極性（「ポジティブ」,「ネガティブ」）のラベルを自動的に付与する手法を提案する．

我々の既提案手法では，抽出すべき業績要因を「共通頻出表現」と「手がかり表現」の 2 つの表現で構成される形態素列と定義した．ここで，手がかり表現を，業績要因獲得のための手がかり的な形態素列と定義し，共通頻出表現を，異なった業績要因に対して共通して頻出する形態素列と定義した（詳細は文献 [Sakai 08] を参照.）．例えば,「ソフト販売の収益が寄与する」では，手がかり表現が「が寄与する」であり，共通頻出表現は「ソフト販売」,「収益」である．既提案手法では，数多くの「手がかり表現」と「共通頻出表現」を自動的に獲得し，それらを使用することで業績要因を抽出する．ここで，業績要因に極性を付与する場合,「共通頻出表現」と「手がかり表現」の組合せに注目して極性付与を行う．しかしながら，例えば「が好調」の手がかり表現が含まれる業績要因には「ポジティブ」のラベルを付与できるが，手がかり表現「が増加」のように,「売り上げが増加」はポジティブであるが「リストラ費用が増加」はネガティブである．このことから，手がかり表現だけでは極性を判定できない．そのため，共通頻出表現と手がかり表現の組み合わせに着目する必要がある．しかし，数百種類の共通頻出表現，手がかり表現が抽出されるため組み合わせ数は膨大な数におよび，人手での極性付与は不可能である．また，機械学習による手法では学習データに存在しない共通頻出表現と手がかり表現の組合せには対応できないため，機械学習手法による業績要因への極性付与は適用できない．そこで，まず，業績発表記事を業績に対する極性で分類し（すなわち，業績が向上したなら「ポジティブ」，業績が悪化したなら「ネガティブ」に分類），その情報を利用して業績発表記事に含まれる業績要因に対して極性を付与する．これは，上記の例では，共通頻出表現「売り上げ」と手がかり表現「が増加」の組合せは，業績が向上した内容の記事に多く含まれ，共通頻出表現「リストラ費用」と手がかり表現「が増加」の組合せは，業績が悪化した内容の記事に多く含まれるという仮定に基づく．この統計情報を使用することで,「売り上げ」と「が増加」の組合せで構成される業績要因に対しては「ポジティブ」,「リストラ費用」と「が増加」の組合せで構成される業績要因に対しては「ネガティブ」の極性が付与できる．評価実験の結果，極性付与の精度はポジティブで 73.8%，ネガティブで 71.8%，再現率はポジティブで 46.1%，ネガティブで 53.1% であった．

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，sakai@smlab.tutkie.tut.ac.jp

*1 http://www.jasme.go.jp/jpn/summary/message041010.html

*2 すなわち,「ポジティブ」の業績要因が多くなれば，景気が回復することが予測できる（図 1 を参照）.

2. 関連研究

関連研究として，複数語で構成される表現（評価表現など）の感情極性を分類する研究がある．高村らは２つの単語から成る表現に対して隠れ変数モデルを用い，機械学習を用いて構成語の属性をクラスタという形で抽出して確率モデルを構築し，複数語表現の感情極性を分類する手法を提案している [高村 06]．この手法では未出現語（学習データに出現しない単語）からなる複数語表現は分類不可能である．それに対して，本手法では業績発表記事をあらかじめ極性分類し，その情報を使用することで業績要因への極性付与を行うため，業績要因への極性付与を行うための学習データを必要としない．よって，未出現語からなる業績要因に対しての極性付与も可能である．Turney は検索エンジンを用いて，複数語表現と “excellent” や “poor” といった極性が分かっている語との共起頻度を取得し，その情報を使用することで極性の付与を行う手法を提案している [Turney 02]．Wilson らは複数語表現を構成する語の極性（語の極性は既知）に基づいた複数語表現の極性分類手法を提案している [Wilson 05]．それらに対して，業績要因を取得するために必要な共通頻出表現，手がかり表現は数多くの種類があるうえに，その各共通頻出表現，手がかり表現の極性は既知ではない．そのうえ,「が増加」のように共起する共通頻出表現の種類によっては極性が変化する場合もあるため，業績要因を構成する語の極性を使用する手法を本タスクに適用することはできない．

Kaji らは，評価文に含まれる評価表現（名詞＋格助詞＋形容詞）に対して極性を決定する際に，人手で作成した手がかり表現リストやパタン，規則を使用して評価文を抽出して極性を付与しておき，評価表現が好評文に出現する頻度，不評文に出現する頻度を使用することで評価表現の極性を決定する手法を提案している [Kaji 07]．この手法では，同一の評価表現が不評文，好評文に３回以上，出現する必要があるが，業績要因は多くの名詞や動詞で構成されるため，同一の業績要因が出現することはまずない．そのため，Kaji らの手法を本タスクに適用することはできない．それに対して，本手法では，複数の名詞や動詞で構成される業績要因を複数の共通頻出表現と１つの手がかり表現で構成されると置き換えることにより簡略化することで，同一の業績要因が出現しない問題を解決している．

Koppel らは企業に関する記事がその企業の株価に影響を与えるかどうかを判別する手法を提案している [Koppel 04]．しかし，この研究では，記事のどの部分 (文や表現) が株価に影響を与えるのかを判定することはできない．それに対して，本研究では，業績発表記事から業績要因を抽出し極性を付与することで，株価に影響を与える原因を業績発表記事の中から取得することができる．

3. 企業の業績発表記事からの業績要因抽出

我々は既に企業の業績発表記事から業績要因を自動的に抽出する手法を提案している [Sakai 08]．本節では，既提案手法について簡単に述べる．なお，業績要因は１つの文中の複数の文節で構成される．既提案手法では，抽出すべき業績要因を「共通頻出表現」と「手がかり表現」の２つの表現で構成される形態素列と定義し，これらを自動的に獲得することで抽出を行う．

Step 1: 少数の手がかり表現（「が好調」,「が不振」）を人手で与え，それに係る節を取得する．

Step 2: 取得した節の集合から，その中で共通して頻繁に出現する表現を共通頻出表現として抽出する．

Step 3: 共通頻出表現が係る節を，新たな手がかり表現として獲得する．

Step 4: 獲得した手がかり表現から，それに係る節を取得する．

Step 5: Step 2 から Step 4 を，新たな手がかり表現と共通頻出表現が獲得されなくなる，もしくは，予め定めた回数まで繰り返す． □

3.1 共通頻出表現の抽出

本節では，共通頻出表現の自動獲得について述べる．まず，手がかり表現に係る文節に対して，それに係る文節を追加することで派生する表現を取得する．そして，既に得られている表現に係る文節を次々に追加することで派生する表現を全て取得する．ここで，手がかり表現に直接係っている文節から助詞を除去した形態素列を c とおく．次に，c から派生した各表現 e に対して，以下の式 1 で表されるスコアを計算する．

$$Score(e,c) \quad = \quad -f_e(e,c)\sqrt{f_p(e)}\log_2 P(e,c) \qquad (1)$$

ただし，$f_p(e)$ は表現 e に含まれる文節の数，$P(e,c)$ は c から派生する表現 e の派生確率，$f_e(e,c)$ は c から派生する表現 e の派生回数である．そして，c から派生する表現の中で，$f_e(e,c)$ の値が 2 以上である表現のうちスコアが最大の表現を共通頻出表現として抽出する．

次に，抽出された共通頻出表現の中から適切な共通頻出表現を選別する．具体的には，様々な手がかり表現に係っている共通頻出表現は適切であるという仮定に基づき，共通頻出表現が手がかり表現に係る確率に基づくエントロピーを式 2 で求め，その値が閾値 T_e 以上の共通頻出表現を選別する．

$$H(e) = -\sum_{s\in S(e)} P(e,s)\log_2 P(e,s) \qquad (2)$$

ただし，業績発表記事集合において，$P(e,s)$ は共通頻出表現 e が手がかり表現 s に係る確率，$S(e)$ は共通頻出表現 e が係る手がかり表現の集合である．閾値 T_e は，以下の式 3 によって設定する．

$$T_e = \alpha \log_2 Ns \qquad (3)$$

ただし，Ns は共通頻出表現を取得するのに使用した手がかり表現の集合，α は定数 $(0 < \alpha < 1)$ である．

3.2 新たな手がかり表現の獲得

共通頻出表現の選別を行った後，その選別した共通頻出表現から新たな手がかり表現を獲得する．まず，抽出した共通頻出表現を含む文を抽出し，その中で共通頻出表現を含む節 P_a が係っている文節 P_b を獲得する．次に，P_a に含まれる助詞を P_b に追加し，それを手がかり表現候補とする．ここで，様々な共通頻出表現が係っている手がかり表現は適切であるという仮定にもとづき，手がかり表現候補に対して共通頻出表現が係る確率に基づくエントロピーを式 4 で求め，閾値以上の候補を手がかり表現として抽出する．

$$H(s) = -\sum_{e\in E(s)} P(s,e)\log_2 P(s,e) \qquad (4)$$

ただし，業績発表記事集合において，$P(s,e)$ は手がかり表現 s に対して共通頻出表現 e が係る確率，$E(s)$ は手がかり表現 s に係る共通頻出表現の集合である．閾値は，共通頻出表現と同様に式 3 によって設定するが，Ns は新たな手がかり表現を獲得するのに使用した共通頻出表現の集合である．

4. 業績要因への極性付与

4.1 提案手法の概要

本手法では,「共通頻出表現」と「手がかり表現」の組合せに着目し，業績要因に対して極性を付与する手法について述べる．本手法では，業績要因への極性付与に共通頻出表現と手がかり表現の組合せを用いる．例えば，共通頻出表現「売り上げ」と手がかり表現「が増加」の組合せは，業績が向上した内容の記事に多く含まれ，共通頻出表現「リストラ費用」と手がかり表現「が増加」の組合せは，業績が悪化した内容の記事に多く含まれる可能性が高い．そのため，まず，業績発表記事を業績に対する極性で分類する（すなわち，業績が向上した内容の記事は「ポジティブ」，業績が悪化した記事は「ネガティブ」に分類）．この業績発表記事の極性分類には SVM を用いる．そして，極性に基づき 2 つに分類された業績発表記事集合における，共通頻出表現と手がかり表現の組合せの頻度分布を使用することで，業績要因への極性を付与する．

4.2 業績発表記事の極性分類手法

以下に，業績発表記事の極性を分類する手法を示す．まず，人手で抽出した業績発表記事に対して極性分類（ポジティブ，ネガティブ）し，SVM の訓練データとする．次に，業績発表記事の極性分類に有効な素性を選択する．具体的には，まず，訓練データのポジティブに分類された記事に含まれる語（名詞，動詞，形容詞）に対して，以下の式 5 で重み付けを行う．

$$W(t_i, S_p) = P(t_i, S_p)H(t_i, S_p) \tag{5}$$

ここで，$P(t_i, S_p)$ はポジティブに分類された記事集合 S_p における語 t_i の出現確率である．また，$H(t_i, S_p)$ は，S_p に含まれる各記事における語 t_i の出現確率に基づくエントロピーを表し，エントロピーが高い語ほど記事集合 S_p 中に均一に分布している語であることが分かる．同様にして，ネガティブに分類された記事集合 S_n に含まれる語（名詞，動詞，形容詞）に対しても，重み $W(t_i, S_n)$ を計算する．そして，以下の条件が成り立つ語 t_i を素性として抽出する．

$$W(t_i, S_p) > 2W(t_i, S_n) \ or \ W(t_i, S_n) > 2W(t_i, S_p) \tag{6}$$

その結果，例えば,「過去最高」や「下方修正」といった語が素性として抽出される．SVM による学習に用いる素性ベクトルの各要素は，訓練データの各文書における素性として選択された語の出現確率とし，カーネルは線形カーネルを使用した．また，実装にあたり，SVM^{light} *3を使用した．

4.3 業績要因への極性付与手法

業績要因に対しての極性付与は，ナイーブベイズに基づいた手法で行った．ここで，共通頻出表現を fp_i，手がかり表現を cp とする．そして，1 つの業績要因は，1 つの手がかり表現と 1 つ，ないし，複数の共通頻出表現で構成されるため，業績要因 $\boldsymbol{x}$ を以下のように定義する．

$$\boldsymbol{x} = (\langle fp_1, cp\rangle, \langle fp_2, cp\rangle, ..., \langle fp_n, cp\rangle) \tag{7}$$

ここで，業績要因の極性を，$c \in \{$ ポジティブ，ネガティブ $\}$ と定義する．業績要因 $\boldsymbol{x}$ に極性 c を付与するには，以下の式 8 で行う．

$$\hat{c} = \arg\max_c \ P(c|\boldsymbol{x}) = \arg\max_c \ P(c)P(\boldsymbol{x}|c) \tag{8}$$

*3 http://svmlight.joachims.org

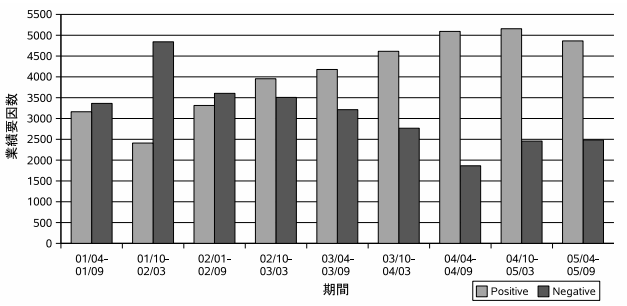

図 1: 業績要因の獲得数

表 1: 抽出された業績要因と付与された極性の例

業績要因：	中国向けの液晶関連の電子部品が回復
共通頻出表現：	電子部品，液晶関連
手がかり表現：	が回復
極性：	ポジティブ

そして，$\boldsymbol{x}$ の各要素 $\langle f_i, cp\rangle$ が独立に生起すると仮定し，条件付確率 $P(\boldsymbol{x}|c)$ を，以下の式で推定する．

$$P(\boldsymbol{x}|c) \approx \prod_{i=1}^{n} \frac{P(\langle fp_i, cp\rangle, c_d)}{P(c_d)} \tag{9}$$

ただし，$c_d \in \{$ ポジティブ，ネガティブ $\}$ を，業績発表記事が分類された極性と定義し，$P(\langle fp_i, cp\rangle, c_d)$ を，c_d に分類された業績発表記事において，$\langle fp_i, cp\rangle$ を含む業績要因が出現する確率，$P(c_d)$ を，業績要因が c_d に分類された業績発表記事において出現する確率とする．また，$c =$ ポジティブ のときの $P(c|\boldsymbol{x})$ を P_p，$c =$ ネガティブ のときの $P(c|\boldsymbol{x})$ を P_n とした場合，$P_p > 2P_n$ のときはポジティブ，$P_n > 2P_p$ のときはネガティブとし，それ以外のときは極性を付与しない．

5. 評価

5.1 実装

実装にあたり，形態素解析器として ChaSen*4，係り受け解析器として CaboCha*5を使用した．訓練データは，00 年の日経新聞記事から人手で業績発表記事を判別し，さらに，極性分類を行って訓練データとした．そして，01 年から 05 年の日経新聞記事集合から取得した 20880 個の業績発表記事に対して極性分類を行い，さらに，その業績発表記事集合から業績要因を抽出し，極性付与を行った*6．表 1 に，抽出された業績要因，および，付与された極性の例を示す．図 1 に，抽出された業績要因のうち，極性としてポジティブが付与された数，および，ネガティブが付与された数を，年度ごとに示す．

5.2 評価実験

正解データは，取得した 20880 個の業績発表記事の中から無作為に 138 個の記事を選び，その中から業績要因を人手で取得し，それらに極性を人手で付与して作成した．さらに，既提案手法によって獲得された手がかり表現と共通頻出表現を使用して 138 個の記事中から業績要因を獲得した後，本手法

*4 http://chasen-legacy.sourceforge.jp/
*5 http://chasen.org/~taku/software/cabocha/
*6 抽出された業績発表記事集合，業績要因集合には不適切なものも含まれるが，人手による選別は行わない．

表 2: 評価結果

α	業績要因の抽出数	$P_{posi}(\%)$	$R_{posi}(\%)$	$P_{nega}(\%)$	$R_{nega}(\%)$	$P_{CE}(\%)$	$R_{CE}(\%)$
0.3	89963	73.8	46.1	71.8	53.1	79.2	66.1
0.25	138678	68.1	58.7	65.9	71.2	72.7	77.6
0.2	178102	60.0	66.9	61.8	77.6	65.9	80.8

表 3: 評価結果（ベースライン手法）

α	$P_{posi}(\%)$	$R_{posi}(\%)$	$P_{nega}(\%)$	$R_{nega}(\%)$
0.3	58.5	43.1	54.9	52.9
0.25	55.0	55.0	51.6	69.2
0.2	48.8	60.2	47.0	75.2

で極性を付与した．そして，業績要因に付与された極性が正解データの業績要因に付与された極性と一致しているとき，その業績要因に付与された極性を正解として，精度，再現率を求めた．評価結果を表 2 に示す．ここで，P_{posi}，R_{posi} は，ポジティブが付与された業績要因の精度と再現率，P_{nega}，R_{nega} は，ネガティブが付与された業績要因の精度と再現率，α は，式 3 における，共通頻出表現と手がかり表現を選別する際の閾値を決定するためのパラメータである．また，参考として業績要因抽出の評価結果を表 2 に示す．ここで，P_{CE}，R_{CE} は業績要因抽出の精度，再現率である．

ベースライン手法として，業績要因の極性付与を，その業績要因が含まれる業績発表記事が分類された極性に従う（すなわち,「自動車の売り上げが好調」を含む業績発表記事がネガティブに分類されていれば，その業績要因にネガティブの極性を付与する.），簡単な手法による結果を表 3 に示す．

6. 考察

表 2 より，既提案手法によって抽出された業績要因に対して極性を付与した結果,「ポジティブ」の極性付与の精度が 73.8%,「ネガティブ」の極性付与の精度が 71.8% であり，良好な結果が得られた．ただし，評価にあたって，既提案手法によって抽出された不適切な業績要因（既提案手法による業績要因抽出の精度は 79.2%）を人手によって選別していないため，不適切な業績要因に対しても極性を付与する可能性があり，それらは必ず不正解と認定される．ここで，既提案手法の業績要因抽出が極性付与の精度に与える影響を除外するために，抽出された業績要因から誤ったものを人手で除去した業績要因の集合に対して極性を付与した場合,「ポジティブ」の極性付与の精度が 87.3%,「ネガティブ」の極性付与の精度が 85.9% であった．そのため，業績要因に極性を付与する本手法は，高い精度を達成できたと考える．また，表 3 から，ベースライン手法で行った場合の精度は「ポジティブ」で 58.5%,「ネガティブ」で 54.9% であった．これは，たとえ「ポジティブ」に分類された業績発表記事にも,「ネガティブ」な業績要因が含まれることも多く，業績要因の極性は，その業績要因が含まれる業績発表記事が分類された極性と必ずしも同一ではないことが分かる．ただし，再現率は「ポジティブ」で 46.1%,「ネガティブ」で 53.1% であり，低い結果となった．これは，既提案手法の業績要因抽出の再現率が 66.1% であるためであり，再現率の向上を今後の課題とする．

7. まとめ

本研究では，企業の業績発表記事から業績要因を抽出し，それらに極性（ポジティブ，ネガティブ）を付与する手法を提案した．具体的には，まず，業績発表記事を業績に対する極性で分類し，業績要因を構成する「共通頻出表現」と「手がかり表現」の組合せの頻度分布を利用して，業績要因に対して極性付与を行った．評価実験の結果，極性付与の精度はポジティブで 73.8%，ネガティブで 71.8%，再現率はポジティブで 46.1%，ネガティブで 53.1% であった．今後の課題として，再現率の向上と，業績要因と企業の事業内容との関連性を推定することを挙げる．本手法と既提案手法を組み合わせることで，自動的に業績要因を抽出し，極性を付与することが可能になったが，例えば「株式評価損を計上した」のような，事業内容との関連がない業績要因も抽出される．しかし，投資家にとって事業内容と関連がある業績要因のほうが重要な業績要因であるため，業績要因と事業内容との関連性の推定を行う必要があると考える．

参考文献

[Kaji 07] Kaji, N. and Kitsuregawa, M.: Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents, in *Proceedings of the EMNLP-CoNLL 2007*, pp. 1075–1083 (2007)

[Koppel 04] Koppel, M. and Shtrimberg, I.: Good News or Bad News? Let the Market Decide, in *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pp. 86–88 (2004)

[Sakai 08] Sakai, H. and Masuyama, S.: Cause Information Extraction from Financial Articles Concerning Business Performance, *IEICE Trans. Information and Systems*, Vol. E91-D, No. 4, pp. 959–968 (2008)

[高村 06] 高村 大也, 乾 孝司, 奥村 学.: 隠れ変数モデルによる複数語表現の感情極性分類, 情報処理学会論文誌, Vol. 47, No. 11, pp. 3021–3031 (2006)

[Turney 02] Turney, P. D.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews, in *Proceedings of the ACL2002*, pp. 417–424 (2002)

[Wilson 05] Wilson, T., Wiebe, J., and Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis, in *Proceedings of the HLT/EMNLP'05*, pp. 347–354 (2005)