社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－FI－84 (7)
2006－NL－175 (7)
2006／9／12

経済新聞記事内容の個々の企業におけるインパクトの判定

酒井 浩之† 増山 繁†,††

† 豊橋技術科学大学 知識情報工学系
†† 豊橋技術科学大学インテリジェントセンシングシステムリサーチセンター
E-mail: †sakai@smlab.tutkie.tut.ac.jp, ††masuyama@tutkie.tut.ac.jp

あらまし 新聞やインターネットなどで1日に配信される個々の企業に関する記事は膨大な数になるが，人間にとって重要な記事とは企業業績に影響を与えるほどのインパクトのある記事である．そのため，本研究では，経済新聞記事を対象とし，新聞に掲載される個々の企業の記事の内容を解析し，企業業績に影響を与えるほどのインパクトのある記事 (以下，インパクト記事) であるかどうかを判定し，そのような記事を抽出する．また，インパクト記事の内容が企業業績にとってポジティブな影響を与えるか，ネガティブな影響を与えるかを自動的に判定する．さらに，本手法で抽出対象としている企業業績発表の記事の内容を解析し，その主要因 (好調な事業，もしくは，不振の事業) が記載されている文を抽出する．本手法を評価したところ，インパクト記事抽出の精度は 85.8%，再現率は 66.8% であり，主要因 (好調な事業，不振な事業) の記述のある文抽出の精度は 82.2%，再現率は 26.3% であった．

キーワード インパクト判定，情報抽出，文抽出

Estimation of Impact Contained in Articles about each Company in Financial Articles

Hiroyuki SAKAI† and Shigeru MASUYAMA†,††

† Department of Knowledge-based Information Engineering, Toyohashi University of Technology,
†† Intelligent Sensing System Research Center, Toyohashi University of Technology,
E-mail: †sakai@smlab.tutkie.tut.ac.jp, ††masuyama@tutkie.tut.ac.jp

Abstract Many articles about each company are distributed on the newspaper or Internet in a day. However, an important article for human is an article containing a story that influences the corporate performance. In this research, we propose a method for identifying an article containing a story that influences the corporate performance and extracting such articles from a newspaper corpus. Our method judges whether the story contained in the extracted article is positive or negative to the corporate performance. Moreover, we target the articles of the announcement on the corporate performance, we propose a method for extracting sentences containing its key factor(good business or bad business). Experimental results showed that our method for extracting articles containing a story that influences the corporate performance attained 85.8% precision and 66.8% recall and our method for extracting sentences containing the key factor attained 82.2% precision and 26.3% recall.

Key words Impact estimation, Information extraction, Sentence extraction

1. はじめに

近年のインターネットの発達により，経済ニュースサイトやインターネットなどで1日に配信される企業のプレスリリースは膨大な数になる．しかしながら，実際に企業業績に影響を与えるほどのインパクトのある記事は，その中ではわずかである．しかし，人間にとって重要な記事とは，まさしく企業業績に影響を与えるほどのインパクトのある記事であり，それらをコンピュータが自動的に判別して抽出できれば，企業の情報収集における人間の手間を激減で

きる．

本研究では，新聞に掲載される個々の記事の内容を解析しインパクトのある記事であるかどうかを判定する．そして，その内容が企業業績にとってポジティブなのか，ネガティブなのかを推定する．ここで，「インパクト記事」を企業業績に影響を与えるほどの内容をもつ記事と定義する．また，実際の企業業績の発表もインパクト記事と定義する．ここで，企業業績にとってポジティブな影響を与えるインパクト記事を「ポジティブインパクト記事」とし，企業業績にとってネガティブな影響を与えるインパクト記事を「ネガティブインパクト記事」とする．以下に，本手法で抽出するポジティブインパクト記事の例を示す．

（1） 業績発表 (増益，業績回復)
（2） 企業間同士の提携，合併
（3） 画期的な新商品，および，技術の開発

また，ネガティブインパクト記事の例を示す．

（1） 業績発表 (減益，赤字転落)
（2） 企業の不祥事
（3） 提携解消

さらに，業績発表の記事に着目し，その主要因（好調な事業，もしくは，不振の事業）が記載されている文を，文中で好調な事業，不振な事業の後に出現する手がかり語から抽出する手法を提案する．(例えば，「インクジェットプリンター向けの吸収体やゴミ焼却炉用フィルターが好調」という記述を含む文を抽出し，この場合，手がかり語として「が好調」を使用する．) ただし，有効な手がかり語を全て人手で用意することは困難であるため，最初の数語を人手で与えた後，統計的情報から自動的に手がかり語を獲得していく．

第 2 章では，経済新聞記事からインパクト記事を抽出する手法について述べる．第 3 章では，第 2 章で抽出されたインパクト記事がポジティブなのか，ネガティブなのかを推定する手法について述べる．第 4 章では，第 2 章で抽出されたインパクト記事から，業績発表に関する記事に含まれる主要因（好調な事業，もしくは，不振の事業）が記載されている文を手がかり語を用いて抽出する手法について述べる．第 5 章では手法の実装について述べ，実際に抽出されたインパクト記事の実例を示す．第 6 章では評価を行い，第 7 章では評価結果を考察する．第 8 章では関連研究について紹介し，関連研究と本研究の違いや本手法の特徴について述べる．

2. インパクト記事の抽出

経済新聞記事からインパクトのある記事を判別して抽出する手法について述べる．抽出には Support Vector Machines(SVM) [6] を用いる．

2.1 訓練データの取得

まず，SVM の学習に用いる訓練データ (すなわち，インパクト記事とそうではない記事) の取得について述べる．本稿では「インパクト記事」を企業業績に影響を与えるほどの内容をもつ記事と定義したが，企業業績に影響を与えるかどうかは株価の変動を見れば推定できる．すなわち，ある企業に関する記事が出た日のその企業の株価が大幅に上昇していれば，その記事内容はポジティブインパクトである．逆に，ある企業に関する記事が出た日のその企業の株価が暴落していれば，その記事内容はネガティブインパクトである．そのため，株価変動率を基準として以下の方法で新聞記事コーパスからインパクト記事を収集する．

Step 1: 新聞記事集合から新聞記事の表題，および，本文に企業名が含まれている記事を抽出する．ここで企業名とは東証一部，二部上場の企業 3,724 社である．

Step 2: 抽出した記事に含まれる企業の，その記事の日付けにおける株価の前日比を得る．表題，および，本文に複数の企業が含まれていた場合は含まれている全ての企業の株価の前日比を求め，絶対値が最も高い前日比を採用する．なお，前日比は以下の式 1 で定義する．

$$\text{前日比} = \frac{\text{その日の終値} - \text{前日の終値}}{\text{前日の終値}} \times 100(\%) \quad (1)$$

Step 3: その企業が属する「業種別分類項目」(注1) の中分類を調べ，それに属する企業の Step 2 における日付けの前日比の平均を得る(注2)．

Step 4: その企業の前日比と属する業種分類の企業の前日比の差の絶対値を求め，その差が 8% 以上の記事を抽出する． □

Step 4 で，その企業の前日比と属する業種分類の企業の前日比の差の絶対値を求め，その差を用いた理由は以下のとおりである．株価の変動は企業活動だけではなく世界情勢等にも大きく影響を受ける．例えば，全体の株価が暴落しているときには個別の企業の株価も暴落しているため，個別企業ごとの前日比のみを基準としてはノイズとなってしまう．そこで，企業の前日比と属する業種分類の企業の前日比の差を求めることで世界情勢等の影響を排除する．つまり，全体の株価が下がっているときでも，その企業の株価がそれ以上に暴落していればネガティブインパクトのある内容が発表された可能性があり，逆に，全体の株価が暴落しているときでも，その企業の株価が上がっていればポジティブインパクトのある内容が発表された可能性がある．

しかし，Step 4 で抽出された記事にも多少のノイズが含まれている．そのため，実際には Step 4 で抽出された記事を人手で判定してノイズを除去した．90 年から 00 年の日経新聞記事に対して上記の手法を適用し，さらに人手でノイズを除去した結果，615 個のインパクト記事を抽出した．

(注1)：http://www2.tse.or.jp/sicc/category/ct_chart.html

(注2)：例えば，対象としている企業がソニーであるならば，ソニーの業種分類は「電気機器」であるので電機機器に属する全ての企業の前日比の平均を求める．

また，インパクト記事ではない記事を負例として取得する必要があるが，これは，記事の表題と本文に企業名が含まれており，かつ，前日比の絶対値が 1% 以下における記事をランダムに 615 個抽出した．

2.2 素性選択

SVM における素性を得るため，インパクト記事における特徴語を抽出する．ここで，特徴語はインパクト記事にのみ多く含まれている内容語 (名詞，動詞，形容詞) とした．(なお，複合名詞の場合は分解せず一つの名詞とした．) まず，インパクト記事に含まれている内容語 (以降，語とする) に対して重み付けを行なう．重み付けには次の式 2 を用いる．

$$W_p(t_i, S_p) = P(t_i, S_p) H(t_i, S_p) \tag{2}$$

$$P(t_i, S_p) = \frac{Tf(t_i, S_p)}{\sum_{t_i \in Ts_{S_p}} Tf(t_i, S_p)} \tag{3}$$

ただし，

$H(t_i, S_p)$: インパクト記事集合 S_p に含まれる各文書における語 t_i の出現確率に基づくエントロピー，

$P(t_i, S_p)$: 訓練データにおけるインパクト記事集合 S_p における語 t_i の出現確率，

$Tf(t_i, S_p)$: インパクト記事集合 S_p に含まれる語 t_i の数，

Ts_{S_p}: インパクト記事集合 S_p に含まれる語の集合，

$H(t_i, S_p)$ は，文書集合 S_p に含まれる各文書における語 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文書集合に均一に分布している語であることが分かる．この指標を導入する理由は，文書集合の中でも少数の文書に集中して出現している語よりも多くの文書に分散して出現している語の方がその文書集合の特徴を表すという仮定に基づく．$H(t_i, S_p)$ は次の式 4 で定義される．

$$H(t_i, S_p) = -\sum_{d \in S_p} P(t_i, d) \log_2 P(t_i, d) \tag{4}$$

$$P(t_i, d) = \frac{tf(t_i, d)}{\sum_{d \in S_p} tf(t_i, d)} \tag{5}$$

ここで，$P(t_i, d)$ は文書 d における語 t_i の出現確率を表す．

次に，インパクトがない記事集合に含まれる語に対して次の式 6 を用いて重み付けを行う．

$$W_n(t_i, S_n) = P(t_i, S_n) H(t_i, S_n) \tag{6}$$

$$P(t_i, S_n) = \frac{Tf(t_i, S_n)}{\sum_{t_i \in Ts_{S_n}} Tf(t_i, S_n)} \tag{7}$$

ただし，S_n は訓練データにおいて，インパクトがない記事集合である．そして，ある語 t_i のインパクト記事集合における重み $W_p(t_i, S_p)$ が，インパクトがない記事集合における重み $W_n(t_i, S_n)$ の 2 倍より大きければ，その語 t_i を特徴語として抽出する．すなわち，以下の条件が成り立つ語 t_i を特徴語として抽出する．

$$W_p(t_i, S_p) > 2W_n(t_i, S_n) \tag{8}$$

表 1 抽出された特徴語の例

見通し	売上高	経常利益	特別損失
好調	抱える	膨らむ	撤退
再建	転落	人員削減	リストラ
資本提携	事業提携	基本特許	企業買収

表 2 選択された素性の例

特別損失→計上	特別利益→計上
過去最高→更新	第三者割当増資→実施
評価損→計上	従来予想→上回る
多額→含み損	借入金→返済
不採算事業→撤退	財務体質→改善

式 2 で表した重みでは，特徴語ではない一般的な語でも高い重みが付与される．しかし，そのような語はインパクトがない記事集合においても高い重みが与えられる可能性が高い．そこで，ある語 t_i における重み $W_p(t_i, S_p)$ と重み $W_n(t_i, S_n)$ を比較し，$W_p(t_i, S_p)$ の方が大きい語を選択することで，一般的な語が特徴語として抽出されることを防ぐ．表 1 に抽出された特徴語を例示する．

素性としては，抽出した特徴語のスキップバイグラムを用いた．スキップバイグラムを用いた理由は，抽出した特徴語の対象として形容詞が含まれているからである．ただし，訓練データ中に 2 回以上，出現したもののみを採用した．表 2 に選択された素性をいくつか例示する．選択された素性を使用して SVM でインパクト記事の識別を行う．素性ベクトルは選択された素性が記事に含まれていれば 1，含まれていなければ 0 を要素として構成し，カーネルは線形カーネルを使用した．

3. インパクト内容の判別

抽出したインパクト記事が企業業績に対してポジティブかネガティブかを判定する．この判定にも SVM を使用したが素性は新たに抽出した．訓練データには，2.1 節で述べたインパクト記事の訓練データに対して，ポジティブかネガティブかを人手で付与したデータを使用する．素性には，2.2 節において抽出された特徴語を選別し，ポジティブ，ネガティブを表す特徴語のみを使用した．すなわち，インパクト記事集合を表す特徴語から，ポジティブ，ネガティブの内容をもつ記事の両方に出現する語を排除し，ポジティブインパクト記事にのみ多く出現する語，および，ネガティブインパクト記事にのみ多く出現する語を選別して素性として使用する．

ポジティブ，ネガティブを表す特徴語を以下のように選別する．まず，訓練データにおけるポジティブインパクト記事に含まれる語に対して重み付けを行う．

$$W_{posi}(t_i, S_{posi}) = P(t_i, S_{posi}) H(t_i, S_{posi}) \tag{9}$$

ただし，$P(t_i, S_{posi})$ は，訓練データにおけるポジティブインパクト記事集合 S_{posi} における語 t_i の出現確率，$H(t_i, S_{posi})$

は，ポジティブインパクト記事集合 S_{posi} に含まれる各文書における語 t_i の出現確率に基づくエントロピーである．

同様に，訓練データにおけるネガティブインパクト記事に含まれる語に対して重み付けを行う．

$$W_{nega}(t_i, S_{nega}) = P(t_i, S_{nega})H(t_i, S_{nega}) \qquad (10)$$

ただし，S_{nega} は訓練データにおけるネガティブインパクト記事集合である．

ここで，2.2 節において抽出された特徴語の集合を WF とし，その要素を $w \in WF$ とする．そして，以下の条件のどちらかを満たす特徴語を，ポジティブ，ネガティブを表す特徴語として選別する．

$$W_{posi}(w, S_{posi}) > 2W_{nega}(w, S_{nega})$$
$$W_{nega}(w, S_{nega}) > 2W_{posi}(w, S_{posi})$$

すなわち，2.2 節において抽出された特徴語の中で，訓練データのポジティブインパクト記事，ネガティブインパクト記事に偏って出現する特徴語のみが上記の条件をどちらか満たす．

4. 「好調な事業，不振な事業」の記述を含む文の抽出

本手法によってインパクト記事として認定される記事の内容で最も多いのは業績修正・発表の記事である．そのような記事には，多くの場合，業績が好調，もしくは不調な主要因が記載されている．例えば，「ショーワ連結純利益９８％増、９月中間。」という表題の記事には，「主要取引先であるホンダ向けに緩衝器やパワーステアリング部品が好調だった。」という文があり，業績９８％増の主要因は「緩衝器やパワーステアリング部品が好調」であることが分かる．

本節では，本手法によってインパクト記事と判定された業績修正・発表の記事の中から，その主要因（好調な事業，不振な事業）が記述してある文を自動的に抽出することを試みる．例えば，「インクジェットプリンター向けの吸収体やゴミ焼却炉用フィルターが好調」という記述や，「公共工事の減少で鉄管などが不振」という記述を含む文を抽出する．抽出は文中で好調な事業，不振な事業の後に出現する表現を手がかり語とし，それを使用することで行う．例えば，好調な事業の記述は「が好調」という手がかり語の前に出現していることが多い．また，不振な事業の記述は「が不振」という手がかり語の前に出現していることが多い．したがって，これらの有効な手がかり語を含む文を抽出すれば，好調な事業，不振な事業が記述してある文を自動的に抽出することができる．しかしながら，これらの記述を抽出するために有効な手がかり語は多く，それらを全て人手で用意することは困難である．そこで，手がかり語の自動取得を行った．

4.1 手がかり語の自動取得

手がかり語の自動取得は，我々が以前，提案した交通事故事例記事からの事故原因表現の獲得のための手法 [4] を本タスクに適用して行った．本手法の概要を以下に示す．

Step 1: 少数の手がかり語を人手で与え，それに係る節を取得する．

Step 2: 取得した節から共通部分 (後述) を抽出し，それを使用して新たな手がかり語を獲得する．

Step 3 獲得した手がかり語から，それに係る節を取得する．

Step 4: Step 2，3 を，新たな手がかり語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． □

Step 1 では，初期の手がかり語として「が好調」,「が不振」を人手で与えた．そして，それらに係る節を獲得する．例えば,「が好調」からは「二月に開業した北千住店など新店の売り上げ」などの節を獲得した．(実際には最後の文節から助詞を除去する．)

4.2 共通部分の抽出

次に，手がかり語から獲得された節から，その節に含まれる共通部分を抽出する．共通部分とは，主要因 (好調な事業，不振な事業) が記述されている節において主要因が異なっている場合でも共通して出現する部分と定義する．上記の「二月に開業した北千住店など新店の売り上げ」という例では「売り上げ」が共通部分となる．共通部分抽出の手法を以下に示す．

Step 1: 手がかり語に係る文節に次々と文節を追加することで派生する表現を全て取得．

Step 2: 各表現のスコアを計算．

Step 3: 手がかり語に係る文節から 2 回以上派生し，かつ，スコア最大の表現を共通部分として抽出． □

Step 1 では，例えば，文書 A に「新店の紳士服の売り上げが好調」という文が存在していたとすれば，手がかり語「が好調」に係る文節 (実際には，助詞「が」を除去して「好調」に係る文節)「売り上げ」と，それに文節を追加して「紳士服の売り上げ」「新店の紳士服の売り上げ」という 3 つの表現が派生する．また，文書 B に「主力のカードゲームの売り上げが好調」という文が存在していたとすれば，この文から「売り上げ」「カードゲームの売り上げ」「主力のカードゲームの売り上げ」という 3 つの表現が派生する．そして，文書 A と文書 B からは「売り上げ」が 2 回,「カードゲームの売り上げ」,「主力のカードゲームの売り上げ」,「紳士服の売り上げ」,「新店の紳士服の売り上げ」が 1 回，派生したことになる．

Step 2 では，手がかり語 c から派生した各表現 e に対して，式 11 で表されるスコアを計算する．

$$Score(e, c) = -pf(e)ef(e, c)\log_2 P(e, c) \qquad (11)$$

$$P(e, c) = \frac{ef(e, c)}{Ne(c)} \qquad (12)$$

表 3 抽出した共通部分

売り上げ	需要	受注
売上高	採算の良い商品	

ただし，

$P(e,c)$: 手がかり語 c から派生する表現 e の派生確率

$ef(e,c)$: 手がかり語 c から派生する表現 e の派生回数

$Ne(c)$: 手がかり語 c から派生する表現の総数

$pf(e)$: 表現 e に含まれる文節の数

例えば，前述の文書 A と文書 B の例では，「売り上げ」の $ef(e,c)$ の値は 2 であり $Ne(c)$ の値は 6 であるため，$P(e,c)$ の値は 2/6 となる．

Step 3 では，$ef(e,c)$ 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部分として抽出する．表 3 に実際に取得された共通部分をいくつか示す．

4.3 共通部分の選別

ある手がかり語から共通部分を抽出しても，中には不適切な表現も抽出される．そこで，手がかり語から抽出された共通部分の中から適切な共通部分を選別する．具体的には，様々な手がかり語に係っている共通部分は適切であるという仮定に基づき，共通部分が手がかり語に係る確率に基づくエントロピーを求め，その値がある閾値以上の共通部分を選別する．共通部分が手がかり語に係る確率に基づくエントロピーは式 13 で求める．

$$H(e) = -\sum_{s \in Se} P(e,s) \log_2 P(e,s) \tag{13}$$

$$P(e,s) = \frac{f(e,s)}{N(e)} \tag{14}$$

ただし，手がかり語を抽出するインパクト記事集合において，

$P(e,s)$: 共通部分 e が手がかり語 s に係る確率

$f(e,s)$: 手がかり語 s に係る共通部分 e の数

$N(e)$: 共通部分 e の総数

Se: 共通部分 e が係る手がかり語の集合

閾値 T_e は，以下の式 15 によって設定する．

$$T_e = \alpha \log_2 Ns \tag{15}$$

ただし，

Ns: 共通部分を取得するのに使用した手がかり語の数

α: 定数 $(0 < \alpha < 1)$

$\log_2 Ns$ は，共通部分が手がかり語に係る確率に基づくエントロピーの最大値を表し，その値と定数 α との積が閾値として設定される．

4.4 新たな手がかり語の獲得

共通部分の選別を行った後，その選別した共通部分から新たな手がかり語を獲得する．まず，抽出した共通部分を含む文を抽出し，その中で共通部分が係っている文節を獲得する．獲得した文節に対して共通部分を構成する最後尾の文節に含まれる助詞を追加し，それを手がかり語候補とする．そして，手がかり語候補が共通部分によって係られる確率に基づくエントロピーを求め，ある閾値以上の手がかり語候補を手がかり語として抽出する．手がかり語候補が共通部分によって係られる確率に基づくエントロピーは式 16 で求める．

表 4 獲得した手がかり語

が拡大する。	が減少。	が届かず、
が増加。	が増えた。	が伸びた
が落ち込んだ。	が回復する。	が低迷。

$$H(s) = -\sum_{e \in Es} P(s,e) \log_2 P(s,e) \tag{16}$$

$$P(s,e) = \frac{f(s,e)}{N(s)} \tag{17}$$

ただし，手がかり語を抽出するインパクト記事集合において，

$P(s,e)$: 手がかり語 s が共通部分 e によって係られる確率

$f(s,e)$: 共通部分 e によって係られる手がかり語 s の数

$N(s)$: 手がかり語 s の総数

Es: 手がかり語 s に係る共通部分の集合

閾値 T_s は，以下の式 18 によって設定する．

$$T_s = \alpha \log_2 Ne \tag{18}$$

Ne は新たな手がかり語を獲得するのに使用した共通部分の数である．また，定数 α は，共通部分選択の閾値を求めるときの定数と同じである．表 4 に，実際に初期の手がかり語から共通部分を取得し，取得した共通部分から手がかり語を獲得する手順を繰り返した結果，獲得した手がかり語をいくつか示す．

5. 手法の実装

本手法を実装した．実装にあたり，SVM^{light} (注3)を使用した．また，形態素解析器として ChaSen(注4)，係り受け解析器として CaboCha(注5)を使用した．訓練データは 90 年から 00 年の日経新聞記事に対して 2.1 節の手法を適用し，さらに人手でノイズを除去した結果，615 個のインパクト記事を抽出し，正例とした．さらに，この訓練データを使用して 90 年から 00 年の日経新聞記事に対してインパクト記事を抽出し，抽出された結果を人手で修正することで最終的に 2,895 個の記事を訓練データの正例とした．負例は，同数の前日比の絶対値が 1% 以下における記事である．

テストデータは 01 年から 05 年の日経新聞記事のうち表

(注3) : http://svmlight.joachims.org

(注4) : http://chasen.aist-nara.ac.jp/hiki/ChaSen/

(注5) : http://chasen.org/~taku/software/cabocha/

表 5　インパクト記事の抽出数

年	抽出数	Positive	Negative
2001	585	310	275
2002	565	312	253
2003	540	333	207
2004	536	407	129
2005	600	438	162

- ニッポン放送株、１５日時点で、ライブドア４９．７８％取得。
- フジテレビ・ライブドア、和解、きょう基本合意――資本・業務提携、取締役会で決定
- ライブドア、中古車のジャックを買収――ネット競売、ノウハウ活用。
- タカラ、トミー統合へ――玩具業界の再編加速。
- 前期、シャープ最高益――２４％増の１５１０億円、液晶好調映す。

図 1　ポジティブと判定されたインパクト記事の表題

- ソニー、電機立て直し急ぐ、今期純利益、一転半減へ。
- ソニー、ＴＶ不振鮮明――モノ作りで利益出せず、今期営業益、１３００億円下方修正
- エプソン６０％減益、今期純利益、再び下方修正――電子部品が不振。
- パイオニア営業赤字に、今期１０年ぶり、プラズマの競争激化、ＤＶＤレコーダー不振
- 三洋電機、白物家電・半導体を縮小、三洋クレ株一部売却、今期２０００億円最終赤字

図 2　ネガティブと判定されたインパクト記事の表題

題と本文に企業名が含まれている記事とし，(企業名とは東証一部，二部上場の企業 3,724 社である．(2006 年 1 月現在)) 26,528 記事がテストデータとして抽出された．そのテストデータに対してインパクト記事の抽出を行い，さらに，ポジティブかネガティブかの判定を行った．表 5 に，各年のテストデータから抽出したインパクト記事数，および，その中からポジティブと判定された記事数，ネガティブと判定された記事数を示す．図 1 にインパクト記事として認定され，さらに，ポジティブと判定された記事の表題をいくつか示す．図 2 にインパクト記事として認定され，さらに，ネガティブと判定された記事の表題をいくつか示す．さらに，抽出されたインパクト記事から好調な事業，不振な事業が記述されている文を得るための手がかり語を自動的に獲得し，その手がかり語を含む文を抽出した．なお，抽出されたインパクト記事には企業間提携や買収といった記事も含まれ，また，インパクトのない不正解な記事も含まれるが，人手による選別を行わず手がかり語の自動獲得を行った．なお，新たな手がかり語が獲得されなくなれば手がかり語獲得の処理を終えるが，定数 α の値によっては処理が終らない場合もある．そのため，手がかり語獲得，

- ホームヘルパーの育成などを手掛ける教育事業が好調だった。
- サッカーのワールドカップ（W杯）効果もあって広告関連が伸び、リストラも寄与した。
- 大型のシステム開発案件が減ったほか、パッケージソフト販売が不振だった。

図 3　抽出された好調な事業，不振な事業が記述されている文

共通部分抽出の繰り返し回数を 5 回と制限した．図 3 に抽出された文の例を示す．

6. 評　　価

6.1 インパクト記事抽出の評価

本手法を，人手による正解データとの比較を行うことで評価した．本稿では，2000 年から 2005 年の 5 年分の日経新聞記事に対してインパクト記事を抽出した．しかしながら，5 年分の記事数は膨大であり，それに含まれる全てのインパクト記事を対象として，精度，再現率を求めるのは困難である．そこで，以下のようにして評価用の正解データを作成し，正解データの範囲での精度，再現率を求めた．

まず，テストデータとして抽出した 01 年の日経新聞記事から 1200 記事を 4 人の被験者に読んでもらい，インパクト記事を人手で抽出した．そして，被験者 4 人のうち 3 人がインパクト記事とした記事を正解として，評価用の正解データを作成した．その結果，244 個のインパクト記事を正解データとして得た．インパクト記事抽出の精度，再現率の定義を以下に示す．

$$Precision(\text{インパクト記事抽出}) = \frac{|Sd \cap Ad|}{|Sd \cap N|}$$

$$Recall(\text{インパクト記事抽出}) = \frac{|Sd \cap Ad|}{|Ad|}$$

ただし，

Sd: 本手法によってテストデータから抽出したインパクト記事を要素とする集合．

Ad: 正解データである人手で抽出した 244 個のインパクト記事を要素とする集合．

N: 正解データを作成するために対象とした 1,200 個の記事．

さらに，正解データのインパクト記事を同じ被験者 4 人に読んでもらい，ポジティブかネガティブかの判定を行った．ポジティブ，ネガティブ判定の精度，再現率の定義を以下に示す．なお，以下ではポジティブの場合の定義を示す．

$$Precision(Positive) = \frac{|Sp \cap Ap|}{|Sp \cap N|}$$

$$Recall(Positive) = \frac{|Sp \cap Ap|}{|Ap|}$$

ただし，

Sp: 本手法によってテストデータから抽出したインパク

表 6 インパクト記事抽出の精度，再現率

	精度 (%)	再現率 (%)
インパクト記事	85.8	66.8
Positive	72.0	55.1
Positive(評価データのみ)	94.7	55.1
Negative	84.3	76.6
Negative(評価データのみ)	91.6	76.6

表 7 ベースライン手法との比較

	精度 (%)	再現率 (%)
本手法	85.8	66.8
$baseline_{20}$	86.4	31.1
$baseline_{50}$	82.5	54.1
$baseline_{100}$	72.8	69.3
$baseline_{w}$	72.1	77.5

ト記事の中で，本手法によってポジティブと判定された記事を要素とする集合．

Ap: 正解データであるインパクト記事の中で，ポジティブと判定された記事を要素とする集合．

N: 正解データを作成するために対象とした 1,200 個の記事．

評価結果を表 6 に示す．なお，(評価データのみ) とは，本手法によってテストデータから抽出したインパクト記事から不正解を除き，評価データに含まれるインパクト記事のみにおける場合の結果である．インパクト記事で不正解となった記事は，必ずポジティブ，ネガティブの判定でも不正解となるため，このような結果も併記する．

6.2 ベースライン手法との比較

インパクト記事の抽出においてベースライン手法と比較評価を行った．比較対象のベースライン手法を以下に示す．

$baseline_n$: 記事の表題，および，第 1 文に特徴語上位 n 個が 2 つ以上含まれていればインパクト記事として抽出

$baseline_w$: SVM の素性として特徴語を使用した場合

評価結果を表 7 に示す．

6.3 「好調な事業，不振な事業」の記述のある文抽出の評価

本手法によって抽出した「好調な事業，不振な事業」の記述のある文抽出の評価を行った．正解データは 244 個のインパクト記事の正解データから人手で「好調な事業，不振な事業」の記述のある文を抽出して正解データとし，精度，再現率を求めた．ただし，抽出される文の数は手がかり語を獲得する手法の定数 α によって大きく変化する．そのため，α を 0.9 から 0.1 まで変化させた場合の精度，再現率を測定した．結果を表 8 に示す．なお，表 8 には獲得した手がかり語の数も併記する．

7. 考　　察

表 6 より，インパクト記事抽出の精度は高いが，再現率が低い．特に，ネガティブインパクト記事に比べてポジティブインパクト記事の再現率が低い．これは，ネガティブインパクト記事の内容は，例えば，ソニーショックのような業績不振を伝える記事が大部分を占め，株価にも反映されやすいため訓練データにも多く含まれていたためである．本手法は，株価が大きく変動した日付の記事を訓練データとしているため，株価が変動しやすい内容が訓練データにも多く含まれる．特に業績発表の記事は株価にも反映されやすく，訓練データにも多く含まれるため，それに関連する素性も多く抽出された．そのため，業績発表の記事は多く抽出され，ネガティブインパクト記事における業績不振を伝える記事の割合が多いため，再現率が高くなった．ポジティブインパクト記事については，業績好調を伝える記事も含まれるが，それ以外にも企業間提携の話題，商品開発や技術開発の話題といった内容も含まれる．企業間提携の話題も株価へ反映されやすいため訓練データに多く含まれ，その結果，テストデータからも抽出されやすいが，商品開発や技術開発の話題が多く抽出できなかったことが再現率低下の原因であった．以下に被験者によってインパクト記事と判定されながらも，本手法によって抽出できなかった商品開発や技術開発の話題に関する記事の表題をいくつか示す．

- 「軽」初のハイブリッド車、ダイハツ、２００３年めどに発売。
- シャープや産総研が開発、ＤＶＤの容量２０倍に、赤色半導体レーザー利用

ただし，商品開発や技術開発の話題は訓練データにも少ない．これは，これらの話題が株価に反映されにくく，新薬の開発など反映されやすい分野もあるものの，株価変動率によって訓練データを取得する方法では多くの訓練データを得ることが困難であると考える．

表 8 「好調な事業，不振な事業」の記述のある文抽出の評価

α	精度 (%)	再現率 (%)	手がかり語の獲得数
0.9	90.2	9.9	2
0.6	90.2	9.9	2
0.5	89.5	11.1	6
0.4	82.2	26.3	60
0.3	54.4	47.1	299
0.2	21.7	78.1	3458
0.1	18.9	83.1	11236

「好調な事業，不振な事業」の記述のある文抽出は，定数 α が 0.4 までは精度が高いが再現率が低い．これは，有効な手がかり語の網羅率が低いからであると考える．例えば，定数 α が 0.4 では「が響く」といった手がかり語が取得できなかった．そのため，例えば「合成ゴムをはじめとする主力の石油化学部門の販売減が響く。」といった文が抽出できなかった．「が響く」は定数 α が 0.2 では抽出され，それによって再現率が上がっているが手がかり語として不適切な表現も多く獲得されたため精度が低下した．獲得で

きる有効な手がかり語の網羅率を上げることが必要であるが，手がかり語として不適切な表現を多く獲得しても精度，再現率を落とさず文抽出を行う方法としては，峠らが [5] で行ったような，文の特徴をパターンとして用意しそれを使用することで文抽出を行うことも有効であると考える．これは今後の課題とする．

8. 関連研究

経済記事の内容をポジティブかネガティブかに判定する研究はいくつか行われている．Koppel らは，企業に関する記事に対して，株価が上昇する内容であるか下落する内容であるかを分類する手法を提案している [1]．Lavrenko らは，企業に関する記事が発表された後の株価動向を推定する手法を提案している [2]．これらの研究では，与えられる記事が株価に変化を与えるほどのインパクトのある記事であることを前提としている．それに対して，本研究で入力として想定しているのは日経新聞記事であり，例えば「ダイエーホークス村松、早くも４０盗塁」といった本来の企業活動とは関係のない記事も多く含まれている．また，株価に変化を与えるほどのインパクトがない記事も多く含まれているため，本研究の方がより難しいタスクである．本手法は，これらの関連研究の入力として与えるインパクト記事を大規模知識源から獲得するための手法であるともいえる．もちろん，本手法によってインパクト記事として判定した記事から株価動向を推定するために，Lavrenko らの手法を使用することができる．

本研究では，「好調な事業，不振な事業」の記述のある文抽出を，少数の手がかり語から新たな手がかり語を自動獲得することで抽出している．峠らは，インターネット掲示板から主観的な評価を表している文を抽出するのに，人の主観的評価を表す単語を評価表現として，評価の対象となる単語を評価表現から２つの規則にあてはめ自動獲得し，さらに評価文のパターンを用いて文抽出を行っている [5]．那須川らは，好評文脈，不評文脈を分析し，好不評表現の性質を利用することでネット上の掲示版から好評表現，不評表現を取得する手法を提案している [3]．那須川らの手法では，種表現として少数の好評表現，不評表現を人手で与え，その種表現から好不評表現の性質を利用して文書中の好不評文脈を推測し，その中からさらに好評表現，不評表現を取得することを繰り返して，ブートストラップ的に多くの好評表現，不評表現を自動的に抽出している．手がかり語の自動取得という点や，人手で種表現を与えてブートストラップ的にあらたな手がかり語を獲得する点は本手法も同様なアプローチであるが，本手法では少数の手がかり語を人手で与えた後は全て統計的情報を使用して新たな手がかり語を獲得している点が異なる．

9. むすび

本研究では，新聞に掲載される記事の内容が企業業績に影響を与えるほどのものである記事を「インパクト記事」と定義し，新聞記事集合からインパクト記事を抽出した．そして，その内容が企業業績にとってポジティブなのか，ネガティブなのかを推定した．抽出には SVM を用い，株価変動率が大きい日の記事を訓練データとした．素性は訓練データに含まれる特徴語のスキップバイグラムとし，統計的情報を使用して特徴語の獲得を行った．さらに，業績発表の記事に着目し，その主要因（好調な事業，もしくは，不振の事業）が記載されている文を手がかり語から抽出した．ただし，有効な手がかり語を全て人手で用意することは困難であるため，最初の数語を人手で与えた後，統計的情報から自動的に手がかり語を獲得した．評価の結果，インパクト記事抽出の精度は 85.8%，再現率は 66.8% であり，精度は高いが再現率は低かった．これは，ポジティブインパクト記事，特に，「商品開発や技術開発」に関する記事が取得できなかったのが原因であった．また，「好調な事業，不振な事業」の記述のある文抽出の精度は 82.2%，再現率は 26.3% であり，精度は高いが再現率は低かった．これは，自動獲得した有効な手がかり語の網羅率が低いからであり，手がかり語の網羅率を上げることが今後の課題である．

謝　辞

本研究の一部は，文部科学省科学研究費特定領域研究 (B)(2)16092213，及び，２１世紀 COE プログラム「インテリジェントヒューマンセンシング」(豊橋技術科学大学) の援助により行われた．また，言語データとして，日経新聞 CD-ROM の使用を許可して頂いた日本経済新聞社に深謝する．

文　献

[1] 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).

[2] Lavrenko, V., Schmill, M., Lawrie, D. and Ogilvie, P.: Mining of Concurrent Text and Time Series, *In Proceedings of the KDD 2000 Conference Text Mining Workshop* (2001).

[3] 那須川哲哉, 金山博, 坪井祐太, 渡辺日出雄: 好不評文脈を応用した自然言語処理, 言語処理学会第 11 回年次大会発表論文集, pp. 153–156 (2005).

[4] 酒井浩之, 梅村祥之, 増山繁: 交通事故事例に含まれる事故原因表現の新聞記事からの抽出, 自然言語処理, Vol. 13, No. 4, pp. 99–124 (2006).

[5] 峠泰成, 山本和英: 手がかり語自動取得による Web 掲示板からの評価文抽出, 言語処理学会第 10 回年次大会発表論文集, pp. 107–110 (2004).

[6] Vapnik, V.: *Statistical Learning Theory*, Wiley (1999).