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

2007－NL－180 (26)
2007／7／25

景気動向を示す根拠表現の抽出と分析

坂地 泰紀† 酒井 浩之† 増山 繁†

† 豊橋技術科学大学 知識情報工学系
〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘1-1
E-mail: †sakaji@smlab.tutkie.tut.ac.jp, ††{sakai,masuyama}@tutkie.tut.ac.jp

あらまし 本研究では, 新聞記事から景気動向を示す「根拠となる表現」を統計的手法を用いて自動的に抽出する手法を提案する. また, 抽出された景気動向を示す「根拠となる表現」を景気が回復することを示すPositive表現と悪化することを示すNegative表現に分類する手法も併せて提案する. 企業や投資家にとって, 株価や商品の売行きを予測するために, 景気動向を知ることは重要なことである. そこで, 我々は景気動向に関する記事から景気動向を示す「根拠となる表現」を抽出し, それを用いることにより, 景気動向の予測が可能ではないかと考えた. 今回, 景気動向を予測するための材料として, 景気動向の根拠となる表現を抽出し, それが景気が回復することを示すのか, 悪化することを示すのかを判定する. 我々は1990年から2005年の日経新聞を用いて実験し, 景気動向を示す「根拠となる表現」の抽出手法と分類手法をそれぞれ評価した. その結果, 景気動向を示す「根拠となる表現」の抽出手法に関しては適合率が71.43%, 再現率が33.33%となり, 分類手法に関してはPositive表現分類のF値が0.695, Negative表現分類のF値が0.849となった.

キーワード 景気, 景気動向, 表現抽出, 表現分類

Extraction and Analysis of Basis Expressions that Indicate Economic Trends

Hiroki SAKAJI†, Hiroyuki SAKAI†, and Shigeru MASUYAMA†

† Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-tyou, Toyohashi-shi, Aichi-ken 441-8580
E-mail: †sakaji@smlab.tutkie.tut.ac.jp, ††{sakai,masuyama}@tutkie.tut.ac.jp

Abstract In this research, we propose a method to automatically extract basis expressions that indicate economic trends from newspaper articles by using a statistical method. We also propose a method to classify them into positive expressions that indicate upbeat, and negative expressions that indicate downturn in economy. It is important to foresee the economic trends for companies, governments and investors to forecast stock places and sales of goods. Therefore, we considered that the companies, governments and investors are able to forecast economic trends by using basis expressions extracted from newspaper articles concerning economic trends. In this time, we extracted basis expressions, and classified them into positive expressions or negative expressions as information to forecast economic trends. We evaluated our methods using NIKKEI newspaper articles from 1990 to 2005. The results showed that the method to extract basis expressions, attained presicion 71.43% and recall 33.33%. The classification method attained F-measure with positive expressions 0.695, and F-measure with negative expressions 0.849.

Key words Economy, Economic Trends, Expression Extraction, Expression Classification

1. はじめに

企業や政府, 投資家にとって, 株価の予想や商品の売行きを予測するために, 景気動向を知ることは重要なことである.

景気動向に関する指標には, 景気動向指数 (Diffusion Index)(注1)がある. 景気動向指数は, 生産, 雇用など様々な経済活

(注1)：http://www.esri.cao.go.jp/jp/stat/di/di.html

動での重要かつ景気に敏感な指標の動きを統合することによって，景気の現状把握及び将来予測に資するために作成された統合的な景気指標である．景気動向指数は3ヶ月ごとに計算して求められているため，前期の景気動向は知ることができる．しかしながら，景気動向指数を用いて景気の予測を行う場合は，情報が古いため正確な予測を行うことが難しい．

そこで，中嶋ら [1] は景気の動向に関する記述がある記事 (以下，景気動向記事と略記) を新聞記事集合から抽出し，抽出した景気動向記事を景気が回復することを示す Positive 記事と，悪化することを示す Negative 記事に自動的に分類する手法を提案した．この手法を用いれば，現時点における景気動向と未来の景気動向を知るための判断材料として，分類した記事を得ることができる．しかしながら，同一記事において，景気の動向に関する異なる見解を含むことがあり，そのような場合には対応できていない．例えば，「愛知県の景気は回復しているが，岐阜県の景気は悪化している」など，異なった地方について異なった見解が含まれている記事がある．

そこで，景気動向記事から景気動向を示す根拠となる表現 (以下，根拠表現と略記) を抽出し，それが Positive を表すのか，Negative を表すのかを分類する手法を提案する．我々は景気動向記事に含まれている根拠表現が，Positive を表すのか Negative を表すのかが判断できれば，景気予測の材料として用いることができると考えた．根拠表現を用いることで，記事内に含まれる異なった見解ごとの分類も可能となる．

本研究では，景気動向記事から根拠表現をブートストラップ的に抽出する．そして，抽出した根拠表現の分析を学習器を用いて行う．

2. 提案手法

この章では，本研究で提案する手法について説明する．

2.1 景気動向記事の抽出

景気動向記事中に含まれる景気動向の根拠となる表現を抽出するために，その前処理として，新聞記事から景気動向記事を抽出する．そして，抽出された景気動向記事から景気動向の根拠を示す表現を抽出する．新聞記事からの景気動向記事の抽出には Support Vector Machines(SVM) [5] を用いる．

2.1.1 特徴語選択

SVM に用いる素性抽出の前処理として，次に示す「特徴語」を選択する．特徴語の選択には酒井ら [2] が用いた式を使う．

特徴語: 景気動向記事集合中にのみ広く分布して出現する語．

まず，記事を形態素解析(注2)し，それから以下の条件のもと，正例 (正例と負例については，3. で後述する) に 3 回以上出現する語を特徴語候補とする．

- 名詞，副詞，形容詞，動詞を対象とする
- 名詞の列は複合名詞として抽出する
- 動詞，形容詞は原形に直して扱う

また，各特徴語候補に対して重み付けを行い，特徴語を選択する．重み付けの式には式 (1) を用いる．

$$W_p(t_i, S_p) = P(t_i, S_p)H(t_i, S_p) \tag{1}$$

$$H(t_i, S_p) = -\sum_{d \in S_p} P(t_i, d) \log_2 P(t_i, d) \tag{2}$$

$$P(t_i, d) = \frac{tf(t_i, d)}{\sum_{d' \in S_p} tf(t_i, d')} \tag{3}$$

$P(t_i, S_p)$: 正例の記事集合 S_p(注3)における語 t_i(注4)の出現確率
$Tf(t_i, S_p)$: 正例の記事集合 S_p に含まれる語 t_i の出現回数
Ts_{S_p} : 正例の記事集合 S_p に含まれる語の集合
$P(t_i, d)$: 記事 d における語 t_i の出現確率
$H(t_i, S_p)$: 記事集合 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
$tf(t_i, d)$: 文書 d に含まれる語 t_i の出現回数

また，正例の場合と同様に，負例に含まれる語に対して式 (1) と同じ重み付けを行う．ある語 t_i の正例における重み $W_p(t_i, S_p)$ が負例における重み $W_n(t_i, S_n)$ の 2 倍より大きければ，その語 t_i を特徴語として選択する．すなわち，以下の条件が成り立つ語 t_i を特徴語として選択する．

$$W_p(t_i, S_p) > 2W_n(t_i, S_n) \tag{4}$$

エントロピーを用いた理由は，正例の記事集合中で多くの記事に分散して出現している語の方が，少数の記事に集中して出現している語と比較して，よりその記事集合の特徴を表し，特徴語としても有効であるという仮定に基づく．また，W_n を 2 倍した理由は，一般的な語が特徴語として選択されることを防ぐためである．

2.1.2 素性抽出

SVM に用いる素性として以下の 2 種類を用いた．

特徴語を用いた素性: 選択した特徴語を素性として用いる．

特徴語 bi-gram: 記事を選択した特徴語列に変換する．その特徴語列から連続して出現する特徴語の順序対を取得したもののうち，正例の記事集合中に 2 回以上出現するものを素性として用いる．

2.2 根拠表現の獲得手法

前処理として抽出された景気動向記事から，根拠表現を自動的に獲得する．まず，**核文節**, **種 (たね) 表現**, **根拠表現候補**, **根拠表現文**, **根拠表現**という用語を以下のように定義する (核文節，種表現については図 1 を参照)．

種表現: 根拠表現候補が係る文節の先頭に，その根拠表現候補を構成する最後尾の文節に含まれる助詞や形式名詞「の」を追加した表現

核文節: 根拠表現候補を構成する文節の最後尾の文節から，助詞や形式名詞「の」を削除したもの

根拠表現候補: 核文節を拡張，縮約して得られた表現

根拠表現文: ある種表現と，種表現に係っている根拠表現候補

(注2)：形態素解析には"茶筌"を使用
http://chasen.naist.jp/hiki/ChaSen/

(注3)：正例の集合か負例の集合かを明示するために添字 p,n をそれぞれ付与している

(注4)：t_i は記事集合に存在する i 番目の単語を表す

が同時に含まれる文

根拠表現: 根拠表現文にある根拠表現候補と根拠表現に係る文節を含めた最長の表現

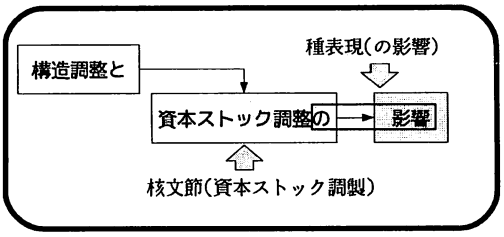

図 1　核文節と種表現の例

根拠表現抽出には酒井らの手法 [2] を拡張した手法を用いる．係り受け解析器には CaboCha(注5)を用いる．本手法の概要を以下に示す．

Step 1: 1 つの初期種表現「の影響」を人手で与え，それに係る根拠表現候補を獲得

Step 2: 獲得した根拠表現候補から，新たな種表現を獲得

Step 3: 獲得した種表現から，新たな根拠表現候補を獲得

Step 4: Step 2, 3 を予め定められた回数だけ繰り返す

Step 5: 獲得した根拠表現候補と種表現を用いて，根拠表現文を抽出

Step 6: 抽出された根拠表現文から，根拠表現候補と根拠表現候補に係る文節を含めた最長の表現を根拠表現として抽出

2.2.1　種表現に係る根拠表現候補の獲得

種表現に係る根拠表現候補は，次に示す「核文節取得」「拡張」「縮約」の 3 つの処理を順番に行うことで獲得される．

核文節取得: 種表現に係っている核文節を取得

拡張: 核文節に文節を追加して表現を拡張

縮約: 拡張された表現から不要な文節を除去して根拠表現候補を生成

2.2.2　縮　　約

Step 1: 核文節に文節を追加することで派生する表現を全て取得

Step 2: Step 1 で取得した各表現のスコアを計算

Step 3: 核文節から 2 回以上派生し，かつ，スコア最大の表現を根拠表現候補として獲得

Step 1 では，図 1 を例にあげて説明すると，核文節「資本ストック調整」と，文節「構造調整と」を核文節に追加することで派生する表現「構造調節と資本ストック調整」の両方を取得する．

Step 2 では，核文節 c から派生した各表現 e に対して，式 (5) で表されるスコアを計算する．

$$Score(e,c) = -pf(e)ef(e,c)\log_2 P(e,c) \tag{5}$$

$$P(e,c) = \frac{ef(e,c)}{Ne(c)} \tag{6}$$

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

ただし，

$P(e,c)$: 核文節 c から派生する表現 e の派生確率

$ef(e,c)$: 核文節 c から派生する表現 e の派生回数

$Ne(c)$: 核文節 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

2.2.3　根拠表現候補の選別

様々な種表現に係っている根拠表現候補が適切であるという仮定に基づき，根拠表現候補が種表現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根拠表現候補を選別する．根拠表現候補が種表現に係る確率に基づくエントロピーは式 (7) に示す．

$$H(e) = -\sum_{s \in Se} P(e,s)\log_2 P(e,s) \tag{7}$$

$$P(e,s) = \frac{f(e,s)}{\sum_{s \in Se} f(e,s)} \tag{8}$$

ただし，

$P(e,s)$: 根拠表現候補 e が種表現 s に係る確率

$f(e,s)$: 種表現 s に係る根拠表現候補 e の数

Se: 根拠表現候補 e が係る種表現の集合

閾値 T_e は，以下の式 (9) によって設定する．

$$T_e = \alpha \log_2 N_s \tag{9}$$

ただし，

N_s: 根拠表現候補を獲得するのに使用した種表現の異なり数

α: 定数 $(0 < \alpha < 1)$

$\log_2 N_s$ は，根拠表現候補が種表現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．ただし，初回は種表現の数が初期種表現「の影響」の 1 つのみなので，根拠表現候補のエントロピー，および，閾値が 0 になる．そのため，初回のみ全ての根拠表現候補が選別される．

2.2.4　種表現の獲得

抽出した根拠表現候補を含む文を抽出し，その中で根拠表現候補が係っている文節を獲得する．そして，その文節に対して，係っている根拠表現候補を構成する最後尾の文節に含まれる助詞や形式名詞「の」を追加し，それを種表現候補とする．そして，種表現候補が根拠表現候補によって係られる確率に基づくエントロピーを求め，ある閾値以上の種表現候補を種表現として抽出する．これは，適切な種表現には様々な根拠表現候補が係っているという仮定に基づく．種表現候補が根拠表現候補によって係られる確率に基づくエントロピーは式 (10) で求める．

$$H(s) = -\sum_{e \in Es} P(s,e)\log_2 P(s,e) \tag{10}$$

$$P(s,e) = \frac{f(s,e)}{\sum_{e \in Es} f(s,e)} \tag{11}$$

ただし，

$P(s,e)$: 種表現 s が根拠表現候補 e によって係られる確率

$f(s,e)$: 根拠表現候補 e によって係られる種表現 s の数

Es: 種表現 s に係る根拠表現候補の集合

閾値 T_s は，以下の式 (12) によって設定する．

$$T_s = \alpha \log_2 N_e \tag{12}$$

N_e は種表現を獲得するのに使用した根拠表現候補の異なり数である．また，定数 α は，根拠表現候補獲得の閾値を求めるときの定数と同じである．

2.2.5 根拠表現文と根拠表現の獲得手法

取得した根拠表現候補と種表現を用いて，根拠表現文を抽出する．抽出した根拠表現文に存在する根拠表現候補と根拠表現候補に係る文節を含めた最長の表現を根拠表現として抽出する．

日銀は二十一日発表した三月の金融経済月報で、国内景気について「輸出の減少を背景に、このところ足踏み状態になっている」との判断を示し、二〇〇〇年七月から続けてきた「緩やかな回復」を削除するなど、景況感を下方修正した

例えば，上記のような文があった場合，根拠表現候補取得では「輸出の減少」と「減少」の 2 つの表現を取得してしまう可能性がある．しかし，「減少」だけでは何を表しているか不明である．そこで，根拠表現候補と根拠表現候補に係る文節を含めた最長の表現を取得することで，意味の通る表現を取得することができる．この最長の表現を根拠表現とする．この例では，「輸出の減少」が根拠表現として抽出される．

2.3 根拠表現の分類

取得した根拠表現から Positive 表現と Negative 表現を分類する．根拠表現は Positive 表現, Negative 表現, Other 表現の 3 つのクラスに分ける．

Positive 表現は「同時拡大を続ける世界経済」や「公共・住宅投資の伸び」などの日本経済が回復することを示す表現である．

Negative 表現は「自動車、建築の不振」や「設備投資や個人消費の鈍化」などの日本経済が悪化することを示す表現である．

Other 表現は，「調査対象変更」や「同二・六%減とマイナス」，または「景気の伸び悩み」などの Positive 表現でも Negative 表現でもないものや，抽出ミスした表現である．上記にあるように「景気の伸び悩み」など，景気に関して直接的な表現は Other 表現とした．これは，本研究で抽出する対象は「景気動向の根拠」であり，「景気動向」自身ではないため，景気に関して直接的な表現は Other 表現とする．

2.3.1 分類方法

3 つの集合 (Positive 表現と Negative 表現と Other 表現の集合) から Positive 表現と Negative 表現を分類する手法を示す．本研究では one-versus-rest [3] をモデルとして用いる．

Positive 表現とそれ以外, Negative 表現とそれ以外の二つの分類器を作成する．one-versus-rest を用いることにより，根拠表現抽出の際のエラーを除去しつつ, Positive 表現と Negative 表現を分類することができる．

2.3.2 SVM に用いる素性

SVM に用いる素性として，文字 N-gram と特徴文字 N-gram を用いる．文字 N-gram, 特徴文字 N-gram ともに，N の値は 1〜3 を用いる．

特徴文字 N-gram : 特定の集合にのみよく表れる文字 N-gram.

酒井ら [2] が素性抽出で用いた式は，訓練データにおける正例, 負例の数を同数にしなければいけない．しかし，本研究では one-versus-rest を用いて多値分類を行うため，正例と負例の数を同数にすることはあまり好ましくない．なぜなら，実際に実験に用いるデータにも正例として分類される表現，負例として分類される表現が同数に入っているとは限らない．そこで，2 値分類を行う場合に正例, 負例を同数にしなくてもよい手法を提案する．

特徴文字 N-gram は全表現中に 2 回以上出現する文字 N-gram に対して，以下の式で重み付けを行ったのち，選択したものである．

$$W_c(t_i, S_c) = Ntf(t_i, S_c) \cdot e^{(1+Ndf(t_i, S_c))} \tag{13}$$

$$Ntf(t_i, S_c) = \frac{tf(t_i, S_c)}{\sum_{t \in Ts_{S_c}} tf(t, S_c)} \tag{14}$$

$$Ndf(t_i, S_c) = \frac{df(t_i, S_c)}{Ds_{S_c}} \tag{15}$$

$Ntf(t_i, S_c)$: 分類したい表現の集合 S_c(注6)における文字 N-gram t_i (注7)の出現確率

S_p : 訓練データにおいて分類したい表現の集合

$tf(t_i, S_c)$: 分類したい表現の集合 S_c に含まれる文字 N-gram t_i の出現回数

Ts_{S_c} : 分類したい表現の集合 S_c に含まれる文字 N-gram の集合

$Ndf(t_i, S_c)$: 分類したい表現の集合 S_c における文字 N-gram t_i が含まれる表現の出現確率

$df(t_i, S_c)$: 分類したい表現の集合 S_c における文字 N-gram t_i が含まれる表現数

Ds_{S_c} : 分類したい表現の集合 S_c における表現数

また，分類したい表現の場合と同様に，その他の表現に対して次の式 (16) を用いて重み付けを行う．

$$W_o(t_i, S_o) = Ntf(t_i, S_o) \cdot e^{(1+Ndf(t_i, S_o))} \tag{16}$$

ある文字 N-gram t_i の分類したい表現集合における重み $W_c(t_i, S_c)$ が分類したい表現以外の集合における重み $W_o(t_i, S_o)$ の 1.5 倍より大きければ，その文字 N-gram t_i を特徴文字 N-gram として選択する．また，逆にある文字 N-gram t_i の分類したい表現以外の集合における重み $W_o(t_i, S_o)$ が，分類したい表現集合における重み $W_c(t_i, S_c)$ の 1.5 倍より大きければ，その文字 N-gram t_i も特徴文字 N-gram として選択する．すなわち，以下の条件が成り立つ文字 N-gram t_i を特徴文字 N-gram として選択する．

$$W_c(t_i, S_c) > 1.5 W_o(t_i, S_o) \tag{17}$$

$$W_o(t_i, S_o) > 1.5 W_c(t_i, S_c) \tag{18}$$

(注6)：分類したい表現の集合か，それ以外の表現の集合かを明示するために添字 c,o をそれぞれ付与している

(注7)：t_i は表現集合に存在する i 番目の文字 N-gram を表す

DF 値を正規化した値を用いることにより，正例と負例を同数にしなければいけないという制約を解消した．また，W_e と W_o を 1.5 倍した理由は酒井ら [2] が用いた式 (4) と同様である．

3. 実　　験

日経新聞記事コーパスを用いて、以下の 3 つの実験を順に行った．

実験 1：正例として人手で収集した 1995 年から 1997 年の景気動向記事 400 と，負例として景気動向記事を除いた 1995 年から 1997 年の新聞記事集合からランダムに収集した記事 400 を用いて，景気動向記事抽出のための素性を抽出した．抽出した素性を用いて景気動向記事を学習し，2002 年から 2004 年の新聞記事のうち「景気」という単語のある 4000 記事に対して景気動向記事抽出を行った．

実験 2：1990 年から 2005 年までの 16 年間の新聞記事のうち「景気」という語を含んでいるものから，提案手法を用いて景気動向記事を抽出した．素性には特徴語 bi-gram を用いた．抽出した景気動向記事から根拠表現を抽出した．

実験 3：実験 2 で抽出した根拠表現を Positive 表現と Negative 表現に分類した．

分類器には，SVM^{Light}(注8) を用いた．カーネルは線形を使用した．

4. 実験結果

2002 年から 2004 年の新聞記事のうち「景気」という語を含んでいる 4000 記事を理系学生 4 人に読んでもらい，景気動向記事を選別してもらった．

抽出された素性を用いて景気動向記事抽出を行い，選別された正解データ 726 記事を用いて評価を行った．結果を表 1 に示す．評価方法は以下の通りである．

$$\text{適合率 (景気動向記事抽出)} = \frac{|Sd \cap Ad|}{|Sd|}$$

$$\text{再現率 (景気動向記事抽出)} = \frac{|Sd \cap Ad|}{|Ad|}$$

ただし，

Sd: 2002 年から 2004 年の新聞記事から，本手法を用いて抽出した景気動向記事を要素とする集合

Ad: 4000 記事に含まれる，人手で抽出した 726 個の景気動向記事を要素とする集合

表 1　景気動向記事抽出の結果

素性の種類	素性数	適合率 (%)	再現率 (%)
特徴語	1497	62.01	60.47
特徴語 bi-gram	2911	76.69	28.10

次に，抽出した景気動向記事からランダムに 100 記事を集め，そこから人手で 91 の根拠表現文と 75 の根拠表現を抽出した．人手で集めた正解データを用いて根拠表現抽出を評価した．評価方法は以下の通りである．

$$\text{適合率 (根拠表現抽出)} = \frac{|Sb \cap Ab|}{|Sb \cap Nb|}$$

$$\text{再現率 (根拠表現抽出)} = \frac{|Sb \cap Ab|}{|Ab|}$$

ただし，

Sb: 16 年分の新聞記事から抽出した根拠表現を要素とする集合

Ab: 100 個の景気動向記事に含まれる 75 個の根拠表現を要素とする集合

Nb: 100 個の景気動向記事に含まれる取得した種表現に係る表現を要素とする集合

α の値を 0.6，根拠表現獲得手法の繰り返し回数を 3 とし，根拠表現抽出を実行して抽出した根拠表現のいくつかを表 2 に示す．

表 2　抽出した根拠表現

構造調整と資本ストック調整，調整の波，
二十四時間体制のフル生産，同時拡大を続ける世界経済，
安値の輸入品，堅調な消費，大雪や寒波，
バブル崩壊に伴う資産デフレ，耐震強度偽装問題，
狂牛病問題，消費税率上げや特別減税廃止，
ＩＴ（情報技術）不況や米同時テロ，メーカーの減産・在庫調整，
好調な企業業績，円高、冷夏、長雨，原油高や大型ハリケーン，

根拠表現獲得手法の繰り返し回数を 3 としたときの，根拠表現抽出の適合率と再現率のグラフを図 2 に示す．

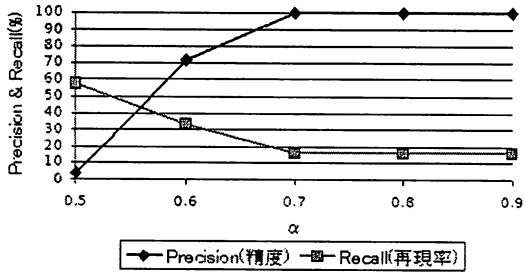

図 2　根拠表現抽出における適合率と再現率

1990 年から 2005 年の日経新聞から，α の値が 0.6，繰り返し回数が 3 回で根拠表現抽出プログラムを実行して抽出された，1620 の根拠表現を Positive 表現と Negative 表現を分類する実験に用いる．1620 の根拠表現には予め人手で，どの集合 (Positive 表現集合か Negative 表現集合か Other 表現集合) に属するかを示すラベルを付与しておく．1620 の根拠表現のうち，1120 の根拠表現を訓練データに，残りの 500 を実験データに用いた．

訓練データから，文字 N-gram と特徴文字 N-gram の 2 種類の素性を抽出した．文字 N-gram，特徴文字 N-gram それぞれについて，Positive 表現とそれ以外を分類する分類器と，Negative

(注8)：http://svmlight.joachims.org/

表現とそれ以外を分類する分類器の，合計 4 種類の分類器を作成した．

作成した 4 種類の分類器を用いて，実験データを分類し，その正解率，適合率，再現率，F 値を求めた．結果を表 3，表 4 に示す．正解率，適合率，再現率，F 値は以下の式で求められる．

$$\text{正解率 (根拠表現分類)} = \frac{|A|}{|S|} \tag{19}$$

$$\text{適合率 (根拠表現分類)} = \frac{|P \cap G|}{|P|} \tag{20}$$

$$\text{再現率 (根拠表現分類)} = \frac{|P \cap G|}{|G|} \tag{21}$$

$$F \text{ 値 (根拠表現分類)} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{22}$$

ただし，

S: 実験データを要素とする集合

A: 分類器により分類された表現の中で，正しく分類された表現を要素とする集合

P: 分類器により分類された表現を要素とする集合

G: 実験データ中の分類したい表現を要素とする集合

表 3　文字 N-gram 素性を用いた結果

	正解率 (%)	適合率 (%)	再現率 (%)	F 値	素性数
Posi	94.4	80.0	61.5	0.695	9021
Nega	83.2	84.3	85.5	0.849	9021

表 4　特徴文字 N-gram 素性を用いた場合

	正解率 (%)	適合率 (%)	再現率 (%)	F 値	素性数
Posi	94.6	79.1	65.4	0.716	2553
Nega	80.8	82.1	83.3	0.827	2456

5. 考　　察

表 1 より，特徴語 bi-gram を用いたほうが，単に特徴語を用いるより景気動向記事抽出の適合率を高めることができた．この結果から，特徴語 bi-gram が単に特徴語を用いたものより，景気動向記事に特化している素性だと考えられる．しかしながら，景気動向記事抽出の再現率は下がってしまった．これは，景気動向記事であるにもかかわらず，文章に景気動向記事に特化している素性が含まれていないことにより，景気動向記事と判断されないものが出現してくるためである．

図 2 より，α の値 0.5 と 0.6 を境に適合率が急激に低くなり，逆に再現率は高くなっている．閾値が下がることにより，種表現と根拠表現候補の中に，種表現や根拠表現候補として正しくない表現が出現してくる．さらに，繰り返し実行することで，正しくない表現から新たな種表現や根拠表現候補として正しくない表現が取得される．結果として，大量の正しくない表現を取得してしまう．この現象が α の値 0.5 と 0.6 の間で顕著に表れる．

表 3，表 4 より，再現率に関しては Positive 表現分類と Negative 表現分類で大きな差がでた．これは，用いたコーパスの時期の景気が悪かったため，訓練データと実験データに含まれる Positive 表現の数が，Negative 表現と Other 表現の数に比べて，圧倒的に少ないためだと考えることができる．また，Positive 表現であるのに，Positive 表現以外だと判断された表現を見てみると，Other 表現や Negative 表現に多く含まれるような語を持つ表現が多数あった．例えば，「米景気回復」は「景気」という語が入っているため，Other 表現と判断された．

表 3，表 4 より，特徴文字 N-gram を用いることで，素性の数を大幅に減らすことができた．素性数を大幅に減らしたのにもかかわらず，文字 N-gram を用いたときの結果と大差のない結果を得ることができた．これは，式 (17), (18) を用いることで不要な文字だけをうまく削除することができたと考えられる．

表 3，表 4 より，適合率をみると，文字 N-gram を用いた場合が，一番良い結果を得ることができた．今回行った実験では，短い表現を対象としたので，文章に比べて単語数が少ない．さらに，今回用いた記事コーパスは英語の文章集合ではなく，日本語の文章集合である．そのため，日本語を対象として，表現を分析する場合は漢字の存在に起因して，文字一つ一つが重要になってくる．よって，式により文字を選別した特徴文字 N-gram を用いるより，文字 N-gram を用いた方が情報量が多いため結果が良くなったと考えられる(注9)．これにより，今回の実験では文字 N-gram を素性に用いた学習器が良い結果を得られることが分かった．

6. まとめと今後の課題

本研究では，新聞記事から根拠表現を抽出，分析する手法を提案した．評価実験の結果，根拠表現抽出手法に関しては適合率 71.43%，根拠表現分類手法に関しては Positive 表現分類の F 値 0.695，Negative 表現分類の F 値 0.849 を達成した．

今後の課題は，本手法と類似した Theresa ら [4] の手法との比較実験を行う．また，根拠表現を用いた景気予測についても考察していく．

文　　献

[1] 中嶋 琢美, 増山 繁, 酒井 浩之, SVM(Support Vector Machine) を用いた経済記事の著者の見解に基づく分類，情報処理学会研究報告「情報学基礎」 No.071-21, 2003.

[2] 酒井 浩之, 梅村 祥之, 増山 繁, 交通事故事例に含まれる事故根拠表現候補の新聞記事からの抽出，自然言語処理，vol.13, No.4, pp.99-124, 2006.

[3] 山田 寛康, 松本 裕治, Support Vector Machine の多値分類問題への適用法について，情報処理学会研究報告，2001-NL-146, pp.33-38, 2001.

[4] Theresa Wilson, Janyce Wiebe and Paul Hoffmann, Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, In *Proceedings of HLT/EMNLP-05*, pp.347-354, 2005.

[5] V.N. Vapnik, Statistical Learning Theory, John Wiley & Sons, 1999.

(注9)：形態素 N-gram でも実験を行い，Positive 表現分類の F 値 0.716，Negative 表現分類の F 値 0.845 という結果を得ている．結果より，表現から文字 N-gram を取得したときの情報量より，形態素 N-gram を取得したときの情報量が少ないため，文字 N-gram のほうが結果が良かったと考えている．