経済新聞記事から抽出した景気動向を示す根拠表現への極性付与手法の提案

谷口 将太 (豊橋技術科学大学　taniguchi@smlab.tutkie.tut.ac.jp)
坂地 泰紀 (豊橋技術科学大学　sakaji@smlab.tutkie.tut.ac.jp)
酒井 浩之 (豊橋技術科学大学　sakai@smlab.tutkie.tut.ac.jp)
増山 繁 (豊橋技術科学大学　masuyama@tutkie.tut.ac.jp)

1 はじめに

企業や投資家にとって，商品の売行きや株価を予測するために，景気動向を知ることは重要である．景気動向を知るための指標としては，内閣府が毎月発表する景気動向指数[*1]がある．景気動向指数は，生産，雇用など様々な経済生活において重要，かつ，景気に敏感な指数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景気指標である．景気動向指数は各指数を 3 ヶ月前の値と比較した結果から求められているため，景気動向を知ることはできるが，景気動向指数を用いて景気の予測を行う場合は直近の景気動向が反映されておらず，正確な予測を行うことが難しい [1].

Sakaji et al[1] は景気の動向に関する記述のある記事 (以降，景気動向記事と略記) から景気動向を示す根拠となる表現 (以降，根拠表現と略記) を抽出し，抽出した根拠表現が，景気の回復を示唆する (以降，極性が Positive と定義) 根拠表現なのか，景気の悪化を示唆する (以降，極性が Negative と定義) 根拠表現なのかを分類する手法を提案した．根拠表現とは，「米国景気悪化」や「公共投資の伸び」などを指し，「米国景気悪化」は極性が Negative，「公共投資の伸び」は極性が Positive な根拠表現となる．この手法を用いることによって，2 つの極性に分類された根拠表現を得ることができる。分類された根拠表現の各極性毎の総数を比較することによって，景気動向予測の材料として用いることができると考えられる．しかしながら，Sakaji et al[1] の極性付与の手法では，Support Vector Machines(SVM) を用いており，学習に用いた素性を含まない根拠表現への極性付与を行うことができない．

そこで, Sakaji et al[1] の手法で抽出した根拠表現に対し, 極性付与の際に機械学習を用いないで極性 (”Positive”,”Negative”) を付与する手法を提案する．

2 関連研究

Sakai et al[2] は，企業の業績発表記事より抽出した業績要因 (例えば，「自動車の売上が好調だった) に対し，業績向上を示す要因に Positive，業績悪化を示す要因に Negative を付与する研究を行っている．Sakai et al は，業績発表記事を極性 (業績が向上する内容の記事は Positive，悪化する内容の記事は Negative) に分類し，分類された業績発表記事集合における共通頻出表現 (例えば，「売上」) と手がかり表現 (例えば，「が好調」) の組み合わせの頻度分布を使用することで，業績要因への極性付与を行う手法を提案している．しかし，経済新聞の景気動向記事では，1 つの記事内に景気の向上を示す内容と景気の悪化を示す内容が混在しているため，記事を景気動向の極性に基づいて分類することは難しい．それに対して本手法では，景気動向記事を段落に分割することで対応する．なぜなら，段落内における極性は，統一されていることが多く，より正確に極性分類が行えるからである．

高村ら [3] は 2 つの単語 (名詞 + 形容詞 or 形容動詞) から成る表現に対して，隠れ変数モデルに基づき，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．この手法では，学習データに出現しない単語からなる複数語表現に対して分類を行うことができない．本手法では，景気動向記事の各段落をあらかじめ極性分類し，その情報を使用して根拠表現への極性付与を行っているため，極性付与の際には学習データを必要としていない．そのため，このような問題は起こらない．

Turney et al[4] は，ウェブの検索エンジンを用いて，複数語表現と”excellent”や”poor”といった極性が明確な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．本手法で扱う根拠表現中の表現は，根拠表現中に共起する表現によって極性が変わるため，極性が既知な表現は少ない (例えば，「企業倒産の減少」の「減少」). そのため Turney et al[4] の手法を本タスクに適応することはできない．

Kaji et al[5] は，人手で作成した手がかり表現リストやパタン，規則から評価文を抽出して極性 (好評，不評) を付与し，各極性中に評価表現 (名詞 + 格助詞 + 形容詞) が出現する頻度から評価表現の極性を決定する手法を提案している．この手法では，同一の評価表現が評価文集合中に 3 回以上出現する必要がある．文字数の少ない根拠表現や抽象的な根拠表現であれば，景気動向記事集合中に 3 回以上出現するが，景気動向記事集合中に 1 回や 2 回しか出現しない根拠表現も数多く存在する．そのため，Kaji et al[5] の手法を本タスクに適応することはできない．本手法では，根拠表現を，部分根拠表現を構成要素とするベクトルに置き換えている．部分根拠表現は，根拠表現の核の部分であり，比較的文字数は少なく，抽象的な表現となっているため出現回数の少ない根拠表現に対しても極性付与を行うことができる．

3 提案手法

3.1 提案手法の概要

本手法では根拠表現に含まれる部分根拠表現 (文献 [1] の”Frequent phrase candidates”) に着目し，極性付与を行っている．

[*1] http://www.esri.cao.go.jp/jp/stat/di/di.html

部分根拠表現とは，景気動向記事から根拠表現を抽出する際に獲得する，根拠表現の核となる表現のことである．Sakaji et al[1] の根拠表現抽出の手法では，手がかりとなる形態素列である clue phase(例えば「を背景に,」) にかかる文節を取得する．取得した文節に係る文節をつなげることで根拠表現を抽出している．clue phase に係る文節のことを本手法では部分根拠表現と定義する (図 1 参照)．例えば，根拠表現「米国景気の悪化」には，「景気」,「悪化」という部分根拠表現が含まれ，根拠表現「設備投資などの伸び」には，「設備投資」,「伸び」という部分根拠表現が含まれている．「悪化」という部分根拠表現は，Negative な極性の根拠表現に出現する可能性が高く，「伸び」という部分根拠表現は Positive な極性の根拠表現に出現する可能性が高い．また，「米国景気」や「設備投資」といった部分根拠表現は，Positive な極性と Negative な極性のどちらの根拠表現にもよく出現する．そのため，まず景気動向記事を段落に分割し，SVM を用いて景気動向に対する極性で分類する．この景気動向の極性に基づき，極性 (”Positive”,”Negative”) 分類された景気動向記事の段落集合における語の出現頻度を用いることで根拠表現への極性付与を行う．

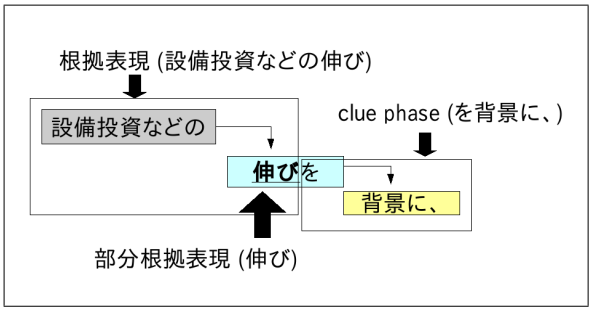

図 1　部分根拠表現の例

3.2　段落の極性に基づく分類

根拠表現への極性付与の前処理として，景気動向記事の段落を極性分類する手法を示す．

まず，景気動向記事を，記事を構成する段落に分割する．各段落を，Positive，Negative，Other の 3 つのクラスに分類する．景気が回復することを示唆する内容の段落は Positive に，景気が悪化することを示唆する内容の段落は Negative に，景気動向に関係ない内容の段落，もしくは Positive，Negative の両方を含む段落は Other に分類する．分類には one-versus-rest[6] をモデルとして用いる．

SVM を用いて，Positive かそれ以外，Negative かそれ以外にそれぞれ分類する 2 種類の分類器を作成する．one-versus-rest を用いることによって，Positive な極性，Negative な極性が入り交じっている段落を除去し，Positive な段落の集合と Negative な段落の集合に分類することができる．

SVM の素性は，人手で極性を付与した訓練データ中の文字 N-gram から選択する．N の値には 1〜3 を用いる．作成した文字列の中から，訓練データ中に 2 回以上出現する文字列を素性候補として重み付けし，素性を選択する．重み付けには文献 [1] の方法を用いる．重み付けの式を以下に示す．

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

$$Ntf(t_i, S_c) = \frac{tf(t_i, S_c)}{\sum_{t \in Ts_{S_c}} tf(t, S_c)} \tag{2}$$

$$Ndf(t_i, S_c) = \frac{df(t_i, S_c)}{Ds_{S_c}} \tag{3}$$

$Ntf(t_i, S_c)$:　分類したい極性の段落集合 S_c における素性候補 t_i の出現確率

S_c:　訓練データにおいて分類したい極性の段落集合

$tf(t_i, S_c)$:　分類したい極性の段落集合 S_c に含まれる素性候補 t_i の出現回数

Ts_{S_c}:　分類したい極性の段落集合 S_c に含まれる素性候補の集合

$Ndf(t_i, S_c)$:　分類したい極性の段落集合 S_c における素性候補 t_i が含まれる段落の出現確率

$df(t_i, S_c)$:　分類したい極性の段落集合 S_c における素性候補 t_i が含まれる段落数

Ds_{S_c}:　分類したい極性の段落集合 S_c における段落数

また，分類したい表現の場合と同様に，それ以外の段落に対して次式 (4) を用いて重み付けを行う．

$$V_o(t_i, S_o) = Ntf(t_i, S_o) \cdot e^{(1+Ndf(t_i, S_o))} \tag{4}$$

全ての素性候補に対して，式 (1)，式 (4) を計算し，以下の式 (5) または，式 (6) を満たす素性候補を，極性 c[*2]かそれ以外に分類する分類器の素性として使用する．これにより，分類したい極性の段落集合，または，それ以外の段落集合中に万遍なく，かつ，高頻度で出現する語を選択することができる．つまり，各段落集合を特徴付ける語を素性として選択することができる．

$$V_c(t_i, S_c) > 2.0 V_o(t_i, S_o) \tag{5}$$

$$V_o(t_i, S_o) > 2.0 V_c(t_i, S_c) \tag{6}$$

Positive かそれ以外，Negative かそれ以外のそれぞれについて素性を選択し，分類器を作成する．景気動向記事の段落を 2 つの分類器にかけ，Positive かそれ以外に分類する分類器において Positive，Negative かそれ以外に分類する分類器においてそれ以外と判断された段落を Positive 段落とする．また，Positive かそれ以外に分類する分類器においてそれ以外，Negative かそれ以外に分類する分類器において Negative と判断された段落を Negative 段落とする．その他の段落を Other 段落とする．Other 段落は極性付与の際には使用しない．

3.3　根拠表現への極性付与

根拠表現への極性付与には，Sakaji et al[1] の手法で抽出された，根拠表現と部分根拠表現を用いる．

まず，3.2 で極性分類した段落の Positive な段落集合と，Negative な段落集合における各根拠表現の出現確率をそれぞれ計算する．ここで，記事中には周辺の語と組み合わせることで極性が反転して出現する根拠表現が存在するため，ルールを与えて対応する．例えば，根拠表現「在庫調整」は Negative 極性であるが，「在庫調整がほぼ終了」のように，「終了」と共に出現することで Positive 極性となり，Positive 段落に多く出現する．それらの根拠表現の出現頻度カウント時には，極性を反転させ

[*2] $c \in \{Positive,\ Negative\}$

る原因の語 (「終了」など) を予め人手で求めておき，表として与え, その根拠表現が出現した段落に与えられた極性とは逆の極性の出現頻度にカウントする．例えば，「在庫調整がほぼ終了」という文が Positive 極性の段落に出現した場合，根拠表現「在校調整」は，Negative 段落に出現したものとし，Negative 段落の出現回数を 1 つ増やす．表として与えている語を表 1，表 2，表 3 に示す．

表 1　Negative な根拠表現の後ろに付き，Positive にする語

完了，終了，進展，峠を，一巡，あるが，こなせば，吸収，懸念，前の駆け込み需要，景気回復の足取りが鈍い，終息，リスク，慎重さ，歯止め

表 2　Positive な根拠表現の後ろに付き，Negative にする語

消えた，期待，が必要，するだろう，下げている，期待，が必要，低くな，過去最低，低下，下落，戻さないと，継続，安定する，追い風，に救われ，下支え，求める，要請，に結びついていない，が不可欠，はあるものの

表 3　文の先頭に付き，極性を反転させる語

だが，ただ，それでも，しかし，一方，それどころか，ものの，そうして，もっとも，が

極性分類した段落における根拠表現の出現確率だけで極性付与を行った場合，段落中に出現しない根拠表現に対して極性付与を行うことができない．そこで，部分根拠表現を用いる．根拠表現は，1 つ，ないし，複数の部分根拠表現を含んでいる (例えば，根拠表現「米経済の悪化」は，部分根拠表現「経済」「悪化」を含む)．そのため，根拠表現 e のベクトル表現 $\boldsymbol{x}_e$ を，部分根拠表現集合 $cp = \{cp_1, cp_2, \cdots, cp_n\}$ を用いて，

$$\boldsymbol{x}_e = (x_1^e, x_2^e, \cdots, x_n^n), \boldsymbol{x}_i^e = \begin{cases} 1, & e \text{ が } cp_i \text{を含む時} \\ 0, & \text{それ以外} \end{cases} \quad (7)$$

と定義することができる．ここで，極性を $c \in \{Positive, Negative\}$ と定義すると，式 (8) で根拠表現 e の極性を推定することができる．

$$\hat{c} = \underset{c}{\operatorname{argmax}} P(c|\boldsymbol{x}_e) = \underset{c}{\operatorname{argmax}} P(c)P(\boldsymbol{x}_e|c) \quad (8)$$

また，$\boldsymbol{x}_e$ の各要素 x_i^e が独立に生起すると仮定し，条件付確率 $P(\boldsymbol{x}_e|c)$ を以下の式で推定する．

$$P(\boldsymbol{x}_e|c) \approx \prod_{i=1}^{n} \frac{P(x_i^e, c_d)}{P(c_d)} \quad (9)$$

ただし，$c_d \in \{Positive, Negative\}$ を，景気動向記事の段落が分類された極性と定義し，$P(x_i^e, c_d)$ を，c_d に分類された景気動向記事の段落において，cp_i を含む根拠表現が出現する確率，$P(c_d)$ を，各部分根拠表現 $cp_i (i = 1, 2, \cdots, n)$ を含む根拠表現が，c_d に分類された景気動向記事の段落において出現する確率とする．ナイーブベイズに基づき，極性を推定した値 $\hat{c}$ と，各段落集合中の根拠表現の出現確率 $P(c, e)$ との積を，重み $W(c, \boldsymbol{x}_e)$ とし，以下の式を計算することにより，極性付与を行う．

$$W(p, \boldsymbol{x}_e) > 2W(n, \boldsymbol{x}_e) \quad (10)$$

$$W(n, \boldsymbol{x}_{\acute{e}}) > 2W(p, \boldsymbol{x}_{\acute{e}}) \quad (11)$$

式 (10) を満たす根拠表現には Positive の極性を，式 (11) を満たす根拠表現には Negative の極性を付与する．どちらの式も満たさない根拠表現は，景気動向予測の根拠として適切ではないとし，極性付与を行わない．

4　評価実験

Sakaji et al[1] の手法で抽出された景気動向記事 (29,525 件)，根拠表現 (1620 件)，部分根拠表現 (273 件) を用いて，以下の 2 つの実験を行った．景気動向記事の抽出は，1990 年から 2005 年までの 16 年間の新聞記事から行っている．

4.1　実験 1 景気動向記事分類

3.2 の手法について実験を行った．実装にあたり，形態素解析器として MeCab[*3]を使用した．また，SVM^{light}[*4]を使用し，カーネルは線形カーネルを使用した．

景気動向記事 29,525 件からランダムに選択した 1,000 件の記事を用いて以下の 3 つについて実験を行った．

パターン 1　景気動向記事 1,000 件の中から 100 件の記事に対して人手で極性を付与．60 記事を学習データ，40 記事をテストデータとして，極性に基づき分類

パターン 2　記事を段落に分割し，400 段落に人手で極性を付与．340 段落を学習データ，60 記事をテストデータとして極性に基づき分類

パターン 3　記事を文に分割し，600 文に人手で極性を付与．500 文を学習データ，100 文をテストデータとして極性に基づき分類

4.2　実験 2 根拠表現への極性付与

景気動向記事 29,525 件を段落に分割し極性分類した結果，Positive な極性の段落が 12,205 段落, Negative な段落が 34,418 段落となった．このうち，Positive，Negative ともに 12,205 段落を用いて，3.3 の手法について実験を行った．人手で極性を付与した根拠表現 1,000 件 (根拠表現として不適切な表現も含まれている) を正解データとして用いた．

また，比較手法として以下の手法による極性付与を行った．

1. ナイーブベイズによる極性の推定値のみを使用した手法 (Sakai et al[2] の手法)
2. 各極性の段落における各根拠表現の出現確率を重みとした手法

5　実験結果

5.1　景気動向記事分類の実験結果

景気動向記事について，記事，段落，文，それぞれの分類を行い，極性 (Positive,Negative,Other) ごとの精度，再現率を計算した結果を表 4，表 5，表 6 に示す．

表 4　実験 1 Positive 極性の結果 (単位：%)

	精度	再現率
パターン 1(記事)	50.0	7.1
パターン 2(段落)	90.0	33.3
パターン 3(文)	42.8	13.0

[*3] http://mecab.sourceforge.net/
[*4] http://svmlight.joachims.org

表 5　実験 1 Negative 極性の結果 (単位：%)

	精度	再現率
パターン 1(記事)	0.0	0.0
パターン 2(段落)	90.9	66.6
パターン 3(文)	71.4	25.6

表 6　実験 1 Other 極性の結果 (単位：%)

	精度	再現率
パターン 1(記事)	28.9	100
パターン 2(段落)	41.0	88.8
パターン 3(文)	40.5	94.1

表 5 において，Negative な記事の分類結果が 0 となっているのは，正解データでは Negative 以外に判定されている記事を 1 つだけ Negative だと判定したからである．

5.2　根拠表現への極性付与結果

本手法を用いて根拠表現への極性付与を行い，精度，再現率，F 値を計算した結果を表 7，表 8 に示す．

表 7　実験 2，Positive な根拠表現の結果（単位：%）

	精度	再現率	F 値
本手法	58.5	54.8	56.6
比較手法 1	54.5	25.0	34.2
比較手法 2	61.2	20.8	31.0

表 8　実験 2，Negative な根拠表現の結果（単位：%）

	精度	再現率	F 値
本手法	67.8	75.8	71.6
比較手法 1	71.0	71.6	71.3
比較手法 2	69.6	20.3	31.4

6　考察

表 4，表 5，表 6 より，記事，段落，文，それぞれの分類において，段落に分割した場合が最も精度よく分類することができた．記事単位での分類では，記事の内容が景気回復を示唆していたとしても，部分的に景気悪化を示唆していることが多かったため，うまく分類することができなかった．また，文単位での分類でも，含まれる素性の数が極端に少なかったため，うまく分類することができなかった．

表 7，表 8 より，比較手法と比べて，本手法の方がよい結果が得られた．出現確率のみによる極性付与の結果が悪かったのは，段落集合内に出現しない根拠表現に対して極性付与できないことが原因である．しかし，出現確率のみによる極性付与においても，精度は高い．これは，極性付与の前準備として段落に分割して極性分類を行った結果，精度が 90% と高かったためであると考えられる．ナイーブベイズによる手法よりも本手法の方が良い結果だったのも，これが原因であると考えられる．

Sakaji et al[1] の手法において，極性付与できなかった根拠表現に対して，本手法では極性付与できていることが確認できた．しかし，逆に，極性付与が行えていた根拠表現に対して，本手法では極性付与できなくなっているものも存在した．極性付与できなかった原因としては，根拠表現の極性と逆の極性を持つ部分根拠表現が含まれ，その部分根拠表現が悪影響を与えているものがあった．例えば，根拠表現「企業倒産の減少」は，Positive 極性であるが，Negative 極性の部分根拠表現「減少」が含まれているため Negative 極性だと判断された．

本手法の性能低下には，根拠表現として不適切な表現が大きく影響している．例えば，根拠表現「はっきりしたマイナス」は，景気動向を示す根拠としては不適切である．しかし，部分根拠表現「マイナス」を含んでいるためネガティブであると判断された．

7　まとめと今後の課題

本研究では，景気動向予測のため，新聞記事から抽出した根拠表現に対して，極性を付与を行う手法を提案した．評価実験の結果，極性付与の精度は，ポジティブで 58.5%，ネガティブで 67.8%，再現率は，ポジティブで 54.8%，ネガティブで 75.8% であった．景気動向予測の材料とするにはさらなる性能の向上が必要な結果となった．

本手法の性能低下には，部分根拠表現が大きく影響していることが分かった．そのため，今後の課題としては，部分根拠表現の代わりをなす表現の検討が挙げられる．また，根拠表現として不適切な根拠表現を取り除く手法が必要であると考えられる．

3.3 において，人手で与えているルールを自動的に与える手法も検討する必要があると考える．

参考文献

[1] Hiroki Sakaji, Hiroyuki Sakai, Shigeru Masuyama, "Automatic Extraction of Basis Expressions That Indicate Economi Trends", The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp977-984(2008).

[2] Hiroyuki Sakai, Shigeru Masuyama, "Assigning Polarity to Causal Information in Financial Articles on Business Performance of Companies", IEICE Trans. Information and Systems, vol.E92-D, no.12(2009).

[3] 高村 大地, 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).

[4] Peter D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", in Proceedings of the ACL2002, pp.417-424(2002).

[5] Nobuhiro Kaji, Masaru Kitsuregawa, "Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents, in Proceedings of the EMNLP-CoNLL 2007, pp.1075-1083(2007).

[6] 山田 寛康, 松本 祐治,「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, pp33-38(2001)