評価表現を利用した高認知度情報の抽出

落谷 亮, 西野 文人
富士通研究所
ochi@flab.fujitsu.co.jp, nisino@flab.fujitsu.co.jp

1 はじめに

新聞や雑誌の記事等のように幅広い読者を対象とした文章では、読者の理解を助けるための表現として、「有名な」、「注目されている」、「知られている」のような記述対象の新しさや重要度、普及の度合いなどの評価を示す表現が使われる。

これらの評価表現は、(1) 読者にも周知の情報を示し提示する情報への親近感を増す、(2) 読者が知らない可能性が高い特定の分野やコミュニティで知られている情報を示し理解を助ける、といった機能を持つと考えられる。(1) のように用いられる場合には周知の事柄を思い起こさせるような情報が示され、(2) の場合は未知の事柄の概要となる情報が示されていると推測できる。

いずれの場合でも、評価表現により概要情報とその情報の認知度に関する簡単な評価が提示されるので、その表現パターンの定型性が利用できれば、高い認知度で評価されている人物や事柄の情報だけをテキスト中から抽出することができると考えられる。

この報告では、新聞記事中に見られる評価表現に対するパターンを用いた抽出処理について述べ、その抽出の結果得られた情報の性質について、文章中の語句出現状況との比較を行うことにより議論する。

2 高認知度情報の抽出

「何が流行っているのか？」、「どういうことに世間の人が関心があるのか？」というような情報は、特に流通や産業でのニーズを取り立てなくとも、多くの人が関心を持つ情報であろう。

このような記事等から得ることを考えると、以下の方法が考えられる。

1. 語句の出現頻度を直接的に調べる定量的手法

   特定の人、物、イベント等の名称などが、どの程度頻繁にテキスト中に現れるかを数える。

2. 語句の表現内容から間接的に推定する手法

   記事の著者の判断を表す表現から推測する。

我々の研究グループでは、企業、商品、人物やそれらに関する事実情報（企業合併・創立、製品の新発売、人物のエピソードなど）を抽出するシステムを開発してきた[西野]。

開発した抽出システムを使うことにより、長期間に渡る大量の新聞記事を対象として抽出を行い、結果として得られた物事や事象の情報の多さを単純に頻度などで比較したり、年ごと、月ごとなどの期間ごとに、それぞれの事柄や事象の出現を集計し、ある特定の事柄についての情報が時間の経過を追って、どう変化するかを調べることなども可能になってきた。

しかし、このような集計処理で得られる情報は、対象となる事柄が記事に出現している度合いを数えているだけであるので、「ある時期に頻繁に記事に出てきている事柄は、その時期に何か特徴的な事が起きている」程度の指標でしかない。世の中のあらゆる事柄が均等に新聞に取り上げられているわけでは無いし、記事の表現に含まれる語句の頻度が語句の表す事柄の重要さに比例するわけでもない。従って、記事表現から抽出した情報が均等な価値を持つと判断したのでは、現実の世界での状況と合わない分析結果が得られる可能性も高い。

そこで、先の方法 2のように、記事中に現れる表現で記者の現実世界の事実に対する評価や判断が現れている部分の情報を集めることが出来れば、機械的に語彙の出現を数えるよりもより、より現実に沿った現実世界の情報が得られるのではないかと考えた。

新聞記事では幅広い読者を対象とするため、記述対象の紹介、確認等の目的で、「～として知られる」、「～として著名な」、「評判の～」など、世間一般での認知度を評価するような表現が頻繁に用いられる。これら認知度の評価を示す表現のうち、単に評価対象Xに対する評価を述べるもの（「注目されているX」、「有名なX」等）、評価対象Xに関連する情報Aまで示すもの（「Aとして注目のX」、「XがAとして注目されている」）の定型パターンによる抽出処理を試みた。

3 抽出パターン

評価表現に例に相当する表現パターンを多く集めるため、以下の手順で顕著な特徴を持つ表現から初めて、荒い条件による抽出処理を実験的に開発し、その抽出結果から得られる表現により段階的にパターンを広げてめて行く作業を行い、記事テキスト中の評価表現を抽出する処理を開発した。

1. 一般に認知されていることを示す表現として、「～とされる」、「として～される」、「として～られる」などの受け身の表現が特徴的に現れるので、これらの表現持った文を集める。次に「～」の部分に高頻度で出現する語句から評価にあたる語句を手作業で選ぶ。これにより表1に示すような語（評価語と呼ぶ）が集まる。

2. 次に先のパターンで「～」部に、選んだ評価語を埋め込みもう一度記事の文を選び直す。この文の集合を対象に、評価の対象やその属性を表す文パターンを集め抽出規則化する。この作業により、「XはAとして注目され」、「XがAとして知られ」などの、助詞「は」、「が」による評価対象情報の表現パターンや「Aとして知られるX」等の連体修飾形の表現パターンに対する抽出処理を作成する。

3. 前のステップのパターンで得られたA、Xの組を用いて、「Aとして～のX」のような「評価語」＋「の」＋「評価対象」のパターンに対する抽出処理を作成する。

4. 以上のパターンでの処理により抽出すると、「注目の的」といった表現から「的」のような語(表2) が評価対象語として沢山得られるので、こ

表 1: 評価語の例

注目、期待、人気、流行、流行り、重視、...

表 2: 高頻度対象語の例

的、きっかけ、秘密、背景、一因、...

のような高頻度の語に対して、これらの語で実際に指される対象語を抽出する「Xは注目の的」のようなパターンの処理を追加する。

日本経済新聞、毎日新聞の記事データ ([毎日],[日経]) を対象とし先のパターンによる抽出を行った。

実際の処理としては、抽出処理の簡略化のため、「話題」、「注目」、「知られる」などの表現毎に分けて開発したので、抽出の前処理として、全文検索でそれぞれの表現を含むテキストデータだけを選び抽出を行った。

例えば、「話題の」を含む列に対する抽出では、検索の結果選ばれた抽出対象は日経、毎日合わせて2653 文であり、抽出処理の結果は 1747 件が得られている。抽出件数の多さからも判るように、今回作成した抽出処理は比較的緩い制約条件により抽出を行っており、抽出結果は「新顔」、「中心」などの意味の無い対象や誤りも多く含んでいる。

このような誤りや無意味な抽出結果を結果をそのまま利用することは難しいので、得られた評価対象情報が人物かどうかの判定、事件名などの判定処理で後処理とする必要があり、抽出規則の絞り込み条件の改良と合わせて今後の課題と考えている。

4 抽出結果の性質

ここでは、抽出により得られた評価対象語の出現状況と、その対象語と同じ語の出現状況が、どのように異なるかの比較について述べる。本来は抽出が出来た語全体での傾向を示すのが望ましいと考えられるが、今回は幾つかの語で選んでの比較だけを行う。

評価表現で抽出された対象情報の中から幾つかの語を選び、特定の期間毎（月、年）に評価表現の抽出頻度と、その評価対象語と同じ単語が出現頻度を集計しグラフ化した。

図 1に示したグラフは、「タイガー・ウッズ」に関する評価表現「注目」（「注目のタイガー・ウッ

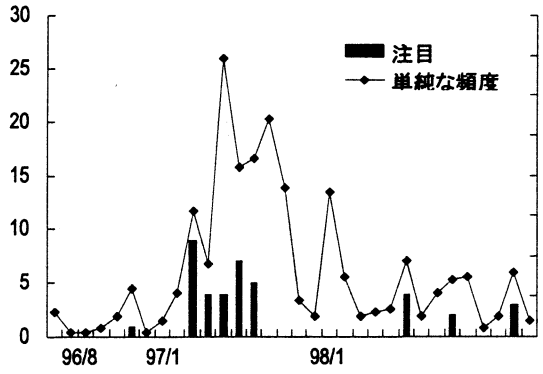

図 1: 「タイガー・ウッズ」

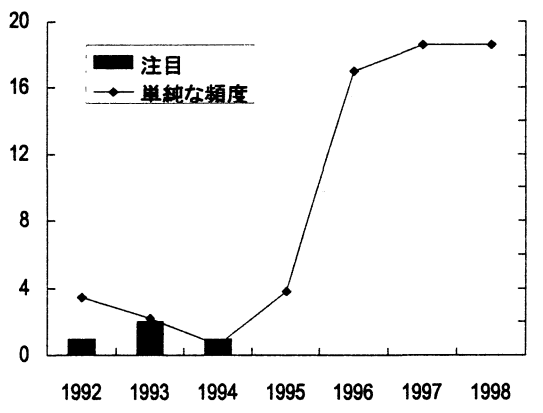

図 2: 「福島晃子」

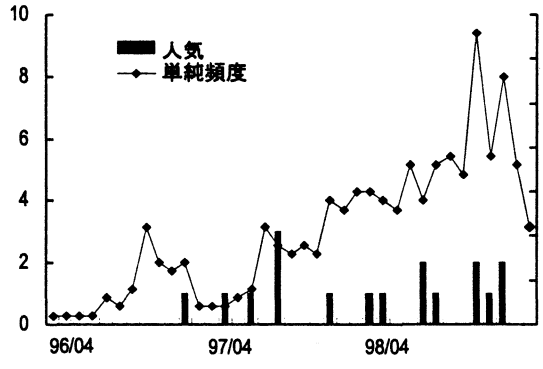

図 3: 「ガーデニング」

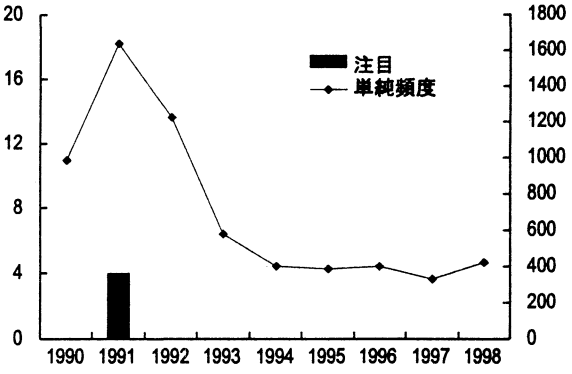

図 4: 「領土問題」

ズは」、「注目されるタイガー・ウッズは」）の出現頻度（棒グラフ）と文字列（「タイガー・ウッズ」）の出現頻度（線グラフ）を示している。（全てのグラフで左側の縦軸が評価表現の頻度、右側の縦軸が条件無しの場合の単語の出現頻度を表している。集計期間は月単位。）

グラフからは、「注目」の評価を示す棒グラフが96年終り頃に集中しており、「タイガー・ウッズ」が、この頃、ちょうど注目されはじめたことがグラフから読み取れる。同様に、図 2は「福島晃子」のグラフを示す（情報が少ないため集計は年単位）。この場合は、注目表現は条件無しの単語出現頻度の上がらない時期に集中している。

このように「注目」評価では、折れ線で示された無条件の単語の出現よりも、棒グラフで示された「注目」評価の方が左側（早い時期）に集まる傾向があり、「注目」評価が新しい事柄や人物の登場の早い時期に出現する傾向がある。

また、人物の場合には、最初に登場した頃に「注目」され、その後、知名度が上がるにつれて「注目」の頻度は減る傾向が見られる。但し、「タイガー・ウッズ」のように試合等のイベントで注目される場合や、「大統領選挙」の行方などのイベントは繰り返し注目される。

図 3は、人物以外の情報への評価として、「ガーデニング」で評価が「人気」の場合の推移を示している（月単位集計）。このように「人気」のような評価も比較的早い時期に集まることが多い。

図 4は「領土問題」で「注目」評価の場合の推移を示している（年単位集計）。このような情報は、

評価表現でも単純な頻度もピークは同じになるが、評価表現に因るものは、どの時期一番注目されたのを推定する判断材料とできると考えられる。

無条件の出現頻度推移でも全ての事柄の頻度推移をみることも不可能ではないが、その中から社会的に注目された情報だけを選ぶ手掛かりとして、評価表現により選ばれた情報が有効である考ともえられる。

5 まとめ

評価表現パターンを利用して、新聞記事データから認知度の高い情報だけを抽出する試みと、抽出結果として得られた情報の性質について単語の出現傾向との比較により議論した。

先のグラフで示したように、高認知度の評価を伴った情報は、無条件の単語の出現傾向よりは時系列上の早い時期に情報が集中する傾向がある、情報の認知度の特に高い時期だけを示す傾向が強い、などの特徴がみられるので、単語の出現傾向で得られる情報を補うなどの用途があるのではないかと考えている。

抽出処理そのものの改良など今後の課題は多いが、特に重要と考える課題について以下に挙げる。

1. 抽出結果の性質の全般的な分析

   現状では幾つかの情報をピックアップして、性質の差がみられるのかを調べているが、抽出結果全体に渡って、先に述べたような出現の傾向があるのかどうかを評価する手段を考える必要があると考えている。

2. 対象による情報の豊富さの差異

   対象情報が人であるか物であるか、具体物であるか抽象物であるかなどにより表現中に含まれる評価情報に差が出る。

   例えば、人物でも芸能人、スポーツ選手などは、知名度が上がると有名なのが当たり前になり「有名な」のような表現は用いられない。非日常的な特定の芸術・学術分野の有名人などの表現に「著名な」のような表現が用いられる。

   幅広い一般分野で抽出を行い認知度を比較するなどの場合には、分野による評価情報の違いを、どう補うかは課題である。

3. 評価表現と著者や読者の理解の関係

   今回の実験では、記事テキストからの抽出結果だけを分析したが、評価対象の認知度を詳細に知るには、その評価表現により、著者がどういう状況を表現しようとしていたか、読者がどういう理解をしたかを、より深く調べる必要もあると考えている。実際の被験者に対して評価表現を提示し、どう理解したかを調べる実験なども興味深いと考えている。

参考文献

[西野] 西野 文人, 落谷 亮
新聞記事からの人物・企業情報の抽出
情処研報, NL127-17, p.125-132 1998.

[落谷] 落谷 亮
組織名抽出のための知識収集
言語処理学会第5回年次大会, A2-2, pp.112-115 1999.

[日経] 日経全文記事データベース 日本経済新聞 CD-ROM 90-98 年版, 日本経済新聞社

[毎日] CD 毎日新聞 91-97 年版, 毎日新聞社