手がかり語自動取得による Web 掲示板からの評価文抽出

峠　泰成　　山本　和英
長岡技術科学大学　電気系
{touge,ykaz}@nlp.nagaokaut.ac.jp

1. はじめに

近年、Web の普及により大規模なテキストデータを入手することが容易となってきた。これにより、テキストマイニングの技術が注目されてきている。これまで企業が自社製品の評判を知ることや、消費者が製品の評判を得るためには、口コミや新聞などの投稿、コールセンターへの問い合わせといった方法でしか情報を得ることができなかった。しかし、インターネットを利用することによって、インターネット掲示板など製品評判を表すテキストデータを入手することが可能となった。掲示板には“Yahoo!”や“2 ちゃんねる”などの多くのサイトがあり、様々な分野について、多くの人が書いている。掲示板の情報を用いることで、購入しようとしている製品の評判を得たり、製品にどのような評価が下されているのかを企業が知ることができる。しかし、評判を得るためには、大量の掲示板記事を読まなくてはならない。掲示板の中には、一覧性が悪いものや、過去ログが調べにくいといったものも、多く見られる。

本研究では、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し、大量にあるデータから有用な情報を抽出するための手法を提案する。

2. 関連研究

Web から評価文を抽出しようとする研究はいくつか行われてきた。インターネット上のページから、評判を述べている部分を抽出する研究として立石ら[1]がある。[1]は、キーワード検索でのキーワード（商品名）と商品の評価である評価表現とを用いて、評判文の含まれるページを抽出している。また村野ら[2]は、携帯電話に関しての掲示板から主観的評価文の抽出を行っている。しかし、これらの研究では、各商品の評価対象の単語（商品名や属性表現）について辞書を用意する必要があること、また対象名（商品名）を手がかりとしているとしているため、対象名が存在しない場合は評価文を抽出することができないという問題があった。特に、一つの商品のみについて書かれた掲示板においては対象名が現れない文がほとんどである。また、小林ら[3]は、ビールやコンピュータの掲示板より、属性表現と評価表現を抽出するため規則を用いて収集し、人手により正例、負例を分けるという手法を提案している。

これに対し本研究では、掲示板より評価の対象である対象表現を自動取得し、製品名が存在しない評価文も抽出することができる手法を提案する。対象表現を自動取得することによって、一つの掲示板に特化することなく、汎用性のある手法を目指している。

3. 対象表現と評価表現

本稿では車の掲示板を例として挙げる。掲示板における評価文とは、“この車の収納は使いづらいですね”や“乗り心地が確かに硬めですね”といった主観的な評価を表す文を指す。本研究では、掲示板から評価文を抽出するために、対象表現と評価表現という手がかり語を用いている。対象表現とは、商品名や、“燃費”、“ブレーキ”といった評価の対象となる単語と定義する。また評価表現は、“良い”、“悪い”といった、人の主観的評価を表す単語と定義する。これらの表現を用いることで、“商品Aの燃費はいいですね。”のような評価文を抽出できる。

しかし、掲示板の表現は個人の自由で書き込むことができ、様々な書き方がなされている。評価文を抽出する際に製品名などが省略されていることもかなり多く見られる。本研究では、最初に評価表現辞書を構築し、それを手がかりに対象表現を抽出した。抽出時には、南瓜<1> により構文解析を行ったテキストを用いている。

3.1 評価表現抽出

本研究の処理の流れを図 1 に示す。まず、Yahoo!掲示板<4>の車に関するカテゴリーより、ある一つの車について書かれている掲示板文書を取得した。データ量は 6371 件(28479 文)の書き込みである。ここから、(1)～(3)の規則に適合した評価表現候補を抽出し、評価表現辞書を構築した。評価表現候補は、主観的表現となる単語が多く含まれる次の品詞とした。動詞-自立、形容詞-自立、名詞-サ変接続、名詞-形容動詞語幹、名詞-ナイ形容詞語幹の５つである。ここで、品詞は南瓜<1>による結果を用いている。取得した評価表現候補から、評価表現として適当な単語を人手で判断し、評価表現辞書として登録した。

(1)（ 名詞-一般 / 名詞-固有名詞 / 未知語 / 名詞-サ変接続)+（ が / は / も / を / に ）+（ 評価表現候補 ）
(2)（ 評価表現候補 ）+（ 名詞-一般 / 名詞-固有名詞 / 未知語 / 名詞-サ変接続 ）
(3)（ 副詞 ）+（ 評価表現候補 ）

評価表現候補となった単語は 2404 単語あり、その中から評価表現となりうる 368 単語を評価表現辞書として半自動構築した。以後評価表現辞書とはこの辞書を指す。

3.2 対象表現抽出

対象表現は、先に収集した評価表現を用いて、用意した二つの規則にあてはめて自動取得する。対象表現の候補となる単語の品詞は、名詞-一般、名詞-固有名詞、名詞-サ変接続、未知語の４つに限定している。更に、ひらがなのみで出現している単語は、評価文を抽出する際に対象表現となることがほとんど考えられないため抽出対象外としている。対象表現の抽出規則を以下に示す。

(1)｛ 対象表現 + が/は/を/も/に + 評価表現 ｝
(2)｛ 評価表現 + 対象表現 ｝

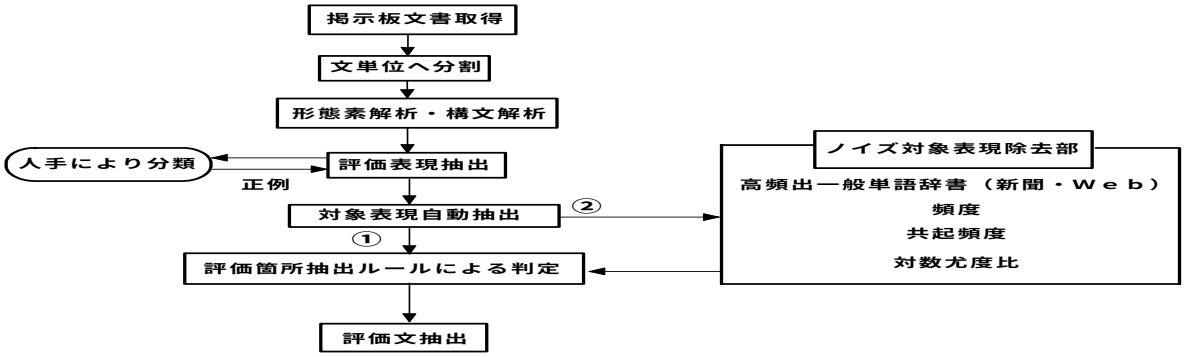

図１　処理の流れ

4. ノイズとなる対象表現の削除

3.2 節の対象表現抽出により自動取得した単語の全てが評価文と関係あるわけではなく、ノイズとなる単語が多数存在する。例えば、車の掲示板でよくみられる対象表現としては、“燃費”、“エンジン”、“ブレーキ”といった単語がある。これらの単語は評価文を抽出する際に十分な手がかり語とすることができる。しかし、高頻出単語として“人”、“年”、“月”などのノイズとなる単語も多数観察した。ノイズとなる単語は評価文を抽出する際に悪影響を及ぼし、適合率の低下につながる。また、低頻度の対象表現と評価表現が共起していたとしても、それが評価文ではない場合が多い。その様な、あきらかに評価文に寄与しない対象表現は削除したい。これらの理由から、ノイズとなる対象表現を削除する。

4. 1　高頻度一般単語の削除

抽出した対象表現の高頻度単語のうち、“人”などのノイズ対象表現を削除し、重要である対象表現を残すことのできる方法を試した。新聞や Web コーパス[4]において高頻度で出現する単語を対象表現から削除することを試みた。新聞や Web コーパスにおいて高頻度に出現する対象表現は、掲示板においてノイズとなるデータを多く含んでおり、評価の対象になりづらい表現であると仮定した。本研究では、新聞と Web コーパスそれぞれにおいて、名詞-一般、名詞-固有名詞、未知語、名詞-サ変接続 、の頻度上位 500、1000 位までの辞書（一般単語辞書）を作成した。辞書作成時に用いたコーパスは次の通りである。

新聞<2> ： 毎日新聞 1999 年、2000 年の 2 年分,43MB
Web コーパス<3> ： Web 中の良質なテキスト, 220MB

一般単語辞書に登録されている単語が、3.2 節で取得した対象表現に含まれていた場合には、その対象表現を削除する。

4. 2　低頻度ノイズ対象表現の削除

自動取得した対象表現において、低頻度でしか出現しない単語が評価文抽出の際のノイズとなってしまうことが多い。従って、あきらかにノイズになってしまう対象表現を削除した方が望ましい。そこで、本研究では、次の３つの方法を試みた。

１）対象表現のみの頻度による削除
２）対象表現と評価表現の係り受け共起頻度による削除
３）対象表現と評価表現の対数尤度比による削除

1)は、頻度のみを用いている方法、2)は共起頻度に基づく方法、3)は対象表現と評価表現との対数尤度比の重みづけによって対象表現を削除する方法である。対数尤度比Gは以下の式より導出した。

$$\frac{G}{2} = a \log \frac{an}{(a+b)(a+c)} + b \log \frac{bn}{(a+b)(b+d)} + c \log \frac{cn}{(a+c)(c+d)} + d \log \frac{dn}{(b+d)(c+d)}$$

a:対象、評価表現がどちらも存在する文の数
b:対象表現のみが存在する文の数
c:評価表現のみが存在する文の数
d:どちらも存在しない文の数

また、4-1 節の方法と組み合わせることによって高頻度と低頻度で現れるノイズ対象表現を削除することも試みた。

5. 評価文抽出パターン

掲示板の文書から評価文を取得するためには、その文が評価文であるかを判断する基準が必要である。本研究では、掲示板でよく見られる評価文の特徴をパターンとして用いることで、評価文を抽出する方法を試みた。これを評価文抽出パターンと呼ぶ。規則の作成に用いたデータは、評価表現辞書を構築する際に用いた掲示板の文書としている。評価文を抽出する際には前節で取得した対象表現と評価表現を手がかり語として用いる。評価文を抽出するための評価文抽出パターンとしては、図 2 に示す 11 のパターンを用いて文単位で評価文を抽出する。ただし、これらのパターンに照合しても、文末が以下に示す疑問形で終わっている形の文は抽出しないことにする。

“？”　“ですか。”　“でしょうか。”

6. 評価実験

評価実験に用いたデータは、調査に用いたデータとは異なる車の掲示板の文書 4523 件（25952 文）とした。この中から、ランダムに 4000 文を選び、人手により評価文を判断した。その結果、538 文を正解の評価文とした。

この掲示板全体からの対象表現自動抽出により、1404 単語を取得した。まず、ベースライン、本手法、先行研究の比較を行った。ベースラインは{ 動詞-自立 / 形容詞-自立 / 名詞-サ変

１） [対象表現] + [が/は/も/でも/で/ので] + [評価表現] + [文末表現]。/ ！
２） [対象表現] + [について / に関して] + [評価表現]
３） [対象表現１] + [の] + [対象表現２] + [が/は/を/も/に] + [評価表現] + [文末表現]。/ ！
４） [対象表現] + [より/の方が/と比べ/と比較] + [評価表現]
５） ～し（、） + [対象表現] + [は/が/も] + [評価表現]
６） ～より(は) + [対象表現] + [は/が/も] + [評価表現]
７） ～と(比べ | 比較 | 比べる) + [対象表現] + [が/は/を/も/に] + [評価表現]
８） ～の方が + [対象表現] + [が/より] + [評価表現]
９） [評価表現] + [対象表現] + [文末表現]。/ ！
１０） 特徴的な副詞 + [動詞/形容詞/名詞-形容動詞語幹/名詞-ナイ形容詞語幹/サ変名詞]
１１） [対象表現１] + 特定の評価表現 + [対象表現２]
※文末表現は次の表現で終わる文としている：｛ です/ます/よ/ね/と思う/と感じる/と思える/と考える/気がする ｝
特徴的な副詞とは次の単語に限定している： ｛ とても/とっても/すごい/全然/かなり/よく/あまり/非常に/結構/やっぱり ｝
特定の評価表現として次の単語を登録している:｛ 想像以上/予想以上/同レベル ｝

図 2. 評価文抽出パターン

接続 / 名詞-形容動詞語幹 / 名詞-ナイ形容詞語幹 }の品詞の単語と{ 名詞-一般 / 名詞-固有名詞 / 名詞-サ変接続 / 未知語 }の品詞の単語が係り受けになっていれば評価文とした。ここで、先行研究は評価表現辞書や属性辞書などをあらかじめ人手によって集めておき、文型パターンに当てはめ、評価文を抽出しようとする研究[2]である。本手法において、対象表現のノイズ削除を行っていない場合の結果を、表 1 に示す。この結果より、評価文抽出パターンは有効に機能しているが、項目ごとに辞書を用いた研究[2]に比べ精度が悪い。

次に、適合率の改善のために、取得した対象表現の高頻度ノイズ削除を行った結果を図 3 に示す。新聞の一般単語辞書 500 単語を対象表現から削除した場合、1404 単語から 139 単語減少し、Web による一般単語辞書 500 単語を用いた場合、226 単語減少した。新聞による削除に比べ、Web による一般単語の削除は、必要となる対象表現をも多く削除する結果となっている。

次に、低頻度ノイズ対象表現削除として行った、頻度、共起頻度、対数尤度比を用いた結果を図 4 に示す。削除する単語の基準として、既知対象試験により一番良い評価であった共起頻度の情報を用いて、上位 500 単語（頻度 5 以上）と上位 800 単語（頻度 2 以上）により行った。頻度、対数尤度比も同様の条件にするため、共起頻度と同じ単語数を用いた。図 4 より、低頻度ノイズ対象表現除去に有効な手法は共起頻度であることがわかった。

さらに、高頻度一般対象表現を一般単語辞書で削除を行い、低頻度ノイズ対象表現を共起尺度で削除する方法も試みた。低頻度ノイズ対象表現除去で良い結果を示した共起頻度を用いた方法の結果を図 5 に示す。この結果より、新聞の一般単語辞書により、対象表現の頻度上位にあった単語を削除でき、共起頻度の情報で低頻度ノイズ対象表現を削除する方法が一番良い。削除された単語も二つの方法で違う単語を削除していることも分かった。

表 1. 評価文抽出結果

	ベースライン	本手法	(先行研究)
適合率(%)	15.1(485/3212)	50.1(310/619)	(66.5)
再現率(%)	90.1(485/538)	57.6(310/538)	(62.0)

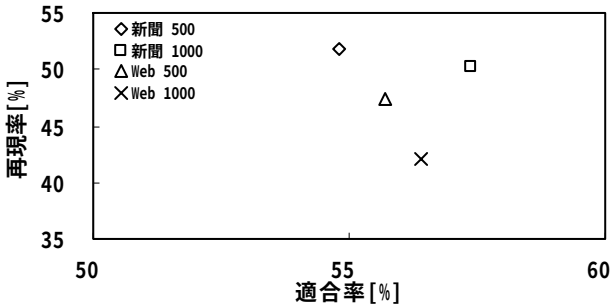

図3.高頻度ノイズ対象表現削除結果

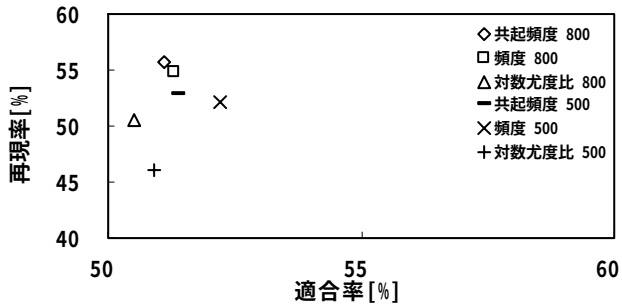

図4.低頻度ノイズ対象表現削除結果

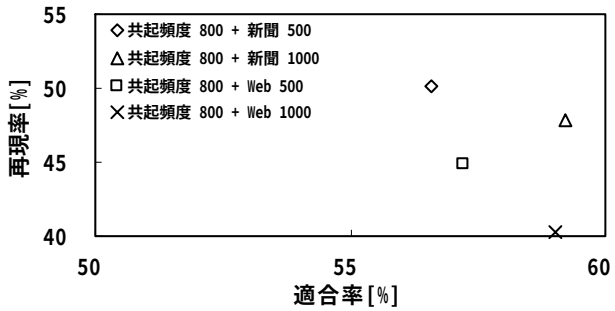

図5.一般単語辞書,共起頻度の組み合わせによるノイズ対象表現削除結果

7. 考察

7. 1 評価文抽出パターンについて

評価文抽出パターンをもとに評価文を抽出する手法を提案したが、今回用いた規則だけでは、6 割程度の再現率となった。ここで、抽出に失敗している例を以下に示す。

1) 外観は決して良いものではありません。
2) 両側スライドドアは広さを犠牲にするぐらいならいらない。

ある(動詞-自立)や、ない(形容詞-自立)などのノイズになりやすい評価表現が評価表現辞書に含まれていないため抽出されなかった。

3)おすすめですよ～。

評価文であるが、対象表現が省略（直前の文に出現）している場合。

4)スタイル△、エンジン△、小回り△、室内の広さ×

評価している箇所が記号などで表されている場合。

5)商品Aを買った理由がこのエンジンとサスペンションだった。

評価文抽出パターン以外の形で出現した場合。

以上のことから、評価表現辞書に含まれていない表現をさらに増やす必要がある。決められた品詞の単語ではなく“腹が立つ”などの一つの表現として自動抽出することができれば良い。

また、ノイズとなる文を抽出せず、これらの状況に対処できるようなもう少し細かい評価文抽出パターンを作成する必要がある。

7. 2 ノイズ対象表現除去について

本手法で行ったノイズ対象表現除去において、“人”、“年”など掲示板に高頻度で出現するが、対象表現にはならない単語は新聞辞書などを用いることで削除することができた。しかし、ノイズとならない対象表現が新聞や Web の上位にくることもあり、削除されてしまう例も見受けられた（“色”、“車”、“エンジン”等)。新聞と Web コーパスそれぞれの一般単語辞書に登録されている単語は半分程度同じ単語であった。新聞の一般単語辞書は経済記事の情報が多く、Web コーパスの一般単語辞書はインターネット上でよく見る単語も多くあった。結果として、Web コーパスでは対象表現を新聞と比べて、より多く削除してしまっており、再現率を低下しまう原因となっていた。また、全体的に一般単語辞書を用いた場合の対象表現自動削除は、再現率がかなり低下する結果となった。いくつかの掲示板と一般単語辞書を用いることで、多くの掲示板で出現するノイズ対象表現を自動で取得することができるのではと考える。

低頻度ノイズ対象表現について、いろいろな尺度を試みた結果、共起頻度の情報を用いる方法が一番良い結果となった。自動抽出した対象表現 1404 単語から共起頻度の高い上位 800 単語を評価文抽出規則に適用した結果より、誤って削除された評価文は 10 文であり、削除された 604 単語の対象表現のほとんどがノイズ単語であることがわかる。これより、ノイズとなる対象表現を効率よく削除できることになる。

対数尤度比によるノイズ対象表現の削除も試みたが、共起頻度よりも良い結果が得られなかった。これは、対象表現と評価表現の結びつきが特徴的ではなかったためであると考えられる。また、一般単語辞書と共起頻度を用いた場合において、それぞれの方法が高頻度ノイズ対象表現と低頻度ノイズ対象表現を削除することに有効であった。しかし、誤って削除されてしまう評価文の数が増えてしまうため、さらに効率的な削除方法を行う必要がある。

8. おわりに

本研究では、対象表現と評価表現を手がかり語とし、掲示板の特徴表現から作成した評価文抽出パターンを用いることでインターネット上の掲示板から評価文を自動抽出する方法を提案した。

実験の結果、評価文抽出パターンのみによる方法での適合率は 50.1(%)、再現率は 57.6(%)となった。また、自動取得した対象表現のうち高頻度と低頻度に存在するノイズとなる対象表現を削除することも行った。新聞から作成した一般単語辞書を用いて、対象表現の高頻度でノイズになっている単語を削除し、対象表現と評価表現の共起頻度により低頻度でノイズとなっている単語を削除する方法が有効であった。

このことより、作成した評価表現が限定した品詞のみの単語だけであり表現が少ないこと、より細かい評価文抽出パターンを作成しなければならないなどの問題が明らかになった。また、効率よくノイズとなる対象表現を削除するために、他の掲示板からの情報を取り入れ、一般単語の自動収集も検討している。これらを適用することで、多くの掲示板から評価文を正確に抽出できるのではないかと考えている。

使用した言語資源とツール

<1> 南瓜 ver.0.4.0 奈良先端科学技術大学院大学 松本研究室 http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/

<2> 毎日新聞全文記事データベース 1999,2000 年版．毎日新聞社

<3> Web コーパス:参考文献[4]における Web コーパスA，220MB

<4> Yahoo!掲示板 日本車(調査用：フィット 評価用：ステップワゴン) http://messages.yahoo.co.jp/yahoo/Recreaion_Sports/Automotive/Makes_Models/Japan/

参考文献

[1]立石健二,石黒義英,福島俊一 ： インターネットからの評判情報検索 ， 情報処理学会研究報告 NL-144-11 ， pp.75-82 (2001)

[2]村野誠治,佐藤理史 ： 文型パターンを用いた主観的評価文の自動抽出 ， 言語処理学会第 9 回年次大会発表論文集 ， pp.67-70(2003)

[3]小林のぞみ,乾健太郎,松本裕治,立石健二,福島俊一 ： テキストマイニングによる評価表現の収集 ， 情報処理学会研究報告 NL-154-12 ， pp.77-84(2003)

[4]関口洋一,山本和英 ： Web コーパスの提案 ， 情報処理学会研究報告 NL-157-17 ， pp.123-130(2003)