ドメイン特徴語の自動取得による Web 掲示板からの意見文抽出

峠 泰成　大橋 一輝　山本 和英
長岡技術科学大学 電気系
E-mail:{touge, ohashi, ykaz}@nlp.nagaokaut.ac.jp

1　はじめに

近年、Web には大量の情報が存在しており、多くの情報の中から有益な情報を発見するテキストマイニング技術が一つの重要な方法となっている。例えば、企業が自社の製品について、消費者がどのような関心をもっているかという情報に対して注目したり、個人が購入しようとしている製品がどのような物であるかを検討するために Web を用いることは一般的である。このことから、多くの人の興味や関心の集まっている情報源に対する研究が多くなってきている。

その中でも、Web 掲示板は、個人の興味・関心が一般の Web 文書に比べて、大量に現れる情報源と考えることができる。しかし、大量の情報があるため、ここから必要な情報を得るためには、多くの掲示板を読むこととなり、多くの時間やコストがかかってしまう原因となる。

本稿では、Web 掲示板から人手を介すことなく意見情報を効率良く取得する手法を提案する。Web 掲示板には、パソコンや旅行など様々なドメインが存在し、自分の意図する情報を得るためには多くの知識が必要となる。これらの知識を、ドメインごとに作成することは非常に労力のかかる作業となってしまう。

そこで、我々は Web 掲示板に大量の書き込みがあるという利点を用いて、ドメインごとに辞書を作成せず、ドメイン特徴語を自動取得し意見文を判別する手法を提案する。

2　関連研究

Web 掲示板から意見情報を取得、分類する研究はいくつか行われている。

意見情報の収集としては、立石ら [1] が表現の 3 つ組による抽出手法を行っている。この研究では、Web 掲示板から意見情報を抽出し、その結果から製品に対する要約提示を行っている。意見情報は、対象・属性・評価の 3 つの表現をもとに抽出する。3 つ組の表現と意見らしさのパターンマッチによって、意見文かどうかを判定している。しかし、この手法では 3 つ組の表現の辞書を人手により構築する必要がある。これは、企業が Web 掲示板などから評判情報を細かく収集する場合には効率も上がり、非常に有効な手段であるが、一般の人が各ドメイン毎に辞書を構築する作業は手間がかかってしまう。一方、我々はこれらのドメインに依存する辞書を作成せずに意見を収集する。

また、藤村ら [2] も評判情報の抽出を行っている。Web 掲示板では、書き込みそのものに肯定や否定のタグが付与されているものがある。この情報を用いて、肯定の評判に現れる単語と否定の評判に現れる単語を学習し、評価表現の分類を行っている。しかし、この表現のみではキーワードを獲得できるが、ノイズも多く文書単位での情報となり文単位の判定のように細かい分類が行えない。

意見情報を抽出するには、人手による辞書の作成の負担の軽減や、作成した辞書がドメインに依存することへの対処が必要となる。また、意見情報を扱う単位が書き込み (テキスト) 単位であると、様々な情報を含んでしまうため、文単位程度の長さで扱う必要があると考える。我々は、この二つの問題を解決するために、人手によるドメイン依存の辞書を作成せず、Web 掲示板の書き込みから意見文を判別する単語の強さを学習することでドメイン依存の問題を解決した意見文の抽出を行う。

3　提案手法

3.1　意見文の定義

本稿で扱う意見文について述べる。意見文は、個人による評価や意見を含んでいる文と定義する。Web 掲示板は、個人による意見が他の Web 文書に比べて多く含まれている。

例 1)　{ エスティマ }$_{対}$ の { 乗り心地 }$_{属}$ は { 良い }$_{評}$ です。

例のように、意見文には、対象表現 (製品名や組織など)、属性表現 (製品の属性)、評価表現 (対象の評価、状態) といった表現の組み合わせによって成り立つことが多い。しかし、実際に意見情報を取得する場合、Web 掲示板では個人が好きなように書き込みを行うため、同じ製品の事であっても、表記揺れがなされていたり、主題となる単語が省略されていたりテキスト自体の問題も多い。人手によって辞書を作成する場合にも、これらの表記揺れに全て対応することは難しい。よって本稿では、これらの表現の自動取得も試みた意見文抽出を行う。

3.2　処理の流れ

本稿での処理の流れを図 1 に示す。

我々は、意見文を抽出するための手法として、入力文に現れた単語が意見文になりやすい単語であるか否かを学習し、この値を用いて意見文であるかの判定を行う。このスコアを計算するため、自分が情報を得たい Web 掲示板と同じドメインの書き込み (タグなしデータ) を取得し学習する。

この学習方法として、提案手法では大きく 2 つの方法の組み合わせから成り立つ。1 つ目は、自分が情報を抽出したいドメインの特徴語となりうる単語を自動取得することを行う。これは、検索エンジン「Google」の検索ヒット数をもとに、抽出した候補単語がドメインにおいて必要な単語であるかを判定する。2 つ目は、掲示板に出現した単語が意見文になりやすいか否かのスコアを計算する。これは同ドメインのデータを用いて、人手により収集した評価表現などの重み付け辞書との共起性をもとに、ドメイン中に出現した単語の意見文へのなりやすさのスコアを意見文判別の値として学習する。そのため、ドメイン依存の辞書を作成せず、そのドメインの特徴語にスコアを付与することが可能となる。

この二つの手法により得られた情報を用いて、1 文ごとに意見文スコアを算出し、スコアの高い上位 10% の文集合と、スコアの低い下位 50% の文集合を学習データとして扱い、単語に対してそのドメインでの意見文になる強さを付与した単語データを作成する。そして、この単語データをドメインに順応させるため、同ドメイン文書を用いて繰り返し学習を行い、新たな単語データを作成する。

実際の意見文取得は、ドメイン特徴語をとらえた単語データを使い意見文判別を行う。

3.3　単語データの作成

意見文を抽出するための情報として、出現した単語が各ドメイン中で、意見文を判別するためにどの程度の強さを持つかを表す単語データを作成する。この単語データの作成には、我々の手法 [3] をもとに行っている。

基本的に、意見情報であるためには、抽出対象の文中に評価表現となりうる単語が存在している。そこで、単語データを作成するために、同ドメインの掲示板文書から学習データとし

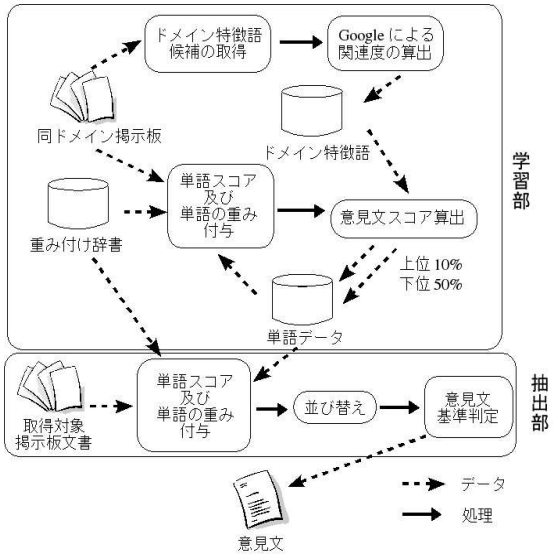

図 1: 処理の流れ

て信頼できるデータを作成する。これは、評価表現などの重みづけ辞書とドメインで特徴的な単語を取得する方法を用いて、掲示板の文が意見文となるか否かを判断し、その結果から実際の単語データを作成するという手法である。

3.3.1 初期単語データ

意見文を判別するスコアを付与するための初期データとして、意見文か否かのタグ付きデータから初期単語データを作成する。タグ付きデータは次の通りである。ここで、単語データ作成の候補に使用する単語の品詞は、名詞、動詞、形容詞、副詞、記号列、未知語となっており、これら以外の単語を含まない。

- 意見文のタグが付与されたデータ : 6000 文
  意見文 : 845 文, 意見文でない : 5155 文
  単語数 : 8044 単語（異なり）

このデータをもとに、ある単語が意見文にどれだけ出現するかによって、単語ごとのスコアを算出する。単語のスコアは式 (1) によって求める。

$$W_s(w_i) = \frac{P_p(w_i)}{P_p(w_i) + P_n(w_i)} \tag{1}$$

$W_s(w_i)$: 単語 w_i のスコア, $P_p(w_i)$: 意見文で単語 w_i が出現する確率, $P_n(w_i)$: 意見文以外で単語 w_i が出現する確率

この処理によって作成される単語データの様式を、表 1 に示す。

表 1: 単語データの様式

単語	意見文	意見文でない	単語スコア
良い	15	4	0.789
快適	10	2	0.833
家族	2	25	0.074
電話	4	9	0.307

この結果を、初期単語データとして、同ドメインの書き込みを学習する際の単語スコア付与のベースとする。

3.3.2 評価表現への重みづけ

ドメインの特徴語をとらえた単語データ作成のために、意見文判別の学習データを作成する。このとき、意見文判別で評価表現は大きな手がかりとなる。評価表現に対して重みを加えることで、一般的な単語と区別することが可能となる。我々は、どのドメインにも現れるような一般的な評価表現を人手により収集し、評価表現辞書として用いた。評価表現辞書の用語数は、1274 表現となっている。

しかし、評価表現辞書の登録数が評価表現のすべてをカバーできているわけではない。そこで、汎化規則を設けて表現の増加を試みた。汎化規則として 20 の規則を作成し、これを満たす表現も評価表現として扱うが、評価表現辞書には登録しない。汎化規則は次に示すような規則となっている。

- 動詞 + やすい (形容詞-非自立)
- 名詞 + 的 (名詞-接尾-形容動詞語幹)
- 名詞 + が + ない

汎化規則では、重みづけしなくてもよいところにスコアの重みづけを高くする可能性があるため、評価表現辞書による重みづけに比べて小さい重みとした。今回は、評価表現辞書の単語には 2 倍、汎化規則による重みには、1.5 倍の重みを加えている。

3.3.3 強調表現への重みづけ

意見文を判別するため、評価表現と同じく、副詞のように表現を強調する単語も手がかりとなる。これらの表現を強調表現と呼ぶことにする。強調表現は、副詞を中心に人手で収集した。強調表現の数は 75 表現である。強調表現は評価表現に比べ、意見文であるかの判断基準としては弱いため、強調表現には 1.5 倍の重みを加えることにしている。

3.3.4 文中での主題判定

意見文を判定するために、評価表現と強調表現に対して重みを加えたが、実際にその文中に評価される対象となる表現が含まれていない場合には、意見文と判断することはできない。

例えば、車の掲示板では、次のような文が見られる。

例 2)$\{$ ハンドル $\}_{\text{主題}}$ に重い分銅を付けているようです。

例 3)$\{$CD$\}_{\text{主題}}$ の使い勝手もなかなか良いですよ。

学習に用いるデータ中のノイズを削減するためにも、主題の含まれないデータはあらかじめ削除しておくべきである。よって、文中での主題表現の有無を判定する必要がある。取得するドメインによって主題となる表現が大きく異なることや、新しい単語へ対応する必要があるため、評価表現のように、あらかじめある程度の量を保持しておくことができない。そこで、掲示板文書から主題を自動抽出するため、検索エンジン「Google」での検索ヒット件数を用いる手法を試みた。

学習に用いている、情報を取得したい掲示板 (以下、対象掲示板) の話題 (例えば「フィット (車名)」) を与え、対象掲示板中に現れる主題候補との関連度を計算し、この関連度が一定以上であった場合にその単語をドメインでの主題と判断した。対象掲示板から取得する主題候補の品詞は、未知語、名詞、記号列 (アルファベット) とした。

話題 Key と対象掲示板から自動取得した主題候補 Word との関連度 R(Key,Word) は次のように計算する。

$$R(Key, Word) = \frac{2 \cdot H(Key, Word)}{H(Key) + H(Word)} \tag{2}$$

H(Key , Word) : 話題と主題候補の共起の検索結果数、H(Key) : 話題の検索結果数、H(Word) : 主題候補の検索結果数

予備実験の結果、関連度は大きく 2 つに分けられる。R $>$ 0.1 の場合、製品名、会社名などが集中的に集まる傾向にある。また 0.1 $>$ R $>$ 0.01 の場合、属性表現が多くみられる傾向に

ある。これより、主題候補が関連度 0.01 以上の単語を主題として扱う。

また、次の例のように、文中に主題が出現しなかった場合には、その文の書き込み以外での主題情報の補完が必要となる。

例 4) かっこいいですね！

本手法では、掲示板の同じ書き込み中にある情報であれば主題となる表現を補完可能であるとし、補完可能な表現であればその文での主題情報であるとする。前文において主題情報を上記の手法で判定してあるため、その助詞をもとにどの情報が現在の文における主題としてあてはまるかの優先順位を定める。助詞は、「は」などは「へ」、「で」などに比べて、主題の単語になりやすいといったようなことを考慮した優先順位のルールを 4 つ作成し、主題情報の補完を行う。また、前文にも主題の候補となる単語が存在しなかった場合には、主題の存在しない文であったとして扱う。

3.3.5 意見文スコアの算出と繰り返し学習

以上の重みを考慮して、単語データを元に意見文スコアを算出する。この時に、単語データのみでは、新出の単語にスコアを付与できない場合もある。この場合、単語データのすべての単語の平均の値を、新出の単語の値として適用する。

ある文 S の意見文スコア S(s) は次のように計算する。

$$S(s) = \frac{\sum_i W_s(w_i)}{Average} \tag{3}$$

Average : 文に含まれる候補の品詞の単語に単語データの平均値を与えた時の総和

意見文スコアは、文ごとに含まれる単語数の違い (文の長さ) を考慮して、単語データの平均値を候補の品詞の単語数だけ足し合わせた文のスコアと、単語スコアを足し合わせた文のスコアの比によって計算した。

文ごとに意見文スコアを付与し、学習に用いた文を降順に並び替える。ここから、意見文スコアの高い文と低い文それぞれに現れる単語を学習し、ドメインに順応する単語データを作成する。意見文スコアの上位 10% は意見文として、下位 50% は意見文でないとして扱い、式 (1) により再計算を行う。これ以外のデータは、意見文と判定することが難しいため扱っていない。ここで、単語データの作成では、意見文には評価表現との共起性と主題表現の有無を考慮して、上位 10% の結果にこれらの表現が存在しない場合は、学習対象とはしない。

再計算の結果を新しい単語データとして、もとの単語データに追加・更新する。もとの単語データを学習により更新することで、同ドメインの単語を学習していくことができる。例えば、車の掲示板から情報を取得したい場合は、車のドメインのテキストの学習を行い、初期単語データではカバーできなかった単語へ意見文を判定するスコアを付与できるようにすることが目的となる。これによって、ドメイン依存の問題を解決する。

3.4 意見文取得

以上の学習を行い作成された単語データを用いて、意見文の取得を行う。

情報を取得したい掲示板の書き込みを入力し、この文書を文分割し、文に対して単語データにより単語にスコア付けを行う。意見文スコアの付与に関する過程は、単語データを作成する場合と同様である。

意見文取得で単語データ作成とは異なる点として、疑問表現や推定表現などの意見文になりにくいと考えられる情報や「ありがとうございます」といった掲示板に特有する表現などは意見文として抽出されない。

評価表現などと共起する強さやドメイン特徴語の取得といった情報を使って、ドメイン依存しない意見文の取得が可能となる。

4 評価実験

情報を取得したい Web 掲示板文書を入力として、ドメインに順応した単語データを用いて、意見文情報を取得できているか評価する。

4.1 実験データ

今回評価に用いたデータは、Yahoo!掲示板の車のドメインの書き込みと携帯電話のドメインの書き込みである。この中より、それぞれのドメインに対する学習データと正解データを用意した。学習データは、ドメインに順応する単語データを作成するためのデータであり、それぞれのドメインから 5 つずつ、同ドメインの別々の製品に対する Web 掲示板文書を用いた。また、正解データとして、学習データとは別にそれぞれのドメインのある製品に対する 100 件の書き込み (車 : 650 文, 携帯電話 : 544 文) に人手により意見文か否かのタグを付与したデータを作成した。これらのデータを用いて次節からの実験を行った。

4.2 評価表現辞書の有効性

人手により収集した評価表現辞書によって、ドメイン依存の関係をどの程度カバーできるか評価する。この評価表現辞書を用いて、辞書に含まれる単語が入力文中に含まれていた場合、意見文を抽出できる対象となるとし、この表現が含まれていた文のみの精度を検討する。この場合に再現率が低ければ、ドメイン依存の評価表現辞書を作成しなければならないため、本手法では評価表現辞書の有効性が重要となる。

意見文抽出に必要である評価表現を、どの程度再現できるかを検討した結果を示す。それぞれのドメインの正解データに適応した場合、評価表現の含まれている文の結果を表 2 に示す。

表 2: 評価表現辞書の有効性

ドメイン名	適合率 (%)	再現率 (%)
車	35.9 (88/245)	88.0 (88/100)
携帯電話	30.0 (65/217)	89.0 (65/73)

※適合率の分母は、評価表現の含まれていた文数

書き込み全体中に評価表現の含まれている文は、約 3 割程度であった。評価表現辞書としてある程度の語数があれば、ドメインの違いにも多くの場合に適応できる傾向があると言える。しかし、収集した評価表現辞書のみでは、約 1 割程度の意見文は抽出もれとなる。これらのドメインに依存する評価表現は、別の手法を用いて抽出する必要がある。

4.3 本手法の有効性

Web 掲示板の書き込みを学習することによって、ドメインの特徴語をとらえ、単語データの単語が、意見文を判別する値を獲得できているか評価する。

ドメインごとに作成した単語データを用いて、実際の意見文抽出結果を検討する。結果は、掲示板文書を文ごとに分割し、意見文スコアの上位から順に出力する。ここで、評価表現辞書の単語と、主題となりうる単語が含まれていなかった場合には意見文としての抽出対象とはならない。

この条件のもと、取得した結果について、適合率と再現率を用いて評価を行う。

ドメインごとの意見文の抽出結果を図 2、3 に示す。

それぞれの図において、左からスコアの値が大きい順に並んでおり、抽出したそこまでのデータ全体の結果を示す。出力結果は、10% ずつのデータ量を加算してプロットしており、一番

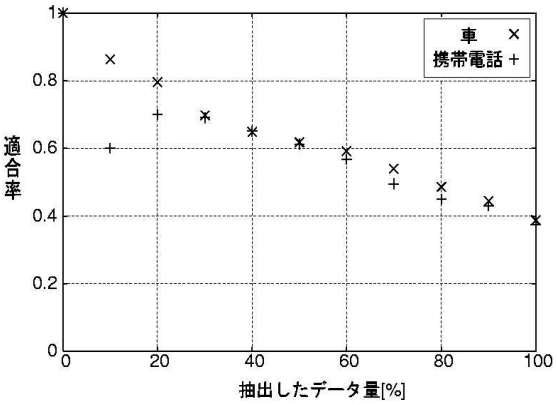

図 2: ドメイン別意見文抽出結果 (適合率)

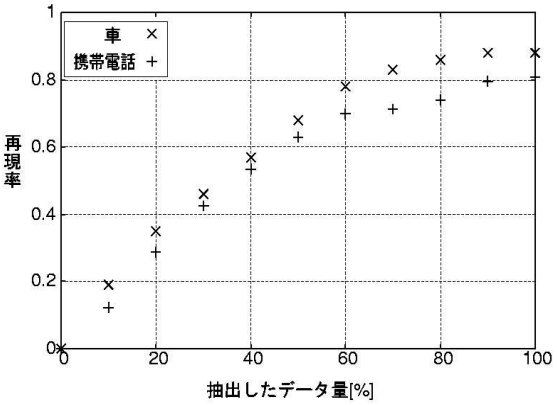

図 3: ドメイン別意見文抽出結果 (再現率)

右端では抽出結果全体の精度となっている。例えば、車のデータの結果として、抽出結果の上位 60% では、6 割の精度で抽出すべき意見文の 8 割の情報を含んでいる。しかし、主題の抽出もれなどによって、評価表現のみの再現率より低くなっていることもあった。結果から、評価表現のみの意見文抽出よりドメイン特徴語も考慮した結果の方が、抽出した文のスコアが上位の場合、比較的高い適合率で意見文を抽出することが可能である。

5 考察

意見文抽出に対して、評価表現辞書の利用と、ドメイン特徴語を自動取得することによる抽出への有効性を実験し、次のようなことが問題点として考えられる。

まず、ドメインに依存しないための評価表現辞書を作成したが、4.2 節の再現率の値から有効に働いていると言える。ここで辞書に存在しなかった単語として、車では「(外観が) きまっている」や「(音が) コトコトなる」などで、携帯電話では「(音が) 割れる」や「(液晶が) 真っ暗」などのドメインに依存してしまう単語であった。また、「ある」、「ない」などの単語は、決まった単語のみが評価表現になるため、全ての網羅が非常に難しい。取得ができていないドメイン依存の評価表現に対応するために、単語データの上位の値の単語との共起を用いて取得する方法を検討する。

次に、評価表現と主題をもとに、ドメイン特徴語をとらえた意見文抽出を行った結果、評価表現のみでの抽出結果と比べ、さらに必要となる情報を上位に収集することができるようになった。

しかし、いくつか問題点も挙げられる。検索エンジンを用いてドメインの主題推定を行っているが、この際にドメイン特徴語となりうる単語のスコアが低い場合 (関連度 < 0.01) には、ドメインに必要ない単語として学習される。関連度を意見文スコアの値に反映する (重みとして) ことで、削除されていた情報を少なくすることが可能となり、誤った学習を少なくすることができる。また、一般的にどの掲示板にも出現するノイズとなるような単語に関しては、あらかじめストップワードとして用意し、対象から外しておくことも必要である。

また、抽出の精度を向上させるためには、補完する単語の意味を考慮するべきである。Web 掲示板には主題の省略が多くみられることや、補完されるべき内容が文全体であることなど様々な問題がある。現在の処理では、前の文の主題となる単語を格関係のみを考慮して補完しているが、さらに大きな範囲と正確な照応解析を行う必要がある。これにより、抽出精度の向上が期待できる。ここで、携帯電話の結果の精度が全体に低くなっているが、車の結果に比べ主題の補完できていない場合がいくつか見られた。細かい精度を確認するには、さらに実験データを増やし検討する。

また、文自体の構造についての考慮をしていないため、Web 掲示板の一つの特徴である、「質問」などサポートセンターのような書き込みに対処できない可能性がある。これは、評価表現とドメイン特徴語の内容の意味を十分判断して取得を行っていないため、意見文としてはノイズとなる文のスコアが高くなる場合が考えられる。

6 おわりに

Web 掲示板に大量の書き込みがある利点を用いて、書き込み文書からドメイン特徴語を学習し、意見文を抽出する手法を提案した。

手法は、同ドメインの掲示板情報から、評価表現など意見文の手がかりになりやすい単語との共起性と、ドメイン特徴語の自動取得の二つの情報を学習した単語データを作成する。この単語データを作成することで、ドメインに依存した辞書を作成することなく意見文抽出を行うことが可能となる。

ドメインに順応した単語データを用いて、2 つの異なったドメインの意見文抽出を行った結果、どちらに対しても評価表現はドメイン依存の辞書を作成しなくとも高い再現率で抽出できることや、本手法により抽出した意見文スコアの上位に、抽出すべき意見文を多く収集することができた。

課題として、ドメイン特有の評価表現を自動取得すること、照応解析を用いたドメインにおける主題の推定などを行う必要がある。さらに、異なったドメインで抽出を行った場合や評価する量を増やした場合に、同様の結果が得られるかを検討することも必要である。

使用した言語資源及びツール

(1) 係り受け解析器「南瓜」,Ver.0.50, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/˜taku/software/cabocha/

(2) Yahoo!掲示板,http://messages.yahoo.co.jp/

(3) 検索エンジン Google,http://www.google.co.jp/

参考文献

[1] 立石健二, 福島俊一, 小林のぞみ, 高橋哲朗, 藤田篤, 乾健太郎, 松本裕治 : Web 文書集合からの意見情報抽出と着眼点に基づく要約生成, 情報処理学会研究報告,NL163-1,pp.1-8,2004.

[2] 藤村滋, 豊田正史, 喜連川優 : 電子掲示板からの評価表現および評判情報の抽出, 人工知能学会全国大会, http://www-kasm.nii.ac.jp/jsai2004_schedule/paper-192.html,2004.

[3] 峠泰成, 大橋一輝, 山本和英:繰り返し学習を用いた話題に順応する意見文抽出, 情報処理学会研究報告,FI-77-5,pp.43-50,2004.