2004－FI－77　(5)
2004／11／26

繰り返し学習を用いた話題に順応する意見文抽出

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

意見文であるか否かのタグつきデータをもとにタグなしデータを学習し、ある話題に対する単語データを作成することによって、Web 掲示板から意見文を抽出する手法を提案する。タグ付きデータを学習した単語データ、評価表現や強調表現などの重みづけや主題の自動取得による重みづけによって、タグなしデータに対して意見文かどうかのスコアを算出する。そのうち上位 5% と下位 50% は、それぞれ意見文であること、意見文でないことが判断できると考えそれぞれを学習に用いた。学習によって作成した単語データを用いることによって、最初のタグ付きデータの単語データのみでの抽出結果に比べて、有効性を確認することができた。また、Web 掲示板を少量ずつ繰り返し学習をすることで単語データを作成する手法の方が、まとめて学習を行う方法よりも良い結果を得られた。

Extracting Opinion Sentence Adapted to Topic using Iteration Learning

Yasunari Touge, Kazuteru Ohashi, Kazuhide Yamamoto

Department of Electrical Engineering, Nagaoka University of Technology

E-mail:{touge, ohashi, ykaz}@nlp.nagaokaut.ac.jp

This paper describes a method of extracting opinion sentences from the Web board using iteration learning. We extracted an opinion sentence by creating the word data to a subject by learning from a Web board. These words are weighted for learning of word data by evaluation expressions, emphasis expressions, and themes which are automatically acquired. We performed scoring to create word data, and both highest 5% and lowest 50% of this data are learned again. Effectiveness was able to be confirmed by learning the word data compared with the extraction result of initial word data.

1　はじめに

近年、Web の普及によって、多くの人が手軽に様々な情報を取得できるようになってきている。この情報の量は日々増加しており、テキストデータのみであっても大量の情報となっている。このような大量の情報の中から有益な情報を発見するマイニング技術が、情報を有効に活用するための一つの重要な方法となっている。例えば、企業は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常に注目している。情報を取得するために企業自身によるアンケートを実施したり、アクセス解析をしたりと様々な手法を用いている。また、個人でも、自分が購入しようとしている製品について、その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。

このような状況から、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。このような情報源の例として Web 掲示板が挙げられる。Web 掲示板では、ある話題に対して多くの人が自由に書き込みを行っている。例えば、車の掲示板などでは次のような文が見られる。

例 1) エンジンも静かでスポーティでいい。

例 2) 最悪なのは、リア、あんな安っぽいリアは他の車でもない。

Web 掲示板を用いることによって、個人の興味、関心やある製品に対して他人がどのような印象をもっているかといった情報を大量に取得できる。企業にとっても、Web を身近に扱う個人にとっても非常に有益な情報となる。しかし、目的の情報を探すために、大量の掲示板や Weblog を読むことになり、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文を効率よく取得する手法を提案する。Web 掲示板で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自動で取得するためには多くの知識が必要となる。この知識としては、ドメイン特有の用語辞書や評価表現などの様々な表現辞書がある。これらの辞書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、意見文を判断するために重要となってくる単語を学習し、ドメインごとに辞書を作成する作業を行わずに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。

まず、意見の収集としては、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、3 つ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。

意見情報の肯定、否定の分類をする研究もいくつか行われており、Dave et al.[2]、藤村ら [3]、Turney[4] が提案している。

Dave et al.[2] は、Web 掲示板から評判抽出を行う方法として Web 掲示板の文書の肯定・否定による分類を行って、その結果を用いて抽出する方法を提案している。同様の手法として、藤村ら [3] 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、分類を行っている。その肯定あるいは否定に分類した結果は、個人の評判情報を持っているという考えによって評判抽出を行っている。 ここで、肯定・否定の分類は良い結果を得られているが、文書による分類を行っているので、文単位での判定が行えない。

また、意見情報を抽出するために重要となってくる評価表現辞書をブートストラップアルゴリズムを用いて作成する研究も行われている [5] [6]。評価表現は、ドメインによって様々であり、人手によってすべて収集することはかなり困難となる。従って、ブートストラップアルゴリズムのように、繰り返し効率的に作業する方法が有用となる。

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 掲示板の書き込みで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する。また、意見文を判別するための単語の強さという指標をもとに、単語による文へのスコアリングを行い、意見文を取得する方法を提案する。

3 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。実際の Web 掲示板からは、次に示すような文が意見文として抽出することができる。

- 対象 , 属性 , 評価 の表現が含まれる
  例 3) { エスティマ }$_{対}$ の { 乗り心地 }$_{属}$ は { 良い }$_{評}$ です。
- 対象 or 属性 , 評価 の表現が含まれる
  例 4) { 荷物 }$_{属}$ も { たっぷり積める }$_{評}$ し、{ 燃費 }$_{属}$ も { 良い }$_{評}$ です。
- 評価 の 表現 のみが含まれる
  例 5) とにかく { 静か }$_{評}$ です。

例 3 や例 4 に示すように、意見文は対象、属性、評価の組み合わせの表現によって表されることが多い。しかし、Web 掲示板では個人が好きなように書き込みを行うため、同じ製品であっても表記揺れがなされていたり (例 : エスティマ = アエラス = グレード L)、主語が省略されていたり、顔文字のみで書かれるといった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、これらの表記ゆれに全て対応することはかなりの負担となる。よって、これらの表現の自動取得も考慮しながら、意見文の抽出を行う必要がある。また、主語が省略されている際にはそれを補う必要があり、文単位で扱う場合にもいくつか難しい課題が含まれている。

今回は、文単位で処理を行い、実際に文そのものは意見文であるが、評価の対象となる表現が含まれていない文については、意見文として判断しないこととした。よって、例 5 のように、評価表現のみで意見文となる文については、意見文として判断していない。

以上の基準によって、次節の提案手法により意見文を抽出することを試みた。

4 提案手法

4.1 処理の流れ

本稿での処理の流れを示す。まず学習部の処理の流れを図 1 に示す。初めに、文単位で意見文か否かのタグのついたデータを用い、そのデータに含まれる単語が意見文を判別するためにどの程度の影響をもっているのかを計算し、初期単語データとして作成する。

この値を初期値とし、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、学習を行う。

初期単語データと評価表現、強調表現、文末表現、主題のそれぞれの重みを用いて文単位で意見文スコアを付与する。意見文スコアが高いほど、その文が意見文である確率が高い。この意見文スコアをもとに、上位 5% に含まれる文と、下位 50% に含まれる文を学習し、初期単語データを更新していく。以後この作成した単語スコアのデータを単語データと呼ぶ。

Web 掲示板には、書き込みが大量に存在するということが大きなメリットである。これをもとにドメイン依存のない話題に順応する単語データを作成することを目標にする。その方法として、同ドメインのデータを一度にすべて学習する方法と、いくつかに分けて繰り返し学習する方法が考えられる。繰り返し学習をすることにより、一度にすべてのデータを学習する方法に比べ、新しく出現した単語に対してのスコアを考慮しながら新たな単語データを作成できると考える。このそれぞれの方法を行う。

学習部において作成した単語データを用いて、意見文取得部で自分の得たい情報の書かれた Web 掲示板から意見文の取得を行う。

これによって、すべての書き込みを読むことなく、意見文を収集することが可能になる。以上の一連の流れによって意見文抽出を行う。

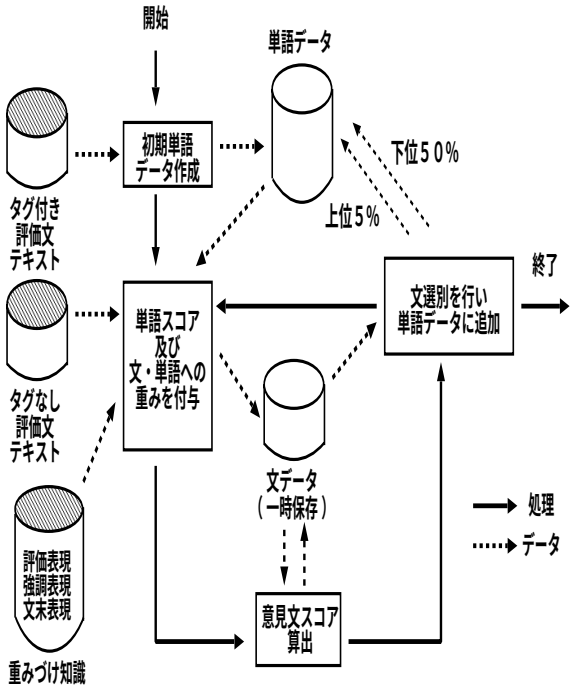

図 1: 学習部の処理の流れ

4.2 単語スコアの算出と単語データの作成

文に出現する単語に対して意見文を判別するスコアを付与するため、意見文か否かのタグ付きデータを用いて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。

- 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 ：7895 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。

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

$W_s(w_i)$: 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例を、表 1 に示す。

表 1: 単語データの例

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学習部での単語スコア付与のベースとする。

4.3 単語、文への重みづけ

単語データをもとに、入力された文・単語に対してスコアを付与する。この時、意見文の判別に大きな手がかりとなる表現に対して重みづけを行う。重みづけは、評価表現・強調表現・文末表現・主題に対して行う。

4.3.1 評価表現への重みづけ

意見文を判別するために、評価表現は大きな手がかりとなる。例えば、

例 6) ハンドリングも { 軽い }$_{評}$ 感じ

例 7) ブレーキの電球が丸見えで { 安っぽくなる }$_{評}$。

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。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、次の通りである。

- 評価表現数 ：510 表現

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。

4.3.2 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を用いる。これらの表現を強調表現と呼ぶことにする。例えば、次に示すようなものが挙げられる。

例 8){ ちょっと }$_{強}$ 足が堅い快適セダンですね。

例 9)TTE は { とっても }$_{強}$ 魅力的ですね

強調表現には、副詞を中心に人手で収集した。強調表現の数は次の通りである。

・強調表現 ： 75 表現

強調表現については、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。

4.3.3 文末表現への重みづけ

意見文を判定する際に、文末を考慮することも重要になってくる。「〜でしょうか？」などの疑問表現、「〜のはず。」などの推定表現を含んでいる文は、意見文にはなりにくいと考えた。よって、これらの文末表現に対しても重みを加えた。文末表現については、表 2 ような 23 表現を用いている。

表 2: 文末表現への重みづけ (例)

文末表現	倍率
〜でしょうか？、ですか？、すみません	0.5
はず (です)、ですよね。、かな	0.7
けれど、 かも 、らしい	0.8

4.3.4 主題に対する重みづけ

意見文かどうかを判定するために、その文での主題となっているものを特定することは、3 つ組による手法 [1] でも行っているように必要になってくる。ドメインに依存しないためには、これらの表現を自動取得することが望ましい。

例えば、車の掲示板では、次のような文が見られる。

例 10){ ハンドル }$_{主題}$ に重い分銅を付けているようです。

例 11){CD}$_{主題}$ の使い勝手もなかなか良いですよ。

例文はある車のハンドルや CD についての意見文である。ハンドルや CD デッキなどは、車の掲示板において多くの人の意見の集まる部分でもある。このようにその文で主題となる単語を含んでいるかどうかということも意見文を判断するのに必要な要素となる。

そこで、掲示板から主題を抽出するため、検索エンジンの「Google」での検索件数を用いる手法を行った。

情報を取得したい掲示板 (以下、対象掲示板) の話題を与え、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。

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

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

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 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 以上の場合には、主題として扱う。そして、主題の出現の仕方によって文全体への重みを加えることにした。文のスコアに対する重みは、表 3 に従って行った。

表 3: 主題への重みづけ

表現	倍率
主題 $\geqq$ 0.1 と 主題 $\geqq$ 0.1 を含み評価表現あり	1.0
主題 $\geqq$ 0.1 と 評価表現あり	0.8
主題 $\geqq$ 0.01 と 評価表現あり	0.8
評価表現あり、主題なし	0.5
主題あり、評価表現なし	0.2
主題、評価表現なし	0.1

4.4 意見文スコアの計算

以上の重みを考慮して、単語データから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。その場合は、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。

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

Average：文に含まれる全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、文に含まれる単語数の違いを考慮して、単語データの平均値を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。

意見文スコアを付与した結果は、図 1 での文データであり、入力されたタグなし学習データは、意見文らしさの値をもった文データとなる。文データの例を表 4 に示す。

表 4: 文データの例

入力文	意見文スコア
静かなのも手伝って、スピード感が殆ど無いですね（良い意味で)。	2.009
ペイントシーラントいいですねぇ。	1.924
逆に今あるストックを提示して貰えば話が早そうですね。	1.120
それともステレオとの組み合わせで決まるのですか？	0.816

4.5 繰り返し学習

意見文スコアを付与し、文を降順に並び替える。意見文として信頼度の高い上位 5% は意見文として、意見文としての信頼度の低い下位 50% は意見文ではないとして扱い、式 (1) により再計算を行う。これ以外のデータは、意見文を判別することが難しいため扱っていない。また、上位と下位の扱いについては、予備実験の結果を元に判断している。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する。

ここで、4.1 節で述べたように、Web 掲示板を学習する方法として、同ドメインの書き込みをまとめて学習する方法と、同ドメインの書き込みを少しずつ繰り返し学習を行う方法がある。

繰り返し学習を行うことによって、まとめて学習を行う方法に比べて、学習していくサイクルの間で、新出してくる単語へのスコア付与が行えるのではないかと考えられる。この考えに基づいて、繰り返し学習を行い、徐々に単語モデルを作成していくことも行った。

また、もとの単語データを学習により更新することによって、同ドメインの単語を学習していくことができる。これによって、ドメイン依存の問題を解決していくことを考えている。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。

4.6 意見文取得部

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。

まず、情報を取得したい掲示板の書き込みを入力する。入力データを文ごとに分割し、文に対して作成した単語データにより単語にスコアを付与する。この過程は、単語データを作成する場合と同様である。今回は、学習により単語データの更新を行う方法がどの程度有効であるかを示すため、作成した単語データのみによる意見文スコアを算出している。よって、評価実験の結果には、重みづけを考慮していない。

単語へのスコア付与を行ったあと、意見文スコアを算出し、文のスコアにより並び替えを行い、意見文の取得を行う。今回、意見文のみを抽出するための閾値をまだ設定していないため、実際に抽出を行う際には設定する必要がある。

5 評価実験

5.1 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である。この中より、 学習データと正解データの二つを用意した。学習データは、話題に順応する単語データを作成するために使用するデータであり、繰り返し学習の有効性を確かめるために、5 分割のデータを用意した。各集合のデータ量は表 5 に示す通りである。それぞれのデータは、Web 掲示板の話題 (イストやハリアーなど) 1 つ 1 つの書き込みを用いている。

また、正解データとして、学習データとは別の書き込みデータから 1064 文を抽出し、意見文か否かのタグを付与した正解データを作成した。意見文は 150 文、そうでない文が 914 文である。これらのデータを用いて次節の実験を行った。

表 5: 学習データ

書き込み A	10476 文
書き込み B	12792 文
書き込み C	12738 文
書き込み D	15740 文
書き込み E	12017 文

5.2 実験方法

Web 掲示板の書き込みを学習することによって、単語データの単語が意見文かどうかを判別する値を獲得できているかを評価するために、学習方法を変えて実験を行った。その方法は次の 3 つである。

方法 1：5 つの書き込みを 1 回で学習し、単語データを作成

方法 2：単語データの単語の増加量が大きくなる順に 1 つずつ学習し、新しい単語データを作成

方法 3：単語データの単語の増加量が小さくなる順に 1 つずつ学習し、新しい単語データを作成

方法 1 は、実験の学習データ 5 つを一回で学習し、初期単語データから単語が増加することにより、同ドメインの単語に対して単語データが順応しているかを検討する。

方法 2,3 は、繰り返し学習の有効性を検討する。5 つの学習データを用いて、単語データの単語の増加量が多い順と少ない順に分けて、どのような結果の違いが現れるのか検討した。

5.3 実験結果

適用した学習方法それぞれについての結果を示す。

方法 1

方法 1 を適用した場合の結果を図 2 に示す。図 2 で、

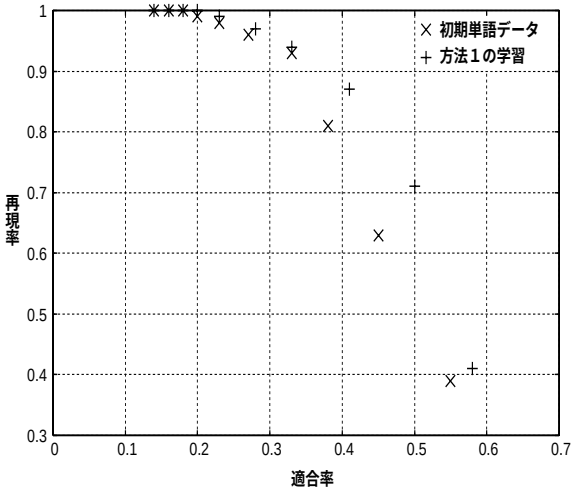

図 2: 方法 1 の適合率、再現率

右のプロットから全体の割合 10% ずつのデータを示している。初期単語データをもとに 5 つの学習データを一度に学習した場合の結果では、初期単語データに比べ上位に意見文が集まり、学習の効果も表している。

表 6: 話題「イスト」との関連度計算結果

主題候補	関連度
ヴィッツ	0.299
カローラ	0.224
プリメーラ	0.114
リアスポイラー	0.069
ワイパー	0.052
トルク	0.047

表 7: 方法 1 の単語データの変化

	初期単語データ	方法 1
総単語数	7895	17424
主題数 (195)	123	171

また、表 6 では、単語データを作成する際に主題を抽出する処理を行い、話題と主題候補との関連度の結果の例を示す。話題はイスト (車名) である。これによって、学習の際にイストとの関連語を主題として扱うことができる結果となった。表 7 では、方法 1 を適用した場合の単語数の変化を示している。初期単語データの単語数が 7895 単語であったが、提案手法によって学習することによって 17424 単語に増加する。

実際に単語データがドメインに順応しているかを確認するため、正解データの中から主題を取得した。その結果、195 単語が正解データから主題として扱われるが、学習した後の単語データでは、初期単語データに比べ正解データの掲示板の書き込みを学習しない段階で、その主題の 171 単語をカバーできている。このように、主題数が増加していることで、ドメインの問題に対応できるようなデータになっていくことも確認できた。よって Web 掲示板学習による単語データの作成は有効であることがわかった。

方法 2

方法 2 を適用した場合の結果を図 3、図 4、表 8 に示す。

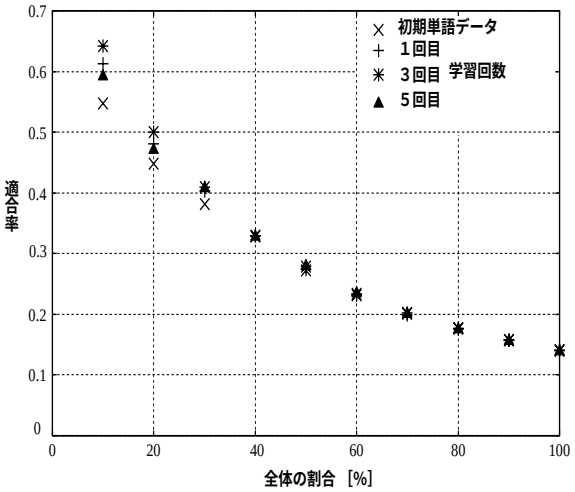

図 3: 方法 2 の繰り返しによる適合率 (小 → 大)

初期単語データをもとに 1 つの学習データを学習をするごとに単語データの単語のスコアを更新していく方法であり、図 3、図 4 では、学習回数別の結果を示している。

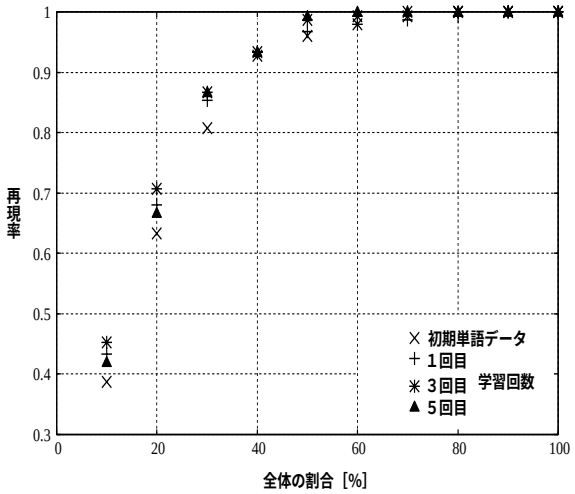

図 4: 方法 2 の繰り返しによる再現率 (小 → 大)

表 8: 方法 2 の単語データの変化

繰り返し数	方法 2 の単語データ数
0	7895
1	9887
2	11950
3	13500
4	15518
5	17339

学習を繰り返すたびに方法 1 と同様に単語数も増えており、主題数も徐々に増えていく。精度に関しては、3 回目の学習までは徐々に適合率も上がっており、学習による結果が現れているようである。しかし、4 回目以降は、精度の向上が起きず、上位の意見文として現れていた文のスコアが下がり、下位の方向にいく傾向が見られた。

方法 3

方法 3 を適用した場合の結果を図 5、図 6、表 9 に示す。

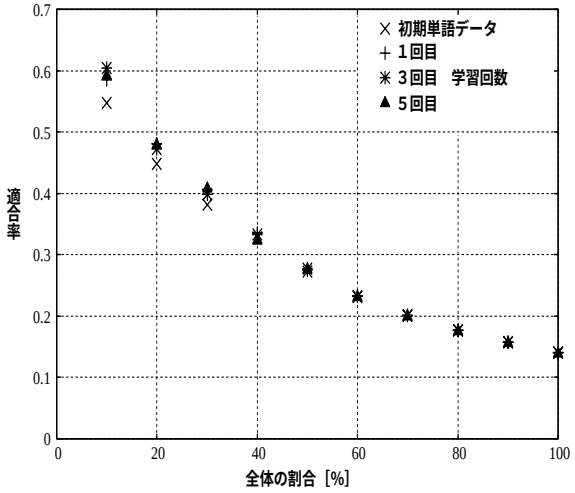

図 5: 方法 3 の繰り返しによる適合率 (大 → 小)

方法 2 と同様にグラフでは、1 回ずつの学習による結果を示している。結果より、学習ごとに方法 2 と同じよ

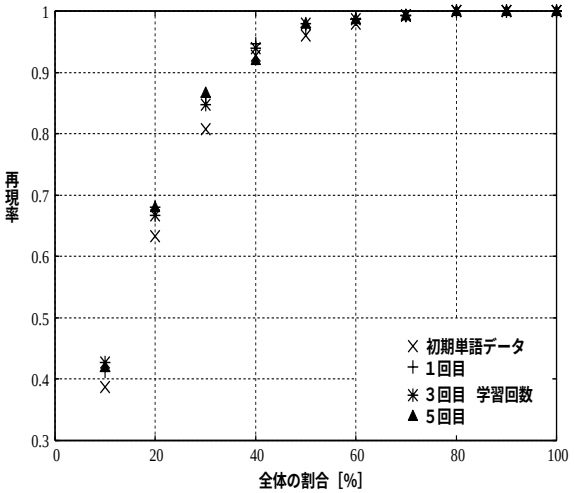

図 6: 方法 (3) の繰り返しによる再現率 (大 → 小)

表 9: 方法 3 の単語データの変化

繰り返し数	方法 3 の単語データ数
0	7895
1	11247
2	13686
3	15148
4	16504
5	17347

うに、単語数も上昇し、精度にも改善が見られた。こちらは、最初に単語データの単語数が増加する量の大きいデータを学習することを行ったが、スコア上位では、3 回を超えてからの学習で精度はあまり向上はしなかった。

これらの結果より、それぞれの学習によって意見文を抽出する精度に向上が見られることが分かった。特に、繰り返し学習をすることによって、意見文を抽出する精度に変化があり、少量を繰り返し学習する方法が良い結果を得られるという傾向がある。

6 考察

6.1 学習について

意見文を抽出するために、意見文か否かのタグを付与したデータとタグなしデータとの二つの情報を用いて意見文判定を行うための単語データの学習を行い、抽出精度の向上を試みた。学習する際に、意見文である信頼性の高い文と意見文である信頼性の低い文に分けて学習を行ったが、ここで、学習の精度の問題が挙げられる。

単語データを作成する際、意見文スコアを単語データやそれぞれの重みにより付与し、上位 5% 以上となった文を意見文、下位 50% となった文を意見文でないとする学習データとして行った。予備実験による結果では、意見文スコアを付与した文の上位 5% を抽出した際の適合率が 8 割前後であった。そのため、学習を行い単語データを作成した場合、2 割の文は、誤ったデータとして扱われてしまう。よって、この 2 割を学習することによる精度の低下が考えられる。

本手法では意見文を抽出する際に、意見文であるとする閾値を設定していないため、抽出する量が設定できない。予備実験では、意見文は、1 つの Web 掲示板の書き込み中で、全体の 1 割〜3 割程度であることが多かった。よって、実際に入力された文書データのうち、どの程度意見文が含まれているか判断するために、閾値を自動で設定できることが良いと考えている。理想的には、意見文スコアを上位から取得していき、意見文が含まれなくなったところを閾値と扱うことができれば、必要な情報のみを取得できることとなり、非常に有用となる。

実際に、上位 5% の文での精度が 8 割程度であり、意見文として扱い学習するデータも信頼性が高いこの辺りのデータまでとなってしまう。さらに多くのデータを学習するためにも、意見文スコア上位の意見文の精度を向上させることが望ましい。

抽出精度の向上には、いくつか方法が考えられる。単語へ付与するスコアに対していくつか重みを加えているが、この重みをヒューリスティックに設定しているため、この与え方を変更することが挙げられる。また、今回は、対象表現や属性表現となる単語を主題という大きな枠で捉えている。この主題については、文単位で処理をする我々の手法では大きな意味をもってくる。抽出精度向上のためには、主題のスコアによる重みをもう少し考慮する必要がある。

また、下位の学習については、意見文のスコアの下位 50% の文を学習している。主題に重みづけを行った際に、「静かです。」「良いですね。」などの主題が現れなかった文については、今回の手法では、評価表現があるが主題がない文として扱われてしまう。実際には、これらの文に評価表現が含まれ、意見文として扱うことができるため、意見文であるとして学習することが正しい。主題を補完することは照応解析などの手法を取り入れる必要があるため、単に文単位で処理するだけでは、正確に捉えることはできない。今回の学習の重みづけでは、このように、意見文であるが下位に集まるという問題となってしまう。下位に集まった意見文を上位に集めるために、主題と評価表現のみの出現を扱うだけでなく、主題や評価表現に付与されている意見文を判別する値も考慮して意見文スコアに重みを加えることにより、下位にある意見文を上位に含むようにすることが可能になると考えている。よって上位のデータの学習方法と同様に、主題の扱いを重要視する必要があると言える。

方法 1 と方法 2,3 の結果から、Web 掲示板の書き込みを学習して新たに単語データに追加していくことで、主題数が増加した。これまで未知語としてスコアが付与されていた単語に対して、意見文を判定するためのスコアとして値を付与できるようになった。この結果より、意見文抽出精度の向上において同ドメインの掲示板の書き込みを学習することによって、単語データのドメイン依存を解消できると考えることができる。単語データの単語のスコアとして、式 (1) のような意見文である強さのみの値ではなく、意見文にならないというスコアも考慮することが考えられる。単語のもつ情報量 (エントロピー) を用いてスコアを付与する方法によって、意見文になるか否かの判定にどの程度の情報があるかを与えることも検討している。

また、単語単位での扱いだけでなく、ある単語と単語の共起性により、意見文となるといった特徴を考慮することも有効な手段であると考える。

6.2 抽出精度について

評価実験の結果より、上位 10% の抽出精度を比較したところ、方法 2 において一番良い結果が得られた。これ

は、学習データを単語データの単語の増加数が小さい順に学習を繰り返すことにより単語データの単語数を増加させていく方法であるが、3 回を超えたあたりから、精度があまり向上しない傾向が見られた。これは、学習データをスコアづけし再計算を行う際に、上位、下位それぞれのデータに誤ったデータが含まれていることや、必要である単語がある程度の量に達しており、過学習になっていることも考える。ただし、それほど大きく精度が下がる訳ではないため、単語のスコアが安定してきているのではないかとも考える。

今回の手法では、初期単語データに含まれている単語によって学習する程度が変化する恐れがあるため、このベースの単語データに依存しないような方法を取り入れることが望ましい。上位、下位それぞれの抽出精度の向上ができれば、このような初期単語データの作成は可能であると考える。学習による有効性は確認できたため、初期単語データを自動作成することも検討している。今回の実験では、単語データを作成し、この有効性を示すため、単語データのみの意見文抽出の結果を示した。方法 1 では、その有効性を示すため主題数の増加を表 7 に示した。これより、書き込みの学習によって単語データの単語数が増加、ドメインに順応するような主題を取得できることを確認した。単語データの学習による作成で、意見文を判別するために有用な情報になると言える。

方法 2,3 において、繰り返し学習により、徐々に単語データの単語数を増加させ、増加させる際に未知語にスコアを付与できるようにすることが可能であるかを検討した。実際に繰り返し学習によって方法 2 の学習方法が良い結果を示したが、すべての学習を終えた結果は、方法 1 と方法 2 の結果がそれほど大きな変化がないことがわかった。よって、最適な学習量を推定するためにも、少量の学習データを繰り返していくことが有効な手段であると考える。また、意見文判別に有用な単語のスコアの算出方法も重要な要素になってくる。

抽出精度であるが、方法 1,2,3 すべてにおいて、全体の 30% を抽出した場合、意見文の 80% が含まれている結果となっている。下位に存在していた意見文も、学習した単語データを用いることによって、上位に集まる傾向もあり、意見文を収集する場合に有効な手段となることがわかる。

今回は、意見文を抽出する際に、単語データによる抽出のみで行っているため、新出している評価表現などへの重みづけがされていない。実際にこれらの重みづけを加えることによって、さらに上位に意見文を集中させることができるのではないかと考えている。

7 おわりに

意見文か否かのタグがついたデータから初期単語データを作成し、いくつかの重みを用いて、単語データの学習を行い、この単語データを用いて意見文を抽出する手法を提案した。学習には、Web 掲示板の書き込みが大量に存在するという利点を用いて、タグなしの Web 掲示板の書き込みを用い、評価表現、強調表現、文末表現、主題の抽出などを考慮して、意見文スコアを付与し、意見文として信頼できるデータを学習することにより、単語データを作成した。

この単語データを用いて意見文の抽出を行った結果、学習により単語データを作成した結果が初期単語データよりも良い結果を得ることができた。学習方法としては、同ドメインの書き込みを学習するということで、そのドメインに現れる単語に対してのスコアが付与できるようになるという利点も確認することができた。

繰り返し学習については、少量の学習データを繰り返し学習していく手法が一番良い結果を示した。

課題として、学習については、学習する際の学習データの信頼性の向上、ヒューリスティックによる重みづけの改善、初期単語データの自動作成、主題への重みを重要視することなどが挙げられる。また、抽出精度の向上については、重みを加える際の評価表現などの語彙の増加、単語の共起性などいくつか考慮する必要がある。

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] Kushal Dave, Steve Lawrence, and Dvid M. Pennock ： Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, In Proceedings of the 12th International World Wide Web Conference,pp.519-528,2003.

[3] 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.

[4] Peter Turney ： Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,pp.417-424,2002.

[5] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi,Toshikazu Fukushima：Collecting Evaluative Expressions for Opinion Extraction, In Proceedings of International Joint Conference on Natural Language Processing,pp.584-589,2004.

[6] 那須川哲哉, 金山博 ： 文脈一貫性を利用した極性付評価表現の語彙獲得, 情報処理学会研究報告,NL162 - 16,pp.109-116,2004.