概念ベースを用いたWebページからの評価項目の自動抽出

廣嶋伸章　山田節夫　奥雅博
日本電信電話株式会社　NTT サイバーソリューション研究所
{hiroshima.nobuaki, yamada.setsuo, oku.masahiro}@lab.ntt.co.jp

製品などの評判が書かれた Web ページから、製品に関する評価項目とその評価内容の組が抽出できれば、製品を購入する際の参考情報などとして非常に有益である。そこで本稿では、評判が書かれた文から評価項目を自動的に抽出する手法を提案する。提案手法では、評価項目とそれに属する属性表現が記述された評価項目辞書を用意しておき、概念ベースを用いて評価項目に属する属性表現と評判が書かれた入力文との意味的な距離を計算し、その距離をもとに文の評価項目を抽出する。このため、辞書に登録された属性表現が文中に出現しない場合でも評価項目を抽出することができるという特長を持つ。

1. はじめに

近年、ブログや掲示板の普及に伴い、個人が自分の意見を情報発信する機会が増加しており、特に個人が所有する製品などに関する評判が書かれることが多くなってきている。このような製品などの評判が書かれた Web ページから、「何について評価しているか」を表す評価項目と、その評価項目に対する評価内容の組を評判情報として抽出し、評価項目ごとに集計して提示することができれば、その製品に対する世の中での評判の傾向を知ることができ、製品を購入する際の参考情報などとして非常に有益である。

従来の評判情報の抽出に関する研究は、大きく二つに分類することができる。一つは、評判情報のうち、「よい」、「すごい」といった評価内容を表す評価表現に着目した研究である[1] [2]。もう一つは、評判情報を「燃費」などといった属性表現と、「よい」などといった属性表現に対する評価表現の組であると考え、属性表現と評価表現の組を抽出するという研究である[3] [4]。後者は本稿の研究に関連が深いため、後者の研究について概観する。

立石らは、属性表現の辞書と評価表現の辞書を用意し、それらの辞書に含まれる属性表現と評価表現が、「属性表現のあとに助詞があり、属性表現を含む節が評価表現に係る」といった抽出ルールに該当する場合に、属性表現と評価表現の組を抽出するという方法を提案している[3]。このように、属性表現と評価表現の組を抽出することができれば、属性表現をもとに集計して表示することができる。しかし、立石らの方法では、あらかじめ収集された属性表現および評価表現が出現しない文から評判情報を抽出することはできないという問題点がある。属性表現や評価表現の網羅性を上げるために、テキストマイニングにより自動的に属性表現や評価表現を獲得する方法も提案されている[4]。評価表現については、分野によらず同じ表現が用いられることが多いため、その収集は比較的容易である。しかし、属性表現は様々な表現が存在し、分野によって異なるだけでなく、新しい表現が作られることも十分あり得る。このような状況で、全ての属性表現をあらかじめ収集しておくことは非常に困難である。

属性表現は収集が困難なだけでなく、書き手によって様々に表される上に細かすぎるので、世の中の評判の傾向を見るには不向きである。前述の立石らも着眼点という属性表現の上位分類を用意し、この上位分類ごとに属性表現を集計してレーダーチャートを作成している。本稿では、属性表現の上位分類を評価項目と呼ぶこととすると、評判情報を評価項目ごとに集計する際、評判が書かれた文から直接評価項目を知ることができれば、詳細な属性表現について知る必要はない。

そこで本稿では、評判が書かれた文を対象に、概念ベースを用いてその文の評価項目を抽出する手法を提案する。本手法は、評価項目とそれに属する属性表現が記述された評価項目辞書を用意しておき、概念ベースを用いて評価項目に属する属性表現と評判が書かれた入力文との意味的な距離を計算し、その距離をもとに文の評価項目を抽出する。このため、辞書に登録された属性表現が文中に出現しない場合でも評価項目を抽出することができる。

以下では、評判が書かれた文から評価項目を抽出する方法と、携帯電話についての評判が書かれた Web ページに含まれる文から評価項目を抽出した場合の評価結果について報告する。

2. 概念ベースを用いた評価項目の抽出

本稿では、概念ベースを利用することにより、あらかじめ収集されていない属性表現が含まれている文から評価項目を抽出する方法を提案する。具体的には、抽出したい評価項目（例えば、携帯電話に関する評価項目としての「音」など）とその評価項目に属する複数の属性表現（例えば、「音」に属する属性表現としての「着メロ」、「スピーカーの音質」など）が記述された評価項目辞書をあらかじめ用意しておき、概念ベースを用いて入力文と評価項目辞書中の各評価項目に属する属性表現との関連度を

算出し、関連度をもとに入力文がどの評価項目について述べられているかを推定する。評価項目辞書はすべての属性表現が登録されている必要はなく、頻繁に出現する属性表現だけを登録すればよい。このため、製品ごとに異なる評価項目辞書の作成コストを削減することができると考えられる。

以下では、提案手法で用いる概念ベースについて概説し、提案手法の詳細について述べる。

2.1 概念ベース

概念ベースは、単語とそれに対応する概念ベクトルとを収めたデータベースである[5]。概念ベクトルは単語の共起傾向をベクトル表現したものであり、概念ベクトルが近い単語同士は関連が高いと考えられる。そのため、単語の概念ベクトル間の距離を計算することで単語の意味的な距離を計ることができる。概念ベクトルを生成するには、まず学習用コーパスを用いて各単語（自立語）間の一文中における共起頻度から単語の共起行列を生成する。共起行列の各行に対応する単語を概念語と呼び、各列に対応する単語を文脈生成単語と呼ぶ。共起行列の各行が各概念語に対する共起パターンのベクトルとなる。ベクトルの次元数の圧縮とデータスパースネスの解消のために特異値分解（SVD）により行列を変換したのち、長さ 1 に正規化したものが概念ベクトルとなる。

2.2 提案手法

提案手法では、入力文がどの評価項目に関わる内容を表しているかを次のように推定する。まず、概念ベースを用いて入力文と各評価項目に属する属性表現との関連度を算出する。次に関連度をもとに入力文に対する評価項目を決定する。

2.2.1 入力文と属性表現との関連度の算出

評価項目Aを、構成する属性表現A_iによって$A=\{A_1,A_2,\cdots,A_l\}$と表す。また、i番目の属性表現A_iを、構成する単語a_{ij}によって$A_i=\{a_{i1},a_{i2},\cdots,a_{im}\}$と表す。入力文$S$を、構成する単語$w_k$によって$S=\{w_1,w_2,\cdots,w_n\}$と表す。

入力文Sと属性表現A_iとの関連度$r(S,A_i)$を以下の(1)式により算出する。

$$r(S,A_i)=\frac{1}{m}\sum_j\{\max_k d(w_k,a_{ij})\} \quad \cdots(1)$$

ここでdは入力文中の単語w_kと属性表現中の単語a_{ij}の概念ベクトル間の距離を表す。(1)式は、属性表現中の各単語に対して文中の単語との関連が最も高いものとの距離を求め、その距離の平均を計算することを表している。本稿では、ベクトル間の距離としてコサイン距離を用いる。

$$d(w_k,a_{ij})=\frac{\vec{v}(w_k)\cdot\vec{v}(a_{ij})}{\left|\vec{v}(w_k)\right|\left|\vec{v}(a_{ij})\right|} \quad \cdots(2)$$

$\vec{v}(w)$は単語wの概念ベクトルを表す。

2.2.2 入力文に対する評価項目の抽出

関連度の高い順に並べた属性表現の集合を表す評価項目$A=\{A_1,A_2,\cdots,A_l\}$に対して、入力文Sと評価項目Aとの関連度$R(S,A)$を以下のようにして算出する。

$$R(S,A)=\frac{1}{T}\sum_{i=1}^{T}r(S,A_i) \quad \cdots(3)$$

(3)式は、属性表現との関連度が高い上位T個の属性表現の関連度の平均を評価項目Aとの関連度とすることを表している。入力文とすべての評価項目との関連度を求め、関連度が最大となる評価項目を選択する。この評価項目を入力文に対する評価項目として抽出する。

3. 実験

提案手法の有効性を検証するため、携帯電話に関する評判が書かれた Web ページを収集し、これを用いて評価項目の抽出実験を行った。以下では、実験データ、実験条件および実験結果について述べる。

3.1 実験データ

3.1.1 Web ページの収集

15 機種の携帯電話に関する評判が書かれた Web ページ 1,069 ページを収集した。収集の際には、携帯電話の機種名と「評価」、「ランキング」などの単語をクエリとしてインターネット検索エンジンにより検索し、検索結果が上位の Web ページの内容を人手によりチェックし、携帯電話に関する評判がどこかに書かれていればそのページ全体を収集するという方法をとった。

3.1.2 概念ベースの作成

通常、概念ベースは新聞や辞典などのテキストから作成されることが多い。しかし、これらのテキストは、評判情報のような主観的な表現はあまり含まれていない。すなわち、新聞や辞典などのテキストは、評判情報に関する意味的な距離を算出する目的に用いる概念ベース作成には向いていないと考えられる。

そこで、3.1.1 節の方法で収集した携帯電話に関する評判が書かれた Web ページをテキストとした。収集したすべての Web ページのタグを除去して文に分割し、約 80,000 文からなるテキストを作成した。このテキストを用いて概念ベースを作成した。概念

表１：作成した評価項目辞書

評価項目	属性表現の例	項目数	評価項目	属性表現の例	項目数
音	着メロの音質	58	速度	処理速度	20
色	ピンク	22	マナー	バイブの強さ	8
デザイン・形	フォルム	33	ラジオ・テレビ	FM ラジオ	8
画像・画質	カメラの画質	94	本体	端末	14
通信	接続	58	ボタン・キー	ジョグダイアル	23
表示・閲覧	Web の閲覧	73	パーツ	電池カバー	38
入力	文字変換	56	ウェブ・アプリ	i モード	28
価格	本体価格	15	メール	送信メール	17
感覚・尺度	ボタンの操作感	134	ソフト	ゲーム	21
サイズ・容量	フォントサイズ	30	インタフェース	メニュー	36
動作	メニューの操作	68	データ	データフォルダ	3
機能・性能	スケジュール機能	115	メーカ	Disney	4

語として高頻度語約 18,000 語を用い、文脈生成単語として上位 50 語を除く高頻度語 1,000 語を用いた。概念語との共起頻度ベクトルを SVD により 100 次元に圧縮し概念ベクトルとした。

3.1.3 評価項目辞書の作成

収集した Web ページから作成した約 80,000 文のテキストから、人手により属性表現を抽出した。抽出の際には、「属性表現のあとに助詞があり、属性表現を含む節が評価表現に係る」などの属性表現を含むパターンに一致した場合に属性表現を抽出するという方法をとった。属性表現は、「バッテリーの持ち」というように複数の名詞が助詞の「の」で接続された形のものや、「通信速度」のように名詞が連続する形のものも抽出した。抽出された属性表現は全部で 976 種類である。これらの属性表現を人手により 24 種類の評価項目に分類して評価項目辞書を作成した。作成した評価項目辞書に関する詳細を表 1 に示す。

3.1.4 正解データの作成

収集した Web ページから作成した約 80,000 文のテキストから、評判が書かれた約 1,000 文を人手によって選択し、正解の評価項目を付与した。これらの文は 3.1.3 節で述べたパターンに一致しないものも含んでいる。

3.2 実験方法

今回の評価実験では、属性表現が評価項目辞書に登録されていない場合にも評価項目を正しく抽出することができるか否かを検証することを目的とした。このため、次のような実験方法をとった。

(a) 評価表現辞書からある数の属性表現を削除した評価表現辞書を作成。
(b) 正解データのうち削除された属性表現を持つ文のみを選択。
(c) (b)の各文を入力文とし、概念ベースを用いて(a)の属性表現が削除された評価表現辞書の属性表現との関連度を算出。
(d) 関連度最大の評価項目を抽出。
(e) 得られた評価項目と正解とを比較し、正解率を計算。

976 種類の属性表現のうち、辞書から削除する属性表現の数を 100 種類から 900 種類まで 100 刻みで変化させて評価項目の抽出を行った。削除する属性表現はランダムに決定した。正解率は実験の対象とする文数のうち抽出に正解した文数の占める割合である。ただし、どの属性表現を削除するかによって対象とする文数は異なってくるため、それぞれの属性表現削減数に対し、対象となる文から評価項目を抽出するという試行を 5 回繰り返し、対象となる文数の合計のうち抽出に正解した文数の合計の占める割合を正解率とした。

3.3 実験結果

実験結果を表 2 に示す。この結果より、およそ半数の文に対し正しく評価項目が抽出されており、おおむね良好な結果が得られていると考えられる。

表２：実験結果

削除した属性表現数	評価対象文数の平均	正解率(%)
100	42.6	51.6
200	93.2	44.8
300	143.4	51.0
400	208.6	49.8
500	277.0	48.2
600	404.8	51.3
700	490.2	51.6
800	651.2	47.9
900	812.2	42.0

表３：評価結果の例

判定	文	正解	システム出力
○	素敵なアイコンが充実してます。	インタフェース	インタフェース
○	とくにメモリースロットカバー、爪が短い人にとっては地獄ですね	パーツ	パーツ
○	しかもサブディスプレイも大きく，これだけでメールが読める。	表示・閲覧	表示・閲覧
×	メールのメニューアイコンをカスタマイズできない	メール	インタフェース
×	覚悟はしていたのだが、メールは非常に打ちにくい。	メール	入力
×	キー配置は一般的な配置だと思われ、特に使いにくいと感じる事は少ないだろう。	ボタン・キー	感覚・尺度

評価結果の例を表 3 に示す。この表のうち判定が×となっている例を観察すると、提案手法による抽出の結果は、人間の直感では正しいと判断できるにも関わらず、正解と異なるために×と判定されてしまっている例が見受けられる。これは、ある一つの評価項目に属する属性表現が複数の評価項目にあてはまるということが原因であると考えられ、評価項目の設定の仕方に問題があったと考えられる。属性表現が複数の評価項目にあてはまることのないように、適切な評価項目を設定して評価項目辞書を構築する必要がある。

表 2 の結果について考察する。今回作成した評価項目辞書は、属性表現の数が 976 種類と比較的多かったが、表 2 を見ると、属性表現の数を大幅に削減しても正解率の低下が起こらないことがわかる。これは、提案手法による評価項目の抽出性能が評価表現辞書の規模によらず常に同程度であるという可能性を示している。つまり、提案手法のために用意する評価項目辞書は小規模でよいかもしれないということを意味している。作成コストを考えると、評価項目辞書はできるだけ小規模であることが望ましい。属性表現を頻繁に出現するものに限定した評価項目辞書を構築し、どの程度まで小規模にできるかについて調査していきたい。

最後に、アルゴリズムの改良について述べる。表 3 の最後の例は、文中の「感じる」という単語よりも「キー」や「配置」という単語のほうが評価項目として特徴的であるため、「感覚・尺度」ではなく「ボタン・キー」が評価項目として適切である。このような誤りに対しては、単語が特徴的かどうかを考慮することによって正解を得ることができるのではないかと考えている。関連度と単語が特徴的かどうかを組み合わせたアルゴリズムを考えていきたい。

4. まとめ

本稿では、概念ベースを用いて評価項目に属する属性表現と評判が書かれた入力文との意味的な距離を計算し、その距離をもとに文の評価項目を抽出することにより、辞書に登録された属性表現が文中に出現しない場合でも評価項目を抽出することができる手法を提案した。提案手法を携帯電話に関する評判に適用して評価項目を抽出する実験を行った結果、評価項目辞書に登録されている属性表現を含まない文の約半数に対して正しく評価項目を抽出できることを示した。

今後の課題としては、評価項目辞書の構築方法についての検討、特に属性表現が複数の評価項目にあてはまらないように評価項目を設定する方法の検討が挙げられる。また、属性表現をクラスタリングするなど、提案手法に適した評価項目辞書を半自動で構築できる手法についても検討課題である。

参考文献

[1] 鈴木泰博, 高村大也, 奥村学. Weblog を対象とした評価表現抽出. 人工知能学会セマンティックウェブとオントロジー研究会, SIG-SWO-A401-02, 2004.

[2] 藤村滋, 豊田正史, 喜連川優. Web からの評判および評価表現抽出に関する一考察. 信学技報, Vol.104, No.177, 2004.

[3] 立石健二, 福島俊一, 小林のぞみ, 上出将行, 高橋哲朗, 乾孝司, 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. 言語処理学会第 10 回年次大会発表論文集, P5-5, pp. 644-647, 2004.

[4] 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, NL154-12, pp. 77-84, 2003.

[5] T.Kato, S.Shimada, M.Kumamoto, K.Matsuzawa. Idea-Deriving Information Retrieval System. Proc. of 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp.187-193, 1999.