

旅行者レビューからの評価情報抽出における評価文型パターンの有効性の調査

An Investigation of Effectiveness of Estimation String Pattern to Extract Sentiment Information from Customer Reviews of Tourists

岡田 真*¹
Makoto Okada

竹内 和弘*²
Kazuhiro Takeuchi

橋本 喜代太*¹
Kiyota Hashimoto

*¹ 大阪府立大学
Osaka Prefecture University

*² 大阪電気通信大学
Osaka Electro-Communication University

Generally, a customer review contains several subjects and their values that show sentiments and estimation of authors of the reviews. Total estimation described by the review should be obtained by these sentimental data comprehensively. Therefore, in order to improve a precision of the sentiment estimation of the reviews, an estimation system must deal with the pairs of the subject and the values appropriately. In this paper, we defined some special patterns of the pair of the subject and the value contained in sentences of the reviews as estimation sentence patterns. After that, we also investigated and discussed effectiveness of the estimation sentence patterns based on experimental results of extraction of information using these patterns from the reviews.

1. はじめに

ウェブ上のさまざまなコマーシャルサイトには製品やサービスに対する利用者からの意見がカスタマーレビューとして多数投稿されている。それらのカスタマーレビューからレビュー筆者の意見や評判情報などを抽出・利用する研究は近年盛んであり、ユーザとサービス提供者の双方から注目されている。

本稿では、カスタマーレビュー中のレビュー筆者の評価に関する意見情報を抽出する評価情報抽出処理のために、評価文型パターンと呼ばれる内容語と付属語を組み合わせたパターンをレビュー文書中の評価情報抽出に用いた場合の有効性について考察する。評価文によく含まれる要素として形容詞と形容動詞がよくあげられるが、我々はそれらが実際のレビュー文書中にどのように含まれているかを調査することで、評価情報抽出の対象がどのような特徴を持つ文であり、それらに対して評価文型パターンがどの程度有効か、また適切な処理のための問題は何かを考察する。

2. 評判分析

評判分析は前述のように近年のウェブ上での活発な商取引により注目されている研究分野である[那須川 2006]。

インターネット上にある商品や料理店やホテルなどについて口コミ情報が存在する。その多くは好評不評の総合的な評価の情報をもち、それは星の個数などの点数であることが多い。

しかし、それらの点数のつけられた基準や経緯は口コミの内容を読まなければわからない場合が多い。口コミの筆者にとって評価の対象や基準となるポイントはどこなのか、加点や減点の理由はなぜかなどは口コミそのものを読んで判断する必要がある。

そこで、文書そのものの内容について分析を行う評判文型の手法が重要となる。

評判分析を実現するうえで最初に必要となるのは好評不評を判断するための表現の定義である。それらは文章中の表現から抽出される。解析の際には、対象中の文書でそれらの表現がどのように表れるかを調査し、それをもとに、文や文章全体がどのような評価をしているか推定する。

一般に評価分析用の表現としては、「良い」「悪い」「最高だ」「最低だ」など形容詞やナ形容詞(形容動詞)が用いられることが多い。そのような表現を判定の中心とし、次にそれらの前後の表現を調べて、文脈に応じて判定を修正する。たとえば評価表現の否定表現が含まれれば好評不評を反転させる処理を行い、条件や仮定の表現が付け加えられていれば、それに応じて好評不評の反転やその強さの調整を行ったりする。

文脈によっては、商品やサービスに対する評価として肯定と否定が入れ替わる場合がある。たとえば映画などにおいては悲しい映画は悲しいと評価されることが好評となるが、ホテルの評価で悲しいとあればそれは否定的な評価と考えるのが妥当である。

このように評判分析を行う対象がどのようなものであるか考慮しつつ評価用の表現の評価値について適宜修正を加えていく必要がある。

一般に文章は文脈一貫性が保たれている場合が多い。特別な記述がなければ、前の文の評価が維持されると推定するのが自然である。文脈が変わる場合には、反転の接続表現が加えられることが多い。そこで、文書の評価情報を適切に判定するために、評価語だけでなく、文書中の接続表現にも注意を払う必要がある。

評価表現をもとにした評判分析手法においては、評価表現の有無に加え、構文情報も重要な要素となる。

評価表現の抽出が終わった段階で処理をやめ、文書データを単語や複合語のみで評価の推定を行おうとすると、評価対象と評価語の対応情報などが欠落し、正しい評価を行えない恐れがある。その解消のためにはもう一段階深く、語と語の関係を反映させた係り受けレベルの内容まで抽出する必要がある。

単語間の係り受け構造を適切に利用することにより、評価表現がどの評価対象について言及しているのかが明確になり、その結果、正確な評価情報の抽出や推定が可能となる。

岡田 真
大阪府立大学 工学域 電気電子系学類 情報工学課程
〒599-8531 大阪府堺市中区学園町 1-1
Tel: 072-252-1161 E-mail: okada@mi.s.osakafu-u.ac.jp

ここで、係り受け解析は文法知識などが必要な解析処理となり、評価表現分析においてその機能を十分に発揮させるためには、利用する側に評価対象に対する知識のみならず、文法構造や係り受け解析手法についての深い理解が求められることになる。このような利用者側への負担の軽減のための一つの手段として、我々は評価文型パターンを利用することを考えた。

3. 文型パターンと評価表現と評価文型パターン

評価情報の抽出処理では、頻出文をもとに構築したパターンを利用する手法が一般的である。

日本語文を構成する要素は、内容的・機能的と言う観点から、主に内容的な意味を表す内容語と、助詞や助動詞といった主に文の構成にかかわる機能語の二つに大きく分類できる。また、複数の語から構成され、全体として一つのまとまった意味をもつ要素もある。これらをまとめて整理すると、表 1 に示すようになる。機能語に関しては、松吉ら[松吉 2007]は機能語と複合辞をまとめて機能表現とし、言語処理において計算機から利用可能な日本語機能表現辞書を編纂している。また、本稿では内容語と複合語をまとめて内容表現とする。

機能表現は、日本語文において内容表現を補助し機能的に働く表現であり、内容表現とともに日本語の文を構成している。文の構造は主語や述語や修飾語などの成分間の関係として考えることができるが、これらの関係と機能表現の結びつき方に特定の類型が認められる。機能表現を中心に、語順を考慮して、機能表現とそれ以外の成分をメタ記号化したものの系列に関して類型化したものを文型パターンと呼ぶ。

機能表現及び文型パターンは、動詞や名詞などの内容語に比べて種類が少なく、新語が生成されにくい。この特徴から、機能表現のみの辞書を整備し、文書中の機能表現部分を特定し、その出現位置を文型パターンに整理する。すなわち、文型パターンは文中の機能表現の出現位置と内容語との文構造中の位置関係の特徴付けられる情報となる。

実際の文書に含まれる頻出する評価文を選び、それらによく見られる表現や単語の組み合わせを文型パターンとして定義する。それらにマッチした語句を抽出し、評価値を求める。評価文型パターンを定義し、それにより評価対象、評価語、評価値など評価関連の諸情報を抽出することができる。以下、評価文型パターンの基礎となる文型パターンおよび評価文型パターンについて述べる。

表 1. 日本語の文を構成する要素

	1 語から構成	複数語の構成
内容表現 (内容的な意味を持つ)	内容語 (名詞, 動詞, 形容詞など)	複合語 (複合名詞, 複合動詞, 慣用句など)
機能表現 (機能的に働く)	機能語 (助詞, 助動詞, 接統詞など)	複合辞 (「ていた」, 「によって」など)

評価文型パターンは以上のような文型パターンの考え方を評価文書分析の目的に限定して整理したものである。具体的には、文書中の筆者の評価に関する表現である形容詞・ナ形容詞に着目し、それらを評価語として、評価表現の文中出現文脈を文書の特徴付けに用いる。

このような評価文型がカスタマーレビューにおける評価語の出現文脈の特徴付けとして有効であるかを調査した。レビューの各文書から句点などで区切られた 1 文を取得し、形態素解析器 MeCab[工藤 2004]と松吉らが編纂した機能表現辞書を用いて、各文ごとに評価文型と比較する。適合した文に印をつけ、そ

の後、レビュー文書全体を人手で調べ、その有効性について調査を行った。図 1 に評価文型パターンと実際の文との比較の例を示す。

関連研究として、評価表現の利用に関する研究があげられ、レビュー文書などのテキスト中における評価表現の分析[乾 2006]や評価表現を利用したクレーム意見の抽出といった研究[乾 2013]など先行研究が存在する。

また、中山ら[中山 2015]は日本語文中に含まれる述語やその他の格を解析する処理である述語項解析において、統語パターンの解析を行っている。中山らは複数の述語項解析を行う

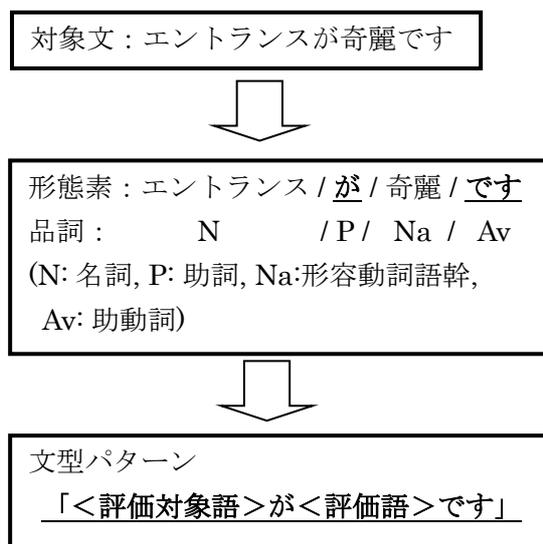


図 1. 評価文型パターンと実際の文の比較例

ためのシステムを提案するために、その前段階として事例の機械的な分類を行っている。その際に分析のターゲットである述語項関係を述語 P と述語への直接かかり語 O と O への直接かかり語 A という 3 つのラベルで表し、述語 P とは直接的係り受け関係にないが意味的に関係がある語 A を P への直接かかり語 O との関係などを利用して関係をとらえることが可能かを、事例をもとに分析している。中村らは一般的な文書に対する調査となっているが、我々の研究の対象はカスタマーレビューであり、対象をより絞り込んだものとなっている。また評価文型パターンは評判分析のために評価表現を中心に機能語を含めた統語情報で構成されたパターンでとなるため、この点でも評判分析に特化したものとなっている。

成田ら[成田 2015]は文章中の事象が実際に起きたことなのか、それとも予想や可能性を述べたものなのかを判断する事実性解析において、機能表現に基づく意味ラベルを設定し、それに基づいてルールベースの事実性解析器を構築して、その有効性について実験をもとに検証している。我々の評価文型パターンも、文書中の評判情報の事実や意見や予想などをパターンとの照合で推定するものであるが、前述のように、評判分析対象のカスタマーレビューの特徴をより強く反映して構築されている点において、一般の文書を対象としている成田らの研究とは異なるものといえる。

4. 旅行者レビュー中の評価表現の特徴

評判情報の抽出のために、旅行者によるホテルなどのカスタマーレビュー中にどのような文が現れるのか調査を行った。

今回は実際のデータとして旅行情報サイト TripAdvisor の日本語レビューを用いた。入出で収集した複数の都市のホテルに

についてのレビュー1,911件について、形容詞または形容動詞を含む評価文型パターンにマッチする文がどの程度含まれるかを調べた。レビューデータの内訳はもともとのレビューにつけられていた総合点が4点と5点のもの肯定的、1点と2点のもの否定的とし、肯定的レビューを1,000件、否定的レビューを911件用いた。その結果、評価表現を含む文は肯定的なレビューでは7,451文中3,862文、否定的なレビューでは9,815文中4,863文得られた。

実際のデータ中の評価文と考えられる文の現れ方を人手により調べたところ、評価表現でとることができた文はそのうちの限られたものだけということが分かった。

評価表現の抽出における問題点として、評価文型パターンの数および種類が実際の文書中の評価文に対応しきれていないということがあげられる。

対応が難しいタイプの文の例として、「***駅からx分です。」のように、形容詞と形容動詞を含んでおらず、事実を挙げている文だが、実際には「駅から近く立地が良い」というような意味を含んでいるタイプの意見文が挙げられる。辻井ら[辻井 2012]はこれらを事実表現と呼んでいる。

これら事実表現に該当する文は、低評価もしくは否定的なレビューでよく見られる。そのようなレビューでは、肯定的な評価のレビューによく見られる「AがBでよかった」というような形容詞を含む評価文をほとんど含まず、不満に感じた事実を羅列し、最後に「もう利用しません」などと言い添えて評価とすることが多い。このように事実表現を羅列するレビューでは、評価と事実を切り分けて扱うことが大変難しい。このようなタイプの文書を適切に評価するには、文脈一貫性などの文脈情報を利用したり、単語や文の内容だけではなく文外の情報を生かした文書外の情報を考慮した評価表現パターンの構築とその利用方法を考案する必要がある。

5. まとめと今後の課題

本論文では、旅行者のカスタマーレビューを用いた評判分析において、旅行者の意見や評判情報を適切に抽出するときに、評価文型パターンを利用する場合の有効性について、調査をもとに述べた。レビュー中の評判情報について説明し、それらを評価文型パターンを用いて抽出する際の問題などについて述べた。

今後の課題として、今回の調査をもとに、適切なパターンを設定する手法や、それらのパターンを用いた評判情報の抽出手法の考案や実験をもとにした有効性の調査などがあげられる。

参考文献

- [那須川 2006] 那須川哲哉: テキストマイニングを使う技術 / 作る技術, 東京電機大学出版局, 2006.
- [松吉 2007] 松吉俊, 佐藤利史, 宇津呂武仁: 日本語機能表現辞書の編纂, 自然言語処理, Vol. 14, No. 5, pp. 123-146, 2007.
- [工藤 2004] 工藤拓, 山本薫, 松本裕治: conditional random fields を用いた日本語形態素解析, 情報処理学会 自然言語処理研究会, Vol. 2004, No. 47, pp.89-96, 2004.
- [乾 2006] 乾 孝司, 奥村 学: テキストを対象とした評価情報の分析に関する研究動向, 自然言語処理, Vol. 13, No. 3, pp.201-242, 2006.
- [乾 2013] 乾 孝司, 梅澤佑介, 山本幹雄: 評価表現と文脈一貫性を利用した教師データ自動生成によるクレーム検出, 自然言語処理, Vol. 20, No. 5, pp.683-706, 2013.

[中山 2015] 中山周, 松林優一郎, 乾健太郎: 日本語述語項構造解析のための統語パターン分析, 言語処理学会第 21 回年次大会, pp. 1028-1031, 2015.

[成田 2015] 成田和弥, 水野淳太, 上岡裕大, 菅野美和, 乾健太郎: 機能表現に基づく日本語事実性解析, 言語処理学会第 21 回年次大会, pp. 1032-1035, 2015.

[辻井 2012] 辻井康一, 津田和彦: テキストマイニングを用いた宿泊レビューからの注目情報抽出方法, 情報処理学会デジタラブラクティス, Vol. 3, No. 4, pp. 289-296, 2012.