2H5-1

評判情報抽出のための評価文型パターンの自動抽出についての考察

A Consideration of a Method to Extract Estimation Patterns Automatically for Sentiment Information Extraction

岡田 真[*1] Makoto Okada
竹内 和広[*2] Kazuhiro Takeuchi
橋本 喜代太[*3] 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
[*3] Prince of Songkla University Phuket Campus

There were a lot of researches about sentiment analysis for costumer reviews on e-commerce sites. In this paper, we devise about a method to extract estimation patterns automatically. We compared experimental results of extracted words using estimation patterns with words used by support vector machine in order to investigate an ability of extraction of the estimation patterns.

1. はじめに

ウェブ上において商業サイトの利用者が購入した商品や利用したサービスの感想をカスタマーレビューとして利用したサイトなどに投稿することが一般的になってきている. これらのレビューには利用者のさまざまな評価や要望が含まれており, これらのレビューデータから評判情報を抽出, 分析する手法に関する研究が盛んになされている[那須川 2006].

本研究では, 評価文型パターンを用いた評判分析を実現するために, 評価文型パターンを分析対象となる文章データ群から自動的に抽出する手法について考察する.

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.

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持ち, 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性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.

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.

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

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の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を特徴付ける情報となる．

実際の文書中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，接続詞など）	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・ナ形容詞に着眼し，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．

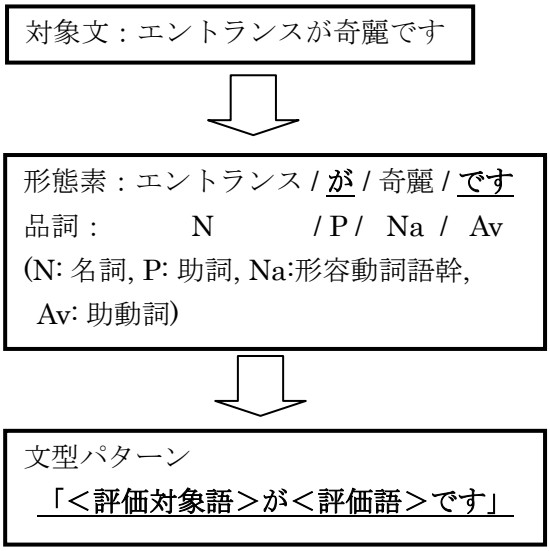

図 1. 評価文型パターンと実際の文の比較例

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. 評判分析と評判情報抽出と評価文型パターンの自動抽出

一般に，評判分析において，評価対象の文書から抽出された評価を種々の手法で集約することで評判分析する．一般的な手法としては，評価語辞書を用いる手法や機械学習を利用する手法などがある．前者は評価語辞書を用いて評価語を抽出して，それらに付与されている評価を集約して推定を行う単語レベルの推定手法である．後者は，一般的には文書を単語単位に切り分け，それらをバッグオブワーズ(BoW)として扱い，その出現頻度などを用いて機械学習による分類器をもちいて推定する手法である．

これら文書を単語単位で扱う手法の問題として，文法情報の喪失とそれに伴う推定誤りがある．先の手法では文を単語単位に切り分けて推定するが，その過程で係り受け情報などの文法

情報は失われる. それにより, 評判情報を含む文中の評価語が, どの評価対象を指しているかという情報も失われ, その結果として, 評価対象と評価の食い違いなどが生じ, 最終的に推定を誤るという恐れが生じる.

上記の文法情報の喪失に伴う問題hに対して, 評価文型パターンによる評判情報抽出は, 文章の係り受けなどの構文的特徴を考慮できる点で有効と考えられる.

評価文型パターンを利用するうえで注意するべき点として, 適切な評判情報パターンの選出と利用があげられる. 評判分析の対象となる文書データ集合は, それぞれの文書に特徴があるため, 評価文型パターンはそれぞれのデータ集合に合わせて調整する必要が生じる. その際に, 適切な評価表現パターンを取捨選択し, 必要ならば新たなパターンを追加しなくてはならない.

評価文型パターンの自動抽出には, 基本的には機械学習手法やテキストマイニングの諸手法の組み合わせで対応できると考えられる. その際に, あらかじめ作成された文型パターンを種として用いて拡張していく半教師あり学習的手法と, あらかじめ文型パターンを準備せず, 単語の共起情報や系列情報などを基に抽出する教師なし学習手法が考えられる.

いずれの場合でも, 頻出する単語情報と, 文型パターンで抽出することができる単語情報の比較が重要となる.

5. 実験

評価文型パターンの自動抽出のため情報抽出のために, 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. 人出で収集した複数の都市のホテル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.

レビューを機械学習手法のサポートベクターマシンにより分類した場合の精度は以下のようになる. 今回はライブラリとしては scikit-learn を用いて分類した. カーネルは RBF, パラメータは C=300, γ =0.001 とした. 5 交差検定を行ったところ, 分類精度は適合率が 0.972, 再現率が 0.989, F 値が 0.98 となった.

その SVM による分類器を基準として, レビュー中のどの単語が有効かを調査することにした. それらの単語と評価文型パターンを用いて抽出された文に含まれる単語との一致率を調べた. サポートベクトルは 1,181 個であり, 9,321 種類の単語が用いられていた. また, すべてのレビューのうち, 評価文型パターンとマッチした文を含むレビューは 1,100 個であり, 9,094 種類の単語が用いられていた. そして, 双方で共通する単語の総数は 7,037 種類であった.

評価文型パターンにマッチする文を含むレビューに現れる単語は, サポートベクトルに用いられている分類に有効と考えられる単語の約 75.5(%)であることが先の結果より示された. ここから, 評価文型パターンは分類に有効なレビューの抽出にある程度は有効であるものの, 約 24.5(%)の単語を抽出可能とすることでさらに改善できる可能性が示された.

また, 評価文型パターンを含む文のみを抽出した場合の単語は, その文を含むレビュー全体より当然ながら少ない. したがって評価文型パターンをより有効に働かせるためには, 評価文型をデータ集合に合わせて調整して, 有益な文をより多く抽出できるようにする必要がある.

評価文型パターンとマッチした文を含むレビューから抽出された単語のうち, 約 25(%)はサポートベクトルに用いられていないことが上記より示されている. この結果も, 評価文型パターンのデータ集合に合わせた調整の必要性を示すものと考えられる.

上記の単語の情報を利用し, 分類に重要である単語を含む文の集合を取得することで, それらを基に頻出文型パターンを抽出することが可能だと考えられる. これを用いることでさまざまな文書データ集合に合わせた評価文型パターンセットの自動生成が可能のとなると予想される.

6. まとめと今後の課題

本論文では評判分析に用いる評価文型パターンの自動抽出手法について, 評価文型パターンにより抽出される単語とサポートベクターマシンによる分類結果を基にして考察した.

今後の課題として, より大規模のデータを用いた評価や, テキストマイニング手法を用いて抽出した頻出単語の集合を利用した場合の有効性についてなどがあげられる.

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.