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

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‡
NEC インターネットシステム研究所†　奈良先端科学技術大学院大学‡

1. はじめに

Web 上の意見は企業における市場調査・新製品開発に重要な情報である。近年、このような Web の掲示板やレビューサイトの意見を抽出・分類する研究が注目されている。これらは主にレビューサイトの意見とそのレーティング情報を学習データとして記事の意見性の判定および肯定・否定の分類の自動化を目的とする[1,2]。

このような先行研究の問題点として、意見の全体像を把握できる機能が存在しないことがある。全体像とは例えば、図 2 のようなレーダーチャートを意味する。図 2 の各軸は意見の着眼点が対応する。軸の値はその着眼点に属する意見全体の内の肯定意見の割合である。言い換えれば、このようなレーダーチャートを生成するのに必要な、着眼点の軸と評価値(肯定/否定)の軸の双方へ同時に意見を分類できる機能が存在しないことが課題である。

我々は、上記の問題を解決する{対象物, 属性, 評価}の 3 つ組を用いた意見抽出分類方式を提案する。本方式では、あらかじめ作成した対象・属性・評価表現辞書と抽出パタンを用いた情報抽出のアプローチを採用して意見を抽出する。抽出した意見は属性表現と評価表現の組み合わせを元に着眼点と評価値の軸に分類し、レーダーチャートを作成できる。

2．意見のモデル

本研究で扱う意見は、表 1 の 3 つ組で定義されるものであり、従来の我々の商品・評価の 2 つ組のモデル[1]を拡張したものである。この定義により意見を抽出する問題は、3 つ組を抽出する情報抽出の問題として扱うことが可能になる。

表 1 意見のモデル

Entity	説明	例
対象物	商品,企業名等	LaVie,NEC
属性表現	評価の着眼点	性能,価格,デザイン,サポート
評価表現	対象物の評価	良い,好き,速い

Web Opinion Extraction and Summarization Based on Product's Viewpoint
Kenji Tateishi, Toshikazu Fukushima, Internet Systems Research Laboratories, NEC Corp.
Nozomi Kobayashi, Masayuki Wade, Tetsuro Takahashi, Takashi Inui, Atsushi Fujita, Kentaro Inui, Yuji Matsumoto, Nara Institute of Science and Technology.

3．システム構成

本システムは、図 1 のように意見抽出部・意見分類部・辞書作成支援ツールから構成される。まず、Web 文書集合からあらかじめ用意された対象物・属性・評価表現辞書と抽出パタンを用いて意見を抽出する（3.2 節参照）。次に、属性・評価表現辞書に付与された分類ラベルを参照して意見を着眼点と評価値の軸に分類し(3.3 節参照)、レーダーチャートを作成する。

属性・評価表現辞書は、[3]で提案した方式を利用して対象物の分野毎に半自動的に作成する(3.1 節参照)。対象物名辞書はシステム利用時にユーザから与えるものとする。

3．1　属性・評価表現辞書作成

辞書作成支援ツールは、属性・評価表現辞書として小規模の初期辞書を用意しておけば、共起パタンを用いて大量文書集合からブートストラッピング的に双方の辞書を交互に増やすことができるツールである。ユーザは、ツールが提示した表現候補の採否を入力するだけで半自動的に大規模な辞書作成を効率的に実施できる。

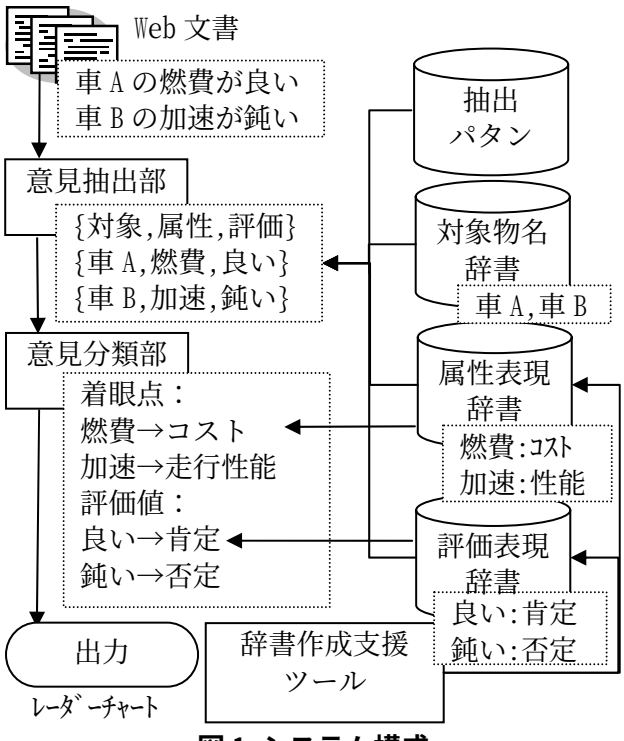

図 1 システム構成

表 2 抽出パタン(属性:属性表現，評価:評価表現，属性は対象物に置き換えたパタンも用意)

抽出パタン	例
(属性:が|は|も|の|に|を|で|φ)→(評価)	デザインが良い,外観良い
(評価:<連体修飾>)→(属性)	良いデザイン
(属性)=(評価)	デザイングッド
(属性:の)→(*:が|は|も)→(評価)	デザインの質が良い
(属性:も|や|と|、)→([*]:も|は|が|で)→(評価)	デザインも広告も良い

このような方式で収集する属性・評価表現に対して、着眼点・評価値の 2 種類のラベルを人手で付与する。着眼点の種類はユーザがその用途に合わせて自由に設定できる。例えば、車の分野では図 2 のような 7 つの着眼点が考えられる。一方、評価値の種類は肯定・否定・中立の 3 種類である。ラベルの付与は、2 節の意見の定義からわかるように原則としては、着眼点のラベルを属性表現(例、燃費：コスト，加速：走行性能)に、評価値のラベルを評価表現（例、良い：肯定、鈍い：否定）に付与する。ただし、共起する属性表現によって評価値が異なる評価表現が存在するため（例、価格が高い：否定，性能が高い：肯定）、それらには属性表現と評価表現の組み合わせに対して評価値を付与する。

３．２　意見抽出方式

意見抽出は 3 つの手順で進める。まず対象物・属性・評価表現の文書内の位置をそれぞれの辞書を参照して検出する。

次に、評価表現を中心として関係する属性表現・対象物を検出し、対象物・属性・評価の 3 つ組を抽出する。この処理は、属性→評価間の関係と対象物→評価間の関係を検出する処理に分けて考える。どちらも、predicate-argument の関係であり、同一の枠組みが適用可能だからである。また、表 2 の係り受け関係を利用した抽出パタンを用いる。これらでは複数の文節に跨る関係も記述している。これらのパタンのいずれかに適合する場合は両者に関係あるとする。

最後に、抽出した 3 つ組の意見性を判定する。抽出した 3 つ組の中には「LaVie のデザインは良いでしょうか？（or 良いらしい。）」や「LaVie のキーボードが少し軽くなれば買いたい。」のように意見とは異なるものも存在する。そこで、評価表現の近傍の条件を示す接続詞、及び質問・伝聞を示す文末表現を考慮し、それらが存在する場合は意見でないとして除外する。

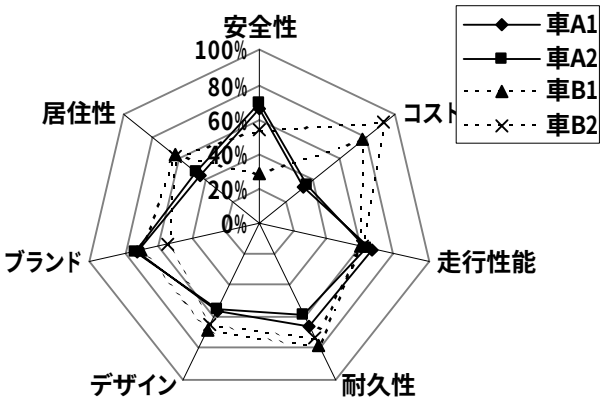

図 2　レーダーチャートの例

３．３　意見分類方式

原則として抽出した意見に付与された着眼点・評価値のラベルに従う。評価値の分類に関して、評価表現の近傍に奇数回の否定表現(例.ない)が存在する場合には評価値を反転する。

４．分析例

3 節の方式を用いて作成したレーダーチャートの例を図 2 にしめす。このチャートは、車に関するレビューサイトの 413 記事から意見を抽出したものである。4 車種を比較しているが、A1 と A2、B1 と B2 は同系列の車であるため類似したチャートの形状を持つことがわかる。A 系列の車は安全性と走行性能、ブランドイメージにおいて B 系列の車よりも評価が高い。B 系列は、コストが特に良く B1 においてはほぼ満点である。このような意見の要点を、商品を比較しながら容易に分析できる。

チャートのために使用した辞書は、同サイトの 1000 記事から約 5 時間で作成できた。1445 表現の属性表現辞書と 1360 表現の評価表現辞書である。評価表現辞書のうち 53 表現は属性表現と評価表現の組に対して評価値を付与した。構文解析ツールには CaboCha(http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/) を使用した。

５．おわりに

本稿では、Web 文書から対象物・属性・評価の 3 つ組の意見を抽出・分類し、レーダーチャートを作成する方式を提案した。今後、意見抽出・分類精度の評価を予定している。

参考文献

[1] 立石健二 石黒義英 福島俊一, "インターネットからの評判情報検索", 情報処理学会研究報告, NL-144-11, pp.75-82, 2001.
[2] Kushal Dave, Steve Lawrence, and David M. Pennock, "Mining the peanut gallery: Opinion Extraction and Semantic Classification of Product Reviews", WWW2003, 2003.
[3] 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, NL154-012, pp. 77-84, 2002.