

Webからの意見抽出システムにおける意見要約方式の評価

立石健二 福島俊一[†]

NECインターネットシステム研究所[†]

松本裕治[‡]

奈良先端科学技術大学院大学[‡]

1. はじめに

インターネットは、誰もが情報を発信できる場であり掲示板や Blog を通してさまざまな人が意見を交換できる。もし、これらの意見を効率的かつ安価に収集・分析できる方法があれば、商品購入の際の情報収集、市場調査等のマーケティング、企業のリスク管理等、幅広い利用が考えられる。

このような期待の元、筆者らは、Web の掲示板等から意見を抽出/分類する意見抽出システムを開発している[1,2]。本システムは、あらかじめ作成した辞書を用いて文書から意見を抽出し、意見の要約を生成する。ここで、辞書の登録率は意見抽出の再現率にかかわるが、人手で作成するため、比較的出現頻度が高い表現は登録しやすいが、存在する全表現を網羅することは困難である。

そこで、本稿では、辞書の登録率がどの程度あれば、本システムの出力である意見の要約の誤差(ぶれ)を許容範囲に収めることができるかを調査する。

2. 意見抽出システム

2.1 方式概要

図1のように、意見抽出システムは意見抽出部とサマリー生成部から構成される[1]。意見抽出部は、下記の3つの手順で意見を抽出し、その結果を意見DBに保存する。

- (1) 3つの辞書(商品名辞書, 属性表現辞書, 評価表現辞書)を参照して掲示板等の文書から、商品名, 属性表現, 評価表現の位置を検出する。これらの辞書は商品分野毎に作成する。
- (2) 検出した3つの表現が正しい対応関係にありかつ、記述者の主観的な評価を示している場合は3つ組{商品名, 属性, 評価}を意見として

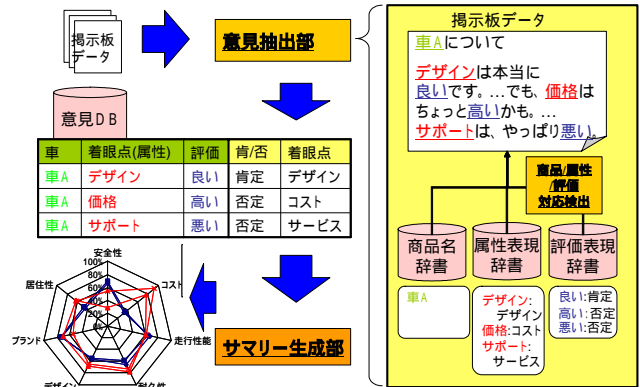


図1 意見抽出システムの方式概要

抽出する。

- (3) 抽出した意見に対して、属性表現と評価表現を手がかりとして着眼点(例, 安全性, コスト, デザイン)のラベルと、肯定/否定のラベルを付与する。

サマリー生成部は、意見DBの意見集合から意見の要約を作成する。ここで、意見の要約とは、ある着眼点に分類される意見全体の内、肯定意見の割合と定義する。これを用いて例えば図1のようなレーダーチャートのような出力が可能である。

2.2 意見の要約の評価

意見抽出システムの出力である要約の精度は下記の3つが要因となる。そのうち、本稿では、(1)の評価実験を行う。なお、(2)については[1,2]で評価している。

- (1) 属性表現辞書と評価表現辞書の登録率
- (2) 意見抽出精度(3つの表現の関係が意見であるかを正確に判定可能か)
- (3) 意見分類精度(肯定/否定及び着眼点の軸へ意見を正確に分類できるか)

3. 実験

3.1 目的

対象分野のすべての属性表現/評価表現を網羅することは困難だが少なくとも意見の要約の誤差すなわち肯定意見の割合の差が一定範囲にお

Evaluation of Opinion Summarization on Web Opinion Extraction System

[†]Kenji Tateishi, Toshikazu Fukushima, Internet Systems Research Laboratories, NEC Corp.

[‡]Yuji Matsumoto, Nara Institute of Science and Technology.

さまるように辞書を準備したい。本実験の目的は、辞書の登録率がどの程度であれば、本システムの出力である意見の要約の誤差を一定範囲に収めることができるかを調査することである。意見の要約の精度は、辞書の規模以外にも、2.2 節にあるように他の要因(意見抽出精度と意見分類精度)も関係するが、本実験では意見抽出と意見分類は完全であるとし、他の要因は考慮しない。

意見の要約の定義は 2.1 節で説明したが、本実験では意見を着眼点ごとに分けず、意見全体の内の肯定意見の割合を意見の要約として用いた。

3.2 実験方法

辞書の登録語を減らしていった場合の意見の要約の誤差を調査するために、実験は下記の手順で実施した。

1. 正解データセットの準備

まず、正解データセットとして[2]で用いたデータを使用した。このデータには車に関する掲示板の 4 車種、計 318 記事に対して、それぞれの意見に該当する箇所に{商品名, 属性, 評価}の 3 種類のタグと、肯定/否定のラベルが付与されている。タグを付与した意見の数は計 2209 件であり、そのうち肯定意見は 1409 件、否定意見は 800 件であった。

2. 属性表現/評価表現辞書の作成

次に、正解データセットから辞書を作成した。属性/評価タグが付与されている箇所を属性表現辞書/評価表現として登録した。

3. 属性表現と評価表現の出現頻度の算出

次に、正解データセットとは別の辞書作成用の記事集合から 2. で作成した辞書の出現頻度を求めた。この記事集合として正解データセットと同一の掲示板の約 14000 記事(正解データセットの記事群も含む)を用いた。

4. 意見の要約の作成

最後に、(3)で求めた出現頻度が閾値以上の表現が辞書に登録されたと仮定した場合の意見の要約を正解データセットから作成した。

3.3 実験結果

図 2 に実験結果を示す。図 2 の横軸は、辞書に登録する表現の出現頻度の閾値を示す。出現頻度が閾値以上に出現した表現を辞書に登録するため、閾値が大きくなるほど辞書の登録数は減少する。一方、縦軸は、意見の要約の値、意見の再現率、属性表現/評価表現の登録率の 4 つを示す。3.1 節でも述べたように、意見の要約とは、意見全体

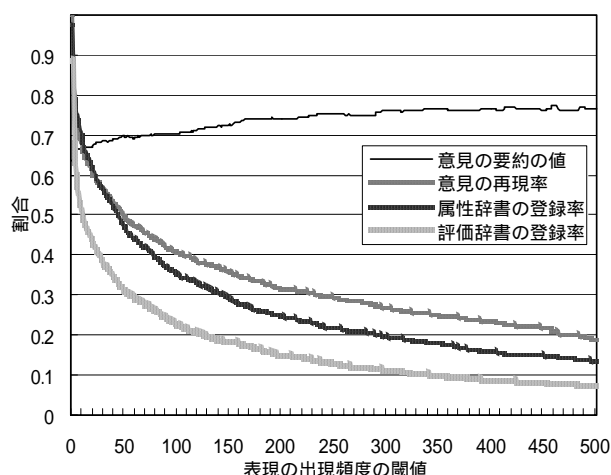


図 2 評価実験の結果

の内の肯定意見が含まれる割合である。

図 2 の出現頻度の閾値が 0 となる箇所は、すべての表現が辞書に登録できた場合を意味する。このときの意見の要約の値は 63.8% である。閾値を大きくするのに従い意見の要約の値が高くなるのは、実験に用いた記事集合では肯定意見の方が否定意見よりも多く含まれるため、肯定的な評価表現は高い出現頻度に集中する一方、否定的な評価表現は低い出現頻度に集中する傾向があるためである。

10%の誤差までを意見の要約の誤差の許容範囲内であると仮定すると、要約が誤差範囲におさまる表現の登録率の閾値は意見の要約の値が 70.0% のときであり、このときの辞書の登録率は属性表現辞書で 40.1%、評価表現辞書で 26.5% であった。つまり、対象分野のすべての属性表現の 40.1%、評価表現の 26.5% がカバーできれば意見の要約を 10% の誤差内に収められる。

4. おわりに

本稿では、意見抽出システムの出力である意見の要約の精度を測るための評価実験の一つとして、辞書の登録率に関する評価を行った。今後は、車分野以外での評価を行うとともに、他の要因(意見抽出精度, 意見分類精度)に関する評価も進める予定である。

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 NL-163, pp 1-8 (2004).
- [2] 飯田龍 小林のぞみ 松本裕治 立石健二 福島俊一, “機械学習を用いた属性と評価値の対の抽出”, 情報処理学会研究報告 NL-165 (2005).