E_006

Webページからの評価表現の抽出

Extraction of Evaluative Expressions on Webpages

石井 基一† 山田 剛一† 絹川 博之†
Motokazu Ishii Koichi Yamada Hiroshi Kinukawa

1. はじめに

インターネット上には掲示板や blog などの Web ページがあり，その中には個人ユーザの商品やサービスに対する意見や感想を含むものがある．それらの意見・感想など情報は消費者にとっては商品やサービスの比較・検討する際の参考となり，また企業などの提供者にとっては消費者のニーズを把握する判断材料になりうる．このような意見情報の抽出には，評価を表す表現が重要な手がかりとなる．評価表現をそれぞれのドメインごとに収集するのは容易ではなく，すべて人手で収集するとなると非常に高コストである[1]．そのため本研究では商品やサービス，もしくはそれらのある側面に対する記述者の主観的な評価を表す評価表現を収集して辞書を構築すること目的とし，Web ページから半自動で評価表現を抽出する評価表現収集支援システムの構築を検討・評価する．

2. 評価表現

本研究では，評価表現を評価対象と評価値表現の組で抽出する．ここで，評価対象とは，広義で製品やサービス，狭義でその製品やサービスの側面である属性のことと定義する．評価値表現とは，評価対象の量や質に関しての値や記述者の主観的な意見とする．

評価対象は階層を取ると考えられる．例えば車のドメインでは「シートの背もたれ」や「シフトレバーの位置」といった表現が出現する．そういった評価対象となるものが評価対象を持つことが多くある．階層で考えることで，それぞれの階層の評価対象についての評価値を収集することができると考えられる．

本研究では評価表現の候補を人手で判定して，精度の高い辞書を構築することを目的とする．

3. 評価表現の抽出手法

本研究での処理の流れを図1に示す．

3.1 前処理

まず，Web から収集したページから本文を抜き出し HTML タグの除去をする．次に，ノイズとなるような記号や特殊文字を除去または置換する．システムへの入力は形態素解析および係り受け解析をしたものを利用した．具体的には GDA(Global Document Annotation)[2]のライブラリを利用した．これは内部的に CaboCha[3]の結果を利用しており，形態素には ChaSen[4]形態素解析結果のつく XML 文書を出力する．

3.2 既登録の表現の除去

辞書に既に登録されている表現は除去する．そうすることにより人手で判定する表現の数が減少する．辞書は評価表現用の辞書と，評価表現ではないと判定された表現の辞書に分かれる．また，初期の段階ではどちらの辞書も何も入っていない状態である．辞書は XML 形式で，1つの評価対象に対して多数の評価値表現をとる形式とした．また，それぞれの表現は形態素情報を保持していて，句はそれを構成する構造を保持している．これは，同音異義語などの区別や，辞書の修正時に役立つであろうとの考えからである．

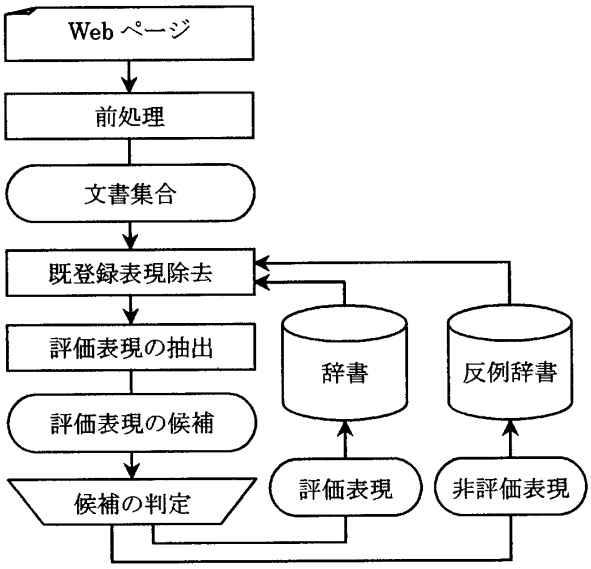

図1. 処理の流れ

3.3 評価表現の抽出

抽出には品詞と係り受け関係の情報を利用した．評価対象か評価値表現の候補になるかを文頭から形態素ごとに品詞により分類していく．どちらかになる語句が出現したら，その係り先を見る．評価対象と評価値表現の組になるもう一方の評価表現を取得し，その組を評価表現の候補として抽出する．句は GDA において主辞の品詞に基づくタグがつけられており，これを利用した．

3.3.1 候補対象外

ノイズを減らすために，非自立(自立語に従属する)タイプ，接頭タイプ，接尾タイプ，代名詞タイプ，副詞可能タイプの名詞，1文字以下のアルファベットは候補の対象外とした．タイプというのは ChaSen の品詞体系での分類である．以下で候補とする語句にはこれらの名詞を含めないとする．

3.3.2 評価対象の候補

次の品詞の語句を評価対象の候補とした．

- 名詞(形容動詞語幹タイプでない名詞，「最」「不」で始まらない名詞)
- 名詞句
- 未知語

†東京電機大学大学院工学研究科

3.3.3 評価値表現の候補

次の品詞の語句を評価値表現の候補とした．

- 形容詞(非自立タイプ以外)
- 形容詞句(用言+非自立形容詞)
- 動詞(自立タイプのみ)
- 動詞句
- 助動詞(「ある」のみ)
- 名詞(形容動詞語幹タイプの名詞，助動詞「だ」が続く名詞，体言止になっている文末の名詞，「最」「不」で始まる名詞)

動詞は自立以外のタイプのものは経験的にノイズとなることが多かったためである．助動詞「ある」は「高級感がある」のような「ある」ことが評価にかかわる語句を候補とするためである．助動詞「だ」が続く名詞および体言止になっている文末の名詞は，一般的な名詞のほかに「満足」などの評価値になる名詞も含まれるため，評価対象，評価値表現のどちらの候補にもするとした．また，接頭詞と分類されない「最高」や「不足」といった名詞を分類するため「最」「不」で始まる名詞は評価値表現の候補とした．

3.4 候補の判定と辞書への登録

評価対象と評価値表現の候補の組の判定は人手で行った．その判定結果により辞書に登録し，そうでないものは反例辞書に登録する．評価表現として採用するかは，人手で判断するため，その判定者の基準にゆだねられることになる．

4. 結果と考察

まず，品詞による分類の効果と，どの程度重複した表現が現れるかを見るために，辞書に既登録の表現を除去しない状態で評価表現の候補を抽出した．対象の Web ページは車のドメインの 100 件のレビュー記事のページとした．

表 1. 候補として抽出された「燃費」のとる評価値表現

評価値表現	主辞の品詞	件数
12Km/1ある	動詞+自立	1
いい	形容詞+自立	2
ネック	名詞+一般	1
のびる	動詞+自立	1
びっくりする	動詞+自立	1
ひどい	形容詞+自立	1
よい	形容詞+自立	3
よくする	動詞+自立	1
悪い	形容詞+自立	10
意外	名詞+形容動詞語幹	1
驚かす	動詞+自立	1
計る	動詞+自立	1
考える	動詞+自立	1
残念	名詞+形容動詞語幹	1
納車する	動詞+自立	1
非の打ち所ない	形容詞+自立	1
付きまとう	動詞+自立	1
有る	動詞+自立	1
良い	形容詞+自立	2
良くなる	動詞+自立	1

表 2. 評価値表現の抽出結果

評価対象	評価値表現
燃費	いい，ネック，のびる，びっくりする，ひどい，よい，悪い，残念，良い
内装	いい，シンプル，安っぽい，豪華，雑，満足

重複を含めて 4478 件の表現の候補の組が得られた．そのなかから評価対象「燃費」がとる評価値表現を表 1 に示す．具体的な数を伴うものや，「燃費」の評価値表現にならないものも含まれている．また，判断に迷うような表現がある．多数重複しているものもあり，このような語については辞書に既登録の表現の除去が有効であると考えられる．

表 2 に評価対象「燃費」と「内装」の取る評価値表現の人手で判断した抽出結果を示す．

5. おわりに

本論文では，Web ページから評価表現を抽出するシステムを実装した．しかし，本論文での実験ではサンプル数が少ないので，さらに実験を重ねる必要がある．今後以下の課題に取り組みたいと考えている．

- 評価対象，評価値表現の分類法を検討する．判定者に提示される表現の精度を上げるために，品詞による分類条件を見直す．
- 評価表現の候補をよりわかりやすいように提示する．例えば，文の形で提示し，評価表現をハイライトするというような方法である．文を提示することで意味を理解し，より正確な分類ができると考えられる．
- 自動化を検討する．今回は精度の高い辞書の構築が目的であるため人手での判定を前提としているが、人手での判定はコストが高いため，可能な範囲での自動化を検討する．
- 評価極性を扱う．収集した評価値表現に対してポジティブ，ネガティブといった評価極性をつける．
- 別のドメインでの有効性を検証する．評価値表現を評価対象(属性)ごとに抽出することで，別のドメインでその評価対象が出現した場合に，その評価対象がとる評価値表現を利用できると考えられる．意味的に近いドメイン(例えば，自動車とバイク) では数多くの表現が共通していて，利用可能だと考えられる．

6. 参考文献

[1] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一: 意見抽出のための評価表現の収集, 自然言語処理,Vol. 12, No.2, pp.203–222, (2005).

[2] 産業技術総合研究所 大域文書修飾 Global Document Annotation (GDA) http://i-content.org/gda/

[3] 奈良先端科学技術大学院大学 松本研究室：日本語係り受け解析器, CaboCha, http://chasen.org/~taku/software/cabocha/

[4] 奈良先端科学技術大学院大学 松本研究室：形態素解析システム, ChaSen, http://chasen.naist.jp/hiki/ChaSen/