

CGM からの評価表現辞書の自動作成に関する基礎研究

中村健二† 田中成典‡ 高橋亨輔† 寺口敏生† 田中裕一† 瓜谷芳樹‡

関西大学大学院総合情報学研究科† 関西大学総合情報学部‡

1. はじめに

近年、消費者が形成するメディアである CGM (Consumer Generated Media) が認知され、コンテンツ量が増加している。その中でも、商品に関する評判情報は、商品購入の意思決定支援に利用される。しかし、膨大な CGM コンテンツから、目的の商品に関する評判情報を取得するのは困難[1]である。そのため、CGM コンテンツから商品ごとの評判情報を自動で抽出する技術が求められている。商品の評判情報を自動で抽出するための基礎研究として、商品の評価している文章に現れる特徴的な語である評価語を収集する研究[2][3]や、評価語と共に商品が持つ属性を収集する研究[4]がある。また、収集した評価語を用いて文章を解析する研究[5]も盛んに行われている。評価語は、評価対象によって適用できるものが異なる。しかし、評価語と属性を収集する研究では、評価対象ごとに人手で精査した CGM コンテンツを入力して評価語を収集しているため、幅広い評価対象に対応した評価語の作成や利用は困難である。そこで、本研究では、CGM コンテンツから動的に評価対象を抽出し、評価対象が持つ属性の類似性に注目して評価対象ごとの属性と評価語からなる評価表現を自動収集することによって、汎用性の高い評価表現辞書の作成を目指す。

2. システムの概要

本研究では、CGM コンテンツを利用した評価表現辞書の作成を目的とする。本システムは、図 1 に示すように評判情報が記載された CGM コンテンツを入力とし、評価表現辞書を出力とする。本システムは、1) 評価表現取得機能、

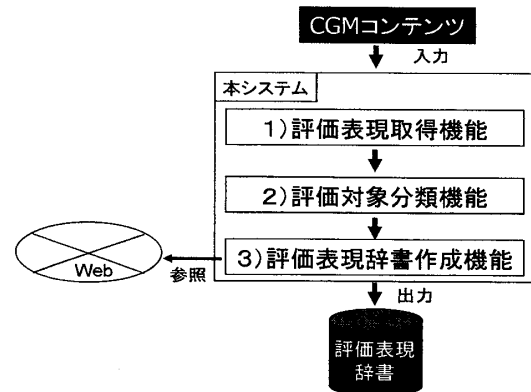


図1 システムの概要

2) 評価対象分類機能、3) 評価表現辞書作成機能により構成される。

2.1 評価表現取得機能

本機能では、CGM コンテンツ中の文から、「評価対象 {の} 属性 {は, が, も, に, を} 評価語」の形で書かれた評価表現候補を取得する。そして、品詞を考慮することで評価表現候補から相応しくない要素を取り除き、評価表現として取得する。

2.2 評価対象分類機能

本機能では、評価対象を評価対象が持つ属性の類似度を用いて分類し、同じ評価表現が適用できるグループを作成する。評価対象が持つ属性の類似度には、同じ属性が含まれる割合を表す被覆率を用いる。また、被覆率を用いて分類することから、多くの評価対象が共通して持つ属性は、分類の精度を下げる原因になると考えられるため、分類には用いない。

2.3 評価表現辞書作成機能

本機能では、まず、評価対象分類機能で作成したグループ内で、評価対象と属性が取り得る全ての組み合わせを新しい評価表現候補として取得する。次に、評価表現候補が評価表現として利用可能であるかについて、Web から取得した文章中における評価対象と属性の共起度合いによって判定し、有意なものを評価表現として取得する。最後に、本機能と評価表現取得機能で取得した評価表現を評価表現辞書に登録する。

Research for Collecting Evaluative Expressions from CGM
 † Kenji Nakamura, Kyouzuke Takahashi, Toshio Teraguchi,
 Yuichi Tanaka
 Graduate School of Informatics, Kansai University, 2-1-1
 Ryouzenji-cho, Takatsuki-shi, Osaka 569-1095, Japan
 ‡ Shigenori Tanaka, Yoshiki Uritani
 Faculty of Informatics, Kansai University, 2-1-1 Ryouzenji-
 cho, Takatsuki-shi, Osaka 569-1095, Japan

3. システムの実証実験と考察

本システムの実行結果を図2に示す。本提案手法の有用性を実証するため、評価表現辞書を作成し、辞書の登録件数および登録内容に関する実験を行った。

3.1 実証実験

本実験では、価格.comのノートパソコンカテゴリにおいて、口コミ件数が多い上位60商品の口コミ情報109,438件を用いて評価表現辞書を作成した。ここで、評価対象分類機能で用いる被覆率の閾値は、実験を複数回試行した結果、経験的に最も分類精度が高くなった4割とした。また、分類に用いる属性は、多くの評価対象が共通して持つ属性を省くため、同じ評価対象に対して2回以上出現しており、かつ、その属性を持つ評価対象が50個未満のものとした。

3.2 結果と考察

実証実験では、15,975件の評価対象に関する評価表現が取得できた。また、取得できた属性は17,608種類で、評価語は378種類だった。このことにより、入力するCGMコンテンツを人手で精査することなく幅広い評価対象についての評価表現が取得できたことを確認した。また、評価表現を取得した評価対象の中で、ノートパソコンに関する主要な評価対象についての結果を表1に示す。表1より、主要な評価対象において、属性と評価語が増加したことが確認できる。また、その他の評価対象についても、評価対象分類機能で分類の対象となった評価対象に関しては、属性と評価語が増加していた。このことから、評価対象が持つ属性の類似性に注目して評価表現を収集する提案手法が自動で幅広い評価対象に対応した評価表現辞書の作成に有効であると考えられる。しかし、キーボードについては、新たに価格、ドライバや感度といった有用な属性が取得できた一方で、録音のような不適切な属性も取得した。これは、キーボードがサウンドという評価対象と同じグループに分類されたために候補として作成され、Web検索時に同音異義語である楽器としてのキーボードとの区別ができなかったために取得されたと考えられる。

4. おわりに

本研究では、CGMコンテンツを利用した汎用性の高い評価表現辞書の作成手法を提案した。そして、システムの実証実験の結果からその有用性を証明した。しかし、取得した評価対象には不適切な語が含まれていた。この問題に対しては、作成した評価表現辞書をWebで公開し、

品名	属性	評価	評価の品詞	文
CPU	DTP			よ色。など。で。
CPU	スワップ			申しCPUのスペック。
CPU	筐体			この時はCPUの。
CPU	クロック			実運部分CPU。
CPU	使用環境			あとはCPUの使。
CPU	製品名			サボートに動かす。
CPU	Pentium4			CPUの起動にPe...
CPU	性能			ほかにもCPUの性。
CPU	必要			CPUの交換が必。
CPU	発熱			CPUの発熱で熱。
CPU	使用率	高い	形容詞	カスカマネーゲー...
CPU	遅い			遅いCPUの遅い。
CPU	故障			原因はCPUの故。
CPU	SON			同じCPUのSON。
CPU	遅い			遅いCPUの遅。
CPU	交換			マザーボードCP...
CPU	性能			CPUの性能やH...
CPU	速度	速い	形容詞	低層設計を施し...
CPU	交換			CPUの交換です。
CPU	クロックアップ			CPUのクロックア...

図2 本システムの実行結果

表1 主要な評価対象についての結果

	属性		評価語	
	A	B	A	B
キーボード	175	272	18	29
スピーカ	3	25	0	3
マウス	67	109	8	16
メモリー	74	148	6	14
バッテリー	118	203	17	29
HDD	187	249	21	28
バックライト	23	1	46	6

A: 評価表現取得機能の出力

B: 評価表現辞書作成機能の出力

人手で評価対象を精査する仕組みを作成することで対応する予定である。また、今後は、同音異義語を区別できない問題を解決するために手法の改良を行う予定である。

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, 人工知能学会誌, 人工知能学会, Vol.19, No.3, pp.317-323, 2004.5.
- [2] 那須川哲哉, 金山博: 文脈一貫性を利用した極性付評価表現の語彙獲得, 自然言語処理研究会研究報告, 情報処理学会, Vol.2004, No.73, pp.109-116, 2004.7.
- [3] Turney, P. D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, pp.417-427, 2002.12.
- [4] Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K. and Fukushima, T.: Collecting Evaluative Expressions for Opinion Extraction, Proc. of the 1st International Joint Conference on Natural Language Processing, pp.584-589, 2004.3.
- [5] 飯田龍, 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一: 意見抽出を目的とした機械学習による属性-評価値対同定, 自然言語処理研究会研究報告, 情報処理学会, Vol.2005, No.1, pp.21-28, 2005.1.