

## データマイニングを利用した口コミの極性分析

坂村華怜<sup>†</sup> 白井 匡人<sup>††</sup> 三浦 孝夫<sup>†</sup>

<sup>†</sup> 法政大学理工学部創生科学科 <sup>††</sup> 法政大学マイクロ・ナノテクノロジー研究センター

### 1. 提案手法

本研究では、同時関係抽出(アソシエーション分析)を用いて極性辞書を拡張させ、特徴語を生成すると同時に、極性を判定する手法を提案する。

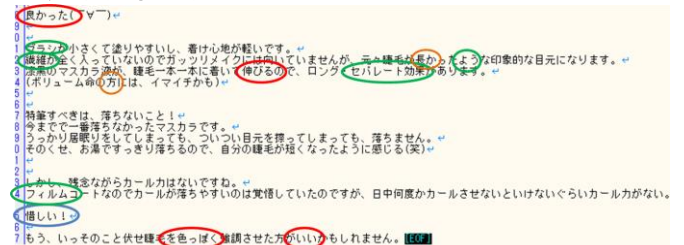
### 2. 実験手順

本研究では化粧品の“口コミ”サイトである、“アットコスメ”を利用する。カテゴリは“マスカラ”、“最新口コミランキング”の上位5位を対象とする。“口コミ”と“レビュー”情報であるおすすめ度を、多く人が参考になったと判断した場合にポイントが増される“Like”という値の高い順に各商品100件収集し、形態素解析ソフト、“Mecab”を利用し、文章の形態素解析を行う。名詞、動詞、形容詞を抽出。この時、極性が反転することを考慮し、付属語“ない”まで含めて一語として取得。解析ツール“R”を用いて、口コミ100件の単語集合を元に、商品ごとに同時関係抽出を行う。この際、設定する閾値は(支持度:0.05、確信度0.6、アイテム数2)とする。相関ルールをエントロピーの値であるリフト値の高い順、1000件を抽出する。同時関係抽出を行い、抽出されたルールを元に辞書更新を行う。極性辞書に登録されている単語と同時関係抽出によって共起された語を新たに辞書に登録し、辞書の拡大を図る。極性語ごとに以上の操作を実施し、辞書の更新は完了する。

評価方法について記述する。本研究では再現率は実際に正である中で正と予測した数、適合率は正と予測したなかで、実際に正である数と定義する。商品ごとに更新された辞書と、各“口コミ”内の極性語の数を算出。“ポジティブ”な単語数と“ネガティブ”な単語数を引き算し総合値とする。総合値が0より大きいものを、商品に対して“良い”、と評価している“口コミ”、総合値が0より小さいものを、商品に対し“悪い”、と評価している“口コミ”とし、極性分析を行う。更新前の辞書でも同様に“口コミ”の極性分析を行う。

各商品には“レビュー”情報の平均である、“おすすめ度平均”が存在し、口コミ100件の各“おすすめ度”と“おすすめ度平均”を比較する。“おすすめ度平均”の小数点以下は無視し、その値より大きい“おすすめ度”の口コミは、商品に対して良い評価を、その値より小さい“おすすめ度”の口コミは、商品に対して悪い評価をしていると定義する。極性分析の結果と“おすすめ度平均”をもとに行った口コミの極性をもとに、適合率と再現率を算出し、更新前の辞書と更新後の辞書で比較する。

### 3. 実験結果と考察



上図は、赤、青(ポジティブ、ネガティブ)が更新前の極性語で緑、オレンジが更新後の極性語である。本手法で辞書の語数が向上する。どちらにも更新された単語は総合値に影響しないので、除外する。再現率は11%向上し、適合率は2.5%下がった。更新された辞書の単語に、“セパレート”や、“フィルム”というカテゴリに依存する単語が確認できる。一方で、ネガティブな極性辞書にはカテゴリに依存する語は見受けられない。おすすめ度平均を基準と見たときに全体で好評価のデータの方が3.2倍もデータ量が多く、低評価のデータでは、同時関係抽出の時点で十分なデータマイニングが行えず、余分な語が更新されたことが原因と考える。

辞書更新前に極性が判断できない“口コミ”に関して更新前の辞書では極性が抽出できないものが42件存在する。更新後の辞書で極性分析可能になったものは29件で、今回の手法を利用することで、45%は極性に分類でき、正しい極性分析が可能となる。

### 4. 結論

従来の極性辞書では判断できなかったものが、同時関係抽出を利用し、カテゴリに依存する特徴語を抽出して辞書に登録、拡張することによって45%向上し、評価可能になった。再現率も11%向上した。適合率に関しては従来の辞書より2.5%と若干下がってしまった。これは取得データ量に依存するデータマイニングの精度に影響されて起きたと考えられる。今後は、データ収集の際に評価の違うデータ数をそろえ、より多くのデータを収集することが課題である。

### 参考文献

- [1] 日本語評価極性辞書 - 乾・岡崎研究室
- [2] Han, J. et al. : Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufman, 2011
- [3] 吉田、北山：商品レビューの極性分析に基づく特徴語抽出手法の評価，電子情報通信学会信学技法 DE2016-5, 2016