ISS-SP-206

評点情報と局所情報の組み合わせによる評価表現辞書の精度向上に関する研究

加藤　さやか†　　吉川　大弘†　　古橋　武†
†　名古屋大学大学院　工学研究科

1. はじめに

近年 web の発展に伴って，膨大なテキストデータが蓄積されるようになった．その例の一つが，商品の評価情報を表すレビューデータである．これを自動で解析する技術は，企業と消費者の双方にとって有用であり，関心が高まっている[3]．レビューを解析する上では，各文について商品への肯定および否定を判別する必要がある．一般的には，単語に対する評価表現辞書が用いられることが多く，またこれを自動で構築する研究も盛んに行われている[1][2]．従来の評価表現辞書の構築には，レビューの持つ評点情報[1]，または文書中に出現する極性が既知の評価表現（局所情報）[2]が用いられている．本研究では，これら両方の情報を適切に用いることで，評価表現辞書の精度を向上させることを目指す．本稿では，評点情報と局所情報をそれぞれ使った場合の精度の比較を行うとともに，これらを組み合わせて評価表現辞書を構築する方法について検討する．

2. 評価表現辞書の作成

例えば「高い」という形容詞に着目すると，「性能が高い」は肯定だが，「値段が高い」では否定表現となる．本稿では，このように，組み合わされる名詞により極性が変化する形容詞（極性不定形容詞）に注目し，評価表現辞書を作成する．また，辞書に登録する評価表現は，これら極性不定形容詞と名詞との組み合わせに対し，極性を登録する．

文書中に評価表現が存在すると，その周囲に評価表現が現れ，また極性が一致する傾向がある[2]ことを，局所情報として利用する．本手法では，極性が既知の単語についての評価表現辞書を用意し，文書中のある名詞と形容詞のペアがその評価表現と共起した場合に，ペアの極性を評価表現の極性と一致させて抽出する．また，評点情報を用いる場合は，名詞と形容詞のペアの極性を，それが出現したレビューの評点に一致させて抽出する．

以上の手順で得られた，各ペアの各極性での抽出回数を用いて，(1)式により評点スコア $Score_G$ を，(2)式により局所スコア $Score_L$ をそれぞれ算出し，各ペアの極性を決める．ここで，(1)式において，$n_{P_G,i}, n_{N_G,i}$:ペア i が評点 5，評点 1 のレビューで出現した数，$P_G=\sum_{i=1}^{n} n_{P_G,i}, N_G=\sum_{i=1}^{n} n_{N_G,i}$:評点 5，評点 1 のペアの数，$k$:スムージング項である．(2)式において，$n_{P_L,i}, n_{N_L,i}$:文書中でペア i が肯定，否定の評価表現と共起した数，$P_L=\sum_{i=1}^{n} n_{P_L,i}, N_L=\sum_{i=1}^{n} n_{N_L,i}$:肯定，否定のペアの数，である．また，次節において，局所スコア 1 は(2)式の $Score_L$ であり，局所スコア 2 は，(2)式における正規化(P_L, N_L で割る)を行わないものである．

$$Score_G=\frac{\frac{n_{P_G,i}}{P_G}-\frac{n_{N_G,i}}{N_G}}{\frac{n_{P_G,i}}{P_G}+\frac{n_{N_G,i}}{N_G}+k} \quad (1)$$

$$Score_L=\frac{\frac{n_{P_L,i}}{P_L}-\frac{n_{N_L,i}}{N_L}}{\frac{n_{P_L,i}}{P_L}+\frac{n_{N_L,i}}{N_L}+k} \quad (2)$$

3. 実験

2 節で示したスコアを，楽天レビュー11 万件と価格.comレビュー1 万 5 千件により算出し，極性の妥当性について評価を行った．ただしここでは，評点情報と局所情報を独立に評価した．極性不定形容詞は，[3]で定義されている 17 語とした．スコアの閾値は，-1≦否定≦-0.5，-0.5＜中性＜0.5，0.5≦肯定≦1 とした．評点情報と局所情報を合わせて 10 回以上抽出されたペアについて，人手により極性を決定し，上述のスコアによる極性との一致度合いに基づいて評価を行った．表 1 から，評点スコアよりも局所スコアの適合率や再現率が高いことがわかる．

表 1　評点/局所情報をそれぞれ使用した結果

		楽天	価格
適合率	評点スコア	0.694	0.8
	局所スコア 1	0.938	0.941
	局所スコア 2	0.926	1
再現率	評点スコア	0.397	0.211
	局所スコア 1	0.476	0.281
	局所スコア 2	0.397	0.263

4. まとめ

本稿では，評点情報と局所情報を用いた評価表現辞書の作成法について述べた．また，レビューデータに対して適用し，局所情報を用いた場合の方が，評点情報を用いた場合よりも適合率や再現率が高いことを確認した．今後の課題としては，評点情報と局所情報の組み合わせ方に対する検討が挙げられる．

参考文献

[1]藤村滋,豊田正史,喜連川優，“文の構造を考慮した評判抽出手法”,電子情報通信学会第 16 回データ工学ワークショップ,6C-i8,2005
[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”,情報処理学会,NL-162,pp.109-116，2004
[3]高村大也,乾孝司,奥村学，“極性反転に対応した評価表現モデル”,情報処理学会,NL-168,pp141-148,2005