1B2-5

評点情報と局所情報を用いた評価表現辞書の構築に関する基礎検討

A Basic Study on Generating Lexicon of Sentiment Expressions Using Rating Information and Local Information

加藤さやか *1 Sayaka Kato
吉川大弘 *1 Tomohiro Yoshikawa
古橋武 *1 Takeshi Furuhashi
奥山賢治 *2 Kenji Okuyama

名古屋大学大学院工学研究科 *1
Graduate School of Engineering Nagoya University
東邦ガス株式会社 *2
TOHO GAS Co., Ltd.

Recently, reviews which have evaluation information of products have been increasing because of diversity of internet. Automatical extraction method of beneficial information is useful for both companies and consumers. When we analyze reviews, we often use lexicon of sentiment expressions to determine positive or negative sentiment which some words or sentences have. There are many researches on generating method of lexicon of sentiment expressions. In this paper, we combine rating information and local sentiment expressions whose polarities are known, and study the generation method of lexicon of sentiment expressions.

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．これら全てに目を通すことは，時間や労力の面で大変困難である．そのため，これら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 [乾 2006]．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 [藤村 2005][那須川 2004]．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 [藤村 2005]，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）[那須川 2004] 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

本節では，評価表現辞書の作成法について説明する．例えば，「美しい」や「酷い」のような評価表現は，どのような名詞と組み合わされても極性は変化しない．このような極性が既知の評価表現は，人手で極性を付け，評価表現辞書を作成すればよいと考えられる．しかし，組み合わされる名詞によって極性が変化するものもある．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．このような形容詞は極性不定形容詞と呼ばれる [高村 2005]．極性不定形容詞と名詞の組み合わせは膨大となるため，それらすべて人手で極性を付けるのは困難である．そこで本稿では，これら極性不定形容詞の極性を，自動で判別し，評価表現辞書を作成することを目的とする．辞書に登録する評価表現は，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．

2.1 評点情報の利用

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．例えば，「性能が高い」という文が評点 5 のレビューから出現した場合，「性能+高い」ペアの極性を肯定とする．抽出する名詞と形容詞のペアは，一文中において，格助詞「が」の前後の名詞と形容詞，または係助詞「は」の前後の名詞と形容詞とする．レビューの評点について，評点 4,5 を肯定とし，評点 1,2 を否定とする．評点 3 については用いない．

2.2 局所情報の利用

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 [那須川 2004] 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．例えば，「値段が高くて残念です」という文について，「残念」の極性は否定であることが既知であり，これにより，「値段+高い」ペアの極性を否定とする．

2.3 極性が既知の単語に対する評価表現辞書の作成

極性が既知の単語についての評価表現辞書は，高村らが作成した既存の評価極性辞書 *1[高村 2006] から，極性不定形容詞を除いた標準形容詞，及びレビューで頻出する極性が既知の評価表現になり得る名詞を登録した．ただし，一部の単語について，極性を人手で修正した．また，登録の際は，「臭い」など，名詞と形容詞を混同してしまうものは削除した．登録件数は，681 件である．さらに，「美しくない」や，「手間がない」など，形容詞や名詞に助動詞の「ない」が付く場合には，極性を反転させて判別する．また，文中に逆接を含むときは，その前後で極性を反転させる．

2.4 スコア算出式

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，[藤村 2005] を基にして，(1) 式により評点スコア $Score_G$

連絡先: 加藤さやか，名古屋大学大学院工学研究科，名古屋市千種区不老町，052-789-2793，052-789-3166，sayaka@cmplx.cse.nagoya-u.ac.jp

*1 http://www.lr.pi.titech.ac.jp/~takamura/pubs/pn_ja.dic

を，(2) 式により局所スコア 1 $Score_{L1}$ を，(3) 式により局所スコア 2 $Score_{L2}$ をそれぞれ算出し，各ペアの極性を求める．スコアは $-1 < Score < 1$ の値をとり，1 に近いほど肯定的，-1 に近いほど否定的である．

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

$$Score_{L1} = \frac{\frac{n_{P_L,i}}{P_L} - \frac{n_{N_L,i}}{N_L}}{\frac{n_{P_L,i}}{P_L} + \frac{n_{N_L,i}}{N_L} + k} \quad (2)$$

$$Score_{L2} = \frac{n_{P_L,i} - n_{N_L,i}}{n_{P_L,i} + n_{N_L,i} + k'} \quad (3)$$

評点スコアと局所スコア 1 は正規化を行っており，これによってレビュー数や出現頻度の偏りを考慮している．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．(1) 式において，$n_{P_G,i}$: ペア i が評点 5 のレビューで出現した数，$n_{N_G,i}$: ペア i が評点 1 のレビューで出現した数，$P_G = \sum_{i=1}^{n} n_{P_G,i}$: 評点 5 のペアの数，$N_G = \sum_{i=1}^{n} n_{N_G,i}$: 評点 1 のペアの数，$k$: スムージング項である．また (2) 式において，$n_{P_L,i}$: 文書中でペア i が肯定の評価表現と共起した数，$n_{N_L,i}$: 文書中でペア i が否定の評価表現と共起した数，$P_L = \sum_{i=1}^{n} n_{P_L,i}$: 肯定のペアの数，$N_L = \sum_{i=1}^{n} n_{N_L,i}$: 否定のペアの数である．なお (3) 式における $k' = k(n_{P_L,i} + n_{N_L,i})$ である．

3. 提案手法

本節では，局所情報を利用する際の工夫点及び評点情報と局所情報の組み合わせ方について述べる．

3.1 局所情報を利用する際の工夫点

接続助詞の「ので」は，因果関係を表す．そのため，以下の二つのような使われ方が考えられる．一つは，“雑穀の種類が多いので，体には良いと思います”である．この例の場合,「良い」は肯定表現であり,「種類が多い」も肯定表現となる．この場合,「ので」の前後で極性が一致している．もう一つは，“純正は高いので，安く買えてよかったです”である．この例の場合,「安い」,「良い」は肯定表現であるが,「純正は高い」は否定表現となる．この場合,「ので」の前後で極性が反転している．このように,「ので」の使われ方は文脈によるため,「ので」を含む文に対しては，局所情報を用いないこととする．

3.2 評点情報と局所情報の組み合わせ方

評点情報と局所情報の組み合わせ方について，二つの方法を検討する．一つは，局所情報で取れるペアに評点を用いる方法である．ここでは，局所情報での抽出回数が 1 回以上のものを対象とする．一文中でペアが評価表現と共起するときは局所情報を用い，共起しないときは評点を用いてペアを抽出する．局所情報での抽出回数，評点での抽出回数をそれぞれ求め，局所スコアの式に評点情報を組み込む．局所情報だけでなく，評点も合わせて用いることで，スコアの精度が高まることが期待される．もう一つは，局所情報が取れないペアに対してのみ評点情報を用いる方法である．局所情報での抽出回数が 0 回だったものに対しては評点を用い，抽出回数が 1 回以上だったものに対しては局所情報を用いる．また，二つの方法で評点を用いるにあたり，各評点のうち少ない方のレビュー数に対して重みをかけた．これは，一般的にレビューは評点が高いものが多いため，そのまま評点を用いると肯定に偏ってしまうためである．実際に，実験に使用するデータセットである米雑穀，プリンタ，季節家電の楽天レビュー 115,649 件の評点は，全体の 97 %が肯定 (評点 5 及び 4) となっている．スコア算出式は，(4)，(5) 式である．正規化した (4) 式を評点局所スコア 1$Score_{GL1}$ と呼び，正規化していない (5) 式は評点局所スコア 2$Score_{GL2}$ と呼ぶ．

$$Score_{GL1} = \frac{(\frac{n_{P_L,i}}{P_L} + \frac{n_{P_C,i}}{P_C}) - (\frac{n_{N_L,i}}{N_L} + \frac{n_{N_C,i}}{N_C})}{(\frac{n_{P_L,i}}{P_L} + \frac{n_{P_C,i}}{P_C}) + (\frac{n_{N_L,i}}{N_L} + \frac{n_{N_C,i}}{N_C}) + k} \quad (4)$$

$$Score_{GL2} = \frac{(n_{P_L,i} + n_{P_C,i}) - (n_{N_L,i} + n_{N_C,i})}{(n_{P_L,i} + n_{P_C,i}) + (n_{N_L,i} + n_{N_C,i}) + k'} \quad (5)$$

ここで，(4) 式において，$n_{P_C,i}$: 文書中で評価表現と共起しないペア i が評点 5 のレビューから出現した数，$n_{N_C,i}$: 文書中で評価表現と共起しないペア i が評点 1 のレビューから出現した数，$P_C = \sum_{i=1}^{n} n_{P_C,i}$: 評価表現と共起しない評点 5 のペアの数，$N_C = \sum_{i=1}^{n} n_{N_C,i}$: 評価表現と共起しない評点 1 のペアの数である．また (5) 式における $k' = k(n_{P_L,i} + n_{P_C,i} + n_{N_L,i} + n_{N_C,i})$ である．

4. 実験

2 章で述べた，評点，局所情報をそれぞれ用いた場合と，3 章で述べた，評点と局所情報を組み合わせた手法を，それぞれレビューデータに適用した．実験には，米雑穀，プリンタ，季節家電に関する楽天レビュー [*2]115,649 件と，家電に関する価格.com レビュー [*3] 15,327 件を用いた．

4.1 実験手順

名詞と形容詞のペアは，格助詞「が」を挟む名詞形容詞，また係助詞「は」を挟む名詞形容詞とする．以下に実験手順を示す．

1. データから名詞と形容詞（極性不定形容詞）のペアを抽出する．
2. 抽出回数より，評点スコア，局所スコア 1 (正規化あり)，局所スコア 2(正規化なし)，評点局所スコア 1(正規化あり)，評点局所スコア 2(正規化なし) を算出する．
3. スコアの値によって極性を判定する．
4. 正解データを用いて，適合率・再現率で評価する．

4.1.1 極性不定形容詞

極性不定形容詞は，[高村 2005] で定義されている 17 語（高い，低い，大きい，小さい，重い，軽い，強い，弱い，多い，少ない，ない，すごい，激しい，深い，浅い，長い，短い）とした．

4.1.2 正解データ

正解データの作成方法について述べる．データから，名詞と極性不定形容詞の全てのペアを抽出した．抽出回数 10 回以上でのものに絞り，3 名で肯定/否定/ニュートラルの極性を付けた．そして，極性が 3 名で一致したもの，3 人中 2 人が肯定/否定で一致し，もう 1 名がニュートラルであったものを正解の極性とし，正解データに登録した．

*2 http://review.rakuten.co.jp/
*3 http://kakaku.com/

4.1.3 適合率

適合率の式を (6) 式に示す．肯定または否定と極性判定された結果が，正解データで肯定または否定と一致しているときのみ正解とした．ニュートラルはカウントしない．また，抽出回数の閾値により，極性が未判定のものはカウントしない．

$$\text{適合率} = \frac{\text{抽出した名詞 + 形容詞のペアの中で極性が正解していた数}}{\text{抽出した名詞 + 形容詞のペアのうち正解データに含まれている数}} \quad (6)$$

4.1.4 再現率

再現率の式を (7) 式に示す．肯定または否定と極性判定された結果が，正解データで肯定または否定と一致しているときのみ正解とした．正解データでニュートラルのものは除いてあるが，正解データが肯定または否定で，判定がニュートラルのものは，不正解となる．また，抽出回数の閾値により，極性が未判定のものも不正解となる．

$$\text{再現率} = \frac{\text{抽出した名詞 + 形容詞のペアの中で極性が正解していた数}}{\text{正解データの数 (ニュートラルを除く)}} \quad (7)$$

4.1.5 F 値

適合率と再現率はトレードオフの関係であるため，適合率と再現率の調和平均である F 値も算出する．F 値の算出式は (8) 式の通りである．

$$F\text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \quad (8)$$

4.2 実験条件

局所情報を用いる際の有効範囲は，一文単位（文末まで）とし，スコア算出式におけるスムージング項のパラメータ k は，0.00001 とした．算出したスコアより極性を決める閾値は，0.1 とした．すなわち，$-1 \leq Score \leq -0.1$: 否定，$-0.1 < Score < 0.1$: ニュートラル，$0.1 \leq Score \leq 1$: 肯定とした．抽出回数は，楽天レビューは評点と局所情報を合わせて 3 回以上，価格.com は 5 回以上とした．これらは事前の予備実験で最も F 値が高かったときの条件である．

4.3 結果と考察

楽天レビューの (a) 適合率，(b) 再現率，(c)F 値を図 1 に示す．また，価格.com レビューの (a) 適合率，(b) 再現率，(c)F 値を図 2 に示す． 楽天レビューの結果より，局所スコア 2（正規化なし）が最も高く，次いで評点局所スコア 2（正規化なし）が高くなっていることがわかる．一方再現率は，評点局所スコア 1（正規化あり）が最も高い．局所情報が取れないものに対してスコアが付くため，再現率が高くなったと考えられる．F 値についても，評点局所スコア 1 が最も高い値となっている．また，適合率，再現率，F 値すべてにおいて，評点スコアが最も低いことがわかる．また，価格.com レビューについても，ほぼ同様の結果となった．これらの結果から，評点情報を用いるよりも，局所情報を用いることで，判別性能が向上することが確認できた．すなわち，局所情報が取れるものは局所情報を用い，取れないものについてのみ評点を用いることで，最も判別性能が高くなることが確認できた．

5. おわりに

本稿では，評価表現辞書の精度の向上を目的として，評点情報と局所情報を組み合わせて辞書を構築する手法について検討した．適合率，再現率，F 値の比較により，局所情報が取れるものは局所情報のみを用い，局所情報で取れないものは評点を用いる方法が，最も精度が高くなることを確認した．今後は，今回考慮していないニュートラルの扱い方について検討し，さらに評価表現辞書の精度を向上させていく予定である．

参考文献

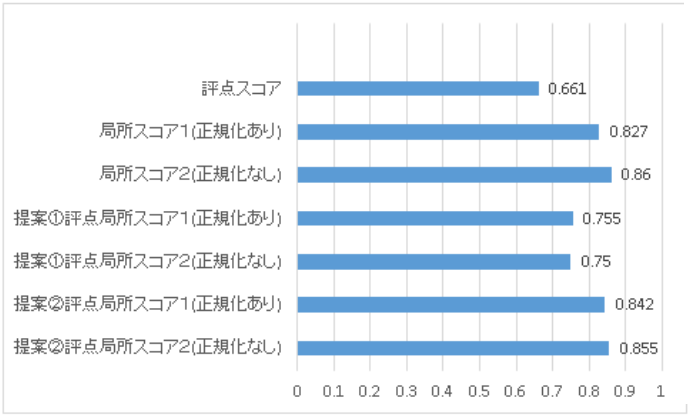
[乾 2006] 乾健太郎, 奥村学: テキストを対象とした評価情報の分析に関する研究動向, 自然言語処理, Vol.13, No.3 (2006)

[藤村 2005] 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)

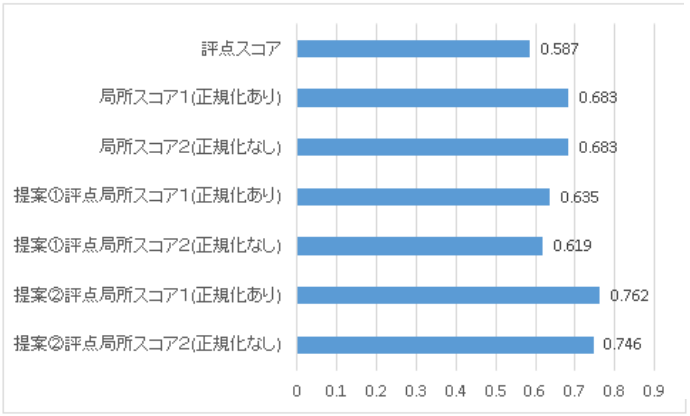
[那須川 2004] 那須川哲哉, 金山博: 文脈一貫性を利用した極性付評価表現の語彙獲得, 情報処理学会, NL-162, pp.109-116 (2004)

[高村 2005] 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, pp.141-148 (2005)

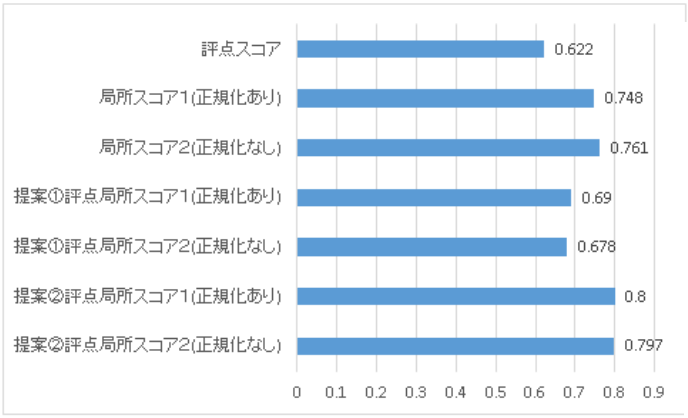
[高村 2006] 高村大也, 乾孝司, 奥村学: スピンモデルによる単語の感情極性抽出, 情報処理学会, Vol.47, No.2, pp.627-637 (2006)



(a) 適合率

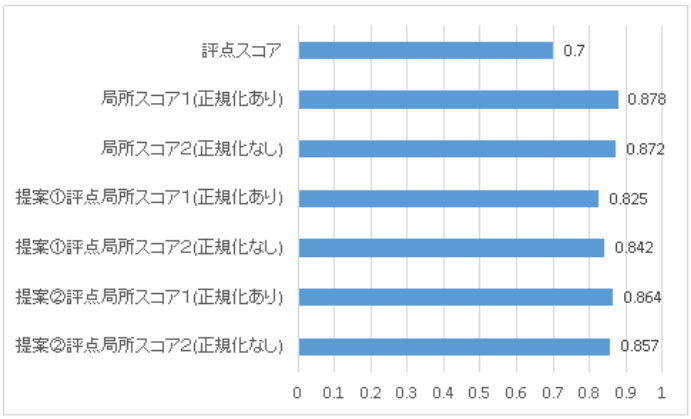


(b) 再現率

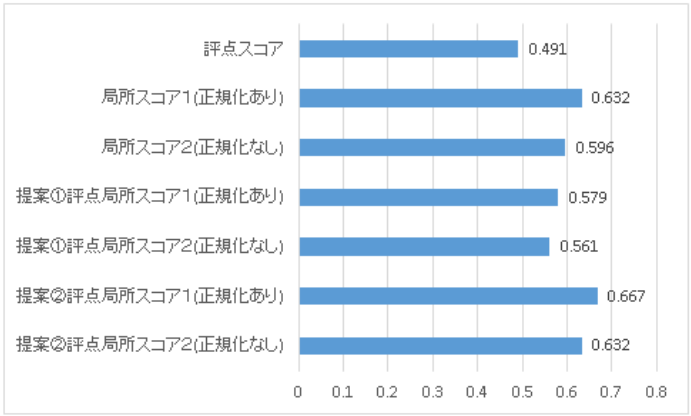


(c)F 値

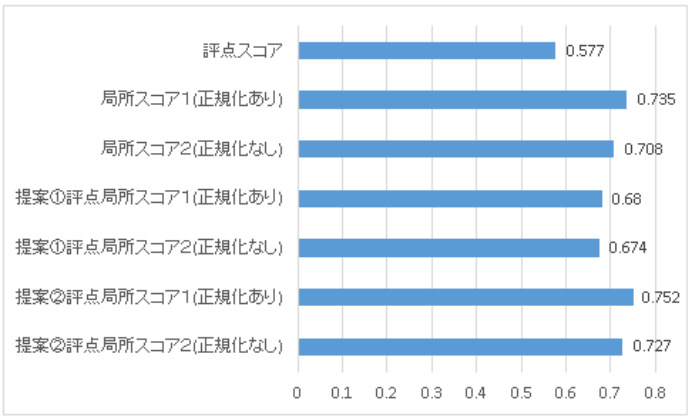
図 1: 楽天レビュー



(a) 適合率



(b) 再現率



(c)F 値

図 2: 価格.com レビュー