自動構築した評価文コーパスからの評価表現辞書の構築

Building Lexicon for Sentiment Analysis from Automatically Constructed Polar Sentence Corpus

鍛治 伸裕♦ 喜連川 優♦
Nobuhiro KAJI Masaru KITSUREGAWA

近年，評価や感情が記述されたテキストを解析する処理が注目を集めている．こうした処理を実現するためには，評価表現とその極性(好評/不評)の組を登録した辞書(評価表現辞書)が必要不可欠である．そのため，大規模な評価表現辞書の構築が重要な研究課題となっている．本論文では，大規模な HTML 文書集合から自動構築した評価文コーパスを用いて，評価表現辞書を自動構築する方法を提案する．実験の結果，提案手法は既存手法と比較して良好な精度を示すことが確認できた．

In recent years, a considerable number of studies have been made on sentiment analysis. Sentiment analysis requires lexicon in which word/phrase and its polarity pairs are listed. Building such lexicon is one of the fundamental issues in sentiment analysis. This paper proposes a method of building such lexicon from polar sentence corpus that is automatically constructed from large collection of HTML documents. Our experimental result showed significant improvement over the previous ones.

1. はじめに

近年，評価や感情が記述されたテキストを解析する処理が注目を集めている．こうした処理を実現するためには，評価表現とその極性(好評/不評)の組を登録した辞書(評価表現辞書)が必要不可欠である．そのため，大規模な評価表現辞書の構築が重要な研究課題となっている．

これまでに，シソーラスや国語辞典などの語彙資源を利用して評価表現辞書を自動構築する手法が提案されている[4] [5] [6] [7]．しかし，そうした語彙資源を利用した場合，そのエントリに登録されている単語しか扱うことができない．そのため，既存の語彙資源に登録されていない新語や口語(「しょぼい」など)に対応できないなど，網羅性に欠けるという問題がある．さらに，句を扱えないということも問題である．例えば「質が高い」は好評極性を持つが「コストが高い」は不評極性を持つというような事例を正しく把握するためには，句とその評価極性が登録された評価表現辞書が必要となる．

一方，少数の種単語との共起度にもとづいて評価表現を獲得しようとする試みもある[1][8]．例えばTurneyは検索エンジンを用いて種単語(「excellent」「poor」など)との共起頻度を求めることで，語句の評価極性の強さを求める手法を提案している．種単語との共起頻度という考え方はシンプルで分かりやすいが，改善の余地があると考えられる．

本論文はHTML文書集合から評価表現辞書を自動構築する手法を提案する．提案手法の概要は以下の通りである．まず，大規模なHTML文書集合から，評価極性を持つ文を自動抽出する(step 1)．以下では，評価極性を持った文を評価文と呼び，抽出された評価文集合のことを評価文コーパスと呼ぶ．次に，評価文コーパスから評価表現の候補を抽出し，その頻度情報を集計する(step 2)．最後に，得られた頻度情報を利用して，候補表現の中から評価表現を選別して辞書に登録する(step 3)．提案手法のポイントは，評価文コーパスを自動構築するときに，再現率を犠牲にして適合率を重視した手法をとることによって，質の高いコーパスの構築を行っている点である．このような厳選された評価文コーパスを使うことによって，既存手法よりも精度よく評価表現を獲得できる効果が期待できる．

2. 評価文コーパスの自動構築

はじめに，HTML文書集合から評価文コーパスを自動構築する方法を説明する．HTML文書中のレイアウト構造やテキスト構造にもとづく手がかりを利用して，評価文を自動抽出するのが基本的なアイデアである．紙面の都合から，ここでは手法の概観を説明することしか出来ないので，詳細ついては文献[3]を参照されたい．

2.1 レイアウト構造の利用

レイアウト構造として，箇条書き形式と表形式で記述された評価文に着目した．まずは箇条書き形式で記述された評価文の具体例を示す．

良い点
- 変に加工しない素直な音を出す．
- 曲の検索が簡単にできる．
- お気に入りのプレイリストを作って楽しめる．

悪い点
- リモコンに液晶表示がない．
- ボディに傷や指紋がつきやすい．
- すぐに電池がなくなる．

図 1 箇条書き形式で記述された評価文
Fig.1 Polar sentences in itemization format.

この箇条書きには「良い点」「悪い点」という見出しが付属しているため，そこに評価文が記述されていることを機械的に判定することができる．この「良い点」「悪い点」のような表現を，以下では手がかり表現と呼ぶ．ある程度の数の手がかり表現を人手でリストアップできれば，このような箇条書き形式から評価文を自動抽出できる．

表形式で記述された評価文も同様である．

燃費（市街地）	7.0km/litter
燃費（高速）	9.0km/litter
満足度	90%
気に入った点	４ドアなのにカッコよすぎる．
イヤな点	シートがぼろくてライトが暗い．色がはげてきている．

図 2 表形式で記述された評価文
Fig.2 Polar sentences in table format.

♦正会員 東京大学生産技術研究所
{kaji,kitsure}@tkl.iis.u-tokyo.ac.jp

この表では，1列目に手がかり表現(気に入った点，イヤな点)が存在していて，これが見出しの働きをしている．そして2列目に評価文が記述されている．具体的な例は省略するが，1列目ではなく1行目が見出しの働きをしているような表も我々は取り扱う．

2.2 テキスト構造の利用

評価文には，以下に示すように定型的なテキスト構造を使って記述されているものがある．

- これの良いところは計算が速いことだ．
- 慣れるまで時間がかかるのが悪いところだ．

上の例文には「計算が速い」，下の例文には「慣れるまで時間がかかる」という評価文が含まれている．いずれも「良いところは～こと」「～ところが悪い点」といった定型的な表現を使って記述されているので，語彙統語パターンを利用して自動抽出することができる．なお，ここでも「良いところ」「悪い点」といった手がかり表現を利用する．

3 評価文コーパスの自動構築

3.1 候補表現の頻度集計

自動構築した評価文コーパスから評価表現の候補(候補表現と呼ぶ)を抽出する．そして，各候補表現の頻度情報を集計する．

これまでの研究でもしばしば指摘されているように，形容詞は評価極性を持ちやすい．ただし，1 単語で評価極性が決まる形容詞(「素晴しい」など)もあれば，決まらない形容詞(「高い」など)もある．後者のような形容詞の評価極性は，係り受けの情報が与えられれば決まる場合が多い．例えば「質が高い」であれば好評極性を持つし「コストが高い」であれば不評極性を持つ．

そこで，全ての形容詞と形容詞句(名詞＋格助詞＋形容詞)を候補表現とした．ただし，一部の機能語表現にたいしては特別な処理を施した．まず，形容詞に否定を表す機能語(「ない」と「ぬ」)が付属している場合は，否定をあらわすタグを形容詞に付与しておき，そうでない場合と区別をする．また，動詞に接尾辞の「やすい」「にくい」が付属している場合，その「動詞＋接尾辞」を 1 つの形容詞として扱った．例えば「使いにくい」「壊れやすい」は 1 つの形容詞と考える．

各候補表現について，それが好評表現として出現した頻度と不評表現として出現した頻度を集計した．単純には，好評文と不評文での出現頻度を求めることになるが，一つの評価文に好評表現と不評表現が混在する場合を考慮し，評価文の主節における出現頻度だけを利用した．

3.2 評価表現の選別

こうして得られた頻度情報を用いて候補表現の評価極性の強さを数値化する．この数値を評価極性値と呼ぶ．そして，それをもとに候補表現の中から評価表現を選別して，評価表現辞書に登録する．

評価極性値は共起度という観点から定義した．候補表現が好評表現(不評表現)として出現した頻度は，候補表現と好評極性(不評極性)が共起した頻度であると解釈することができる．こう考えると，得られた頻度情報を用いて，候補表現と評価極性の共起度を算出できる．

共起の尺度として Pointwise Mutual Information (PMI) を用いた[1]．PMI を用いると，候補表現 c と好評極性 pos または不評極性 neg との共起度は次のように表すことができる．

$$PMI(c, pos) = \log \frac{P(c, pos)}{P(c)P(pos)}$$

$$PMI(c, neg) = \log \frac{P(c, neg)}{P(c)P(neg)}$$

そして，2 つの PMI を用いて候補表現 c の評価極性値 $PV(c)$ を以下のように定義した．

$$PV(c) = PMI(c, pos) - PMI(c, neg)$$

この式を変形すると

$$PV(c) = \log \frac{P(c, pos)}{P(c)P(pos)} - \log \frac{P(c, neg)}{P(c)P(neg)}$$

$$= \log \frac{P(c, pos)/P(pos)}{P(c, neg)/P(neg)}$$

$$= \log \frac{P(c \mid pos)}{P(c \mid neg)}$$

が得られる．

評価極性値を求めるには $P(c|pos)$と $P(c|neg)$の値が必要になるため，これらは次のように推定する．

$$P(c \mid pos) = \frac{f(c, pos)}{\sum_{c'} f(c', pos)}$$

$$P(c \mid neg) = \frac{f(c, neg)}{\sum_{c'} f(c', neg)}$$

式中の $f(c, pos)$は，候補表現 c と好評極性 pos の共起頻度で，前節で求めた頻度情報を使う．$f(c, neg)$も同様である．分母では全ての候補表現に対して $f(c', pos)$の和をとっている．これは好評極性 pos と共起した候補表現 c の述べ数に相当する．

候補表現が評価表現であるかどうかかは，評価極性値と閾値 θ (>0)を用いて以下のように決定して，評価表現と判定されたものだけを辞書に登録する．まず評価極性値が閾値 θ より大きければ，それは好評表現として辞書に登録する．次に，評価極性値が$-\theta$ より小さければ，不評表現として辞書に登録する．上記以外の表現は辞書には登録しない．θ は適合率と再現率のトレードオフを調節するパラメータであり，θ が大きい（小さい）ほど適合率（再現率）重視で評価表現を獲得することになる．なお，低頻度語は頻度情報の信頼性が低いと考えて，好評文または不評文に 3 回以上出現した候補表現だけを用いた．

共起尺度としては，PMI 以外にも様々な尺度を考えることできる．詳細は割愛するが，我々は PMI 以外にもカイ自乗値も試して比較実験を行った．そして，その結果 PMI のほうが優れた精度を示すことを確認している[9].

4. 実験結果

4.1 評価文コーパスの構築

約 10 憶件の HTML 文書を用いて評価文コーパスの構築を行った．その結果，約 50 万文からなる評価文コーパスを構築することができた[1]．その内訳は好評文が 220,716 文，

[1] http://www.tkl.iis.u-tokyo.ac.jp/~kaji/acp

不評文が 288,755 文である．なお構文解析には KNP を用いた．表に実際に抽出された好評文の例を示す．

- 順応性が素晴らしくある．
- 使い方がわかりやすい．
- なんと言っても料金が良心的だ．

同様に不評文の例を示す．

- 費用が高い．
- いい加減な意見，ふざけた意見などが出てくる．
- エンジンが非力で少々うるさい．

自動構築されたコーパスの質を確認するため，コーパス中の 500 文を 2 人の被験者(被験者 A，B と呼ぶ)が個別に調べた．その結果，被験者 A は 91.8%(459/500)の文を適切である判断した．同様に被験者 B は 92%(460/500)の文を適切であると判断した．被験者間での判断の一致率は 93.4%(467/500)であり，これは精度の上限であると考えることができる．このことから，提案手法は非常に高い精度で評価文が獲得できたと結論づけることができる[3]．

不適切であると判断された評価文を観察した結果，そのほとんどは，評価極性が文脈に依存する文であった．例えば，コーパスには「何しろ情報量が多い」が好評文として登録されていたが，被験者は 2 人ともこれを不適切と判断していた．これは，上記の文の評価極性は文脈依存するため，単独では評価極性を決定できないからであると考えることができる．このような文の扱いは今後の課題である．

4.2 評価極性判定

自動構築した評価表現辞書[2]を用いて，形容詞句(名詞＋格助詞＋形容詞)の評価極性を判定する実験を行った．

4.2.1 実験設定

実験のための正解データは以下のような手順で作成した．まず，ウェブ上のテキストから 500 の形容詞句を無作為に抽出して，解析誤りや重複などを人手で除去した．この結果 408 の形容詞句を得た．次に，各形容詞句に「好評」「不評」「中立/不定」の 3 種類のラベルを付与した．「中立/不定」というラベルは，評価極性を持たない形容詞句(中立)と，評価極性を持つが好評極性なのか不評極性なのかは文脈依存する形容詞句(不定)に割りあてる．各ラベルを付与された形容詞句の数は，それぞれ 156，123，129 であった．なお，自動構築した評価文コーパスと，正解データを作成するのに用いたウェブテキストに重複がないことは確認している．

自動構築した評価表現辞書を用いて，正解データ中の形容詞句の評価極性を判定し，適合率と再現率を測る．基本的には辞書に登録されている評価表現との照合を行うだけである．照合に失敗した形容詞句は「中立/不定」と判断する．ただし，形容詞句そのものが辞書に登録されていなくても，形容詞が単独で登録されていれば，照合に成功したと判断する．例えば，正解データに「景色が素晴らしい」という形容詞句があったとする．このとき評価表現辞書に同じ形容詞句が登録されていなくても，形容詞「素晴らしい」が好評表現として登録されていれば，「景色が素晴らしい」は好評極性を持っていると判定される．

比較のため，Turney の提案する評価極性値[1]を用いて評価表現辞書を構築し，それを用いて上記と同じ実験を行った．ただし，本実験では Turney の評価極性値をそのまま使うのではなく，以下のような修正を加えた．まず「excellent」と「poor」の代わりに「最高」と「最低」を使った．そして AltaVista の NEAR 検索の代わりに Google の AND 検索を用いた．

4.2.2 結果

閾値 θ の値を変化させながら，正しく評価極性を判定できた好評表現と不評表現の適合率と再現率を調べた(図 3 と図 4)．図 3 は好評表現の再現率-適合率曲線であり，図 4 は不評表現の再現率-適合率曲線である．

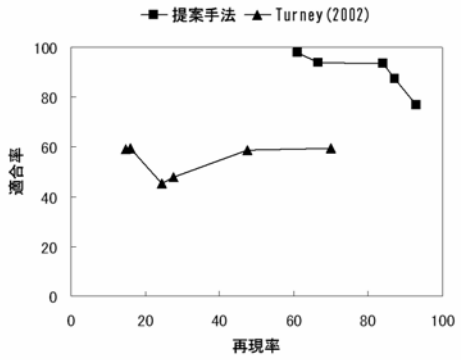

図 3　好評表現の再現率-適合率曲線

Fig.3 Recall-precision curve for positive phrases

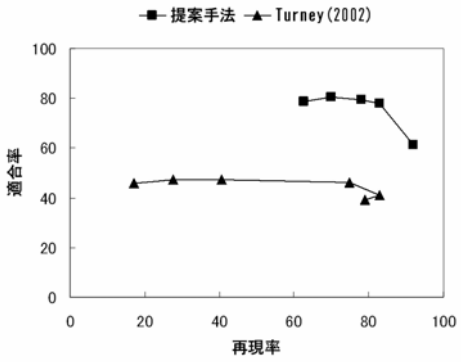

図 4 不評表現の再現率-適合率曲線

Fig.4 Recall-precision curve for negative phrases

4.2.3 結果

実験の結果，提案手法は良好な結果を示した．また提案手法と Turney の手法と比較しても，提案手法のほうが良い精度を示すことが分かった Turney の手法は種単語の取り方次第で精度が向上する可能性はあるものの，提案手法の有効性は確認できたと考える．

表に，獲得された評価表現の総数を示す．閾値のとりかたによるが，約 8,000 から 9,000 の評価表現が獲得されていることが分かる．辞書の規模はまだ拡大する余地があるが，形容詞と形容詞句だけを対象にしていることを考慮すれば，決して小さくないと考えている．

[2] http://www.tkl.iis.u-tokyo.ac.jp/~kaji/polardic

表 1 獲得された評価表現数
Table 1 Number of polar phrases acquired

閾値	1	2	5	10	20
好評表現	3,140	2,928	2,794	2,567	2,422
不評表現	6,455	5,861	5,659	5,486	5,371

5. まとめと今後の課題

本論文では，大規模な HTML 文書集合から自動構築した評価文コーパスを用いて，評価表現辞書を自動構築する方法を提案した．実験の結果，提案手法は既存手法と比較して良好な精度を示すことが確認できた．

最後に今後の課題についていくつか述べておく．提案手法では，形容詞以外の評価表現を取り扱っていない．しかし，動詞や名詞であっても，評価表現になりうるものは存在する(「役に立つ」「絶品だ」など)．こうした評価表現の扱いが重要な課題の一つであると考えている．

細かな問題点は残っているものの，我々はある程度実用的な評価表現辞書を構築することができたと考えている．現在，この評価表現辞書を用いて，大規模なウェブアーカイブからの評価情報抽出実験を進めている．評価情報抽出を行うには，評価表現辞書を用いてテキストの評価極性を判定するだけでなく，評価者や評価対象，または対象の属性も把握する必要があると考えており，そうした課題にも今後取り組む予定である[2].

[謝辞]

本研究は文部科学省リーディングプロジェクト e-society: 先進的なウェブ解析技術によって支援されている．本研究にあたり，生産技術研究所協力研究員田村孝之氏に大変お世話になりました．感謝いたします．

[文献]

[1] Peter D. Turney: “Thumbs up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews”, In Proceedings of ACL, pp.417-424 (2002).

[2] 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 **12**(3), pp.203-222, (2005).

[3] Nobuhiro Kaji and Masaru Kitsuregawa: “Automatic Construction of Polarity-tagged Corpus from HTML Documents'', In Proceedings of COLING/ACL (Poster Sessions), pp.452-459 (2006).

[4] Jaap Kamps and Maarten Marx and Robert J. Mokken and Maarten de Rijke: “Using WordNet to Measure Semantic Orientations of Adjectives”, In Proceedings of LREC, (2004).

[5] Hiroya Takamura and Takashi Inui and Manabu Okumura: “Extracting Semantic Orientation of Words using Spin Model”, In Proceedings of ACL, pp.133-140 (2005).

[6] Andrea Esuli and Fabrizio Sebastiani: “Determining the Semantic Orientation of Terms through Gloss Classification”, In Proceedings of CIKM (2005).

[7] Andrea Esuli and Fabrizio Sebastiani: “Determining Term Subjectivity and Term Orientation for Opinion Mining”, In Proceedings of EACL, pp.193-200 (2006).

[8] Hiroshi Kanayama and Tetsuya Nasukawa: “Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis”, In Proceedings of EMNLP, pp.355-363 (2006).

[9] Nobuhiro Kaji and Masaru Kitsuregawa: “Building Lexicon for Sentiment Analysis from Massive HTML Documents”, In Proceedings of EMNLP-CoNLL, to appear (2007).

鍛治 伸裕 Nobuhiro KAJI

2005 東京大学情報理工学系研究科博士後期課程修了．現在，東京大学生産技術研究所特任助教．自然言語処理の研究に従事．情報処理学会，言語処理学会，ACL，日本データベース学会各会員．

喜連川 優 Masaru KITSUREGAWA

1978 東京大学工学部電子工学科卒業．1983 同大大学院工学系研究科情報工学専攻博士課程修了．工学博士．同年同大生産技術研究所講師．現在，同教授．2003 より同所戦略情報融合国際研究センター長．データベース工学，並列処理，Web マイニングに関する研究に従事．現在，本会理事，情報処理学会，電子情報通信学会各フェロー．ACM SIGMOD Japan Chapter Chair，電子情報通信学会データ工学研究専門委員会委員長歴任．VLDB Trustee (1997-2002), IEEE ICDE, PAKDD,WAIM などステアリング委員．IEEE データ工学国際会議 Program Co-chair(99),General Co-chair(05)..