自動構築した評価文コーパスからの評価表現の獲得

鍛治　伸裕†　　喜連川　優†

† 東京大学 生産技術研究所
〒 153-8505 東京都 目黒区 駒場 4-6-1
E-mail: †{kaji,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし　近年，自然言語処理やウェブマイニングにおいて，Sentiment Analysis や評価情報分析と呼ばれる処理が注目を集めている．こうした処理を行うためには，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．しかし，そうした語彙資源を利用した手法には，網羅性に欠けるという問題や，句を扱えないという問題がある．一方，検索エンジンを使って種単語との共起頻度を求めるという方法も提案されているが，こちらは計算コストが大きく，大規模な評価表現辞書を構築するのには不向きある．また，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，その精度には疑問が残る．そこで，本論文では，HTML 文書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．

キーワード　自然言語処理，評価情報分析，評価極性，HTML 文書

Learning Polarity of Phrases from HTML Documents

Nobuhiro KAJI† and Masaru KITSUREGAWA†

† Institute of Industrial Science, the University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8505 Japan
E-mail: †{kaji,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract　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. In order to build such lexicon, some researchers porposed methods that exploit thesaurus or dictionary. One flaw of this approach is scalability. It cannot handle such words that are not listed in thesaurus or dictionary. In addition, this approach cannot deal with phrases. Co-occurence based method, which uses the number of hits that is obtained by issuing queries to search engine, is another strategy to build lexicon for sentiment analysis. The demerit of this method is computation cost. Since search engine is exploited, it is difficult to process large number of words. It is also problematic that co-occurrence based method relies on naive assumption. This paper proposes a method of building lexicon for sentiment analysis by using polar sentnce corpus that is automatically constructed from HTML documents.

Key words　NLP, Sentiment Analysis, Polarity, HTML Document

1. はじめに

近年，自然言語処理やウェブマイニングにおいて，Sentiment Analysis や評価情報分析と呼ばれる処理が注目を集めている [10]．こうした処理を行うためには，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 [13] [19] [4] [5] [6]．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」

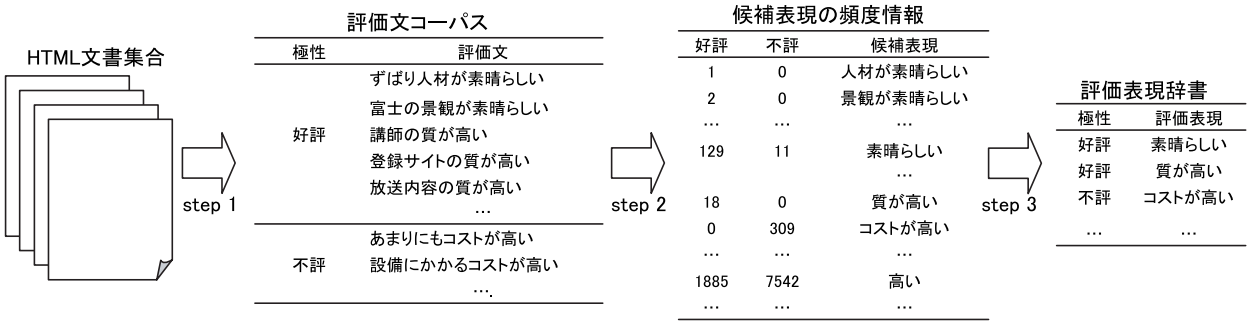


図 1　評価表現辞書構築の流れ

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．

この問題に対する解決策の 1 つとして，Turney の提案する方法を挙げることができる [21] [22]．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．しかし，共起頻度を取得するためには検索エンジンを使わなくてはならないため計算コストが大きく，大規模な評価表現辞書を構築するのには不向きある．また，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，その精度には疑問が残る．

そこで，以下のような辞書構築手法を考案した (図 1)．まず，Kaji ら [11] の考案した手法を用いて，評価極性を持つ文を大規模な HTML 文書集合から自動抽出する (step 1)．以下ではこのような文を評価文と呼び．抽出された評価文は，好評極性を持つ文 (好評文) と不評極性を持つ文 (不評文) に分けて保持される．こうして作られたデータセット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だけを選別して辞書に登録する (step 3)．手法の詳細は 2. 節と 3. 節で説明をする．

この方法のポイントは，厳選された評価文コーパスだけを使うところである．Turney のように検索エンジンを使う必要がないため，手軽に頻度情報を取得できるし，構文解析などの深い言語解析も適用できるようになる．さらに，種単語との共起頻度よりも，評価文コーパスにおける頻度情報のほうが，語句の評価極性の強さを適切に反映していると考えられる．

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から説明をする (step 1)．

2.1　アイデア

コーパス構築のための基本的なアイデアは，定型文/箇条書き/表という形式に着目するというものである [11] [12]．まずはそのアイデアを説明するために，これら 3 種類の形式で記述された評価文の具体例を示す．

2.1.1　定　型　文

評価文には，以下に示すように定型的な表現を使って記述されているものがある．

(1) a.　これの 良いところは 計算が速い ことだ.
　　b.　慣れるまで時間がかかる ところが，悪い点だ.

例文 (1a) には「計算が速い」，例文 (1b) 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，単純な規則で自動抽出することができる．なお，ここでは「良いところ」「悪い点」のように，評価文の存在を示唆している表現のことを手がかり表現と呼ぶこととする．

2.1.2　箇 条 書 き

次に着目したのは，図 2 のような箇条書き形式で記述された評価文である．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が，いずれも手がかり表現である．そのため，箇条書き内に評価文が存在することを機械的に判定できる．

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.
- ライトを点灯し続けると直ぐに電池がなくなる.

図 2　箇条書き形式で記述された評価文

2.1.3　表

表形式で記述されている評価文も，箇条書きの場合とほぼ同様である (図 3)．

燃費 (市街地)	7.0km/litter
燃費 (高速)	9.0km/litter
満足度	95%
気に入った点	４ドアなのにカッコよすぎる.
イヤな点	シートがぼろくライトが暗い，色がはげてきてる.

図 3　表形式で記述された評価文

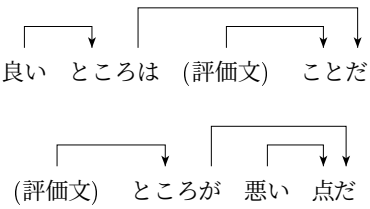
この表では，1 列目に手がかり表現 (気に入った点，イヤな点) が存在していて，これが見出しの働きをしている．そして 2 列目に評価文が記述されている．

2.2 評価文の抽出

これら 3 種類の評価文は単純な規則を用いて自動抽出することができる．

2.2.1 定 型 文

定型文からの抽出には語彙統語パターンを用いた．例えば (1a) と (1b) から評価文を抽出するためには，文を依存構造木に変換したのち，以下のような抽出パターンを用いればよい．



矢印は文節間の依存関係を表していて，(評価文) はマッチした部分木が評価文として抽出されることを表す．なお抽出された評価文の極性は自明である．

実際のコーパス構築では，上記の語彙統語パターンをそのまま使うのではなく，手がかり表現の部分 (良いところ，悪い点) を汎化したものを使用した．以下のような手がかり表現リストを人手で作成しておき，リスト中のどの表現にもマッチするようにした．

良い点，善い点，利点，メリット
悪い点，改善してほしい所, 難点，デメリット

2.2.2 箇 条 書 き

箇条書き形式からは，前述の手がかり表現リストと HTML タグを用いれば，容易に評価文を抽出できる．まず，手がかり表現が見出しになっている箇条書きを見つけて，そして，その箇条書きの項目から文を取り出せばよい．

例えば図 2 からは「変に加工しない素直な音を出す」「曲の検索が簡単にでる」「お気に入りのプレイリストを作って楽しめる」が好評文として,「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」「ライトを点灯し続けると直ぐに電池がなくなる」が不評文として取り出される．

箇条書きからの評価文抽出で問題となるのは，1 つの項目に複数文が記述されている場合の処理である (図 4 の 3 番目の項目)．このような場合は，1 つの項目内に好評文と不評文が混在している可能性があるため，抽出に使わないことにした．すなわち，図 4 からは「発色がものすごくよい」と「撮っていくうちに楽しくなる」の 2 文だけを抽出する．

よい点
- 発色がものすごくよい.
- 撮っていくうちに楽しくなる.
- 400 万画素という高画素．200 万画素では物足りなかった.

図 4 1 項目に複数文が記述されている箇条書き

2.2.3 表

表形式から評価文を抽出するさいには，処理のしやすさを考えて，図 5 のような 2 つのタイプの表だけを考えた．タイプ A は，1 列目に手がかり句があって，その横に評価文があるタイプである．図 3 はこのタイプに相当する．タイプ B は，1 行目に手がかり句があって，その下に評価文があるタイプである．図中の＋と－は好評文と不評文を表し，C_+ と C_- はそれらに対応する手がかり表現を表す．

タイプ A

C_+	＋
C_-	－

タイプ B

	C_+	C_-	
	＋	－	

図 5 評価文を抽出するときに考慮する表タイプ

評価文抽出を行うには，まず与えられた表のタイプを決定しなくてはならない．表のタイプは，1 列目 (1 行目) を調べて手がかり句が出現していればタイプ A(タイプ B) であると判定する．表のタイプが決まれば，あとは図の＋と－に対応するマスから評価文を抽出すればよい．

3. 評価表現の獲得

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 (step 2)．

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，一部の機能語表現を特別扱いをすることにした．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，ここでは以下のような場合を考慮した．

(2) a. 値段が高いですが、そんなことも忘れさせてくれるほどシルエットが美しい.
b. 面倒な準備やテクニックは不要で，非常に簡単です.

いずれも文全体としては肯定的な評価を示しているため，これらは好評文である．しかし，その中に「値段が高い」「面倒だ」といった不評表現が出現している．そこで「好評文 (不評文) の主節には，好評表現 (不評表現) が出現しやすい」と考え，主節における頻度のみを集計した．例えば，例文 (2a) と (2b) からは「シルエットが美しい」「美しい」「簡単だ」の頻度が 1 回ずつ集計される．

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 (step 3)．

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) [1] [21] [22] 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．

$$PMI(c, pos) = \log_2 \frac{P(c, pos)}{P(c)P(pos)} \quad (1)$$

$$PMI(c, neg) = \log_2 \frac{P(c, neg)}{P(c)P(neg)} \quad (2)$$

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．

$$PV(c) = PMI(c, pos) - PMI(c, neg) \quad (3)$$

この式を変形すると

$$PV(c) = \log_2 \frac{P(c, pos)}{P(c)P(pos)} - \log_2 \frac{P(c, neg)}{P(c)P(neg)} \quad (4)$$

$$= \log_2 \frac{P(c, pos)/P(pos)}{P(c, neg)/P(neg)} \quad (5)$$

$$= \log_2 \frac{P(c|pos)}{P(c|neg)} \quad (6)$$

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，これらを次のように推定する．

$$P(c|pos) = \frac{f(c, pos)}{\sum_{c'} f(c', pos)} \quad (7)$$

$$P(c|neg) = \frac{f(c, neg)}{\sum_{c'} f(c', neg)} \quad (8)$$

式 (7) の分母 $f(c, pos)$ 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．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．式 (8) も同様である．

結局 $P(c|pos)$ と $P(c|neg)$ は，好評文と不評文 (の主節) における候補表現 c の出現確率となる．すなわち，上記のように評価極性値 $PV(c)$ を計算するということは，この 2 つの確率の比を求めることに相当するため，これは直感的に自然である．また評価極性値 $PV(c)$ は $P(c|neg) < P(c|pos)$ のときに正の値をとり，逆のときは負の値をとることになる．

候補表現が評価表現であるかどうかは，評価極性値と閾値 $\theta(\geqq 1)$ を用いて以下のように決定する．そして，評価表現だけを辞書に登録する．

$\log_2 \theta < PV(c)$	$\rightarrow$	評価表現 (好評極性)
$PV(c) < -\log_2 \theta$	$\rightarrow$	評価表現 (不評極性)
$-\log_2 \theta \leqq PV(c) \leqq \log_2 \theta$	$\rightarrow$	非評価表現

θ の値を変化させることによって，適合率と再現率のトレードオフを調節することができる．ただし，低頻度語は頻度情報の信頼性が低いと考えて，$f(c, pos)$ または $f(c, neg)$ のいずれかが 3 以上である候補表現だけを用いた．

4. 実験結果

4.1 評価文コーパスの構築

約 10 億件の 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(注1)．その内訳は好評文が 220,716 文，不評文が 288,755 文である．表 Table1 に実際に抽出された評価文の例を示す．なお構文解析には KNP(注2)を用いた．以下の実験でも同様である．

表 1 評価文の例

極性	評価文
好評	順応性が素晴らしくある．
	使い方がわかりやすい．
	何と言っても，料金が良心的だ．
不評	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．

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．

4.2 評価極性判定

自動構築した評価表現辞書を用いて，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．

(注1)：http://www.tkl.iis.u-tokyo.ac.jp/~kaji/acp/
(注2)：http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

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 [21] の提案する評価極性値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．Turney の評価極性値は以下の式で与えられる．

$$\log_2 \frac{hits(c\ NEAR\ \text{“}excellent''\text{)}hits(\text{“}poor''\text{)}}{hits(c\ NEAR\ \text{“}poor''\text{)}hits(\text{“}excellent''\text{)}} \qquad (9)$$

$hits(query)$ は検索語 $query$ の AltaVista のヒット数であり，NEAR は 2 つの検索語の NEAR 検索を行うこと意味する．ただし，実験では「excellent」と「poor」の代わりに「最高」と「最低」を使い，AltaVista の NEAR 検索の代わりに Google の AND 検索を用いた．

4.2.2 実験結果

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 6)．図中の再現率-適合率曲線は，左が好評表現で右が不評表現に対応する．閾値などの詳細な値は表 2 と表 3 に示すとおりである．

表 2　適合率と再現率 (提案手法)

θ	1	2	5	10	20
好評表現					
適合率	76.6	87.2	93.6	93.7	97.9
再現率	92.9	87.2	84.0	66.7	60.9
不評表現					
適合率	61.1	77.9	79.3	80.4	78.6
再現率	91.9	82.9	78.0	69.9	62.6

4.3 議　　論

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 の手法と比べると，提案手法のほうが良い結果を示した．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．Turney の手法との詳しい議論は 5. 節で行う．

表 3　適合率と再現率 (Turney(2002))

θ	1	1.2	1.4	1.6	1.8
好評表現					
適合率	59.4	58.7	47.7	45.2	59.5
再現率	69.9	47.4.	27.6	24.4	16.0
不評表現					
適合率	39.1	41.1	46.0	47.1	47.2
再現率	78.9	82.9	74.8	40.7	27.6

表 4 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．

表 4　獲得された評価表現数

θ	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

閾値 $\theta = 5$ のときに，実際に獲得された評価表現とその評価極性値の例を表 5 に示す．単語だけでなく「支障が無い」「魅力が無い」のような句の評価表現も獲得されていることが分かる．また「しょぼい」や「ダサい」といった口語もうまく扱えている．こういった単語は既存の語彙資源を使った手法では獲得が困難である．

表 5　評価表現の具体例

評価表現	評価極性値
謙虚だ	7.55
支障が無い	7.45
エキサイティングだ	6.51
漏れが少ない	6.13
能力が高い	4.06
ダサい	-2.86
厄介だ	-3.33
消耗が早い	-3.67
魅力が無い	-3.76
しょぼい	-4.47

提案手法は，好評文と不評文を利用して評価表現辞書を構築している．つまり，評価極性を持たない文の情報は使っていないことになる．評価極性を持たない表現の重要性はこれまでにも指摘されており [24] [17] [10] [5]，極性を持たない文を使うことによって，より高い精度で辞書を構築できる可能性がある．しかし，問題は極性を持たない文のコーパスが入手困難なことであり，そのような言語資源の整備が待たれる．

5. 関連研究

5.1 語彙資源に基づく辞書構築

シソーラスや国語辞典などの語彙資源を活用する場合，最

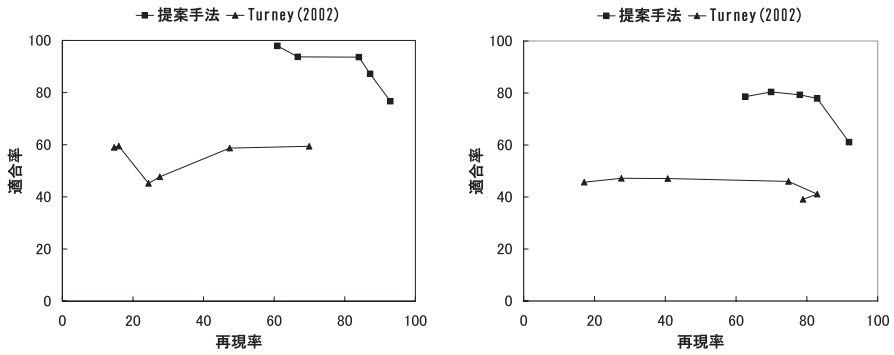


図 6　再現率-適合率曲線 (左:好評表現，右:不評表現)

も広く用いられるのが，語彙ネットワークを構築する手法である [13] [9] [19]．例えば，Kamps [13] は，シソーラスを用いて類義語同士を結んだ語彙ネットワークを構築し，そのネットワーク上での経路長にもとづいて，語の評価極性を判定している．この方法は，類義語同士は同一の評価極性を持つという考えにもとづく．

Esuli らは，WordNet [7] の gloss を分類することによって単語の評価極性を決定する手法を考案し，評価表現辞書 SentiWordNet(注3)を構築した [4] [5] [6]．彼らは少数の訓練事例を準備

すでに議論したように，語彙資源に基づく手法は，そこ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，網羅性に欠けるという問題と，句を扱うことができないという問題がある．

5.2　コーパスに基づく辞書構築

Hatzivassiloglou らは，前述のような語彙ネットワークをコーパスから構築して，単語の評価極性を判定する手法を提案している [8]．しかし，この方法も句を扱うことができないため，我々が指摘した問題を解決されていない．

Kanayama らは，評価表現同士の接続関係に着目して，初期評価表現辞書 (3,275 表現が登録されている) をもとに新たな評価表現を獲得する手法を考案している [14]．Kanayama らが議論しているのは，辞書の自動構築というよりも正確には辞書の自動拡張であり，我々とはやや異なる問題設定であると言える．我々の問題設定は，初期辞書に評価表現が全く登録されていない場合に相当するが，彼らはそのような場合について議論をしていない．また Kanayama らがドメイン依存した評価表現の獲得を主に議論しているのに対して，本論文や上で紹介した研究ではドメイン非依存な評価表現の獲得を議論しており，研究の方向性にも違いがみられる．こうした 2 つの研究の方向性は相補的なものであるため，互いの手法を組み合わせることがおもしろいと考えている．例えば，自動構築した辞書を初期辞書として使えば，提案手法では獲得できなかったドメイン依存した評価表現を獲得できるようになる可能性がある．

Turney は検索エンジンを用いて種単語 (「excellent」「poor」) との共起を測ることで，語句の評価極性を求める手法を提案している [21] [22]．Turney の手法も提案手法も，PMI に基づいて評価極性値を定義しているため，2 つの評価極性値には類似点が見られる．実際，式 (6) の $P(c|pos)$ と $P(c|neg)$ に

$$P(c|pos) = \frac{hits(c\ NEAR\ \text{“}excellent\text{''})}{hits(\text{“}excellent\text{''})} \tag{10}$$

$$P(c|neg) = \frac{hits(c\ NEAR\ \text{“}poor\text{''})}{hits(\text{“}poor\text{''})} \tag{11}$$

を代入すれば，Turney の提案する評価極性値である式 (9) が得られる．つまり．2 つの評価極性値の相違点は $P(c|pos)$ と $P(c|neg)$ の推定の仕方であると解釈することもできる．我々が自動構築した評価文コーパスに基づいてこれらの値を求めているのに対して，Turney は種単語との共起を利用している．実験の結果 2 つの手法には差が出たが，単純な種単語との共起よりも，自動構築した評価文コーパスのほうが精度が高いからであると考えられる．

Turney の方法を使って辞書を構築するときの議論しなくてはいけないことは，候補表現をどのように設定するかである．すでに指摘したことであるが，Turney の方法は検索エンジンを使うため計算コストが高い．そのため，大量の候補表現を処理することは現実的ではなく，候補表現を十分に絞りこむ必要がある．

5.3　その他の研究

評価表現辞書の自動構築という観点からは外れるが，本論文に関わりのある研究をいくつか紹介する．

小林らは評価表現を半自動で収集する手法を提案している [18]．実験では，人手で収集を行った場合と作業時間の比較を行い，手法の有効性の検証を行っている．ただし，彼女らは評価表現の極性判定は行っていない．

Wilson らや Takamura らは，人手でタグが付与されたデータを用いて句の評価極性の学習を行っている [20] [23]．評価表現辞書は，こうした研究で利用されるタグ付きデータに近い位置付けのものである．

6.　今後の課題

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も重要な課題であると考えている．

(注3)：http://sentiwordnet.isti.cnr.it/

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が構築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いる [9] [2] [3] [15] [16].

7. おわりに

評価情報分析のためには，大規模な評価表現辞書が必要であり，その構築方法は重要な研究課題である．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自動構築する手法が提案されているが，新語や口語に弱く網羅性に欠けるという問題や，句を扱えないという問題がある．一方，検索エンジンを用いて種単語との共起頻度を求めるような手法にも，計算コストが大きいなどの問題が存在した．

このことを踏まえ，本論文では HTML 文書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手法を提案した．この方法のポイントは，厳選された評価文コーパスだけを使うところである．検索エンジンを使う必要がないため，手軽に頻度情報を取得できるし，構文解析などの深い言語解析も適用できるようになる．さらに，種単語との共起頻度よりも，評価文コーパスにおける頻度情報のほうが，語句の評価極性の強さを適切に反映していると考えられる．実験の結果，既存の手法と比較して十分高い精度で評価表現を獲得できることを確認した．

文　献

[1] Kenneth Ward Church and Patric Hanks, “Word Association Norms, Mutual Information, and Lexicography”, In Proceedings of ACL, pp. 76-83, 1989.

[2] Yejin Choi and Claire Cardie and Ellen Riloff and Siddharth Patwardhan, “Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns”, In Proceedings of HLT/EMNLP, 2005.

[3] Yejin Choi and Eric Breck and Claire Cardie, “Joint Extraction of Entities and Relations for Opinion Recognition”, In Proceedings of EMNLP, pp. 431-439, 2006.

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

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

[6] Andrea Esuli and Fabrizio Sebastiani, “SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining”, In Proceedings of LREC, 2006.

[7] Christiane Fellbaum, “WordNet: An Electoronic Lexical database”, MIT Press, Cambridge, 1998.

[8] Vasileios Hatzivassiloglou and Katheleen R. McKeown, “Predicting the Semantic Orientation of Adjectives”, In Proceedings of ACL, pp.174-181, 1997.

[9] Minqing Hu and Bing Liu, “Mining and Summarizing Customer Reviews”, In Proceedings of KDD, pp.168-177, 2004.

[10] 乾孝司，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 **13**(3), pp.201-242, 2006.

[11] 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.

[12] 鍛治伸裕，喜連川優, “WWW を用いた評価極性タグ付きコーパスの自動構築”, 言語処理学会第１２回年次大会, pp.61-64, 2006.

[13] 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.

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

[15] Soo-Min Kim and Eduard Hovy, “Identifying and Analyzing Judgement Opinions”, In Proceedings of NAACL-HLT, 2006.

[16] Soo-Min Kim and Eduard Hovy, ‘’Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text‘’, In Proceedings of Workshop on Sentiment and Subjectivity in Text, pp.1-8, 2006.

[17] Moshe Koppel and Jonathan Schler, ‘’The Importance of Neutral Examples for Learning Sentiment‘’, In Proceedings of Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations, 2005.

[18] 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.

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

[20] Hiroya Takamura and Takashi Inui and Manabu Okumura, “Latent Variable Models for Semantic Orientation of Phrases”, In Proceedings of EACL, pp.201-108, 2006.

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

[22] Peter D. Turney and Michael L. Littman, “Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus”, National Research Council, Institute for Information Technology, Technical Report ERB-1094(NRC #44929), 2002.

[23] Theresa Wilson and Janyce Wiebe and Paul Hoffmann, “Recognizing Contextual Polarity in Phrase-level Sentiment Analysis”, In Proceedings of HLT/EMNLP, 2005.

[24] Hong Yu and Yasileios Hatzivassiloglou, “Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences”, In Proceedings of the EMNLP, 2003.