HTML 文書からの評価表現辞書の自動構築

鍛治伸裕　　喜連川優

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

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技術が注目を集めている．そのようなテキストを解析するためには，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) が必要不可欠となる．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．提案手法の概要は図 1 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)．

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を自動抽出する手法について説明する (step 1)．自動抽出には，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する [1]．

2.1　レイアウト構造の利用

レイアウト構造は箇条書き形式と表形式の 2 種類を利用した．例えば，図 2 のような箇条書きは「良い点」「悪い点」という見出しを持っているため，箇条書きに評価文が記述されていることを判定できる．本論文では「良い点」「悪い点」のような，評価文の存在を示唆する表現を手がかり表現と呼ぶ．手がかり表現リストを人手で作成して，それと HTML タグを利用して評価文を自動抽出した．以下に手がかり表現リストの一部を示す

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

表形式も箇条書き形式の場合とほぼ同様である (図 3)．表の 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 列目には評価文が記述されている．

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

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

2.2　テキスト構造の利用

次に，定型的なテキスト構造に着目した．

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

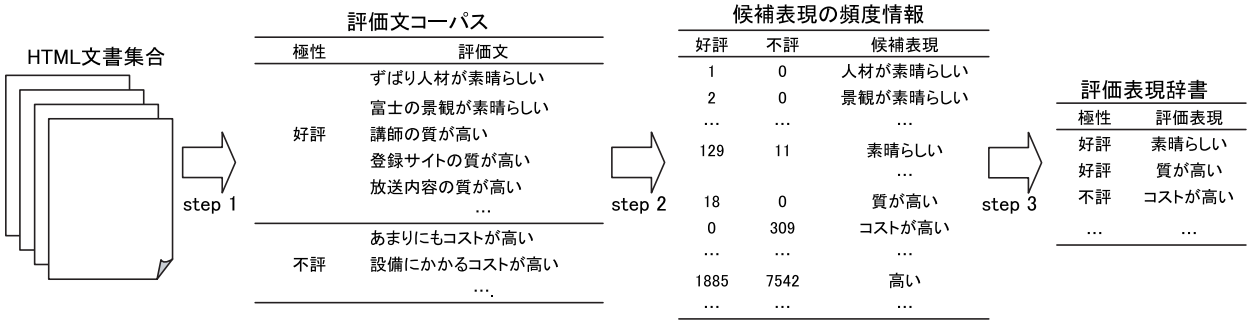
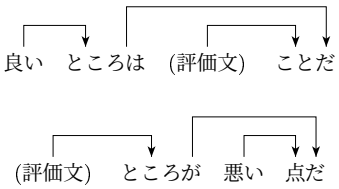


図 1: 評価表現辞書構築の流れ

例文 (1) には「計算が速い」，例文 (2) 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[1]．いずれも「良いところは〜こと」「〜ところが悪い点」といった定型的なテキスト構造を使って記述されている．

このような評価文は，以下のような語彙統語パターンを用いて自動抽出する．



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．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用いるのでなく，手がかり表現の部分 (良いところ，悪い点) を，前述の手がかり表現リストを用いて汎化したものを使った．

2.3 評価文コーパス

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[2]．その内訳は好評文が 220,716 文，不評文が 288,755 文である．表 1 に実際に抽出された評価文の例を示す．構文解析には KNP[3]を用いた．以下の実験でも同様である．

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 [1]．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，このことから，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．

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．

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．

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)．

しばしば指摘されるように，形容詞は評価極性を持ちやすい．また，1 単語で評価極性が決まる形容詞もあれば，そうでない語 (「高い」など) も存在する．これらを踏まえて，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処理をする．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，

[1] 厳密には文ではなく節と呼ぶべきだが，レイアウト構造を用いて抽出される評価文との整合性を考えて文と呼ぶ．
[2] http://www.tkl.iis.u-tokyo.ac.jp/~kaji/acp/
[3] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

否定をあらわすタグを形容詞に付与する．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」は 1 つの形容詞と考える．

各候補表現について，好評文と不評文における出現頻度を集計した．単純には，好評表現 (不評表現) は好評文 (不評文) に多く出現すると考えられるが，以下のような例外的な場合が存在する．

(1) 面倒な 準備やテクニックは不要で，非常に簡単です．

この例文は文全体としては肯定的な評価を示しているため，好評文である．しかし，その中に「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」と仮定して，主節における頻度のみを集計した．

3.2 評価表現の選別

候補表現の評価極性の強さを数値化して (この数値を評価極性値と呼ぶ)，それにもとづき候補表現の中から評価表現だけを選別する (step 3)．

各候補表現 c に対して，次のような分割表を作成することができる．

表 2: 分割表

	pos	neg
c	$f(c,pos)$	$f(c,neg)$
$\neg c$	$f(\neg c,pos)$	$f(\neg c,neg)$

$f(c,pos)$ は候補表現 c の好評文における頻度，$f(\neg c,pos)$ は c 以外の候補表現の頻度の和である．$f(c,neg)$ と $f(\neg c,neg)$ も同様である．この分割表から c の評価極性値を規定する．実験では比較のため，次の 2 種類の手法を試した．

χ^2 値にもとづく評価極性値　候補表現 c の出現の偏りを見積るために χ^2 値を利用した．表 2 から求めた χ^2 値は次のようになる．

$$\chi^2(c) = \sum_{x\in(c,\neg c)} \sum_{y\in(pos,neg)} \frac{\{f(x,y) - \hat{f}(x,y)\}^2}{\hat{f}(x,y)}$$

$\hat{f}(x,y)$ は，候補表現 c の出現確率が好評文と不評文で独立であると仮定したときの $f(x,y)$ の期待値である．

$\chi^2(c)$ には，c が好評文と不評文のどちらに多く出現しているのかという情報は反映されていない．そこで $\chi^2(c)$ を用いて，以下のように評価極性値を設定した．

$$PV_{\chi^2}(c) = \begin{cases} \chi^2(c) & \text{if } P(c|neg) < P(c|pos) \\ -\chi^2(c) & \text{otherwise} \end{cases}$$

$P(c|pos)$ は c の好評文における出現確率であり，$P(c|neg)$ は不評文における出現確率である．

$$P(c|pos) = \frac{f(c,pos)}{f(c,pos)+f(\neg c,pos)}$$
$$P(c|neg) = \frac{f(c,neg)}{f(c,neg)+f(\neg c,neg)}$$

PMI にもとづく評価極性値　PMI (Pointwise Mutual Information) を用いると，候補表現 c と好評文 pos (不評文 neg) の関連の強さは次のように定義できる．

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

この 2 つの数値の差を評価極性値とした．これは Turney と同様の考え方である [2]．

$$\begin{aligned} PV_{PMI}(c) &= PMI(c,pos) - PMI(c,neg) \\ &= \log_2 \frac{P(c,pos)/P(pos)}{P(c,neg)/P(neg)} \\ &= \log_2 \frac{P(c|pos)}{P(c|neg)} \end{aligned}$$

評価表現の選別　上記のように定義した評価極性値と閾値 $\theta(>0)$ を用いて，ある候補表現 c が評価表現であるかどうかを判定する．まず $\theta < PV(c)$ であれば，その候補表現は好評表現と考える．同様に，$PV(c) < -\theta$ であれば不評表現とする．それ以外は評価表現ではないと考える．

4 実験結果

自動構築した評価表現辞書を用いて，テストデータから評価表現を抽出する実験を行った．

テストデータは，ウェブテキストから無作為に抽出した 405 の形容詞句に評価極性 (好評/不評/中立) をタグ付けして作成した．タグ付けの結果，好評/不評/中立の数はそれぞれ 158/150/97 であった．同一データを二人の被験者がタグ付けしたところ Kappa 値は 0.73 であった．

表 3: 評価表現辞書の大きさ ($PV_{\chi^2}(c)$)

θ	0	10	20	30	40	50	60
評価表現数	9,670	2,056	1,047	698	533	423	335

表 4: 評価表現辞書の大きさ ($PV_{PMI}(c)$)

θ	0	0.5	1.0	1.5	2.0	2.5	3.0
評価表現数	9,670	9,320	9,039	8,804	8,570	8,398	8,166

このテストデータから評価表現を抽出する．基本的には，テストデータに含まれる形容詞句を辞書引きしていくことになる．ただし，辞書に登録されている形容詞 (「素晴しい」など) は，その形容詞を含む全ての形容詞句 (「景色が素晴しい」など) とマッチさせる．

比較のために，Turney[2] の提案する評価極性値を用いて評価表現辞書を構築し，同様の実験を行った．Turney の手法は「excellent」「poor」のような種単語が必要となるがここでは「最高」「最低」を用いた．検索エンジンには我々の研究室で開発したローカル検索エンジンと Google の 2 つを試した．前者は約 1 億 5,000 万件の HTML 文書をインデックスしている．

評価表現抽出の結果を適合率と再現率で評価した (図 4)．上のグラフは好評表現抽出の適合率と再現率を，閾値 θ を変化させながら観察したものである．下のグラフは同様のことを不評表現に対して行った結果である．実験の結果，提案手法は Turney の手法よりもうまく働くことが確認できた．また，PMI にもとづく評価極性値は，χ^2 値にもとづくものよりも優れていることが分かった．表 5 に評価表現の具体例を示す．また，獲得された評価表現数を表 3 と表 4 に示す．

表 5: 評価表現の具体例

評価表現	$PV_{\chi^2}(c)$	$PV_{PMI}(c)$
謙虚だ	38.3	11.9
エキサイティングだ	13.5	10.4
能力が高い	113.0	6.9
ダサい	-2.9	-3.1
消耗が早い	-17.7	-4.3
しょぼい	-55.3	-9.1

5 おわりに

本論文では 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した．そして実験を行い手法の有効性を検証した．今後は，この辞書を用いて，実際のテキストから評価情報抽出を行う予定である．

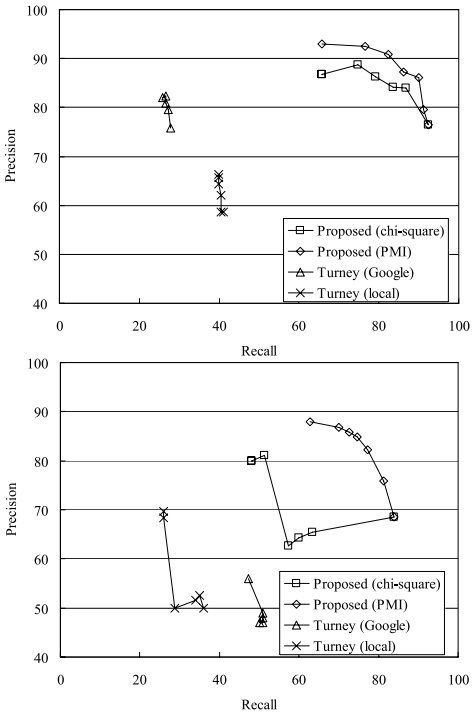


図 4: 再現率-適合率曲線 (上: 好評表現，下: 不評表現)

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の田村孝之氏に大変お世話になりました．感謝致します．

参考文献

[1] 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.

[2] Peter D. Turney. Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL*, pp. 417–424, 2002.