

HTML 文書集合からの評価文の自動収集

鍛冶 伸裕[†]

喜連川 優[†]

本論文では大規模な HTML 文書集合から評価文を自動収集する手法を提案する。基本的なアイデアは「定型文」「箇条書き」「表」といった記述形式を利用するというものである。本手法に必要なのは少数の規則だけであるため、人手をほとんどかけずに評価文を収集することが可能である。また、任意の HTML 文書に適用できる手法であるため、様々なドメインの評価文を収集できることが期待される。実験では、提案手法を約 10 億件の HTML 文書に適用したところ、約 65 万の評価文を獲得することができた。

キーワード: 評価情報分析, 評価極性

Acquiring Polar Sentences from HTML Documents

NOBUHIRO KAJI[†], MASARU KITSUREGAWA[†]

This paper represents a method of acquiring polar sentences from HTML documents. The basic idea is to exploit three lexico-syntactic patterns and two layout structures of HTML documents. The method requires only a small amount of hand-crafted rules and can be implemented in low cost. In our experiment, the method was applied to one billion documents and 650 thousands polar sentences were acquired.

KeyWords: *Sentiment Analysis, Polarity, Semantic Orientation*

1 はじめに

近年、自然言語処理において評価情報処理が注目を集めている(乾孝司, 奥村学 2006)。評価情報処理とは、物事に対する評価が記述されたテキストを検索、分類、要約、構造化するような処理の総称であり、国家政治に対する意見集約やマーケティングといった幅広い応用を持っている。具体的な研究事例としては、テキストから特定の商品やサービスに対する評価情報を抽出する処理や、文書や文を評価極性(好評と不評)に応じて分類する処理などが議論されている(小林のぞみ, 乾健太郎, 松本裕二, 立石健二, 福島俊一 2005; Pang, Lee and Vaihyanathan 2002; Kudo and Matsumoto 2004; Matsumoto, Takamura and Okumura 2005; 藤村滋, 豊田正史, 喜連川優 2005; 箆島郁子, 嶋田和考, 遠藤勉 2005; McDonald, Hannan, Neylon, Wells and Reynar 2007)。

評価情報処理を行うためには様々な言語資源が必要となる。例えば、評価情報を抽出するためには「良い」「素晴らしい」「ひどい」といった評価表現を登録した辞書が不可欠である(小林のぞみ他 2005)。また、文書や文を評価極性に応じて分類するためには、評価極性がタグ付けさ

[†] 東京大学生産技術研究所, Institute of Industrial Science, University of Tokyo

れたコーパスが教師あり学習の訓練事例として使われる (Pang et al. 2002) . 我々は , 評価情報処理のために利用する言語資源の一つとして , 評価文コーパスの構築に取り組んでいる . ここで言う評価文コーパスとは , 何かの評価を述べている文 (評価文) とその評価極性を示すタグが対になったデータのことである (図 1) . タグは好評と不評の 2 種類を想定している . 大規模な評価文コーパスがあれば , それを評価文分類器の訓練事例として利用することや , そのコーパスから評価表現を獲得することが可能になると考えられる .

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

図 1: 評価文コーパスの例 . + は好評極性 , - は不評極性を表す .

This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.

図 2: 不評文書に好評文が出現するレビュー文書

評価文コーパスを構築するには , 単純に考えると以下の 2 つの方法がある . 人手でコーパスを作成する方法と , ウェブ上のレビューデータを活用する方法である . 後者は , 例えばアマゾン¹のようなサイトが利用するというものである . アマゾンに投稿されているレビューには , そのレビューの評価極性を表すメタデータが付与されている . そのため , メタデータを利用することによって , 好評内容のレビューと不評内容のレビューを自動的に収集することができる .

しかしながら , このような方法には問題がある . まず , 人手でコーパスを作るという方法は , 大規模なコーパスを作ることを考えるとコストが問題となる . また , レビューデータを利用する方法には , 文単位の評価極性情報を取得しにくいという問題がある . 後者の具体例として図 2 に示すレビュー文書 (Pang et al. 2002) を考える . これは文書全体としては不評内容を述べているが , その中には好評文がいくつも出現している例である . このような文書を扱う場合 , 文書単位の評価極性だけでなく , 文単位の評価極性も把握しておくことが望ましい . しかし , 一般的にレビューのメタデータは文書に対して与えられるので , 文単位の評価極性の獲得は難しい . さらに , レビューデータを利用した場合には , 内容が特定ドメインに偏ってしまうという

¹ <http://amazon.com/>

問題もある。

こうした問題を踏まえて、本論文では大規模な HTML 文書集合から評価文を自動収集する手法を提案する。基本的なアイデアは「定型文」「箇条書き」「表」といった記述形式を利用するというものである。本手法に必要なのは少数の規則だけであるため、人手をほとんどかけずに大量の評価文を収集することが可能となる。また、評価文書ではなく評価文を収集対象としているため、図 2 のような問題は緩和される。さらに任意の HTML 文書に適用できる方法であるため、様々なドメインの評価文を収集できることが期待される。実験では、提案手法を約 10 億件の HTML 文書に適用したところ、約 65 万の評価文を獲得することができた。

2 アイデア

提案手法は「定型文」「箇条書き」「表」という 3 つの記述形式を利用して評価文を自動抽出する。本節では、これら 3 つの形式で記述された評価文の例を概観して、基本的な考え方を説明する。手法の詳細は次節で述べる。

2.1 定型文

まず我々が着目したのは定型的な評価文である。

- (1) a. この 良いところは 計算が速い ことです。
- b. 悪い点は、慣れるまで時間がかかる こと。

いずれの評価文も「良いところ/悪い点は～なこと」という定型的な表現を使って記述されている。そのため、下線部にマッチするような語彙統語パターンを用意すれば、四角で囲まれたテキストを評価文として抽出することができる²。

以下では「良いところ」「悪い点」のように、評価文の存在を示唆する表現のことを手がかり句と呼ぶ。特に好評文の存在を示す手がかり句を「好評手がかり句」と呼ぶ。例えば「良いところ」は好評手がかり句である。同様に、不評文の存在を示す手がかり句を「不評手がかり句」と呼ぶ。

2.2 箇条書き

次に着目したのは、図 3 のように箇条書き形式で列挙された評価文である。この箇条書きは手がかり句(良い点, 悪い点)を見出しに持つため、各項目に評価文が含まれていることが分かる。

² 四角で囲まれたテキストは正確には文ではなく句と呼ぶべきであるが、箇条書きと表から抽出される評価文との整合性を考えて、ここでは文と呼ぶことにする。

<p>良い点</p> <ul style="list-style-type: none"> ● 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 . <p>悪い点</p> <ul style="list-style-type: none"> ● 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 .
--

図 3: 箇条書き形式で記述された評価文

2.3 表

箇条書きと同様に，図 4 のような表形式からも評価文を自動収集することができる．この表は左側の列が見出しの働きをしているが，ここにも手がかり句（気に入った点，イヤな点）が使われているので，表中に評価文が記述されていることが分かる．

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 4: 表形式で記述された評価文

3 評価文の自動収集

HTML 文書から評価文を収集する手続きは次のようになる．

1. 手がかり句のリストを作成する．
2. HTML 文書をタグとテキストに分割する．一部の箇条書きと見出しは HTML タグを使わないで記述されているので，ルールでタグを補完する．
3. 手がかり句のリストを利用して「定型文」「箇条書き」「表」から評価文を抽出する．

以下では，まず実験で用いた手がかり句について説明する．そして「定型文」「箇条書き」「表」から評価文を抽出する方法を順に説明する．

3.1 手がかり句

実験で用いた手がかり句の一覧を表 1 に示す。これらは予備実験を通して人手で選定した。表中の動詞, 形容詞, 形容動詞(「良い」など)は「所」「点」「面」という 3 つの名詞と組み合わせる。例えば「良い」は「良い所」「良い点」「良い面」の 3 つを手がかり句として使うことを意味する。「長所」や「メリット」のような名詞は, 単語そのものを手がかり句として使う。なお, 詳細は省略しているが「駄目な所」と「ダメな所」または「良い所」と「良いところ」のような表記揺れも網羅的に人手で記述している。

表 1: 実験で使用した手がかり句

好評手がかり句	不評手がかり句
良い, いい, 素晴らしい, 楽しい, 面白い, 最高だ, 満足だ, 便利だ, グッドだ, プラスだ, 良好だ, 大好きだ, 好きだ, 優れる, 長所, 利点, メリット	悪い, 最悪だ, 残念だ, 不満足だ, 不満だ, 嫌だ, 大嫌いだ, 嫌いだ, 駄目だ, いまいちだ, いまひとつだ, パッドだ, マイナスだ, 不便だ, 困る, 短所, デメリット, 問題点, 不満点, 欠点, 難点, 弱点

3.2 定型文からの抽出

定型文から評価文を抽出するために, 3 種類の語彙統語パターンを人手で作成した。各パターンとそれにマッチする定型文の具体例を表 2 に示す。

最初のパターン(表 2 上)は, 主部が手がかり句(良いところ)で述部が評価文であるような定型文にマッチする。パターン中の矢印は文節間の依存関係, 「手がかり句」は手がかり句をそれぞれ表している。また「評価文」は, この部分にマッチしたテキストが評価文として抽出されることを意味する。表の右側に評価文が抽出される様子を示す。

残り 2 つのパターンも同様である。それぞれ, 主部が評価文で述部が手がかり句である定型文にマッチするパターン(表 2 中)と, 評価文と手がかり句が同格になっている定型文にマッチするパターン(表 2 下)である。

3.3 箇条書きからの抽出

箇条書きからの評価文抽出は, 手がかり句リストと HTML タグを利用すれば容易に実現できる。すなわち, 手がかり句が見出しになっている箇条書きを見つけて, その箇条書きの項目を順に取り出していけばよい。例えば, 前節の図 3 からは「変に加工しない素直な音を出す」

表 2: 語彙統語パターンと評価文抽出の例

語彙統語パターン	抽出例
<p>手がかかり句 は 評価文 ことだ</p>	<p>この「良いところ」は「計算が速い」ことです。</p>
<p>評価文 ところが 手がかかり句</p>	<p>「お料理の盛り付けがイマイチな」ところが「難点」ですが。</p>
<p>評価文 という 手がかかり句</p>	<p>「所要時間が短縮される」という「メリット」があります。</p>

「曲の検索が簡単にできる」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付きやすい」が不評文として取り出される。

ここで問題となるのは、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5の3番目の項目)。このような場合は1項目に好評文と不評文が混在している可能性がある。各文の評価極性を自動判定することは難しいので、1つの項目に複数文が存在した場合、その項目は抽出に使わないことにした。例えば図5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。」「カメラ背面の液晶画面が大きく、見やすい。」が好評文として抽出される。

<p>よい点</p> <ul style="list-style-type: none"> ● 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。 ● カメラ背面の液晶画面が大きく、見やすい。

図 5: 1項目に複数文が記述されている箇条書き

3.4 表からの抽出

最後に、表から評価文を抽出する方法を述べる。基本的には手がかかり表現と <table> タグを利用すればよいのだが、HTML 文書には多種多様な表が出現するので、あらゆる表に対応し

た抽出規則を作成することは難しい．そこで 2 種類の表だけを考えることにした (図 6)．図中の C_+ と C_- は好評手がかり句と不評手がかり句を表し，+ と - は好評文と不評文を表す．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．前節で紹介した図 4 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．

タイプ A						タイプ B			
C_+	+					C_+	C_-		
C_-	-					+	-		

図 6: 評価文を抽出するときに考慮する表タイプ

与えられた表のタイプは，1 列目 (1 行目) を調べて，好評手がかり句と不評手がかり句の両方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．ただし，1 つのマ스에複数文が記述されている場合は抽出対象としない．これは箇条書きの 1 項目に複数文が記述されている場合と同様の理由からである．

4 実験

約 10 億件の HTML 文書集合を用いて評価文の収集実験を行った．この数字にはミラーサイトなどの重複文書も含まれている．表 3 に収集された評価文の数とその内訳 (好評文か不評文か) を示す．ただし，ミラーサイトから全く同じ評価文が繰り返し収集される可能性を考慮して，同一の評価文が複数回抽出された場合は集計に入れていない．定型文は，3 種類の語彙統語パターン (表 2) から抽出された評価文数を分けて示している．定型文 1, 2, 3 というのは，それぞれ表 2 の上, 中, 下に記述されたパターンに相当する．表 4 に自動収集された評価文の一例を示す．

収集された評価文から 250 文 (各抽出法ごとに 50 文ずつ) を無作為に取り出し，妥当な評価文が集められているかどうかを手で調査した．評価者には収集された文のみを提示して，それが好評であるか不評であるかを 5 段階で分類するように指示した．5 が最も好評で 1 が最も不評，そして好評か不評か判断に迷う場合を 3 とした．提案手法で好評文として収集された文は，評価者が 4 または 5 と判断していれば正解とした．不評文も同様である．評価者が 3 と判断した文は全て不正解とした．評価の結果を表 5 に示す．手法によって精度にややばらつきが見られるものの，おおよそ 70% から 80% の収集精度で評価文が収集できたことが分かる．評価者が 3 と判断した文を不正解としているので，これは決して低い精度ではないと考えている．さら

表 3: 収集された評価文の数

	好評文	不評文	合計
定型文 1	49,486	29,160	78,646
定型文 2	64,865	236,083	300,948
定型文 3	98,074	63,395	161,469
箇条書き	45,621	27,309	72,930
表	18,724	18,234	36,958
合計	276,770	374,181	650,951

表 4: 収集された評価文の例

	画像が綺麗になった 内部ノイズは P L L の割には少ないです .
好評文	みんなに , さりげなく自慢できる . 初心者でもポイントが分かり易い . 弾みすぎず , コントロールがよい .
	でも , ボディの塗装が褪せるのが早かった様な気が . 操作がわかりづらいのと面倒だ
不評文	ブレーカーがよく落ちる . とりまとめが煩雑 . ゲームバランスが最悪 .

に表 6 に , 評価者が 3 と判断した文を除いた場合の精度を示す . 表 5 の結果と比べて , 精度は大きく向上している . この結果から , 表 5 で不正解に数えられている事例は , 好評と不評を取り違えているような場合は少なく , ほとんどが人間でも判断に迷うような微妙な (= 評価者が 3 を与えていた) ケースであったことが分かる . 人間でも判断に迷う典型例は , 文脈情報が欠如している文であった . これについては次節で詳しく議論する .

表 5: 評価文の収集精度 1

定型文 1	0.72 (36/50)
定型文 2	0.70 (35/50)
定型文 3	0.76 (38/50)
箇条書き	0.80 (40/50)
表	0.82 (41/50)

表 6: 評価文の収集精度 2

定型文 1	0.92 (36/39)
定型文 2	0.95 (35/37)
定型文 3	0.88 (38/43)
箇条書き	0.95 (40/42)
表	0.91 (41/45)

5 議論

実験の結果から, 提案手法は非常に良い精度で評価文を獲得できることが分かった. これは実用的な精度であると考えている. 現在のところ, このコーパスを用いて評価表現辞書の構築を行っているが, 良好な結果を得ている. また, 評価文コーパスの規模の拡大にも並行して取り組んでいる. 具体的な方法としては, より多くの HTML 文書に適用する. また, 定型文から収集された評価文が多かったことから, 新しい語彙統語パターンを考案することも効果が大きいと考えている.

すでに述べたように, 十分な文脈情報が与えられていない評価文の評価極性は, 人間でも判断が困難であり, このことが精度に悪影響を与えていた. 例えば, コーパスには「何しろ情報量が多い」という文が好評文として登録されていたが, この文の評価極性は文脈に依存するため, この文単独で評価極性を決定することは困難である. したがって, この評価文を人手で評価した場合には, 不正解にカウントされる可能性が高い. しかし, 実際にこの文が抽出された元テキストを調べてみると, 次のようなものであった.

(2) このガイドブックのいいところは 何しろ情報量が多い ところです.

これを見ると「何しろ情報量が多い」という文は, 少なくとも原文では好評文として使われていたことが分かる. そのため, 評価極性が文脈に強く依存する文の中には, 今の評価法では不正解と判断されてしまっている, 完全な間違いとは言えないものが多く含まれていると考えている.

表 5 の結果を見ると, 定型文から収集された評価文は, それ以外のものよりも精度が悪い. この原因の一つは, 定型文からは短い評価文が収集されやすいため, 文脈が欠落しやすいからであると考えている. 実際, 上で示した「何しろ情報量が多い」という評価文も定型文から抽出されている. さらに, 表 6 の結果では, 定型文とそれ以外の手法に大きな差は見られない.

利用した記述形式によって, 収集される好評文と不評文の割合に大きな差が見られた. 特に定型文 2 の内訳を見ると, 不評文の数が圧倒的に多いことが分かる (表 3 の定型文 2). この理由を調べたところ「～なのが難点」という言い回しが頻出していることが原因であった. 本コーパスを利用するとき, このような好不評の偏りがどのような影響を与えるのかはまだ明らかではないが, 我々がこのコーパスを使っている範囲では大きな問題は起きていない.

本実験では HTML 文書から評価文の収集を行ったが、提案手法自体は HTML 文書に特化したものではないと考えている。たしかに、箇条書きと表形式を利用した抽出処理は、HTML 文書の特性を利用している。しかし、表 3 から分かるように、抽出された評価文のうち 80%以上が定型文から抽出されている。このことから、提案手法は HTML 文書以外のコーパスに対しても有効に働くと考えられる。

6 関連研究

我々の知る限り、評価文の自動収集を行ったという研究はこれまでに報告されていない。最も関連が深いのは、評価語や評価句の自動獲得に関する研究である。これには主に 2 つのアプローチがあり、1 つはシソーラスや国語辞典のような言語資源を利用して評価語を獲得する手法である。Kamps らは、類義語/反義語は同一/逆極性を持ちやすいという仮定にもとづき、WordNet を利用して評価語を獲得する手法を提案した (Kamps, Marx, Mokken and de Rijke 2004)。同様の考え方にもとづく手法はこれまでに多数提案されている (Hu and Liu 2004; Kim and Hovy 2004; Takamura, Inui and Okumura 2005)。一方で、Esuli らは語の評価極性の判定を定義文の分類問題として解いている (Esuli and Sebastiani 2005, 2006)。

評価語や評価句を獲得するためのもう 1 つのアプローチは、評価表現同士の共起関係を利用する方法である。Turney は、評価表現は同一ウィンドウ内に共起しやすいとことに着目し、語句の評価極性を判定する手法を提案した (Turney 2002; Turney and Littman 2002)。共起頻度を求めるときに既存のコーパスを利用するのではなく、検索エンジンを利用してウェブという大規模コーパスでの頻度を見積もり、データスパースネスの問題に対処している点が特徴である。Hatzivassiloglou らは、コーパス中で 2 つの形容詞が *and* などの順接を表す接続詞で結ばれていれば同一評価極性を持ちやすく、逆に *but* のような逆接で結ばれていれば逆極性を持ちやすいという観察にもとづき、コーパスから評価語を獲得する手法を提案した (Hatzivassiloglou and McKeown 1997)。この考え方は Kanayama らによってさらに拡張されている (Kanayama and Nasukawa 2006)。

語句の評価極性ではなく主観性に着目した研究報告も存在する。Wiebe は、人手でタグ付けされた訓練事例と単語の類似度を利用することによって、主観的形容詞を学習する手法を提案している (Wiebe 2000)。Riloff らは主観的名詞の獲得を行っている (Riloff, Wiebe and Wilson 2003)。Riloff らの手法では、主観的名詞とその抽出パターンが交互に学習される。まずシステムには少数の主観的名詞が入力として与えられる。そして、それを利用して主観的名詞の抽出パターンを学習する、学習されたパターンで新たな主観的名詞を獲得する、という処理が繰り返される。これにより大量の主観的名詞の獲得が行われる。(Riloff et al. 2003) では名詞が対象とされていたが、(Riloff and Wiebe 2003) では名詞以外の句も獲得対象となっている。同様のブートストラップ的な手法は (Wiebe and Riloff 2005) でも議論されている。

提案手法のように語彙統語パターンやレイアウト (箇条書きや表) パターンを用いて知識獲

得を行う手法は古くから研究されてきている。Hearst は, *such as* のようなパターンに着目して, 単語間の上位下位関係をコーパスから獲得する手法を提案した (Hearst 1992)。Hearst の手法は英語を対象としているが, 日本語においても安藤らが同様の手法を試している (安藤まや, 関根聡, 石崎俊 2003)。一方, 新里らは, 同一箇条書きに出現する単語は共通の上位語を持ちやすいという仮説にもとづき, 上位下位関係を獲得する手法を提案している (新里圭司, 鳥澤健太郎 2005)。これ以外にも, 全体部分関係にある単語対の獲得や, 属性と属性値の獲得といったタスクにも, 同様の手法が用いられている (Berland and Charniak 1999; Chklovski and Pantel 2004; 吉永直樹, 鳥澤健太郎 2006)。我々の知る限り, 評価情報処理に同様の手法を適用したという報告はない。Hu らや Kim らの手法ではレイアウトパターンが利用されているが, これらの研究では, レイアウトは特定サイトに固有の手がかりとして議論されているため意味合いが異なる (Liu, Hu and Cheng 2005; Kim and Hovy 2006)。

7 おわりに

本論文では大規模な HTML 文書集合から評価文を自動収集する手法を提案した。提案手法を約 10 億件の HTML 文書に適用したところ, 約 65 万の評価文を獲得することができた。収集精度はおおよそ 70% から 80% であったが, 文脈依存する評価文の存在を考慮すると, 良好な結果であると考えている。今後は, このコーパスからの評価表現の収集に取り組む予定である。

参考文献

- Berland, M. and Charniak, E. (1999). "Finding Parts in Very Large Corpora." In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 57–64.
- Chklovski, T. and Pantel, P. (2004). "VERBOCEAN: Mining the Web for Fine-Grained Semantic Verb Relations." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Esuli, A. and Sebastiani, F. (2005). "Determining the Semantic Orientation of Terms through Gloss Classification." In *Proceedings of Conference on Information and Knowledge Management (CIKM)*, pp. 617–624.
- Esuli, A. and Sebastiani, F. (2006). "Determining Term Subjectivity and Term Orientation for Opinion Mining." In *Proceedings of European Chapter of the Association of Computational Linguistics (EACL)*, pp. 193–200.
- Hatzivassiloglou, V. and McKeown, K. R. (1997). "Predicting the Semantic Orientation of Adjectives." In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 174–181.

- Hearst, M. (1992). “Automatic Acquisition of Hyponyms from Large Text Corpora.” In *Proceedings of International Conference on Computational Linguistics (COLING)*, pp. 539–545.
- Hu, M. and Liu, B. (2004). “Mining and Summarizing Customer Reviews.” In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 168–177.
- Kamps, J., Marx, M., Mokken, R. J., and de Rijke, M. (2004). “Using WordNet to Measure Semantic Orientations of Adjectives.” In *In proceedings of the International Conference on Language Resources and Evaluation (LREC)*.
- Kanayama, H. and Nasukawa, T. (2006). “Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis.” In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 355–363.
- Kim, S.-M. and Hovy, E. (2004). “Determining the Sentiment of Opinions.” In *Proceedings of International Conference on Computational Linguistics (COLING)*, pp. 1367–1373.
- Kim, S.-M. and Hovy, E. (2006). “Automatic Identification of Pro and Con Reasons in Online Reviews.” In *Proceedings of International Conference on Computational Linguistics, Poster Sessions (COLING)*, pp. 483–490.
- Kudo, T. and Matsumoto, Y. (2004). “A Boosting Algorithm for Classification of Semi-Structured Text.” In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Liu, B., Hu, M., and Cheng, J. (2005). “Opinion Observer: Analyzing and Comparing Opinions on the Web.” In *Proceedings of WWW*.
- Matsumoto, S., Takamura, H., and Okumura, M. (2005). “Sentiment Classification Using Word Sub-sequences and Dependency Sub-tree.” In *Proceedings of Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp. 301–311.
- McDonald, R., Hannan, K., Neylon, T., Wells, M., and Reynar, J. (2007). “Structured Models for Fine-toCoarse Sentiment Analysis.” In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Pang, B., Lee, L., and Vaihyanathan, S. (2002). “Thumbs up? Sentiment Classification using Machine Learning Techniques.” In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Riloff, E. and Wiebe, J. (2003). “Learning Extraction Patterns for Subjective Expressions.” In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Riloff, E., Wiebe, J., and Wilson, T. (2003). “Learning Subjective Nouns using Extraction Pattern Bootstrapping.” In *Proceedings of the Seventh Conference on Natural Language*

Learning (CoNLL).

- Takamura, H., Inui, T., and Okumura, M. (2005). "Extracting Semantic Orientation of Words using Spin Model." In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 133-140.
- Turney, P. D. (2002). "Thumbs up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 417-424.
- Turney, P. D. and Littman, M. L. (2002). "Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus." Tech. rep., National Research Council Canada.
- Wiebe, J. and Riloff, E. (2005). "Creating Subjective and Objective Sentence Classifiers from Unannotated Texts." In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)*.
- Wiebe, J. M. (2000). "Learning Subjective Adjectives from Corpora." In *Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI)*.
- 安藤まや, 関根聡, 石崎俊 (2003). "定型表現を利用した新聞記事からの下位概念単語の自動抽出." 情報処理学会研究報告 2003-NL-157, pp. 77-82.
- 乾孝司, 奥村学 (2006). "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).
- 吉永直樹, 鳥澤健太郎 (2006). "Web からの属性情報記述ページの発見." 人工知能学会論文誌, 21 (6).
- 小林のぞみ, 乾健太郎, 松本裕二, 立石健二, 福島俊一 (2005). "意見抽出のための評価表現の収集." 自然言語処理, 12 (3).
- 新里圭司, 鳥澤健太郎 (2005). "HTML 文書からの単語間の上位下位関係の自動獲得." 自然言語処理, 12 (1).
- 藤村滋, 豊田正史, 喜連川優 (2005). "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.
- 箴島郁子, 嶋田和考, 遠藤勉 (2005). "系列パターンを利用した評価表現の分類." 言語処理学会第 11 回年次大会発表論文集, pp. 448-451.

略歴

鍛治 伸裕: 2005 年東京大学大学院情報理工学系研究科博士課程修了. 情報理工学博士. 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.

喜連川 優: 1978 東京大学工学部電子工学科卒業. 1983 同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, 日本データベ

ス学会理事，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).

(2003 年 7 月 3 日 受付)
(2003 年 9 月 25 日 再受付)
(2003 年 10 月 23 日 採録)