自然言語処理を用いたWeb情報の評価

リスク工学グループ演習 9 班 (2018 年度)
芦田佳樹　池田遼　村山喬則　渡邊竣
アドバイザー教員　高安亮紀　遠藤靖典

1　はじめに

Web 上に多数存在する通信販売サイト, ブログ, 口コミサイトや SNS などといったコミュニティでは, 個人ユーザから寄せられた大量の文章が蓄積されている. それらの中には商品やサービスなどの対象に対する評価や, 組織や個人に対する批評などが多く含まれている [1]. 消費者が発信する情報は, 消費者目線での商品の評価が含まれているため, 企業等にとってはマーケティング分析の対象となることも多い [2]. こうしたネットワーク上でアクセス可能なテキストに対する自然言語処理技術の需要が高まってきており, その一つとして評判分析の技術 [3] などがある. Web における消費者が記述した各種製品, 映画, 旅行先, 音楽などに関する情報は, 消費者にとって対象の購入や行き先として検討する際に貴重な判断材料となり得る. しかし実際に Web 上にある情報には, 書き込みを行なった消費者の先入観や価値観によって評価基準がばらつくという問題や, 点数付けなどの評価システムを採用しておらず対象の評価が一見して判別できないものが多数存在する. これらの問題に対する現状の閲覧者側の対策はひとつひとつの評価情報に目を通し, 意見に対する考察を行うことであるが, それには多くの時間と労力が必要となる. 以上のことから Web 情報にある自然言語によって記述された評価を定量化し, 五つ星評価やレーダーチャートなどの一見して評価の良し悪しを理解できる客観的評価指数を与えることを本研究の目的とした.

2　研究手法

2.1　自然言語処理

本研究の対象は, 人間が日常的に使っている自然言語であり, Web 上にある自然言語を取得し, コンピュータに処理をさせる必要がある. この際用いるのが自然言語処理の技術である. 本研究の対象は, 日本語によって記載されている情報としているため, 日本語を処理する自然言語処理の技術を用いる. 現在日本語を処理する自然言語の基礎技術として, 形態素解析, 構文解析, 語義の曖昧性解消, 照応解析といった手法があるが, 本研究では形態素解析, 構文解析の手法を用いる.

2.1.1　形態素解析

形態素解析は自然言語のテキストデータから文法や, 辞書と呼ばれる単語の品詞等の情報データに基づき, 形態素と呼ばれる言語で意味を持つ最小単位の列に分割することで, それぞれの形態素の品詞等を判別する手法である. 日本語は語の区切りに空白を挟む記述法である「わかち書き」がされない言語であるため, 単語の区切りを特定することが必要であり, 非常に複雑な作業である. 現在ではフリーライセンスで使用可能な日本語の形態素解析エンジンが複数あり, 本研究では Janome(Python) や, 動作環境によっては MeCab といったエンジンを使用して解析を行う. 図 1 は Janome を用いた形態素解析の例であり, 「リスク工学グループ演

習 9 班が発表します」というテキストデータを形態素に分解している．

```
>>> for token in t.tokenize(u'リスク工学グループ演習9班が発表します'):
...  print(token)
...
リスク  名詞,一般,*,*,*,*,リスク,リスク,リスク
工学    名詞,一般,*,*,*,*,工学,コウガク,コーガク
グループ        名詞,一般,*,*,*,*,グループ,グループ,グループ
演習    名詞,サ変接続,*,*,*,*,演習,エンシュウ,エンシュー
9       名詞,数,*,*,*,*,9,*,*
班      名詞,接尾,助数詞,*,*,*,班,ハン,ハン
が      助詞,格助詞,一般,*,*,*,が,ガ,ガ
発表    名詞,サ変接続,*,*,*,*,発表,ハッピョウ,ハッピョー
し      動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
ます    助動詞,*,*,*,特殊・マス,基本形,ます,マス,マス
```

図 1: Janmoe による形態素解析の例

図 1 の通り, テキストデータが品詞ごとに分解されていることがわかる．形態素解析エンジンを利用することで, Web 上にあるテキストデータを取得できればデータ内の自然言語の分解が可能となる．

2.1.2 構文解析

構文解析は自然言語を前述の形態素に切り分け, その間にある修飾-被修飾関係などの関連を解析する手法である．テキストデータを構文解析を行う機構である構文解析器を用いて解析すると, 構文木という形態素の関係性を図式化した形で得られる．構文解析器としては CaboCha が利用可能であり, MeCab や Janome といった形態素解析エンジンと連動して構文解析が行われる．本研究における構文解析では CaboCha を利用してテキストデータの形態素間の係り受け関係を取得する．CaboCha を用いた構文解析の例が図 2 である．例では構文木状に表記されているが, 形態素ごとの係り受け関係を数値データとして抽出することも可能であり, 後述の意見抽出に用いることができる．

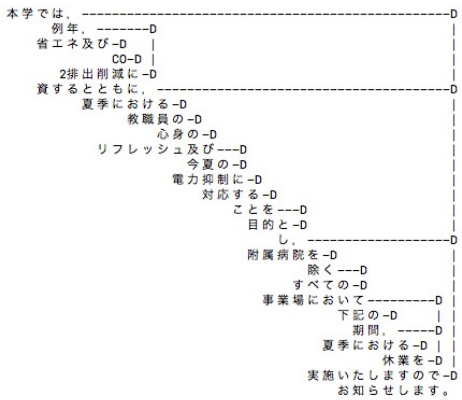

図 2: CaboCha を用いた構文解析結果の構文木

2.2 極性辞書

2.1.1 節にて記述した形態素解析に関連して極性辞書と呼ばれる形態素ごとにポジティブ, ネガティブの 2 極性値が対応しているデータベースを利用する．極性辞書には様々な種類があり, 辞書によって特化した対象などの性格付けがなされている．本研究で用いたものは PN Table[4] と呼ばれる東工大高村研究室によって作成されたものであり, 形態素に-1〜1 の極性値 (PN 値) が割り振られており, 言語学的知見に基づき, 人手により評価の確かさが保証されているものである．図 3 は PN Table の一部であり, 図のように形態素に関する点数付けがなされている [5].

優れる:すぐれる:動詞:1
良い:よい:形容詞:0.999995
喜ぶ:よろこぶ:動詞:0.999979
褒める:ほめる:動詞:0.999979
めでたい:めでたい:形容詞:0.999645
賢い:かしこい:形容詞:0.999486
善い:いい:形容詞:0.999314
適す:てきす:動詞:0.999295
天晴:あっぱれ:名詞:0.999267
祝う:いわう:動詞:0.999122
功績:こうせき:名詞:0.999104
賞:しょう:名詞:0.998943
嬉しい:うれしい:形容詞:0.998871
喜び:よろこび:名詞:0.998861
才知:さいち:名詞:0.998771

図 3: PN Table の一部（[4] から抜粋）

2.3 評価指数の付与

本研究では二つの手法で Web 情報中にある意見の定量化を試みる．一つは形態素解析の後, PN Table を参照することで極性値を与え, 意見全体の平均の極性値を評価点数とする手法 (2.3.1 節参照). もう一つは構文解析を行い, 意見文中に記述されている対象の評価属性とその良し悪しである評価を抽出し, 評価属性ごとに定量化, レーダーチャートとして要約する手法である (2.3.2 節参照). この手法は立石らが Web 上の意見集合からの意見抽出と要約生成を行った際に用いた手法と同様の手法 [6] であり, 本研究で用いる際に一部の工程を変更したものである.

2.3.1 PN 値を用いたテキストデータの定量化

この手法は比較的簡単に評価指数を与えることが可能なものであり, 対象に関する多種多様で大量にある意見の集約として, それぞれの意見に含まれる形態素の極性値を平均することで一つの点数をつける．まずは対象となる意見の集合をテキストデータとして収集する．このテキストデータは形態素解析エンジンを利用することで形態素に分けることが出来る．次にテキストデータ内に含まれて高い頻度で現れるが, 固有名詞などで点数の付与に不適な語 (商品名や記事の題名, URL など) を取り除く．残ったテキストデータは形態素に分解されているため PN Table の参照が可能である．PN Table はデータベースであるため, PN Table に含まれていない語については極性値を 0 とし, 全体の評価指数を与える際の計算には含まない. 以上の過程を経て, 意見の記述者のテキストデータごとに極性値を平均することで, 対象に評価指数を与えることが出来る．対象全体についての評価点数を求めたい場合には, 意見全体の極性値を平均化することで算出が可能である.

2.3.2 意見抽出と着眼点に基づく要約生成

この手法は対象についての意見情報から, レーダーチャートの形式で複数の評価軸 (属性) を用意した評価指数の付与を行うものである．多種多様な意見の集合から評価軸となる属性要素を自動的に選択することは難しく, システムの構成に多大な時間を要することから, 本研究では評価軸の選択は人手により行う方法をとる．まずは評価指数を与える対象を決め,　対象についての意見が記述されている意見の集合 (通販サイトのレビュー等) を二つ以上用意する．　一つの意見集合を用い, 構文解析を行うことで表 1 のような意見抽出ルールの文関係となっている箇所を抽出し, レーダーチャートの評価軸となり得る属性表現をリストアップする．この属性表現の類似性と出現頻度を考慮し, いくつかの評価軸を設定することで一つの評価軸に対する複数の属性表現という対応付けが作成できる．これを属性表現辞書とすることで, 他の意見集合にも適用することが可能なデータベースとなる．また, 良し悪しの記述がなされている評価表現についても良い, 悪いに対応する表現を集めることで評価表現辞書となるデータベースを作製する．ここで作成した属性表現辞書と評価表現辞書を用いて, 他の意見集合について同様の意見抽出ルールで評価されている箇所を抽出, 評価軸ごとの評価がリストアップされるため, 評価軸ごとの良い, 悪いの比率を求めることが出来る．これらをレーダーチャート状で表記することで多種多様な意見の要約を行う.

表 1: 意見抽出ルール

抽出ルール	例
(属性:が/は/も/の/に/を/で)→(評価)	デザインが良い, 外観も好き
(評価:＜連体修飾＞)→(属性)	良いデザイン
(属性)＝(評価)	デザイングッド
(属性:の)→(∗:が/は/も)→(評価)	デザインの質が良い
(属性:も/や/と/、)→(∗:も/は/が/で)→(評価)	デザインも広告も良い

3　結果と考察

3.1　PN 値を用いた結果

評価指数を与える対象として，「バッドデザイン賞を勝手にノミネートしてみた-2017 年度版-」[7] というブログ記事を選択した．記事では，執筆者が日常で見かけたり，SNS で発見した案内板やボタンの扱いづらい意匠のものが紹介されている．この記事にはコメント機能が搭載されているが点数付け等の機能はない．また，SNS でブログ記事を引用しての意見投稿数が多く，記事についての評価指数を与える手法の対象としては適している．SNS に引用して投稿された意見のテキストデータを収集し，2.3.1 の手順で評価指数付けを行った．その結果の意見を点数ごとにまとめたものが図 4 である．

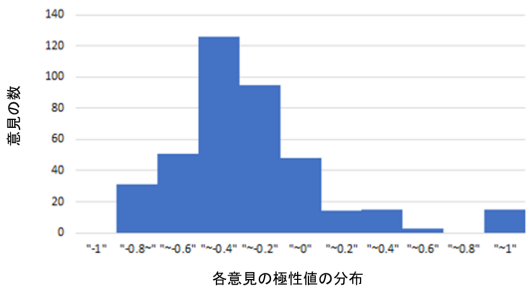

図 4: 対象ブログ記事についての意見の極性値分布

この結果では全体の 9 割近くがネガティブな意見として判別されている．これは対象となるブログ記事がバッドデザイン賞についての記事であり，テーマが商品を批判するような内容であることが理由の一つとして考えられる．また，今回用いた極性辞書である PN Table は辞書に内包されている語の極性値は全体的にネガティブに偏っていることがあり，PN Tbale を用いた極性値付与はネガティブ寄りの結果となりやすいことが確認されている [8][9].

3.2　評価属性ごとに定量化した結果

次に 2.3.2 節で紹介した手法で Web 上の口コミに対して評価付けを行う．分析対象は Web 上で公開されている通販サイトのレビュー情報とした．特に今回は評価軸が複数存在すると考えられる「イヤホン」に関するレビューを対象とし，具体的な商品として「パナソニック カナル型イヤホン ホワイト RP-HJE150-W」のレビューを利用した．この商品に対して，2018 年 10 月 13 日までに 3623 件のレビューが投稿されており，5 段階評価で 3.9 の評価が付けられている．この評価は各ユーザーがレビューを投稿する際に，商品を 5 段階評価したものの平均である．分析手順としては，最初に 2.3.2 節で挙げた抽出ルールに基づき，商品に対する属性表現と評価表現の候補を抽出する．次に属性表現と評価表現のカテゴリ分けを人手で行う．その際，候補は出現頻度が 10 回以上のものを対象にする．今回はイヤホンの属性表現を比較的言及が多かった「総合」「音質」「デザイン」「付け心地」「コストパフォーマンス」の 5 つに分類した．評価表現は「positive」「negative」の 2 種類に分類した．属性表現と評価表現の分類結果を以下の表 2 と表 3 に示す．

表 2: 属性表現の辞書

属性カテゴリ	属性
総合	'イヤホン', '商品', '製品', '買い物', '質', 'クオリティ', '性能'
音質	'音', '音質', '高音', '重低音', '中音', '音響', '音色', '音域', 'サウンド', '音色'
デザイン	'デザイン', 'デザイン性', '色', '見た目', '形', '形状', 'カラー', 'レッド', 'ブルー', 'ピンク色'
付け心地	'フィット感', 'フィット', 'フィット性', '装着感', '装着', '心地', 'つけ心地'
コスパ	'コストパフォーマンス', 'コスパ'

表 3: 評価表現の辞書

評価カテゴリ	評価表現
positive	'良い', 'いい', 'よい', '高い', '素晴らしい', 'すばらしい', '最高', '好き', '十分', '綺麗', '凄い', '豊富', '抜群', '可愛い', 'シンプル', 'かわいい', '良好', '嬉しい', '感動'
negative	'悪い', '微妙', '悪意', '安っぽい', '弱い', '少ない', '低い', '残念', '最低', '酷い', '荒い', '不安', 'ダメ', '貧弱', 'おかしい', 'イヤ', '最悪'

以上の結果を属性表現辞書, 評価表現辞書とする. これらの作成した辞書と 2.3.2 節の抽出ルールに基づいて評価を行う. 評価方法は positive を 1, negative を-1, と評価し, 全体の平均を商品の評価とした. レビューの具体的な評価例と商品のカテゴリ別の評価を以下の表 4 と図 5 に示す。

表 4: レビューの評価例

レビュー	総合	音質	デザイン	付け心地	コスパ
安い割に乗りのいい音で良い. 装着感も良好で, デザインも安っぽくない.	0	1	1	1	0
デザインはいいが, 安っぽい音質です.	0	-1	1	0	0

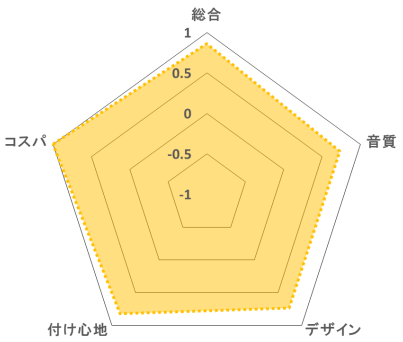

図 5: 商品のカテゴリ別評価

最後に全 3623 件のレビュー中の何割が評価に用いられたかを示すために, 評価に用いられたレビューの数を次の表 5 に示す.

表 5: 評価に用いられたレビューの数

	総合	音質	デザイン	付け心地	コスパ
評価されたレビューの数	360	817	90	103	54
評価されたレビューの割合	0.10	0.23	0.02	0.03	0.01

以上の結果から, ルールベースで辞書を作成し評価を行うことでカテゴリ別の評価が可能なことが確認できた. しかし, 現状では 3600 件あるレビューの 35%ほどしか評価に用いられていない. この問題は辞書単語の追加やルールの追加によってある程度の改善が見込めるが, 基本的に手作業による辞書作成であるため限界がある. また, 複雑な文法や口語的なレビューなどは, ルールベースでの抽出は困難である.

4 まとめ

本稿では, Web 情報に溢れる多種多様な意見の集合から, 対象への評価の定量化を 2 つの手法で試みた. PN 値を用いたテキストデータの定量化は既存の極性辞書を用いた定量化の手法であり, 比較的簡単に解析が可能であるため, 幅広い対象の解析に対応が可能であるというメリットが考えられる. 一方で極性辞書への依存性が高く, 辞書に含まれていない語の多さや, 辞書に含まれる語の極性値の偏りに大きく結果が左右される問題がある. 現在における自然言語の分析手法では機械学習による感情分析が主流であり, 一般向けの簡易で高性能な分析ツールも公開されている [10].

また, 意見抽出と着眼点に基づく要約生成では構文解析をもちいた意見の抽出を行い, 評価項目と評価の割合を算出することでレーダーチャート状に定量化した評価指数を与えた. 一定の工程は人手による判別が必要ではあるが, 多種多様で多量な意見から対象の評価を定量化し, 要約するという目標には達している. この手法の改善点としては人手による工程の削減が挙げられ, 構想としては属性表現の単語の類似性を分析し, その結果をもとに評価軸を選択する方法があると考えられる.

最後に本稿では自然言語処理を用いたWeb情報の評価を行なったが, 日本語という言語の特徴でもある複雑さや, 独自性もあり, 単一の手法だけでWeb情報の評価を行うことは難しいと感じた. 本研究をより発展させていくには, より多くの手法を調査し, 組み合わせて解析を行なっていく必要があると考えられる.

参考文献

[1] 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. 2005, vol. 12, no. 3, p. 203-222.

[2] 田中成典, 北野光一, 寺口敏生, 今城彰子, 柳田尚明. 広告の特徴に基づく口コミの分類に関する研究. 情報処理学会論文誌データベース (TOD). 2011, vol. 4, no. 3, p. 22-32.

[3] 那須川哲哉, 金山博, 日本IBM(株) 東京基礎研究所. 文脈一貫性を利用した極性付評価表現の語彙獲得. 情報処理学会研究報告自然言語処理 (NL). 2004, no. 73, p. 109-116.

[4] 東京工業大学 高村大也.“単語感情極性対応表”, 高村大也, http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html, (参照 2018-06-04).

[5] 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. 情報処理学会論文誌. 2006, vol. 47, no. 2, p. 627-637.

[6] 立石健二, 福島俊一, 小林のぞみ, 上出将行, 高橋哲朗, 乾孝司, 藤田篤, 乾健太郎, 松本裕治. Web文章集合からの意見情報抽出と着眼点に基づく要約生成. 情報処理学会研究報告情報学基礎 (FI). 2004, no. 93, p. 1-8.

[7] おり. “バッドデザイン賞を勝手にノミネートしてみた-2017 年度版-”, note. https://note.mu/ori_io/n/n721e9694788b,(参照 2018-06-04).

[8] midnightseminar(id:midnightseminar). “【Python】MeCab と極性辞書を使ったツイートの感情分析入門”, Stats-Beginner: 初学者の統計学習ノート. https://www.statsbeginner.net/entry/2017/05/07/091435, (参照 2018-06-10).

[9] Aidemy Inc.(id:aidemy-blog). “花火大会における Twitter 民の感情分析”,Aidemy Blog. https://blog.aidemy.net/entry/2017/08/10/170715, (参照 2018-06-10).

[10] ヤフー株式会社. “Yahoo リアルタイム検索”, Yahoo!JAPAN. https://search.yahoo.co.jp/realtime, (参照 2018-10-04).