

製品レビュー文に基づく評判情報コーパスの作成とその特徴の分析

宮崎林太郎

横浜国立大学 大学院
環境情報学府

E-mail:{ rintaro, mori}@forest.eis.ynu.ac.jp

森辰則

横浜国立大学 大学院
環境情報研究院

本稿では、1 万文の製品レビュー文に対して、人手による注釈付けを行い作成した評判情報コーパスについて、その特徴を分析した。

我々は、評判情報抽出タスクにおいて、評判情報コーパスが必要不可欠と考え、作成を行ってきた。作成されたコーパス中において、評判情報がどのように出現するかを調べることで、今後の評判情報抽出への取り組み方を検討する。

本稿では、製品の様態とそれに対する評価を分けるために項目(item)、属性(attribute)、属性値(value)、評価(evaluation)の 4 つ組から成る評判情報モデルを用いている。コーパス中における各構成要素の表層表現の出現頻度、省略されている要素について統計的な調査を行った。また、evaluation が出現する場合と出現しない場合の比較、同一の attribute-value に対して異なる極性の evaluation が出現する場合を調べることで、4 つ組で評判情報をモデル化する際の効果を検証することが出来た。

Creation of a Corpus for Sentiment Analysis based on Product Reviews and Analysis of its Features

Rintaro MIYAZAKI

Tatsunori MORI

Graduate School of Environment and Information Science,
Yokohama National University

In this paper, we analyzed features of a corpus for sentiment analysis that was made by manually annotating product reviews, which consist of 10000 sentences.

We have been making the corpus because it is one of essential resources for automated extraction of sentiment information. By analyzing the corpus from the viewpoint of how sentiment information appears, we will examine the direction of sentiment information extraction in our future work.

We have proposed a model of sentiment analysis, in which one unit of sentiment information consists of four elements, namely, item, attribute, value and evaluation. In this model, descriptions of attributes of a product are explicitly separated from the evaluations that reviewers made. We conducted statistical investigation about each specified element and each omitted element in the corpus. We also compared the following cases with each other: the case that evaluation elements appeared and the case that they did not, and the cases in which the same attribute-value pairs appear but the evaluation elements have different polarities. According the comparison, we confirmed the effectiveness of our model.

1. はじめに

近年、CGM(Consumer Generated Media)と呼ばれるサービスが広く利用されるようになってきている。CGMには利用者が製品やサービス等に対するレビューを多数公開している。また、

製品の情報やレビューを集約するAmazon¹や価格.com²等のサイトも日常的に利用されるようになってきている。このような、ネット上に存在するレビューを評判情報として収集・分析することで、製品の同一様態に対するユーザごとの評価の違

¹ <http://www.amazon.co.jp>

² <http://www.kakaku.com>

いや、同じ評価に関する様々な理由などを分析することが期待される。

我々は、評判情報抽出タスクに関する研究を推進するにあたって必要不可欠であると考えられる評判情報コーパスの作成を行ってきた。

本稿では、人手による注釈付けで作成された評判情報コーパスの特徴を調べることで、レビューにおいて評判情報が出現する傾向を把握すると同時に、我々が提案する4項組による評判情報のモデル化の効果を確認する。

2. 関連研究

2. 1 評判情報のモデル化と抽出

評判情報抽出タスクの中には、評判情報をいくつかの構成要素からなるものとする研究が存在する。立石ら[1]は評判情報を「対象物、評価軸、満足度」からなるものとし、その後「対象、属性、属性値」の3つ組からなるものとする研究が行われている[2]。村野ら[3]は評判情報の構成要素をさらに細かく分け、「対象、比較対象、評価、項目、様態」と分類している。また、Kobayashiら[8]は評判を「Opinion holder, Subject, Part, Attribute, Evaluation, Condition, Support」からなるものとし、意見主や条件などにも言及している。

上記の研究では各構成要素の抽出も行っているが、評価表現辞書などのあらかじめ用意された知識を用いているものが多い。

2. 2 評判情報に関連するコーパス

評判情報に関連するコーパスとして、意見コーパスがあげられる。意見コーパスにはMPQAコーパス[4]やNTCIR-6における意見コーパス[9]がある。これらは、新聞記事に対して主に事態に対する意見について注釈付けされている。また、日本語における評判情報コーパスとしては、小林ら[5]はWeblogの記事に対して評判情報を構成要素ごとに注釈付けを行い、意見タグ付きコーパスの試作を行っている。また、kajiら[7]はHTML文書から評価文コーパスの自動構築を行っている。

上記の研究に対して、本研究の貢献は1)製品の様態とそれに対する評価を分けたモデルについて、2)Webから収集した文書に対して、注釈事例を用いた人手による注釈付けによりコーパスを作成し、3)コーパス中の構成要素について統計的な調査を行うことでレビュー文中における評判の出現の仕方を分析する点にある。

3. 評判情報コーパスの作成

3. 1 用いたモデル

本研究で用いた評判情報のモデルはitem, attribute, evaluation, valueの4つ組モデルである。これは、同一の製品の様態に対して、複数のレビュアーが異なる評価を持つ場合に対応したものである。各タグの説明とタグセットを図1に示す。

```
<item> 製品を構成する要素を意味する概念クラス
        やそのインスタンスを指示する表層表現
class: この item が属する概念番号
<attribute> item の様態を表す観点
pair: attribute-value の組番号
target: attribute-value を持つ概念番号
<value> attribute に対する様態の内容
pair: attribute-value の組番号
target: attribute-value を持つ概念番号
<evaluation> item に対する主観的見解
target: evaluation の対象になる概念番号
reason: 評価理由となる attribute-value の組番号
orientation: 評価が肯定・中立・否定のどれか
```

図 1. 注釈付けに用いたタグセット

3. 2 コーパスの作成

評判情報の作成に用いた文書はAmazonから収集したレビュー文1万文である。Amazonの製品分類を元に、電化製品(ジャンル1)、ホーム・キッチン(ジャンル2)、ホビー・おもちゃ(ジャンル3)、映像・音楽(ジャンル4)、ソフトウェア(ジャンル5)、の5ジャンルについて、各2000文ずつを収集した。

注釈者は情報工学を専攻する大学院生、学部生計10名であり、事前説明として、モデルの説明、注釈揺れの事例、モダリティを表現する表層表現等を各要素を注釈付ける範囲に含めるか否かという点について説明をおこなった。なお、今回の注釈付けにおいては、注釈者に対し「valueかevaluationを見つけてもらい、見つかった場合には、それと組となる要素を探す」方法で注釈付けを行ってもらった。

また、注釈付けの際には注釈事例を参照しながら注釈付けを行えるツールを試作し注釈付けに用いた[6]。

こうして得られた注釈付きコーパスについて、注釈情報の記入漏れや、注釈付けツールの操作ミスと思われる注釈ミスの修正作業を行った後のコーパスを今回の分析に用いた。

今回のコーパスに注釈付けされた評判情報の構成要素の数を表1に示す。

また、注釈付けの一致率を確認するために、コーパス作成に用いた文書の一部(5ジャンルから35文ずつ、185文)を、今回の注釈付けに参加した作業者の内の4名に改めて注釈付けしてもら

表 2. コーパス中の各構成要素の表層表現の出現頻度上位 20 個

attribute		value		evaluation	
表層表現	出現回数	表層表現	出現回数	表層表現	出現回数
音	124	有る	192	良い	190
デザイン	103	出来る	99	いい	121
値段	97	高い	78	勧める	82
機能	64	無い	75	善い	64
価格	58	便利だ	57	気に入る	63
音質	47	する	51	素晴らしい	55
内容	42	可愛い	50	残念だ	48
色	42	面白い	50	嬉しい	34
大きさ	37	大きい	48	満足	33
サイズ	31	使える	47	満足する	31
使い勝手	29	安い	47	最高だ	28
操作	27	楽しめる	47	驚く	27
見た目	24	就く	43	オススメ	24
パフォーマンス	22	出る	38	好きだ	24
出来	22	重い	38	悪い	24
雰囲気	22	問題	37	満足だ	24
作り	21	多い	36	勧めする	21
動作	21	綺麗だ	36	魅力	21
性能	19	軽い	36	ビックリ	19
声	18	十分だ	35	星 4 つ	19

い、注釈付けの一致率を確認した。なお、ここで使用した文書は 4 名の注釈者がコーパス作成の段階では見えていない文書である。一致率の平均は κ 値で 0.5 程度であった。

表 1. コーパス中の各構成要素の数

	item	attribute	value	evaluation
ジャンル 1	1375	1075	1747	513
ジャンル 2	1078	874	1838	363
ジャンル 3	1252	780	1687	436
ジャンル 4	1188	614	1027	367
ジャンル 5	1245	818	1586	302
計	6138	4161	7885	1981

4. 作成された評判情報コーパスの特徴

本節では、作成したコーパスの特徴について、分析を行う。モデル中の各構成要素について、主に表層表現の出現頻度を調べることで、レビューにおける評判情報の記述のされかたを分析する。

4. 1 各構成要素の出現頻度

最初に、attribute, value, evaluation について、表層表現の出現頻度を調べた。表記揺れなどを考慮した頻度を調べるために、品詞については Chasen⁵による形態素解析結果の品詞が、名詞、形容詞のものは全て、動詞については「動詞-自

立」を採用し、表層表現についてはさらなる表記揺れに対応するために Juman⁶の形態素解析結果の代表表記を用いた。

コーパス中に存在した attribute, value, evaluation が注釈付けられている表層表現の頻度上位 20 個を表 2 に示す。

attribute については、製品ジャンルを横断してレビューが興味を持つと推測される属性が上位にきている。

value については、「ある」「ない」とそれに類するものが頻出した。製品の機能の有無に関する言及をする際に汎用的に用いられることが頻出した原因として考えられる。

evaluation については「いい」「良い」に関連するものが頻出している。形態素解析結果の間違いなどにより、表記揺れを併合しきれていない部分もあるが、肯定的な記述が非常に多いことがわかる。

なお、今回の分析を行うに際して item の表層表現についても、出現頻度を調べた。しかし、item に関しては、コーパス作成後に修正作業を行って、item と思われる部分には value, evaluation の出現に関係なく網羅的に注釈付けを行った。このため、レビューが特定の商品名を、連呼するだ

⁵ <http://chasen-legacy.sourceforge.jp/>

⁶ <http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html>

表 3. 評判情報構成要素の省略パターンとコーパス中の数

省略されている要素のパターン					全文	ジャンル 1	ジャンル 2	ジャンル 3	ジャンル 4	ジャンル 5
①	item	attribute	value	evaluation	653 組	238	157	117	51	90
②	item	attribute	value	φ	3025 組	650	643	560	487	685
③	item	attribute	φ	evaluation	656 組	245	114	138	82	77
④	item	φ	value	evaluation	648 組	186	190	144	43	85
⑤	item	φ	value	φ	3895 組	768	900	945	503	779
⑥	item	φ	φ	evaluation	649 組	99	64	147	235	104
⑦	計				9526 組	2186	2068	2051	1401	1820

表 4. ①における表層表現の出現頻度上位

attribute		value		evaluation	
デザイン	31	出来る	25	気に入る	42
値段	26	有る	17	勧める	42
音	25	簡単だ	12	良い	39
機能	17	高い	12	いい	36
スキャン	11	シンプルだ	11	満足する	22
音質	9	会う	11	善い	18
価格	8	無い	11	残念だ	17
色	8	安い	10	満足	17
操作	7	多い	8	満足だ	11
サイズ	6	少ない	8	好感もてる	10
作り	6	する	7	驚く	10
吸引力	5	対応	7	魅力	10
声	5	小さい	7	素晴らしい	9
性能	5	出る	6	星 4 つ	8
演技	5	手頃だ	5	オススメ	7
外観	4	楽しめる	5	嬉しい	7
存在感	4	綺麗だ	5	勧める品	6
完成度	4	解ける	5	最高だ	6

表 5. ②における表層表現の出現頻度上位

attribute		value	
音	88	有る	158
値段	56	出来る	75
デザイン	47	高い	55
機能	40	無い	49
内容	33	簡単だ	35
色	30	出る	29
サイズ	28	する	23
音質	28	多い	20
価格	26	大きい	20
動作	21	就く	20
操作	20	問題	19
大きさ	19	少ない	19
雰囲気	19	綺麗だ	19
ストーリー	17	可能だ	18
作る	16	気成る	18
時間	16	十分だ	17
声	15	シンプルだ	16
価値	14	変わる	15

表 6. ③における表層表現の出現頻度上位

attribute		evaluation	
デザイン	29	良い	107
音	29	いい	60
価格	24	素晴らしい	36
出来	23	善い	26
勝手	21	悪い	20
大きさ	14	満足出来る	18
値段	13	最高だ	15
音質	13	気に入る	13
パフォーマンス	11	勧める	11
操作性	11	満足	10
バランス	9	驚く	10
機能	9	好きだ	8
質感	8	星 5 つ	8
内容	7	満足する	8
可動範囲	7	満足だ	8
形状	7	不満が有る	7
サイズ	6	イイ	6
性能	6	難点	6

表 7. ④における表層表現の出現頻度上位

value		evaluation	
使える	9	いい	36
安い	9	気に入る	36
コンパクト	8	残念だ	28
便利だ	6	勧める	27
入る	6	良い	26
収録さ	6	嬉しい	20
有る	6	善い	19
無い	6	オススメ	13
軽い	6	満足する	13
収録	4	満足だ	11
可愛い	4	欠点	11
よく再現さ	4	満足	10
大きい	4	星 4 つ	8
楽しめる	4	難点	8
重い	4	ビックリする	7
面白い	3	素晴らしい	7
うるさい	3	魅力	7
コードレス	3	ビックリ	6

表 8. ⑤における表層の
出現頻度上位

value	
便利だ	43
面白い	38
楽しめる	36
可愛い	32
使える	31
楽しい	30
重い	29
就く	27
快適だ	25
軽い	24
する	21
小さい	21
大きい	20
最適だ	19
使いやすい	18
十分だ	18
有る	18
凄い	17

表 9. 各構成要素の組と出現順の関係

組	出現順	個数
attribute - value - evaluation	att - val - eval	466
	att - eval - val	13
	val - att - eval	97
	val - eval - att	4
	eval - att - val	69
	eval - val - att	4
attribute - value - ϕ	att - val	2612
	val - att	413
attribute - ϕ - evaluation	att - eval	583
	eval - att	73
ϕ - value - evaluation	val - eval	575
	eval - val	73

けで、出現回数が増えてしまう結果となった。そのため、今回の考察の対象からは除外した。

次に、attribute, value, evaluation について、出現回数と、その回数出現した表層表現の種類数を調べた。attribute に関しては出現回数 1 回のものが 1533 種類あり、文書中で attribute が注釈付けされた部分の 35%程度であった。また、出現回数 2 回のものは 201 種類、3 回以上のものが 268 種類であり、attribute が注釈付けされた表層表現の種類で見ると 75%以上が出現回数 1 回であった。value に至っては出現回数 1 回のものが 3589 種類あり、value が注釈付けされた部分の 43%以上であった。value においては出現回数 2 回のものは 341 種類、3 回以上のものは 394 種類であり、value が注釈付けされた表層表現の種類数で見ると 83%以上が出現回数 1 回であった。

最後に evaluation については、出現回数 1 回のものが 384 種類、文書中で evaluation と注釈付けされた箇所内 20%以下であった。出現頻度 2 回のものも 98 種類、3 回以上のものが 102 種類であり、表層の種類数からみても、出現頻度 1 回のものは 70%以下であった。evaluation は attribute, value に比べると、表層表現の種類が少ない上に、同じ表層表現が多用される結果となっている。一方で、attribute, value については 1 回しか出現しない表層表現が非常に多くなっている。特に、value は全体の数も多い上に、種類数も多くなっている。value は、製品の様態がどのようになっているかを言い表すものであるために、記述のされ方が多様になり、このような

結果になったと考えられる。

我々は、今後の研究において機械学習手法による評判情報の構成要素抽出を想定している。value における出現頻度 1 の表層表現の多さは、今後の自動抽出を困難にする原因になると予測される。

4. 2 構成要素の組における傾向

本節では各構成要素の省略のされ方について分析する。今回のコーパス作成に用いたモデルでは、4 つ組中の各構成要素について省略を許している。なお、事前説明において注釈者には「value, evaluation に着目して組を探す」ように説明をしたため、value, evaluation が共に省略されている組はない。

また、本研究のモデルでは各構成要素を組にする際に、1 対多の関係を確認している。例として、図 2 に、構成要素が 1 対多関係になっている文の例を示す。図 2 の例では「しっかりしている」という 1 つの value に対して「デザイン」「操作性」という 2 つの attribute が組になっている。

<item>この製品</item>は<attribute>デザイン</attribute>、<attribute>操作性</attribute>が<value>しっかりしている</value>。

図 2. 構成要素間が 1 対多関係になっている例

本節以降で述べる組数は、1 対多関係を全て展開した後の、延べ組数である。表 3 に、省略された要素ごとの組数と総組数を示す。表中の各構成要素について、上から順に出現頻度の上位の表層

表現である。横方向については組などの対応関係を意味するものではない。また、 ϕ が省略されている要素である。

本モデルでは **evaluation** はそれ単体で肯定・否定などの評価を確実に表すものとしたため、**evaluation** を持たない組が多い結果となった。

次に、それぞれの構成要素が省略された場合について、各構成要素の出現頻度上位を見てみる。表 4 は省略要素のパタン①における各構成要素について表層表現の出現頻度上位を示している。特に **attribute** について、コーパス中の出現頻度とは少し異なる結果になっている。「スキャン」、「吸引力」、「演技」などが高い順位となっているのは、製品への評価まで言及する場合に、その製品に固有の特に重要となる **attribute** に着目することが多いためと考えられる。

パタン②の場合の **value** と **evaluation** の出現頻度上位を表 5 に示す。パタン①と比べて、コーパス中に出現する頻度上位に近い結果となっている。これは、組数自体が多いために、全体の結果に近くなっていることもあるが、**attribute**, **value** 共に、様々な製品に広く関連しそうな一般的なものが上位になりやすい特徴があるためと考えられる。

次に、省略要素のパタン③の場合の表層表現の出現頻度上位を表 6 に示す。**value** を数値などと言い表すのが難しい **attribute** については **value** に言及することなく **evaluation** が記述されるのではないかと予想をしていたが、全体の傾向と大きく違わない結果となり、はっきりとした傾向は見られなかった。

一方でパタン④の場合の表層表現の出現頻度上位を示した表 7 では特徴があらわれている。全体では出現頻度 1 位だった「有る」の順位が下がり、「使える」、「安い」、「コンパクト」などが上位にきている。これらの **value** は **attribute** と共に現れなくても **value** のみから **attribute** の特定がしやすいものであると考えられる。

同様にパタン⑤の場合の **value** の表層表現の出現頻度上位を表 8 に示す。この場合はパタン④と同様に **attribute** が省略されているのだが、**value** の傾向が、より主観的なものとなっている。本モデルでは **evaluation** を「それ単体で肯定・否定・中立を明らかにするもの」としており、主観性の高い **value** であっても、必ずしも肯定的や否定的に決まるとは限らないものは **value** として扱っている。ここに現れている **value** は特に数値などと言い表すことが難しく、**attribute** も簡単に一語で表せないものが多く現れている。また、表 7 と比較して考えると、このような **value** については **evaluation** まで言及する場合が少ないと

も考えられる。

次に、要素が省略される場合には、文中での要素の出現順に傾向があるのではないかと考え、調べた。例えば、「この製品の良い点は○○、□□、・・・」のように、**evaluation-attribute** の順番で出てくる場合には **value** が省略される場合がある、などのような特徴が統計的に有るかということである。表 9 に出現順とその組数を示す。

表 9 を見ると、**attribute-value-evaluation** の構成要素順に出現している場合が最も多くなっている。我々が評判情報のモデル化を行う際に想定していた出現順が、レビュワーがレビューを記述する際にも順番通りに記述することが多いことがわかる。また、**attribute** と **value** は関連性が強いいため、間に **evaluation** が記述される場合は少なくなっている。

4. 3 evaluation からみた傾向

本節では構成要素の中でも特に **evaluation** に着目してコーパスの特徴を分析する。

4. 2 節の観察によれば **evaluation** が省略されている組が多数存在した。この中には、同じ **attribute**, **value** が記述されているにもかかわらず、**evaluation** までとはっきりと言及する場合と、そうでない場合がある。そこで、**evaluation** を持つ組の **attribute-value** に注目したときに、その **attribute-value** が現れているが **evaluation** が現れていない場合があるかどうかを調べた。結果、**evaluation** を含む組 2606 組の内 521 組が **evaluation** を含まない組に同一の **attribute**, **value** を持つことが分かった。例を図 3 に示す。

操作 - 簡単だ - 不満だ
操作 - 簡単だ - ϕ
低音 - 出る - オススメ出来る
低音 - 出る - ϕ

図 3. **evaluation** の有無が異なる場合

表 1 のパタン①②を合わせたものの中に出現する **attribute-value** の表層表現の出現頻度上位について、**evaluation** まで言及される場合と、**evaluation** が省略される場合、それぞれの数を表 10 に示す。

表 10 を見ると、**attribute-value** の中では頻出する組に対しても、必ずしも **evaluation** が組となっているわけではないことが分かる。また、表 10 で興味深い点として、「値段・手頃だ」の場合には **evaluation** が現れているのに対して「価格・手頃だ」の場合には **evaluation** が現れてない点がある。表 6 を見ると「価格」という **attribute** と **evaluation** が組になりにくいわけではなさそ

うであるにも関わらず、出現頻度に大きな差は無く、似た内容の attribute-value である両者に対して evaluation の現れ方が異なっている。これについては、収集した文書の網羅性によるものであるかをさらに調査する必要があると思われる。

表 10. attribute-value に対する evaluation の有無

attribute	value	出現回数	eval 有り
値段	高い	16	3
値段	安い	15	6
デザイン	シンプルだ	12	8
音	静かだ	11	4
価値	有る	9	0
時間	架かる	9	0
完成度	高い	8	2
低音	出る	7	1
価格	安い	7	3
値段	手頃だ	7	5
注意	必要だ	7	0
価格	手頃だ	6	0
組み立てる	簡単だ	6	0
音	うるさい	6	0
音	する	6	0

さらに、パタン③④を合わせたもののの中の value の出現頻度上位について、evaluation の有無を調べた結果を表 11 に示す。

表 11. φ-value に対する evaluation の有無

value	出現回数	eval 有り
便利だ	49	6
面白い	42	4
使える	40	9
楽しめる	40	4
可愛い	36	4
重い	33	4
楽しい	32	2
軽い	30	6
就く	28	1
快適だ	27	2
大きい	24	4
安い	24	9
有る	24	6
する	22	1
使うやすい	21	4
入る	21	6
小さい	21	0
凄い	19	2
十分だ	19	1
最適だ	19	0

表 11 の結果を見るとほとんどの value が少ないながらも evaluation と組になっている。また、

特定の value の場合には evaluation と高確率で組になるなどの特徴を期待したが、そのような傾向はみられなかった。

次に、evaluation の極性を中心にコーパスの傾向を分析する。

本コーパス中で evaluation と注釈付けされた表層表現には注釈者により positive, negative, neutral のいずれかの極性を付与されている。コーパス中の全 evaluation の極性分布は positive:1894, negative:172, neutral:172 となっている。4. 1 節の頻出する evaluation から分かる結果と同様に、肯定的な記述が大部分を占めている。これは、製品レビューという文書が、製品を他人に勧めるという性質を持つことが原因と考えられる。

ここでは、同じ attribute-value に対して異なる極性の evaluation が組になっている場合を調べた。これは、同じ製品の様態についても注釈者により評価が異なる点であり、今回のコーパス作成に用いた評判情報のモデルが効果的に用いられている点である。表 3 のパタン①の中の attribute-value の中で、極性の異なる evaluation が組になっているものを図 4 に示す。図 4 では attribute-value の組と、それに対する肯定的否定的それぞれの極性を持つ評価を示している。

操作・簡単だ・(不満だ↔よい)
機能・有る・(満足だ↔若干敬遠する)
音質・解る・(良い↔失格です)
故障の原因・解る・(良い↔失格)
値段・高い・(満足↔残念な点)
音・小さい・(良い点↔残念だ)
違和感・無い・(非常に気に入る↔落胆する)
視野率・95%・(星 4 つ半↔残念だ)

図 4. 極性の異なる evaluation と組になっていた attribute-value

図 4 を見ると、一般的には肯定的と思われる「機能-有る」や、一般的には否定的と思われる「値段-高い」についても極性の異なる評価が付与される場合があることがわかる。同様に表 3 のパタン④についても、11 種類の value に対して極性の異なる evaluation が付与されていた。図 5 に 11 種類の value を示す。

今回の調査では表記揺れを吸収しきれていない部分があるが、筆者がコーパス中を目視で確認したところ、似た内容の value に対して極性の異なる evaluation が組になっている場合はまだ存在している。例えば「忠実に再現されている(忠実に再現している)」や「予想以上(予想外)」などの value に対して、極性の異なる evaluation

が組になっている場合が確認されている。このことから、製品の様態から必ずしも評価の極性が決まるわけではなく、この2つを分けて扱う必要があると考えられる。

読みやすく・(ヨカッタです↔不満)
有る・(善い↔不安を感じる)
安い・(心配だ↔オススメ)
無い・(オススメ↔弱点)
安っぽい・(いい↔がっかりする)
多い・(とてもよい↔オススメしません)
収録され・(残念です↔大満足です)
入る・(残念だ↔大喜びだ)
含む・(良ポイント↔欠点)
違う・(気に入っている↔残念だ)
神秘的な感じ・(好きだ↔残念だ)
使用できる・(オススメ↔落とし穴)

図 5. 極性の異なる evaluation と組になっていた value

5. おわりに

本稿では、人手により注釈付けされた1万文の評判情報コーパスについて、コーパス中における評判情報の現れ方を統計的に調査・分析した。結果として、各構成要素が省略される場合の特徴について知見を得ると共に、製品の様態と評価を分離するモデルにより正確に捉えることが出来た評判情報の記述を確認することが出来た。

今後、コーパスを用いて評判情報の自動抽出を行うと同時に、コーパスの規模が十分であるかどうかを確認する必要がある。また、今回作成した評判情報コーパスの公開について検討している。Web から収集した文書については2次配布が困難なため、原文書の収集法について検討中である。

謝辞

本研究の一部は、独立行政法人情報通信研究機構の委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」プロジェクトの成果である。

参考文献

- [1] 立石健二, 石黒義英, 福島俊一: インターネットからの評判情報検索, 情報処理学会研究報告 NL144-11, pp.75-82 (2001)
- [2] 小林のぞみ, 乾健太郎, 松本祐治, 立石健二, 福島俊一: テキストマイニングによる評価表現の収集, 情報処理学会研究報告 NL-154-12, pp.77-84 (2003)
- [3] 村野誠治, 佐藤理史: 文型パターンを用いた主観的評価文の自動抽出, 言語処理学会第9回年次大会発表論文集, pp.67-70 (2003)

- [4] Janyce Wiebe, Theresa Wilson and Claire Cardie: Annotating Expressions of Opinions and Emotions in Language, Language Resources and Evaluation, Vol.39, No.2-3, pp.165-210 (2005)

- [5] 小林のぞみ, 乾健太郎, 松本祐治: 意見情報の抽出/構造化のタスク仕様に関する考察, 情報処理学会研究報告 NL171-18, pp.111-118 (2006)

- [6] 宮崎林太郎, 前田直人, 森辰則: 人手による評判情報注釈付けにおける揺れの分析と注釈付け支援ツール, 情報処理学会研究報告 NL176-21, pp.143-150 (2006)

- [7] Nobuhiro Kaji and Masaru Kitsuregawa, Automatic Construction of Polarity-tagged Corpus from HTML Documents, Proceedings of the 21st International Conference on Computational Linguistics (COLING/ACL), pp.452-459 (2006)

- [8] Nozomi Kobayashi, Kentaro Inui and Yuji Matsumoto: Opinion Mining from Web Documents: Extraction and Structurization, 人工知能学会論文誌, Vol.22, No.2, pp.227-238 (2007)

- [9] 関洋平, David Kirk Evans, Hsin-His Chen, Lun-Wei Ku, 神門典子: 意見分析タスク-多言語テキストを対象とした意見抽出技術の評価, 情報処理学会研究報告 NL183-8, pp.51-58 (2008)