

電子掲示板からの比較文およびその構成要素の抽出

尾崎 啓太[†] 木村 文則[‡] 前田 亮[‡]

[†] 立命館大学 情報理工学研究科 〒525-8857 滋賀県草津市野路東 1-1-1

[‡] 立命館大学 情報理工学部 〒525-8857 滋賀県草津市野路東 1-1-1

E-mail: [†] is011084@ed.ritsumei.ac.jp, [‡] fkimura@is.ritsumei.ac.jp, [‡] amaeda@media.ritsumei.ac.jp

あらまし 本稿では、掲示板に投稿された発言からの比較文およびその構成要素の抽出手法について述べる。比較文とは、任意の2つ以上の対象間における優劣を表す文である。この比較文を構成する要素である「対象、属性、評価」の組を文節単位で抽出する手法を提案する。具体的には、実際の掲示板を観察した知見から得た手法により、我々が定めた比較文の定義に合致する比較文を抽出し、比較文の各構成要素における特徴の違いを利用して解析することにより、比較文の構成要素である対象—属性—評価を決定する。提案手法を用いた評価実験を行い、その結果に基づいた今後の展望を述べる。

キーワード 比較文, 電子掲示板, 評価表現

Extraction of the Comparative Sentence and its Components from BBS

Keita OZAKI[†] Fuminori KIMURA[‡] and Akira MAEDA[‡]

[†] Graduate School of Information Science and Engineering, Ritsumeikan University

1-1-1 Nojihigashi, Kusatsu-shi, Shiga, 525-8857 Japan

[‡] College of Information Science and Engineering, Ritsumeikan University

1-1-1 Nojihigashi, Kusatsu-shi, Shiga, 525-8857 Japan

E-mail: [†] is011084@ed.ritsumei.ac.jp, [‡] fkimura@is.ritsumei.ac.jp, [‡] amaeda@media.ritsumei.ac.jp

Abstract This paper describes a method to extract the comparative sentence and its components from statement that was posted on the bulletin board. The comparison statement is a statement representing the relative merits between any two or more targets. Propose a method to extract a set of target-attribute-evaluation which is elements that make up this comparative statement. Specifically, by the method we have obtained from findings that observed actual forums, we have extracted the comparative sentence that meets the definition of a comparative sentence that we determined. Next, by utilizing to analysis the difference of the characteristic of each component of the comparative sentence, we determined the object, attribute, evaluation that is a component of comparative sentence. Conducted experiments using the proposed method, I describe the outlook for the future based on the results.

Keyword Comparative Sentence, Forum, Expression of evaluation

1. はじめに

現在、インターネットを介して複数の人数で交流できるコミュニティが増加している。スマートフォンなどの新たなインタフェースの登場は利用者が拡大する要因となっている。利用者にはtwitterやFaceBookなどのSNSを利用する人が多い。そのようなソーシャルメディアの中でも特に匿名掲示板に注目する。掲示板にはあるテーマ（スレッドタイトル）に関する発言が書き込まれるスレッドが存在している。スレッドでは始めに設定されたテーマについて複数の人と情報交換や会話・議論ができる。基本的に文字のみで他人とやりとりすることになるため、議論が活発になるケースが

多く、テーマに関する物事の優劣や特徴の違いが議論の対象となっていることも多い。これらの議論文中には、「ある特定の観点における物事の優劣」を判定するための情報が含まれている。これらの情報を自動的に取得することで、注目している対象の優劣の判断をする際の補助を目的とする。

そのため、本研究では前段落に述べたような物事の優劣や特徴の違いを表す発言、つまり比較文をより多く発見し、その発見された文に含まれる比較構成要素を抽出する。比較文が多いと考えられる掲示板を対象とし、本稿で定義した比較文の定義に基づいた文を抽出する。比較文とは“この店のほうがあの店よりもオ

ムライスはおいしい”といった文である。比較文であると判断できる発言の特徴やその特徴に対する係り受け解析を用いて、スレッドに書き込まれている比較文を抽出する手法を提案する。

次に、比較文の比較構成要素である「対象、属性、評価」を文節ごとに抽出する。文節単位である理由は、掲示板に投稿される発言は主に口語表現が用いられており、単語単位における比較構成要素の抽出は困難であるからである。例えば、「これだけなのか」という語を単語単位で抽出すると「これ」「だけ」「な」「の」「か」のように分割されてしまう。しかし、文節単位で抽出するとそのような語句を一つの塊として抽出することができるため、掲示板の口語表現に適していると考えた。

上文の例文から比較文の構成要素を抽出すると、対象は「この店のほうが」、属性は「オムライスは」評価は「あの店よりもおいしい」となる。比較文抽出の際に、このような要素を抽出する手法も提案する。

本論文では、実験対象として掲示板「2ちゃんねる」[1]のスレッドを用いている。比較文やその構成要素の抽出結果に対する評価実験と、その評価に対する考察を行った。

2. 関連研究

本節では、比較文や評価観点の抽出や分類に関する関連研究について述べる。

比較文の抽出では、倉島ら[2]は日本語の比較関係の対象間の優劣（A店とB店のどちらが安いかわいいかといった文）のみに着目し、比較表現に対する知見から得た規則を用いて、文集合から比較文を構成する要素「基準、対象、属性、評価」を抽出している。また、Jindalら[3]は人手で作成した英語の比較を表す手がかり語句のリストを用いて比較文候補を網羅的に収集し、同リストから作成した Class Sequential Rules を素性として用いる比較文分類器を提案している。

評価観点の抽出に関して、飯田ら[4]は複数の評価候補と1つの属性の組を取り出し、属性と評価の最適な組み合わせを教師あり学習で判定し、「属性—評価」対を抽出している。鈴木ら[5]は、ナイーブベイズ分類器とEMアルゴリズムを組み合わせることにより、「対象、属性、評価語」候補を抽出している。

本研究では、比較文の抽出の際、特定の表現に依存しない方法を用いており、評価観点の抽出では、主に口語表現で発言が投稿される掲示板を対象としている点と、対象が口語表現による発言が多いため、文節ごとに観点を抽出している点で、上記の論文と異なる。

3. 提案手法

掲示板に対する比較文とその構成要素の抽出手法の提案と比較文の定義について述べる。

本手法では、掲示板のスレッドから対象とする比較文を取り出すために、比較文の定義を定める必要がある。実際の掲示板に投稿される発言と比較の意味を考慮した比較文の定義を定め、その定義に合致した比較文を抽出する。それらの比較文からその構成要素となる「対象、属性、評価」を抽出する。

3.1. 比較文の定義

定義1.

1つの発言中に2つの比較され得る対象（比較対象）がある

定義2.

1つの発言中に比較対象①・比較対象②・比較対象①or②に対する評価表現（評価）が含まれている

定義3.

各発言から、比較対象間の関係として「優劣・同等・最上級・特徴の違い」のいずれかが読み取れる

上記のそれぞれの定義が掲示板に投稿された発言に一致する一例を紹介する。

図1に示す発言のように、比較され得る対象としてサッカーゲームで代表的な「ウイイレ（ウイニングイレブン）」、「FIFA」の2つの比較対象があり、両対象に任意の評価が下され、関係性として対象間の優劣を読み取ることができる。よって、この発言は比較文である。

324 : 名無しさん@お腹いっぱい。[sage] : 投稿日: 2012/10/30 21:58:02 ID:aA5gw08I [1/1回]
ウイイレの選手ってゾンビだね
FIFAのフェイスの方が方向性としては正解

図 1: 対象間に優劣の関係を持つ発言

図2に示す発言は同等の意味を表す比較文の例である。比較対象は国の名前である。「日本」と「クロアチア」「スウェーデン」の両対象に任意の評価が下され、同程度の評価を読み取ることができる。

137 : あ:2012/09/10(月) 13:05:14.67 ID:pMFbnpco
日本はデンマーク以上フランス以下のレベル
この中に当てはまるクロアチアやスウェーデンなどと同等って事だよ

図 2: 対象間に同等の関係を持つ発言

図3に示す発言は最上級の意味を表す比較文の例である。対象は「イルカ」などスマートフォンのアプリケーション名である。各対象に任意の評価が下され、「ガラパゴスブラウザ」が一番良いという最上級の評価を

読み取ることができる。

246 : 名無しさん@お腹いっぱい。 : 2011/01/27(木) 11:04:27 ID:Klva+LG4 [1/1回発言]
イルカやオペラ、アンドロメダなどいろいろ使ってみたがガラパゴスブラウザが神だと思わ
れる

図 3: 対象間に最上級の関係を持つ発言

図 4 に示す発言は特徴の違いの意味を表す比較文の例である。対象は「Andmade share」などスマートフォンのアプリケーション名である。各対象に任意の評価が下され、「設定しやすい」「設定しにくい」のように対象間の特徴の違いを読み取ることができる。

37 : 名無しさん@お腹いっぱい。 : 2012/08/13(月) 13:36:15.71 ID:L/aE9fr8
Andmade shareは見た目スッキリ設定しやすい
Share Custmizerは黒背景(!)見た目スッキリ設定しにくい
I HUBは見た目カスタムし放題設定しやすい

図 4: 対象間に最上級の関係を持つ発言

3.2. 比較文抽出

実際の掲示板を見て、比較文の定義に合致する発言を調査した。「2ちゃんねる」掲示板からジャンルが異なる色々なスレッドから比較文を手で収集した。調査した発言数の合計は約 1 万である。13 個のスレッドに対して調査を行い、各スレッドにおけるジャンルの内訳は「iphone」、「お菓子・甘味」、「スポーツ」など様々である。結果として収集できた比較文の総数は 340 文である。

掲示板から収集した比較文と比較文ではない文を見比べた結果、図 5 のような 1 つの評価表現が 1 つの比較対象に掛かっている組みを 2 つ持つ関係性を持った発言が比較文に多いことが分かった。

30 : iPhone774G : 2012/09/21(金) 17:09:36.08 ID:OV1WoMp5P
裏は黒はよかった。白はいまいち。
でも、白はエッジがキラキラしててすごくキレイに見える。黒は同化し
ちゃった。

図 5: 実際の掲示板の比較発言

図 5 の赤文字「白・黒」は比較されている 2 つの比較対象である。青色下線部は比較対象に掛かっている評価表現「よかった・いまいち」であり、1 つの発言中に比較対象の両方に掛かっている。このとき、「裏」「エッジ」などの属性が発言中に含まれているが、比較文の抽出の際には判定に用いない。以上の発言を抽出する際の流れを図 6 に示す。

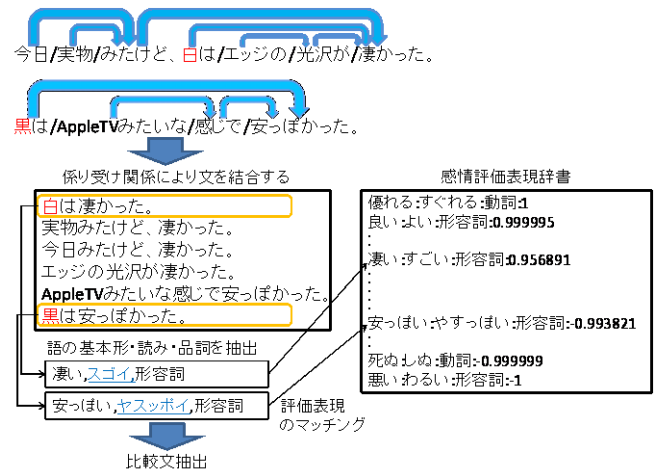


図 6: 比較文抽出の流れ

比較文に対する必要な関係性として、比較対象と評価表現の係り受け関係を判定する必要がある。対象と評価は必ずしも隣り合った文節間で係り受け関係が形成されている訳ではないため、各文節間で係り受け関係にあるすべてのパターンで繋げる。ここでは、係り受け関係取得に Cabocha[6]を用いている。Cabocha とは、解析対象の文から各単語を区切る形態素解析と文節の区切り済みデータを出力するツールである。次に、対象と評価のマッチング処理を行う。対象に掛かる評価表現が適切であるかを判定するため、感情評価表現辞書[7]を用いる。その辞書に含まれる単語の基本形、読み、品詞の 3 つ全てにマッチした場合のみ、評価表現であると判定される。図 6 では、「凄い、スゴイ、形容詞」の 3 つの要素全てに一致する項目が辞書に含まれている。このマッチング処理を各発言に対して行い、2 つの比較対象と評価表現がこの処理でマッチしたとき（図 6 の黄色で囲われている 2 つの文が存在するとき）、その発言を比較文として抽出する。

3.3. 比較文の構成要素抽出

3.3.1 比較文の前処理と tfidf 法

3.2 節で取り出した比較文の構成要素として、「対象、属性、評価」を抽出する。基本的には、対象と係り受け関係となる属性や評価を取り出すが、対象が出現しない文においても、対象に対する属性や評価が存在することがあるため、どの文が対象に対する評価文であるかを判断する必要がある。その処理を図 7 の青色矢印の段階に示す。

14: 名無しさん@お腹痛いっばい。[a] : 投稿日:2012/10/17 16:32:09 ID: +jTsjk2m [1/3回]

ウイイレってオンライン30000人しかいないんだな ... (1)
初週22万売れてこれだけかオン対戦やってるの ... (2)
話になんないなw ... (3)

FIFA12なんて多い時で何時も18万ぐらいいたけど

FIFA13も日本版発売でその日の内に軽く10万は超えるよ

やっぱ試合が楽しくないんだらうなウイイレw

対象に対する属性や評価を含む文を抽出

ウイイレって/オンライン/30000人しか/いないんだな
初週/22万/売れて/これだけか/オン対戦/やってるの
話に/なんないなw

FIFA12なんて/多い/時で/何時も/18万ぐらい/いたけど
FIFA13も/日本版発売で/その日の/内に/軽く/10万は/
超えるよ

やっぱ/試合が/楽しくないんだらうなウイイレw

対象・属性・評価を抽出

[対象]:ウイイレって[属性]:オンライン[評価]:30000人しか/いないんだな
[属性]:オン対戦/やってるの[評価]:これだけか
[評価]:話に/なんないなw

図 7: 比較文構成要素の前処理と抽出例

図 7 の青矢印のように発言中の先頭の一文から順に比較対象を含む一文を起点として解析する。発言を一文ずつ順に解析するとき、その一文中に別の対象が存在しない場合、起点とした文に含まれる比較対象に対する発言として取り出す。例えば、図 7 の上の(1)文中に「ウイイレ」という比較対象が存在するとき、次の(2)文に比較対象が存在しない場合、比較対象「ウイイレ」に対する属性や評価が含まれている文であると判定し、(1)文に(2)文を足し、これを(1+2)文とする。次の(3)文も比較対象が含まれていないため、(1+2)文に(3)文を足す。この処理によって分割された文の組から対象とその属性、評価を抽出する。このとき、掲示板の投稿発言は口語表現であることが多く、文語表現のように文法的に正しい日本語となっていないパターンが多いという理由から、単語単位で比較文の構成要素を取り出すのではなく、文節単位で取り出す。

この処理の後、比較文の構成要素を文節ごとに取り出す。そのためには、それぞれの文節が、対象・属性・評価・その他（構成要素に当てはまらない文節）のいずれかに該当するかについて判定する必要がある。このとき掲示板における各構成要素の特徴は以下のとおりである。

- 対象... 特定のスレッドにのみ出現する可能性が高い文節
- 属性... 特定のスレッドにのみ出現する可能性が高い文節
- 評価... 多くのスレッドに出現する可能性が高い文節
- その他... 対象・属性・評価以外の文節

更に、対象は特定の人、物、組織、製品など固有名詞である可能性が高く、属性に比べてより他スレッドで使用される頻度が少ないと予想される。そのため、各構成要素の間に「対象>属性>評価」のように重要

度の違いが現れると考えられ、各構成要素である文節内の単語も同様に考えた結果、単語の重要度を求めるために tfidf 法を用いた。掲示板の 1 つのスレッドを 1 文書とみなし、tfidf 値の算出を行う。

このとき、単語の tfidf 値算出に利用したスレッド数（文書数）は 108 である。また、発言数は 104,544、単語数は 3,072,131 である。tfidf 値を求める際に「全角記号、半角記号、数字、スペース」を除いている。

3.3.2 比較構成要素抽出に用いる品詞情報

次に単語の品詞情報を活用する。比較構成要素を抽出する際、解析対象とする文節内の品詞も用いる。各品詞が現れた時にどの構成要素（対象、属性、評価、その他）として現れやすいかの出現確率を用いれば、比較構成要素の抽出精度を向上させることができると考えた。

活用する品詞情報として、文節内に含まれる一単語の品詞と文節内の品詞の並びを用いる。用いる品詞情報の例を以下に示す。下の(1)文節の「名詞」や「助詞」といった 1 つ 1 つの品詞と、(1)文節内の品詞の並びである「名詞」・「助詞」の並びを品詞情報として用いる。

(1)文以下の文節も同様に品詞情報を用いる。

- 「ウイイレ/って」 名詞+助詞 (1)
- 「3/0/0/0/0/人/しか」 名詞+名詞+名詞+名詞+名詞+名詞+助詞
- 「オン/対戦」 名詞+名詞
- 「これ/だけ/か」 名詞+助詞+助詞

各構成要素の文節内の単語単位における品詞の出現確率と文節内の全ての単語における品詞の並びの出現確率を用いる。出現確率算出にはベイズの法則を適用する。このとき、求められる確率は、各品詞が各要素に属する確率である。この出現確率を求めるために、人手で各構成要素のデータを収集した。データ収集の際の判断基準は以下の通りである。

1. 各構成要素を目視で判断する
2. 品詞は考慮しない
3. 属性がある場合はそれに対する評価を収集する
4. 属性がない場合は評価のみを収集する
5. その他ができるだけ少なくなるように収集する

上に載せた判断基準一覧を詳細に説明する。1 の条件は、比較文を見たときに直感的に各構成要素を集める条件である。2 では、属性は名詞、評価は形容詞などといった制約を付けない条件である。3 と 4 は評価が文中に含まれていても属性は必ずしも含まれているとは限らない場合の条件付けである。

[対象]:ウイイレって[属性]:オンライン[評価]:30000人しか/いないんだな
[属性]:オン対戦/やってるの[評価]:これだけか
[評価]:話に/なんないなw

図 8:比較文の構成要素

図 8 に示す文節をみると「オンライン 3 0 0 0 0 人しか/いないんだな」のように属性と評価を含む文もあれば、「話に/なんないなw」のように評価のみを含む文も存在するためである。5 については属性や評価をできるだけ詳細に抽出することを目的としているため、その他が少なくなるようにデータを収集している。正解データを収集した結果の一部を以下の表 1 と表 2 に示す。表 1 は各品詞であるときの各比較構成要素の出現頻度を並べている。表 2 は各品詞の並びであるときの各比較構成要素の出現頻度である。

表 1: 正解データをまとめた表 (各品詞)

品詞	対象	属性	評価	その他
名詞,一般,*	80	105	107	28
助詞,係助詞,*	33	37	28	9
助詞,格助詞,一般	19	75	59	19
名詞,固有名詞,組織	18	6	6	2
助詞,連体化,*	16	21	13	6

表 2: 正解データをまとめた表 (各品詞の並び)

品詞	対象	属性	評価	その他
名詞-助詞	59	83	53	32
名詞-名詞-助詞	8	33	17	3
名詞	7	8	9	8
名詞-助動詞-助詞	4	1	4	2
名詞-助詞-記号	2	1	0	0

表 1 や表 2 のデータから各構成要素であるときの各品詞である出現確率を求める。例えば、構成要素が対象、品詞が名詞であるときの同時確率 $P(A \cap B)$ (事前確率) は以下の(1)式で求められる。

$$\text{例. } P(\text{対象} \cap \text{名詞}) = P(\text{名詞} | \text{対象})P(\text{対象}) \quad (1)$$

一般的には以下の(2)式で表される。

$$P(A \cap B) = P(A | B)P(B) \quad (2)$$

式(2)からベイズの法則を用いると、各品詞がどの要素に属するかを確率で表すことができる。以下の(3)式がその確率値 (事後確率) を求める式である。実際にベイズの法則を適用すると、 A_0 から A_3 まで「対象, 属性, 評価, その他」といった各構成要素であり、 B は名詞や形容詞といった各品詞を表す。

$$P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_{i=0}^n P(B | A_i)P(A_i)} \quad (3)$$

求めた確率値を tfidf 値に適用することを考える。tfidf 値は「対象, 評価, 評価, その他」の間で「対象」から「その他」まで重要度の高い順番になっているため、適用する値として、「対象>属性>評価」の順に重み付けする必要がある。重み付けは各比較構成要素間の性質を考慮する必要がある。ある品詞における対象や属性の出現確率が高ければ高く、評価やその他の出現確率が高ければ低くなるように重み付けをする。

こうした考えの元、以下の(4)式のように比較文の比較構成要素における品詞の重要度を 1 つの値 $W(B)$ で表現した。

$$W(B) = \frac{3 \times P(A_0 | B) + 2 \times P(A_1 | B) + P(A_2 | B)}{P(A_1 | B) + 2 \times P(A_2 | B) + 3 \times P(A_3 | B)} \quad (4)$$

3.3.3 tfidf 値と品詞情報を用いた比較構成要素抽出

本節では、3.3.1 節で求めた tfidf 値と 3.3.2 節で求めた品詞情報を用いて算出した値を組み合わせ、比較文の比較構成要素を抽出する手法を説明する。

$$I = \frac{\sum_{x=1}^t tf \times \log \frac{N}{n} \times W_1(B) \times W_2(B)}{t} \quad (5)$$

(5)式は tfidf と品詞情報を組み合わせて文節内の重要度を求める式である。tf はあるスレッド内のある単語の出現頻度である。idf はある単語を含むスレッドの数を全スレッド数で割った数である。t は文節内の単語数を示す。 $W_1(B)$ は文節内の単語単位の品詞情報を使い、(1)式から(3)式までの課程で求めた数値である。 $W_2(B)$ は文節内の全ての単語における品詞の並びを使い、(1)式から(3)式までの課程で求めた数値である。I は文節内の重要度を示す数値である。

品詞情報から求めた数値は確率値であることを考慮し、tfidf 値と掛け合わせている。また、文節内に含まれる単語数 t により、文節の重要度 I に差が開いてしまうため、文節内の単語数 t で正規化している。

表 3 の左から 4 列目までの数値はベイズの法則を用いた事後確率であり、右端の列に並ぶ数値は左 4 列の同じ品詞の数値を用いて求めた $W_1(B)$ である。

表 3: 事後確率と $W_1(B)$ (一部)

品詞	対象	属性	評価	その他	$W_1(B)$
名詞,固有名詞,組織	0.919307	0.04733	0.025389	0.007974	23.58402
記号,括弧開,*	0.80789	0	0.050201	0.141909	4.702025
助詞,係助詞,*	0.551358	0.321166	0.098661	0.028814	3.959244
助詞,連体化,*	0.485177	0.387275	0.07961	0.047938	3.345878
名詞,一般,*	0.429383	0.342739	0.190921	0.036957	2.590867

最後に、各構成要素間において重要度 I に差が付くことから、対象と属性、属性と評価、評価とその他のそれぞれの間に閾値を手動で設定し、各構成要素を抽出する。

4. 実験

本節では、3 節で提案した手法を用いた評価実験の結果について述べる。

1 つのスレッドから比較文を抽出する実験と、その比較文から構成要素を抽出する 2 つの段階でそれぞれ評価する。

4.1. 比較文抽出の評価

実験対象として使用する掲示板は「2ちゃんねる」であり、その掲示板の 3 つのスレッドに対する実験を行う。以下の表 4 に示す掲示板のスレッド A, B, C の合計 1215 個の発言に対して 3.1 節で提案した手法を用いてスレッド中に存在する比較文を抽出する。システムに対して最初に 2 つの対象を入力する必要があり、実際に入力した対象は表 4 の「対象」欄に示す。

表 4: 比較文抽出の実験結果

スレッド	対象	発言数	抽出 発言数	適合率
A	白, 黒	355	90	0.867
B	iPad, Nexus7	422	11	0.909
C	FIFA, ウイイレ	438	49	0.857

この実験結果として、表 4 に適合率を示す。本来、比較文抽出の際、再現率も重要な指標となるが、本研究の目的である比較文の構成要素抽出の際に、ノイズとなるデータが重要であることや、正解文を収集するために多くの時間が掛かることを考慮して、再現率は算出していない。

4.2. 比較文構成要素抽出の評価

4.1 節で抽出した比較文の構成要素を抽出する。今回の実験では、4.1 節で対象としたスレッド C を対象としている。実験に用いる正解データは人手で集めており、収集した正解データの内訳を表 5 にまとめる。

表 5: 収集した正解データ

対象	発言数	抽出発言数	合計
FIFA, ウイイレ	438	49	744

スレッドから抽出した比較文の構成要素のうち、正解データとなる文節の抽出数を正解抽出数とする。表 5 の合計は、対象・属性・評価・その他の文節数を足した数である。このとき、全抽出数のうちの正解抽出数の割合を適合率、正解データ全 744 個の文節のうち、正解抽出数が含まれている割合を再現率として評価する。

表 6: 比較文構成要素抽出の実験結果

手法の種類	適合率	再現率	F 値
tfidf	0.5334	0.6653	0.5921
tfidf+品詞	0.6056	0.7554	0.6722

表 6 では tfidf のみを用いた手法と tfidf と品詞情報を組み合わせた手法による適合率、再現率を示す。表 3 から品詞情報を加えることによって、各指標とも約 8 % 向上していることがわかる。

表 6 の「tfidf+品詞」による手法で抽出できた実際の文を図 9 に示す。図の下の方角欄に抽出した文節と提案手法により判定された構成要素を載せている。構成要素の横には、正しい構成要素であれば○、間違いであれば×を付けている。このとき、文中の語尾に「？」を含む疑問文は明確に発言投稿者自身の意見として捉えることが困難であるため削除している。この比較文中に含まれる構成要素を抽出することができた。対象ごとに属性や評価が抽出でき、その後も判定することができた。対象や属性に現れやすい品詞の並びである「名詞+助詞」などの組み合わせが発言中の対象や属性に用いられているため、抽出できたと考えられる。

9 : 名無しさん@お腹いっぱい。[sage] : 投稿日:2012/10/17 15:03:47 ID:SmkYDWF [2/4回]
>>8
FIFAはまだ体験版しかやってないからわからんけど、ウイイレは今回は駄目だな。
発売日からのこのトラブル続きはあり得ない。
PS3って2008からだっけ？
高塚の頃はこんななかったのに、ゲームよくなっても動かなきゃ意味ないじゃんwww

FIFAは-対象○/まだ-その他○/体験版しか-属性○/やってないから-評価○/わからんけど、-評価○
ウイイレは-対象○/今回は-属性○/駄目だな。-評価○
発売日からの-属性○/この-その他○/トラブル続きは-属性○/あり得ない。-評価○
高塚の-属性○/頃は-属性○/にんなの-評価○/なかったのに、-評価○/ゲーム-属性○/よくなっても-評価○
動かなきゃ-評価○/意味ないじゃんwww-評価○

図 9: 比較文の構成要素抽出例 (1)

表 6 の「tfidf+品詞」による手法で一部正確に抽出できなかった比較文を図 10 に示す。間違っ

は対象と判定すれば正解となる。また、「やってるの」は「ユーザ」の意味を表す属性として抽出するべきであるが、評価として抽出してしまうことがわかった。

14: 名無しさん@お腹いっぱい。[a]: 投稿日: 2012/10/17 16:32:09 ID: +jt5jk2m [1/3回]
ウイイレってオンライン30000人しかいないんだな
初週22万売れてこれだけかオン対戦やってるの
話になんないなw
FIFA12なんて多い時で何時も18万ぐらいいたけど
FIFA13も日本版発売でその日の内に軽く10万は超えるよ
やっぱ試合が楽しくないんだろなウイイレw

ウイイレって-対象○/オンライン-属性○/30000人しか-評価○/いないんだな-評価○
初週-その他○/22万-評価×/売れて-評価×/これだけか-評価○/オン対戦-属性○/や
てるの-評価×
話に-評価○/なんないなw-評価○
FIFA12なんて-対象○/多い-評価○/時で-評価○/何時も-その他×/18万ぐらい-評価○/い
たけど-評価○
FIFA13も-対象○/日本版発売で-属性○/その日の-評価○/内に-評価○/軽く-評価○/10
万は-属性×/超えるよ-評価○
やっぱ-その他○/試合が-属性○/楽しくないんだろなウイイレw-対象×

図 10: 比較文の構成要素抽出例 (2)

5. 考察

4 節の評価実験に対する考察を述べる。

適合率が低くなった要因として、係り受け解析が機能しない場合がある点が挙げられる。「楽しくないんだろなウイイレ」のような発言が 1 つの文節として判定され、適切な評価、対象を抽出することができなかった。語尾に当たるような語「だろな」と名詞「ウイイレ」を間隔なしで書き込まれている文は正確に係り受け解析ができないことがわかった。

「やってるの」は形態素解析により単語に分割すると「やっ」「てる」「の」となり、単語単位の tfidf 値や品詞情報による重要度が低くなってしまふ点が原因であることが分かった。

文単位における文の抽出方法も問題であり、不必要な文も抽出してしまう問題があった。

再現率については「FIFA みたいな」といった比較対象を含む評価が対象として抽出される問題がある。対象自体が重要度を高くするため、～みたいな、などの表現のパターンを用いるなど、適切な対処が必要である。

6. おわりに

本稿では、係り受け解析を用いた構文解析と、比較対象と評価表現のマッチング処理による比較文の判定手法を提案し、さらに単語の tfidf 値と品詞情報を用いた比較文の比較構成要素の抽出手法を併せて提案した。これらの提案手法を用いて、掲示板のスレッドを対象とした評価実験を行った。

今後、精度改善のために、全体を通して扱えていない特徴の利用を考えている。文章内の構成要素の出現位置、文のカテゴリ（スポーツ、芸能、政治、...）な

どの特徴である。今後、これらの特徴を利用した手法による実験を行うことを考えている。

参考文献

- [1] 2ちゃんねる
<http://menu.2ch.net/bbstable.html>
- [2] 倉島 健, 別所 克人, 内山 俊郎, 片岡 良治, "比較評価情報の抽出とそれに基づくランキング手法の提案", DEWS2007.
- [3] N. Jindal and B. Liu. Identifying comparative sentences in text documents. In Proc. of SIGIR 2006, 2006.
- [4] 飯田 龍, 小林 のぞみ, 乾 健太郎, 松本 裕治, 立石 健二, 福島 俊一, "意見抽出を目的とした機械学習による属性-評価値対同定(属性抽出)", 情報処理学会研究報告. 自然言語処理研究会報告 2005(1), 21-28, 2005.
- [5] 鈴木泰裕, 高村大也, 奥村学, "Weblog を対象とした評価表現抽出", 人工知能学会研究会資料, SIG-SW&ONT-A401-02, 2004.
- [6] CaboCha/南瓜: Yet Another Japanese Dependency Structure Analyzer
<http://code.google.com/p/cabocha/>
- [7] H. Takamura, T. Inui, M. Okumura, "スピンモデルによる単語の感情極性抽出(自然言語)", 情報処理学会論文誌 47(2), 627-637, 2006.