Vol.2011-IFAT-101 No.5
Vol.2011-NL-200 No.5
2011/1/28

競合事物間における比較関係認識

山崎義隆[†1]　乾健太郎[†2]　松本裕治[†1]

人々が，ある商品やサービスについてウェブ上で評判情報を探す場合，そのものに関する情報を収集するだけではなく，たいてい，競合する商品やサービスについての評判情報も収集する．この時，当初の目的の事物とそれと競合する事物を比較した意見や感想は，非常に有益である．なぜなら，このような競合事物の比較文から，目的の事物の優位点や欠点，相違点等を明示的に知ることができるからである．本稿では，ユーザーが入力した競合する 2 つの事物に対して，それらを明示的に比較した比較文をテキスト集合から抽出し，その文において競合事物対がどのような比較関係 (有差，同等，最上級，特徴) にあるか判定する手法を提案する．提案手法では，日本語の比較表現を観察することにより人手で構築した規則集合と，教師あり機械学習手法を併用する．9 組の競合事物対を用いて提案手法の評価実験を行い，その有効性を確かめた．

Identifying and Classifying Comparative Sentences in Japanese

Yoshitaka Yamasaki,[†1] Kentaro Inui[†2,†1]
and Yuji Matsumoto [†1]

When Web users search for opinions and reviews about a product or a service, they usually collect opinions and reviews about competitors for it as well. In such a case, it is very helpful for the users to show sentences where the product or the service is compared to its competitors, because a comparative sentence clearly expresses an advantage or a disadvantage of the former over the latter, or a difference between the former and the latter. In this paper, we propose a method of identifying comparative sentences from Japanese texts and classifying these sentences into the following four classes: *Equative*, *Non-equal gradable*, *Non gradable* and *Superlative*. By observing examples of comparative sentences in Japanese, we have manually constructed rules for identifying comparative sentences and non-comparative sentences. Our proposed system uses a machine learning method in addition to these rules. Experimental evaluation using nine pairs of competing products or services shows that the system has relatively high performance.

1. はじめに

近年，人々がウェブ上に意見や感想を投稿する機会が増え，ウェブ上に情報が溢れている．ブログや口コミサイトだけでなく，Twitter[⋆1]等，人々の意見や感想が投稿される場が増え，ウェブ上の情報は益々増加する一方である．商品やサービス等の事物に対する意見や感想等は**評判情報**と呼ばれる．このような評判情報は，事物の製造者や事物に関心のある潜在的な消費者にとって有益である．例えば，製造者は，製造・発売した商品やサービスが消費者にどのように評価されているかを知ることができる．一方，潜在的な消費者は，関心のある商品やサービスについて，実際に利用した消費者がどのような意見や感想を持っているかを知ることができる．一般に，人々がある事物についてウェブ上で評判情報を閲覧する場合，そのものに関する情報を収集するだけではなく，たいてい，競合する他の事物についての評判情報も収集し，それらを比較閲覧することで各事物の特徴や価値を認識している．その中でも，関心のある事物とその競合事物を比較している文は最も有益である．なぜならば，比較文は複数の事物間の相対的な価値や特徴を表現するからである．関心のある事物を含む比較文を収集することにより，他の事物との優位点や相違点を知ることができる．

そこで本稿では，評判情報のうち**比較文**に着目した．近年，情報抽出の分野において，評判情報からの意見抽出に関する研究が盛んに行われている．概して，意見は単一の事物に対する主観的な表現であるのに対して，比較文は複数の事物に対する主観的もしくは客観的な表現である．例えば，次の例文 (1) は意見であるが，例文 (2) は比較文である．

(1)　iPod nano はかなり可愛い．
(2)　iPod touch は iPod nano よりサイズが 5cm 大きい．

この例でも見られるように，比較文は特徴的な言語表現や構造を持っている．例えば，この文は次のような構造を持っている．

＜対象＞は＜基準＞より＜属性＞が 5cm 大きい．

ここで，＜対象＞=「iPod touch」，＜基準＞=「iPod nano」，＜属性＞=「サイズ」であ

†1 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
†2 東北大学
Tohoku University
⋆1 http://www.twitter.com

Vol.2011-IFAT-101 No.5
Vol.2011-NL-200 No.5
2011/1/28

る．本稿では，＜対象＞と＜基準＞を合わせて**競合事物対**と呼ぶ．Jindal ら[2] は，比較文に対して次の 4 つの比較関係ラベルを定義し，比較文を分類した．

有差，同等，最上級，特徴

例えば，上の例文 (2) は有差の比較文である．我々もこの比較関係ラベルを利用する．これらの比較関係については 3.2 節で詳しく述べる．

本稿では，与えられた競合事物対に対して，それらを明示的に比較した比較文をテキスト集合から抽出し，その文において競合事物対がどのような比較関係にあるか判定する手法を提案する．競合事物対が一文に含まれる場合，その文は比較文である可能性が高いと考えられる．そこで，比較文抽出の第 1 ステップとして，テキストから競合事物対を含む文をすべて抽出する．比較文候補抽出後,「より」や「ほうが」等の比較表現に基づき，その比較文候補が比較文であるかどうかと，競合事物間の比較関係を判定する．

本稿は次のように構成される．まず，2 章で関連研究について述べる．次に，3 章で比較に関する定義と 4 つの比較関係について説明し，英語と日本語における比較文の特徴を述べる．4 章で，競合事物間における，人手規則と教師あり学習を併用した比較関係認識手法について述べ，続く 5 章でその手法を評価する．6 章で全体をまとめる．

2. 関 連 研 究

倉島ら[10] と佐藤ら[12] は，日本語の比較関係の有差のみに着目し，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．文集合から比較文を抽出する過程において，我々の研究は，比較関係の有差，同等，特徴，最上級を対象にしている点が彼らの研究と異なる．

Jindal ら[2] 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．Yang ら[8] は，まず，人手で作成した韓国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する．このとき，抽出精度が高い手がかり語句から得られた比較文候補は無条件に比較文と見なす．そうでない語句から得られた比較文候補には，手がかり語句の前後の単語を素性として用いる比較文分類器を適用し，比較文候補が比較文であるかどうか判定している．4 章で述べる我々の手法は Yang らの手法と類似しているが，必ず競合事物対を含む文を抽出している点と比較関係の認識も行っている点が異なる．

Xu ら[7] は，機械学習手法を用いて比較関係の有差と最上級を認識している．Jindal ら[3] は，主に人手で作成した Label Sequential Rules を用いて，比較関係の有差，同等，最上級を認識している．我々はこれらに加えて比較関係の特徴も対象にしている点が異なる．

Ganapathibhotla ら[1] は，ウェブ上の評価極性付きユーザーレビュー内の頻度を用いて，レビュー者がより好んでいる事物を判定する．我々の研究は，競合事物間の詳細な比較関係認識が目的である．

Li ら[6] は，ユーザーレビューに含まれる比較の質問文から，ブートストラップ法を用いて頻繁に比較される事物対を取得している．我々の研究では，競合事物対は利用者の入力により与えられるため，目的が異なる．

3. 比較に関する定義と比較関係

3.1 比較文の定義

比較とは，競合事物対 (＜対象＞と＜基準＞) が持つ共通の属性に関して類似や相違に基づいた関係を示すことである[2]．本稿で取り扱う比較文は，以下の 2 つの条件を満たす．

(i) その中に競合事物対を含む．

(ii) 競合事物対に共通する属性において，優劣や類似，相違の関係が明示的に汲み取れる．

比較文と比較文でない文 (以下，本稿では非比較文と呼ぶ) の例を示す．

(1) iPod touch よりも iPod nano の方が見た目が可愛い

(2) iPod touch と iPod nano の音質は同じくらい良い

(3) iPod touch と iPod nano の音質は良い

例文 (1) は，競合事物対である「iPod nano ＜対象＞」と「iPod touch ＜基準＞」を含む．ゆえに，上の条件 (i) を満たす．また，共通の属性である「外見」に関する優劣の関係を明示的に汲み取ることができるので，上の条件 (ii) を満たす．以上，2 つの条件を満たすので，例文 (1) は比較文である．例文 (2) は競合事物対を含み，かつ，共通の属性である「音質」に関して明示的に類似の関係を示しているので，比較文である．

その一方で，例文 (3) は比較文ではない．例文 (3) は条件 (i) を満たすが，条件 (ii) は満たしていない．なぜならば，例文 (3) では，競合事物対を直接的に比較しておらず，文外の他の事物と比較して「音質が良い」という評価をしている可能性があるので，直接的な関係が明示的に汲み取れないからである．

3.2 比 較 関 係

Jindal ら[2] は，比較文に対して次の 4 つの比較関係ラベルを定義し，比較文を分類した．

有差　ある共通の属性に関して競合事物間に 有意な順序差がある

Vol.2011-IFAT-101 No.5
Vol.2011-NL-200 No.5
2011/1/28

同等　ある共通の属性に関して競合事物対が 等しい
最上級　ある共通の属性に関して 一方の競合事物が最も高い順序にある
特徴　ある共通の属性に関して競合事物間に 差はあるが，順序差はない

有差の比較文の例を以下に示す．有差の比較文は「より」や「上回る」のような表現によって特徴付けられる．

(4) iPod touch よりも iPod nano の方が可愛い．
(5) iPod touch が iPod nano の売れ行きを上回った．

同等の比較文の例を以下に示す．同等の比較文は「同じくらい」や「並ぶ」のような表現によって特徴付けられる．

(6) iPod touch と iPod nano の音質は同じくらい良い．
(7) iPod touch と iPod nano の売上が並んだ．

最上級の比較文の例を以下に示す．最上級の比較文は「1 番」や「優勝」のような表現によって特徴付けられる．

(8) iPod touch と iPod nano と iPod mini の中では，iPod touch が 1 番売れている．
(9) 巨人が阪神に勝ち，優勝した．

特徴の比較文の例を以下に示す．特徴の比較文は「のに対し」や「けど」のような表現によって特徴付けられる．

(10)金閣寺が金箔を張った建物であるのに対し，銀閣寺には銀箔を張った痕跡はない．
(11)iPod touch には，カメラが付いてるけど，iPod nano には付いてない．

最初の 3 つの有差，同等，最上級は，競合事物間で順序関係を示すことができるが，最後の特徴だけは，競合事物間で順序関係を示さない．

本稿では，これらの比較関係を採用する．

3.3 英語と日本語における比較文の特徴

英語の形容詞や副詞には比較級と最上級が存在するため，英語において比較文候補を抽出することは容易である．一方，日本語には形容詞や副詞の比較級・最上級が存在しないため，日本語において比較文候補を網羅的に抽出することは困難である.「より」や「ほうが」や「上回る」など，比較に特有の表現が存在するため，日本語でも比較文候補を抽出することはできるが，その再現率は英語に比べてかなり低いと思われる．

そこで，我々は，日本語において，手がかり語句を用いてどの程度比較文を収集できるのかについて予備実験を行った．まず，**表 1** に示す 9 組の競合事物対のそれぞれをクエリーとして，ブログからこれらの対を含む文を抽出した．このような文は 1,950 文得られ，これを 600 文の規則開発用データと 1,350 文の評価用データに分割した．次に，これらの各文に対して，その文が比較文であるかどうか人手で判断した．その結果を**表 2** に示す．この予備実験では，この規則開発用データを用いた．

表 1　競合事物対リスト

セブンイレブン	ローソン
金閣寺	銀閣寺
日産自動車	トヨタ自動車
マクドナルド	ロッテリア
自民党	民主党
ドコモ	ソフトバンク
液晶	プラズマ
任天堂	ソニー
巨人	阪神

表 2　収集した比較文と非比較文のデータ

	比較文	非比較文	計
規則開発用	250	350	600
評価用	542	808	1,350
計	792	1,158	1,950

ならぶ，負け，対照，衰退，倍，変更，低迷，最大，オススメ，切り替え，おすすめ，追い抜く，抜く，ぬく，凌駕，引き離す，破る，大勝，大敗，圧勝，惨敗，勝ち，同様，同率，番，優位，有利，優勢，以下，最も，逆転，初めて，差，異なる，に対し，筆頭，勝てる，互角，似る，追い越す，並ぶ，上回る，移る，勝利，位，首位，ほう，優勝，近い，より，方，比べる，〜対し…，〜対して…，〜一方…，〜と…の差，〜と… の違い，〜をやめて…に，〜と変わらない

図 1　日本語の手がかり語句の例

次に，Jindal ら[2] が英語の比較文を網羅的に収集するために利用した手がかりを調査した．そして，上記の規則開発用データを観察し，Jindal らの手がかりに対応する日本語表現を人手で収集した．収集した語句リストを日本語 WordNet[⋆1]の synset を用いて拡張した．最終的に，160 の語句が得られた．得られた語句の一部を**図 1** に示す．この予備実験では，規則開発用データからこれらの語句を 1 つ以上含む文を抽出し，それを比較文であるとみなす．

⋆1 http://nlpwww.nict.go.jp/wn-ja/

Vol.2011-IFAT-101 No.5
Vol.2011-NL-200 No.5
2011/1/28

表 3 手がかり語句による比較文抽出の評価結果 [英語は Jindal ら[2] からの引用]

言語	再現率	精度	全手がかり語句数	データに含まれていた手がかり語句数
日本語	0.61	0.73	160	88
英語	0.94	0.32	83	N/A

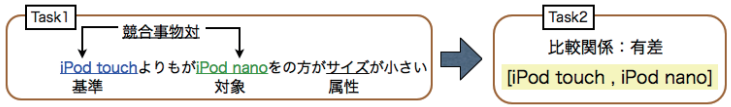

図 2 比較関係認識処理 [本研究は Task2 のみを対象とする]

我々の日本語における実験結果と，Jindal ら[2] から引用した英語における実験結果を**表 3** に示す．英語において再現率は 0.94 と非常に高い値であり，この主な要因は比較級と最上級が存在するためであると考えられる．一方，日本語において，精度は 0.73 と英語に比べかなり高い値であるが，再現率は 0.61 と相対的に低い値であった．この再現率を上げるためには，手がかり語句だけに頼らない比較文抽出法が必要である．

4. 規則と教師あり機械学習を併用した比較関係認識手法

本稿では，比較文に含まれる競合事物間の比較関係を認識することを目的とする．この時，文内における競合事物対が特定されていることを仮定する．つまり，**図 2** において，Task1 は完了しているものとし，Task2 のみを行う．

提案手法の処理手順は次のとおりである．

(1) ユーザーの入力により与えられた競合事物対を含む比較文候補をテキスト集合から抽出する．
(2) 比較文抽出規則を用いて，比較文候補から比較文を抽出する．
(3) 非比較文棄却規則を用いて，比較文候補から非比較文を除去する．
(4) 2 種類の規則に適合しない比較文候補を比較文分類器に入力し，比較文かどうか判定する．

以下，下 3 つについて詳しく述べる．

4.1 比較文抽出規則の適用

比較文抽出規則は，日本語において典型的な比較を表現する単語やフレーズであり，この規則に適合した文は比較文と見なす．比較抽出規則は，日本語の比較表現を示す手がかり語句の部分集合であり，比較文を抽出する際に精度の低い手がかり語句をフィルタリングしたものである．今回は，フィルタリングの閾値は精度 80%とした．図 1 比較文抽出規則の例で示した通り，典型的な比較表現である「より」「方が」と事物間の順序関係を示す「上回る」だけでなく事物間の差異を表す「〜に対し…は」等の複雑なフレーズも含まれている．本稿では，比較候補文が以下のいずれかの条件を満たした場合，比較文抽出規則に適合したとする．

(1) 比較文抽出規則が単語の場合，係り受け解析[4] の結果より，1 つ以上の事物が，比較抽出規則に係っている．
(2) 比較文抽出規則が単語の場合，係り受け解析[4] の結果より，比較文抽出規則が 1 つ以上の事物に係っている．
(3) 比較文抽出規則がフレーズの場合，比較文候補が正規表現を満たす．

比較文抽出規則を用いて，競合事物対を含む比較文候補から比較文を抽出する予備実験の結果を 7 に示す．**表 7** から比較文抽出規則を用いた比較文抽出の精度は 87%だが，再現率は 48%と精度に比べて低いという結果が得られた．この理由として，精度の高い手がかり語句の部分集合を比較文抽出規則と利用したためだと考えられる．再現率が低いため，比較文抽出規則に適合しない比較文を抽出し，精度を保ちつつ再現率を上昇させる必要がある．

また，**表 4** から比較抽出規則では，比較関係の有差，同等，最上級と比較して，特徴は抽出困難であることが示されている．なぜなら，有差，同等，最上級は事物間の順序関係を示す．例えば,「より」,「同じ」,「上回る」等の規則は，他の事物間との明示的な順序関係を示す．しかし特徴は，競合事物間の明示的な順序関係を示さないため，日本語の典型的な比較表現である比較文抽出規則では抽出が困難だと考えられる．

比較文抽出規則に適合しない比較文を抽出する場合，比較の表現は多種多様であるため，比較文抽出規則を補充し続けていくのは困難である．そこで本稿では教師あり学習を利用し，ラベル付き学習データから比較の手がかりを機械的に取得する．ここで我々は，機械学習の問題を簡単にするために，学習データに含まれる明らかな負例を除去することを考えた．競合事物対の並列表現を含む比較文候補は，非比較文である可能性が高いという知見を得た．

以下は，比較抽出規則に適合したが，非比較文の例である．

- 特徴としては液晶、プラズマテレビよりも価格が安いことがあげられる．
- ドコモとソフトバンクの方が可愛い．

比較抽出規則には適合するが，直接的に競合事物対を比較していないため，比較文ではない．

Vol.2011-IFAT-101 No.5
Vol.2011-NL-200 No.5
2011/1/28

表 4 比較文抽出規則を用いた各比較関係の抽出割合

比較関係	再現率
有差	0.60 %
同等	0.49%
最上級	0.65%
特徴	0.30 %

表 5 各比較関係をラベル付けした比較文抽出規則の例

比較文抽出規則	比較関係
より	有差
同じ	同等
上回る	有差
同じ	同等
最も	最上級
差	特徴
違い	特徴

本稿では，比較文抽出規則により抽出された比較文のみ，比較関係の有差，同等，最上級，特徴のいずれかに属するかを決定する．**表 5** に各比較関係をラベル付けした比較文抽出規則を示す．

4.2 非比較文棄却規則の適用

競合事物対の並列表現を含む比較文候補は非比較文である可能性が高いという知見から，非比較棄却規則を作成した．非比較文棄却規則は，日本語において並列を表す典型的な表現であり，この規則に適合した文は非比較文と見なす．並列表現を含む比較文候補は，文外の事物と比較している可能性があるため，競合事物間の明示的な関係を示す可能性が低い．そのため，3.1 章の定義から，並列表現が含まれる比較候補文は非比較文とする．非比較文棄却規則の作成方方として，まず競合事物対を含む文集合である規則開発用データ（5 章で説明）を観察し，並列表現を示唆する単語とフレーズを人手で抽出する．次に，非比較文を抽出する際に精度の低い規則を人手でフィルタリングする．今回は，フィルタリングの精度を 80%とした．**表 6** に非比較文抽出機足の例を示す通り，並列助詞である「と」や「や」だけでなく，比較候補文が以下のいずれかの条件を満たした場合，非比較文棄却規則に適合したとする．

(1) 非比較文棄却規則が単語の場合，非比較文棄却規則の前後に競合事物対が含まれている．

表 6 非比較文棄却規則の例

とか，やら，や，か,,，と，も，〜でもなく…でもない， 〜にしても…にしても，〜であれ…であれ，〜だろうが…だろうが，〜にも…にも

表 7 各規則の評価

規則 / 人手	比較文抽出規則	非比較文棄却規則	その他	合計
比較文	263	72	207	542
非比較文	39	572	197	808
合計	302	644	404	1350

(2) 非比較文棄却規則がフレーズの場合，比較文候補が正規表現を満たす．

今回は比較候補文において，非比較文棄却規則の前後 3 単語に競合事物対が含まれている場合，非比較文と見なす．表 7 から非比較文棄却規則を用いて非比較文を精度は 83%と再現率は 74%と抽出できていることが分かる．この理由として，競合事物対とは頻繁に比較されるであろう名詞対であるため，対等な関係である可能性が高い．そして並列助詞は，名詞と名詞を対等な関係で結びつけるため，競合事物対を含む並列表現の文を精度よく分類できたと考える．

以下は非比較文棄却規則に適合したが，比較文である例である．

- マイクロソフトから始まり、任天堂、ソニーという順番だ
- ドコモとソフトバンク迷ったけど、ドコモに決めた．

非比較文棄却規則である並列助詞「と」に適合しているが，比較を示唆する単語である「順番」や「決めた」を用いて競合事物間において直接的に比較している．

4.3 教師あり機械学習に基づく比較文抽出

比較文抽出規則の比較文を抽出する際の再現率が低いため，機械学習の目的は比較抽出規則と非比較抽出規則に適合しない比較文を抽出することである．特に比較文抽出規則では抽出が困難である比較関係が特徴である比較文を抽出する必要がある．本稿では，BACT[5] と呼ばれる木構造をマイニングするアルゴリズムを利用する．BACT は，木構造のデータを入力とし，学習によって判別に効果のある構造を規則として学習できる．

機械学習における学習データと学習に利用した素性について述べる．機械学習では，比較文抽出規則と非比較文棄却規則に適合しない比較文候補を分類することを目的としている．比較文であるかどうかを決定する手がかりは，比較関係の有差，同等，最上級の比較の場

Vol.2011-IFAT-101 No.5
Vol.2011-NL-200 No.5
2011/1/28

表 8 評価用データにおける比較関係

比較関係	文数
有差	269
同等	31
最上級	43
特徴	199
計	542

合，競合事物間の順序関係を示す表現であるため,「上回る」や「越える」等の比較を示唆する単語である．他方，比較関係の特徴の場合，競合事物間の差異を表現する「A は〜に対し，B は…」の様に，文の構造が手がかりとなる．

我々は，競合事物対の周辺の単語は並列助詞の出現や競合事物を形容する単語が出現するため，比較の手がかりとなる単語や構造が出現する可能性が高いと考えた．そのため，競合事物対の前後の単語を素性として学習に利用する．以下に，素性の作成手順を示す．

(1) 比較文候補に形態素解析[11)]

(2) 評価表現辞書[9)] に収録されている語をラベルに変換

(3) 競合事物対の前後の単語を素性として抽出

単語を品詞に置換することにより，素性を汎化させる，今回は，最も性能が高かった競合事物対の前後 4 単語を利用する．

5. 評 価 実 験

5.1 実験データ

実験データは，3.3 節で述べた評価用データを用いた．この評価用データの各文に対して比較関係を人手で付与した．その結果を**表 8** に示す．

5.2 実 験 内 容

(1) 2 種類の規則と教師あり学習を用いて競合事物対を含む比較文をどの程度収集できるのかを示す．

(2) 比較文抽出規則を用いて，競合事物間の比較関係をどの程度認識できるのかを示す．

実験では，9 組の競合事物対のうち，8 組で学習，残りの 1 組で評価を行った．これを 9 回実行した．

表 9 に実験内容 1 の結果を示す．**表 10** に実験内容 1 の結果を示す．

表 9 各手法における比較実験

比較文抽出規則	○	×	×	○	○	○
非比較文棄却規則	×	○	×	○	×	○
機械学習	×	×	○	×	○	○
精度	0.87	0.67	0.53	0.67	0.65	0.70
再現率	0.49	0.78	0.86	0.87	0.79	0.80
F 値	0.62	0.72	0.65	**0.75**	0.71	**0.75**

表 10 提案手法における各比較関係の抽出割合

比較関係	取得割合
有差	77%
同等	61%
最上級	77%
特徴	64%

5.3 比較文抽出の考察

表 9 から考察する．比較文抽出規則と機械学習を組み合わせた手法より，提案手法である非比較文棄却規則を含めた手法の方が F 値が高いことが示された．これは非比較文抽出規則が有効であることを示している．

2 種類の規則を用いた手法と提案手法の F 値が等しいということは，機械学習を用いて規則に適合しない文集合から比較文を同定することが困難であることを示している．その理由は，今回の学習データが少ないからだと考えられる．学習器が獲得した素性を観察したところ，比較の手がかりとなりえる素性をうまく抽出できていなかった．学習データを増やすことでこれに対応できる可能性がある．

5.4 比較文抽出規則に基づいた比較関係認識

システムの精度は 76%，再現率は 48%であった．各比較関係を人手でラベル付けした比較文抽出規則が持つラベルを適合した比較文候補に割り当てる．典型的な比較表現を比較文抽出規則に利用しているため，精度は 76%と高い．しかし，精度の低い規則はフィルタリングをしているため，再現率は，48%と低い結果である．

5.5 比較関係認識の考察

比較候補文が複数の比較関係を持つ場合がある．例えば，

- セリーグでは巨人も阪神も勝利し，同率首位をキープしてますね

の様に比較関係が最上級の比較文は，有差の比較関係にも属することが多い．なぜなら，競

Vol.2011-IFAT-101 No.5
Vol.2011-NL-200 No.5
2011/1/28

合事物対を含む文を対象にしているため，比較関係の最上級は有差を含意しているからである．

6. おわりに

本稿では，一般に比較されるであろう競合事物対を含む文集合から比較文を抽出し，競合事物間の比較関係を認識する手法を提案した．比較文抽出規則と非比較文棄却規則の 2 種類の規則を利用することで，比較文を精度よく抽出できることを示した．しかし，規則に適合しない比較文候補を機械学習を用いて分類することを試みた結果，効率的に比較文を抽出することが困難であると示された．

今後の課題として，規則と機械学習の最適化，規則の補充，機械学習に用いる学習データの収集，最適な素性選択をする必要がある．また，競合事物間の比較関係認識において，規則だけでは再現率が低いため機械学習と組み合わせた比較関係認識を試みる必要がある．

謝辞 本研究を遂行するにあたり、奈良先端科学技術大学院大学の松吉俊特任助教に多大な助力を頂いた．

参 考 文 献

1) Murthy Ganapathibhotla and Bing Liu. Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 241–248, 2008.
2) Nitin Jindal and Bing Liu. Identifying comparative sentences in text documents. In *Proceedings of SIGIR*, pp. 244–251, 2006.
3) Nitin Jindal and Bing Liu. Mining comparative sentences and relation. In *American Association for Artificial Inteligence 2006*, 2006.
4) Taku Kudo and Yuji Matsumoto. Japanese dependency analysis using cascaded chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pp. 63–69, 2002.
5) Taku Kudo and Yuji Matsumoto. A boosting algorithm for classification of semi-structured text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP2004)*, 2004.
6) Shasha Li, Chin-Yew Lin, Young-In Song, and Zhoujun Li. Comparable entity mining from comparable questions. In *Proceedings of the Joint Conference of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 650–658, 2010.
7) Kaiquan Xu, Stephen Shaoyi Liao, Raymond Y.K. Lau, Heng Tang, and Shanshan Wang. Building comparative product relation maps by mining consumer opinions on the web. In *AMCIS 2009*, 2009.
8) Seon Yang and Youngjoong Ko. Extracting comparative sentences from korean text documents using comparative lexical patterns and machine learning techniques. In *ACL-IJCNLP 2009*, pp. 153–156, 2009.
9) 小林のぞみ, 乾健太郎, 松本裕治. 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 2006-NL-171, pp. 111–118, 2006.
10) 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. In *DEWS 2007-L1-5*, 2007.
11) 工藤拓. 形態素解析器 mecab. `http://chasen.org/~taku/software/mecab/`, 2005.
12) 佐藤敏紀, 奥村学. blog からの比較関係抽出. 情報処理学会研究報告 2007-NL-181, pp. 7–14, 2007.