2007－NL－181 (2)
2007／9／25

blogからの比較関係抽出

佐藤 敏紀† 奥村 学††

† 東京工業大学 大学院 総合理工学研究科 〒226–8503 神奈川県横浜市緑区長津田町 4259–R2–7
†† 東京工業大学 精密工学研究所 〒226–8503 神奈川県横浜市緑区長津田町 4259–R2–7
E-mail: †satou@lr.pi.titech.ac.jp, ††oku@pi.titech.ac.jp

あらまし 本稿では日本語の比較表現に対する知見から得られたルールと，構文情報とセンタリング理論を用いることでblog記事中に含まれる比較表現から比較関係を抽出する手法を提案する．比較関係は〈対象，基準，属性，評価〉の4つ組，または属性が非明示な〈対象，基準，評価〉の3つ組で構成される．提案手法ではこれらの組を抽出する．比較関係は構成する要素が全て単文中に存在する場合と複数文にまたがって存在する場合がある．提案手法は両方の場合に対応する．実験により提案手法はblog記事中の単文または複数文にまたがる比較表現から比較関係の4つ組を高い精度で抽出できることがわかった．

キーワード 比較関係，比較表現，テキストマイニング，ブログ

Extraction of Comparative Relations from Japanese Weblog

Toshinori SATOU† and Manabu OKUMURA††

† Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
4259, Nagatsuta-cho, Midori-ku, Yokohama-shi, Kanagawa, 226–8503 Japan
†† Precision and Intelligence Laboratory, Tokyo Institute of Technology
4259, Nagatsuta-cho, Midori-ku, Yokohama-shi, Kanagawa, 226–8503 Japan
E-mail: †satou@lr.pi.titech.ac.jp, ††oku@pi.titech.ac.jp

Abstract In this paper, we propose a new method for extracting comparative relations from comparative expressions in Japanese Weblogs. A comparative relation is expressed with <object, criteria, attribute, evaluation>, or <object, criteria, evaluation> when the attribute is not explicitly shown. Our proposed method extracts relations of both types. We can observe the fact that all elements of a comparative relation are in a simple sentence or range over multiple sentences. Our proposed method can apply to both cases. Experimental results show that our proposed method can extract the comparative relation with high precision.

Key words Comparative Relation, Comparative Expression, Text Mining, Weblog

1. はじめに

近年，人々はblogやSNSを使って積極的に情報を発信している．それらの情報源は一般大衆による生の声として広く盛んに活用されている．リアルタイムに生まれ続ける大量のWeb文書から有益な情報を抽出する際には，情報の速報性や鮮度を失わないために，人手による情報抽出ではなく自動的に情報を抽出する手法を用いるほうが望ましい．

Web文書からの自動的な情報抽出は，意見抽出の分野で盛んに行われている．その中でも評判情報抽出はもっとも一般的なタスクだといえる．このタスクでは評判情報を〈対象，属性，評価〉の3つ組と，それに付与される極性で表わす場合がある．このような評判情報は非常に有益である．たとえば消費者が商品の購入を検討する際に集積された評判情報を参照することで，商品の価値や特徴を容易に確認することができる．しかし，一般に人々は単一の実体に関する評判情報を眺めるだけでなく，商品比較サイト等で複数の実体の評判情報を比較することで，各実体の価値や特徴を確認している．

人々がblogやSNSを使って発信する情報には図1に示した例のように，複数の実体間の共通の属性に着目し，

(A) 単文　東京より大阪は雰囲気が暖かい
(B) 複数文　大阪は良い．
なぜなら東京よりも雰囲気が暖かい

図 1　比較表現を含む文の例

図 2　比較表現を構成する要素

個人の主観に基づき相対的な評価をしている表現が含まれている．このような表現は一般に比較表現と呼ばれている．従来のタスクで扱われてきた評判情報が単一の実体を評価しているのに対し，比較表現は複数の実体の関係性を表している．比較表現は各実体の相対的な価値や特徴を知るうえで有益な情報だといえる．

ここで図 1(A) の単文に着目する．この文には比較表現を構成する要素である対象，基準，属性と評価が全て含まれている．この文に含まれる比較表現の要素を図 2 に示す．図 1(A) の文は，対象を示す実体「大阪」の属性「雰囲気」が，相対評価の基準となる実体「東京」の属性「雰囲気」よりも，「良い」と評価している，と言える．

本稿では日本語の比較表現に対する知見から得られたルールと，構文情報とセンタリング理論を用いることで，日本語の blog 記事中に含まれる比較表現から比較関係を抽出する手法を提案する．比較関係は〈対象，基準，属性，評価〉の 4 つ組，または属性が非明示な〈対象，基準，評価〉の 3 つ組で構成される．従来の日本語の比較関係抽出手法は〈対象，基準，評価〉の 3 つ組を抽出することに焦点を当てていたが，提案手法では両者を抽出する．比較表現には比較関係を構成する要素が全て単文中に存在する場合と，複数文にまたがって存在する場合がある．従来の日本語の比較関係抽出手法では，要素が全て単文中に存在する場合に着目していたが，提案手法は両方の場合に対応する．

提案手法では取得した比較関係が，評判情報を含んでいるかを判定できない．しかし，評判情報の有無の判定や，評判情報の極性の判定は，本手法を実世界で応用する際に必須であるため，今後取り組む．

実験により提案手法は blog 記事中の単文，または複数文にまたがる比較表現から，比較関係の 4 つ組を高い精度で抽出できることがわかった．

2. 関連研究

本研究で扱う比較関係抽出は意見文から情報を抽出するタスクである．このタスクのおもな課題は文書から評判情報として〈対象，属性，評価〉の 3 つ組を抽出し，その組の極性を判定することである．鈴木ら [1] は 3 つ組における評価の抽出と，その評価表現が肯定的か否定的かを判定する処理をブートストラッピング的に実現するために，教師付き学習手法を EM アルゴリズムで補強する semi-supervised な手法を提案した．Liu ら [2] は相関マイニングシステムを意見文からのルールのマイニングに

(1) 入力　東京より大阪は雰囲気が暖かい
(2) 出力　〈大阪，東京，雰囲気，暖かい〉

図 3　比較関係抽出処理の入出力の例

利用することで，3 つ組における属性を抽出する手法を提案した．これらの研究では単一の実体がもつ属性に対する評価を扱っていた．

一方で複数の実体がもつ共通な属性に対する相対的な評価を含む意見文に着目した研究が増えている．Jindal ら [3] はルールマイニングとナイーブベイズ分類器による分類手法を組み合わせて，評価文書中の比較文を同定する手法を提案した．さらに，Jindal ら [4] は比較文を SVM によって比較の型に基づき分類し，比較の型ごとに比較関係の要素を抽出するためのパターンを自動生成する手法を提案した．Jindal らは提案した手法により自動生成したルールを用いて 4 つ組の比較関係 (〈対象，基準，属性，評価〉) を抽出した．倉島ら [5] は人手で作成した比較関係抽出用のパターンと，比較関係における対象または基準となりうる実体のリストとの照合を組み合わせることで，Jindal らと同程度の精度で 3 つ組の比較関係 (〈対象，基準，評価〉) を抽出した．

本研究は比較関係抽出の処理に倉島ら [5] と同様，人手で作成した比較関係抽出用のパターンおよび，事前に作成した実体のリストを使用する．我々が提案する手法ではこれらの情報以外に，必要に応じて構文情報やセンタリング理論を用いることで，図 3(2) の例のような 4 つ組の比較関係を抽出する．また，本研究では図 1(B) の例のような複数文にまたがる比較関係の抽出を実現している．

3. 提案手法の概要

3.1 日本語の比較表現とその要素

本研究では森山 [6] と同じく，「比較は複数の実体を特定の共通する属性の程度から位置づける表現だ」と考える．人々は比較をおこなうことで，複数の実体の間に優劣や相違を見出すことができる．

日本語の比較表現を構成するおもな要素は以下の 4 種類である．比較表現には少なくとも以下の対象，基準と評価の 3 つの要素が含まれている．

- 対象：比較表現において評価される実体
- 基準：対象を評価する際の基準となる実体
- 属性：対象と基準がもつ共通の観点
- 評価：対象の属性を基準に基づき位置づける表現

比較関係抽出タスクは上記の比較表現の要素分のスロットを用意し，比較表現から抽出した要素で，それらのスロットを埋めるタスクと考えることができる．本研究では処理の結果，スロットを埋めた要素の組のうち少なくとも対象，基準と評価の 3 要素が含まれている組を比較関係として扱う．人間が比較関係を見出せる日本語の比較表現の型や構文は多岐に渡るものの，型については大きく以下の 3 種類に分けられる．

- 有差：複数の実体間の相対的な差を示す
  (例) 沖縄は北海道より平均気温が高い．

(1) PS3 は Wii より値段が高い.
(2) Wii と比較すると PS3 は値段が高い.
(3) PS3 と Wii では PS3 の方が値段は高い.
(4) PS3 の方が Wii より値段は高い.
(5) PS3 と Wii では PS3 の方が高い.
(6) PS3 の値段は高いが Wii の値段は安い.

図 4 日本語の有差比較表現の構文例

対象の比較特有表現	[(と∣に)/助詞]+[比べ/動詞]
	[(と∣に)/助詞]+[比較/名詞]
基準の比較特有表現	[より/助詞]
	[方/名詞]+[(は∣が)/助詞]
	[ほう/名詞]+[(は∣が)/助詞]

図 5 比較特有表現の例

- 同等, 同程度 : 複数の実体が同等・同程度なことを示す

(例) そばとうどんの値段は同じ位だ.

- 最上級 : ある実体に対する評価の絶対値が複数の実体間において相対的に最大であることを示す

(例) 富士山は国内の山の中で一番標高が高い.

本研究では狭義の比較として 2 実体間の有差比較を扱う.

3.2 日本語の有差比較の構文

日本語には英語のように, 比較級と呼ばれる形容詞や副詞の形式がない. そのため日本語文では, 形容詞や副詞の表層に基づく比較関係抽出が難しい. 一方で日本語の有差比較表現を含む文を観察すると, 様々な有差比較表現の構文を見出せる.

図 4 に 2 実体間で共通な属性「値段」の評価が「高い」となる有差比較の例を示す. 日本語の有差比較で一番多く使われているといわれる構文は, 図 4 の (1) から (2) のように基準となる実体に対して, 基準からの分離を意味する表現「より」や「比べ」などを付与する型である.

日本語の有差比較表現には図 4(3) のように対象と基準の組を示す構文もある. この例では 2 つの実体の間に並立助詞をはさみ併記したうえで, 対象を表す実体に方向を指し示す意味をもつ表現「方が」を付与している. このような構文では併記された 2 実体のうち, どちらかを対象として指し示すことで, 基準となる実態が決まるため比較を表現できる. 比較において対象を表す実体に付与される表現「方が」などは, 基準からの分離を意味する表現「より」などと共起することができるため, 図 4(4) のようにも使うことができる. その場合には対象を表す実体の周囲に基準を表す実体が存在することで比較が成立する.

比較は複数実体間の共通な属性に着目しておこなわれるが, その属性が明示されない場合がある. 図 4(5) では 2 つの実体に共通な属性が何も表記されていない. そのため, この例文は比較が行われていないと考えることもできるが, このような場合には文の表層に属性を発見できないだけで, 実際には隠れた属性が存在していると考えるのが自然である. この例では評価の形容詞「高い」の用例や 2 実体に共通な属性を考えることで, 属性の範囲をある程度は絞ることができる. しかし実際には隠れた属性が「値段」か「描画性能」かその他なのかは分からない. また, 評価として抽出した形容詞節などから属性を特定できるとは限らない. 森山 [6] によると属性を表す表現が見つからなかった場合に, 事態の選択を表す形式, 話者の希望をあらわす形式, 命令文などの, 動きを策する比較表現の場合には「(話者にとっての) 適切性の量」が隠れた属性であることが多い. しかし, 提案手法では文の表層から比較関係の属性を抽出できなかった場合でも, 森山の知見などに基づく隠れた属性の推定はしない.

日本語の比較表現には図 4(6) のように各実体についての評価文を連結して複数の実体を比較する構文がある. このような構文で 2 実体間の有差比較を表現するには, 各評価文が等位接続されており, 接続された各評価文が文中の実体に共通の属性に着目しており, さらに一方の実体に対する評価がもう一方の実体に対する評価の反意か否定の意味をもつ必要がある. 図 4(6) は, これら 3 つの条件を満たしている. ここで新たな例文「PS3 は高いが Wii は安い. 」について考える. この例文は上述した 3 つの条件を満たしていないため, 2 実体間の有差比較表現になるとは言い切れない. なぜなら 2 つの実体と周囲の文脈に出現した実体とを比較している可能性や, 前後の評価文が異なる属性に着目している可能性があるからである. 4(6) のような構文から高い精度で比較関係を抽出するためには評価の形容詞節が互いに反義かの判定など, 本研究テーマの初期段階で十分な成果を得られないタスクに取り組む必要がある. そのため本稿では複数の評価文を等位接続して実体を比較する構文からの比較表現抽出を対象外としている.

日本語比較表現には他にも対象とすべき構文が有るが本研究では主に図 4(1) から (5) のような構文に着目した.

3.3 比較関係抽出

前節で述べたように日本語の比較表現では, 基準を表す実体に付与される表現と, 対象を表す実体に付与される表現が多く見られる. 本研究では便宜上, これらの日本語の比較表現に特有な表現を比較特有表現と呼ぶ. とくに前者を基準の比較特有表現と呼び, 後者を対象の比較特有表現と呼ぶ. 比較特有表現の例を図 5 に示す.

人々は日本語の文から文中の比較特有表現を目印に, 比較関係を抽出することができる. 例えば人は「新宿駅の方が八王子駅より乗降客が多い. 」という日本語文を見たときに, 文中の比較関係の各要素のうち対象は「新宿駅」, 基準は「八王子駅」, 属性は「乗降客」で評価は「多い」であると理解することができる. これは対象と基準を各実体に付与された比較特有表現から一意に決定できるうえに, それらの実体が共通に係っている評価と, 評価に係ってる属性も一意に決定できるからである.

本研究では比較特有表現が比較表現から比較関係を抽出する際の重要なヒントになると考え, 比較特有表現が使われている比較表現を観察した. そして, その観察結果から得た知見に基づき比較の対象や基準を抽出・補完するルールを作成した. そのルールの一部を図 6 に示す.

対象抽出　[実体/名詞][の/助詞][方/名詞][(は|が)/助詞]
　　[実体/名詞][の/助詞][ほう/名詞][(は|が)/助詞]
基準抽出　[実体/名詞][より/助詞]
　　[実体/名詞][(と|に)/助詞][比べ/動詞]
　　[実体/名詞][(と|に)/助詞][比較/名詞]

図 6　比較関係の対象や基準を抽出するルールの例

基準補完　[実体 A/名詞][と/助詞][実体 B/名詞]
実体補完　[実体/名詞][(は|って|が|に|を|の)/助詞]

図 7　比較関係の対象や基準を補完するルールの例

提案手法では最初に図 6 のような抽出ルールを使い, 比較の対象と基準の抽出を試みる. その際に比較関係を取得したい実体組 (固有名詞 A, B) の表層を当てはめたルールを原文に適用することで, 目標に適合する対象や基準を抽出する. しかし抽出ルールのみでは比較特有表現が隣接していない実体を抽出できない. そのような場合は評価 (形容詞節など) を抽出した後, 図 7 のような補完ルールを用いて評価に係る実体を抽出する. 構文情報に基づき実体を抽出できない場合は, センタリング理論により取得した文の主題を実体のスロットを補完する候補とする. その後, 評価に係る副詞や属性 (名詞節) を抽出する.

本研究で提案する比較関係抽出手法は, 比較関係のあるスロットを埋めることで, さらに別の要素の抽出を試行できるようになるボトムアップな手法だといえる. ボトムアップな手法では初期段階での精度低下は, 処理全体の精度を大きく下げる. そこで比較特有表現を含まないために対象や基準の抽出が困難な比較表現の抽出は, 本研究における処理で対象外としている. また複数の評価表現を等位接続して 2 つの実体を比較する型の比較表現も, 2 実体間の有差比較かを判定する手法の考案が課題となるため対象外とした. 比較関係の抽出のための詳しい手法は次節で述べる.

4. blog からの比較関係抽出

本節では我々が提案する比較表現抽出手法の詳細を述べる. 以後, 処理の際に着目する文をターゲット文と呼ぶ. また同文書におけるターゲット文の 1 文前の文を「直前の文」と呼ぶ. 図 8 に提案する比較関係抽出手法のフローチャートを示す. 本手法は大きく以下の 5 ステップで構成される.

（1） 前処理
（2） 実体の抽出ルールによる対象と基準の抽出
（3） 抽出した実体と構文情報に基づく評価の抽出
（4） 対象または基準が未抽出な場合の実体補完処理
（5） 評価の文節と構文情報に基づく副詞と属性の抽出

処理の最初に対象, 基準, 属性, 評価と副詞の 5 スロットを用意し, 処理の過程で文から抽出した要素は順次スロットに格納する. すべての処理が終わった時点で対象, 基準と評価のスロットが埋まっていれば, スロットに格納済みの要素から比較関係を構築し出力する.

(1) 直前の文の文頭からターゲット文の文末までに両実体を含む
(2) ターゲット文中では最低一方の実体が比較特有表現と隣接する
(3) 実体間に並列を表す助詞・記号を挟んだ文字列を含まない
(4) 実体組を単に連結した文字列を含まない（例, 実体 A 実体 B）
(5) 1 文中で両実体が同時に基準の比較特有表現と隣接しない
(6) 1 文中で両実体が同時に対象の比較特有表現と隣接しない

図 9　実体組に関する比較表現を含む文に関する仮定

4.1 前 処 理

本手法では, 抽出処理の最初に図 6 のような対象や基準の抽出ルールを用いて, ターゲット文から実体を抽出する. 前処理ではその準備として以下の処理を順次おこなう.

（1） 比較したい固有名詞をリスト化
（2） 固有名詞リストから 2 つの固有名詞の非順序な組をすべて作成
（3） 未処理の固有名詞組を選択
（4） 選択した固有名詞組の各固有名詞から blog 検索用のクエリを作成
（5） blog 検索の検索結果中の URL に基づき, blog 記事を収集し本文を抽出, 本文を正規化し文単位で分割
（6） 以後の処理に適さない文を取り除き文集合を作成

複数の実体間にある比較関係を網羅的に取得するには, 実体のリストからすべての非順序な実体組を作成し逐次処理する必要がある. 本研究で対象とする実体は固有名詞である. そこで人手で用意した固有名詞リストから, 固有名詞の非順序な組をすべて作成する. 以下では, リストから, ある固有名詞の組（固有名詞 A, B）を選択して処理を続けたものとする

次に固有名詞に関する比較表現を含んでいることを期待できる blog 記事を取得する. そのために固有名詞組と比較特有表現などを組み合わせたパターンからクエリを作成する. このパターンは事前の実験に基づき作成した. 実験では比較特有表現を含むパターンと固有名詞から文字列を作成し, その文字列の blog 検索エンジンにおけるヒット数を調べた. その結果から様々な固有名詞と組み合わせた際にヒット数が多かったパターンを選んだ. 以下に, ある固有名詞組を選択した時に, その組の各固有名詞 (NE) から作る blog 検索用のクエリの一部を列挙する.

- NE より, NE の方が, NE のほうは, NE と比べ, NE に比較, より NE, よりも NE, 比べ NE, 比べて NE

本手法では 2 実体間の有差比較表現を含む文は, 図 9 に示した条件を全て満たすと仮定している. これらの仮定は処理対象とする 2 実体間有差比較表現の構文を含まない文を見つけることを狙っている. 作成したクエリを順次使用し, さらにいくつかの処理を経て文集合を取得した後, 文集合から図 9 の条件をすべて満たさなかった文は文集合から排除する.

前処理の結果, 図 1 のような, 文集合中には比較表現を含む単文か複数文のみが残る.

4.2 実体の抽出ルールによる対象と基準の抽出

前処理で作成した文集合から比較関係の基準と対象を

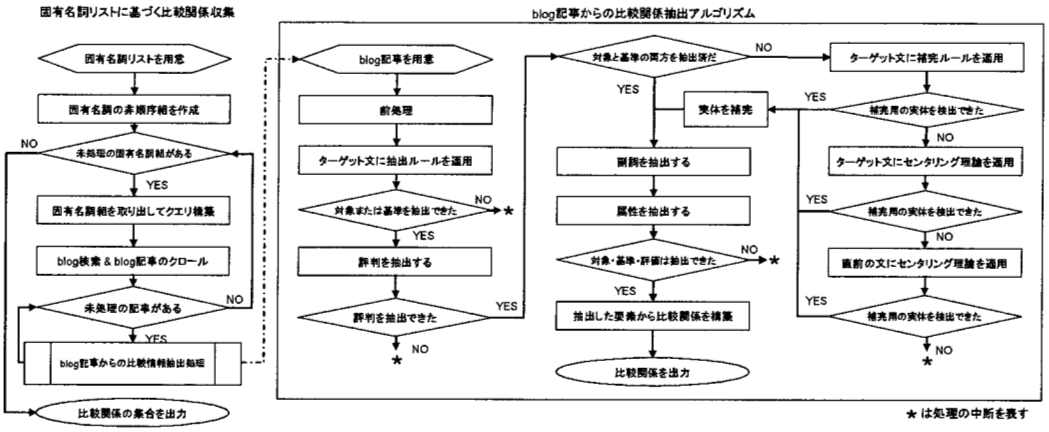

図 8　提案する比較関係抽出手法のフローチャート

抽出する．はじめに文集合からターゲット文を取得する．また，文集合に直前の文が含まれていた場合には，それも取得する．文集合から取得した文はすべて形態素解析器 MeCab(注1) と係り受け解析器 CaboCha(注2) で処理し，構文情報を取得する．

次に図 6 のような対象・基準の抽出ルールをターゲット文の構文情報に適用し対象と基準を抽出する．ルールと組み合わせる固有名詞組は前処理で使用した固有名詞組と同じである．ターゲット文には blog 検索用のクエリ文字列が何かしら含まれていることを期待できるので，抽出ルールによって少なくとも対象か属性のうち一方の実体を抽出できると期待できる．

4.3　抽出した実体と構文情報に基づく評価の抽出

前節の処理により，ターゲット文から対象か基準となる実体を抽出できたら，抽出した実体とターゲット文の構文情報を用いて，評価を含む文節を抽出する．

ここで比較表現の要素の出現順序に関する知見を述べる．山本 [7] は「比較表現において基準を評価の前に置く言語は OV 語順をとる言語である，という普遍的な傾向がある」と述べている．日本語は OV 語順をとる言語である．確かに本研究の作業中には評価が基準に係る比較表現の構文は見られず，一方で対象と評価の順序には普遍的な傾向を確認できなかった．

以上の知見をふまえ本手法では，比較の実体は両方とも評価に係ると仮定した．本手法の評価抽出では，対象か基準のうち一方が抽出できた場合に，構文情報に基づき抽出済みの実体を含む文節が係る文節を文末に向けて順に辿る．さらに基点となった実体と文節間の距離が一番近い形容詞節を取得する．文末まで文節を辿っても形容詞節が存在しない場合には，一番近い動詞節を取得し，動詞節も存在しない場合には実体と一番近い「(だ｜です)/助動詞」を含む名詞節を取得する．そして，取得できた文節を評価として抽出する．対象と基準の両方を抽出できている場合には，はじめに対象と基準について同時に上述した評価の探索をおこなう．その結果から，対象と基準が共通で係る両実体と一番文節間の距離が近い評価となりうる文節を抽出する．

- 基準が抽出済みな場合，対象と，評価の傾向「選択」を補完
  - [実体/名詞][派/名詞]
  - [抽出済な実体][より/助詞][実体/名詞][(だ｜です) 助詞]
- 基準が抽出済みな場合，対象を補完
  - 文末が [実体/名詞] で体言止め

図 10　blog 記事中の比較表現に対応するための補完ルール

4.4　対象または基準が未抽出な場合の実体補完処理

評価を抽出した後，対象か基準の一方を取得できている場合は，取得できていない実体のスロットを補完する．

処理には，抽出した評価の文節とターゲット文の構文情報を用いる．ターゲット文の構文情報中の評価の文節を基点にし，評価の文節に係る文節を文頭に向けて順に遡る．そして固有名詞組のうち対象か基準に割り当てられていない固有名詞を含む文節を探す．その結果，見つかった文節に対して図 7 のような補完ルールを適用する．

図 7 のような補完ルールに適合した文節のうち，評価の文節との文節間距離が一番近い文節を埋まっていない実体のスロットを補完する文節として扱う．

本手法では blog 記事に頻出する比較表現の構文に対応するために，図 10 に挙げる例のような補完ルールも使用した．blog 記事には属性だけでなく，評価も明示されていない比較表現の構文が頻出するので，そのような構文に対応し必要に応じて評価を補完することは必要である．そこで本手法では，4.3 節の処理で評価が取得できなかったターゲット文についても，図 10 のような実体以外に評価も補完するルールを適用し，補完できれば以後の処理の流れに戻している．評価も補完できる補完ルールでは，属性が明示されにくい比較表現の構文のみを対象として

(注1)：Web サイト：http://mecab.sourceforge.net/
(注2)：Web サイト：http://chasen.org/~taku/software/cabocha/

いる．属性が明示されにくい比較表現の評価を補完する際には，評価の表層を補完するのではなく，森山 [6] の知見を基に，「選択や希望を表す評価が行われている」という評価の傾向のみを補完している．

構文情報と補完ルールに基づいて対象や基準のスロットを補完できなかった場合は，ターゲット文に対してセンタリング理論を適用し，言語学的な知見によって補完用の実体を探索する．本手法では Nariyama [8] の手法を用いた．先行詞の候補リストには Salience Reference List(SRL) を用いた．SRL は先行詞候補を抽出し蓄えるため，先行詞らしさの選好の要素 (助詞など) ごとにスロットをもつ記憶領域である．この選好は主題であるほど省略されやすく，主題は助詞を用いて記載されやすいという日本語文の観察に基づいている．センタリング理論では文章の最初から文をたどり「名詞 + 選好」の要素を文節単位で検出してスロットを埋める．スロットがすでに埋まっている場合はスロットの内容を上書きする．これらの処理を指定した位置までおこない，その間スロット挿入を繰り返す．本研究で用いた SRL のスロットとその選好の要素は，Walker ら [9] の，センタリング理論に基づく日本語照応解析処理のモデルの先行詞らしさの選好を，以下のように拡張したものである．

- 対象・基準の比較特有表現 > は，って > が > を > に > その他

Walker らの選好を拡張した理由は，比較表現を含む文の主題が，おおむね比較特有表現と隣接するからである．また，blog 記事において助詞「って」は，助詞「は」と同じ意味で頻繁に使われるからである．

ここでターゲット文にセンタリング理論を用いる処理を最後まで行ったとする．また，センタリング理論により文を探索する際の条件として比較特有表現に基づく抽出ルールで抽出できた実体は SRL に格納しないとする．その結果，要素が挿入されている SRL のスロットのうち一番優先度の高いスロットに，対象か基準に割り当てられていない固有名詞を含む文節が含まれていれば，そのスロットの内容を抽出できていない比較関係の実体とする．

ターゲット文における主題の探索範囲は，文頭から評価の文節の周辺までが最適であることが，事前に行った観察から分かった．本研究では，対象が評価の直後に出現する場合を考慮し，ターゲット文における主題の探索範囲を文頭から評価の文節の 2 文節後の文節までとした．

本研究では図 1(B) のように，比較関係の要素が複数文にまたがって存在する場合についても考慮する．

本研究ではターゲット文に対するセンタリング理論の適用で実体を補完できなかった場合には，直前の文のみにセンタリング理論を適用し実体の補完を試みる．直前の文にセンタリング理論を用いる場合には，文頭から文末まで探索をおこなった．

4.5 評価の文節と構文情報に基づく副詞と属性の抽出

本手法では比較表現中の副詞を活用できていないが，副詞の抽出はおこなっている．比較表現中の副詞として，4.3 節の処理で抽出した評価の文節に直接係る副詞節を，ターゲット文の構文情報に基づきすべて抽出している．

また上述までの処理で属性を一切抽出できなかった場合には，属性の抽出処理をおこなう．この処理では，評価の文節に直接係る名詞節のうち，いままでの処理において比較表現の要素として抽出されていない評価の文節に，文節間距離が一番近いものを評価表現中の属性として取得する．条件にあう名詞節が存在しない場合には，属性を取得しない．

表 1　評価につかう固有名詞の組

拡張固有名詞タイプ名	クエリ A	クエリ B
政党名	自民党	民主党
国籍名	日本人	韓国人
国内地域名	関東	関西
神社寺名	金閣寺	銀閣寺
製品名その他	DS	PSP
製品名その他	ドラクエ	FF
便名	のぞみ	ひかり
言語名	英語	スペイン語
競技名	サッカー	野球
昆虫類	カブトムシ	クワガタ
植物名	桜	梅
魚類	鰻	穴子

5. 評価実験

本節では評価実験によって提案手法の有効性を以下の点に着目し調べた結果について述べる．

- 4 つ組の比較関係をどの程度正確に抽出できるか
- さまざまなタイプの固有名詞に対して有効か
- 複数文からの比較関係抽出にどの程度有効か

今回は blog 記事を提案手法で処理した結果を人手で確認した．具体的な実験方法を以下に述べる．

5.1 評価データ

はじめに評価データを取得するための固有名詞組を用意した．その際に Sekine ら [10] の拡張固有表現階層(注3) から選んだ，11 種類の固有名詞タイプに含まれる固有名詞組として，表 1 に示した計 12 組を取得した．固有名詞のタイプは拡張固有表現階層の葉からランダムに選択した後，階層を人手で観察したところ，選択したタイプが製品名分野に偏っていたので，偏りがあった分野については，人手で固有名詞のタイプを分散するように選択した．各固有名詞タイプに属する固有名詞組は，4.1 節で述べたクエリによって blog 検索で文書を取得できるかを確認し，ある程度の記事が取得できた固有名詞組を採用した．取得した固有名詞の表記は「ドラクエ」のように blog 検索で検索結果が多く得られるよう省略したものもある．

表 1 の固有名詞組を 4.1 節で述べた前処理に用いて，各固有名詞組に対応する文集合を取得した．文集合は 2007 年 8 月 20 日に取得した Yahoo!ブログ検索(注4) の検索結果 RSS 中の URL をクロールして取得した．

(注3)：Web サイト：http://nlp.cs.nyu.edu/ene/
(注4)：Web サイト：http://blog-search.yahoo.co.jp/

表 2　比較表現を含まない文の除去処理を評価

除去処理の平均精度	0.880
除去処理の平均再現率	0.726

表 3　抽出された比較関係の評価

	精度	再現率	F 値
4 つ組	0.922	0.680	0.783
3 つ組	0.825	0.734	0.776
4 つ組 +3 つ組	0.847	0.722	0.780

その後, 文集合から図 9 に示したすべての仮定が真にならない文を除去した. この作業は提案手法の処理対象となり得ない構文を含む文を文集合から取り除くことを狙っている. 特に図 9(1) の条件を満たしていない文からは比較関係をとることが不可能である. そこで図 9(1) の仮定に適合した 939 文を, 図 9(2) から (6) の仮定に基づき処理し, 処理にかけた全ての文を人手で評価した. 比較表現を含まない文を除去できた場合を正解とした. 再現率を計算するために, 人手で確認した文は図 9(1) の仮定に適合した 939 文で, 仮定に適合しなかった文は確認していない. 結果, 表 2 に示すような性能で以後の処理に不要な文を取り除けることが分かった.

その後, 今回の実験結果の考察に寄与しないと考えた文区切りが不正確な文と図 9 の仮定では除去できないが人が見れば対象外と分かる文を各文集合から取り除いた. 12 組の固有名詞組とそれらに対応する各文集合に最後まで残った文 (計 561 文) を評価データとして取得した.

5.2 評価結果

はじめに提案手法によって評価データの文集合から抽出した 4 つ組と 3 つ組の比較関係を人手で評価した. 比較関係は合計で 478 組抽出できた. 判定時には提案手法によって出力した比較関係と原文とを人間が見比べた. 本実験では評価を著者がおこなった. この評価時には以下の条件のうち, どちらかを満たす比較関係を正解とした.

- 人間が 4 つ組を取得できると判断した文から人間が正しいと感じる 4 つ組を取得できた場合
- 人間が 3 つ組を取得できると判断した文から人間が正しいと感じる 3 つ組を取得できた場合

また, 再現率を計算するために評価データとして取得した 561 文を人手で確認した.

表 3 が実験結果である. 実験結果から 4 つ組, 3 つ組抽出ともに高い精度を達成していることが分かる. 表 4 は表 3 の固有名詞組別の詳細である. 表 3 から, 3 つ組も 4 つ組も極端に精度の低い固有名詞タイプは少ない. 本手法はさまざまな分野の固有名詞に関して, 高い精度で比較関係を抽出できたといえる. しかし再現率については課題を残した. 提案した比較関係抽出手法による出力の全体の F 値は 0.780 であった. 評価方法が異なるため既存の手法の結果との比較ができず, 一概には言い切れないが, 提案手法は良い成果をあげたと考えている.

次に, 実体のスロットを補完する際にセンタリング理論を適用したことの有効性や, 直前の文から主題を取得することの有効性を確認する. 提案手法では, 比較関係の対象と基準のスロットを埋めるための方法の一つとして, ターゲット文または直前の文にセンタリング理論を使用している. センタリング理論によって文の主題を見つかり, それが目標としている実体であれば, その実体を空の対象か基準のスロットを補完する実体として扱う. 提案手法ではターゲット文には比較特有表現と隣接する実体が最低 1 つ出現することを見込んでおり, 一つ目の実体は抽出ルールによって抽出できる. そこで今回の評価実験中に, 正しく対象と基準を抽出できた場合に着目し, 正しく対象と基準を抽出できた際の二つ目に取得した実体の取得方法の割合を調べた. その結果を表 5 に示す.

表 5　2 つめの実体の取得時に使った方法の割合

使用した手法	割合
実体の抽出ルール	0.536
評価の文節番号+係り受け結果+実体の補完ルール	0.337
ターゲット文に対するセンタリング理論	0.109
直前の文に対するセンタリング理論	0.019

表 6　各要素の抽出を試行した際の正解率

抽出を試行した要素	正解率
対象・基準組	0.852
評価	0.839
属性	0.908

表 5 の結果から, ターゲット文にセンタリング理論を適用することで, 対象と基準の抽出に関する正解率が約 11%向上していることが分かる. また, 直前の文にセンタリング理論を適用することで 2%程度向上していることが分かる. 今回の実験において, 直前の文の主題を比較の対象として補完しなければならない場合は, 478 文中 9 文であったが, 今回は 9 文全てから補完用の実体を取得できた.

次に比較表現抽出手法の各処理段階における処理の正確さを調べる. 処理ごとに正解率を確認することで, 処理のボトルネックを見つけ今後の課題としたい. 比較関係の各要素の抽出をおこなった際の正解率を表 6 にまとめた. この評価では, ある要素の抽出が終わった時点で, その要素を正しく抽出できていれば正解とした. 一連の基準と対象を取得する処理で 2 つ組を正しく取得することができたのは, 評価データ中の比較表現のうち 85.2%であった. また評価抽出の精度は 83.9%であった. 結果から特に実体組と評価の抽出処理の正解率が, 表 3 の再現率を下げる原因となっていることがわかった.

6. 考　察

6.1 比較関係抽出の再現率

提案手法の比較関係抽出処理は十分な精度の高さを達成した. 一方で再現率が低いために F 値では 0.8 を超えられなかった. 比較関係抽出の再現率が低かった原因のひとつは, 表 6 で示した比較関係の各要素の抽出を試行した際の正解率が十分に高くないことだと考えられる. 表

表 4　抽出した比較関係の固有名詞組別の評価

固有名詞組		総処理文数	総抽出組数	4 つ組抽出数	3 つ組抽出数	抽出した比較関係の人手による評価結果					
						4 つ組		3 つ組		4 つ組 +3 つ組	
						精度	再現率	精度	再現率	精度	再現率
野球	サッカー	86	77	30	47	0.967	0.763	0.894	0.875	0.922	0.826
関東	関西	60	55	26	29	0.885	0.605	0.690	0.909	0.782	0.717
ドラクエ	FF	38	34	3	31	1.000	0.500	0.903	0.875	0.911	0.815
自民党	民主党	89	70	29	41	0.931	0.771	0.693	0.519	0.786	0.618
英語	スペイン語	20	15	1	14	1.000	0.333	0.786	0.647	0.800	0.600
金閣寺	銀閣寺	21	18	7	11	0.571	0.800	0.909	0.625	0.778	0.667
DS	PSP	108	90	23	67	0.957	0.489	0.597	0.634	0.689	0.574
日本人	韓国人	44	33	7	26	1.000	0.467	0.731	0.656	0.788	0.591
ひかり	のぞみ	6	4	1	3	1.000	0.750	1.000	0.500	1.000	0.667
クワガタ	カブトムシ	24	23	5	18	1.000	0.833	0.889	0.889	0.913	0.875
桜	梅	60	54	16	38	0.750	0.857	0.816	0.674	0.796	0.717
鰻	穴子	5	5	3	2	1.000	1.000	1.000	1.000	1.000	1.000

6 の結果から, 対象・基準組の抽出と評価の抽出が終わると 3 割程度の比較表現からは属性抽出を正しく行えないと分かる. 比較関係抽出処理の再現率を向上するためには, 対象・基準組と評価の抽出の改良が必要である.

6.2　センタリング理論の有効性

評価実験から表 5 に示すようにセンタリング理論が対象・基準組の取得に寄与することがわかった.

センタリング理論が比較表現からの実体抽出に有効な理由は, 比較表現を含む文には複数の主題が含まれており, 比較関係を構成する実体は比較表現を含む文の主題となっていることが多いからだと考えられる. センタリング理論で文を探索する際に, 比較特有表現に基づく抽出ルールで抽出できた実体は SRL に格納しない, という制約を加えることで, SRL に格納される主題のうちどれかは抽出ルールで抽出できなかった実体であることが期待できる.

6.3　複数文にまたがる比較表現への対応

本研究では図 1(B) のように, 比較表現が複数文にまたがって存在する場合についてもセンタリング理論を用いることで対処した. ターゲット文中の実体とともに比較関係を構成する実体が, 直前の文における主題の一つとなっていれば, 人間はそれらの実体間に, 比較関係を見出せる可能性がある. 表 5 に示したとおり, 複数文に対応したことの対象・基準組を取得する処理に対する寄与は約 2%であった. 現実には, 単文中に含まれる比較表現を処理すれば, 複数文を処理する前に実体組を抽出できる場合が大幅に多いため, この数字は妥当な値だといえる. 本稿では複数文という言葉の範囲をターゲット文の 1 文前までとしていたが, 1 文前にセンタリング理論を適用した際に, SRL のスロットが一切埋まらなかった場合には, それ以前の文から 1 文前の文の文末までの範囲にセンタリング理論を適用していれば, より多くの複数文にまたがって存在する比較表現を取得できたと考えられる.

7. おわりに

本研究では blog 記事から比較関係を抽出する手法を提案した. 従来提案されていた日本語の比較関係抽出は 3 つ組 (〈対象, 基準, 評価〉) を単文から抽出する手法であった. それに対し提案手法では, 単文中または複数文にまたがる 4 つ組 (〈対象, 基準, 属性, 評価〉) および 3 つ組の比較関係を抽出できる手法を提案した. センタリング理論で抽出した文の主題によって, 比較関係の実体を補完することは構文情報に基づく抽出ができなかった実体を見つけるうえで有効であると分かった. 比較関係中の評判情報の抽出と, その評判情報の極性判定は本手法を実世界において活用するために解決すべき課題である.

文　献

[1] 鈴木 泰裕, 高村 大也, 奥村 学: Semi-Supervised な学習手法による評価表現分類, 言語処理学会 第 11 回年次大会, pp. 668-671, (2005).

[2] Liu, B., Hu, M., Cheng, J.: Opinion Observer: Analyzing and Comparing Opinions on the Web, *Proceedings of the 14th International World Wide Web Conference (WWW)*, pp. 342-351, (2005).

[3] Jindal, N., Liu, B.: Identifying Comparative Sentences in Text Documents, *Proceedings of the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR)*, pp. 244-251, (2006).

[4] Jindal, N., Liu, B.: Mining Comparative Sentences and Relations, *Proceedings of 21st National Conference on Artificial Intelligience(AAAI)*, (2006).

[5] 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, L1-5 (2007).

[6] 森山卓郎: 日本語における比較の形式, 月間言語 2004 年 10 月号, pp. 32-39, 大修館書店, (2004).

[7] 山本秀樹: 世界諸言語の地理的・系統的語順分布とその変遷, 溪水社, (2003).

[8] Nariyama, S.: Grammar for ellipsis resolution in Japanese, *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation* , pp. 135-145, (2002).

[9] Walker, M., Iida, M. and Cote, S.: Japanese discourse and the process of centering, *Computational Linguistics*, Vol. 20, No. 2, pp. 192-233, (1994).

[10] Sekine, S., Sudo, K., Nobata, C.: Extended Named Entity Hierarchy, *3rd International Conference on Language Resources and Evaluation(LREC-2002)*, pp. 1818-1824, (2002).