E-006

技術論文からの比較情報抽出

Mining Comparative Relations from Technical Papers

蜂谷 和士[†] 寺邊 正大[‡] 橋本 和夫[†]
Kazuto Hachiya Masahiro Terabe Kazuo Hashimoto

1 はじめに

現在では，論文のデータベース化が進み非常に多くの論文にネットワークを介してアクセスすることができる．しかし，論文内容の分析が技術サーベイのボトルネックとなっており，技術サーベイの自動化と支援は今後必要となる技術である．そこで本論文では，このための基盤となる，技術論文からの比較情報の抽出について検討する．

テキストからの比較情報の抽出については，Jindal[1] は，比較文をその意味によって最上級・同等・有差の３種類に分類した．また，系列ルールマイニングで得られた頻出パターンをナイーブベイズ学習の属性に組み込み，英文テキスト中の比較情報の元となる比較文の位置同定を行った．さらに Jindal[2] は，英語形態素の系列ルールマイニングによって，英語比較文から４つ組の比較情報を非常に高い精度で抽出している．日本語テキストを扱った研究では，倉島ら [3] が，Web 上で収集した比較文の集合から，形態素のパターンや対象集合を用いることによって，比較情報のうち，対象，基準，評価の３つ組の抽出に高精度で成功している．また，比較の２項関係をもとにランキングを構築する手法も提案している．同様に形態素パターンを利用した佐藤ら [4] は，blog 記事の集合から比較文を抽出し，さらにその中から比較情報の４つ組全てを抽出する手法を提案しており，高い精度を得ている．

本論文では，手法比較を行うため Jindal[1] に倣い，対象・基準・属性・評価の４つの要素により比較情報を構成する．「手法 A は手法 B よりも精度が 高い」という文を例にとると，対象は比較される対象となる「手法 A」，基準は比較の基準となる「手法 B」，属性は比較の観点を表す「精度」，評価は比較の評価結果を表す「高い」をそれぞれ示す．このように比較情報を構成することにより，属性についての２項関係を得ることができる．属性に関する２項関係の自動抽出・分析が，本論文の目的である．

本論文では，技術論文における比較文の統語構造を分析し，これを受理する構文規則の検討を行った．以下では，２章に技術論文と blog における比較文の相違点を述べ，３章で技術論文中の比較文の統語構造分析ならびに構文規則の提案を行い，４章でまとめを述べる．

2 技術論文と blog における比較文の相違点

係り受けによって比較関係抽出を行なう佐藤らの手法 [4] は．技術論文の比較文において特定の統語構造における係り受けの特徴が評価できるため，まずこれについて説明する．

佐藤らは，Jindal[1] の分類による有差を表す比較文を対象とし，Jindal[1] に従い比較情報は対象，基準，評価，属性の４要素から構成されるものとした．また，対象と基準のことをまとめて実体と呼称した．この前提をもとに，佐藤らは，比較文抽出，前処理，対象・基準抽出，評価抽出，実体・基準補完，属性抽出の６つのプロセスから構成される比較関係抽出手法を提案した．佐藤らの手法は表１に示す統語構造を持つ比較文を対象に，係り受け関係に基づいて，比較の４つ組である対象，基準，属性，評価を抽出する．

佐藤らの手法の技術論文への適用可能性は，技術論文に現れる比較文の統語構造に依存する．そこで，本節では佐藤らが前提とした比較文の統語構造が，技術論文でどの程度頻繁に現れるかを調査した．結果を表

[†]東北大学大学院　情報科学研究科，Graduate School of Information Sciencies, Tohoku University
[‡]株式会社　三菱総合研究所，Mitsubishi Resarch Institute, Inc.

表 1: blog に頻出の比較文の統語構造

統語構造	比較文→対象句, 基準句, 属性句, 評価句
係り受け関係	（実体１）は （実体２）よりも 値段が 高い

3 に示す.

実験に使用したデータは，実際の技術論文に含まれている手法性能の比較文 106 件と，blog に現れた比較文 73 件である．前者は人工知能学会学会誌から無作為に論文 30 件を選び，その中から手作業で抽出した．後者は佐藤らに倣い，表 2 に示す対象・基準組をクエリとして，google ブログ検索 [5] を利用し，その中から手作業で抽出した．

blog では (1)(3) の構文が多いのに対し，技術論文では (3) の構文が多いことが分かる．

表 2: 佐藤らが用いた実体（対象・基準）組

分類	実体 1	実体 2
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
昆虫名	カブトムシ	クワガタ
植物名	桜	梅
魚類	鰻	穴子

表 3: 比較文の統語構造の分析

統語構造	blog (72 文)	技術論文 (107 文)
(1) 比較文→対象句, 基準句, 属性句, 評価句.	46%	25%
(2) 比較文→基準句, 対象句, 属性句, 評価句.	27%	0%
(3) 比較文→対象句, 基準句, 評価句, 属性句.	0%	42%
(4) 比較文→属性句, 対象句, 基準句, 評価句.	3%	11%
(5) その他	24%	22%

次に，佐藤らの手法 6 工程のうち，比較文の係り受け関係を利用しているプロセスである評価抽出，属性抽出を，blog と技術論文双方に適用し，それらの抽出精度を求めた．なお，累積誤差を無くすため，ここでの評価抽出は正しい対象及び基準が既に抽出されている条件の下で行い，属性抽出も正しい評価が抽出されている条件下で行った．また，統語構造毎に評価と属性の双方が抽出できたものとそうでない場合の割合を調べた．実験結果を表 4 および表 5 に示す．技術論文で頻出する (3) の構文は，表 6 に示すような係り受け関係であると考えられるが，佐藤らの手法は十分に機能していない．これは，佐藤らの手法が blog に頻出する比較文の統語構造のパターンを前提としているためであり，技術論文に適用するためには，前提とする統語構造を拡張する必要がある．そこで，３章では技術論文に現れる比較文の統語構造を分析し，これを受理するために必要となる文法規則を明らかにする．

表 4: 評価および属性の抽出結果

	blog(71 文)	技術論文 (106 文)
評価抽出	75%	50%
属性抽出	86%	36%

表 5: 統語構造毎の評価・属性抽出

統語構造	抽出可	抽出不可
(1) 比較文→対象句, 基準句, 属性句, 評価句	38%	12%
(2) 比較文→基準句, 対象句, 属性句, 評価句	0%	0%
(3) 比較文→対象句, 基準句, 評価句, 属性句	6%	56%
(4) 比較文→属性句, 対象句, 基準句, 評価句	24%	5%
(5) その他	32%	27%

表 6: 技術論文に頻出の比較文の統語構造

統語構造	比較文→対象句, 基準句, 評価句, 属性句
係り受け関係	（実体１）は （実体２）よりも 高い 精度を もつ

3 提案手法

3.1 統語構造の分析

２章の実験で用いた技術論文の比較文106件に対して，それぞれの統語構造の特徴を分析した．結果を以下に示す．

1. 評価と属性で名詞句を構成する場合 (18件/106件)
   - 従来型分布推定アルゴリズムである $UMDA, MIMIC, EBNA_{対象}$ は従来型進化計算である $SSGA_{基準}$ より よい$_{評価}$ 性能$_{属性}$ を示した
2. 評価と属性で名詞句を構成し，且つ比較文が名詞節に埋め込まれる場合 (13件/106件)
   - このように，提案手法 $PAGE_{対象}$ は Royal-Tree 問題において，既存の手法$_{基準}$ と比較して 高い$_{評価}$ 探索能力$_{属性}$ を有していることが分かる
3. 評価と属性で名詞句を構成し，対象を示す名詞句が省略される場合 (2件/106件)
   - しかし，$EHBSA/WT_{基準}$ よりは 悪い$_{評価}$ 結果$_{属性}$ を示した
4. 評価と属性で名詞句を構成し，対象を示す名詞句が省略され，且つ比較文が名詞節に埋め込まれる場合 (1件/106件)
   - 一方，騙し構造が強くなるほど，$ANS(List3)_{基準}$ の方が 良い$_{評価}$ 結果$_{属性}$ が得られるものと考えられる
5. 評価と属性で構成される名詞句が，格助詞「で」を伴って手段を示す副次補語となり，且つ比較文が名詞節に埋め込まれる場合 (5件/106件)
   - この表から分かるように，$UPAGE_{対象}$ は $PAGE_{基準}$ と比較して 少ない$_{評価}$ 適合度評価回数$_{属性}$ で最適解を獲得していることが分かる
6. 複数の属性に関する比較表現が１つの文を構成する場合 (3件/106件)
   - この両者で $*opt/UB$ へ達した回数の平均$_{属性\ 1}$ を比較すると，$ANS(List3)_{対象\ 1}$ の結果が 良い$_{評価\ 1}$ が，近似解の精度の平均$_{属性\ 2}$ を比較すると，$ISM_{対象\ 2}$ の方が僅かに 良い$_{評価\ 2}$ 結果となっている

     属性 $*opt/UB$ へ達した回数の平均$_{属性\ 1}$ に対して対象 $ANS(List3)_{対象\ 1}$ は文中に現れるが，基準となる ISM は文中に現れない．また，属性 近似解の精度の平均$_{属性\ 2}$ に対して対象 $ISM_{対象\ 2}$ が現れるが，基準となる ANS(List3) は現れない．しかし，この両者で比較していることから，省略された基準が何を指し示すかは明らかである．
7. 属性が取り立てられて，対象・基準より先に出現する場合 (15件/106件)
   - 計算コスト$_{属性}$ という点から見ると，$MSGP_{対象}$ は通常の $GP_{基準}$ よりも 高い$_{評価}$
8. 属性が取り立てられて，対象･基準より先に出現し，比較文が名詞節に埋め込まれる場合 (4件/106件)
   - 計算時間$_{属性}$ においては明らかに 提案手法$_{対象}$ が $EAX-HSGA_{基準}$ と比較して 少ない$_{評価}$ ことがわかる
9. 属性が取り立てられて，対象･基準より先に出現し，基準が対象よりも先に出現する場合 (2件/106件)
   - 逆に，図 7, 図 8 の FS-MOGA と，図 18, 図 19 を比較すると，局所パレート最適解乗り越えのための探索$_{属性}$ においては，$UNDX_{基準}$ よりも $SPX_{対象}$ が 適している$_{評価}$ ことがわかる
10. 対象・基準・属性・評価の語順で比較文が構成される場合 (16件/106件)
    - また，図25から，$DS-GA+MGGGS_{対象}$ はいずれのサブ集団数においても $IGA+MGG_{基準}$ より 収束速度$_{属性}$ が 速い$_{評価}$ ことが分かる
11. 対象・基準・属性・評価の語順で比較文が構成され，対象が省略されている場合 (1件/106件)

- いずれの選択方法においても，$DS-GA+MGGGS$ や $IGA+MGG$基準 に比べて 最適解に収束する比率属性 が 低かった評価

対象を示す主語は文脈から明らかであることから省略され，比較文中に現れていない．

12. 対象・基準・属性・評価の語順で比較文が構成され，対象が省略され，且つ比較文が名詞節に埋め込まれる場合 (3 件/106 件)
    - また，適切なサブ集団を用いることで $sGA+MGG$基準 より 最適解収束比率属性 が 高い評価 ことが示された
13. 対象・基準・属性・評価の語順で比較文が構成され，基準と属性で名詞句となり，且つ比較文が名詞節に埋め込まれる場合 (1 件/106 件)
    - さらに，提案手法対象 が TSP の近似手法として従来提案されている性能の良い GA基準 の 性能属性 も 上回る評価 ことを示した
14. 属性がサ変動詞化し，評価が動詞を修飾する副詞として現れる場合 (2 件/106 件)
    - 成功回数を比較しても，FLIP は安定化に優れる SLIP with Feedback より多く成功している
15. 属性がサ変動詞化し，評価が動詞を修飾する副詞として現れ，且つ比較文が名詞節に埋め込まれる場合 (11 件/106 件)
    - そのため，複数解が多ければ多いほど $MSGP$対象 が通常の GP基準 と比較して，効率的評価 に 複数解を探索属性 することが出来る
16. 属性がサ変動詞化し，評価が動詞を修飾する副詞として現れ，基準が省略されている場合 (2 件/106 件)
    - そのため，一つの大谷を解くのであれば従来の GA対象 の方が 効率よく評価 最適解を発見属性 できるであろう

    基準となる名詞句が文脈から明らかであることから省略され，比較文中に現れていない．
17. 属性がサ変動詞化し，評価が動詞を修飾する副詞として現れ，対象が省略され，且つ比較文が名詞節に埋め込まれる場合 (1 件/106 件)
    - しかしその一方で，最適解がコーナー付近に位置する場合は，他の手法基準 よりも 速く評価 最適解に到達属性 する場合が多く見られた
18. 対象・属性・基準・評価の語順で比較文が構成される場合 (2 件/106 件)
    - $UNDX$ を用いた既存手法対象 は，GD に関しては FS-MOGA に近い値を得ているが，$D1R$ の改善属性 が $FS-MOGA$基準 よりも 遅い評価
19. 対象となる名詞句が格助詞「の」を伴って属性の連体修飾語となる場合 (2 件/106 件)
    - 一方,f2 の曲線の稜は座標系と無関係であるために UX は改善解を発見できず,f2 での $UNDX-m+UX$対象 の 性能属性 は $UNDX-m+EDX$基準 と比較し大きく 劣る評価．
20. 対象となる名詞句が格助詞「の」を伴って属性の連体修飾語となり，且つ比較文が名詞節に埋め込まれる場合 (1 件/106 件)
    - 図 26 から $DS-GA+MGGGS$対象 の 収束速度属性 は $IGA+MGG$基準 より 速い評価 ことが分かる
21. 基準となる名詞句が格助詞「の」を伴って属性の連体修飾語となり，且つ比較文が名詞節に埋め込まれる場合 (1 件/106 件)
    - 4 章では，GA では精度の高い解を得るのが困難であること，GA と $GAwithLS$基準 よりも $GAthenLS$対象 の 性能属性 が よい評価 ことを実験によって確認する

3.2 比較文を受理する構文規則の提案

3.1 節で網羅した比較文の統語構造を受理する日本語文法を以下に定義する．

$\{A_i|i=1,\ldots,n\}$ は，要素 A_i が任意の順番で一回ずつ生起することを示す．(A) は，要素 A が 1 回もしくは 0 回生起することを示す．$[A_i|i=1,\ldots,n]$ は，要素 A_i のいずれかが生起することを示す．

比較文の構文規則

文 → (前置表現), 比較文, 読点. (1)
文 → (前置表現), 比較埋め込み文, 読点 (2)
比較文 → { 対象句, 基準句, 属成句 }, 評価句. (3)
比較文 → { 対象句, 基準句 }, 評価・属性句, 提示動詞句. (4)
比較文 → { 対象句, 基準・属性句 }, 評価句. (5)
比較文 → { 基準句, 対象・属性句 }, 評価句. (6)
比較文 → 比較文$_{-基準}$, 接続詞, 比較文$_{-基準}$. (7)
提示動詞句 → [示す, 示した，示される, 有している, 表す, 示唆する, 達成する]. (8)
比較文$_{-対象}$ → 属性句, 基準句, 評価句. (9)
比較文$_{-基準}$ → 属性句, 対象句, 評価句. (10)
比較埋め込み文 → 比較文, 非自立名詞, 格助詞, 動詞句. (11)

一般構文規則

後置詞 → 係助詞. (12)
後置詞句 → 名詞句, 後置詞. (13)
後置詞句 → 対象句, 後置詞. (14)
後置詞句 → 基準句, 後置詞. (15)
後置詞句 → 属性句, 後置詞. (16)
前置表現 → 接続詞, (句点). (17)
前置表現 → 動詞句, [と, ても], (句点). (18)
前置表現 → 副詞句, (句点). (19)
前置表現 → 名詞句, 格助詞, (係助詞), (句点). (20)
前置表現 → 連用節, (句点). (21)
連用節 → (後置詞句), (名詞句), (格助詞), 動詞句, [副助詞, 接続助詞, 非自立名詞 + 格助詞, 助動詞], (句点). (22)

対象を抽出する構文規則

対象 → 名詞. (23)
対象句 → (連体節), 対象, ([格助詞, 対象固有表現]), (句点). (24)
対象固有表現 → [のほうが, のほうは, の方が, の方は]. (25)

基準を抽出する構文規則

基準 → 名詞. (26)
基準句 → ([後置詞句, 連体節]), 基準, ([格助詞, 基準固有表現]), (句点). (27)
基準固有表現 → [より, よりは, と比較して, と比較すると, に比べて, と比べると, に比較して, に比べて]. (28)

属性を抽出する構文規則

属性 → 名詞. (29)
属性句 → (連体節), 属性, ([格助詞, 属性固有表現]), (句点). (30)
属性句 → 名詞句, (格助詞$_{を}$), 属性, [する]. (31)
属性固有表現 → [においても, については, という点から見ると, には, は]. (32)

評価を抽出する構文規則

評価 → [形容詞, 形容動詞, 比較動詞]. (33)
評価句 → (副詞句), 評価. (34)
評価$_{連体形}$ → [形容詞$_{連体形}$, 形容動詞$_{連体形}$, 比較動詞$_{連体形}$]. (35)
比較動詞 → [上回る, 優る, 優れる, 劣る]. (36)

属性と名詞句を構成する他の要素を抽出するための構文規則

対象・属性句 → 対象句, 格助詞$_{の}$, 属性句. (37)
基準・属性句 → 基準句, 格助詞$_{の}$, 属性句. (38)
評価・属性句 → (副詞), 評価$_{連体形}$, 属性句. (39)

3.3 構文規則の実装例

3.2 節の構文規則は，DCG を用いて Program 1 のように記述することができる．この文法により，係り受け解析では難しかった比較文の解析が可能となる．

3.1 節の 6 番目に示した統語構造の特徴では，複数の属性に関する比較表現が 1 つの文を構成しており，これは対象句が欠けた比較文が接続詞で並列する場合と基準句が欠けた比較文が接続詞で並列する場合として記述する事が出来る．

統語構造を明確に規定することにより，非明示的に要素が示される比較文においても，比較情報の 4 つ組の抽出を行う事が出来る．

4 まとめ

本論文では，技術論文からの比較情報の抽出を目的として，小規模な実際の技術論文のデータセットを用いた比較文の分析を行い，頻出する統語構造を明らかにした．さらに，これらの統語構造を受理するための構文規則を提案した．本論文で提案した文法の有効性については，今後大規模なデータを用いて検証を行う予定である．

Program 1 技術論文中の比較文を受理するための DCG 文法

```
% 通常の比較文
比較文 ([L]) --> 比較文_{full}(L).
比較文_{full}([Obj, Ref, Att, Val]) -->
        対象句 (Obj), 基準句 (Ref),
        属性句 (Att), 評価句 (Val).

% 対象句が欠けた比較文が接続詞で並列する場合
比較文 ([[Obj|L1][Obj|L2]]) -->
        対象句 (Obj),
        比較文_{-対象}(L1),
        接続詞,
        比較文_{-対象}(L2).
比較文_{-対象}(Ref,Att,Val) -->
        基準句 (Ref), 属性句 (Att),
        評価句 (Val).

% 基準句が欠けた比較文が接続詞で並列する場合
比較文 ([[Obj1, Obj2, Att1, Val1]
            [Obj2,Obj1, Att2, Val2]]) -->
        比較文_{-基準}(Obj1, Att1, Val1),
        接続詞,
        比較文_{-基準}(Obj2, Att2, Val2).
比較文_{-基準}(Obj, Att, Val) -->
        属性句 (Att), 対象句 (Obj),
        評価句 (Val).
```

参考文献

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

[2] Jindal, N., Liu, B., :Mining Conparative Sentences and Relations Proceeding of 21th National Conference on Artifitial Intelligence, 2006.

[3] 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.

[4] 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，自然言語処理研究会報告 Vol.2007，No.94，pp. 7-14，2007.

[5] http://blogsearch.google.co.jp/