構文規則に基づく特許文書からの比較情報抽出に関する検討†

蜂谷 和士‡　　橋本 和夫‡
Kazuto Hachiya　　Kazuo Hashimoto

1　はじめに

情報氾濫の時代と呼ばれる現代において，情報の全体像を見極めることは極めて難しくなっている．技術サーベイや特許審査など，大量のテキストから動向を分析する際，検索エンジンや文書分類の利用が一般的である．しかし，いずれも利用者が読むべき文書の取捨選択が目的であり，最終的には人手による内容分析に頼らざるを得ない．

こうした背景から，筆者は学術論文の内容分析を補佐する手法を研究している．学術論文中によく現れる，手法性能の善し悪しを既存手法と比較している表現を利用して，性能という観点から内容理解の補佐を行うものである．既に，学術論文から比較の表現を抽出するためのツールとして構文規則を利用したものを作成している．

現在，特許文書への応用も視野に入れ，技術文書と学術論文間の統一的な動向調査の手法を研究しているが，特許文書中に性能の比較文がどれだけ含まれているのか，実際のデータで検証がなされていない．

そこで本論文では，特許文書と学術論文中の各性能の比較文の現れ方について比較し検証する．そして，性能の比較文という観点から特許文書の動向分析を行う事の可能性について論じる．

2　学術論文からの比較情報抽出

構文規則を用いて学術論文から比較情報を抽出する手法について紹介する．

2.1　比較文と比較情報の定義

この手法で扱っている比較文は，Jindal[1] の分類による *Non-Equal Gradable* と定義している．また，比較文を構成する比較情報として，佐藤ら [2] に倣い，対象，基準，属性，評価の4つ組からなると定義した．

これら4つ組について，「手法Aは手法Bよりも高い精度を示している」という比較文を例に取ると，対象は比較の主体となる「手法A」，基準は比較の対となる「手法B」，属性は比較の観点である「精度」，評価は比較の評価結果を表す「高い」となる．

手法 $A_{対象}$ は 手法 $B_{基準}$ よりも 高い$_{評価}$ 精度$_{属性}$ をもつ

表 1: 性能に関する比較文と4つ組の例

2.2　構文の実装

比較文からの比較情報抽出を実現するため，学術論文に頻出する比較文構文の分析を行い，それらを受理する構文規則を Prolog の DCG 機能を用いて実装した．図1は構文規則による解析イメージである．

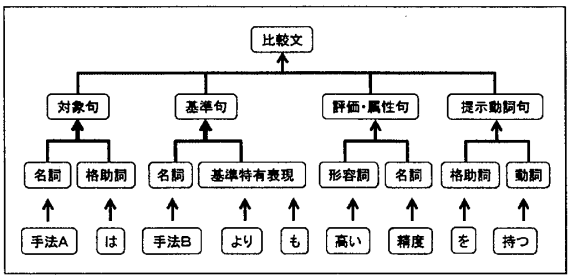

図 1: 構文規則による解析木

図の例では最終的に，対象句，基準句，評価・属性句という形で4つ組が抽出される．

2.3　スケーラビリティの評価

受理対象の比較文を限定した都合上，大規模データでは規則数の収束性に懸念がある．そこで，受理できる比較文の増加に対して規則数がどのように変化するかを調査した．結果を図2として示す．

受理可能な比較文の数に対して構文規則の数は安定しており，構文規則を用いるというアプローチはスケーラビリティの面で信頼できるものであるといえる．

†A Study Of Syntactic Analysis For Extracting Comparative Relations from Patent Documents

‡東北大学大学院　情報科学研究科，Graduate School of Information Sciencies, Tohoku University

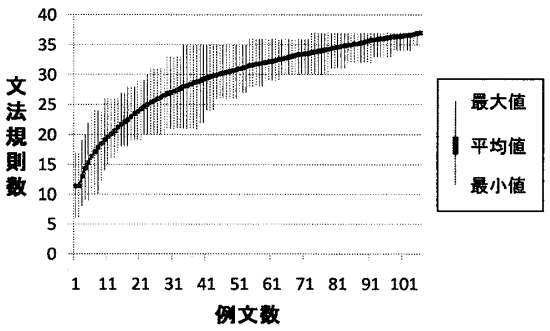

図 2: 受理可能な比較文の数に対する構文規則数の変化

3 比較情報抽出の特許文書への適用可能性

前章の学術論文から比較情報を抽出する手法を，特許論文に適用することを検討する．

3.1 特許文書における比較文件数の調査

まず，特許論文において比較文の出現頻度を調査し，学術論文と比較を行う事で特許論文に対し，比較文という観点で内容分析を行うことが適当であるかを判定する．

実験データとして，学術論文では，人工知能学会論文紙から無作為に選んだ論文 16 件を選び，その中から手作業で比較文 106 件を分析対象とした．一方，特許文書は，NTCIR-7[3] コーパスの 2002 年度特許全文データから 100 件無作為に選んだものを使用した．

その両データに対して，比較文件数および性能についての比較文件数を調査したところ，表 2 のようになった．

表 2: 1 文書あたりの比較文件数の比較

	比較文件数	性能の比較文件数
学術論文	10.50	5.25
特許文書	3.22	0.37

特許文書では，1 文書に性能の比較文件数が 1 件を下回っている．性能の比較文を抽出する本来の目的は，性能の比較情報の手がかりとするためであり，1 文書に少なくとも 1 つ以上の性能の比較文が存在しなければ，その文書の評価が得られず，性能の比較文単体で比較情報を構築する事が出来ない．

つまり，特許文書の場合，性能の比較文だけでは性能の比較情報の手がかりとして不十分であることがわかる．

3.2 比較情報抽出の代替案についての考察

性能の比較文とは別の手がかりを使って，比較情報を構築する方法を考える．

特許文書において，手法の性能は比較文という形で語られるよりは，寧ろ「〜ができる」「〜が可能になる」といった特徴表現で説明される傾向がある．これらの表現は，明示的には比較となる技術が示されていないが，文書の別の箇所に【発明の属する技術分野】が示されている．つまり，対象の特許は，提案された時点の所属分野すべての技術に対して，「〜ができる」という点の比較であるとみなすことができる．

なお，できる・できないを述べる比較は，Jindal の分類では *Non Gragable* に相当する．つまり，前提としている比較情報の定義を，*Non-Equal Gradable* から拡張する必要がある．

よって，比較情報として特許の性能をまとめるためには，抽出対象を比較文に限定しないこと，比較情報の定義を拡張すること，これら 2 点について検討しなければならない．

4 まとめ

本論文では，特許文書と学術論文間の統一的な動向調査の手法構築を目的として，比較情報の抽出という観点で両文書の比較を行った．

特許文書と学術論文中に現れる各比較文の現れ方について検証した結果，比較文単体では特許文書から性能に関する比較情報の抽出は難しい事が示された．

今後，比較文とは別の特徴表現を用いて，性能に関する比較情報を構築する手法について検討を行う．

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] 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.

[3] NTCIR Project, http://research.nii.ac.jp/ntcir/index-ja.html