博士（工 学） 桝 井 文 人

学 位 論 文 題 名

大規模コーパス知識を利用した 比較構造解析に関する研究

学位論文内容の要旨

自然言語処理技術は，機械翻訳や情報検索，質問応答など，序々に応用面での成果を上げつつある．しかし，人間にとっては容易に対処できるが，システムにとっては対処困難な言語表現は依然数多く残っている．例えば，"as..as" や "more..than" のような比較構文は省略や倒置を伴って長文の原因となるし，「…のような…」や「…のように…」などは比喩表現を生成して慣用化の原因となる．従来，これらの表現は抽象化が難しく，イディオムや慣用パターンとして表現毎に処理知識を用意していた．しかしながら，表現毎の対応では，用意した知識と実際に出現する表現との間に不整合が生じたり，処理の汎用化を考えた場合にコスト的な問題が生じる．

上記であげた特殊な表現は，二つの要素を比較する構造 (比較構造) として捉えることができる．比較構造は，表層レベルの比較によるものと意味レベルの比較によるものに大別できる．"as..as" や "more..than" などの構文に代表される，表層パターンや統語構造などの表層レベルの比較では，単語列や統語構造の規則性に基づいた比較構造の分類モデルを構築することが可能である．構築モデルを用いて比較構造の機能を判定することで，個々の比較構造と関連の強い省略や倒置，慣用化を特定でき，それらの補正処理や曖昧性解消にも繋がる．

これに対し，「…のような…」や「…のように…」といった形で出現する意味レベルの比較は，比喩表現を生成する場合と例示を生成する場合が存在するが，その違いは表層情報からは判断できず，より高度な処理が必要となる．特に統語面で自由度が高い日本語文では，その傾向が顕著である．意味レベルの比較については，まず，対象とする比較が，比喩・例示・無意味 (比較とは成り得ない) のいずれの意味として用いられているのかの判別が有効である．このような判別問題は，比較構造における比喩性検出の問題と考えることができる．例えば，「砂のような雪」という直喩 (属性比喩) の解釈では，source 概念 (たとえる概念)「砂」の顕現属性値「細かい」が，target 概念 (たとえられる概念)「雪」で強調される属性値としてクローズアップされ，「細かい」や「さらさらした」という特徴が理解されると説明できる．この解釈過程を計算機上に実装できれば，比喩性の検出，すなわち，比喩，例示，無意味の判別が可能となる．

また，上記判別処理の精度を確保するには，処理に用いる知識ベースの構築方法をも考慮する必要がある．従来研究では，知識ベースは心理学実験に基づいて構築することが基本であった．知識の大規模化，汎用化を考慮すると，知識ベースの自動構築は必須であるが，このような課題を扱った研究事例は過去ほとんど報告されていない．また，比喩性検出精度を下げる主要因として，概念を表す属性値集合 (のランキング) の歪みによる属性値クローズアップ誤りが挙げられる．これは，心理学的実験に基づいた知識を用いた場合でも，コーパスを利用した知識を用いた場合でも同様に生じ得る問題である．この問題を低減するためには，高精度の知識ベース構築手法あるいは知識補正手法が必要である．

本論文では，上で述べたような問題点を解決することを目標とする．比較構造を表層レベルの比較と，意味レベルの比較に分けて考える．前者については，統語構造の規則性に関する統計情報を利用した機能分類と，それぞれの機能に対応した解析手法について考える．後者については，意味的な語

彙比較の判別を属性比喩における比喩性の検出という問題として捉え, 属性値に基づいた比喩性検出手法について考える.

まず第一に, 表層パターンや統語構造を利用して比較構造を区別, 復元する手法を提案する. 英語長文中に多い"as..as" や"more..than" などの比較構造について, 文法書から得られる知識とコーパスの統計的傾向から得られる特性を整理統合し, 比較構文のモデルおよびそれを利用した解析処理の実現手法について述べ, システムが備える文法規則や辞書の適用が不可能な場合にも柔軟に対応できる比較構造判別モデルを構築する. 英字新聞に対する本方式と商用機械翻訳システムの解析結果の比較実験によって, 本方式の解析正解率が 80% を超え, 従来方式を大きく上回ことを確認し, 提案手法の有効性を示した.

第二に, 比喩や例示として出現する「…のような…」や「…のように…」など, 意味レベルの比較構造に対する処理手法を提案する. テキスト中に出現する比較構造の判別, 特に比喩表現の認識を重視し, 確率的な尺度を用いて, 概念 (単語) 間の比喩性を検出する手法について述べる. 比喩性を検出するための確率的な尺度として,「顕現性落差」と「意外性」を設定する.「顕現性落差」は, 概念対を比較したときに, クローズアップされる顕現特徴の強さをはかる尺度であり, 概念同士が理解可能か否かの判断に用いる.「顕現性落差」は, 確率的なプロトタイプ概念記述の枠組を用いて, 概念の共有属性値集合が持つ冗長度の差で定量化する.「意外性」は, 概念の組み合わせがどれほど稀であるかをはかる尺度であり, 概念同士が例示関係であるか否かの判断に用いる.「意外性」は, 単語間の意味距離を用いて定量化する. 二つの尺度を併用することによって, 比喩関係を持つ概念対, すなわち, 比喩性の判定が可能となる. 二つの尺度を計算するために, コーパス中から抽出した語の共起情報を利用して知識ベースを利用する. 両尺度を用いた比喩性検出手法を検証するために,1 年分の新聞記事コーパスから構築した知識ベースと, 比喩関係・例示関係・無意味の各単語対が混在するデータ 100 組を用いて, 単語対の判別実験を行い,70% 強の適合率で比喩関係単語対が判別できることを確認し, 提案手法の有効性を示した.

第三に, 上記手法で用いる知識の洗練手法を提案する. 比較構造判別処理における評価分析作業の効率化のために, 判別処理過程でクローズアップされた属性値の適合性判定と属性値集合への判定結果フィードバックを自動的に行う手法を提案する. 提案手法は, 対象概念とクローズアップ属性値を用いて生成した特定表現について,World Wide Web(WWW) 上の出現状況を調べることによって, クローズアップ属性値の適合性判定を行う. 不適合と判定された場合は,WWW から取得した属性値知識に基づいて属性値集合を再ランキングすることでフィードバックを行う. 実験の結果, 自動判定結果と人間による判定結果の間では, 約 80% の一致率が得られ, 十分な判定性能を確保できることを示すとともに, フィードバックにおいても, 属性値のランキング精度を約 20% 向上させることが可能であることを示した.

学位論文審査の要旨

主　査　教　授　荒　木　健　治
副　査　教　授　山　本　　強
副　査　教　授　宮　永　喜　一

学位論文題名

大規模コーパス知識を利用した
比較構造解析に関する研究

自然言語処理技術は，これまで要素技術に関する研究成果を蓄積し，近年，序々にその応用面，実用面において成果を積み上げつつある．しかし，そのほとんどは，典型的で素直な言語表現を対象としたものであり，いかなる文章，表現に対しても十分な処理性能を発揮するわけではない．現在，比較構文や比喩表現のように，人間のコミュニケーションにおいて非常に重要であることが認識されているにもかかわらず，現時点では対応が難しく，処理困難とされる高度な言語表現にも対応できる柔軟な自然言語処理技術の発展が待たれている状況にある．

本論文は，このような状況にある高度な言語表現の典型例である，比較構文や比喩表現/例示表現について，言語学や認知科学に基づく理論的モデルと大規模なコーパスを利用する統計的知識抽出の枠組みを用いて，比較構文の処理，比喩表現の検出，知識の適合性判定とフィードバック処理による知識の精緻化，またその応用に関して，「比較構造」という観点から統一的に研究を進め，自然言語処理が対応可能な言語表現の対象範囲を拡大することを目的としたものである．

第一に，比較表現を表層レベルの比較に基づく言語表現として扱い，英語長文中に多く出現する比較構文について，文法書から得られる処理規則と，言語理論に基づいた統合構造の基底構造 (D 構造) への抽象化および構成要素の復元モデルと，コーパス中の言語表現の統計的傾向を関連付け，組み合わせることによって，比較構文の解析精度が 43% から 84% へと大きく向上することを明らかにした．

上記比較構文処理では，意味レベルの比較に基づく言語表現である比喩表現への対応は難しく，さらなる考察と処理機構の検討が必要であった．そこで，第二に，テキスト中に出現する比喩表現と他の表現の判別を目的として，比喩性検出のモデル化を行った．認知科学と情報理論に基づく確率的なプロトタイプ概念記述の枠組みと，相互作用説に基づく顕現性落差と意外性という確率的判定尺度を定義した．顕現性とは，概念の典型性を決定する尺度であり，意外性とは，概念の組み合わせの新鮮さを決定する尺度である．本論文では，これらの尺度を定量化するモデルを構築し，コーパス中の連体修飾関係の統計的傾向に基づく名詞概念とその属性値の統計的知識を組み合わせることによって，概念対を比喩概念対，例示概念対，無意味概念対に判別することを可能とした．その結果，比喩概念対の検出を 70% 以上の精度で検出できることを明らかにした．

上記の検出性能をさらに精緻化させるためには，精度の良い知識を大量に獲得する過程が必要となる．しかし，大量の知識を獲得する過程，および，それらの知識の適合性を適宜判断し，修正・精緻化する作業コストが大きいため，これらの過程は自動化されることが望ましい．そこで，第三に，連体修飾語の装定機能が持つ顕現性に関する曖昧性を排除するために，既に獲得された知識 (名詞概念と属性値) と定型パターンを用いて比較表現を生成し，World Wide Web(WWW) 中の比較表現の

統計的傾向を調べることによって知識の適合性を判定し, 不適合である場合はさらに知識を抽出してフィードバックを行い, 知識を修正する処理機構について考察した. これによって, 大規模な知識に対して, 知識中の属性値集合の確率分布を自動的に補正することが可能となり, 適合性判定について人間判断の約 80% 程度をシミュレートして知識を精緻化できることを明らかにした.

第四に, 上記の研究結果を, より実用的な側面, 例えば, 質問応答の定義タスクや情報検索のクエリ拡張などへ応用することを目的として, 物事の概念を指し示す Entity を他の言葉で描写・叙述して表現する descriptor(記述表現) という定義と, 与えられた概念について比較表現を生成してコーパス中のそれらの統計的傾向を用いて知識が収集可能であるという統計的手法を組み合わせ,WWW からクエリ概念の descriptor を動的に取り出して視覚化表示する応用技術について考察し, 実験によって 74% 程度の精度で妥当な descriptor を獲得提示することが可能であることが明らかになった. ことによって, 本手法がクエリ概念のイメージを連想的に理解する支援手段として有効であることを示した.

これを要するに, 著者は, 自然言語処理において高度かつ難解とされる言語表現について有効な計算処理機構に関する新知見を得たものであり, 工学において貢献するところ大なるものがある。よって著者は, 北海道大学博士 (工学) の学位を授与される資格あるものと認める。