連体形形容詞に関する係りの特性の分析

氏名・(本籍)	菊　池　浩　三 (愛媛県)
学位の種類	博　士　(工　学)
学位記番号	工博甲第　189　号
学位授与の日付	平成11年 3 月24日
学位授与の要件	学位規則第4条第1項該当
研究科・専攻の名称	電子科学研究科　電子応用工学
学位論文題目	**連体形形容詞に関する係りの特性の分析**
論文審査委員	(委員長) 教授　水野忠則　　教授　吉田敬一 教授　新妻清三郎　　助教授　伊東幸宏 教授　下平美文

論文内容の要旨

インターネットの急速な発達により、人々の活動はますますバリアフリーなものとなり、日々膨大な情報が人々の間で交換されるようになった。また、交換される情報も、技術者中心の技術文のみでなく、一般の人が扱う日常的な感情表現を含む文へと拡大した。このような状況下で、一層便利で高品質な自然言語処理システム(機械翻訳システム)の出現が強く求められている。しかし、現在の機械翻訳システムは、一文の長さが短い技術文に対しては実用的な精度で翻訳することが可能になっているが、係りの場合の数が級数的に増加する長い文章の解析率は依然として低い。

そこで、本論文は、一般の日常活動でよく利用される用語である形容詞に関し、長い文での出現形態を考慮し、連体修飾を構成しうる連体形について係りの側面より詳細に分析を加え、その係り特性を自然言語処理で利用可能なルールとして抽出し、現行のシステムへの適用可能性を検討するものである。

自然言語の研究手法において、近年のコンピュータの飛躍的な性能向上は、大量の電子化されたテキストデータ(コーパスと呼ぶ)の蓄積を可能とし、それを利用した多面的な自然言語分析を可能とした。大規模コーパスの利用により、得られた結果の信頼性が増すだけでなく、統計的な分析が可能となり、今まで見えなかった言語の持つ特性が見える可能性が出てきた。大規模コーパスを利用した研究は以下のように多岐に渡っている。

・共起関係の抽出

・多訳語の同定

・統計確率に基づく出現用語の品詞推定

・構文規則の自動抽出

・出現共起によるルールの補強

ただし、大規模コーパスを利用し、統計的な処理を行った研究でも、得られた結果を言語現象の側面から詳細に分析した研究はあまり多くない。

一方、言語現象の側面からの研究には、人間の直感を必要とするが、このような研究では、長文の係りに関して次の2つの考え方に基づく研究がある。

・大域的分析：文全体の骨格を決定し、係りの曖昧性をなくす方式。

・局所的分析：特定構文を狭い範囲でできるだけ正確に分析し曖昧性を排除する方式。

これらの研究においては、表層情報(文字列)の中にも、未利用の多くの情報が存在することが指摘されている。

そこで本研究では、連体形形容詞という古くて新しい問題の解析に向け、以下の様な方針で取り組むこととした。

①大量のコーパスを利用する。

②分類キーとして形容詞の表層情報を使用し、大量分析を可能とする。

③実用性を考え、分析の中では意味情報はなるべく使わない。

④局所的分析を徹底的に行う。

⑤統計的分析の実施と、得られた結果の詳細な分析を行う。

⑥既存システムとの親和性を考え、ルールベースの方式で分析する。

本論文では、連体形形容詞を分析するにあたり、現状のシステムの適用可能性の観点から、まず現状システムが翻訳の主ターゲットとしている技術文においてよく利用される形容詞について詳細に分析した。これは、この範囲でも高い精度で解析できれば、あまり困難なく現状のシステムに組み込むことが可能となるからである。分析においては、まず一般に考えられる単純なルールを使った処理での解析精度をまとめ、それのみでは十分な精度が得られないことを示す。そして、精度向上のための詳細な分析を行い、得られた係りを規定する7つの規則(ルール)について、3つのカテゴリに分類し、ルールの適用順位を考慮しながら説明する。そして、最後に統計的確率分析に基づく係りのデフォルト属性について説明する。これらのルールを評価文に適用し評価したところ、形容詞の係りを97%以上の精度で特定できることを説明する。

技術文でよく利用される形容詞に対しては高い精度で係り解釈が可能なルールを検出できたので、これらのルールの汎用化を試みる。ルールが一般の形容詞に適用可能かどうか調べるため、形容詞を網羅的に調べ上げることとし、国立国語研究所での分析に基づき形容詞を体系化し、その分類に従って形容詞を抽出する方式をとった。言葉のスパース性のため、分類上の漏れをなくすことを目標にすえ、類似語や反意語等の視点から抽出ルールの妥当性検証を行い、その結果について説明する。またルールの汎用化や拡張について説明する。これらの結果、追加分析した形容詞に対しても95%以上の

精度で係りが特定できたことを説明する。

最後に、現在実用化されているシステムとの解析精度比較を行い、本方式が優れていることを示すと共に、実用システムへの具体的な組み込み手順について検討する。

以上、長年の課題の一つであった連体形形容詞の振る舞いに対して、机上でのシュミレーションとはいえ実用性のある一つの解析方法を提示することができた。これらは、連体形形容詞の一般的な係りの振る舞いを規定するものであり、得られた結果はシステムの構築方法に依存するものではない。それゆえ、本論文で提案した方式はいかなる翻訳方式を採用するに当たっても根底に流れる係りの振る舞いとして利用可能なものであると確信している。

論 文 審 査 結 果 の 要 旨

本論文は、自然言語処理における構文解析での課題の一つであった連体形イ・ナ形容詞の係りの振る舞いに対し、実用的な解析手法を提案している。提案の手法では係りの特性を計算機に組み込み易いルールとして表現している。また、一般的なルール化が困難な部分に対しては、大規模コーパスを用いた統計的手法で処理し、各形容詞の特性をデフォルト特性として整理している。この方法により、分析対象構文「名詞1＋が／の格＋形容詞＋名詞2」に対して、97%の解析精度で係り受けを判定できることを示している。

本論文は全6章からなっている。

第1章では、研究の背景と目的について述べている。

第2章では、日本語の持つ特徴と曖昧性について説明し、それらの解決のために現在までどのような研究がなされてきたかについて説明している。そして、本研究の位置づけを明確にするために、これらの方式と本論文で採用した方式との関係について述べている。

第3章では、技術文で使用頻度の高い形容詞に限定して詳細な分析を行い、係りを規定するルールとその精度について説明している。具体的には、まず、形容詞の係りを規定すると考えられている単純なルールを使った場合の係り解釈の精度をまとめ、それでは十分な精度が得られないことを示している。そして詳細な分析により得られた係りを規定するルールについて、3つのカテゴリに分類し、ルールの適用順位を考慮しながら説明している。また、統計的確率分析に基づく係りのデフォルト属性についても説明している。そして、検出したルールを評価文で評価し、連体形形容詞に関連する係りが97%以上の精度で特定できることを説明している。また先行類似研究との比較も行っており、本研究の方式が実用面ではるかに優れていることを示している。

第4章では、3章で得られたルールが形容詞全般に対して利用可能かどうかを検証するために、国立国語研究所で実施された分類に基づき形容詞を網羅的に調べ分析している。分析では、分類上の漏れをなくすよう用語を選択し、類似語や反意語等の視点から3章のルールの妥当性を検証し、その結果について説明している。またルールの汎用化や拡張について説明している。これらの結果、新規に分析対象とした形容詞群に対しても、95%以上の精度で係りを特定できることを説明している。

第5章では、現在実用化されているシステムとの係り解釈の精度比較を行い、本方式が優れていることを示すと共に、実用システムへの具体的な組み込み方法について段階的な組み込みを提案している。

第6章では、本研究のまとめを行っている。

以上により、本研究で提案した係り解析の有効性と既存システムへの適用可能性が検討できており、博士(工学)の学位を与えるものにふさわしいと認定する。