青山学院大学理工学研究科経営工学専攻

日本語文章の意味解析システムの精度と効率の向上手法の考案と評価実験

原田研究室　水野 高宏（35598014）

原田研究室ではこれまで、日本語で記述された問題文章からオブジェクト指向設計要素を自動抽出するオブジェクト指向分析の自動化システムの開発研究が行われている。この最初の工程として、日本語の意味解析システム SAGE98 が昨年度までに開発されている。しかし、SAGE98 はプロトタイプであり、自動的に処理は行われるが、解析時間が長い、解析精度が悪いといった問題点がある。そこで本研究では、性能的にも精度的にも実用レベルの意味解析システム SAGE99 を開発することを目標とする。開発方針としては、SAGE98 の基本的なアーキテクチャは踏襲するが、個々のコンポーネントに対して新しい解析アルゴリズムを考案することで、解析時間の短縮と解析精度の向上を目指す。

係り受け木が与えられると、SAGE98 は係り受け関係にある 2 文節の中心語の語意と 2 語間の深層格を EDR 辞書で検索し、語意-格組の統計的尤もらしさを割り当てる。そして、係り受け木の各枝に対して語意-格組とその統計的尤もらしさを割り当てることで解釈木を構築する。次に SAGE98 は、割り当てられた語意と格の統計的尤もらしさの和である総合評価値が最大となる木を探索する。最後に、SAGE98 は結果として得られた解釈木を、各語と語間の関係の情報を含んだ意味フレーム群に変換する。SAGE98 で解析時間が最もかかっているのは、EDR 辞書検索と解釈木構築の部分である。そこで、SAGE99 では辞書検索に C ライブラリを利用することで高速化を実現した。また、解釈木を構築する過程で無駄な枝を刈り取ることで探索空間を縮小する手法を考案したり、分枝限定法を適用することで、高速化を実現した。

解析精度が低いことは、具体的には次の２つの問題に分けられる。(A)意味フレームが出力されない。(B)出力意味フレームが誤っている。(A)の問題は概念記述辞書で語意-格組が見つからなかった場合に起こる。SAGE99 では、辞書に依存しない方法で語意-格組を決定することでこの問題を解決した。(B)の問題は、次の３つの方法で解決した。①それぞれの語意-格組の統計的確からしさを求める新しい計算法を導入した。②解釈木構築において、概念記述辞書に格情報が無い場合には、全ての語意の組み合わせに統計的尤もらしさ 0 を割り当てることにした。③用言間の係り受け関係において、深層格から表層格を推定する経験的ルールを利用した。SAGE99 では以上の３つの精度向上手法を実装し、評価した。

エレベータ問題を用いて SAGE98 と SAGE99 を比較評価したところ、文節数が 7 以下である文の 1 文あたりの平均解析時間が約 59 分→約 0.25 分と大幅に短縮された（PentiumⅢ550MHz RAM 320MB）。また、出力意味フレーム行数は 175 行→226 行となり、全ての文に対する意味フレームが出力された。解析精度（正解率）も、語意が 57.0%→82.1%、格が 43.3%→77.8%と、大幅に改善された。総合的に考えて、SAGE99 はプロトタイプシステムであった SAGE98 と比べて、解析速度的にも精度的にもより実用可能な意味解析システムになったといえる。