青山学院大学理工学部経営工学科

EDR 電子化辞書の
オンメモリ検索による意味理解の高速化

伊澤　友輔（15597009）　　大島　正樹（15598029）　　濵﨑　友子（15598091）
原　田　研　究　室

１．はじめに

原田研では、日本語文を意味解析し、格フレーム群に自動変換するシステム SAGE を開発してきた。このシステムは、EDR 電子化辞書に記載された情報を元に、係り受け関係にある２単語間の語意とその間の深層格に対する統計的な優先率（語意-格総合評価値）を算出し、これを元に格フレームを生成する。昨年度までの研究により、最適な解釈木構築にかかる解析速度は大幅に向上した結果、現状では全体の解析時間に占める辞書検索時間の割合が 90%以上になっている。そこで、本研究では、EDR 電子化辞書から SAGE による意味解析に必要なデータをメモリ上に展開し、メモリ上で辞書を検索することにより、更なる意味解析全体の高速化をめざす。

２．システム構成

本研究では EDR のテキスト形式の辞書を展開し、メモリ上のハッシュテーブルに格納する。

使用するメモリを最小限にするために、以下のような手段をとることにする。

・SAGE の解析で必要なデータのみを抽出する。

・日本語単語辞書の品詞および、概念体系辞書、概念記述辞書、日本語共起辞書の概念関係子（深層格）は対応する int 型の列挙型定数として格納する。

・概念関係子 CID（語意）については、概念体系辞書と概念記述辞書を同一の概念辞書テーブル上に展開し、各概念の上位概念、概念関係子（係り先の格）、係り先概念を容易に得ることができるようにする。

・日本語単語辞書および日本語共起辞書における概念識別子としては、概念辞書テーブルの該当概念辞書構造体のアドレスを格納する。

・日本語単語辞書は、同一語に対する多様な表記から検索できるようにするため、活用語については語尾付き単語見出しと不変化部をキーとして登録した。

利用にあたっては、辞書構築関数を呼び出し、辞書をメモリ上に構築してから、個々の文章を SAGE の解析部に渡すことにした。

３．語意-格総合評価値の求め方

先に述べたメモリ上の辞書テーブルを検索して語意-格総合評価値を求める。語意-格総合評価値とは、以下に示す語意-格確率と共起語意-格確率の和である。

３．１．語意-格確率

日本語単語辞書を用いて係り先単語と係り元単語の概念識別子を求め、概念体系辞書を用いてこの概念識別子の各上位概念とそれとの間の概念距離を求める。２単語の各上位概念を含めて、２単語間にどのような格関係が考えられるのかを概念記述辞書を用いて検索し、2 語の語意とその間の格の３つ組の尤もらしさを、語意-格確率として算出する。

３．２．共起語意-格確率

まず、係り先単語・助詞・係り元単語（助詞付き２単語）をキーにして共起辞書を検索し、その助詞と単語が共に出現した場合の２語の語意と格の出現確率を次式のように求める。

助詞付き２単語引きでの共起語意-格確率＝
該当共起レコードの共起項目頻度／該当共起レコードの表層共起頻度

助詞付き２単語引きでの共起辞書データがなかった場合、助詞・係り元単語（助詞付き１単語）をキーにして共起辞書を検索する。係り先概念識別子と共起レコードの係り先概念識別子の上位概念を比較して類似度を計算し、共起語意-格確率を求める。

４．結論

本システムを、メモリ容量 2GB のマシン（富士通 CELSIUS460）で比較実験したところ、意味解析全体にかかる時間は従来の SAGE の 20%以下に短縮された。この結果、多量の文章の解析にも耐えうる現実的な解析速度に至ったといえる。

５．参考文献

[1] 水野 高宏,原田実: “ EDR を用いた日本語意味解析システム SAGE ", 人工知能学会論文誌, Vol.16, No.1, pp.85-93 (2001.1).

[2] 田淵和幸,原田実: “ 日本語意味解析システム SAGE の高速化・高精度化と精度評価 ”,言語処理学会第７回年次大会発表論文集,pp.486-489（2001.3）.