青山学院大学理工学部経営工学科

日本語文章からの意味フレーム自動生成システム SAGE(Ｓｅｍａｎｔｉｃ frame Automatic GEnerator) の開発研究

尾見　孝一郎　(15594035)　佐々木　毅　(15594057)　林田　和久　(15594094)

原　田　研　究　室

１．初めに

過去に原田研究室内において、日本語要求仕様を分析し、オブジェクト指向設計図を自動生成する研究を行っていた。このシステムは CAMEO と呼ばれる。設計要素の抽出や設計図への配置は既に完成しているが、日本語文を分析するのに必要なフレーム形式への変換は手作業であった。そこで、日本語文章を解析し、その内容を文中の語の意味と他の語との関係を、統一した表現形式である意味フレーム形式として自動生成するシステム「SAGE(Semantic frame Automatic GEnerator)」を開発する事にした。

なお、日本語文章を形態素に分解していく作業には Chasen（奈良先端大学松本研）、日本語文章を文節に分解し各文節と述部との従属関係を解析する作業には BEST（奈良先端大学松本研）を、それぞれ使用する。

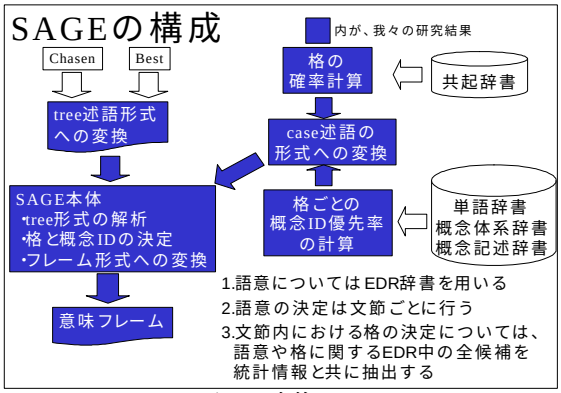
２．意味フレーム自動生成システム -ＳＡＧＥ-

日本語文章を解析するためには、巨大な辞書が必要となる。我々はこれについて、良質の語意や格(語彙間の関係)に関する文例を多く持つＥＤＲ電子化辞書（以下ＥＤＲ）を利用する事にした。

ＳＡＧＥは

- Tree 述語形式への変換ルーチン
- EDR 辞書検索
- SAGE 本体

より成り立ち、ＣＡＭＥＯへ受け渡す意味フレームを生成する。



２．１．ＳＡＧＥデータ変換ルーチン

Ｃｈａｓｅｎ/BEST からの出力は、そのままではＳＡＧＥへの入力には使えない。そこで我々は、Prolog の述語形式に対応するデータ構造として Tree 述語を考案し、これに変換する事にした。

２．２．ＥＤＲ検索ツール

これは二つのルーチンから成り立つ。

1) 格の確率計算ルーチン

文節内の助詞と述部との格関係をＥＤＲから検索、統計を取る。これをある格を取る確率、即ち「格の確率」とする。

2) 格ごとの概念 ID 優先率計算ルーチン

・単語と品詞の情報から、EDR の単語辞書を検索して概念 ID を得る。

・概念体系辞書を検索し、概念 ID の上位概念を得る。

・概念記述辞書を検索し、概念間の格、及び概念間の距離を得る。

・概念間の距離から、同一格間における、その組み合わせが成りうる概念 ID 優先率を求める。

２．３．メイン推論ルーチン

一つの形態素は通常複数の概念 ID 持つため、これを特定した文の構造表現(解釈木)は一つになるとは限らない。そこで、SAGE は最適な解釈木を得る指標として、確信度という概念を考えた。

確信度は、ＥＤＲ検索ルーチンにより求められた概念 ID 確率（Pc）を基に、確信度 CP＝ΣＰｃとして定義する。

各解釈木の確信度を比較し、確信度が最大となる解釈木を意味フレームの抽出に用いる。

なお、意味フレームは、

frame(フレーム番号、文中単語、読み、EDR 品詞情報、基本語、日本語品詞情報、活用形、概念 ID、格情報、文章番号)

で構成される。

３．終わりに

我々は、ＳＡＧＥにより日本語文章から意味フレームを自動生成することに成功した。この事により、ＳＡＧＥはＣＡＭＥＯを運用する際の有用な支援ツールになった。さらに、その他の自然言語処理問題に関しても適用が容易であることが示せた。

今後の課題としては、文脈解析(複文・重文解析)の精度を上げることにより、より多様な日本語文章に対処していくことが可能である。

４．参考文献

長尾　真 編、「自然言語処理」、岩波書店、1996

『茶筌』 ホームページ奈良先端大学松本研究室

http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html