意味解析システム SAGE の精度向上

青木 洋（15803003）　川口 純一（15803022）
原　田　研　究　室　　指導教員　原田 実 教授　松田 源立 助手

1．背景・研究目的

近年の IT の急激な発展によって、大量の文書データからの知識の発掘（テキストマイニング）などの分野で、文章の意味解析への期待が高まっている。

原田研究室では、文章中の単語の語意の決定および係り受け関係にある 2 文節間（主辞同士）の深層格の決定を行う意味解析システムSAGE[1]の研究を行ってきた。今年度の研究では、EDR辞書の整備としてひらがな語の登録、辞書中の不整合の解消を行い、また解析部においては、辞書引きにおける速度向上と、主辞・副主辞決定ルール、格決定ルール、ムード決定ルールの詳細化に重点を置いた精度向上を目的とする。

2．基本的考え方

SAGE は日本文を意味解析し、結果を文節や形態素ごとにそれらの意味や品詞や深層格（他の文節との役割的関係）などを保持したリストの集合として表現する。これは、文節を頂点、文節間の深層格を辺と考えると、図 1 のような意味グラフとして表現される。図 1 において、紫色の丸は文節を、白色の丸は文末を表し、文節間の矢線は係り受け関係および深層格を表示している。格の向きは、係り先→係り元とし、黒い辺は係り受け関係にある文節間の深層格関係、緑の辺は並列の深層格関係を示す。表示される語意は各文節の主辞（主要となる形態素）に基づき、文中の最後の文末節には主述語の文節への main 格を付与している。

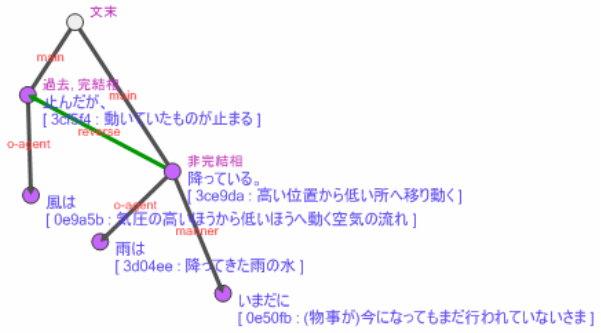

図 1.　SAGE の意味解析結果を示す意味グラフ

3.システム概要

SAGE の解析手順を図 2 を用いて説明する。

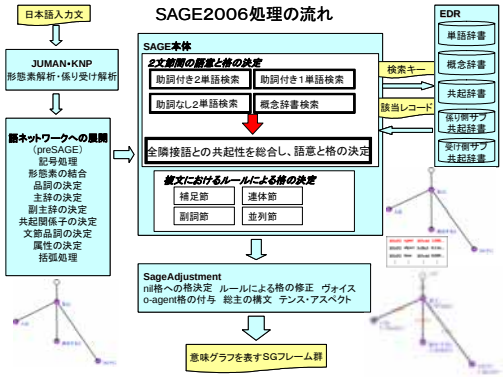

図 2.　SAGE2006 における処理の流れ

SAGEの処理に移る前に、まずJUMANとKNP[2]によって解析する日本語文章の形態素解析および係り受け解析を行う。その後、語意・深層格を決定するための処理を各文節・形態素ごとに行うことで調整する。SAGE本体ではEDR辞書を用いて、全隣接語との共起性を総合し、語意と格を決定する。また複文においてはそれぞれのルールに従って深層格決定を行う。

出力結果は図 3 のように表示される。文節、形態素ごとに概念 ID、品詞 ID、深層格 ID 等必要な情報を出力している。

```
[sg_v100]
f: 1,風は,は,ME,2,,[],[],[],,[]
    s: 2,風,カゼ,,0e9a5b,FTM,JN1,,
    s: 3,は,ハ,,3ca448,FJJ,JJO,,
f: 4,止んだが、,が,DO,5,,[oa1],[],[],,[過去,完結相]
    s: 5,止んだ,ヤンダ,止む,3cf5f4,DOS,JVE,子音動詞マ行,タ形
    s: 6,が,ガ,,3ca448,SEJ,JJO,,
    s: 7,、,,,2621d7,TOT,JSY,,
f: 8,雨は,は,ME,9,,[],[],[],,[]
    s: 9,雨,アメ,,3d04ee,FTM,JN1,,
    s: 10,は,ハ,,3ca448,FJJ,JJO,,
f: 11,いまだに,ゞ,HU,12,,[],[],[],,[]
    s: 12,いまだに,イマダニ,,0e50fb,FUK,JD1,,
f: 13,降っている。,,DO,14,15,[rv4,oa8,ma11],[],[],,[非完結相]
    s: 14,降って,フッテ,降る,3ce9da,DOS,JVE,子音動詞ラ行,タ系連用テ形
    s: 15,いる,イル,,0e52f0,DOB,JAX,母音動詞,基本形
    s: 16,。,,,2621d8,KUT,JSY,,
e: 17,null,null,[mn13]
```

図 3.　SAGE 格フレーム出力結果例

4．SAGE の精度向上

SAGE2005 では語意・格の精度が非常に信頼できる数値にまで達してきた。今年度は更なる精度向上を目指し、以下にあげる 4 点において特に重点的に研究した。

4.1 深層格精度の向上

「3.システム概要」で述べたように深層格は 2 文節間の関連性を共起事例を基にして決定されるが、全

てをこのルールのみで判定するのは非常に困難である。例えば、「ハ」格は「太郎は困った」のように主格を表す場合や、「この本は読みにくい」のように目的格を表す場合、「いつも朝は早く起きる」のように述部の対象や状況（総主）を表す場合など、多種多様な格になる性質がある。また、「も」「まで」「しか」といった取立て助詞は文脈によって主格にも目的格にもなる可能性があり、2 文節の共起だけでは判断が極めて難しい問題である。従来の SAGE では「神戸は夜景が綺麗だ」のように「～は、…が××だ」という形式の構文（総主の構文）のルールは実装されていたが、今年度はさらに時詞（朝、明日など）の処理や二重の object 格に関する処理を追加し、また受け側、係り側の品詞を基に主格、目的格を決定するルールの改善、取立て助詞における主格、目的格判定ルールの追加を行った。

また、SAGE2005 において精度に不安のあった並列格のルール修正をはじめ、様々な例文を基に、正しい格決定ルールの作成に力を注いだ。

4.2 語意精度向上

文節の語意を決定する際に用いる材料として主辞・副主辞は不可欠な要素である。そのため、主辞決定ルールに曖昧さがあるとその後の処理に支障をきたす可能性が高くなる。特に、「記念式典」のように 2 語以上の名詞（複合名詞）からなる文節ではより詳細なルールが必要となる。従来では、複合名詞に関しては文節の最後にくる名詞に重きが置かれていたが、「イージス艦飛翔」など前の名詞を後ろの名詞で言い換えているような複合名詞も存在するため、様々な場合を想定しルールの追加・改良を行った。また人名や商品名などでは未知のカタカナ語、アルファベット語が使用されることがあるが、従来の SAGE では辞書に登録されていない語に対して nil を返していた。SAGE2006 ではそのような語および「モトール・ワーゲン」のように”・”で連なる語に関して主辞・副主辞を含め「呼び名」という語意を与えることにした。

また、事態や相手に対する話し手の判断・態度を表す文法形式（ムード）を決定するルールの改良を行い、より詳細な解析を実現した。

4.3 解析速度向上

元来の助詞付き 1 単語検索による辞書引き方法は係り側の単語辞書レコードと受け側の共起辞書レコードにおいて一致した数を全数検査により類似度計算していた。しかし、例えば「～すること」では係り側の「する」の語意も、受け側の「こと」と@rentai による共起レコードも膨大に存在しており、従来の方法では多大な計算時間を費やさざるを得ない状況であった。SAGE2006 では、「4.4 EDR 辞書の改良」で後述するサブ共起辞書を作成し、共起事例中の係り側或いは受け側におけるある程度似通った語意を共通上位概念としてまとめた結果を利用することで、解析時間が約 1/20 に（JUMAN と KNP の処理を含めると約 1/4 に）短縮された。

4.4 EDR 辞書の改良

SAGE2005 での EDR 辞書は単語・概念・共起の 3 つの辞書で構成されていた。今年度はまず、「4.3 解析速度向上」で述べたように解析速度の向上のため、2 つのサブ共起辞書を共起辞書から作成した。これらの辞書は表 1 のように構成されている。すなわち、共起辞書を、係り側（あるいは受け側）＋共起関係子のペア毎に検索し、その結果のレコード群を、さらに、語意と深層格のペアで束ねた細群を作る。この際相手側の語には様々あるがそれらの概念の共通上位概念を保持させた。上位概念があまり上位レベルになっては相手にする語の概念がぼけてしまう。そのため、この共通上位概念の深さをある閾値以上とし、そのような上位概念が複数個になるときはリストとして保持させることにした。

次に SAGE2005 での単語辞書にはひらがなで表記された形容詞・動詞があまり登録されていなかった。そこで漢字で表記されている形容詞・動詞の単語レコードに対して、その漢字表記を語の読みを基にひらがな表記にし、新たに単語レコードを複製した。

また、EDR 辞書の不変化部の統一や、EDR 辞書の基となっているコーパス例文[3]より再度頻度を計算しその頻度に最も近い値を残して単語辞書の重複レコードの統合、語意決定の不都合を修正するレコードの登録・更新・削除を行い、EDR 辞書の改良に努めた。

表 1.　SAGE 格フレーム出力結果例

集合名	データ項目	データサイズ(byte)	type	繰り返し
mkj(muj).slz	係り(受け)側サブ共起列情報	不定	-	ファイル終端まで
係り(受け)側サブ共起列情報	辞書見出しサイズ(byte数)	4	signed	1
	※辞書見出し	1	utf-16	辞書見出しサイズ
	レコードサイズ(byte数)	4	signed	1
	レコード数	4	signed	1
	係り(受け)側サブ共起レコード	レコードサイズ-4	-	レコード数
※辞書見出しの形式は、"共起関係子¥係り(受け)側見出し"				
係り(受け)側サブ共起レコード	係り側概念ID	4	unsigned	1
	深層格名サイズ(byte数)	4	signed	1
	深層格名	1	utf-16	深層格名サイズ
	相対頻度サイズ(byte数)	4	signed	1
	相対頻度	1	utf-16	相対頻度サイズ
	相手側概念の数	4	unsigned	1
	相手側概念ID	4	unsigned	1
	・			
	・	相手側概念の数だけ繰り返し		
	・			

5.実験及び評価

本研究の評価実験では、無作為で抽出した WWW 上のニュース記事 101 文を使用した。語意、格、主辞・副主辞の精度及び解析速度について、SAGE2005 と比較した結果を表 2 に示す。

表 2.　SAGE2006 の評価実験の結果

	語意の正誤	格の正誤	主辞・副主辞選定の正誤	解析時間(sec)
SAGE2005	92.7%	84.0%	91.6%	92
SAGE2006	94.3%	87.2%	98.5%	3

SAGE2006 では主辞・副主辞の精度と解析速度が著しく向上した。また、語意・格においても高い精度を維持しており、より効率的な意味解析システムを実現したといえる。

今後の課題として、引き続き辞書データを整理すること及び連体節、同格、受身表現の深層格精度向上に焦点をあて、更なる利便性の向上を目指す。

参考文献

[1] 山本 哲哉, 小林 寛之, 米澤 太一, 原田 実：”意味解析システム SAGE の精度向上と利便性の向上”, 青山学院大学 2005 年度卒業論文
[2] 京都大学情報学研究科知能情報学専攻
知能メディア講座言語メディア研究室（黒橋研究室）
http://nlp.kuee.kyoto-u.ac.jp/
[3] (株) 日本語電子辞書研究所: EDR 電子化辞書仕様説明書（第 2 版）,(株) 日本語電子辞書研究所(1995)