意味解析システム SAGE の高精度化と概念グラフへの変換

前澤 敏之† 面来 道彦† 上野 雅和† 韓 東力‡ 原田 実‡
†,‡青山学院大学 理工学部 情報テクノロジー学科

1.はじめに

原田研究室はこれまで，EDR 電子化辞書を元に，日本語文章を意味解析し，格フレーム群に変換するシステム SAGE[1]を開発してきた．SAGE は，係り受け関係にある 2 文節の語意と格を EDR 辞書の情報によって決定している．このため，EDR 辞書に事例の少ない固有名詞，疑問詞，括弧による修飾，断定表現を含む文章の語意と格を決定する精度が低かった．また，文章が 1 文節のみで構成されている場合，共起情報が得られず語意が決定できなかった．本研究では，このような場合にも正しい語意と格を決定する規則を提案し，SAGE の高精度化を行った．また SAGE の格フレーム群を Sowa の提案する概念グラフ[2]に変換するためのシステム AEGIS を開発した．AEGIS は形式の変換に際して，述語の時制や様相の判別，概念とそのインスタンスの明確な区別を行う．

2.意味解析システム SAGE

SAGE における意味解析とは，係り受け関係にあるすべての 2 文節の語意とその主辞同士の深層格を決定し，語毎の格フレームからなる意味ネットワークを生成することである．ここでいう主辞とは，文節の主要な語を指す．ただし，複合語の場合は，主辞を明確に判別するために，主辞自身には head 格を付与している．また，文の主要な概念として文中における最後の文節の主辞には main 格を付与する．これは，日本語では一般に最後の文節がその文の主要な主張を表していることが多いためである．格の向きは，係り先→係り元とする．（例：「写真を撮る」撮る→object 格→写真）

3.SAGE2003 のシステム構成

SAGE は形態素解析と係り受け解析のされた文章を入力として意味解析を行い格フレーム群を生成する．なお，形態素解析には「茶筅」「JUMAN」を，係り受け解析には「南瓜」「KNP」を用いている．SAGE 本体では EDR 辞書による語意と格の決定が行われるが，このとき本研究で開発した SAGE2003 では，入力された文章が固有名詞，疑問詞，括弧による修飾，断定表現を含んでいる場合や，文章が 1 文節からなる場合，新規に作成した決定規則が適用される．このようにして生成された格フレーム群は AEGIS によって概念グラフ形式のフレーム(CG フレーム)群へと変換される．

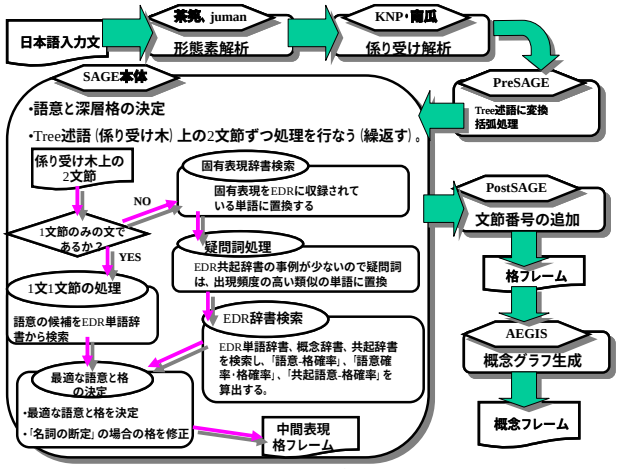

図 1.SAGE2003 の処理の流れ

4.SAGE の精度向上

ここでは従来 SAGE の問題点を解決し，正しい語意と格を決定する規則を提案する．これらの試みの結果については「6.結論」でまとめる．

4.1.固有表現の解析

SAGE2003 では EDR 辞書に未収録の固有名詞(以下，EDR 未知語)の語意を決定するための固有名詞辞書を用意している．SAGE は EDR 未知語を一時的に EDR 辞書に存在する語に置換し，その語と係り受け関係にある語との格を決定した後，固有名詞辞書により決定した語意を与える．例えば，「スナフキンは行く」の「スナフキン」は EDR 未知語なので，固有名詞辞書を検索する．その結果，「スナフキン」が人名であると判別され「人間」という語意が決定する．次に，表 1.に示す固有名詞置換表によって一時的に「スナフキン」を「私」に置換し，「私」と「行く」の間の格を object 格と決定した後，「スナフキン」に対して固有名詞辞書によってすでに決定されている「人間」という語意を与える．これによって正しい語意と格が決定できる．

表 1.固有名詞置換表

未知語分類	置換される語	語意
人名	私	人間
地名	渋谷	物事が行われたり，存在したりする物理的空間
組織名	マイクロソフト	組織体

4.2.補足節を伴う疑問表現の解析

疑問詞についても「4.1.固有表現の解析」で用いたものと同様の手法で解析できる． 例えば，「それは何故だ」の「何故」は疑問詞であるから疑問詞置換表によって「理由」に置換し，格が決定された後で「なぜ・どうして」という語意を与える．

4.3.名詞の断定表現解析

「A は(or が)B だ」というような断定表現のうちで，「柊は花だ」のように A と B が共に名詞である場合(以下，

Improvement of the precision of the semantic analysis system SAGE, and generation of Conceptual Graph
Toshiyuki Maezawa†, Michihiko Menrai†, Masakazu Ueno†, Dongli Han‡ and Minoru Harada‡
†Department of Integrated Information Technology, Aoyama Gakuin University.
‡Department of Integrated Information Technology, Faculty of Science and Engineering, Aoyama Gakuin University.

名詞の断定)，A は B という属性を持っていると考えられる．そこで AB 間の格を a-object 格とすることにした．

4.4.半角文字の処理

「南瓜」は全角文字による文章しか解析できない．そこで「南瓜」の前処理として文章中の半角文字を全角文字へと変換するルーチンを追加した．

4.5.括弧を用いた修飾の解析

従来の SAGE は括弧を含む文章に対して，単純に括弧を取り除く処理を行っていた．この処理方法では，「(私の個人的な意見ではあるが，)彼は天才である．」のような文章は括弧を除いても意味的な違いは生じないが，「田中氏(46)は東京に行った．」という文章では括弧を除くと「田中氏 46 は」となり意味が通らない．よって「田中氏」と「46」の間の深層格を正しく決定できない．そこで括弧を含む文章を以下の表 2.に示すように 5 つのパターンに分類し，それぞれに相応しい格を与えるようにした．

表 2.括弧の分類と対処法

種類	判別ルール	出力格・処理方法
補足・年齢	括弧内が名詞句のみ	modifier格を付与
換言・代替	括弧内が名詞句のみであり，括弧内と括弧前の概念IDが等しい	or格を付与
補完	括弧内の最後の形態素が助詞	括弧を取り除く

4.6.一文が一文節からなる文章の解析

SAGE は 2 文節間の共起情報から語意と格を決定する．しかし，この手法では「君だ」「まさか」といった 1 文節からなる文章(以下，1 文 1 文節の文)は解析できない．そこで SAGE2003 では 1 文 1 文節の文が入力として与えられた場合，共起情報による語意と格の決定を行わずに EDR 辞書の単語情報から語意を決定するようにした(1 文 1 文節の文では係り受け関係が存在しないため格の決定は行わない)．この結果，『「この本を持ち出したのは誰だ？」「私だ」「君か…」』といった口語に近い文章の解析が可能となった．

5.SAGE フレームから概念グラフへの変換

SAGE の格フレームを利用したシステムには質問応答や文書分類などが挙げられるが，これらのシステムでは与えられる知識文から推論によって新しい知識文を得ることが必要となる．また，語意についても，より詳細な記述が必要である．例えば，「猫の目が光った」という場合に「猫」が一般的な概念としての猫であるのか，ある特定の猫であるのかを区別して表現したい．また，「光った」に対しても，「光ること」という意味の語意に加えて，それが過去の出来事であったことを追加情報として表現したい．検討の結果，Sowa の提案する概念グラフがこれらの要求を満足することがわかった．そこで，本研究では SAGE の格フレーム群から概念グラフへの変換を行うシステム AEGIS の開発を行った．概念グラフは「語意」をノード，「語間の深層格」または「時制や様相などの属性」をアークとするグラフ表現であり，「語意」を語の指す実体が属する型(type label)と，実体そのものに対する参照(referent)の組で表すグラフであると定義されている．ここで referent は，語が特定の実体を指す場合は個体識別子としてシリアルナンバー「#n(n は一意の数)」もしくは一意な文字列をとり，特定の実体を指さない一般の概念である場合は一般識別子「*」をとる．即ち，概念グラフの一般形は

```
((属性)→[type label: referent])
←(深層格)←((属性)→[type label: referent])
```

となる．これに対して，AEGIS は概念グラフをフレーム形式で表し(以下 CG フレーム)，その一般形を

```
frame(フレーム番号,type label(ID 表記),
      type label(日本語表記),referent,
      文節日本語表記,品詞,共起関係子,
      [深層格],[属性],文節番号).
```

とした．CG フレームは概念グラフと SAGE フレームの双方の情報を持たせてある．これは概念グラフが知識表現に重点をおいており自然語文章への変換が困難であるのに対し，SAGE フレームは自然語文章を構成するための情報を豊富に持っているからである．CG フレームによる知識表現の例を以下に示す．

「私は東京を知らない．」

```
frame(1,'4449ee','首都である地名','東京',
    '東京','JN2','を',[],[],[1,2]).
frame(2,'0e7e95','わたし','*',
    '私','JN1','は',[],[],[1,1]).
frame(3,'1e855e','存在を認める','*',
    '知らない','JVE','none',
    [[object,1],[agent,2],[main,3]],
    [[not,3]],[1,3]).
```

AEGIS が type label と referent を決定する際の規則を以下の表 3.にまとめた．

表 3.type label と referent の決定規則

	type label	referent
一般概念	主辞の概念ID	＊
固有概念	主辞の一階上位概念ID	文節の日本語表記
同一概念	主辞の概念ID	シリアルナンバー

ここで，照応解析によって先行詞，ゼロ代名詞，限定詞とみなされた語とそれらが指す語を同一概念と呼び，一意なシリアルナンバーを与える．

6.おわりに

本研究で開発した SAGE2003 の意味解析の精度を調べる評価実験を EDR コーパスおよびインターネットや新聞記事などから無作為に抽出した文章に対して行った．

表 4.SAGE2003 の評価実験の結果

	コーパス101文		新聞・インターネット記事	
	語意の正誤	格の正誤	語意の正誤	格の正誤
SAGE2002	89.3%	89.2%	57.5%	57.4%
SAGE2003	90.2%	90.0%	87.0%	86.8%

表 4.に示したように本年度の研究によって，SAGE が一般の文章の解析に耐え得る精度を持つことが証明された．

謝辞

本研究の一部は文部科学省科学研究費基盤研究 C『日本語文章の常識を用いた意味理解・文脈理解システムの開発研究』の補助金を用いて行われました．ここに感謝いたします．

参考文献

[1] 井村 裕，沓掛 俊樹，佐藤 直美，原田 実: “意味解析システム SAGE の Web 化と連体・使役・受身における解析精度向上", 情報処理学会自然言語処理研究会研報, 03-NL-153, pp.27-63(2003).

[2] John F. Sowa: “Conceptual Structures: Information Processing in Mind and Machine”, Addison-Wesley, Reading, MA(1984).