青山学院大学　理工学部　情報テクノロジー学科

2007（平成 19）年度卒業論文要旨

意味解析システム SAGE の精度向上と モダリティの付与と辞書更新支援系の開発

梅澤　俊之（15804013）　西尾　華織（15804054）
原　田　研　究　室

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，EDR 電子化辞書[1]に記載された情報を元に，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[2]の研究開発を行ってきた．

本研究においては，クレーム分類，要約などの応用研究において，文中の語意には現れない話し手の認識や態度など文の意図を把握する必要性があることから，単文を対象に，文の発話者の命題（文の主要部分）に対する認識や，発話態度を表すといったモダリティの分類と付与を行う．さらに Web 文書には顔文字や記号などが多く見られるが，従来の SAGE では顔文字が正しく解析されなかった．しかし，これらは話し手の感情を表しているため無視できない．そこで顔文字への語意付与に対応した精度向上を行う．また，辞書保守の面では，従来の辞書更新作業の煩雑さや誤入力の可能性を解消する為に，統合辞書更新支援ツールの開発を行う．

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．

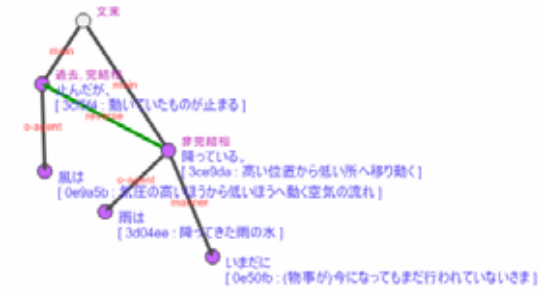

図 1：SAGE の意味解析結果を示す意味グラフ

3. システム概要

SAGE の解析手順を，図 2 を用いて説明する．

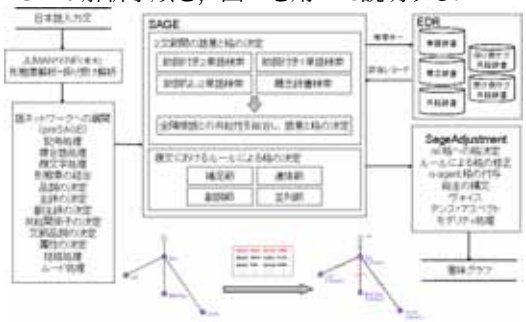

図 2：SAGE2007 における処理の流れ

SAGE の処理に移る前に，まず JUMAN と KNP[3]によって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い，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,ゆ,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 格による意味グラフ出力結果例

4. SAGE の精度向上

SAGE2006 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 3 点において特に重点的に研究した．

4.1. モダリティの付与

4.1.1. モダリティの定義

益岡[4]や仁田[5]によると，文は「命題」と呼ばれる客観的な事柄を表す領域と，「モダリティ」と呼ばれる話し手の命題に対する主観的認識や発話態度を表す領域から構成される．モダリティは命題述部の語尾に現れる．

今日は、雨が降る　らしいよ。
命題　モダリティ

本研究においては，モダリティを大きく，話し手の命題に対する主観的認識を表す「判断のモダリティ」と発話態度を表す「発話のモダリティ」，命題実現の程度を表す「程度のモダリティ」という三つのカテゴリに分ける．判断のモダリティは，さらに二つのカテゴリに分かれ，命題を確かなものとして捉えるか不確かなものとして捉えるかといった「真偽判断のモダリティ」と，命題の実現を望ましいものとして捉える「価値判断のモダリティ」から構成される．

判断のモダリティについては益岡を基に 7 種類，発話のモダリティのついては仁田を基に 18 種類，一方，程度のモダリティについては本研究で 5 種類に分類する．

4.1.2. モダリティの判定

まず，モダリティを付与する文節だが，基本的にモダリティが現れるのは文の述語節である．従って，判断のモダリティと程度のモダリティについては主節と接続節(副詞節，連体節，補足節，並列節)に付与する．一方，発話のモダリティは接続節には現れない，これは発話行為の基本単位として文が存在しているからである．よって，発話のモダリティについては主節の述語節にのみ付与する．SAGE における述語節には，文節品詞として動詞節，動名詞節，形容詞節，形容動詞節，断定節，形容名詞節，形容動名詞節，断定名詞節がある．

次に，具体的なモダリティの判定方法について説明する．判断・程度のモダリティは述語語尾に現れる為，述語語尾を構成する形態素およびその品詞・活用形からモダリティ表現形式を抽出し，判定する．

発話のモダリティでは述語の語尾に加え，聞き手の存在が重要になるため，表現形式の抽出に加え，主格の人称，主題の有無により判定する．

これらは JUMAN・KNP の形態素解析と係り受け解析，及び SAGE の意味解析の結果を用いることにより，ルールによる判定が十分に可能である．動詞の活用形，助動詞，接尾辞は JUMAN 出力の品詞，活用形から，主格の人称は SAGE 出力の述語節から agent 格，o-agent 格または a-object 格でつながる文節の主辞の概念から，主題の有無は主格の助詞の種類によりそれぞれ判断する．

以下に例を示す．

・外は寒いだろう。
　助動詞：だろう　⇒　断定保留
・私が行きましょう。
　活用形：意志形　主格：１人称　主題：無
⇒ 肯定の意志
・この話は信頼しがたい。
　接尾辞：がたい　⇒　困難

4.2. 解析精度向上

4.2.1. 顔文字への語意付与

「(^-^)」など，WEB 文章に多く見られる顔文字は，喜怒哀楽などの感情を持つと考えられる．その為，本研究では，辞書に概念（「喜びを表す顔文字」など）8 種類と，それらの概念を持つ顔文字 178 種類を追加する．

また，JUMAN・KNP で，顔文字の形態素が分割して解析された場合は，顔文字を構成する形態素を結合し，前文節につなげる．

4.2.2. 記号，顔文字の終止符としての役割

SAGE2006 では，「。」と「．」を終止符として認識し，JUMAN・KNP で解析をする前に，入力文を区切っている．本研究では，顔文字や記号（例：☆）も終止符の役割を持つと考え，区切り文字として処理する．

4.3. EDR 辞書更新支援系の開発と辞書の改良

本研究では，EDR 辞書ファイルを SQL データベースに展開して，更新・管理し，そこから SAGE で必要とする辞書データを含む EDR シリアライズ辞書を生成している．辞書データは，誤りや不要レコードなど，追加・更新を要し，これを簡易に行う必要がある．

本研究で，開発した統合辞書更新支援システムは，従来の辞書更新処理で使用していた解析結果視覚化ツール，概念辞書視覚化ツール，そして，辞書情報検索ツールを一つの画面上で利用できるよう統合し，更新ファイル作成を行う修正フォームを追加した．修正フォームは，更新情報の入力を制限できる．また，概念辞書視覚化ツールに，２概念間の共通上位概念検索機能を追加した．

この他，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．

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，SAGE2006 と比較した結果を表 1 に示す．

表 1：SAGE2007 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)
SAGE2006	95.2%	87.0%	99.4%	3
SAGE2007	95.7%	87.9%	99.8%	3

SAGE2007 では語意，格，主辞・副主辞選定の精度が向上した．解析速度においては，十分な速度を維持しており，効率的でより有用性が高い意味解析システムを実現したといえる．

また，モダリティの付与により，文中の語意に現れない意図把握の精密化が可能となり，応用研究での有用性が高まった．

辞書更新支援系の開発においては，煩雑だった更新作業を解消し，更新情報の誤入力を防止することで，より簡易な辞書データの更新が可能となったといえる．

今後の課題として，語意・深層格の更なる精度向上，及び引き続き辞書データを整理することに焦点をあて，利便性の向上を目指す．

参考文献

[1] (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(2002)
[2] 青木 洋, 川口 純一, 原田 実：”意味解析システム SAGE の精度向上”，青山学院大学 2006 年度卒業論文
[3] 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/
[4] 益岡隆志：”日本語モダリティ探求”，くろしお出版，(2007)
[5] 仁田義雄：”日本語のモダリティと人称”，ひつじ書房(1991)