1J-2

意味解析システム SAGE の精度向上と利便性の向上

杉村 和徳† 山本 哲哉† 木村 健太郎† 鳥居 隼† 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

本研究では，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を生成するシステム SAGE[1]の大幅な改訂を行い精度向上と利便性の向上を図った．具体的には，単語の語意と深層格を決定する際の辞書引きを大幅に改善した．特に単語間の辞書引き後の各語の語意と格に関する曖昧性を係り受け関係にある語との共起頻度を基に総合的に決定する高速のアルゴリズムを開発した．また，従来は無視していた引用括弧や丸括弧などの記号も解析の対象にし，副詞節の深層格の決定なども詳細化した．さらに，システム的には，従来バッチ型処理だったのをクライアントサーバ型にして他システムからも通信で解析処理を受け付けられるようにした．

2. 意味解析システム SAGE2004

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 文節の主辞および他の主要な形態素の語意と主辞間の深層格を決定し，下記に示すような形で，文節に対する frame と形態素に対する sframe からなる意味グラフを生成することである．ここで主辞とは，文節の主要な語を指す．また，文の主要な概念として文中における最後の文節(並列節がある場合は，最後の文節と並列関係にある全ての文節)には文末フレームを示す eos より main 格を付与する．格の向きは，係り先→係り元とする．例えば，「摘む。」は下記の事例の様に展開される．

frame出力形式：
frame(FrameID, 解析器種, 文中表記, 共起部表記, 文節語意ID, 語意説明, 文節品詞,主辞FrameID, 副主辞FrameID, [構成要素のFrameID(,)+], [[深層格, 係り側FrameID]*],Referent, Attributes, [文ID, 文節ID]).
sframe出力形式 ：
sframe(FrameID, 解析器種, 文中表記, 読みカナ, 基本形, 語意ID, 語意説明, 解析器依存の詳細品詞, 解析器依存の品詞大分類, EDR詳細品詞, EDR品詞大分類, 解析器依存の活用型, 解析器依存の活用形, [文ID, 文節ID, 形態素ID]).
事例：
frame(4, J, 摘む, ツム, 摘む, 0febdb, (生えている物の一部を)つまみ取る, 動詞節, 5, -, [5, 6], [[object, 1]], [], -, [], [1, 2]).
sframe(5, J, 摘む, ツム, 摘む, 0febdb, (生えている物の一部を)つまみ取る, -, 動詞, 動詞, 動詞, 子音動詞マ行, 基本形, [1, 2, 1]).
sframe(6, J, 。, 。, 。, 000000, -, 句点, 特殊, 記号, 記号, -, -, [1, 2, 2]).
eos(7, [[main, 4]], [1, 3]).

3. SAGE2004 のシステム構成

SAGE2004 は，図 1 に示すように日本文の形態素解析と係り受け解析をされた結果を入力し意味解析を行い概念グラフ[2]を拡張した意味グラフを生成する．SAGE2004 では，EDR 辞書を引く際の手法に改良を加え，入力された文章が補足節，副詞節，記号による修飾，括弧による修飾を含む場合，新規に作成した決定規則を適用することで精度向上を図った．

Improvement of the precision of the semantic analysis system SAGE, and generation of Semantic Graph
Kazunori Sugimura†, Tetsuya Yamamoto†, Kentarou Kimura†, Hayato Torii†, 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.

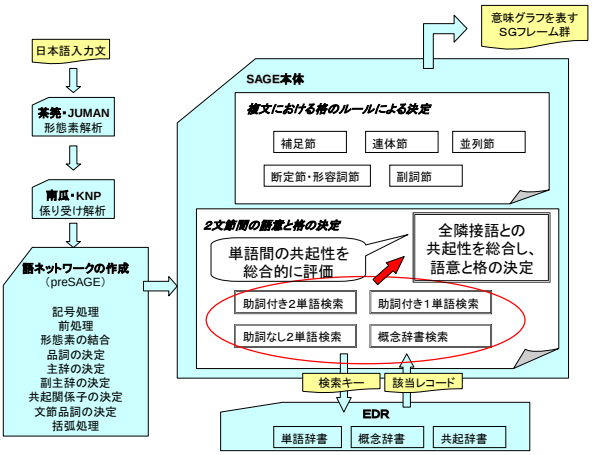

図 1. SAGE2004 の処理の流れ

4. SAGE の精度向上

4. 1. 語ネットワークの作成

係り受け解析結果を，文節がノード，係り受け関係がアークのネットワークに展開する．なお，文節を構成する形態素も文節の子ノードとして展開する．展開後，各文節に対し，図 1 左下ボックス内に示すような処理を行う．

4. 2. 辞書引きの精度の向上

係り受け関係にある 2 単語と助詞を対象として，以下のように EDR 辞書を検索し，2 語の語意とその間の深層格からなる 3 つ組と，その組み合わせの頻度に基づく優先率からなる語意-格総合評価値配列 P を求める．

1)[助詞付き 2 単語検索]共起辞書を，受け側文節主辞，係り側文節主辞，係り側文節の共起関係子で検索し，該当レコードの出現頻度を元に語意格組の語意-格総合評価値配列 P’を作成する．

2)[助詞付き 1 単語検索]上記 1)での該当レコードの有無にかかわらず，共起辞書を受け側文節主辞と係り側文節の共起関係子で検索し，仮の語意-格組を得る．この係り側語意と，係り側文節の主辞について単語辞書を表記と読みで検索したレコード上の語意の類似度を用いて仮の優先率を修正し，語意-格総合評価値配列 P1 を作成する．なお，この際，共起事例での係り側語意と単語辞書での係り側語意の類似度に，1/総組み合わせ数と，語意の頻度を乗じる．同様に，係り側文節主辞と係り側文節の共起関係子で助詞付き 1 単語検索による語意格組の語意-格総合評価値配列 P2 を計算する．P1 と P2 の両方が存在した場合は，その和集合である P”を作成し，どちらか一方のみ存在する場合は，それを P”とする．なお，単語辞書引きの際に，表記，読みをキーにして検索するが，これらでレコードが存在しない時は，読みで検索し，該当するレコードを求める．この結果，語意-格総合評価値配列 P は 1)で得られた P’と 2)で得られた P”の単純和とする．

3)[助詞なし 2 単語検索]1)と 2)で語意-格総合評価値配列が得られなかった場合，共起辞書を，受け側文節主

辞と係り側文節主辞で検索し，該当レコードの出現頻度を元に語意-格総合評価値配列Pを作成する．

4)[概念辞書検索]上記の 1)，2)および 3)で，語意-格総合評価値配列が得られなかった場合，単語辞書，概念体系辞書と概念記述辞書を用い P を計算する．受け側文節主辞と係り側文節主辞をそれぞれ単語辞書で検索し，該当レコードに対して概念体系辞書を使い，上位概念まで調べ，それらの語意の持つ概念体系を元に語意-格総合評価値配列Pを作成する．

4.3. 語意の決定

語ネットワークを，根からスタートして縦型探索し，各文節頂点にてその頂点を退出する際に，辞書引きにより作成された語意-格総合評価値配列を元に，訪問した頂点が表す文節の主辞の語意を決定する．

4.4. 格の決定

係り受け関係にある 2 文節間毎に作成された語意-格総合評価値配列 P を元に，4.3 で確定された語意を持つ語意-格組を切り取り，その組で与えられる格を採用する．この時，切り取った結果，語意-格組が 1 つに定まらない場合は，P 内の優先率の高い組の格を採用する．また，語意-格組が 1 つもない場合は，切り取るのではなく，概念体系辞書を使い，4.3 で確定された語意と P 内の組の語意の類似度の最も高い語意-格組の格を採用する．

4.5. 補足節の解析の向上

補足節には形式名詞「こと･ところ･の･とおり」を伴う形式と，疑問表現「か･かどうか」を伴う形式の２つがあり，語意を EDR の共起辞書で正しく定めることは共起事例が少なく難しいので下記のルールで決定する．なお，補足節を構成しているかは，この後の形態素が助詞であって(もし助詞がなければ副詞節となることが多い)，その前の形態素の品詞や活用形などが以下のようであることから判断する．

・ [rule01]「こと･ところ･の･とおり」が受け側
辞書引きする前に，EDR 品詞が名詞である受け側の「こと･ところ･の･とおり」の語意を補足節を形成する形式名詞の語意”0e84ad”に固定し，助詞付き 1 単語検索と概念辞書検索を行い，係り側の語意を決定し，深層格を「modifier」に固定する．

・ [rule02]「こと･ところ･の･とおり」が係り側
係り側の語意を rule1 と同様の語意に固定し，受け側の語意･深層格を決定する．

・ [rule03]「か･かどうか」が受け側
受け側の語意を疑問表現の補足節を形成する助詞の語意”2621d5”に固定し，rule1 と同様の処理を行う．

・ [rule04]「か･かどうか」が係り側
係り側の語意を rule3 と同様の語意に固定し， 受け側の語意･深層格を決定する．

4.6. 連体節の解析の向上

連体節とは，名詞を修飾する働きをするものであり，“この小説を書いた作家”のような場合，被修飾名詞“作家”が連体節中の述語“書いた”に対する補足語の関係にあり，修飾関係を表す modifier 格(作家→modifier→書いた)と被修飾子の修飾子に対する役割を表す格(書いた→agent→作家)の両方を深層格として求める．

4.7. 副詞節の解析の向上

副詞節は，文節の係り受けの観点から見れば，述語節が述語節に係るということになるが，EDR の共起辞書には述語節間の深層格のレコードは極めて少なく，辞書引きでこの間の深層格を決定できない．よって，係り側と受け側双方が述語節の場合，係り側の共起関係子あるいは文節の末尾の形態素が特定の条件を満足する場合に，すでに確定している格を置き換える．(例:「けれど」を含む場合，深層格を「reverse」に置き換える．)

4.8. 記号を含む文章の修飾の解析

「南瓜」や「KNP」による係り受け解析の結果，記号が単独で文節となる場合がある．記号は EDR 共起辞書で検索できず，記号のみで構成されている文節があると不具合を起こす可能性がある．そこで“昨日は何時間寝ましたか？”の場合“？”が独立した文節になってしまうのを“寝ましたか？”のように，直前の文節の末尾に記号を移す処理を行い，辞書引きの際は記号を除去する．

文を区切る”;”や”:”を含む文の係り受け解析結果を“例：／この／川の／幅は／５０～６０ｍです。”のように記号の後で文節が切れるようにし，“;”や“:”で終わる文節の共起関係子を“として”に設定する．

4.9. 括弧を用いた修飾の解析

括弧に関する処理を表１のように分類した．例えば分類番号 1～4 に対しては，「田中氏(46)は東京に行った。」では“田中氏”が“（46）は”に係り，“(46)は”が“行った”にかかっていたところを，“(46)は”から“田中氏”に係り，“田中氏”から“行った”に係るように変更すると共に，“田中氏”の共起関係子を“(46)は”の共起関係子にし，“(46)は”の共起関係子を“の”に変更する処理を行う．

また，5～8 に対しては，“「楽しく音楽を作る」と言った。”では“「楽しく”と“音楽を”が“作る」と”に係り，文節“作る」と”が“言った.”に係るので，解析器の出力結果をそのまま採用しても係り受けに問題は生じないので，特別な処理は行わない．

なお，9，10 については“(1)序章”のように括弧記号が用いられることを発見したに留める．

表 1.括弧の分類と判別方法

番号	種類	判別ルール
1,2,3,4	補足,属性, 換言,代替	括弧内が名詞節のみで括弧の直前の EDR 品詞が名詞か数詞
5,6	名称,強調	括弧内が名詞節のみで括弧の直前の EDR 品詞が助詞
7,8	発言,引用	括弧内が名詞節･動詞節などの文を構成する諸節が混合して成り立っている
9,10	箇条, 箇条引用	括弧内が数詞のみで箇条書きの文章で用いられる

5. おわりに

本研究で開発した SAGE2004 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 2.SAGE2004 の評価実験の結果

	コーパス101文		インターネット101文	
	語意の正解率	格の正解率	語意の正解率	格の正解率
SAGE2003	90.2%	90.0%	87.0%	86.8%
SAGE2004	96.1%	93.1%	92.5%	90.2%

表 2.に示したように本年度の研究によって，SAGE の解析精度が実用域に達したことが検証された．

参考文献

[1] 前澤 敏之，面来 道彦，上野 雅和，韓 東力，原田 実:“意味解析システム SAGE の高精度化と概念グラフへの変換”，情報処理学会第 66 回全国大会論文集,6U-05,第 2 分冊 pp.177-178 (2004.3).
[2] John F. Sowa:“Conceptual Structures: Information Processing in Mind and Machine”, Addison-Wesley, Reading, MA(1984).