

日本語文における連体修飾要素の意味構造解析手法

納 富 一 宏¹・石 井 博 章¹

¹ 情報工学科

Semantic Analyzing Method of Adjective Parts in Japanese Sentences

Kazuhiro NOTOMI¹⁾, Hiroaki ISHII¹⁾

Abstract

In this article, we propose a new semantic analyzing method of adjective parts in Japanese documents. The method uses a recursive structure, as a semantic representational structure, to recognize adjective parts in a sentence. The data structure, which was defined by us, has some attributes of semantic features and syntactic features of adjective parts. One of these is called '*Semantic ID*', and the other is called '*Grammar ID*'.

With the method, we developed an analyzing system called '*Modifier Analyzer*', which can detect modifiers, their target words, and relations of them in Japanese sentences. And we implemented this system on a web server as a CGI application, which uses HTML tags for user-interface, therefore the '*Modifier Analyzer*' can be used with a general web browser on any platforms. Then we attempted evaluating the performance of the '*Modifier Analyzer*'. As a result, it was confirmed that the method has a high reliability for semantic and syntactic analyzing.

Key Words : Natural Language Processing, Semantic Analysis, Adjective Parts, Case Analysis, Proofreading

1. はしがき

日本語における係り受け関係には、副詞の呼応や接続詞の呼応の他に、連体修飾、連用修飾などがある。

一般に、自然言語文において、係り受け関係の構造を調べることは、精緻な構文解析を実現する上で必要である。特に、続く後段の意味解析フェーズに対し、有効なデータ構造をパスする上で重要となる。

従来の統語的な解析手法では、多くの別解を与えるような係り受け関係の構造は、多義性問題として意味解析フェーズに頼るという方針をとってきた。この理由は、複雑な構文規則を構文解析フェーズでは扱わないようにするためである。

例えば、①「赤い彼の顔」、②「赤い顔の彼」、③「彼の赤い顔」、④「顔の赤い彼」という4例文は、いずれも3文節4品詞（形容詞×1、名詞×2、格助詞「の」×1）からなり、同じ意味構造を持つと考えられるが、それぞれ異なる構文木 (parse-tree) を与える。

極端に多くの非終端記号の仮定なしには構文規則の記述

が困難である状況では、意味解析フェーズを早い段階で導入し、係り受け関係の多義性を極力排除していくべきである。

特に、連体修飾の構造は、意味解析において文に出現する名詞概念を制限する働きを持つことから、意味表現構造の定義およびその利用が重要であると考えられる。

筆者らは、以前から日本語を対象とする文書校正・推考支援システムの開発を行ってきた^{[1]~[4]}。この中で、日本語の文節表現構造として「JFK 構造」を提案した^[3]。

本稿では、日本語文における連体修飾要素の意味構造解析を実現するためのデータ構造として、JFK 構造を内部に含み、連体修飾関係を扱うためのモデルとなる「意味表現構造」を定義した上で、アルゴリズムを提案する。

また、この手法を用いた解析システムの設計および実装例を示す。さらに、システム動作例と評価実験の結果を示して、本手法の有効性について考察する。

以下、最初に連体修飾の原理について触れる。

2. 連体修飾

2.1 連体修飾関係

修飾語と被修飾語の関係のうち、被修飾語が体言となる場合を「連体修飾関係」と呼ぶ。連体修飾関係を構成する際の修飾語となり得る品詞(文法属性)を列举すると、①形容詞(連体形)、②形容動詞(連体形)、③連体詞、④体言+格助詞「の」、⑤体言+助詞類、⑥文形式(連体形)となる。

これらを例と共に表 2.1 にまとめる。

表 2.1 修飾語となり得る品詞(文法属性)

No.	品詞(文法属性)	例
1	形容詞(連体形)	丸い → 石
2	形容動詞(連体形)	静かな → 部屋
3	連体詞	大きな → リンゴ
4	体言+格助詞「の」	日本+の → 国旗
5	体言+助詞類	デバイス+からの → 信号 回路+における → 電圧
6	文形式(連体形)	太郎が買った → 本

表 2.1 において、No.1~3 については単純な依存関係として捉えられるので特に問題はない(2.2)。しかし、No.4~6 については意味表現構造が複雑になる場合が多く、解析の際に注意が必要となる。これらの点については後述する(2.3, 2.4)。

2.2 制限属性としての連体修飾関係

連体修飾関係の基本型を図 2.1 に示す。ただし、ノードは自立語要素を、また、リンク(矢印)の根元は修飾語を、リンクの先は被修飾語をそれぞれ表す。

例：大きなリンゴ

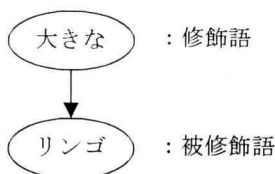


図 2.1 基本的な連体修飾関係

修飾語は被修飾語に対する意味の限定であり、一種の制限属性であると考えられる。

従来の統語的な解析では、最初に与えた構文規則に合致した構造を与えるだけであり、制限属性がどの語要素に対して働くかといった依存構造関係をうまく表現できない場合がある。よって、意味解析というレベルでは、別の表現手法が必要になる。

文の要素は1次的に配置されるので、依存関係は局所的に見れば、必ず図 2.1 に示す線形構造(1対1対応)をとること

になるが、意味表現構造としては2次的な広がりを持つ。

例えば、連体修飾関係では、前置される修飾語が列举された並立的な構造をとる場合がある。これを図 2.2 に示す。

例：丸い大きな赤いリンゴ

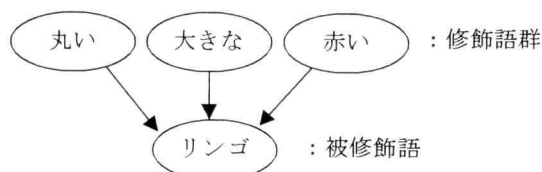


図 2.2 列举型の連体修飾関係

また、複数のターゲットを並列的に修飾する構造をとる場合がある。これを図 2.3 に示す。

例：大きなリンゴとミカン

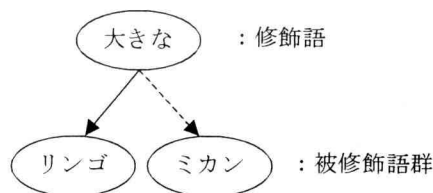


図 2.3 並列型の連体修飾関係

上図に示した構造では、修飾語と隣接した第1ターゲット(「リンゴ」)に強い依存関係があり、隣接していない第2ターゲット(「ミカン」)に弱い依存関係があるものとしている^{※1}。

さらに、表 2.1 における No.4, 5 のタイプの連体修飾関係では、直列階層型の構造を与える場合がある。これを図 2.4 に示す。

例：デバイスからの信号の有無による判定

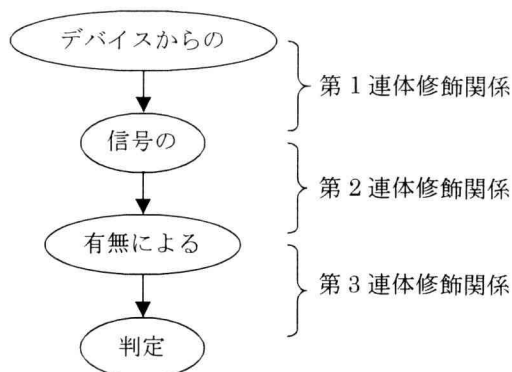


図 2.4 直列階層型の連体修飾関係

※1 意味解析の常として、局所的な部分構造だけで発話内容が判断できるわけではない。「大きな」は「リンゴ」のみを修飾する場合もある。また、「リンゴ」と「ミカン」の両方を同程度に修飾する場合もある。

図 2.4 に示した構造では、複数の連体修飾関係の連鎖が構成されており、それぞれの係り受け関係は、図に示した以外の意味構造をとらない点が特徴である。

2.3 意味の曖昧性

一般に、連体修飾関係は、先に示した構造（図 2.1～2.4）のいずれかになるわけだが、文脈情報を利用しなければ、構造を一意に決定できない場合がある。構造が一意に決定できなければ、複数の意味構造が存在することになり、したがって、曖昧性（多義性）の問題が生じる。これを文脈依存性問題と呼ぶ。

意味の曖昧性（多義性）は、意味構造の違いとして捉えることができる。図 2.5 に連体修飾に関する曖昧性の例を示す。

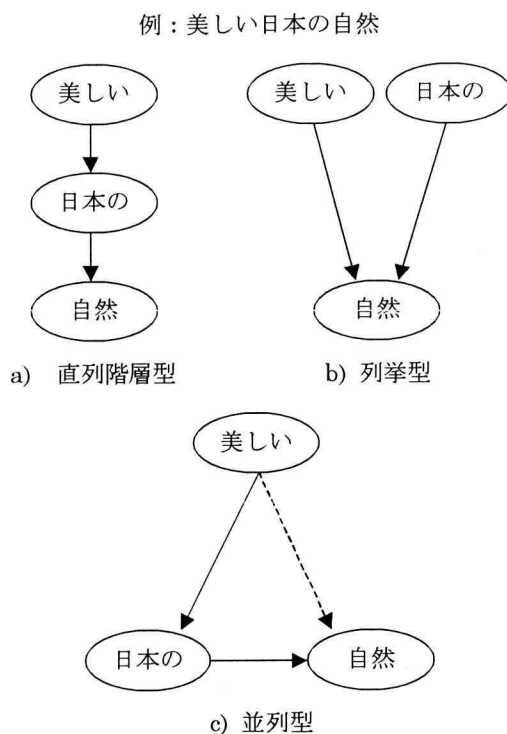


図 2.5 意味の曖昧性

この例では、「美しい日本の自然」という表現が、どのような連体修飾関係を構成し得るかを示しており、修飾語「美しい」が、どの自立語に係るのかによって意味構造に違いが生じる。

すなわち、①「日本」にのみ係る場合（直列階層型）、②「自然」にのみかかる場合（列挙型）、③「日本」および「自然」にかかる場合（並列型）、の 3 種類の解釈が可能である。

2.4 格助詞「の」による連体修飾

国文法における格助詞「の」の連体修飾用法は、一般に、複雑であり、意味的には表 3.1 のように分類することができる。

表 3.1 格助詞「の」の連体修飾用法の意味的分類

機能	意味	例
所有	～が所有する	私の教科書
所属	～に属する	大学の学生
同格	～である、～という	教育者の父
性質の状態	～である	緑色の表紙
存在の場所・時	～における	夏の風物詩
材料	～でできている	金の指輪
目的	～のための	試験の勉強
原因	～による	台風の被害
関係	～に関する	コンピュータの話
体言化要素の主語	～が	母の帰りを待つ

これら「AのB」という形で表現される意味構造の差異を正しく解析するためには、名詞意味素性を用いて細分類されたソーラス情報が必要であり、一般には、解析コストがかかりすぎるため、何らかの妥協が必要な場合が多い。

3. データ構造と解析手法の提案

自然言語で記述された入力文字列に対応する統語的および意味的構造を「意味表現構造」と呼ぶ。

ここでは、連体修飾関係の意味表現構造として、計算機への実装を考慮したデータ構造について述べる。また、解析手法として、入力文（自然言語）からデータ構造への素性値取得および格納手順について言及する。

3.1 意味表現構造

先に述べた「連体修飾関係」を表現するための意味構造は、①形態素レベルの表層ラベル、②意味属性、③文法属性、および、④他の要素へのリンク関係とそのリンク強度、により表現することができると考えられる。

ひとつの意味表現構造は、最小文節を単位としたひとつの文節に対応づけ、連体修飾関係そのものは、複数の文節をリンクで結合した表現をとる。文節の表現構造としては、筆者らが以前から提案している JFK 構造^{[1]~[4]}を用いる。

JFK 構造における J は「自立語要素部分」を意味し、意味属性を取得するためのターゲットとなる。また同様に、F は「付属語要素部分」を意味し、文法属性を取得するためのターゲットとなる。

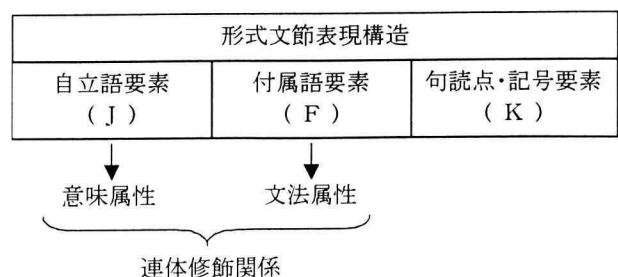


図 3.1 JFK 構造による解析

ここで提案する意味表現構造の具体例を図 3.2 に示す。

例：美しい日本の自然

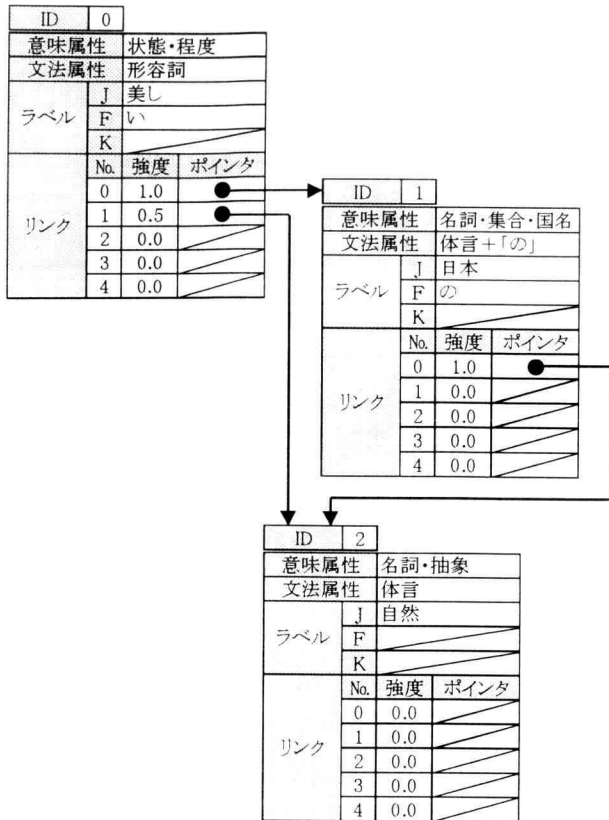


図 3.2 連体修飾関係の意味表現構造の例

図 3.2 について補足する。

それぞれの意味表現構造は、個別に ID 番号を持ち、この番号で管理される。また、表層情報(入力文字列)の格納領域としてラベルがあり、JFK 構造^{[1]~[4]}を保持する。ラベル J, F, K には、自立要素、付属要素、記号要素がそれぞれ格納される。さらに、自立要素に対する意味属性および文法属性が格納される。これにより意味表現構造としての参照性を上げることができる。

他の意味構造へのリンクは、No.、結合強度、ポインタを1セットとした単位で保持される。リンク結合は、連体修飾関係として、直列階層型、列挙型、並列型の3つの構造(図 2.1~2.5 参照)を表現することができる。

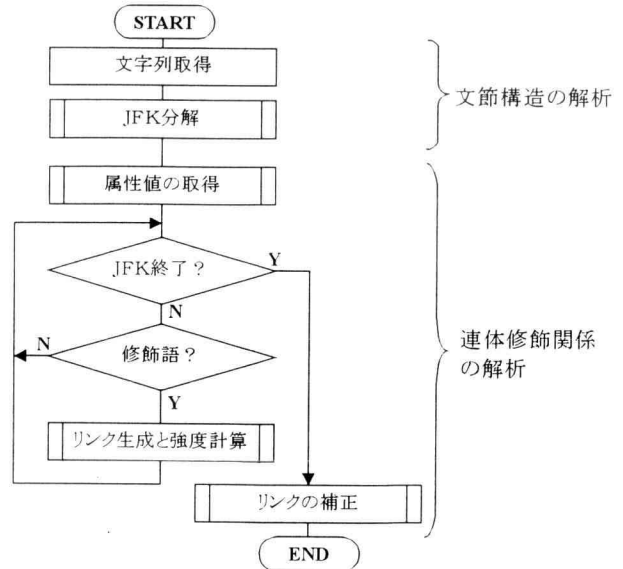
結合強度は 0 から 1 までの実数値をとり、連体修飾の意味的依存優先度を表現する。この値が大きいほど連体修飾は強い依存関係を持つと考える。

結合強度を採用する理由は、連体修飾関係の多義性に対応するためである。解析手法としては多義性の解消を目的にすべきだが、データ構造の表現能力としては、逆に汎用性が要求される点に注意しなければならない。

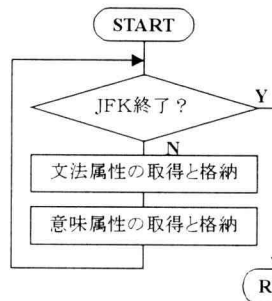
3.2 意味解析手法

最初に、3.1 で述べた意味表現構造を用いた解析アルゴリズムのフローチャートを図 3.3 に示す。

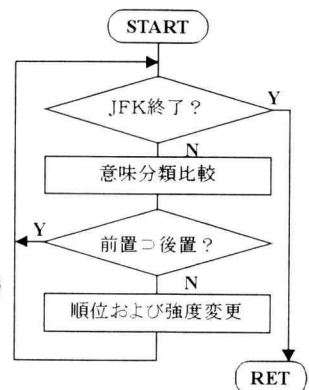
a) メイン処理



b) 属性値の取得



c) リンクの補正



d) リンク生成と強度計算

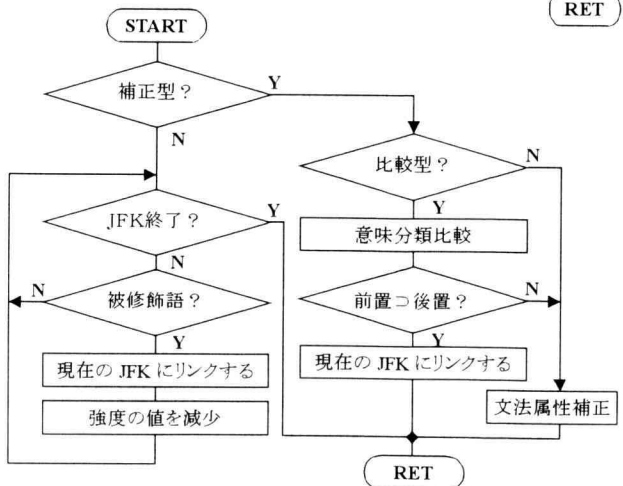


図 3.3 解析手順

以下、図 3.3 について補足する。

a) メイン処理

形態素解析に相当する JFK 分割を行なって、入力文字列 (日本語文) から、 $n (\geq 2)$ 個の文節構造を抽出する。抽出された全ての文節に対して属性値ペアを決定する。この際、 i 番目の文節要素 C_i の属性値ペアを $(S_i | G_i)$ と表記する。

次に、修飾語条件を満たす文節 C_i と、リンクターゲットとなる文節 $C_j (i < j)$ を検索し、リンクを生成すると共に、結合強度を計算して、データを格納する。この操作を全ての文節について行なう。

b) 属性値の取得

文節構造 (JFK構造) は、自立語要素、および付属語要素から構成されるので、前者より意味属性値 (S_i) を、また後者より文法属性値 (G_i) を決定することができる。

c) リンクの補正

最初に作成されたリンクおよび強度を「意味分類比較」(後述)を用いて補正する。

これは、連体修飾関係の解析で問題となる格助詞「の」による多義性を解消するための処理である。例えば、

1. 「長い太郎の脚」
2. 「長い脚の太郎」

という表現において、形容詞「長い」は、統語的な解析では、前置体言か後置体言のいずれか一方に係るという構文規則でしか扱えない。よって、どちらの例文においても、「長い」が「脚」に係るという結果を得るためには、意味分類による前置体言と後置体言の優先順位比較を行ない、その結果により修飾関係を決定しなければならない。

また、2.3 で述べた並列型の連体修飾関係では、結合強度によるリンクターゲットの優先順位が重要となっている。このため、意味分類比較を用いてリンクを変更した場合に、同時に強度も変更する必要がある。

d) リンク生成と強度計算

基本的には、被修飾語を検索し、リンクリストに保持する動作を行なう。リンクが形成される場合は、同時に強度 σ_k が決まり、通常は 1.0 の値を持つ。並列型の連体修飾関係では、修飾語に隣接する被修飾語が最も強度が高く、それ以降のターゲットは、出現位置が後ろになるほど、強度値が減少する。実際には、1.0 から始まり、以降、値が 1/2 倍ずつとなる。JFK 分割により得られた文節数が n のとき、評価式は次のようになる。

リンク強度:

$$\sigma_k = \left(\frac{1}{2}\right)^k \quad 0 \leq k \leq n$$

被修飾語の検索では、文法属性から得られる修飾文節の統語パターンにより、リンクターゲットの判定条件が異なる。

大別すると、補正型と比較型とがあり、前者は連体修飾関係を構成しない可能性がある。また、後者は意味分類比較を用いた判定によりリンクが作成される場合がある。

連体修飾関係を構成しない、すなわちリンクが作成されない場合とは以下の通りである。

- ①修飾語となる文節 C_i が格助詞「の」を付属語要素として持ち、かつ隣接文節 C_{i+1} が動詞類連体形である場合

例: 太郎の読んだ本

- ②修飾語となる文節 C_i が格助詞「の」を付属語要素として持ち、かつ隣接文節 C_{i+1} が形容詞または形容動詞連体形であり、さらに後置文節 C_{i+2} が体言を自立語要素として持つ場合で、しかも C_i の自立部が C_{i+2} の自立部に含まれる意味分類を持つ場合

例: 表紙の赤い本

上記のいずれの条件も、格助詞「の」は主格用法となり、格助詞「が」に置き換えられるため、直接的な連体修飾関係は成立しない。

4. システム設計

入力された日本語文字列から連体修飾関係を自動的に検出し、これを解析した結果を出力するシステム「Modifier Analyzer」を構築した。ここでは、システム構成、および実装について述べる。

4.1 システム構成

システムは、UNIX 上の CGI (Common Gateway Interface) 対応アプリケーションプログラムとして実装した。このため、クライアント側に Web ブラウザを用意すれば、インターネットを経由して本システムを利用することができる。

システム構成を図 4.1 に示す。

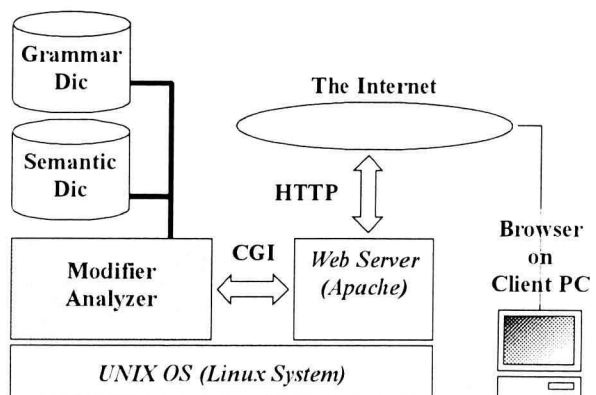


図 4.1 システム構成

4.2 C++による実装例

システムの記述には、C++言語を用いた。連体修飾構造要素のクラス定義の一部(データ部のみ)をリスト4.1に示す。

```
//=====
// クラス: CModifLinkItem
//=====
class CModifItem;
class CModifLinkItem
{
public:
    double      Value; // 結合強度
    CModifItem* Next; // 被修飾要素へのポインタ
};

//=====
// クラス: CModifItem
//=====
class CModifItem
{
public: JFK // JFK データ構造の継承 ...①
{
public:
    enum
    {
        MAX_LINKS = 5 // 最大リンク数
    };
    CSemanticID  Sid; // 意味属性 ...②
    CGrammarID   Gid; // 文法属性 ...③

    // 結合強度による順位が変更されたか否か
    bool IsSwapped;

    // この語が被修飾語であるか否か
    bool IsModified;

    // 被修飾要素の個数
    int Num;

    // 被修飾要素へのリンク
    CModifLinkItem Links[MAX_LINKS];
};
```

リスト 4.1 連体修飾構造要素のクラス定義

リスト 4.1 について補足する。クラス JFK(…①)は、自立部、付属部、句読点・記号部のそれぞれ3つの文字列を格納するデータ構造である。クラス CSemanticID(…②)、および CGrammarID(…③)は、共に列挙型(enum)で定義された値を保持するデータ構造であり、実際にはint型をメンバに持つ。クラス CModifItem およびクラス CModifLinkItem は、互いに依存関係にあり、再帰的なデータ構造を提供している。

5. 評価と考察

ここでは、実装した解析システム「Modifier Analyzer」の動作例を示すと共に、先に提案した手法の評価を行なう。

5.1 動作例

動作画面を図 5.1 および図 5.2 に示す。

システムへの入力としては、文または連体修飾表現を扱うこ

とができる。また、解析オプションおよび結果表示オプションをダイアログから指定することができる。

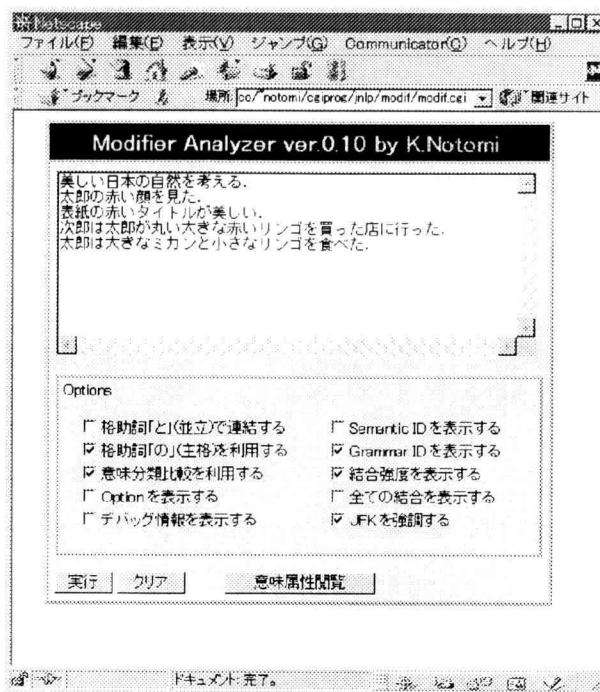


図 5.1 システム (Modifier Analyzer) の入力画面

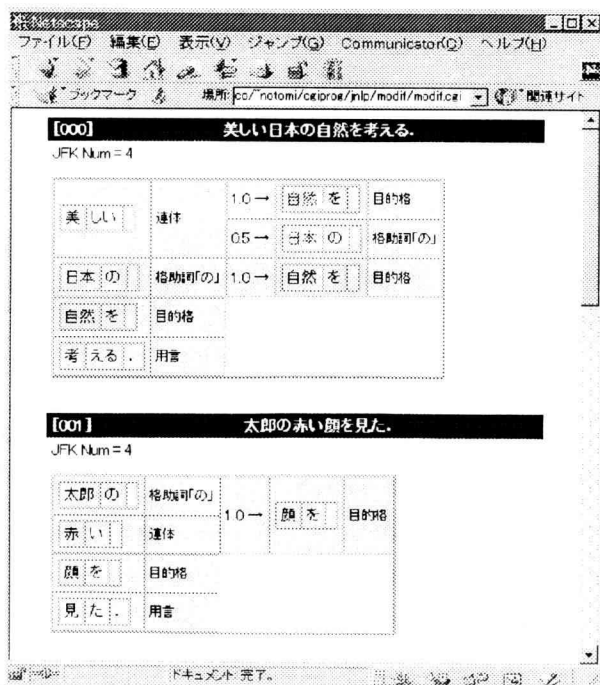


図 5.2 結果出力画面

5.2 評価実験

自然言語処理に関する和文学術論文3本から、「文」を構成する部分のみを抽出し、3つの入力データファイルを作成し、

システムによる解析結果と、人手による解析結果とを比較して評価した。

比較には、次の3点に留意し、評価値として適合率および再現率を計算した。

- ①修飾語要素が正しく検出されたか否か
- ②被修飾語要素が正しく検出されたか否か
- ③連体修飾関係が正しく検出されたか否か

表 5.1 に入力データに含まれる文の数、および形式文節の数を示す。また、表 5.2 に入力データにおける連体修飾関係の種類別出現頻度の割合を、表 5.3 に解析結果を、表 5.4 に評価結果をそれぞれ示す。表の各項目についての詳細は 5.3 で述べる。

表 5.1 入力データのサイズ

File	総文数	総文節数	平均文節数
ki97	157	1,460	9.30
ki98	55	604	10.98
ki99	91	842	9.25
合計	303	2,906	

表 5.2 連体修飾関係の種類別出現頻度の割合

File	連体修飾関係の種類		
	直列階層型	列挙型	並列型
ki97	84%	12%	4%
ki98	81%	12%	6%
ki99	88%	10%	2%
平均	85%	11%	4%

表 5.3 解析結果

File	出力文節数		エラー数		
	Human	System	過剰検出	未検出	誤リンク
ki97	767	719	55	61	16
ki98	337	363	37	18	5
ki99	402	387	38	40	10
平均	502.0	489.7	43.3	39.7	10.3

表 5.4 評価結果

File	評価			
	検出		総合	
	適合率	再現率	適合率	再現率
ki97	0.924	0.920	0.901	0.900
ki98	0.898	0.947	0.884	0.932
ki99	0.902	0.900	0.876	0.876
平均	0.908	0.923	0.887	0.902

5.3 考察

表 5.1～ 5.4 について補足する。

表 5.1 において、1 文あたりの平均文節数は約 9 以上であり、比較的 1 文が長い文章であることが分かる。

表 5.2 において、2.3 で述べた連体修飾関係の意味表現構造を種類別にカウントし、その出現頻度の割合を求めた結果、平均で直列階層型(図 2.5 a)が全体の 85%を占め、列挙型(図 2.5 b)、並列型(図 2.5 c)はそれぞれ 11%、4%であった。この割合は文書ファイルによる差は見られなかった。

表 5.3 において、出力文節数とは、修飾語を含む文節数、および被修飾語を含む文節数の合計値である。この値は、格助詞「の」による連体修飾が連鎖するような場合、同じ文節が修飾文節にも被修飾文節にもなるので、延べ数となっている(異なり数ではない)。

また、エラー数とは、修飾語、もしくは被修飾語の検出において、正しく検出できなかった個数を示している。

過剰検出とは、修飾語、もしくは被修飾語となり得ない部分を誤検出した場合の個数であり、未検出とは、検出されるべき部分が検出されなかった場合の個数である。さらに、誤リンクとは、修飾-被修飾関係のリンクが誤っている場合の個数である。いずれも、人手による判定結果と異なる部分をエラーとした。

表 5.4 において、検出の適合率、および再現率はそれぞれ次式で計算した。また、総合適合率および再現率は、誤リンクを解析ペナルティと考え、以下のそれぞれの式で計算した。

$$\text{検出適合率} = \frac{\text{System 総検出数} - \text{過剰検出数}}{\text{System 総検出数}}$$

$$\text{検出再現率} = \frac{\text{Human 総検出数} - \text{未検出数}}{\text{Human 総検出数}}$$

$$\text{総合適合率} = \frac{\text{System 総検出数} - \text{過剰検出数} - \text{誤リンク数}}{\text{System 総検出数}}$$

$$\text{総合再現率} = \frac{\text{Human 総検出数} - \text{未検出数} - \text{誤リンク数}}{\text{Human 総検出数}}$$

評価値(総合)を見ると、適合率、再現率共に高い値を示しており、本手法の有効性が確認されたものと考えられる。

さらに、検出および総合において、平均して再現率が適合率に比して高い数値を示している。このことから、本システムは検出漏れが比較的少ない解析手法として、その信頼性が示唆されていることが分かる。

本手法、および本システムは、連体修飾関係の構造解析、およびその可視化に役立つものであり、意味解析フェーズを早期に導入することで、係り受け関係の多義性を解消すべく動作する。これは解析コストの低減、および処理時間の短縮に寄与するものと考えられる。

6. むすび

連体修飾要素の意味構造解析について、連体修飾関係の例、およびデータ構造としての意味表現構造と実装例を示し、解析アルゴリズムについて述べた。

また、本手法を用いたシステム「Modifier Analyzer」を UNIX (Web サーバ) 上に実装し、評価実験を行った。評価結果から本手法が日本語文における連体修飾関係を構成する諸要素の検出、およびその関係解析に有効であると結論された。

本手法は、従来の形態素解析から構文解析に至るフェーズに比して、非常に簡便である。

今後は、本手法を連用修飾関係の解析に適用できるよう拡張することが課題である。

参考文献

- [1] 納富, 他:「日本語文書における共起格情報を用いた助詞要素の訂正」, 神奈川工科大学研究報告, B理工学編, 第 23 号, pp.101, (1999).
- [2] 納富, 他:「日本語文書校正支援ツールの開発—共起格情報による助詞要素の訂正—」, 情処第56回全大, (1998).
- [3] 納富, 他:「日本語文書校正支援システムにおける高速統語解析手法」, 神奈川工科大学研究報告, B理工学編, 第 20 号, pp.165-175, (1997).
- [4] 納富:日本語文書校正支援ツール HSP の開発, 情報処理学会デジタルドキュメント研究会報告, (1997).