

オンライン辞書定義文の解析と知識ベース化

酒井 桂一, 中村 順一, 長尾 真

〒606 京都市左京区吉田本町
京都大学工学部電気工学第二教室

要約

現在, 辞書を始めとする自然言語データの機械可読化が盛んに行われている。もし, 既存の辞書やハンドブックなどのデータから各種の知識が(半)自動的に抽出できれば, 自然言語処理のための辞書や知識ベースの作成に有効である。そこで, 辞書データを知識ベースの形に整理する第1歩として, ロングマン現代英英辞典 LDOCE の名詞の定義文の構文解析システムを作成し, それを用いて意味ネットワークを作成する実験を行った。

定義文の構文解析システムとしては, 並列左隅解析法の一つである SAX に優先規則の考え方を導入したもの (SAX+p) を用いた。本稿では, 名詞の定義文の解析のための優先規則と意味ネットワーク作成の実験の結果について述べる。

Parsing Definitions in On-Line Dictionary: toward the Development of Lexical Knowledge-Base

Keiichi SAKAI, Jun-ichi NAKAMURA and Makoto NAGAO

Department of Electrical Engineering, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, 606, Japan

Abstract

Many projects are now carried out to make various natural language data, like dictionaries and handbooks, accessible as an on-line database. If we can extract structured 'knowledge' from such on-line data (semi-)automatically, it helps us to develop a knowledge-base for a natural language processing system. As a first step for this extraction, we have developed the system to construct semantic networks by parsing definitions of noun in LDOCE.

For this purpose, we have used SAX+p as a parser, which is an augmented version of the parallel left-corner parser SAX with a preference rule mechanism. This paper discusses the preference rules to parse definitions of nouns, and the result of construction of semantic networks.

1 はじめに

大規模な自然言語処理システムを実用化するには、特に、そのシステムが使用する単語辞書（一種の知識ベース）の量と質とを向上することが重要である。このことから、大規模な辞書データを作成することが、各所で行なわれている [3]。

この場合、辞書を始めとする自然言語データの機械可読化が現在盛んに行なわれている点を考慮すると、既存の可読辞書やハンドブックが言語データの作成に活用できれば、辞書もしくは知識ベースの大規模化に有効であろう。そこで、既存の辞書を活用する研究が各種行なわれている。例えば、辞書データを構文解析に応用する研究 [1]、辞書を計算機で扱い易い構造に変換する研究 [17]、複数の辞書を統合した大規模辞書データベースを作成する研究 [5]、辞書定義文を利用して前置詞句の係り受けの曖昧さを解消する研究 [4]、日本語の辞書定義文を利用して、シソーラスを作成する研究 [16] などがある。

筆者らは従来から、既存辞書としてロングマン現代英英辞典 (Longman Dictionary of Contemporary English, LDOCE) 1978 年版 [6] を用い、その活用手法についての研究を行なって来た [9, 10, 13]。例えば、文字列のパターン・マッチングにより、LDOCE の名詞定義文から、見出し語と主に上位/下位関係にある「中心名詞」及び、見出し語と「中心名詞」の関係を示す「機能語」を抽出し、その結果を用いて、名詞の意味的階層関係を求めた [11, 12]。しかし、単純なパターン・マッチングによる方法には、当然ながら限界があり、以下の点が問題となった。

1. 「中心名詞」が正しく求まる保証がない [12]。
2. LDOCE 以外の辞書・ハンドブックなどに適用できるかどうかはわからない。
3. 定義文には、中心名詞や機能語以外の情報も含まれているが、それらを取り出すことができない。

そこで、定義文を構文解析するためのシステムを作成し、それを用いて、定義文から知識ベースを作成するための第 1 歩として、一種の意味ネットワークを作成する実験を行なった [14]。

具体的には、LDOCE のデータを Prolog 専用マシン PSI に移植し、PSI 上で構文解析システムを作成した。この構文解析システムには、並列左解析法の一つである SAX [7] に優先規則の考え方 [15, 2] を導入したもの (SAX+p) を用いた [8]。また、構文解析等の処理用辞書として、LDOCE そのもの (見出し語、品詞、文法コード) を用いた。本稿では、名詞の定義文の解析のための優先規則と意味ネットワーク作成の実験結果について述べる。

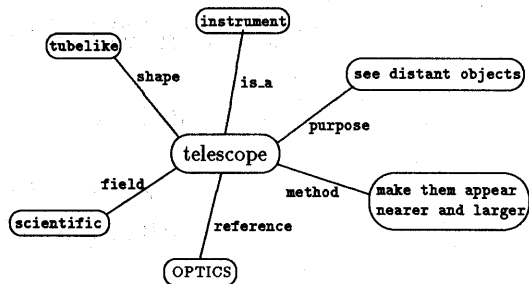
2 LDOCE 名詞定義文の解析

2.1 LDOCE 名詞定義文のリンクの種類

2.1.1 リンクの種類

文献 [11, 12, 13] で名詞定義文から抽出した中心名詞及び機能語は、見出し語と主として上位/下位関係にある単語であった。しかし、名詞定義文には、もっと多様な意味関係が記述されている。そこで、定義文中の各種の意味的関係を表現する方法として、意味ネットワークの考え方を採用することにした。

例えば、名詞 'telescope' の定義文を読めば、図 1 に示すようなネットワークを想定することができる。図は、'telescope' は、上位概念 (is_a) が 'instrument' で、分野 (field) が 'scientific' で、形状



telescope : a tubelike scientific instrument used for seeing distant objects by making them appear nearer and larger see picture at OPTICS.

図 1: 'telescope' の定義文と対応する意味ネットワーク

(shape) が 'tubelike' で、用途 (purpose) が "see distant objects" で、手段 (method) が "make them appear nearer and larger" で、参照すべき定義文 (reference) が 'OPTICS' を表わしている。なお、以下では、見出し語と定義文中の一定の属性を持った語句との関係を「リンク」と呼び、「リンク」に付けたラベル (shape, purpose など) を「リンク名」、「リンク名」を属性として持つ語句を「リンク値」と呼ぶことにする。

リンクの種類をあらかじめ決定しておくことは、簡単なことではない。そこで、修飾語を比較的多数持ち、共通の中心名詞を持つ、主に具象名詞である名詞群のいくつかに関して、リンクの種類とその記述形式の調査を行なった。リンクには、調査した名詞群に共通と考えられるものと名詞群毎に特有なものがあつた。以下に調査結果 (一部) を示す。

2.1.2 調査した具象名詞に共通なリンク

is_a リンク: 見出し語と中心名詞とのリンクを is_a リンクとする。

colour リンク: 見出し語と中心名詞を修飾する色属性を示す形容詞とのリンクを colour リンクとする。リンク値とする形容詞は、LDOCE 定義文中に colour が現れる形容詞とした¹。

size リンク: 中心名詞を修飾する形容詞 'large', 'small' は、見出し語と中心名詞との相対的な大きさの関係を示している。そこで、このリンクを size リンクとする。

purpose リンク: 中心名詞を修飾する "used for" に続く動名詞、及び中心名詞を直接修飾する to-不定詞は、見出し語の「用途」を示している。そこで、このリンクを purpose リンクとする。

function リンク: 中心名詞を直接修飾する現在分詞節、あるいは中心名詞を先行詞とする能動関係代名詞節は、見出し語の「機能」を示している。そこで、このリンクを function リンクとする。

¹第 1 定義文中で colour が用いられている全形容詞を調査した結果、それらはすべて、色を表す形容詞であつた。

表 1: 物質名詞の定義文の例

(a)	plutonium	a man-made simple substance (ELEMENT) that is used esp. in the production of atomic power
(b)	pectin	a sugar-like chemical compound substance found in certain fruits
(c)	metal	any usu. solid shiny mineral substance of a group which can all be shaped by pressure and used for passing an electric current, and which share other properties
(d)	gamma globulin	a natural substance found in the body, a form of ANTIBODY, which gives protection against certain diseases
(e)	jelly	a sweet soft food substance that shakes when moved, made with GELATINE
(f)	amber	a yellowish brown hard clear substance used for jewels, ornaments, etc.
(g)	glue	a sticky substance which is obtained from animal bones or fish and used for joining things together
(h)	celluloid	a strong easily burnt plastic substance made mainly from CELLULOSE and formerly used for making photographic film
(i)	rouge	a red substance used by women and actors for colouring the cheeks
(j)	sweetener	a substance which is used instead of sugar to make food and drink taste sweet
(k)	charcoal	(pieces of) the black substance made by burning wood in a closed container with little air, burnt in fires to give heat or used in sticks for drawing with

substance リンク: 中心名詞を修飾する “made of” に続く名詞句は、見出し語の「材料」を示している。そこで、このリンクを substance リンクとする。

2.1.3 特定の名詞群に関するリンク

物質名詞群 (中心名詞が ‘substance’ である名詞群)

中心名詞が ‘substance’ である名詞を物質名詞とする。表 1 に物質名詞の定義文の例を示し²、以下に物質名詞群特有のリンクを示す。

attrib リンク: 形容詞 ‘simple’, あるいは ‘compound’ は、見出し語が「単体」であるか「化合物」であるかを示している (表 1(a),(b)).

field リンク: 形容詞 ‘chemical’, ‘mineral’, ‘natural’, ‘man-made’, ‘food’ は、物質名詞の「分野 (field)」を示している (表 1(b),(c),(d),(a),(e)).

colour リンク: 物質名詞の場合は、共通の colour リンクのリンク値とする形容詞の他に ‘clear’ (透明な), ‘shiny’ (光沢のある) を付け加える (表 1(f),(c)).

hardness リンク: 中心名詞を修飾する形容詞 ‘soft’, ‘hard’, ‘sticky’ (粘性のある), ‘plastic’ (熱可塑性の) は、物質名詞の「硬度 (hardness)」を示している (表 1(e),(f),(g),(h)).

occasion リンク: “used in (on)” に続く名詞句 (動名詞を含む) は、物質名詞が「使用される状況 (occasion)」を示している。例

²以下の表中の定義文は、すべて、各見出し語の最初の定義文である。

表 2: 道具に関する名詞の定義文の例

(a)	telescope	a tubelike scientific instrument used for seeing distant objects by making them appear nearer and larger see picture at OPTICS
(b)	violin	a type of 4-stringed wooden musical instrument, supported between the left shoulder and the chin and played by drawing a BOW (2) across the strings see picture at STRINGED INSTRUMENT
(c)	trephine	a special medical instrument with a sharp fine-toothed circular cutting edge used in trephining (TREPHINE)
(d)	card	a comblike instrument used for combing, cleaning, and preparing wool, cotton, etc., for spinning
(e)	cathode ray tube	a glass instrument in which streams of ELECTRONS from the CATHODE (cathode rays) are directed onto a flat surface where they give out light, as in a television receiver
(f)	bugle	a brass musical instrument, played by blowing, like a TRUMPET but shorter, used esp. for army calls see picture at WIND INSTRUMENT
(g)	ammeter	an instrument for measuring, in AMPERES, the strength of an electric current see picture at SCIENTIFIC

例えば、‘plutonium’ が使用される状況は、“the production of atomic power” である (表 1(a)).

used_by リンク: “used by” に続く名詞句は、物質名詞を「使用する主体」を示している。例えば、‘rouge’ を使用する主体は、“women” である (表 1(i)).

instead_of リンク: “used instead of” に続く名詞句は物質名詞の「代用元」を示している。例えば、‘sweetener’ の代用元は、“sugar” である (表 1(j)).

made_from リンク: “[made, obtained] from” に続く名詞句は、物質名詞の「原料」を示している。例えば、‘celluloid’ の原料は、“CELLULOSE” であり (表 1(h)), ‘glue’ の原料は、“animal bones or fish” である (表 1(g)).

made_by リンク: “made by” に続く動名詞は、物質名詞の「作成方法」を示している。例えば、‘charcoal’ の作成方法は、“burn wood in a closed container with little air” である (表 1(k)).

found_in リンク: “found in” に続く名詞句は、物質名詞の「存在先」を示している。例えば、‘pectin’ は、“certain fruit” 中に存在する (表 1(b)).

道具に関する名詞群 (中心名詞が ‘instrument’ である語群)

中心名詞が ‘instrument’ である名詞の定義文の例を表 2 に示す。この名詞群に特有のリンクを以下に示す。

field リンク: 中心名詞を修飾する形容詞 ‘scientific’, ‘musical’, ‘medical’ は、道具の「分野 (field)」を示している (表 2(a),(b),(c)).

shape リンク: 中心名詞を修飾する ‘-like’ の形の形容詞は、道具の「形状 (shape)」を示している (表 2(a),(d)).

表 3: 衣服に関する名詞の定義文の例

(a)	skirt	a woman's outer garment that fits around the waist and hangs down with one lower edge all round
(b)	surplice	a garment made of white material worn over a darker garment during religious services by some priests and CHIRBOYS see also VESTMENT
(c)	sweatshirt	a loose cotton garment for the upper part of the body
(d)	jersey	a tight KNITted woollen garment for the upper part of the body
(e)	singlet	a man's garment without sleeves worn as a VEST or as an outer shirt when playing some sports
(f)	apron	a simple garment worn over the front part of one's clothes to keep them clean while working or doing something dirty or esp. while cooking
(g)	bathrobe	a loose garment (usu. made of a material that takes in water easily) worn before and after bathing esp. by men

substance リンク: 共通の substance リンクのリンク値の他に, 中心名詞を前から修飾する形容詞として 'wooden', 名詞として 'glass', 'brass' を付け加える (表 2(b),(e),(f)).

purpose リンク: 共通の purpose リンクのリンク値の他に, 中心名詞の直後の 'for' に続く動名詞を付け加える。例えば, 'ammeter' の用途は, "measure the strength of an electric current" である (表 2(g)).

method リンク: "played by" に続く動名詞は, 道具 (特に楽器 'musical instrument') の「使い方 (method)」を示している。例えば, 'violin' の使い方は, "draw a BOW across the strings" であり (表 2(b)), 'bugle' の使い方は, 'blow' である (表 2(f)).

衣服に関する名詞群 (中心名詞が 'garment' である語群)

中心名詞が 'garment' である名詞の定義文の例を表 3 に示す。この名詞群に特有のリンクを以下に示す。

agent リンク: 中心名詞を修飾する所有格, あるいは "worn by" に続く名詞句は, 衣服を着る「主体 (agent)」を示している。例えば, 'skirt' を着るのは, 'women' であり (表 3(a)), 'surplice' を着るのは, "priests and CHOIRBOY" である (表 3(b)).

substance リンク: 共通の substance リンクのリンク値の他に, 形容詞として 'woollen', 名詞として 'cotton' を付け加える (表 3(c),(d)).

time リンク: 'when', 'while', 'after', 'before', に続く動名詞, 及び 'during' に続く名詞句は, 衣服を着用される「時間帯」を示している。例えば, 'singlet' は, "play some sports" の際に着用され (表 3(e)), 'apron' は, "work or do something dirty", 特に "cook" の際に着用される (表 3(f)). また, 'bathrobe' は, "bathe" の前後に着用される (表 3(g)).

2.2 優先規則 (p ルール)

前節では, 具象名詞の定義文を分析した結果について述べた。この調査に基づき, SAX+p [8] を用いて名詞定義文の解析用の文法 (DCG 形式) を作成した。この文法は,

np --> det, prempg, ng, postm

といったように³, 定義文のパターンを尊重したものとした。以下では, 名詞定義文を解析するための構文的優先規則 (preference rule) についてのみ述べる。詳細は, 文献 [14] を参照のこと。なお, 優先度は点数 (スコア) で表現した。

2.2.1 前置修飾語句の優先度を上げる

LDOCE のような大規模の辞書を用いると, 多くの語 (例えば, 'white') が品詞として名詞と形容詞の両方を持っている。従って, そのような語が別の名詞の直前に現われた場合, 例えば "white substance" では, 図 2 に示す 2 通りの解釈が存在する。このような場合には, (b) の形容詞の解釈が正しいことが多い。そこで, 前置修飾語句のスコアを上げることにした。p ルールを図 3 に示す。この記述の 3 行目の 'pref_CAT is prefs(1)*2+prefs(2)' が, prems のスコアの 2 倍と ng のスコアを加えたものを, np のスコアとすることを示している。

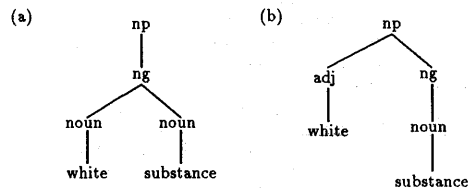


図 2: "white substance" の 2 通りの解析木

```
np(np(ng,[Pre]),N,thd,[sub,obj],and(Spr,SW)) -->
  prems(Pre,Spr), ng(ng(N,SW))
  &&{pref_CAT is prefs(1)*2+prefs(2)}.
np(np(ng,[ ]),N,thd,[sub,obj],SW) --> ng(ng(N,SW)).
ng(ng(NG),N,and(SW1,SW2)) --> noun(N1,SW1), noun(N2,N,SW2),
  extC(append(N1,N2,NG)).
```

図 3: 前置修飾語句の優先度を上げる p ルール

2.2.2 並列名詞句よりも並列名詞群の優先度を上げる

名詞群は単独で名詞句となり得るので, detw, ng1, coconj, ng2⁴ の順序で現われた場合には, 図 4 に示す 2 通りの解釈が存在する。このような場合, LDOCE 名詞定義文では, 名詞群が並列する (a) の解釈が正しいことが多い (例 'chlorophyll' (葉緑素) の定義文の一部 "the stems and leaves"). そこで, 並列名詞群のスコアを上げる (図 5 の 5 行目の '+5' がこれを表現している)。

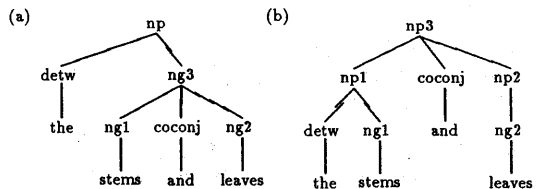


図 4: detw, ng1, coconj, ng2 の 2 通りの解析木

³ この例のシンボルの意味は, 以下のとおり。det: 冠詞, prempg: 前置修飾語句 (列) (形容詞など), ng: 名詞群 (列), postm: 後置修飾語句 (前置詞句など)
⁴ detw: 冠詞相当語句 (冠詞を含む), coconj: 等位接続詞

```

ng(ng({Conj,[NG1,NG2]}),N,and(SW1,SC,SW2)) -->
ng(ng(NG1),N1,SW1), coconj([Conj],SC),
ng(ng(NG2),N2,SW2),
extC(check_number(Conj,N1,N))
&&{pref_CAT is prefs(1)+prefs(2)+prefs(3)+5}.

np(np({Conj,PWP}),N,P,C,and(SW1,SC,SW2)) -->
np(NP1,N1,P1,C,SW1), coconj([Conj],SC),
np(NP2,N2,P2,C,SW2),
extC(check_number(Conj,N1,N),check_person(P1,P2,P),
append([NP1],[NP2],PWP))
&&{pref_CAT is prefs(1)+prefs(2)+prefs(3)}.

```

図5: 並列名詞句よりも並列名詞群の優先度を上げる p ルールの例

2.2.3 動詞句の必須格要素の優先度を上げる

例えば, verb(ing,t1), ng が名詞句を形成する場合, 前の動詞が現在分詞として名詞群を修飾する解釈と, 動名詞句とする解釈がある。このような場合, LDOCE 名詞定義文では, 動名詞とする解釈が正しいことの方が多い(例 'alum'(ミョウバン)の定義文の一部 "used in preparing leather")。そこで, 動名詞の必須格要素のスコアを上げる⁵(図6)。

```

vp(vp(verb(V),[Npo,Npc],[ ],T,and(SV,SW0,SWC)) -->
verb(V,x1,T,SV), np(Npo,N,P,C1,SW0), np(Npc,N,P,C2,SWC),
extC(member(C1,obj),member(C2,sub))
&&{pref_CAT is prefs(1)+prefs(2)+prefs(3)+2}.

```

図6: 動詞句の必須格要素を上げる p ルールの例

2.2.4 vp からなる動名詞の優先度を上げる

LDOCE 定義文では, vp-o⁶ も動名詞となることがある(例 'bleach'(漂白剤)の定義文 "a substance of BLEACHing") ので, 文法規則ではこれを認めている。しかし, 多くの動詞は文法コードとして 't1' と 'i0' の両方を持っているので, そのような場合,

1. vp --> verb(i0)
2. vp_o --> verb(t1)

の2通りの解釈が存在してしまう(例 'foodstuff'(食糧)の定義文の一部 "foods for eating")。そこで, 一般の解釈である vp の動名詞のスコアを上げる(図7)。

```

gerundp(grd(vp(VP)),and(SV)) --> vp(VP,ing,SV)
&&{pref_CAT is prefs(1)+2}.
gerundp(grd(vp_o(VP)),and(SV)) --> vp_o(VP,ing,SV).

```

図7: vp の動名詞の優先度を上げる p ルール

2.3 解析結果

物質名詞群の定義文 171 文について, 構文解析を行なった結果を表4に示す。得られた優先解中のほとんどに正しい解析結果が含まれていた。なお, 解が得られなかった原因としては, 文中に

⁵但し, これらの前に冠詞相当語句が存在する場合 ("the preparing leather") には, DCG ルールにより現在分詞となる解釈のみが得られる。

⁶vp_o は, 目的語の欠けた vp で, 主として, 関係代名詞節に現われる。

表4: 物質名詞の定義文 (171 文) の解析結果

解が得られた文	語数	平均	13.8	113	(66.1%)
	優先解数	平均	3.45	最多	30
	全解数	平均	63.8	最多	114
解が得られなかった文				48	(28.1%)
メモリが不足し, 処理できなかった文				10	(5.8%)

挿入句が含まれていたもの, 副詞が助動詞と本動詞の間にあったもの, などがあった。

3 解析結果からの情報抽出と知識ベースの作成

3.1 情報抽出アルゴリズム

3.1.1 構文解析結果からのリンクの決定

前節で述べた SAX+p による構文解析の出力 (以下構造引数と呼ぶ) は,

top(冠詞相当語句, 前置修飾句群, 中心名詞, 後置修飾句群)

の形式にした。ここで, リンクの抽出を容易にするため, 前置修飾句群および後置修飾句群は, 木構造ではなく, リスト形式にした。意味ネットワークのリンクは, 2.1 節で示したように, 主として, これらの修飾群から抽出される。そこで, 前置修飾句群, 後置修飾句群を以下に示す方法で解析してリンクを決定する。なお, 現在は, 冠詞相当語句の処理を行っていない。

前置修飾句群の解析

構文解析結果の前置修飾句群のリストの各要素は,

prem(カテゴリ (主要語, 副詞群のリスト))

の形式になるようにした。ここで, カテゴリは, 現在分詞, 過去分詞, 形容詞のいずれかである。以下の手順によって, 各前置修飾句を順次, リンクセットに変換していく。

1. カテゴリが現在分詞の場合

リンク名を function とし, 主要語をその動詞の原形, 必須格要素を空リスト, 主要語の修飾語リストとして副詞群のリストとする。

2. その他の場合

主要語を 3.1.2 節で述べる抽出情報照会クラスに照会して, リンク名を決定し, 副詞群のリストを主要語の修飾語リストとする。

後置修飾句群の解析

構文解析結果の後置修飾句群のリストの各要素は,

postm(カテゴリ (引数並び))

の形式とした。この場合, カテゴリは, 現在分詞句, 過去分詞句, などである (以下を参照)。以下の手順によって, 各要素を順次, リンクセットに変換していく。

1. カテゴリが現在分詞句及び主格の能動関係代名詞節の場合
リンク名を function とし、主要語をその動詞の原形とする。
必須格要素リスト、主要語の修飾語リストは、動詞句の構造
引数として得られているものとする。例えば、'adrenalin' の
定義文の一部 "causing action" の構造引数は、

```
vp([causing,cause],[np(action,[ ]),[ ]])
```

であり、変換結果は、

```
function(cause,[np(action,[ ]),[ ]])
```

となる。

2. カテゴリが過去分詞句及び主格の受動関係代名詞節の場合
過去分詞と副詞句リスト中の各要素のうち、前置詞句中の前
置詞を、順次、抽出情報照会クラスに照会して、リンク名を
決定する。リンク値は、前置詞句内の名詞句とする。例えば、
'pectin' の定義文 (表 1(b)) の一部 "found in certain fruit"
の構造引数は、

```
vp([found,find],[ ],[pp(prep([in]),  
np(fruit,[prem(certain,[ ]))]))])
```

であり、'found' と 'in' を抽出情報照会クラスに照会した結果、
リンク名 found_in が得られ、変換結果は、

```
found_in(fruit,[prem(certain,[ ]))])
```

となる。

なお、過去分詞と前置詞の組が照会クラスに登録されてい
ない場合は、現在、捨てている。これは、3.1.3 節で示すように、
最優先解を決定するのに利用している。

3.1.2 抽出情報照会クラス

抽出情報照会クラスは、構造引数解析によって得られた主要語
あるいは過去分詞と前置詞の組からリンク名を決定するように、
そのそれぞれに登録するクラスである。抽出情報照会クラスには、
全範疇に共通のものと (図 8)、範疇毎に作成するもの (クラス名:
範疇名.link.info) の 2 種類がある。

```
class lex_info has
: create(Class,Inst) :- new(Class,Inst);
instance
: premlink(Inst,Word,Attr) :- premlink(Word,Attr);

: postmlink(Inst,PP,Prep,Attr) :- postmlink(PP,Prep,Attr);
local
premlink(black,colour) :- !;
premlink(blue,colour) :- !;
...
premlink(yellow,colour) :- !;
premlink(large,size) :- !;
premlink(small,size) :- !;

postmlink(used,for,purpose) :- !;
postmlink(null,such_as,example) :- !;
postmlink(null,like,example) :- !;
postmlink(made,of,substance) :- !;
end.
```

図 8: 具象名詞に共通の抽出情報照会クラスの定義

これらのクラスに用意している述語は、前置修飾句群用として、

premlink(主要語, リンク名);

後置修飾句群用として、

postmlink(過去分詞, 前置詞, リンク名);

がある。

全範疇に共通のクラスには、2.1.2 節で述べたリンクに関するも
のを登録する。範疇毎のクラスには、2.1.3 節で述べた範疇毎のリン
クに関するものを登録し、共通クラスを継承する。知識ベース
作成の呼び出しの際に、範疇毎のクラスが存在する場合には、そ
のクラスが呼び出され、存在しない場合には、共通クラスが呼び
出される。

3.1.3 最優先解の決定

SAX+p の優先解が 1 つであり、構造解析によってリンクセット
が得られた場合は、もちろんそれを最優先解として知識ベースに
登録する。しかし、現段階では、各単語の意味情報が分からないた
め、語の意味によって係り受けが変わるような曖昧さに関して優
先規則が書けない。このような場合には、SAX+p の優先解が複数
個存在する。例えば、図 9 に示すように、定義文中に、verb(pp,
prep1, ng1, prep2, ng2 が現われる場合には、prep2, ng2 で構成
される pp2 が、verb を修飾する解釈と ng1 を修飾する解釈の 2
通りが SAX+p の解析結果として得られる。

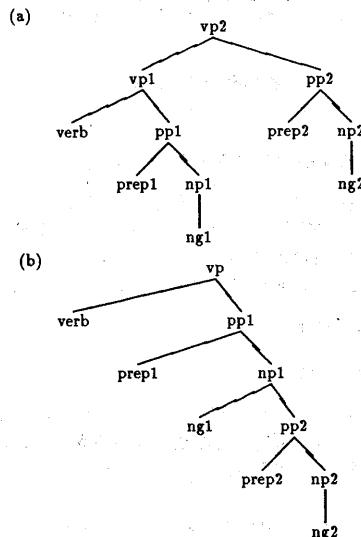


図 9: verb, prep1, ng1, prep2, ng2 の 2 通りの解析木

そこで、複数個の優先解それぞれの構造引数を解析して得られ
たリンクセットに対し、

1. リンク数が多いものを優先する。
2. リンク数が同じであれば、修飾語数の合計が多いものを優先
する。

ことによって最優先解を決定する。

例えば、'kimono' の定義文の一部

a garment worn in Japan by women

```

1. [inder([a]), prem([ ]), kn([garment]),
   postm([pastpp([worn, wear]), [ ]),
   pp([in], [ng(['Japan']), [ ]]),
   pp([by], [ng([women, woman]), [ ]], [ ]])]
2. [inder([a]), prem([ ]), kn([garment]),
   postm([pastpp([worn, wear]), [ ]),
   pp([in], [ng(['Japan']), [ ]], [ ]], [ ]],
   [ng([women, woman]), [ ]], [ ]])]

```

図 10: “a garment worn in Japan by women” の SAX+p/LDOCE の解析結果

の SAX+p の解析結果は、図 10 に示す 2 通りが得られる。この解析結果に適用できる過去分詞と前置詞の組として、抽出情報照会クラスには、

```

postmlink(worn, in, worn, in);
postmlink(worn, by, worn, by);

```

の両方を登録しているので、3.1.1 で述べたように、1. の構造引数から抽出されるリンクセットは、

```

[worn, in('Japan', [ ]),
 worn, by(woman, [ ])]

```

となり、2. の構造引数から抽出されるリンクセットは、

```

[worn, in('Japan',
 [postm(pp([by], [ng([woman]), [ ]], [ ])])])

```

となる。そこで、リンク数の多い 1. を最優先解とする。

```

1. [inder([a]), prem([ ]), kn([garment]),
   postm([pastpp([worn, wear]), [ ]),
   pp([by], [ng([citizens, citizen]), [ ]]),
   pp([of], [ng(['Rome']), [ ]], [ ]])]
2. [inder([a]), prem([ ]), kn([garment]),
   postm([pastpp([worn, wear]), [ ]),
   pp([by], [ng([citizens, citizen]),
   [postm(pp([of], [ng(['Rome']), [ ]], [ ])])])

```

図 11: “a garment worn by citizens of Rome” の SAX+p/LDOCE の解析結果

一方、‘toga’ の定義文の一部分

a garment worn by citizens of Rome

の SAX+p の解析結果は、図 11 に示す 2 通りが得られる。この解析結果に適用できる過去分詞と前置詞の組として、抽出情報照会クラスには

```

postmlink(worn, by, worn, by);

```

だけを登録しているので、3.1.1 で述べたように、2. の構造引数から抽出されるリンクセットは、先の例と同様に

```

[worn, by(citizen,
 [postm(pp([of], [ng(['Rome']), [ ]], [ ])])])

```

となる。しかし、1. の構造引数から抽出されるリンクセットは、抽出情報照会クラスに過去分詞と前置詞の組 worn, of を登録していないので、

```

[worn, by(citizen, [ ])]

```

表 5: 情報抽出の結果 (形容詞句)

(a)	plutonium	field(man-made, [])
		attrib(simple, [])
(b)	pectin	field(chemical, [])
		attrib(compound, [])
(f)	amber	colour(brown, [yellowish])
		hardness(hard, [])
		colour(clear, [])
(g)	glue	hardness(stickey, [])
(h)	celluloid	hardness(plastic, [])
(i)	rouge	colour(red, [])

となる。そこで、リンク数は同じであるが、修飾語数の多い 2. を最優先解とする。

なお、上記の処理を行っても複数の優先解が存在する場合は、現在、人間が選択している。また、SAX+p の解が存在しなかった場合及び、SAX+p の解が存在し、かつ全ての優先解のリンクセットが得られなかった場合も実際にはある。

3.2 抽出結果

また、以下に表 1 に示した定義文からのリンクの抽出結果をリンク値のカテゴリ別に述べる。なお、以下で、(a), (b) などは、表 1 中のそれに対応している。

attrib リンク、field リンクなどリンク値が形容詞句のもの

表 5 にリンク値が形容詞句のものを示す。これらは、主要語である形容詞を個別に抽出情報紹介クラスに登録している。

- 例えば、(a) plutonium は、正しく抽出できた。
- 形容詞句のほとんどが形容詞 1 語なので、第 2 引数の修飾語リストが空リストとなっているが、(f) amber の colour リンクには ‘brown’ の修飾語リストとして [yellowish] が抽出できた⁷。

found_in リンク、made_from リンクなどリンク値が名詞句のもの

表 6 にリンク値が形容詞句のものを示す。これらは、過去分詞と前置詞の組からリンク値を決定したものである。

表 6: 情報抽出の結果 (名詞句)

(b)	pectin	found_in(fruit, [prem(certain, [])])
(g)	rouge	used_by(and([women, actor]), [])
(h)	celluloid	made_from('CELLULOSE', [])
(i)	glue	made_from(or([bones, fish]), [prem(animal)])
(j)	sweetner	instead_of(sugar, [postm(toinf(make, ...))])

- 例えば、(b) pectin の found_in リンクは、正しく抽出できた。
- (g) glue の made_from リンクのリンク値は、構文解析の際、名詞群 ‘bones’ と ‘fish’ の並列を優先させた (2.2 節参照) ので、‘or’ を分解すると “animal bones or animal fish” という値が登録されていることになる。

⁷但し、‘yellowish’ は、LDOCE の見出し語になかったため、別途、副詞として登録した。

表 7: 情報抽出の結果 (動名詞句)

(g)	glue	purpose(join, [things], [adv([together])])
(h)	celluloid	purpose(make, [film], [prem(photographic, [])])
(i)	rouge	purpose(colour, [cheeks], [def(the)], [])

- (a) plutonium の定義文中には, occasion リンク (used in に続くもの) があるが, 現在のプログラムでは, 関係代名詞節の構造引数を解析できないので, 抽出できなかった。

リンク値が動詞句 (動名詞) のもの

動詞句の例として, purpose リンクを表 7 に示す。purpose リンクは, “used for” に続く名詞句が動名詞であった場合, 及び to-不定詞である。

- (g) glue 及び (h) celluloid の purpose リンクは正しく抽出されている。
- (j) sweetner の purpose リンク (to-不定詞) は, 優先解の中に直前の名詞 ‘sugar’ を修飾するものしか現われなかったため, 抽出できなかった。

4 おわりに

本報告では, 既存辞書 - LDOCE - の名詞定義文を意味関係に重点を置いて解析し, その解析結果から情報を抽出して, 知識ベースを作成するシステムについて述べた。LDOCE の意味関係の記述方法としては, 意味ネットワークの考え方をを用いた。

今後の課題としては, 以下のものが考えられる。

- 優先規則を含め, 定義文解析用の文法を改良し, 構文解析の精度を上げる。
- 調査した以外の定義文についても解析を行ない, 意味関係を表すリンクの種類と知識ベースを拡充する。
- 作成した知識ベースを他の自然言語処理システムに利用することにより, リンクの種類, および, 知識ベースそのものの有効な点を実証し, 不十分な点を明らかにする。

参考文献

- [1] B. BOGRAEV AND T. BRISCOE, Large Lexicons For Natural Language Processing: Utilizing The Grammar Coding System of LDOCE, *Computational Linguistics*, Vol.3, No.3-4 (1987), p.203-218.
- [2] 池田裕治, 武藤幸好, 辻井潤一, 長尾真, ユニフィケーションに基づくパーザ KGW, 電子通信学会, 言語処理とコミュニケーション研究会 (1987)。
- [3] 石崎俊, 内田裕士, 多言語間翻訳のための中間言語について, 情報処理学会, 情報処理学会研究報告 89-NL-70-3 (1989)。
- [4] K. JENSEN AND JEAN-LOUIS BINOT, Disambiguating Prepositional Phrase Attachments by Using On-line Dictionary Definitions, *IBM Research Report*, RC 12148 (1986)。

- [5] J. KLAVANS, COMPLEX: A Computational Lexicon for Natural Language Systems, *Proc. of COLING88* (1988), p.815-823.
- [6] Longman Dictionary of Contemporary English, Longman Group Ltd. (1978)。
- [7] 松本裕治, 杉村領一, 論理型言語に基づく構文解析システム SAX, コンピュータソフトウェア, Vol.3, No.4 (1986), p.4-11.
- [8] 松本裕治, 杉村領一, 構文解析システム SAX のための文法記述言語, 日本ソフトウェア科学会, 第 5 回全国大会論文集 (1988), p.77-80.
- [9] 長尾真他, 機械翻訳に対するロングマン辞書データベースの応用, 情報処理学会, 自然言語処理研究会資料, 29-5 (1982)。
- [10] 中村順一他, Longman 辞書データベースと情報の抽出, 数理科学講義録, 693 (1986)。
- [11] 中村順一, 酒井桂一, 長尾真, 英々辞典を用いた名詞の意味関係の分析, 電子情報通信学会, NLC86-23 (1987), p.17-24.
- [12] 中村順一, 酒井桂一, 長尾真, 英語辞書における名詞の意味分類について, 情報処理学会, 第 36 回全国大会 3U-8 (1988), p.1253-1254.
- [13] J. NAKAMURA, AND M. NAGAO, Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation, *Proc. of COLING88* (1988), p.459-464.
- [14] 酒井桂一, 辞書定義文の解析とその知識ベース化, 京都大学大学院工学研究科修士論文 (1989)。
- [15] 辻井潤一, 池田裕治, 武藤幸好, 長尾真, KGW+p の制御方式, 日本ソフトウェア科学会, 第 4 回全国大会論文集 (1987), p.355-358.
- [16] 鶴丸弘昭他, 単語間の上位-下位関係の自動抽出, 情報処理学会, 情報学基礎研究会 3-1 (1986)。
- [17] Y. WILKS, Machine Tractable Dictionaries as Tools and Resources for Natural Language Processing, *Proc. of COLING88* (1988), p.750-755.