日本語自然文処理における概念ネットワークデータの構築手法

高 田 明 典

1．はじめに

Langacker（2002）は、集合概念が形成されていく様子を概念ネットワークモデルを用いて説明した。そこでは、

（1）**ある概念枠組み**（schema）**が、典型的事例**（prototype）**の提示によって形成される**

（2）**その概念枠組みに基づいて、ある事例**（instance）**が、ある集合概念のもとに属するという判断が行われる**

（3）**ある拡張事例**（extention）**の提示によって概念枠組みが変化する**

という手続きによって概念ネットワークが形成されていく過程が提示されている（Fig.1）

このLangackerのモデルは自然知能における概念形成のよい模式図となっており、自然言語処理システムの構築においても、その処理アルゴリズムの設計・構築において参照しうる有用なモデルであると言える。しかしながら､具体的な自然文処理の過程において､どのような処理アルゴリズムによってこのような概念ネットワークが形成されていくのかに関しては不明な点も多い。特に物語構造分析の前処理段階としての自然言語処理を想定する場合においては、入力された具体的な自然文の羅列から集合概念などが形成されていく過程のアルゴリズムを明確にしなければならないが、Langackerに代表される認知言語学のモデルにおいては、システムとして実装可能なレベルでの処理アルゴリズムに関する言及は極めて少ない。

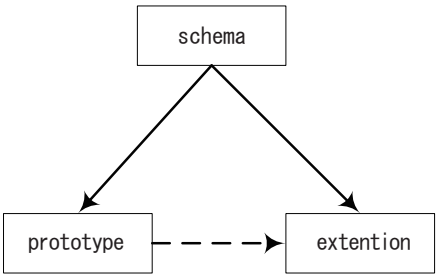

Fig.1

さらには、物語構造分析の前処理段階としての自然文処理においては、少々特殊な事情が存在し、効率的な概念クラスの形成の妨げとなっている。特に、その自動化処理を実現する場合においては、対象となる登場人物と名詞・代名詞との照応関係を解析し、あるクラスやインスタンスのもとに統合することは重要であるが、困難も存在する。たとえば、物語の当初においてアヒルだった対象が後に白鳥であると判明したり、白鳥や野獣が王子であったり、王子がカエルになったりもするという場合の処理は容易ではない。

本研究においては、上記の問題を考慮しつつ、主として物語構造分析のための自然文処理システムにおいて使用することを前提とした概念ネットワークの構築手法を検討した。

またそのため、入力された自然文から概念ネットワークが形成されていく過程について検討し、そのデータ構築に関しての具体的方法を提案する。

2．データ構築手法の概要

本研究においては、Langacker の概念ネットワークモデルを基礎として用いるが、そのデータ構築に際しては必ずしも Langacker によるスキーマモデルによらない。Langacker は、入力された事例からスキーマが抽出されるモデルを考えたが、前述のように、具体的な処理過程においては、この方法の実現は容易ではないと考えたことによる。実際の自然文処理の過程においては、事例についての自然文入力が先行し、それらの事例と概念の関係についてのさらなる自然文入力によって、概念ネットワークが構築されていく。

たとえば、Fig.2（a）に示したように、「高田は男である。」という文においては、まず「高田」という名称属性を持つオブジェクト（Obj_1）と、「男」という名称属性を持つオブジェクト（Obj_2）の二つが構築される。ここでオブジェクトとは「対象物」という意味で用いられているが、必ずしも事物のみを指すものではなく、集合概念や固有の対象事物をも指し、それらは「オブジェクト」の属性として登録される。オブジェクトの属性としては、Class（**集合**）／Instance（**事例**）／Meta Instance（**上位事例**）／Pronoun Class（**代名詞集合**）

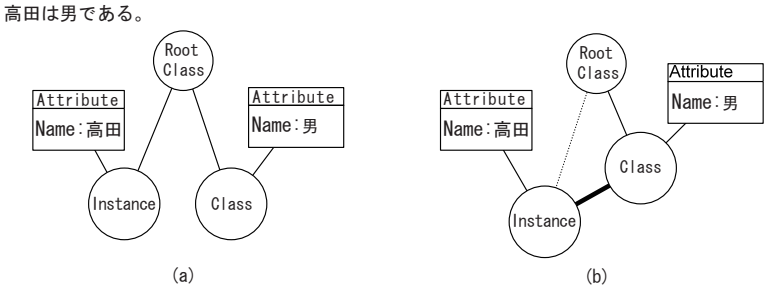

Fig.2

が想定されている。

処理対象となるすべての概念は、当初“Root Class”の下に置かれる。このとき「高田」という名称属性が固有名詞を示すものであることが予め知られていれば、それによって示されるオブジェクトは“Instance”として登録される。Instance とはオブジェクトの種類の一つであり、ある固有の事物の記述に対応する概念である。また「男」が一般名詞であることが知られていれば、それは「男」という名称属性を持つ“Class”を示すオブジェクトとして登録される。その後、Obj_1 と Obj_2 の間のリンクが形成され、オブジェクト同士の関係が調整される（Fig.2（b））。

名称属性などの属性は、Fig.2 において四角形で示した Attribute データとして登録され、ある特定のオブジェクトとの間にリンクを形成する。このリンクは、線形リストで実装される。また、オブジェクト同士の関係も同様のデータ構造によって管理される。

Attributeとして管理されるデータには、様態／行為、および、属性／状態の組み合わせが登録される。つまり、

①様態属性　②様態状態　③行為属性　④行為状態

の4種類となる。ここで、様態とは「名称属性」や「形容詞属性」などで示されるものを指す。また行為とは、原則として動詞句を伴う表現によって示されるものを指す。さらに属性（property）とは基本的に不変のものを指し、状態（state）とは変化するものを指す。

たとえば、「鳥は飛ぶ」という場合、「鳥」によって示されたクラスの「行為属性」として、「飛ぶ」が登録される。これらの処理は、係り受けの文構造に基づいて行われる。たとえば、「高田は男である。」という文は、

$[Obj_1\ Obj_2]$

という係り受け構造を持っている。さらに「高田は大学の教員である。」であれば、

$[Obj_1\ [Obj_2\ Obj_3]]$

という係り受け構造となる。

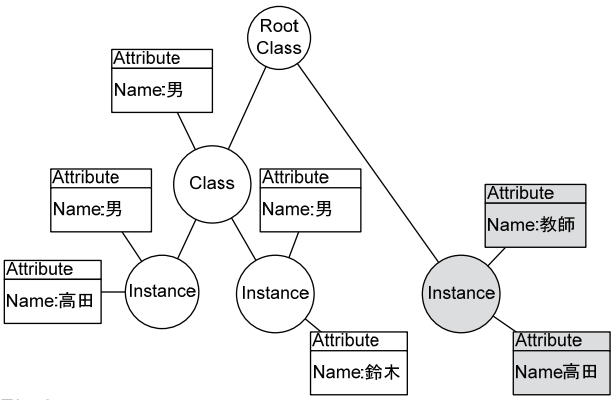

Fig.3

もちろん、文を構成する要素はオブジェトに限られるわけではなく、形容詞、動詞などによって構成される様態／行為なども同様に扱われる。係り受け構造において内側に位置するデータから順に、二つの構成要素が一つにまとめられていくという処理が行われる。

また、名詞や代名詞によって指示されたオブジェクト（Instance）が同一のものを指し示していると判定された場合には、それらのInstanceの上位に、それらを統合するものととして“Meta Instance”のオブジェクトが作られる。この判定は、同一性判定として知られているものであるが、本研究においては、文脈情報をさかのぼることによって、(1) 同一の

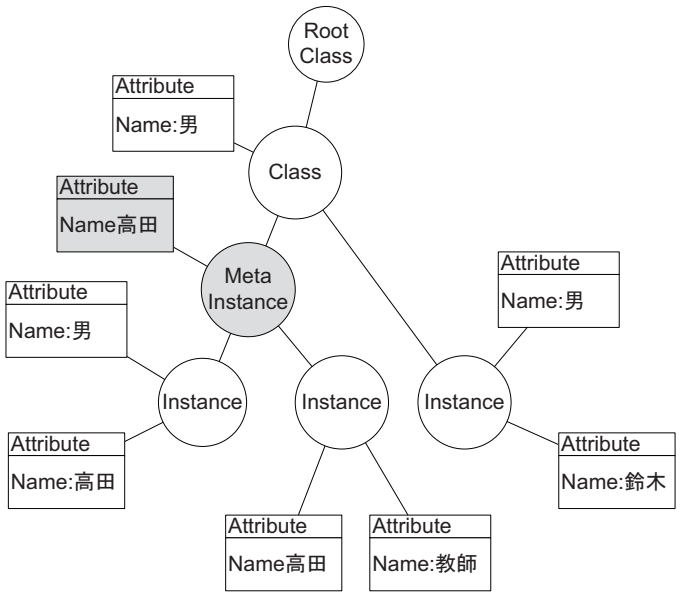

Fig. 4

固有名詞の名称属性を有している場合、(2) 既に統合されたクラスの下位に位置している場合、(3) 代名詞で指示されているオブジェクトが直近のオブジェクトと同一のクラスの下位に属する場合、に「同一」であると判定している。

Fig.3 に示したように、新規 Instance（図中の灰色の円）が既存の「高田」という同一名称属性を有している場合、Fig.4 のように、それらの二つの Instance の上位に Meta Instance が作成される。

Meta Instance を置くことにより、新たに入力された文情報によって統合の状況や照応関係が変化したとしても、付け替えによって対応することが可能となる。文入力が進行することによって、概略として Fig.5 に示したようなオブジェクト構造が構築されることを想定している。

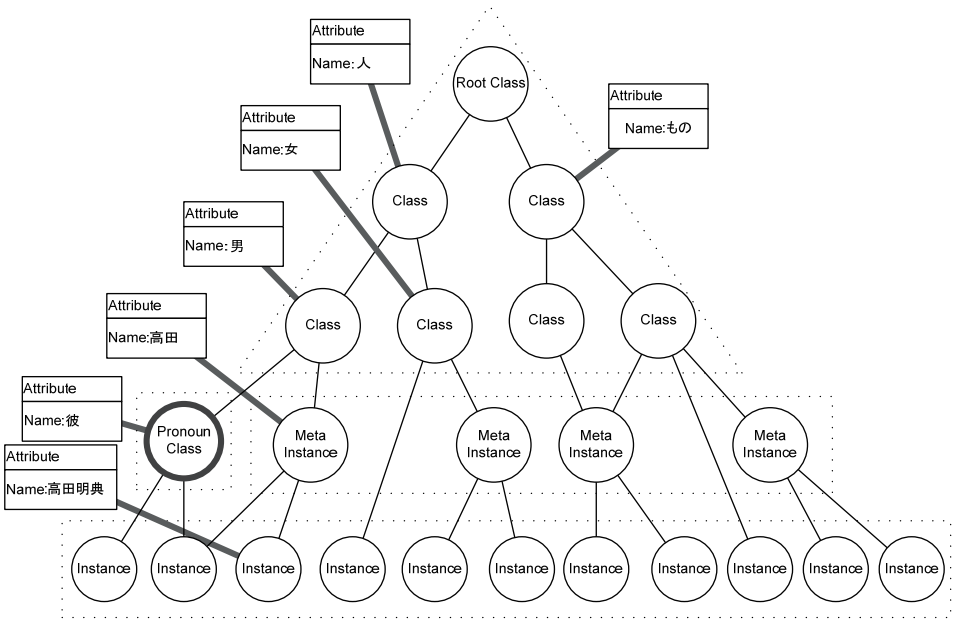

Fig.5

3．システムの実装

本手法は、シミュレーター ides-J の一部であるオブジェクト処理ルーチンとして実装された。おおまかな処理の流れを Fig.6 に示した。形態素解析／構文解析システムとしては「茶筅／南瓜」を使用した。また、概念ネットワークの構築状態の可視化のために Graphviz を使用した。システムは、Fedora Core7 上で gcc を用いて作成された。

例として、「高田は男である。彼は眠っている。」の２文を入力した場合の構文解析データの例を Table 1 に示す。このデータには、すでに係り受け関係の情報が含まれているが、さらに Ides-J　Parser によって補完的処理を行っている。

「高田は男である」という場合、係り受けに基づくデータ構造は、

［obj_1 obj_2］

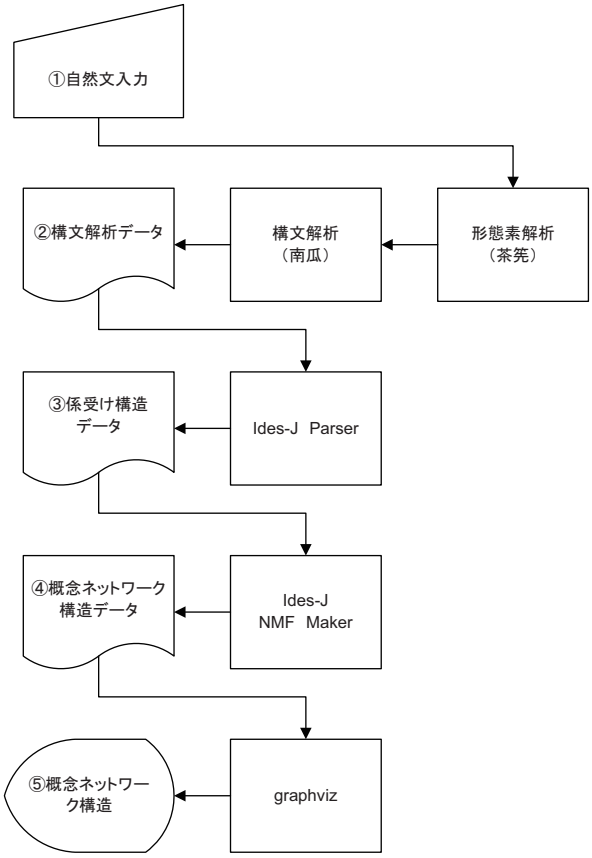

Fig. 6

Table 1

```
  * 0 1D 0/1 0.00000000
高田    タカダ  高田    名詞-固有名詞-人名-姓    B-PERSON
は      ハ      は      助詞-係助詞              O
* 1 -1O 0/2 0.00000000
男      オトコ  男      名詞-一般                O
で      デ      だ      助動詞  特殊・ダ 連用形  O
ある    アル    ある    助動詞  五段・ラ行アル  基本形  O
。      。      。      記号-句点                O
EOS
* 0 1D 0/1 0.00000000
彼      カレ    彼      名詞-代名詞-一般         O
は      ハ      は      助詞-係助詞              O
* 1 -1O 0/2 0.00000000
眠っ    ネムッ  眠る    動詞-自立      五段・ラ行  連用タ接続  O
て      テ      て      助詞-接続助詞            O
いる    イル    いる    動詞-非自立    一段        基本形  O
。      。      。      記号-句点                O
EOS
```

という単純なものになる。「a はｂである」「a であるｂ」「a のｂ」などという表現において、その係り受け構造は、

［a　b］

として表現されるように、基本的に二つの要素の関係によってデータ構築が行われる。ただし、並列関係にある要素に関しては、() を用いて、

［a　(b　c)］

のように表現される。

これらの係り受け構造に基づき、オブジェクトの階層構造が構築される。オブジェクトデータ構造は Graphviz のデータとして出力され表示される。前述した例の出力を、Fig.7 に示す。

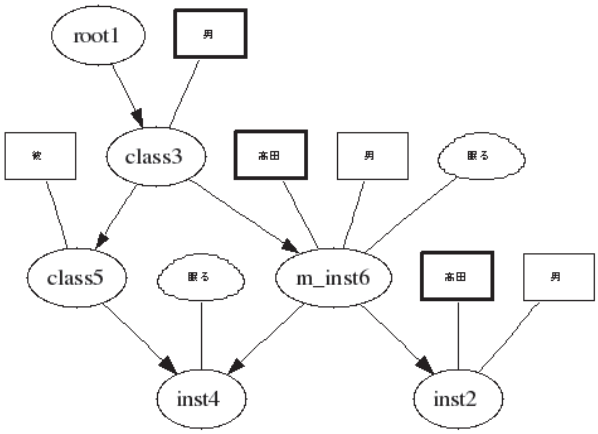

Fig. 7

本システムは、まだ実装が完了したばかりであり、評価をなしうる段階ではないが、比較的単純な文であれば､羅列的に入力された場合でも概念ネットワーク構造を構築しうる。以下に、単純な文例での出力例を示す（Fig.8)。

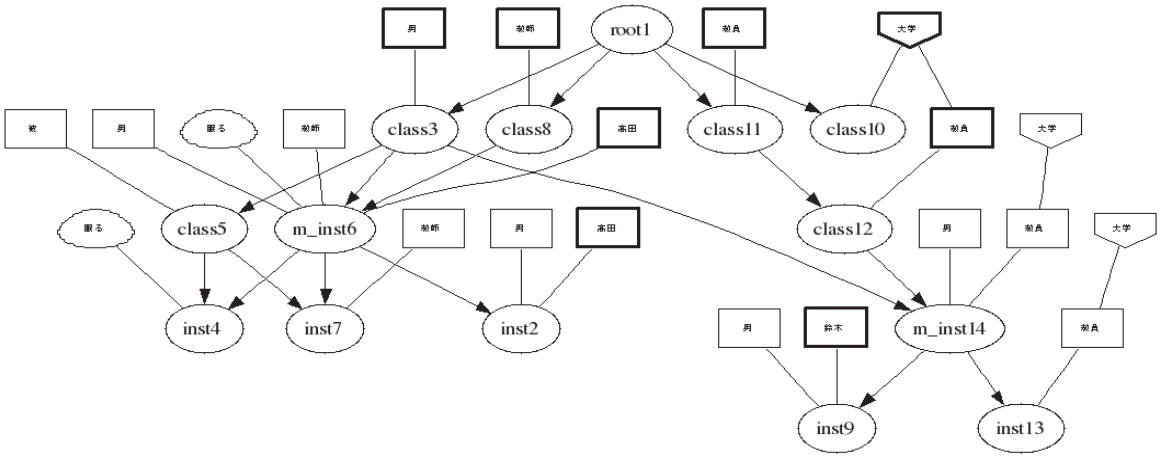

Fig. 8

（入力文）

「高田は男である。

彼は眠っている。

彼は教師である。

鈴木は男である。

彼は大学の教員である」

ここで、五角形で示された概念は、その概念が「～の」などの連体修飾語句となっていることを示す。また、それぞれ太枠で示された名称属性は、そのオブジェクト（クラスやインスタンスなど）の「主たる属性（多くの場合は、最初に登録された名称属性）」であると判断されたことを意味している。

上記例においては、Root Class（root1）のもとに、以下の 4 つのクラスが構築されている。

class 3：男

class 8：教師

class10：大学

class11：教員

ただしここで、数字はオブジェクト番号を示しているので、class ごとには連番にならない。“class3”とは、

［object No.3（class 属性）］

という意味である。

さらに、class11 の下には、class12（「大学の教員」）が構築されている。これは、「～の」という連体修飾語句で示された概念によって下位クラスが生成されたことを意味している。class12 の下には、meta instance（m_inst14）が生成されており、instance9 と instance13 を統合する概念となっている。m_inst14 は、「鈴木」という名称属性を持つオブジェクトであり、事例としては二度出現しているが、それらの二度の言及が「同一人物」を示していると判断されている。

同様に meta instance である m_inst6 は、名称属性「高田」を持つオブジェクトであり、「彼」という代名詞で指示されたものを含め三つの instance の上位にあり、それらを統合している概念である。卵型で示された「眠る」は、行為状態を示しており、inst4（object No.4（instance 属性））の状態記述である。

4．おわりに

本手法は未だ試験実装の段階であり、日本語自然文処理に基づく物語構造分析において十分な機能を有しているとは言い難い。しかしながら、上述のように、具体的な自然文入力から概念ネットワークが構築されていく過程のモデルの実装の一事例としては、ある程度の成果を得ていると考える。今回参照した Langacker に代表される認知言語学的モデルは、モデルとしては説得的であると考えられているものの、それが具体的な処理システ

ムとしての実装を通して検討されることは、そう多くない。しかしながら、実用的なモデルとして検討するためには、システムとして実装しつつ、その問題点を具体的に検証していくことが、最も近道であると考えられる。本システムは、物語構造分析の自動化処理の一環として構築されたものであるが、その枠内に留まらず、日本語自然文理解のモデルもしくは認知言語学的モデルの検証のための一つの具体的方法論として、有用な道筋を示しうるものであると思われる。換言するならば、モデルが妥当であるなら、処理システムとして実装することが可能であるはずである。逆に、実装が困難であれば、それはモデルに欠陥があることが示唆されていると考えるべきであり、また、そのような示唆によってモデルを精緻化していくことが可能であると考える。

本研究は概念ネットワーク構築に留まるものではなく、自然文によって表現された物語の構造をメタ形式に変換することを最終的な目的としているが、それは同時に、自然文のメタ形式構築モデルを模索する研究的営為でもある。現時点においては、上述のオブジェクト概念のデータ構築システムの作成がある程度終了し、それに基づいて物語シーケンス（物語の話素の連鎖）のデータを構築する段階に差し掛かっているが、概念ネットワークに関しても、さらなる研究が必要である。

【参考文献】

1）Langacker, Ronald W. 2002. Concept, Image and Symbol：The Cognitive Basis of Grammar（2nd ed.）. Berlyn･New York：Walter De Gruyter Inc.

2）山岸謙治，村松孝彦，原田実　語意に基づく深層レベルの指示代名詞照応解析システム Anasys/D，情報処理学会研究報告．自然言語処理研究会報告 IPSJ SIG Notes Vol.2003，No.4，pp.33-40