情 報 学 基 礎 34－5
（1994．9．14）

図鑑の解説文から内容抽出を行なうための専門知識の構築

渡辺 靖彦　長尾 真
京都大学 工学部 電気工学第二教室

要旨

我々は画像データから直接抽出するのが困難な画像の内容についての情報を、画像データに付属する解説文から自然言語処理を行なって抽出することを目指している。本稿では対象として植物図鑑を選び、その解説文を解析するのに必要な専門知識の獲得の方法について述べる。さらに図鑑のテキストから無作為に抽出した名詞述語文200文に対し、この専門知識を用いて用例にもとづいた意味解析と係り受け解析を行なった。その結果、図鑑の解説文において画像の内容に関する重要な情報を表現する名詞述語文の意味関係の87%、係り受け関係の96%の解析に成功した。

Technical Knowledge Acquisition for Information Extraction from Pictorial Book of the Flora

Yasuhiko Watanabe　Makoto Nagao
Department of Electrical Engineering II, Kyoto University

Abstract

It is difficult to extract contents of images from image data itself. So we took a different way. That is, we intend to extract these information from the explanation texts of image data. In this thesis, we described a new method of technical knowledge acquisition for information extraction from pictorial book of the flora. We experimented our method for 200 copular sentences extracted from pictorial book by the modification structure analysis and semantic analysis in example-based method. Then, we obtained the correct recognition scores of 96% in the modification structure analysis and 87% in the semantic analysis.

1 はじめに

マルチメディアシステムへの期待が高まっている今日、画像と言語という異なるドメインの情報を統合して計算機で扱うことが今後ますます重要になると考えられる。そのためには画像と言語の関係性をどのようにつけるかが重要な問題である。我々はこの関係性を概念、すなわち意味によってつけることを考えている。そのため、画像の内容がどのような概念に対応し、それがどのような意味で言語情報に結びついているのか、あるいは画像内容を解説するテキストからそれらの情報をどのように抽出するのかについて研究を行なっている[1]。

テキストから情報抽出や知識獲得を行なう方法には、辞書の語義文などを対象にして表層表現を手がかりにしたパターンマッチングによる方法[2]が研究されている。対象テキストや抽出すべき情報が限定されているときにはこの方法によってかなりの精度で情報を抽出できることが明らかになっている。しかし、情報抽出のためのパターンを作成するには人間がその文章表現を詳細に調査しなければならず、かなりの労力を必要とする。

そこで本研究では、画像の内容を解説するテキストとして植物図鑑のテキストを選び、そのテキストから画像の中の対象がもつ属性についての表現を名詞述語文という単純なパターンによって抽出し、抽出した表現をその中で用いられている専門用語の類似度によって意味分類することで、画像の中の対象がもつ詳細な属性情報を獲得する方法を提案する。さらに得られた属性情報と人手で作成した概念間の関係情報を用いて画像内容を解説するテキストを対象に用例にもとづく方法で意味解析および係り受け解析を行ない、提案した方法の有効性を検討する。

2 図鑑の意味理解

2.1 画像情報と言語情報の特徴

発話されたり書かれたりした文章による自然言語の情報と図や写真などの画像情報をうまく組み合わせて用いると、受け手にわかりやすくかつ正確に情報を伝達することができる。画像と言語という異なるドメインの情報によって説明される対象の理解について考えるため、画像情報と言語情報の特徴を表1に示す。こうした相補的な特徴のため、画像と言語の情報を統合して扱うとより深い理解が得られると考えられる。そこで、画像と言語によって対象を説明している例として植物図鑑を選び、その意味理解について次節で考察する。

2.2 図鑑の意味理解に対する考え方

人間が言語と画像を統合的に理解する過程では、非常に複雑で多様な情報処理が行なわれていると考えられる。しかし、その詳細はまだ明らかになっていない。そこで画像と言語を用いて図鑑で説明される対象を理解するということを次の2つにしぼる[3]。

画像情報	v.s.	言語情報
複雑で微妙な非論理的な情報、言葉では表現しにくい物理的な情報を直感的にわかりやすく表現できる	長所	具体的な事象から抽象的な概念まで正確かつ詳細に表現でき、画像にあらわれない背景的・解釈的な情報も表現できる
画像情報だけで読み手がその意味を正しく理解するのは難しい	欠点	表現が冗長かつ複雑で、直感的な理解が難しい
トップダウン的な理解 全体的な理解から詳細な理解へ	理解の方向	ボトムアップ的な理解 個々の理解から全体の理解へ

表 1: 画像情報と言語情報の相補的な特徴

画像とテキストの対応、相互参照に関する理解 画像と言語のそれぞれの要素がどのように対応しているか、お互いにどのような情報を付加しあっているかについての認識

説明の対象に関する理解 言語および画像の情報がどんな意味をもっているのか、その構成(格フレームや空間的関係)が何を表しているのかについての解釈

我々は図鑑の意味理解を図1のように考える。画像、言語それぞれのドメインには要素(画像の場合は構成要素、言語の場合は単語およびその他の構造)が存在する。画像の場合は空間的な関係や空間的属性(位置、形状、色など)によって、言語の場合は格フレーム内での関係などによって要素は互いに関係づけられている。画像とテキストの対応、相互参照に関する理解は、画像とテキストのそれぞれの要素間にリンクをはることに対応する(図1の太線)。説明の対象に関する理解は、要素間の関係づけ情報と図鑑の対象分野についての情報によって意味を解釈することに対応する(図1の破線)。図鑑の対象分野についての情報とは、概念のもつ意味、概念とその空間的属性との対応関係、概念間の上位－下位・全体－部分関係および空間的な位置関係や位相的接続関係などがあり、図鑑の意味理解には不可欠である。

そこでこうした情報の中から、図鑑のテキストの内容を理解し、画像の中の対象についての情報やその意味を抽出するのに必要な図鑑の対象分野の専門的な情報について検討する。

2.3 図鑑のテキストを理解するために必要な情報

図鑑のテキストでは写真や図などの中の要素やそれに対応する概念がもついろいろな属性情報が並べられ、それらの要素や概念の意味を説明する。しかしそこで用いられている概念の意味や概念間の関係、概念がもつ属性などの図鑑の対象分野についての情報はテキストには陽に記述されていない。例えば植物図鑑では「葉は楕円形」「高さ20メートル」といったそれぞれの植物体に固有な属性情報はテキストに記述されているが、葉が《形》という属性をもっているとか根、茎、葉が植物の基本

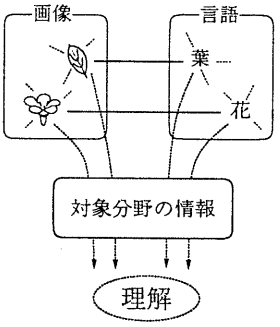

図 1: 図鑑の意味理解

セイヨウミザクラ
Prunus avium L.
いわゆるサクランボ、オウトウ(桜桃)で、明治初年に果樹として導入された。落葉高木で、高さ 20m にもなり、ピラミッド状の樹形をもつ。葉身は倒卵状長楕円形、先は短い鋭尖形、基部は広いくさび形、長さ 6-12cm で、ふぞろいな鈍鋸歯があり、基部(または葉柄の上部)に蜜腺があり、裏面は脈に沿って伏毛を生じる。花期は 4-5 月で、花は葉とほぼ同時に開く。花序は散形状。萼筒は長さ約 5mm、萼裂片は長楕円形、先は鈍形で、萼筒とほぼ同長になる。花弁は白色、倒卵形で先は円形、長さ 11-13mm。花柱は無毛。果実は球形、径 15-25mm、原種では黄赤色に熟す。西アジア原産であるが、ヨーロッパ東部には野生状態で生えているという。日本ではおもに山形・福島・山梨県などで栽培される。

図 2: 植物図鑑のテキストの例

的な器官であるという情報は説明されていない。そこで図鑑のテキストの内容を理解するために必要な対象分野の情報を

1. どのように獲得するのか
2. どのように表現するのか

という 2 つの問題を解決しなければならない。本研究ではそうした情報の獲得方法として、文章表現を詳しく調べて詳細なパターンを設定するという従来の方法を用いない。我々は簡単なパターンで重要な情報を含む表現を取り出して蓄積し、それを自動的に意味分類して詳細な情報抽出を可能にするメタな情報を獲得するという方法を考えた。したがって図鑑の対象分野の情報を表現する方法も情報の追加が容易な表現形式であることが重要である。

3 図鑑のテキストの特徴とその解析に必要な情報

植物図鑑では図 2に示すように【ミヤマザクラ】【イチイ】といった植物の種を単位にして、その植物について以下の情報がテキストに記述されている。

形態情報 植物体の外部的形態のありさまについての情報

生態情報 植物の分布状況や生育環境についての情報

その他の情報 用途、名前の由来、分類上のエピソードなど画像の内容とは関わりのない背景的な情報

形態情報と生態情報は説明の対象となっている画像の中の要素の属性を表し、他の植物と区別することを可能にする重要な情報である。そこで形態情報と生態情報が図鑑のテキストでどのように表現されているのかを明らかにする。

3.1 図鑑のテキストの表現パターンとその意味

植物図鑑のテキストで形態および生態上の情報を表現するパターンは次の 2 つにまとめられる。

用言を中心にした文 植物体の状態(変化)、あるいはサ変動詞で表現される属性についての情報を用言を中心にした格フレームで表現する。

(例文 1) 葉には鋸歯がある (状態)

(例文 2) 花序は軸が袋状になる (状態変化)

(例文 3) 花序は頂生する (属性 [花序の種類])

名詞述語文 以下の例文のように名詞述語文とは述語が名詞で形成されている文である。植物体の形態・生態についての属性情報の多くはこの名詞述語文によって表現される。

(例文 4) 果実はほぼ球形だ (属性 [形状])

(例文 5) 花序は長さ 1． 5 — 3 c m (属性 [長さ])

名詞述語文の基本的な意味関係は、1. 下位・上位関係、2. 同一関係、3. 対象・属性関係、4. 対象・事象関係、5. 要素・集合関係、6. うなぎ文関係、7. 比喩関係、8. 同語反復同一関係である[4]。しかし、植物図鑑のテキストでは植物のもつ属性情報を中心に記述されているので、その意味関係は 3 の対象・属性関係が大部分を占める。我々は、属性情報は画像の中の対象を理解するのに重要な情報であり、また図鑑のテキストから属性情報を表現する名詞述語文を抽出するのは容易であると考え、植物の形態・生態についての属性情報の抽出対象を名詞述語文にしぼった。以下、名詞述語文における単語、並列構造の特徴について述べる。

3.2 名詞述語文にあらわれる単語の特徴

名詞述語文の主語も述語もその大部分が専門用語である。すなわち、名詞述語文の主語が表す植物の形態的あるいは生態的特徴も、述語が表すその属性情報も専門用語で表現される。これは専門用語を用いると多義性の問題をさけ、情報を正確かつ簡潔に表現できるからである。しかし、専門用語は一般的な辞書(シソーラス)には登録されていない。

専門用語は一般に複数の名詞から構成される複合名詞であるので、意味的に近い専門用語には以下のような特徴がある。

- 全般的に字面が似ている。すなわち一致する文字が多い。
- 共通する構成要素の名詞(構成語)をもつ。日本語の単語の平均長は 2 文字程度なので、共通する構成語は連続して一致する文字列としてとらえることができる。
- 一般に最後の構成語が専門用語全体の意味を支配するので、語尾が一致する。

以上の特徴から、専門用語間の意味的な類似度の評価には文字

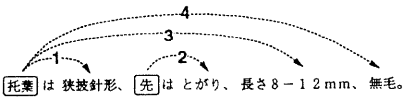

図 3: 直前の主語ではなく、もっと前の主語が述語にかかる例

列の一致を考慮した方法が有効であると考えられる。

3.3 名詞述語文の述語の並列

図鑑の名詞述語文では主語が表す概念はふつう複数の属性をもつので、複数の属性値を表すために述語は並列構造を含むことが多い。述語の並列の表現には次の３つがある。

「で」による並列 異なる意味分類に属する属性値が並ぶ。

(例文 6) 果実は核果で球形

「または」による並列 同じ意味分類に属する属性値が並ぶ。「または」以外に述語内で同じ意味分類の属性値を結ぶものに「あるいは」「および」「から」「ないし」「まれに」「や」「～」がある。

(例文 7) 葉は卵形または長楕円形

読点「、」による並列 異なる意味分類の属性値の並列と、同じ意味分類の属性値の並列がある。

(例文 8) 葉は単葉、革質

(例文 9) 葉は単葉、３出または奇数羽状複葉

最初の２つの並列は述語内の並列であるが、「、」による並列は述語の並列である。植物図鑑のテキストでは「、」による述語の並列が多く含まれ、１つの主語が複数の述語にかかる。

3.4 図鑑のテキストにおける主語と述語の係り受け関係

１文中に複数の主語があると、述語にかかる主語も直前にあらわれた主語だけでなく、それよりも前にあらわれた主語がかかる場合もある (図 3)。このような係り方は階層的な説明が行なわれている表現でみられる。階層的な説明とは説明の途中で説明対象の下位・部分概念の対象の説明を挿入し、もとの対象の説明を補足する説明の方法である。下位・部分概念の主語の説明が終わってから再びもとの主語の説明が行なわれると、図3のような係り受けがおこる。図 3では「先」は「托葉」の先端という部分概念で、「先はとがり」は「托葉」の説明を補足している。主語が省略されている述語にかかる主語は非交差条件が成り立つので文法的には、

1. 直前の述語にかかる主語
2. 文中で 1より前にあらわれ、その上位・全体概念である主語

のいずれかである。意味的にも複数の主語がかかる可能性がある場合は、述語に位置的に近い主語がかかる傾向が高い。

4 名詞述語文を解析するための情報の獲得

4.1 名詞述語文の意味解析と係り受け解析

名詞述語文から形態・生態情報を抽出するには名詞述語文の主語と述語の間の意味関係を明らかにして、それにしたがって意味を決定する方法が必要である。3.1節で述べたように、図鑑のテキストにおける名詞述語文の基本的な意味関係はその大部分が対象・属性関係である。本研究では図鑑のテキストにあらわれる名詞述語文の意味解析とは前述した８つの基本的な意味関係の決定ではなく、どんな属性を表しているのかを解析することを意味するものとする。

本研究では名詞述語文の意味解析には用例にもとづく方法を用いる。そのため名詞述語文の主語が表す概念がもつ属性の種類とその属性値の例から構成されている属性情報を用意する。意味解析は、属性値の例の中から述語に最も類似している属性値を文字列間の類似度にもとづいて判定し、それが属する属性を名詞述語文の意味関係と決定することで実現する。属性値の例は名詞述語文の述語を文字列間の類似度にもとづいて自動分類したものを用いる。この方法の有利な点は、規則にもとづく意味解析に比べシステムの作成や拡張が容易なことである。

係り受け解析には意味解析の結果を用いる。概念間の上位－下位・全体－部分関係の情報を用いて述語にかかる可能性のある主語を取り出す。そして文中での位置が述語に近いものから順に意味解析を行ない、属性値の例と述語の類似度の最高得点が一定値以上であるならその係り受けが正しいと判定することで係り受け解析を実現する。この処理には意味解析で用いる属性情報の他に、述語にかかる可能性のある主語を取り出すための概念間の上位－下位・全体－部分関係の情報が必要である。

そこで、名詞述語文の主語になる単語のもつ属性情報と上位－下位・全体－部分関係の情報を作成する手順を以下で説明する。

4.2 上位－下位および全体－部分関係の作成

対象を理解するのに必要な概念の上位－下位および全体－部分関係の情報は図鑑のテキストでは説明されていない。しかし概念自体はそれがもつ属性情報や状態 (変化) を表す文の主語としてテキスト中にあらわれている。そこでテキストからこれらの概念を提題の助詞「は」を手がかりにして取り出す。本研究では「葉」の理解に重要な概念を取り出すために「葉は」という文字列を含む文から「は」を手がかりに名詞 29 個を取り出した。そして生物学辞典の語義文や生物図説の解説図などを参考にして、抽出した名詞に人手で上位－下位および全体－部分の関係づけを行なった。

4.3 主語になる概念の属性情報の獲得

植物図鑑のテキストから属性を表す名詞述語文を大量に抽出し、述語内の並列構造の検出と文字列間の類似度にもとづいて述語の名詞を自動的に分類する。

4.3.1 名詞述語文の抽出

名詞述語文の述語の自動分類の前処理として図鑑のテキストから以下の手順で名詞述語文を抽出し、用例として蓄積する。

1. 形態素解析を行ない[5]、句読点で文を分割し、(a) 提題の助詞「は」を含み(ただし「には」「では」は除く)、(b) 用言を含まず、(c) 「に」など他の格要素をとらないものを名詞述語文として抽出する。したがって用言による連体修飾がある名詞述語文は抽出しない。また「、」による並列で主語が省略されている名詞述語文も抽出しないので、例文8、9からは「葉は単葉」の部分だけを抽出する。
2. 「で」による並列を以下に示すように分割し、述語が表す属性が1つになるように異なる意味分類に属する属性値の並列を解消した。
   a. AでBはC ⇒ BはC
      羽状脈で側脈は12–16対 ⇒ 側脈は12–16対
   b. AはBでC ⇒ AはB, AはC
      果実は核果で球形 ⇒ 果実は核果, 果実は球形
   c. AはBでCはD ⇒ AはB, CはD
      花弁は卵形で先は円形 ⇒ 花弁は卵形, 先は円形
3. 属性の種類とは関係のない「やや」などの副詞を述語から取りのぞく。

この抽出方法によって葉およびその下位概念を表す29個の名詞を主語にとる923種類の名詞述語文を抽出した。その99.1 %(915個)の意味関係は対象・属性関係であった。

4.3.2 文字列の一致による専門用語間の類似度

植物図鑑では名詞述語文の主語、述語は一般に専門用語である。3.2節で専門用語間の類似度の評価には文字列の一致を考慮した評価が有効であると述べた。そこで2つの専門用語の文字列を $A = a_1 a_2 \cdots a_x$、$B = b_1 b_2 \cdots b_y$ とすると、1. 共通する文字の数、2. 連続した文字列の一致、3. 語尾の一致を評価できるようにその類似度 $SCORE(A,B)$ を次のように定義した[6]。

$$SCORE(A,B) = \begin{cases} \mathrm{s}(x,y) & \text{if } m(x,y)=0 \\ \mathrm{s}(x,y)+2 & \text{if } m(x,y)=1 \end{cases}$$

$$s(i,j) = \begin{cases} 0 & \text{if } i=0 \vee j=0 \\ max \begin{pmatrix} s(i-1,j-1) \\ +min\,(cm(i,j),W), \\ s(i-1,j), \\ s(i,j-1) \end{pmatrix} & \text{if } (1 \le i \le x) \wedge (1 \le j \le y) \end{cases}$$

$$cm(i,j) = \begin{cases} 0 & \text{if } i=0 \vee j=0 \\ (cm(i-1,j-1)+1) \cdot m(i,j) & \text{if } (1 \le i \le x) \wedge (1 \le j \le y) \end{cases}$$

$$m(i,j) = \begin{cases} 1 & \text{if } a_i = b_j \\ 0 & \text{if } a_i \neq b_j \end{cases}$$

W は連続する文字列の一致に対するボーナスの最大値。本研究では佐藤と同様に $W=4$ とする。語尾の一致にもボーナスを設定し2点を与える。なお、上の定義では「長さ 7cm」のように数字が構成語である場合について考慮していない。そこで本研究では述語から数字を取り除いて類似度を計算する。

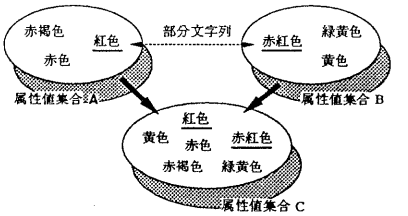

図 4: 末尾の部分文字列による分類の融合

4.3.3 名詞述語文の述語の自動分類

述語として用いられている専門用語を主語ごとに以下の手順で自動分類する。

1. 述語内の並列構造による分類
   3.3節の例文7のように述語には「または」などの接続詞によって同じ意味分類の属性値が並列に記述されている場合がある。そこでそれらの接続詞を手がかりに並列構造を検出し、並列する専門用語(例文2では「卵形」と「長楕円形」)を抽出し、同じ属性値集合に分類する。
2. 専門用語間の類似度による分類
   4.3.2節で定義した専門用語間の類似度を用いて述語の専門用語を分類する。属性値集合の初期状態には述語内の並列構造による分類で得た属性値集合を用いる。未分類の述語がなくなるまで2-1 ～ 2-3の処理をくりかえす。
   2-1. 未分類の専門用語と分類済みのすべての専門用語との類似度を4.3.2節の式を用いて計算する。
   2-2. 類似度の最高得点が3点以上の場合、すなわち、(a)3文字以上が一致、(b)2文字以上の連続一致、(c)最後の1文字が一致のいずれかの条件をみたす場合、最高得点を獲得した専門用語の属する属性値集合に未分類の専門用語を分類する。
   2-3. 類似度の最高得点が3点未満の場合は、新しい属性を表す属性値があらわれたとみなし、新しい属性値集合を作成し、そこに分類する。
3. 末尾の部分文字列による分類の融合
   図4の属性値集合A、Bのように1つにまとまるべき属性値集合が2つ以上に分裂している場合がある。この原因は、1の述語の並列構造だけでは不十分な分類しかできず、2の専門用語間の類似度による分類の初期状態ですでに属性値集合が分裂しているためである。

   そこで、ある属性値集合の要素aが他の属性値集合の要素bの末尾の部分文字列であるなら、bはaの特殊化された概念であるとみなし、bの属する属性値集合はaの属する属性値集合の部分集合であると考え、2つの属性値集合を1つにする。図4では「赤紅色」が「紅色」の特殊化された概念であるとし、

表 2: 自動分類と人手による分類の比較

述語の自動分類の結果	集合数
人手による分類と一致する属性値集合	96
人手による分類と一致しない属性値集合	25
合計	121

人手による分類と一致しない属性値集合の内訳	集合数
1つになるべき属性値集合が複数に分裂	23
本来独立であるべき属性値集合が1つに融合	2
合計	25

1つになるべき属性値集合が複数に分裂した理由	集合数
数量の単位の不一致 (例: 葉の枚数 ⇒ 枚, 個, 対)	11
属性値集合の初期状態がわかれすぎ	6
専門用語の字面が似ていない	4
一般的な名詞で属性が表現されている	2
合計	23

属性値集合 A、B をまとめて新しい属性値集合 C をつくっている。

4.3.4 名詞述語文の述語の自動分類の結果

名詞述語文 915 文を対象に述語の自動分類を行ない、121 個の属性値集合を得た。人手による分類では 109 個であった。分類の比較の結果を表 2に示す。自動分類の結果の 79% が人手による分類と一致した。自動分類した結果は人手で修正し、それぞれの属性値集合にはその集合が表す意味のラベルをつけた。分類に失敗した属性値集合を検討すると、1つにまとまるべき属性値集合が複数に分かれてしまった失敗が多い。本研究では部分文字列を用いた属性値集合の融合を行なっているが、より強力な融合方法を検討する必要があるかもしれない。しかし、名詞述語文の用例をもっと大量に蓄積することで改善できる可能性もある。

5 図鑑のテキストの解析

4章で作成した「葉」に関する概念がもつ属性情報と概念間の関係の情報を用いて図鑑のテキストにあらわれる名詞述語文に対して

1. 主語の係り受け解析
2. 述語が表す属性の意味解析

を行なうシステムの概要を説明する。さらにシステムの評価のため図鑑のテキストから無作為に選んだ 200 文に対して実験を行ない、システムの有効性を評価する。

5.1 システムの構成

作成したシステムの概要を図 5に示す。システムは前処理を行なう部分と名詞述語文の係り受け解析と意味解析を行う部分の2つから構成されている。

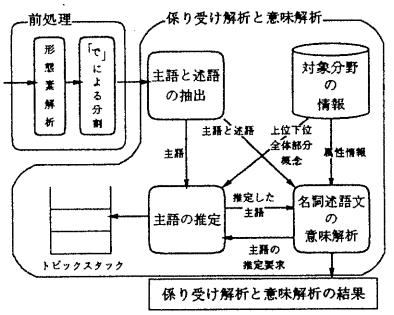

図 5: システムの概要

5.1.1 前処理

前処理は形態素解析と「で」による並列の分割の2つの処理から構成されている。形態素解析の結果からは名詞述語文であるかの判定に必要な品詞情報を得る。「で」による分割は述語が表す属性を1つにするためのものである。下に示すように主語が省略された名詞述語文の述語も分割する以外は 4.3.1節の名詞述語文の抽出処理における「で」の分割と同じである。

- A で B ⇒ A, B

  楕円形で鈍頭 ⇒ 楕円形, 鈍頭

5.1.2 名詞述語文の係り受け解析と意味解析

この処理は植物に関する概念間の関係と概念のもつ属性についての情報、主語の履歴を保存する動的な記憶 (以後トピックスタックとよぶ)、そして以下で説明する 3 つのモジュールから構成されている。図 6は図 3の例文の解析を行なっているときのシステムの動作状態を示している。

1. 主語と述語の抽出モジュール

   このモジュールでは主語と述語の抽出と名詞述語文であるかどうかの判定を行なう。前処理の結果を句読点で分割し、主語があるなら「は」を手がかりにそれを取り出す。取り出した主語は主語の推定モジュールへ送る。図 6では 1 と 2で主語が送られている。名詞述語文でなくても主語の抽出は行なう (図 6の 2)。次に (a) 用言を含まず、(b) 「に」などの他の格を含まないものを名詞述語文と判定し、主語と述語を名詞述語文の意味解析モジュールへ送る。もし主語が省略されていたら述語のみを送る。図 6の 1では主語と述語を、3 と 4では述語のみを送っている。

2. 主語の推定 (トピックスタックの操作) モジュール

   このモジュールはトピックスタックの操作と述語にかかる可能性のある主語を意味解析モジュールに送ることを行なう。トピックスタックは後入れ先出し (LIFO) で、最も最近格納された主語ほど早く取り出せるようになっている。名詞述語文の意味解析モジュールへはトピックスタックに最後に格納した主語から順に意味解析モジュールの要求に応じて送る。

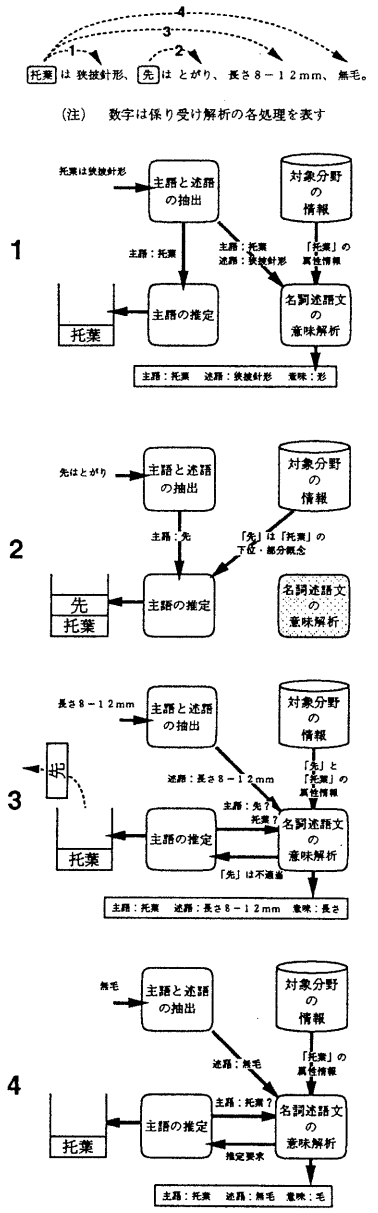

図 6: 解析システムの動作例

スタックの操作は新しい主語が送られてきた場合と主語の再推定の場合に行なう。新しい主語が 1の主語と述語の抽出モジュールから送られてくると、すでにスタックに格納している主語から新しい主語の上位・全体概念ではないものを取り除く。これは別の主語を飛び越して述語にかかることができる主語の情報を更新することに対応している。図 6の 3の「先」のように推定した主語が意味的に不適当と判定され、名詞述語文の意味解析モジュールから主語の再推定の要求された場合は、最後に格納した主語 (図 6の 3では「先」) の説明が完了したと判定してこれを捨て、スタックから次に取り出せる主語 (図 6の 3では「托葉」) を名詞述語文の意味解析モジュールへ送る。もし新たに送る主語がない時は名詞述語文の意味解析モジュールへは主語を送らない。この処理は述語に近い主語から順に係り受けが意味的に適切かどうか試すことを実現している。

3. 名詞述語文の意味解析モジュール

主語と述語の抽出モジュールから送られてきた主語と述語を入力にして主語のもつ属性情報を用いて名詞述語文の意味解析を行なう (図 6の 1)。主語が省略されているときは主語の推定モジュールに主語の推定を要求し、推定された主語で意味解析を行なう (図 6の 3、 4)。意味解析の妥当性によって係り受け解析も実現する。

名詞述語文の意味解析には用例にもとづく方法を用いる。自動分類された属性値の例と述語の類似度を 4.3.2節の式で計算し、最高得点が 3 点以上の場合は最高得点を獲得した属性値が属する集合の意味のラベルを解析対象である名詞述語文の意味関係とする。 3 点未満の場合は、別の係り受けの可能性があれば主語が妥当でなかったと判定し、主語の推定モジュールへ主語の推定要求を送り、送られてきた新しい主語で再び意味解析を行なう。図 6の 3では「先」の属性情報では最高得点が 3 点未満だったので、「托葉」を主語にして再び意味解析を行なった。別の係り受けの可能性がない場合には新しい属性概念があらわれたと判定する。

5.2 実験と結果

植物図鑑のテキストから葉について記述している文を 200 文無作為に抽出し、作成したシステムで意味解析と係り受け解析を行なった。

意味解析の対象となる名詞述語文は 200 文中に 384 個存在した。意味解析の結果を表 3に示す。意味解析の失敗は 17 個で、その内訳を表 4に示す。意味解析の失敗のうち、誤った主語の属性情報を用いて意味解析を行なった失敗を例をあげて説明す

成功	失敗	新しい属性概念	合計
335	17	32	384

表 3: 名詞述語文の意味解析の結果

意味解析に失敗した理由	名詞述語文
誤った主語を推定し、意味解析を行なった	4
主語が上位－下位・全体－部分関係に含まれていない用語	1
主語の一部が省略されている (例:「多く (の托葉) は」)	3
主語が代名詞であり、指示対象がわからない	2
主語が「植物体の一部 + 空間語」(例:「葉脈上は」)	4
主語が「植物体の一部 + 属性名」(例:「葉の形は」)	3
合計	17

表 4: 意味解析の失敗の理由の内訳

新しい属性概念の検出の内訳	名詞述語文
主語が葉とは関係のない用語だった	24
抽出した用例からでは獲得できなかった属性概念	8
合計	32

表 5: 新しい属性概念の検出の内訳

る。

(例文 10) 葉は 1 小葉からなり、葉柄は長さ 3–10mm、楕円形

「楕円形」の主語は「葉」なのだが、文法的には「葉」「葉柄」のどちらも「楕円形」に主語としてかかることができる。「葉柄」の属性情報に「円柱形」という属性値の例があるので「楕円形」の主語は「葉柄」であると判定し、そのため意味解析と係り受け解析の両方とも失敗した。新しい属性概念の検出が行なわれたのは 32 個で、その内訳を表 5に示す。4.3節の処理で抽出した名詞述語文の用例からは獲得できなかった属性概念の例として、「頂小葉」の《葉縁の種類》がある。これは図鑑のテキストでは「頂小葉」の《葉縁の種類》を表す述語が「、」による並列の 2 番目以降にしかあらわれなかったからである。このため名詞述語文の用例として蓄積できなかったので、属性概念も獲得できなかった。

次に係り受け解析について検討する。 384 個の名詞述語文のうち、主語が省略されていて係り受け解析が必要な名詞述語文は 168 個であった。係り受け解析の結果は表 6に示す。そのうち図 3のように主語の飛び越えが発生した表現での係り受け解析の結果を表 7に示す。係り受け解析の失敗は意味解析の失敗が原因である。失敗の内訳を表 8に示す。

5.3 検討

作成したシステムで解析を行なった結果、名詞述語文の意味関係の 87%、係り受け関係の 96% について解析に成功した。属性情報のかなりの部分を自動的に作成したことを考えると、これはかなり良い結果である。一方、次のような問題が明らかになった。

1. 述語に複数の主語がかかる可能性があるとき、どの主語の係りを優先するのかは主語の位置のみで決めている。係り受け解析の失敗の原因の半分はこの優先規則によるものなので、係り受け解析の精度を向上させるため属性値の例と述語の類似度による評価などを導入する必要がある。

成功	失敗	合計
161	7	168

表 6: 名詞述語文の係り受け解析の結果

成功	失敗	合計
16	7	23

表 7: 主語の飛び越えの解析結果

係り受け解析の失敗の原因	名詞述語文
誤った主語を意味解析で適当と判断した	4
主語が代名詞であり、指示対象がわからない	2
主語が「植物体の一部 + 属性名」(例:「葉の形は」)	1
合計	7

表 8: 係り受け解析の失敗の原因

2. 主語の一部が省略されたり、代名詞であったりして指示対象が何であるか明らかでない場合は意味解析に失敗する。この問題を解決するためには、指示対象を明らかにする研究を行なわなければならない。
3. 「頂小葉」の《葉縁の種類》のように属性情報が十分に獲得できなかった概念がある。しかし、係り受け解析が非常に精度良く行なえたので、その結果を用いて名詞述語文の用例を追加すればより精密な属性情報が獲得でき、この問題を解決できると考えている。

6 おわりに

意味解析と係り受け解析によって新たに得た名詞述語文を用いてより精密な属性情報を作成して、より高精度の意味解析、係り受け解析を実現し、画像の内容についての情報を抽出する予定である。また、抽出した画像の内容についての情報によって図鑑に収められている画像の内容検索を実現する方法についても研究を行なう。

謝辞 植物の専門用語について助言していただいた京都大学理学部植物学教室の丹下晴美氏に感謝いたします

参考文献

(1) 渡辺, 中村, 長尾: 絵画解説文の対象情報・感性的情報の抽出, 情報処理学会研究報告 93–CH–20 (1993).

(2) 鶴丸 他: 単語の釈義文を利用した単語間の階層関係の抽出について, 情報処理学会, 自然言語処理研究会資料 45–4 (1984).

(3) 中村, 古川, 長尾: 概念図理解を目的としたパターン情報と自然言語情報の統合, 信学技報, AI93–67 (1993).

(4) 荒木, 桃内: 名詞述語文における意味概念の学習, 情報処理学会研究報告 90–NL–79 (1990).

(5) 松本 他: 日本語形態素解析システム JUMAN 使用説明書 ver.1.0., 京都大学長尾研 (1992).

(6) 佐藤: 用例検索による日英翻訳支援システム CTM2, 研究報告, 北陸先端科学技術大学院大学 (1993).