自然言語処理 108－4
(1995. 7. 20)

植物図鑑のテキストと図による植物用語間の階層関係の獲得

渡辺 靖彦　　長尾 真
京都大学工学部 電気工学第二教室
〒606-01 京都市左京区吉田本町
e-mail: watanabe@kuee.kyoto-u.ac.jp

あらまし
われわれは画像データから直接取り出すのが困難な画像の内容情報を、画像の内容を解説するテキストから自然言語処理を行なって取り出すことについて研究を行なっている。われわれは画像とその内容を説明するテキストの例として植物図鑑を選び、そのテキストを理解することをめざしている。
本稿では図鑑のテキストを理解するのに重要な用語間の (1) 上位－下位関係、(2) 全体－部分関係、(3) 属性の情報を図鑑の図とテキストから獲得する方法を説明する。そして、これらの情報を用いて図鑑のテキストにおける係り受けのあいまいさを解消する方法を説明し、実験でその有効性を確かめた。

キーワード　用語の階層関係, 専門用語, 説明の順序, 図の理解, 意味処理, パターン情報と自然言語情報の統合

Automatic Acquisition of Hierarchical Information of Words Using Diagrams and Text of Pictorial Book of Flora

Yasuhiko Watanabe　　Makoto Nagao
Department of Electrical Engineering II, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto 606-01, Japan
e-mail: watanabe@kuee.kyoto-u.ac.jp

Abstract　It is difficult to extract contents of images from image data itself. To avoid this difficulty, we intend to extract the content information of images from the explanation texts of image data. We select a pictorial book of flora as the explanation texts of image data.
In this paper, we described a new method of acquisition for hierarchical information of words: (1) IS-A information, (2) PART-OF information, (3) property information. We obtained these information from the explanation texts of image data and diagrams in the pictorial book of flora. Then, we used these information for the analysis of modifier phrases and the semantic analysis of copular sentences in the pictorial book of flora.

key words　hierarchical information of words, technical term, explanation order, diagram understanding, semantic analysis, integration of pattern information and natural language information

1 はじめに

より高度で柔軟な情報処理を実現するためには、画像情報と自然言語情報を統合して計算機で扱うことが重要になるとわれわれは考えている。画像情報と自然言語情報を相補的に用いれば、いずれか一方の情報だけでは解決が困難な問題が取り扱えるようになることがあるからである。例えば、画像内容の理解や画像の内容検索には、画像情報と自然言語情報の統合が重要な役割をはたすと考えられる。なぜなら、画像の内容情報を画像データから直接取り出すことは現在の画像処理の技術では困難だからである。そこでわれわれは画像の内容を解説しているテキストを利用することを考えた。すなわち、画像データの内容情報を画像データそのものからではなく、画像の内容を解説しているテキストから自然言語処理によって画像の内容情報を取り出すのである。このようにわれわれは、画像情報と自然言語情報の統合の研究の１つとして、画像の内容を解説するテキストを理解するための研究を行なっている [渡辺 94][渡辺 95A][渡辺 95B]。

われわれは画像データとその内容を解説するテキストの例として植物図鑑を選んだ。図鑑のテキストには

- 空間的な構造をもつ対象を説明する
- 対象分野の専門用語を多用する

という特徴があり、一般的なシソーラスや単語意味辞書では図鑑のテキストを正確にまた詳細に理解することは困難である。このため、専門用語のシソーラスなどが必要であるが、それらを人手で作成するには膨大な時間と対象分野の専門知識が必要である。そこで図鑑の対象分野について説明しているテキストから図鑑のテキストを理解するのに必要な情報を取り出すことを考えた。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が研究されている [鶴丸 84]。本研究では、図鑑のテキストを理解するのに必要な情報を

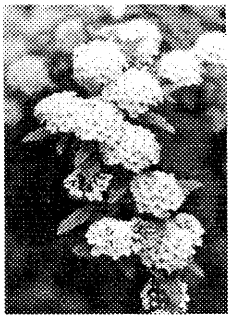
- 図鑑のテキスト
- 図鑑の図

の２つから取り出すことにした。

本論文では、最初に、図鑑の図で説明されている情報を考察し、その結果から図鑑のテキストを理解するのに重要な情報が用語間の

1. 上位－下位関係
2. 全体－部分関係
3. 属性関係

の３つの情報であることを明らかにする。次に、その用語間の関係情報を図鑑の図とテキストから獲得する方法を説明する。そして、獲得した情報を用いると図鑑のテキストにおける係り受けのあいまいさを解消することができることを明らかにし、獲得した情報の有効性を示す。



コデマリ

spiraea cantoniensis

落葉低木、高さ 1.5m に達する。枝は細く、円柱形、無毛、幼時に暗紅褐色を呈し、しばしば先が枝垂れる。葉はひし状狭卵形またはひし状楕円形、鋭頭で基部はくさび形、長さ 2-5cm、幅 6-20mm、中部以上に不整鋸歯がつき、両面無毛、裏面はしばしば粉白色で、脈が突出する。葉柄は長さ 4-7mm で無毛。花序は今年枝に頂生し、散房状、無毛。花は 4-5 月に咲き、白色で、径 7-10mm、単弁。萼裂片は３角形で鋭頭、花後に反曲しない。花托は内面に白短毛が密生する。袋果は無毛。中国 (中南部) 原産。日本では広く庭園に栽培される。名は球形の花序の様子を小型の手毬に見たてたもの。

図 1: 植物図鑑の写真とテキストの例

2 図鑑のテキストを理解するのに必要な情報

植物図鑑では読み手に植物を理解させる手段として

- 植物の生態・環境を表す写真 (あるいはスケッチ画)
- 植物の特徴を解説するテキスト
- 植物の器官の名称や形態的特徴を解説する図

の３つを一般に用いる。写真とテキストの例を図 1に、植物用語を解説する図 (以後、**植物用語の解説図**とよぶ) の例を図 2に示す。図鑑のテキストを理解するにはさまざまな情報が必要であるが、その中でも特に重要な情報は植物用語の解説図によって説明されているとわれわれは考えた。なぜなら、解説図を参照しながらテキストを読むと、テキストで用いられている植物用語 (専門用語) 間の関係が明らかになり、より深い理解がえられるからである。このため、植物用語の解説図で表現されている情報を調査すれば、植物図鑑のテキストを理解するのにはどんな知識が必要なのか明らかになると考えた。

[渡辺 95A] で行なった調査の結果、植物図鑑の図には以下の５つの情報が説明されていた。

1. 植物の器官の種類 (例: 図 2の上の図の「袋果」「ミカン状果」など)
2. 植物の器官の名称 (例: 図 2の左下の図の「托葉」など)
3. 植物の器官の属性 (例: 図 2の右下の図の「楕円形」など)
4. 種の名前 (例: 図 2の上の図の「カツラ」「ナツミカン」など)
5. 補足説明

植物の器官の種類とは、同じ機能をもつが、植物の種や環境

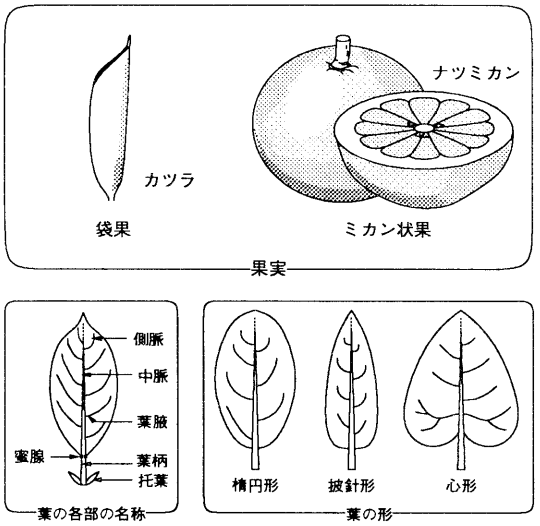


図 2: 植物用語の解説図の例

の違いによって形が異なる器官を説明する情報である。

上の5つの情報のうち、植物の器官の名称、種類、属性の3つは植物用語間の関係を表現している。図2の上の図(「果物」の図)の「袋果」「ミカン状果」は植物の器官の種類を表している。この図から「袋果」「ミカン状果」は「果物」の一種であり、「果物」と「袋果」「ミカン状果」の間には上位－下位の関係があることがわかる。図2の左下の図(「葉の各部の名称」の図)の「托葉」は植物の器官の名称を表している。この図では「托葉」は絵で表現されている「葉」と矢印によって接続され、「托葉」が「葉」の一部であることが示されている。このことから「葉」と「托葉」の間には全体－部分の関係があることがわかる。図2の右下の図(「葉の形」の図)の「楕円形」は属性を表している。この図から「楕円形」は「葉」の「形」という属性の属性値であることがわかる。

以上の調査結果から、図鑑のテキストを理解するには植物用語間の

1. 上位－下位関係
2. 全体－部分関係
3. 属性関係

の情報が重要であることがわかる。そこで3章ではこれらの情報を図とテキストを用いて獲得する方法について説明する。

3 図とテキストを用いて用語間の関係を獲得する方法

本章では図とテキストの情報を用いて、植物用語間の(1)上位－下位関係、(2)全体－部分関係、(3)属性の情報を獲得する方法について述べる。

3.1 用語間の上位－下位関係を獲得する方法

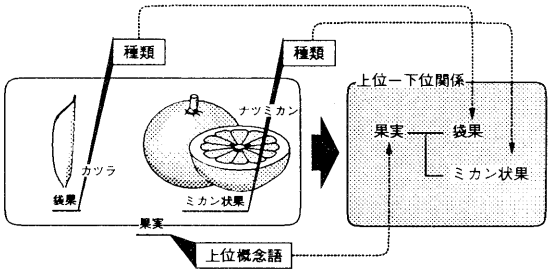


図 3: 解説図から用語間の上位－下位関係を獲得する例

植物用語間の上位－下位関係は図鑑の解説図から以下の手順で獲得した。その概要を図3に示す。

1. 図鑑の解説図から用語間の上位－下位関係を表す図を取り出す

   用語間の上位－下位関係を表す図は、植物の器官の種類を表す単語がその図の中で用いられている図である。例えば、図3の「袋果」と「ミカン状果」は植物の器官の種類を表している。そこで、図の中で用いられている単語が植物の器官の種類を表している図を用語間の上位－下位関係を表す図として取り出す。単語が表している意味は[渡辺 95A]で提案した方法によって判定する。[渡辺 95A]で提案した方法とは、図のパターン情報とテキストの自然言語情報を統合して図の内容を理解するものである。[渡辺 95A]で行なった実験では、図の中で用いられている単語の意味を95%という非常に高い精度で判定できた。

2. 図の表題から上位概念に相当する用語を取り出す

   植物の器官の種類を表す植物用語(図3では「袋果」「ミカン状果」)の上位概念語が何であるかは、図そのものにではなく、図の表題の中で示されている。そこで、図3で示すように、表題から上位概念語を取り出す。表題は「果実」のように1つの名詞か、「葉の各部の名称」のように助詞「の」によって複数の名詞が結ばれた名詞句であることが多い。表題から上位概念語を取り出すには、表題を構成する名詞から植物の器官を表す名詞を取り出せばよい。なぜなら植物の器官の上位概念は植物の器官だからである。植物の器官は図鑑のテキストでは説明の対象であるので、図鑑のテキストで「は」格の格要素になる名詞を上位概念語として表題から取り出す。例えば、図2の表題の「果実」は例文1のように説明対象となり、「は」格の格要素になる。

   (例文1) 果実は 核果

   そこで、図3に示すように、表題から「果物」を「袋果」と「ミカン状果」の上位概念語として取り出した。

以上の方法で33個の用語間の上位－下位関係が得られた。

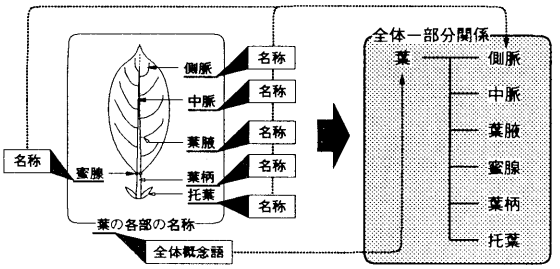


図 4: 解説図から用語間の全体－部分関係を獲得する例

3.2 用語間の全体－部分関係を獲得する方法

植物用語間の全体－部分関係は図鑑の解説図とテキストを用いて獲得する。

最初に図鑑の解説図から全体－部分関係を獲得する方法について述べる。われわれは以下の手順で用語間の全体－部分関係を解説図から獲得した。その概要を図 4に示す。

1. 解説図から用語間の全体－部分関係を表す図を抽出する
   用語間の全体－部分関係を表す図は、植物の器官の名称を表す単語がその図の中で用いられている図である。例えば、図 4の「托葉」などの単語は植物の器官の名称を表している。そこで、図の中の単語が植物の器官の名称を表している図を植物用語間の全体－部分情報を表す図として抽出する。図の中の単語の意味は [渡辺 95A] の方法で判定する。
2. 図の表題から全体概念に相当する用語を取り出す
   図 4では「托葉」などの単語の全体概念「葉」は絵で表されている。このように用語間の全体－部分関係を表す図では全体概念が単語ではなく絵で表現されているので、そのままでは用語間の関係情報はえられない。一方、図の表題には全体概念を表す名詞が含まれている。全体概念語も植物の器官を表すので、テキストで説明対象となる名詞 (「は」格の格要素になる名詞) を表題から取り出す。

以上の方法で 75 個の用語間の全体－部分関係が得られた。

次に図鑑のテキストから全体－部分関係を獲得する方法について説明する。図鑑のテキストには例文 2のような複文が非常に多い。

(例文 2) 托葉は卵形、先はとがり、長さ 10cm、早落する。

さらに、例文 2で説明の対象が「托葉」から「先」に変わったように、図鑑のテキストでは 1 文中に複数の「は」格があらわれて何度も説明の対象が変わることも多い。そこでわれわれはこうした説明対象の変化の情報、すなわち説明する順序の情報を用いて用語間の全体－部分関係の情報を獲得することを考えた。なぜなら、空間的な構造をもつ対象を説明する場合、その説明は大まかな構造から細かい部分的な構造

先 先端 両端 上部 上半 上半部 下部 基部 底部 触部
筒部 露出部 表面 裏面 背面 両面 外面 断面 横断面
外側 内側 向軸側 軸 中軸 着点

図 5: 本研究で用いた空間語

へと進んでいくのが一般的だからである。これは図鑑のテキストにもあてはまる。実際、いくつかのテキストを図鑑から取り出して調べたが、その説明は全体－部分の概念階層における上位の用語 (全体概念語) から下位の用語 (部分概念語) へと進められ、逆に下位の用語から上位の用語にさかのぼる例はなかった。しかし、全体から部分へという説明の順序は一般的ではあるが、常に成り立つ規則ではない。そこでわれわれは、何度も用いられる説明の順序は一般的な説明の順序で、全体から部分へという説明の順序にしたがっていると考えた。本研究では図鑑のテキストで 2 回以上用いられた説明の順序の例を対象に用語間の全体－部分関係の情報を取り出すことにした。

全体－部分関係の概念階層を下位から上位にさかのぼる説明がないのなら、新しい「は」格があらわれて説明対象が変化したとき、新しい説明対象は直前の説明対象の

- 部分語
- 直接の上位の用語 (全体語) ではない用語

のいずれかである。そこで新しい説明対象が直前の説明対象の部分語である例をテキストから大量に抽出し、それらをまとめると植物全般についての用語間の全体－部分関係がえられるとわれわれは考えた。この方法で用語間の全体－部分関係を取り出すには、新しい説明対象が直前の説明対象の部分概念語であるかどうかの判定が重要である。本研究では、新しい説明対象 B が直前の説明対象 A の部分語であるのは以下の 4 つの条件のいずれかを満たす場合とする。

(a) 図で用語 B が用語 A の部分語であることが明示されている
(b) 「A の B」という表現が図鑑のテキストに存在する
(c) 「A(に) は B がある」という表現が図鑑のテキストに存在する
(d) A が非空間語、B が空間語

「A の B」「A(に) は B がある」という表現は、名詞 A と B が図鑑のテキストで説明対象であるなら、A と B の間に全体－部分関係があることを示す (例文 3,4)。

(例文 3) 葉の先

(例文 4) 表面には毛がある

空間語とは「基部」「先端」など各器官の部位の位置を明示的に表す語である [山田 90]。本研究では図 5に示す 25 語を空間語として用いた。

以上の考え方にもとづいて図鑑のテキストから用語間の全

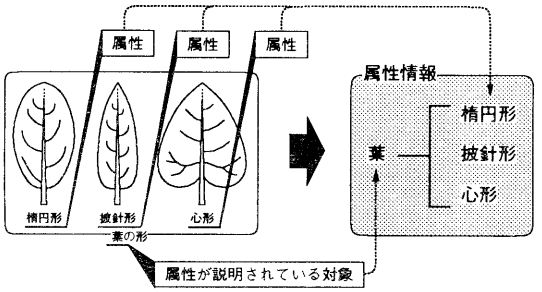


図 6: 解説図から属性情報を獲得する例

体－部分関係を以下の手順で取り出した。

1. 図鑑のテキストを形態素解析する [松本 92]。
2. 助詞「は」を手がかりに説明対象を検出し、テキストで 2 回以上用いられた説明の順序の例を取り出す。
3. 取り出した説明の順序の例が上で述べた 4 つの条件 (a) ～ (d) を満たすものを取り出す。

以上の処理で 130 個の用語間の全体－部分関係がえられた。

図とテキストから得られた用語間の全体－部分関係の情報を検討したが、不適当な全体－部分関係はなかった。そこでその結果をまとめ、植物用語間の全体－部分関係を木構造で表現した。この全体－部分関係の情報が植物図鑑のテキストを理解するのに有効な規模と内容をもつかは 4章で検討する。

3.3 用語間の属性関係を獲得する方法

植物図鑑を理解するのに必要な属性情報とは、植物の各器官の性質や状態を表すのにどのような用語が用いられるのかという情報であると考えた。植物の各器官の属性情報は解説図とテキストから獲得する。

最初に解説図から属性情報を獲得する方法を説明する。われわれは以下に示す手順で属性情報を解説図から獲得した。その概要を図 6に示す。

1. 解説図から属性情報を表す図を取り出す

   植物の器官の属性情報を表す図は、植物の器官の属性を表している単語がその図の中で用いられている図である。例えば、図 6の「楕円形」は植物の器官の属性を表している。[渡辺 95A] の方法を用いて単語の意味を判定し、植物の器官の属性を表している図を取り出す。
2. 属性情報が説明されている対象を表す用語を表題から取り出す

   図 6の「楕円形」などの単語で属性が説明されている「葉」は絵で表現されている。このように属性情報を表している図では属性が説明されている対象は単語ではなく絵で表現されているので、そのままでは属性情報をえられない。一方、図の表題には図で属性情報を説明している説明の対象を表す名詞が含まれている。そこで、上位－下位・全体－部分関係の場合と同じ方法で、表題から図で属性情報が説明されている対象を表す名詞を取り出す。

以上の方法で 112 個の属性情報が解説図から得られた。この情報は属性情報をテキストから獲得する処理で類義語情報として利用する。

次に、テキストから属性情報を獲得する方法について説明する。図鑑のテキストでは説明対象の性質や状態は主に

- 名詞述語文
  - (例文 5) 葉は鋭尖頭
  - (例文 6) 葉は卵形または長楕円形
  - (例文 7) 葉は鈍頭
- 用言を中心にした文
  - (例文 8) 葉はとがる
  - (例文 9) 葉は大きい
  - (例文 10) 葉は互生する

の 2 つで表現されている。そこでわれわれは、図鑑のテキストから主語と述語の係り受けにあいまいさがない名詞述語文と用言を中心にした文を取り出し、そこで用いられている述語を植物の器官の属性を表す述語の例として蓄積し、それらを属性情報とした。[渡辺 94] と [渡辺 95B] ではそれぞれ葉およびその下位・部分概念語の属性を表す名詞を 915 語、用言を 247 語取り出した。

名詞述語文の述語になる名詞は以下の 3 つの情報を用いて意味の近さを評価できる。

1. 名詞述語文の述部における並列構造

   名詞述語文の述部で並列する名詞は類義語である。例えば例文 6の述部では「卵形」と「長楕円形」が並列しているので、それらは類義語とみなす。
2. 名詞述語文の述部の名詞の字面の近さ

   名詞述語文の述語に用いられる名詞は専門用語であることが多い。そして意味が近い専門用語は一般に字面が似る。例えば例文 7の「鈍頭」は、例文 6の「卵形」「長楕円形」に比べ、より意味の近い例文 5の「鋭尖頭」に字面が似ている。これは、専門用語の多くが複合語で、意味が近い専門用語は共通する構成要素の単語 (構成語) をもつからである。そこで字面の近さを用いて名詞述語文の述語の名詞が表す意味の近さを評価する。字面の近さの評価には [渡辺 94] で提案した文字列の一致による評価を用いる。
3. 図から獲得した単語の類義関係

   属性情報を表す解説図から取り出した情報を用いる。同じ図の中で用いられる単語で、属性情報を表すものは類義語である。例えば、図 6の「楕円形」「披針形」「心形」は類義語とみなす。

そこで、この 3 つの情報を用いて説明する対象ごとにその属

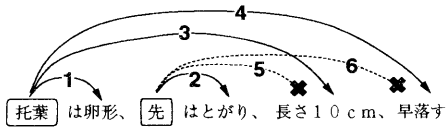


図 7: 図鑑のテキストにおける係り受けの例

性情報を表す名詞を自動分類した。その実験結果は人手で分類した結果と 82% まで一致した [渡辺 95A]。自動分類した結果は人手で修正し、それぞれの意味のまとまりごとにそれらが表す意味のラベルをつけた。

4 用語間の関係情報を用いた係り受け解析

本章では最初に図鑑のテキストにおける主語と述語の係り受けのあいまいさについて説明する。次に、そのあいまいさを植物用語間の関係情報を用いて解消する方法を提案する。そして、3章で獲得した植物用語間の関係情報を用いて図鑑のテキストの係り受け解析を行ない、その結果から提案した方法の有効性と 3章で獲得した専門用語間の関係情報の妥当性を示す。

4.1 図鑑のテキストにおける主語と述語の係り受け関係

図 7は例文 2の主語と述語の係り受け関係を示している。そこでは「托葉」は「先」をとびこえて「長さ 10cm」「早落する」にかかっている。このように植物図鑑のテキストでは、直前にあらわれた主語ではなく、それよりも前にあらわれた主語が述語にかかる場合がある。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・部分概念にあたる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。例えば図 7では「先」は「托葉」の部分概念で、「先はとがり」は「托葉」の説明を補足している。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 7の 3および 4のような係り受けがおこる。図鑑のテキストではこうした階層的な説明がたびたび行なわれているので、主題とそれを説明する述語の係り受けに図 7の 3〜6のようなあいまいさが生じる。こうした図鑑のテキストを正しく理解するには、このあいまいさを解消する方法を明らかにしなくてはならない。

4.2 図鑑のテキストにおける係り受けのあいまいさの解消

4.1節で述べた係り受けのあいまいさを用語間の (1) 上位 – 下位関係、(2) 全体 – 部分関係、(3) 属性の情報を用いて解消することを考えた。それぞれの情報をどのように用いるかは以下で説明する。

上位 – 下位関係の情報 上位 – 下位関係の情報は以下の 2つの推定に用いる。

1. 全体 – 部分関係の概念階層における位置の推定

   上位 – 下位関係にある概念は全体 – 部分関係の概念階層では同じ階層に位置する。したがって全体 – 部分関係の概念階層における位置が不明な用語は上位 – 下位関係を用いてその位置を推定する。

2. 属性情報の推定

   上位 – 下位関係にある概念は属性情報を継承する。したがって属性情報が未知、あるいは不完全な用語は上位 – 下位関係を用いてその属性情報を推定する。

全体 – 部分関係の情報 図鑑のテキストを解析する際、述語にかかる可能性のある主語を推定するのに用いる。係り受け関係は交差しないので、主語が省略されている述語にかかる可能性がある主語は

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、その全体概念語である主語

である。そこで用語間の全体 – 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推定する。例えば図 7の場合、「托葉」が「先」の全体概念語であるので、「長さ 10cm」にかかる可能性がある主語は「托葉」と「先」であると推定できる。

属性情報 属性情報は以下の 2つの処理に用いる。

1. 主語と述語の係り受けの妥当性の判定

   属する全体 – 部分関係の概念階層が異なる対象はその性質や状態が異なる。したがって、それらを説明する述語もまた異なる。例えば「先」という概念は「長さ 10cm」で説明される属性をもたないが、「托葉」はもつ。また「先」は「早落する」という用言で説明される状態をとらないが、「托葉」はとる。このため「長さ 10cm」「早落する」の主語として「先」は妥当ではないが、「托葉」は妥当であると判定できる。このように、主語になる概念がとる述語の情報があれば、係り受けの妥当性が判定できる。この述語の情報に 3章で獲得した属性情報を利用する。

   係り受けの妥当性の判定は、文中での位置が述語に近い主語から順に用例にもとづく方法で行なう。述語が名詞の場合は、その述語と説明対象がとりうる述語の例と比較し、その類似度から係り受けの妥当性を判定する。類似度の評価には [渡辺 94] で提案した文字列の一致による専門用語の類似度の評価を用いる。述語が用言の場合は、その述語が説明対象がとる用言の例そのものか、その類義語である場合に妥当であると判定する。用言の類義語情報には [渡辺 95B] で獲得したものを用いる。

2. 名詞述語文の意味解析

   名詞述語文の述語が表す意味は、その主語の属性を表す名詞の例 (属性情報) の中から最も意味が近いと判定した名詞が表す意味と同じであると考えた。属性を表す名

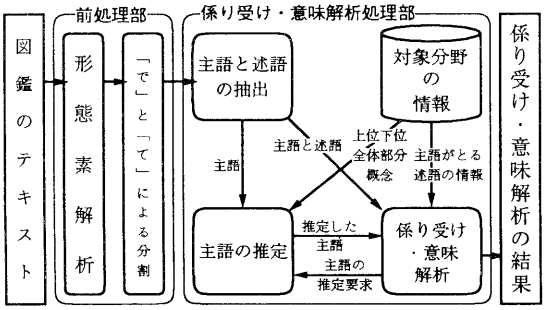


図 8: システムの概要

詞の例はそれが表す意味にしたがって分類されている。このため、この意味情報と名詞間の意味の近さの情報によって名詞述語文の意味解析が実現できる。名詞間の意味の近さは [渡辺 94] の文字列の一致による専門用語の類似度を用いて評価する。

図とテキストから獲得した属性情報が係り受け解析および名詞述語文の意味解析に有効であることはすでに [渡辺 94][渡辺 95B] で実験を行ない確かめた。

4.3 図鑑のテキストの係り受け解析システム

本節では用語間の (1) 上位 – 下位関係、(2) 全体 – 部分関係、(3) 属性の情報を用いて図鑑のテキストの係り受け解析を行なうシステムについて説明する。作成したシステムの概要を図 8に示す。このシステムに

- 人手で作成した用語間の上位 – 下位・全体 – 部分関係の情報
- テキストから自動抽出した属性情報

を与えて図鑑のテキストを係り受け解析すると、係り受けにあいまいさのある文を 77% の精度で正しく解析できた [渡辺 95B]。システムは前処理を行なう部分と係り受け・意味解析を行う部分の 2 つから構成されている。

4.3.1 前処理部

前処理は形態素解析と、「で」および「て」による並列の分割の 2 つの処理から構成されている。「で」および「て」による並列の分割は、述語が表す属性・状態 (変化) の情報を 1 つにするための処理である。

- A で / て B は C ⇒ B は C
  卵状長楕円形で先はとがる ⇒ 卵状長楕円形, 先はとがる
  小さくて長さはふつう 3–6cm ⇒ 小さく, 長さはふつう 3–6cm
- A は B で / て C ⇒ A は B, A は C
  表面は緑色で光沢がある ⇒ 表面は緑色, 表面は光沢がある
  葉は大きくて厚い ⇒ 葉は大きい, 葉は厚い
- A は B で / て C は D ⇒ A は B, C は D
  表面は無毛で葉脈はくぼむ ⇒ 表面は無毛, 葉脈はくぼむ
  先はとがって先端はへこむ ⇒ 先はとがる, 先端はへこむ

表 1: 用言を中心にした文と名詞述語文の内訳

文の種類	出現総数	主語がない文	主語があいまいな文
用言を中心にした文	302	132	61
名詞述語文	385	193	62

表 2: 係り受け解析の結果

文の種類	情報の獲得方法	成功	失敗	総数
用言を中心にした文	人手で作成	45	16	61
	自動獲得	43	18	
名詞述語文	人手で作成	50	12	62
	自動獲得	47	15	

- A で / て B ⇒ A, B
  楕円形で鈍頭 ⇒ 楕円形, 鈍頭

4.3.2 係り受け解析処理部

係り受け解析の処理は図 8に示すように以下の 4 つのモジュールから構成されている。4 つのモジュールは互いに情報をやりとりし、全体で係り受け解析および意味解析を実現する。

対象分野の情報 植物用語間の上位 – 下位・全体 – 部分関係および主語になる概念が述語としてとる名詞および用言についての情報 (属性情報) を格納している。概念間の上位 – 下位・全体 – 部分関係の情報は主語の推定に、属性情報は係り受けの妥当性の判定と意味解析に用いる。

主語と述語の抽出モジュール 前処理の結果を句読点で分割し、提題の助詞「は」を手がかりに主語を、文末の自立語を述語として取り出す。

主語の推定モジュール 処理の対象の述語にかかる可能性のある主語を推定する。述語にかかる主語の推定には概念間の関係情報 (上位 – 下位・全体 – 部分) と非交差条件を用いる。

係り受け・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、用例にもとづく方法で係り受け解析を行なう。述語が名詞の場合 (名詞述語文の場合) は意味解析も行なう。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、推定された主語で係り受け解析および意味解析を行なう。

4.4 実験と検討

4.4.1 実験と結果

図鑑の図とテキストから自動的に獲得した用語間の関係情報を用いて図鑑のテキストの係り受け解析と意味解析を行なう実験を行なった。この実験では、図鑑の図とテキストから自動的に獲得した用語間の上位 – 下位・全体 – 部分関係の情報の妥当性を評価するために、人手で作成した用語間の上位 – 下位・全体 – 部分関係の情報を 4.3節のシステムに与え、

表 3: 係り受け解析の失敗の原因

用言を中心にした文の係り受け解析の失敗の原因	人手	自動
「は」格だけの格情報では不十分. 他の格情報が必要	5	5
主語がとる用言の例が十分ではなかった	4	4
主語が指示詞で、指示対象がわからない	4	4
主語が上位－下位・全体－部分情報に含まれていない用語	2	4
主語が「植物の一部＋空間語」(例:「葉脈上は」)	1	1
合計	16	18

名詞述語文の係り受け解析の失敗の原因	人手	自動
妥当でない主語が述語に似た属性値の例をもっていた	5	5
先行する係り受け解析が誤り、正しい係り受けが禁止された	2	2
主語が指示詞で、指示対象がわからない	3	3
主語が「植物の一部＋属性名」(例:「葉の形は」)	1	1
主語がとる属性値の例が十分ではなかった	1	1
主語が上位－下位・全体－部分情報に含まれていない用語	0	3
合計	12	15

表 4: 名詞述語文の意味解析の結果

用語間の関係	成功	失敗	合計
人手で作成	355	30	385
図とテキストから自動作成	352	33	385

その結果を比較した。人手で作成した用語間の上位－下位・全体－部分関係の情報は、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作成した。実験は図鑑のテキストから「葉」について記述している 200 文を対象に行なった。これは人手で作成した用語間の関係情報が、葉とその下位・部分概念を表す 29 語に限られているためである。実験対象の 200 文における用言を中心にした文および名詞述語文の出現総数、主語が省略されている文の数、さらに主語の係り受けにあいまいさがある文の数を表 1に示す。主語の係り受けにあいまいさがある文に対する係り受け解析の結果を表 2に示す。表 3には係り受け解析の失敗の内訳を示す。解析対象が名詞述語文の場合は意味解析も係り受け解析と同時に行なう。その結果を表 4に示す。表 5には名詞述語文の意味解析の失敗の内訳を示す。

4.4.2 検討

4.4.1節のテキストの係り受け解析および名詞述語文の意味解析の実験結果から人手で作成したものとほぼ同等の内容をもつ用語間の上位－下位・全体－部分関係の情報を自動的に獲得できたといえる。

係り受け解析の過程を詳しく調べると、用語間の上位－下位・全体－部分関係が不完全なため、述語にかかる可能性がある主語を正しく推定できなかった文が人手で作成した情報を用いた係り受け解析では 5 例、自動的に獲得した情報を用いた解析では 7 例あった。このうち、表 3でも示しているように、人手で作成した情報を用いた係り受け解析では 2 例、

表 5: 意味解析の失敗の理由の内訳

意味解析に失敗した理由	人手	自動
係り受け解析が失敗した	12	15
主語が上位－下位・全体－部分関係に含まれていない用語	1	1
主語の一部が省略されている (例:「多く (の托葉) は」)	3	3
主語が代名詞であり、指示対象がわからない	3	3
主語が「植物体の一部 ＋ 空間語」(例:「葉脈上は」)	1	1
主語が「植物体の一部 ＋ 属性名」(例:「葉の形は」)	2	2
主語がとる属性値の例が十分ではなかった	8	8
合計	30	33

表 6: 用語間の関係情報が獲得できなかった原因

用語間の関係情報が獲得できなかった原因	件数
説明の順序の例がテキストで 1 回しかあらわれなかった	4
説明の順序が全体－部分関係にしたがっていることを 4 つの判定規則では判定できなかった	3

自動獲得した情報を用いた解析では 7 例すべてが解析に失敗した。この 7 例が自動獲得できなかった原因を表 6に示す。人手で作成した用語間の上位－下位・全体－部分関係の情報は生物学辞典や生物図説を参考にして精密に作成したが、それでもこのように不完全であった。これは、植物の葉という狭い範囲に限ってもなお、用語間の精密な関係情報を人手で作成することが難しいことを示している。

5 おわりに

今後は獲得した用語間の関係情報を用いて図鑑のテキストから知識ベースを構築する。さらにその知識ベースを利用し、図鑑の写真を対象にした画像検索システムを作成する予定である。

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の影山貴子氏、久松ユリ氏ならびに丹下晴美氏に感謝いたします。

参考文献

[国研 64] 国立国語研究所: 分類語彙表, 秀英出版, (1964).

[松本 92] 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).

[鶴丸 84] 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).

[渡辺 94] 渡辺, 長尾: 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, 情報処理学会研究報告, 94–FI–34 (1994).

[渡辺 95A] 渡辺, 長尾: パターン情報と自然言語情報の統合による植物図鑑の図の理解, 電子情報通信学会技術研究報告, NLC95–2, (1995).

[渡辺 95B] 渡辺, 長尾: 図鑑の解説文から内容抽出を行なうための用言の類義関係の獲得, 言語処理学会第 1 回年次大会, C5–3, (1995).

[山田 90] 山田 他: 自然言語における空間描写の解析と情景の再構成, 情報処理学会論文誌, Vol.31, No.5, (1990).