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

渡辺 靖彦　長尾 真
京都大学 工学部 電気工学第二教室

1 はじめに

写真や絵画などの内容情報を画像データから直接抽出するのは困難である。そこで我々はそれらの情報を画像の内容を解説するテキストから抽出することを目指している。そのため、画像の内容がどのような概念に対応し、言語によってどのように表現されているのか、あるいは画像内容を解説するテキストから画像の内容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 [渡辺 93][渡辺 94]。

[渡辺 94] では画像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んだ。植物図鑑では、属性情報の多くが名詞述語文という単純なパターンで記述されている。そこで名詞述語文の述語の名詞を文字列の類似性にもとづいて自動分類し、図鑑の中で写真や図によって表されている植物の属性情報を獲得する方法を提案した。さらに獲得した情報が意味解析および係り受け解析に有効であることを明らかにした。

植物図鑑の写真あるいは図では、植物の形態・生態上の属性情報だけでなく状態の情報も表されている。植物の状態(変化)の情報はテキストでは用言を中心にした文によって表現されることが多い。本稿では最初に、植物図鑑における用言を中心にした文および名詞述語文での係り受けのあいまいさを用例にもとづいて解消する方法を提案する。次に、用例にもとづく係り受け解析に必要な用言間の類義関係を獲得する方法を説明する。そして獲得した類義関係を用いて係り受け解析を行ない、提案した係り受け解析の方法と用言の類義関係の獲得が有効であることを示す。

2 図鑑のテキストの特徴

2.1 図鑑のテキストに記述されている情報

植物図鑑のテキストは図1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まとめられていて、それぞれの植物について以下の情報が記述されている。

形態情報 植物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植物の属性および状態(変化)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。

2.2 植物の形態・生態情報の表現パターン

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。

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。植物の形態・生態上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。

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 1: 植物図鑑のテキストの例

(例文1) 果実はほぼ球形だ (属性 [形状])

(例文2) 花序は長さ１．５－３ｃｍ (属性 [長さ])

用言を中心にした文 文の中心になる述語は動詞、形容詞、あるいは動詞化したサ変名詞である。植物の形態・生態上の状態(変化)の情報を主に表すが、属性情報も表す。

(例文3) 葉には鋸歯がある (状態)

(例文4) 花序は軸が袋状になる (状態変化)

(例文5) 花序は頂生する (状態, 属性 [花序の種類])

名詞述語文は [渡辺 94] で取り扱ったので、本稿では植物の状態(変化)を表す用言を中心にした文を扱う。

2.3 植物の状態(変化)を表す述語の並列

植物は多様な属性をもち、さまざまな状態をとる。そのため植物図鑑のテキストではそれらの情報を表す述語が並列することが多い。植物の状態(変化)を表す用言が関わる述語の並列には次の4つの表現がある。

「で」による並列 属性を表す名詞の述語と状態(変化)を表す用言の述語が並ぶ。

(例文6) 表面は緑色で光沢がある

「て」による並列 述語の単純な並列と、前の用言による後ろの用言への修飾がある。

(例文7) 葉は大きくて厚い

(例文8) 葉は枝の上方に集まってつく

「または」による並列 類義関係にある述語の並列と類義関係にある文の並列がある。「または」以外にこれらの文または述語を結ぶものに「あるいは」「か」「ないし」「また」「まれに」「や」がある。

(例文9) 葉は互生または対生する

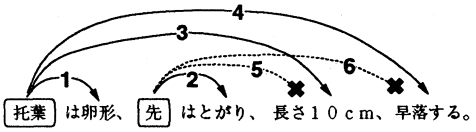

図 2: 1つの主語が複数の述語にかかる例

(例文 10) 縁には鋸歯があるか浅裂する

例文 10は類義関係にある文の並列の例で、そこで用いられている述語「ある」と「浅裂する」には類義関係はない。

読点「、」による並列 一つの主語を共有する述語が並ぶ。

(例文 11) 托葉は卵形、長さ 10cm、早落する

最初の 3 つの並列は単文中の述語内の並列であるが、「、」による並列は単文の並列である。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、図 2の「托葉」のように 1 つの主語が並列する複数の文の述語にかかる。

2.4 図鑑のテキストにおける主語と述語の係り受け関係

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 (図 2の 3 および 4 の係り受け)。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。例えば図 2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。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 2の 3 および 4 のような係り受けがおこる。

3 図鑑のテキストにおける係り受けのあいまいさの解消

図 2のように 1 文中に複数の主語があると、主語と述語の係り受けにあいまいさが生じる。本研究では以下に述べる方法でこうした主語と述語の係り受けのあいまいさを解消する。最初に、解析対象の述語にかかる可能性のある主語を取り出す。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、その上位・全体概念である主語

で、これらを概念間の上位 – 下位・全体 – 部分関係の情報を用いて取り出す。例えば図 2の「長さ 10cm」、図 3の「縁に鋸歯をつける」にかかる可能性がある主語は「先」とその上位・全体概念の「托葉」である。

次に、取り出した主語がとる述語の情報を用いてその係り受けが妥当かどうかを判定し、係り受け関係を決定する。取り出した主語は概念階層が異なるので、それぞれがもつ属性およびとりうる状態 (変化) は異なり、それらを説明する述語も異なる。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」

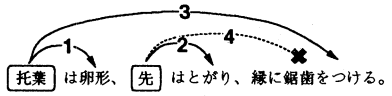

図 3: 用言の種類の情報によって係り受けの曖昧さを解消した例

は「つける」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」「縁に鋸歯をつける」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。したがって、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。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行なう。述語が名詞の場合は [渡辺 94] の係り受け判定方法を用いる。すなわち、主語がとる属性値の例を述語と比較し、その類似度から係り受けの妥当性を判定する。述語が用言の場合は、それぞれの主語がとる用言の例と用言間の類義関係を参照し、その主語が述語として妥当かどうか判定する。

以上で述べた係り受け解析を実現するために、本研究では以下の 3 つの情報を用いる。

1. 主語になる概念間の上位 – 下位・全体 – 部分関係情報
2. 主語になる概念が述語としてとる名詞の例の情報
3. 主語になる概念の説明としてとる用言の例の情報

1 と 2 の情報には [渡辺 94] で作成したものを用いる。3 の情報は、それぞれの主語がとる用言の例と用言間の類義関係から構成される。それらの情報の獲得方法は 4章および 5章で述べる。

4 主語になる概念が述語としてとる用言の例の蓄積

図鑑のテキストから用言を中心にした文を抽出し、主語になる概念とそれがとる用言の例を獲得する。

1. 形態素解析 [松本 92] を行ない、形式動詞「する」の前にある数字あるいは漢字連続をサ変名詞の候補として取り出す。これは、名詞がサ変名詞かどうかを判定するのに利用する。
2. 形態素解析結果を句読点で分割し、(a) 提題の助詞「は」を含み (ただし「には」「では」は除く)、(b) 最後にあらわれた自立語が用言である文を抽出する。したがって図 2、3の例文から取り出されるのは、「先はとがり」の部分だけである。
3. 「で」および「て」による並列を以下のように分割し、サ変名詞以外の名詞を述語にとる文を取り除く。

   a. A で / て B は C ⇒ B は C
   卵状長楕円形で先はとがる ⇒ 先はとがる
   小さくて長さはふつう 3–6cm ⇒ 長さはふつう 3–6cm

   b. 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

表 1: 抽出できた用言とできなかった用言の例

品詞	抽出できた例	抽出できなかった例
動詞	くぼむ つく とがる もつ 枯れる 伸びる 生える 帯びる 落ちる 裂ける	からみつく ねじれる 囲む 乾く 切れ込む 波打つ 反る 敷く 包む 乱れる
形容詞	かたい 鋭い 円い 厚い 細長い 多い 大きい 幅広い	悪い 深い 痛い 等しい
サ変名詞	2 裂 凹入 下垂 互生 斜上 宿存 尖裂 早落 発達 分岐 落葉 隆起 湾曲	3 浅裂 羽状中裂 急尖 欠刻 紅葉 黒変 硬生 退化 肥厚 萌芽 匍匐

c. 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

4. 提題の助詞「は」を手がかりに主語を取り出し、その文の最後の自立語を述語として取り出す。ただし、最後の自立語が「する」の場合は、その前のサ変名詞を取り出す。

以上の方法によって葉およびその下位概念を表す 29 個の名詞が述語としてとる 138 個の用言 (動詞 61 個、形容詞 23 個、サ変名詞 54 個) を抽出した。

5 用言間の類義関係の獲得

図鑑のテキストでは 104 個の動詞、27 個の形容詞、116 個の動詞化したサ変名詞が葉およびその下位概念の属性および状態 (変化) を表す述語として用いられている。このため、4章で抽出した用言だけでは係り受け解析を行なうのに不十分である。表 1に抽出できた用言とできなかった用言の例を示す。そこで本研究では (1) 分類語彙表の分類、(2) 国語辞典の語義文、(3) 文字列間の類似度を用いて用言間の類義関係を獲得し、用言の用例の不足を補う。分類語彙表には 7 レベルの階層的な意味分類があるが、その最小の意味分類は段落とよばれる。我々は分類語彙表で同じ段落に属する用言は類義関係にあると判定することにした。しかし、サ変名詞とそれ以外の用言の類義関係を分類語彙表を用いて評価するには以下の問題がある。

1. 図鑑で用いられているサ変名詞は専門用語であることが多く、分類語彙表のような一般的なシソーラスでは分類されていない。
2. 分類語彙表では用語をまず品詞で分け、それから意味分類を行なっている。分類語彙表で分類されているサ変名詞でも、品詞が異なる他の用言との意味の近さを評価することは難しい。

そこで以下の方法で類義関係を獲得する。

1. 動詞化したものが分類語彙表で分類されている 14 個のサ変名詞は、動詞化したものが分類されている段落に属すると判定する。
2. 1 以外のサ変名詞は国語辞典の語義文から類義関係にある動詞を取り出すことを試みる。国語辞典の語義文では

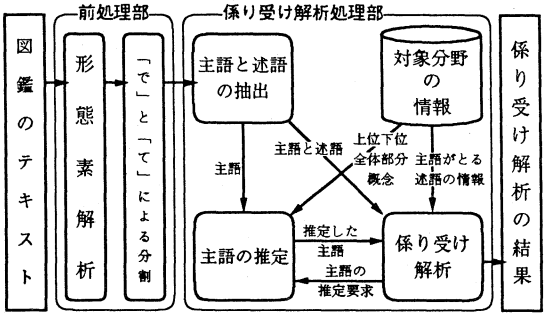

図 4: システムの概要

類義関係にある動詞を名詞化してサ変名詞の意味を説明することが多い。そこでサ変名詞の語義文を形態素解析し、以下の表層の表現パターンと品詞情報を手がかりに類義関係にある動詞を取り出す。

a. 「(動詞) こと」(「〜になること」「〜のあること」はのぞく)
黒変: 色が黒く かわること。 ⇒ かわる
b. 「(動詞) いること」
残存: 残っていること。 ⇒ 残る

取り出した動詞が分類されている段落にサ変名詞は属すると判定する。この処理で 54 個のサ変名詞の属する段落を判定した。

3. 1、2 の処理でどの段落に属するか判定できなかったサ変名詞は、1 または 2 の処理でどの段落に属するのか明らかになったサ変名詞の中から最も意味が近いものを文字列の類似度を用いて判定し、それが属する段落に属するものとする。分類語彙表および国語辞典を用いても類義関係が得られないサ変名詞は専門用語であると考え、意味が近い専門用語は字面が似るというヒューリスティックスをこの処理では利用した。文字列の類似度は [渡辺 94] の定義を用いる。この処理で 42 個のサ変名詞の属する段落を判定した。
4. 1、2、3 の処理でどの段落に属するのか判定できない 6 個のサ変名詞は人手でどの段落に属するか決めた。

以上の方法によって葉およびその下位概念を説明する用言 247 個を、165 個の意味的なまとまりに分類した。

6 図鑑のテキストの係り受け解析システム

本章では概念間の関係情報と主語になる概念がとる述語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 4に示す。システムは前処理を行なう部分と係り受け解析を行う部分の 2 つから構成されている。

6.1 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」

表 2: 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	308	136	61
名詞述語文	405	196	61

表 3: 係り受け解析の結果

文の種類	成功	失敗	総数
用言を中心にした文	45	16	61
名詞述語文	49	12	61

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。この処理では、4章で述べた3種類の並列の分割の他に、主語が省略された文の述語も以下のように分割する。

- A で / て B ⇒ A, B
  楕円形で鈍頭 ⇒ 楕円形, 鈍頭

6.2 係り受け解析処理部

係り受け解析の処理は図4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を実現する。

対象分野の情報 植物に関する概念間の関係と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を格納している。概念間の関係情報は主語の推定に、述語の情報は係り受けの妥当性の判定に用いる。

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。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を行なう。

7 実験と検討

植物図鑑のテキストから「葉」について記述している200文を無作為に抽出し、作成したシステムで係り受け解析を行なった。

7.1 実験

実験対象の200文中での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2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3に示す。さらに係り受け解析の失敗の内訳を表4に示す。

7.2 検討

作成したシステムで解析を行なった結果、主語の係り受けにあいまいさがある文の77%の解析に成功した。主語がとる述語の例と用言間の類義関係の大部分を自動的に獲得したことを考えると、これはかなりよい結果である。

表 4: 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	文数
「は」格だけの格情報では不十分. 他の格情報が必要	5
主語がとる用言の例が十分ではなかった	4
主語が指示詞で、指示対象がわからない	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
主語が「植物の一部 + 空間語」(例:「葉脈上は」)	1
合計	16

名詞述語文の係り受け解析の失敗の原因	文数
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
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
主語が指示詞で、指示対象がわからない	3
主語が「植物の一部 + 属性名」(例:「葉の形は」)	1
主語がとる属性値の例が十分ではなかった	1
合計	12

係り受け解析で用言間の類義関係の情報を参照しなければならない文は15例あった。そのうち14例は類義関係の情報によって正しく判定でき、判定を誤ったのは1例だけだった。この結果から獲得した用言間の類義関係は妥当なものであるといえる。

「は」格の格情報だけでは係り受けを正しく判定できなかった5例のうち、3例の用言が「ある」であった。「ある」は他の格の情報も係り受けの判定に用いる必要がある。

8 おわりに

今後は用言を中心にした文が表す情報を詳細に調査・分類し、それらを抽出する方法を検討する予定である。

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ならびに丹下晴美氏に感謝いたします。

参考文献

[渡辺 93] 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).

[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).

[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).

[国研 64] 国立国語研究所: 分類語彙表, 秀英出版, (1964).