情報処理分野専門用語概念体系の開発

1M－4

崔進　横田英司+　安原宏
（株）日本電子化辞書研究所

1 はじめに

日本電子化辞書研究所（以下、EDRと略す）の概念体系は、概念間の上位下位関係を与えるものであり、概念間の類似性を判断する基礎情報になるものである[1]。

本稿では、EDR概念体系の開発方針に基づき、EDR情報処理分野専門用語の語義を対象に、半自動的に体系化する方法について述べる。体系化は類似概念のグループ化とそのグループ毎への上位概念の設定という二つの作業からなる。

専門用語の多くは、複合語であり、かつ、単語義である。日本語の複合語の多くは、右端の構成語が、代表概念を担っている。始めからすべての専門用語を相手にするのは困難であるので、基本となる用語（項目語）をもとにした体系作成法がある[2]。これと同様に、われわれは、日本語専門用語のほとんどに対して項目語が右端にくることに着目し、以下の手順で体系化作業を行なった。

1) 構成語の逆ソートにより下位部分木を作成する
2) 非複合語専門用語および複合語専門用語の右端構成語から上位体系構築用代表概念を選択する
3) 代表概念に基づき上位体系を構築する
4) 個々の部分木と上位体系を結合し、体系化する

以下、2節は下位部分木の作成、3節は代表概念の選択、4節は代表概念に基づいた上位体系の構築、5節はEDR基本語概念体系と専門用語概念体系の統合について説明する。

2 下位部分木の作成

EDR日本語専門用語の一部（46,820語）に対し、まず、構成語情報の逆ソートをすることによって、4,706個の単語集合を作成した。次に、単語を概念に置き換え、人間のチェックも加え、下位部分木を作成した。これらの下位部分木は概念体系の最下位部分になる。図2.1は一つの部分木の例を示す。

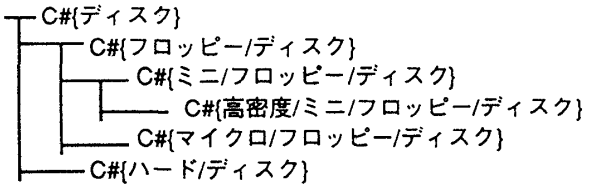

図2.1 下位部分木の例

3 代表概念の選択

すべての下位部分木をまとめ、一つの体系にするために、体系の上位部分を構築する必要がある。体系の上位部分を構築するときに使う代表概念は、非複合語専門用語、および、個々の下位部分木の最上位項目である単語の概念から選出する。代表概念を選出する際に、以下の2つに注意した。

3.1 同義語の削除

部分木の中には、「電子計算機」と「コンピュータ」のような同じ概念の持つ単語対がある場合、片方の単語概念だけを代表概念として選出する。

3.2 多義性単語の対応

例えば、単語「キー」は<キーボードのキー>と<識別子のキー>の二つの概念を持つ。このような多義性単語に対して、次のように代表概念を作成する。

C#{ キー$_1$ } = キーボードのキー
C#{ キー$_2$ } = 識別子のキー

4 上位体系の構築

本節では、3節で作成した5,250の代表概念をもとに上位体系の構築方法について述べる。これは基本的に人間の内省による作業である。

本研究における専門用語概念体系はすべての概念をモノ、コト、属性に大分類する。3つの項目にある概念数を表4.1に示す。以下、モノ体系とコト体系の構築法について述べる。

大項目	モノ	コト	属性
概念数	2,913	1,322	1,015

表4.1 大項目の概念数

Developing a Classification Dictionary of Technical Concept. Jin Cui, Eiji Yokota+, Hiroshi Yasuhara
Japan Electronic Dictionary Research Institute, Ltd.
+現　三菱電機東部コンピュータシステム（株）