情報学基礎 7－4
（1987・11・9）

電子化辞書の構成について

内田裕士

（株）日本電子化辞書研究所

電子化辞書は従来の辞書の機械可読版ではなく、コンピュータが自然言語を理解できるようにするために全く新しく設計されたものである。

電子化辞書には、単語が表わしている概念と、その単語がそういった概念を表わすときの文法的な特徴（特性）、その概念をコンピュータが理解するために必要な知識が格納される。

本報告では、電子化辞書の構成、電子化辞書の開発法および電子化辞書の利用法について述べる。

The Composition of Electronic Dictionaries

Hiroshi Uchida

Japan Electronic Research Iustitute Ltd.

MITA KOKUSAI BUILDING ANNEX 1F 4-28, MITA 1-CHOME, MINATO-KU, TOKYO 108

An electronic dictionary is not a machine readable version of conventional dictionaries. It is a completely new dictionary designed from scratch. It is a dictionary developed for a computer to understand a natural language.

In an electronic dictionary, the concept expressed by a word, grammaticalcharacteristics of the word when it expresses the concept, and the knowledge necessary for understanding the concept will be written in an from understandable by the computer.

In this paper, composition of electronic dictionaries, development method and the use of dictionaries are described.

1．はじめに

電子化辞書の研究開発は、自然言語処理技術開発の基盤となるものであり、２０世紀人類の文化遺産として次世代に継承されるべきものである。この辞書開発には計算機科学、言語学、辞書学等の幅広い分野の英知を結集する必要がある。しかしながら、これらの分野の専門家の人数的制約や、辞書の規模、辞書に要求される質等を考えると単独の企業ベースでの開発はほとんど不可能である。

このような状況のもとで、（株）日本電子化辞書研究所は、基盤技術研究促進センターおよび民間企業８社（富士通（株）、日本電気（株）、（株）日立製作所、シャープ（株）、（株）東芝、沖電気工業（株）、三菱電機（株）、松下電器産業（株））によって１９８６年４月に設立された。（株）日本電子化辞書研究所（ＥＤＲ）の目的は、コンピュータが自然言語を理解するため辞書を研究開発するだけでなく、電子化辞書の標準化を図ることもある。

2．電子化辞書

電子化辞書は従来の辞書の機械可読版でなく、コンピュータが自然言語を理解できるようにするために全く新しく設計されたものである。

人間にしろ、コンピュータにしろ、自然言語で書かれた文章を理解するためには、言語の意味を知っていなければならないし、またその言葉が使われた文脈の中での意味を把握できなければならない。電子化辞書には、単語が表している概念と、その単語がそういった概念を表すときの文法的な特徴（特性）、その概念をコンピュータが理解するために必要な知識が格納されることになる。一方コンピュータ側には、単語の使われ方を理解するのに必要な情報が文法規則の形で格納されることになる。

3．電子化辞書の構成

電子化辞書は基本的には単語辞書と概念辞書から構成される。単語辞書は単語そのものに関する情報が格納されており、概念辞書は概念に関する情報が格納されている。

単語辞書には次のような情報が入っている。

① 概念の表層表現としての単語の見出し

② その見出しで表される概念

③ 単語がある概念を表すときの文法的特性

概念の定義は文章でなされる。これは、ある概念を人間が他の概念から識別できるようにするためである。この定義文は概念辞書の見出しとして使用される。複合語の場合は、その複合語で表される概念が構成語の持っている概念を使用して表さ

れる。

概念辞書は２つの概念の間に成立し得る関係を定義したものである。複合概念は、より基本的な概念を用いて、概念間の二項関係の集合として定義される。概念間の関係としては格関係、因果関係、同義関係、類義関係、上位－下位関係などがある。図１に単語辞書と概念辞書の関係を示す。

図１において“eagle”という単語は２つ以上の概念をもっている。一つは“a score of two below on any hole”，であり、もう一つは“a bird called eagle”である。どちらの場合でも文法特性は名詞である。また“fly”という単語も二つ以上の概念を持っている。一つは“to move through the air with wings”である。この場合の文法的特性は自動詞である。もう一つは“an insect called fly”であり、この場合の文法的特性は名詞である。

概念辞書はこれらの概念間に成立する可能な関係を定義するわけであるから“a score of two below on any hole”という概念は“to reach the end of an activity”という概念の程度（degree）を表すことができるし“a bird called eagle”という概念は“to move through the air with wings”という概念の動作主（agent）になれるということが定義される。

単語辞書には２種類あり、基本語辞書（２０万語）と専門用語辞書（１０万語）である。開発対象言語は日本語と英語であり合計８つの辞書が作られることになる。（図２参照）

概念辞書も２種類あり、上位－下位関係だけを定義した概念体系と、それ以外の関係等を定義した概念記述に分けられる。（図２参照）

４．開発方法

大規模な電子化辞書を作成するときの最も大きな問題は情報の一様性と正確さをいかにして保証するかである。

ＥＤＲでは辞書の作成に際して図３のようなワークシートを使用している。このワークシートはある単語が表している１概念毎に記述され、その後コンピュータに入力される。コンピュータ上の辞書エディタを使用することにより、複合語に関する構成語とのリンクや、訳語のリンク付けなどがなされる他、一般の校正作業もこの辞書エディタを使用して行われる。情報の正確さに関しては、大規模なテキストのコンピュータによる分析をもとにして検証することによって保証しようとしている。

概念辞書は大規模テキストベースを分析した結果を使用して作成している。また辞書の正確さと効果を検証するために、機械翻訳システムや音声認識システム、情報検索システム等を作成している。これらのシステムを使用して評価した結果は辞書作成にフィードバックされる。現状と開発予定を図４に示す。

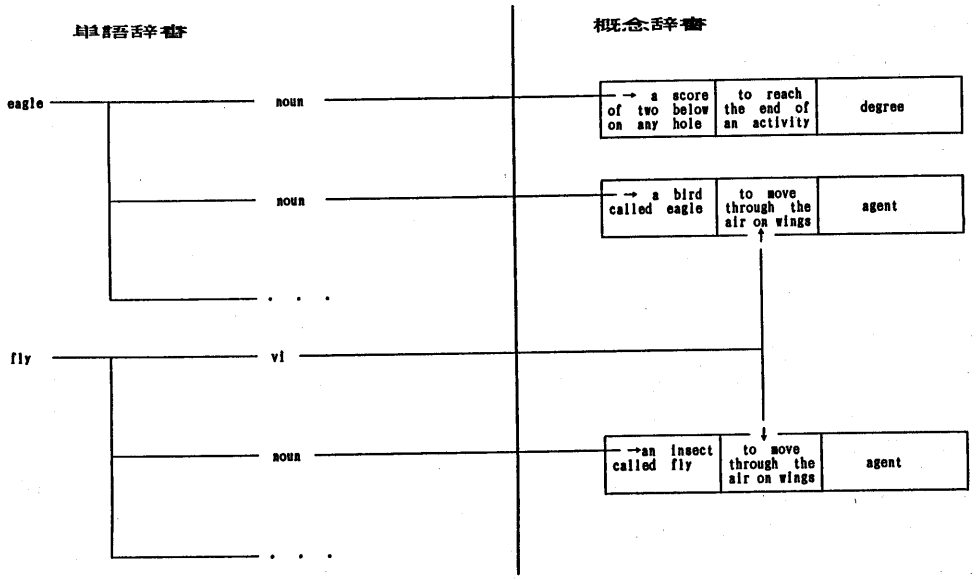

図1　単語辞書と概念辞書の関係

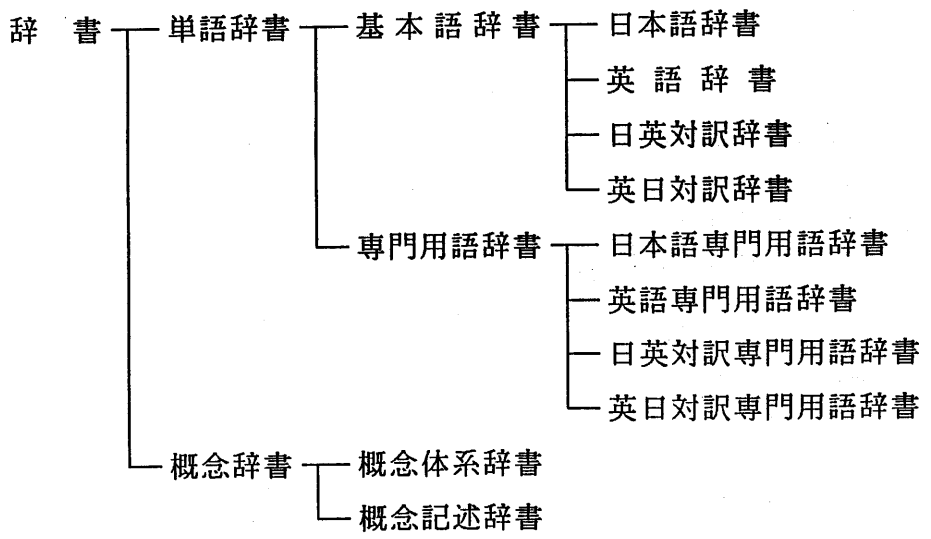

図2　電子化辞書の種類

英語・英日ワークシート（V10L20）
㈱日本電子化辞書研究所

＊記入者名							
＊記入月日							
見出し語情報 No.	（右づめ，多品詞・多意味語は番号のみ記入）	(8)					
単語表記		□ (64)					
異表記		□ (64)					
品詞情報 No.	（右づめ，多意味の単語は番号のみ記入）	(4)					
★異表記		□ (64)					
発音		□ (64)					
☆構成語		□ (64)					
慣用句情報		□ (64)					
品詞・活用		□ (64)					
使用分野	■	□ (30) □ (32)					
用法	■	□ (30)					
意味情報 No.	（右づめ）	(4)					
概念見出し		□ (192)					
★発音		□ (64)					
★異表記		□ (64)					
★構成語		□ (64)					
★慣用句情報		□ (64)					
★活用		□ (64)					
表層情報	■ ■ ■	□ (90) □ (32)					
★使用分野	■	□ (30) □ (32)					
★用法	■	□ (30)					
訳語情報		(1) □ (64)					

＊：記入の形式は任意　　★：上位情報の同項目と内容が同じならば省略可
☆：単語表記が一語ならば省略　　□：項目欄が不足の場合にContinueを記入

図3　ワークシート

5．辞書の利用

ＥＤＲで作成されている電子化辞書はコンピュータが自然言語を理解するために必要な情報を含んでいるので、機械翻訳、質問応答、情報検索等さまざまの応用システムに適応可能であるが、最も中心的な応用システムは、中間言語方式による機械翻訳システムである。

形態素解析や構文解析などの文解析は辞書の文法特性を利用して行われるが、解析過程で起こる曖昧性は概念辞書によって与えられるモデルを利用して解消あるいは減少させられることになる。

サブテーマ ＼ 年度	初年度	62年度	63年度	64年度	65年度	66年度	67年度	68年度	69年度
基本語辞書									
専門用語辞書	試作	試作	試作	評価	改良・拡張	改良・拡張	改良拡張評価		
概念体系		試作	試作	試作	評価	改良・拡張	改良拡張評価		
概念記述	実験・設計	第1次試作	第1次試作	評価	第2次試作	改良・拡張	改良拡張評価		
データ管理システム	実験・設計	第1次試作	評価	第2次試作	第2次試作	第2次試作	評価	改良拡張評価	改良拡張評価
実証評価システム	設計・試作	改良・拡張	設計・試作	試作	改良・拡張	改良・拡張	改良・拡張		
		設計・試作	第1次試作	評価・試作	第2次試作	第2次試作	実証・評価	実証・評価	実証・評価

図4　電子化辞書の現状と開発予定