EDR電子化辞書の検索システム

2B－2

小出 東洋　松山 尚市　竹田 正幸　松尾 文碩
九州大学工学部

1. はじめに

EDR電子化辞書を使って自然言語処理を効率的に行うためには，処理プログラムから高速に辞書データを検索できる機能とEDR電子化辞書の高速ブラウザが必要である．そこで，EDR電子化辞書を著者らが開発したSeepを使ってデータベース化した．

2. EDR電子化辞書の構成

EDR電子化辞書は，単語辞書，対訳辞書，概念辞書，共起辞書，専門用語辞書とEDRコーパスから構成されている．表1にEDR電子化辞書の全体図を示す．

表１ EDR電子化辞書

単語辞書	日本語単語辞書	25万語
	英語単語辞書	19万語
対訳辞書	日英対訳辞書	23万語
	英日対訳辞書	16万語
概念辞書	概念見出し辞書	—
	概念体系辞書	40万概念
	概念記述辞書	40万概念
共起辞書	日本語共起辞書	90万句
	英語共起辞書	46万句
専門用語辞書（情報処理）	日本語専門用語単語辞書	12万語
	英語専門用語単語辞書	8万語
	その他	—
EDRコーパス	日本語コーパス	22万文
	英語コーパス	16万文

単語辞書は，単語と概念(意味)との対応関係を記述する．対訳辞書は，日本語と英語の単語の見出し語間の対応関係をあたえる．

概念辞書は，各概念を言葉で説明する概念見出し辞書と，概念間の上下関係を規定した概念体系辞書，それ以外の関係を規定した概念記述辞書から成り，単語辞書，対訳辞書，共起辞書の各辞書から参照される．

共起辞書は，コーパス中での共起状況の情報に基づいて妥当な単語の組合せ方の実例を示す．

コーパスは，実際の文に対して構成要素の認定を行い，それらの構成要素がどのようにまとまって文を形態的，構文的，意味的に構成するかを示す．

専門用語辞書(情報処理)は，情報処理分野の専門用語を記述している単語辞書，対訳辞書，概念見出し辞書，概念体系辞書，共起辞書から成る．

3. Seepのデータベース管理システム

Seepのデータベース管理システムは，C言語で記述されたデータ操作のための関数群であり，この関数がデータベースの基本操作の集合であるDMP (data manipulation primitives)を実現している．DMPはアプリケーションソフトから関数として呼び出すことができる．表2にSeepで使用できるデータ操作機能を示す．

表２ 基本的なデータ操作機能

操作名	機能
define	データベース定義
show	データベース定義情報表示
open	データベース使用開始
close	データベース使用終了
create	関係・関数の創成
eliminate	関係・関数の消去
insert	タップル追加
delete	タップル削除
update	タップル更新
add	タップル一括追加
retrieve(n)	タップル集合の検索
find	データベースキー集合の検索
eval(n)	データベースキーによる タップル集合の検索
getn	タップルの後続部分獲得
value	関数値を求める

Seepのデータベースは関係･関数の集合であり，デー

Retrieval system for EDR electronic dictionary
Haruhiro Koide, Shouichi Matsuyama, Masayuki Takeda and Fumihiro Matsuo
Faculty of Engineering, Kyushu University 36, Hakozaki, Fukuoka, 812-81 Japan

```
DATABASE(EDR)
        ADMINISTRATOR(SOMEONE)
        MASTER_PASSWORD(******)
                ⋮
DEF(CPC < CPCREC*UP*DOWN)
        L('概念体系辞書検索')
        PATH(/EDRrel/)
DEF(CPH : CONC -> JHC)
        L('日本語概念見出し')
        PATH(/EDRrel/)
                ⋮
DEF(CPCREC) T(L4) L('CPC RECORD ID')
DEF(JHC)    T(C*) L('JP. HEADCONCEPT')
                ⋮
```

図 1 EDR のデータベース定義

タベースを作成するためには，データ定義言語を用いてデータベースの定義をあたえる必要がある．図1にEDR電子化辞書のデータベース定義の一部分を示す．図1において，

```
DEF(CPC < CPCREC*UP*DOWN)
DEF(CPH : CONC -> JHC)
```

はそれぞれ関係

$$CPC \subseteq CPCREC \times UP \times DOWN$$

および関数

$$CPH : CONC \rightarrow JHC$$

を表している．

Seepでは関係の特別な場合として関数を区別する．関係では，すべての属性に対し転置索引(B^+tree)を作成する．一方，関数では定義域の属性についてのみ転置索引を作成する．

作成したデータ定義ファイルをもとに，DMPであるDEFINEを実行し，データベース定義を行う．次に，各関係・関数ごとにDMPであるCREATEを実行して，関係・関数の創成を行う．構築した関係のタップルの検索には，DMPのRETRIEVEを実行する．条件を満たすタップルが複数ある場合は，DMPのRETRIEVENを複数回実行して残りのタップルを順次求める．関数の求値には，DMPのVALUEを実行する．

4. 関係・関数の構成

4.1 属性・型に関する全体的な変更

日本語単語辞書，日英対訳辞書で使われている単語見出しは見出し語とその読みの組であるので，この組を分割していずれか一方だけでも検索可能とした．また，活用語の読みには活用語尾の直前に中点があるのでこれを取り除き，終止形で検索することにした．語幹のみで検索する場合は，Seepのトランケーション処理で代用することにした．

レコード番号は，アルファベット部分を取り除いて4バイトの整数とした．

概念識別子は16進数8桁以下のものが殆どであるため，8桁を超えるもの(概念識別子4種類)については変更を行った．

4.2 各辞書の構成

日本語単語辞書は，レコード番号，単語見出し，カナ見出し，概念識別子から成る関係を作成し，これらの属性で検索できるようにした．その他の項目はそれぞれレコード番号を定義域とする関数にした．日本語と英語による概念見出しと概念説明は，概念説明辞書と重複するので省略した．

英語単語辞書は，レコード番号，単語見出し，概念識別子から成る関係を作成した．他は日本語単語辞書と同様にした．

日英対訳辞書は，レコード番号，単語見出し，カナ見出し，概念識別子を定義域とし，対訳情報を値域とする関数とした．

英日対訳辞書は，レコード番号，単語見出し，概念識別子を定義域とし，対訳情報を値域とする関数とした．

概念見出し辞書は，概念識別子から，英語概念見出し，日本語概念見出し，英語概念説明，日本語概念説明への関数とした．概念体系辞書は，上位概念識別子と下位概念識別子から成る関係とした．概念記述辞書は，概念識別子1，関係子，概念識別子2，記述区分，真偽値から成る関係とした．

5. まとめ

EDR電子化辞書をデータベース化し，自然言語処理プログラムから高速に検索できるようにした．現在，これをもとにブラウザの設計を行っている．

また，概念体系辞書を使用するとき，上位概念識別子や下位概念識別子の閉包を効率的に求めるデータ構造の研究も必要である．

参考文献

1) 日本電子化辞書研究所: EDR電子化辞書利用マニュアル，第2版，1995.