汎電子化辞書：言語知識のアーキテクチャ

横井 俊夫 安原 宏 村木 一至 原田 千秋 丸山 冬樹
日本電子化辞書研究所

1 はじめに

今後の言語処理の研究開発のために言語にかかわるデータや知識を統合的に集積したものを汎電子化辞書と呼ぶ。いわゆる辞書ばかりではなく、通常、シソーラス、コーパス、テキストデータと呼ばれるものも対象にしたものである。EDR 電子化辞書 [1] の位置付けを明確にし、この分野の今後の方向性を明らかにするためにこの汎電子化辞書なるものを提案する。EDR 電子化辞書は統合的な言語データの観点から電子化辞書を実現しようとする試みであるが、汎電子化辞書の立場からすれば第一歩を試みたにすぎない。しかしながら、EDR 電子化辞書プロジェクトにおいては、成果としてまとめられるもの以外にも多くの知見の蓄積がなされており、これらが汎電子化辞書を考察するに際しての大きな拠り所となる。また、同種の試みが国内外の各所で行われている [2]。欧州における代表事例としては、辞書側からの試みとして、Acquilex, Genelex, Multilex, コーパスにおける試みとして British National Corpus がある。米国における代表事例としては、辞書側からの試みとして Comlex, シソーラス側からの試みとして WordNet, コーパスにおける試みとして Penn Tree Bank がある。日本における代表事例としては辞書側からの試みとして、ALT 辞書、IPAL 辞書、JICST 対訳辞書等がある。さらに、中国、韓国、タイなどアジア諸国においても試みが始められている。

2 汎電子化辞書の要件

汎電子化辞書が満たすべき要件を列挙すると以下のようになる。

(1) 有用性

自然言語処理の今後の技術動向に沿うこと。

形態素・構文処理に対してはロバストネスの達成、意味処理・文脈処理への本格的な取組み、コーパスベースなど大量言語データ解析に基づく言語処理、文書処理としてみる言語処理などという技術動向にそうこと。

(2) 柔軟性・汎用性

柔軟な構造を持ち、状況に適切に対応できること。

言語処理全般を視野に入れること。言語理論や処理方式に対し極力中立であること。十分に予測できない将来の技術変化に極力対応可能であること。検証可能であること。応用に対して高い応用性を持つこと。多言語に対する共通性を目指すこと。使用目的に合わせ自由に加工できること。大規模知識ベースの代表事例となること。

(3) 実現可能性

大規模で高精度のものを低コストで実現できること。

コンピュータ処理にとって必要な精度向上が達成できること（参照頻度とのかね合い、処理精度とのかね合い等）。基本的な素材の蓄積がなされていること。構造がモジュール化されており漸進的な開発が可能であること。ある程度の規模や精度が達成されたところで十分な有用性が生ずること。低熟練作業者も活用できること。コンピュータによる強力な開発支援機能が実現できること。規模の拡大、精度の向上にともなって開発支援機能の能力が増大していくこと。

3　汎電子化辞書の基本構造

記述の単位、記述のレベル、言語の種類の３軸によって特徴付けられるサブ辞書が互いに関係し合うという情報構造を形作る。情報構造の概観を図１に、EDR 電子化辞書がどう対応するかを図２と図３に示す。

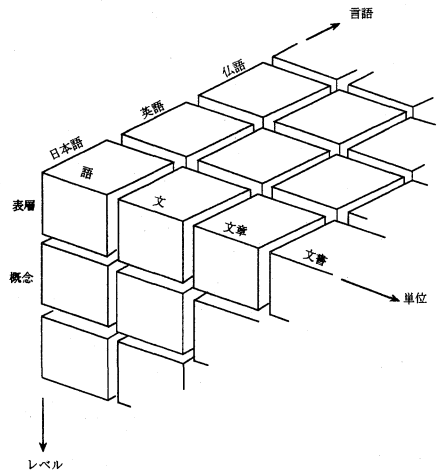

図１．電子化辞書の基本構造

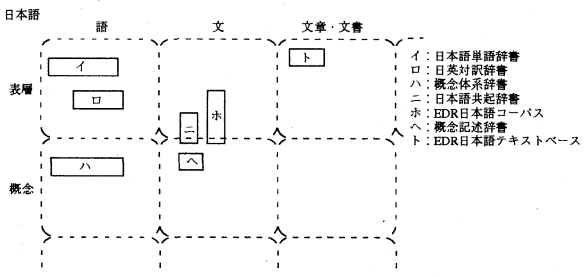

図２．EDR電子化辞書（日本語）の位置付け

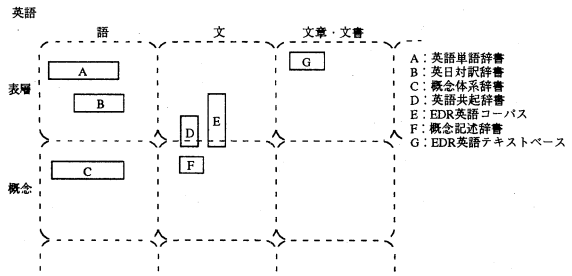

図３．EDR電子化辞書（英語）の位置付け

（1）情報内容の特性

サブ辞書がどのような内容を記述しているのか、サブ辞書の情報内容が次の３点によって特徴付けられる。

(a) 記述単位：語、句、文、文章、文書という言語情報（言語によって表現された情報）の構成単位。

(b) 記述のレベル：表層的なレベル [3] から深層の意味記述にかかわるレベル [4]。意味の記述方式によって色々なレベルが設定される。

(c) 言語の種類：記述対象となる言語の種類。日本語、英語等々。極力、言語に共通となる構造を設定する。ただし、情報構造の基本的な部分は同じであるが、細部は言語の違いを反映する。深層レベルにいたるにしたがって相対的な違いはなくなる。

（2）情報内容間の関係

サブ辞書は辞書項目の集合であるが、各サブ辞書内の辞書項目は、３つの特徴にそって関係づけられる。互いに関係し合うことによって辞書項目内の情報の定義がなされることになる。

(a) 記述の単位

この軸にそって２つの基本関係が定義される。

構成関係（コンテクスト関係）：上位の言語情報の中でどの言語情報のどの部分の構成要素となっているか。あるいは、上位の言語情報がどのような形でコンテキストとなっているかを示す。

説明関係：サブ辞書内のレコードの定義説明自身も言語情報によってなされることから生ずる関係。語と語義説明文との関係、文書と要約文（文章）、表題（句、文）、キーワード（語）との関係などである。

(b) 記述のレベル

意味表現関係：表層よりの構成単位と深層よりの構成単位との間の意味表現の対応関係。意味表現もある種の言語による意味情報の記述であると考えると表層言語と深層言語の間の対訳関係と見なすことができる。

(c) 言語の種類

対訳関係：それぞれの言語の言語情報の間に同義（ほぼ同義）であるという事実が認められる関係。原則として同じ構成単位の間での対応をとる。この関係はすべての言語対の間に定義される。また、記述のレベルのそれぞれに対しても定義される。

4 言語処理・文書処理との対応

言語処理（文書処理）の基本プロセスと汎電子化辞書の基本構造との対応付けを行う。これは、汎電子化辞書の利用・応用イメージを明らかにし、さらに、要件 (1)、(2) の観点から汎電子化辞書の妥当性を検証するためのものである。

言語処理も問題解決の一つである。問題解決の最も一般的なプロセスは、与えられた問題を極力独立な副問題に分割し、それぞれの副問題に対する解を得、それらを適切に合成し、所与の問題の解に達するというものである。実際には副問題が独立なものになることはまれなことである。言語処理においてはかなり複雑にからみ合う副問題に対処しなければならない。工学的には、いかに独立なものへと近似するかが要点となり、そのような適切な近似が得られたものから順次実用化されることになる。

(1) 言語処理における問題

判定問題：言語情報がある条件に合うか否かを判定する。あるいは、合う度合いを求める。条件としては、形態的、構文的、意味的に正しいか（ほぼ正しいか）否か、可読性の良さ等。
（例）スペルチェック、スタイルチェック、可読性判定。

比較問題：言語情報同志がある条件に合うか否かを比較する。あるいは、比較した結果の度合いを求める。条件としては同一であるか否か、類似しているか否か等である。
（例）キーワード検索（語と語）、フルテキスト検索（語と文書）、類似語検索（語と文書）、類似文検索（文と文書）、連想検索。

単純変換問題：言語情報を同種（同系統）のメディアの言語情報に等価変換する。
（例）かな漢字変換、点字変換、解析・生成。

翻訳変換問題：言語情報を別種（別系統）のメディアの言語情報に等価変換する。
（例）機械翻訳（対外国語）、自然言語インタフェース（対形式言語）、自然言語理解（対形式言語）。

要約変換問題：言語情報を同種（同系統）のメディアの言語情報に縮約変換する。
（例）キーワード抽出、カテゴリ抽出、知識フレーム抽出、アグストラクト生成。

(2) 基本プロセス

何段階かの規則ベースプロセスと事例ベースプロセスが適切に組合わされ基本プロセスを形作る。

〈規則ベースプロセス〉

ステップ１：要素となる構成単位からなる解析構造、すなわち構成要素と構造情報を生成する。
ステップ２：構成要素の解の集合を構造情報の変更と併せつつ全体を調整して解を生成する。

〈事例ベースプロセス〉

ステップ１：構成単位に対応する類似の解事例を検索する。

ステップ２：解事例を変更して解を生成する（解析構造は保存）。

汎電子化辞書においては、各構成単位について上位の解析構造を構成する時どのような性質のものとして振舞うのか、下位の解析構造にどのように分解されるのかの情報が記述されている。これらの情報をどのような規則にまとめ、どのようなアルゴリズムを適用するかは、応用・利用の自由度に委ねられるのが通常である。

(3) 基本構造との対応

記述の単位：この軸にそって基本的な副問題への分割が行われる。

記述のレベル：プロセスが曖昧度の低い、よりコンパクトな、場合によってはより精度の高いものになる。しかし、処理の負荷は増す。

言語の種類：翻訳変換問題に利用。

5 開発プロセスとの対応

開発の基本プロセスと汎電子化辞書の基本構造との対応付けを行う。これは、要件 (3) の観点から汎電子化辞書の妥当性を検証するためのものである。言語処理システムも開発プロセスにおいて重要な役割をはたす。

基本構造に照らし合わせ、どの順番で着手していくか、サブ辞書のレコード内のどの項目から記述していくかの基本手順が定まる。また、記述内容の相互の依存・制約の関係から規模の拡大や精度向上の基本プロセスが定まる。基本手順を決める要因は記述内容の安定性、網羅性、容易性である。表層的なものほど安定した記述となる。構成単位が小さなものほど網羅性が達成される。母国語の周辺になるほど作業者の確保や既存の蓄積の利用が容易になる。

規模の拡大や精度向上の基本プロセスは、言語直観に基づく作業者の記述作業と相互の依存・制約の関係をチェックする支援システムとの協調作業となる。基本的にはある言語処理機能（言語処理のある問題）を実現するために必要な記述内容に対しては、その機能を実現するシステムそのものが必要になる。このジレンマを解く漸進的、増殖的なメカニズムが必要である。

6 おわりに

汎電子化辞書の提案はこのようなものを実際に実現しようという提案ではない。言語知識のアーキテクチャ、すなわち、言語知識の情報構造モデルを明らかにすることを目的としたものである。この情報構造モデルによって、言語データばかりではなく、ハイパーテキストや大規模知識ベースの構造まで見通しの良いものにすることができる。

参考文献

[1] 日本電子化辞書研究所：”EDR 電子化辞書マニュアル”, http://www.iijnet.or.jp/edr (1995).

[2] 日本電子化辞書研究所：” 電子化辞書関連の研究開発の動向”, TM-044 (1995).

[3] 横井、木村、小泉、三吉、川田：” 汎電子化辞書：表層レベルの構造”, 言語処理学会全国大会 B3-3 (1995).

[4] 横井、田中、仲尾、荻野、野口：” 汎電子化辞書：深層レベルの構造”, 言語処理学会全国大会 B3-4 (1995).