専門用語辞書からの情報抽出と翻訳支援への適用

1P−5　高山泰博　森口 修　鈴木克志　丸山冬樹
三菱電機（株）情報電子研究所

1．はじめに

電子化された辞書からの情報抽出に関する種々の研究が行なわれている。これらのほとんどは国語辞典、英英辞典等に記載されている基本語に関する研究である。実際に自然言語処理を応用したシステムを利用する場では、基本語よりもむしろ個々の分野の専門語に関する情報を必要とする場合が多い。専門語辞書は、通常の辞書に比べて言語知識に関する情報（文法情報）が少ない代わりに、専門知識に関する情報を潜在的に保有している。本稿では、機械可読化した専門語辞書から情報を抽出し、構造モデリング手法であるＩＳＭ法(文献２)を応用した言語表現間の関連情報を増強する手法を提案し、翻訳支援への適用について述べる。

2．原データの作成と情報抽出

2．1　Prolog項形式による機械可読辞書の作成

中規模な専門語辞書の例として和英金融用語辞典（約１万７千見出し）（文献１）を機械可読化した。紙上の各辞書項目の形式をＢＮＦで以下に示す。

```
<辞書項目>::=<ローマ字索引（読み）>
              <見出し語>
              {<訳語列>}
              {<例文>}
              {<子見出し項目>}
              {<関連語>}
  <子見出し項目>::=
              <見出し語>
              {<訳語列>}
              {<例文>}
              {<関連語>}        {}:0個以上の要素
```

図１ 辞書項目の形式

紙に印刷された辞書を電子化し段階的に情報を抽出して実験を行なう場合に、情報抽出を容易にするためProlog項形式にタグ付けした原データを作成した(図２参照)。一見出し項目は一つのリストである。項目内のﾌｨｰﾙﾄﾞが複数要素のものはリストで表現する。そのﾌｨｰﾙﾄﾞの要素が存在しない場合には空リストとなる。

図２に示す例のように、複数語からなる英訳語列をquoted atom(複数の語を引用符で囲む)で記述し、例文をstring(文全体を２重引用符で囲む)で記述しておくことで、まとまった言語表現の抽出が容易になる。

```
[kashitsuke, 貸付,                    ← 読み,見出し
 [loan,advance,lending],              ← 訳語フィールド
 ["to extend a {loan} on bills",
  "to make an {advance}"],            ← 例文フィールド
 ref(貸付金) ↑                        ← 参照フィールド
].          string
[cho-kikashitsuke, 長期貸付,          ← 読み,子見出し
 isa(貸付),                           ← 上位語(親見出し)
 ['long-term loan'].                  ← 訳語フィールド
 [],          ↑                       ← 例文フィールド
 []       quoted atom                 ← 参照フィールド
].
```

図２ Prolog項形式による原データの例

2．2 機械可読化した専門語辞書からの情報抽出

原データのフィールドから大見出し−小見出しの関係を「上位−下位関係」として抽出し、参照フィールドから「参照関係」を抽出した。抽出した情報を再編集し、実際に検索に用いるファイル形式を作成した。抽出した上位−下位関係は4482個(上位見出し648個)、関連語の参照関係は321個であり、一見出しに対する訳語個数の最大は18個であった。また、約2760の例文が含まれていた。

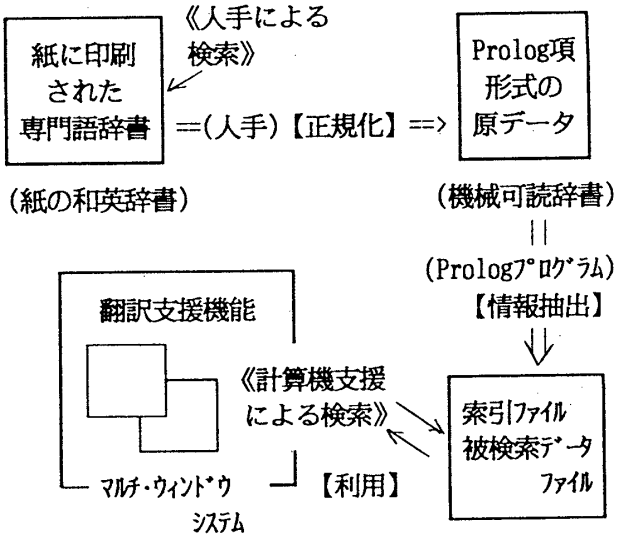

図３ 機械可読辞書の作成と情報抽出および利用

索引ファイルには、読み、見出しなどの主索引、副索引に加えて、上位-下位関係のための「階層索引」と参照関係のための「参照索引」とを格納する。これらの索引は、辞書情報の利用時に上位語・下位語や関連語の検索に使用する。

Information extraction from a domain specific dictionary and the application to translation support
Yasuhiro TAKAYAMA,Osamu MORIGUCHI,Katsushi SUZUKI,
Fuyuki MARUYAMA　MITSUBISHI Electric Corp.

3．言語表現間の関係情報の増強

2章で述べた参照関係の情報は豊富なものではない。実験に用いた辞書の場合には参照関係は全見出し数の2%弱しか記述されていない。また、辞書中の上位-下位関係および参照関係は辞書編集者の観点から整理されたものである。辞書情報を利用する際には辞書検索者の観点から関係情報を増やしていく必要がある。

そこで、構造モデル化手法の一つであるＩＳＭ法(Interpretive Structural Modeling)(文献2)を応用して関係情報の増強を図る。この手法は、断片的な知識に潜在している階層構造を顕在化させる。関係に疑似的な順序関係が存在するものと仮定して、大まかに関係を把握する。次に、顕在化した関係を選択的に保存することにより、抽出･蓄積された知識を再利用する。

ＩＳＭの基本処理は以下のステップで行なう。

(1)関係付け対象の要素(Vi:i=1,...,n)の設定
repeat
(2)関係行列X(ｎ×ｎの正方2値行列)の作成
(3)可到達行列X'の計算　　(入力･修正)」
(4)構造化行列の計算
(5)階層関係の表示
until(表示された関係付けに満足する)
(6)階層索引、参照索引の更新

(1)では、新しく追加する語と一対比較を行なう関連付け対象の候補となる語を、形態素解析を利用して辞書中から取出し提示する。(2)の関係行列は各要素Vi間の2値関係を入力する。(3)の可到達行列はWarshall法を用いて計算する。(4)の構造化行列は可到達行列X'=[xij]から各要素Vi(i=1,...,n)に対して、

可到達集合(Viから到達可能な全ての要素を含む集合)

$$R(Vi)=\{Vj|xij=1\}$$

と先行集合(Viへ到達可能な全ての要素を含む集合)

$$A(Vi)=\{Vj|xji=1\}$$

を求め、　$R(Vi)=R(Vi)\cap A(Vi)$

となる要素の集合(その集合外のどの要素からも到達できない要素からなる集合)をそのレベルの集合とし、全ての要素に対してレベルを決定することで計算する。

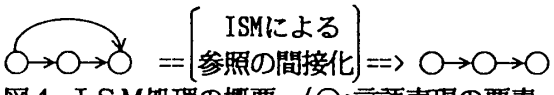

図4　ＩＳＭ処理の概要　(○:言語表現の要素

上記手法は(a)語と句,語と文等異なる単位の言語表現、(b)複数の知識源の情報の関連付けにも適用できる。更に、2値関係だけでなく多値化したり、属性情報を付加するなどの拡張も可能である。

4．辞書情報の検索と検索履歴の保存

辞書検索は複数回繰り返して行なうのが普通である。オンラインで辞書情報を柔軟に利用するには、一般的な検索･登録･削除に加えて以下の機能が必要である。

(1)検索語の前後の情報の表示

利用者が追加した語は、一つの見出しに一つの訳語だけを付与する場合が多いと推定される。一見出しの全ての訳語を表示するより検索語とその前後の見出しの第一訳語を同時に表示する方が有効な場合がある。

(2)検索履歴の保存・参照

検索を繰り返しながら、検索履歴を保存したり、必要な情報を追加することにより「"しおり"をはさむ」のと同等な機能を実現する。検索セッションが終了したら「"しおり"をはずす」ように検索履歴を削除する。

ＩＳＭでは間接的にも直接的にも関連付けられている要素を間接的に扱う(図4)。これらの間接的に参照されている要素同士を直接結び付けたい場合には、検索履歴の中の要素を指定して直接関係付ける(図5)。

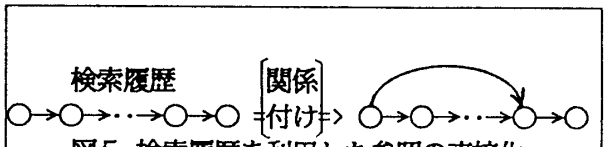

図5　検索履歴を利用した参照の直接化

5．翻訳支援への適用

言語表現の間の参照関係を動的に間接化したり直接化したりする、検索者の要求に応じた関連付け機能を以下の機能と共に翻訳支援に適用する。

(1)対訳文書の管理・編集　　(2)対訳例文の検索
(3)慣用句(複数の言語表現の組み合せ)の統合的検索
(4)単語辞書の検索

(1)～(4)は、それぞれ文書,文,句,単語といった言語表現の単位に対応する機能である。翻訳支援では、訳語･言い回しの決定までに必要な反復作業を支援し、現在の機械翻訳では取り扱えない文のレベルを越えた言語表現に関する情報を提供することが必要である。

特に、以下の点に考慮を要する。

(a)表現の追加・削除の容易性(b)選択的提示(c)再利用

6．おわりに

従来も翻訳支援のための用例検索等の研究(文献3,4)があるが、翻訳業務に携わっている利用者の実用に耐える情報をどのように提供するかについては課題が多く残されており今後研究を進めていく必要がある。

参考文献

(1)花田實:「和英･金融用語辞典」,ｼﾞｬﾊﾟﾝﾀｲﾑｽﾞ(1985).
(2)大内ほか:2項関係理論による知識獲得ﾂｰﾙとしての階層構造分析法の構成,
電学論,Vol.107-C,No.2,pp.135-140(1987).
(3)中村:用例検索支援ｼｽﾃﾑ,38回情処全大4E-5(1989).
(4)隅田,堤:翻訳支援のための類似用例の実用的検索法信学論,Vol.J74-D-II,No.10,pp.1437-1447(1991).