9B−4

データベース日本語検索システムのための日本語表記からの対象分野知識獲得方式

谷 幹也　市山 俊治

NEC 関西C&C研究所

1 はじめに

自然言語のインタフェースを様々なデータベースに適応させるためには、それぞれが背後に持っている対象分野の知識ベースを構築することが必要である。この知識ベースは、自然言語世界の概念と対象分野世界の概念の対応づけで構成される.従来、この知識ベースの構築には、知識表現および自然言語処理に対する知識が必要とされ、対象分野のみのエキスパートでは登録することが困難であった.本報告では、対象分野のエキスパートがデータベーススキーマに付与した日本語表記を言語解析し知識ベースを構築することで登録者の負担を軽くする実現方式を提案する.現在開発を進めている自然言語インタフェース構築キット"IF-Kit"[谷 91]上で本方式を実現した結果、従来手作業だった対象領域辞書作成作業を不要とすることができた.

2 対象分野知識ベース

自然言語の概念を対象世界上の概念にマッピングするためには、自然言語上での揺れを解消するための語彙－自然言語世界概念素(以降 CP と呼ぶ)の組合せと、対象世界上での意味を確定するための自然言語世界概念素－対象世界概念素(以降 DP と呼ぶ)の組合せを記述する必要がある.前者を対象領域辞書と呼び、入力文中で出現する語彙の表層と文法的情報から、対象となっている語彙の CP を選択するための情報を持つ.また、後者を、対象領域意味ネットと呼び、複数の CP の組合せからなる概念ネットを複数の DP の組合せにマッピング情報を持つ.この対象領域辞書と対象意味ネットを構成するに必要な知識を図1にまとめる.

	対象領域辞書	対象領域意味ネット
名詞句	CP 意味分類	CP-DPマッピング (対象世界解釈) DP-DPマッピング (同義語のリンク) 複合概念 (CPネット-DPネット)
用言	CP 意味分類　活用パタン 必須格パタン	
その他 未登録語	自立語への接続パタン	なし

図1. 必要知識

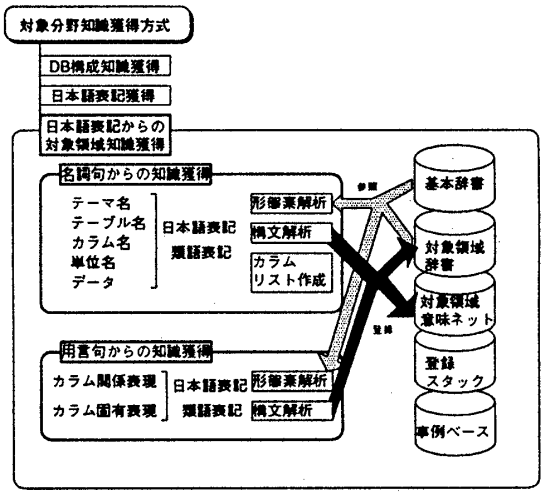

図2. 獲得方式

3 対象領域知識獲得方式

自然言語インタフェースでは未登録語獲得などのユーザ適応機能が存在するため、ある程度の対象領域知識を導入すれば、その後ユーザが使用しながら知識を増加させていくことができる.従来は、この導入時の対象領域辞書と対象領域意味ネットを作成するために、予想されるコーパスから単語の切り出しを行ない、その一つ一つに対して対象領域意味ネット上でリンク付ける必要があった.

図2で示す獲得方式によって、導入時の対象領域辞書と対象領域意味ネットを作成する際に、人が判断する必要性を減少させることができる.以下に、本方式の要件を説明する.

(1) 名詞句からの知識獲得

導入時の対象領域辞書には、最低限それぞれの DB 構成要素を識別できる最小の日本語表現があれば良い.そのためには、DB 構成要素の日本語表現を解析できるだけの対象領域辞書を作成する必要がある.

この DB 構成要素の日本語表現を得るため、[久保 92]で述べたように、DB 構成知識獲得と日本語表記獲得の両機能を用いて、対象領域の専門家が日本語表記を付与する.知識工学や言語処理の知識を用いなくてもこの作業は可能であり、データベースの構成要素間の関係は、データベースにアクセスすることで自動的に取得できる.

The domain knowledge acquisition method from japanese description for japanese information retrieval system

Mikiya Tani and Shunji Ichiyama

Kansai C&C Research Lab., NEC Corp.

ここで登録したDB構成要素の日本語表現を 基本辞書を用いて形態素解析し、既登録語に対しては基本辞書のエントリをそのまま使用し、未登録語に関してはその属性を推定する.この時、下位概念の表層には、上位概念の表層が含まれやすいため、上位概念から下位概念の方向に語彙の推定と登録を行ない、登録した語彙を利用して形態素解析を行なうことで、推定回数を減少させることができる.

そのため、名詞句からの知識獲得ではテーマ→テーブル→カラム→単位→データの順にそれぞれの形態素解析と構文解析を行なう.この構文解析の中で出現した、名詞句全てにその解析結果を登録することで、対象領域意味ネットの最小セットを作成することができる.名詞句の場合、対象領域辞書で必要となる情報はCPと意味分類だけであり、CPは表層で代用できるため、未登録語に対する意味分類を推定することが本方式のポイントとなる.

意味分類を推定できるのは、次の場合に限られる.

1. 同じ表層に関して意味分類の登録が終了している場合
2. その未登録語の前後に現れる語彙が既に現れていて、その場合の未登録が来ている位置にくる語彙の意味分類がわかっている場合
3. 一つのDB構成要素の意味分類が一つであるので、既に現在着目しているDB構成要素に対する意味分類が決まっていて、そのDB構成要素に対する日本語表記全体が未登録語となっている場合
4. そのDB構成要素にくる単位が定まっている場合

上記の場合以外は意味分類を推定することができないため、登録者に問い合わせる必要がある.この問い合わせも、未登録語の一つ一つの意味分類を仮定した場合の例文を作成して、その文意が正当か否かを問い合わせるなどの方式をとることで、言語処理知識を必要とすることなく、意味分類を決定することができる.

(2) 用言からの知識獲得

対象領域世界上では概念素の関係を表すことになるとともに、自然言語世界においても活用変化と格パタンを持つため、対象世界上に概念素の存在する名詞概念とは異なり推定が困難である.

そのため、用言概念の登録時には、以下の表現形式のように用言語彙を中心とした文に対して解析を行なう.

「(カラム1)が(カラム2)を(カラム3)に(用言語彙)」

この際に、カラム1～3の意味分類と格助詞から格パタンを、文末の活用語尾から活用パタンを推定するようにしているので、用言概念の登録までに、カラムの意味分類が必要である.図2のカラムリストは、

(カラム名, 日本語表層, 意味分類)

の3つ組で表されるリスト構造で、カラムに関する対象領域知識を作成する段階で作成し、上で述べたように用言概念の格パタンを決定する際に用いる.

4 実現方式

前章で提案した方式を自然言語インタフェース構築キット"IF-Kit"の対象分野知識入力支援ツール[久保92]の言語知識獲得部として実現した. 支援ツールでは、図3の知識ベース作成ツールを用いての日本語表記を専門家に入力してもらう.

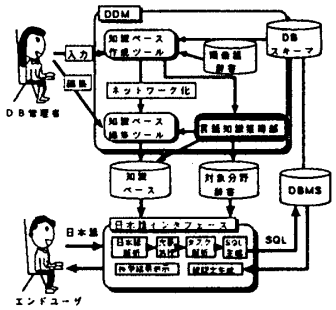

図3. 対象分野知識入力支援ツール

5 おわりに

従来、知識工学や自然言語処理の知識を必要とした対象領域知識の構築を簡単に行なえる方式を提案した.本方式は、データベーススキーマの日本語表記を未登録語を含む文として構文解析し、未登録語獲得をデータベースの構造を利用して行なうことで、対象領域辞書、意味ネットを漸進的に拡張していくものである.そのため、対象領域の専門家はデータベースの日本語表記と普通に使用する用言表現を入力し、簡単な質問に答えることで対象領域知識の獲得が可能である.実際に本方式を自然言語インタフェース構築ツール(IF-Kit)上で実現したところ、対象領域辞書作成作業が不要となり、対象領域意味ネットのプロトタイプを作成することができた.

今後は、基本辞書の拡張と基本的な言語概念に対する対象領域知識を拡張することで高精度な知識獲得を行なっていく予定である.

[参考文献]

[市山91] 市山俊治, 村木一至：自然言語インタフェースの構築キットの提案, 43回情処全大1H-2, 1991.

[谷91] 谷幹也, 飯野香, 山口智治, 市山俊治：自然言語インタフェース構築キット:IF-Kit, 信学技法NLC91-62, 1991.

[久保92] 久保加奈子, 市山俊治：自然言語によるデータベース検索のための対象分野知識入力支援ツール45回情処全大2F-10, 1992.