自然言語処理 97−13
（1993. 9. 17）

日本語による対象分野知識の獲得

谷 幹也　　市山 俊治
NEC 関西 C&C 研究所
m-tani@obp.cl.nec.co.jp

自然言語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．本研究では、データベースの日本語表記を解析し、部分文字列に対する概念及びその組合せ部分木を知識ベースに登録することにより、登録者が行なった数少ない日本語表記から、より広い入力表現をカバーするとともに、スキーマ情報と文法知識を用いることで、登録に不足した情報に対する登録者への問い合わせを少なくした対象領域知識獲得方式を提案する．本方式により、従来、知識表現や自然言語に対する専門的な知識を要した知識ベース構築作業を、対象分野のエキスパートが容易に構築することが可能になった．

Domain Knowledge Acquisition from Japanese DB-Scheme Description.

Mikiya TANI and Shunji ICHIYAMA
Kansai C&C Research Laboratory, NEC Corp.
NEC Kansai Bld., 1-4-24 Shiromi Chuo-ku Osaka 540 Japan.
baseand

Natural language interface needs domain-specific knowledges to work properly in a specific domain. Building up a domain-specific knowledge-base requires a great deal of labor. To use natural language interfaces in various domains becomes a significant issue. We propose an easier method to build up domain-specific knowledge-bases without any special skills in natural language processing and knowledge processing. This method has a feature:the system holds all sub trees of Japanese DB-sheme descriptions by users as domain-specific knowledge base, and can deal with various expressions which correpond to combination of these sub-trees. So, domain-specific knowledge experts only needs to answer minimum queries from this system, build up domain-specifc knowledge bases and dictionary.

1 はじめに

グラフィカルユーザインタフェース、マルチメディアインタフェースなど様々なヒューマンインタフェースが提唱される中でも、自然言語インタフェースは利用前の学習を要しないという点で、親和性の高いインタフェースとしての重要性を失っていない。

自然言語インタフェースの研究は、自然言語理解の主要な応用として古くから行なわれ、いくつかの実験システムが開発された [Winograd72,Bobrow et al.77, Woods et al.72]。これらのシステムでは、それぞれの対象領域に適した知識表現法を採用することで、強力な言語解析機能を実現したが、その結果限定された対象領域しか扱えず、異なる対象領域への適応性が十分ではなかったため実用化には至らなかった。その後、米国を中心に大規模なデータベースを対象として、対象領域移行性を備えた自然言語インタフェースの研究が進められ [Walts78,Hendrix et al.78,Kaplan84]、ここ数年製品化されるようになってきた。日本においても、データベース検索や事務処理操作などのユーザ親和性を増加させるために対象分野適応性を向上させた自然言語インタフェースの開発が活発化してきている [牧之内 他 88,絹川 86,難波 他 92]。

しかし、対象領域の知識ベースの構築に手間がかかりすぎるという問題は解決されていない。また、対象領域知識ベースが言語知識と複雑に組み合わされていたり、高度な知識表現を用いていたりするため、言語処理や知識処理の専門的知識が必である問題も解決されていない。そこで本稿では、データベースの日本語表記を解析し、部分文字列に対する概念及びその組合せ部分木を知識ベースに登録することにより、登録者が行なった数少ない日本語表記から、より広い入力表現をカバーするとともに、スキーマ情報と文法知識を用いることで、登録に不足した情報に対する登録者への問い合わせを少なくした対象領域知識獲得方式を提案する。現在、本方式を当社で開発中の自然言語インタフェース構築キット IF-Kit の対象領域知識獲得部として実装し評価を行なっている。c2章では自然言語インタフェースの能力を決定する対象領域知識 (自然言語表現 - 対象システム言語表現間の対応関係) と知識獲得における問題点について述べる. 3章では知識獲得についての問題点を解決する手段の一つとして、データベース日本語表記の解析と入力情報の最小化による知識獲得方式について述べる. 4章では自然言語インタフェース構築キット:IF-Kit の対象領域知識獲得部としての本知識獲得方式の実現方式について述べる. 5章では、現在の方式の問題点について述べる. なお、本文中の "ユーザ" はシステムユーザ (データベース管理者) を指すものとし、ユーザは対象領域及び着目しているデータベースの構造に関する知識を有するものとする.

2 対象領域知識と従来の問題点

2.1 対象領域知識

データベースに関する対象領域の知識とは、

1. データベースの構造に関する知識
2. データベースの構成要素と自然言語との対応に関する知識

の 2 種類に分けられる. 1. に関しては [久保 他 92] で詳細に説明した. 2. について、自然言語世界からの立場で分類したものが図 1である. 図 1においてカラム固有表現 (★) とは、数値カラムにおいて「値が大きい／小さい」時に用いるカラムに固有な形容詞的表現であり、カラム関係表現 (☆) とは、複数のカラムの間の関係を示す動詞的表現のことである [久保 他 92].

K1)	名詞句表現	
	K11)	テーマ名の日本語表現及び類義語
	K12)	テーブル名の日本語表現及び類義語
	K13)	カラム名の日本語表現及び類義語
	K14)	カラム値の日本語表現及び類義語
K2)	用言句表現	
	K21)	カラム固有表現 (★)
	K22)	カラム関係表現 (☆)
K3)	複合表現	
	K31)	マクロ表現

図 1: DB 構成要素と自然言語の対応付け知識

2.2 従来の問題点

自然言語インタフェースの利点である入力表現の多様性は、知識獲得の際に問題となる. つまり、図 1の K11),K12),K13),K14) のような DB スキーマの各構成要素に対する自然言語の対応付けの知識として、入力表現として想定し得る全ての自然言語入力に対する語彙の登録が必要となり、次のような問題が生じる.

1. 一つの構成要素に対する類義語の登録数が膨大
2. 登録に必要な情報の獲得手段が冗長
3. 複数の概念の組合せで表される日本語表記の場合
   - 一部が変化したものでも解釈不能
   - 複数の概念の間に別の表現があれば解釈不能

例えば 3. では、一つのカラム名の日本語表記に「3カ月の平均株価」があった場合、この日本語表記をそのまま登録しただけでは「3カ月の日本電気の平均株価」や「3カ月平均の株価」「3カ月の株価の平均」を解釈することができない.

3 日本語表記の解析と情報入力の最小化

前章で述べた対象領域知識の獲得における問題点を解決するために、次の 2 つを特徴とする知識獲得方式を提案する.

登録した表記より広い入力表現の確保　自然言語世界の概念と対象分野世界の概念の対応付けて構成される知識ベース上で、各データベーススキーマの日本語表記及びその類義語を、一語として各データベーススキーマに連結するだけではなく、その語の構成要素と構成要素間の関係をも知識ベースに登録することにより、ユーザが登録した数少ない日本語表記から、より広い入力表現をカバーする知識ベースを作成する.

不足情報の入力の最小限化　文字列の構成要素と構成要素間の関係を解析する際に、その時点までに登録された知識ベース上の情報を再利用し不足している情報を、スキーマ情報と文法知識を最大限に利用することで推定し、推定できなかった時にもユーザの入力が極力少なくなるように問い合わせを行なう.

3.1　解析による部分登録

ここでは [谷 他 93] で簡単に触れた自然言語解析を用いた対象領域知識獲得について詳しく述べる. 日本語表記記述部によって図 1の記述をそれぞれについて解析を行ない対象領域知識の獲得を行なう.

3.2　獲得方式

データベース構成上、カラム名の日本語表記にはテーブル名の日本語表記が、またテーブル名の日本語表記にはテーマ名の日本語表記が含まれる場合が多く、不足情報に対する情報の入力を最小とするため、図 1の K11),K12),K13),K14) の順に, 次の言語解析処理を行ない、カラムに関する意味分類情報を用いて K21) カラム固有表現、 K22) カラム関係表現の解析を行なう。

1. 対象領域辞書を使用して語切りを行なう.
2. 既登録語と未登録語からなる形態素列をその文法情報と解析ルールを用いて、ボトムアップにまとめ上げ概念表現に変換する.
3. 登録語、名詞句、用言句としてまとまった部分文字列に対しては 3.3の登録ルールを適応する.
4. 全体としてまとまった場合も 3.3の登録ルールを適応する

カラム名、テーブル名、テーマ名の日本語表記、類義語は名詞あるいは名詞句であるので、その解析木は名詞概念に関するネットワークとなる。ボトムアップにまとめ上げる際に全ての名詞句を辞書とネットワークに登録しながら解析を進めることにより、自然言語インタフェースの入力として、日本語表記、類義語から派生した概念の組合せも受理できることなり、インタフェースの入力表現の範囲をより広くすることができる。

例えば、

「3 カ月の平均株価」

の場合、これを解析して対象領域知識に登録することで (図 2)、2.2節であげた「3 カ月の日本電気の株価」のように、カラムの日本語表記文字列の中に異なった概念を示す文字列が入っているような表現や、「3 カ月平均の株価」「3 カ月の株価の平均」など一部に変更があった表現に関しても意味解釈を行なうことが可能となる。

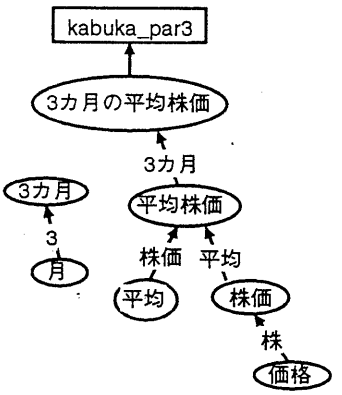

図 2: 獲得した対象領域知識の例

3.3　登録ルール

登録ルールは図 3と表現できる. 登録ルールの詳細について述べる.

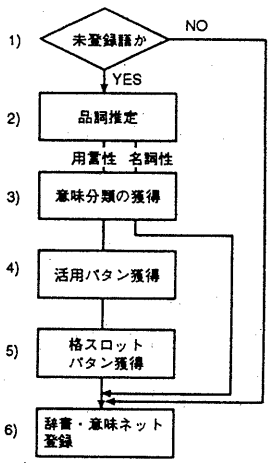

図 3: 登録ルール

1) 未登録語チェック　対象となる構造が未登録語でない場合、その構成要素の一部が既登録語であるた

め、対象領域知識ベースとして必要な情報を推定することが可能である。従って、その構造情報を対象領域知識ベースに登録する。

2) 品詞推定 未登録部分の左右の形態素から判断を行なう. 例えば、「設立される」の「設立」が未登録語であった場合、図4のように語切りできるので、未登録語に後接している「活用語尾＋連用接続助動詞」より未登録語を「動詞語幹」と推定することができる.

形態素	設立	さ	れる
品詞	未登録語	活用語尾	未然形接続助動詞

図 4: 用言性未登録語の例

3) 意味分類の獲得

1. **着目スキーマの意味分類の利用**

   未登録部分がある DB スキーマの日本語表記又は類義語の文字列全体で、既にその DB スキーマに対して意味分類が指定されている場合 (例えば、図5のように k_name の日本語表記の一つである「会社」に対する意味分類が決定されている場合)「企業」という未登録の日本語表記の意味分類として、同じスキーマに登録してある意味分類を登録する.

カラム名	k_name	
解析状態	日本語表記	意味分類
済	会社	or(組織, 場所, 抽象)
未	企業	-

図 5: 着目スキーマの意味分類利用

2. **共起関係からの推定**

   未登録部部の前接続語、後接続語との共起関係がそれまでに存在すれば、そこから意味分類を推定できる. 例えば、図 6) において、「営業利潤」と「営業利益」が同じカラムに対する記述であることから、未登録語「利潤」の意味分類として、「利益」の意味分類を登録する。

3. **例文を用いたユーザへ問い合わせ**

   意味分類が推定できなかった場合、その表層を用いて意味分類を仮定した例文を作成し、ユーザに対して質問を生成し意味分類を獲得する (図7).

共起 DB		
解析状態	日本語表記	共起 DB への登録
済	営業利益	営業 {ako or(動作, 抽象)} , 利益 {ako or(抽象, 金)}
未	営業利潤	-

図 6: 共起関係による意味分類の決定

次の表現のうち、
おかしくないものを選択して下さい。

1) 私は会社に会う。(人)
2) 私は会社を持ち上げる。(物)
3) 私は会社に属している。(組織)
4) 私は会社に行く。(場所)
5) 私は会社について考える。(抽象)
6) 私はそれに会社を払った。(金)
7) 会社が起こったのはいつか？(事象)
8) 4 会社。(単位)

番号を入力して下さい [1-8]>

図 7: ユーザに対する例文を用いた質問

4) 活用パタン獲得 これは用言性の未登録語に対してのみ必要な機能である。未登録語に右接続している附属語から推定し、曖昧性が残ってしまった場合、候補の活用語尾を連体形にして表示し、ユーザへ確認する. 例えば、図4の場合、“サ行変格活用” か “サ行五段活用” かわからないため図8のように、ユーザに対して問い合わせを行なう.

次の表現のうち、
おかしくないものを選択して下さい。

1) 設立する時
2) 設立す時

番号を入力して下さい [1-2]>

図 8: 活用パタンに対する質問

5) 格スロットパタン獲得 これは用言性の未登録語に対してのみ必要な機能である。

1. 共起関係からの獲得

   共起関係から同様の位置に出現した用言がある場合はその格スロットを使用する.

2. 前後の名詞からの獲得

前後に接続している名詞から格スロットを決定し、格スロットの条件を例文の形式でユーザに格にする.

4 知識獲得方式の実現

4.1 自然言語インタフェースの構成

処理単位と対象領域へのの依存性を明確にするために、構文解析部や文脈処理部、アプリケーション言語生成部などをモジュール化する [谷 他 91]。（以下、対象アプリーケーション言語を単にコマンドと呼ぶ）

自然言語インタフェースの構成を大まかに次の3つの構成要素でとらえることができる。

- 言語解析部
- タスク解析部
- コマンド生成部

言語解析部では、入力文に表された要求を概念素 (CP) とその格関係からなる概念表現に変換するため、対象領域に依存した語彙に関する辞書が不可欠であり、

タスク解析部では、概念表現からアプリケーションで実行するタスクを決定するために概念素 (CP) と対象領域上の概念素 (DP) 間のマッピングである対象領域知識の利用が不可欠である。

図?? で示す知識は、このように図?? で示す複数の対象領域知識として使用される.

そこで、対象領域適応時の変更範囲を最小にとどめるために、自然言語文の解析過程と、コマンドの生成過程から分離する (図 9) ことで、対象領域依存なモジュールを明確にし、ユーザによる変更が必要な範囲を明確にする。

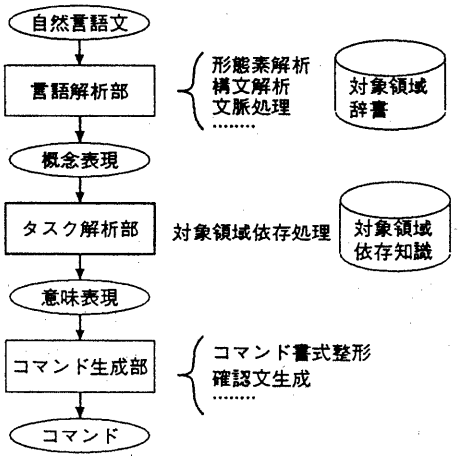

図 9: 自然言語インタフェースの構成

タスク解析部は、言語解析部の出力である概念表現を入力とし、アプリケーションを用いて遂行するタスクの表現を出力とする。対象アプリケーションの動作によって自然言語インタフェースにおける「意味」が定義できるので、この出力表現を意味表現と呼ぶ。意味表現は対象アプリケーションシステムが受理するコマンドやオプション、パラメタなどで構成される。

概念表現と意味表現はいずれもタスク解析内部ではネットワーク表現として扱っている。概念表現のノードに相当する概念表層を概念記述子、リンクによってつくられる構造を概念構造と呼ぶ。同様に意味表現のノードとリンクについても意味記述子、意味構造と呼ぶ。

「売上額が最も多い会社の名前を教えて」という簡単な文を例に概念表現と意味表現をそれぞれ図 10、図 11に示す。

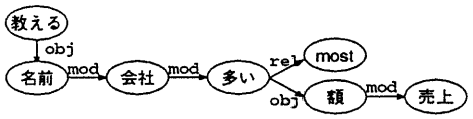

図 10: 概念表現の例

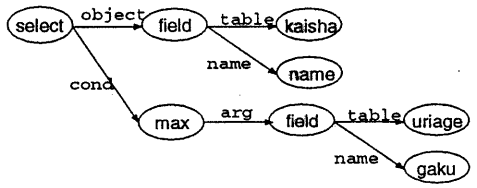

図 11: 意味表現の例

4.2 システム構成

図 12における構成のうち、DBMS スキーマ抽出部、日本語表記記述部に関しては、[久保 他 92] で述べられている知識ベース作成ツールをそのまま使用している.

4.3 DBMS スキーマ抽出部

データベース管理システム (DBMS) が持っているテーブルとカラムの一覧テーブルからテーブル名、カラム名、カラムのデータ型、幅など各 DB スキーマを自動的に獲得する. 同じ SQL インタフェースを持つ関係データベースであっても、詳細な部分 (例えば、データベースに含まれるテーブルの一覧表を持つテーブル名や、そのテーブルのスキーマ名称) は、データベース毎に異なるため、自然言語インタフェース:IF-KIt では、インストール時にデータベースに固有な情報は獲得する枠組を持っている.

4.4 日本語表記記述部

獲得したスキーマ情報を用いて、スキーマ情報を表形式で提示することで、図 1の情報をユーザから獲得

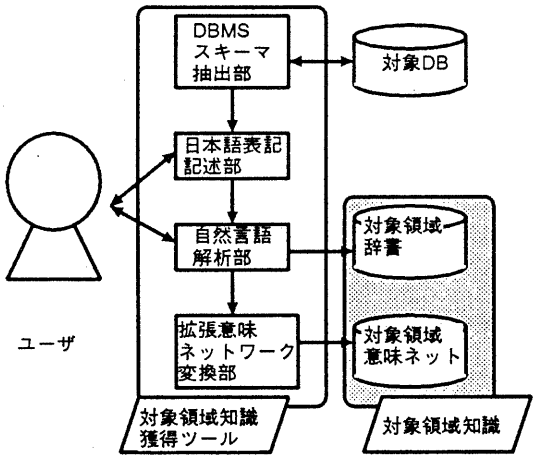

図 12: システム構成

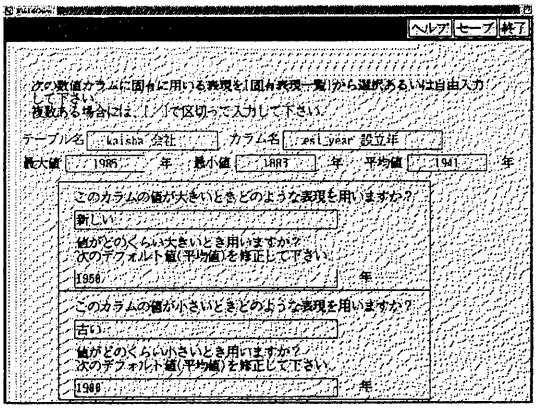

図 13: カラム固有表現定義

する [久保 他 92].

4.5 自然言語解析部

3.1節で述べた獲得方式を実現した. この際、辞書への登録／検索、共起パタンの追加／検索、対象領域知識ベースへの登録／検索、大域的なスタックへの登録／検索／削除などをすべてルールの中で指定できるようにして、登録形式の設定、登録の是非の決定の記述に柔軟性を持たせた.

名詞句表現 図 1の K1) で表される名詞句表現については 3.1節で述べた獲得方式である. カラムの日本語表記を解析する時に獲得されるカラムの日本語表記及び類義語から知識獲得の意味分類の一覧は、次の場合に使用するため対象領域知識として登録する.

1. カラム関係表現における未登録用言の格スロット条件の決定
2. 対象領域意味ネット上で疑問詞の曖昧性の絞り込み

用言句表現 用言句表現には図 1の K21) カラム固有表現と 22) カラム関係表現が存在する.

K21) カラム固有表現 カラム固有表現は、日本語表記記述部において、図 13のように定義される.

カラム固有表現は、“一つの自然言語の語彙” を対象領域世界の複数の概念の複合体にマッピングする知識であり、複合体の構造はインタフェースとして解析を行なった概念表現と同じものとするので、対象領域知識ベースとして図 14のようなネットワークを生成する.

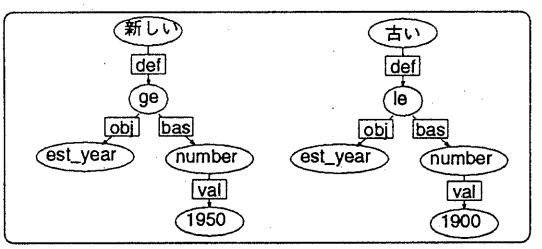

図 14: カラム固有表現の意味ネット表現

現在、カラム固有表現に出現する表現を次の表 1に絞って獲得することにしている. それぞれの右接続して助動詞が存在する場合も獲得対象とする.

表 1: カラム固有表現の表現形態

品詞	例
形容詞	「新しい／古い」
形容動詞	「静かな／きれいな」
名詞＋体接助動詞	「国際的だ」
名詞＋助詞＋形容詞	「儲けの多い」
名詞＋助詞＋形容動詞	「音が静かな」
名詞＋助詞＋名詞＋体接助動詞	...
未登録語	「近い」
部分的未登録語	「影響力が大きい」

K22) カラム関係表現 カラム関係表現は、日本語表記記述部によって図 15のように獲得できる。カ

ラム関係表現では、「構成要素間の関係」を表す用言表現を獲得する.

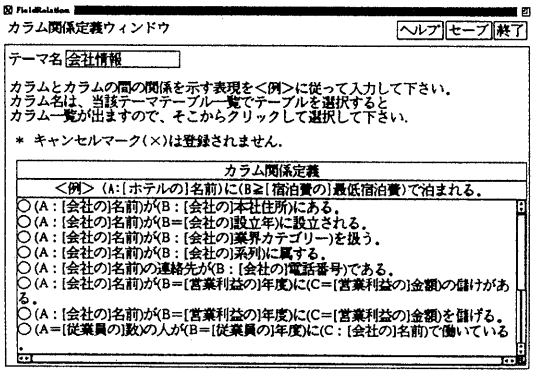

図 15: カラム関係表現定義

カラム関係表現の解析時には、図 15の関係表現の最後の用言に着目して、最後の用言が表 2の場合に絞って獲得することにしている. それぞれの右側に連用接続助動詞が接続する場合も獲得対象とする.

未登録語の場合は、3.3節の登録ルールに準じる. ただし、格スロットパタンの獲得は、関係表現のカラム表現から格要素の意味分類条件を切り出して対象領域知識を作成する。図 15の最後のエントリに対する対象領域知識は図 16のようになる。図 16は、自然言語インタフェースの言語解析部が出力した概念表現において、3),4),5) で表される深層格で「人が働いている」につながっている部分を、それぞれ 3),4),5) の位置で展開した斜線部のネットワークに置き換えることを意味している. つまり、「４５０００人の人が１９９２年に日本電気で働いている。」という入力文に対する概念構造 (図 17) は、タスク解析部によって、意味構造 (図 18) となることを意味している。

表 2: カラム関係表現の表現形態

品詞	例
動詞	「ある／いる／属する」
名詞＋体接助動詞	「である」
名詞＋助詞＋動詞	「儲けがある」
未登録語	「位置している」
部分的未登録語	「営業利益をあげる」

4.6 意味ネットワーク変換部

日本語表記記述ツールによって、獲得した (データベーススキーマ - 日本語表記) 対応情報と自然言語解析部によって生成された名詞句、用言句に対する概念

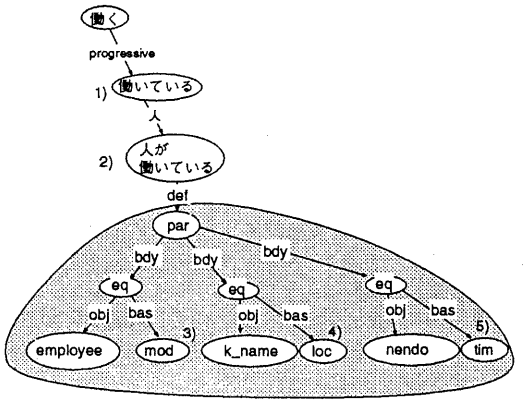

図 16: カラム関係表現の意味ネット表現

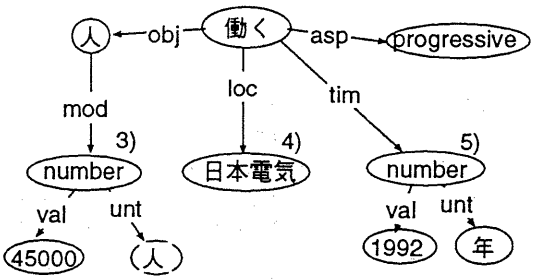

図 17: 例文の概念構造

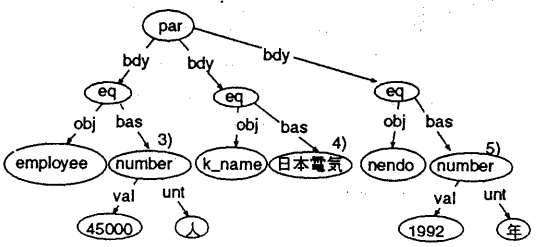

図 18: 例文の意味構造

のネットワーク記述を拡張意味ネットワークで表現される対象領域意味ネットに登録する [山口 他 92].

5 考察

4章で実現した対象領域知識獲得方式に関する問題点について述べる.

自然言語の単一の概念に対して対象領域世界の複数の概念の複合体をマッピングさせる知識も、カラム固有表現の枠組 (図 14) のように表現することで獲得することができる。しかし、カラム関係表現のように用言概念と格要素のような複数の概念の複合体を対象領域世界での複合体にマッピングする枠組の汎用的な記述方式を定式化していない. この問題の代表例を 2 つ上げる.

自然言語による定義 一般に概念を定義する際には、自然言語で定義することが多い. プリミティブとなる十分な対象領域知識が獲得されていれば、ある概念を自然言語で汎用的に記述することができる.

例えば「豊かな会社＝売上が 2 兆円以上で利益率が XX %以上である業界カテゴリーが保険か金融・証券の会社」のような自然言語による定義文を解釈する枠組を実装していない.

複雑なカラム関係表現 カラム関係表現では、図 16のように、自然言語上の格要素にあたる対象領域世界上の部分ネットワークを展開する場所を指示できるネットワークとして記述される。この際、展開する場所の指示子は用言表現への格関係を表すラベル (obj,agt,tar,sor,mod など) であるため、同じ格要素を関係表現に対しては、場所を特定できない

6 おわりに

自然言語インタフェースを様々な対象領域に適応するためには、対象領域知識ベースの構築に手間がかかりすぎるという問題があった。

本稿では、この問題に対するアプローチとして、対象領域内のデータベースの日本語表記を解析し、部分文字列に対する概念及びその組合せ部分木を知識ベースに登録することにより、登録者が行なった数少ない日本語表記から、より広い入力表現をカバーするとともに、スキーマ情報と文法知識を用いることで、登録に不足した情報に対する登録者への問い合わせを少なくした対象領域知識獲得方式及びその実現方式について述べた。

現在、自然言語インタフェース構築キット IF-Kit の対象領域知識獲得部として実装し、複数のデータベースに対して評価を行なっている。

参考文献

[Bobrow et al.77] D. G. Bobrow et al. GUS:a frame-driven dialog system. *Artificial Intelligence*, Vol. 8, pp.155–173, 1977.

[Hendrix et al.78] G. G. Hendrix et al. Developing a natural language interface to complex data. *ACM Trans. on Database Syatems*, Vol. 3, No. 2, pp.105–147, 1978.

[Kaplan84] S. J. Kaplan. Designing a portable natural language database query system. *ACM Trans. on Database Syatems*, Vol. 9, No. 1, pp.1–19, 1984.

[Walts78] L. D. Walts. An engulish language question answering system for a large relational database. *Comm. of ACM*, Vol. 21, No. 7, pp.526–539, 1978.

[Winograd72] T. Winograd. *Understanding Natural Language*. Academic Press, 1972.

[Woods et al.72] W. A. Woods et al. The lunar science natural language information system. Final Report, BBN Report 2378, BBN Inc., 1972.

[久保 他 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. 第 *45* 回全国大会予稿集, pp. 2F–10. 情報処理学会, 1992.

[絹川 86] 絹川 博之. 表階層モデルに基づく自然言語インタフェース処理方式. 情報処理学会 論文誌, Vol. 27, No. 5, pp.499–509, 1986.

[山口 他 92] 山口 智治, 市山 俊治. ドメイン適応性に優れた自然言語インタフェースのための知識表現と要求理解機構. 技術研究報告 AI92-102 ～ 109, 電子情報通信学会, 1992.

[谷 他 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. 技術研究報告 NLC91-62, 電子情報通信学会, 1991.

[谷 他 93] 谷 幹也, 市山 俊治. 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. 第 *45* 回全国大会予稿集, pp. 9B–4. 情報処理学会, March 1993.

[難波 他 92] 難波 康晴, 辻 洋, 絹川 博之. 自然語インタフェースにおける操作対象と操作条件の表現. 第 *45* 回全国大会予稿集, pp. 3.137–138(2F–8). 情報処理学会, 1992.

[牧之内 他 88] 牧之内 顕文, 吉野 利明, 泉田 義男. 移行性のあるデータベース自然言語インタフェース. 情報処理学会 論文誌, Vol. 29, No. 8, pp.749–759, 1988.