実用性を重視した日本語検索システムの試作

4C−4

笠　晃一　　小林修二　　横田将生
(㈱日本データベースネットワーク研究所)　　(福岡工業大学)

１．はじめに

関係データベースにアクセスするには、ＳＱＬ[1]などの形式的問合せ言語を利用するのが一般的であるが、形式的問合せ言語は抽象的かつ複雑であるので、修得するのに時間がかかるし、また、仮に修得できたとしても、検索内容によっては、かなり複雑な問合せ文になる可能性がある。このため、自然言語によってデータベースにアクセスできるようなシステムが、いくつも考えられてきた[2]-[4]。しかしながら、いままでの自然言語によるデータベース検索システムは、そのほとんどが、大型計算機やワークステーション上で開発されたものであり、パーソナル・コンピュータ上で動かせるようにはなっていなかった。

我々は、自然言語によるデータベース検索システムを広く普及させるには、パーソナル・コンピュータ上での動作が不可欠と考え、研究を行なってきた。どのような応用プログラムにも言えることであるが、パーソナル・コンピュータのもつ二つの制限、つまり、速度的な制限と記憶容量（特に主記憶）の制限のために、できるだけ速く、できるだけメモリを消費しないようなシステムを作る必要がある。しかしながら、よく言われるように、速度と消費メモリ量とは、互いにトレードオフの関係にあり、この両方を満足させることはなかなか難しい。我々は、ＤＣＧ[5]で記述された構文・意味規則をＳＡＸアルゴリズムのＣプログラムへ変換するＳＡＸ−Ｃトランスレータを開発し、この問題をある程度まで解決することができた。

２．システムの全体構成

試作したシステムは、図１に示すような構成になっている。構文解析と意味解析を同時に行なっているため、処理の組合せ的爆発を抑えることができ、したがって、処理時間を短く、作業領域を少なくすることができる。

２．１．形態素解析部

日本語問合せ文として許されるのは、漢字かな混じりのべた書き文であるため、普通にワード・プロセッサを使っているような感覚で入力できる。また、ヒストリ機能を持っているので、以前に入力した文をいつでも呼び出すことができ、タイピングの手間を最小限に抑えることができる。形態素解析部では、この入力された日本語問合せ文を最長一致法を用いて単語に分解し、各単語の構文的・意味的情報を次の構文・意味解析部に引き渡している。解析用辞書の構造として、ＴＲＩＥ構造[5]に改良を加えたものを使っている。

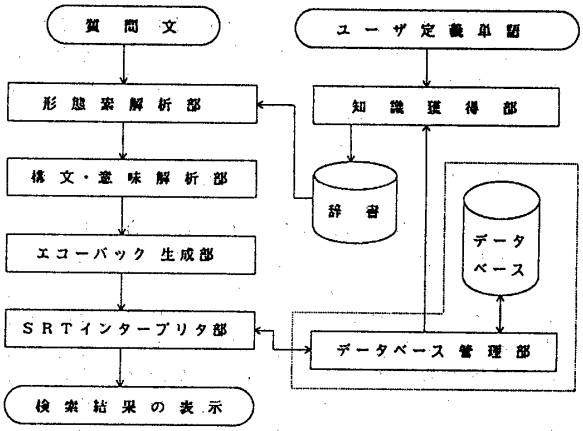

図１．システムの全体構成

２．２．構文・意味解析部

構文・意味規則の記述にはＤＣＧ記法を用いている。この構文・意味規則は、ＳＡＸ−Ｃトランスレータと呼ばれるツールによって、ＳＡＸアルゴリズムのＣプログラムに変換されてから実行される。ＳＡＸ−Ｃトランスレータと言うのは、松本らによって開発されたＳＡＸトランスレータ[6]を改良したもので、ＳＡＸトランスレータがＰｒｏｌｏｇプログラムを生成するのに対し、ＳＡＸ−ＣトランスレータはＣプログラムを生成する。したがって、ＳＡＸ−Ｃトランスレータの方がより実用的だと言える。なお、ＤＣＧ記法を用いているため、構文解析と意味解析は同時並行的に行なわれる。そのため、構文解析木を経ずにいきなり意味表現を生成でき、この点も、処理時間の短縮と作業領域の縮小に寄与している。

ここで使われている意味表現は、ＳＲＴと呼ばれるものである。ＳＲＴというのは基本的には、形式的問合せ言語ＳＱＬと同じものであるが、多少の拡張を施してあり、また、生成や解釈がしやすいように木構造の形をしている。ただし、意味解析部によって直接作り出されるものは、深層ＳＲＴと呼ばれるもので、これは、ＳＱＬを圧縮したような構造をしている。これを、意味トランスレータによって、表層ＳＲＴ（あるいは単にＳＲＴ）に変換している。

２．３．エコーバック生成部

ここでは、構文・意味解析部で作られたＳＲＴを曖昧性のない日本語になおして表示する。エコーバック生成部の本来の働きは、問合せ文がユーザの意図通りに解釈されたかどうかを、ユーザに確認してもらうと

Experimental and Practical Retrieval System with a Japanese Inerface
Kouichi Ryu , Shuji Kobayashi, Masao Yokota
Japan Database Network Laboratory Co.,Ltd., Fukuoka Institute of Technology

いうものである。理想的なシステムであれば、問合せ文はユーザの意図通りに解釈されるはずであるが、現在の技術レベルにおいては、どうしても微妙な点で解釈の食い違いが出てくるので、エコーバック生成部を省略することはできない。エコーバック生成部のもう一つの働きは、構文的曖昧性の解消である。構文的曖昧性が生じたとき、係り受け関係を示して、ユーザに正しい方を選択させるというやり方もあるが、多少とも文法知識が必要になり、あまり現実的ではない。ここでは、生成された複数個のＳＲＴをそれぞれ曖昧性のない日本語で表示して、その中から正しいものを選ばせるというやり方をとっている。

２．４．ＳＲＴインタープリタ部

データベース管理部が扱えるのは、関係データベースの関係表が一つに限定された形で、いわゆるカード型データベースといわれるものである。したがって、データベース管理部が受けつけるＳＲＴも単純な条件式を持つものに限られ、たとえば、異なる条件で二度以上検索しなければならないものなどは、受けつけることができない。したがって、ここでは、ＳＲＴを単純な条件式に分解して、データベース管理部へ送っている。検索結果を保持する機能を持っているので、一度検索して、その結果を使って再び検索しなければならないような場合にも対処できる。

２．５．知識獲得部

問合せ文の中で使われる名詞のうち、属性名と属性値の情報は自動的に対応するデータベースから取り込まれるので、ユーザが入力してやる必要はない。また、問合せ文の中で使われる動詞のうち、「住む」とか「生まれる」のような属性名に関連のある動詞は、学習によって自動的に属性名と関連づけられる。したがって、システムに何も教えないでも、最初からかなりの範囲の文を使用することができる。なお、システムに同義語やキー・フィールドなどを教えたい場合を考えて、問合せシステムをユーザが教育することもできるようになっている。たとえば、「少年」を「１５才以下の男性」のように定義しておけば、「福岡に住んでいる１５才以下の男性は誰か」と質問する代わりに、「福岡に住んでいる少年は誰か」と聞くことができるし、「名前」フィールドと「所属」フィールドをキー・フィールドに指定しておけば、「３０才以下の人の趣味は何か」のように聞いたときでも、「趣味」フィールドだけでなく、「名前」フィールドと「所属」フィールドも表示してくれるようになる。

３．システムが受理可能な文

システムが受理可能な日本語検索文の一つは次のような形をしている。

検索文　－－＞　制約，主題，［は］．

たとえば、「福岡に住んでいる社員は」と言ったとき、「福岡に住んでいる」の部分が制約、「社員」の部分が主題である。制約は検索条件を表す部分であり、主題は検索対象のデータベースが何についてのものかを表している部分と考えることができる。制約には次のようなものがある。

①一つの属性を一つの属性値と比較するもの

「趣味が読書である」、「年齢が２０以上の」、「福岡に住んでいる」などは、ある一つの属性を一つの属性値と比較するもので、最も単純な制約と考えられる。

②一つの属性を複数の属性値と比較するもの

「趣味が読書とピアノである」、「年齢が２０以上３０以下の」、「福岡か熊本に住んでいる」などは、ある一つの属性を複数の属性値と比較している。

③一つの属性を組集合の属性と比較するもの

これは、「山川さんの年齢よりも若い」とか「山川さんより若い」などが当てはまる。

④以上の３つが組み合わさったもの

①から③までを、連言的もしくは選言的に組み合わせて、「趣味が読書とピアノで福岡に住んでいる」とか「山川さんより若いか、あるいは年齢が２０以上の」のような制約を作ることができる。

４．まとめ

現在作成中の日本語によるデータベース検索システムについて述べた。現在の構文・意味規則数は約５００であるが、４０文字程度の文であれば、パーソナル・コンピュータ（ＰＣ－９８０１ＲＡ）であっても、たいてい、１秒以内に解析を終えることができ、たとえ構文・意味規則の追加分を入れたとしても満足のできる速度が得られるものと思われる。なお、今後の課題として、次のようなものが挙げられる。

①ＤＣＧ規則を直接アセンブラ・プログラムに変換してしまうトランスレータを開発したい。
②文脈に依存した処理を加えたい。
③ＤＣＧをＣ言語に変換すると、Ｐｒｏｌｏｇに変換したときに比べ、デバッグがやりにくくなる。したがって、専用デバッガも開発したい。

《参考文献》

［１］Date,C.J.: An Introduction to Database Systems, Addison-Wesley, 1975
［２］藤崎他：データベース照会システム「ヤチマタ」と名詞句データ模型，情報処理学会論文誌，Ｖｏｌ．２０，ｎｏ．１，１９７９
［３］服部他：データベースの日本語インターフェースにおける対話処理について，電子通信学会技術研究報告，Ｖｏｌ．８６，ｎｏ．２１６，１９８６
［４］伊藤，高橋：データベース用自然言語インタフェース，情報処理学会第３８回全国大会講演論文集，ｎｏ．２，１９８９
［５］Aho,A.V., Hopcroft,J.E. and Ullman,J.D.: Data Structures and Algorithms, Addison-Wesley, 1983
［６］松本，杉村：論理型言語に基づく構文解析システムＳＡＸ，コンピュータソフトウェア，Ｖｏｌ．３，ｎｏ．４，１９８６