解　説

日本語によるデータベース照会†

渋谷政昭†† 藤崎哲之助†† 鷹尾洋一††

まえがき

本稿ではデータベースについての照会を行うのに，日本語により十分目的を達せられることを示す．重要な点は，意味を正しく形式化することであり，このために名詞句データ模型と呼ぶ新しい概念を導入した．この概念と照会文との対応関係について議論する．

1. 日本語によるデータベースの照会

日本語　「……日本の国語程，不完全で不便なものはないと思ふ．……私は此際，日本は思ひ切って世界中で一番いい言語，一番美しい言語をとって，その儘，国語に採用してはどうかと考へてゐる．それはフランス語が最もいいのではないかと思ふ．(志賀直哉)」

このような極端な発言は例外的としても，日本語は複雑，曖昧で含意が多く，とうてい機械的な分析はできない，という意見はかなり広い．しかしながら日本語は，データベースの照会に使用するには十分「論理的」であるし，むしろ西欧語よりも解析には便利な面がある．語順ではなく格助詞などにより語の格が定まること，前の語が後の語を修飾すること，動詞の変化が規則的であること，などである．かな書き入力とすると，同形異義語が多く，分かち書き規則に不定部分が多い，など手間のかかることはあるが，これらの欠点も致命的ではない．本稿では日本語について述べるが，基本内容はすべての自然言語に共通なものであることを始めに強調しておく．

質問解答システム　自然言語により計算機に質問をし，計算機がこれに答える質問解答システム (QA system) は昔から人工知能の主要テーマであり，いくつかの方向の試みがあった．データベース照会の先駆としても，たとえばMITで開発されたBASEBALL (1961) は，アメリカン・リーグの1シーズンの月日，チーム，得点に関する質問に答えるもので，入力文にたいする比較的簡単なパターンの当てはめで，良い解答をするものであった．60年代末からCAL-TECHのF. B. Thompsonが開発したRELは構文規則の定義を簡便化して諸種の言語開発を容易にしたものであり，その主要成果であるREL-Englishの上ではザンビア住民の人類学調査データの照会システムが作られた[1]．この頃から単なる人工知能研究にとどまらず実用化を探るための研究が始められた．BBNのA. Woodsが開発したLUNARはアポロが採集した岩石標本に関する照会システムで地質学者の使用実験を試みたりしている[2]．

データベース研究開発が進んで関係形式データ模型が確立され，構文分析の良いプログラムが作られるようになって，より本格的な照会システムが開発された．最近ではIBM T. J. Watson Res. Ctr. のREQUEST[3]，IBM Heidelberg Sci. Ctr. のUSL[4]，Toronto大学のTORUS[5]，SRIのLADDER/INLAND[6]など類似のシステムが動いている．

現実世界の限定　質問解答システムの中でデータベース照会がもっとも成功しているのは十分な理由がある．一般に，計算機との対話を行おうとするときに話題の範囲をどこまでに限るかを定めることが困難である．意味の形式化を明確にするために対象を狭くすればつまらぬものとなり，少し拡げようとすると大きな困難に遭遇することが多い．

データベースの照会では話題が明確に規定され，問い方も自然に限定される．もちろんデータベースから得られる情報をすべて得るには汎用の計算言語を使用する他はないし，いわゆるデータ準言語 (data sub language) にも，いくつかの水準のものがありえるが，基本検索演算は日本語で十分に覆える．

ヤチマタ　上記のいくつかの照会システムは，特定のデータベースを対象として開発されている．各システムには当然，適用するデータベースに依存する部分としない部分があるが，依存しない部分を明確にとり

† A Japanese Language Query System by Masaaki SIBUYA, Tetsunosuke FUJISAKI, and Yoichi TAKAO (Tokyo Scientific Center, IBM Japan).

†† 日本アイ・ビー・エム(株)東京サイエンティフィック・センター

出すにはどうすればよいか．著者たちが開発した実験システム「ヤチマタ」では，この問題を解決するために「名詞句データ模型」の概念を導入した．このシステムと概念についてはすでに報告している[7),8)]ので，ここでは自然言語処理としての側面をより詳しく議論する．

2. 名詞句データ模型と照会文

名詞句データ模型　一般に，利用者が大量データを見るときの視点を定め，データの論理的構造を規定する方式を，一般にデータ模型とよぶ．「名詞句データ模型」は特に，自然言語とデータベースの橋渡しをするためのものであり，すべての名詞，名詞句は一群のデータを表わし，逆に一群のデータには適当な名詞句が対応している，とみなす模型である．ただし一群のデータとは，後述のように1枚の表によって表わされるものである．

たとえば大学における情報システムで「学生」とは「在学生」と同義語であって，受験生ではないし，全県，全日本の学生でもない．さらにデータ「学生」は氏名，学生番号，学年，学科，性別のような必要なデータ全体を表わすものである．ここでは，「妹のいる学生」，などは意味のない言葉として排除することになる．長期休学者や退学者をいつまで「学生」に含めておくかはデータベース運営の問題であって照会システムの処理することではない．利用者は対話者＝計算機の回答可能範囲を大体心得て聞くことになる．

「……は何か，どこか，……」という質問も，答が肯定否定の質問も，本質的には「<名詞句>は？」という形の質問に等しい．

修飾と検索　諸種の条件を満たすデータを照会文で指定するためには，いくつかの修飾句をもつ名詞句を用いることになり，各修飾はデータベースにおける一つの検索演算に対応する．**図-1** はその例で，「価格」，「部品」，「在庫量」という普通集合名詞が図中の表（「名詞表」と呼ぶ）のように定義されていることを前提としている．文中の「5号倉庫」は名詞表の要素として現われる固有名であるが，「5万円」という数字とともに1行1列の名詞表とみなされている．図の中では2種類3個の修飾が行われている．基本検索関数は **表-1** の通りである．

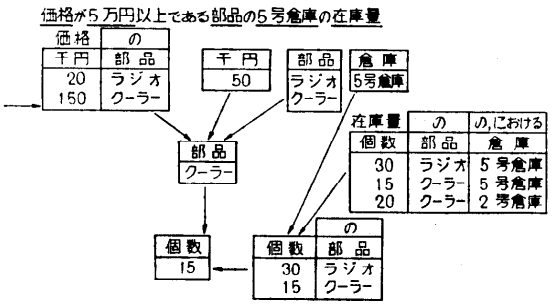

図-1　修飾句と検索との対応

表-1　基本検索関数と構文（下線は一例を示す）

構　　文	基本検索関数
<名詞句><後置詞><名詞句>	正順制約
<名詞句>が<名詞句>である<名詞句>	逆順制約
<名詞句>である<名詞句>	論理積
<名詞句>と<名詞句>（接続詞）	論理積，論理和
<名詞句>＋<名詞句>	2項演算
<名詞句>の平均	統計関数

表-1の<後置詞>とは格助詞などの集りであって，どのようなものが許されるかを名詞表に明記しておく．名詞表の列には，図-1の部品，倉庫のような定義域と呼ぶ，いわば集合名詞の小範疇を記入し，その列の要素の同，不同，大小比較可能を指定する．

表-1以外に用言による連体修飾句も許される．**図-2** のように二つの意味が異なる（！）「ある」を定義すれば，図-1の照会を「5号倉庫にある……の在庫量」と書けるし，「港区にある倉庫」という句も許される．

多様さと省略可能性　名詞表，動詞表はデータベースのサブスキーマと呼ばれるものに対応し，実際，実働化のときはサブスキーマの機能により定義できる．これらの表を増すことにより，図-1，2でも示したように同一の意味を多様に表現できるし，必要なデータを指すのに「何々表の何列と何列にある……」などの細かい記述なしに，簡単に直接に指定できる．

<名詞句><後置詞><名詞句>の構文は，二つの表の結合演算（join）であるが，これも「何々表と何表をどの列に関して結合する」ことを簡単に表現している．定義域と後置詞の等しい列を対応させることにより，いわばジグソーパズルのように正しい表を正しく合わせる．この機能は隠れた所でも働いて，**図-3** のような二つの名詞表の演算・比較をするときには，言外の制約条件を合わせることになる．

名詞句データ模型では，このように自然言語の多様

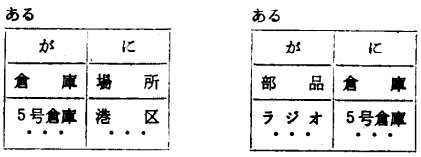

図-2　動詞の定義，同形異義の定義

歩合給が固定給より多い社員

歩合給	の，における	の
金額	年度	社員
140万	昭53	山田
150万	昭52	山田
・・・	・・・	・・・

固定給	の，における	の
金額	年度	社員
160万	昭53	山田
130万	昭52	山田
・・・	・・・	・・・

図-3　言外の制約条件（統合）

さと省略可能性を積極的に取り入れている．このような特徴を計算言語一般に活用しようという提案はいくつかされており，実用化もされている．

統語論　既製の統語論の中にはきわめて秀れているものもあるが，結局は言語の一面を表わしているに過ぎない．データ模型により形式化した意味は異なる面を把えているのであるから，その統語論として既製のもの，あるいはその一部を用いることはできない．

名詞句データ模型で，実は名詞と名詞句の区別はない．名詞表に対応するか，その要素に対応するかが問題で，修飾された名詞句もそうでない名詞（名詞表）も同じものである．「部品」と並んで「単車用の部品」を定義してよいし，「大学院合格学生」という名詞表の同義語として，「大学院に行く学生」を用いても，動詞「行く」が定義されていなければ支障はない．

定義域という概念も，統語論で用いる「生物／非生物」のような言語的範疇や，物理的次元に類似しているが，より細かく現実性を含めたものである．同じ金額（円）でも，取引の額と日用消耗品の額を比較することはないと判断すれば異なる定義域とすべきである．

後置詞も，格（case）を定める語であれば助詞に限ることはない．たとえば「東京（から，発，を出て）大阪（へ，行き，に向かう）（の）便」の括弧の中の語は名詞句データ模型では同等の役割を果たす．

3. 名詞句データ模型の問題点

曖昧さ　曖昧さが，多様な表現が許されているときに省略をし過ぎて生ずることは，自然言語でも，名詞句データ模型でも同じである．対話者間で話題についての了解があり，それを前提とする省略を行ったとき，了解の食い違い，判断の誤りで文が曖昧となる．このようなときは，話者へ問い正して解消するのが基本的な解決法である．名詞句データ模型では次のような種類の曖昧さが生じ得る．

a　多義な名詞句を指示したとき：「（販売員）の（年度）の販売額」と「（支店）の（年度）の販売額」の両方が定義されていると，常に「昭和53年度の販売額」と言ったのでは曖昧である．

b　修飾の曖昧さ：「AのBのC」という型の文はB，Cを限定する定義域が共通のときに「(AのB)のC」か「Aの（BのC)」か曖昧となる．たとえば，「30才以下の課長の秘書」，「港区にある倉庫に納入している業者」．前者は，人間にとって曖昧でない表現に変えれば解消する．後者は，「倉庫に納入している」という句が「どこかの倉庫に納入している」ことを意味して曖昧となる．修飾句は原則として近い方の句を修飾するという規則を設けたり，コンマやセミコロンで区別したりもできるが，常用的でない．

接続詞を含んで「AとBのC」，「AのBとC」などとなると，類似のものが並列されて，修飾範囲が不確実となることが多い．逆に「(社員）の建物」と言えば必ず「(社員）の課の建物」を意味するときには，前者の省略形を許すことが考えられる．しかし，これを濫用すると思いがけない解釈をされる危険を生じる．

c　統計関数の計算：「(人）の（年度）の収入」が定義されているときに，「収入の平均が400万円以上の人は？」という文では年度に関する平均と定まり曖昧さはないが，「収入の平均は？」という文では人に関する平均，年度に関する平均，の二つの意味がある．

d　比較における省略：「社員数が人事より多い課」という句だと問題はないが，「北海道にある出張所の数がA社より多い会社」となると，修飾の範囲の曖昧さのために何を比較するか不明となる．比較するものを省略せずに書くしかないが，一般に構文分析は困難である．

拡張の可能性　データベースのほかに知識ベースを設け，そこでの推論により照会の柔軟性を増そうという試みがある．上位・下位概念について3段論法を適用するなどというのがもっとも初歩的である．名詞句モデルでは名詞表を適当に定義すれば同等の機能を発揮できる．秘書は社員であり，社員は給与を受け取るから秘書は給与を受け取る，という事実は，社員の職務という限定属性に秘書という要素を置くか，秘書という名詞を定義してその定義域を社員として表現できる．

推論も見方によっては省略であり，その可能性は設計段階で慎重に考慮する方が安全である．しかし使用者が，使用中に簡便な表現の使用を望むことは十分に

考えられるので，言葉により新しい言葉を定義する機能は望ましい．

たとえば「首都圏＝東京と神奈川と千葉と埼玉」，「首都圏人口＝首都圏の人口の和」……と上位概念を定義し，同様に下位概念を定義することを許す．さらに，「"港区"の会社＝所在地が"港区"の会社」のように，いわば変数を含む定義が考えられる．

実時間処理のためのデータベースでは処理時刻を記録することが望ましい．それ自身は難しくないが，時刻に関連した照会文では時制を正しく処理しなければならない．「……時までの」，「……時に終っていた」「……と……の間に」のような修飾を扱うことになり，構文規則はずっと豊富にしなければならない．前述のUSL[4]などでは時制を取り入れている．

名詞句模型の限界 名詞句が1枚の表に対応するというのは強い制限であって，「部品の価格と業者の住所」などは許さない．実用上は照会を分割すれば処理できるし，「大阪の昭和50年の人口と人口密度は？」の結果のように，構造の同じ二つの表を結合して出すことは難しくない．名詞表の定義法を変えて，一つの名詞表に関連する量をすべて詰め込んでおく手段も考えられる．

上記の「人口と人口密度」は出力形式の問題とも考えられる．何をどの順序に表示したいか，は場合によって異なり，それを言葉で定義するのは適切でない．画面上のメニュー選択などで別に処理すべきであろう．

限定詞 (quantifier)，つまり「すべての，一つでも，それぞれの，何人かの，……」の処理が困難であることが認められている．構文分析に手間がかかるし否定を含むと意味が曖昧になる．それがうまくいっても検索に手間がかかる．努力の割に得ることが少ないので敬遠されるのであるが，破るべき壁である．

4. おわりに

データベース照会言語として，どのようなものが理想的であるかは，使用者層の範囲，使用目的，費用などに依存する．自然言語に近ければ，多数の人が簡単な学習で，ほとんど手引き書なしに照会できるが，やはり費用は高い．絶対的な価格が下がり，広汎な層の需要が高まったときに実用化されるであろう．

参 考 文 献

1) Thompson, F. B. et al.: REL——A rapidly extensible language system, ACM Natnl. Conf., No. 24, pp. 399-417 (1969).
2) Woods. W. A.: Progress in natural language understanding——An application to lunar geology, AFIPS Conf. Proc., Vol. 42, pp. 441-450 (1973).
3) Plath, W. J.: REQUEST——A natural language question-answering system, IBM J. Res. Develop., Vol. 20, pp. 326-335 (1976).
4) Lehmann, H.: Interpretation of natural language in an information system. IBM J. Res. Develop., Vol. 22, pp. 560-572 (1978).
5) Mylopoulos, J. et al.: TORUS——A step towards bridging the gap between data bases and the casual user, Inf. Syst., Vol. 2, pp. 49-64 (1976).
6) Hendrix, G. G. et al.: Developing a natural language interface to complex data, Proc. 3rd Conf. VLDB, pp. 292-302 (1977).
7) 藤崎哲之助他：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, pp. 77-84 (1979).
8) Sibuya, M. et al.: Noun-phrase model and natural query language, IBM J. Res. Develop., Vol. 22, pp. 533-540 (1978).

(昭和54年6月4日受付)