問い合わせ文の言い替えと複数化

長谷川 守寿

1. 目的

本稿の目的は、問い合わせの仕方に対応した表現である基本検索条件がどのように言い替えられたり、複数化したりするかを分析することである。言い替えられた文や複数 化された文を調べ、検索条件に働く規則を明らかにすることは、問い合わせ文の解析システム作りに役立つと思われる。

2. 先行研究

藤崎他(1979)では、複数の検索条件を「集合演算」和積に対応させ、名詞表をもとに 説明しているが、実際にどのような構文があり、言い替えや制約があるかに欠ける。絹 川(1986)でも「論理演算子」に対応させているが具体的な構文や言い替え、制約に欠け る。加藤・中川(1984)では、並列の処理として「or並列」「and並列」に分類し、判定 の問題を論じているが、これは属性値に関する考察が不足していると思われる。また、 動詞に関して、藤崎他(1979)では動詞表を使い動詞を扱っているが、使える構文は「 < 名詞句>は?」の形だけでWH疑問詞は名詞句内に入らないが、本研究では、WH疑問 詞と動詞が名詞句に含まれるものも考察する。長谷川(1993)では、基本検索条件が1つ で、しかも何の言い替えもなされていないもののみを対象としているが、本研究では基 本検索条件が動詞等を使い、言い替えられたものと、基本検索条件が複数の場合の間い 合わせ文を考察する。

3. 方法

対象とする問い合わせ文は、筆者が先行研究の例文等を参考に作文した1400文程の中の、内部に埋め込みを持たず、属性値が明示されている文約400文である。さらに、長谷川(1993)で制限したように、問い合わせ文を表示文(該当するデータの出力を求める文)に限定した。問い合わせ文は、基本検索条件に述語(この場合「出せ」等)が加わったものが、言い替えられたり、複数化してできたものという観点から分析を行う。分析のもととなる基本検索条件は、以下の形で表現できるものである(対応する検索の仕方や、語彙等に関する詳細は長谷川(1993)参照)。

- 基本検索条件 ① <データ名>
 - ② <データ名>の<属件>
 - ③ (<属性>が)<属性値><制限表現><データ指示語>
 - ④ (《属性が・の》)《一番》<形容詞><データ指示語>
 - ⑤ <属性>

また、本研究でモデルとして使用するデータベースは、CDと曲のデータベースで、 以下のような属性(下線部)と属性値(波線部)を持つものとする。

CD={(CD番号 TOCT-6676)(発売された年 1992)

 $(CD91hN A, O, R)(P-F1Xh \Delta-v919-X)$

曲={(<u>曲番 4</u>)(<u>曲名 さよならを手に</u>)(<u>時間 4:31</u>)(<u>作詞家 鈴木慶一</u>) (<u>作曲家 阿田</u>徽)(<u>編曲家 阿田</u>徽)(<u>追加情報 なし</u>)}

よって、この研究の場合、<データ名>は「A.O.R」や「さよならを手に」等の CDタイトルや曲名、<データ指示語>はどのデータかを指示する言葉で、「CD」か 「曲」である。

4. 言い替え

4.1. 言い替えの種類

問い合わせ文の言い替えについて調べる。問い合わせ文を検索条件と述語がらなると する観点からいくと、言い替えもまた以下のように分類される。

- ・検索条件だけに言い替えが働き、述語には影響を与えないもの
- ・述語にだけ言い替えが働き、検索条件には影響を与えないもの
- ・文全体に言い替えが働くもの

4.2.検索条件の言い替え

検索条件を言い替えた文には、まず語彙を言い替えた文がある。属性や、相対的な数値を表す属性値を同義語である別の言葉に言い替えた文である。この言い替えが可能な検索条件は②から⑤である。①は<データ名>なので、言い替えはない。

- (1)「My Boy」のレコード会社 → 「My Boy」の発売元
- (2) 昨年のCD → 去年のCD

次に検索条件内のデータと属性の関係から、動詞を使って言い替えられたもの(以後動詞化と呼ぶ)があるが、これには述語、全体の言い替えもあるので、4.5.でまとめて述べる。

また、検索条件④のみに可能な言い替えとして、<属性が・の><一番><形容詞> <データ指示語>の語順の文の、<一番>+<形容詞>の部分を、「最長、最多、最古」 等に「の」をつけた表現に言い替えたものがある。

- (3)収録時間が最も長い C D → 収録時間が最長の C D
- (4)曲数が一番多い C D → 曲数が最多の C D

4. 3. 述語の言い替え

次に述語だけを言い替えた場合を考える。まず、これには述語をWH疑問詞に言い替えたものがある。使われるWH疑問詞については属性等によって決定されるので、4.5.の動詞化で詳しく述べる。またこれには、WH疑問詞を検索条件と述語との間に挟み(6)のように言い替えたものもある。

- (5)「誕生」の発売された年を出せ。→「誕生」の発売された年はいつか?
- (6)「誕生」の発売された年はいつか、出せ。

また述語「出せ・表示せよ」等が「出して下さい・見たい・見たいのですが」等にかわって、命令文から願望・依頼文に言い替えられた文もある。この際、願望を表す助動詞「たい」がつく場合、今までの助詞が「を」から「が」にかわった文もある。

(7)1990年のCDを表示せよ! →1990年のCDを表示してください。 →1990年のCDが見たいのですが。

4.4.文全体の言い替え

全体を言い替えたものでは、検索条件④に4.3.で述べたWH疑問詞(この場合「何」のみ)を使う言い替えをあわせ、WH疑問詞を文頭に置く語順の入れ替えを行った文がある。これは④が唯一のものを決定するため可能な文である。この場合④の属性は省略されるのが自然と思われ、(9)のような文はやや不自然かと思われる。

検索条件④は何か → 何が検索条件④か

- (8) 最も古いCDは何か? → 何が最も古いCDか?
- (9)収録時間が最も長いCDは何か? → *何が収録時間が最も長いCDか?

表示文につき、希望件数を表す副詞「全て、全部」は基本検索条件⑤の場合、文頭に 名詞として位置をしめ、基本検索条件②のような「全て・全部+の<属性>」の形をと ることもある。

(10)作詞家を全て出せ。 → 全ての作詞家を出せ。

4.5.動詞化を伴う言い替え

動詞化が加えらえた文を見ると、基本検索条件①と⑤に関するものがない。これらは < データ名>又は < 属性 > のみを指定するもので、動詞化の余地がないからと思われる。 ④ (属性値の最大、最小を出す)に関するものは、 < 形容詞 > が副詞にかわった動詞化で、これは③からの類推で構文の説明ができるので、結果の最後で述べる。②(データ「X」と属性を指定し、その属性値を出力させる。構文は「Xの < 属性 > 」)、③(ある属性と属性値「Y」を指定し、そのペアを持つデータを出力させる。構文は「 < 属性 > が Y の < データ指示語 > 」)に関して、どのような動詞化の加わった言い替えがあるかを調べる。

本研究のデータベースに対する問い合わせ文には、以下のようなものがある。より広範なデータと属性の関係からの動詞化について考察を加えるのは、別の機会に行う。

4.5.1.表記法

結果を示す前に表記法を示す。 "DSG"はデータ指示語か、その上位概念の言葉で、この場合「CD・曲」または「もの」である。 "HV"は「出せ、出力せよ」のような表示を求める動詞を指す。 "属性" は属性を表す言葉そのものかまたはその上位概念の言葉を指し、「作詞家・作曲家・編曲家・アーティスト・ヴォーカル」のような場合は、「人」、「発売された年」の場合は「年」、「レコード会社」の場合は「会社・ところ」というようになる。 "RARE"は助動詞「れる・られる」を指し、 "(A)"は括弧内が省略可能なことを、 "《A》《B》"はAB、BAの語順が可能なこと、 "A・B"はAかBのどちらか一方が現れることを示す。

4.5.2. 個別結果

まず、共通していえることは、WH2は共通で、全て「何」である。これは基本検索 条件③が、本研究の場合全てCDか曲を指示しているものであるためと思われる。WH は②で使われるため、属性によって異なると思われる。

次に個別の結果として、まず、曲と人名の属性(作詞家・作曲家・編曲家・アーティスト・ヴォーカル)、CDとレコード会社のような、属性がデータの主体と考えられる場合には次のような文型がある。

曲と人名 (V=作詞した・作曲した・編曲した・歌っている) (VN=詞・曲 VS=書いた・作った) (WH=誰)

- CDとレコード会社 (V=発売した・出した) (WH=どこ)
 - ②XをV+属性'+をHV・はWHか

WHがXをV(の)か

XはWHがV(DSG)か

XのVNはWHがVS(の)か

XはWHに(よって) V+RARE(DSG)か

XのVNはWHに(よって) VS+RARE(の)か

③ Yが・のV+DSG+をHV・はWH2か 《Yが》《VNを》VS+DSG+をHV・はWH2か Yに(よって) V+RARE+DSG+をHV・はWH2か YはWH2をVか?

さらに、CDとレコード会社のように属性がデータの源泉と考えられる場合には以下の文型がある。Vは「出た・発売された」、WHは「どこ」である。

- ②XはWHからV(DSG)か WHからXはV(の)か
- ③YからV+DSG+をHV·はWH2か

次に、CDと発売された年のように属性がデータの日時と考えられる場合、以下の文型がある。Vは「出た・発売された」、WHは「いつ・何年」である。

- ③Y【に】V+DSG+をHV・はWH2か

(【に】は前件が"今年·去年·昨年·X年前"の場合以外)

次に、CDと曲数、CDと収録時間のように属性がデータの数量と考えられる場合以下の文型がある。なお、曲数と収録時間は3.の方法の使用データベースには入力されていないが、曲の数、曲の時間の合計を得ることでわかる属性である。

- CDと曲数 (V=入っている・収録されている WH=何曲)
- CDと収録時間(V=入っている・収録されている WH=何分)
 - ②《X(に)は》《WH》Vか XにV+属性'+をHV・はWHか
 - ③Y+V+DSG+をHV·はWH2か

さらに上記の場合の全てにおいて、「V+DSG・属性'」を、「V+の」に言い替えたものがある。

- (11)「月夜」を作詞した人は誰か?→「月夜」を作詞したのは誰か?
- (12) CBSが発売したCDを出せ。→CBSが発売したのを出せ。

また、述語では「+はWH・WH2か」の部分を「+は」だけに省略したものがあるが、これは語用的にその前までの部分で何を要求しているかわかるためと思われる。

- (13) CBSSONYが発売したCDは?
- (14) 「月夜」を作詞したのは?

また、WH、WH2の値が固有名詞になる場合、以下のように言い替えたものもある。 WH→何という(名前の)属性' WH2→何という(名前の)DSG

- (15)「月夜」を作詞したのは誰か? → 「月夜」を作詞したのは何という人か?
- (16)「REPLICA」を発売したのはどこか?
 - →「REPLICA」を発売したのは何という名前のレコード会社か?

また、Vを「VNをVS」に形に言い替えることができる時、「VNをVS」に言い替えた文の場合、以下のようにVの前の助詞がかわる。

- * & V → * O V N & V S
- *? V → *? V N を V S (*は仟意の名詞 ?は「を」以外の助詞)
- (17)秋元康は何を作詞したか? →秋元康は何の作詞をしたか?
- (18)東芝EMIが発売したCDを出せ。→東芝EMIが発売をしたCDを出せ。

ここで、今まで除外してきた④の動詞化であるが、これは数値に関係ある「CDと発売された年」等の場合で、③のYの部分(下線部)を<一番><形容詞の連用形>に替えたものと考えられる。

(19)1987年に発売されたCDを出せ。→最も新しく発売されたCDを出せ。

本研究で対象としたデータベースの場合、以上のような言い替えられた文型が見られるので、解析にはこれらに対応した処理が必要となる。この場合「XはWHへ行く」のような属性値がデータのゴールと考えられる構文は見られなかったが、使用するデータベースとその属性から、この研究のように処理に必要な動詞化の構文を設定することで検索条件の言い替えには対処できると思われる。

5. 複数の検索条件

基本検索条件を複数持つ文を分析するには、基本検索条件間に生じる関係とそれを表す接続語を用いて分析を行う。ここで使用する関係は論理演算子に対応し、その論理演算子の種類とそれに対応すると思われる表現を以下に挙げる。

論理和(OR) : < と>:と、か、もしくは、または、あるいは、・・・・ 論理稿(AND): < で>:で、でかつ、でしかも、であり、の中で、・・・・

5.1,論理和の基本文型

論理和の基本文型は以下のような形で、使われる基本検索条件は①から⑤である。 基本文型 <検索条件1><と>、・・・、<検索条件N>

(Nは含まれる基本検索条件の数 N≥2)

5.1.1. 論理和の言い替え

論理和において、同じタイプの基本検索条件を繰り返す時、省略がなされるものとそのまま繰り返されるものに分かれ、省略が行われるものは②③④であり、①⑤はそのまま繰り返すだけである。異なるタイプの基本検索条件を使う場合(例えば②と③)は(20)のようにそのまま<と>で結ばれるのみである。

- (20)「最後の晩餐」の発売された年と、発売された年が90年のCD
- ②を繰り返すと以下のようになる。

<デ'-9名1>の<属性1><と>、・・・・、<デ'-9名N>の<属性N> この場合、データ名が同じ時は(21)に、属性が同じ時は(22)の形に省略したものがある。

- (21)「MOON」の発売された年とCD番号
- (22)「MOON」と「PANT」の発売された年

そして、<データ名><属性>ともに以下のように複数になったものもある。(M≥2) →<デ'-タ名1><と>、・・・・、<デ'-タ名N>の<属性1><と>、・・・・、<属性M>

- (23)「MOON」と「PANT」の発売された年とCD番号
- ③を繰り返すと以下のようになる。

(<属性1>が) <属性値1><制限表現1><データ指示語1><と>、・・・・、

(< 属性N>が) < 属性値N> < 制限表現N> < データ指示語N>

(24)のように基本検索条件を<と>で結んだ形の他に、<と>で連結させられたものの 属性とデータ指示語が同一の(25)のような場合、(26)のように(N-1)までのデータ 指示語とNの属性が省略される。

- (24)発売された年が89年のCDと、レコード会社が東芝のCD
- (25)発売された年が89年のCDと、発売された年が90年のCD
- (26) 発売された年が89年と90年のCD

またこの文型で、属性値が人名の時、解釈において曖昧性の問題が生ずる。例えば(27) のような例である。これには(28)と(29)の解釈の可能性が生ずるからである。

- (27)作曲家が鈴木慶一と渚十吾の曲
- (28)作曲家が鈴木慶一の曲と、作曲家が渚十吾の曲
- (29)作曲家が鈴木慶一と渚十吾のペアの曲
- ④の場合、4.2.で述べた言い替えを経て「最長の曲」のような形になったもの(30)は、 ③と同様に言い替えられた文があるが、その言い替えを経ていないもの(31)は繰り返し のみとなる。
 - (30)時間が最短のCDと、最長のCD → 時間が最短と最長のCD
 - (31) 最も短いCDと、最も長いCD

5.2. 論理稽の基本文型

論理積の検索条件の基本形は、複数個の基本検索条件を<で>で結んだ形とする。使用される基本検索条件は③のみである。

基本文型 <検索条件1> <で>、・・・<検索条件N>

- (32)50分以上のCDで、曲数が10曲以上のCD
- (33)作詞家が中島みゆきの曲で、時間が6分以上の曲

また、論理積を表す検索条件と一緒に使うことができる基本検索条件は④であり、④ は最後に来る。④は(34)のように<と>を伴ない、複数個現れるものもある。④を含む 時は基本検索条件③は(35)のように1つの場合もある。

<論理稿>+4

- (34)87年以降のCDで、中島みゆきのCDで、一番短いCDと一番長いCD
- (35)作詞が秋元康の曲で、一番長い曲

5. 2. 1. 論理積の言い替え

論理積に関する言い替えで、まず語順に関するものをみていく。この場合検索条件 P $(2 \le P \le N)$ を先頭に移動する言い替えが加えられたものがある。これは助詞も変更を加えられ、以下のような形である。この場合のみ説明が分かりやすくなるように < 述語>を加えた。

<検索条件1><で>、・・、<検索条件P><で>、・・、<検索条件N>を出せ。

- →<検索条件P>を、<検索条件1><で>、・・・、<検索条件N>(の中)から出せ。
 - (36) 45分以上のCDで、87年に出たCDで、10曲以上のCDを出せ。
 - (36a)→87年に出たCDを、45分以上のCDで、10曲以上のCDから出せ。
 - (37) 89年のCDで、CBSのCDで、時間が最長のCDを出せ。
 - (37a)→ *89年のCDを、CBSのCDで、時間が最長のCDの中から出せ。
 - (37b)→時間が最長のCDを、CBSのCDで、89年のCDの中から出せ。

この時、④を含む場合、Pの位置に来ることができるのは④のみである。それは④がその段階で唯一のものを決定してしまうからと思われ、例えば(37)の場合「時間が最長のCD」の部分で唯一のデータが決定されるので、(37a)のような文は不自然と思われる。

次に、語形が変わった言い替えをみる。(32)(33)からもわかるように、複数個の基本 検索条件の中に現れるデータ指示語は、共通であるという論理積の特徴から次のように 言い替えられたものがある。

まず、<検索条件1>の中のデータ指示語を残し、それ以後の検索条件の中のデータ 指示語を「もの」に言い替えたものがある。

- (38)50分以上のCDで、10曲以上のもの
- (39)レコード会社が東芝のCDで、1987年のもの

次に、検索条件のN番目のデータ指示語を残し、その前までの基本検索条件中のデータ指示語を削除したものがある。

(40)50分以上で、10曲以上のCD

この場合には、<で>のかわりに「の」でつないだものがある。「<属性>が」は文の先頭のみに位置する。

(41)87年以降で、10曲以上のCD → 87年以降の10曲以上のCD

これにはまた、4.5.の検索条件の言い替えで述べた動詞化が加えられて、<で>でつなぐかわりに、以下の方法1・2でつながれたものがある。

方法 1. (N-2)までを連用形、(N-1)を連体形にする

方法2. (N-1)までを連用形、Nを連体形にする

- (42)50分以上入っていて、10曲以上収録されている、東芝のCD
- (43)87年に発売されて、60分以上入っているCD

また、数値に関する同一の属性を含み、基本検索条件が2つで、数値の幅を指示する 検索条件の場合は以下の形をとり、(44)のような文である。この場合属性とデータ指示 語が同一のため(下線部)、前件のデータ指示語と後件の<属性が>が現れずに、(45) のように言い替えられた形をとることもある。

(<属性が>)<属性値><大小関係表現語><データ指示語><で>、

<u>(<属性が>)</u><属性値><大小関係表現語><u><データ指示語></u>

- → (< 属性が >) < 属性値 > < 大小関係表現語 > < で > < 属性値 > < 大小関係表現語 > < データ指示語 >
 - (44)87年以降のCDで、90年以前のCD
 - (45)時間が3分20秒以上で、3分40秒以下の曲

さらに、この場合くで>も削除され、以下の形をとることもある。

- → (<属性が>)<属性値><大小関係表現語>
 - < 属性値><大小関係表現語><データ指示語>
 - (46)3分20秒以上3分40秒以下の曲

5.3.論理和と論理積に関する制限

基本検索条件③において、同じ属性を複数使った場合、属性の中には、論理積と論理和を構成できるものと、論理積を構成せずに論理和のみ構成するものに分かれる。

属性値が文字型で、データの一部に"(作曲家 鈴木慶一・渚十吾)" のように属性値に複数の固有名詞を持つことがある属性(本研究では、作詞家・作曲家・編曲家・ボーカルの人名)の場合、(47)(48)のように論理和も論理積も構成することができる。(47)は鈴木慶一と渚十吾が組んで作曲した曲である。

- (47)作曲家が鈴木慶一の曲で、作曲家が渚十吾の曲
- (48)作曲家が鈴木慶一の曲と、作曲家が渚十吾の曲

しかし属性値が文字型で、しかも属性がレコード会社・アーティスト名・CDタイトル・曲名のような、属性値に複数の固有名詞を持つことがない場合、(49)のように論理和は構成するが、(50)のような論理積は構成しない。

- (49)レコード会社がキャニオンのCDか、レコード会社がCBSのCD
- ×(50)レコード会社がキャニオンのCDで、レコード会社がCBSのCD

次に属性値が数値型で、しかも大小関係表現語を伴う属性値が2つある場合のみ数値の幅を決定するので(範囲関係表現形成)、(51)や(52)のように論理和も論理積も構成する。それ以外の数値の幅を決定しないもの(53)や、範囲関係表現語を含むもの(54)は論理積を構成できず、(55)(56)のように論理和のみ構成する。

- (51)10曲以上のCDと、12曲以下のCD
- (52)10曲以上のCDで、12曲以下のCD
- ×(53)87年のCDで、88年のCD
- ×(54)3分台の曲で、4分台の曲
 - (55)87年のCDと、88年のCD
 - (56)3分台の曲と、4分台の曲

以上をまとめると図1になる。属性は全て論理和を構成できるが、論理積は条件を満 たす時のみである。

属性値の 型	論理和構成 の可能性	論理積構成 の可能性	条件
文字型	0	0	データの属性値がX・Yのように2つ 以上の固有名詞からなる場合[注1]
文字型	0	×	データの属性値が固有名詞 1 つからなる場合
数值型	0	0	大小関係表現語を伴う属性値が2つの 場合のみ
数值型	0	×	大小関係表現語を伴わない属性値 それ以外

図1 論理和と論理積に関する制限

6. まとめ

以上のような言い替えや複数化があることがわかったので、それに対応する処理系を用意することで、これらの文は処理可能となる。言い替えに関しては、使用するデータベースに固有の属性の同義語を定義し、処理としては、言い替えられる前の形(矢印の左)に替える処理を設けることで可能となる。また、検索条件の複数化に関しては、使用するデータベース固有の<属性>と<データ指示語>を指示し、検索条件の動詞化においては、使用データベースの<属性>と<データ指示語>の関係から考えられる、動詞(V)、疑問詞(WH1、WH2)を設定し、解析部に定義することで、この形式化はこのデータベース以外にも広く使用が可能と思われる。

また紙面の都合上、基本検索条件が複数の場合に生じる曖昧さについて述べられなかったが、曖昧さに対する処理も必要となってくる。

7. 今後の課題

この形式化の汎用性を確かめるためにも、ほかのデータベースを対象とした検索システムに使ってみることが必要となる。

また、動詞化については、属性が1つのものしか見ていないが、属性が複数で、論理 積を構成している場合、どういった文型が可能かの考察が必要である。

さらに、この研究では扱うことがなかった、属性値を"O"または"1"と表現できるような属性(CDを新品で購入したか、中古で購入したか)の場合、どのような言語表現があり、この研究とどのような関係にあるかといった視点での研究も必要となる。

[注]

- 1) 前述の(27)((57)として再び載せる)のように、属性値を「と」で結んだ場合、 論理和・論理積両方の可能性があるので、曖昧である。このため曖昧性解消のため にベアであることを表す「一緒に、ベアの、二人で、組んで」等をつけた(58)のよ うな文がある。
 - (57)作曲家が鈴木慶一と渚十吾の曲
 - (58)鈴木慶一と渚十吾が二人で作曲した曲

【総合文献】

加藤恒昭・中川優(1984) 日本語DB検索システムQuestにおける意味解析 自然言語処理vol.46 no.5

絹川博之(1986) 表階層モデルに基づく自然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

草薙裕(1985) 『自然言語とコンピュータ言語』 講談社

長谷川守寿(1993) 間い合わせ文の構造の形式化 計量国語学vol.19 no.3

藤崎哲之助・間下浩之・諸橋正幸・渋谷政昭・鷹尾洋一(1979) 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

謝辞

この論文をまとめるにあたり、有益なご助言を下さった、筑波大学文芸・言語学系の 草薙裕教授に御礼を申し上げます。

Paraphrase and Pluralization of Query Language

Morihisa HASEGAWA

In this paper, the manner in which basic query units are used and the forms in which they are expressed is analyzed from the point of view of paraphrase and cases in which more than one such unit is used at the same time.

Paraphrase is classified into four types,

- 1) paraphrase of a subject using synonymy.
- 2) paraphrase of a predicate using an auxiliary and an interrogative,
- 3) paraphrase of a whole sentence, changing the word order,
- 4) paraphrase using a verb whose form is determined by the relation between the data and its attributes.

Corresponding expressions are analyzed in each case. When multiple basic query units are used, two relations emerge: disjunction(OR) and conjunction(AND). In each of these cases, basic sentence styles are determined and the paraphrases of these are considered. Limitations which arise when the same attributes are used disjunctively or conjunctively are analyzed.