論文

ファジィデータベース検索のための 自然言語インタフェースの処理能力†

前田 博*1 吉田 和明*2

ファジィ表現を許容した自然言語によるファジィデータベース検索について、筆者らは、平仮名検索と最適モデルを組み合わせた自然言語処理手法を提案した。しかし、その手法がどの範囲の自然言語問い合わせ文を処理可能なのか明確でなかった。本論文では、自然言語問い合わせ文をいくつかのクラスに分類し、この手法及び、それに拡張処理を付加することによって処理可能な問い合わせ文のクラスを明らかにする。同時に、未知語を含む文や省略文に対してもこれらの手法が有効であり、頑健で柔軟な自然言語処理手法であることを示す。

1. はじめに

データベースシステムはコンピュータの記憶機能を積極的に活用して大量のデータを管理する情報処理技術として発展してきた。今日のDSS, SIS, CADといった高度な情報システムは、あいまい、複雑、大規模といった形容詞が冠される問題解決の支援を目指しており、そこでのデータベースは情報システムの一部品として位置付けられている。さらに、それらの情報システムは人間の問題解決すなわち意思決定を支援するための道具であるとの認識から、人間に対する情報システムの優しさや柔軟さをいかに実現していくかが大きな課題となっている。これに対する一つのアプローチとして、人間特有の発見的能力を活用するための対話型システムの導入¹⁾がある。

筆者らはその考えをさらに進め、自然言語対話

方式が将来的に有効であるとの視座から、その一環として自然言語問い合わせによるファジィデータベース検索を可能とするための自然言語インタフェースを開発してきた²⁾。そこでは、自然言語の持つ3種類のあいまいさ、すなわち、(1) 検索条件のあいまいさ、(2) 問い合わせ文の解釈のあいまいさ、(3) 未知語や語の多義性に関するあいまいさ、をいかに処理するかが大きな問題であり、平仮名検索と最適モデルを組み合わせた自然言語処理手法を提案した。しかし、その手法がどの範囲の自然言語問い合わせ文を処理可能なのか明確でないという課題があった。

本論文では、自然言語問い合わせ文をいくつかのクラスに分類し、本手法及びそれを拡張することによって処理可能な問い合わせ文のクラスを明らかにする。同時に、未知語を含む文や省略文に対してもこれらの手法が有効であり、頑健性を持った柔軟な自然言語処理手法であることを示す。

2. 自然言語問い合わせ文とその処理手法²⁾

2.1 問い合わせ文の制約

システムに入力される自然言語は、漢字仮名混じりの平易なデータベースへの問い合わせ文であることを前提とし、単文に対して以下の検索情報

† Ability of a Natural Language Interface for Fuzzy Database Retrieval

Hiroshi MAEDA and Kazuaki YOSHIDA

*1 九州工業大学 工学部情報工学教室

Department of Computer Science, Faculty of Engineering, Kyushu Institute of Technology

*2 九州工業大学 工学部電気工学専攻 博士前期課程2年

Postgraduate of Computer Science Department, Faculty of Engineering, Kyushu Institute of Technology

を持つものとする。単文を接続語により結合した複合問い合わせ文は、各単文に分解されて処理されるので、以後の記述は特に断らない限り単文に関するものである。

- 1) 検索対象 (E): 検索要求によって特徴づけられ、最終的に出力される対象。
- 2) 検索属性 (T): 検索対象を特徴づける種々の属性。
- 3) 比較演算子 (R): 検索対象間の関係。
- 4) 比較の対象 (O): 比較演算子における比較の対象。
- 5) ファジィ集合 (F): 検索属性上で使われるあいまいな語。
- 6) ファジィ修飾詞 (M): 比較演算子、ファジィ集合を修飾するあいまいな語。
- 7) 否定 (N): 否定を表す語。
- 8) 並列語 (LW): 検索属性、比較の対象を複数接続する語。

自然言語問い合わせ文からデータベース検索に必要な情報だけを抽出し、それらを形式化した中間表現を形式言語と呼ぶことにする。これは、8個の検索情報のうち、事前に指定される検索対象 E を除く7項目の組 $Q(T, R, O, M, F, N, LW)$ で定義される。問い合わせ文中に、該当する情報が存在すればその記号を、存在しなければ記号 ϕ を代入するとすれば、形式言語と自然言語問い合わせ文は、例えば

「給料が極めて高い。」 $\rightarrow Q(T, \phi, \phi, M, F, \phi, \phi)$
(T : 給料, M : 極めて, F : 高い)

のように対応する。

2.2 自然言語処理

自然言語処理は以下の5種類の辞書を用いた5つのステップから成る。

1) 平仮名検索辞書

問い合わせ文中の文字列が形式言語のどの項目に該当するかを、文字列に含まれる平仮名の文法的特徴を記述することによって推定できるようにした辞書。

2) ファジィ集合(比較演算子)辞書

問い合わせ文中で、検索情報になり得るあいまいな語を50音順に登録した辞書。

3) ファジィ修飾詞辞書

ファジィ集合辞書内の語を修飾するあいまいな語に登録した辞書。

4) 接続語辞書

接続語として用いられる語を連言、選言、含意に分けて登録した辞書。

5) 不要語辞書

データベースの対象とする世界に依存して問い合わせ文中に現れる語で、検索情報となり得ない人称代名詞などの一般語や検索対象を表現する語に登録した辞書。

Step 1) 字種切り。入力文を字種の変わり目で切断する。

Step 2) 複合文の単文分解。複合問い合わせ文は、単一問い合わせ文間の接続関係(連言、選言、含意)を調べ、複数の単一問い合わせ文に分解する。

Step 3) 不要語除去。各単文について、不要語を検索し、記号 D に置き換える。

Step 4) 平仮名検索。平仮名文字列の先頭の文字から終端まで順に平仮名検索辞書を参照し、辞書に記述された処理を行う。この時、各平仮名文字に対して、可能な全ての場合について処理がなされる。辞書のどの条件にも適合しない場合は、未知語 U として扱われる。

Step 5) 形式言語の確定。複数の形式言語候補に対して、最適モデルを適用して1つの形式言語を確定する。

ここで、後述の理解のために、step 4)の平仮名検索辞書とその検索について、文中に良く使われる「が」を例に述べておく。

平仮名検索辞書では、例えば

① $(Bz=x)\{T \leftarrow BX\} < (bX=LW)\{T \leftarrow LW\} > [Y \leftarrow Y-y]$

② $(Y \in DIC) \{M \leftarrow J\} [Y \leftarrow Y-J]$

のような記述が見られる。この意味は、

①「が」の直前の1文字が非平仮名ならば、直前の非平仮名列を検索属性とだけ、前の非平仮名列が接続語で結合されていれば、それも検索属性とし、この処理を全ての接続語について繰り返し、平仮名列から「が」を除き、平仮名列を更新せよ。

②「が」を含んだ検索対象平仮名列の全てあるいは一部がファジィ修飾詞辞書にあれば、それをファジィ修飾詞とし、平仮名列を更新せよ。

のようである。このように平仮名検索辞書は、平仮名の付随した語について、平仮名の文法的意味や他の辞書の検索から推定される形式言語中の項目との関係、及び文字列の処理の仕方について記述したものである。

2.3 最適モデル

前向き探索のみを行う平仮名検索は、問い合わせ文に対して、複数の形式言語候補を生じさせる。この中から、最適な形式言語を一意に決定するのが0-1計画法を用いた最適モデルである。最適モデルは、平易な問い合わせ文に対して、形式言語の各項目の存在条件や項目間の位置関係（語順）を制約条件とし、問い合わせ文の持つ情報を最大にするという基準で定式化される。

問い合わせ文から、 n 個の文字列が検索情報として切り出されたとし、 i 番目の文字列を S_i とする。形式言語の7項目から接続語情報を除く（接続された語の1つの役割が分かれば十分である）6項目、及び新たに未知項目 U を加えた7項目、 T, R, O, M, F, N, U 、について、制約条件は以下のように表される。

- 1) 文字列 S_i ($i=1 \sim n$) には必ず1項目が当てはまる。
- 2) 未知項目 U 以外の項目には高々1文字列しか当てはまらない。
- 3) 検索項目 T は必ず1つ存在する。
- 4) 比較演算子 R と比較の対象 O は必ず同時に存在し、 O が R より前に存在する。

5) 比較演算子 R とファジィ集合 F は同時に存在しない。

6) ファジィ修飾詞 M は、 R あるいは F の前に存在する。

これらの制約の下で、問い合わせ文から最大の情報を引き出すという基準を導入する。これから、未知項目の個数を最小にするという最適基準の最適問題が構成される。

最適モデルによって選択された形式言語が妥当な意味を持つ時、最適モデルは解を持つということにする。最適問題の解が存在しても妥当な意味を持たない場合は、例えば、 $Q(T, \phi, O, \phi, \phi, \phi, \phi)$ や $Q(T, \phi, \phi, \phi, \phi, N, \phi)$ などがある。これらは、不完全な問い合わせ文となっている。従って、妥当な意味を持つ形式言語は、少なくとも T と F または R のいずれかを含まなければならない。さて、最適モデルの解について次の命題が成立する。

[命題]

問い合わせ文から切り出された文字列が、未知語でなく、少なくとも1つの形式言語中の項目に該当するとき、最適モデルの解は一意である。

[証明]

平仮名検索より、文字列が n (≤ 5) 個切り出されたとする。

- (1) $n=1$ の時、意味のある形式言語は存在しないので最適モデルは解を持たない。
- (2) $n=2$ の時、意味のある形式言語は $Q(T, \phi, \phi, \phi, F, \phi, \phi)$ のみであるので最適モデルの解は一意である。
- (3) $n=3$ の時、意味のある形式言語は3通り考えられる。

a) $Q(T, R, O, \phi, \phi, \phi, \phi)$

b) $Q(T, \phi, \phi, M, F, \phi, \phi)$

c) $Q(T, \phi, \phi, \phi, F, N, \phi)$

ここで、制約条件5)よりa)とb)、a)とc)は同時に成り立たない。また、文法的に一つの文字列が同時に M と N に当てはまらないことから、b)とc)は同時に成り立たない。よって、最適モデルの解は一意である。

(4) $n=4$ の時, 意味のある形式言語は 3 通り考えられる.

a) $Q(T, R, O, M, \phi, \phi, \phi)$

b) $Q(T, R, O, \phi, \phi, N, \phi)$

c) $Q(T, \phi, \phi, M, F, N, \phi)$

ここで, 制約条件 5), 6) より, a) と c), b) と c) は同時に成り立たない. また 3) と同様の文法的理由から a) と b) は同時に成り立たない. よって, 最適モデルの解は一意である.

(5) $n=5$ の時, 意味のある形式言語は $Q(T, R, O, M, \phi, N, \phi)$ のみであるので, 最適モデルの解は一意である.

[証明終]

3. 自然言語インタフェースの処理能力

平仮名検索と最適モデルから成る自然言語処理手法 (これを従来法と呼ぶ) がどの程度の処理能力を持つかを考察するために, 問い合わせ文を以下のクラスに分類する.

- (1) 完全な問い合わせ文
- (2) 不完全な問い合わせ文
 - a. “未知語を含む”として認識される文
 - b. 主格が省略された文
 - b.1. 解釈が一意である文
 - b.2. 解釈が複数できる文
 - c. 述語の活用語尾が省略された文
 - d. 述語が省略された文
 - e. 名詞句
 - f. 有効情報が認識できない文
 - g. 無効な文

従来法が各クラスに対して処理可能かどうか検討し, 不可能なものについては, 従来法を拡張した処理手法を提示する.

(1) 完全な問い合わせ文

完全な問い合わせ文とは, 従来法によって, 問い合わせ文からすべての有効な情報が取り出せる文のことである.

(2) 不完全な問い合わせ文

a. “未知語を含む”として認識される文

未知語を含む文とは, 平仮名検索によって, 問

い合わせ文中に未知項目に該当する文字列が存在すると認識された文のことである. 平仮名を部分的に含む単語が使われる場合に生じる. 例として, 「休みが多い会社。」を考えてみる.

平仮名検索では, 「休み」のように非平仮名文字列に平仮名が付随して一つの語を形成する場合, 平仮名「み」の情報から非平仮名文字列の役割が定まらず, 未知語 U と認識される. また「い」の検索によって, 「多い」がファジィ集合辞書中に存在すると認識され,

「休み」→ U , 「多い」→ R または F ,

となる. 最適モデルにおける制約条件「検索属性は必ず 1 つ存在する。」より, 「休み」は検索属性となる. よって, 従来法によって形式言語は以下のように正しく確定する.

$Q(T, \phi, \phi, \phi, F, \phi, \phi)$,

(T : 休み, F : 多い)

b. 主格が省略された文

これには 2 つの場合がある.

b.1) 解釈が一意である文

例として, 問い合わせ文「A 社の休暇より多い。」を考えてみる.

平仮名検索では, 「の」が連体修飾を表す格助詞, 「より」が比較を表す格助詞と見なされ,

「A 社の休暇」→ O , 「多い」→ R または F ,

となる. ここでは, 検索属性 (T) または未知語 (U) に該当する文字列は検出されない. この場合, 最適モデルの制約条件「検索属性 (T) は, 必ず 1 つ存在する」に反し, 従来法では処理できない.

そこで, この問い合わせ文が有効であるためには, どこかに検索属性情報が存在するはずであるから, 切り出された文字列について検索属性情報を検索する. これは, 文字列「A 社の休暇」と検索属性に関するシソーラス辞書とのマッチング処理³⁾によってなされる. ここでは, 比較の対象 (O) 中の「休暇」が検索属性情報と認識され, 「休暇」を除いた「A 社」が比較の対象に置き換えられる. よって, もとの文字列を S_1 : 休暇, S_2 : A 社に分

割し、最適モデルを適用すると形式言語は以下の
ように正しく確定する。

$$Q(T, R, O, \phi, \phi, \phi, \phi),$$

(T : 休暇, R : 多い, O : A社)

b.2) 解釈が複数できる文

これは、上記の特殊な場合で、「AのB」という
形式でBが未知語の場合に、「の」解釈により複数
の解釈が可能となる。例として、b.1)の例と同一
内容である問い合わせ文「A社の休みより多い。」
を考えてみる。

従来法では、2通りの解釈ができる。

- (1) 「の」は連体修飾を表す格助詞、「み」はa.と
同様に未知語となる。「より」は前の語が未知
語であるから比較の格助詞とは成らず、ファジ
ィ修飾詞と認識される。従って、

「A社の休み」→ U , 「より」→ M ,

「多い」→ R または F

形式言語: $Q(T, \phi, \phi, M, F, \phi, \phi)$,

(T : A社の休み, M : より, F : 多い)

- (2) 「の」は主格を表す格助詞、他は(1)と同様
に見なされ、

「A社」→ T , 「休み」→ U , 「より」→ M ,

「多い」→ R または F

形式言語: $Q(T, R, O, M, \phi, \phi, \phi)$,

(T : A社, R : 多い, O : 休み, M : より)

いずれも最適モデルの解の条件を満たしている。
このような場合は、ユーザとの対話によって妥当
な解を選択するが、この例では、候補の中に妥当
なものがない。

このような場合には、制約条件3)を緩和し、新
たな制約条件「検索属性(T)、比較の対象(O)
の少なくともいずれか1つは存在する。」を導入す
る。すると、(2)は緩和された制約条件も満たして
いるが、(1)では、 T の存在に対して妥当な解が存
在しなかったため、新たに

「A社の休み」→ O , 「多い」→ R または F
が得られる。これは、b.1)と同じであり、検索属
性情報の探索により形式言語は正しく確定する。

$$Q(T, R, O, \phi, \phi, \phi, \phi)$$

$$(T: \text{休み}, R: \text{多い}, O: \text{A社})$$

c. 述語の活用語尾が省略された文

例として、問い合わせ文「給料が良好。」を考え
てみる。従来法では、平仮名語尾を持たない文字
列「良好」が検索されず、妥当な意味を持たない
形式言語

$$Q(T, \phi, \phi, \phi, \phi, \phi, \phi), \quad (T: \text{給料})$$

となり、最適モデルは解を持たない。「良好」につ
いては、平仮名検索の際、非平仮名文字列の存在
のみ認識される。この文が有効であるためには、
検索条件が問い合わせ文中に存在しなければなら
ないので、この中に検索条件が存在すると考えら
れる。

そこで、「良好」を未知語(U)とおいて、最適
モデルを適用すると、「良好」はファジィ集合と判
断される。「良好」がファジィ集合であるかどうか
をファジィ集合辞書で確認し、辞書中に存在すれ
ば、形式言語は以下のように正しく確定する。

$$Q(T, \phi, \phi, \phi, F, \phi, \phi), \quad (T: \text{給料}, F: \text{良好})$$

辞書中に存在しなければ、d.の文として処理する。

d. 述語が省略された文

述語が省略された文とは、問い合わせ文中に形
容詞や形容動詞が存在しない文である。例として、
問い合わせ文「勤務地が東京。」を考えてみる。

c.と同様の拡張処理によって、

$$Q(T, \phi, \phi, \phi, F, \phi, \phi), \quad (T: \text{勤務地}, F: \text{東京})$$

と判断されるが、「東京」はファジィ集合ではな
い。そこで、述部「である」が省略された可能性
があるとみなし、文「勤務地が東京である。」とし
て再び従来法を適用すると、平仮名検索より、

「勤務地」→ T , 「東京」→ O ,

「ある」→ F または R

となり、形式言語は以下のように正しく確定する。

$$Q(T, O, R, \phi, \phi, \phi, \phi)$$

(T : 勤務地, O : 東京, R : ある)

e. 名詞句

名詞句とは、名詞あるいは名詞の前に連体修飾語が付随するもので、例として、「A社より多い給料。」を考えてみる。

平仮名検索より、「より」が比較の格助詞、「い」がファジィ集合を表す形容詞の語尾、「給料」はc.と同様の拡張処理によって未知語となる。従って、

「A社」→ O , 「多い」→ F または R ,

「給料」→ U

となる。最適モデルにより、形式言語は以下の様に正しく確定するため、従来法で処理可能である。

$Q(T, O, R, \phi, \phi, \phi, \phi)$

(T : 給料, O : A社, R : 多い)

f. 有効情報が認識できない文

この問い合わせ文は、これまで述べてきた方法を用いても有効情報が取り出せないものである。例として、問い合わせ文「給料が約20万円より多い会社。」を考えてみる。

平仮名検索より、

「給料」→ T , 「約20万円」→ O ,

「多い」→ F または R

となり、この文は、

$Q(T, O, R, \phi, \phi, \phi, \phi)$

(T : 給料, O : 20万円, R : 多い)

と解釈される。文字列「約20万円」が比較の対象(O) (数値データ)としてみなされ、ファジィ修飾詞“約”が認識できない。よって、この文は「給料が20万円より多い。」と同義とみなされる。このように、異なる情報の非平仮名文字列が連結した文字列は適切に分解できない。

g. 無効な文

無効な文とは、妥当な形式言語を持たない文のことである。例として、検索属性(T)を持たない問い合わせ文「A社より多い会社。」を考えてみる。

平仮名検索より、

「A社」→ O , 「多い」→ F または R

となり、検索属性も未知語も検出できないので、b.1)の方法を適用する。しかし、検索属性情報が検出できず、問い合わせ文中に検索属性が存在しないと判断される。このような文は、情報が不足している問い合わせ文として、ユーザとの対話により不足部分が再入力される。

以上のようなユーザとの対話も含めた拡張処理と従来法とを統合した自然言語処理アルゴリズムをFig.1に示す。図中の記号は問い合わせ文の分類記号である。①～⑤は従来法であり、その後、⑥で解の存在を確かめ、解が存在しない時は⑥aでb.1)のクラスに属するかどうか調べ、属しなければg.のクラスとして再入力となる。解が存在するときは、⑦で解の妥当性を確かめる。解が妥当でなければ、⑦aでクラスc.d.とクラスg.が弁別され、⑦bで更にc.とd.が弁別される。解が妥当である時は⑧でユーザの承認を求める。解が承認されなければ、⑧aで他に適合する平仮名検索を確かめ、全ての場合を尽くしていれば、クラスf.,b.2)として処理される。クラスf.の文は、有効情報が抽出されないためユーザの承認を得ることができず、最終的には再入力となる。

従来法に拡張処理を付加した自然言語処理手法は、クラスa.の未知語を含む文やクラスb.～d.の不完全ではあるが良く使われる省略文に対しても柔軟に対応できる。

4. おわりに

本論文では、データベースへの自然言語問い合わせに対して、平仮名検索と最適モデルを組み合わせた処理手法について、問い合わせ文をクラス化し、クラス別にその処理能力を評価した。その結果以下のことが明らかになった。

(1) 従来法に拡張処理を付加することによって、異なる情報の非平仮名文字列が連結した文字列を含む問い合わせ文については有効情報が欠落するが、それ以外の問い合わせ文については全て処理可能である。

(2) 無効な問い合わせ文の判定が可能である。

(3) 従来法に拡張処理を付加した自然言語処理手

法は、未知語を含む文や省略文についても処理可能であり、頑健で柔軟な自然言語処理手法といえる。

今後、本自然言語インタフェースを一部に持った柔軟な意思決定支援システムの構築を目指したい。

参考文献

- 1) 榎木, 中山, 中森: 新しいシステム工学入門, オーム社 (1988)
- 2) 前田, 石飛: ファジィデータベース検索のための最適モデルを用いた自然言語インタフェース, 日本ファジィ学会誌, 3-1, 108/119 (1991)
- 3) 前田, 村上: ファジィ理論の対話型ファジィ意思決定支援システムにおける自然言語処理への適用, 計測自動制御学会論文集, 26-9, 92/98 (1990)

(1991年7月4日 受付)

(1991年9月20日 再受付)

Fig.1 拡張された自然言語処理アルゴリズム

[問い合わせ先]

〒804 北九州市戸畑区仙水町1-1

九州工業大学 工学部 情報工学教室

前田 博

TEL: 093-871-1931 (内線) 434

FAX: 093-871-5835

—— 著 者 紹 介 ——

前田 博 (まえだ ひろし)

九州工業大学 工学部情報工学教室

1973年, 九州工業大学 工学部制御工学科卒業。同年4月いすゞ自動車入社。1977年九州工業大学工学部助手を経て, 1987年4月より同助教授。ファジィ意思決定理論, ファジィデータベース検索, 社会システムのモデリングなどの研究に従事。工学博士。日本ファジィ学会, 計測自動制御学会, 日本OR学会などの会員。

吉田 和明 (よしだ かずあき)

九州工業大学 工学部電気工学専攻 博士前期課程2年

1990年, 九州工業大学工学部情報工学科卒業, 現在に至る。柔軟かい意思決定支援システムの研究に従事, 計測自動制御学会会員。