日本語質問文からデータベース質問文への翻訳

2B－8

井谷 真　　中川 圭介

電気通信大学

1　概要

関係データベースを日本語文によって検索するシステムを試作したので、それについて報告する。

日本語による関係型データベース検索システムについてはすでに多くのものが開発されている [1][2]。

ここで報告するシステムは、質問文を構文解析し、データベースシステム Postgres の言語に翻訳した後、データベースを検索し結果を返すものである。そして、翻訳は関係型データベースの属性間の従属関係と属性の値域を利用する翻訳規則を使って行なっている。

2　翻訳の方法

処理系は、字句解析、構文解析、翻訳、データベース検索の順に仕事を進める。

構文解析の規則は助詞によって、名詞あるいは動詞を結合するという考えを基本に作られている。

また、関係型のデータベースはには、関係の形式と関係 (データ) がある。関係の形式は、属性の集合として定義され、属性には名前とその値域が定義されている。また、属性間には従属関係がある。例えば、属性 (科目名、学生名、評価、学科) で構成される関係を考えると、その従属関係は

科目名,学生名 → 評価

学生名 → 学科

である。このデータベースに対して、

「石山一郎の国語の成績を教えよ。」

という質問を行なうと

retrieve (評価)

where 学生名 = “石山一郎”

and 科目名 = “国語”

を生成する。

翻訳は、構文解析の結果として得られる木に対してボトムアップに行なわれる。例の文に対して構文解析から得られる木は

((石山一郎の (国語の成績)) を教えよ)。

となる。ここで、「石山一郎」は、「学生名」に属する属性値、「国語」は「科目名」に属する属性値」であるが、これを、以下のように表してみよう。

(属性値 1 の 属性名 2)

付加情報

((属性値 1:国語:科目名に属する) の

(属性名 2:評点))

そして、(属性値 の 属性名) については、属性値と属性名の間に何らかの従属性があるときに、規則 (属性名 ← 属性値の属性名) を適用して (属性名) にする。

(属性値 3 の 属性名 2)

((属性値 3:石山一郎:学生名) の

(属性名 2:評価:科目名 = “国語”))

これも、規則 (属性名 ← 属性値の属性名) を適用して (属性名) にする。

(属性名 2 を 指示 1)

((属性名 2:評価:学生名 = “石山一郎”

and 科目名 = “国語”)

を (指示 1:教えよ:retrieve))

規則 (句 ← 属性名を指示) を適用して質問が完成する。

3　翻訳規則

翻訳規則の一部を表 1に示してある。各々の節点のにおいて与えられた「入力」を、データベース条

Translation of queries from Japanese to a database query language

Makoto ITANI, Keisuke NAKGAWA

University of Electro-Communications

件が成り立つ時に翻訳を行ない、「出力」をその節点の種類とする。

出力	入力	データベース条件
文	句 結び	T
条件	属性名 が 属性値	属性名 ∋ 属性値
属性名	条件 の 属性名	属性名→→ 条件
属性名	属性名 (1) の 属性名 (2)	属性名 (1) →→ 属性名 (2)
句	属性名 を 指示	T
属性名	属性値 の 属性名	属性値→→ 属性名 または 属性名→→ 属性値

表 1: 翻訳規則の一部

注:データベース条件内では、「属性値」は属する属性名に、「条件」は、条件のかかっている属性名に置き換えて検査する。

これらの規則で翻訳を行なうが、これらは構文解析で得られる木の中で翻訳できない木を落すように作られている。2章の例では

((石山一郎 の 国語) の 成績)

という部分木が出力されることもあるが、これは、(石山一郎 の 国語) に対して、(属性値 の 属性値) という節点を翻訳する規則を定義しないことで正しくないとする。

データベースへの質問では、未知のもの (属性名) を出力するよう質問するのであり、既知のもの (属性値) を出力するようには質問しない。よって、助詞「の」の後に属性名が現れるような規則は定義しても、属性値が現れるような規則は定義しない。

また、「情報工学科の学生の成績を教えよ。」という質問では、((情報工学科 の 学生) の 成績) は翻訳できるが (情報工学科 の (学生 の 成績)) は翻訳できない。(情報工学科 の 成績) は (属性値 の 属性名) であるが、このように助詞「の」で結合する時はそれらの間に何らかの関連性がなければならない。この関連性をデータベースの従属性と考え、従属性がない時には変換を行なわないように規則を作成している。

上記の規則は、名詞と名詞を助詞で結合する構文規則に対応するもので、一般的な性質に対するものであるが、これ以外に、名詞と動詞に関する構文規則に対応するものなど、言葉を定義するものも追加されている。これらの規則は、個々のデータベースに対して別々に与えられなければならない。

4 処理系

処理系は最初に各々の処理で使われる表を作成してから使用される。表の作成は、字句解析規則、構文解析規則、翻訳規則、データベースの形式、データベース、といくつかの追加情報から生成される。

字句解析プログラムは、辞書と字句解析規則の表を使って前処理を行なう。辞書内では、各単語に、品詞などの文法情報が入っている。

構文解析も、文脈自由文法の構文規則表によって行なわれる。現在、約50の規則が使われている。構文解析にはアーリー法を用いている。

翻訳は、2章の方法によるもので、これも辞書と表を使って行なわれる。辞書には、各単語に対して属性名、属性値などのデータベースの形式などに関連した情報が入っている。

翻訳部は、第1正規形に対する質問を出力するが、これを第4正規形に対する質問に変換するフィルタプログラムで後処理を行なっている。

表を使用しているので翻訳規則などは入れ換えが可能な柔軟性のあるシステムである。現在、処理はワークステーション上で行なっているが十分な速度で翻訳が可能である。

5 謝辞

システムの開発に御協力して頂いた榊原崇氏、小向孝典氏、余語宣幸氏に深く感謝致します。

参考文献

[1] 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, 1979

[2] 牧之内顕文, 他: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, 1988