日本語文によるデータベース検索システム

5Aa－4

余語 宣幸* 中川 圭介* 井谷 真†

電気通信大学* 東京大学大学院経済学研究科†

1 はじめに

日本語による関係型データベース検索システムはすでにいくつか存在しているが、今回報告するシステムは日本語質問文を構文解析し、データベースの従属関係やデータ型、属性値の値域などの性質を元に作った翻訳規則を使ってデータベース言語の質問文に翻訳し、データベースの検索結果を返すというものである。このシステムは以前に試作したシステム [1] の機能を拡張し、問題点を改善したものである。

2 構文規則

構文規則の一部を表 1 に示してある。
構文解析は文脈自由文法の構文規則を表で持ってお

ID	左辺	右辺
1	文	述語@句点
2	数量名詞	数詞@助数辞
3	名詞句	名詞
4	名詞句	数量名詞
5	名詞句	名詞句@名詞接続助詞@名詞句
6	述語	名詞句@格助詞@述語
7	述語	名詞句@判定詞

表 1: 構文規則の一部

り、アーリー法を用いて解析を行なっている。
現在約 60 の規則が使われている。

3 翻訳

翻訳は構文木を帰りがけ順でたどって、節点ごとに子が持つ意味情報から翻訳後の文字列を生成し、新たな意味情報とともに親に渡していくという方法をとっている。

複数の構文木から正しい木を選び、質問文の翻訳結果を得るために、データベースの従属関係が役立つと考えている。

データベースに対する日本語の質問文は従属関係と深いつながりがあり、例えば成績のデータベースでは「学生名, 科目名→成績」という従属関係があり、この従属関係と次のような種類の日本語文が対応する。

- 助詞「の」: 国語の点数
- 判定詞: 点数が 80 点である学生
- 動詞: 石山が履修している科目
- 比較形容詞: 点数が 80 点より高い学生

このように助詞あるいは述語を中心として複数の属性が結合される。
現在のところ、従属関係の左辺と右辺、あるいは左辺同士の関係に対応する日本語文が存在すると考えている。

質問の性質 「国語 を 教えよ」は意味の上では
(属性値 を 指示)
であるが、データベースでは未知のものを質問するのでこういう形のものは登場しない。したがってこの様な質問は翻訳対象にならない。

属性間の関係 「情報工学科 の 学生 の 成績」は
1. ((情報工学科 の 学生) の 成績)
2. (情報工学科 の (学生 の 成績))
の 2 通りの構文木がでる。成績のデータベースには「学生名→学科名」「学生名, 科目名→成績」という従属関係があり、上で述べた様に従属関係から導かれる組合わせだけ「の」による結合ができるとし、2. は「学科名」と「成績」の間に関係が無い事により落としている。

属性値の値域 「点数 が 80 点である」という文は
(属性名 が 数値)
であるが、「80 点」の「点」は辞書を引くと単位であり、「点数」という属性に属していることが判るので、質問文として正しいかどうか判断できる。また、「国語の成績」の様に属性名が省略された場合にも補完する事ができる。

動詞 動詞と主語、目的語となる属性とそれに続く助詞の組合せは対象のデータベースごとに異なっている。例えば成績のデータベースでは「取る」に対して「学生名 が, 科目名 で, 点数 を」の様に属性と助詞の組を定義しておく。

ドメイン 比較の形容詞とその言葉が使われる属性は対象のデータベース固有のものである。例えば

Database Query System in Japanese
* Nobuyuki YOGO, Keisuke NAKGAWA
University of Electro-Communications
† Makoto ITANI
Faculty of Economics, University of Tokyo

成績のデータベースでは「高い、低い」は「点数」に対して使われるので、「点数: 高い, 低い」の様に形容詞と属性の組を定義しておく。

4 翻訳規則

構文木をたどって翻訳するにあたり、翻訳のための付加情報を節点ごとに持つわけだが、現在のシステムは以下の6つの情報を持っている。

品詞：「属性名」「属性値」などのデータベースでの役割を表す翻訳用の品詞。

クエリー：これまでに生成されたSQLの文字列。

標準化表現：助詞、動詞(現在形)、形容詞の単語そのものが入り、辞書を参照する時に利用する。

意味：対象の節点を代表する単語がデータベース上で対応する文字列。主に属性値、数値、演算子などで利用する。

属性：対象の節点を代表する単語が属する属性名が入り、従属関係や動詞のチェック、SQLの生成に使う。

使用済み属性：質問文で既に使われた属性名を記憶し、同じ属性名が2度使われていないか調べるために使う。

翻訳規則の一部を表2に示してある。表の「規則」の欄は表1で示した構文規則との対応を表している。翻訳規則は日本語の構文規則それぞれに対して、翻訳する場合だけが登録されていて、構文規則が同じものの中から「右辺」が一致する翻訳規則を探し、翻訳条件が満たされている時に翻訳して「左辺」の品詞を付加情報とともに親に渡す。

規則	左辺	右辺	翻訳条件
1	文	句@結び	なし
6	句	属性名@を@指示	なし
2	数	数詞@単位	なし
6	動詞	数@助詞@動詞	動詞辞書を検査
6	動詞	属性値@助詞@動詞	動詞辞書を検査
6	属性名	動詞@属性名	動詞辞書を検査
5	属性名	属性値@の@属性名	属性名→属性値 or 属性値→属性名

表 2: 翻訳規則の一部

翻訳条件には従属関係、データ型、動詞と助詞の対応、属性の比較、ドメイン検査などの条件を盛り込むことができる。

5 翻訳例

以下の様なスキーマをもつ成績のデータベース(下線はキー属性)を対象にして、このシステムによる実際の日本語の翻訳例を示す。

学生 (学生番号, 学生名)
科目 (科目番号, 科目名)
成績 (学生番号, 科目番号, 点数)

解析経過は省くが、以下のような結果が得られる。

- 「石山一郎の履修している科目を教えよ。」
```
SELECT 科目.科目名
WHERE 学生.学生名=' 石山一郎' AND 科目.科目番号=成績.科目番号 AND 成績.学生番号=学生.学生番号
```
- 「科目ごとの平均点を示せ。」
```
SELECT avg(成績.点数)
WHERE 科目.科目番号=成績.科目番号
GROUP BY 科目.科目番号
```
- 「国語と数学の点数が80点以上の学生を教えなさい。」
```
SELECT 学生.学生名 FROM 科目 Tmp0_0, 成績 Tmp0_1, 科目 Tmp1_0, 成績 Tmp1_1
WHERE Tmp0_0.科目名=' 国語' AND Tmp0_1.点数>=80 AND Tmp1_0.科目名=' 数学' AND Tmp1_1.点数>=80 AND Tmp0_1.科目番号=Tmp0_0.科目番号 AND 学生.学生番号=Tmp0_1.学生番号 AND Tmp1_1.科目番号=Tmp1_0.科目番号 AND 学生.学生番号=Tmp1_1.学生番号
```

※ Tmp?_? は範囲変数を表している。

6 処理系

字句解析には「茶筌」[2]を使った。連結規則は構文規則に合わせてシステム標準文法を少し変更したものを使っている。

構文解析については2節で述べた通りである。

翻訳部は3、4節で述べた通りであり、第1正規形に対する質問をSQLを独自に拡張した表現で出力している。そのまま正規のSQLに翻訳することが望ましいが、後に続く文まで解析しないと翻訳が一意に決められない場合があるので拡張表現を使用している。

SQL拡張部分の展開や範囲変数の割り当て、第4正規形に対する質問への変換は、後処理のフィルタプログラムで行なっている。

7 おわりに

現在、翻訳規則は約70で、通常の質問に加え、集約関数を使った質問、範囲変数を用いた同一クラス内の複数のインスタンスに対する質問、GROUP BY句を含む質問など、ほぼ全ての種類の検索に対応できている。

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。

参考文献

[1] 井谷真, 中川圭介: 日本語質問文からデータベース質問文への翻訳, 情報処理学会第52回全国大会講演論文集 (3), 3-33, 1996

[2] 松本裕治, 他: 日本語形態素解析システム「茶筌」version 1.5 使用説明書, Technical Report NAIST-IS-TR97007, 奈良先端科学技術大学院大学, 1997