动物数据库自然语言前端的设计与实现

张永奎　刘　红　罗栋梁
(山西大学计算机科学系)

[摘要] 哺乳动物数据库是从百科全书《哺乳动物》的描述文本中建立的数据库的集合。我们开发了关于这个哺乳动物数据库的自然语言前端，它允许用户用英语的自然语言形式同系统进行交互作用。本文叙述了这个自然语言前端的设计与实现。

1 引　言

目前，计算机的许多环境为用户提供的人工语言只能使用有限的菜单或肖像选择、编程命令和数据库询问语言。这些方法使用户易于在很多情况下做出有效的选择，但它们要求用户只对可见目标实施操作，而无法对具有复杂性质、内容频繁变动的目标操作，这无疑会增加用户的负担，并需要用户掌握如何用计算机语言表达自己的要求。而使用自然语言可以基本消除这类情况，使用更加方便、灵活。

自然语言主要用于：机器翻译、专家系统、自然语言分析及生成和自然语言前端(FRONT END)，其中自然语言前端为我们提供了一种用自然语言表达专家推理的人机界面，而且它使用户可以不考虑各个不同软件包之间的差别、内部结构和具体操作，仅仅注意所使用的人机界面，降低了用户使用时应具备的专业要求。

本系统就是一个对哺乳动物数据库进行处理的自然语言前端，主要用于对动物知识用自然语言提问，并用自然语言给出恰当回答。

2 动物数据库的简介

哺乳动物数据库是从百科全书《哺乳动物》[1]的描述文本中建立的数据库的集合[2]。百科全书《哺乳动物》描述了遍布于世界各地的426种哺乳动物，从最小的鼠类到最大的鲸。全书共426篇文章，每一篇包括一种动物的名称、分类、形态描述、分布地区、栖息地及行为特征、饮食习惯和繁殖情况等信息。

本系统目前可使用的动物数据库的内容全部由Prolog事实组成。下面是数据库中的部分Prolog事实的结构：

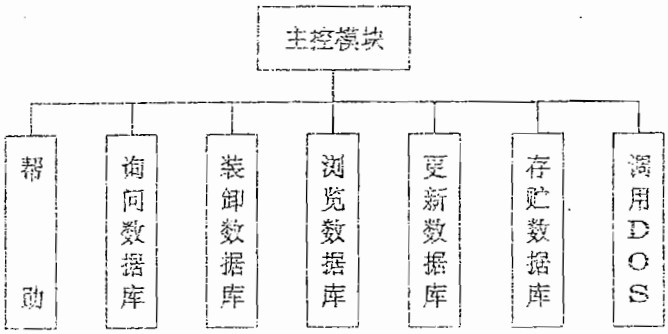
```
classification(No,Common name,Species,Genus,Family,Order)
habitat(No,Habitat,Distribution)
altitude(No,Low altitude,High altitude)
obj structure(
    mammal( general( shape(Gen1), size(Gen2), colour(Gen3) ),
        head( shape(Hea1), size(Hea2), colour(Hea3) ),
        snout( shape(Sno1), size(Sno2), colour(Sno3) ),
        )).
```

其中，classification事实描述了动物的编号(No)、俗名(Common name)、种(Species)、属(Genus)、科(Family)、目(Order)等内容；habitat描述了动物的生活环境(Habitat)、分布地区(Distribution)；

altitude记录了哺乳动物所生活的海拔高度；结构模板obj structure描述了动物的形态。

3 自然语言前端的总体设计

哺乳动物数据库自然语言前端总体结构框图如下所示：



各个模块功能分别介绍如下：

帮助模块包括 SD-PROLOG 系统文件[3]和本系统所涉及的哺乳动物数据库的相关信息，主要用于帮助用户熟悉系统环境和操作，了解用户可使用的提问句型及使用方法。

询问数据库模块为用户开设窗口，用于对数据库进行询问实现用自然语言提问和回答。

装卸数据库模块可实现对数据库文件进行有选择的装入和调出内存的功能。

浏览数据库模块向用户提供子菜单，使用户便于了解系统数据库中的具体内容。

更新数据库模块可使用户根据需要对数据库进行增加、删除和修改。

存贮数据库模块用于把所选择的数据库存贮到文件中，并将原数据库文件转换为.bak 文件。

调用 DOS 模块实现对 DOS 命令的调用。在本系统的实现过程中，以询问数据库这一自然语言前端功能的实现难度最大，本文的以下部分将着重阐述这一功能的实现。

4 自然语言的句法分析器

4.1 DCG 方法

自然语言数据库前端中有两个主要的成分；即：句法分析器和回答生成器。一个句子的语法分析依赖于句子的含义，因而语法必须和系统知识结构一致，并通过句法分析器调用数据词典，使回答生成器能根据输入的句子回答，同时与句法分析器共享信息。本系统采用定义子句语法（Definite Clause Grammar, 简称 DCG）模拟人类语法。

4.2 可分析的句子类型

为了避免语言过于庞大和复杂，必须对系统可识别的句子结构做出限制，预知哪些类型的问题能满足大多数人的需要。本系统将处理四类句型：陈述句、一般疑问句、特殊疑问句、祈使句，基本覆盖了英语中的常用提问句型。

在确定了系统可识别的句子类型后，也就确定了系统所需要的词汇表。本系统建立的词汇表中包括了五类单词：名词和动词定义系统识别的对象和动作，量词、形容词和代词决定用户所需要信息的性质，助动词表明这个句子是问句还是陈述句。这些词汇均作为 PROLOG 事实存贮在数据库中。

当我们要扩充系统可识别的句型时，应首先将出现的新词汇追加到数据库中。

4.3 对于句型的语法分析

系统需要的词汇和句子结构确定后，就能为每个正确的句型编写语法规则，采用的技术则依赖于语言的要求。如果语法需要在句子的各部分间传递信息，就必须使用上下文相关语法分析技术。

这里，DCG 假定了语法定义的输入是切分后的词表，这种定义能通过一些可识别的词表来检测表头元素的合法性。当语法检测词表是否符合一个正确句子的定义时，它也对句子构造了一个恰当的数据库询问。为了构造这个询问，句法分析器必须对作为 DCG 附加参数的句子的各部分间进行信息传递，附加参数提供的信息包括：句子所确定的变量，当前被例示的参数，句子的歧义性等。

4.4 语法的顶层

语法的顶层用来检验句子是问句还是陈述句，由问句形成的询问以带有 test 的结构返回，而由陈述句形成的询问以带有 assert 的结构返回，若系统无法从语法上分析句子，那么就发出一条错误信息，并且程序用"!"和 fail 退出。

```
semantic(test(Sent)) -->
        aux, sentence(Sent),!.
semantic(assert(Sent)) -->
        sentence(Sent),!.
semantic( ) -->
        {write('Cannot understand your question.')},nl,!,{fail} .
```

根据句型的不同结构，谓词 sentence 将可能遇到的句子分成了六种结构，而动词短语必须向可能遇到的任何介词短语传递歧义性信息

4.5 句子的语义分析

在语法分析时只要几个句子之间的语法结构相同，将不区分句子的具体含义，按照同一方式来处理。但在语义分析时，对于结构相同，含义不同的句子会采取不同的分析方法，形成不同的语义信息，并传递给生成器。nphrase 的定义把名词词组解释为一个或多个目标表以实现对句子的语义分析。下面是用于处理查找动物"目"的名词短语的一个语义分析规则。

```
nphrase([order(ORD)],ORD,args(X,Y,Z,A),not amb) -->
        mamname(MAMMAL),
        prep(P),
        [N], { is noun(N,order) } ,
        mams order(ORD).
```

5 生成自然语言回答

通常，一个用自然语言理解的系统也需要用自然语言回答。在自然语言理解中遇到的问题在生成自然语言的过程中也会碰到。当分析一个自然语言句子时，我们必须假定用户可能想问的问题的类型，而执行由问题生成询问的回答时，又必须假设合适的回答类型。最重要的是，我们构造回答时，必须试图让回答能反映用户提问的意图。

5.1 问题的求解及数据库访问

问题求解时，必须首先决定句法分析器传送来的语句形式是 assert 还是 test，然后才能分别处理。当系统遇到一个陈述句，它将可能做两件事之一：如果语句不存在就向数据库追加此新信息，或者把陈述句当作问题来处理。当追加信息时，系统要保证追加完整信息且不会与已存在的信息发生冲突。因此，插入信息时，第一步是把陈述句当作问题，以此检测信息是否已存在或信息之间是否会发生冲突。

下面给出语义解释含有量词句的方法的一个例子：

```
?-semantic(S,[tell,me,the,name,of,mammal,m33],[ ]).
S = assert([num(m33)]).
```

在这些目标执行前，它们必须由中间形式转化为一个可执行子句。转换工作将由 execute 谓词完成。如果用户要想向数据库追加数据扩充系统，就应该在程序中同时扩充 execute 谓词。

```
execute(test(X)):-
        construct(X).
execute(assert(X)):-        construct(X).
execute( ).
```

execute 把目标传送给谓词 construct，construct 可以识别目标可能采用的两种形式。对于歧义句将构造两个子句，一个符合 --> 左边的目标表，而另一个符合 --> 右边的目标表。对于非歧义句 construct 只产生一个子句。

```
construct( for each X: (P --> Q)):-
        make clause fe(P,C1),
        make clause fe(Q,C2),
        check for negative(C1,P),!,
        for each(C1,C2,P,Q).
construct([H   T]):-
        make clause(T,H,Clause),
        check for negative(Clause,[H   T]),
        gen(Clause,[H   T]).
```

谓词 make clause fe 和 make clause 递归地把每个目标加到可执行子句中。

```
make clause fe([H   T],Clause):-
        make clause(T,H,Clause).
make clause([ ],Temp,Temp).
make clause([H   T],Temp,X):-
        make clause(T,(Temp,H),X).
```

construct 同时也控制这些子句的执行，将规则联结传送给谓词 gen，然后就可以在循环中找到数据库的全部答案。

```
gen(Clause,L):-
        not(call(Clause)),
        ans(Answer,L,[ ]), fail.
gen( , ).
```

5.2 自然语言回答的生成过程

这个系统不仅能对可直接回答的问题生成回答，而且可以对不能直接找到答案的问题生成回答。目标的一个例示失败了，系统也能这样回答：

Question: Which mammal lives on altitude 50000 ?

Answer : No mammal lives on altitude 50000 .

若目标的一个例示成功了，系统则会回答：

Question: Which mammals live in area China ?

Answer: The mammals that live in area China are :

rhesus macaque	north african jird	harvest mouse
raccoon dog	asiatic black bear	hog badger
indian muntjac	barking deer	common goral
bharal	blue sheep	

11 solutions.

只要用户的问题是用 DCG 分析的，系统就可以用 DCG 生成回答，这恰好是用 semantic 执行语法语义分析的逆过程。语言生成器不是接受构成句子的字符，而是接受构成询问的目标表，答案的每一部分都将作为系统追加给回答的其余部分的单词表而返回，所有 DCG 定义的可能的目标结合方式都要与 semantic 定义的可能的结合方式相一致。

在语义分析中，系统把大量可能的句子形式转换成少量询问形式，同时还定义了同义词，这就使可能有的问题数量比回答所针对的问题数要大得多。下面是八种回答生成中的一个例子，它用于生成满足给定条件的所有动物名字的回答。

```
ans(Ans) -->
    num(MAM),
    { string_atom(MAMMAL,MAM) },
    { append([The,name,of,mammal,MAM,is,':'],[ ],Ans1)},
    { output(Ans1) },nl,
    { bagof(NAME,db("name","of","mammal",NAME,MAMMAL),L)},
    { (write list1(1,L)) },
    { length(L,N) },
    { write solutions(N) },nl .
```

由以上程序可见，gen 每次调用子句，表中的变量就被例示为与子句中变量相同的值，并将此表传递给 ans，把找到的回答当作一个词表返回，同时传递给一个简单的输出程序。

6 实验系统的实现

6.1 实验系统的软硬件环境

本系统的实现是在 IBM PC / AT 及其兼容机上完成的。

本系统的软件环境是 SD-PROLOG[3]。它可以利用 IBM PC 机及其兼容机上的硬件、窗口和图形能力编出图文并茂、独树一帜的 SD-PROLOG 程序。同时 SD-PROLOG 系统本身具有虚拟存贮功能，允许对大量数据进行操作。

6.2 目前的规模

这个系统所使用的数据库容量为 110K，程序容量为 39K。

目前已经能解决的问题的类型有以下 13 种：

1. Give me the name of mammal m222 .
2. Please tell me the mammals of order Marsupialia .

3. I need the mammals of family Solenodontidae .
4. Find out all mammals in distribution Australia .
5. Could you tell me the habitat of mammal rats ?
6. Where is the distribution of mammal foxes ?
7. What is the habitat of mammal wolves ?
8. Which mammal lives in area Australia ?
9. Is there any mammal in area Asia ?
10. Which mammals are there in region India ?
11. How many mammals live in the continent America ?
12. How many mammals live on altitude 10000 ?
13. What is the living altitude about mammal gibbons ?

当系统要扩充时，只须对词汇表、名词短语和动词短语的规则分别加以扩充。

7 结束语

目前，本系统已具有一定的实用性。它除了完成了人们通常所做的语法分析工作外，在语义分析和自然语言回答生成方面也做了一些研究探讨工作。哺乳动物数据库咨询系统的实现将为动物学方面的各类用户提供参考工具，带来极大的便利。下一步要解决的问题主要是扩大对歧义句的分析功能。

参考文献

〔1〕 L Boitani and S Bartoli:The Macdonald Encyclopedia of Mammals, Macdonald, London,1983

〔2〕 Yongkui Zhang and J R Cowie, Building a Mammalsbase from an Encyclopedia, Technical Report TR82, University of Stirling, 1992

〔3〕 The SD-Prolog Programmer's Reference Manual, Pembroke House, Camberly,Surrey, 1987

THE DESIGN AND IMPLEMENTATION OF THE NATURAL LANGUAGE FRONT END TO MAMMALSBASE

Zhang Yongkui, Liu Hong, and Luo Dongliang
(Department of Computer Science, Shanxi University)

Abstract

Mammalsbase is a collection of databases built from descriptions from a mammals encyclopedia. A natural language front end to the Mammalsbase has been developed, which enables users to interact with the system using natural English language. In this paper, the design and implementation of the natural language front end to Mammalsbase are described.