第34卷 第10期 计 算 机 工 程 2008年5月
Vol.34 No.10 Computer Engineering May 2008
·博士论文· 文章编号：1000—3428(2008)10—0025—03 文献标识码：A 中图分类号：TP391

受限领域问答系统的中文问句分析研究

陈 康[1]，樊孝忠[1]，刘 杰[1]，余正涛[2]
(1. 北京理工大学计算机科学技术学院，北京 100081；2. 昆明理工大学信息工程与自动化学院，昆明 650051)

摘 要：对用户所提问句的理解是受限领域问答系统实现的关键，该文提出一种基于本体和问句句型模板规则的中文问句分析方法，研究如何使用问句语义表征来表示问句分析的结果，将该方法应用于某受限领域问答系统中。实验结果表明，使用该方法进行中文问句分析，准确率达90%以上，可以在实际的问答系统中使用该方法。

关键词：本体；受限领域问答系统；中文问句分析；问句语义表征

Study on Chinese Question Parsing of Restricted-domain Question Answering System

CHEN Kang[1], FAN Xiao-zhong[1], LIU Jie[1], YU Zheng-tao[2]
(1. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081;
2. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051)

【Abstract】 Question parsing is one of the most important steps in the implementation of restricted-domain question answering system. This paper puts forward a new method to parse Chinese questions based on ontology and question semantic model, and explores how to use question semantic representation to represent the results of question parsing. A restricted-domain question answering system adopting this question parsing method is developed. Experimental results show that the precision of the method can exceed 90%, and it is feasible to use the method to develop restricted-domain question answering system.

【Key words】 ontology; restricted-domain question answering system; Chinese question parsing; question semantic representation

1 概述

问答系统(Question Answering, QA)是目前自然语言处理领域中的研究热点，它既能让用户使用自然语言提问，又能为用户返回简洁、准确的答案。按照问答的范围可以将QA分为受限领域QA和通用领域QA。在国际文本检索会议(Text Retrieval Conference, TREC)的支持下，面向大规模文本的通用领域QA取得了很大的进展[1]；在受限领域QA方面，英语、日语和德语的QA系统已获得了应用。在国内，许多科研机构都投入了相当大的精力开展汉语问答系统的研究，出现了一批汉语问答系统，如中科院计算所的红楼梦人物关系问答系统[2]、清华大学的校园导航系统[3]、哈工大的基于常问问题集的问答系统[4]、北京理工大学的银行领域问答系统[5]。

起源于哲学的本体论(ontology)近年来受到信息领域的广泛关注，并得到广泛应用[6]。本文提出一种基于本体和问句句型模板规则匹配的中文问句分析方法，该方法使用问句语义表征表示问句分析的结果，成功应用于某医院问答系统中。

2 系统框架

2.1 系统设计思想

本文所述基于本体的受限领域 QA 系统，其结构如图 1 所示。问句分析是一个至关重要的模块，其结果对后续处理过程有很大的影响，系统采用问句语义表征表示问句分析的结果。答案生成模块采用多策略方式进行实现：对于常见的问题采用 FAQ 库和问句匹配技术实现问答；对于领域常识知识的问答通过领域本体知识库和逻辑推理技术实现；其他问题通过信息检索技术直接从领域文本中提取答案。基于 FAQ 库的问句匹配技术主要采用计算问句语义相似度的方法进行处理；基于领域本体的逻辑推理技术主要根据本体关系进行推理；基于领域文本的答案提取技术主要依据问句语义表征与领域文本标记进行匹配，并适当地辅以信息检索技术。

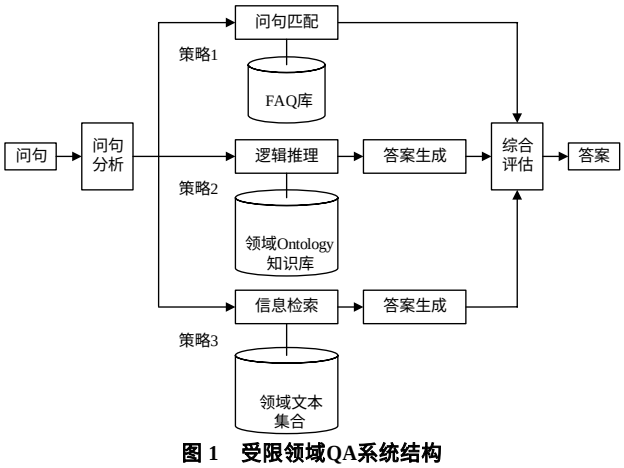

图 1 受限领域QA系统结构

基金项目：国家自然科学基金资助项目(60663004)；教育部高等学校博士学科点专项科研基金资助项目(20050007023)
作者简介：陈 康(1982－)，男，博士研究生，主研方向：自然语言处理，自动问答系统；樊孝忠，教授、博士生导师；刘 杰，博士研究生；余正涛，教授
收稿日期：2007-05-25 **E-mail：**chenkang@bit.edu.cn

2.2 知识资源

在系统研究过程中，使用了如下知识资源：

(1)通用词库：系统使用了中科院分词程序 ICTCLAS 的通用词库。

(2)领域词库：包括类词库、实例词库、属性词库，分别存放本体知识库中的类名、实例名、属性名。

(3)知网：是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。系统利用知网辅助问句语义块的识别。

(4)客气词库：用于过滤问句中的客气词。

(5)语义块规则库：系统采用基于规则的方法进行语义块识别，库中存放实体块、属性块等语义块的识别规则。

(6)问句句型模板规则库：存放根据语义块之间的搭配关系和次序建立的基于语义块的问句句型模板规则。

(7)领域本体知识库：通过分析领域知识，提取领域本体概念、概念之间的关系、实例、属性等本体元素构成本体库。采用 W3C 国际标准语言 OWL 构建领域本体知识库。

3 基于本体和问句句型模板规则的中文问句分析

汉语语言学把中文问句分为疑问句、设问句和反问句，疑问句又分为是非问句、选择问句和特指问句。根据统计，在信息咨询方面是非问句不足 10%，选择问句不足 1%，余下的约 90%都是特指问句。除另有说明外，本文讨论的问句均指特指问句，而且主要是针对咨询信息的问句。

从受限领域 QA 系统的需要出发，把问句分为实体类、属性值类、定义类、事件类、事件角色类和关系类。对每一类问句分析其表示形式，总结出相应问句规则。本文根据问句中语义块之间的搭配关系和次序构建了基于语义块的问句句型模板规则。

对于简单问句，采用问句句型模板规则匹配提取其问句语义表征，对于复杂和无规则问句，采用基于问句中一些关键词的语义进行联想的策略，概率推测其问句语义表征。因此，在 2 个层次上构建简单问句的问句句型模板规则库，第 1 个层次是对问句中的语义块(包括问点块)进行统计分析，建立语义块组成规则库。第 2 个层次是根据问句中语义块之间的搭配关系和次序建立问句句型模板规则库，并为每一个句型模板建立对应的问句语义表征，只要能正确地识别出问句的句型模板，就能直接得到其问句语义表征。问句分析的流程如图 2 所示。

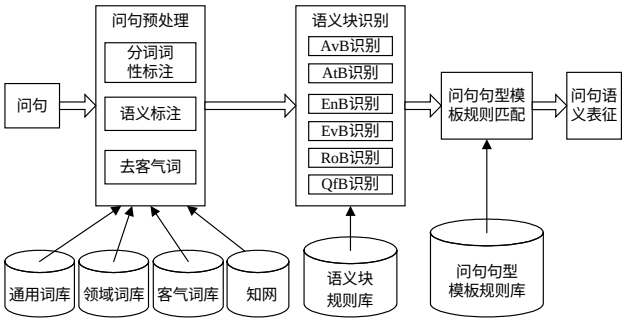

图 2 中文问句分析流程

3.1 客气词过滤

用户在进行提问时，往往会使用一些客气词，例如“请问”、“请问一下”、“请您告诉我”等。客气词对分析问句的语义表征没有帮助，在系统处理的第 1 步将问句中的客气词过滤掉。

3.2 分词和词性标注

本文使用了 ICTCLAS 系统的源码，并在原有的基础上进行了局部的改动，增加了领域专业词库，在词库查询时，领域词库比通用词库具有更高的优先级。ICTCLAS 系统使用 VC 编写，而本文的问句分析是采用 Java 语言开发，使用 JNI 技术实现了对 ICTCLAS 系统的 Java 调用。

3.3 语义块识别

语义块是指句子中具有固定语义的单词或者多词单元。定义了属性块(AtB)、属性值块(AvB)、实体块(EnB)、事件块(EvB)、角色块(RoB)共 5 种语义块，每一种语义块都由其语义组成规则。例如 AvB 的 2 个组成规则：

(1)AvB=Av，如漂亮、迅速、二、确实；

(2)AvB=AvB+[Fw(DEF={DeChinese})]+Av，如非常的漂亮、确实很迅速、两千两百。

其中，En 表示实体；Ev 表示事件；At 表示属性；Av 表示属性值；Fw 表示功能词；()表示约束要求；[]表示可选；|表示“或”；DEF 表示概念定义。

问句中还有一个特殊的语义块——问点块(QfB)，它完整地描述问句的问点。问点块通常由疑问词或疑问词和相关的词组合而成。通过统计发现，不同问句类型的问点块的语义组成不同，但都有一定的规则。例如询问时间、数量值、定义的问点块的组成规则如下：

(1)QfB=Qw(什么|啥) +En(DEF=时间) 什么时间；

(2)QfB=Qw(几)+Av(DEF=qValue) 几十、几百万；

(3)QfB=Qw(什么|啥|何|何物|何人|何事) 什么。

3.4 问句句型模板规则匹配

根据问句中语义块之间的搭配关系和次序建立基于语义块的问句句型模板规则(Question Semantic Model, QSM)。2 个询问定义的问句句型模板规则为

```
<QSM>
<QSMID>QSM_Attribute_Definition_2</QSMID>
<CONTENT>
QfB(QfBID=QfB_Attribute_Definition_1)+
EvB(DEF=是)+EnB|EvB|AtB|AvB
</CONTENT>
<Qf_relate>K(3)</Qf_relate>
<Answer_Model>K(3)+的定义如下：</Answer_Model>
<EXAMPLE>
什么+是+试管婴儿，何+为+前列腺炎
< /EXAMPLE>
</QSM>
<QSM>
<QSMID> QSM_Attribute_Definition_3</QSMID>
<CONTENT>
EnB+EvB(DEF=是)+
QfB(QfBID=QfB_Attribute_Definition_1)
</CONTENT>
<Qf_relate>K(1)</Qf_relate>
<Answer_Model> K(1)+的定义如下：</Answer_Model>
<EXAMPLE>
试管婴儿＋是+什么，前列腺炎+是+啥
< /EXAMPLE>
</QSM>
```

其中，<QSM>和</QSM>表示一个问句句型模板规则的开始和结束；<QSMID>和</QSMID>表示问句句型模板规则的编号；<CONTENT>和</CONTENT>表示问句句型模板规则的具体组成；QfBID 表示问点块(QfB)的标号；<Qf_relate>和</Qf_relate>表示对生成问句语义表征有用的语义块的标识；<Answer_Model> 和 </Answer_Model> 表示应答模板；<EXAMPLE>和</EXAMPLE>表示问句实例。

3.4.1 问句句型模板规则的组织

为了提高速度，充分利用空间，系统借鉴了ALICE的知识组织方法，采用节点复用技术，把所有的问句句型模板规则根据<CONTENT>标签的内容以树的形式来组织。假设r_i为规则树中的一个节点，k为一个词语或者语义块，则$R(r_i, k)$要么尚未定义，要么返回r_i后继节点r_j的值，$R_r = \{k : m \mid R(r, k)=m\}$则定义了以$r$为直接父节点的所有子节点值的集合。例如，设*root*为规则树的根节点，则*R_root*就表示所有规则中第1个词语或者语义块的集合。

在规则加载时，按顺序从根节点*root*到末节点t依次存入树中，同时把问句模板规则的其他信息记录在末节点t之中。设$k_1, k_2, \cdots, k_n$为一问句句型模板规则的<CONTENT>标签下的内容，要把该规则插入到树中，系统首先验证$m=R(r, k_1)$是否已经存在，如果已经存在，则以递归的方式继续把$k_2, \cdots, k_n$插入以m为根的子树中，直到$R(n, k_i)$返回未定义时为止。此时，系统就会创建一个新的节点$m=G(n, k_i)$，插入规则树中，并把余下的语义块以同样的方式处理，直至结束[7]。

3.4.2 规则搜索

系统在规则搜索时采用了带回溯的递归过程。用户输入的问句首先经过前面介绍的相关处理，然后在规则树中按层次逐个查找。如果与用户问句相匹配的终端节点中包含模板信息，则中止搜索，取出模板的内容进行后处理，返回问句分析的结果。

3.4.3 规则匹配后处理

问句分析的目的是为了提取问句的语义信息，称之为问句语义表征(Question Semantic Representation, QSR)。问句语义表征是问句语义信息的形式化表示，剔除了问句中无关或者干扰的信息，是问句语义的必要表示。答案提取模块直接依据它进行答案的提取。

一个简单的问句(一个包含多个问题的问句可以拆分为多个不同的简单问句)通常只对应一个问句语义表征，但是一个问句语义表征可以有多种不同的问句表示形式。问句语义表征的组成与问句类型直接相关，如询问实体定义的QSR={Qf=属性，Content=定义，Entity=<实体名>}；询问实体属性的 QSR={Qf=属性，Content=<属性名>，Entity=<实体名>}。例如：

Q_1：内分泌失调是什么意思？QT=定义

QSR={Qf=属性，Content=定义，Entity=内分泌失调}

Q_2：什么情况下不能做人工授精？QT=属性

QSR={Qf=属性，Content=禁忌症，Entity=人工授精}

Q_3：请告诉我得了不孕症如何治疗？QT=实体角色

QSR={Qf=事件角色，Content=方式，Event=治疗，Entity=不孕}

其中，QT 表示问句类型；Qf 表示问点；Event 表示事件；Entity 表示实体；Content 表示具体的属性内容或者事件角色。

4 实验结果

对该系统进行了 2 方面的测试：一是面向问句语料库(共 1 600 个问句)进行测试；二是面向实际用户现场测试，实验结果如表 1 所示。对第 1 种测试方法的结果进行分析，没有正确回答的问题分为 2 种类型：一种是问句分析结果正确，但本体知识库中没有知识(44 个)；另一种是没有正确提取问句语义表征(144 个)，而这种类型又可分为缺少领域同义词、缺少相应的问句句型模板规则等。因此，问句分析的正确率为(1 840+44)/2 028=92.899%。第 2 种方法的准确率比第 1 种方法稍低，通过分析，原因主要有：用户输入的问句中存在错别字、缺少领域同义词、缺少相应的问句句型模板规则、本体知识库中缺少相应的知识。实验结果表明，通过这种方法设计的受限领域 QA 系统是可行的。

表 1 受限领域 QA 系统实验结果

测试方法	问题数	正确回答数	准确率/(%)
1	2 028	1 840	90.72
2	500	379	75.80

5 结束语

本文根据语义块组成规则和基于语义块的问句句型模板规则库，采用规则匹配的方法对问句进行分析，比传统的基于关键词的方法有一定的提高。提出了问句语义表征的概念，问句语义表征是问句所表达的语义信息的形式化表示。本文探索了问句语义表征的表示形式，以及如何提取问句的语义表征。在该方法的基础上，将继续深入开展如下研究：

(1)问句语义块的自动识别。

(2)使用统计的方法来研究问句语义表征的自动生成。

(3)研究如何进行复杂问句语义表征的自动分析。

(4)使研究成果能够方便地移植到其他领域中。

参考文献

[1] 郑实福, 刘 挺, 秦 兵, 等. 自动问答综述[J]. 中文信息学报, 2002, 16(6): 46-52.

[2] 王树西, 刘 群, 白 硕. 一个人物关系问答的专家系统[J]. 广西师范大学学报, 2003, 21(1): 31-36.

[3] 黄寅飞, 郑 方, 燕鹏举, 等. 校园导航系统 EasyNav 的设计与实现[J]. 中文信息学报, 2001, 15(4): 35-40.

[4] 秦 兵, 刘 挺, 王 洋, 等. 基于常问问题集的中文问答系统研究[J]. 哈尔滨工业大学学报, 2003, 35(10): 1179-1182.

[5] 樊孝忠, 李宏乔, 李良富, 等. 银行领域汉语自动问答系统 BAQS 的研究与实现[J]. 北京理工大学学报, 2004, 24(6): 528-532.

[6] 邓志鸿, 唐世渭, 张 铭, 等. Ontology 研究综述[J]. 北京大学学报, 2002, 38(5): 730-738.

[7] 夏 天, 樊孝忠, 刘 林, 等. 基于 ALICE 的汉语自然语言接口[J]. 北京理工大学学报, 2004, 24(10): 528-532.