基于领域本体的自动问答系统关键技术研究

高俊杰　李茹　李双红

山西大学计算机与信息技术学院，太原 030006

E-mail: gaojunjie_verve@foxmail.com

摘　要：基于领域本体的自动问答系统的性能主要取决于本体知识库的结构、问句分析结果以及从本体中进行答案查询推理的策略。本文首先构建了一个结构良好的本体知识库；然后对问句进行面向领域本体的问题分类，并在CFN标注的基础上提取问句的结构化语义信息；最后通过规则将问句结构化语义信息映射到本体中，运用规则推理在本体中查询答案，并对答案进行整合处理。实验结果表明，本文的技术路线应用于自动问答系统是可行的。

关键字：自动问答系统，领域本体，问句分析，汉语框架网，规则推理

Research on the Key Technology of Automatic Question Answering System Based on Domain Ontology

GAO Junjie, LI Ru, LI Shuanghong

School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, China

E-mail: gaojunjie_verve@foxmail.com

Abstract: The performance of the automatic question answering system based on domain ontology depends largely on the following three points: the structure of ontology knowledge base, the results of the question analysis, as well as the strategy of query and inference answers from the ontology. In this paper, we build an ontology knowledge base with reasonable structure firstly. Then we classify the questions according to the domain ontology and extract the structured semantic information of questions depending on CFN label. Finally, we map the structured information of questions to ontology by heuristic rules, and query answers in the ontology using rule-based inference mechanism, and arrange the answers. The result shows that, the technical route in the paper for automatic question answering system is effective.

Key words: automatic question answering system, domain ontology, question analysis, Chinese FrameNet, rule inference

1 引言

自动问答系统是目前自然语言处理领域一个比较热的问题，它既能够让用户用自然句子提问，又能够为用户返回一个简洁、准确的答案。因此，自动问答系统和传统的依靠关键字匹配的搜索引擎相比，能够更好地满足用户的检索需求，更准确地找出用户所需的答案，具有方便、高效、准确等特点。

基于知识库的问答系统[1]是一类重要问答系统。本体作为一种能在语义和知识层次上描述信息系统的概念模型建模工具，本质就是领域知识的共享和复用，因而基于本体的问答系统是基于知识库的问答系统研究中的一个重要方向。

基于本体的自动问答系统的关键技术有以下三个：一是本体知识库的结构；二是问句分析理解方法；三是在本体中的答案查询推理策略。

目前，基于领域本体的自动问答系统的研究工作还不多。骆正华等[2]研究的金融领域汉语自动问答系统BAQS中，通用本体（知网）起到联系语言学知识和领域知识的中介作用。该系统用知网中少量的义原定义了金融领域内的概念，并通过义原联系领域内的概念。

基金项目：国家 863 高技术研究发展计划资助项目(2006AA01Z142)，国家社会科学基金青年项目(07CYY022)

作者简介：高俊杰(1984—)，男，硕士生，主要研究方向为智能信息处理；李茹(1963—)，女，教授，主要研究方向为智能信息处理；李双红(1984—)，男，硕士生，主要研究方向为计算语言学。

在传统自动问答系统的问句分析过程中，有些系统把问句分成两个部分：问句焦点、疑问词，然后基于模式进行匹配，通过统计，然后进行评估，确定答案[3]；有些系统除了问句焦点和疑问词，还考虑到问句的一些语法特征[4]；余正涛等提出基于问句语料库的问句分析方法，对问句的关键词、主题词、问点等语义信息进行标注[5]。曹志娟等在问题分析过程中问句分类结合了疑问词短语分类、问题标准型以及特征词分类[6]。

在传统自动问答系统的信息检索模块，一般是根据问句分析得到的关键词调用Internet上的搜索引擎或是调用已有的检索系统。而在答案抽取模块则依据问题的类型（如，句子作为答案、词语作为答案或文摘作为答案）采取不同的抽取策略[7]。

本文通过设计基于领域本体的旅游自动问答系统，构建了旅游本体知识库，提出了面向领域本体的问题分类方法以及基于汉语框架网（Chinese FrameNet, 简称CFN）的问句结构化语义信息提取方法，并利用问句结构化语义信息进行本体的查询推理，对基于本体的自动问答系统的三个关键技术进行了研究。

2 旅游本体知识库的构建

尽管构建本体的方法有多种，但是仍然没有一种标准的方法或者流程[8]。对于构建领域本体，一般来说需要领域专家的指导，而且其构建目标需要基于使用本体的目的。

根据旅游业定义中的要素，并结合旅游自动问答系统的应用需要，本文构建了旅游领域本体。到目前为止，共建立了69个本体类，83种属性和关系，添加有6400多条实例数据。

2.1 本体的类层次结构

根据对旅游业的旅游资源、旅游设施和旅游服务三大要素的考察，在本文构建的旅游本体中，设有“旅游资源”(TouristResources)和“服务设施”(ServiceFacilities)两个一级类目。

为“旅游资源”和“服务设施的”设置子类的过程中，本文参考了《旅游资源分类、调查与评价》（GB/T18972-2003）中的《旅游资源分类表》以及旅游的“吃、住、行、游、购、娱”六要素。最终本体中主要的类层次如下面图1所示，各个类中进一步划分的子类在这里不再详述。

图1　旅游本体主要类层次

2.2 本体类的属性

在构建本体的过程中，合理地确定每个类的属性、属性的约束以及类之间的关系非常困难。如果属性的概念太过细化，则属性的数量势必大量增加，人工构建本体的工作量将增大，且对于问答系统来说，系统回答一些概括性问题需要将好多属性值进行组合；反之，若属性的概念太笼统，属性值则包含大量信息，一般成为一大段文本，这将导致当提问一些具体问题时系统不能直接进行回答，必须得借助信息抽取等技术从一个属性值中进行答案过滤。

本文在确定各个类属性的过程中，一部分属性根据领域的客观知识确定，一部分属性根据领域专家的建议确定，还有一部分属性是根据对所收集领域知识文本和真实问句语料的分析结果确定。某些情况下，在属性的客观性与易用性之间进行了折中处理。

2.3 本体的实例数据

结构化信息有利于信息的检索与提取，因此本文构建本体时选择将领域知识表示成结构化的本体实例数据。本文构建的旅游本体知识库有着完整的实例数据，它并不只是一个领域概念框架（或者说是术语本体），这有别于现有的一些基于本体的应用。

如Textpresso本体[9]是利用本体中的概念来标注领域知识文本的，它将一个完整的文本内容（包括其中的每个词以及标点）完全用本体中的术语标注。而本文在构建本体时将抽取领域知识文本中的信息然后组织成为本体中的实例。

3 问句分析

问句分析作为整个问答系统的子模块，其效能直接影响问答系统后续处理流程的展开，具有极其重要的作用。在问句分析模块，一般要对问句进行分词、词性标注、问题分类、问句信息提取等等，在这里主要介绍面向领域本体的问题分类和结构化语义信息提取。

3.1 问题分类

问题分类的目的是确定答案的语义类别及其搜索分析策略。通过将用户问句映射至不同的问题类型，从而确定答案搜索、答案抽取的策略，以提高系统的效率和准确率。

目前流行的基于开放域的问答系统提出了多种问题分类标准[10,11]，一般分为：人物、地点、数字、时间等6-7大类，大类中又分若干小类，但这些分类对于领域问题不太适合。

本文参考已有的问题分类体系和已建立的旅游领域本体模型，结合了问句的答案类型和问句的领域主题概念类型，总结出以下的分类体系（如表1所示）。

表1 问题分类体系

大类（Coarse）	小类（Fine）
人物(HUM)	特定人物 团体机构 人物描述 人物列举 人物别名 人物其他
地点(LOC)	所在地区 景点方位 网址 地址 名人住地 特定景点 艺术源地 地点其他
数字(NUM)	电话号码 邮编区号 门票价格 车票价格 旅游花费 景点气温 数量 距离 重量 温度 年龄 面积 体积 频率 速度 范围 顺序 数字其他
时间(TIME)	起源时间 发车时间 历史年代(范围) 旅游最佳期 交通时间 游览需要时间 时间其它
实体(OBJ)	动物 植物 食物 颜色 货币 语言文字 物质机械 交通工具 宗教 娱乐 庙宇 人造神 旅游景点 实体列举 实体其他
描述(DES)	定义 意义 方法 命名原因 其它原因 交通路线 日程安排 景点描述 历史 景点特色 建筑特征 故事 传说 典故 描述其他
未知(Unknown)	未知

3.2 基于CFN标注的问句结构化语义信息提取

为了从本体知识库中获取用户提问问句的答案，必须对问句进行分析并提取其结构化语义信息，进而根据这些信息从本体知识库中进行查询推理以得到问题的答案。本文在问句分析过程中使用CFN[12]对问句进行了标注，然后提取特定的CFN框架元素作为问句的语义信息。

虽然在一个问题类型中不同问句的句式各异，并且它们所涉及到的目标词也有许多，但一个CFN框架涉及的不仅仅是个别的词，而是一类词语，它们都具有共同的认知结构，支配相同类型的语义角色。通过对已经分类并加以CFN标注的问句分析表明：一个问题类别中的所有问句标注后所涉及的CFN框架只有少数几个。

另一方面，不同框架中所拥有的核心框架元素和非核心框架元素是不同的，因此提取语义信息需对同类型的问句再按标注的框架分类，进行不同的处理。

通过对特定问题类别中特定框架的问句进行总结，分析其各个框架元素的语义，就可以制定规则来提取合适的框架元素作为问句语义，所提取的框架元素必须是回答该类别问题必要的语义限制关键词，因为有了这一结构化信息就能对该类问题进行回答。下面举一例来说明。

“交通路线”类问句提取规则：“交通路线”类问句，由[到达]框架标注的提取[src]、[goal]和[mot]三个框架元素；由[位移]框架或[出发]框架标注的提取[src]、[goal]和[car]三个框架元素。

上面的规则之所以提取那些框架元素是因为：根据[到达]、[出发]、[位移]这几个框架的定义，在“交通路线”类问句中，[src]表示“路线起点”、[goal]表示“路线终点”，而[mot]或[car]表示“交通方式”；而对于“交通路线”类问句，只要知道路线起点、路线终点和交通方式这三点信息就能够对问题做出回答。

到目前为止，本文总结出了“起源时间”、“发车时间”、“人物别名”等19类问句中涉及不同框架共62种问句语义信息提取规则，具体规则在这里不再详述。

4 本体的查询推理

问句分析服务于在本体中提取答案。当一个问句经过问句分析后得到了问题类别并提取到结构化语义信息，接下来就可以运用问句的这些信息在本体中去提取答案。

4.1 问句结构化语义信息到本体的映射规则

从本体知识库中检索答案，必须将从问句中提取的结构化语义信息定位到本体中。本文通过规则联系问句语义与本体语义，将问句结构化语义信息中的词语定位到本体之中。然后通过查询推理，从本体知识库中得到问题的答案。

概括地说，从问句结构化语义信息定位到本体中的一般规则为：将问句结构化语义信息中的关键词语映射为本体中的概念、属性或关系，或是将结构化语义信息中的关键词语作为本体中某些属性的属性值，并且把问句的答案也作为本体中某些实例的一个属性值（未知，待查询）。

下面举例来说明问句与本体的映射过程。

在本体中，Traffic类的一个实例为一条具体的从地点A到地点B的交通路线，因此它有诸如“出发地”(traffic_startplace)、“到达地”(traffic_destination)、“交通需要时间”(traffic_needTime)等属性。同时Traffic又有“自驾车”(Selfcar)、“长途大巴”(Coach)、“火车”(Train)等子类。

当我们对“交通时间”类问句的答案进行查询推理时，从问句提取的语义信息映射到本体的规则是：

(1)“交通方式”对应本体中Traffic类的子类；（映射为概念）

(2)“路线起点”对应本体中Traffic类的traffic_startplace属性；（作为属性值）

(3)“路线终点”对应本体中Traffic类的traffic_destination属性；（作为属性值）

(4) 返回本体中符合上述条件的实例的traffic_needTime属性值。

4.2 本体中的查询推理规则

前面4.1节中简单的映射规则并不能解决所有的用户问句查询。本节来讨论这种情况并给出解决策略。

在本体中traffic_startplace属性值和traffic_destination属性值都是地名，对问句“从太原到北岳恒山自驾车怎么走？”进行问题分类并提取语义后得出的“路线终点”是“北岳恒山”或是“恒山”（这不同于traffic_startplace的属性值）。有时，提取的“路线起点”也可能不是个地名，如问句“从乔家大院到常家庄园怎么走？”，这都将导致仅用上节中那些简单规则进行查询时匹配不到路线实例，需要运用本体的推理机制。

在旅游领域本体中，“北岳恒山”、“乔家大院”都是景点类的实例，而旅游景点有“景点所在地”(sight_location)属性。因此当4.1节中规则查询推理没有结果时，就对“路线起点”、“路线终点”作为旅游景点去查询其sight_location属性值，然后再执行以上规则。这个过程的推理规则形式化如下：

规则1：path(a, b, p)←sight_location(a, A)∧sight_location(b, B)∧traffic_path(A, B, p)

对规则1直观解释为：从景点a所在地A到景点b所在地B的路线即为从景点a到景点b的路线。

又如，本体中景点类之间虽然定义有“邻近关系”(sight_neighbours)，但人工建立本体时也许会遗漏掉一些景点的邻近关系声明，这种情况下若对某景点的邻近景点进行查询就会漏掉一些事实上应有的结果。为了解决这个问题，本文在本体中加入以下的规则：

规则2：sight_neighbours(a, b)←sight_location(a, A)∧sight_location(b, A)

对规则2的直观解释为：同一地点的两个景点是邻近的。

本体的基础是描述逻辑，描述逻辑提供的推理能力是有限的，描述逻辑不提供关系间的组合推理[13]，已知hasFather(Tom, Jack)和hasSpouse(Jack, Rose)，却无法推出hasMother(Tom, Rose)（假设已有hasMother这个关系的声明）。本节中规则2所在的例子就类似这种情况，虽然有关系的声明但不能通过描述逻辑的推理来完成；而规则1所在的例子较之更为复杂，因为两个景点的path根本没有在本体中声明。

研究基于领域本体的自动问答系统时，从本体知识库中查询答案可以看作是一种复杂的关系间的组合推理。多数情况下，这种关系未在本体中声明，甚至声明这种关系是不可能的。因此从本体中获取知识时必须合理地利用规则。

4.3 利用Jena工具实现对本体的查询推理

在本文的实验系统中，对本体的查询推理主要通过Jena来完成。Jena是来自于惠普实验室语义Web研究项目的开放资源，是用于创建语义Web应用系统的Java框架结构，它为RDF、RDFS、OWL、SPARQL提供了一个程序开发环境，并且包括一个基于规则的推理引擎。

对于4.1中简单的映射规则本文通过构造SPARQL语句，然后交给Jena去执行查询；对于4.2中所述的情况本文通过把自定义规则加入到Jena推理机中，让其产生推理的本体模型，然后再利用SPARQL语句进行处理。

5 实验及结果分析

本文在J2SE平台开发了实验系统，测试数据为搜集的2011句真实问句语料（记为数据集A）。实验过程中按本文的问题分类体系用最大熵模型对这些问句进行分类，封闭测试得大类分类准确率为90.38%，小类分类准确率为84.24%，这个结果略低于传统分类体系得出结果，这表明本文的分类体系仍是一种可取的选择。

为了验证问句语义提取规则以及查询推理规则的有效性及其效果，对数据集A中的693句问句（所有属于已总结出问句语义提取规则的19类问句，记为数据集B）进行了单独测试，自动分

类正确的为594句，不正确的为99句（数据集B的分类准确率为85.71%，比数据集A的结果稍高）。

由于CFN自动标注技术仍不成熟，提取问句语义信息时测试系统直接调用问句对应的人工CFN标注结果。另一方面，本体知识库中已保证了数据集B中所有问句的答案。

实验最终结果为，系统只对数据集B中的548句问句返回了答案，且返回答案的问句都是自动分类正确的问句。分析实验结果可得到如下几点：

（1）如不考虑问题分类以及CFN标注的影响，问句语义提取规则和答案的查询推理规则的准确率达92.25%（在594句分类正确的问句中有548句返回答案），表明规则的制定是比较合适的。

（2）系统对分类错误的句子不能给出答案。这是因为问句的语义信息主要根据CFN框架元素来提取，而不同类型的问题往往没有相同的框架元素，即使有相同的框架元素但意义也不同。所以分类错误时规则失效。

（3）进一步分析分类正确但得不到正确答案的问句，主要由以下两个原因造成：①从框架元素中提取语义信息时提取的词语不当；②问句中缺少一定的语义角色导致提取的结构化语义信息不完整，如问句“到五台山旅游坐汽车怎么走？”中没有“路线起点”的信息。

6 小结与展望

实验的结果表明，运用本文所述的问句分析方法以及查询推理方法可以达到令人满意的效果，本文的问句分析方法和查询推理方法是行之有效的。

但是，本文的工作仍有不足，在现有的基础上可以继续开展以下研究：（1）构建更加合理的旅游领域本体，进一步讨论其在理论上的合理性以及在应用上的便利性和可行性。（2）在问句语义信息的提取过程中，一方面可以进一步研究如何将CFN提供的语义信息更好地应用于结构化语义提取；另一方面可以尝试在规则的基础上结合其它方法，以克服规则的局限性。（3）旅游领域本体的推理规则需要进一步完善，以使推理框架能更加方便、准确地获取本体知识库中的知识，来满足自动问答系统的需要。

参 考 文 献

[1] 陆汝钤. 人工智能[M]. 北京: 科学出版社, 2000.

[2] 骆正华, 樊孝忠, 刘林. 本体论在自动问答系统中的应用[J]. 计算机工程与应用. 2005, 41(32): 229-232.

[3] Radev D, Fan W. G, Qi H. Probabilistic Question Answering on the Web. 2002.

[4] Zukerman I, Horvitz E. Using Machine Learning Techniques to Interpret WH-question. 2001.

[5] 余正涛, 樊孝忠, 宋丽哲. 基于问句语料库的受限领域自动应答系统[J]. 计算机工程与应用. 2003: 28-30.

[6] 曹志娟, 李祖枢, 刘朝涛. 自动问答系统中的问题理解研究[J]. 计算机科学. 2005, 32(11): 158-160.

[7] 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.

[8] Noy N. F, Mcguinness D. L. Ontology Development 101: A Guide to Creating Your First Ontology. 2007.

[9] Müller H, Kenny E. E, Sternberg P. W. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature[J]. PLos Biology. 2004, 2(11).

[10] 文勖, 张宇, 刘挺, 等. 基于句法结构分析的中文问题分类[J]. 中文信息学报. 2006, 20(2): 33-39.

[11] 贾可亮, 樊孝忠, 许进忠. 基于KNN的汉语问句分类[J]. 微电子学与计算机. 2008, 25(1): 156-158.

[12] 郝晓燕, 刘伟, 李茹, 等. 汉语框架语义知识库及软件描述体系[J]. 中文信息学报. 2007, 21(5): 96-100.

[13] Staab S, Horrocks I, Angele J, et al. Trends & controversies - Where are the rules?[J]. Intelligent Systems, IEEE. 2003, 18(5): 76-83.