北京工业大学

硕士学位论文

基于问答系统的人机对话平台设计与实现

姓名：丁杰

申请学位级别：硕士

专业：计算机应用技术

指导教师：庄梓新

20090401

摘 要

北京 2008 年奥运会是中国历史上第一次主办的奥运会，届时来自世界各地的有着不同文化背景的中外友人将云集北京，参与体育竞技、组织工作及观光活动，北京将成为世界关注的焦点。根据“科技奥运促进人文奥运”的重要理念和《北京奥运行动规划》制定的重要目标，北京在确保奥运会顺利进行的同时，也要为社会公众提供丰富、便捷、易于获取的信息服务。因此，集成了自然语言理解、对话管理、信息抽取等技术的对话系统将以奥运为契机，成为首都信息服务建设的重要组成部分,这也是本课题的研究内容。

课题的主要研究目的是利用首都信息发展股份有限公司在对话管理与信息处理方面的技术积累，通过新的对话管理流程，实现一个以自然语言问答形式为公众提供多主题信息查询服务的人机对话平台。本文是在奥运多语言综合信息服务系统前期研究成果的基础上进行的阶段性研究，依靠奥运综合信息资源库的资源支持，课题成果最终服务于北京 2008 年奥运会，为大众提供全面周到的城市信息与奥运信息服务。

本文通过深入研究对话系统所涉及的关键技术，仔细分析用户和系统需求，制定了人机对话平台的总体设计方案。根据形式语言与自动机理论，在系统中实现了基于规则匹配与参数提取的自然语言理解方法。凭借首信公司在对话管理方面的研究成果，结合树结构的主题管理方法与可追溯的历史管理策略，完成了基于槽和任务的对话管理模块设计，建立了规则库、参数库、语料库和信息数据库，并采用 Web Service 方式发布服务。最终，通过系统测试给出了性能评价和问题分析。

关键词 对话系统；自然语言理解；对话管理；信息抽取；自然语言生成

Abstract

The 2008 Beijing Olympic Games is the first Olympic Games hosted by China. People from all over the world under different cultural background will gather in Beijing to engage in athletic sports, service work or for tourism at the appointed time. The eyes of the world will be on Beijing. According to the concept of "Technology Olympics promote Cultural Olympics" and the aim established by "Beijing Olympic Action Plan", Beijing needs to provide rich, convenient and easily available information services for public in ensuring successful Olympics. Therefore, dialogue system, which integrates natural language understanding, dialogue management and information extraction will be an important part of informatization construction in Beijing through Olympic Games. This is also the content of the research.

The main purpose of the research is to implement a human-machine dialogue platform using new flow of dialogue management based on technique accumulation of Capinfo Co., Ltd in dialogue management and information processing. It will provide information query service through question-and-answer form in natural languages. This dissertation is a research on the basis of the initial achievements of the Olympic multilingual integrated information service system in phases, depending on resources support of Olympic integrated information resource database. This research achievement serves for the 2008 Beijing Olympic Games to provide information of city and Olympic Games for public.

This dissertation makes an intensive study of key technologies in dialogue system, analyses requirement of user and system carefully, establishes integral design scheme of human-machine dialogue platform. A method of natural language understanding has been carried out in system based on rule matching and parameter extraction, according to the theory of formal language and automata. Depending on achievements of Capinfo Co., Ltd in dialogue management and combining the subject management method of tree structure and traceable history management strategy, this research accomplishes module design of dialogue management based on slot and task, establishes rule database, parameter database, corpus database and information database, and use web service to release. Finally, performance evaluation and problem analysis are put forward through system test.

Keywords dialogue system; natural language understanding; dialogue management; information extraction; natural language generation

独 创 性 声 明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京工业大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签名：丁杰 日期：2009.5.18

关于论文使用授权的说明

本人完全了解北京工业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵守此规定）

签名：丁杰 导师签名：[illegible] 日期：2009.5.18

第 1 章 绪 论

1.1 课题来源及研究目的

2007 年 4 月，首都信息发展股份有限公司（以下简称首信公司）与北京奥组委签约成为奥运历史上第一个多语言服务供应商，负责承建奥运多语言综合信息服务系统。该系统属于国家 863 计划重大课题《嵌入式分布式语音技术在城市综合信息服务系统中的应用示范》中需要实现的重要指标。

课题主要研究目的是利用首信公司多年来在对话管理与信息处理方面的技术积累和奥运多语言综合信息服务系统的前期研究成果，集成了自然语言理解、对话管理、信息抽取、自然语言生成等技术，实现了一个采用自然语言查询方式的人机对话系统，并通过该系统为人们提供方便、快速、丰富、自然的信息服务。课题成果最终将服务于 2008 年北京奥运会，为公众提供多领域、多应用场景的城市信息与奥运信息查询服务。

1.2 研究本课题的重要意义

1.2.1 理论意义

采用问答形式的自然语言对话系统是当前对话系统领域的研究热点。本课题通过对自然语言理解、对话管理、信息抽取、自然语言生成等技术的探索与研究，建立了基于问答系统的人机对话平台系统模型，制定了规则库、参数库、语料库和信息数据库的结构标准，给出了一个易于主题扩展的对话管理流程，对人机对话系统的应用起到了示范作用，也为首信公司人机对话系统的开发积累了宝贵的经验。

1.2.2 直接意义

(1) 为 2008 年北京奥运会提供综合信息服务

北京 2008 年奥运会是中国历史上第一次主办的奥运会。在奥运会期间，将会有来自世界各地的国际奥委会委员、各国代表、体育界人士、运动员、教练、奥运工作者、新闻记者，以及数以百万计的国内外游客云集北京，参与奥运会比赛、组织工作及观光活动。届时，北京将成为世界关注的焦点，在确保奥运会顺

利进行的同时，也要为社会公众提供丰富、便捷的信息发布、信息查询、人机交互等多种形式的信息服务。

根据北京奥申委提出的“绿色奥运、科技奥运、人文奥运”理念和“到2008年，基本实现任何人、在任何时间、任何场所都能够安全、方便、快捷、高效地获取可支付得起的、丰富的、无语言障碍的、个性化的信息服务”的承诺，本课题紧密结合人们在奥运会期间的信息服务需求，运用国家863计划以及首信公司承建的奥运多语言综合信息服务系统所支持的各种信息资源和先进技术研发了一个采用自然语言问答形式的人机对话系统，并利用综合信息资源库提供的信息资源建立有较强扩展性的规则库和语料库，使其具有更加人性化的交互信息，能够在2008年奥运会期间提供多领域、多应用场景的信息查询服务。

(2) 在线应答服务

当前，国内有许多网站都在自己的系统中加入了在线应答服务，但是在人力配置上往往无法满足即时回复用户提问的要求，例如某一时间段用户访问量的剧增，往往会导致部分用户长时间得不到答复；另外，大多数网站都无法保证全天24小时提供服务，如果用户在工作时间外提出问题，回复将会被延后。基于特定领域的问答系统可以有效的弥补人工应答服务上的不足，克服人力配置的缺陷，具有问题处理时间短，受时间因素影响小的优点。将自动问答系统与人工应答服务相结合，可以节省人力资源，提高工作效率。

(3) 公共场所的信息查询终端

信息查询终端是公共场所服务设施的一部分，目前已经在机场、车站、银行、邮局等场所广泛使用。通常，这类查询终端采用的是基于常见问题解答（Frequently Asked Questions，FAQ）的查询方式[1]，查询内容固定，无法适应特殊情况。特别是遇到很少使用这类设施的用户，往往无法很快的切入或是不能切入到想要查询的内容点上，影响了使用效率。采用了问答系统的信息查询终端可以降低用户对设备熟悉程度的依赖，以自然语言对话的方式引导用户发现需要查询的重点信息。

(4) 移动信息查询终端

与公共场所的信息查询终端类似，不过移动设备的用户组成更为广泛，设备使用更加频繁，且不会受时间和所在地区的影响。从应用角度上来说，采用自然语言查询方式的移动信息查询终端将会在较大程度上提高系统的便捷性和灵活性。

1.3 国内外研究现状及分析

1.3.1 国外研究进展

近年来，基于问答系统的人机对话技术及其产品引起了国内外许多科研机构和公司的兴趣。国外一些知名的大学和科研机构都对特定领域自然语言理解、智能人机对话系统进行了深入的研究和系统开发，其中比较著名的如麻省理工学院（MIT）、密歇根大学（MU）、卡内基-梅隆大学（CMU）、科罗拉多大学（CU）等均进行了问答式人机对话系统的研发。一些公司如 Microsoft、IBM、Roussinov 等也推出了自己的智能对话系统[2~4]。下面是国外科研机构及公司在问答式人机对话系统上的一些典型应用：

(1) Start 自动问答系统

麻省理工学院人工智能实验室的 Start 系统是第一个基于 Web 的自动问答系统，也是在线问答系统中的佼佼者。它以自然语言提问为检索入口，采用基于知识标注（Knowledge Annotation）和数据挖掘（Knowledge Mining）的核心技术将结构化半结构化的数据与自由格式文本区别处理，向用户提供准确的信息，而不是提供一堆相关信息让用户自己挑选[5]。采用了基于知识库与信息检索混合模式的 Start 系统，在用户查询信息时，若从知识库中可以找到答案，则直接反馈；否则将通过搜索引擎检索并处理后反馈给用户[6]。

(2) AnswerBUS 问答系统

密歇根大学的 AnswerBus 系统是一个面向开放领域的问答系统，它接受自然语言的提问方式，从 Internet 上提取问题可能的答案（一个或多个），其特点是能支持包括英语、法语、德语、意大利语、西班牙和葡萄牙语在内的多种语言提问方式[7,8]。AnswerBus 自动问答系统自 2001 年研究开发完成并开始在 Internet 上运行以来，现已成为互联网上重要的智能信息检索工具，广泛应用于科学研究、文化娱乐等许多领域。

(3) CATCH2004 项目

由欧盟 IST（Information Society Technology）计划资助的 CATCH 2004 项目于 2000 年 1 月启动，研究周期为两年半。该项目的目标是支持用户以多模态、多种语言、比较自由的口语对信息系统进行访问。CATCH 2004 项目采用了包括电话、手持无线通讯设备和信息亭的三张接入方式，支持英语、德语、希腊语和芬兰语四种语言，并计划在雅典 2004 年奥运会上实现城市事件信息和运动会信息两个应用系统[9,10]。尽管由于该项目 2002 年结束后的后续应用示范和商业试用没有跟上，影响了研发成果在 2004 年雅典奥运会上的正式开通使用，但 CATCH 2004 的成果仍可为以后的综合信息系统研发提供技术指导。

(4) 知名公司的研究成果

2005 年，IBM 公司在其非结构化信息管理架构（Unstructured Information Management Architecture，UIMA）的软件架构平台上展开了基于语言分析、知识库、问答系统、机器翻译等功能的自然语言搜索研究，并计划在此基础上构建第三代信息检索引擎，通过 UIMA 架构使应用程序可以提取多媒体数据中的文档信息，并将这些文档视为“人类语言的表达”而不是匹配文字模式，最终组织成更加结构化的信息，实现智能化信息检索[11,12]。

2005 年，Microsoft 公司将 Internet 信息检索技术和基于微软在线百科全书的知识库整合到 MSN 即时通讯服务中，通过用户与虚拟机器人 Encarta 的交互实现了 IM 方式的人机对话系统。

2007 年，Dmitri Roussinov 提出了一种新型问答式应用系统，旨在信息检索时返回问题对应的准确答案，而不是包含分类结果的若干相关网页[13]。

1.3.2 国内研究进展

在国内也有许多大学和研究所在进行问答式人机对话系统的研究工作。不过中文对话系统的起步较晚，相对于国外的技术还不够成熟，其主要原因是：中文对话系统除了要具有一般对话系统的功能外，还需要考虑到汉语的特性。在自然语言处理中，中文的语法语义等方面都与西方语言有着很大的区别，其结构特点决定了它的句法分析和语义理解要更加复杂，词与词之间没有空格分界符也使得系统在信息处理时要先对句子进行切分，因此中文对话系统往往无法直接利用国外一些成熟技术和研究成果。另外，中文对话系统的知识库、评测标准、评测平台等语言处理基础资源缺乏，也在一定程度上影响着中文对话系统的发展[14~16]。下面是国内科研机构及公司在中文问答式人机对话系统上的一些典型应用：

(1) NKI 知识问答系统

中科院计算所智能信息处理实验室研发的大规模知识处理科研项目“国家知识基础设施”（National Knowledge Infrastructure，NKI）是一个庞大的、可共享的知识信息平台。它包含 16 个学科的 580 多个专业本体，各学科本体按照继承和实现等关系形成了相对独立的体系结构。该系统包含大约几百万条的专业知识信息，通过一个基于 NKI 知识库的中文问答系统 HKI，向用户提供多领域的知识信息服务。HKI 系统的主要特点是支持自由的提问方式，并向用户提供准确的回答信息[17,18]。

(2) 小 i 机器人

赢思公司开发的小 i 机器人是当前国内关注度较高的中文对话系统，它同时提供了开放域和针对特定领域的多种人机交互环境，可以根据关键词将用户引导

到不同领域的信息服务系统中，与用户进行颇具人性化的人机交互。小i机器人具有Web和IM两种用户接口，同时提供了可自定义领域信息的系统扩展接口，使用户可以定制自己的问答机器人。

1.4 主要研究内容

本论文从人机对话系统方便、快速、丰富且易于扩展的实际需求出发，对基于问答系统的人机对话平台的总体框架、设计思想、所需要涉及的技术等进行研究。在此基础上，详细讨论了人机对话平台的系统实现，其中包括：平台体系结构的设计思想；重要数据结构与数据文件的定义；自然语言处理方法的选择；对话管理策略的设计；信息内容的获取方式；规则库、参数库、语料库和信息数据库的结构设计；人机对话平台的系统测试等。

1.5 论文的组织结构

本文分三个部分来阐述：

第一部分：本文的第1章。介绍选题的背景及意义，描述了目前问答式人机对话系统的现状，并说明了本文的主要研究内容。

第二部分：此部分是论文的主题。本文的第2章描述了人机对话平台所涉及的关键技术。第3章描述了人机对话平台的系统需求、层次框架和结构模型，详细介绍了人机对话平台的总体设计。第4章描述了人机对话平台各功能模块的详细设计与实现。第5章对人机对话平台的系统功能与性能进行了测试分析，并对测试中出现的问题给出了解决办法。

第三部分：全文总结。

第 2 章　人机对话平台的关键技术

2.1　中文分词

自然语言处理是计算机科学领域与人工智能领域中的一个重要方向，它主要研究人与计算机之间进行自然语言交互的基本理论和方法。中文分词技术是自然语言处理系统的重要组成部分，它是计算机理解自然语言信息的基础[19,20]。

中文分词是将按照自然语言规范组合的句子划分成词序列的过程。在英文文本中，空格是单词之间的自然分界符，无需对句子的词边界进行确认。而中文在句子构成上没有一种明显的词边界符，也就是说中文只是字、句和段可以通过明显的分界标志来划分边界。所以对于中文来讲，确定词的划分是理解自然语言的第一步[21~23]。

现有的中文分词可以分为基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法三大类[24]：

(1) 基于字符串匹配的分词方法。这种方法又叫做机械分词方法，它是按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行配，若在词典中找到某个字符串，则匹配成功（识别出一个词）。按照扫描方向的不同，串匹配分词方法可以分为正向匹配和逆向匹配；按照不同长度优先匹配的情况，可以分为最大（最长）匹配和最小（最短）匹配；按照是否与词性标注过程相结合，又可以分为单纯分词方法和分词与标注相结合的一体化方法。常用的机械分词方法有正向最大匹配法、逆向最大匹配法和最少切分法等等。

(2) 基于理解的分词方法。这种方法的基本思想就是在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。它在分词消岐过程中模拟了人对句子的理解过程，这就需要使用大量的语言知识和信息。

(3) 基于统计的分词方法。这种方法的基本思想是对语料中字组频度进行统计，计算出字组之间的互信息量，该值反映了字组之间结合关系的紧密程度，当它达到一定的阀值时，可以确定字组是一个词。基于统计的分词方法就是通过对语料中字组频度的统计获得分词信息的，它不需要切分词典，因此也叫做无词典分词法或统计取词方法。

中文分词是其他中文信息处理技术的基础，其技术成果主要应用在信息检索、机器翻译、语音合成、自动分类、自动摘要、自动校对等方面。目前已有清华、北大、哈工大、中科院、北京语言学院、东北大学、IBM 研究院、微软中国研

究院等科研院校积极参与中文分词研究，市场上也出现了海量分词系统等商业产品。随着更多的团队参与研究，中文分词技术将会变得更加成熟。

2.2 上下文无关文法与下推自动机

在形式语言与自动机理论中，若定义一个文法 G=(V，T，P，S)，其产生式规则均符合：A→w，其中 A∈V，w∈(V∪T)*，则称此文法是上下文无关的。这里，上下文无关文法（Content-Free Grammar, CFG）其“上下文无关”的原因是变量A总可以被终结符w所替换，因此就无需考虑变量A出现的上下文内容。如果一个形式语言是由 CFG 生成的，那么该形式语言就是上下文无关语言（Content-Free Language，CFL）[25,26]。

经验表明，高级程序设计语言的绝大多数语法结构都可以用 CFG 来描述。因此，高级程序设计语言的规范说明及其编译是 CFG 的一个重要应用领域。用来描述高级程序设计语言的巴克斯范式（Backus Normal Form，BNF）就是 CFG 的一种特殊形式。CFG 的这种表达能力，以及计算机系统对于处理 CFG 的适应性，使得 CFG 和 CFL 在计算机的各种相关语言的处理、相关理论的研究中占有非常重要的地位。通常，我们可以构造分析器（LL 分析器和 LR 分析器）来检验一个给定字串是否是由某个 CFG 产生的[27,28]。

下推自动机（Pushdown Automata，PDA）是形式语言与自动机理论中定义的一种抽象计算模型，它的形式定义为：PDA M=(Q，Σ，Γ，δ，q_0，Z_0，F)。其中，Q 是一个有穷状态集合；Σ是一个字母表，称为输入字母表；Γ是一个字母表，称为栈字母表；$\delta:Q\times(\Sigma\cup\{\varepsilon\})\times\Gamma\rightarrow Q\times\Gamma^*$是 M 的动作函数；$q_0$属于 Q，是初始状态；$Z_0$属于Γ，是一个特殊的栈符号，称为栈起始符号；F 包含于 Q，是终结状态集合[25]。

下推自动机比有限状态自动机复杂：除了有限状态组成部分外，还包括一个长度不受限制的栈；下推自动机的状态迁移不但要参考有限状态部分，也要参照栈当前的状态；状态迁移不但包括有限状态的变迁，还包括一个栈的出栈或入栈过程。下推自动机可以形象的理解为，把有限状态自动机扩展使之可以存取一个栈。CFG 与用空栈接受语言的 PDA 等价或与用终态接受语言的 PDA 等价，因此可以将 PDA 作为 CFL 的分析器[25,29]。

2.3 对话管理

对话管理在对话系统中处于非常重要的地位，其设计优秀与否关系到整个对话系统的性能。对话管理的主要任务是控制对话流程，帮助用户高效自然地完成对话。在进行人机交互时，对话系统的对话管理模块会充分考虑语境因素，通过

一定的控制策略，指导对话顺利进行。人机之间可以通过间接或直接的言语行为、新对话轮次的发起、对话的澄清和纠正、上下文历史记录和语言信息等因素获得相互理解[30,31]。在人机交互过程中，当对话识别出现错误或者用户提供的信息不完整，对话管理模块会根据预先定义好的应答模板对用户进行引导，使对话向着正确的方向推进。

通常，为了提高对话系统的对话管理质量，其对话策略往往是限定在一个特定的、词汇量有限的主题中。这些主题作为特定的应用领域还可以划分成若干应用场景，比如交通信息查询包括交通换乘和公交线路等。通过限定对话范围，可以提高对话系统的性能。一般来说，对话系统主要由三部分组成：自然语言理解模块，将用户的输入解析为语义信息；对话管理模块，根据对话策略确定应答模板；自然语言生成模块，组织生成应答信息和应答格式。

对话模型是对话管理的理论基础，它主要包括语法模型和规划模型两种。语法模型就是把对话看作一个状态机，状态有先后次序和限制条件，用类似语法描述的方法来表示。规划模型则假设人们在和别人或计算机系统对话时，头脑中早有一些目标和规划，对话的作用只是逐步确定或实现这些目标和规划，最后得到所期望的结果，因此对话由规划和规划识别构成[32]。

对话管理系统是对话模型理论的具体实现。目前常见的对话管理系统的设计方法主要包括以下三种[30,31]：

(1) 基于自动机的设计：这种方法把对话过程看成是自动机的状态转移过程，主要工作是设计自动机的状态和状态转移条件。这种对话管理结构要求程序设计人员在开发系统时预先给出所有可能的对话状态和用户操作，即所有状态之间的转移条件。如果用户的反应超出了系统给定的状态范围，将会导致对话无法正确进行。

(2) 基于槽的设计。这种方法把对话过程看作是对槽的填充过程，通过不断的人机交互，填充所有必要的信息槽，直至对话目标的实现。这种对话管理结构所实现的对话过程比较机械，人机交互的自然度较低，但实现复杂度较低，易于开发成熟的商业实用系统。

(3) 基于任务的设计：这种方法把对话过程看作是实现任务目标的过程，通过用户提出的新需求不断的更新系统的进度表，直至最终完成任务。这种对话管理结构的适用范围比较广泛，易于开发需求相对复杂的系统。

2.4 Web Service

Web Service 是一种轻量级的通讯技术，它可以从 Internet 上接收其它系统传递过来的请求并给出响应。开发者可以用任何编程语言，在任何平台上实现 Web

Service，只要这些 Web Service 符合标准，其他使用者就能够通过 Web 调用 API 进行查询和访问。

Web Service 广泛使用 XML 数据格式，通过 SOAP 在 Web 上提供的软件服务，使用 WSDL 文件进行说明，并通过 UDDI 进行注册。它们是组成 Web Service 平台的主要技术[33,34]：

(1) XML：扩展型可标记语言（Extensible Markup Language）。它是 Web Service 平台中表示数据的基本格式，具有易于建立、易于分析、平台无关、厂商无关等优点，是 SOAP 的基础。

(2) SOAP：简单对象存取协议（Simple Object Access Protocol）。它是 Web Service 的通信协议，用于在 Web 上实现服务器端与客户端的数据交换。SOAP 是 XML 文档形式调用方法的规范，它可以支持不同的底层接口，像 HTTP(S)或者 SMTP。

(3) WSDL：Web 服务描述语言（Web Service Description Language）。WSDL 文件是一个 XML 文档，用于说明一组 SOAP 消息以及如何交换这些消息。大多数情况下由软件自动生成和使用。

(4) UDDI：统一描述、发现与集成协议（Universal Description, Discovery, and Integration）。它是一个主要针对 Web 服务供应商和使用者的新项目，通过一种根据描述文档来引导系统查找相应服务的机制，可以有效的浏览并发现 Web Services 以及它们之间的相互作用。UDDI 利用 SOAP 发布和查找注册信息，采用 XML 格式封装各种不同类型的数据，以完成与注册中心的数据交换。

Web Service 的主要目标是跨平台的可互操作性。为了实现这一目标，Web Service 完全基于 XML、XSD（XML Schema）等独立于平台、独立于软件供应商的标准，是创建可互操作的、分布式应用程序的新平台。因此，Web Service 具有跨防火墙通信、不同语言应用程序集成、B2B 电子商务集成、软件与数据重用等优点，它为全球范围内的多系统、多应用领域的数据共享提供了可能[35]。

2.5 本章小结

本章详细介绍了研发基于问答系统的人机对话平台所涉及的关键技术，包括：中文分词、上下文无关文法与下推自动机、对话管理和 Web Service，并对它们的发展现状和技术细节进行了阐述。

第 3 章 人机对话平台的总体设计

人机对话平台根据系统的实际需求，通过首信公司综合信息资源库的资源支持和友好的对话管理策略，使用户能够采用自然语言问答形式进行智能、友好、自然的人机交互，获得丰富便捷的信息查询服务。

3.1 系统需求规定

3.1.1 系统功能需求

- 人机对话平台的用户接口要求操作方便、简洁美观。
- 系统采用模块化设计，可在不改变总体设计的情况下根据需要实现各模块的功能升级。
- 系统可根据不同需求定制具体应用，规则库、参数库、语料库和信息数据库应具备良好的扩展功能。
- 提供信息缓存功能，实现远程信息的本地化积累，以提高信息查询效率。
- 提供日志记录功能，保证对系统性能的维护。

3.1.2 系统输入输出需求

- 用户输入：自然语言文本
- 系统输出：符合 html 格式的自然语言文本，其中包括导航信息。

3.1.3 系统数据需求

- 规则库：为人机对话平台提供对自然语言信息进行句法分析和语义分析的匹配规则。
- 参数库：以参数词典形式存放，包括参数类型、参数标记符和参数词表，用于表示人机对话过程中的关键信息。
- 语料库：保存人机对话平台所有应答信息的模板，采用两级表格式存储。第一级存放应答信息的分类及对应索引值，第二级存放具体语言模板。
- 信息数据库：包含查询信息、消岐信息、导航信息等信息资源。

3.1.4 其他需求

人机对话平台通过 Web Service 提供服务，系统可以应用在 Windows XP、Windows Mobile、Linux 等操作系统中，用户能够通过桌面终端、移动终端、城市信息亭等方式进行访问。

3.2 层次框架模型

基于问答系统的人机对话平台的层次框架模型如图 3-1 所示，整个平台可以划分为三个层次：访问适配层、智能交互层和资源适配层，其主要任务为：

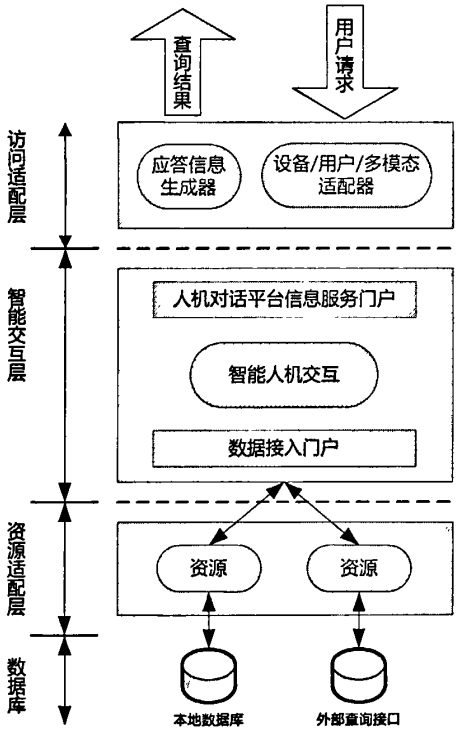

图 3-1 人机对话平台的层次框架

Figure 3-1 Hierarchy frame of human-machine dialogue platform

(1) 访问适配层要解决的问题是用户访问人机对话平台时可能存在不同设备、不同输入模态所带来的问题。

(2) 智能交互层为用户提供智能化、人性化的信息服务，并完成对人机对话平台信息服务门户和数据接入门户的标准和协议设计、研究与制定工作。

(3) 资源适配层是按照统一的协议和标准对不同来源的信息资源，以及规则库、参数库、语料库信息进行组织和管理，使智能交互层能够按照一套标准的协议对这些信息进行访问与操作。

3.3　系统结构模型

基于问答系统的人机对话平台包括自然语言理解、对话管理、自然语言生成、信息抽取、顶层控制和用户接口六个主要功能模块，其系统结构如图 3-2 所示。

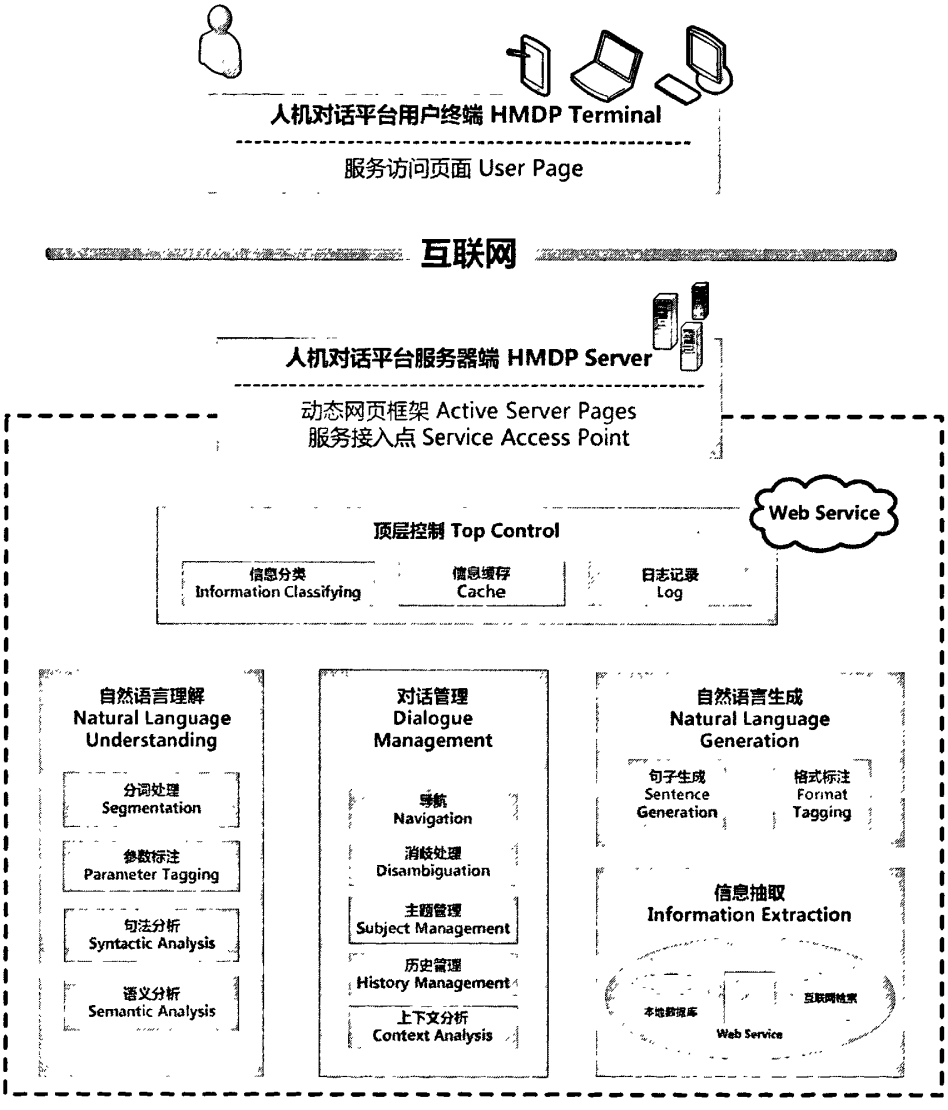

图 3-2 人机对话平台的系统结构

Figure 3-2 System structure of human-machine dialogue platform

3.4 本章小结

本章详细介绍了基于问答系统的人机对话平台的系统需求和总体设计思想，通过分析系统所提供的功能，以层次结构和模块结构的角度给出了人机对话平台的设计方案，为详细设计和具体实现提供了清晰的思路。

第 4 章　人机对话平台的实现

基于问答系统的人机对话平台按照功能划分可以包括自然语言理解、对话管理、自然语言生成、信息抽取、顶层控制和用户接口六个模块，下面分别介绍各功能模块的详细设计和具体实现。

4.1　自然语言理解模块

4.1.1 模块设计

4.1.1.1　主要任务

自然语言理解模块的主要任务是将顶层控制模块接收到的自然语言信息解析为计算机可以理解的参数信息，发送给对话管理模块。

4.1.1.2　模块结构及功能

本模块由分词处理、参数标注、句法分析和语义分析四个子模块组成。自然语言信息所构成的字序列在经过四个子模块后分别变成分词序列、带参数标注的分词序列、参数序列、带领域与场景信息的参数序列，最终以参数信息形式传递给对话管理模块。自然语言理解模块的结构模型如图 4-1 所示。

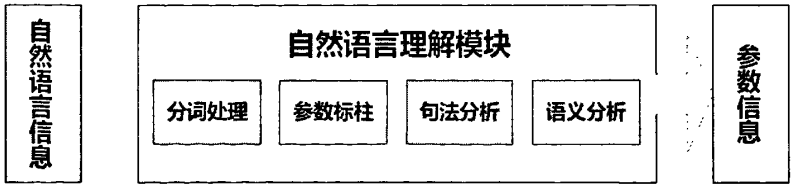

图 4-1 自然语言理解模块的结构模型

Figure 4-1 Structure model of natural language understanding module

4.1.1.3　处理流程

自然语言理解模块的处理流程如图 4-2 所示。

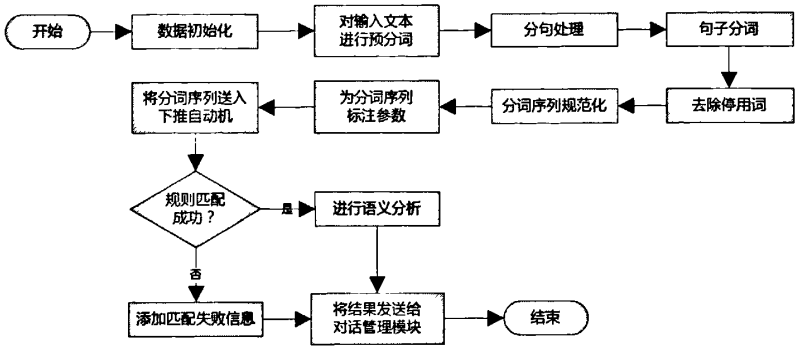

图 4-2 自然语言理解模块的处理流程图

Figure 4-2 Flow chart of natural language understanding module

4.1.2 分词处理

分词处理是对中文句子划分词边界的过程，在开始中文分词之前，系统会对字序列进行预处理，其主要目的是对字序列进行标记与拆分，以提高分词速度和准确率。在分词结束后，系统还会对分词结果进行后处理，其主要目的是优化分词结果，提高机器对信息的识别能力[36~39]。

4.1.2.1 预分词

在字序列中，经常会出现一些不易被分词算法正确切分的特殊信息，比如浮点数、IP 地址、电子邮件地址、时间和日期等。这些信息可能是影响计算机理解的重要参数，且通常具有固定的书写格式。为了使这些信息不被分词算法错误切分，需要提前对其进行分词处理，也就是进行预分词。

正则表达式是通过某种模式去匹配一类字符串的一种公式。使用正则表达式对字序列中常见的特殊信息进行匹配，将匹配成功的词按照信息类型标注上参数，分词算法就不用再对这些词进行切分了。预分词处理的匹配规则如表 4-1 所示。

表 4-1 预分词处理的匹配规则

Table 4-1 Matching rule of pre-segmentation

信息类型	正则表达式
浮点数	(-?\d+)(\.\d+)?
IP 地址	((2[0-4]\d|25[0-5]|[01]?\d\d?)\.){3}(2[0-4]\d|25[0-5]|[01]?\d\d?)
电子邮件	\w+([-+.]\w+)*@\w+([-.]\w+)*\.\w+([-.]\w+)*
带区号的电话号码	(\(\d{3,4}\)|\d{3,4}-)\d{7,8}

考虑到时间和日期的表示方法较多，可能出现中文数字混合表示或数字符号混合表示的情况，因此在预处理过程中使用 DateFind 和 TimeFind 方法专门对时间和日期信息进行预分词。

4.1.2.2　分句

在中文信息中，汉语词是不包含符号的，所以预先将字序列做分句处理不会破坏自然语言信息中的词结构。将具有断句功能的标点符号（逗号、句号、冒号、分号、问号、叹号、引号、书名号、括号、省略号、破折号）作为分句依据，对字序列进行句子拆分，可以减少每次中文分词过程要处理的信息量，加快分词速度。

由于一些可以预分词处理的信息（如 IP 地址）包含着影响分句的标点符号，所以预分词处理应该在分句处理之前完成，并且被预分词提前分好的词也将作为句子拆分的依据，以提高分句效率。

4.1.2.3　中文分词

(1) 传统分词方法存在的问题

对于基于字符串匹配的分词方法来说，传统的正向最大匹配算法和反向最大匹配算法在处理一些存在多元歧义的句子时缺乏良好的支持，比如正向分词产生的歧义切分“长春市/长春/节/贺词”和反向分词产生的歧义切分“长春/市长/春药店”。单纯的靠扩充词典信息量是无法解决这个问题的。

(2) 分词算法的选择

鉴于最大匹配算法在消岐方面的缺点，本系统选用了 KTDictSeg 开源分词组件，它采用的是一种改进型的正向分词方法—KaiToo 搜索开发的基于字典的中英文分词算法。该算法将匹配单词数和未匹配字数作为反映分词性能的依据，藉此产生一套针对分词结果的评价标准，以评价得分来调整分词结构，可以在一定程度上消除歧义。

下面是该分词算法的应用举例，“长春市长春节贺词”按照 KTDictSeg 的匹配顺序可以出现如下的分词组合：

a) 长春市/长春/贺词　　匹配单词数 3, 未匹配字数 1
b) 长春市/春节/贺词　　匹配单词数 3, 未匹配字数 1
c) 长春/市长/春节/贺词　匹配单词数 4, 未匹配字数 0
d) 长春/长春/贺词　　匹配单词数 3, 未匹配字数 2

可见组合 c 的未匹配字数最少，因此选择组合 c 的分词结果。

“长春市长春药店”按照 KTDictSeg 的匹配顺序可以出现如下的分词组合：

a) 长春市/长春/药店　　匹配单词数 3, 未匹配字数 0
b) 长春市/春药店　　匹配单词数 2, 未匹配字数 1
c) 长春市/春药　　匹配单词数 2, 未匹配字数 2
d) 长春市/药店　　匹配单词数 2, 未匹配字数 2
e) 长春/市长/春药店　　匹配单词数 3, 未匹配字数 0
f) 长春/市长/春药　　匹配单词数 3, 未匹配字数 1

g) 长春/市长/药店　　匹配单词数 3, 未匹配字数 1

h) 长春/市长/药店　　匹配单词数 3, 未匹配字数 1

i) 长春/长春/药店　　匹配单词数 3, 未匹配字数 1

可见组合 a 和组合 e 未匹配字数最少，匹配单词数相等，但组合 a 匹配顺序靠前，因此选取组合 a 的分词结果。

(3) 对 KTDictSeg 分词组件的一些修改

为了保证模块功能的正常运行，对该组件的开源代码做了以下修改：

- 删除组件中对电子邮件格式和浮点数格式的预处理，统一在预分词模块中对特殊格式的句子进行处理。
- 采用双词典结构的词条载入方式，除了具有大量基本词条的分词词典外，还包括一个由参数表构成的参数词典作为附加词典。
- 将词性标注的数据结构替换为参数标注的数据结构。

4.1.2.4 停用词处理

停用词是指文本中出现频率很高，但实际意义又不大的词，主要指副词、虚词、介词、语气词、连接词等，如“啊”、“呢”、“以及”、“这些”等。通常自然语言处理系统会在开始句法分析之前，将这些停用词去掉，降低它们对信息理解的干扰。

在人机对话平台中，系统维护了一张停用词表。它在中文分词结束时，对切分好的词序列进行遍历，以去掉其中包含的停用词。

4.1.2.5 规范化处理

由于汉语语法的复杂性、词汇的广泛性和常用语的不规范性，通常会导致同一个意思的表达有多种方式，比如表示“今天”可以说：今天，今日，今儿，今儿个...甚至汉语中表示“我”的词有几十种。同义词的广泛使用是影响自然语言处理性能的一个重要因素，不对其进行处理，会使自然语言处理系统的复杂度大幅增加。

为了降低同义词使用对系统性能的影响，可以在分词处理阶段引入同义词规范化操作，通过规范化映射表将分词结果中的同义词替换为机器便于识别的标准词。规范化映射表在内存中采用 Hashtable 存储结构，不规范的词信息可以通过 Hash 映射进行替换。规范化映射表的物理存储结构采用集合表的形式，即映射表中的每一行都是一个同义词组的集合，标准词是这个集合里的第一个词。下例为规范化映射表的一个元组表示：

(今天，{今天，今日，今儿，今儿个})

它包含 2 个分量，分别对应了表属性中的标准词和同义词集合。

此外，通过预分词处理得到的一些分词信息，比如日期、时间等，也会被规范化为标准格式，方便系统识别。

4.1.3 参数标注

参数标注的目的是为采用文法规则匹配的句法分析方法提供必要的参数标记。对于不同领域不同应用场景的用户提问来说，系统需要从中抽取出查询答案的关键信息，这些信息的集合就是系统预定义的参数集，比如在查询天气时，可以问“今天北京天气怎么样？”，这句话包含了两类参数，它们分别是时间参数“今天”和城市参数“北京”。

在基于规则的句法分析方法中，标记参数对计算机理解自然语言信息起到了较大的辅助作用，所以参数标注是句法分析前的重要准备工作。参数标注与词性标注类似，不过标注的内容不是词性，而是词所包含的参数类型。比如“天安门”在进行参数标注时将被标注上两个参数：[地点]和[景点]，分别对应“交通”和“旅游”两个领域的信息查询；“怎么样”则不会被标注为参数。一个词可能不具备任何参数类型，也可能具备多种参数类型，这是由领域和应用场景决定的。

人机对话平台使用参数词典来记录参数信息，它包含若干个参数表，每个参数表保存了一种参数类型的全部参数记录。根据系统对参数表存储格式的定义，它应包括参数类型、参数标记符和参数词表，如表 4-2 所示。

表 4-2 参数词典的存储格式

Table 4-2 Storage format of parameter dictionary

字段名	说明
Para_id	参数标记符，由参数词典定义的与参数类型等价的数字符号
Para_name	参数表所属的参数类型
Word_id	参数词表中的参数索引号
Word_name	参数词表中的参数记录

当系统初始化时，首先会载入分词词典中的词条，随后读取参数词典中的参数表信息，记录参数类型与参数标记符，并对每一个出现在参数词表中的词添加参数记录。通过参数词典来维护参数表，可以集中管理参数信息，便于多领域多应用场景的参数扩展。参数词典的存储结构及应用举例如图 4-3 所示。

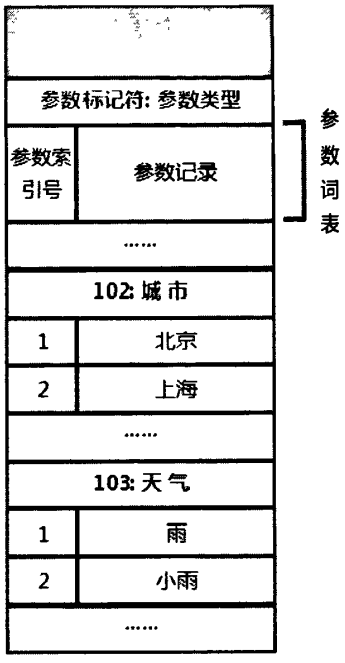

图 4-3 参数词典的应用举例

Figure 4-3 Example of parameter dictionary

4.1.4 句法分析

4.1.4.1 基于文法规则匹配的句法分析思想

自然语言是由无限多的句子构成的集合，而计算机的存储结构决定了它不可能对每一个句子都独立的给出识别方法，因此计算机需要记录的是有限的词集和自然语言的语法规则。

乔姆斯基范式中的上下文无关文法是一种被广泛应用于自然语言句法分析的文法，它具备很强的自然语言生成能力，可以用来描述任何一种递归可枚举的语言。根据特定需求设计的上下文无关文法，可以将自然语言按照句法结构归纳成符合分析器识别要求的规则，这些规则的集合组成了自然语言处理系统的规则库。通过定义好的文法规则可以将处理自然语言信息的过程看做是一颗句法树的规约过程。规则库中存放的每一条上下文无关文法的产生式规则就是一颗句法树的生成规则[40,41]。

根据形式语言与自动机理论，任何上下文无关语言都存在与之等价的下推自动机，因此系统可以通过设计好的上下文无关语言生成识别该语言的下推自动机。当一段语言信息作为输入进入了下推自动机，就会根据状态转移条件推动自动机执行，识别结束时所停留的状态结点若是终止结点则表示成功的将输入信息规约成规则，信息识别成功，否则即识别失败[42]。

在人机对话平台中，将对话管理模块各领域及应用场景所需要的重要信息以参数形式标注在分词序列的属性中，下推自动机把参数类型和一些具有标志意义的词作为状态转移条件，完成自然语言的文法规则匹配。这种方法通过文法产生式生成可以识别特定语言信息的规则，并将规则作为生成下推自动机结点和状态转移条件的依据，利用不断扩充的规则集来调整自动机的结构和状态，以提高对语言信息的识别效率。这就是基于文法规则匹配的句法分析方法[43,44]。

4.1.4.2 文法设计

定义文法 G = ({S，E, W, L}，{(,), [,], {, }, <, >, $, num, word}，P，S)

其中 P = {

S→E，

E→EE | (W) | [] | [L] | {} | {L} | <E> | word，

W→E$E | W$E，

L→num }

L(G)={w|w∈T*,S⇒w}为 G 产生的语言（Language），∀w∈L(G)，w 为 G 产生的一个句子（Sentence）。在人机对话平台中，将 G 产生的句子称为规则，并通过使用括号等标志来确定规则匹配的优先级，实现对上下文无关文法二义性问题的处理。

4.1.4.3 关于产生式的解释

根据上述文法 G 所包含的产生式 P，可以对自然语言集进行规则抽象。对于符合文法 G 的规则 $A_1,A_2,...,A_n$ ，包括以下几种信息匹配情况：

(1) 连接：“A_1A_2”表示两个规则 A_1 和 A_2 连接成一个新的规则 A_1A_2。

(2) 选择：“($A_1$$A_2$$...$$A_n$)”其中 n≥2，表示匹配时只需满足 $A_1,A_2,...,A_n$ 其中一条规则即可。

(3) 可去除：“<A_1>”表示可以满足规则 A_1，也可以不满足 A_1。

(4) 参数标记：“[L]”表示一个参数类型为 L 的词。

(5) 任意参数标记：“[]”表示一个任意参数类型的词。

(6) 组标记：“{L}”表示一组参数类型均为 L 的词。

(7) 任意组标记：“{}”表示一组任意参数类型的词。

通过文法产生式可以定义相应的规则表达式，比如

([101][102]$[102][101])<的>天气<怎么样>

其中[101]代表日期参数，[102]代表城市参数，($A_1$$A_2$)表示在进行规则匹配时 A_1 和 A_2 只能选择一个来匹配，<>表示所括选内容是可以去除的。由此，前面给出的规则范例可以匹配到的句子包括：

(1) 北京今天天气怎么样

(2) 明天天津的天气怎么样

(3) 广州昨天的天气

(4) 后天上海天气

……

通过文法规则，可以用少量的规则表达式来识别大量的句子组合。

4.1.4.4 规则表达式的派生过程

在为人机对话平台定义规则时，需要满足文法产生式的要求，即这条规则可以通过文法产生式派生出来。

下例给出了一条规则的派生过程：

规则：([101][102]$[102][101])<的>天气<怎么样>

S => E => EE => (W)E => (E$E)E => (EE$E)E => ([L]E$E)E

=> ([101]E$E)E => ([101][L]$E)E => ([101][102]$E)E

=> ([101][102]$EE)E => ([101][102]$[L]E)E

=> ([101][102]$[102]E)E => ([101][102]$[102][L])E

=> ([101][102]$[102][101])E => ([101][102]$[102][101])EE

=> ([101][102]$[102][101])<E>E

=> ([101][102]$[102][101])<的>E

=> ([101][102]$[102][101])<的>EE

=> ([101][102]$[102][101])<的>天气 E

=> ([101][102]$[102][101])<的>天气<E>

=> ([101][102]$[102][101])<的>天气<怎么样>

与上述派生过程相对应的派生树如图 4-4 所示。

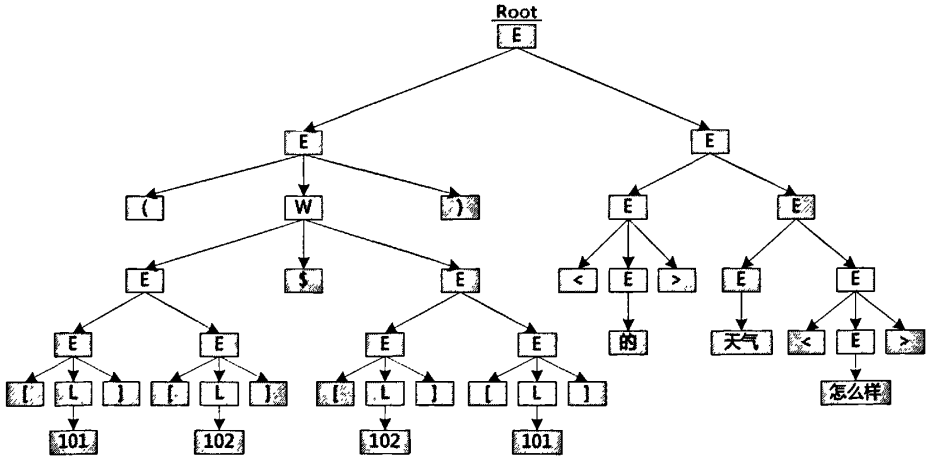

图 4-4 规则派生树的举例

Figure 4-4 Example of rule derivation tree

4.1.4.5　下推自动机的生成方法

当系统在加载规则时，会先将规则信息保存在一个二叉树结构中，该二叉树采用孩子兄弟表示法存储，如图 4-5 所示。下推自动机由一个初始状态结点开始，通过对规则存储树的深度优先遍历完成状态结点的添加。

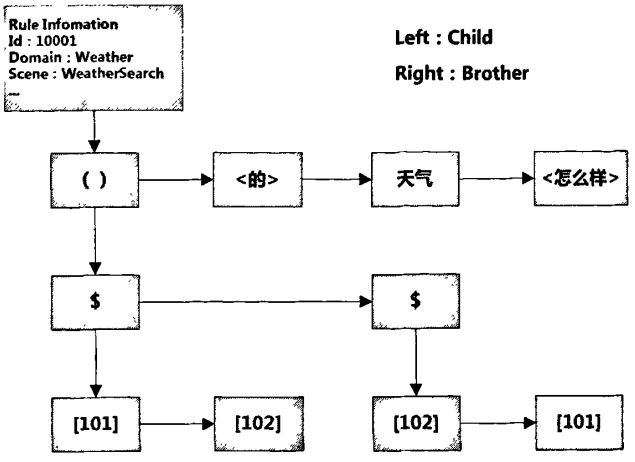

图 4-5 规则存储树的举例

Figure 4-5 Example of rule storage tree

在添加状态结点时，生成函数需要读入所有可以作为当前结点的前置结点的 ID 列表 Head（Arraylist），返回可以作为后续结点的前置结点的 ID 列表 Tail（Arraylist），由此完成状态结点的连接。下推自动机依靠结点间的连接关系，通过状态转移方法实现状态转移。根据产生式的定义，状态结点的连接方法包括三种情况：连接、选择、可去除，如图 4-6 所示。

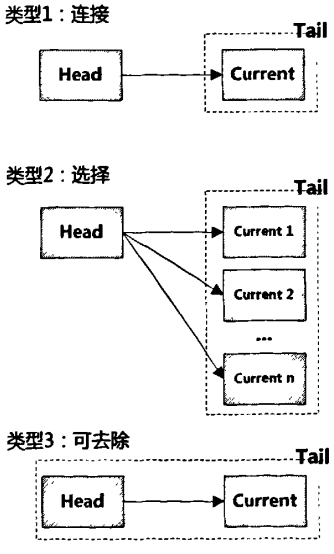

图 4-6 状态结点的连接方法

Figure 4-6 Connecting method of state node

4.1.4.6 状态转移方法的选择

句法分析模块的下推自动机采用带结点自回环的树状结构，除组标记生成的状态结点自回环以外，没有其他的环形结构。因此在自然语言信息识别成功时，词序列所经历的状态转移路径是唯一的，即只对应一条规则。对于一个结点的一次词匹配，下推自动机会按照固定的顺序选择状态转移方法，如果一种状态转移方法的词匹配失败，将通过回溯法回到前面结点，选择新的状态转移方法继续匹配，直至匹配成功或所有结点的状态转移方法均匹配失败。

对状态转移方法的选择顺序为：

(1) Symbol，具有断句功能的标点符号，由于规则中不包含符号，这里是通过一个全局符号集对输入进行判断，符号为输入终止标志。

(2) Word，输入词

(3) {}，任意组标记

(4) {L}，与输入词参数类型一致的组标记

(5) []，任意参数标记

(6) [L]，输入词的参数类型

系统会在选择方法时跳过当前结点不包含的状态转移方法。

4.1.4.7 参数提取方法

当输入结束时，下推自动机若停留在终止结点上，则表示成功将输入信息规约成规则，即该信息可以被计算机理解。此时系统将会得到唯一的一条状态转移路径，对于这条路径来说，其终止结点只能包含唯一的一个规则标识。若在生成下推自动机时，同时出现两条规则拥有同样的状态转移路径，则表示它们是等价的，系统会根据覆盖原则使用新的规则信息来覆盖旧的。

下推自动机通过回溯法将在终止结点获得的规则标识传递给本次识别产生的状态转移路径上的每一个状态结点。这些结点会从自己的参数表中查询该规则是否登记了参数提取信息，若发现存在登记信息，自动机将会把该结点的输入词及参数类型添加到参数序列中。

4.1.5 语义分析

一条规则代表了一组确定的没有歧义的句子，因此在规则的附带信息中会包含用于机器理解的语义信息。规则匹配成功时，下推自动机的输出包括规则标识和参数序列。语义分析模块会根据规则标识在规则集中找到该规则携带的语义信息，将其和参数信息一起提供给对话管理模块。

语义信息包括领域信息和应用场景信息，以“领域|应用场景”格式存放，例如 Weather|WeatherSearch。

4.1.6 模块实现

4.1.6.1 类结构图

自然语言理解类（NLU）、正则表达式类（myRegex）、数据加载类(myDict)和下推自动机类(PDA)的类结构图如图 4-7 所示。

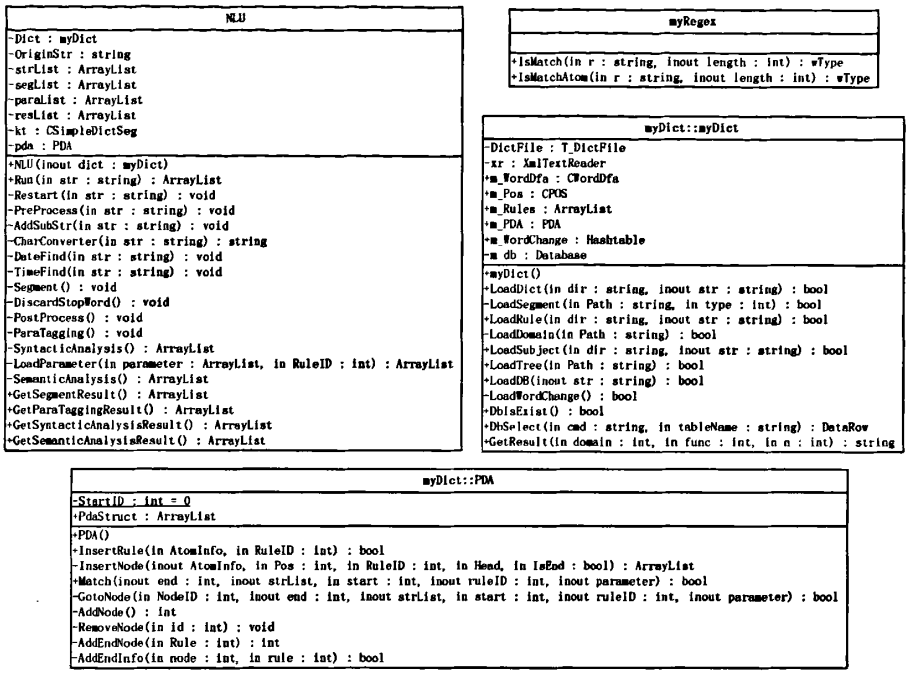

图 4-7 自然语言理解模块的类结构图

Figure 4-7 Class diagram of natural language understanding module

4.1.6.2 主要数据结构

自然语言理解模块的主要数据结构如表 4-3 至表 4-9 所示。

表 4-3 RuleAtomInfo：规则元素的数据结构

Table 4-3 RuleAtomInfo：data structure of rule element

变量名	变量类型	说明
id	int	规则元素号
parent	int	该规则元素的前置元素
Children	ArrayList	该规则元素的后置元素
parameter	string	该规则元素的携带参数信息
info	string	该规则元素内容
AtomType	RuleAtomType	该规则元素的类型

表 4-4 WordInfo：带参数标注的分词信息数据结构

Table 4-4 WordInfo：data structure of segmentation information with parameter tagging

变量名	变量类型	说明
str	string	词信息
paratype	Hashtable	该词包含的所有参数类型

表 4-5 Rule：规则的数据结构

Table 4-5 Rule：data structure of rule

变量名	变量类型	说明
id	int	规则号
RuleInfo	ArrayList	规则中各元素的数组表示
semantic	string	语义信息

表 4-6 ParameterInfo：参数信息的数据结构

Table 4-6 ParameterInfo：data structure of parameter information

变量名	变量类型	说明
para	string	该参数信息的参数类型
info	string	该参数信息对应的词信息内容
type	RuleAtomType	该参数信息对应的状态转移类型

表 4-7 PdaNode：状态结点的数据结构

Table 4-7 PdaNode：data structure of state node

变量名	变量类型	说明
ID	int	状态结点号
End	bool	是否是终止结点
WordJump	Hashtable	状态转移的词映射信息
TypeJump	Hashtable	状态转移的参数类型映射信息
TypeLoop	int	状态转移的任意组的映射信息
TypeLoopJump	Hashtable	状态转移的组映射信息
RuleID	int	该结点若是终止结点，所对应的规则号
Parameter	Hashtable	该结点所包含的所有规则的参数提取信息

表 4-8 DateType：日期信息的数据结构

Table 4-8 DateType：data structure of date information

变量名	变量类型	说明
date	DateTime	日期信息的绝对值
dr	int	日期信息的相对值
tr	string	日期信息的近期表述，如今天、明天、后天等
month	int	日期信息的月份信息
year	int	日期信息的年信息
holiday	string	日期信息的节日信息
ad	string	日期信息的相关信息，如上午、下午等

表 4-9 TimeType：时间信息的数据结构

Table 4-9 TimeType：data structure of time information

变量名	变量类型	说明
time	DateTime	时间信息的绝对值
mr	int	时间信息的分钟相对值
hr	int	时间信息的小时相对值
tp	int	时间信息中表示一段时间的时间间隔绝对值
timetype	bool	时间信息类型，属于不确定时间，该值为 false

4.1.7 规则的存储格式设计

根据文法产生式生成的规则是以表达式形式存在的句子，这种形式的字符串不能被系统直接载入，需要将其转换为 XML 标签格式，由 XML 分析器载入系统。使用 XML 标签可以为规则添加额外的标记，作为帮助系统进行参数提取的依据。另外，也可以对一些具体词添加参数提取标记，增加规则应用的灵活性。在规则标签尾部包含了与规则对应的语义分析信息。XML 标签类别如表 4-10 所示，与系统的规则元素枚举类型 RuleAtomType 的对应关系如表 4-11 所示。

表 4-10 XML 的标签定义

Table 4-10 Tag definition of XML

标签	说明
<Rules>	规则集标志
<Rule>	规则标志
<w>	词信息
<t>	参数类别信息
<tl>	组类别信息
<g>	选择结构标志
<o>	选择结构的选择项标志
<e>	可去除结构标志
<goto>	语义信息
除 rules、rule 和 goto 以外的标签均可以加 para 参数指定参数提取信息类别。	

表 4-11 XML 标签与 RuleAtomType 的对应关系

Table 4-11 Relationship between XML tag and RuleAtomType

类型	说明
word	<w></w>
type	<t></t>
typeloop	<tl></tl>
group	<g></g>
or	<o></o>
except	<e></e>
root	<Rule></Rule>

下面是规则的表达式格式和 XML 标签格式的应用范例：

表达式格式：([101][102]$[102][101])<的>天气<怎么样>

XML 格式：<Rule><g><o><t para="Date">101</t><t para="City">102</t></o><o><t para= "City">102</t><t para="Date">101</t></o></g><e><w>的</w></e><w para= "Weather">天气</w><e><w>怎么样</w></e><goto>Weather|WeatherSearch</goto></Rule>

在识别句子“北京今天天气怎么样”时，系统根据上述规则得到的信息为：

(1) 句法分析的参数提取：City|北京；Date|今天

(2) 语法分析：Weather|WeatherSearch（领域|应用场景）

4.1.8 XML 辅助生成工具设计

XML 辅助生成工具是用来将规则表达式转换为 XML 标签的工具。在工具生成标签过程中，往往需要对规则所包含的确定词进行切分操作，因此 XML 辅助生成工具使用了与人机对话平台一致的分词处理方法。

辅助生成工具的主界面上提供了规则输入、XML 输出、参数提取信息添加和语义信息添加四个功能，如图 4-8 所示。在生成 XML 标签之前，工具会通过协议栈对用户输入规则的书写正确性进行校验。

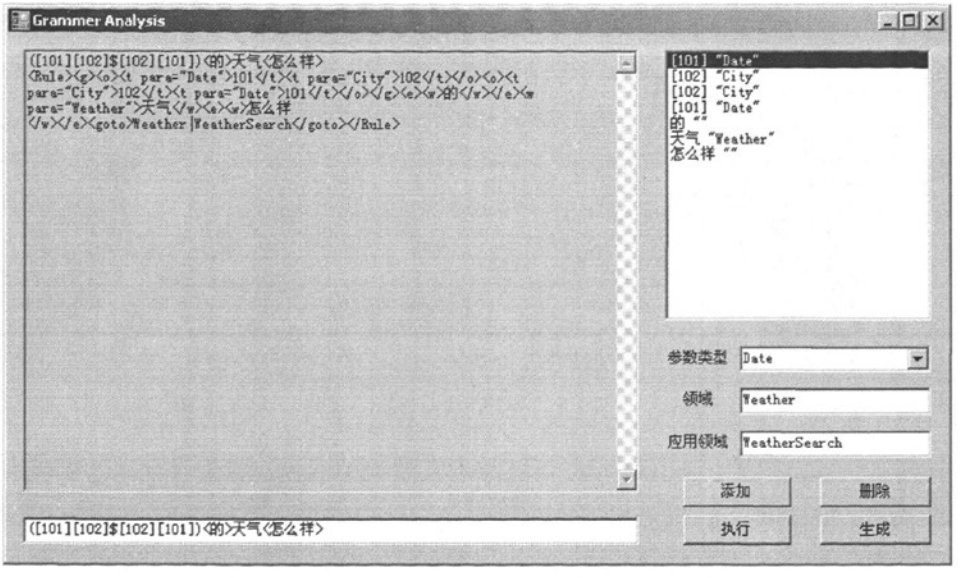

图 4-8 辅助生成工具的主界面

Figure 4-8 Main interface of assistant create tool

当辅助生成工具对规则格式进行匹配校验时，系统会根据匹配情况给出相应的提示，如表 4-12 所示。若规则表达式输入错误，系统会指出错误位置。

表 4-12 规则匹配的提示信息

Table 4-12 Notices of rule matching

提示信息	说明
请输入规则表达式	执行格式校验时，规则输入栏为空
规则中存在非法字符	规则中只允许出现数字、英文、中文以及[][{}<>()$符号
所输入参数类型超出范围	规定参数类型标记符从 100 开始至参数词典中定义的最大值
$之前的表达式格式错误	(与$之间或两个$之间表达式规约错误
()中未包含分隔符$	()中未出现$符号
存在未被()包含的分隔符$	$未被包含在()之内
以数字表示的参数类型前面要加[或{	数字字符前面不是[或{符号
[]中只能包含以数字表示的参数类型或为空	[]中存在非数字字符
{}中只能包含以数字表示的参数类型或为空	{}中存在非数字字符
存在未匹配的左侧符号	缺少)]}>与已输入的([{<匹配
存在未匹配的右侧符号	缺少([{<与已输入的)]}>匹配
表达式匹配成功	规则表达式匹配成功

4.2 对话管理模块

4.2.1 模块设计

4.2.1.1 主要任务

根据自然语言理解模块识别出的提问信息，通过信息提取模块找到查询答案，将其发送给自然语言生成模块。对于查询信息不全的提问，对话管理模块会使用相应的对话管理策略引导用户补全信息，返回答案。

4.2.1.2 模块结构及功能

本模块由上下文分析、主题管理、历史管理、消岐处理和导航五个子模块组成。对话管理模块的功能是控制和指导计算机与用户的对话过程，提供人机交互的对话策略，使计算机可以根据识别到的信息做出对话引导、历史参照、信息确认、信息查询、主题切换等响应。当一次人机交互开始后，对话过程会基于轮次进行（turn-based），即用户输入一句，对话管理模块就会做出一次处理，并结合对话历史判断用户意图，给用户产生合理的应答，直到对话结束。对话管理模块的结构模型如图 4-9 所示。

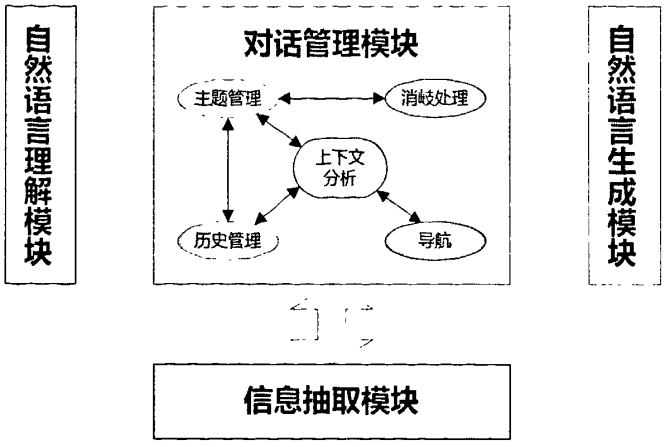

图 4-9 对话管理模块的结构模型

Figure 4-9 Structure model of dialogue management module

4.2.1.3　处理流程

对话管理模块根据用户当前输入的内容以及内部对话历史来决定是进行信息查询，还是对用户进行对话引导。其处理流程为：

(1) 从自然语言理解模块中接收参数信息和语义信息；

(2) 处理输入信息，保存对话状态和对话信息，判断对话流程；

(3) 明确用户提问后，对查询信息进行消岐处理；

(4) 向信息抽取模块查询答案；

(5) 将应答信息发送给自然语言生成模块，使对话沿着用户期望的方向前进。

通常，一次交互过程需要经过几轮的（1）、(2)、(5）的对话引导与信息确认，才能最终从信息抽取模块返回确定的查询结果，如图 4-10 所示。

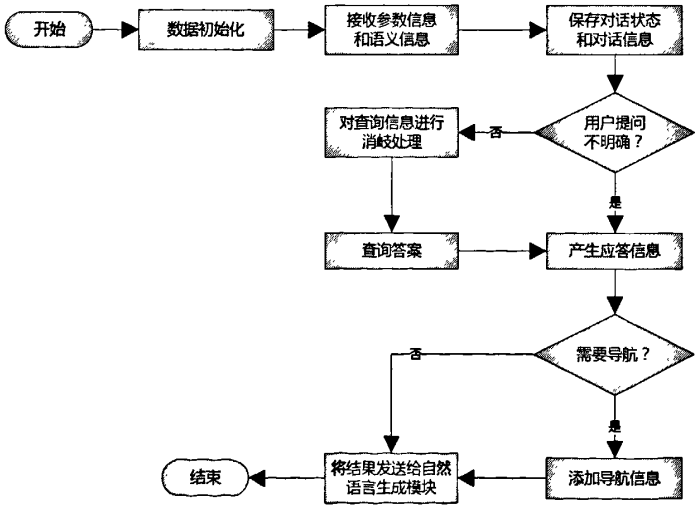

图 4-10 对话管理模块的处理流程图

Figure 4-10 Flow chart of dialogue management module

4.2.1.4 对话管理的应答方式

对话过程中会遇到不同的应答情况，主要包括以下几种：

(1) 提示信息。人机交互过程中，若需要用户对歧义信息做确认或提示用户补全信息的关键参数时，系统就会根据上下文调用相应的提示信息模板，生成提示信息来引导对话。

(2) 查询信息。人机对话平台的主要任务就是向用户提供信息查询功能，系统通过对话引导获得了用户确切的查询意图，就可以向信息抽取模块发起本轮对话的信息查询，并将查询结果反馈给用户。

(3) 错误信息。对话过程并不总是能够顺利进行，其间可能会出现一些问题，比如用户输入的问题不能被计算机理解，信息查询错误等，对话管理模块需要对这些错误信息做出响应，向用户给出错误提示，将对话引导回出错前的状态中。

(4) 帮助信息。在对话管理模块中的每个主题都有相应的帮助模板，当计算机与用户经历了多轮交互或用户提出帮助申请时，系统将会给出帮助信息，提高用户的查询效率。

(5) 导航信息。一些由人机对话平台提供的以引导对话为目的的系统信息，它们以链接形式出现，通过链接附带的对话标识可以直接返回系统能够理解的对话信息。

应答方式的确定与生成是对话管理模块和自然语言生成模块共同完成的，对话管理决定了对话动作与参数，自然语言生成则负责将其组合成完整的应答语句，与用户进行对话。

4.2.2 上下文分析

上下文分析是对话管理的核心部分，它通过与对话管理模块中其他几个子模块的互动，完成以下几项工作：

(1) 分析人机对话过程的交互历史，确定用户本轮对话的主要意图；

(2) 根据主题选择策略和切换方法，确定当前的对话主题；

(3) 确定是否对应答信息提供导航支持；

(4) 提供逐步回退的对话取消机制。

4.2.2.1 主题选择策略

由于人机对话平台是多领域多应用场景的多主题对话系统，每次对话都是围绕某个特定主题展开的，所以首要工作是对话的主题选择。对话主题的选择是通过交互历史信息和当前语义信息来决定的，在用户意图没有发生明显转变时，对话将会默认在交互历史中记录的前次对话主题中进行。

当用户通过人机对话平台发起对话时，系统首先根据用户的提问信息来确定主题，主题选择依据是由自然语言理解模块的语义分析提供的，这种主题确定方式给了用户较高的自由度。当语义分析无法得到用户意图或是分析结果与对话过程的交互逻辑不匹配，系统会给出帮助信息向用户提示人机对话平台的所有领域查询功能并给出领域选择的导航信息，以保证用户能够快速展开人机对话。主题选择流程如图 4-11 所示。

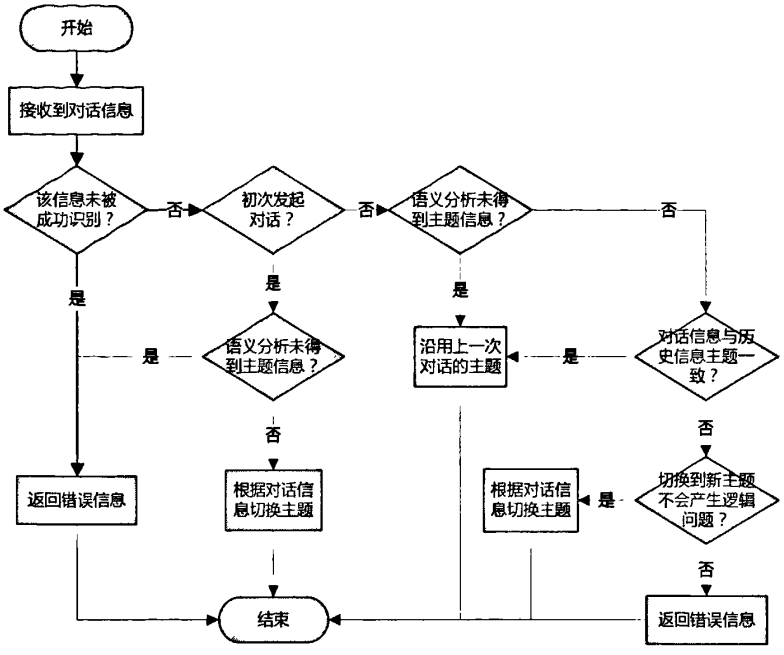

图 4-11　主题选择的流程图

Figure 4-11 Flow chart of subject choice

4.2.2.2　主题切换方法

人机对话平台采用了统一的历史管理方法，各主题的历史信息保存在一个全局的数据结构中。对话过程中用户有可能从一个主题过渡到另一个主题，上下文分析模块通过历史管理模块提供的主题切换方法，对历史信息中的相关主题标记和记录信息进行修改，实现主题切换。

4.2.2.3　对话取消机制

根据历史管理模块提供的对话信息记录，用户可以通过逐步回退的取消机制来取消某次输入的对话信息。取消机制如图 4-12 所示。

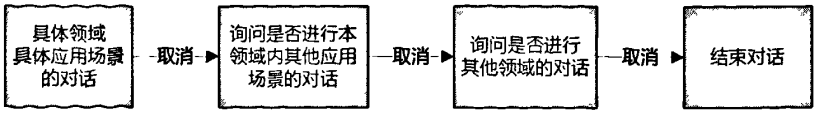

图 4-12　对话取消机制的流程图

Figure 4-12 Flow chart of dialogue cancel mechanism

4.2.3 主题管理

4.2.3.1 知识表示方法

现有的对话管理系统一般采用有限状态网格、表格结构和树结构三种知识表示方法来支持对话管理。本系统选用树结构的知识表示方法实现对多领域多应用场景信息的表示。

在树结构的知识表示方法中，领域及应用场景所属的主题由主题树来表示，叶子结点是对特定领域特定应用场景信息项的描述，因此主题树的叶子集就是该主题的概念组成，记录了该主题所有影响信息查询的参数内容。这种表示结构不仅可以直接给出各信息项的状态，还可以在对话过程中对信息项状态进行归纳总结，根据对特定子树的遍历结果确定是返回用户查询信息还是引导用户给出缺少的信息项。由于采用主题树进行主题管理的方法是在人机交互过程中动态的获取信息项，无需像有限状态网格那样预先定义状态和转移条件，所以它在设计上相对简单，便于添加新的领域和应用场景，适合多主题人机对话系统的主题信息扩展[45, 46]。

4.2.3.2 知识信息的存储结构

主题管理模块的系统结构是主题森林结构，即每个对话领域对应于一颗主题树，不同的主题树组成了主题森林，主题树结构如图 4-13 所示。

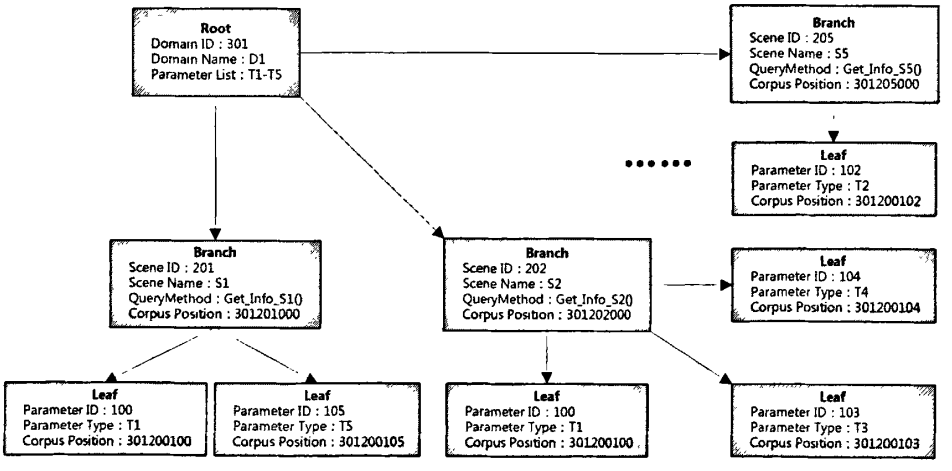

图 4-13 主题树的结构图

Figure 4-13 Structure of subject tree

每个主题树对应了一个领域，管理与该领域相关的所有对话。每个分支结点对应了一个应用场景，管理与该场景内容相关的所有对话，接收与该场景相关的用户输入，对与该场景相关的信息进行消岐处理，给出相应的提示信息。对于主题树来说，不同类型的结点有着不同的作用：

(1) 根结点：表示该主题树的领域类型。结点信息中记录了领域标识符和该领域涉及的所有参数类型。

(2) 分支结点：表示该领域的一个应用场景。结点信息中记录了该场景参数信息收集完毕时向信息抽取模块查询答案的方法，以及对应的应答信息模板在语料库中的位置。

(3) 叶子结点：表示该应用场景的一个信息项，也就是能够确定用户对话意图的一个参数信息类型。结点信息中记录了参数类型标识和缺少该信息时的应答信息模板在语料库中的位置。

4.2.3.3 主题管理的工作内容

人机对话过程中，主题管理的工作是结合历史管理来进行的。当对话内容围绕一个领域的具体场景展开时，主题管理通过对该场景子树的遍历检查各叶子结点的信息获取情况，历史管理将获取到的新信息项添加到历史信息表中，并对已存在的旧信息项进行更新。当对话内容在相同领域不同场景间切换时，主题管理需要将对话焦点指向目标场景子树，历史管理仍然是对历史信息表进行添加和更新操作。当对话内容从一个领域跳转到另一个领域时，主题管理需要将主题焦点指向目标主题树，历史管理则是在历史信息表中清除所有新领域不需要的信息项。

当主题管理模块发现某个场景的信息项已收集完毕时，会向消岐处理模块申请对信息项进行消岐，之后根据主题树分支结点上携带的信息查询方法，将消除歧义的参数信息发送到信息抽取模块查询答案。

4.2.4 历史管理

4.2.4.1 历史信息管理的必要性

对话系统的交互历史记录了人机对话的语义交流过程，通过对历史信息的管理与分析，可以加强系统的信息识别能力。在人与人的对话过程中，人们交流信息时总会保留一些对历史信息的记忆，并在后续对话过程中默认之前提到的这些信息已被对方了解，这是人际间交流的一种普遍现象。因此，在人机对话系统中加入历史管理模块，是实现基于问答系统的人机对话平台必不可少的一个功能。

计算机自身特点决定了其具有较强的信息存储功能，可以将交互历史信息从对话一开始就完整的保存下来。但对于人的记忆特点来说，往往只会记住较近的一些历史信息或是与对话内容最紧密的信息。因此在人机交互过程中，保留全部的历史信息并不一定有意义，需要对如何管理历史信息制定策略。

4.2.4.2 历史管理策略

人机对话平台采用的是一种可追溯历史的主题相关历史管理策略。其关键点

有两个，分别是主题相关和可追溯历史。

(1) 主题相关

主题相关是指系统维护了一张历史信息表，表中记录了人机对话过程在进行到当前主题以后所有提及的参数信息。这些历史记录在对话出现信息不全时，可以作为引导对话进行的重要依据来使用。

当围绕一个主题进行对话时，对历史信息表的维护操作包括两种：添加和替换。添加操作是将对话中新提到的信息添加到历史信息表中，以提高系统引导对话的能力。替换操作是用新提到的信息替换类型相同的旧信息，以确保对话管理的时效性。当人机交互需要切换主题时，历史管理模块会对历史信息表进行检索，保留新主题包含的参数信息，删除其他的历史信息。历史信息表实现了对当前对话主题的历史管理，其结构如图 4-14 所示。

<table>
<tr><td colspan="5">焦点信息</td></tr>
<tr><td colspan="5">主题焦点ID</td></tr>
<tr><td colspan="5">当前对话领域</td></tr>
<tr><td colspan="5">当前对话场景</td></tr>
<tr><td colspan="5">上次对话信息</td></tr>
<tr><td colspan="5">用户提问</td></tr>
<tr><td colspan="5">系统应答</td></tr>
<tr><td colspan="5">1.等待用户答复(信息补全、消岐确定)
2.完成查询
3.出现错误
4.给出帮助
5.存在导航信息</td></tr>
<tr><td colspan="5">历史参数信息</td></tr>
<tr><td>参数1</td><td colspan="2">参数2</td><td>参数3</td><td>……</td></tr>
<tr><td>语义1-1</td><td>语义2-1</td><td>语义2-2</td><td>语义3-1</td><td>……</td></tr>
</table>

图 4-14 历史信息表的结构图

Figure 4-14 Structure of history information list

在记录历史参数信息时，参数信息由参数类型与具体语义二重映射结构存储，这是因为在一个主题对话中，同一种参数类型可能表示多种语义，比如地点参数既可以表示出发地也可以表示目的地。

(2) 可追溯历史

可追溯历史是指系统提供了一个用于追溯历史信息的栈，用于记录所有的交互历史。栈元素的存储结构如表 4-13 所示。

表 4-13 栈元素的追溯信息存储结构

Table 4-13 Storage structure of trace information of stack element

变量名	变量类型	说明
SubjectID	string	焦点主题的 ID
Domain	string	对话领域
Scene	string	对话场景
UserSay	string	用户输入
Reply	string	系统应答输出
ReplyInfo	Reply_Enum	应答输出的枚举类型
ReplyAttach	string	应答附加信息，比如枚举类型为等待用户补全信息，附加信息为待补全参数类型[城市]
Para_add	arraylist	本次添加的历史信息，数组元素采用 string 类型，其格式为(参数类型|具体语义|参数内容)
Para_del	arraylist	本次删除的历史信息，数组元素采用 string 类型，其格式为(参数类型|具体语义|参数内容)
Para_update	arraylist	本次更新的历史信息，数组元素采用 string 类型，其格式为(参数类型|具体语义|新参数内容|旧参数内容)

对于人机交互过程中的每一次对话，栈元素会记录用户的对话内容，并根据历史信息表的变化记录添加、删除和替换的参数，以及主题变更信息。通过记录的变更信息和对话内容，可以还原前次对话的历史信息表，实现对交互历史的追溯功能。

4.2.5 消岐处理

4.2.5.1 消岐处理的目的

由于人机对话平台的输入信息是采用自然语言表达方式，往往会出现对某一事物的表达存在歧义的现象，而计算机数据库中存储的信息都是规范的标准数据，因此必须在对话管理过程中对信息进行消岐处理。通过消岐，可以使信息抽取模块的查询功能更加有效，使系统与用户交互起来更加友好，从而适应用户不同风格的提问。

在人机对话平台的自然语言理解模块中，分词处理过程的后处理阶段有针对分词结果的规范化处理，那是为了提高参数标注和句法分析性能而采取的有限数据的同义词替换策略，消岐功能只是针对词法。而在对话管理阶段进行的消岐处理，则是针对语义和机器理解上的消岐，其功能更加复杂。

4.2.5.2 歧义类型的定义

在消岐处理模块中，歧义类型主要包括以下四种：

(1) 别名歧义。由于一个标准名在日常生活中往往会存在同义的别名，而且有的还对应多个别名，所以需要一个多对一的映射表，将对话中的部分信息项转化为标准项。消岐处理中的映射表结构与自然语言理解模块中规范化映射表的结构一致，主要针对查询信息中的关键参数，如地点别名、场馆别名、赛事别名等，进行处理。

(2) 同名歧义。同名歧义包括两种情况，一种是完全同名，例如“王涛”是个比较常见的名字，我国有踢足球的“王涛”也有打乒乓球的“王涛”。另一种是别名同名，例如北京的“图书大厦”是一个习惯性的别名，它可能是“海淀图书大厦”也可能是“西单图书大厦”。对这类歧义信息，系统需要通过提示信息进行确认。

(3) 包含歧义。在自然语言中，一些词汇本身就包含有多项含义，比如查询赛事信息时，“短跑”这个信息就包含了 100 米、200 米、400 米，而且每一项还分为男子和女子项目；查询地点时，一些大学包含有分校，一些场馆包含有分馆。这些存在的情况是一对多的映射关系，需要对细节进行确认。是否存在包含歧义是根据具体对话内容分析出来的，比如查询内容是“短跑比赛成绩”，就需要确定比赛的具体类别；查询内容是“什么是短跑”，则是给出概括性的介绍。

(4) 从属歧义。与包含歧义相反，对从属歧义的处理是由细节确定所属类别的过程。比如“我要去看男篮决赛”，这里“男篮”属于“蓝球”，系统在信息查询时，可以通过数据库查询到“奥运篮球比赛在五棵松篮球馆进行”，并给出查询结果。

4.2.5.3 消岐处理机制

消岐处理模块根据信息数据库所包含的消岐库来完成消岐工作，在向数据库写入新的知识信息时，应对可能存在歧义的内容添加相应的消岐信息。消岐库中的表按领域划分，表名与领域名保持一致，并额外包含一张适用于所有领域的通用消岐表。消岐表结构如表 4-14 所示。

对于申请消岐的信息项，消岐处理模块会在消岐表中查询记录信息，处理方式包括以下三种：

(1) 不存在歧义，返回原信息项。

(2) 存在歧义但可通过主题信息与消岐表直接消岐，返回无歧义的信息项。

(3) 存在歧义且无法通过系统直接消岐，消岐处理模块会生成相应的提示信息引导用户完成消岐工作。

表 4-14 消岐表的存储结构

Table 4-14 Storage structure of disambiguation list

字段名	说明
scene	应用场景名，all 表示适用于全部场景
content	输入的非标准信息
para_type	输入信息的参数类型
type	歧义类型
standard	非标准信息对应的标准信息格式
difference	可以区分歧义信息的区别信息

4.2.6 导 航

4.2.6.1 导航的任务

导航模块的主要任务是为用户提供一些便于操作的引导信息，它是以链接形式出现在人机对话平台前端的用户界面中。导航信息是针对用户查询内容给出的额外信息引导，其链接格式有特殊的系统标识，顶层控制模块可以通过区分这些标识来判断接收信息是用户输入的自然语言信息还是点击链接回复的导航信息。顶层控制模块会将导航信息转换为参数信息直接送入对话管理模块，而无需经过自然语言理解模块。

4.2.6.2 导航信息的类型

导航信息的类型包括以下三种：

(1) 列表导航：当信息抽取模块返回的查询结果过多时，将所有结果全部返回给用户，会降低人机对话平台的智能性。因此将返回结果按照一定规则排序，返回固定数量的信息，然后在信息末端给出带有明显导航性质的链接，比如“下一页”、“首页”等，可以提高信息查询效率。

(2) 细节导航：对于一些应用场景的查询结果，导航模块会提供一些细化项来增强人机对话平台的查询性能。比如在查询奥运信息时，信息抽取模块返回了若干场馆名，而这些场馆在信息数据库中有简要介绍，则导航模块会在应答信息中将场馆名以链接格式返回，为用户提供更多的信息查询方法。

(3) 扩展导航：导航模块可以对与用户查询信息间接相关的内容提供导航链接。比如在查询奥运信息时提到了信息项“篮球”，系统会通过信息抽取模块查询近期与篮球相关的热点新闻，以导航链接形式提供给用户。另外，系统还可以根据信息数据库中的统计信息提供一些热点查询问题的导航。

4.2.6.3 导航信息的管理策略

导航信息的管理策略是由导航信息配置文件来设定的。配置文件包括全局参数和主题参数两部分，全局参数是影响整个系统导航策略的配置参数，主题参数是影响某个特定主题导航策略的配置参数。对于前面所述的三种导航信息类型，列表导航是在全局参数中进行配置，细节导航与扩展导航则是在不同的主题中采取不同的配置方法。下面是人机对话平台的导航信息配置文件 NaviConfig.xml 的应用举例：

```
<NaviConfig>                                    //配置文件标志
<Global>                                        //全局配置信息
    <Enable>true</Enable>                       //允许系统使用导航
    <ListNum>5</ListNum>                        //列表导航信息为每页 5 条
    <Snum>3</Snum>                              //本配置文件包含 3 个主题配置项
    <Sname>Weather|Traffic|Olympic</Sname>//各主题配置项的标签名
</Global>
<Weather>
    <Detail>part</Detail>                       //对指定参数类型使用细节导航
    <Dname>City</Dname>                         //细节导航指定参数类型 City
    <Extend> part </Extend>                     //对指定参数类型使用扩展导航
    <Ename>City</Ename>                         //扩展导航指定参数类型 City
</Weather>
< Traffic>
    <Detail>all</Detail>                        //对所有参数类型使用细节导航
    <Extend>none</Extend>                       //不使用扩展导航
</Traffic>
<Olympic>
    <Detail>all</Detail>                        //对所有参数类型使用细节导航
    <Extend>all</Extend>                        //对所有参数类型使用扩展导航
</Olympic>
</ NaviConfig >
```

4.2.7 模块实现

4.2.7.1 类结构图

对话管理类（DM）的类结构图如图 4-15 所示。

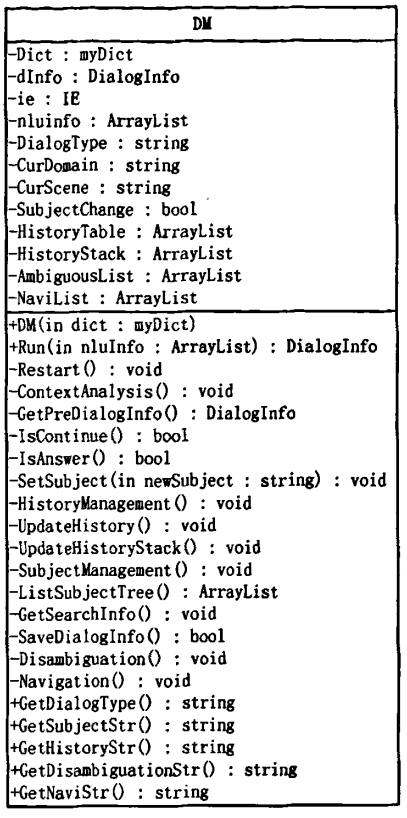

图 4-15 对话管理模块的类结构图

Figure 4-15 Class diagram of dialogue management module

4.2.7.2　主要数据结构

对话管理模块的主要数据结构如表 4-15 和表 4-16 所示。

表 4-15 DialogInfo：应答信息的数据结构

Table 4-15 DialogInfo：data structure of reply information

变量名	变量类型	说明
TempletID	string	应答消息的语言模板号
ReplyType	string	应答类型
ParaList	ArrayList	查询参数信息序列
ResList	ArrayList	查询结果信息序列

表 4-16 ListInfo：信息序列的数据结构

Table 4-16 ListInfo：data structure of information sequence

变量名	变量类型	说明
Word	string	信息序列的词项
Parameter	string	词项所属参数
NaviFlag	bool	是否对该词项导航
Link	string	导航信息内容

4.2.7.3 天气信息查询

天气查询主题通过人机对话平台向用户提供中国各大城市的天气信息查询服务，主要包括三类应用场景：第一类是天气预报查询，系统根据用户输入的城市和日期，为其提供温度、风力等天气状况；第二类是天气预报的细节信息查询，用户可以针对某项细节进行单独查询；第三类是与天气相关的指标信息查询，如穿衣指数、感冒指数等建议性信息。天气信息是通过信息数据库查询的，其来源是 Internet 上基于 Web Service 的天气查询服务。天气查询的主题树结构如图 4-16 所示。

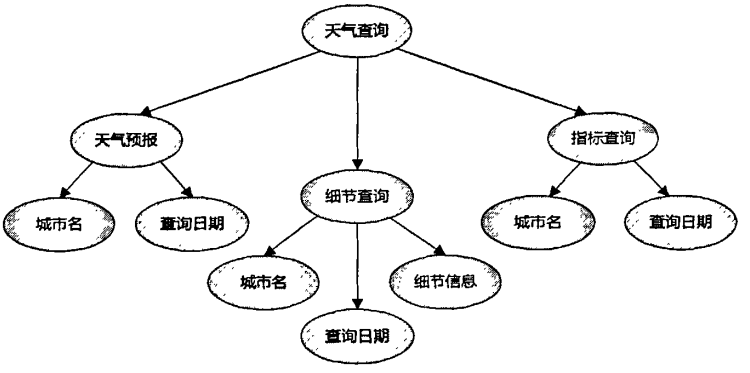

图 4-16 天气查询的主题树结构

Figure 4-16 Subject tree of weather query

城市信息和日期信息是天气查询的通用参数信息，细节信息则是天气预报细节查询独有的参数信息。天气查询可接收的信息项如表 4-17 所示，其所属参数类型如表 4-18 所示。

表 4-17 天气查询的接收信息

Table 4-17 Receive information of weather query

信息项	说明
城市名	接收用户在天气查询时提供的城市信息
查询日期	接收用户在天气查询时提供的日期信息
细节信息	接收用户在细节查询时提供的天气查询细节

表 4-18 天气查询的参数类型

Table 4-18 Parameter type of weather query

参数类型	说明
城市	查询城市的参数类型
日期	查询日期的参数类型，需要分析参数的 DateType 数据结构
天气	细节信息的参数类型

天气查询主题的消岐处理主要是针对少量城市的别名歧义进行的，可以直接通过消岐表来完成。系统在执行天气信息查询之前，需要对查询日期是否合法进行判断，即查询日期范围应在最近几日内（由前天至后天）。

4.2.7.4 交通信息查询

交通查询主题通过人机对话平台向用户提供北京地区的道路交通信息服务，主要包括两类应用场景：一类是交通换乘查询，系统根据用户输入的出发地和目的地，为其提供交通换乘方案；另一类是公交线路查询，系统根据用户输入的公交车次，为其提供该公交线路的详细站点信息。其主题树结构如图 4-17 所示。

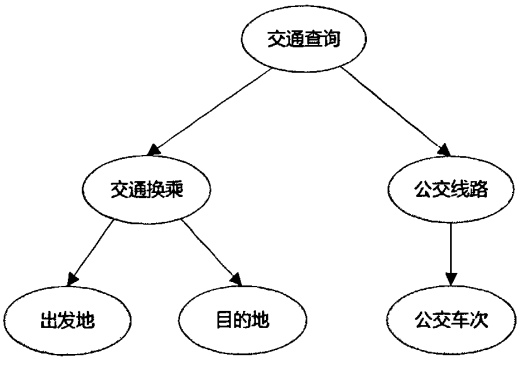

图 4-17 交通查询的主题树结构

Figure 4-17 Subject tree of traffic query

由于人机对话平台主要服务于北京奥运会，因此将北京设置为交通查询主题的默认查询城市，系统不会再对城市信息参数进行额外的判断。交通查询可接收的信息项如表 4-19 所示，其所属参数类型如表 4-20 所示。

表 4-19 交通查询的接收信息

Table 4-19 Receive information of traffic query

信息项	说明
出发地	接收用户在交通换乘查询时提供的出发地点信息
目的地	接收用户在交通换乘查询时提供的目的地点信息
公交车次	接收用户在公交线路查询时提供的车次信息

表 4-20 交通查询的参数类型

Table 4-20 Parameter type of traffic query

参数类型	说明
地点	出发地点和目的地点信息的参数类型
景点	该参数类型一般也可看做地点参数类型来使用
车次	车次信息的参数类型
车次前缀	车次前缀信息的参数类型
车次后缀	车次后缀信息的参数类型

交通查询主题的消岐处理包括以下两种情况：

(1) 交通换乘查询：在查询交通换乘方案之前，需先对用户输入的出发地与目的地信息进行同名歧义与别名歧义的消岐处理。由于同名歧义出现的频率较低，消岐操作可以直接通过消岐表来完成的。但交通查询的地点信息数据量比较大，别名歧义出现的频率较高，其消岐方法则由北京奥运交通信息查询系统封装的 Web Service 来实现。Web Service 会返回与当前输入地点最接近的地点信息，如果存在多个相似信息，则会返回相似度最高的前 10 个信息由用户来确定，消岐方法的结构如表 4-21 所示。通过 Web Service 获得的消岐信息会被保存在消岐表中，以后可以直接根据表中存在的别名歧义信息进行消岐处理。

(2) 公交线路查询：车次信息一般包含三部分，分别是前缀（如运通等）、车次（如 300 等）和后缀（如支线、专线等）。消岐方法会对用户输入的车次进行判断，如果该车次存在则直接进行查询，否则会通过添加或删除车次的前缀或后缀对新生成车次重新进行判断，若仍然不存在，对话管理模块会返回出错信息，否则返回消岐提示信息引导查询。

表 4-21 GetAddr：交通查询消岐方法的参数结构

Table 4-21 GetAddr：Parameter structure of disambiguation method of traffic query

变量名	变量类型	说明
addr	string	用户输入的地点信息
num	int	返回相似信息的数量上限
消岐结果以 ArrayList 格式返回		

4.2.7.5 奥运消息查询

奥运查询主题通过人机对话平台向用户提供奥运期间的场馆信息服务与赛事信息服务，主要包括两类应用场景：一类是场馆信息查询，系统根据用户输入的奥运信息，为其提供奥运场馆信息；另一类是赛事信息查询，系统根据用户输入的奥运信息、查询内容、细节信息、查询日期和查询时间，为其提供相应的奥运赛事信息。其主题树结构如图 4-18 所示。

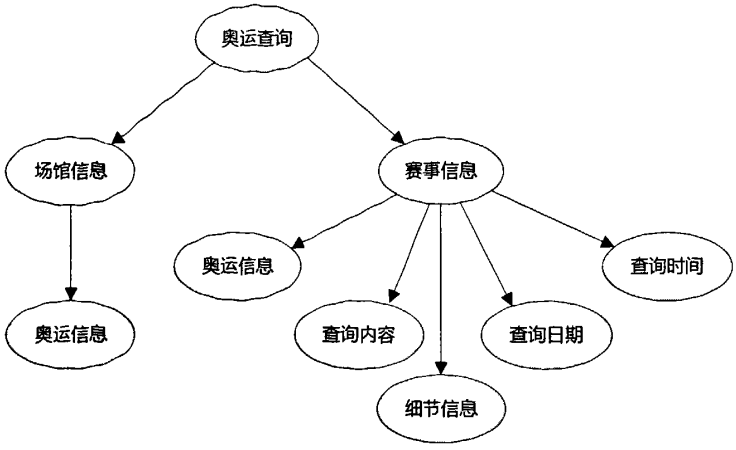

图 4-18 奥运查询的主题树结构

Figure 4-18 Subject tree of Olympic query

奥运信息包括运动员姓名、赛事名和场馆名三种类型，通常采用其中一种类型来确定查询，奥运信息的数据结构如表 4-22 所示。细节信息包括赛事等级（如半决赛）和赛事子项（如男单）。奥运查询可接收的信息项如表 4-23 所示，其所属参数类型如表 4-24 所示。

表 4-22 OlympicInfo：奥运信息的数据结构

Table 4-22 OlympicInfo：data structure of Olympic information

变量名	变量类型	说明
content	string	奥运信息内容
type	OlympicType	奥运信息类型的枚举

表 4-23 奥运查询的接收信息

Table 4-23 Receive information of Olympic query

信息项	说明
奥运信息	接收用户在奥运查询时提供的奥运信息，如人名、赛事名或场馆名
查询内容	接收用户在赛事查询时提供的查询内容，如新闻查询、成绩查询等
细节信息	接收用户在赛事查询时提供的细节信息，如赛事等级、赛事子项等
查询日期	接收用户在赛事查询时提供的日期信息
查询时间	接收用户在赛事查询时提供的时间信息

表 4-24 奥运查询的参数类型

Table 4-24 Parameter type of Olympic query

参数类型	说明
人名	运动员姓名的参数类型
场馆	场馆名称的参数类型
赛事	赛事名称的参数类型
赛事等级	赛事等级的参数类型，如半决赛、决赛等
赛事子项	赛事子项的参数类型，如男单、女单、男 100 米等
赛事查询类型	赛事查询类型，如新闻查询、成绩查询等。
日期	查询日期的参数类型，需要分析参数的 DateType 数据结构
时间	查询时间的参数类型，需要分析参数的 TimeType 数据结构

奥运查询主题的消岐处理主要是针对奥运信息进行的，奥运信息所包含的运动员信息、场馆信息和赛事信息可能存在的歧义类型如表 4-25 所示。其中，由于几个比赛项目可能在同一个场馆中举行，一项比赛可能包含若干赛事子项和运动员，所以在信息数据库中包含“场馆名-赛事名关系表”和“赛事名-赛事子项-运动员关系表”。

表 4-25 奥运信息的歧义类型举例

Table 4-25 Example of ambiguity type of Olympic information

信息项	歧义类型	举例
运动员信息	同名歧义	踢足球的“王涛”和打乒乓球的“王涛”
	别名歧义	“小巨人”和“姚明”
场馆信息	别名歧义	工体可以是工人体育场，也可以是工人体育馆
	包含歧义	北京工业大学体育馆包括羽毛球比赛和艺术体操比赛
赛事信息	包含歧义	“短跑”包含 100 米、200 米、400 米，还区分男女项目
	从属歧义	“男篮”属于“蓝球”

4.3 自然语言生成模块

4.3.1 模块设计

4.3.1.1 主要任务

自然语言生成模块的主要任务是根据语料库提供的信息模板将对话管理模块返回的应答信息组合成自然语言形式的回答，并添加格式标记传送给顶层控制模块。其输入是以信息序列格式传递进来的应答信息，输出是带格式的自然语言文本。

4.3.1.2　模块结构及功能

本模块由句子生成和格式标注两个子模块组成。句子生成模块是根据指定的语料库位置调用语言模板，再将应答信息序列填到模板槽中生成自然语言文本。格式标注模块是根据应答信息的属性和语言模板的类型在自然语言文本生成过程中添加格式标记，使其可以在用户终端上以指定格式显示对话，提高人机交互的友好性。自然语言生成模块的结构模型如图 4-19 所示。

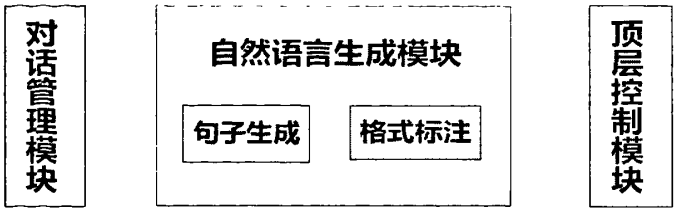

图 4-19 自然语言生成模块的结构模型

Figure 4-19 Structure model of natural language generation module

4.3.1.3　处理流程

自然语言生成模块的处理流程如图 4-20 所示。

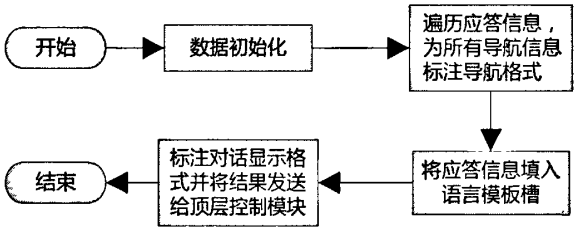

图 4-20 自然语言生成模块的处理流程图

Figure 4-20 Flow chart of natural language generation module

4.3.2 句子生成

句子生成是自然语言生成的核心部分，其性能取决于模板的完善程度以及对模板处理的健壮性。为了实现多样化、个性化的人机对话服务，句子生成模块会从语料库中的指定位置选取语言模板并确定模板内容，将对话管理模块传递来的应答信息序列填入模板槽中生成自然语言文本。

4.3.2.1　句子生成的多样性

在人机对话平台中，语言模板的多样性主要体现在三个方面：

(1) 同一种对话在语料库中可能存在多种语言模板，每次对应答信息进行句子生成时，系统会从中随机选择一种模板作为本次对话的应用模板。比如系统返回询问时间的提示信息可以说“请问您想了解哪天的天气信息？”，也可以说“您想查询哪天的天气情况呢？”。

(2) 一个语言模板中的词汇可能有多种选择，系统可以随机选取词汇，生成对应的句子。比如一个语言模板“请问您想[<refer>][<when>]的天气”，这里[<refer>]可以填入[咨询|查询|查找|询问|想知道|了解]，[<when>]可以填入[什么时间|什么时候|哪一天|哪天|哪一日]。

(3) 语言模板中还包括填入查询结果的模板槽，不同的查询结果可以生成不同的句子。比如通过信息抽取模块的查询结果和对话管理模块的分析得知明天北京地区不会下雨，于是选择一个语言模板“[@City][@Date]没有雨”，并将传递来的参数“City|北京”、“Date|明天”填入模板槽中，组成句子“北京明天没有雨”。

4.3.2.2 句子生成类型

根据对话管理模块能够返回的五种应答信息，句子生成模块可以产生的对话类型主要有以下几种：

(1) 欢迎信息：用户第一次使用系统时，系统向用户发送的欢迎词。

(2) 告别信息：用户结束使用系统时，系统向用户发送的告别词。

(3) 提示信息：系统给予用户提示性的信息，以引导对话。

(4) 查询结果信息：系统发送给用户的查询结果。

(5) 帮助信息：用户请求帮助或系统认为应该给予用户帮助时，发送的帮助性信息，告诉用户如何使用系统以及系统可以实现的功能。

(6) 建议信息：系统给予用户建议性的信息，供用户参考。

(7) 错误信息：系统发生错误时，向用户发送的错误提示信息。

(8) 扩展查询信息：系统在用户获得所需信息后，额外提供的间接相关信息。

4.3.2.3 语料库与语言模板

语料库采用两级表结构，第一级是对应答信息语言模板类别的确定。第二级是根据语言模板的多样性原则随机选取该类别的一个模板作为本次对话的应用模板。语料库结构如表 4-26 和表 4-27 所示。

表 4-26 语料库的第一级表结构

Table 4-26 The first level structure of corpus database

字段名	说明
id	语言模板类别，以“领域-场景-内容”格式标记，内容为 0 表示默认模板
domain	领域名
scene	场景名
content	内容标识
info	内容注释
num	该模板类别包含的同义语言模板数

表 4-27 语料库的第二级表结构

Table 4-27 The second level structure of corpus database

字段名	说明
id	具体语言模板，以“领域-场景-内容-编号”格式标记，系统会随机选择一个编号的语言模板作为当前的语言模板
content	语言模板内容

语言模板由自然语言文本和模板符号组成，模板符号会根据多样性原则在模板槽填充过程中替换为不同的词汇，其符号定义如表 4-28 所示。

表 4-28 语言模板的符号定义

Table 4-28 Symbol definition of language templet

符号	说明
<>	来自语料库中的替换词
@	来自用户提供或查询结果的参数
[]	必须出现一次或一次以上
{}	可以出现一次也可以不出现
|	并列选择符，表示多选一

4.3.3 格式标注

格式标注是对应答信息的自然语言文本做格式上的修饰，主要工作是为导航信息添加链接标签和链接内容，以及一些格式符号的使用。人机对话平台允许在生成语言模板时就使用标签来预先添加模板格式，而对导航信息的格式标注是在该信息填入模板槽之前完成的。格式标注的标签类型如表 4-29 所示。

表 4-29 格式标注的标签类型

Table 4-29 Tag type of format tagging

标签	说明
 	换行符
	空格符
、	加粗符，标签之间的文字加粗
、 、 、 、 、	字号标签，标签之间的文字设置为指定大小
	导航链接起始标签，参数 href 记录了回复给系统的导航信息
	导航链接终止标签，起始标签与终止标签之间的信息将被显示为导航链接

4.3.4 模块实现

4.3.4.1 类结构图

自然语言生成类（NLG）的类结构图如图 4-21 所示。

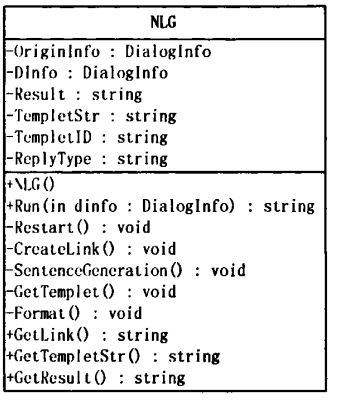

图 4-21 自然语言生成模块的类结构图

Figure 4-21 Class diagram of natural language generation module

4.3.4.2 主要数据结构

自然语言生成模块的主要数据结构如表 4-30 所示。

表 4-30 NaviInfo：存储导航信息的数据结构

Table 4-30 NaviInfo：data structure of navigation information

变量名	变量类型	说明
word	string	导航链接显示内容
link	string	导航链接回复内容

4.4 信息抽取模块

4.4.1 模块设计

4.4.1.1 主要任务

信息抽取模块的主要任务是为对话管理模块提供用户提问信息的查询方法，通过多种信息查询方式获取答案。

4.4.1.2 模块结构及功能

信息抽取模块提供了 3 种类型的查询接口，它们分别是数据库查询、Web

Service 查询和 Internet 信息检索。其中数据库查询接口是从人机对话平台的信息数据库中查询信息，另外两种查询接口是通过网络远程查询信息。系统在调用信息抽取模块的查询方法时，会按照数据库、Web Service、Internet 信息检索的顺序进行。信息抽取模块的结构模型如图 4-22 所示。

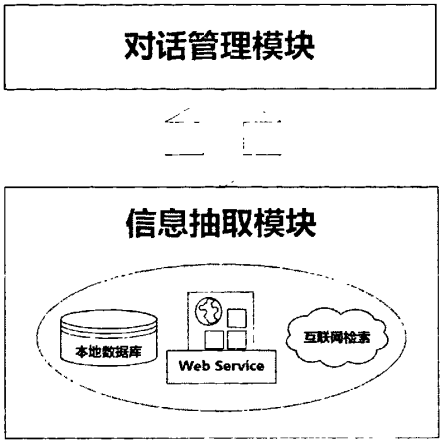

图 4-22 信息抽取模块的结构模型

Figure 4-22 Structure model of information extraction module

4.4.1.3 处理流程

信息抽取模块的处理流程如图 4-23 所示。

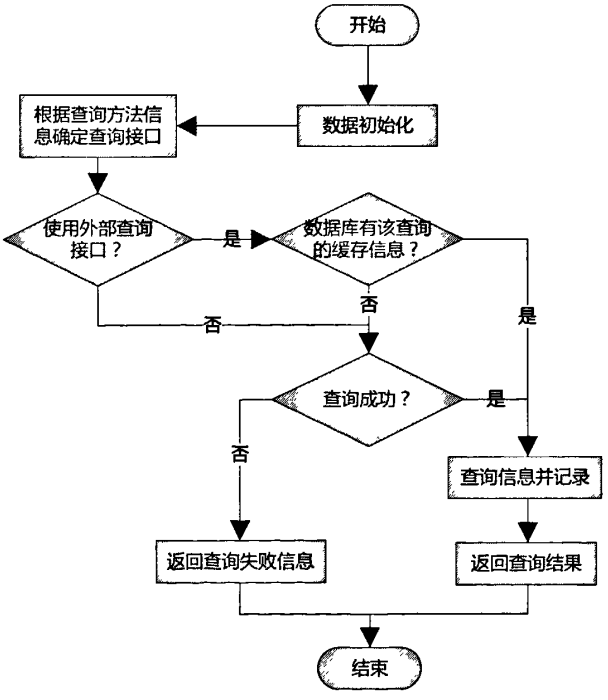

图 4-23 信息抽取模块的处理流程图

Figure 4-23 Flow chart of information extraction module

4.4.2 信息查询接口

人机对话平台的信息查询接口包括以下三种：

(1) 数据库查询。数据库查询是人机对话平台的主要查询方式，在调用信息抽取模块的查询方法时，会优先通过信息数据库来查询结果。查询失败后，才使用其他接口进行查询。

(2) Web Service 查询。包括两类查询方法，分别是首信公司奥运综合信息服务系统提供的 Web Service 和 Internet 上的共享 Web Service。

(3) Internet 信息检索。人机对话平台预留了 Internet 信息检索接口。

4.4.3 模块实现

4.4.3.1 类结构图

信息抽取模块定义了信息查询接口（InfoCollection），所有主题的查询类都必须实现这个接口，这样可以便于扩展主题查询方法。信息抽取类（IE）及信息查询接口的结构图如图 4-24 所示。

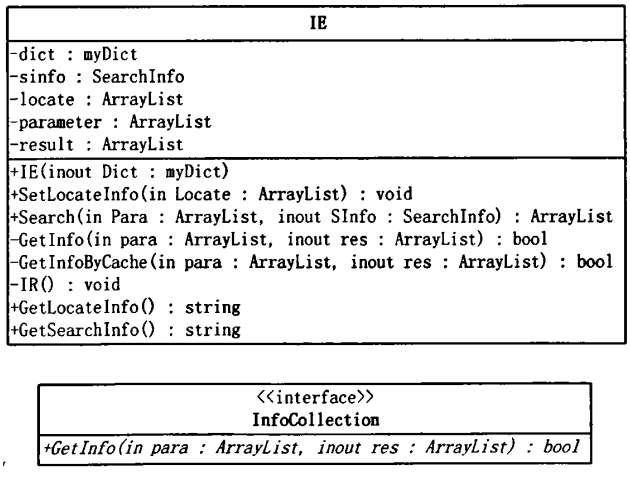

图 4-24 信息抽取模块的类结构图

Figure 4-24 Class diagram of information extraction module

4.4.3.2 主要数据结构

信息抽取模块的主要数据结构如表 4-31 所示。

表 4-31 SearchInfo：信息查询结果的数据结构

Table 4-31 SearchInfo：data structure of navigation information

变量名	变量类型	说明
domain	string	领域信息
scene	string	场景信息
para	string	查询参数信息
result	string	查询结果信息
type	string	查询接口类型

4.4.3.3 天气信息查询

天气信息的主要来源是 Internet 上基于 Web Service 的天气查询服务。人机对话平台通过天气信息抓取工具，可以将 Web Service 提供的天气信息抓取到信息数据库中。天气信息抓取工具是独立于人机对话平台的辅助更新程序，它利用操作系统的任务计划功能，每天定时（Am 1:00）为人机对话平台抓取天气信息。因此信息抽取模块在查询天气信息时，只需对信息数据库进行查询，而不必去调用 Web Service。天气信息的存储结构如表 4-32 所示。

表 4-32 天气信息表的存储结构

Table 4-32 Storage structure of weather information list

字段名	说明
city	城市名
date	日期
weather	天气
low_temp	最低温度
high_temp	最高温度
wind	风力
index1	穿衣指数
index2	感冒指数
index3	晨练指数
index4	交通指数
index5	中暑指数
index6	公园指数
index7	防晒指数
index8	旅行指数
update_time	信息更新时间

天气信息抓取工具的更新原则有以下三点：

(1) 将从 Web Service 上抓取到的最近三日的天气信息同时写入信息数据库，并记录更新时间。

(2) 在更新数据库时若某日的天气信息已存在，则以更新时间为依据，用新的天气信息覆盖旧的天气信息。

(3) 考虑到用户一般只对近期的天气情况比较关心，因此在更新数据库时会删除三天以前的天气信息，以提高检索效率。

4.4.3.4 交通信息查询

交通信息的主要来源是首信公司的北京奥运交通信息查询系统（采用 MapBar 地图引擎），通过该系统封装的 Web Service，人机对话平台可以使用交通查询方法获得详细的交通信息。对于每次交互得到的信息查询结果，系统会将其缓存到信息数据库，下次再出现同样的查询问题时，可以不用访问 Web Service 而直接通过数据库查询答案。交通信息查询方法的结构如表 4-33 和表 4-34 所示。

表 4-33 GetWay：交通换乘信息查询方法的结构表

Table 4-33 GetWay：structure of information query method for traffic transfer

变量名	变量类型	说明
start	string	出发地信息
end	string	目的地信息
num	int	返回交通换乘信息的数量上限
换乘方案以 ArrayList 格式返回		

表 4-34 GetBus：公交线路信息查询方法的结构表

Table 4-34 GetBus：structure of information query method for bus line

变量名	变量类型	说明
no	string	公交车次信息
公交线路信息以 string 格式返回		

4.4.3.5 奥运信息查询

奥运信息的主要来源是首信公司的奥运综合信息资源库，通过资源库封装的 Web Service 为人机对话平台提供了奥运信息查询方法。与交通信息类似，奥运信息的查询结果也会被系统缓存到信息数据库中，以降低对远程查询接口的依赖。奥运信息查询方法的结构如表 4-35 和 4-36 所示。

表 4-35 GetVenue：场馆信息查询方法的结构表

Table 4-35 GetVenue：structure of information query method for venue information

变量名	变量类型	说明
info	OlympicInfo	奥运信息
奥运场馆信息以 string 格式返回		

表 4-36 GetGame：赛事信息查询方法的结构表

Table 4-36 GetGame: structure of information query method for game information

变量名	变量类型	说明
info	OlympicInfo	奥运信息
content	string	查询内容
detail	ArrayList	细节信息，格式为“类型|内容”
date	string	查询日期的标准格式
time	string	查询时间的标准格式
num	int	返回赛事信息的数量上限
赛事信息以 ArrayList 格式返回		

4.5 顶层控制模块

4.5.1 模块设计

4.5.1.1 主要任务

作为系统与用户交互的顶层控制模块，主要任务有：将从终端接收的信息传递给自然语言理解模块；对接收信息进行分析，区分是用户输入信息还是导航信息，采取不同的处理策略；对从远程查询接口获得的查询信息进行信息缓存；返回自然语言生成模块提供的应答信息；记录日志信息。

4.5.1.2 模块结构及功能

本模块包括信息分类、信息缓存和日志记录三个子模块，进行信息分类与信息记录工作。顶层控制模块为人机对话平台提供了自然语言形式的输入接口和输出接口，使其提供的信息查询服务可以被封装为 Web Service 供用户终端调用。顶层控制模块的结构模型如图 4-25 所示。

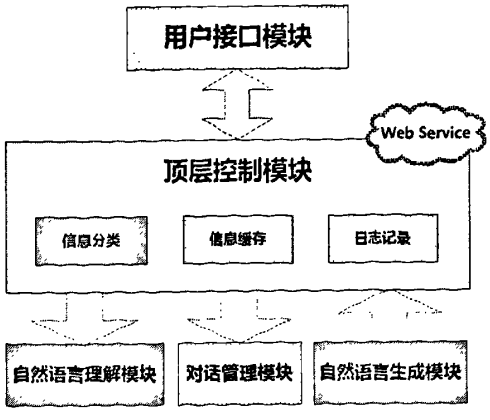

图 4-25 顶层控制模块的结构模型

Figure 4-25 Structure model of top control module

4.5.1.3 处理流程

顶层控制模块的处理流程如图 4-26 所示。

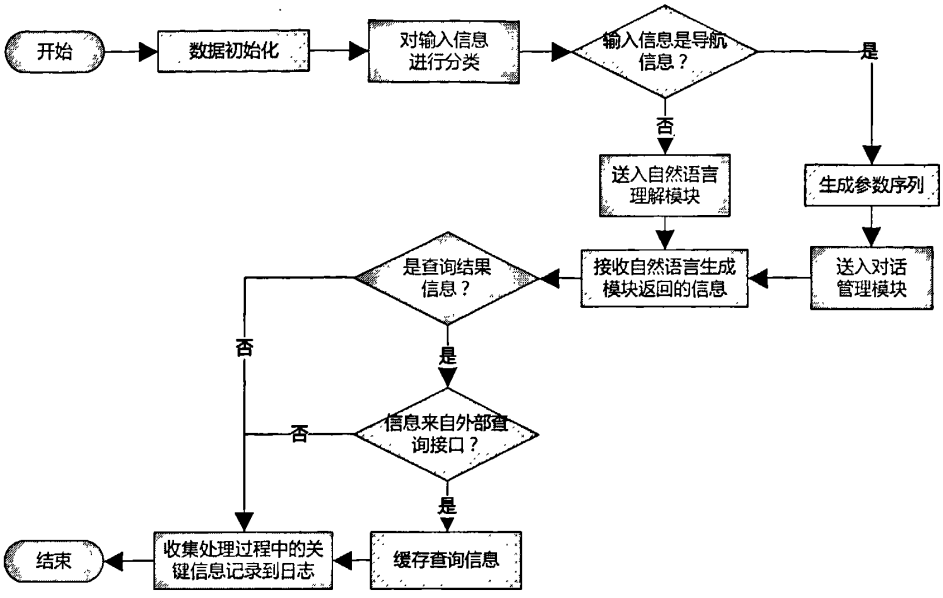

图 4-26 顶层控制模块的处理流程图

Figure 4-26 Flow chart of top control module

4.5.2 信息分类

通过分类标记符，顶层控制模块可以对接收的信息进行分类。用户输入信息在从终端发送时，会将一个“#”符号添加到信息首位置，作为用户输入标记。从终端返回的导航信息，其首位置用“@”符号作为导航标记。用户输入信息在去除分类标记符后，将被送入自然语言理解模块。导航信息则是在去掉分类标记符后，转换成参数序列格式送入对话管理模块。

4.5.3 信息缓存

系统通过信息抽取模块查询信息后会返回一个查询记录，通过该记录提供的信息可以知道本次查询是使用的本地数据库还是远程查询接口。信息缓存模块可以将远程查询接口得到的信息按照规定格式缓存到信息数据库中，使系统再遇到同样提问时，可以从缓存信息中查询到该问题的结果，从而减少人机对话平台对远程信息的依赖，提高查询速度。信息缓存表按领域划分，表名与领域名保持一致，其结构如表 4-37 所示。

表 4-37 信息缓存表的存储结构

Table 4-37 Storage structure of information cache list

字段名	说明
scene	所属场景
parameter	信息查询所需的参数。存在多个参数时，按照系统内部的参数排序顺序以"|"符号将其连接成一个参数
result	从远程查询接口返回的查询结果。存在多个结果时，按照返回顺序用"|"符号将其连接成一个结果
update_time	缓存信息的更新时间
source	缓存信息的来源

4.5.4 日志记录

日志记录模块负责记录人机对话平台各主要模块之间的信息传递，特别是对自然语言理解失败的对话和信息抽取查询失败的对话做详细记录，以便发现系统在规则定义、参数定义或信息抽取方面的问题。根据日志记录，系统维护人员可以为识别性能不好的对话添加新的规则与参数，提高人机对话平台的交互能力。日志信息内容由日志配置文件决定，下面是人机对话平台的日志配置文件 LogConfig.xml 的应用举例：

```
<LogConfig>                          //配置文件标志
    <Enable>true</Enable>            //允许系统记录日志
    <OnlyError>false</OnlyError>     //记录所有对话，不局限于错误对话
    <Turn>search</Turn>              //以一次查询交互过程作为一个记录段
    <Time>true</Time>                //记录对话时间
    <Domain>true</Domain>            //记录领域信息
    <Scene>true</Scene>              //记录场景信息
    <Content>                        //0 表示不记录，1、2、4、8...表示记录对应子项的
信息，可以通过"或"操作记录多个子项的信息
        <Top>0</Top>                 //不记录顶层控制模块的过程信息
        <NLU>5</NLU>                 //记录分词结果和句法分析结果
        <DM>11</DM>                  //记录应答类型、主题信息和消歧结果
        <IE>0</IE>                   //不记录信息抽取模块的过程信息
        <NLG>0</NLG>                 //不记录自然语言生成模块的过程信息
    </Content>
</ LogConfig >
```

4.5.5 模块实现

4.5.5.1 类结构图

顶层控制类（TopControl）的类结构如图 4-27 所示。

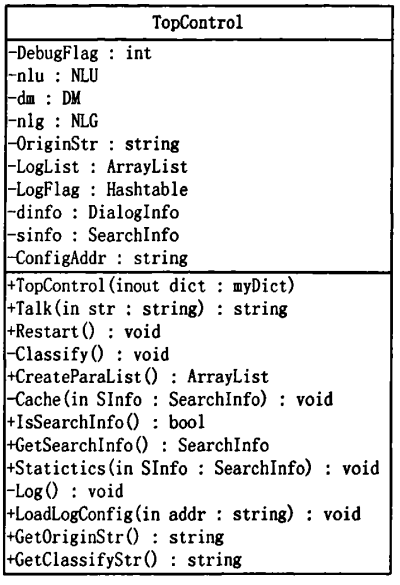

图 4-27 顶层控制模块的类结构图

Figure 4-27 Class diagram of top control module

4.5.5.2 主要数据结构

顶层控制模块的主要数据结构如表 4-38 所示。

表 4-38 LogInfo：日志信息的数据结构

Table 4-38 LogInfo：data structure of log information

变量名	变量类型	说明
content	string	日志信息内容
type	string	日志信息类型

4.5.6 人机对话平台的 Web Service 封装

将人机对话平台封装为 Web Service 可以减少对用户终端环境的要求，便于对系统进行跨平台移植，使其可以应用在不同的操作系统环境中，以及移动终端中。顶层控制模块是系统为用户提供服务的最上层模块， 其输入和输出就是 Web Service 的输入和输出。

4.6　用户接口模块

4.6.1 模块设计

4.6.1.1　设计思想

用户接口即用户终端的服务界面，应操作方便、简洁美观，可以及时与服务器端进行对话交互。人机对话平台选用 B/S 架构，以 ASP.NET 优秀的动态交互能力来实现对用户接口的设计。

4.6.1.2　模块结构及功能

本模块包括服务器端和用户终端两部分，服务器端通过与人机对话平台的 Web Service 进行信息交互，在动态页面的框架（iframe）中生成符合 html 标准的应答信息，并在历史列表中添加对话历史。用户终端通过获取生成好的访问页面，来展开人机对话平台的用户体验。用户接口模块的结构模型如图 4-28 所示。

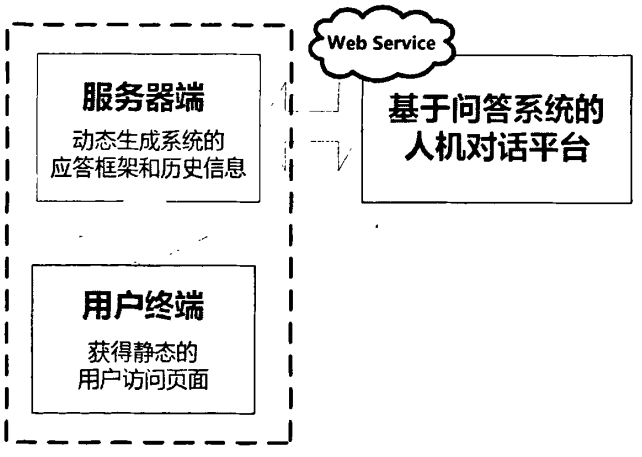

图 4-28 用户接口模块的结构模型

Figure 4-28 Structure model of user interface module

4.6.1.3　处理流程

用户接口模块的处理流程如图 4-29 所示。

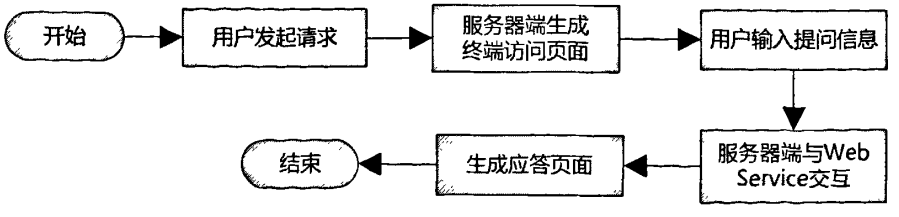

图 4-29 用户接口模块的处理流程图

Figure 4-29 Flow chart of user interface module

4.6.2 界面设计

人机对话平台的用户界面如图 4-30 所示。

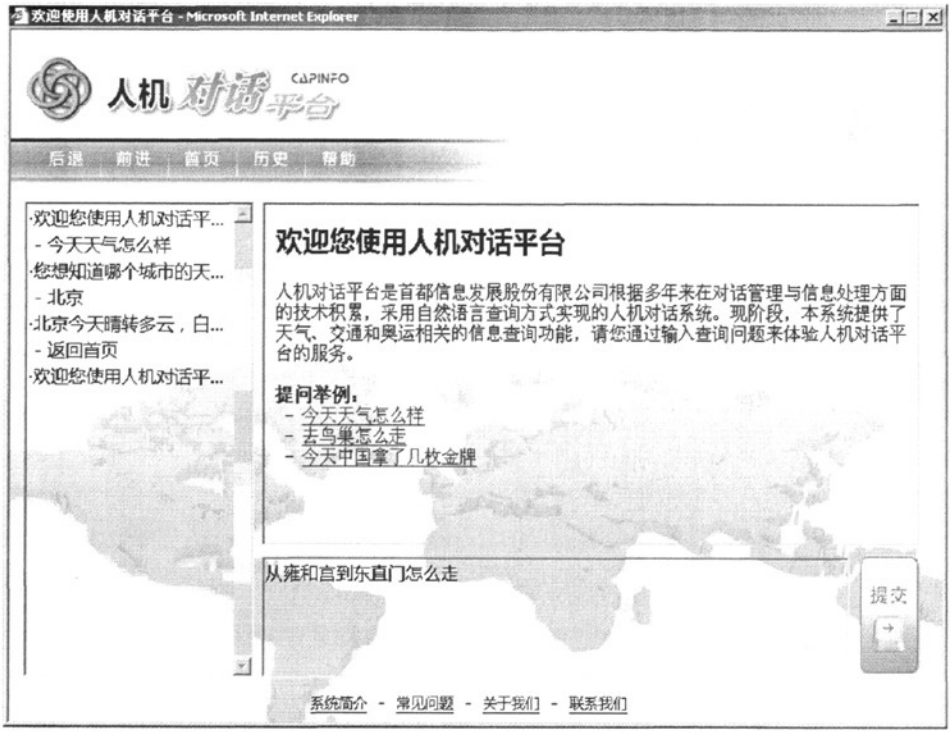

图 4-30 用户终端的服务访问页面

Figure 4-30 User page of terminal

4.7 本章小结

本章详细介绍了基于问答系统的人机对话平台的具体实现，其中包括各功能模块的详细设计及三个对话主题的应用设计，所述内容基本涵盖了人机对话平台开发时需要考虑的问题和实现方法。

第 5 章 人机对话平台的系统测试

5.1 测试目的

人机对话平台系统测试的主要目的包括：确保系统需求得到满足；验证各模块的功能实现，找出系统 Bug 并改正；检验各主题查询服务是否能够正常运行；根据统计信息分析系统性能。

5.2 测试内容

人机对话平台的测试内容包括以下几个方面：

(1) 功能测试：采取与开发同步的测试模式，对系统的每个模块原型和流程能否顺利工作进行检测，也包括对模块输入输出数据的校验和非法数据的处理，对各模块逻辑功能的测试以黑盒测试为主，白盒测试为辅。

(2) 应用测试：通过用例测试系统对规则的处理能力，检验人机对话能否按照设计的方向推进。应用测试采用黑盒测试方法，测试场景包括系统的应答响应，以及天气信息查询、交通信息查询和奥运信息查询三个主题的应用示范。

(3) 性能测试：根据对话过程中产生的日志记录，对系统的识别正确率、查询完成率等信息进行统计分析，并给出性能评价。

5.3 测试环境

人机对话平台的测试环境与实际运行环境保持一致，包括：

(1) 服务器端 PC 一台，Windows XP，Internet Information Server 5.1

(2) 测试终端 PC 六台，Windows XP，Internet Explorer 7

5.4 测试用例

5.4.1 用例来源

人机对话平台测试用例的来源包括：

(1) 根据主题类别和应答响应类别选用的应答测试用例。

(2) 首信公司在对话系统方面所积累的适用于人机对话平台应用示范的部分测试用例。

(3) 根据应用示范中各主题的规则集与交互要点补充添加的测试用例。

5.4.2 用例内容

测试用例包括应答响应用例和应用示范用例，其中每个应用示范用例都对应于一轮完整的交互过程，用例内容如表 5-1 至 5-4 所示。

表 5-1 应答响应的测试用例表

Table 5-1 Test case of reply

测试类别	测试条件	系统状态	预期结果	用例举例
错误信息，测试系统对无法识别信息的反应	未开始交互，焦点主题为空	服务初始化	返回错误信息以及介绍系统全部服务的帮助信息	只要是系统不支持的规则输入即可
	已开始交互，存在焦点主题	上一次对话返回了查询结果	返回错误信息，连续多次出错后返回该主题的帮助信息	
		上一次对话返回了提示信息	返回错误信息和上一次的提示信息，连续多次出错后返回提示信息及该主题的帮助信息	
帮助信息，测试系统的帮助信息是否完备	未开始交互，焦点主题为空	服务初始化	返回介绍系统全部服务的帮助信息	我可以做什么
	已开始交互，存在焦点主题	上一次对话返回了查询结果	返回当前主题服务内容的帮助信息	我可以查询什么信息
		上一次对话返回了提示信息	返回当前主题服务内容的帮助信息，以及上一次的提示信息	请求帮助
其他应答信息相关的测试用例在后表中以应用主题来区别记录				

表 5-2 天气查询主题的测试用例表

Table 5-2 Test case of weather query subject

测试类别	测试条件	预期结果	用例举例
天气预报	日期+城市	返回查询结果信息	北京今天的天气怎么样
			告诉我明天上海的天气
	日期	返回询问城市的提示信息	说说今天天气
	城市	返回近期天气情况	北京天气如何
	无	返回询问城市和日期的提示信息	告诉我天气状况
细节查询	日期+城市+天气情况	返回查询结果信息	明天天津有雨吗
	日期+天气情况	返回询问城市的提示信息	明天温度多高
	城市+天气情况	返回近期天气情况	北京下雪吗
	天气情况	返回询问城市和日期的提示信息	有风吗
指标查询	日期+城市	返回查询结果信息	说说明天上海的穿衣指数
	日期	返回询问城市的提示信息	告诉我最新的穿衣指数
	城市	返回近期指数信息	北京的穿衣指数

表 5-3 交通查询主题的测试用例表

Table 5-3 Test case of traffic query subject

测试类别	测试条件	预期结果	用例举例
交通换乘	出发地+目的地	返回查询结果信息	我要从中国科技馆到天安门怎么走
	出发地	返回询问目的地的提示信息	我从西单出发
	目的地	返回询问出发地的提示信息	我要去复兴门
	无	返回询问出发地和目的地的提示信息	帮我查询下换乘信息
公交线路	公交车次	返回查询结果信息	告诉我运通 104 路的线路信息
	无	返回询问公交车次的提示信息	查查线路信息

表 5-4 奥运查询主题的测试用例表

Table 5-4 Test case of Olympic query subject

测试类别	测试条件	预期结果	用例举例
场馆信息	奥运信息	返回查询结果信息	篮球在哪比赛
	无	返回询问奥运信息的提示信息	我要查询场馆信息
赛事信息	奥运信息+查询内容+细节信息+日期/时间	返回查询结果信息	查查今天上午男篮决赛的结果
	奥运信息+查询内容+细节信息	返回询问日期/时间的提示信息或返回查询结果	我想知道男篮决赛的结果
	奥运信息+查询内容+日期/时间	返回询问细节信息的提示信息或返回查询结果	我要看看短跑比赛成绩
	查询内容+细节信息+日期/时间	返回询问奥运信息的提示信息	明天决赛在哪里进行
	奥运信息+查询内容	返回询问缺省项的提示信息或返回查询结果	查询乒乓球比赛日程
	查询内容+细节信息	返回询问奥运信息和日期/时间的提示信息	查看决赛地点
	查询内容+日期/时间	返回询问奥运信息和细节信息的提示信息或返回查询结果	明天有什么比赛
	奥运信息	返回询问缺省项的提示信息	篮球比赛；刘翔；工人体育馆
	查询内容	返回询问缺省项的提示信息	查看赛事新闻
	无	返回询问缺省项的提示信息	我要查看赛事信息

5.5 测试结果分析

5.5.1 功能测试

功能测试是在Visual Studio 2005开发工具的Debug模式中对人机对话平台各模块的逻辑实现和输入输出内容进行的调试，测试过程与模块开发同步。测试时，首先对已实现的类方法进行白盒测试，确定其逻辑正确；再使用测试用例对模块的输入输出内容进行黑盒测试，确定其功能正确。通过对测试结果的分析，可以发现并改正突出的逻辑错误，以及功能策略上的缺陷。

根据功能测试的需要，在各模块的类设计中添加了便于检测结果的测试方法，如GetSegmentResult、GetSubjectStr等，在正式发布版本中这些方法将被保留，作为日志记录模块的功能方法为系统提供日志信息。通过同步进行的功能测试，人机对话平台达到了总体设计的基本需求。

5.5.2 应答响应测试

应答响应测试是对系统容错能力和帮助能力的检验，考量的是各主题对应的错误信息和帮助信息是否完备，测试重点在于：

(1) 错误信息：根据应答响应测试用例表指出的不同系统状态分别输入错误对话，检测语料库中的相关语言模板是否齐全，以确保系统对用户的非法输入能够做出正确的处理。

(2) 帮助信息：人机对话平台提供了7条询问服务内容的规则，可扩展为多条请求帮助的用户提问。系统在频繁接收到错误信息时也会自动给出帮助提示和能够直接选定目标领域或应用场景的导航信息，使用户可以快速定位查询内容。

根据测试用例表的测试条件与系统状态，应答响应测试检验了11种状态下（包括未选定主题1种、选定主题3种和选定对话场景7种）的错误信息反馈，并选用23条用例完成了对帮助信息的测试，其结果均达到了预期的效果。经此测试表明，人机对话平台可以满足各主题所需的错误处理和服务信息提示功能，为第一次使用本系统的用户提供最基本的对话引导，使其能够尽快熟悉各主题的信息查询功能。

5.5.3 应用示范测试

应用示范测试是针对规则库和具体交互内容的测试，通过由规则生成的测试

用例来确定人机对话平台是否可以正确完成用户的查询需求。测试重点有以下几个方面：

(1) 完整性：一个测试用例代表了一次用户提问，但有些用例并不能使系统直接确定查询内容并返回结果，这就需要人机对话平台引导用户提供额外的查询信息。由此，一轮查询过程会在切换场景或查询完毕时结束。这里对于每一条用例的测试，均要求完成完整的查询过程，即包括系统引导补全查询信息的过程。

(2) 全面性：由于规则是系统正确识别用户输入信息的最关键要素，因此在进行应用示范测试时对规则库中的每条规则都生成了至少一条测试用例。这样可以较大程度的发现系统在语言理解和对话逻辑方面的错误并予以修正，同时也可以找出一些设计不合理的对话流程。

(3) 消岐信息：对不同领域不同应用场景存在的全部歧义类型，选取至少一条用例进行测试，检验消岐信息是否正确。

(4) 场景切换：在对话过程中输入其他领域的查询信息，检验不同领域不同应用场景间的对话切换状况。

(5) 查询失败：对于一些测试用例，虽然提问格式没有问题，但其查询结果是不存在的，比如查询北京奥运会在 2008 年 6 月 1 日的赛事结果。对于这些在查询范围之外的提问用例，需要检验系统信息的反馈状况。

应用示范测试根据前面用例测试表中给出的测试条件对人机对话平台进行分主题测试，检验其反馈信息是否与表中预期结果一致，在完成查询时是否可以正确获得查询结果。本次测试共使用用例 232 条，其中天气查询用例 68 条，交通查询用例 57 条，奥运查询用例 107 条。根据测试结果，可以正确完成用户查询的用例有 222 条，正确率约为 95.69%。其主要原因有以下两点：

(1) 测试用例的来源是规则库中存在的规则，这些用例可以被自然语言理解模块正确识别为语义信息，避免了系统识别能力对应用测试的影响。(对识别能力的评估将由后续性能测试完成)

(2) 大多数场景的交互流程已经在各模块的开发环节中进行了局部调试，尤其是对话管理模块的对话策略与对话流程的调试，更是涵盖了系统的全部应用场景。

(3) 测试时出现的错误主要集中在规则定义、历史管理和消岐判断上，这些问题已在测试结束时予以改正。

上述测试结果表明，人机对话平台已经具备了较好的交互能力和信息查询能力。通过应用示范测试，进一步明确了系统存在的程序设计错误和对话逻辑错误。对这些问题进行改进后，系统已基本满足了信息查询应用的功能要求，为发布后续版本展开性能测试打下了良好的基础。

5.5.4 性能测试

在对话系统的研究中，由于系统功能、应用领域、语料库等设计的不同，其评测形式与内容也会有所不同。鉴于对话系统在性能评测上的复杂性，这里采用试用系统的形式获取用户在对话过程中的日志记录，并根据日志信息的统计分析给出人机对话平台的性能评价。

试用过程中，我们要求测试者完成全部三个主题的人机交互，各主题需要约10 组对话，每组对话以查询完毕、提问出错或场景切换为终止标记。通过对日志信息的统计，共收集到 18 名测试者的对话记录，包括：进行了 507 组对话，其中天气查询对话 157 组，交通查询对话 141 组，奥运查询对话 209 组；完成对话输入（包括回复的确认信息和导航信息）共 1478 句，其中天气查询有 301 句，交通查询有 389 句，奥运查询有 711 句，与主题无关的有 77 句，平均用户输入信息约 82 句。统计结果如表 5-5 和表 5-6 所示。

表 5-5 日志记录的句子统计信息

Table 5-5 Sentence statistics of log

类别	句子总数	正确句数	正确率
天气查询	301	286	95.02%
交通查询	389	363	93.32%
奥运查询	711	651	91.56%
正确输入	1401	1300	92.79%

表 5-6 日志记录的对话统计信息

Table 5-6 Dialogue statistics of log

类别	对话总数	完成查询数	查询完成率
天气查询	157	144	91.72%
交通查询	141	126	89.36%
奥运查询	209	181	86.60%
全部交互	507	451	88.95%

在对话系统的性能测试中，用户输入能否被系统正确识别是评价其系统性能的重要指标。未被正确识别的句子一般包括两类：未被识别的，即属于系统应用领域但未被识别的用户输入；错误识别，即不属于系统应用领域却被系统错误识别的句子。表 5-5 是关于人机对话平台句子识别能力的统计分析，其中正确句数就是指排除了未被识别和错误识别句子后的成功识别句数。通过信息统计发现，人机对话平台对于用户输入的大部分内容都可以正确识别并响应，而关于统计结果有以下几点分析：

(1) 由于系统采用的是基于规则匹配和参数提取的自然语言处理方法，交互失败的主要原因是用户输入未被识别，即规则库中缺少相应的对话规则。

(2) 天气信息查询的识别结果较好，这是因为属于该主题的大部分用户提问都可以在规则库中找到对应的规则。

(3) 交通信息查询的识别问题主要是地点名称未被正确提取为参数所导致的规则匹配错误，该问题可以通过不断积累地点信息参数来改善。

(4) 奥运信息查询会涉及到大量具有包含歧义和从属歧义的用户提问，这是该主题识别正确率低于其他两种主题的原因。通过进一步的积累和完善消岐库，可以较好的改进此类提问的识别能力。

表 5-6 是关于人机对话平台对话完成度的统计分析。一般来说，导致查询服务无法完成的主要原因是用户的提问重点无法被系统正确识别，即系统不能确定用户要选择哪项查询服务，因此系统的对话完成效率受其识别能力影响较大。在日志记录统计中，因识别失败而未完成查询的对话有 52 组，即查询完成率应小于（507-52）/507=89.74%。根据实际查询完成率 88.95%可见，对话管理模块的响应准确率已达到较满意的效果。而对于试用过程中识别失败的用户提问，将通过扩充规则库、语料库和消岐库信息来改进。

在性能测试中，人机对话平台的良好表现反映了其所提供的信息查询服务在一定领域、一定程度上达到了实用水平。对于测试者所反映的问题及给出的建议，将会在下一个版本中重点修改和改进。

5.6 本章小结

本章详细介绍了基于问答系统的人机对话平台的系统测试过程，根据功能测试、应用测试和性能测试的结果做出了系统分析和综合评估，并对测试中出现的问题给出了解决办法。

结 论

随着人机交互技术的不断发展，对话系统在信息服务领域的应用也越来越广泛。本课题以服务 2008 北京奥运会为主要目的，根据科技奥运促进人文奥运的重要理念，依靠首信公司在对话管理与信息处理方面的技术积累和资源支持，设计实现了一个提供天气信息、交通信息和奥运信息查询服务的人机对话平台。

本课题将对话系统作为研究对象，针对系统所涉及的关键技术和主要问题进行展开，采用新的对话管理流程，建立了规则库、参数库、语料库和信息数据库，使人机对话平台具有操作方便、引导丰富和易于主题扩展的特点。课题的主要成果包括：

(1) 将形式语言与自动机理论运用到人机对话平台中，实现了基于规则匹配与参数提取的自然语言理解方法。该方法改进了中文分词算法，通过加入参数词典对影响语义理解的关键信息进行参数标注。在句法分析与语义分析模块中使用了可接受文法规则的下推自动机，并定义了基于 XML 标签的规则存储格式，使规则表达式可以携带更多的语义信息，以便于语言识别和对话管理。

(2) 对话管理模块采用基于槽和任务的设计方法，将树结构的主题管理方法与可追溯的历史管理策略结合起来，作为对话管理的核心技术。其中还加入了首信公司在消岐处理与导航信息方面的技术成果，建立了对应的消岐库和导航配置文件，并设计了两级表结构的语料库，使系统的应答信息更加自然。

(3) 根据开发语言的反射机制和工厂模式的设计思想定义了信息查询接口，使系统可以在不改变原有结构的基础上，完成信息查询方法的复用和扩展。同时，人机对话平台还提供了对远程查询结果的本地缓存，以降低对远程信息的依赖。

(4) 以 Web Service 方式向用户提供信息查询服务，使系统能够在不同的平台上实现应用。通过标签来表示应答响应文本的格式，可在用户界面的 Frame 中生成符合 html 规范的应答页面，使其更加美观。

本课题是以首信公司在对话系统方面的技术成果为基础进行的阶段性研究，取得了一定的成果，同时也存在一些有待改进的地方：

(1) 由于时间有限，资源数据库的信息量不够完备。

(2) 人机对话平台的稳定性和健壮性还有待提高。

(3) 信息抽取模块的 Internet 信息检索接口需要完善。

上述内容将作为人机对话平台下一步的研究目标，在后续版本中逐步改进。

参考文献

1 叶正, 林鸿飞, 杨志豪. 基于问句相似度的中文 FAQ 问答系统. 计算机工程与应用. 2007 年 43 卷 9 期:161~163

2 J. Lin, B. Katz. Question Answering Techniques for the World Wide Web. The 11th Conference of the European Chapter of the Association of Computational Linguistics(EACL), Budapest, 2003. EACL Press:24~40

3 杜玮, 邸书灵, 孙树静. 基于互联网技术的问答系统研究. 微计算机信息. 2007 年 36 期: 124~126

4 王树西. 问答系统:核心技术,发展趋势. 计算机工程与应用. 2005 年 41 卷 18 期:1~3

5 J. Prager, E. Brown, A. Coden, D. Radev. Question-answering by Predictive Annotation. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, 2000. ACM Press:103~112

6 B. Katz. From Sentence Processing to Information Access on the World Wide Web. The Proceedings on Natural Language Processing for the World Wide Web, AAAI Spring Symposium, California, 1997. AAAI Press:77~94.

7 刘里, 曾庆田. 自动问答系统研究综述. 山东科技大学学报:自然科学版. 2007 年 26 卷 4 期:73~76

8 Zhiping Zheng. AnswerBus Question Answering System. Proceeding of HLT Human Language Technology Conference (HLT 2002), San Diego, 2002:53~66

9 T. Macek. Speech in New Generation of User Interfaces. Proceedings of the 3rd annual conference on Task models and diagrams, 2004

10 M. Abad, A. A. Sorzabal, M. T. Linaza, F. Humanidades. NOMENCLATOR- Innovative Multilingual Environment for Collaborative Applications for Tourists and Cultural Organizations. Information and Communication Technologies in Tourism, Innsbruck, 2005: 21~45

11 D. Ferrucci, A. Lally. UIMA:an Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. Natural Language Engineering. Issue 10, 2004: 327~348

12 李子臣. 搜索技术的现状及发展前景. 情报科学. 2006 年 24 卷 3 期:468~474

13 D. Roussinov, J. A. Robles. Applying Question Answering Technology to Locating Malevolent Online Content. Decision Support Systems. Elsevier. Volume 43, Issue 4, 2007: 1404~1418

14 张亮, 黄河燕, 胡春玲. 中文问答系统模型研究. 情报学报. 2006 年 25 卷 2 期: 197~201

15 王宇, 战学刚, 蔡建山. 基于网络的中文问答系统的研究. 计算机工程与应用. 2006 年 42 卷 7 期:162~165

16 余正涛, 邓锦辉, 韩露, 毛存礼, 郑志蕴, 郭剑毅. 受限域 FAQ 中文问答系统研究. 计算机研究与发展. 2007 年 44 卷 z2 期:33~36

17 曹存根. NKI-21 世纪的科技热点. 计算机世界报, 1998 年 50 期:1~3.

18 曾庆田, 段华, 梁永全. 基于 Web 的数学概念知识问答系统研究. 山东科技大学学报:自然科学版. 2006 年 25 卷 1 期:60~63

19 刘迁, 贾惠波.中文信息处理中自动分词技术的研究与展望. 计算机工程与应用. 2006 年 42 卷 3 期:175~177

20 余希田, 李丹亚, 胡铁军. 汉语自动分词歧义处理研究. 医学信息学杂志. 2007 年 28 卷 6 期:541~544

21 黄昌宁, 赵海. 中文分词十年回顾. 中文信息学报. 2007 年 21 卷 3 期:8~18

22 牛洪波, 丁华福. 基于文本分类技术的信息过滤方法的研究. 信息技术. 2007 年 31 卷 12 期:100~102

23 张钹. 自然语言处理的计算模型. 中文信息学报. 2007 年 21 卷 3 期:3~7

24 邓宏涛. 中文自动分词系统的设计模型. 计算机与数字工程. 2005 年 33 卷 4 期: 138~139

25 蒋宗礼, 姜守旭. 形式语言与自动机理论. 清华大学出版社, 2003:194~279

26 Peter Linz. An Introduction to Formal Languages and Automata. 4th Edition. Jones & Bartlett Publishers, 2006: 125~218

27 R. J. Kate, Y. W. Wong, R. J. Mooney. Learning to Transform Natural to Formal Languages. Proceedings of the Twentieth National Conference on Artificial Intelligence, Pittsburgh, 2005. AAAI Press:1062~1068

28 S. C. Reghizzi. Formal Languages and Compilation. Springer Berlin, 2009: 3~5

29 R. Alur, S. Chaudhuri. Branching Pushdown Tree Automata. Springer Berlin, 2006: 26~27.

30 黄民烈, 朱小燕. 对话管理中基于槽特征有限状态自动机的方法研究. 计算机学报. 2004 年 27 卷 8 期:1092~1101

31 王菁华, 钟义信, 王枞, 刘建毅. 口语对话管理综述. 计算机应用研究. 2005 年 22 卷 10 期:5~8

32 M. A. Walker. An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email. Journal of Artificial Intelligence Research, Issue 12, 2000:387~416

33 N. Milanovic, M. Malek. Current Solutions for Web Service Composition. IEEE Internet Computing. Volume 8, Issue 6, 2004:51~59

34 J. Q. Chen, R. D. Heath. Building Web Applications. Information systems management. Volume 18, Issue 1, 2001:213~221

35 J. D. Bruijn, H. Lausen, A. Polleres, D. Fensel. The Web Service Modeling Language WSML: An Overview. Springer Berlin, 2006:336~337

36 路永刚, 赵伟. 一种改进的 MM 分词方法的研究与实现. 长春工业大学学报:自然科学版. 2006 年 27 卷 4 期:320~323

37 傅立云, 刘新. 基于词典的汉语自动分词算法的改进. 情报杂志. 2006 年 25 卷 1 期: 40~41

38 吴涛, 张毛迪, 陈传波. 一种改进的统计与后串最大匹配的中文分词算法研究. 计算机工程与科学. 2008 年 30 卷 8 期:79~82

39 C. L. Goh, M. Asahara, Y. Matsumoto. Chinese Word Segmentation by Classification of Characters. Computational Linguistics and Chinese Language Processing. Volume 10, Issue 3, 2005:381~396

40 J. Pustejovsky, B. Boguraev. Lexical Knowledge Representation and Natural Language Processing. Taylor & Francis, 2004:71~207

41 杨晓明, 罗振声. 模式匹配在中文问答系统中的应用研究. 科学技术与工程. 2006 年 6 卷 3 期: 319~322

42 张继军, 吴哲辉. 下推自动机的状态转换图与下推自动机的化简. 计算机科学. 2006 年 33 卷 3 期:271~274

43 P. Jackson, I. Moulinier. Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization. John Benjamins Publishing Company, 2007:23~162

44 CH. Wu, JF. Yeh, YS. Lai. Semantic Segment Extraction and Matching for Internet FAQ Retrieval. IEEE Transactions on Knowledge and Data Engineering. Volume 18, Issue 7, 2006: 930~940

45 邬晓钧. 对话管理和可定制对话系统框架的研究. 清华大学, 2003:33~45

46 R. J. Brachman, H. J. Levesque . Knowledge Representation and Reasoning. Morgan Kaufmann, 2004:135~204

攻读硕士学位期间发表的学术论文

1 丁杰. 基于文法规则匹配的自然语言处理系统研究与实现. 电脑知识与技术. 2009 年 5 卷 4 期

2 陈卓, 丁杰, 张良, 陈道新, 付磊, 尚鑫, 史正昕. 《面向公安应用的多语言智能移动终端——警译通软件 V2.0[简称:警译通]》软件著作权. 编号:软著登字第 BJ11649 号. 登记号: 2008SRBJ1343. 2008 年 04 月 08 日

致 谢

在本文即将结束之际，我衷心地感谢所有帮助过我的师长、同事和同学。

感谢我的导师庄梓新教授，为我提供了一个良好的科研环境，使我可以把全部精力都放到研究工作和学业任务中去。我的每一点进步都离不开导师的教导、鼓励和支持，他渊博的知识和严谨的态度使我在短暂的研究生学习期间受益匪浅。

感谢周小兵教授在这三年里给予我的支持和帮助，使我能够顺利的完成课题研究和论文撰写。感谢我的师兄张良，他总是在学习和工作中给予我无微不至的指导。感谢我的同学陈卓、郭艳庆、吴树国和刮俊杰，大家互相学习、互相帮助，在三年的研究生生活中共同进步。

感谢首信公司的领导为我提供了难得的实践机会，使我能够参与公司的正式项目，培养了我的团队合作意识，也锻炼了专业技能。感谢各位同事对我实习工作的帮助和支持。

感谢我的父母，他们的关怀和支持是我前进的动力，我所取得的一切成绩都是与他们的付出分不开的。

衷心感谢在百忙之中评阅论文和参加答辩的各位专家、教授。