面向高考招生咨询的问答系统设计与实现

王陈阳

院 （系）：计算机学院　　专　　业：计算机科学与技术

学　　号：1150310609　　指导教师：赵铁军

2019 年 6 月

哈爾濱工業大學

毕业设计（论文）

题　　目 面向高考招生咨询的问答系统设计与实现

专　　业 计算机科学与技术

学　　号 1150310609

学　　生 王陈阳

指 导 教 师 赵铁军

答 辩 日 期 2019 年 6 月 11 日

摘　要

自动问答是自然语言处理领域中的重要技术，它综合使用了自然语言处理领域的众多技术。中国每年有大量的高考考生面临着选择学校，招生咨询的问题，而通过网络搜索得到的数据无疑是大量的，繁杂的，需要用户进行自我筛选，很难获取到准确满足用户要求的信息，就此，提出了面向高考招生咨询的自动问答系统的设计与实现此题，本文的研究重点就是针对高考招生咨询领域构建自动问答系统软件。

本文主要从数据获取，问题分类，问题处理和答案生成以及软件设计四个方面论述自动问答系统的设计与实现。

数据获取涉及到网络数据抓取技术，数据库存储与操作，本文使用了 python 网络爬虫和 MySQL 数据库；问题分类的任务是识别问题的类型，根据系统能回答的问题类型做出相应的回答，本文使用了三种问题分类策略，基于关键词的、基于问句模板的、基于 fastText 分类模型的，三种策略优先级递减，进入不同的问题处理分支；问题处理和答案生成主要基于问答句模板匹配，问句关键词抽取，之后，使用模板和关键词进行了 SQL 语句的构造，数据库查询结果后，使用答句模板构造答句，同时，本文采用自我标注的招生咨询数据对问题分类和问题处理的模块进行了测试；在软件设计中，本文使用了 PyQt5 进行了系统界面的设计，并使用百度语音接口实现语音问答。

关键词：自动问答系统；招生咨询；问题分类；信息检索；答案生成

Abstract

Automated Q&A is an important technology in the field of natural language processing, which combines many techniques in the field of natural language processing. Every year, a large number of college entrance examination candidates face the problem of choosing schools and admission consultation. The data obtained through web search is undoubtedly a large number of complicated and requires users to conduct self-screening. It is difficult to obtain information that accurately meets user requirements. This paper proposes the design and implementation of the automatic question answering system for college entrance examinations admission consultation. The focus of this paper is to build an automatic question answering system software for the college entrance examination admission consultation field.

This paper discusses the design and implementation of automatic question answering system from four aspects: data acquisition, problem classification, problem processing and answer generation, and software design.

Data acquisition involves network data capture technology, database storage and operation. This article uses python web crawler and MySQL database; the task of problem classification is to identify the type of problem, and respond according to the type of questions that the system can answer. Three problem classification strategies, keyword-based, question-template-based, and based on the fastText classification model, the three strategies are reduced in priority and enter different problem processing branches; problem processing and answer generation are mainly based on question-and-answer sentence template matching, question keyword extraction, after that, using the template and keywords to construct the SQL statement, after the database query results, use the answer template to construct the answer sentence , at the same time, this paper tests the problem classification and problem processing modules with self-labeled enrollment consultation data;in the software design, this article used PyQt5 to design the system interface, and use Baidu voice interface to achieve voice question and answer.

Keywords: Automatic Question Answering System, Enrollment Consultation, Problem Classification, Information Retrieval, Answer Generation

目录

摘　要 I
Abstract II
第 1 章 绪　论 1
1.1 课题背景及研究的目的和意义 1
1.2 国内外在问答系统上的研究现状及分析 1
1.3 本文的主要研究内容 2
1.4 本文的组织结构 3
1.5 本章小结 3
第 2 章 领域数据的获取解析与存储 4
2.1 引言 4
2.2 数据获取 4
2.3 数据解析 5
2.4 数据存储 6
2.5 数据获取结果 8
2.6 本章小结 9
第 3 章 问题分类 10
3.1 引言 10
3.2 基于关键词的方法 10
3.3 基于问句模板的方法 10
3.4 基于分类模型的方法 11
3.3.1 fastText 模型介绍 11
3.3.2 fastText 模型训练与测试 14
3.5 本章小结 14
第 4 章 问题处理和答案生成 15
4.1 引言 15
4.2 问题预处理 15
4.2.1 分词与词性标注 15
4.2.2 自定义词典与实体扩展 15
4.2.3 去除停用词 16
4.2.4 关键词抽取与规范 16

4.3 查询语句生成 17
4.3.1 抽象问句匹配模板句式 17
4.2.2 SQL 语句构造 18
4.4 数据查询和答案生成 19
4.4.1 SQL 语句查询数据库 19
4.4.2 查询结果同模板答句构造答案句 19
4.5 问题处理和答案生成的测试结果 19
4.6 本章小结 20
第 5 章 面向高考招生咨询的自动问答系统设计 21
5.1 系统设计 21
5.2 系统功能 21
5.2.1 问答 21
5.2.2 数据库查询 22
5.2.3 问题模板 23
5.3 实验结果及分析 25
5.3.1 问题分类模块测试结果 25
5.3.2 问答模块测试结果 27
5.4 本章小结 27
结 论 28
参考文献 29
哈尔滨工业大学本科毕业设计（论文）原创性声明 30
致 谢 ..... **– 31 –**

第 1 章 绪　论

1.1 课题背景及研究的目的和意义

中国每年都有大量的高考考生面临高考志愿填报的问题，这是一个信息整合的过程，考生需要获取相关高校在招生方面的信息，从而选择一所适合自己的大学。互联网时代，获取信息最便捷最高效的方式就是依靠网络，但正是因为这样，网络上的信息良莠不齐，很多信息的可信度无法确定，同时，网络获取的信息会很繁杂，需要用户进行自我筛选，有时进行筛选后也很难获取到满足用户相应需求的信息；当然，考生还能通过招生咨询的方式向相关高校的老师进行咨询，但这种采用人工咨询的方式往往很难及时地获得回复，而且对于需要较多信息进行对比的要求是比较难实现的。我们针对这样一个实际问题，面向高考招生咨询领域，设计并实现了该领域的自动问答系统。

本课题的研究旨在通过设计面向高考招生咨询领域的自动问答系统，基于大学高考招生的相关信息数据库，用户通过输入自然语言的问句，系统通过一系列的自然语言处理技术对用户问句进行处理，然后根据已有的数据库给出用户准确的回答，若数据库中有与该问题相关的数据，则进行回答，若没有相关的数据，则根据具体情况（问题不在回答范围内、数据库中没有该条数据）返回相应提示。

本课题所研究的自动问答系统可以为考生和家长带来帮助，在进行相关学校的招生咨询时能够通过自动问答系统获取相关的数据，并进行一定程度的对比，在对学校招生信息有一定的了解后能够做出适合自己的选择。

1.2 国内外在问答系统上的研究现状及分析

自动问答系统（Automatic Question and Answering System），简称问答系统（QA），是指接受用户以自然语言形式描述的提问，并从大量的异构数据中查找出能回答该提问的准确、简洁答案的信息检索系统[1]。信息检索的过程是通过关键词匹配相关的文档，例如基于大量文档集合的搜索引擎系统，而自动问答系统则比信息检索更进一步，它需要给出精确的回答，如：哈工大 2015 年在湖南是否招生？则应回答是或否。问答系统最初设计为针对一些特殊的问题领域，如专家系统，这些问答系统的构建往往需要相关专业的人员的参与，专业性很强。随着网络和信息技术的发展，人们能够通过极为便捷的方式从互联网获取信息，与此同时，互联网上的信

息繁多且复杂，人们很难直接准确地获取到满足相应需求的信息。面对这样的问题，自动问答系统研究和发展得到了各个公司和研究机构的关注。1999 年，文本信息检索会议（Text Retrieval Conference,TREC）第一次把 Automatic Question Answering Track 设为评测专项。

目前，国外相对成熟的问答系统有，麻省理工大学(MIT)的 Start[2],密歇根州立大学的 AnswerBus[3],美国的 AskJeeves 自然语言检索系统，IBM 基于统计的问答系统[4]。Start 是第一个基于 Web 的自动问答系统，能够向用户提供准确的回答信息，Start 基于知识库和信息检索的混合模式，先在知识库中检索，若能检索到则直接输出，若不能检索到，则采用搜索引擎检索处理后输出。AnswerBus 是个多语种的问答系统，它不仅可以回答英语的问题，还可以回答法语、西班牙语、德语、意大利语、和葡萄牙语的问题。

国内也有很多进行自动问答系统研究的机构，哈尔滨工业大学开发了基于常用问题集的中文问答系统，该系统先根据用户提问句建立候选问题集，然后通过句子语义相似度计算，在候选问题集中找到相似的问句，然后将答案返回给用户。除此，还有中科院自动问答系统、百度知道、北京理工大学的银行领域汉语自动问答系统 BAQS[5]。

1.3 本文的主要研究内容

本课题的研究内容主要是针对特定领域，即高考招生咨询领域，进行自动问答系统的设计与实现。

问答系统一般包括三个主要部分：问题分析、信息检索和答案抽取[6]。

对于问答系统来说，接受的是用户自然语言描述的问题，首先要做的就是分析和理解问题，例如“哈工大 2017 年计算机类在河北招多少人？”，问题分析模块通过问题分析可以知道该问题询问招生计划问题。中文的问句分析一般包括的基础工作有分词、词性标注、句法分析、命名实体识别、关键词提取与扩展，并在此基础上完成对问句的分类和语义分析等。[7]

经过问题分析后，我们能够获得问题的关键词，对于基于文档的问答系统，我们需要对文档进行检索，然后按照相关性进行排序；基于数据库的问答系统，我们需要进行数据库查询语句的构造，保证能够查询到相应的字段，返回查询结果。

最后，答案抽取模块将从信息检索模块返回的数据进行处理，对于基于问答的问答系统，我们返回了若干候选文档，在进行词法、句法、语义等分析并根据问题分析模块的问题类型，我们返回一个词、短语、或一句话的答案；基于数据库的问

答系统返回了数据库的查询结果，我们需要对数据库的查询结果进行分析，然后构造相应的答案句进行回答。

针对特定领域的自动问答系统，需要先构建相应的数据集，基于数据进行自动问答系统的设计与实现，对本文来说，我们采用了数据库进行数据的存储，需要的数据有：高校的历年招生计划和录取分数数据（以 C9 高校为例），高考招生常见问题集数据，中国大学专业目录数据。然后使用问题分类的方法对问题进行处理，判断该问题是否是系统能够回答的问题，对问题进行分析和处理，转化为对应的查询语句，获得查询结果后根据相应的规则构造答句，返回给用户，还应设置相应的常见问题（FAQ）库，用于存储用户经常询问的问题，当有新的问题时，先在 FAQ 库中搜索是否有与之匹配的问答对，若有，则输出 FAQ 库中与该问题相对应的答案，若没有的话，再进行之前的三个模块进行回答。FAQ 库可以快速处理用户常问的问题，不经过复杂系统处理且可以保证正确性。

1.4 本文的组织结构

本文的组织结构如下：

第 1 章 绪论，介绍本课题的背景及研究的目的和意义，国内外在自动问答系统上的研究现状及分析，本文的研究内容、组织结构。

第 2 章 数据的获取解析与存储，介绍了本文的数据集构建过程，包括网络爬虫技术、数据解析技术和数据存储技术。

第 3 章 问题分类，本章利用基于关键词，基于问题模板，基于 fastText 分类模型的方法对输入问题进行分类，并介绍了三种方法的具体实现过程和优先级。

第 4 章 问题处理与答案生成，本章主要通过对问题进行处理，相应地生成其对应地 SQL 语句，使用查询结果生成答案。

第 5 章 自动问答系统设计，结合第 2-4 章的设计过程，利用 PyQt5、MySQL、语音接口等技术实现了完整的问答系统客户端软件设计，并对软件功能进行展示。

结论，总结本文的主要工作，并进行展望和设想。

1.5 本章小结

本章讲述了高考招生咨询领域的自动问答系统的背景及研究的目的和意义，本文的研究具有现实意义；其次，简要介绍了国内外在问答系统上的研究现状，最后，说明了本文的组织结构。

第 2 章 领域数据的获取解析与存储

2.1 引言

自动问答系统并不能凭空构造答案，同时一个有效实用的自动问答系统一定针对某一个特别的领域，对领域内的问题进行回答。这样，本文的自动问答系统也是基于一个确定的数据集，从而研究在该数据集上的自动问答系统设计与实现。所以，数据是问答系统的基础，只有基于该领域的真实有效数据，同时分析该领域的数据特点，才能使用最合适的自然语言处理技术，达到预期的效果，并且在系统优化过程中能够做出有效的提升。本章主要论述领域数据的获取、解析和存储。

2.2 数据获取

针对本文研究的高考招生咨询领域，我们首先要确定该领域有什么类型的问题，这样便于我们构建与之相对应的数据集来支撑问答系统的构建工作。

我们的第一步工作是获取数据，在互联网网页上有许多有用的信息，而通过互联网网页获取信息的过程就是网络爬虫，它通过网页的 URL 获取到网页源代码，并从源代码中解析到需要的信息，这也是我们获取数据的主要手段。

Python 语言由于其易用性，能够快速学习和入手成为了网络数据抓取的便捷工具，本文在网络数据抓取过程中，主要使用了 Python 的 requests 库和 selenium 库，通过 requests 库获取静态目标网页源代码，selenium 库获取动态目标网页源代码，然后进行相应的数据解析，获得合适的数据。

本文针对高考招生咨询领域，所以我们通过教育部阳光高考网各院校的咨询论坛，抓取了中国所有 985、211 大学的高考咨询常见问题，共 125 所院校，有 123 所具有咨询信息，经过简单分句后，得到 642415 条句子，对部分数据的问题类别进行人工标注后，我们总结出了 20 类问题，同时我们发现，在高考招生咨询领域，高校的招生计划和录取分数是最多被提问的部分，且这两个部分的数据能够通过各大高校的官方招生网站获取，具有比较高的可信度。

但由于各大高校招生网页不同，不能使用同一个数据抓取程序进行数据获取，而本文的研究重点并不在此，故本文选取了 C9 高校（北京大学、清华大学、复旦大学、浙江大学、上海交通大学、哈尔滨工业大学、南京大学、中国科学技术大学、西安交通大学）的招生计划和录取分数作为问答系统构建的数据集，最终的问题测试也将基于该数据集进行测试，所以最终数据获取主要是 C9 高校的招生计划和录

取分数，数据来源是 C9 高校招生官网。

根据以上分析，需要获取 985、211 院校的高考招生咨询信息，在网络爬取过程中，因为我们的访问会比较频繁，所以需要对访问请求进行一系列的处理，常见的处理方法有：模拟 User-Agen、设置一定的爬取频率、使用多线程的方法进行爬取等。[8]我们使用 python 爬虫脚本通过多线程的方式获取如下数据：985、211 院校的常问问题集数据，C9 高校的招生计划数据，录取分数数据。

2.3 数据解析

数据解析是数据获取过程中的重要步骤，互联网信息的多样性同时也带来了一定程度的问题，由于在数据获取过程中，我们得到的是相应 URL 界面的源码，一般都是 xml 文档的形式，而我们需要的信息却是隐藏在 xml 的标签中，由于每一个页面的结构都有可能不一样，这就导致并不会有一个通用的数据解析方式，需要根据具体的页面结构和想要提取的信息进行相应的数据解析，以便获取合适，纯净的数据。

对于网页中的数据，由于网页大多是以 xml 文件的形式展现，而我们需要解析的数据在 xml 的某一个标签中，同时我们也可以将 xml 的标签看作是纯文本，这样就衍生出了两种网页数据解析方式：解析 xml 文件和解析纯文本。解析 xml 文件主要依靠一些 xml 文件解析库，如比较常用的 lxml 库、BeautifulSoup4 库等，通过将 xml 文档转化为类对象，通过类成员变量访问对应的节点；解析纯文本主要依靠文本匹配查询库，如 Python 自带的 re 库，支持使用正则表达式对目标数据进行匹配。

两种匹配解析方法各有优劣，解析 xml 文件的方法比较适用于 xml 标签完整，能够通过足够的标签信息直接定位到相应的位置获取信息，同时，BeautifulSoup4 库也支持使用正则表达式匹配标签内的文本；解析纯文本的方法比较适用于不能完整转换为 xml 对象的文本，可以通过观察目标信息位置，写出相应的正则表达式对目标信息进行匹配，适用性强，但正则表达式的编写是一个难点。在实际的文档解析过程中，这两种方法并不是非此即彼的，通常使用 BeautifulSoup4 库将目标信息所在的标签进行快速定位，如有需要使用正则表达式匹配其中的文本，湖区信息，发挥 BeautifulSoup4 库快速定位的特点和 re 库快速匹配的特点。

但是，真实网页中的数据并不是都能从页面代码中直接获取，若数据可以直接从网页中获取，直接解析网页源码即可获取规则化的数据，但有时从网页中获取到的是文件链接，主要以 excel 文件、pdf 文件为主，这时，对于网页数据的解析就

转为了对文件数据的解析。

在数据解析的过程中我们还会遇到 excel 和 pdf 类型的数据，Python 提供了解析 excel 文件的库 xlrd 和解析 pdf 中表格的库 pdfplumber，通过调用库中的函数将待解析的文件中的数据转化成已处理的数据格式，再从中获取相应的字段，完成文件的解析。

2.4 数据存储

本文的数据获取依赖于网络爬虫，通过 C9 高校的官方招生网站，我们获取到了 C9 高校的招生咨询相关信息。包括：C9 高校的招生计划数据，录取分数数据（分省和分专业）。同时我们对获取到的信息格式进行了规范，规则有：

（1）、年份统一使用 4 位阿拉伯数字进行表示，21 世纪的年份均为“20xx”；

（2）、地区使用省份、直辖市、自治区的普遍名称，如“贵州”而不是“贵州省”，“重庆”而不是“重庆市”，“广西”而不是“广西壮族自治区”；

（3）、 命名中省份名称可以多个的取最长的进行表示，采用了″黑龙江″，″内蒙古″的标识，而″龙江″，″内蒙″不采用。

由于获取到的 C9 高校数据比较多，需要使用数据库进行存储，这里，我们选用了 MySQL 数据库进行数据的存储，一共有三张数据库表。

admission_plan(招生计划)，表的详细字段如表 2-1，部分数据如图 2-1。

表 2-1 招生计划表的字段说明

字段	类型	字段解释
id	int(11)	招生计划 id
university	varchar(30)	大学名称
district	varchar(10)	地区名称
year	int(11)	年份
major	varchar(100)	专业名称
classy	varchar(10)	文理科
numbers	varchar(10)	招生人数

id	university	district	year	major	classy	numbers
23014	哈尔滨工业大学	重庆	2018	自动化类	理工	2
23015	哈尔滨工业大学	重庆	2018	工科试验班（含工程力学、复合材料与工程）	理工	1
23016	哈尔滨工业大学	重庆	2018	飞行器设计与工程	理工	2
23017	哈尔滨工业大学	重庆	2018	电子信息类	理工	2
23018	哈尔滨工业大学	重庆	2018	电气工程及其自动化	理工	2
23019	哈尔滨工业大学	重庆	2018	工科试验班（信息与通信工程）	理工	2
23020	哈尔滨工业大学	重庆	2018	计算机类	理工	3
23021	哈尔滨工业大学	重庆	2018	机械类	理工	4
23022	哈尔滨工业大学	重庆	2018	材料科学与工程	理工	2
23023	哈尔滨工业大学	重庆	2018	能源动力类	理工	4
23024	哈尔滨工业大学	重庆	2018	工科试验班（仪器工程及智能化）	理工	4
23025	哈尔滨工业大学	重庆	2018	工科试验班（功能新材料与化工）	理工	3
23026	哈尔滨工业大学	重庆	2018	环境科学与工程类	理工	2
23027	哈尔滨工业大学	重庆	2018	给排水科学与工程	理工	1
23028	哈尔滨工业大学	重庆	2018	土木类	理工	3
23029	哈尔滨工业大学	重庆	2018	建筑学	理工	2
23030	哈尔滨工业大学	重庆	2018	城乡规划	理工	2

图 2-1 招生计划数据举例

admission_score_pro(录取计划-分省)表字段如表 2-2，部分数据如图 2-2。

表 2-2 录取分数分省表的字段说明

字段	类型	字段解释
id	int(11)	录取分数-分省 id
university	varchar(30)	大学名称
year	int(11)	年份
district	varchar(10)	地区名称
batch	varchar(30)	批次
classy	varchar(10)	文理科
line	varchar(30)	分数线

id	university	year	district	batch	classy	line
610	北京大学	2017	广西	一批	理工	670
611	北京大学	2017	海南	一批	文史	881
612	北京大学	2017	海南	一批	理工	865
613	北京大学	2017	重庆	一批	文史	646
614	北京大学	2017	重庆	一批	理工	689
615	北京大学	2017	四川	一批	文史	650
616	北京大学	2017	四川	一批	理工	688
617	北京大学	2017	贵州	一批	文史	683
618	北京大学	2017	贵州	一批	理工	669
619	北京大学	2017	云南	一批	文史	676
620	北京大学	2017	云南	一批	理工	690
621	北京大学	2017	陕西	一批	文史	668
622	北京大学	2017	陕西	一批	理工	699
623	北京大学	2017	甘肃	一批	文史	626
624	北京大学	2017	甘肃	一批	理工	660
625	北京大学	2017	青海	一批	文史	608

图 2-2 录取分数（分省）数据举例

admission_score_major(录取计划-分专业)表的详细字段如表 2-3，部分数据展示如图 2-3。

表 2-3 录取分数分专业的字段说明

字段	类型	字段解释
id	int(11)	录取分数-分专业 id
university	varchar(30)	大学名称
district	varchar(10)	地区名称
year	int(11)	年份
major	varchar(100)	专业
classy	varchar(30)	文理科
highest	varchar(10)	最高分
average	varchar(10)	平均分
lowest	varchar(10)	最低分
amount	varchar(10)	录取人数

id	university	district	year	major	classy	highest	average	lowest	amount
16007	哈尔滨工业大学	贵州	2015	土木工程	理工	610	607	604	2
16008	哈尔滨工业大学	贵州	2015	给排水科学与工程	理工	600	599.5	599	2
16009	哈尔滨工业大学	贵州	2015	环境工程	理工	594	592	591	3
16010	哈尔滨工业大学	贵州	2015	建筑环境与能源应用工程	理工	587	585	583	2
16011	哈尔滨工业大学	贵州	2015	建筑学	理工	604	604	604	1
16012	哈尔滨工业大学	贵州	2015	城乡规划	理工	595	593.5	592	2
16013	哈尔滨工业大学	贵州	2015	道路桥梁与渡河工程	理工	610	602	594	2
16014	哈尔滨工业大学	贵州	2015	交通工程	理工	588	587	586	2
16015	哈尔滨工业大学	贵州	2015	计算机科学与技术	理工	601	596	592	3
16016	哈尔滨工业大学	贵州	2015	化学工程与工艺（电化学方向）	理工	586	585	584	2
16017	哈尔滨工业大学	辽宁	2015	数字媒体艺术	文史	599	596.5	594	2
16018	哈尔滨工业大学	辽宁	2015	市场营销	文史	599	599	599	1
16019	哈尔滨工业大学	辽宁	2015	会计学	文史	604	604	604	1
16020	哈尔滨工业大学	辽宁	2015	国际经济与贸易（经济与管理学院）	文史	600	600	600	1
16021	哈尔滨工业大学	辽宁	2015	金融学	文史	602	602	602	1
16022	哈尔滨工业大学	辽宁	2015	国际经济与贸易（人文与社会科学学院）	文史	598	597.5	597	2

图 2-3 录取分数（分专业）数据举例

2.5 数据获取结果

通过数据获取过程，我们从 C9 高校的官方招生网站上获取了招生计划和录取分数的数据，经过统计后，三张表中的数据条数如表 2-4：

表 2-4 C9 高校数据获取情况

高校名称	招生计划	录取分数-分省	录取分数-分专业
北京大学	8171（2008-2016）	717（2008-2018）	—
清华大学	6563（2009-2014）	749（2006-2013）	6920（2006-2013 ）
复旦大学	5378（2006-2015）	—	5737（2004-2013 ）
上海交通大学	—	315（2014-2018）	—
浙江大学	—	78（2017）	—
南京大学	559（2018）	185（2015-2017）	762（2017-2018）
中国科学技术大学	442（2018）	420（2004-2017）	—

表 2-4（续表）

高校名称	招生计划	录取分数-分省	录取分数-分专业
哈尔滨工业大学	2040（2017-2018 ）	—	5695（2014-2018 ）
西安交通大学	1785（2016-2018）	—	2694（2015-2018）
北京大学医学部	440（2016-2017）	138（2014-2017）	889（2014-2017）
上海交通大学医学部	—	155（2014-2018）	—
复旦大学上海医学部	404（2013-2015）	—	296（2013-2017）
总计	25782 条	2757 条	22993 条

2.6 本章小结

本章主要讲述了高考招生领域的数据获取工作及数据来源，数据解析部分分析了不同情况下的数据获取方式，最后使用 MySQL 数据库对抓取的数据进行存储。

第 3 章 问题分类

3.1 引言

对用户输入的问题进行分类实质上是在进行用户意图识别。[9]通过识别用户问题的意图能够有效地提取问句中的关键词信息。本章介绍系统中使用的三种问题分类方式，我们通过对问句的分类，能够确定该问题的答案在哪一张表中，从而在使用关键词信息构造 SQL 语句时能够更加有效地

我们使用的三种问题分类方法中，基于关键词的方法能够有效处理含关键词的问句判断，我们将它的优先级放到最高，这样有助于直接识别比较标准的问题。基于问句模板的方法能够通过问句的抽象模式判断问句类型，它的优先级次之，能够通过人工添加模板的方式对问题类型判断进行一定程度的补充。基于分类模型的方法适用性最好，能够通过标注数据进行模型训练，对问句类型进行判断，可以对句式不那么标准的问句进行识别，优先级最低。

3.2 基于关键词的方法

针对高考招生咨询领域，问题中经常出现该领域的专有名词，且该专有名词能够唯一的对应一个问题类型，这样我们能够使用比较简单的方式确定问题的类型，便于后续的问题处理。

经过统计我们发现以下词汇出现在问句中，能够唯一确定问句的类型：

（1）、“招生计划”对应招生计划的问题类型；

（2）、“录取分数”对应录取分数的问题类型。

3.3 基于问句模板的方法

同时，对常问问题集的研究中我们总结出了一些常见的句式，满足这些常见句式的句子能够直接确定其句子类型，这里使用了模板匹配的方法。

通过人工统计出的句式如下：

（1）、招生计划：(university)(year)(major)(district)(classy)招生人数/招多少人等；

（2）、录取分数（分省）：(university)(year)(district)(batch)(classy)录取分数是多少等；

（3）、录取分数（分专业）：(university)(year)(major)(district)(classy)分数线等。

在进行问句模板匹配的过程中，不能直接将原句同模板进行匹配，因为有可能

两个句子表达的是一个意思，但表述却不同，如：“哈尔滨工业大学 2017 年在河北招收计算机类多少人？”，“北京大学 2015 年在湖南招收机械工程多少人？”，这两句话表达的问题类型是一致的，句子模板也是一样的，但由于问题的具体条件不一样，导致句子大不相同，所以需要对问句进行如下的处理。

首先我们需要对问句进行分词和词性标注，具体使用了开源汉语处理包 hanlp，同时通过自定义了学校词典和专业词典，在分词过程中对关键词进行个性化切分，并进行相应词性的标注，在此过程中我们进行了实体扩展，主要是针对学校的全称和简称，详细的说明见 4.1.1，4.1.2 小节。

其次，我们对得到的词序列和词性标注序列抽取关键词，通过词性判断能够有效地识别其中的学校，年份，专业等字段，但识别出的关键词并不能直接使用，我们需要对其进行规范，对同一类关键词转化成预先规定的形式，详细说明见 4.1.4 小节。

最后，我们将问句中的关键词部分使用特定词汇进行替换，得到了抽象问句，使用该抽象问句同问句模板进行相似度计算，衡量标准是归一化的编辑距离，计算公式见式 4-1，通过计算能够得到与之最相似的问句模板，相应地能够确定问句类型，详细说明见 4.2.1 小节。

3.4 基于分类模型的方法

3.3.1 fastText 模型介绍

fastText 是 FAIR(Facebook AIResearch)在 2016 年推出的一款文本分类和向量化工具[10]，来自官网(fasttext.cc)的描述：fastText 是一个开源的、免费的、轻量级的库，允许用户学习文本向量化表示和文本分类。它适用于标准的通用硬件。模型可以经过压缩后可以运行在移动设备上。

fastText 同深度神经网络模型相比，它在分类精度等指标相同的情况下，能够将模型训练时间和预测时间降低了几个数量级，针对本文的文本分类任务，一次训练的时间在秒级别。

下面介绍 fastText 的预备知识，包括词袋模型、哈夫曼编码树、Word2Vec 的 CBOW 模型，fastText 能够做到效果好、速度快，主要依靠两点：利用词内的 n-gram 信息和层次化 Softmax 回归，我们介绍了

词袋模型（Bag of words），广泛应用于自然语言处理、信息检索和图像分类等方面。词袋模型忽略了文本中的语法、语序、语义等要素，将文本看作是若干个词

汇的集合。同时，词袋模型假设文本中每个词的出现是独立的，这样就能使用一组无序的词频来表示这段文本。如下，有两段已分词的文本：

“元芳/你/怎么/看”，“还能/怎么/看/趴/窗户/上/看”，基于两段文本的分词结果构造词典{“元芳”:1, “你”:2, “怎么”:3, “看”:4, “还能”:5, “趴”:6, “窗户”:7, “上”:8}，这样我们构建了一个有 8 个词的词典，其中每一个词在其中都有唯一的索引值，则现在我们能够使用一个 8 维的向量来表示一个文本，向量中对应位置的元素是对应词典中的词语在文本中出现的频数，如下：

[1,1,1,1,0,0,0,0]表示“元芳/你/怎么/看”，[0,0,1,2,1,1,1,1]表示“还能/怎么/看/趴/窗户/上/看”，词袋模型输出的向量同原文本中词的顺序没有关系，只能表征该词在原文本中的频数特征。

哈夫曼编码树，又被称为最优二叉树。是一类带权路径长度最短的树。假设有 n 个权值{w1,w2,⋯,wn}，如果构造一棵有 n 个叶子节点的二叉树，而这 n 个叶子节点的权值是{w1,w2,⋯,wn}，则所构造出的带权路径长度最小的二叉树就被称为哈夫曼树。带权路径长度：如果在一棵二叉树中共有 n 个叶子节点，用 Wi 表示第 i 个叶子节点的权值，Li 表示第 i 个叶子节点到根节点的路径长度，则该二叉树的带权路径长度：WPL=W1∗L1+W2∗L2+...Wn∗Ln

Word2Vec 的 CBOW 模型，Word2Vec 是 Mikolov 等人 2013 年提出的训练词向量的模型[11]。CBOW 是任务模型之一，如图 3-1，对于每一个词，使用该词周围的词来预测生成当前词的概率，即基于上下文预测当前词的概率。

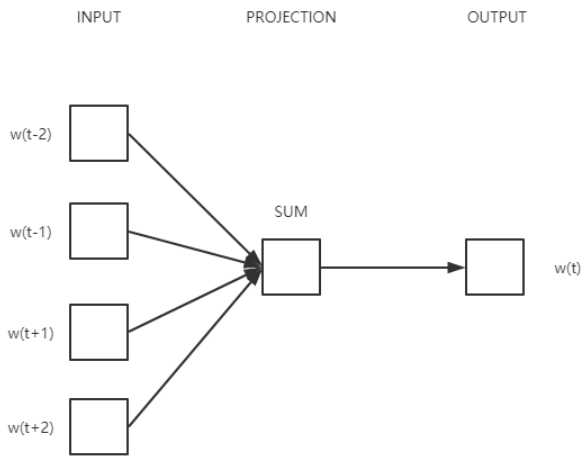

图 3-1 CBOW 模型结构

fastText 的词内 N-gram 特征，以词作为基本单元进行训练学习，然后表示文本，是比较直观的想法，但语料中会出现一些低频词、罕见词或未登录词，这时对于这几类词的训练就会达不到预期的效果。fastText 引入了 subword n-gram 的概念，

它将一个词降到字符级别，使用字符级别的 n-gram 信息来共同表征该词的信息，例如：单词“where”,先设置单词边界为“<where>”，考虑 n=3 的情况，得到集合{“<wh”, “whe”, “her”, “ere”, “re>”}，通常情况下会考虑 n=2/3/4/5 的情况，这样在训练过程中，每一个 n-gram 都会对应一个向量，最后，该单词的向量表征由它所有的 n-gram 的向量求和得到。

fastText 的层次 Softmax，Softmax 是逻辑回归(LogisticRegression)在多分类任务上的推广，是我们训练的神经网络中的最后一层。一般地，Softmax 以隐藏层的输出 h 为输入，经过线性和指数变换后，再进行全局的归一化处理，找到概率最大的输出项。当词汇数量 V 较大时（一般会到几十万量级），Softmax 计算代价很大，是 O(V)量级。层次化的 Softmax 的思想实质上是将一个全局多分类的问题，转化成为了若干个二元分类问题，从而将计算复杂度从 O(V)降到 O(logV)。

fastText 的模型架构，fastText 算法是一种有监督的模型，它的模型架构同 Word2Vec 的 CBOW 架构是一致的，不同的是，CBOW 使用上下文来预测中间词出现的概率，而 fastText 使用上下文来预测标签，fastText 结构如图 3-2。

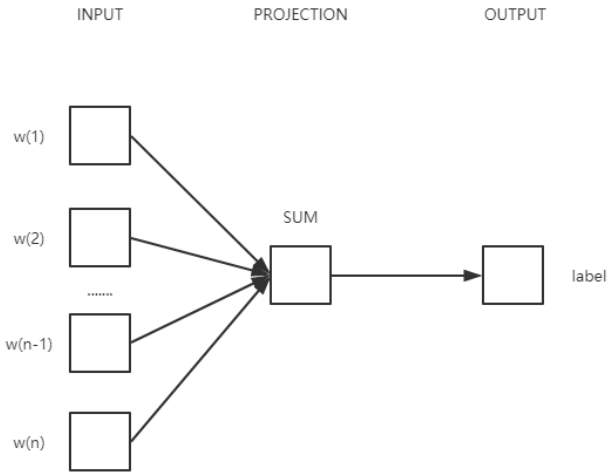

图 3-2 fastText 的结构

fastText 模型的输入是一个词序列，输出的是这个词序列属于不同类别的概率。在序列中的词和词组构成特征向量，特征向量通过线性变换映射到中间层，再由中间层映射到标签。fastText 在预测标签时使用了非线性激活函数，但在中间层不使用非线性激活函数。

输入层：CBOW 的输入是目标单词的上下文并进行 one-hot 编码，fastText 的输入是多个单词 embedding 向量，并将单词的字符级别的 n-gram 向量作为额外的特征；

从输入层到隐藏层，CBOW 会将上下文单词向量叠加起来并经过一次矩阵乘法（线性变化）并应用激活函数，而 fastText 省略了这一过程，直接将 embedding 过的向量特征求和取平均；

输出层，一般的 CBOW 模型会采用 Softmax 作为输出，而 fastText 则采用了 Hierarchical Softmax，大大降低了模型训练时间；

CBOW 的输出是目标词汇，fastText 的输出是文档对应的类标。

3.3.2 fastText 模型训练与测试

经过人工标注，得到了 20 类数据，共 64999 条文本，从中对数据进行了随机划分，每类随机选取 10 个句子得到了 200 个测试集，其余数据作为训练集，进行了模型的训练。

FastText 支持自定义超参数，我们选择了 ngram 和训练轮数 epoch 作为调整的量，根据官方文档的建议使用了 ngram=1，2，epoch=5-50，进行了模型的训练，分别对每一个模型进行了测试，测试结果详见 5.3.1 节，我们发现在构造的测试集上，ngram=1 时 f1 值在 0.9 左右，ngram=2 时，f1 值在 0.95 左右。

对输入问题进行预处理后调用已训练好的模型进行预测，得到相应的问题类型，进入问题的具体处理和答案生成过程。

3.5 本章小结

本章主要介绍了问题分类的三种方法，基于关键词的，使用问句关键词对问句进行问句类型预测；基于问句模板的，对问句进行抽象处理后得到抽象问句，计算抽象问句与模板句的相似度，得到相应的问句类型和问句模板；基于 fastText 分类模型的，使用人工标记的数据建立分类模型，对输入问句预处理后，调用模型进行问题类型预测。三种方法优先级递减，且每种方法进入的程序分支不同，基于关键词的和基于分类模型的输出问句类型后进入问句分析处理，而基于模板的能够在模板匹配过程中获得足够的信息，可以直接进入 SQL 语句构造的过程。

第 4 章 问题处理和答案生成

4.1 引言

本章是问答系统的核心部分，说明了本文对自然语言问句的分析过程，我们通过问题分类和问题预处理过程，得到问题类型和问题关键词信息，从而构造对应的 SQL 查询语句，进行数据查询，最后通过预设的答句模板进行答案的生成。

4.2 问题预处理

本文处理的问题中，用户输入的是自然语言问句，同编程语言和机器语言相比，自然语言是在人类长期的生产活动中慢慢总结演化而来的，具有其独特性，从本质上来说是一种上下文相关语言，而编程语言和机器语言都是人类设计出的机器能无二义理解的语言，所以自然语言的预处理至关重要，能否准确地从自然语言中提取出我们需要的信息对于我们想要解决的问题非常重要。在本文中自然语言的预处理包括以下几个部分。

4.2.1 分词与词性标注

中文自然语言和英文自然语言在预处理上一个最大的不同就是英文以词为句子最基本的单位，且词与词之间有明显的空格间隔，不需进行分词即可开始后面的任务，而中文自然语言在书写上以字为单位，但语义层面上是以词作为基本单位，这就决定了中文分词任务是中文自然语言处理领域最基础的任务。同时在分词后我们还应对分出的词进行词性标注，这有助于我们后续任务的开展和实现。

现在的分词系统很多，比如 Hanlp 分词，jieba 分词，LTP 分词，它们都能根据用户需要选择相应的分词方式进行分词，本文中采用的是 Hanlp 分词中的 NLP 分词器，它能执行词性标注和命名实体识别，由结构化感知机序列标注框架支撑，能够有效应对自然语言处理中的任务。

4.2.2 自定义词典与实体扩展

使用分词模型的好处是能够有效识别大多数常见词，但一般的自然语言处理任务都是针对某一个领域的，而该领域一定有其领域内的专业名词，由于通用的分词模型在训练时不能确定具体的使用领域，使用的训练语料并没有领域性，所以针对

特定的领域问题，我们想要达到较好的分词效果，还需要根据领域的特性，设置领域的专属词汇表，再结合分词模型，能够达到预期的分词效果。

本文的领域是高考招生咨询领域，同时侧重解决的是考生的询问的招生计划和录取分数的问题，所以本文中预定义的词表有：学校词表，专业词表，同时定义词表中词的词性分别为 nuniversity、nmajor。学校词表的获取主要依靠阳光高考网上的学校名单，专业词表通过百度百科“大学专业”词条获取，同时根据获取的数据中的专业字段进行一定的整合组成专业词表，存储在文本文件中，每行一个词，忽略词性，然后在 hanlp 的配置文件中对词典路径进行指定，同时设置整个词典的词性。

自然语言中，人们对同一个事物的表达通常有多种不同的方式，这在进行分词时也应进行相应的考虑，如“哈尔滨工业大学”，自然语言中通常以“哈工大”代称，但表达相同的意义，再如“计算机科学与技术”，也可表达为“计科”，这样的全称、简称，在高考招生领域比较常见，这里，我们采用了人工处理的方式列举学校和专业的常见简称，将之作为词典中词的实体扩展，一并进行分词和词性标注工作。

4.2.3 去除停用词

自然语言的另一个特点就是并不具有很强的规范性，人们在表达时通常会使用助词、虚词等并不包含具体意义的词进行语言的组织，这在我们处理相关领域的问题时回带来一定的干扰，为了后续工作的进行，我们需要去除自然语言中的停用词。

对于中文来说，停用词一般指标点符号、助词、副词、介词、连接词等，同时在特定领域中经常出现的高频词也可以作为停用词，在本文的研究领域中，如“老师”、“请问”、“谢谢”、“贵校”等都应加入到停用词表中，在进行句子分类之前对句子中的停用词进行剔除能够使分类模型更加关注问题的核心部分，分类效果更好。

4.2.4 关键词抽取与规范

当分词和词性标注结束，我们得到了一系列带有词性标签的词，我们需要对其中的关键信息进行抽取和识别，便于进行查询语句的转化工作。

我们需要根据能构造的查询语句进行相关类型关键词的抽取，主要是以下几 zh 种类型：

（1）、学校：通过分词序列中的 nuniversity 属性的词进行识别抽取；

（2）、专业：通过分词序列中的 nmajor 属性的词进行识别抽取；

（3）、地区：NLP 分词器可以自动识别地点词为 ns 属性，进行抽取即可；

（4）、年份：NLP 分词器能够识别地点词为 t 属性，数词为 m 属性，进行抽取后再处理即可；

（5）、批次、类别：分别使用关键词匹配的方式进行识别。

很多情况下我们识别并抽取的关键词并不规范，而我们对应的数据记录中的相应字段只有一个描述，这时，就需要对抽取的关键词进行一定程度的规范，具体是以下几种转换方式：

（1）、学校：全部使用学校的全称，利用之前词典中的实体扩展对学校词进行规范；

（2）、专业：使用专业的全称，规范方式同学校字段；

（3）、地区：统一规范到省一级别，因为我们数据集中地区字段为省份，这里，我们从国家统计局官网收集了中国行政区划信息，对句子中识别出的信息进行规范，如“哈尔滨市”转换为“黑龙江”；

（4）、年份：主要出现的几种类型“2015 年”、“15 年”、“2015”、“二零一五年”、“一五年”、“二零一五”、“今年”、“去年”、“前年”，使用正则转换的方式将它们转换为 4 位数字表示的年份；

（5）、批次、类别：相应转化为“一批”、“提前批”、“文史”、“理工”等。

4.3 查询语句生成

本文针对高考招生咨询领域获取的数据以 MySQL 数据表的形式存储在 MySQL 数据库中，我们对应的查询语句也是 SQL 查询语句，本节主要介绍由已获取的关键词信息构建查询语句的过程。

4.3.1 抽象问句匹配模板句式

当我们使用模板匹配的方式进行问句类型判断时，可以通过替换原句中关键词的方式得到抽象问句，替换如下：

学校关键词→(university)、专业关键词→(major)、地区关键词→(district)、年份关键词→(year)、批次关键词→(batch)、类别关键词→(classy)。

得到抽象问句后，我们使用句子相似度计算方法将该抽象问句同模板句式进行相似度计算，并对相似度进行排序，取出最相似的句子及其对应的模板答句。

计算句子相似的方法很多，本文中采用的是编辑距离的度量方式。编辑距离是

针对两个字符串的差异程度的量度，核心思想是一个字符串至少需要多少次操作才能变成另一个字符串。其中，最常使用的操作定义是莱文斯坦距离，它允许删除、加入和取代字符串中的任意一个字符。例如，将 kitten 变为 sitting，使用以下操作步骤：

①、kitten→sitten（取代：k→s）；②、sitten→sittin（取代：e→i）；③、sittin→sitting（加入：g）。

这样字符串 kitten 与 sitting 之间的编辑距离就是 3，需要注意的是，这是最少的操作步骤，但不一定是唯一的操作路径，同时编辑距离是可逆的，即字符串 a 变为字符串 b 最少的的操作次数是 x，则反过来，字符串 b 变为字符串 a 最少的操作次数也是 x。

但是，绝对的编辑距离只能衡量两个字符串间的相似度，编辑距离为 0，说明是同一个字符串，但这个绝对的编辑距离并不能很好地说明它们的相似程度，因此，我们需要将编辑距离进行归一化操作。我们知道，两个字符串进行编辑距离比较时，不妨设两个字符串中最长的那个的长度为 long_length，最坏情况下，即两个字符串完全不一样，则我们需要对其中一个字符串进行 long_length 次操作才能使它变为另一个字符串，所以我们的编辑距离归一化公式（4-1）设为：

$$edit\_distance_{normalized} = \begin{cases} 1.0 & if\ str1 == str2 \\ 0.0 & if\ len1 == 0\ or\ len2 == 0 \\ edit\_distance(str1, str2)/\max(len1, len2) & else \end{cases} \quad (4\text{-}1)$$

公式中，str1、str2 表示两个字符串，len1、len2 表示字符串长度。

这样，我们计算出每一个模板句与输入问句的相似度，并进行排序，得到了与输入问句最相似的模板问句及其对应的模板答句。

4.2.2 SQL 语句构造

SQL 语句构造的关键在于确定 SQL 语句要查找的表名，以及 where 子句中的各个约束字段值。

在问题分类一章中，我们能够通过问题分类的最终结果确定该问题需要查询哪一张数据表，针对录取分数有两张表的情况（分省和分专业），我们通过问句的关键词进行判断，若专业关键词不为空，则默认查询分专业录取分数表，否则查询分省录取分数表。

我们使用本章中问题预处理的方法获取问句的关键词作为构造 SQL 语句的依据，进行 SQL 语句构造时分为以下两种情况：

（1）针对使用模板匹配的方式进行问句类型判断时，我们通过 4.2.1 节可以得

到问句匹配的模板，而模板中含有一定的关键词槽位，我们借助这些槽位和模板对应的问句类型构造缺省关键词的 SQL 语句，然后使用关键词中对应的词进行填充，得到 SQL 语句。

（2）、针对使用关键词和分类模型判断问题类型的方法，我们使用对应的问题类型和从句子中抽取关键词按照一个统一的模板进行进行聚合，填充相应的槽位，得到对应的 SQL 语句。

4.4 数据查询和答案生成

4.4.1 SQL 语句查询数据库

使用构造的合法 SQL 语句对数据库进行访问，返回查询到的记录，以字段的键值形式进行返回。

4.4.2 查询结果同模板答句构造答案句

得到查询结果后，我们需要使用相应的规则对查询结果进行构造，主要的规则如下：

（1）由问题模板匹配解决的问题，能够在模板匹配过程中得到其模板问句对应的模板答句，使用查询结果中对应的值对模板答句中的相应槽位进行填充就获得了构造的答案句；

（2）由关键词和分类模型解决的问题，得到查询结果后，通过每张数据表对应的记录句式对每一条查询结果进行填充，得到答案句。

4.5 问题处理和答案生成的测试结果

本章是问答的核心部分，涉及到从问句分析、查询语句构造和答句生成的过程，通过分析常用问题集数据中招生计划类和录取类问题，同时根据第二章中获取到的数据，我们进行了问题构造，生成了 100 个问句作为问答系统的测试数据，100 个问句中，我们划分为能回答的且有数据的问题、能回答的但没有相应数据的问题和不能回答的问题类型，它们的比例是 5：3：2，其中能回答的问题中，招生计划问题和录取分数问题之比为 1：1。

具体的测试结果详见 5.3.2 节，通过测试、统计，回答的准确率为 76%，经过分析，出现错误的主要地方是：问句类型判断错误、模板匹配错误。

4.6 本章小结

本章主要介绍了输入问句的预处理过程，涉及到了中文自然语言处理的基本任务：分词与词性标注，并通过自定义词典与实体扩展的方式增强了系统的领域适应性，使用去除停用词的方法排除句中的干扰，方便分类任务的进行，然后使用关键词抽取和规范的方法从句子中抽取了问题的主干信息，接下来使用模板构造的方法生成了 SQL 语句，并进行数据查询，同样使用模板构造的方法生成了问句对应的答案。

第 5 章 面向高考招生咨询的自动问答系统设计

5.1 系统设计

本文设计并实现的问答系统主要有以下模块：问题分析模块、问题查询模块、问题回答模块、模板加载模块、语音模块、系统 UI 模块，系统结构图如图 5-1。

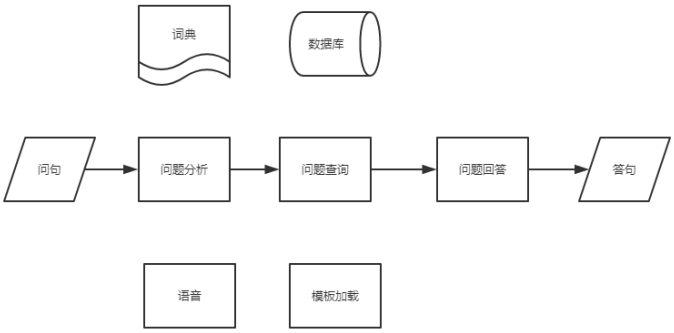

图 5-1 问答系统结构图

5.2 系统功能

PyQt 是一个创建 GUI 应用程序的工具包，能够支持快速便捷的应用程序界面开发。本文中实现的问答系统界面使用了 Python 的 PyQt5 工具包进行设计，主界面包括菜单栏界面和工具栏界面，能够通过点选或快捷键的方式使用这些功能，功能界面包括：问答界面，数据库查询界面，模板查看与创建界面。

5.2.1 问答

问答界面内嵌在主界面中，由当前状态框，问题输入框，结果框和相应的控制按钮组成，控制按钮有开启语音模式、关闭语音模式、清空、回答提问按钮。实现的功能有两个：

一是用户在问题文本框输入问题，按下回车或点击回答提问按钮，系统给出相应的答案，同时在状态框显示当前工作状态，如图 5-2。

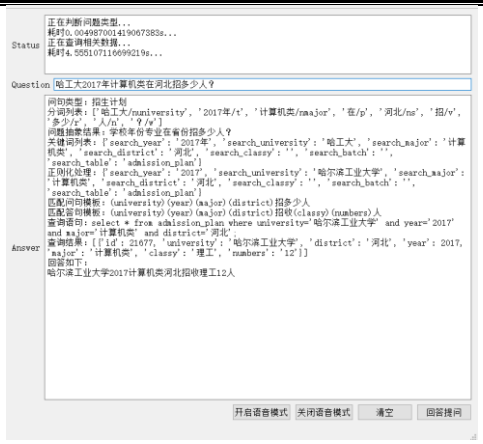

图 5-2 输入文本进行提问的效果图

第二种是用户点击开启语音模式按钮，并开始录音，录音结束后，自动将语音转化为问句显示在问题文本框中，同时开始查询并生成答案，输出到答案文本框中，同时调用语音合成接口，对答案进行语音播报，过程中的每个状态都会在状态文本框中显示当前状态和任务耗时，效果如图 5-3。

图 5-3 语音问答的效果图

5.2.2 数据库查询

数据库查询界面的功能主要是提供直接访问查询数据库的接口，由字段下拉选择框（包括：数据表类型，学校，地区，年份，专业，批次，科别）、MySQL 语句显示框、查询结果显示框和查询按钮组成，能够实现的功能如下。

通过依次选择下拉菜单框，不断确定要查询的数据，同时使用相应的 SQL 语

句构造规则构造 SQL 语句在显示框中，点击查询按钮后，将查询结果显示在结果框内，如图 5-4。

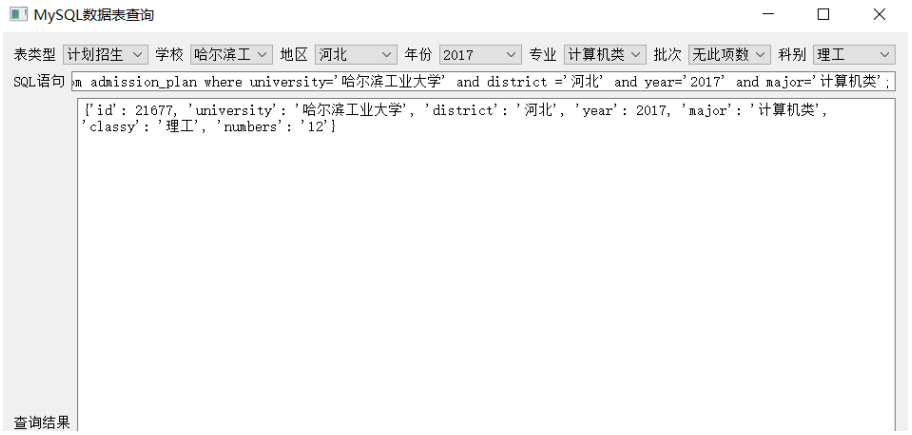

图 5-4 数据库下拉框查询结果效果图

直接在 SQL 语句编辑框中输入合法的 SQL 语句进行查询，如图 5-5。

图 5-5 输入 SQL 语句查询数据库效果图

5.2.3 问题模板

在主页面选择菜单 File 进入 Template 查看界面，或使用快捷键 Ctrl+T 进入，点击查看当前模板可查询到当前的模板文件，通过下拉框选择相应的模板即可看到模板的相关信息（问句条件词、问句目标词、模板问题句、模板答案句），显示在下方的文本框中，如图 5-6。

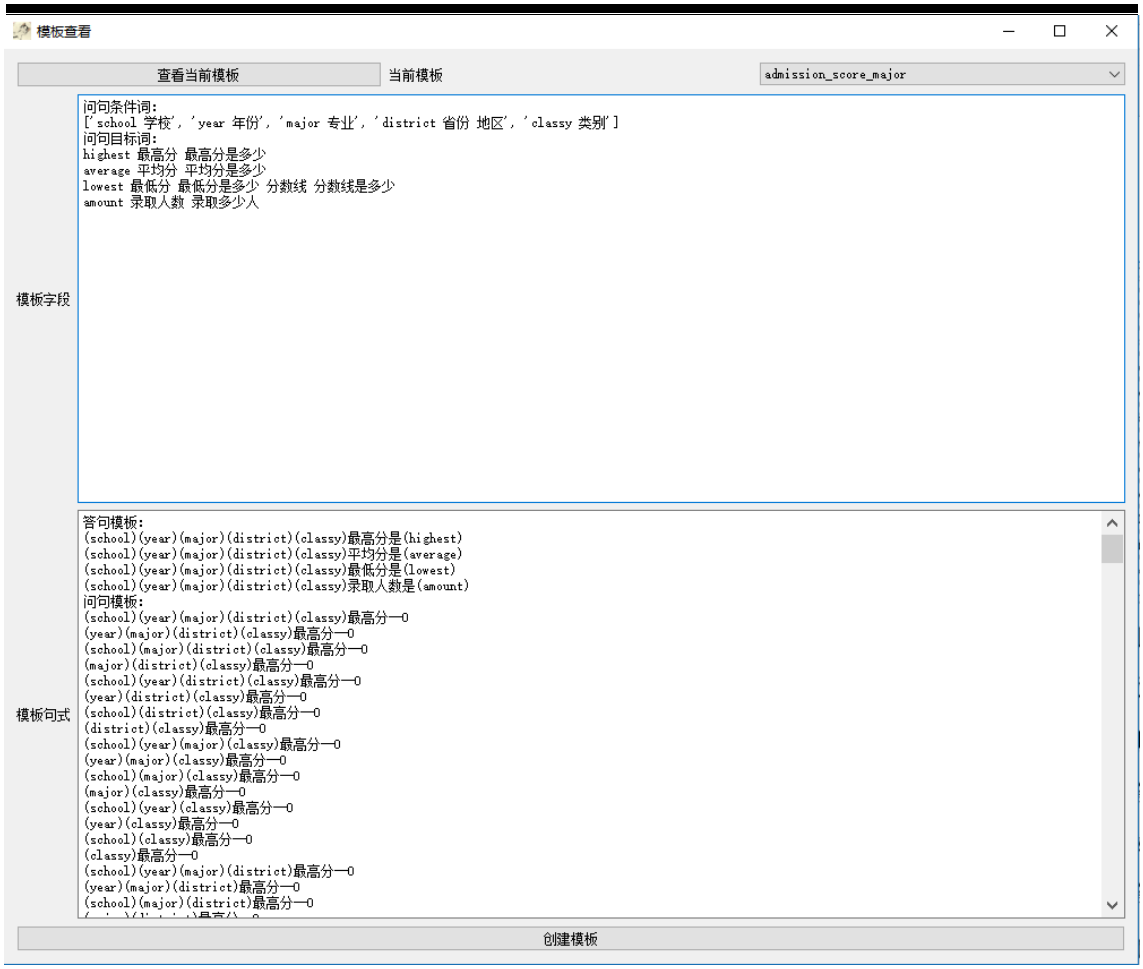

图 5-6 模板查看示例

在模板查看界面点击创建模板按钮进入模板创建界面，按照相应的格式填入相关信息：模板名、问句条件词、问句目标词、模板问答句，点击输入以上信息按钮，返回字段分析结果（是否有为空的字段），确认信息无误后，点击模板构造，可在模板构造框中看到已构造好的模板信息。如图 5-7。

图 5-7 模板创建示例

5.3 实验结果及分析

5.3.1 问题分类模块测试结果

基于关键词和模板的问题分类比较依赖于问句是否含有关键词和相应的模式，所以在问句分类中，我们主要对基于分类模型的问题分类方法进行了测试。最终我们建立了 60000 余条 20 个分类的 fastText 分类模型，分类如表 5-1。

表 5-1 训练数据各类数据量统计

问题类别	问题数量	问题类别	问题数量
专业培养方向	608	文科考生	921
专业学习内容	2085	服从调剂与录取提档	4822
专业学费	525	特长生事宜	1064
专业推荐与选择	5365	理科考生	2113
加分	1912	能否被专业录取	9081

表 5-1（续表）

问题类别	问题数量	问题类别	问题数量
就业情况	932	能够被学校录取	12250
录取分数	4748	自主招生	1475
志愿优先与分数优先	482	转专业事宜	3136
招生联系咨询	3921	重点专业于排名	1390
招生计划	4246	高考成绩与排名	3723

我们在每一类中随机选取了 10 个问题作为测试数据，共 200 个句子进行测试，测试结果如图 5-8，5-9。可以看到取 ngram=1 时 f1 值在 0.9 左右，ngram=2 时 f1 值在 0.95 左右。

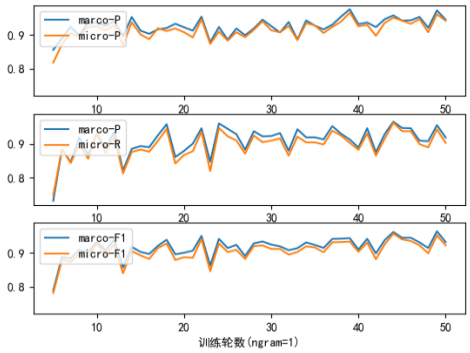

图 5-8 ngram=1 时的测试结果

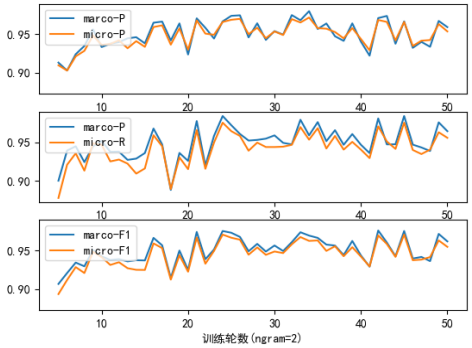

图 5-9 ngram=2 时的测试结果

我们选取 ngram=2，epoch=24 时的模型进行具体的错误分析，这时测试结果如表 5-2。

表 5-2 ngram=2,epoch=24 时测试结果

	精确率	召回率	F_1
宏平均	0.951971	0.955586	0.953775
微平均	0.946099	0.952857	0.949466

检查测试结果中错误的情况即该句子属于 A 类，却预测成了 B 类，如问题“老师您好，请问北京地区自主招生加 5 分的考生有多少？谢谢!”模型预测该问题属于加分类型的问题，但实际该问题是自主招生类型的问题，预测错误的原因是因为在该句中出现了“加 5 分”的字串，导致预测出现了偏差；再比如“贵校管理科学与工程类是学什么的好不好了？”模型预测该问题属于招生联系咨询的问题，但实际该问题是专业学习内容的问题，预测错误的原因是该句中出现了“好不好”，通常咨询招生联系方式的句子中经常出现“告诉我 xxx 好不好”的句式，导致出现了偏差。

5.3.2 问答模块测试结果

通过对常问问题数据中招生计划类和录取分数类问题的分析，结合本系统中有的数据，构造了 100 个问题，其中包括：能回答的问题且有数据、能回答的问题但没有相应的数据、不能回答的问题类型，比例为 5：3：2，能回答的问题中招生计划与录取分数之比为 1：1。测试结果如表 5-3 所示。

表 5-3 问答系统测试结果

类别	回答正确	回答错误	总计
能回答有数据（招生计划）	19	6	25
能回答有数据（录取分数）	15	10	25
能回答无数据（招生计划）	13	2	15
能回答无数据（录取分数）	12	3	15
不能回答	17	3	20
总计	76	24	100

计算得到回答的准确率为 76%，出现错误的位置主要是：问题类型判断，模板匹配部分。如以下两个例子：“请问获得全国奥赛省二等奖，可以报贵校自主招生吗？谢谢”，这是一个自主招生类型的问题，因为使用了关键词匹配的方法判断问题类型，判断为招生计划，但缺少相应的关键词，无法进行查询，导致出现错误；“2009 年复旦在江西招收名额是多少？”，这是一个能回答的招生计划类型的问题，问题类型判断正确了，但在匹配模板时，匹配到了“(year)(major)(district)招生计划是多少”，漏掉了学校字段，导致查询的结果时所有学校的结果。

5.4 本章小结

本章主要介绍了问答系统的界面设计，以及每个界面的控件设计，在软件流程中介绍了本文开发的问答系统的功能的使用流程，包括最主要的问答功能，数据库查询和模板的查询与创建。

结　论

本文针对高考招生咨询领域的问答系统设计与实现，做了以下工作：

（1）、高考招生咨询领域的数据获取，包括 C9 高校的招生计划和录取分数数据，阳光高考网各院校招生咨询论坛常见问题，学校名词典，专业名词典，中国行政区划数据；

（2）、设计并实现了基于关键词和基于模板的问题分类器；

（3）、通过标记数据，建立了基于 fastText 的问题分类系统，能够对问题进行有效分类；

（4）、通过分词及词性标注、去停用词、关键词抽取和规范等工作，建立了较为完整的问句到 SQL 语句的转化系统；

（5）、使用 PyQt5 设计并实现了问答系统软件的完整设计。

本文的研究尚处在探索和尝试阶段，还存在很多不足的地方：

（1）、本文的研究重点在于问答系统的设计与实现，在数据的选择上做了让步，只选取了 C9 高校进行研究，并且其中某些高校的数据无法完整获取；

（2）、受到数据量的限制，本文的研究中只解决了招生计划和录取分数的问题，问题类型不够全面，对于输入较为规范的问句，系统能够给出相对满意的答案，但对于不够规范的问句，会产生错误的回复；

（3）、受到技术能力，时间等的限制，本文的系统实现是在本机完成的，只能够实现本机的访问功能，不能接受多人同时访问。

今后进一步在本研究方向进行研究工作的展望和设想：

（1）、获取更多高校的数据，丰富数据库内容，如高校的专业介绍的丰富，能够通过高校门户网站的数据抓取，获得高校专业介绍的相关信息，同时能够对该校内的重点专业进行总结和存储；

（2）、使用 Web 端框架构建问答系统，使得问答系统能够支持多用户访问；

（3）、对于问题分类部分可以在数据量足够且标注完成的情况下构建一个更加全面完整的分类系统，对问题类型进行预测；

（4）、核心的 SQL 语句构造技术可以进行相应的改进，使得问句的构造更加合理，且能适应多种情况，如多个年份的数据查询等。

参考文献

[1] 刘里,曾庆田.自动问答系统研究综述[J].山东科技大学学报(自然科学版),2007(04):73-76.

[2] The START Natural Language Question Answering System. Boris Katz,Gregory Marton,Gray Borchardt,et al. http://start.csail.mit.edu. 2006

[3] AnswerBus Question Answering System. Zhiping Zheng. http://answerbus.coli.uni-saarland.de/index.shtml. 2006

[4] IBM’s Statistical Queston Answering System. A Ittycheriah, S Roukos. Proceedings of the TREC-11 Confrence. 2002

[5] 樊孝忠,李宏乔,李良富,叶江.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(06):528-532.

[6] 郑实福,刘挺,秦兵,李生.自动问答综述[J].中文信息学报,2002(06):46-52.

[7] 张宁,朱礼军.中文问答系统问句分析研究综述[J].情报工程,2016,2(01):32-42.

[8] 胡俊潇,陈国伟.网络爬虫反爬策略研究[J].科技创新与应用,2019(15):137-138+140.

[9] 张晓娟. 查询意图自动分类与分析[D].武汉大学,2014.

[10] Joulin A, Grave E, Bojanowski P, et al. Bag of Tricks for Efficient Text Classification[J].Computer Science,2016,7:427-431.

[11] Mikolov T,Chen K,Corrado G,et al. Efficient Estimation of Word of Representations in Vector Space[J].Computer Science,2013:1-12.

哈尔滨工业大学本科毕业设计（论文）原创性声明

本人郑重声明：在哈尔滨工业大学攻读学士学位期间，所提交的毕业设计（论文）《面向高考招生咨询的问答系统设计与实现》，是本人在导师指导下独立进行研究工作所取得的成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明，其它未注明部分不包含他人已发表或撰写过的研究成果，不存在购买、由他人代写、剽窃和伪造数据等作假行为。

本人愿为此声明承担法律责任。

作者签名：　　日期：　　年　　月　　日

致　谢

回顾本科毕设期间的工作和经历，我的导师赵铁军教授对我进行了精心的指导，选择本文课题后，导师与我进行了深入的探讨和沟通，确定了开题的方向和研究的方案，在系统设计、数据获取、论文创作的过程中，导师都没有松懈对我的指导，正是有了导师在该课题方向上的指导，使得我的课题研究过程少走了很多弯路，导师对系统设计和软件效果的严格要求加深了我对课题的理解程度，提高了我的编程能力。导师和蔼可亲、平易近人、认真严格，他的言传身教将使我受益终生！

感谢深智科技史桦兴师兄和张春越师兄的帮助，在课题研究中期，两位师兄在我的系统设计和改进上提出了很多建设性的意见，同时为我提供了安静舒适的工作环境，我也从他们身上学到了很多自然语言处理算法研究和软件开发的知识。另外，特别感谢与我同组的黄道龙同学、武德浩同学在数据抓取和系统实现上的帮助，以及蔡妙钰同学在我毕设设计过程中对我的支持和鼓励。