面向大学招生的自动问答系统设计与实现

专业：计算机科学与技术

姓名：王陈阳

指导老师：赵铁军

目录

- 工作说明
- 系统实验设计及结果
- 论文撰写工作

工作说明

- 原工作是面向大学招生的自动问答系统的设计与实现，在系统设计与实现的过程中根据抓取到的数据和相应的问题领域进行了细化，修改为面向高考招生咨询的问答系统的设计与实现，主要是以C9高校数据以及招生计划和录取分数的问题类型进行设计和实现。

系统工作--数据搜集

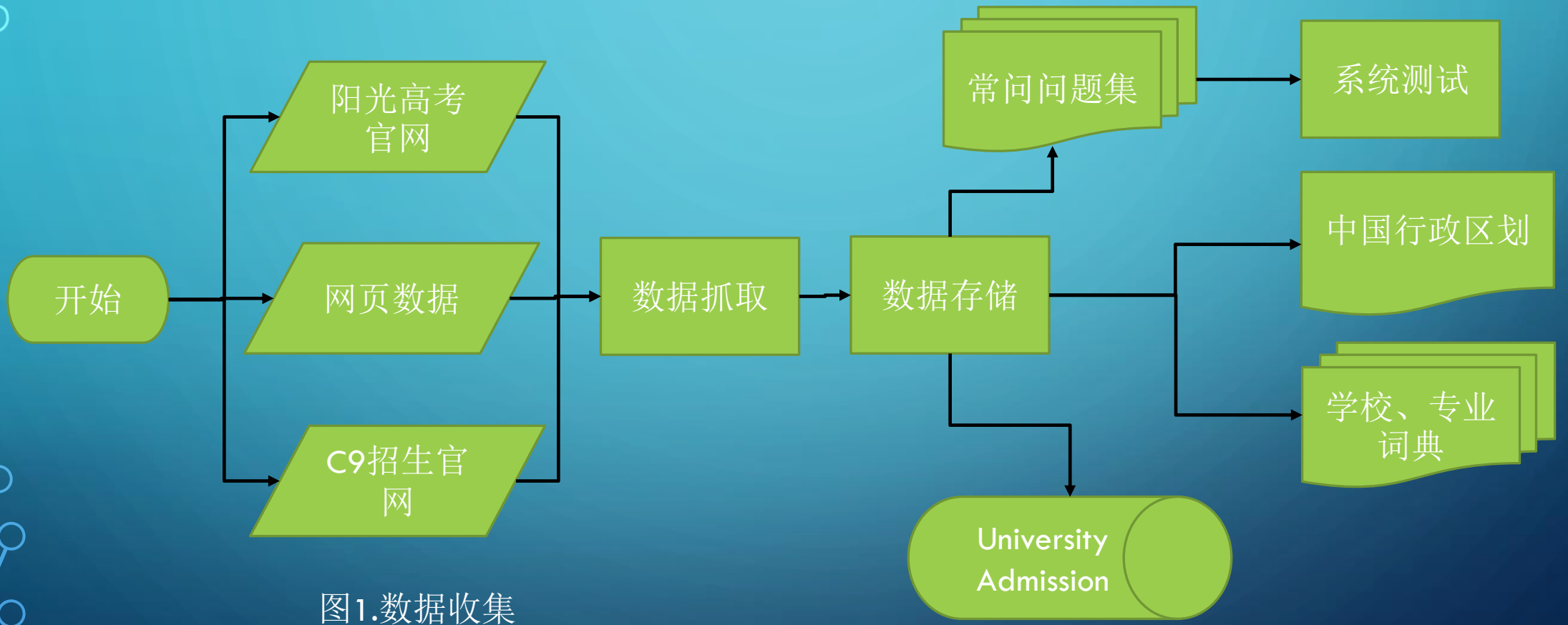


图1.数据收集

数据搜集结果—招生数据

学校	招生计划	录取分数（省份）	录取分数（专业）
北大	2008-2016	2008-2018	无
北大医学部	2016-2017	2014-2017	2014-2017
清华	2009-2014	2006-2013	2006-2013
复旦	2006-2015	无	2004-2017
复旦医学部	2013-2015	无	2013-2017
上交	无	2014-2018	无
上交医学部	无	2014-2018	无
中科大	2018	2004-2017	无
哈工大	2017-2018	无	2014-2018
西交大	2016-2018	无	2015-2018
南京大学	2018	2015-2017	2017-2018
浙大	无	2017	无
总计	25782条	2757条	22993条

表1.招生数据收集情况表

数据来源：C9高校本科生招生官网

数据存储--招生数据

- MySQL数据库存储
- 招生计划表： admission_plan（25782条数据）
  - id，学校，地区，年份，专业，类别，招生人数
- 录取分数（省份）表： admission_score_pro（2757条数据）
  - id，学校，年份，地区，批次，类别，分数线
- 录取分数（专业）表： admission_score_major（22993条数据）
  - id，学校，地区，年份，专业，类别，最高分，平均分，最低分，录取人数

问答系统设计—模块

- 问题分类
- 问题处理
- 查询语句构造与数据查询
- 答句构造
- UI设计

问答系统设计--系统流程图（自动问答）

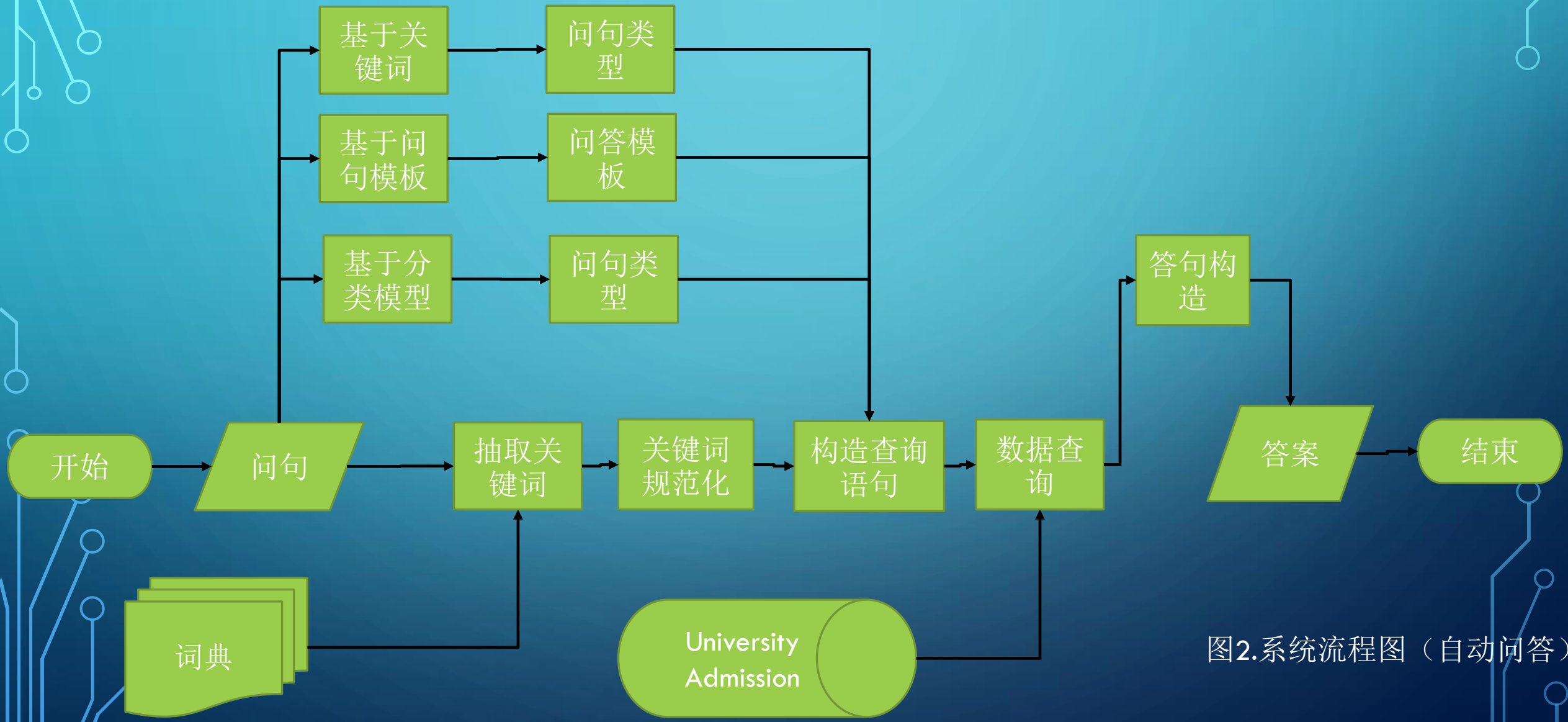


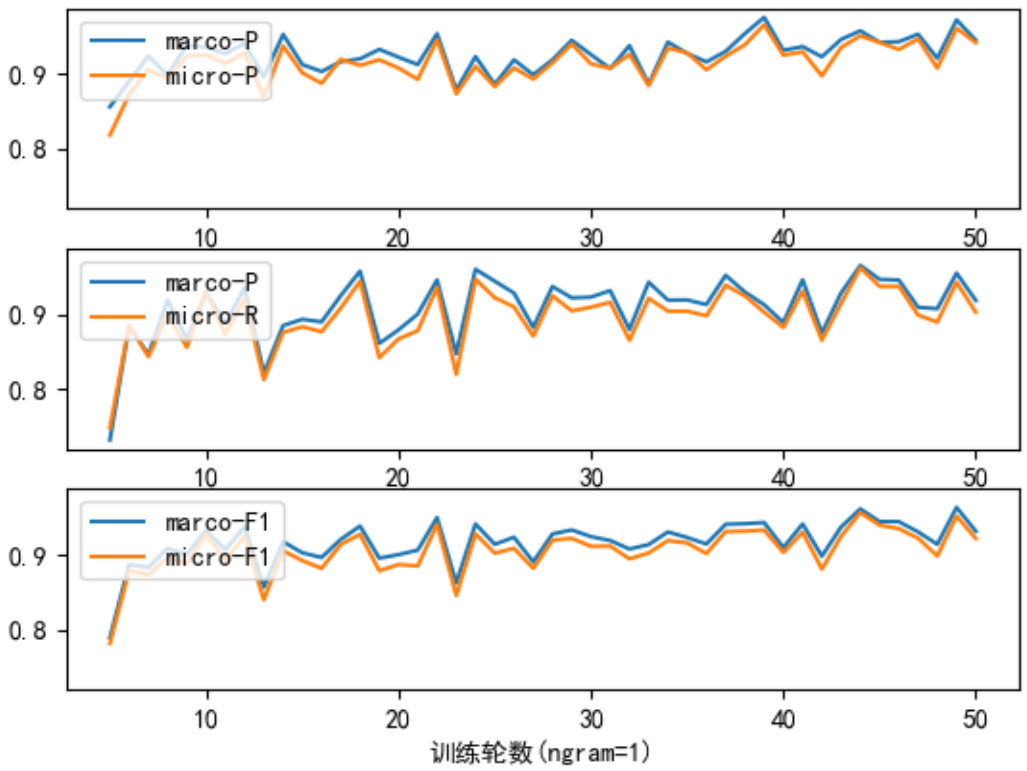
图2.系统流程图（自动问答）

问答系统设计--问题分类

- 基于关键词
  - 使用问句中的关键词（“招生计划”、“录取分数”）进行问句类型匹配
- 基于问句模板
  - 使用事先设定的问题模板同问句处理后得到的抽象问句进行相似度计算，得到问答句模板
- 基于fastText文本分类模型
  - 通过阳光高考网抓取985、211院校的招生咨询问题，分句后进行人工标注
  - 使用人工标注的问题数据60000余条20个分类训练fastText分类模型
  - 使用分类模型对输入的问句进行分类

问题分类测试结果

- 在每一类中随机选取10个问题作为测试数据，共200个句子进行测试，测试结果如下，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：



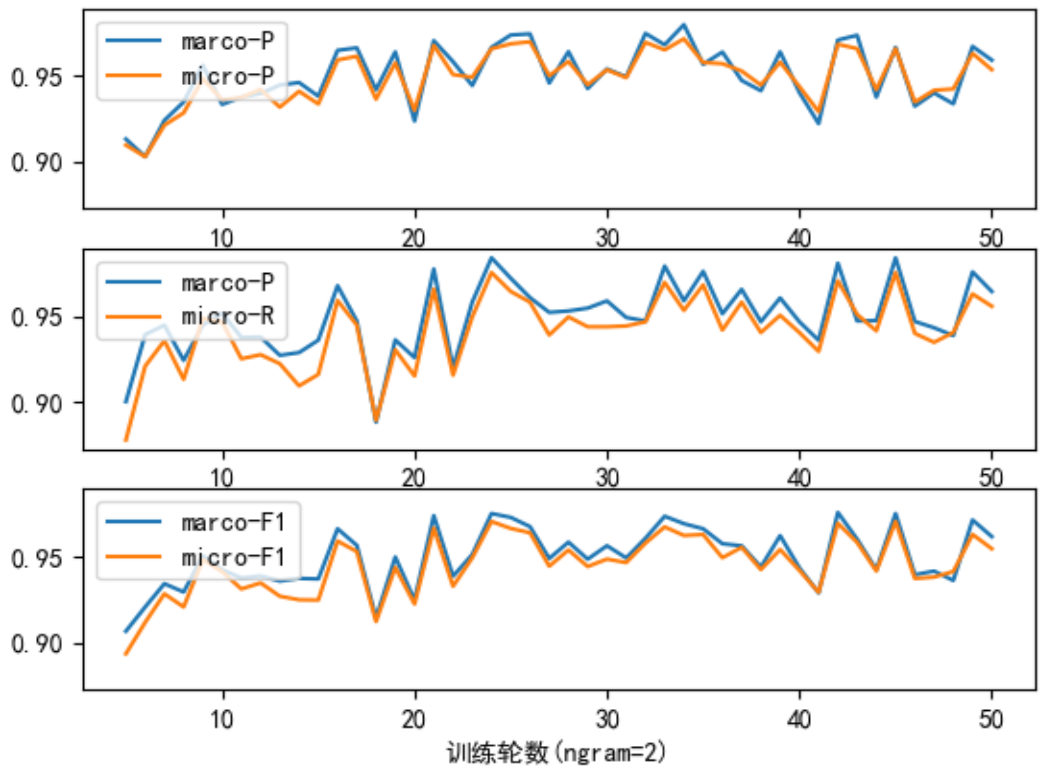


图3 问题分类测试结果

问答系统设计--问题处理

- 分词和词性标注：使用hanlp开源工具包的NLP分词器。
- 自定义词典与实体扩展：定义学校、专业词典，对学校名进行扩展。
- 去除停用词：标点符号、助词等虚词，加上领域常见词（“请问”、“贵校”、“谢谢”等）
- 关键词抽取与规范：学校词、地点词、时间词。
- 问句抽象与问句模板匹配：使用归一化的编辑距离进行衡量。

问答系统设计--查询语句构造与数据查询

- 查询语句构造
  - 基于问句模板的的问题类型判断在过程中能够得到关键词信息和问句模板，使用这两者进行SQL语句构造，问句模板对应带槽位的SQL语句，填入对应的关键词
  - 基于关键词和基于分类模型的问题类型判断得到问句类型，通过关键词抽取可以得到关键词，每一种问句类型对应有相应的SQL语句模板，使用关键词进行填充得到对应的SQL语句
- 数据查询：使用SQL语句查询得到带字段名的键值查询结果

问答系统设计—答句构造

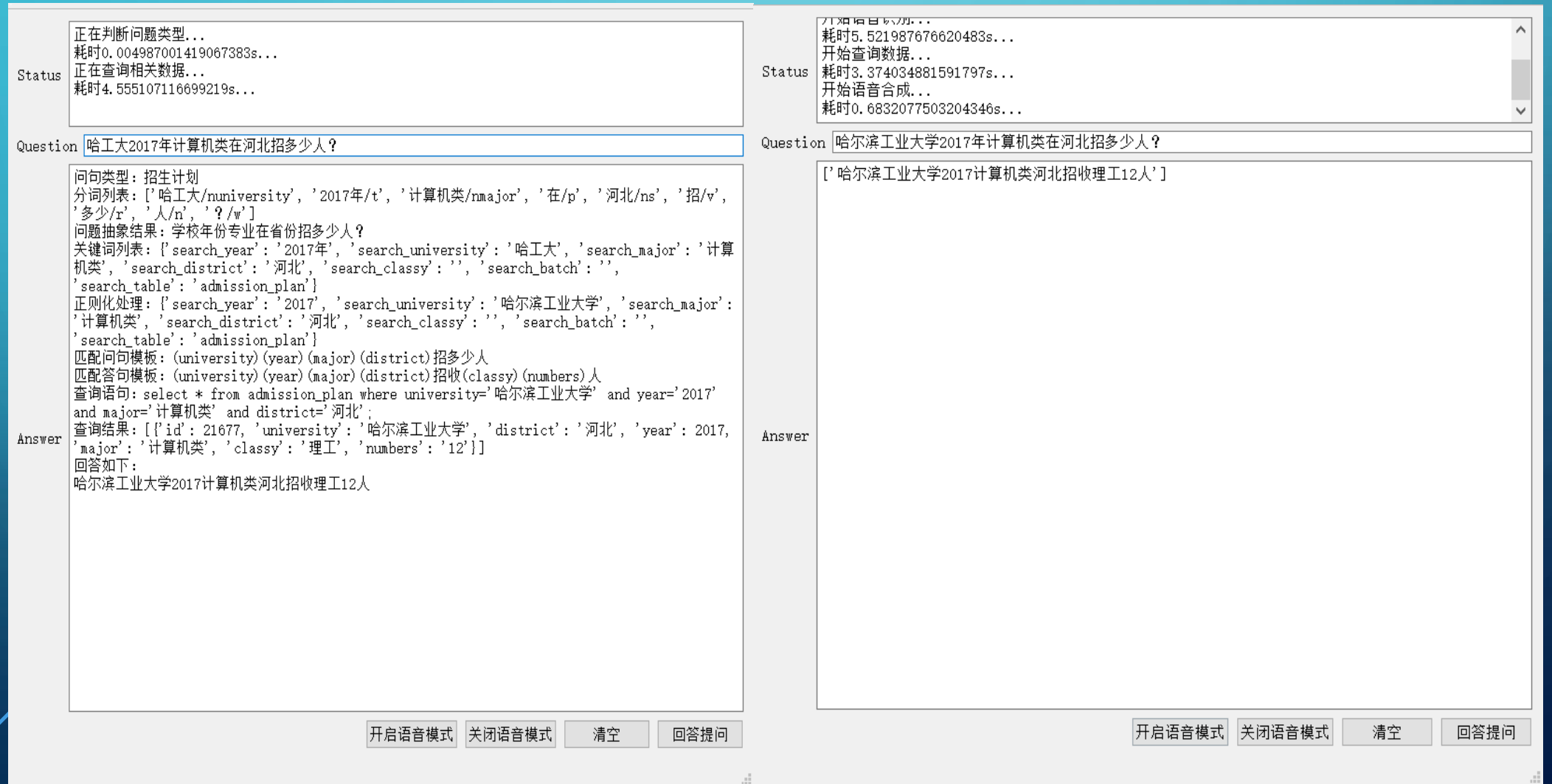
- 答句构造
  - 基于问句模板的问题类型判断在匹配模板时能够获得与之相对应的答句模板，最后使用查询结果和答句模板进行答句构造获得答句。
  - 基于关键词和基于分类模型的问题类型判断能够得到问题类型，问题类型能够与一个答句模板相对应，与查询结果相结合获得答句。

问答系统设计—UI设计

- 问答
  - 输入问句
  - 语音问答
- 数据库查看
- 模板查看与创建

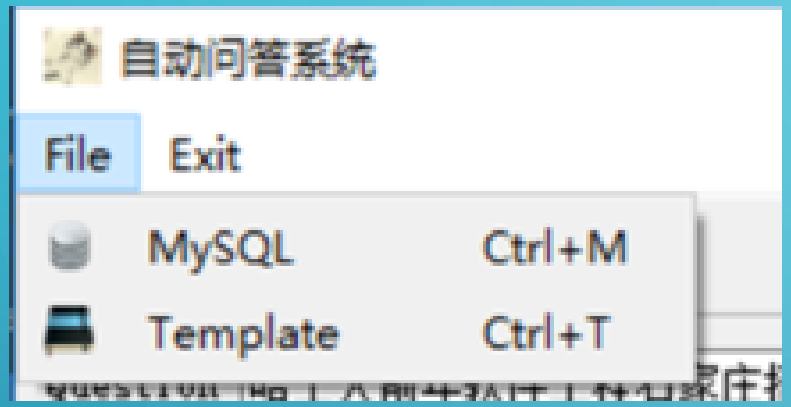
UI界面设计--问答界面

- 通过Question编辑栏输入问题或通过语音输入问题，得到查询结果或语音回复



UI界面设计--数据库按表项查询界面

进入方式①、菜单File→MySQL；②Ctrl+M

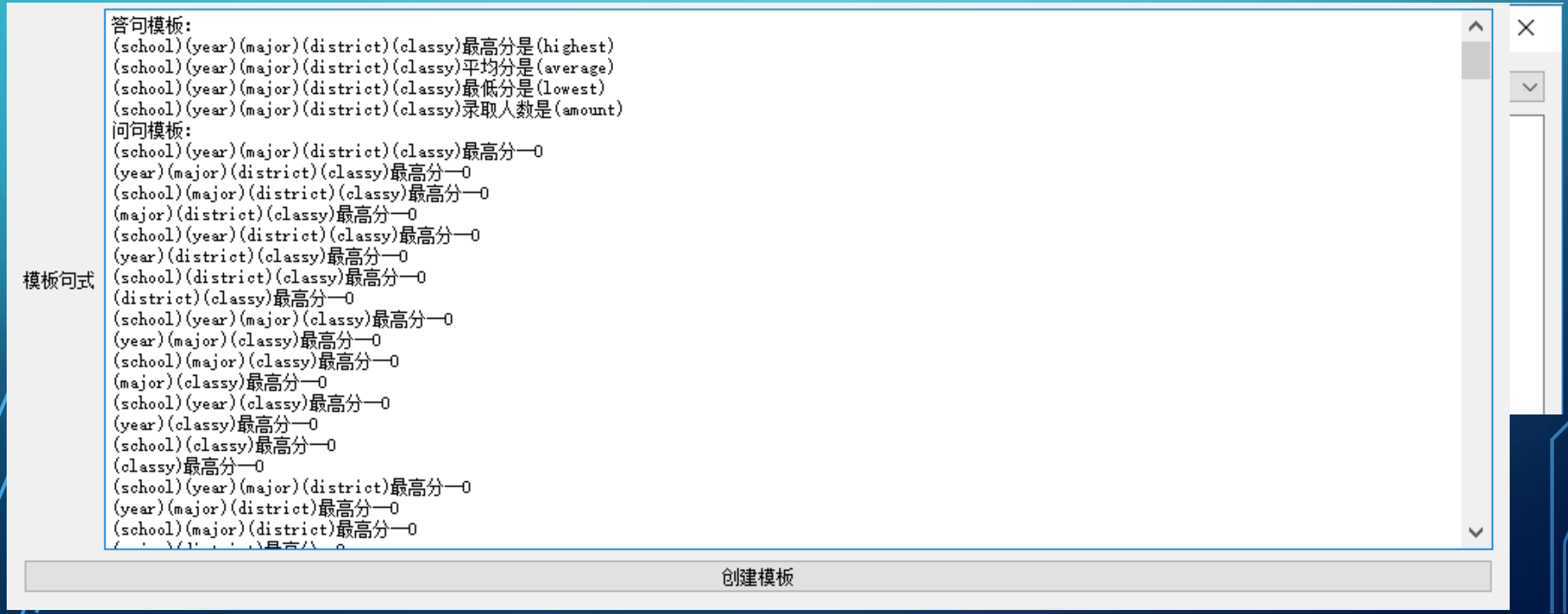


使用方式：①、依次选择下拉框选项，动态生成SQL语句；
②、直接写入合法的SQL语句；点击“查询”按钮，结果如下：

UI界面设计--模板查看界面

进入方式①、菜单File→Template；②Ctrl+T

点击“查看当前模板”，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、答句模板、问句模板）



UI界面设计--模板创建界面

- 点击“模板创建按钮”进入模板创建界面，输入相关信息创建模板，确认信息无误后即可进行模板创建，返回模板创建结果

问答系统设计—测试结果

- 测试数据：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。

问答系统设计—测试结果

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。

论文撰写情况—第一章 绪论

论文撰写已基本完成，正在调整和修改中

- 1.1 课题背景及研究的目的和意义
- 1.2 国内外在问答系统上的研究现状及分析
- 1.3 本文的主要研究内容
- 1.4 本文的组织结构
- 1.5 本章小结

论文撰写情况—第二章 领域数据的获取解析与存储

- 2.1 数据获取
  - 2.1.1 网络数据抓取技术
  - 2.1.2 领域数据来源分析
  - 2.1.3 数据获取的程序说明
- 2.2 数据解析
  - 2.2.1 网页数据解析技术
  - 2.2.2 解析excel文件和pdf文件
- 2.3 数据存储
  - 2.3.1 数据说明
  - 2.3.2 数据库表信息
- 2.4 数据规模
- 2.5 本章小结

论文撰写情况—第三章 问题分类

- 3.1 基于关键词的方法
- 3.2 基于问句模板的方法
  - 3.2.1 分词与词性标注
  - 3.2.2 关键词扩展
  - 3.2.3 关键词抽取与规范
  - 3.2.4 问句抽象
  - 3.2.5 问句模板匹配
- 3.3 基于分类模型的方法
  - 3.3.1 fastText模型介绍
  - 3.3.2 fastText模型训练
  - 3.3.3 fastText模型分类
- 3.4 问题分类方法的优先级与使用场景
- 3.5 本章小结

论文撰写情况—第四章 问题处理和答案生成

- 4.1 问题预处理
  - 4.1.1 分词与词性标注
  - 4.1.2 自定义词典与实体扩展
  - 4.1.3 去除停用词
  - 4.1.4 关键词抽取与规范
- 4.2 查询语句生成
  - 4.2.1 抽象问句匹配模板句式
  - 4.2.2 SQL语句构造
- 4.3 数据查询和答案生成
  - 4.3.1 SQL语句查询数据库
  - 4.3.2 查询结果同模板答句构造答案句
- 4.4 本章小结

论文撰写情况—第五章 面向高考招生咨询的自动问答系统设计

- 5.1 系统设计
- 5.2 系统功能
  - 5.2.1 问答
  - 5.2.2 数据库查询
  - 5.2.3 问题模板
- 5.3 实验结果及分析
  - 5.3.1 问题分类模块测试结果
  - 5.3.2 问答模块测试结果
- 5.4 本章小结

谢谢聆听

Q&A

学生：王陈阳
指导老师：赵铁军