基于维基百科的智能答题机器人的设计与实现

专业：计算机科学与技术

学生:陈强　　指导老师:刘正熙

随着互联网的发展和普及，互联网有着越来越多的用户。互联网上由用户自主产生的一些资料，信息也越来越多。为了更好地组织这些信息和资料，有像维基百科，百度百科，互动百科这样的网站出现，这些网站可以方便用户更好地管理，组织，共享知识。除此之外，还有数不胜数的网站包含着我们需要的信息，资料，为了方便用户更好地找到想要的资料，有像 Google, Bing, 百度一类的搜索引擎出现 。为了加快用户获取信息的速度，问答系统越来越受到研究者的关注，早在图灵测试的提出之时，人们就开始在往这方面努力。英文的问答系统的研究比较详尽，比较成功的案例有，IBM 的 Watson 问答系统，她曾经在一个问答性质的节目《危险边缘》中，答题速度与正确率超过人类。相对来说，现在成熟的中文问答系统比较少，中文问答系统遇到的困难要比英文的多的多，首先，英文的单词有空格作为单词与单词的分隔，而中文却不是这样，对中文信息进行处理的第一步就是要解决分词问题，另外中文的句子结构，语法相对英文来说会更加不规律，更加松散，这对中文信息的分析也增加了一定的难度。这里我们利用维基百科中的中文内容，以及现有的中文分词工具，尝试着设计和实现一个可以回答简单问题的答题机器人，也就是问答系统。这里设计的问答系统主要包括三个模块，问题分析模块，文档检索模块，答案抽取模块。这三个模块之间相互作用，配合，最后完成答题的工作。问题分析模块，采用贝叶斯分类器对问题进行分类，文档检索模块，运用了现在比较成熟的语言向量模型，BM25 文档与查询语句相似度的计算方式，答案抽取模块采用了隐马尔可夫模型进行对答案进行抽取。

关键词：维基百科；问答系统；信息检索；答案抽取

Design and implementation of answer Robot Based Wikipedia data

Computer Science and Technology

Student： Chen Qiang **Adviser:**Liu Zhengxi

With the development and popularization of the Internet, the Internet has more and more users. Some information on the Internet by the user self-generated information is also increasing. To better organize the information and materials, there are like Wikipedia, Baidu Encyclopedia, interactive encyclopedia such sites appear, it can help user to organize knowledge. In addition, there are numerous sites contain information we need, the information, in order to facilitate the users to find the desired information, there are like Google, Bing, Baidu a kind of search engine appears. In order to accelerate the speed of user access to information, more and more researchers focus on question answer system as early as at the time proposed the Turing Test. Study of English Question Answer system are more detailed, and the most one of successful cases is IBM's Watson question answering system, Who won human beings in a quiz show "Jeopardy". Relatively speaking, the Chinese QA System relatively less and weaker. Chinese QA system encountered more difficulties than English. First, the English word has a space as separate between words, but Chinese is not in the case. The first step in natural language processing is to solve word segment problems. Second, Chinese sentence structure and syntax relatively speaking are more irregular and more loosely than English. Here we try to use Wikipedia Chinese content and developed Chinese Word Segment Tool to design and implement a robot that can answer simple

questions, which is called question answer system. The designed system includes three modules, problem analysis module, document retrieve module, answer extraction module. The interaction between these three modules help to get the final answer. The problem analysis module uses Bayesian classifier to classify question type. The document retrieve module base language space vector model, use BM25 as document and query similarity calculation model. The answer extraction module contain Hidden Markov model.

Key Words: Wikipedia; Question Answer System; Information Retrieve; Answer Extraction.

目录

1 引言

近年来,随着信息技术的快速发展，网站数目不断增加，网络中的资料越来越丰富，为了帮助人们从这些数量巨大的资料库中找到所关心的资料，产生了 google 等一系列帮助人们获取相关信息资料的网站，我们称之为搜索引擎。用户给搜索引擎提供相关的关键字，搜索引擎可以帮助用户找出用户所需要的文档，用户再从文档当中找到所需要的信息。这个过程，可能出现两个问题，一是用户必须选择合适的关键字提供给搜索引擎，二是用户必须从文档中，找到自己需要的信息。为了更好，更精确地为用户提供信息，更进一步实现问答系统成为许多研究者的关注点，问答系统的输入不再是关键字，他可以是日常的一些问句，系统的输出也不再是一系列相关的文档，而是精确的答案。这样的系统具有更好的用户体验，可以帮助人们迅速得到关键的信息，相关的技术可以应用到医疗，客服，等领域。

我们这里，利用维基百科中相对丰富，结构清楚，干净的文本信息数据，构建一个可能回答简单的基于事实的问题的问答机器人。

现有搜索引擎提供商在不断进步，一些问题已经可以被回答了，比如说在百度中搜索，“上海的简称是什么”，可以有如下，图 1.1 的结果，这表明现在的搜索引擎服务提供商已经在向问答系统这方向发展了。

图 1.1 百度搜索截图

同时百度知道，等问答相关的互联网产品，也在提高搜索引擎帮助用户解决问题的能力，因为有这么多答案问题对，前期有一些基于问答问题对的问答系统的研出现。

为了利用网上其它丰富的资料，我们有必要研究如何利用这些资料，建造问答系统。

我们这里做的是基于信息检索的问答系统，IBM Waston, Google 就有包含基于信息检索的问答系统模块，这种问答系统模块无法进行推理的工作。

简要说一下，基于知识的问答系统。这种系统把用户的问题，用一种语义的方式进行重新表示，然后利用这种表示方式，到现有的结构化的数据或者资料中进行答案的查询，例如，时空数据库（geospatial databases），ontologies（Wikipedia infoboxes, dbPedia, WordNet, Yago），Restaurant review sources and reservation services，Scientific databases，典型的代表有：IBM Waston, Apple Siri, Wolfam Alpha, True Knowledge Evi.。

当然还有一种是混合方式的问答系统，即包含了信息检索的模块，也包含了知识推理的模块，IBM Waston 就是这种混合结构。

1.1 定义

什么是问答系统？问答系统的输入是日常生活中的自然语言的问题，输出是精简，准确的答案。比如，输入，“第一代 iPhone 是什么时候发布的？”，问答系统应输出，“2007 年 1 月 9 日”。

1.2 分类

问答系统的分类，从是否能够进行推理上可以分，一是拥有推理能力的问答系统，这种系统，具有推理能力，可以回答现有资料中没有直接给出答案的问题，这是一种高级水平的问答系统，二是无法进行推理的简单问答系统，这种问答系统，只能对问题进行简单分析，之后，从已有的资料中抽取相应的答案。

我们这里实现的基于维基百科的答题机器人，是属于后者，它不具有推理能力，其拥有的所有资料来自于维基百科中的中文内容。

同时这时原答题机器人还不能回答列表有关的问题，比如说，“戏剧的表

演形式多种多样，常见的包括哪些？”，也不能回答，怎么做一类的问题，比如“从四川大学怎么去省体育馆？”。

1.3 国内外发展现状

随着 TREC 的举办，国外英文版的问答系统，已经比较深入和完善，比较成熟的系统包括，IBM’s Waston，他于 2011 年 2 月 16，在美国的问答电视节目，《危险边缘》，中战胜了人类。同时还有苹果公司的 Siri，它可能直接与人进行对话，此外还有 WolframAlpha，START，等等。

国内这方面研究来处于起步阶段，在这方面有研究的机构和成果包括，复旦大学的问答系统，哈工大的问答系统组，中科院计算所等。

1.4 本文组织结构

第二部分，介绍了整个系统的设计方案，以及实现原理。第三部分介绍了实际进行的工作，第四部分展示了实验的结果，第五部分对整个系统进行了简要的评价，以及分析了系统可改进的方向。

2 系统结构设计

此问答系统与一般的问答系统结构相同，只不过数据来源，以及系统细节的处理上面会有一些差异。

整体的系统结构，包括文档检索模块，问题分析模块，答案抽取模块。

其中，问题分析模块，对问题进行处理，得到关键字，以及问题对应答案类型，关键字提供给文档检索模块，文档检索模块根据关键字，对已经索引过的文档进行检索，提取出相关度最高（我们认为相关度高，意味着其文档中更有可能包括着问题最终的答案）的几个文档，文档检索模块把检索出来的文档，提供给答案抽取模块。答案抽取模块，同时根据，问题分析模块中给定的问题对应答案类型的信息，从文档中抽取相应的答案。整个系统的工作流程可以通过下图 1.2 进行表示。

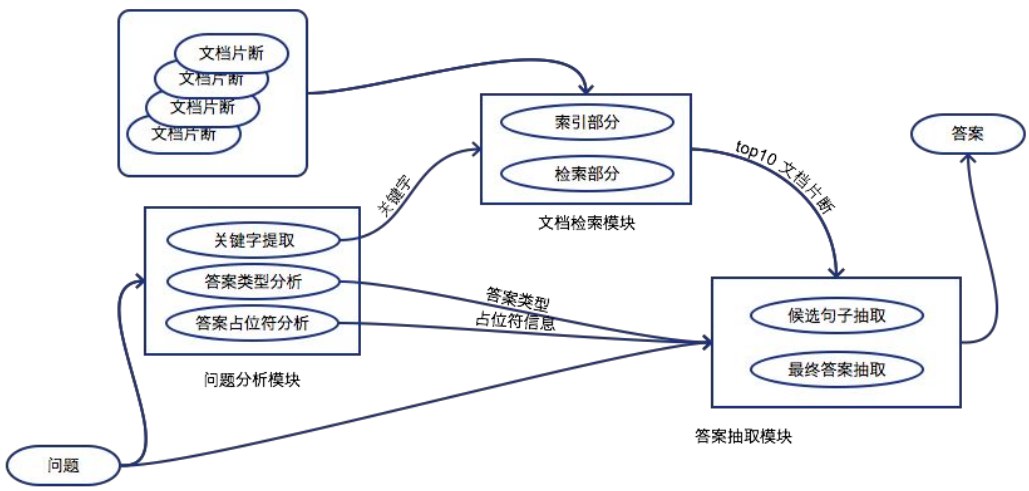

图 1.2 问答系统，三大模块结构图表

2.1 问题分析模块

此模块至关重要，其中关键字的准确性，直接影响到文档检索模块中检

索结果的准确性，问题类型的分类正确与否，也直接影响最后答案抽取模块的成功与否。

分词是对中文信息进行处理的基础，我们要对中文问题进行分析的第一步就是对中文句子进行分词，自 2003 年国际中文分词评测比赛 Bakeoff[1]举办以来，中文分词技术有了很大的进步。有基于词典，基于规则，基于统计等等方法，其中基于统计的分词方法的慢慢占据分词领域的上峰，至今，分词技术的准确率上召回率都可达 0.95 以上，更多分词历史的回顾，可以参考此[2] 。

因为分词技术相对成熟，现有的很多开源的分词器可以被借鉴。这里，我们直接使用现有的分词工具，python jieba 分词[3]，对中文句子进行分词,这个系统基于前缀词典实现高效的词图扫描， 采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

答案类型分类的实现方案，有手动规则的匹配分类方案，也有运用机器学习方法的分类方案。

手动规则的匹配举例，如果问题中出现“谁”这个关键字，那么这个问题的答案类型很有可能就是人名。出现“什么时候”对应的答案类型很可能是时间。

但是一般基于手动规则的方法，费时费力，一般会采用机器学习的方法对问题进行分类，我们这里采用的方法是运用语言词袋模型，基于贝叶斯概率模型的分类方法，后面会讲到详细的工作内容。

关键字提取中，利用分词工具进行分词，去除停用词，然后，利用索引模块中统计词语的词频，我们选择问题中包含的词语中词频较低的词语作为关键字。

2.2 文档检索模块

此模块的功能是为了快速地从大量的文档之中，快速抽取出我们需要的相关文档，此模块属于信息检索领域的范畴，空间向量模型和语言概率模型是这个领域最常用的两种方法，这里我们运用已经发展成熟的空间向量模型，实现我们的文档检索模块的功能，成熟的向量模型包括，Pivoted Length

Normalization VSM[11]和 BM25/Okpi 空间向量模型[10]。

这里利用的是 BM25/Okpi 空间向量模型[10]。

2.3 答案抽取模块

为了从候选的文档中抽取正确答案，我们首先可以通过，实体识别技术对候选文档中的实体进行识别，把与问题分析模块中分析出来的答案类别相同的实体类型的实体作为我们答案的候选词。我们根据候选词与问题中的各个词语的相对位置的关系，对各个候选词进行打分，分数最高的为我们最后抽取的答案。

3 相关的工作

3.1 数据准备

3.1.1 数据的获取

维基百科为对维基百科中的内容有使用兴趣的人提供了完整内容的电子档案[11]。这里选择的是，20141009 版本的镜像。我们只需要，维基百科内容中的文字信息，以及最新一版本的数据情况，而不需要维基百科内容的历史修改记录，因此我们选择下载页面中的，zhwiki-20141009-pages-articles-multistream.xml 其容量大小为 4.9G，其压缩版下载网址在这里[13]。

问题集的获取：这里使用的问题集来自万小军语义计算与知识挖掘课程[5]中的中文智能问答系统的问题集。

3.1.2 数据的整合与处理

先对下载之后的维基百科文件数据进行解压，得到了 xml 格式的文件，通过 xml2sql 工具[9]，把 xml 的文件格式转换成 SQL 格式的文件，导出的 SQL 格式文件大小为 4.0G ，之后利用 SQL 文件建立数据库。

维基百科的数据库结构可以参见维基百科官网[8]。

之后，我们对结构化的数据处理，清除不必要的信息，比如，在数据库文章正文中，用”''italic''”（这种两个单引号把文字围住）表示斜体，为了使文本更为干净，需要去除单引号，类似的情况还有很多，更多维基百科的字体规则可以参见维基百科给出的规范[7]。

我们对文章处理的过程如下：

1. 忽略表格，也就是 “{|” 和 “|}”中的内容。
2. 忽略，“<”和“>”中的内容，因为里面的都是 html 样式。
3. 忽略，“{{”和“}}”中的内容，因为里面是其它注释的内容。
4. 忽略，”[[File:”，和”]]”，以及其中间的内容。
5. 忽略，”[[Category:”，和”]]”，以及其中间的内容。
6. 忽略，“<ref>”和“</ref>”中的内容，因为里面是关于引用的内容。

7. 移除，句首的“----”，因为他表示水平线。

8. 移除，成对的五个，三个，两个，“’ ”，因为，成对的若干个“’ ”都是对文字做特定的格式显示。

9. 移除“[[”, “]]”中的内容，因为他们表示，他们中间的文字上有链接，这些链接可以调转到维基百科的其它页面。

10. 如果位于句首有多个连续的，”*”, 或者”#”，把位于句子首的“*”，以及”#”， 把第一个去掉，第二个以后的换成逗号。因为，“*”，以及”#”表示的是列表中的一个个条目，其中“*”表示无序条目，“#”表示有序条目。

11. 去掉空白行。

把文章切分成片断：此外，我们把文章中”==”和”==” 和其中的 文字，做为片断与片断的分隔符号，把中间的文字，以及当前 wiki 条目的标题做为片断的名称（也有可能是一个，或者多个”=”）。切割出来的片断一共有 3469833 个。

3.2 文档检索模块的实现

这里采用的是 BM25 空间向量模型[10]，其关键的查询语句与文档片断的相关度计算方法的计算函数表示如下：

$$f(q,d)=\sum_{w\in q\cap d}\frac{c(w,d)(k+1)}{c(w,d)+k(1-b+b\frac{|d|}{avdl})}\log\frac{M+1}{df(w)} \qquad (1\text{-}1)$$

其中，q 表示查询语句，d 表示文档,w 表示词语。c(w,d)表示词语 w 在文档 d 中的数目，k 和为方程的参数，df(w) 表示包含词语 w 的文档中数目。|d|表示文档的长度，avdl 表示所有文档的平均长度。M 表示所有文档的数目。

下面是为了加快计算相关度，给文档片断进行倒排索引的过程：

为了计算相关度，我们需要获取词语在文档中出现的次数，以及包含词语的文档数目，这些我们可以提前准备好。为了方便快速对这些数据进行储存和查询，我们建立一个包含 3 张数据库表的数据库。

使用的数据库为开源的 mysql 数据库,版本为 14.14。

下面几个表，表示了 3 张数据库表的结构。

表 1-1 词语表

term 表（也就是词语表，用来记录词语的信息）		
（字段名称）	（数据类型）	（说明）
term_id	INT	自增类型的 id，方便做索引
term	VARCHAR(100)	词语的内容
doc_frequency	INT	包含该词的文档片断数目
备注，我们在 term_id 上面做了主键索引，在 term 上面做了,unique 索引		

表 1-2 文档片断表

doc 表(也就是文档片断表，用来记录文档片断的信息)		
（字段名称）	（数据类型）	（说明）
doc_id	INT	自增类型的 id，方便做索引
doc_len	INT	文档的长度
doc_path	VARCHAR(200)	文档所在的路径
备注：把 doc_id 作为 主键索引		

表 1-3 词语文档关系表

doc_term 表（用来记录，词语与文档关系的数据）		
（字段名称）	（数据类型）	（说明）
doc_id	INT	文档 ID
term_id	INT	词语 ID
count	INT	该文档 ID 对应的文档路径中文档内容中含有该词语 ID 对应词语的数目
这里，对 term_id 做 B+ tree 索引，把(doc_id,term_id)作为组合主键索引		

我们对所有的文档片断进行扫描。

对于每个文档片断，把此文档片断信息插入 wiki_doc 表中，之后再对其进行分词处理，对于每个词语，如果它不在 wiki_term 表中，我们向 wiki_term 表中插入一条这个词语的数据，并初始化其 doc_frequency 为 1，如果它在 wiki_term 表中，我们把相应的 doc_frequency 增加 1。

停用词的获取方式：在上面的文档扫描处理过程处理完之后，我们选取在 term 表中选取 doc_frequency 最大的 300 个词语做为停用词。

3.3 问题分析模块

问题分类我们参考 TREC 相关的英文文献中提到了问题分类方法[4]，同时，针对我们自己的问题集，我们对之做相应的修改，同时表格分析了问题对应答案的词性，我们可以用候选文本中相应词性的文字作为候选答案，如果没有候选词性的词语，我们将考虑另外的方法获取答案。这里利用 jieba 工具对候选文本进行词性标注 。

表 2-1 问题分类表

问题类别	问题子类	例子	可能包含的词语	答案的词性
HUM,与人相关的问题	HUM_PERSON,人名	谁是湖南省的省长?	哪位，哪一位	nr 或 n
	HUM_ORG,组织名称	哪个公司最先拥有核武器?	机构，公司	nt 或 n
LOC,与地点有关的问题	LOC_UNIVERSE，行星	太阳系中唯一一颗没有磁场的行星是?	行星，星球	nz 或 n
	LOC_CITY ，	黄牛一词来源于哪里?	城市	ns 或 n
	LOC_CONTINENT	肯尼亚是属于五大洲中的哪个大洲的国家?	洲，大洲	n

	LOC_COUNTRY	芭蕾起源于哪个国家？	国家，国	n
	LOC_ COUNTY	哪个县人口最多？	县	ns
	LOC_STATE	美国哪些州发生过雪崩伤亡事件	州	ns
	LOC_PROVINCE	元谋人化石发现于中国的哪一省份？	省份，省	ns
	LOC_TOWN	中国古代大商人沈万三的故居位于苏州昆山的哪个古镇？	镇，古镇	ns
	LOC_RIVER	流过伊甸园的四条河中现存的有哪些？	河，河流	ns
	LOC_LAKE	世界上最大的淡水湖是哪个湖泊	湖，湖泊	n
	LOC_MOUNTAIN	江西第一个世界自然遗产，被誉为“江南第一仙山”的是哪座山？	山，山脉	n
	LOC_OCEAN	世界上最北的洋是哪一个？	洋，海洋	n
	LOC_ISLAND	玻利维亚位于哪一块大陆？	陆地，大陆	
	LOC_BUILDING	18 世纪末期，北欧海盗袭击了哪所修道院		
	LOC_BASIC	黄牛一词来源于哪里？		ns
NUM	NUM_COUNT	央视拥有多少个电视频道？	个	m
	NUM_MONEY	图灵奖目前的奖金有多少？	钱，奖金，金币，钞票	m
	NUM_PERCENT	飞机头等舱的票价通常至少为普通舱位票价的百分之多少？	百分比，百分，	m

	NUM_DISTANCE	磁悬浮列车速度最高可达每小时多少公里?	公里，里，长，长度，米，厘米，毫米	m
	NUM_WEIGHT	埃菲尔铁塔的结构使用了多少公吨的熟铁?	吨，公斤，斤，kg,g,克，公吨	m
	NUM_DEGREE	水的沸点是多少?	度	m
	NUM_AGE	美国最年轻的总统多大年纪?	年纪，年龄，岁数，岁	m
	NUM_RANGE	汉朝存在的时间范围是多少	范围	m
	NUM_SPEED	CDMA手机上网速度有多快?	快，速度，速	m
	NUM_FREQUENCY	RDRAM能够以怎样的频率工作	频率	m
	NUM_SIZE	金星的体积有多大	容量，体积	m
	NUM_AREA	这块木板的面积是多大?	面积，多大，平方米，平方	m

	NUM_BASIC	光速是多少		
	NUM_CODE	北京大学的联系电话是什么		
	NUM_PERIOD	《双子神偷》要拍多久		
	NUM_RANK	这是《北斗神拳》中的哪一集		
TIME	TIME_YEAR	自行车大约于哪一年传入中国?	年，年份	m
	TIME_MONTH	《十面埋伏》几月份全国上映?	月，月份，	m
	TIME_DAY	毛泽东是几号出生的?	号，日	
	TIME_SEASON	古代的死刑在什么季节行刑	季节	
	TIME_BASIC	Adobe Audition第一版于什么时候发布?	时候，时间	m
OBJ	OBJ_CURRENCY	赞比亚使用哪种货币	货币，币	m
	OBJ_MUSIC	威廉史密斯演唱的那首歌颂父母的歌曲叫什么名字	歌，音乐，名字	n
	OBJ_MOVIE	哪部电影是唯一一部获得奥斯卡最佳外语片的华语电影?	电影，哪部	n
	OBJ_ANIMAL	加拿大的官方动物是什么?	动物	n
	OBJ_COLOR	南京市市旗的下部三份之一是什么颜色?	颜色，色，色彩	n
	OBJ_BASIC	武汉腐乳是用什么发酵的?	什么	n

	OBJ_FOOD	俄罗斯人冬天的主食是什么	什么	
	OBJ_FLOWER	人称“花中的隐士”指的是什么花?		
	OBJ_ACADEMIC	请问自考经济学有哪些课程		
	OBJ_EVENT	在霸桥上经历的重大历史事件分别是什么事件		
	OBJ_INSTRUMENT	最动听的乐器是什么		
DES	DES_ABB	简称CPR代表什么	代表，什么，简称	nz
	DES_MEANING	近视指什么	指，是	?
	DES_REASON	什么原因意致使霸桥闻名古代中国?	原因，为什么	?
	DES_POEM_NEXT	陶渊明《饮酒》中的诗句“采菊东篱下”的下一句是什么?	下一句	
	DES_BASIC	MEP中文怎么翻译		
	DES_EXPRESSION	MEP中文怎么翻译		
	DES_JUDGE	矿泉水和果汁哪个容易结冰		
	DES_MANNER	DES_MANNER		

问题分类方法：

这里采用词袋模型对问题进行建模，用贝叶斯后验概率模型对问题进行分类。

$$P(C_i|D_j)=\frac{P(C_i)P(D_j|C_i)}{P(D_j)}=\frac{P(C_i)\prod_{k=1}^{D_j}P(W_k|C_i)^{TF(W_k)}}{P(D_j)} \quad (2-1)$$

公式中，C_i 表示问题的某一类别，D_j 表示问题 j，W_k 表示问题 j 中出现的词语 k，TF（W_k）表示词语 k 在问题 j 中出现的次数，$P(C_i)$和 $P(W_k|C_i)$都可以从分类器的训练过程中获取，我们选择概率最大的问题类型做为问题 j 的问题类别。

下面就是上述公式的实现过程：

首先，这里从问题集中选出 500 个问题，一一对这 500 个问题进行类别的标注以及核查。

第一步，我们计算，各个类别的词语出现的概率向量<$w_1,w_2,w_3...$>,

$$w_k=P(W_k|C_j)=\frac{1+\sum_{i=1}^{|D|}N(W_k,d_i)}{|V|+\sum_{s=1}^{|V|}\sum_{i=1}^{|D|}N(W_s,d_i)} \quad (2-2)$$

其中，|V|表示词语的数量,|D|表示，问题数量，$N(W_k,d_i)$表示，词语 k 在问题 i 中出现的次数。

第二步：用训练好的数据，对新的问题进行分类，此问题 i 属于类型 j 的概率：

$$P(C_j|d_i)=\frac{P(C_j)\prod_{k=1}^{n}P(W_k|C_j)^{N(W_k,d_i)}}{\sum_{r=1}^{|C|}P(C_j)\prod_{k=1}^{n}P(W_k|C_j)^{N(W_k,d_i)}} \quad (2-3)$$

其中，$P(C_j)=\frac{\text{类别}C_j\text{的问题数量}}{\text{参与训练的问题数量}}$，$P(W_k|C_j)$就是问题类型 j 中向量 w_k的值。

选择，$P(C_1|d_i),P(C_2|di)$.....中概率最大的类型，在计算时，可以把（2-3）式中的分母忽略不计，因为计算过程中，所有的分母都相同。

另外我们还参照了张宇等人的论文[14]，对此贝叶斯的模型进行的改进。

答案占位符的提取：根据疑问句语法规则，以及对我们样本集中 100 个问

题的观察，我们发现，所有的问题都中都会包括下列词语的其中之一：什么，哪，谁，多少，多，几，何，是？，为？，为什么。

我们对占位符的提取算法如下：

如果发现问题是以“是？”，或者，“为？”结尾，我们把就把对应的，“是？”或者“为？”作为，占位符。如果不是，我们先对问题进行分词处理，按照词语的顺序序列，我们对词语序列扫描两遍，只要扫描到符合条件的占位符，算法就结束：

第一遍，如果发现词语为“谁”，则找到占位符，“谁”并返回结果，如果发现，词语为，“什么”，“哪”，“多少”，“多”，“几”，“何”是的任何一个，我们把这个词语，以及下一个词语组成的新的词语返回，如果存在下一个词语的话，如果不存在，直接返回当前词语。

第二遍，如果发现词语以“什么”，“哪”，“多少”，“多”，“几”，“何”任意一个词打头，则返回该词语。

如果两遍扫描之后没有结果，则返回空。

关键字提取：提取问题中的关键字。先从问题中去掉答案占位符，再对去掉答案占位符的问题进行分词处理，得到一系列的词语，根据停用词表去掉停用词，去 term 表中查询每个词语的词频（词频就是包含这个词语片断的数目），如果 term 表中没有这个词语，则我们设置这个词语的词频为 0，我们选取词频最小的 3 个作为问题的关键字，如果不足 3 个，我们把所有的词语都作为关键字。

3.4 答案抽取模块

我们前面通过问题分析模块得到关键字，通过文档检索模块，用关键字在进行档案检索，我们得到前 10 个文档，现在我们从文档中提取答案。

首先，我们先从 10 个文档中选取最有可能包含答案的句子。

我们还要对问题进行分词处理，另外还要把答案占位符特别地标记出来，同时保持词语和词语之间的位置信息。

我们对 10 个候选文档做如下的处理：

把问题中各个词语，匹配到候选文档中，匹配到的问题中的词语越多，匹配到的词语的词频越低越好，匹配之后各个词语的位置信息与原来问题的

位置信息越接近，我们认为这个文档中越有可能包括正确的答案。

我们选取了最有可能包含正确答案的文档之后，就要对答案进行提取了。

首先，我们通过上面问题类型对应的答案词性把，所有相应的词性的所有词语提取出来，之后，我们给各个词语进行打分，分数最高的就是最有可能的答案。

打分的过程如下：同上面从 10 个候选文档中选择最有可能的包含答案的文档一样，我们先把问题中的词语，匹配到选择出来的文档中，把答案占位符与文档中各个词语位置距离分别除以候选答案到文档中各个词语位置的比例之和做为，此候选词的得分。

之后，我们选择得分最高的词语做为答案。

同样我们要对获取到的资料进行分词处理，这次我们同时利用 jieba 工具给词语标注上词性，我们根据问题分析模块中各个答案类型对应的词性，从候选文本中抽取出相应的词性的词语，我们同时记录，各个词语的位置信息，用相应的数字表示。

4 实验结果

下面是各个模块的实验结果

4.1 问题分析模块

我们利用利用手工标注的问题及问题类别，对贝叶斯模型进行训练，用训练之后的模型对新出现的问题进行分类。为了测试模型的分类正确率，我们进行如下的操作：

1.手工标注的 497 个问题中，随机抽取 397 个问题用来，训练贝叶斯模型，用训练之后的模型对剩下的 100 个问题进行分类测试，把分类结果与人工标注的问题类别进行对比，计算正确率。

2.重复步骤 1 十次，得到 10 个正确率

3.我们对步骤 2 得到的 10 个正确率进行，求正确率平均值作为我们实验最后的结果。

我们手工标注了 497 个问题，下面我们展示了问题标注数据的一部分，其中，左边的代码为问题类别的代码，右边为问题数据，问题类别的代码与表 2-1 问题分类表，中的数据相对应，U 表示在 2-1 问题分类表中无法找到相对应的问题类别。

2-2 问题分类数据表

问题类别	问题
DES_ABB	国民生产总值的英文简写是什么
DES_EXPRESSION	从地图上看，黄河干流像一个巨大的汉子，是什么字？
DES_EXPRESSION	数学公式中表示角度的变量一般是使用哪种字母
DES_MANNER	现代工业制备氧化性最强的单质的方法是？
DES_MEANING	在古文中，“汝”、“尔”是什么意思？
DES_POEM_NEXT	形容自己经历的事情，别人不宜插嘴的俗语“如人饮水”的下一句是什么？
DES_REASON	台湾为什么要独立
HUM_ORG	杨魏玲花、曾毅是哪个以“民族风”“接地气”为特点的流行组合的成员？
HUM_ORG	《哈利波特》系列电影中哈利就读的魔法学

	校叫什么名字？
HUM_ORG	NBA 联盟中唯一主场没有设在美国境内的是哪支球队？
HUM_PERSON	金庸小说《笑傲江湖》中，爱慕令狐冲的恒山派小尼姑叫什么？
HUM_PERSON	日本动画片《名侦探柯南》中少年侦探团初期 4 名成员中唯一的女性成员是谁？
HUM_PERSON	我国男子羽毛球双打主教练，在悉尼和雅典奥运会上与高崚搭档两度获得奥运混双冠军的是哪位运动员？
LOC_BASIC	150001 是哪里的邮编
LOC_BASIC	第一例 SARS 在哪里被发现
LOC_BASIC	世界上石油存储最多的是哪个地区
LOC_BUILDING	世界最大的金字塔是什么金字塔？
LOC_BUILDING	中国神话当中，嫦娥在月亮上住的行宫叫什么？
LOC_BUILDING	2008 年北京奥运会的主场馆被人们称为“鸟巢”，那么 2012 年伦敦奥运会的主场馆被俗称为什么？
LOC_CITY	哪个城市拥有世界上最长的地铁
LOC_CITY	网球运动中最古老和最具声望的赛事温布尔登网球公开赛是在哪个城市举办的？
LOC_COUNTRY	目前已知的世界上第一部比较完整的成文法典《汉摩拉比法典》成书于哪个古代文明古国？
LOC_COUNTRY	世界上最后一个独立的殖民地，旧名西南非洲的是哪个国家？
LOC_COUNTRY	火药是哪国人发明的
LOC_ISLAND	世界上最大半岛是哪个
LOC_ISLAND	历来冲突不断，有“欧洲火药桶”之称的是欧洲的哪个半岛？
LOC_ISLAND	世界第二大岛是什么
LOC_LAKE	我国最大的淡水湖是什么湖？
LOC_LAKE	我国最大的淡水湖是哪个淡水湖
LOC_MOUNTAIN	江西第一个世界自然遗产，被誉为“江南第一仙山”的是哪座山？
LOC_MOUNTAIN	美国准备在哪座山上建立军事基地
LOC_MOUNTAIN	少林武术的发祥地少林寺位于我国河南的

	哪座山？
LOC_OCEAN	世界上流量最大的海洋是哪个
LOC_OCEAN	丹江流向哪个海
LOC_OCEAN	世界上最北的洋是哪一个
LOC_PROVINCE	中国面积最大的省份是哪个省
LOC_PROVINCE	汽锅鸡以鸡肉滋嫩.汤汁鲜美.富于营养而广为流传，请问它是我国哪个省独有的高级风味菜？
LOC_RIVER	卢沟桥在哪条河上面？
LOC_STATE	曾举办过冬奥会的犹他州盐湖城是哪一支NBA 球队的所在地？
LOC_UNIVERSE	太阳系九大行星中位于最外侧的是什么星
NUM_AGE	张三今年多大了
NUM_AGE	我国《道路安全法》规定，在道路上驾驶自行车，三轮车必须年满多少周岁？
NUM_AGE	李小华多大年纪了呀
NUM_AREA	多大的房子算大
NUM_AREA	这块木板的面积是多大
NUM_BASIC	1982 年宪法是我国建国以来第几部宪法
NUM_BASIC	罗马数字中的“X”表示阿拉伯数字中的几？
NUM_BASIC	四川的面积是香港的多少倍
NUM_CODE	长沙的邮编号是多少
NUM_CODE	国家主席的联系电话是多少
NUM_COUNT	我们俗称的“季度”，按时间划分几个月为一个季度？
NUM_COUNT	堪培拉有多少人口
NUM_DEGREE	多少摄氏度时原子停止震动
NUM_DEGREE	黄瓜的内部比外面的空气低多少度
NUM_DISTANCE	本垒打要跑多远
NUM_DISTANCE	大气层厚度约为多少？
NUM_DISTANCE	月亮的直径是多少
NUM_DISTANCE	从地球到火星有多远
NUM_FREQUENCY	地球发射地磁波的频率为多少
NUM_FREQUENCY	在什么频率范围出现数百伏/米的辐射
NUM_MONEY	上海到南京的火车票多少钱一张
NUM_MONEY	一台笔记本大概多少钱
NUM_MONEY	比喻想要隐瞒掩饰，结果反而暴露是说此地

	无银多少两？
NUM_PERCENT	中国农村儿童的失学率是多少
NUM_RANGE	中度酒的酒精含量在多少范围之内？
NUM_SPEED	火星上的重力加速度是多少
NUM_SPEED	高速公路的行车最低时速是多少
NUM_SPEED	空气中的声速能达到多少
NUM_WEIGHT	女运动员掷铅球重量是多少公斤？
OBJ_ACADEMIC	理科综合是指高中学习科目的物理、化学和什么？
OBJ_ACADEMIC	文科综合是指高中学习科目的政治、历史和什么？
OBJ_ANIMAL	雌驼鹿的学名是什么
OBJ_ANIMAL	动画片《蓝猫淘气三千问》中的“甜妞”这一形象是以什么动物为原型的？
OBJ_BOOK	我国收入字最多的字典是哪一部
OBJ_BOOK	黄仁宇有一本书围绕公元 1587 的明代历史展开，这本书是？
OBJ_COLOR	你认为组成橙色的主要颜色是什么
OBJ_COLOR	微软的 word 软件中，保存按钮图标默认是一个什么颜色的磁盘？
OBJ_EVENT	动漫《七龙珠》中由富人出钱举办的评出天下第一的武斗比赛是？
OBJ_EVENT	1895 年 5 月，康有为率梁启超等数千名举人联名上书清光绪皇帝的事件，史称什么？
OBJ_FLOWER	有“凌波仙子”“玉玲珑”美称的花是什么？
OBJ_FLOWER	成语中把什么花短暂开放指代美好的事物出现的时间很短？
OBJ_FOOD	最香的食物是什么
OBJ_FOOD	通常所说的“三高食品”指高蛋白、高脂肪和高什么的食品？
OBJ_INSTRUMENT	最常用的乐器是什么
OBJ_MOVIE	“西边的太阳快要落山了”这句歌词是出自哪部电影的插曲？
OBJ_MUSIC	陈奕迅的哪首歌歌词由 65 首歌名组成，第一句就包含了《我》《以为》《用心良苦》三个歌名？
OBJ_MUSIC	信乐团与 beyond 乐队有一首歌曲的名字一样的是哪首歌曲？

OBJ_MUSIC	歌词“没有你的日子里，我会更加珍惜自己”出自着名歌手齐秦的哪首歌？
OBJ_PLANT	北方最多的树叫什么树
OJB_FLOWER	宋代词人李清照的名句“人比黄花瘦”、“满地黄花堆积”中，黄花指的是哪种花？
TIME_BASIC	第一届奥运会什么时候举行
TIME_BASIC	二十四节气中，表示春季到来是对应哪个节气？
TIME_DAY	端午节是哪天
TIME_DAY	《独立宣言》是哪天签署的
TIME_DAY	德怀特·D·艾森豪威尔的生日是哪一天
TIME_MONTH	愚人节在几月份
TIME_RANGE	一塌糊涂 BBS 从哪年建立到哪年被关闭
TIME_SEASON	古代的死刑在什么季节行刑
TIME_YEAR	2003 年日本经济景气一致指数哪几个月份低于 50%
TIME_YEAR	美国参议院代表团访问中国在哪一年
U	用五条等距离的平行线来记录音符的形式叫什么谱？
U	单字在姓名里面读什么？
U	机动车在道路上通告须要携带的有机动车行驶证，车牌，检验剑桥标志，强险标志和什么？

手工标注数据中，一共标注的问题类别一共有 52 个类别，问题类别最多的是 HUM_PERSON 类别，问题分类数据统计结果如下：

2-3 问题类别数据统计

问题类型	问题数量	问题数量百分比
U	103	20.724%
HUM_PERSON	98	19.718%
DES_POEM_NEXT	25	5.030%
LOC_COUNTRY	21	4.225%
LOC_CITY	18	3.622%
LOC_BASIC	17	3.421%
NUM_COUNT	15	3.018%

HUM_ORG	14	2.817%
TIME_BASIC	11	2.213%
LOC_PROVINCE	10	2.012%
OBJ_ANIMAL	8	1.610%
OBJ_FOOD	8	1.610%
DES_MEANING	8	1.610%
LOC_BUILDING	7	1.408%
OBJ_BOOK	7	1.408%
TIME_YEAR	7	1.408%
OBJ_EVENT	7	1.408%
LOC_LAKE	7	1.408%
DES_ABB	7	1.408%
TIME_DAY	6	1.207%
OBJ_MUSIC	6	1.207%
OBJ_MOVIE	5	1.006%
NUM_PERCENT	5	1.006%
NUM_BASIC	5	1.006%
OBJ_FLOWER	4	0.805%
OBJ_COLOR	4	0.805%
NUM_DISTANCE	4	0.805%
DES_EXPRESSION	4	0.805%
NUM_CODE	4	0.805%
TIME_MONTH	4	0.805%
OBJ_INSTRUMENT	3	0.604%
NUM_AGE	3	0.604%
DES_MANNER	3	0.604%
LOC_ISLAND	3	0.604%
LOC_MOUNTAIN	3	0.604%
NUM_MONEY	3	0.604%
LOC_RIVER	3	0.604%
NUM_SPEED	3	0.604%
DES_REASON	3	0.604%
LOC_OCEAN	3	0.604%
TIME_RANGE	3	0.604%
NUM_AREA	2	0.402%
OBJ_ACADEMIC	2	0.402%
NUM_DEGREE	2	0.402%

NUM_FREQUENCY	2	0.402%
LOC_STATE	1	0.201%
OBJ_PLANT	1	0.201%
NUM_WEIGHT	1	0.201%
LOC_UNIVERSE	1	0.201%
NUM_RANGE	1	0.201%
TIME_SEASON	1	0.201%
OJB_FLOWER	1	0.201%

10 次测试，问题分类的正确个数分别是 47，38，41，42，42，38，37，38,42，42。分类平均正确个数为 40.7,分类的平均正确率为 40.7%。

4.2 文档检索模块

100 个问题答案对作为测试数据,利用关键字抽取程序从问题中抽取的关键字，输入到文档检索系统，进行文档的检索，如果检索出来的内容中，包含了答案，则我们认识此次检索是成功的，如此来计算检索成功率。

我们的部分问题样本如下：

2-4 问题答案部分样本

哪部电影是唯一一部获得奥斯卡最佳外语片的华语电影？	卧虎藏龙
武汉腐乳是用什么发酵的？	枯草杆菌
哪种食物在 2011 年 6 月被美国 CNN 的 iReport 栏目评为“世界最恶心的食物”头名？	皮蛋
自行车大约于哪一年传入中国？	1875 年
谁是右手中国式直拍弧圈结合两面快攻打法选手？	王皓
黄牛一词来源于哪里？	上海
周星驰其名出自于《滕王阁序》中的哪一句？	雄州雾列，俊彩星驰
《愤怒的小鸟》（芬兰语：Vihainen Lintu，英语：Angry Birds）是哪个公司推出的一款益智游戏？	芬兰 Rovio 娱乐
第一代 iPhone 是什么时候发布的？	2007 年 1 月 9 日
魔方是由谁发明的机械益智玩具？	比克·艾尔内
Adobe Audition 第一版于什么时候发布？	2003 年 8 月 18 日

央视拥有多少个电视频道？	45
冯小刚的现任妻子是谁？	演员徐帆
王宝强在哪一年首次参加中央电视台“春晚”？	2008
芭蕾起源于哪个国家？	意大利
戏剧的表演形式多种多样，常见的包括哪些？	话剧、歌剧、舞剧、音乐剧、木偶戏等
在美国，已知最早的十字绣样品目前在存放在马萨诸塞州普利茅斯的哪里？	Pilgrim Hall 博物馆
北京大学现有几个校区？	六个
义务教育的三个基本原则是什么？	强制、普遍与免费
2011 年 8 月，莫言创作的什么长篇小说获得第 8 届茅盾文学奖？	《蛙》
截拳道是谁创立的一类现代武术体系？	李小龙
哪个是唯一两次获选为世界三大夜景的城市？	香港
埃菲尔铁塔的结构使用了多少公吨的熟铁？	7,300
《天空之城》是何时推出的一部长篇动画电影？	1986 年 8 月 2 日
乒乓球球台被一个多高的球网分为两部分？	15.25 厘米
邓亚萍多大时夺得了首个世界冠军？	16 岁
在历史上乌镇曾出过多少名举人？	161
南京市市旗的下部三份之一是什么颜色？	粉绿色
全世界约有多少佛教信众？	5 亿
《本草纲目》撰成于哪一年？	万历六年（1578 年）
元曲共计多少支曲牌？	447
人的岁数等于 12 的倍数称为什么？	本命年
服务业目前于发达国家的产业比重约占多少？	70%以上
姚明的父亲姚志源多高？	6 英尺 10 英寸（2.08 米）
马尔代夫的第一大支柱产业是什么？	旅游业
澳大利亚是南半球面积第几大的国家？	二
阿根廷国家足球队赢得过多少次美洲杯冠军？	十四
平均人口密度最高的大洲是？	欧洲

中国历史上最杰出的浪漫主义诗人是？	李白
甲午战争爆发的标志是？	丰岛海战
《华英字典》的作者是？	马礼逊
《资治通鉴》的撰写一共耗时多少年？	19
世界上首位从南坡登上珠穆朗玛峰的女性的国籍是？	日本
最后宣布废除奴隶制度的行政地区是？	毛里塔尼亚
国际海洋法法庭的总部位于哪个国家？	德国
《苏德互不侵犯条约》的签订地点是？	莫斯科
莫言获得诺贝尔文学奖的年份是？	2012
元谋人化石发现于中国的哪一省份？	云南
美国的当代人类学通常划分为几大分支？	四
“中华民族”一词最早由谁提出？	梁启超
欧洲的第一所大学是？	博洛尼亚大学
图书馆学最早由谁提出？	马丁·施莱廷格
约翰纳什的博士学位是在哪所大学获得的？	普林斯顿大学
现存所有老虎亚种中最小的亚种是？	苏门答腊虎
目前的北回归线位置是由哪个组织确定的？	联合国教科文组织
首先分离出元素氟的化学家是？	亨利·莫瓦桑
第一位获得菲尔兹奖的女性是？	玛丽安·米尔札哈尼
薛定谔方程的提出者的国籍是？	奥地利
太阳系中唯一一颗没有磁场的行星是？	金星
位于双子座和狮子座之间的星座是？	巨蟹座
世界最大的流动性沙漠是？	塔克拉玛干沙漠
人类合成的第一种抗菌药是？	磺胺
第一种以人工合成无机物质而得到的有机化合物是？	尿素

检索结果中，有 33 个检索内容中包括正确答案，检索成功率为 33%。部分检索结果如下：

2-5 部分检索结果

问题	答案	关键字	包含答案	包含答案的文档语句
芭蕾起源于哪个国	意大利	芭蕾,起源于	包含	“芭蕾”起源于意大利，兴盛于法国，其部分手势可追溯至古埃及的祭祀舞蹈。

家？				
北京大学现有几个校区？	六个	北京大学,校区,现有	不包含	
义务教育的三个基本原则是什么？	强制、普遍与免费	义务教育,原则,三个	不包含	
2011 年 8 月，莫言创作的什么长篇小说获得第 8 届茅盾文学奖？	《蛙》	茅盾文学奖,莫言,长篇小说	包含	2011 年 8 月，莫言凭长篇小说《蛙》获第八届茅盾文学奖。
南京市市旗的下部三份之一是什么颜色？	粉绿色	市旗,三份,下部	不包含	
全世界约有多少佛教信众？	5 亿	信众,全世界,佛教	不包含	
《本草纲目》撰成于哪一年？	万历六年(1578 年)	撰成,本草纲目	包含	《本草纲目》是一部集 16 世紀以前，中國本草學大成的著作。作者是明朝的李时珍，撰成于万历六年（1578 年），万历二十三年（1596 年）在金陵（今南京）正式刊行。
服务业目前于发达国家的产业比重约占多少？	70%以上	发达国家,服务业,比重	不包含	
姚明的父亲姚志源多高？	6 英尺 10 英寸（2.08 米）	姚志源,姚明,父亲	包含	姚明是独生子，父亲姚志源身高 6 英尺 10 英寸（2.08 米），母亲方凤娣身高 6 英尺 2 英寸（1.88 米）。
马尔代夫的第一大	旅游业	支柱产业,马尔代夫	包含	马尔代夫是个岛国，陆地面积相当狭窄。工业、农业水平低下。渔业、航运和旅游是三

支柱产业是什么？				大经济支柱。近年来，随着旅游业的发展，旅游收入已达到了国民总产值的 70%以上。成为马尔代夫的第一支柱产业。
中国历史上最杰出的浪漫主义诗人是？	李白	浪漫主义,杰出,诗人	包含	李白，字太白，号青莲居士，中国唐朝诗人，剑南道绵州昌隆县（今四川省江油）人，自言祖籍陇西郡|陇西成纪县|成纪（今甘肃省天水市秦安县），另說郭沫若考证唐代大诗人李白出生于吉尔吉斯斯坦碎叶河上的碎叶城,屬唐安西都護府(今楚河州托克马克市)。有“詩仙”、“詩俠”、“酒仙”、“谪仙人”等称呼，公认为是中国历史上最杰出的浪漫主义诗人。
甲午战争爆发的标志是？	丰岛海战	甲午战争,标志,爆发	不包含	
国际海洋法法庭的总部位于哪个国家？	德国	海洋法,法庭,总部	包含	国际海洋法法庭（InternationalTribunalfortheLawoftheSea，简称 ITLOS），是根据《联合国海洋法公约》建立的的一个法律组织。始建于 1996 年，总部位于德国汉堡市，是专门审理海洋法案件的国际组织。
《苏德互不侵犯条约》的签订地点是？	莫斯科	互不侵犯条约,苏德,签订	包含	《苏德互不侵犯条约》是 1939 年第二次世界大战爆发前苏联与纳粹德国在莫斯科所秘密签订之互不侵犯條約。
欧洲的第一所大学是？	博洛尼亚大学	第一所,欧洲,大学	包含	博洛尼亚大学是一所坐落在意大利艾米利亚-罗马涅大区首府博洛尼亚的综合性公立大学，是广泛公认的西方最古老的大学，建立于公元 1088 年神圣罗马帝国时期。
图书馆学最早由谁提出？	马丁·施莱廷格	图书馆学,最早,提出	包含	图书馆学最早由德国图书馆学家马丁·施莱廷格于 1807 年提出,是一门不断变化和发展的新兴学科，
约翰纳什的博士学位是在哪所大学获得的？	普林斯顿大学	纳什,博士学位,约翰	包含	关于纳什均衡的普遍意义和存在性定理的证明等奠定非合作博弈理论发展基础的重要成果，是约翰·纳什在普林斯顿大学攻读博士学位时完成的

现存所有老虎亚种中最小的亚种是?	苏门答腊虎	亚种,老虎,现存	包含	苏门答腊虎是现存所有老虎亚种中最小的亚种，雄性体重 100-150kg，雌性体重 75-100kg，指间有蹼。
第一位获得菲尔兹奖的女性是?	玛丽安·米尔札哈尼	菲尔,兹,第一位	包含	2014 年玛丽安·米尔札哈尼成为第一位获得菲尔兹奖的女性。
位于双子座和狮子座之间的星座是?	巨蟹座	狮子座,双子座,星座	包含	巨蟹座位于双子座和狮子座之间、北方是天猫座、南面则是小犬座和长蛇座，是一個暗淡細小的星座，没有亮于 3 等的恒星。
世界最大的流动性沙漠是?	塔克拉玛干沙漠	流动性,沙漠,最大	包含	塔克拉玛干沙漠位于中国新疆的塔里木盆地中央，是中国最大的沙漠，也是世界第二大沙漠，同时还是是世界最大的流动性沙漠。
人类合成的第一种抗菌药是?	磺胺	抗菌药,第一种,合成	包含	人类合成的第一种抗菌药是磺胺，1932～1933 年间德国病理与细菌学家格哈德·多馬克发现其具有体内抗菌活性，他因此获得 1939 年诺贝尔生理学或医学奖。
第一种以人工合成无机物质而得到的有机化合物是?	尿素	人工合成,无机,第一种	包含	尿素是第一种以维勒尿素合成无机物质而得到的有机化合物。
磁悬浮列车速度最高可达每小时多少公里?	550	磁悬浮列车,可达,小时	不包含	
JPEG2000 格式文件扩展名为?	.jp2、.j2c	JPEG2000,格式文件,扩展	不包含	?

4.3 答案抽取模块

答案抽取模块先对从文档检索模块检索出来的 10 个文档进行处理，抽取

出最可能包含答案的句子，一个文档可能拥有很多个句子，于是我们要对将近上百个句子进行打分，得分最高的，就认为这个句子最有可能包含我们的答案。之后从句子中抽取候选答案，对每个答案进行打分，得分最高的为最终的答案。

同样，以文档检索中使用的 100 个问题为例，部分问题可参见，表 2-4 问题答案部分样本。

2-6 部分答案抽取结果

问题	关键字	候选句子	是否包含标准答案	标准答案	系统答案
谁是右手中国式直拍弧圈结合两面快攻打法选手？	直拍,弧圈,快攻	王皓的身高是 1.75 米，體重 78 公斤,打法是右手直拍弧圈结合快攻，特点是用球拍的反面击球，亦即直拍横打	包含	王皓	78
黄牛一词来源于哪里？	黄牛,一词,来源于	黄牛一词来源于 20 世纪的上海，是指票贩子们聯群搶購票時常“有如黃牛群之騷動”,故将他们稱为黄牛或黄牛黨	包含	上海	上海
姚明的父亲姚志源多高？	姚志源,姚明,父亲	姚明是独生子|独生子女，父亲姚志源身高 6 英尺 10 英寸（2.08 米），母亲方凤娣身高 6 英尺 2 英寸（1.88 米）	包含	6 英尺 10 英寸（2.08 米）	姚明
欧洲的第一所大学是？	第一所,欧洲,大学	logo=,footnotes=,}}博洛尼亚大学（;; 尊称：大学之母；又譯-{;}-）是一所坐落在意大利艾米利亚-罗马涅大区首府博洛尼亚的综合性公立大学，是广泛公认的西方最古老的大学，建立于公元 1088 年神圣罗马帝国时期	包含	博洛尼亚大学	logo
约翰纳什的博士学位是在哪所大学获得的？	纳什,博士学位,约翰	关于纳什均衡的普遍意义和存在性定理的证明等奠定非合作博弈理论发展基础的重要成果，是约翰·纳什在普林斯顿大学攻读博士学位时完成的	包含	普林斯顿大学	关于
现存所有老虎亚种中最小的亚种是？	亚种,老虎,现存	苏门答腊虎是现存所有老虎亚种中最小的亚种，雄性体重 100-150kg，雌性体重 75-100kg，指间有蹼	包含	苏门答腊虎	苏门答腊虎

位于双子座和狮子座之间的星座是？	狮子座,双子座,星座	巨蟹座位于双子座和狮子座之间、北方是天猫座、南面则是小犬座和长蛇座，是一個暗淡細小的星座，没有亮于 3 等的恒星	包含	巨蟹座	巨蟹座
世界最大的流动性沙漠是？	流动性,沙漠,最大	塔克拉玛干沙漠（/Teklimakanqumluqi）位于中国新疆的塔里木盆地中央，是中国最大的沙漠，也是世界第二大沙漠，同时还是是世界最大的流动性沙漠	包含	塔克拉玛干沙漠	塔克拉玛干沙漠
人类合成的第一种抗菌药是？	抗菌药,第一种,合成	百浪多息（）是世界上第一种商品化的合成抗菌药（SyntheticAntibacterialAgent）和磺胺类抗菌药（Sulfonamideantibacterial），是由德国法本公司下属拜耳实验室的研究人员在 1932 年发现的	包含	磺胺	CAS
第一种以人工合成无机物质而得到的有机化合物是？	人工合成,无机,第一种	Section2=,Section3=,Section7=,Section8=,}}尿素()是由碳、氮、氧和氢组成的有机化合物，又稱脲（與尿同音）	包含	尿素	Section2
路由器处于 OSI 模型的哪一层？	OSI, 路由器,一层	网络层的作用是决定如何将发送方的数据传到接收方	包含	网络层	网络层
央视拥有多少个电视频道？	电视频道,央视,拥有	北京市复兴路 11 号邮编 1=100859,电话 1=,场所 2=新台址（央视总部）,地址 2=北京市光华路甲 1 号，邮编 2=100789, 官网=[http//cctvenchiridion.cctv.com/index.shtml 中国中央电视台],链接=,沿革链接=北京电视台,时期 1=1958 年 5 月 1 日,机构 1=中国中央电视台#历史|北京电视台,时期 2=1978 年 5 月 1 日，机构 2=中央电视台 ,image1=Beijingguangbodasha1959.jpg,caption1=广播大厦（老台址、现为国家新闻出版广电总局）,image2=ChromaticTelevisionCenter,CCTV,Beijing.jpg,caption2=彩电中心（旧台 址）,image3=Beijingskyscraperpic5croprotatelighten.jpg,caption3=总部大楼（新台址）,a=中央人民广播电台,b=中国国际广播电台,c=中国教育电视台,d=苏联中央电	不包含	45	北京市

		视台,e=朝鲜中央电视台,f=越南电视台,}}]]中国中央电视-{台}-（简称央视;，英文缩写为“”），是中华人民共和国的官方电视媒体之一（另有直属于中华人民共和国教育部的中国教育电视台）			
王宝强在哪一年首次参加中央电视台“春晚”？	王宝强,春晚,中央电视台	6岁时开始练习武术，8至14岁在嵩山少林寺做俗家弟子，法号“恒志”	不包含	2008	6
芭蕾起源于哪个国家？	芭蕾,起源于	芭蕾是一种轻盈，舒缓，优雅的舞蹈	不包含	意大利	芭蕾
2011年8月，莫言创作的什么长篇小说获得第8届茅盾文学奖？	茅盾文学奖,莫言,长篇小说	莫言，原名管谟业，山东高密县人，中国当代著名作家，1980年代中以乡土作品崛起，充满着“怀乡”以及“怨乡”的复杂情感，被归类为“寻根文学”作家	不包含	《蛙》	莫言
在历史上乌镇曾出过多少名举人？	乌镇,出过,在历史上	简称=,设立始年=,行政区类别=鎮,区划=,全镇面积=71.19平方公里,总人口=6萬（2007年）,本土语言=吳語,电话区号=0573,邮政编码=,经纬度=,毗邻乡镇=濮院镇石门镇龙翔街道练市镇（湖州南浔区）桃源鎮（江蘇），官方网站=http/www.txwzw.com.cn/www.txwzw.com.cn/,}}乌镇位于浙江省桐乡市，是江南著名古镇之一	不包含	161	简称
《本草纲目》撰成于哪一年？	撰成,本草纲目	《本草纲目》是一部集16世紀以前，中國本草學大成的著作	不包含	万历六年（1578年）	《
人的岁数等于12的倍数称为什么？	岁数,倍数,等于	生肖的一輪周期为12年	不包含	本命年	生肖
马尔代夫的第一大支柱产业是什么？	支柱产业,马尔代夫	马尔代夫是个岛国，陆地面积相当狭窄	不包含	旅游业	马尔代夫
平均人口密度最高的大洲	大洲,平均,最高	（冰覆蓋區域:；無冰區域:）population=非永久居民-約1,000	不包含	欧洲	（

是？		人 ,density=,demonym=,countries=,list_countries=,dependencies=,unrecognized=,languages=,time= 無 ， 唯 葛 拉 汉 地（GrahamLand）屬於 UTC-3			
中国历史上最杰出的浪漫主义诗人是？	浪漫主义,杰出,诗人	从 1789 年法国大革命的爆发，直到 1815 年拿破仑政权的最终颠覆，法国人都专注於对外界事物的观察	不包含	李白	从
最后宣布废除奴隶制度的行政地区是？	奴隶制度,废除,最后	法蘭西第二共和國，简称第二共和，是 1848 年 11 月 4 日到 1852 年 12 月 2 日间统治法国的共和政体	不包含	毛里塔尼亚	法蘭西
国际海洋法法庭的总部位于哪个国家？	海洋法,法庭,总部	国际海洋法法庭（InternationalTribunalfortheLawoftheSea，简称 ITLOS），是根据《联合国海洋法公约》建立的的一个法律组织	不包含	德国	国际
《苏德互不侵犯条约》的签订地点是？	互不侵犯条约,苏德,签订	1939 年 8 月 23 日，苏联与德国签订苏德互不侵犯条约|互不侵犯条约，同年 9 月，德国国防军|德军闪电战|闪击波兰，苏联根据苏德之间的秘密协定出并占领了波兰的西乌克兰和西白俄罗斯	不包含	莫斯科	1939
图书馆学最早由谁提出？	图书馆学,最早,提出	整理说是最早的关于图书馆学研究对象的学说	不包含	马丁·施莱廷格	整理
首先分离出元素氟的化学家是？	氟,化学家,分离	因为氟的化合物很稳定，所以氟单质很难分离	不包含	亨利·莫瓦桑	因为
第一位获得菲尔兹奖的女性是？	菲尔,兹,第一位	1954 年讓-皮埃爾·塞爾獲菲爾茲獎時年僅 27 歲，他至今仍是最年輕的得主	不包含	玛丽安·米尔札哈尼	1954
太阳系中唯一一颗没有磁场的行星是？	磁场,太阳系,行星	阿尔文学说是瑞典物理学家阿尔文提出的一种关于太阳系起源的学说	不包含	金星	阿尔文
马达加斯加位于哪个大洲？	马达加斯加,大洲	一些生态学家因这种独特的生态系统而将马达加斯加称为“第八大洲”，保护国际也将该岛列为生物多样性热点地区	不包含	非洲	马达加斯加
经典力学的基础是？	经典力学,基础	它是工程和日常生活中最常用的表述方式，但并不是唯一的表述方式：約瑟夫·拉格朗	不包含	牛顿运动定律	瑟夫

		日、威廉·哈密顿、卡爾·雅可比等发展了经典力学的新的表述形式，即所谓分析力学			
转基因植物常用哪些方法？	转基因, 常用, 方法	}}转基因就是将人工分离和修饰过的基因导入到目的生物体的基因组中，从而达到改造生物的目的	不包含	花粉管通道法、农杆菌介导转化法、基因枪介导转化法、细胞融合法	}
被道教奉为太上老君的人物姓什么？	太上老君, 道教, 奉	道教，是发源于古代本土中国春秋战国的方仙道，是一個崇拜諸多神明的多神教原生的宗教形式，主要宗旨是追求長生不死、得道成仙、濟世救人	不包含	李	道教
柴油和酒精的沸点哪个高？	沸点, 酒精, 柴油	不同组分液体具有不同的沸点，是蒸馏最基本的理论依据	不包含	柴油	不同

100 个问题中，有 33 个问题，在文档检索系统的工作下成功的把含有答案的文档检索出来，候选句选择程序，从 33 个包含正确答案的文档列表中，成功抽取出 11 个包含正确答案的候选句子，成功率为 33.3%（11/33 = 33.3%）。最终的答案抽取程序，从候选句中抽取出最终的正确答案，成功抽取出的结果，有 5 个与标准答案相符合，抽取的成功率为，45.5%（5/11 = 45.5%）。

5 总结与展望

5.1 总结

总的来说，这个问答系统还是很粗糙，每个模块都有改善的空间。问题分类模块中，采用的是简单的语言词袋模型，忽略了词语与词语之间的顺序信息，后续可以采用另外的模型可以提高问题分类的成功率，另一方面人工标注的问题类别数量也会影响最后模型的精确度，后续可以增加人工标注问题的数量。

5.2 展望

整个系统的分词和词性标注工具,用的是开源的python模块 jieba 分词，但是在实验中发现，在答问系统中一般的分词工具的分词效果并不是很理想，问答系统会有自己的偏好，比如说，“2014 年 10 月 10 号”，jieba 分词会把这个句子，切分成，“2014”，“年”，“10”，“月”，“10”，“号”，而对于，问答系统，它需要的单词粒度不需要这么细致，它更喜欢，这么分词，“2014 年”，“10 月”，“10 号”，如果还要进行优化的话，自行构建分词系统是一个方向。

有利用到知网，对关键字进行扩展，比如说问题中有关系字，“中国”，没有进行关键字扩展的话，无法把包含，“中华人民共和国”的文档检索出来，但其实，“中国”和“中华人民共和国”的含义是一样的，我们可以利用现有的词网数据，或者现有的词典，搜索引擎等等工具对一些词语进行扩展，提高对文档的理解度，提高抽取正确答案的能力。

维基百科中，网页与网页之间的链接没有被使用，或者这方面的数据可能帮助我们对文档的重要程序进行辅助地评价。

分词过程，我们并没有未登录词语，因此可以在分词方面最额外的优化工作。

现成的语料库比较少，比如问题类别标注数据。

参考文献

[1] Sproat, R. and Emerson, T. The First International Chinese Word Segmentation Bakeoff[A]. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing[C]. Sapporo , Japan: July 11-12 , 2003 ,133-143.

[2] 黄昌宁,赵海. 中文分词十年回顾[J]. 中文信息学报,2007,03:8-19.

[3] 结巴中文分词[EB/OL]https://github.com/fxsjy/jieba,2015.05.09/2015.05.09

[4] The Integration of Lexical Knowledge and External Resources for Question Answering

[5] Web Data Mining 2014 Fall - PKU » 互联网数据挖掘 [EB/OL] http://www.icst.pku.edu.cn/lcwm/course/WebDataMining2014/

[6] 黄翼彪. 开源中文分词器的比较研究[D].郑州大学,2013.

[7] Help:Formatting - MediaWiki[EB/OL] http://www.mediawiki.org/wiki/Help:Formatting,2015.04.12/2015.05.03

[8] Manual:Database layout[EB/OL] http://www.mediawiki.org/wiki/Manual:Database_layout, 2015.03.12/2015.05.09

[9] Data dumps/xml2sql[EB/OL] http://meta.wikimedia.org/wiki/Data_dumps/xml2sql, 2013.3.5/2015.05.09

[10] Robertson, Stephen, and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. [M] Now Publishers Inc, 2009

[11] Singhal, Amit, Chris Buckley, and Mandar Mitra. "Pivoted document length normalization." [C] Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1996.

[12] 维基百科:数据库下载[EB/OL]http://zh.wikipedia.org/wiki/Wikipedia:数据库下载, 2014.10.15/2015.05.09

[13] 下载页面[EB/OL] http://download.wikipedia.com/zhwiki/ /20141009/zhwiki-20141009-pages-articles-multistream.xml.bz2,2015.05.09/2015.05.09

[14] 张宇,刘挺,文勖. 基于改进贝叶斯模型的问题分类[J]. 中文信息学报,2005,02:100-105.