一个标准中文问答系统的研究与实现

Research and Implement on A Prototype Chinese Question-Answering System

李 季 （辽宁 鞍山师范学院 计算中心 114001）

摘要：本文介绍了有关中文问答系统的一些研究。问答系统是集知识表示、信息检索、自然语言处理技术于一体的难度很高的研究课题。本文从问题处理、检索系统、答案抽取三个方面进行了详细说明，并提出了具体的实现方案。

关键词：问答系统 关键词 信息检索 自然语言处理 答案抽取 TREC(文本检索会议)

1 引言

问答系统是根据用户的问题从大量文本集合中找出确切的答案。换句话来说，这个答案不是跟问题有关的整个文献，而是更精确地满足用户需求的文献的一部分或一句话。

然而，目前的信息检索系统只能为我们定位出相关文献并把它们按相似性排序，却把从这些文献中抽取确切信息的任务留给了我们自己。近一段时间以来，互联网的广泛使用使文献的数目呈指数增加，我们迫切需要一个能检索有用信息而不是整篇文献的系统即问答系统。

自然语言问答系统作为自然语言处理的一个应用领域的研究，现在还处于起步阶段。自从文本检索会议(TREC)在1999年的TREC-8会议上引入了对问答系统的评测后，人们对基于自然语言的问答系统产生了浓厚的兴趣，国外很多研究机构和公司对自然语言问答系统的研究已经取得了一些成果，并且一些机构陆续参加了TREC（文本检索会议）的评测，我国一些科研院所（如复旦大学、清华大学、微软中国研究院）参加了2000年第九次以及后续的文本检索会议，并在一些项目中（如CLIR（跨语言检索）、Filtering（文本过滤）等）取得了较好的成绩，但是由于问答系统是集自然语言处理、知识表示、信息检索于一体，它的发展将大大取决于这些领域的进步，所以目前对它的研究还非常有限。

相对来说，国内对中文问答系统的研究更是少之又少，因为中文问答系统比自然语言问答系统对一些有关领域的研究要求更高，比如中文词语之间没有空格，因此在操作之前需要进行词语切分，与自然语言相比，汉语句法分析和语义理解更为困难，这些都造成了中文问答系统的发展缓慢。我们研究并开发了此中文问答系统，此系统在结合相关文献及实现过程中实际情况确定此系统结构包括三个部分：问题处理部分、检索系统部分、答案抽取部分。以下详细介绍了此系统。

2 系统概述

系统主要有以下三个部分组成：问题处理部分，基于句法分析语义映射；检索系统部分，基于信息检索技术；答案抽取部分，利用问题和文档段落的相似度确定正确的答案。系统的结构如下图所示：

3 研究与实现

3.1 问题处理

在问题处理阶段，其主要工作是系统把用户输入的问题转化为一组关键词以备在检索系统中使用及系统由问题类型确定答案类型以备在抽取答案时使用。问题处理部分经过细化需要完成以下几部分工作：对问题进行分词和词性标注、确定问题的类型（Question Types）、答案类型（Answer Types）识别、确定问题的焦点（Question Focus）、提取出问题的关键词（Keywords）、对关键词进行适当的扩展。

3.1.1 问题-答案语义映射表

答案类型识别它是系统的重点和难点，

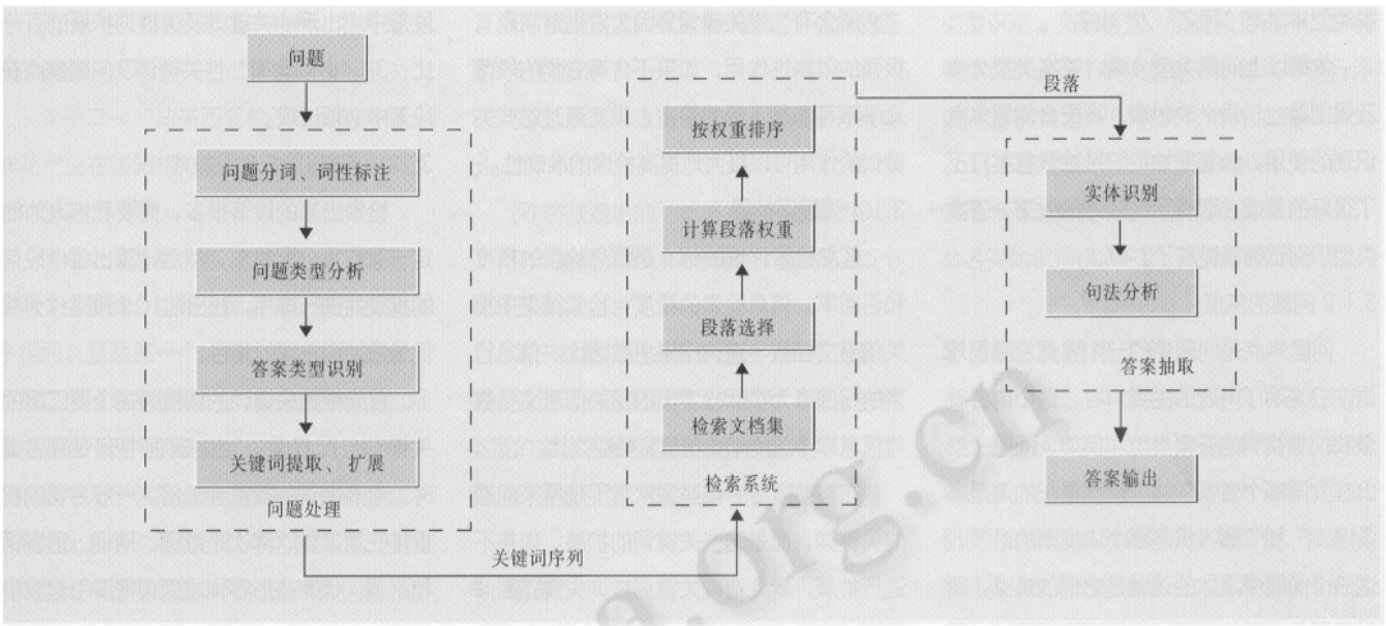

图 1　系统结构图

它对答案抽取起着至关重要的作用。答案类型识别主要依据答案类型分类和句法分析模式匹配规则。不同的问题类型确定不同的答案类型。答案类型每一个分类它都有自己的一套规则，一般来说，一套规则可以查出特殊关键词的存在，专有名词的存在及属于哪种给出的语义分类。

传统的问题分类一般分为以下几种，如：询问人（谁）、询问时间（什么时间）、询问地点和位置（哪里）、询问数量（多少）、询问原因（为什么）、询问定义（什么是）、其他等。使用这种问题分类方法来确定答案类型比较笼统，答案抽取的准确率和召回率都不理想，针对这种情况系统建立了问题-答案语义映射表，在其中包括了几十种答案类型分类及规则。如其部分分类和规则如下：

(1) 询问地点。在用户提出的问题中有“什么地方/哪/哪里/何地…”疑问词，答案的类型肯定和“地点”和“位置”有关，如国家、省、市、区等。

(2) 询问时间。在用户提出的问题中有“什么时间/什么时候/何时/哪年…”疑问词，答案的类型肯定和“时间”或“日期”有关，如：年、月、日、星期等。

(3) 询问人。在以往的系统中，在用户提出的问题中有“谁”疑问词，你会说答案的类型肯定是单个“人”，如“谁发现了北美洲？”，系统便去寻找单个人的实体，这样的答案分类有些笼统，造成了答案类型识别错误率比较高.在问题-答案语义映射表中系统把以下两种情况都考虑进去。

第一种情况：问题“谁是伽俐略？”就是一个“谁”问题的特例，它要寻找的答案类型不是人而是描述这个关键词“伽俐略”的一段话。针对这种情况，系统定制了一套规则：

（谁）是（专有名词（人名））？/（（专有名词（人名））是（谁）？

这样问题的答案类型系统把它定为描述类型（DESCRIPTION）.在答案抽取中应抽取“Answer=(专有名词（人名）+是+(一段描述))”。该规则可以解释为问题的答案是由(专有名词（人名）)加“是”加后面的“一段描述”所构成。

第二种情况：“谁”问题对应的答案类型也可以是一组人。这“一组人”可以是一个专有名词的实体：“谁是最可爱的人？（人民志愿军）”，也可以是一组人的列表：“在100米短跑中谁击败了李平？（孙力、魏征和赵军）”或是两个专有名词实体的联合：“在接力比赛中谁击败了英国？（美国和加拿大）”。（此种情况在体育比赛中用的比较多）

(4) 询问定义。在用户提出的问题中有“是什么”疑问词，它属于“询问定义”问题类型，定制的规则如下：“Answer=behind(Noun+是/称为/定义为)或front(是/称为/叫做/定义为+Noun)”这里的Noun指的是被询问定义的名词，因此该规则就可以解释为“问题的答案是在由该名词加“是”、“称为”或“定义为”等所构成的串的后面部分，或者在由“是”、“称为”、“叫做”或“定义为”等加被询问名词所构成的串的前面部分。

(5) “哪”疑问词。“哪”疑问词和英语中的“which”很相似，单凭这个词不能确定答案的类型，需要确定问题的焦点（详见3.1.2）。如”哪/哪里…”答案类型肯定是“地点”或“位置”；“哪个人”答案类型肯定是“人”；“哪年/哪月/哪日/哪时”答案类型肯定是“时间”；“哪条河/哪座山”答

案类型肯定是“河名”/“山名”。

依据以上问题类型分类、答案类型分类及规则建立了语义映射表，再配合问题焦点识别的使用，为答案抽取阶段抽取答案打下了很好的基础，取得了比较好的效果，答案类型识别正确率提高了3.4%。

3.1.2 问题的焦点

问题焦点是问题中的一个名词或名词短语，它说明了问题的主要内容，这个内容就是我们想找到的答案类型。例如，问题“泰山在我国哪个省？”、“世界最长的河是那条河？”和“蒸汽机是哪个人发明的？”，这三个问题中黑体的词就是它们的焦点，通过问题焦点系统就可以很明确知道第一、第二、第三个问题的答案类型分别是“省名”、“河名”和“人名”。问题焦点对找到答案非常重要，系统通过一定的规则找到问题焦点，对问题焦点进行分析，最终确定出问题的答案类型。对于本系统来说，一个通用的规则就是在问题中疑问词后面的第一个名词或名词短语成为焦点的优先率最高。

3.1.3 关键词提取

问题关键词的提取影响到后面的检索效果，一般来说关键词主要有名词、动词、形容词、限定性副词等组成。但在实际应用中可以把问题中除了疑问词以外的大部分词都作为关键词以提高检索的精度。

关键词可以分为两种：一般性关键词、“必须含有”性关键词[1]。所谓“必须含有”性关键词指的是这些关键词必须在答案段落中含有。而一般性关键词可以不被答案段落所包含。

关键词被赋予不同的权重，在检索段落时这些权重用来计算段落的权重。通常问题焦点、“必须含有”性关键词、具有限定性作用的副词会有比较高的权重。“必须含有”性关键词主要由专有名词、限定性副词（如：最大、最高、最远、最快、第一等）、时间、数词等组成。之所以要制定“必须含有”性关键词是因为它们对问题有极强的限制性作用，如果不含有它们的段落几乎不可能是正确的答案，因此通过这些关键词的作用可以极大地提高检索的准确性。

3.1.4 关键词扩展

信息检索评价的标准是信息检索的精度和召回率。信息检索的精度为检索结果中相关信息文档数与查询结果总数之比。信息检索的召回率为实际检索出的相关信息文档数与信息库中总的相关信息文档数之比。

在答案段落中某些词常常不是原来问题的关键词，而是这些关键词的扩展。如果不进行扩展，就会造成关键词查询失败，因此需要对关键词进行适当的扩展，以提高系统的召回率。系统采用了以下方式对关键词进行扩展：1. 名词同义词扩展和语义蕴涵扩展。2.动词同义词扩展（意义用法相同，这样出现歧义的可能性很小）。3. 根据问题类型扩展。对于根据问题类型扩展就需要根据问题的类型制定不同的扩展规则，如询问数量类型可以把一些表示数量的单位加入。对于这种关键词的扩展不仅能提高答案的召回率，而且还可以提高正确率。

3.2 检索系统

在此部分中系统实际用了两个时期的检索：文献检索和段落选择。

3.2.1 文献检索

在文献检索中应用的主要技术是信息检索技术。信息检索技术可以根据问题处理部分产生的关键词序列进行查询递交出符合查询关键字的所有文献（取前100篇文献），而且可以根据文献的相关性对它们进行排列。

3.2.2 段落选择

在段落选择时期，系统从文献检索时期已检索的文献中选择最相关的100个段落。在候选段落选择中注意以下三个方面：1）被抽取的段落的长度；2）根据TREC问答评测，每个答案的长度不超过125个汉字。3）在候选段落中应出现的关键词及关键词扩展的百分比；3）“必须含有”性关键词及问题焦点在段落中必须出现。

3.2.3 计算段落权重及排序

检索出来的段落很多，需要我们人为制定一些权重分配机制，然后计算出每个段落的权重并进行排序，找出前10个段落作为我们最后抽取答案的集合。

首先根据关键词的词性为每个要匹配的关键词分配权重，当关键词中有专有名词时，专有名词的权重会加倍。一般名词的权重排在第二位，其次是数词、动词、形容词和副词。但有些形容词或副词需要引起我们特殊的关注，如“第一”,“最好”,“最长”,“最高”等形容词或副词应赋予加倍的权重。另外，问题焦点和“必须含有”性关键词也应赋予加倍的权重。

其次，还要根据类似IDF（Inverse Document Frequency）的权重公式为每个关键词分配权重，每个关键词都具有类似IDF形式的权重：

IDFi = log（N / fi）[2]

其中，fi为关键词i在段落中被匹配的次数，N是段落的的总长度。

那么计算段落权重的公式：

W = ∑KWi ®w IDFi （其中i=1~m）

其中：KWi是在问题处理阶段给出的第i个关键词的权重，IDFi 是该关键词在段落中的idf值。

一个段落的权重除了和关键词的存在有关外，在很大程度上还和关键词在段落中的顺序、距离有关。尤其对于使用以上公式计算权重相同或相似的段落应采用另外一种权重计算方法重新计算进行调整。此权重计算方法主要考虑关键词密度分布情况（即考虑关键词在段落中的顺序、距离情况）。一般的，关键词分布越集中，段落的权重越大。例如问“中国的国花是什么？”对于这个问题系统提取的关键词为“中国”、“国花”，检索的结果有这样两个句子：

句子一："洛阳市盛产牡丹，牡丹是中国的国花。…"

句子二："马来西亚位于中国的南部，鲜花产业在该国相对发达，扶桑花是马来西亚的国花。…"

这两个段落都含有"中国"、"国花"这两个关键词，应用上面公式计算的权重几乎相同，显然第一个段落含有问题的答案，而第二个段落是错误的。造成这种错误的原因主要在于两个关键词"中国"、"国花"在问题中几乎邻接出现，而且只有邻接时才能表达出正确的意思。针对这种情况，应按关键词分布密度权重计算公式计算权重，公式如下：

$W = 1 / (\sum(Kposi - Kposi\text{-}1))$ (其中 i=2~m)

其中：Kposi为第i个关键词在段落中的位置。

3.3 答案抽取

找到10个候选段落后，我们就需要从这些候选段落中检索出可能包含答案的5个段落，即候选答案。

在答案抽取部分为了找到答案系统主要依据关键词、命名实体识别、句法模式匹配。

答案抽取主要按以下步骤进行抽取答案：

(1) 确定从问题处理部分得到的答案类型；

(2) 对候选段落进行实体名识别，从中得到人名、地名、日期和时间等实体名；

(3) 在候选段落中确定实体名是否和答案类型分类相匹配；

(4) 扫描候选段落找出与答案类型匹配的候选答案；

(5) 扫描候选段落找出与问题焦点匹配的候选答案；

(6) 给每个可能的答案分配一个初始的分数；

(7) 根据指定的权重规则将这个可能的答案的分数加分；

(8) 从段落中选出分数最大的可能答案输出；

(9) 将被选出的可能答案的分数置为0；

(10) 重复步骤8）和步骤9）直到选出5个候选答案。

其中权重规则主要依据以下几个因素：每个关键词的匹配程度之加权和（符合答案类型的实体名和问题焦点的权重较大）；关键词的覆盖程度；句法结构的匹配程度。

4 结果

在一个小的问题集内我们对4个关联的独立模块进行了手工分析。这4个模块是：答案类型识别，文献检索，候选段落选择，命名实体识别。每个模块根据它的错误率进行评估，答案类型识别模块它的评估错误率为26%，文献检索模块它的评估错误率为54%，候选段落选择模块它的评估错误率为44%，命名实体识别模块它的评估错误率为61%。

5 结论及将来工作

如上所述，本文介绍了中文问答系统的一些重要技术，并给出了一些具体有效的实现方案。在以后的具体实现中对答案类型识别使用的语义映射表还需完善细化，候选段落选择中使用的权重算法、答案抽取中的命名实体识别还需进一步提高。

参考文献

1 Bernardo Magnini, Matteo Negri, Roberto Prevete and Hristo Tanev. Multilingual Question/Answering: the Diogene System. Proceedings of the tenth Text Retrieval Conference(TREC-10), Gaithersburg Maryland: Dept of Commerce, National Institute of Standards and Technology, 2001,313-321.

2 C. L. A. Clarke, G. V. Cormack, D. I. E. Kisman, and T. R. Lynam. Question Answering by Passage Selection. Proceedings of the ninth Text Retrieval Conference (TREC-9), Gaithersburg Maryland: Dept of Commerce, National Institute of Standards and Technology, 2000,673-678.

3 R. J. Cooper and S. M. R ü ger. A simple Question Answering System. Proceedings of the ninth Text Retrieval Conference(TREC-9), Gaithersburg Maryland: Dept of Commerce, National Institute of Standards and Technology, 2000,249-255.

4 Enrique Alfonseca, Marco De Boni, Suresh Manandhar. A prototype Question Answering system using syntactic and semantic information for answer retrieval. Proceedings of the tenth Text Retrieval Conference(TREC-10), Gaithersburg Maryland: Dept of Commerce, National Institute of Standards and Technology,2001, 680-685.

5 Jianping Chen, Anne R.Diekema, Mary D.Taffet, Nancy McCracken, Necati Ercan Ozgencil, Ozgur Yilmazel, and Elizabeth D.Liddy. Question Answering: CNLP at the TREC-10 Question Answering Track. Proceedings of the Tenth Text Retrieval Conference (TREC-10), Gaithersburg Maryland: Dept of Commerce, National Institute of Standards and Technology, 2001,485-494.