# 中文问答系统中基于主题和焦点的问题理解①

陈永平[1]，杨思春[2]，毛万胜[1]，苏 新[1]，刘 俞[1]

[1](马鞍山职业技术学院 计算机系，马鞍山 243000)
[2](安徽工业大学 计算机学院，马鞍山 243002)

**摘 要**：问答系统应该能够用准确、简洁的语言回答用户提出的问题。问题理解是问答系统的首要的分析工作，分析的结果的正确率直接影响后续处理。提出了一种基于用户问题的主题和焦点的问题理解的方法，该方法以问题的疑问词、句法分析、问题焦点、问题主题在知网中的首义原作为问题分类特征。实验结果证明，该方法对提高分类的准确率有较好的效果。

**关键词**：问答系统，问题理解，问题分类，焦点，主题

## Question Interpretation Baesd on Theme and Focus in Chinese Question Answering System

CHEN Yong-Ping[1], YANG Si-Chun[2], MAO Wan-Sheng[1], SU Xin[1], LIU Yu[1]

[1](Department of Computer, Ma'anshan Vocational Technology College, Ma'anshan 243000, China)
[2](Computer College, Anhui University of Technology, Ma'anshan 243002, China)

**Abstract:** Question answering system can answer the users' questionsin the application of the precise and concise language. Question interpretation is the primary work of question answering. The precision of Question analysis has a great effect on the following processing work. In this paper, we present a new method of question interpretation based on the users' question theme and focus. This method by which the interrogative words, syntax analysis, question focus words and the first semenes of the question topic in HowNet have used as classification feature. Experiment result indicate that this method can bring about a better effect on improving the accuracy of the classification.

**Keywords:** question answering; question interpretation; question classification; focus; theme

## 1 引言

随着计算机网络技术的发展，各种信息愈来愈多地通过互联网为人们所利用，人们经常借助搜索引擎进行搜索。然而，目前的搜索引擎存在不少的弊端，传统的搜索引擎已不能满足人们的需求，自动问答系统正是在这种情况下提出的。自动问答系统是当今信息检索领域的一大热门话题。它的输入是用自然语言描述的句子，返回的结果是用户所需的直接答案。例如，如果我们想知道谁是美国现任总统，在现在的搜索引擎中，我们首先输人关键词“美国总统”，然后再在返回的文档中，查找究竟是谁是现任的美国总统。而在自动问答系统中，我们输入的是“谁是现任的美国总统?”，系统返回的便是其精确答案。可见，自动问答系统能更好的满足用户的需求，能更快的找出用户所需的信息。可以说自动问答系统是未来的新一代搜索引擎。

自动问答系统主要包含三个模块[1]:问题理解，文档检索和答案抽取。其中问题的理解是问答系统首先进行的分析工作，这个阶段分析的结果对后续阶段的处理有很大的影响。目前，问答系统回答问题的准确率都不是很高，其中一个很重要的因素就是问答系统“理解”用户提问不准确。在中文问答系统中，问题理解首先要对问题进行词法分析，然后根据问题句所询问的内容，确定问题的类型、提取出问题的关键词、

① 基金项目:安徽省教育厅自然科学基金(KJ2010B223)
收稿时间:2010-09-23;收到修改稿时间:2010-11-08

依据问题的类型等因素对问题进行适当的扩展。本文首先对用户问句的预处理及进行词法分析和句法分析，找出用户问句的主题和焦点，然后根据用户问句的主题在知网中的义原和焦点完成对用户问句的分类。

## 2 知网

知网（HowNet）是一个以汉英双语来表示概念与概念之间以及概念的属性之间关系的知识库[2],它是一个网状的知识系统。知网将客观世界中的词汇所代表的概念分为四大类：实体、事件、属性、属性值，并通过义原来标注概念。在知网中，义原是最基本的、不易于再分割的意义的最小单位，在知网中每一个概念是通过一组义原来表示的。义原间存在 8 种关系：上下位关系、同义关系、反义关系、对义关系、属性—宿主关系、部件—整体关系、材料—成品关系、事件—角色关系。这些义原以上下位关系为主干，形成树状结构分别存放于相应的义类文件中。在知网中对于概念的定义采用知识描述语（Knowledge Database Mark –up language,KDML） 来描述。KDML 对概念的定义采用 DEF 语义表达式，DEF 描述了词语详尽的语义特征，如：生日：DEF={time|时间：Timesect={day|日},{ComeToWorld|问世：time={~}}}。词语在知网中的首义原是指该词语在 DEF 定义中出现的第一个义原，例如，“生日”的首义原就是“time|时间“。它能较好地表达出该词语所对应概念的主要语义信息[3]。

## 3 用户问句的预处理

关于用户问句的预处理，主要的操作就是对用户问句进行词法分析。 “词是最小的能够独立活动的有意义的语言成分”[4]，但汉语的语素和单字词，合成词和短语之间没有清晰的界限。它是以字为基本的书写单位，因此，中文词语分析是用户问句处理的基础与关键。而中文词语分析一般包括以下几个过程:词语粗切分，切分排歧与未登录词识别、词性标注、句法分析[5]。

目前中文词语分析采取的主要步骤是[6]:先采取最大匹配、最短路径、概率统计方法、全切分等方法，得到一个相对最好的粗分结果，然后进行歧义消解，关于歧义消解，从构成形式上看，歧义有两种[7]：交集型歧义，组合型歧义。设 A、B、C 为汉字字符串，交集型歧义是指在汉字字符串 ABC 中，AB 和 BC 都可以组成一个词；组合型歧义是指切分 AB 和 A / B 都能分作词。单从分词结果上看歧义字段还可以分为以下两类；第一类是确定分法的歧义切分字段：第二类是不确定分法的歧义切分字段，即在不同的语境种，会出现不同的切分结果。比如：字段：“集中”在不同的语境种会有不同的切分结果。在句子“全校师生集中在学校操场上开会”中“集中”是一个词；但是在句子“自然数集中的所有质数构成的集合”中“集中”就不是一个词。歧义消解完成进行未登录词识别，未登录词是指那些在词典中没有收录的但又确实是词的词，又称新词。如人名、地名、机构名、产品名、简称、省略语等。目前对未登录词处理的方法是采用预处理策略，对词先进行分析处理，经分析是人名、地名还有机构名，对他们进行适当地标记处理。

词性标注的任务就是根据一个词在某个特定句子中的上下文，为这个词标注正确的词性。其实质是研究词语所表现的语法功能的聚合关系，它要解决的主要问题是词性歧义(词性兼类)和未登录词词性的确定. 词性自动标注是自然语言处理中的一个基本问题，因为在汉语中广泛存在着一个词语的词性多于一个的歧义现象。目前词性标注方法主要有三种[6]：基于规则的方法、基于统计的方法、混合方法。在上述三种方法中，基于统计的词性标注方法用的最为普遍，是目前词性标注的主流方法。

本文没有设计自己的词法分析系统 ，而是采用了哈尔滨工业大学的词法分析系统。

对用户问句进行句法分析是指在给定的文法下来分析自然语言的层次结构，它是自然语言处理中的中心问题之一，它是在词法分析的基础上进行的。本文利用哈尔滨式业大学信息检索研究室提供的汉语句法分析器，对用户问句进行句法分析，找出用户问句的主语和宾语。如“谁是小偷”的句法分析结果如下：


其中“谁”是主语，“小偷”是宾语。本文根据需要将利用句法分析所得问句的主语或宾语作为问句的主题。

## 4 问句的主题和焦点的确定

### 4.1 疑问词和疑问词短语

在汉语疑问句中包含疑惑和询问两种意义，通常既有疑惑也有询问，但也可以无疑而问，如反问句、

设问句。提问的手段，有语调、疑问词、语气词或疑问格式等，有时只用一种手段，有时兼用两三种，但其中语调是不可少的。疑问句可以根据上述表示提问的手段特点，一般分成四类：是非问句、特指问句、选择问句和正反问句。判断一个句子是不是疑问句要涉及到语调，语气助词，疑问词等[8]。其中最重要的是疑问词。下表列出了常见的的疑问词类型表：

表 1　疑问词类型表

| 类型 | 疑问词及疑问词短语 | 例句 |
|---|---|---|
| 人物 | 什么人 /谁/哪个人/何人/哪些人 | 谁发明了电灯 |
| 地 点 | 什么地方/什么地点/哪里/哪儿/何处 | 黄山在哪里？ |
| 时间 | 什么时间/什么时候/何时/哪个时候/何时 多长时间/多 少时间 | 中华人民共和国是什么时候成立的？ |
| 数量 | 多少/几 | 中国有多少人？ |
| 原因 | 什么原因/哪些原因/什么因素/哪些因素/为什么 | 为什么会发生大地震？ |
| 方式 | 哪些方法/哪些方式/哪些算法/哪些途径<br>什么方法/什么方式/什么算法/什么途径<br>怎样/怎么/怎么样/如何 | 什么方法可以准确地预测将要发生大地震？ |
| 其他 | —— | —— |

有时我们将一些询问目的明确的词语与疑问词合并，称为疑问词短语。在本文中我们将一个问句中的的疑问词和紧跟疑问词后没有别的修饰词的名词一起作为疑问词短语[8]。

比如：哪个人提出了人工智能？

在这个问句中将“哪个人”作为疑问词短语。

### 4.2 问句的主题和焦点的确定

疑问句中的焦点是指由问题找到的相关性质和实体，它往往就是问句的主要内容，它能比较准确地反映该问句涉及的知识范围[9]。抽取提问焦点，有助于确定问句的类型和答案的内容范围，综合其他信息为问题检索出相对准确的答案。那么问题焦点是怎么确定的呢？通过对大量的疑问句的分析，可以得出疑问句中的问题焦点就是该疑问句中的疑问词或疑问词短语构成的。当疑问句中的疑问词后是非名词或没有别的词时，则该疑问句的问题焦点就是由疑问词构成的；若紧跟疑问词后是名词，则该疑问句的问题焦点就是疑问词和疑问词后的名词构成的疑问词短语构成的。比如：

2009 年诺贝尔和平奖的奖金是多少？

问题焦点：多少

中国第一次卫星发射发生在什么时候？

问题焦点：什么时候

疑问句中的主题是指问题的对象或者事件，它是一个问句的概念主体[9]。它的作用是能够完整、深刻地揭示问题的内容和中心。我们可以从问句论述的主题直接检索所需资源，主题可以根据所研究问题的内容直接查找，凡是和所研究问题的主题内容相关资源都会被集中起来，反馈给用户，而且这一查找十分方便快捷。一般疑问句中的主题可根据以下方法确定：

(1) 在疑问词位于句子末尾的问句中，主题是由问句的主语构成的。如：

2009 年诺贝尔和平奖的奖金是多少？

问题主题：奖金

中国人第一次太空慢步发生在什么时候？

问题主题：太空慢步

(2) 而对于“谁是……？”、“什么是……？”、“哪个是……？”等疑问词后为动词“是、为”等问句中主题是由该问句中的宾语构成的。如：

什么是人工智能？

问题主题：人工智能

谁是现任美国总统？

问题主题：现任美国总统。

## 5　问题分类

对不同类型的问题，往往有不同的处理方法，所以不论是英文自动问答系统还是中文自动问答系统一般都有问题分类。在英语中问题分类相对简单，而对于汉语，问句的提问方式灵活多变，这就使得汉语中的问题分类要复杂得多。在中文自动问答系统中，常用的问题分类主要依靠的是句子中的疑问词，但各个疑问词对问题的辨别能力是不同的。例如，如果问句中出现疑问词“哪里”“哪儿”，就可以很容易的判断出问题类型为“询问地点”；如果问句中含有“谁”，就可以判断出问题类型为“询问人”。为了叙述方便，将这样的疑问词为专有疑问词。但如果问句中出现“什么”“哪”“哪个”“哪些”等疑问词，因为很多问题类型中都可能出现这些疑问词，称这些疑问词为通用疑问词。由通用疑问词构成的疑问句，如果依靠疑问词就不能判断出问题类型[10]。例如对于问句:人工智能是谁提出的？不同的人有不同的提问方式：

1）人工智能是谁提出的?

2）什么人提出了人工智能?

3）哪个人提出了人工智能?

4）人工智能是何人提出的?

显然，上述 4 个问句问的是相同的问题，但如果采用常用的依靠疑问词的问题分类的方法，它们将被划分为不同的类型，返回的答案也可能因为采用不同的搜索策略而不同。而本文提出了依据问句的焦点和主题对问题进行分类就不会发生上述问题分类的错误。比如在上述问句中，问句 1 的焦点是“谁”，主题是“人工智能”

问句 2 的焦点是“什么人”，主题是“人工智能”

问句 3 的焦点是“哪个人”，主题是“人工智能”

问句 4 的焦点是“何人”，主题是“人工智能”

我们可以依据这四个问句的焦点就可直接判断出它们是同一类型的问句。有一些问句仅依靠问句的焦点还不能确定问句的类型，要将问句的焦点和主题相结合才能判断出问句的类型。比如：林肯、奥巴马、里根这三人中哪一个是美国现任领导？在该问句中焦点是“哪一个”，而主题是“美国现任领导”。显然根据该问句的焦点并不能确定问句的类型，但是根据该句的主题“领导”在知网中的首义原（领导的首义原是：human|人）可以判断出问题类型为“询问人”。还有一些由通用疑问词位于句尾的问句，比如：地球上地震的原因是什么？这一类的问句的类型也是由问句的主题和焦点共同确定的。

通过对大量问题的观察和统计，本文归纳出了根据问句的焦点和主题判断问题类型的判断规则如下：

(1) 如果问句是由专用疑问词构成的，则可直接由问句的焦点确定问句的类型。

(2) 如果问句是由通用疑问词构成，并且疑问词后紧跟着名词，则可由问句的焦点确定问句的类型。

(3) 如果问句的通用疑问句位于句子末尾或疑问词后紧着动词，则可由主题和焦点共同确定问句的类型。

## 6 实验和实验分析

### 6.1 实验

本系统使用了中国科学院自动化研究所模式识别国家重点实验室和哈尔滨工业大学信息检索实验室提供的问句集，共选取了 4500 个问句。从中选取了 900 个问句，并对一些问句进行人工扩展，共 1600 个问句作为测试集，剩下的 3600 个问句作为训练集。对训练集中的问句进行人工类型标注，并抽取出大类特征模型和小类特征模型，人工分类共分为 7 大类，每个大类根据实际情况再定义了一些小类，共 60 个小类[11]，如下表：

表 2 本文的问题分类体系

| 大类 (Coarse) | 小类(Fine) |
|---|---|
| 人物 (HUM) | 特定人物 团体机构 人物描述 人物列举 人物其他 |
| 地点(LOC) | 星球 城市 大陆 国家 省 河流 湖泊 山脉 大洋 岛屿 地点列举 地址 地点其他 |
| 数字 (NUM) | 号码 数量 价格 百分比 距离 重量 温度 年龄 面积 频率 速度 范围 顺序 数字列举 数字其他 |
| 时间 (TIME ) | 年 月 日 时间 时间范围 时间列举 时间其他 |
| 实体(OBJ) | 动物 植物 食物 颜色 货币 语言文字 物质 机械 交通工具 宗教 娱乐 实体列举实体其它 |
| 描述(DES) | 简写 意义 方法 原因 定义 描述其它 |
| 未知 (Unknown) | 未知 |

另外还通过人工方式构造了专用疑问词集和通用疑问词集。

实验性能采用的评价标准是分类准确率，可用如下公式求得:

$$\text{分类准确率}=\frac{\text{测试集中正确分类的问题数}}{\text{测试集中总的问题数}}*100\%$$

在实验中，我们利用了哈尔滨工业大学信息检索研究室提供的词法分析器和句法分析器,对测试集中的每一问句,通过分析器的分析,得到问句的每一个词语及其词性和问句的主语、谓语、宾语。然后根据前文提到的方法，利用训练集、专用疑问词集、通用疑问词集等，顺序完成以下操作：

① 焦点的确定

如果疑问词后紧跟着名词，则焦点是由疑问词和名词组成，否则焦点就是疑问词。

② 主题的确定

如果疑问词位于句子的末尾，则主题是句子的主语，否则主题是句子的宾语。

③ 问句的分类

如果疑问词是专用疑问词或由通用疑问词后紧跟名词，则可由问句的焦点，根据人工标注的大类特征模型和小类特征模型得也该问句所属的大类和小类；否则，由问句的主题、主题在知网中的义原和问句的焦点，根据大类特征模型和小类特征模型判断出该问句的所属的大类和小类。

通过实验，利用上述分类准确率公式分别对表 2 中的大类和小类的分类准确率进行评价，得出的实验结果如下表：

表 3 基于焦点和主题的分类的实验结果

| 类别 | 准确率 |
|---|---|
| 7 大类 | 88.36% |
| 60 小类 | 80.19% |

### 6.2 实验结果分析

由表 3 可以看出，当使用基于焦点和主题对用户问句进行分类时，能够取得较好的实验结果，其中 7 个大类的分类准确率最高可以达到 88.36%，而 60 个小类的分类准确率可以达到 80.19%。同时对实验中出现错误问题分类进行分析后发现，主要由以下原因造成的：

(1) 分词和词性标注造成的错误。对于一些新词和岐义词等常出现分词或词性标注的错误。比如问句：“自然数集中的所有小于 1000 的质数集中在一起是多少个？”，分词系统将该问句中的两个“集中”都划分成一个词，显然这是不正确的。

(2) 知网（HowNet）是一个以汉英双语来表示概念与概念之间以及概念的属性之间关系的知识库，它是一个语义词典。一方面由于它上面的词语是有限的，所以有一些词在知网上是没有的，从而不能确定其首义原；另一方面本文在一些问句类型的确定是利用主题词在知网中的首义原，这也会产生错误。比如问句：“怎样才能取得博士学位？”中的“博士”一词在知网中的首义原是：human|人，显然在该问句它并不是指人。

(3) 句法分析析造成的错误。由于句法分析到目前还没有达到百分之百的准确，所以也会造成分类的错误。比如对于一些接近于口语含有多个动词的复杂问句就容易造成分析的错误。

(4) 对于一些在问题集很少出现的特殊问句，比如“黄山凭什么闻名于世？”，在分类时就很容易出现错误。

## 7 总结与展望

从实验的结果可以看出，本文提出的基于焦点和主题的问题理解能够取得较好的性能，不管是大类的分类准确率，还是小类的分类准确率都取得较好的效果。目前，问题理解仍然是自动问答系统中的重要而关键的一步。下一步的我们将对本文提出的方法进行进一步的改进，特别是对实验中出现错误的分类进行分析和改进，使分类的准确率得到进一步的提高。

致谢 本文使用了哈尔滨工业大学信息检索研究室和中科院自动化研究所模式识别国家重点实验室提供的资源。在此，对他们表示诚挚的感谢。

## 参考文献

1 Voorhees EM. Overview of the TREC 2003 Question Answering Track. The Twelfth Text Retrieval Conference. Gaithersburg, Maryland, 2003,54–69.

2 董振东,董强.知网.[2010-03-12].http://www.keenage.com/c_zhiwang.html

3 孙景广,蔡东风,吕德新,董燕举.基于知网的中文问题自动分类.中文信息学报,2007,21(1):90–95.

4 朱德熙.语法讲义.北京:商务印书馆,1982.

5 张华平,刘群.基于 N-最短路径方法的中文词语粗分模型.中文信息学报,2002,16(5):1–7.

6 安玉璞.自然语言问答系统的设计与实现[硕士学位论文].哈尔滨:哈尔滨工业大学,2003.

7 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,2006,42(3):175–177.

8 王开扬.汉语的自动理解与汉语文本的改进.术语标准化与信息技术,2006,(4):36–40.

9 唐娟,杜亚军,王可亮.一种基于形式形式概念分析的问答系统答案抽取的研究.计算机应用,2007,27(3):653–655.

10 金砚硕.中文问答系统中答案提取的研究[硕士学位论文].鞍山:辽宁科技大学,2008.

11 文勖,张宇,刘挺,马金山.基于句法结构分析的中文问题分类.中文信息学报,2006,20(2):33–39.