自动问答系统中的问题理解研究

曹志娟　李祖枢　刘朝涛
（重庆大学智能自动化研究所　重庆 400044）

摘　要　问题理解是问答系统的首要的分析工作，分析的结果对后面的处理，以至找到问题的正确答案都有很大的影响。本文将对常规的问题理解方法进行改进，从而使系统能够较准确地回答用户的提问。实验证明新的方法对提高系统性能有显著作用，尤其针对性强、意思表达清晰的提问，回答准确率有很大提高。

关键词　问答系统，虚拟信息顾问，问题理解，分类，扩展

Study of Question Analysis in Question-Answering System

CAO Zhi-Juan　LI Zu-Shu　LIU Chao-Tao
(Intelligence and Automation Laboratory, Chongqing University, Chongqing 400044)

Abstract　Question analysis is the primary task of Question answering System. The result of Question analysis has a great effect on following processing work, even on finding the correct answer. In this article, we improve general method of Question Analysis that our system can now answer user questions with high precision. The new method has been proved most useful in improving the performance of the system, especially in direct and clear questions.

Keywords　Question answering system, Virtual information consultant, Question analysis, Categorization, Expansion

1　引言

随着网络和信息技术的快速发展，人们想更快地获取信息，然而传统的搜索引擎似乎已不能满足人们的需求，这就促进了一种新技术的发展——自动问答系统，也称为问答系统或 QA 系统，允许用户输入一个问题，目标是返回一个简短而准确的答案。

每年一度的文本信息检索（TREC）会议上，自动问答（Question Answering Track）是最受关注的主题之一。越来越多的大学和科研机构参与了 TREC 会议的 Question Answering Track。

目前，国外已经开发出一些相对成熟的问答系统。麻省理工（MIT）开发的问答系统 Start，从 1993 年开始发布在 Internet 上，可以回答一些有关地理、历史、文化、科技、娱乐等方面的简单问题[1]。另一个比较成熟的问答系统 AnswerBus，它是个多语种的自动问答系统，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[2]。

国内也有一些研究机构参与了自动问答技术的研究：中科院计算所、复旦大学、香港科技大学，哈尔滨工业大学[3~5]。但是参与中文自动问答技术研究的科研机构比较少，而且基本没有成型的中文自动问答系统。

问题理解阶段是自动问答系统执行的开始，本文将对这一部分目前存在的问题进行比较分析，提出一种更为有效的方法，对用户的提问进行详尽的分析和判断，这将大大提高系统后期工作的准确率，从而提高系统的性能。

2　虚拟信息顾问系统

虚拟信息顾问系统（Virtual Information Consultant）是一个开放域中文问答系统，它利用自然语言处理技术，理解用户提出的问题，然后利用搜索引擎自主搜集相关资料，并处理抽取的文档，得到明确答案。系统主要包括五个部分：问题理解、信息检索、信息处理、答案抽取、FAQ 系统。系统结构如图 1 所示。

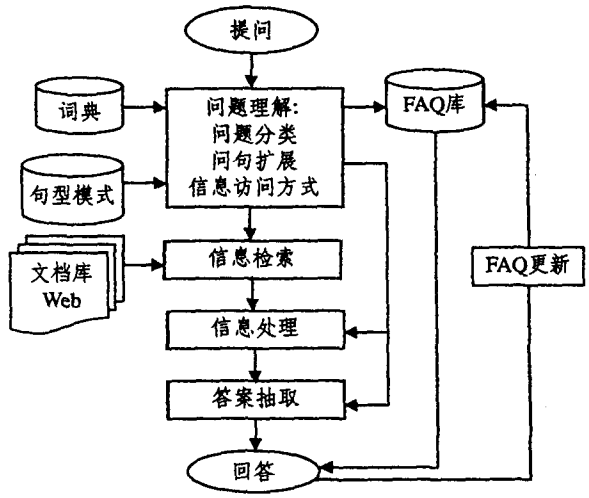

图 1　虚拟信息顾问系统框图

其中三个主要部分是：问题理解、信息检索、答案抽取。如何在问题理解阶段充分理解用户的提问意图，如何在信息检索模块中把相关的文档找出来，如何在答案抽取模块中准确地把答案从相关文档中抽取出来，这三个问题是自动问答技术要解决的核心问题。

3　问题理解

问题理解（Question Analysis）模块首先要对问句分词和标注，也就是词法分析，国内词法分析技术研究已很成熟，因此目前问题理解研究的重点将是两大任务：问题分类和问题扩展。目的是从问句中提取关于提问主旨的重要信息和细节

曹志娟　硕士研究生，主要研究方向：人工智能与模式识别，自然语言理解；李祖枢　教授，博士生导师，主要研究方向：人工智能与模式识别，智能控制与智能自动化，智能机器人，仿人智能控制理论及应用；刘朝涛　博士研究生，主要研究方向：人工智能与模式识别，自然语言理解。

特征，用于抽取可能包含答案的段落。

3.1 问题分类

传统的问题分类是基于语义的分类，即根据答案对象的类型进行划分，如询问人物、地点、时间、数量等[6]。这种方法的好处在于人可以直观地知道问题所指向的对象，但是让计算机只通过一些规则或算法，一次性准确识别提问的对象却难以实现，尤其对于表达形式丰富的中文。所以我们将在传统分类的基础上增加疑问词短语分类、问题标准型、特征词分类，使计算机对问题的理解更详尽，也使后期信息检索针对性更强。

在问题分类模块中，系统首先识别问句中包含的疑问词短语，根据疑问词短语找到对应的句型模式集，然后与模式集中的句型规则进行匹配，从而得到问题标准型，由此得知问题的类型，再根据特征词确定问题领域，得到搜索答案时所需要的访问方式，确定搜索的数据源。

3.1.1 *疑问词短语分类* 通过对大量问题的统计发现，用户提出的问题可以分为若干种类型，下表列出了常见的问题类型。

表 1 问题类型表

问题类型	疑问词短语	例子
人	谁/什么人/哪个人/哪一个人/何人	人工智能是什么人提出的?
地点	什么地方/什么地点/哪里/哪儿/何处	第 28 届奥运会的举办城市是哪里?
具体时间	什么时间/什么时候/哪个时候/何时	人类第一次登上月球是什么时间?
持续时间	多久/多长时间/多少时间	火车从北京到重庆要多久?
数量	多少/几	中国国民最低收入是多少?
原因	为什么/什么原因/什么因素	为什么手机的辐射很大?
方法	哪些方法/哪些方式/哪些算法/哪些途径/什么方法/什么方式	可以通过哪些方法改善环境?
其他	—	—

表 1 与常规的问题类型表有些细微的差异：一般问答系统都是选取分词得到的单个词作为判断问题类型的疑问词，我们则进一步将一些联合比较紧密、询问目的明确的词语与疑问词合并，生成新的疑问词短语，同时把完全可以互相替换的疑问词短语归为同一组，其中的一个疑问词短语作为“关键疑问词”。例如表 1 中的“什么人/哪个人/哪一个人/哪些人/何人”为一组，“什么人”为该组的关键疑问词。

这样就可以使属于同一类型的问题都使用相同的分析算法，而不会因为疑问词的不同，重复制定算法规则，进行重复分析。

例：人工智能是谁提出的?

不同的用户可能有不同的提问形式：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。显然以上 4 个问句问的是相同的问题，但使用的疑问词和句子的表达形式却各不相同，如果采用传统的问题分类的方法，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。

但是通过与表 2 的匹配，我们就可以把它们归为同一类型的问题。通过对表中“句型模式”的不断扩充，系统就可以接受用户各种形式的提问，进而理解和回答。

表 2 问题模式匹配表

疑问词短语	关键疑问词	句型模式	问题标准型
谁	什么人	np+是+~+vp	np— vp—who
什么人		~+vp+np	
哪个人		~+vp+np	
何人		np+是+~+vp	

其中“句型模式”属于汉语语言学的研究内容，只有对大量的语料进行统计和分析才能得到较完备的句型模式集，这也是汉语语言学长期研究的方面，自动问答系统性能将会因为语言学研究在这一方面取得的成果而得到极大的提高。

3.1.2 *问题标准型* 识别疑问词短语只能得到问题的基本类型，不足以确定搜索策略，还需要依据句型规则进一步划分。这些规则可以用形式语言表示出来[7]。系统实现了对这种形式语言的解释，可以动态地对规则解释执行。在实际的处理过程中，还存在相当一部分问题很难确定问题类型，这样的问题称为“其它”类型。对于这样的问题不能制定具体的规则，将采用概率分类的方法[8]，概率分类方法需要收集大量的问题作为训练语料，通过程序统计出问题句型属于各种问题类型的概率，然后选择概率值最大的问题类型作为该问句的类型。

当系统能够接受用户各种形式的提问，进行简单分类之后，又如何保证相同类型的提问都能得到相同处理呢？于是我们引入“问题标准型”，通过问题标准型，实现多(多种提问方式)对一(问题标准型)，一(问题标准型)对多(多种答案抽取规则)的映射。

问题标准型不仅是同种类型问题的唯一标识，也是某一问题类型的代表性结构表达式，可看作是一个直观的三元表达式：对象-属性-值。“属性”是对象的属性，“值”是属性的值，就是系统要搜索的问题答案。

当然，可能存在很多类型的问题不能用“对象-属性-值”的模式表达，例如关于两个对象之间关系的问题(北京到重庆有多少公里?)。然而，我们的实验发现，能够用“对象-属性-值”的模式表达的问题在实际提问中出现很频繁。

表 3 是前面例子对应的包含“问题标准型”的答案匹配表。

表 3 答案模式匹配表

问题标准型	答案抽取规则	回答形式
np— vp—who	who+ vp + np	〈NAME〉+提出+人工智能
	np+是+who+vp	人工智能+是+〈NAME〉+提出
	vp +np+是+who	提出+人工智能+是+〈NAME〉

使用问题标准型，就可以建立起多种提问方式与多种答案形式的联系，根据问句构建对应答案的结构形式，也即答案抽取规则，为后面抽取包含答案的段落提供很好的搜索依据。

3.1.3 *特征词分类* 目前网络提供的信息大多数是无结构化表达形式，让人类使用是可以的，但要让计算机理解这些数据就很困难，这就迫使我们在没有完全实现语义网络之前，如果要让计算机有效地搜索信息，那么最好的方法就是有针对性地访问网络，而不是漫无目的地搜索所有网站。

假设用户提问：《珍珠港》的导演是谁? 如果系统知道“珍珠港”是电影名字，那么就可以到一个专门的电影网站上进行搜索，而不像 Google 那样返回数以万计的网页。“珍珠港”就

是这个问句中的特征词。

系统数据库中包含各种类型的特征词，这就使系统能够识别问句中的特征词。特征词确定了搜索的对象，同时也明确了数据源和访问方式。由于不同数据源的信息组织结构不同，因此访问数据的方式也不一样，对于不同的数据源，系统将使用不同的访问模块。如果网站的知识很完备，那么只需要很少的网站就可以较好地回答出现频繁的提问。下面是两个例子：

www. weathercn. com 是一个包含地区较多的天气查询网站。大多数的天气网站只有大城市的预报，而该网站都覆盖各中小城市的天气情况，使系统能较好地解决用户关于气候方面的提问。

CIA World Factbook 提供了各个国家的各种情况，如：人口、面积、首都、经济等内容。一个网站就可以回答各个国家地理、文化、经济等常识问题。

特征词的选取要求针对性强，特征明显，有代表性，容易区分。特征词的扩充是一个长期的、连续的工作。

3.2 问句扩展

原始的问句一般都比较简练单一，不可能包含查找相关文档需要的所有词语，也就是说会导致查全率较低，因此就需要对原始的问句进行扩展，建立问句与文档的充分联系，从而帮助问答系统找到正确的答案。在我们的问题理解模块中问题扩展主要通过两种方式实现：问句重写和关键词扩展。

3.2.1 *问句重写* 每一种问题类型对应一条详细的重写规则，重写规则有 1～5 种形式。输出的重写模式是三元形式：

[string, L/R/-, weight]

其中，“string”表示重写的问句，“L/R/-”表示我们期望找到答案的位置是在问句的左侧还是右侧，“weight”表示我们希望在这种特殊问句（重写问句）中找到答案的期望值。引用“weight”的意义在于：在一个期望值高的问句中比在一个期望值低的问句中更有可能找到正确的答案。

系统的句子重写是简单的字符串操作，一些问句的重写句子是通过移动主要动词，例如句子“地震是怎么产生的?”的重写句子“地震产生是……，产生地震是……”，只要分析句子的语法我们就能确定动词可能移到什么地方。假定有简单问句“谁提出 w1w2…wn?”，这里 wi 表示词，我们重写的句子将重新定位动词，例如：w1 提出 w2…wn；w1 w2 提出…wn，等等。然而这里面很多的重写的句子没有意义，不符合语法，用这种不符合语法的句子搜索文档是不会产生与答案无关的段落，而恰当的移动得到的合理的重写句子就可以被查找到，因而也就找到答案的所在。如果仅仅依赖于分析器，我们将得到很少的重写句子，错误的分析将导致恰当的重写句子不能找到。

问题“什么是相对湿度?”的重写句子如下：

[“＋是相对湿度”, LEFT, 5]

[“相对＋是湿度”, *RIGHT*, 3]

[“相对湿度＋是”, RIGHT, 3]

[“相对湿度”, NULL, 2]

[“相对” 和“湿度”, NULL, 1]

3.2.2 *关键词扩展* 有的时候问句中的关键词显然存在同义词，而包含答案的段落正是包含了关键词的同义词，而不是关键词本身，这种情况下，如果我们在搜索答案之前对关键词进行扩展，则可能很快定位到包含答案的段落，提高系统的查全率和查准率。例如，我们预先对词语“发生”进行扩展，得到“产生”，那么，对于问句“海市蜃楼是怎么发生的?”，明确包含答案的段落是“海市蜃楼是光在密度分布不均匀的空气中传播时发生全反射而产生的……”，这样我们就很快找到了答案，从而提高系统的执行效率。

目前我们主要是对名词、动词的关键词做扩展，主要通过查询语义词典和语料库，以及观察统计实现。关键词是在问题分类时，从句型模式匹配成功后得到的核心词组中抽取的。

问句扩展模块同时对句式与词汇扩展，采用重写句子和关键词扩展的权值排序得到的搜索策略，将增大搜索的覆盖面，同时又提高这些文档包含答案的可能性。

4 实验结果与分析

为了验证我们采用的问题理解技术的可行性，我们征集了 90 个简单问句做了一次仿真实验，下面是系统的执行步骤：

- 问句分词和词性标注，去掉停用词；
- 根据语料库，识别问句中的疑问词短语；
- 根据句法分析的结果，匹配句型模式；
- 抽取特征词，同时得到数据源和访问方式；
- 问句扩展，得到“答案搜索模式”；
- 搜索策略排序，进行搜索。

表 4 实验结果

提问对象	问句数量（个）	问题类型识别正确数量（个）	查准率（检索结果前 30 项）
人	30	26 (0.912)	0.767
地点	20	19 (0.905)	0.733
具体时间	10	12 (0.857)	0.600
天气	30	28 (0.933)	0.892

可以看出该系统找到包含答案信息的查准率在 70％以上，尤其针对性较强的“天气”类型查准率是最高的。可以预测，后期工作从这些相关文档中抽取答案，正确率就不会很低，这个结果还是比较好的。因为从 Trec 会议上看，国际上一般的问答系统最后回答的查准确都小于 40％[11]。

下面的因素对测试带来一些影响：1）分词和词性标注的错误；2）句法分析的错误导致不能匹配到正确的句型模式，甚至没有符合的句型与问句匹配；3）句型模式通用性不强和数量缺乏。

从实验数据还可以看出，系统的查准率在问题理解执行的每个阶段都有所下降，就是说后一阶段的查准率不会比前一阶段高。那么提高前期工作的查准率将能有效地提高系统的性能。

结束语 问题理解是自动问答系统首先进行的分析工作，分析的结果就是后期要处理和使用的信息。本文提出的问题理解方法从系统执行初期就已对后阶段的工作进行了充分的考虑，对用户的提问进行了明确的分析，保证了后阶段能得到有效的分析结果，从而显著地提高了系统的性能。

中文自动问答系统不仅可用作智能搜索引擎，还可应用

（下转第 230 页）

议〉

系统结构(SystemStructure):包括系统整体结构和构件关系结构,系统整体结构定义构件与连接件的整体关系,构件关系结构定义不同构件子集合中的构件之间的关系。TADL中的系统结构可以表述成:

系统结构::=〈系统整体结构,构件关系结构〉

系统整体结构::=〈构件子集合,连接件子集合,基础构件子集合,组织规范〉

构件关系结构::=〈原子结构,指针〉

原子结构::=〈构件名,所处子集合名,关系定义〉

关系定义::=〈单向关系|双向关系〉

单向关系::=〈方向标志,原子结构,指针〉

双向关系::=〈方向标志,原子结构,指针〉

构件语义约束(Semantic Map):描述了构件之间的语义关系。系统结构描述的是构件之间的拓扑关系,而构件语义约束则从构件功能定义出发,在构件的不同子集合内部,给出构件使用时的语义连接顺序和规则。在TADL中,根据构件子集合本身的功能特点,为必要的构件子集合定义了构件语义关系。如果构件子集合有 n 个构件,则此构件子集合的构件语义约束为二维的 $n*n$ 的表结构,构件之间的语义关系是有方向的,此结构为自动生成领域应用程序提供了生成规则。

```
SemanticMap ={
        ComponentSetName;
        Table ={ a_ij }n * n // 当第 i 个构件到第 j 个构件由一
                个顺序规则时,a_ij =1
        }
```

系统模式(Styles):表示一类体系结构族,或称风格。目前的体系结构风格主要有:客户端/服务器式体系结构,管道式体系结构,对象连接式体系结构,接口连接式体系结构,插头插座式体系结构和分层递阶式体系结构[11]。在TADL中,目前只定义了一类结构,就是分层的结构模式。

系统模式::=〈构件子集合,连接件集合,系统结构,构件语义约束,系统规范〉

5 样本程序生成工具设计

本文定义的形式化描述语言TADL可用于描述系统的模型,同时提供了丰富的连接件类型来表达系统构件之间的通信关系,对每一个系统元素的定义都是基于实现的,所以为样本程序生成提供了方便的数据结构。

对于构件系统样本程序的生成,要解决的主要问题是:①如何从众多的构件集合中选择需要的构件;②按照怎样的结构来集成构件,以形成样本程序。本文提出的TADL通过定义构件的体系结构给出了集成构件,形成样本程序的方案;通过定义构件子集合间的结构关系和构件子集合内的语义关系,给出了选择构件的方案。具体过程包括:①建立应用领域的领域模型;②建立Use Case模型;③解析Use Case模型、划分构件集合、对应构件基础层;④定义构件子集合内的语义关系;⑤生成不同构件子集合间的结构关系树;⑥集成构件,形成样本程序。

结束语 本文在分析构件系统的特点的基础上,提出了一种基于结构模式的测试方法,增加对构件系统内部不同层次的信息描述,包括不同构件子集合间的结构关系描述,以及构件子集合内部的语义关系描述,为构件的使用者集成构件生成测试用样本程序提供必要的信息。同时给出了本方法的形式化描述,用于对待测构件系统的模型描述,便于转化为可执行代码,实现样本程序的生成工具。本方法同样适合构件的开发者进行集成测试和系统测试,为构件系统提供安全保证,使得这种开发模式发挥更大的作用。

参 考 文 献

1 Weyuker E J. Testing Component-Based Software: A cautionary tale. IEEE Software, Sept.-Oct. 1998
2 Kozaczynski W, Booch G. Component-based Software Engineering, IEEE Software, Sept.-Oct. 1998
3 Voas J. Maintaining Component-Based System. IEEE Software, July-Aug. 1998
4 Rosenblum D S. Challenges in Exploiting Architectural Models for Software Testing. In: Intl. Workshop on The Role of Software Architecture in Testing and Analysis, Marsala, Sicily Italy, 1998
5 Harrold M J. An Approach To Analyzing and Testing Component-Based system. In: ICSE'99 Workshop on Testing distributed Component-Based System, May 1999
6 Orso A, Harrold M J. Component Metadata For Software Engineering Tasks. Proc. of EDO 2000, LNCS Vol 1999
7 Harrold M J. Testing: A Roadmap. In: 22nd Intl. Con. on Software Engineering, June 2000
8 Rothermel G, Harrold M J. Analyzing Regression Test Selection Techniques. IEEE Transactions on Software Engineering, Aug. 1996
9 Rothermel G, Harrold M J. Empirical Studies of a Safe Regression Test Selection Technique. IEEE Transactions on Software Engineering, June 1998
10 Binder R V. 面向对象系统的测试. 人民邮电出版社, 2002
11 Shaw M, Garlan D. Software Architecture. Prentice Hall, April 1996
12 Allen R J, et al. Formal Modeling and Analysis of the HLA Component Integration Standard. In: Proc. of the Sixth Intl. Symposium on the Foundations of Software Engineering (FSE-6), Nov. 1998
13 Allen R, Garlan D. A Formal Basis for Architectural Connection. ACM Transactions on Software Engineering and Methodology, July 1997

(上接第160页)

在远程教育、企业客户咨询等方面。广阔的应用前景正推动着自动问答技术的快速发展,相信在不久的将来问答系统的研究将会取得重大的突破并得到广泛的应用。实现计算机理解自然语言的提问正是计算机理解人类语言的开始。

参 考 文 献

1 http://www.ai.mit.edu/projects/infolab/
2 http://misshoover.si.umich.edu/~zzheng/qa-new/
3 Chang Yi, Xu Hongbo, Bai Shuo. TREC 2003 Question Answering Track at CAS-ICT. In: The Twelfth Text Retrieval Conf. Gaithersburg, Maryland, 2003. 147~152
4 Wu Lide, Huang Xuanjing, Zhou Yaqian, et al. FDUQA on TREC2003 QA task. In: The Twelfth Text Retrieval Conf. Gaithersburg, Maryland, 2003. 246~254
5 郑实福,刘挺,秦兵,李生. 自动问答综述. 中文信息学报, 2002, (6): 46~53
6 Li X, Roth D. Learning Question Classifiers. In: Proc. of the 19th Intl. Conf. on Computational Linguistics (COLING). Taipei, Taiwan, 2002. 556~562
7 Zhang D, Lee W S. A Language Modeling Approach to Passage Question Answering. In: The Twelfth Text Retrieval Conf. Gaithersburg, Maryland, 2003. 489~496
8 Ravichandran D, Hovy E. Learning Surface Text Patterns for a Question Answering System. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia, July 2002. 41~47
9 Wang G, Chua T-S, Wang Y-C. Extracting Key Semantic Terms from Chinese Speech Query for Web Searches. In: The Eleventh Text Retrieval Conf. Gaithersburg, Maryland, 2002. 389~396
10 Wu Min, Zheng Xiaoyu, Duan Michelle, et al. Question Answering By Pattern Matching, Web-Proofing, Semantic Form Proofing. In: The Twelfth Text Retrieval Conf. Gaithersburg, Maryland, 2003. 578~586
11 Voorhees E M. Overview of the TREC 2003 Question Answering Track. In: The Twelfth Text Retrieval Conf. Gaithersburg, Maryland, 2003. 54~69