自动问答综述

郑实福 刘挺 秦兵 李生
（哈尔滨工业大学信息检索研究室 哈尔滨 150001）

摘要：自动问答技术是自然语言处理领域中一个非常热门的研究方向，它综合运用了各种自然语言处理技术。本文介绍了自动问答技术的发展现状和自动问答系统中常用的技术。自动问答系统一般包括三个主要组成部分：问题分析、信息检索和答案抽取。本文分别介绍了这三个主要组成部分的主要功能和常用的方法。最后还介绍了自动问答系统的评价问题。

关键词：自动问答，问题分类，信息检索，答案抽取

Survey on Question-Answering

Zheng Shi-fu Liu Ting Qin Bing Li Sheng
(Information Retrieval Laboratory, Harbin Institute of Technology, Harbin 150001)

Abstract: Question-Answering is a hot research field in Natural Language Processing, which includes many kinds of NLP technology. This paper introduces the current research status and the methods that are often used in Question-Answering. In general, a Question-Answering system is make up of three parts: Question Analysis, Information Retrieval and Answer Extraction. This paper describes the main function of these three parts and the common approach used in these parts in detail. At last, this paper introduces the evaluation of Question-Answering system.

Keywords: Question-Answering, question classification, information retrieval, answer extraction

1. 引言

随着互联网的普及，互联网上的信息越来越丰富，现在人们能够通过搜索引擎方便的得到自己想要的各种信息。比较有名的搜索引擎有 Google、Sohu、Yahoo 等。无论哪方面的内容，这些搜索引擎都能帮助人们快速地找到相关的网页。用户只需输入一些关键字，它们马上就会搜索出相关的网页。

但是这些传统的搜索引擎存在很多不足的地方，其中主要有三个方面：一是相关性信息太多。传统的搜索引擎返回的相关网页太多，用户很难快速准确地定位到所需的信息。例如，用户在 Google 上输入几个关键字，它有可能返回成千上万个网页，用户将浪费很多时间在这些网页中查找自己所需要的信息。二是以关键词的逻辑组合来表达检索需求，因为人们的检索需求往往是非常复杂而特殊的，是无法以几个关键词的简单组合来表达的，这样用户都没有将自己的检索意图表达清楚，搜索引擎自然也就没有办法找出令用户满意的答案了。三是以关键词为基础的索引、匹配算法尽管简单易行，毕竟停留在语言的表层，而没有触及语义，因此检索效果很难进一步提高。

为了克服传统搜索引擎的弊端，国外一些有实力的科研机构和大公司正在探索新的检索技术，在这方面最成功的检索系统是美国 AskJeeves 公司的检索系统，网址为：http://www.askjeeves.com/。AskJeeves 最突出的特点是允许用户用自然语言句子提问，检索系统会自动分析用户的提问，然后通过反问，即人机交互方式，准确地辨识用户的意图，这样用户就能够充分表达他的检索需求，这比 Yahoo 的关键词检索方式有了明显的进步。香港科技大学参考 AskJeeves 的思路正在做中文的提问式搜索引擎 Weniwen，网址为：http://www.weniwen.com/。100 多个学生被组织起来对 Internet 上的各个网页进行提问，这些提问被记录下来作为网页的索引，在实际使用时，如果用户的某个提问与作为索引的某些提问在语义上非常接近，那么就把与这些提问相连的网页返还给用户。AskJeeves 和 Weniwen 提供了自然语言句子的提问方式，这和关键词的提问方式相比，无疑是一个进步，但是 AskJeeves 和 Weniwen 的返回结果仍然是网页，而不是问题的直接答案。

和 AskJeeves 以及 Weniwen 不同，自动问答系统既能用自然语言句子提问，又能为用户直接返回所需的答案，而不是相关的网页。所以，问答系统能更好的满足用户的检索需求，能更快地找出用户所需的答案。可以说，问答系统就是新一代的搜索引擎。对于问答系统，用户不需要把自己的问题分解成关键字，用户可以把整个问题直接交给问答系统。问答系统结合自然语言处理技术，通过对问题理解，能够直接提交给用户想要的答案。问答系统就像一个知识渊博的专家，可以快速准确地回答任何问题。比如，用户提交一个问题“上海的简称是什么？”问答系统将会直接给出答案“上海的简称是沪”。可以看出，问答系统要比传统的搜索引擎方便、快捷、高效。

2. 研究概况

早在 60 年代人工智能研究刚开始的时候，人们就提出了让计算机用自然语言来回答人们的问题，这就是指自动问答系统。问答系统在 80 年代的自然语言处理领域曾风行一时，因为 Turing 实验告诉人们如果计算机能够象人一样与人进行对话，就可以认为计算机有智能，所以研究者们为了探索语言理解技术，纷纷研究自然语言问答系统。但是，由于当时的条件限制，所有的实验都是在非常受限的领域，甚至是固定段落上进行的，所以自动问答一直被限制在特殊领域的专家系统。此后，由于大规模文本处理技术的兴起，问答系统的研究受到了冷落。

最近几年，随着网络和信息技术的快速发展，同时人们想更快地获取信息的愿望也重新促进了自动问答技术的发展。最近有越来越多的的公司和科研院所参与了自动问答技术的研究。比如，微软和 IBM 等著名的跨国公司。在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。在 2000 年 10 月召开的 ACL2000 国际计算语言学学术会议上，有一个专题讨论会，题目是”Open-Domain Question Answering”。

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） 就开发出一个问答系统 Start，从 1993 年开始发布在 Internet 上，网址如下：http://www.ai.mit.edu/projects/infolab/。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。比如：对于问题“What is the longest river in the world?”Start 将会回答“With a length of 4,180 miles, the Nile River is the longest river in the world.”另外还有一个比较成熟的问答系统 AnswerBus 的网址是：http://misshoover.si.umich.edu/~zzheng/qa-new/。AnswerBus 是个多语种的自动问答系统，它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。问答系统一般包括三个主要部分：问题分析、信息检索和答案抽取。如图 1 所示：

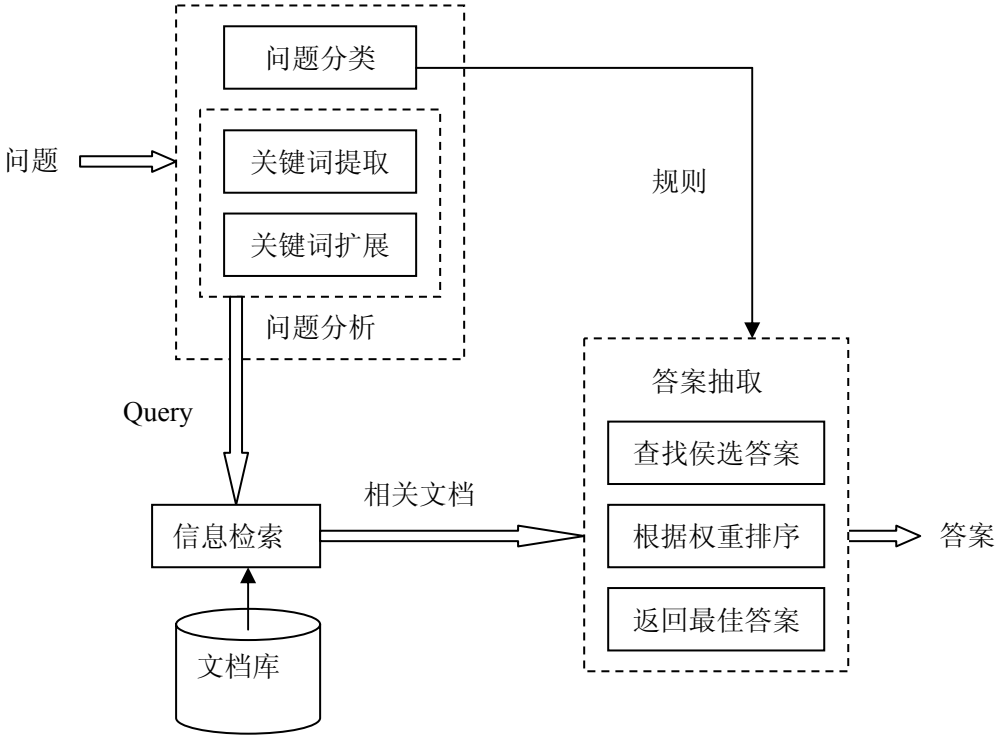

图 1：自动问答系统结构

对于用户提交的问题，首先要对问题进行分析，要理解用户的要问的是什么。比如，“华山在哪里？”问题分析模块通过对这个问题的分析，就可以知道用户是在问华山的地理位置。问题的分析一般包括问题的分类、关键词的提取和关键词扩展。如果是中文，还需要进行分词处理。

通过问题分析而得到的关键词集需要提交给信息检索模块来查找相关的文档。检索系统的任务就是在已有的文档库中搜索和关键词集相关的文档。为了保证对任何问题都能找到相关的文档，文档库必须足够大。文档库中可以从互联网上下载。也可以把百科全书加到文档库中。

信息检索模块返回的是一堆相关的网页。然后答案抽取模块从这些相关的网页中找出相关的答案（一句话，或者是一段）提交给用户。答案抽取是问答系统的最后一部分，也是难度最大的一部分。如果答案抽取模块不能准确地把正确答案抽取出来，将严重影响整个问答系统的准确性。

除了上述三个模块之外，有的问答系统还包括了一个常问问题（FAQ）库，把用户经常问的问题及其答案保存起来。有了 FAQ 库之后，对用户问的问题先在 FAQ 库中搜索，看看有没有相同的问题，如果有，就可以直接把 FAQ 库中这个问题的答案返回。这样，对于用户常问的问题，问答系统就可以很快给出答案，而不需要经过复杂的处理，而且还能保证答案的正确。所以有了 FAQ 库之后，既能提高问答系统的效率，又能提高准确性。

3. 问题分析

问题的分析是问答系统首先进行的分析工作，这个过程分析的效果对后面的处理过程有着重要的影响。问题分析部分需要完成以下几部分工作：确定问题的类型、提取出问题的关

键词、依据问题的类型等因素对关键词进行适当的扩展。如果是汉语的问答系统，首先要对问题进行分词以及词性标注等。有一部分的问答系统还对问题进行了语法分析和语义分析。

3.1 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这个过程。这里我们以中文问答系统为例。一般的问答系统都按照疑问短语来对问题的进行分类。下表列出了常见的问题类型：

问题类型	疑问词	例子
询问人	谁	谁发现了北美洲？
询问时间	什么时候 / 何时 / 那年…	人类哪年登陆月球？
询问数量	多少 / 几 / 多大 / 多高…	茉莉花每年能开花几次？
询问定义	是什么 / 什么是	什么是氨基酸？烟碱是什么？
询问地点或位置	哪 / 哪里 / 什么地方	黄山在哪个省？
询问原因	为什么	天为什么是蓝的?
其它	—	—

表 1：常见的问题类型

针对于不同类型的问题制定相应的答案抽取规则，以便在答案抽取阶段应用这些规则来抽取问题的答案。比如对于询问地点的问题，我们就可以规定，答案中必须含有位置信息。

大部分的自动问答系统都是按照事先规定好的类别进行分类。但是这种分类还是存在很多不足的地方，太多人为的因素，而且分类太粗，并不能完全符合实际的要求。所以也有一部分研究人员提出对问题自动分类的思想。首先收集大量的问题作为训练语料，然后通过程序统计出经常出现的疑问短语。比如通过统计，“什么颜色”这几个词经常出现在问题中，那我们就可以把“什么颜色”当作一个疑问短语。然后凡是含有“什么颜色”这个短语的问题都当作一类问题。

3.2 关键词提取

我们需要在用户提问的问题中，提取出对后面检索系统有用的关键字。并不是在问题中的每个词都可以提取出来作为检索系统的关键词。比如，疑问词和一些常用的“吧、了、的”等词就应该被过滤掉，为此，需要一个停用词表来过滤这些词。

关键词主要由名词、动词、形容词、限定性副词等组成。有一些问答系统还把关键词可以分为两种：一般性关键词、“必须含有”的关键词。所谓“必须含有”的关键词指的是这些关键词必须在答案句子中含有，而一般性关键词可以不被答案句子包含。关键词被赋予不同的权重，在检索句子时这些权重用来计算句子的权重。通常名词、具有限定性作用的副词会有比较高的权重。“必须含有”的关键词由专有名词、限定性副词（如：最大、最高、最快等）、时间（如：1997 年）组成。之所以要制定“必须含有”的关键词原则是因为他们对问题有极强的限定性作用，如果不含有它们的句子是几乎不可能是正确的答案。例如：问题是“世界上最高的山峰是哪座山？”而检索的结果却出现“乔戈里山是世界第二高峰”，这显然不是用户想得到的结果，之所以出现这种情况的原因就在于非常重要的关键词“最高”

没有被答案句子所含有。如果加上“必须含有”的关键词这个限制，那么这个答案就不会被检索出来，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。

3.3 关键词扩展

为了提高检索系统的召回率，一般的问答系统都对关键词进行扩展。在答案句子中某些词常常不是原来问题的关键词，而是这些关键词的同义扩展。例如：问题是“Who is the first American astronaut to do a space walk?”，答案的句子是“Edward White was the first American cosmonaut to do a space walk.”在问题中使用的是“astronaut”，而在答案中却采用了“cosmonaut”这个词汇。这就造成关键词查询失败，因此需要对关键词做适当的扩展。

关键词扩展虽然提高了系统的召回率，但如果扩展不适当会极大地降低了检索的正确率，因此一般的问答系统对关键词的扩展都是很谨慎的。所以这些问答系统都对关键词的扩展添加了很多限制条件，比如只对名词的关键词进行扩展。可用 Wordnet 或者其他的同义词词典来扩展关键词。还有一些问答系统通过统计的办法来扩展关键词。这种方法需要大量的问题和答案语料来进行训练。每一类的问题所对应的答案一般都有某种共同的特性。例如，对于询问地点的问题，答案中经常会出现“在、位于、地处”等关键词。所以通过统计，找出这些词后，就可以把它们加到 query 当中。另外还有一些问答系统是用检索返回来得相关文档来对关键词进行扩展。

扩展后的关键词的重要性往往比从问题中提取的关键词的重要性低，为了提高系统的准确性，很多问答系统又对关键词附了权重，以此来区分他们之间的重要性。

4. 信息检索模块

信息检索的任务就是用前面提取出来的关键字到文档库中查找相关的文档。信息检索模块返回的是一些最相关的文档。在问答系统中的信息检索模块也可以直接调用已有检索系统，比如 Smart 系统，或者也可调用 Internet 上的搜索引擎比如 Google。在 TREC 会议中就不要求每个问答系统都要有自己的信息检索模块，因为 TREC 会议会为每个问题提供最相关的 1000 个文档。这些相关的文档就是用 Smart 检索出来的。信息检索模块的输入一般都是关键字的组合，如果是英文的问答系统，还需要对关键字进行词根操作（Stemming）。

要建立一个信息检索模块，需要对文档库建立索引。这样才能快速地找到包含特定关键词的文档。在建立索引之前，有必要对语料进行预处理，比如去除重复的文档，如果是英文的语料需要进行词根操作（Stemming），如果是汉语语料则需要分词。如果是汉语的语料库，还需要进行分词处理。

信息检索模块中的关键是对文档权重的确定和对文档进行排序。文档的权重可以按照如下公式来计算：

$$Wd = \sum_{i=1}^{n} (KW_i \times TF_i \times IDF_i) + D$$

其中：KW_i 是该文档包含的第 i 个关键词在问题分析阶段的权重，TF_i 是该关键词在这篇文档中出现的频率，IDF_i 是该关键词在文档中出现的反频率，D 是指关键字在文档中的分布密度。关键词在该文档中出现的频率越高则它的 TF 就越大，关键词在越多的文档中出现则

它的 IDF 就越小，反之越大，关键词在这篇文档中分布的越集中，则 D 值越大。TF*IDF 值从一个方面反映了该关键词的重要程度，通常在一个文档中经常出现（TF 大）的词，而很少现在其他文档中的词（IDF 大），该词所含有的信息量就越多，这个词也就越重要。另外如果关键词在文档中的分布越密集，则这篇文档包含相关答案的可能性越大，这篇文档的权重就越大。对文档计算完权重后，就可以按照权重对文档进行排序，把权重最大的那些文档返回给答案抽取模块。

一般信息检索模块返回的都是文档，但是应用于问答系统的信息检索模块返回的可以是文档，也可以是段落，甚至还可以是句子。信息检索模块返回的相关文档中，一般只有文档中一小部分才是问题的答案。在这么多的相关的文档中查找答案还是个很复杂的过程。所以有的信息检索模块返回的是相关的段落，这样，答案的查找就更快了。

5. 答案抽取

一般搜索引擎返回的是一堆网页，而问答系统需要返回的是简短的答案。这样，通过信息检索模块搜索出来的相关文档就要提交给答案抽取模块来提炼答案。答案可以是一句话，或者是几句话，也可以是几个词或者短语。对于那些问时间地点的问题，就可以用很短的语句来回答，而对于询问原因、事件的问题就需要较长的语句才能回答。比如对于问题“9.11 事件的是怎么回事？”就不可能用一句话就能回答的。所以答案的抽取还需要依据问题的类型。

5.1 以句子作为答案

为了处理的方便，很多的问答系统返回的是句子作为答案。在这种系统中，答案的抽取的步骤如下：

（1） 把检索出来的文档分成句子

（2） 按照一定的算法，计算每个句子的权重

（3） 对句子按照权重进行排序

（4） 根据问题的类型对候选答案重新排序

在第二步中，计算句子的权重需要考虑如下方面：句子中含有的关键词、和关键词有相同语义的词、句子中不包含的关键词。具体的计算公式如下：

$$Ws = \sum_{i=1}^{n} (KW_i \times TF_i \times IDF_i) + D$$

其中：KW_i 是该句子中包含的第 i 个关键词在问题分析阶段的权重，TF_i 是该关键词在文档中出现的频率，IDF_i 是该关键词在文档中出现的反频率，D 是指关键字在句子中的分布密度。根据权重进行排序后，还需要对依据问题的类型对候选答案进行重新排序。每类问题对答案都有特殊的要求，所以每类问题都有自己特定的答案抽取规则。第四步中的重新排序就是根据这些规则进行的。对于问时间的问题，答案中就必须含有时间信息。对于问数量的问题答案中必须含有数字信息。否则就不可能是正确答案。这就需要对候选答案进行语义分析，才能识别这些时间信息、数字信息等。经过重新排序后，排在最前面的那个句子就是问答系统返回的最终答案。

5.2 以词或短语作为答案

如果以句子作为答案，处理起来相对简单一些。但是，对于那些问时间地点的问题，其答案就比较简短，而用不着一句话。比如，对于问题:“中华人民共和国是什么时候成立的？”我们可能检索出这样的一句话：“自从*１９４９年１０月１日中华人民共和国成立*以来至１９９４年底止，我国已经同世界上的约１６０个国家建立了外交关系，而且还同更多的国家和地区发展了经济贸易关系和文化往来。”。从这个例子可以看出，我们所要的答案只是这句话中的一小部分，如果我们能把这整句话作为答案都提交给用户的话，显然冗余信息太多。所以有些问答系统希望直接把包含答案的那段话抽取出来。

5.3 以文摘作为答案

对于有些问题，简短的一个短语或者一句话很难说清楚，比如对于问题“9.11 事件的是怎么回事？”。像这种问题，在互联网上有许多相关的报道，如果把这些相关报道都交给用户的话，那么用户将要花很多时间来阅读。如果能把这些相关报道做成一个简短的文摘，让用户只要看文摘就能知道整个事件的前因后果，那么将会为用户带来很大的方便。这就需要用到多文档自动文摘技术。多文档自动文摘模块把信息检索模块检索出来的相关文档做成文摘，再把这个文摘作为答案返回给用户。

6. 评价

问答系统需要一个评价机制来衡量问答系统的性能。首先需要建立一个测试集，这个测试集是人工做出来的问题和答案对的集合。把这个测试集中的问题提交给问答系统，让问答系统自动的给出答案。然后把问答系统自动找出的答案和测试集中的答案，进行人工的对比。如果问答系统给出的答案通过人工的对比基本正确，则可以判断这个答案是正确的，否则可以判断这个答案是错误的。这样就可以计算出问答系统的准确率，公式如下：

准确率＝答对的问题数／问题总数

Trec 会议每年都会提供一个测试集，让参加 Trec 的研究人员来评价自己的问答系统。Trec 允许对每个问题给出 5 个答案。如果一个答案是对的，那么这个问题就得 5 分，如果第二个答案是对的，那么这个问题得 4 分，如果第三个问题是对的，那么这个问题得 3 分，依此类推。把每个问题所得的分加起来就可以得到问答系统所得的总分。总分越高，说明该系统的准确率越高。

7. 结论

早在图灵就提出了著名的图灵测试，如果计算机能通过这个测试，就可以说计算机已经具有了人类的思维。但是，目前计算机科学离这个目标还很遥远。对于目前的问答系统来说，还不能像人类一样能自如地回答用户提出的各种问题。问答系统并不具备任何思维和推论能力，它只能从已有的文档库中搜索相关的答案。所以问答系统所能回答的问题受限于文档库。

如果文档库中没有相关的内容，那么问答系统就不能正确地回答出用户提的问题。而且，目前问答系统的准确率还比较低，在 Trec 会议中，一般的问答系统的准确率都在 30%左右。

虽然问答系统离我们理想的目标还很远，自动问答技术还处于刚刚起步阶段，但是自动问答技术在最近这几年得到了很大的发展。已经有越来越多的相对成熟的问答系统问世。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将会取得重大的突破。

参考文献

[1] Ittycheriah, M. Franz, W-J Zhu, A. Ratnaparkhi. “IBM's Statistical Question Answering System**”.** Proceedings of the night Text Retrieval Conference (TREC-9)

[2] D. Elworthy. “Question Answering Using a Large NLP System”. Proceedings of the night Text Retrieval Conference (TREC-9)

[3] L. Wu, X-j Huang, Y. Guo, B. Liu, Y. Zhang. “FDU at TREC-9: CLIR, Filtering and QA Tasks”. Proceedings of the night Text Retrieval Conference (TREC-9)

[4] R. J. Cooper, S.M. Rüger. “A Simple Question Answering System”. Proceedings of the night Text Retrieval Conference (TREC-9)

[5] C.L.A. Clarke, G.V. Cormack, D.I.E. Kisman, T.R. Lynam. “Question Answering by Passage Selection”. Proceedings of the night Text Retrieval Conference (TREC-9)

[6] S-M Kim, D-H Baek, S-B Kim, H-C Rim. “Question Answering Considering Semantic Categories and Co-Occurrence Density”. Proceedings of the night Text Retrieval Conference (TREC-9)

[7] Richard J Cooper, Stefan M Ruger. “A simple Question Answering System”. Proceedings of the night Text Retrieval Conference (TREC-9)

[8] Ulf Hermjakob. “Parsing and Question Classification for Question Answering”. Proceeding of the workshop on Open-Domain Question Answering at ACL-2001

[9] Eugene Agichtein, Steve Lawrence, Luis Gravano. “Learning Search Engine Specific Query Transformations for Question Answering”. ACM 1-58113-348-1/01/0005.

[10] Soo-Min Kim, ae-Ho Baek, Sang-Beom Kim, Hae-Chang Rim “Question Answering Considering Semantic Categories and Co-occurrence Density”. Proceedings of the night Text Retrieval Conference (TREC-9)