·博士论坛·

问答系统:核心技术、发展趋势

王树西[1,2]
[1](中国科学院计算技术研究所软件研究室,北京 100080)
[2](中国科学院研究生院,北京 100039)
E-mail:wangshuxi@software.ict.ac.cn

摘 要 该文首先给出问答系统的定义,并简要回顾了问答系统的历史;然后对现有各类问答系统进行了介绍,并对其核心技术、评测机制进行了分析;最后对问答系统的发展方向进行了展望。

关键词 问答系统 聊天机器人 知识库 问答式检索系统 TREC QA Track

文章编号 1002-8331-(2005)18-0001-03 **文献标识码** A **中图分类号** TP18

Question Answering System: Core Technology, Application

Wang Shuxi[1,2]
[1](Institute of Computing Technology, the Chinese Academy of Sciences, Beijing 100080)
[2](Graduate School, the Chinese Academy of Sciences, Beijing 100039)

Abstract: This Paper firstly proposes the definition of Question Answering System (QA), and gave a survey on the history of Question Answering System. Then, this Paper introduces current Question Answering Systems, analyzes the technologies which are used by them. At the end, this paper gives the evaluating method of Question Answering System.

Keywords: Question Answering System (QA), ChatBot, Knowledge Base (KB), Question Answering Retrieving System, TREC QA Track

1 引言

问答系统(Question Answering System,QA),又称为人机对话系统(Human-machine conversation,HMC),是指这样一个机器系统:对于用户通过自然语言输入的问句,它能够给出简洁、准确、人性化的回答,这种回答通常是指一小段文本。

在目前的自然语言处理领域,问答系统是一个热门话题,因为它既允许用户用自然语言提问,又能够为用户返回一个确切的答案,而不是一些相关的网页。

该文首先简要回顾了问答系统的历史;然后介绍了现有各类问答系统,对其核心技术进行了分析,对其性能进行了测试;最后提出一种新的问答系统评测机制。

2 图灵测试和早期的问答系统

问答系统的历史,可以追溯到1950年。1950年,著名的英国数学家图灵(A.M.Turing)发表了里程碑式的论文"Computing Machinery and Intelligence"。在文中,图灵提出"机器能思考吗(Can machines think)?",并提出了判定机器能否思考的方法——图灵测试。其方案为:由测试人A与另一房间中的两个对象B和C对话,B和C中有一个是人,另一个是计算机。如果经过一段时间的对话之后,A不能断定B和C中谁是人,谁是计算机,则认为计算机已经具备了人的智能[1]。

由于计算机能否具备和如何具备人的智能,是计算机科学中的一个根本问题,所以历来争论激烈。John Searle提出一个中国人房子问题(Chinese Room),质疑图灵测试对测定计算机智能的意义有多大。John Searle试图通过这个"中国人房子问题"来证明:非生物智能机的概念是不正确的[2]。

但Searle否定不了图灵测试。图灵测试有着重要的意义,它的难度很大,要做好图灵测试,仅仅编出一个"有智能的"程序是不够的。现在有一种误解,认为只要计算机在某些方面做得很好,很有智能,甚至让人看不出来这是计算机做的,就算是通过了图灵测试。这种看法是不对的,实际上,计算机在计算方面早就做的比人强了。图灵测试就其本质来说,测试的是一种特定的智能,称之为言语智能,而言语智能要求计算机必须具备普通人的常识。陆汝钤院士认为,图灵测试永远不可能在图灵定义的层面上真正实现[3]。

一般认为,Jaseph Weizenbaum在1966年左右实现的"Eliza",是第一个问答系统。Eliza扮演一个心理学专家的角色,它采用启发式的心理疗法,通过反问来应对精神病人的提问,诱导病人不停地说话,从而达到对病人进行心理治疗的目的[4]。

1971年左右,Terry Winograd利用MACLISP语言,开发了"SHRDLU"问答系统[5]。SHRDLU包括解析器、英语语法识别器、语义分析器、一般问题解答器等。该系统在一项内容受限的话题——儿童积木话题上,取得了更加惊人的成果。

基金项目:国家973重点基础研究发展规划项目(编号:G1998030507-4,G1998030510)资助

作者简介:王树西,男,博士生,主要研究领域为计算语言学。

回顾历史可以看出，问答系统是伴随着“图灵测试”诞生的，在计算机学界，它称得上历史悠久。在过去的几十年里，许多学者在这个研究方向上做出了诸多有益的探索。

3 现有问答系统

现有问答系统，大致可以分为：聊天机器人、基于知识库的问答系统、问答式检索系统、基于自由文本的问答系统等。

3.1 聊天机器人(ChatBot)

所谓聊天机器人，是指这样的问答系统：它模仿人的语言习惯，给出的答案较为人性化。上文提到的 ELIZA、SHRDLU、PARRY，都是聊天机器人。

3.1.1 典型的聊天机器人[8]

3.1.1.1 ALICE

ALICE 是由 Richard S.Wallac 开发的，遵循 GNU Public License 的标准开放源码，有 300 多人对其发展做出了自己的贡献。ALICE 分别于 2000 年、2001 年、2002 年，三次获得“Loebner Prize”比赛冠军。

ALICE 使用 AIML 表示其知识，而使用 Java 作为引擎对用户输入进行分析，在知识库中寻找最合适的回答并返回给用户。由于 ALICE 定义了丰富的标签，所以它具有各种拟人的智能。

ALICE 背后并没有复杂的算法，事实上，ALICE 有 40,000 多个模板，也是采用了模式匹配的方法来检索最合适的回答。但 ALICE 采用了一种很好的扩充机制，AIML 文件可以进行内联，许多包含特殊领域知识的 AIML 文件可以方便地合并成一个更大的知识库。并且，ALICE 通过对聊天记录进行分析，可以得到尚且没有明确回答的问题，并给出建议的模式。

3.1.1.2 Cyber Ivar

Cyber Ivar 是 Jaczone 公司开发的一个聊天机器人。测试结果表明，Cyber Ivar 响应速度快，在回答 UML、WayPointer 和 Jaczone 之类的问题时，Cyber Ivar 给出的答案相当准确、全面；对于常识性问题，给出的答案也比较贴切。令人惊异的是，对于用户提出的问题 “Who is Maozedong?”，Cyber Ivar 竟然回答：“He was a Chinese communist who reigned from 1949 until his death in 1983.”。在 Cyber Ivar 的知识库中，竟然有这么一条关于“Maozedong(毛泽东)”的知识(虽然这是一条错误的知识)，这说明 Cyber Ivar 的知识库还是挺大的。

3.1.2 小结

所有的这些聊天机器人程序，它们的背后几乎没有复杂的算法，事实上它们几乎全部采用模式匹配的方法，来寻找问题最合适的答案。它们有一个共同的特点，那就是在与用户的交谈过程中，都是基于谈话技巧和程序技巧，而不是根据常识。在它们的对话库中，可以存放多个句型、模板，但几乎没有常识库。

3.2 基于知识库的问答系统

拥有一个或多个知识库，并利用检索、推理等技术，来理解与求解用户问题的问答系统，称为基于知识库的问答系统。和聊天机器人不同的是，这类系统擅长于知识问答，对于不能回答的问题，就老实回答说“不知道”，而非故意转移话题。

3.2.1 知识库

一般说来，知识的数量与质量，是这类问答系统性能是否优越的决定性因素，因此，这类问答系统的主要特征，是拥有一个或者多个知识库，其中存储一个或者多个领域的知识。

以 Feigenbaum 提出来的“知识原则”为理论基础，Doug Lenat 于 1984 年发起了 CYC 研究项目，于 1995 年结束。CYC 耗费了 200 人年的工作量，建立起一个拥有 50 万断言的知识库，并在此基础之上研究了自然语言理解、学习、问题求解等人类智能活动的机理[9]。

在国内，1998 年以来，陆汝钤院士通过研究常识知识本体论及其与人的智能的关系，提出“Agent 和本体是常识知识库的两大支柱”的观点，并且建立了一个大型的常识知识库“Pangu”。由曹存根研究员率领的课题组，正在建立 NKI(国家知识基础设施)海量知识库。

3.2.2 基于知识库的问答系统分类

3.2.2.1 基于本体的问答系统

上述 Pangu、NKI，都是基于本体构建的。但是，当前建立本体大部分还是采用手工方式，建立本体还远远没有成为一种工程性的活动，每个本体开发组都有自己的原则、设计标准和定义的开发阶段。目前，对本体的共享、重用和互操作还难以实现。

3.2.2.2 自然语言界面的专家系统

如果专家系统采用自然语言问答的人机接口方式，那么可以看作是一个问答系统。目前的专家系统自动获取知识能力差，存在知识获取的瓶颈问题。

3.2.2.3 基于受限语言的的数据库查询系统

基于受限语言的数据库查询系统，是指使用受限的自然语言，对数据库进行查询的系统，它的关键步骤是要将中文查询句转换为数据库的 SQL 语句。近年来，国内研制出很多相关系统，如 RCHIQL、NCHIQL、NLCQI 等。他们所用的是类似于语法和模板的技术，由于查询的对象是数据库，所以大部分系统都充分利用了 ER 模型。但基于受限语言，必然带来对查询句的很多限制和语法分析的时间损耗，从而阻碍系统较大规模的应用。

3.2.2.4 基于 FAQ 的问答系统

与产生式、语义网络和框架等传统知识表示相比，FAQ 库中的知识，虽然也是经过人工处理的，但是处理方法却并非基于上述符号处理机制，而是采用自问自答的方式，知识描述的颗粒(粒度)很大、很粗，属于半结构化文本。现有的很多问答系统，都是基于 FAQ 知识库开发的，特别是一些企业，为了快速回答客户的一些业务咨询，往往基于 FAQ 知识库，开发一个关于本企业产品信息的在线问答系统。

3.2.3 小结

这类系统的优点是，对于用户提出的许多问题，回答准确，甚至可以进行一定程度的推理计算，并且由于是基于知识库的，所以系统具有良好的可扩展性。但是，如果用户的问题超出系统的知识库范围，系统性能很快下降为零。从知识库的角度分析其弱点的来源，可以发现系统的知识库规模不足、知识获取困难，存在知识库的瓶颈问题。

3.3 问答式检索系统

根据以自然语言方式提交的用户查询，从系统文档集合或 WWW 中，检索出相关文本或网页，并将其返还给用户，这种系统称为问答式检索系统，也称问答式搜索引擎、智能搜索引擎。问答式检索系统需要正确理解自然语言形式的用户查询，充分领会用户的查询意图，并检索出与用户需求最相关的文本或者

网页。

3.3.1 现有典型的问答式检索系统

3.3.1.1 Start(http://www.ai.mit.edu/projects/infolab/start.html)

Start 是 MIT 人工智能实验室开发的，是世界上第一个基于 Web 的问答系统，自 1993 年 12 月开始，它持续在线运行至今。现在，Start 能够回答数百万的多类英语问题，包括“place”类(城市、国家、湖泊、天气、地图、人口统计学、政治和经济等)、电影类(片名、演员和导演等)、人物类(出生日期、传记等)、词典定义类等。

Start 包含两个知识库(“START KB”、“Internet Public Library”)以及一个搜索引擎。如果通过这两个知识库就能回答用户的问题，那么系统立刻给出准确的答案；否则，首先解析用户输入，得到其中的关键词，然后利用这些关键词，通过系统自身的搜索引擎进行检索，最后将得到所有相关文本，以链接的形式提交给用户，供用户点击并在打开的网页中自行寻找答案。测试结果表明，Start 是一个优秀的问答系统。

3.3.1.2 Encarta(http://encarta.msn.com/)

Encarta 是由 Microsoft 公司开发的。作为一个在线百科全书式的问答式搜索引擎，也提供了多语种的支持。测试表明，Encarta 回答问题较快，答案也比较准确。

3.3.2 核心技术分析

一般而言，问答式检索系统主要需要两种技术：用户查询处理技术、信息检索(IR)技术。由于信息检索(IR)技术目前已经比较成熟，所以不再赘述，这里主要讨论问答式检索系统所面临的第一个问题：如何正确理解用户用自然语言提出的查询。

第一种方法，对用户查询进行浅层分析，识别出其中的关键词，然后利用查询扩展技术，借助 HowNet、WordNet 等语义词典，将关键词的同义词、近义词，一并提交给后继的检索系统。这种方法，虽然允许用户使用自然语言查询，但并没有充分利用用户查询的信息，特别是语义信息，其能力等同于一般的词表法。许多号称自然语言查询的系统其实都是这么做的。

第二种方法，就是使用问句模板。如果系统面向的领域非常狭窄，那么这种方法的好处是显而易见的，数量很少的问句模板就可以覆盖绝大多数的用户提问方式。但如果系统面向的领域非常广阔，甚至是开放领域，那么仍然利用这种方法的话，所需模板库和模板答案的数量就非常多，由于模板库和模板答案一般是人工产生和维护的，所以工作量非常大。

3.4 基于自由文本的问答系统

所谓自由文本，又称原始文本、非结构化文本，是指未经人工处理的原始文档、网页等。现有基于自由文本的问答系统，一般采用单通道管状的体系结构：用户查询处理→自由文本检索→答案抽取。现有基于自由文本的问答系统，很多都是参加 TREC QA Track 比赛的系统。

3.4.1 典型的系统

3.4.1.1 Textract[10]

Textract 是 Cymfony 公司的产品 IE，Textract 参加 TREC-8 QA Track 比赛，获得较好的成绩，回答问题的正确率达到 66%。并且研究结果表明：IE 能为 QA 提供坚实的基础。

3.4.1.2 Webclopedia[11]

Webclopedia 基于的理论是：用户问题肯定属于某一个自然语言类型，基于用户问题所期待答案的类型，得到用户的答案。Webclopedia 包括如下几个模块：CONTEX 问题分析器；查询构成模块；MG 信息检索器；3 个文本分离器；BBN 的命名实体识别器 IdentiFinder。在 2000 年的 TREC QA Track 比赛中，Webclopedia 获得第二名；在 2001 年的 TREC QA Track 比赛中，Webclopedia 获得第四名。

3.4.2 小结

基于自由文本的问答系统，涉及到 IR、IE、模式推理技术，和几乎所有的 NLP 技术，是这些技术的集大成者。由于不需要建立大规模知识库，而是基于自由文本进行知识问答，所以节省了大量的人力物力；由于系统返还给用户的，是用户问题的具体答案，所以方便了用户而受到更多的欢迎。应该说，基于自由文本的问答系统，特别是基于 Web 的开放领域问答系统，代表着问答系统的发展方向。

4 问答系统的评测

如何客观而科学地评测问答系统的性能，是该研究领域一个很重要的问题。应该说，目前的 QA 测评标准，并不太成熟，就连 TREC QA Track 的评测标准，也有相当的主观成分在里面。

综合现有问答系统的评测指标，提出问答系统的评测指标如下：

(1)系统给出的答案应该是准确的。

(2)系统响应时间必须能够让人接受，响应时间越小越好，最好是实时响应。

(3)系统给出的答案应该是全面的。

(4)系统给出的答案应该是语句流畅、简短扼要，而非生硬拗口、长篇大论。

(5)对于每个问题，标出其难度系数，难度系数正比于系统得分。

5 结论

该文回顾了问答系统的历史，综述了现有各类问答系统，并分析了其核心技术，最后提出了问答系统的评测指标。

(收稿日期：2005 年 5 月)

参考文献

1.A M Turing.Computing Machinery and Intelligence[J].Mind,1950;59(236):433~460

2.John R Searle.Minds,brains,and programs[J].Behavioral and Brain Sciences,1980;3:417~424

3.陆汝钤.世纪之交的知识工程与知识科学[M].北京：清华大学出版社，2001:317~500

4.Weizenbaum,Joseph.ELIZA-A Computer Program for the Study of Natural Language Communication between Man and Machine[J].Communications of the ACM,1966;9(1):36~45

5.http://hci.stanford.edu/~winograd/shrdlu/

6.ROBERT F SIMMONS.Natural Language Question-Answering Systems:1969[J].Communications of the ACM,1970;13(1):15~30

7.Nils J Nilson 著.郑扣根，庄越挺译.潘云鹤校.Artificial Intelligence A new Synthesis[M].北京：机械工业出版社，2000

8.Rohini Srihari,Wei Li.Information Extraction Supported Question Answering.http://trec.nist.gov/pubs/trec8/papers/cymfony.pdf

9.Webclopedia.http://www.isi.edu/natural-language/projects/webclopedia