DOI:10.16452/j.cnki.sdkjzk.2007.04.022
第26卷 第4期 Vol.26 No.4
2007年10月 Oct.2007

Journal of Shandong University of Science and Technology
Natural Science

自动问答系统研究综述

刘 里,曾庆田
(山东科技大学 信息科学与工程学院,山东 青岛 266510)

摘 要:自动问答系统是自然语言处理领域中一个非常热门的研究方向,它综合运用了多种自然语言处理技术。本文综述了国内外自动问答技术的发展现状,对系统三个主要组成部分:问题分析、信息检索和答案抽取进行了深入的分析,通过比较,得出了自动问答系统运用各类技术之间的优势与不足。在此基础上,提出了自动问答系统的发展方向。
关键词:自动问答系统;问题处理;信息检索;答案抽取
中图分类号:TP391.3 文献标志码:A 文章编号:1672-3767(2007)04-0073-04

An Overview of Automatic Question and Answering System

LIU Li, ZENG Qing-tian
(College of Information Science and Technology, SUST, Qingdao, Shandong 266510, China)

Abstract: The automatic question and answering system (QA) is a hot research field in natural language processing (NLP), which includes many kinds of NLP technologies. This paper introduces the current development of QA at home and abroad and analyzes the three parts of QA: question analysis, information retrieval and answer extraction. By comparing the advantages and weaknesses of various technologies, this paper prompts the development direction of QA.
Key words: automatic question and answering system; question processing; information retrieval; answer extraction

互联网的普及使人们对网上信息的应用需求不断提高。目前人们主要利用搜索引擎和分类目录来进行网络信息查询,通常这种查询主要基于关键字匹配进行。这种用词汇信息表述问题的不足是缺少上下文背景信息,检索得到的结果往往是数以万计的相关网页,而这些网页中能够满足用户需求的信息却只有一小部分。用户必须逐个阅读这些网页去寻找自己真正感兴趣的信息,是否存在没有被检索出来的相关网页也无从知晓。面对庞大的信息量,如何以最快的速度准确而详尽地找到用户感兴趣的信息已成为信息时代的一个重要研究课题。用户需要一个更加高效和人性化的搜索引擎,自动问答系统应运而生。

自动问答系统(Automatic Question and Answering System),简称问答系统(QA),是指接受用户以自然语言形式描述的提问(例如:世界上最高的山是哪座?),并从大量的异构数据中查找出能回答该提问的准确、简洁答案(例如:喜马拉雅山)的信息检索系统。因此,问答系统和根据关键词检索并返回相关文档集合的传统搜索引擎有着根本的区别。问答系统的目标是精确回答用户用自然语言提出的问题。与传统搜索引擎相比,问答系统更强调精确性。回顾问答系统研究历史,总结问答技术研究现状,将有助于推动问答系统的发展。

1 国内外研究现状

问答系统的相关技术及其产品引起了国内外许多科研机构和公司极大的兴趣。自从1999年文本信息检索会议(Text Retrieval Conference, TREC)第一次把Automatic Question Answering Track设为评测专项以来,QA Track逐渐成为最受关注的TREC评测项目之一。

收稿日期:2006-12-17
基金项目:国家自然科学基金(60603090);山东省优秀中青年科学家奖励基金(2006BSB01171);山东省泰山学者专项基金
作者简介:刘 里(1983—),男,山东德州人,硕士研究生,主要从事面向Web的信息处理技术研究.

国外开发的相对成熟的问答系统有麻省理工大学人工智能实验室的 Start[1]、密歇根大学的 AnswerBus[2]、IBM 基于统计的问答系统[3]等。Start 是第一个基于 Web 的自动问答系统，其特点是向用户提供准确的信息，而不是提供一堆相关信息。Start 采用基于知识库和信息检索的混合模式，如果用户查询在它的知识库中可以找到，则直接反馈；如果没有，则通过搜索引擎检索并处理后反馈给用户。AnswerBus 是一个面向开放领域的问答系统，它接受自然语言的提问，从 WEB 中提取问题可能的答案(一个或多个)，其特点是能支持包括英语、法语、德语、西班牙、意大利语和葡萄牙语在内的多种语言提问方式。

国内也有不少大学和研究所正在进行问答系统的研究。复旦大学和中科院都参加了 QA Track 的竞赛[4]，哈工大也在这方面做了一些研究[5]。中科院计算所正在进行的大规模知识处理科研项目 National Knowledge Infrastructure(简称 NKI)中的一个具体应用就是 NKI 知识问答系统——HKI[6]。HKI 以 NKI 知识库为基础，向用户提供各个领域的知识服务，其特点是向用户提供准确的信息，支持自由的提问方式。相对英文问答系统来说，中文问答系统起步较晚，不够成熟，这和中文的语法、语义复杂性等多种因素有关。

2 问答系统的原理与关键技术

2.1 问答系统的基本原理

开放领域问答系统是包含知识存储、知识表示、信息抽取、自然语言处理等多方面研究技术的综合性应用系统。其体系结构一般包括三个主要部分：问题处理、信息检索和答案抽取[7-8]。其基本原理如图 1 所示：

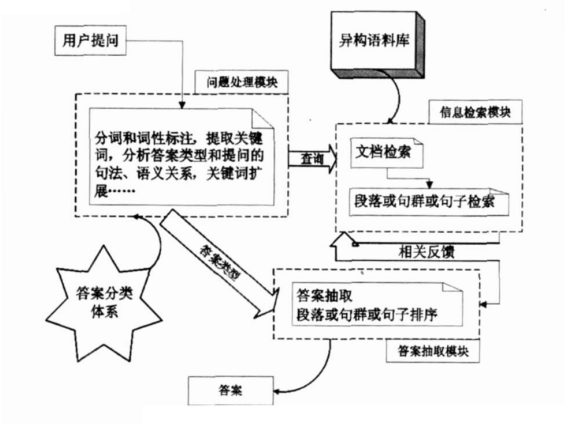

图 1 问答系统基本架构图

Fig. 1 Basic framework of QA

问题处理部分是对用户用自然语言提出的问题进行预处理，包括词法、句法、语义等方面的分析，得到用户查询的关键词、查询句的关注焦点和用户问题所属类型。

信息检索部分是通过传统信息检索技术获得答案可能所在的文档，并对文档进行排序。

答案抽取部分对信息检索得到的候选文档进行词法、句法、语义等方面的分析，并根据查询问题所属类别，抽取出答案，返回给用户。

无论采用何种分类方式，问答系统都是采用相似的体系结构，其中涉及几种关键的技术。本文将对这几种常用技术做系统的综述。

2.2 分词及词性标注

中文文本的分词与词性标注是中文信息处理中特有的基础性问题[9]。中文信息处理要以“词”为基础，但汉语书面语不像西方文字那样通过天然的切分标志——空格切分。让计算机将等间距排列的汉字字串序列按词切分开，并打上切分标志，这就是中文文本自动分词问题。

词性标注是在给定句子中判定每个词的语法范畴，确定词性并加以标注的过程，是对切分所得的词进行分析、运算、确定词在上下文中合适的词性并加以标注的过程[10]。关于词性标注规则可参考北京大学语言学研究所《汉语文本词性标注标记集》[11]。

现在正在研究的中文问答系统普遍采用中国科学院计算技术研究所的汉语词法分析系统(Institute of Computing Technology, Chinese Lexical Analysis System，简称 ICTCLAS)[12]。ICTCLAS 的最主要特点在于采用了层叠隐马尔可夫模型(Hierarchical Hidden Markov Model)，将汉语词法分析的主要问题(汉语分词、未定义词识别和词性标注)都统一到了一个完整的理论框架中，切分的词语包含词语本身和它的词性。实践表明，选用 ICTCLAS 作为词法分析组件，的确达到了时间效率和词法处理准确率之间很好的平衡，为问答系统的后续处理提供了良好的基础。

2.3 关键词抽取与扩展

关键词抽取常用和经典的方法是统计法，它通过确定候选词的权重，从中筛选出权重较大者作为最终的关键词[13]。因此，候选词权重的确定就成为文献关键词抽取的核心。候选词的权重由它反映文献主题的重要性决定，能够较好反映文献主题的词语将被赋予较大的权值。图 2 是一个正文关键词抽取方法的例子[14]。

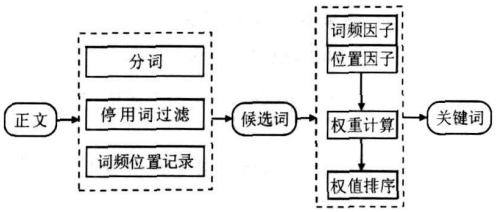

图 2 关键词抽取流程图

Fig. 2 Flowchart of keywords extraction

关键词抽取的基本流程包括:

1) 将正文进行分词处理。

2) 过滤掉停用词。一般将停用词确定为所有虚词以及标点符号。

3)记录词语的位置信息。这样,当软件逐词扫描统计词频时,就可以将每个词的位置信息加入到词频统计表中。

4) 根据统计的词频和位置信息,分别计算词语的词频因子和位置因子。

5) 利用具体的词语权重函数计算出词语的权重。词语的权值确定以后,进行排序,取权值较大者作为最后的抽取结果。

另外,关键词抽取过程中还应进行命名实体的识别,以便对句子的主语或宾语进行判断[15]。

关键词的扩展也是问答系统中问题处理的必要步骤,其目的是为了提高检索系统的召回率。一些面向英语的问答系统利用 WorldNet 中语义相关信息进行扩展;中文问答系统通常利用《同义词词林》对关键词进行扩展。对于一些专业领域的关键词作扩展时,可能还要建立相应的专业领域词汇知识词典,以便于某些专业术语的扩展。对扩展的关键词应该赋予较低的权值,以防止扩展的关键词造成检索主题漂移。

例如,对于问题"山东科技大学在哪里?",抽取关键词"山东科技大学"。如果考虑到此问题的答案句式可能是"山东科技大学位于……"、"山东科技大学地处……"或者"山东科技大学地址是……"等,把查询串扩展为"山东科技大学 AND 位于(OR 地处 OR 地址)",将有助于搜索到潜在的答案。

2.4 基于模式的句子相似度计算

由于问答系统的最终目的是对用户的提问给出准确答案,而给出一系列的候选答案句子集(或段落、文摘,甚至整个文档)的排序仅仅是在无法明确具体答案时一种过渡的策略[16]。基于这个原因,以获得准确答案为目标的模式匹配方式在英文问答系统得到重视和应用[17-18]。

基于模式的句子之间相似程度的衡量,一般应该综合考虑二句的组成词汇语义信息与整体框架结构信息,也就是通常所说的句子相似度等于语法相似度与语义相似度之和[19]。实际进行句子相似度衡量时,在句子的组成词汇语义方面需考虑一些基本的特征词汇语义;在句子的整体框架结构方面需考虑一些主要的骨架成分。而介于规则和例句之间的一种抽象的句子结构形式——模式,既反映了一个句子的主要骨架成分,又表达了一个句子的基本特征语义[20]。因此,在定义和计算句子相似度时,可以直接根据二句的模式对其进行相似判断,即把二句的相似判断等价地转化为二句模式的相似判断。

基本算法思想是,首先将汉语常用句的谓词中心词等价分类,并将例句库扩展为例句模式库。在比较二句相似时,先比较二句模式的谓语中心词是否位于同一等价类,然后再比较二句模式的其它成份是否一一对应。仅当谓语中心词等价且其它成分均一一对应时,才认二句模式相似,进而判定二句亦相似。可见,定义一个高效的例句模式库,可以显著提高相似度的计算速度。

3 当前问答系统的研究热点与难点

3.1 问题理解

问题内容的正确理解是问题处理过程乃至整个问答系统处理过程的前提和基础,问题理解的效果对整个问答系统的性能具有至关重要的影响,错误的理解必然导致错误的操作。只有首先确定用户问题中所要表达的正确含义以后,才能够对识别问题的问点进行问题分类,生成用于信息检索的查询表达式,确立答案抽取规则和约束条件。

用相同的形式表示不同的含义是自然语言区别于形式语言的一大特点,也是导致自然语言中歧义现象非常普遍的原因。自然语言歧义给自然语言语句的准确理解带来相当大的难度。汉语句法成分的构成十分灵活,而且缺乏形态变化,这使得汉语的句法结构分析十分困难,而且句法分析得到的合法的句法结构不一定有正确的逻辑意义[21]。语义知识作为当前的研究热点,可以用来对句法结构进行语义检验,以排除意义不正确的句法结构。在语义分析中,语义知识可以帮助我们获得语言片段各成分间的语义关系,更好地理解用户的目的。

3.2 问题相关性推荐

问题相关性推荐是问答系统需要扩充的另一个方面,这一点在有些问答系统中已经实现。比如

MIT的START问答系统,当用户提问"How many people live in China?",START将回答"1,313,973,713(2006-12-16)",同时产生一个链接,用户点击时就能看到世界上各个国家的人口列表。

问题相关性推荐可以通过对问题进行语义分析,对用户的"问点"进行扩充,从而得到相关性的问题。对于基于模式的问答系统或者是拥有知识库的问答系统来说,这一点更容易实现。将问题替换为相同模式的相似度较高的问题,或者在问题集中选择相似问题就可以实现问答系统的问题相似度推荐。

然而,如何计算两个问题之间的语义相关性是一个非常复杂的问题,至今仍然没有很好的解决方法。

3.3 用户自适应性答案

随着问答系统的发展,能够满足用户问题的答案越来越多。虽然现在有的问答系统可以对检索结果进行打分并按照得分高低呈现给用户,但它根本没有考虑用户个人的兴趣爱好,不同用户使用同样的问题检索出的答案是相同的,这样做并不能完全满足用户的要求。

在现有的问答系统中,如何表示用户的兴趣模型,并且根据用户的浏览信息正确建立用户的兴趣模型,是一个需要解决的问题。一旦获取用户的兴趣模型,问答系统就可以根据用户的知识背景、兴趣爱好等特征组织个性化的答案反馈给用户,不同的用户便会得到符合自己兴趣特点的问题答案。

4 结束语

目前,问答系统的准确率还比较低,在TREC会议中,一般问答系统的准确率都在30%左右。虽然当前问答系统仍处在起步阶段,自动问答技术还不能满足需要,但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。

参考文献:

[1] Boris Katz, Gregory Marton, Gary Borchardt, et al. The START Natural Language Question Answering System [EB/OL]. [2006-12-16]. http://start.csail.mit.edu.

[2] Zhiping Zheng. AnswerBus Question Answering System [EB/OL]. [2006-12-16]. http://answerbus.coli.uni-saarland.de/index.shtml.

[3] A Ittycheriah, S Roukos. IBM's Statistical Question Answering System[C]//Proceedings of the TREC-11 Conference. Gaithersburg: NIST Special Publication, 2002: 394-401.

[4] 黄萱菁. 复旦大学媒体计算与Web智能实验室信息检索和自然语言处理(IRNLP)组研究介绍[EB/OL]. [2006-12-16]. http://www.yssnlp.com/yssnlp2004/report/Fudan-Huangxuanjing.pdf/.

[5] 盖节. 基于Ontology的自动问答系统关键技术研究[D]. 南京:南京大学, 2004.

[6] 曹存根. NKI-21世纪的科技热点[J]. 计算机世界报, 1998, 5(2): 1-3.

[7] 吴友政, 赵军, 段湘煜, 等. 问答式检索技术及评测研究综述[J]. 中文信息学报, 2005, 19(3): 1-13.

[8] 戴谢宁. 基于Web的中文自动问答系统的研究[D]. 广州:华南理工大学, 2005.

[9] 刘迁, 贾惠波. 中文信息处理中自动分词技术的研究与展望[J]. 计算机工程与应用, 2006, 42(3): 175-182.

[10] 陈晓文. 自动词性标注方法的比较[J]. 温州大学学报, 2006, 19(1): 53-57.

[11] 俞士汶. 现代汉语语料库加工—词语切分与词性标注[M]. 北京: 北京大学计算语言学研究所, 1999.

[12] 张华平. 中科院计算所汉语词法分析系统[EB/OL]. [2006-12-16]. http://www.nlp.org.cn/project/project.php?proj_id=6.

[13] 谭伟. 面向网络的中文问答系统相关技术的研究与系统初步实现[D]. 北京:清华大学, 2005.

[14] 郑家恒, 卢娇丽. 关键词抽取方法的研究[J]. 计算机工程, 2005, 31(18): 194-196.

[15] 李保利, 陈玉忠, 俞士汶. 信息抽取研究综述[J]. 计算机工程与应用, 2003, 39(10): 1-5.

[16] 杨晓明, 罗振声. 模式匹配在中文问答系统中的应用研究[J]. 科学技术与工程, 2006, 3(6): 319-322.

[17] Wu Min, Zheng Xiaoyu, Duan Michelle, et al. Question answering by pattern matching, Web proofing semantic form proofing[C]//Proceedings of the TREC-12 Conference. Maryland: NIST Special Publication, 2003: 165-169.

[18] Dell Zhang, Wee Sun Lee. Web based Pattern Mining and Matching Approach to Question Answering[C]//Proceedings of the TREC-11 Conference. Gaithersburg: NIST Special Publication, 2002: 97-101.

[19] 管小艳, 郑家恒. 一种改进的句子相似度计算方法[C]//第二届全国信息检索与内容安全学术会议论文集. 北京:清华大学出版社, 2005: 401-407.

[20] 杨思春, 程节华, 陈家骏, 等. 一种基于模式的汉语句子相似度计算方法[J]. 微型机与应用, 2001, 20(8): 52-53.

[21] 王鹏, 戴新宇, 陈家骏, 等. 基于规则的汉语句法分析方法研究[J]. 计算机工程与应用, 2003, 39(29): 63-66, 169.