问答式检索技术及评测研究综述*

吴友政　赵军　段湘煜　徐波

（中国科学院自动化研究所 模式识别国家重点实验室，北京 100080）

摘　要：问答式检索系统（简称问答系统）是集自然语言处理技术和信息检索技术于一身的新一代搜索引擎。它的出现旨在提供更有力的信息获取工具，以应对信息爆炸带来的严重挑战。经过这几年的发展，问答系统已经成为自然语言处理领域和信息检索领域的一个重要分支和新兴的研究热点，其 “通过系统化、大规模地定量评测推动研究向前发展” 的发展轨迹，以及某些成功的启示，如基于字符表层的文本分析技术（模板技术）的有效性，快速、浅层自然语言处理技术的必要性，都极大地推动了自然语言处理研究的发展，促进了 NLP 研究与应用的紧密结合。回顾问答系统研究的历史，总结问答技术的研究现状，将有助于这方面工作向前发展。

关键词：问答系统；问答评测；信息抽取；信息检索；自然语言处理

中图分类号：TP391　　**文献标示码：**A

Research on Question Answering & Evaluation: A Survey

WU You-zheng　ZHAO Jun　DUAN Xiang-yu　XU Bo

(National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing 100080)

Abstract: Question Answering (QA) is the next generation of search engine which is related to natural language processing, information retrieval and etc. QA aims at providing more powerful information access tools to help users overcome the problem of information overloading. In the last decade, QA has become an important subfield of NLP and IR. Its development track, i.e. accelerating research via systematical and large scale evaluation, and some successful experiences, such as the effectiveness of partial-parsing techniques based on character surface (Pattern Matching) and the importance of fast NLP tools, have made it a great and most important impetus to the research of NLP. Moreover, QA has built a more effective connection between NLP research and NLP application. It will be helpful to review the history and investigate state of the art of QA.

Key words: Question Answering; Evaluation of QA; Information Retrieval; Information Extraction; Natural Language Processing

一、引言

互联网的迅猛发展和广泛普及，使人们可以方便地从网络上获得信息；但网络信息的爆炸性增长，又使人们准确、快速地获得有价值信息的难度大大增加。英国莫里（MORI）调查公司的民意调查[I]结果显示，只有 18%的用户表示总能在网上搜索到需要的信息，68%的用户说他们对搜索引擎很失望，28%表示还可以，其余 5%为不知道。

从这些调查数据中不难看出，尽管一些优秀的搜索服务提供商（Google、Yahoo、百度等）在研发搜索技术方面已经花费了大量的时间和精力，但目前的搜索引擎仍然存在不少的局限性，比如信息丢失、返回信息太多、信息无关等。这使得网络用户对于现有的搜索技术仍然不满，期盼更完美的搜索技术的出现。

为了克服传统搜索引擎的弊端，研究人员正尝试探索一种更高效、更人性化的搜索引擎技术-问答系统（Question Answering）。

基金项目：国家自然科学基金项目（60372016, 60272041），北京市自然科学基金项目（4052027）.

作者简介：**吴友政**(1976－)，男，博士生,主要研究领域为自然语言处理,信息检索,自动问答技术等.

[I] http://www.sowang.com/9238/meiri/7.htm

问答系统是指系统接受用户以自然语言形式描述的提问（例如：世界上最大的宫殿是什么宫殿？），并从大量的异构数据中查找出能回答该提问的准确、简洁的答案（例如：“紫禁城”或者“故宫”）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。可以说，问答系统能够提供用户真正有用、精确的信息，它将是下一代的搜索引擎的理想选择之一。

经过这几年的发展，问答系统已经成为自然语言处理领域和信息检索领域的一个重要分支和新兴的研究热点，其 “通过系统化、大规模地定量评测推动研究向前发展”的发展轨迹，以及某些成功启示，如基于字符表层的文本分析技术（例如模板技术）的有效性，快速、浅层自然语言处理技术的必要性，都极大地推动了自然语言处理研究的发展，促进了 NLP 研究与应用的紧密结合。回顾问答系统研究的历史，总结问答技术的研究现状，将有助于这方面工作向前发展。

本文是这样组织的：第二、三节分别介绍问答系统的研究现状和分类；第四节是国际上三个典型的问答技术评测平台，即英语问答评测平台 TREC、日语问答评测平 NICIR、多语言问答评测平台 CLEF 以及本研究小组初步建立的汉语问答系统评测平台 EPCQA；第五、六节分别综述了问答系统的基本原理和三种有代表性的问答技术，即基于信息检索和信息抽取的问答技术、基于模式匹配的问答技术和基于自然语言处理的问答技术；第七节总结了可以应用于问答系统的各种自然语言处理技术，这包括命名实体识别技术、短语和依存结构分析技术、Paraphrase 技术、词汇链和逻辑形式转换等；论文的最后分析了三种代表性问答技术的优缺点及问答系统未来的发展。

二、问答系统研究现状

自 1999 年文本检索会议（Text Retrieval Conference，简称 TREC）引入问答系统评测专项（Question Answering Track，简称 QA Track）后，人们对基于自然语言的问答系统再次产生了浓厚的兴趣，在近些年的 TREC 比赛中，QA Track 是最受关注的评测项目之一。

从第一个英文问答系统STUDENT[40]，到早期著名的LUNAR系统[42]， MURAX系统[22]，DARPA支持的HPKB工程[32]和现今由美国NIST组织的TREC QA Track [9, 10, 11, 12, 13]，英文问答技术已经获得长足的发展，研究领域也从初期的限定领域（Moon Rock, Crisis Management）拓展到如今的开放领域；研究对象从当初的固定语料库拓展到互联网。目前，比较成功的英文问答式检索系统有Ask Jeeves[II]，AnswerBus[III]和START[IV]等等。

近年来，国内从事问答系统的研究机构也在不断地增加。在往届的TREC QA Track评测中，复旦大学[26, 27]、中科院计算所[3][17]都获得了良好的成绩。此外，中科院计算所[V]、哈尔滨工业大学[47]、复旦大学[44]等[37, 43, 46, 48]在汉语问答技术的研究中也作了有益的探索。但是，和国际研究相比，国内从事问答系统尤其是汉语自动问答技术研究的科研机构还是很少，而且基本没有成型的汉语自动问答系统问世。一个很重要的原因是：缺乏一个公认的，相对成熟的汉语问答系统评测平台。

三、问答系统分类

问答系统根据用户自然语言的提问，从大量异构数据中查找提问的答案。根据不同的分类标准，我们把问答系统分为如图 3.1 所示的体系。

[II] http://www.ask.com
[III] http://www.answerbus.com/about/index.shtml
[IV] http://www.ai.mit. edu/projects/infolab
[V] http://www.nki.net. cn/

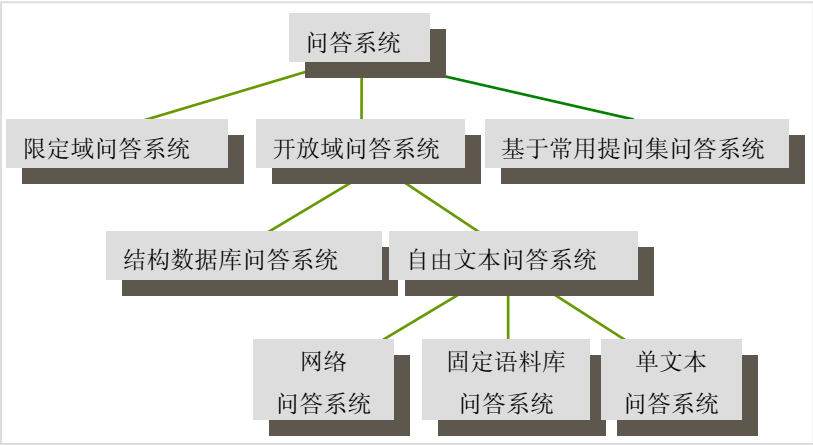

图3.1 问答系统分类体系

在不同的应用环境，我们需要设计不同类型的问答系统。其中，基于大规模真实文本固定语料库问答系统是从预先建立的大规模真实文本语料库中查找答案的，类似于 TREC QA Track。这类问答系统的缺点是无法涵盖用户所有类型提问的答案，却能够提供一个优良的算法评测平台，适合我们对不同问答技术的比较研究。基于网络的问答系统是从互联网中查找提问的答案。虽然它是在真实环境下研发的问答技术，却不适合评价各种问答技术的优劣。单文本问答系统，也可以称之为阅读理解式的问答系统，它是从一篇给定的文章中查找答案。系统在“阅读”完一篇文章后，根据对文章的“理解”给出用户提问的答案。这种系统非常类似于我们在学习英语时做的阅读理解。基于结构数据库的问答系统是从一个预先建立的结构化的数据库中查找提问的答案。所以该系统可以具有较强推理能力，但建立大规模的结构知识库是一个非常困难的问题。另外，基于常用提问集的问答系统是在已有的“提问—答案”对集合中找到与用户提问相匹配的提问，并将其对应的答案返回给用户。而限定领域问答系统的用户提出只能限定在某一特定领域。

四、问答系统评测

所有问答技术的研究者在设计、研发问答技术的时候都会遇到同样一个问题：如何比较不同问答技术的优劣？问答系统评测平台即是完成这个任务，它对问答技术的发展有着很大的推动作用。目前，对问答系统进行评测的国际会议有：英语问答评测平台TREC QA Track[VI]、日语问答评测平台NICIR[VII]和多语言问答评测平台CLEF[VIII]。

4.1 TREC-英语问答评测平台

关于问答系统评测机制，不得不提到 TREC QA Track。从 1999 年 TREC-8 到 2004 年 TREC-13，QA Track 已经成功举办了 6 届。TREC QA Track 每年的评测任务和评测指标都在不断地变化，大致包括分为以下几类：

- Factoid 任务测试系统对**基于事实、有简短答案的提问**的处理能力。例如，Where is Belize located？而那些需要总结、概括的提问不在测试之列。例如，如何办理出国手续？如何赚钱？等。
- List 任务要求系统列出满足条件的几个答案。在 TREC2003 之前，任务要求被测试系统给出不少于给定数目的实例，如：Name 22 cities that have a subway system。TREC2003 要求系统要给出满足条件的尽可能多实例，如：List the names of chewing gums。
- Definition 任务要求系统给出某个概念，术语或现象的定义、解释。例如：What is Iqra?等。
- Context 任务测试系统对相关联的系列提问的处理能力，即对提问 i 的回答还依赖对提问 j（i>j）的理解。例如：a、佛罗伦萨的哪家博物馆在 1993 年遭到炸弹的摧毁？b、这次爆炸发生在那一天？c、有多少人在这次爆炸中受伤？

[VI] http://trec.nist.gov
[VII] http://research.nii.ac.jp/ntcir/workshop/
[VIII] http://clef.iei.pi.cnr.it/

- Passage 任务是 TREC2003 提出的新任务。和其他任务不同的是，它对答案的要求偏低，不需要系统给出精确答案，只要给出包含答案的一个字符序列(a small chunk of text that contains an answer)。
- Other 任务是 TREC2004 才定义的任务。TREC2004 的测试集是包括 65 个目标（Target），每个 Target 由数个 Factoid 问题，0~2 个 List 问题和一个 Other 问题组成。其中，Other 问题的返回答案应该是一个非空的、无序的、无限定的关于这个 Target 的描述，且不包括 Factoid、List 问题已经回答的内容。

TREC QA Track 的评测指标主要有平均排序倒数（Mean Reciprocal Rank，简称 MRR）、准确率（Accuracy）、CWS（Confidence Weighted Score）等，计算公式分别如(4.1)~(4.2)。

$$MRR = \sum_{i=1}^{N} 1 \Big/ \text{标准答案在系统给出的排序结果中的位置} \tag{4.1}$$

如果标准答案存在于系统给出的排序结果中的多个位置，以排序最高的位置计算；如果标准答案不在系统给出的排序结果中，本题得 0 分。

$$CWS = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{前}i\text{个提问中被正确回答的提问数}}{i} \tag{4.2}$$

公式(4.1) ~(4.2)中的 N 表示测试集中的提问个数。

4.2 NICIR-日语问答评测平台

日语问答评测平台 Question Answering Challenge（QAC）是从 2002 年开始的，每两年举办一届。NICIR 定义了三个子任务[23]。

[任务 1] 每个提问，系统给出五个按概率大小排列的答案列表；采用 MRR 打分标准；系统必须给出支持每个答案的文档。

[任务 2] 每个提问，系统只能给出一个答案；如果某个提问在语料中有几个答案，系统须给出所有答案，且必须给出支持每个答案的文档。

[任务 3] 这个任务评测系统对关联提问的处理能力；关联提问是指提问之间可能有互指关系、省略等，类似 TREC 中的 Context Task；系统必须给出支持每个答案的文档。

2002 和 2003 年日文问答系统的评测情况基本如表 4.1 所示。

		任务 1	任务 2	任务 3	语料库
QAC1 2002	测试提问数	200	200	40	Mainichi Newspaper（1998~1999）
	最佳系统性能	0.61	0.36	0.17	
QAC2 2004	测试提问数	200	200	200	Mainichi Newspaper Yomiuri Newspaper（1998~1999）
	最佳系统性能	0.607	0.321	约 0.21	
评测指标		MRR，MF，MMF			

表 4.1 各届日文问答评测情况

4.3 CLEF-多语言问答系统评测平台

由IST Programme of the European Union资助的Cross Language Evaluation Forum (CLEF)在 2003 年设立第一届多语言问答系统评测（Multilingual Question Answering）项目，并计划每年举办一次。CLEF QA Track定义了单语和多语两个任务，具体情况如表 4.2。

[单语言任务] 指输入提问是某种语言，输出的答案就是某种语言。

[多语言任务] 指输入提问可以是任何一种语言，但是系统给出的答案必须是英语文本。

		CLEF2003	CLEF2004
单语言任务	语种	Dutch Italian Spanish	Dutch French German Italian Spanish
	最佳系统性能	0.422	0.455
多语言任务	语言	Dutch, French, German, Italian, Spanish	Dutch, French, German, Italian, Spanish, others
	最佳系统性能	0.393（bilingual Italian）	0.35
评测指标		MRR	CWS ， Accuracy

表 4.2 CLEF QA Track 每届基本情况

4.4 EPCQA-汉语问答系统评测平台

目前，汉语问答技术的研究还处于起步阶段。国际上也没有一个公开、公认的汉语问答系统测试集以及评估方法。作为尝试，本研究小组已初步建立一个汉语问答系统评测平台(简称 EPCQA)。其中，EPCQA 语料库、测试集和打分标准的建立基本参考 TREC QA Track、NICIR 和 CLEF 的成功经验，并针对汉语的特点进行适当的修正。

现阶段，EPCQA 的答案源语料库约 1.8GB，主要来自互联网网页，分布于国内、国际、娱乐、体育、社会和财经等领域的新闻报道。为评测需要，我们还对语料库进行了一定程度的深加工。

EPCQA 测试集的建立遵循全面性、真实性和无歧义性三个原则，且已从多个不同的渠道（自然语言搜索网站日志、百科知识问答题库、实验室工作人员，英语提问的翻译等）收集了 4250 个涉及事实、列表和描述等三大类型的测试提问集合。

从自然语言搜索网站的日志中提取的提问很多不是现阶段我们问答系统研究的重点。例如：省略了疑问词的提问、表达模糊的提问、要求回答的是完成某件事的程序而非简短答案的提问，等等。我们对这些提问进行人工剔除。例如：“如何网上赚钱？”、“女朋友过生日送什么？”、“如何申请免费空间？”、“成龙的近况如何?”、等等。还有一些符合要求但表达不当的提问，我们对此进行了一定的修改。

而百科知识问答题库中的提问描述得都很书面化，不能够反映用户真实提问的方式。对此我们进行了一些口语化的处理。例如百科知识问答题库中的提问：“香港电影《花样年华》最近在第 5 3 届戛纳国际电影节上获最佳男主角奖，在该片中饰演男主角的哪一位演员？”，我们把它修改成：“谁在香港电影《花样年华》中饰演男主角？”，我们认为这样更能反映系统在实际使用中的情况。

对英语提问句的“翻译”是我们获取汉语问答系统测试集的另一个非常重要的途径。其中，英语提问句的来源主要是往届的 TREC 比赛的测试集。这里的“翻译”并不全是对英语提问的直译，而是对于部分可能在中文中找不出答案的提问在不改变提问类型的情况下，进行了适当的修改，例如：

英语提问：Who wrote "East is east, west is west and never the twain shall meet"?
中文提问：名著《红楼梦》是谁的作品？
英语提问：What is the name of CEO of Apricot Computer?
中文提问：联想公司的 CEO 叫什么名字？

EPCQA 针对不同类型的提问采用不同的打分标准。初步拟定，事实提问采用 MRR 打分标准。列表提问采用事例召回率（IR）、事例准确率（IP）和 F-Measure（IF）等打分准则。而对每一个描述问题，评测员列出一个基本信息和可接受信息的表单。基本信息是指这一问题的答案中不可缺少的描述部分。可接受信息是指可以构成一个正确的答案的，但还不是必需的信息。超出基本信息和可接受信息的部分将在评分体系中给予扣分。这样可以用片断召回率（NR）、片断准确率（NP）和 F-Measure 来评测一个描述提问的得分。

EPCQA 目前还不成熟，训练和测试集的规模（4250 个用户提问）也有点小。我们计划在下一步除了逐步扩大现有测试类型提问的测试规模外，主要研究如何逐步扩大用户提问的广度和深度。我们计划按照

如图 4.1 所示的阶段研发我们的汉语问答系统。

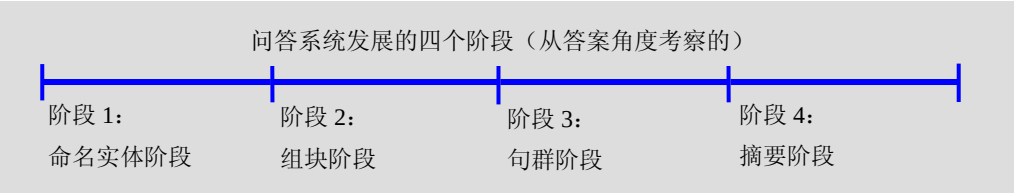

图 4.1 我们的汉语问答系统研发阶段

五、问答系统的基本原理

典型的问答系统通常由提问处理模块，检索模块和答案抽取模块三部分组成，如图 5.1 所示。其中，提问处理模块负责对用户的提问进行处理；生成查询关键词（提问关键词，扩展关键词，...）；确定提问答案类型（PER, LOC, ORG, TIM, NUM, ...）以及提问的句法、语义表示等等。检索模块根据提问处理模块生成的查询关键词，使用传统检索方式，检索出和提问相关的信息。返回的信息可以是段落、也可以是句群或者句子。答案抽取模块则从检索模块检索出的相关段落、或句群、或句子中抽取出和提问答案类型一致的实体，根据某种原则对候选答案进行打分，把概率最大的候选答案返回给用户。

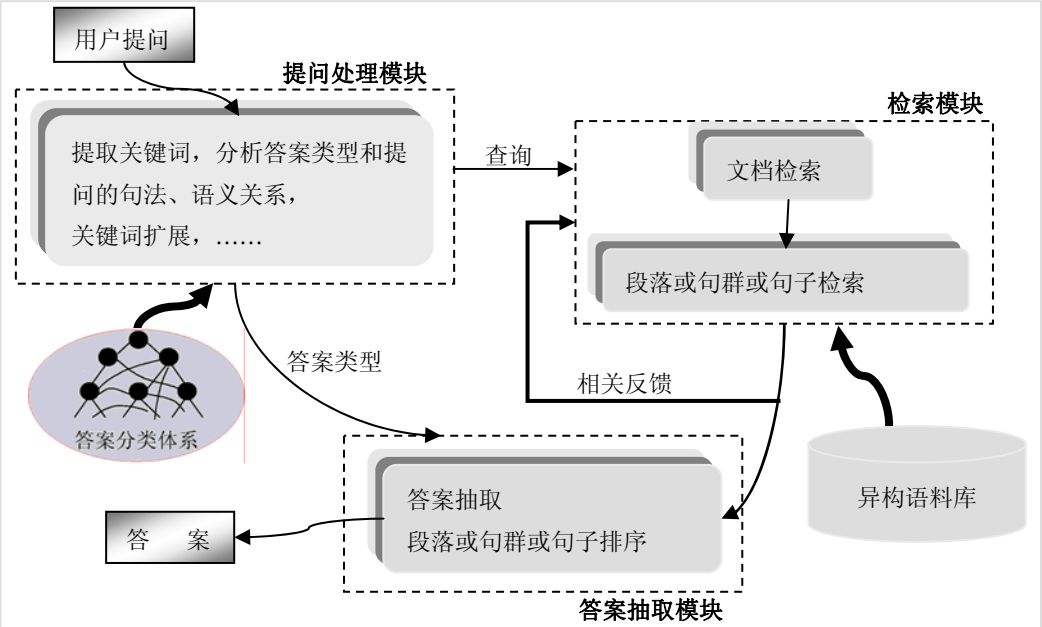

图 5.1 问答系统原理图

六、三种典型的问答技术分析

根据问答技术的技术特色，我们把问答技术分为三大类：基于信息检索和信息抽取的问答技术（IR + IE）、基于模式匹配的问答技术（IR + Pattern Matching）和基于自然语言处理的问答技术（IR + IE + NLP）。

6.1 基于信息检索和信息抽取的问答技术

候选答案的排序是这类技术的核心，排序的依据通常是提问处理模块生成的查询关键词。由于不同类别的关键词对排序的贡献不同，算法把查询关键词分为几类，即普通关键词（O）：从提问中直接抽取的关键词；扩展关键词（E）：从 WordNet 或者 Web 中扩展的关键词；基本名词短语（B）；引用词（Q）：通常是引号中的词；其他关键词（T）等等。公式(6.1)给出常用关键词的一种加权方法。

$$Score = wo \times O + we \times E + wb \times B + wq \times Q + wt \times T + ... \quad (6.1)$$

式(5.1)中的 *wo*，*we*，*wb*，*wq*，*wt* 分别是普通关键词、扩展关键词、基本名词短语、引用词和其他关键词的加权因子，他们体现各种关键词的重要程度。通常，*wo* > *we*，*wq* > *wb* > *wt*。式(5.1)中的 *O*、*E*、*B*、

Q、*T* 是关键词本身的得分，系统[19]使用答案关键词和提问关键词的覆盖度来表示；而[1]使用 ISF（Inverse Sentence Frequency）表示。基于信息检索和信息抽取的问答技术代表系统参见新加坡国立大学 Hui Yang 等人研发的系统[19]。

6.2 基于模式匹配的问答技术

如何自动获取某些类型提问（某人的出生日期、某人的原名、某物的别称等）的尽可能多的答案模式是基于模式匹配问答系统的关键技术。也就是说，如果我们能够获得某类提问答案所有可能的答案表达方式（模式），问答系统的设计将会变得相对简单。

基于模式匹配的方法往往是先离线获得各类提问答案的模式[7, 8, 35, 14]，在运行阶段，系统首先判断当前提问属于哪一类，然后使用这类提问的所有模式来对抽取的候选答案进行验证。例如，询问“某人生日年月日”类提问的部分答案模式如下：

1.0 <NAME>（<ANSWER> -）
0.85 <NAME> was born on <ANSWER>,
0.6 <NAME> was born in <ANSWER>
0.59 <NAME> was born <ANSWER>
0.53 <ANSWER> <NAME> was born
0.50 – <NAME>（<ANSWER>
0.36 <NAME>（<ANSWER> -

值得一提的是，在信息抽取领域，人们已经开始把注意力从原来的基于深层文本分析方法转移到基于字符的表层的文本分析技术上[28, 29]。基于模式匹配的问答技术代表系统参见俄罗斯InsightSoft-M公司 Martin Soubbotin等人研发的系统[28, 29]。

6.3 基于自然语言处理的问答技术

虽然前两种方法相对简单、有效，在 TREC2001、TREC2002 中获得了良好的成绩。但是，人们普遍认为：要想改进或者说更大程度地提高问答系统的性能，必须引入自然语言处理的技术[15]，前两种方法有它本身的缺陷性。现阶段，自然语言处理的技术还不成熟，对句子的深层句法、语义分析还不能达到实用的效果。因此，大多数系统都是基于对句子进行浅层分析，获得句子的浅层句法、语义表示，作为对前两种方法的补充和改进。这方面代表性工作有[4, 7, 18, 30, 31, 36]。

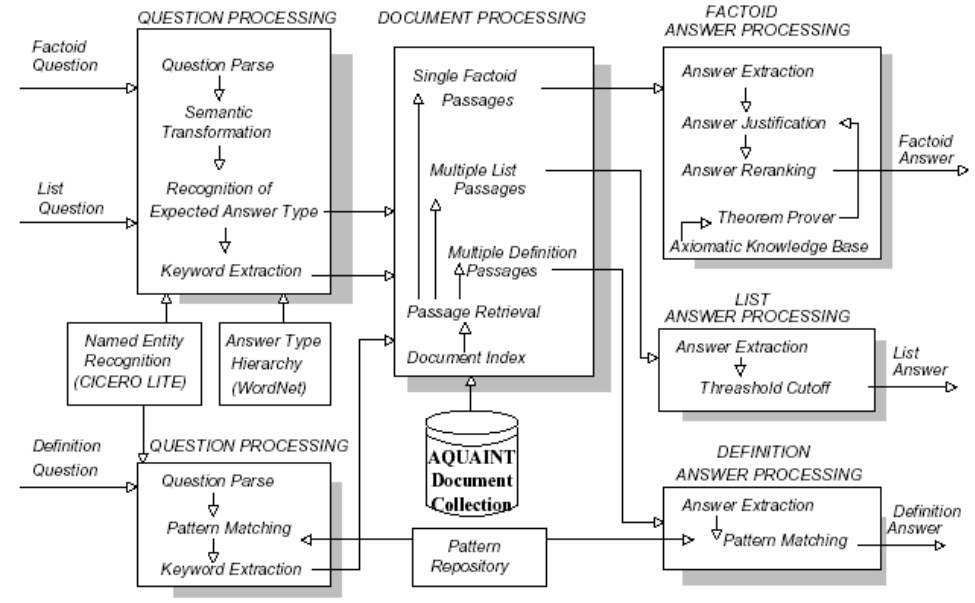

图 6.1　LCC 系统原理图

基于自然语言处理的问答技术典型系统参见美国 Language Computer Corporation 公司 Sanda Harabagiu 等人研发的系统[36]。该系统在 TREC QA Track 2001 ~ 2004 连续三年的评测中均获得第一名的成绩，且具有较大的领先优势。LCC 系统的原理图如 6.1 所示。该系统一个最大的特点是使用了逻辑形式转换技术实现答案的推理验证，这在 7.3 节有详细的描述。

6.4　三种代表技术的比较分析

本小节总结 IR + IE、IR + Pattern Matching 和 IR + IE + NLP 这三类问答式检索技术的优缺点，以及他们在各届 TREC 比赛中的成绩。

基于信息检索和信息抽取的问答技术相对简单，容易实现。但它以基于关键词的检索技术（也可被称为词袋检索技术）为重点，只考虑离散的词，不考虑词之间的关系。因此无法从句法关系和语义关系的角度解释系统给出的答案，也无法回答需要推理的提问。

基于模式匹配的问答技术虽然对于某些类型提问（如定义，出生日期提问等）有良好的性能，但模板不能涵盖所有提问的答案模式，也不能表达长距离和复杂关系的模式，同样也无法实现推理。

基于自然语言处理的问答技术可以对提问和答案文本进行一定程度的句法和语义分析，从而实现推理。但目前自然语言处理技术还不成熟，除一些浅层的技术（命名实体识别，汉语分词、词性标注等）外，其他技术还没有达到实用的程度。所以，这种技术的作用非常有限，只能作为对前两种方法有效的补充。

我们认为：基于字符表层的文本分析技术（例如，模板技术）必须和快速、浅层自然语言处理技术有效结合，才能获得性能优良的问答系统。

TREC 评测结果也说明，系统整体性能的优劣在很大程度上依赖于对 NLP 技术和资源的有效利用。表 6.1 给出了三类方法的典型系统以及他们在各届 TREC Factoid 子任务中取得的名次。

		IR + IE	IR + Pattern Match	IR + IE + NLP
代表系统		代表系统[19]	代表系统[28, 29]	代表系统[4, 31]
代表方法在各届 TREC 中的名次	2000	-	-	1
	2001	-	1	2
	2002	3	2	1
	2003	2	-	1

表 6.1 三类问答技术代表系统在 TREC 评测中的成绩比较

七、应用于问答系统的自然语言处理技术

本节将详细介绍被问答系统广泛使用的自然语言处理技术，这包括：命名实体识别技术、短语或依存结构分析技术、Paraphrase 技术、词汇链和逻辑形式转换等等。

7.1　命名实体识别技术

相对于其他技术来，命名实体识别技术是使用最广泛的一项自然语言处理技术，对问答系统有着十分重要的影响。命名实体识别技术主要被用在问答系统的段落或句子级排列和答案抽取两个阶段。

■　段落或句子排列

问答系统首先根据查询关键词进行检索，然后对于检索出来的段落或句子重新进行排序：当某个句子包含所期望的实体时，则给句子适当的加分。系统[1]采用如式(7.1)的加分策略。

$$Sne, i = Si + (E + (3 - dne) \times Dne) \tag{7.1}$$

式(7.1)还使用了距离信息，即期望实体和提问关键词之间的距离 *Dne*。*Dne* 是对距离的惩罚因子。当期望实体和提问关键词之间的距离越远，所加的分值就越小。

■　答案抽取

大多数的问答系统都是在答案抽取阶段使用命名实体的技术[4, 14, 19, 31]，答案抽取模块只抽取和期

望答案类型一致的实体作为答案，而命名实体不参与句子或段落的排序。

7.2 短语结构分析或依存结构分析技术

短语结构分析或依存结构分析的结果是得到句子的短语结构句法树或依存结构句法树。在句子排序或答案抽取阶段，使用更合理的句法信息。举个例子：

提问：Who killed Lee Harvey Oswald?

文本：Belli's clients have included ***Jack Ruby***, who killed ***John F. Kennedy*** assassin Lee Harvey Oswald and Jim and Tammy Bakker.

对于候选答案 Jack Ruby 和 John F. Kennedy，如果采用基于词袋的方法，系统很有可能返回 John F. Kennedy，因为 John F. Kennedy 和查询关键词 killed、Lee Harvey Oswald 的距离更近。但是，如果引入句法信息，系统只会返回答案 Jack Ruby。因为 Jack Ruby 在文本中是 killed 的逻辑主语，Lee Harvey Oswald 是 killed 的逻辑宾语，这和问句的句法结构完全相似。

算法[5, 7, 18, 24, 39, 41, 36]在使用句法树（短语句法树或依存句法树）的细节上有所不同，但他们的目的都是比较提问句法树和文本句法树的相似性，使系统给出的答案有句法上的解释。

7.3 逻辑形式转换（Logic Form Transformation）技术

通过比较提问和文本的句法树来抽取答案虽然提高了系统的性能，但这种基于句法树分析的方法还是非常浅层的。因为对句法树的分析基本上就是合一（Unification）运算，比较两棵句法树的相似性，无法回答那些需要推理才能回答的提问。举个例子：

提问：Who is the first Russian astronaut to walk in space?

文本：The broad-shouldered but paunchy Leonov, who in 1965 became the first man to walk in space, signed autographs.

提问依存树和文本依存树分别如图 7.1 和 7.2 所示。

图 7.1 提问依存关系数

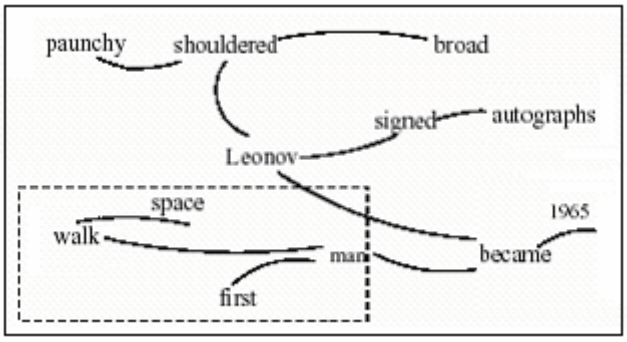

图 7.2 文本依存关系树

图 7.2 中的虚线框是问句依存树和文本依存树合一的结果。可以看到，合一的结果并不能给出提问的答案 Leonov，因为 Leonov 和 man 没有依存关系。这个时候，必须使用语义信息（Leonov 和 man 在这里是指同一个实体）才能给出正确答案。

[4, 31]提出 Logic Form 来解决这个问题，即把问句和文本同时转化成统一的 Logic Form（QLF 和 ALF），通过对 QLF 和 ALF 的运算来抽取答案。Logic Form 最大的特色是它结合词汇链可以表达语义知识，实现推理功能，这也是 LCC 系统成绩优异的主要原因。举个例子：

Q：When did Lucelly Garcia, former ambassador of Colombia to Honduras, die?

A：Several gunmen on a highway leading to the Colombian city of Ibague murdered Colombian Ambassador to Honduras Lucelly Garcia in 1994.

QLF：Lucelly_Garcia(x1) & former(x1) & ambassador(x1) & of(x1, x2) & Colombia(x2) & to(x1, x3) &

Honduras(x3) & die(e1, x1) & TIME_STAMP(e1)

ALF：gunman(x2') & murder(e1', x2', x1') & Colombian(x1') & ambassador(x1') & to(x1', x3') & Honduras(x3') & Lucelly_Garcia(x1') & TIME_STAMP(e1')

从 WordNet 中获取的推理知识库：

Rule1: Colombian(x1) ---→ of(x1, x2) & Colombia(x2)

Rule2: murder(e1, x2, x1) ---→ kill(e1, x2, x1) & intentionally(e1)

Rule3: kill(e, x1, x2) --→ cause(e1, x1, e2) & die(e2, x2)

根据 Rule1，Rule2 和 Rule3，由 ALF 可以推理出 ALF1。

ALF1：gunman(x2') & cause (e2', x2', e3') & die (e3', x1') & intentionally (e1') & of(x1', x7') & Colombia(x7') & ambassador(x1') & to(x1', x3') & Honduras(x3') & Lucelly_Garcia(x1') & TIME_STAMP(e2') & TIME_STAMP(e3')

通过合一运算 QLF 和 ALF1，系统给出正确答案"in 1994"。在整个推理过程中，词汇链知识库起到了非常重要的作用。

7.4 词汇链技术

很多情况下，提问关键词和文本关键词是不一致，但他们却表达相同的意思。词汇链对于解决这类提问非常的重要。7.3 节利用 WordNet 构建词汇链，连接提问关键词和答案关键词，实现推理。例如，WordNet 对 kill 的一个解释：cause to die，这样我们就可以把 kill 和 die 连接起来，即 kill 分解为 cause 和 die 两个动作，而且 kill 的宾语使 die 的主语，其 LF 为：Kill(e, x1, x2) --→ cause(e1, x1, e2) & kill(e2, x2)。

7.5 Paraphrase 技术

Paraphrase 是指用不同的词汇-句法结构表达同样的意思。7.4 节描述的词汇链就是一种特殊的 Paraphrase – 词汇 Paraphrase。我们认为 Paraphrase 可以解决因提问和答案的表述不同给问答系统的设计带来的麻烦。举个例子：

When did Colorado become a state?

(1a) Colorado became a state in 1876.

(1b) Colorado was admitted to the Union in 1876.

Who killed Abraham Lincoln?

(2a) John Wilkes Booth killed Abraham Lincoln.

(2b) John Wilkes Booth ended Abraham Lincoln's life with a bullet.

如果上述两个提问的答案都是以(1a)(2a)的形式来表述的，问答系统可以使用非常简单的技术（命名实体识别技术）就可以找出答案。但是如果答案以(1b)(2b)的形式出现，问答系统要找到答案将是非常困难的。但是通过 Paraphrase 技术获得如下的 Paraphrase 规则：

X became a state in Y ⟺ X was admitted to the Union in Y

X killed Y ⟺ X ended Y's life

问答系统就能容易地找出提问的答案。将 Paraphrase 技术应用于问答系统的代表工作有[2, 6, 7, 16, 40]等。

八、结束语

基于自然语言的问答式检索系统经过这几年的发展，已经成为自然语言处理领域的一个重要分支和新兴的研究热点，其"通过系统化、大规模地定量评测推动研究向前发展"的发展轨迹，以及某些成功启示，都极大地推动了自然语言处理研究的发展，促进了 NLP 研究与应用的紧密结合。

但是，目前的问答技术也不成熟，问答系统能够处理的提问非常有限，系统的性能离实用的目标还很远。作者认为，在问答系统（尤其是汉语问答系统）的发展过程中，我们应该注意以下一些问题。

1. 处理好问答系统研究和问答系统实用之间的关系。目前我们的问答系统基本上都是针对具有简短答案的事实问题研发的，但这样的系统在实际应用中到底能够解决用户真正关心问题的百分之多少，或者说我们应该研究哪种类型问答系统，非常值得我们去研究。
2. 重视大规模的公开评测技术，以评测推动问答技术的发展。现阶段对于汉语问答技术的研究，我们迫切需要一个公开、公认、合理的问答评测平台；
3. 从问答技术的研究角度看，我们需要重视基于字符表层的文本分析技术和基于自然语言处理技术的有效结合，扬长避短。

自然语言问答系统是集自然语言处理技术和信息检索技术于一身的新一代检索引擎，随着人们研究的逐步深入[21]及其广阔的应用前景，相信在不久的将来问答系统将会取得重大的突破。

致谢

两位评审专家对论文提出了非常中肯和宝贵的意见，在此表示衷心感谢！

参 考 文 献

[1] Ittycheriah and S. Roukos. IBM's Statistical Question Answering System-TREC 11[C]. In the Eleventh Text Retrieval Conference (TREC 2002), Gaithersburg, Maryland, November 2002.

[2] Ali Ibrahim, Boris Katz, and Jimmy Lin. Extracting structural paraphrases from aligned monolingual corpora[C]. In Proceedings of the Second International Workshop on Paraphrasing (IWP-2003), 2003.

[3] B. Wang, H. Xu, Z. Yang, Y. Liu, X. Cheng, D. Bu, S. Bai, TREC-10 Experiments at CAS-ICT: Filtering, Web and QA[C]. In The Tenth Text REtrieval Conference (TREC 10), page 109, 2001.

[4] D. Moldovan and V. Rus. Logic Form Transformation of WordNet and its Applicability to Question Answering[C]. In Proceedings of 37th Meeting of Association of Computational Linguistics (ACL'2001).

[5] D. Mollá. Towards Semantic-Based Overlap Measures for Question Answering (2003) [C]. Proc. ALTW03, Melbourne, December 2003.

[6] DUCLAYE F., YVON F. & COLLIN O. Learning paraphrases to improve a question answering system[C]. In Proceedings of the Natural Language Processing for Question Answering Workshop at EACL (EACL'03), Budapest, Hungary.

[7] Dekang Lin and Patrick Pantel. Discovery of inference rules for question-answering[C]. In Natural Language Engineering, volume 7, pages 343–360.

[8] Dell Zhang, Wee Sun Lee. Web based Pattern Mining and Matching Approach to Question Answering[C]. In Proceedings of the 11th Text REtrieval Conference (TREC), NIST, Gaithersburg, MD, Nov 2002.

[9] Ellen M. Voorhees, Dawn M. Tice. The TREC-8 Question Answering Track Evaluation[C]. The Eighth Text REtrieval Conference (TREC-8), Spec Pub 500-246, Washington DC: NIST, 1999, 77-82.

[10] Ellen M. Voorhees. Overview of the TREC 2003 question answering track[C]. In Proceedings of the Twelfth Text REtrieval Conference (TREC 2003), 2003.

[11] Ellen M. Voorhees. Overview of the TREC-9 Question Answering Track[C]. The Ninth Text REtrieval Conference (TREC-9), Spec Pub 500-249, Washington DC: NIST, 2000, 77-82.

[12] Ellen M. Voorhees. Overview of the TREC2001 Question Answering Track[C]. The Tenth Text REtrieval Conference (TREC-01), Spec Pub 500-250, Washington DC: NIST, 2001, 42-51.

[13] Ellen M. Voorhees. Overview of the TREC2002 Question Answering Track[C]. The Eleventh Text REtrieval Conference (TREC-02), Spec Pub 500-251,Washington DC: NIST, 2002.

[14] Eric Brill, Jimmy Lin, Michele Banko, Susan Dumais and Andrew Ng. Data-Intensive Question Answering[C]. In Proceedings of the Tenth Text Retrieval Conference (TREC2001), November, 2001.

[15] Eric Nyberg, Teruko Mitamura, Jaime Carbonell, Jaime Callan, Kevyn Collins-Thompson, Krzysztof Czuba, Michael Duggan, Laurie Hiyakumoto, Ng Hu, Yifen Huang, Jeongwoo Ko, Lucian V. Lita, Stephen Murtagh, Vasco Pedro, David Svoboda. 2003. The JAVELIN Question Answering System at TREC 2002[C]. In TREC 2002 Proceedings.

[16] F. Rinaldi, J. Dowdall, K. Kaljurand, M. Hess, D. Mollá. Exploiting Paraphrases in a Question Answering System (2003) [C]. Proc. Workshop in Paraphrasing at ACL2003, pp.25-32. July 11, Sapporo, Japan.

[17] Hongbo Xu, Hao Zhang, Shuo Bai. ICT Experiments in TREC-11 QA Main Task[C]. In the Eleventh Text REtrieval Conference (TREC 11), 2002.

[18] Hovy, E.H., U. Hermjakob, and Chin-Yew Lin. 2001. The Use of External Knowledge of Factoid QA. In Proceedings of the 10th Text Retrieval Conference (TREC 2001) [C], Gaithersburg, MD, U.S.A., November 13-16, 2001.

[19] Hui Yang, Tat-Seng Chua. The Integration of Lexical Knowledge and External Resources for Question Answering[C]. In the Proceedings of the Eleventh Text REtrieval Conference (TREC'2002), Maryland, USA, 19-22 Nov 2002, page 155-161.

[20] Ian Soboro, Donna Harman. Overview of the TREC 2003 Novelty Track[C]. Text Retrieval Conference (TREC12), NIST, Maryland, USA, 2003.

[21] John Burger et al. 2001. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A) [C]. http://www.ai.mit.edu/people/jimmylin/papers/Burger00-Roadmap.pdf

[22] Julian Kupiec. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia[C]. In Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 181–190, 1993. Special issue of the SIGIR FORUM.

[23] Junichi Fukumoto, Tsuneaki Kato and Fumito Masui. Question Answering Challenge (QAC1): An Evaluation of QA Tasks at the NTCIR Workshop 3[C]. In Proc. of AAAI Spring Symposium: New Directions in Question Answering, pp.122-133, 2003.

[24] Kenneth C. Litkowski. Question-Answering Using Semantic Triples[C]. Eighth Text REtrieval Conference (TREC-8). Gaithersburg, MD. November 17-19, 1999.

[25] Kenneth C. Litkowski. Syntactic Clues and Lexical Resources in Question-Answering[C]. Ninth Text REtrieval Conference(TREC-9). Gaithersburg, MD. November 13-16, 2000.

[26] Lide Wu et al.. FDU at TREC-10: Filtering, QA, Web and Video Tasks[C]. 10th Text REtrieval Conference, Gaithersburg, USA, Nov. 2001

[27] Lide Wu, Xuanjing Huang, Junyu Niu, Yingju Xia, Zhe Feng, Yaqian Zhou. FDU at TREC2002: Filtering, Q&A, Web and Video tasks[C]. 11th Text REtrieval Conference, Gaithersburg, USA, Nov. 2002

[28] M. M. Soubbotin, S. M. Soubbotin. Patterns of Potential Answer Expressions as Clues to the Right Answers. Tenth Text REtrieval Conference (TREC-10) [C]. Gaithersburg, MD. November 13-16, 2001.

[29] M.M. Soubbotin, S.M. Soubbotin. Use of Patterns for Detection of Likely Answer Strings: A Systematic Approach[C]. In the Eleventh Text Retrieval Conference (TREC 2002), Gaithersburg, Maryland, November 2002.

[30] Marius Pasca. A Relational and Logic Representation for Open-Domain Textual Question Answering[C]. ACL (Companion Volume) 2001: 37-42

[31] Moldovan, D., Harabagio, S., Girju, R., Morarescu, P., Lacatsu, F., Novischi, A., et al. LCC Tools for Question Answering[C]. NIST Special Publication: SP 500-251 The Eleventh Text Retrieval Conference (TREC 2002).

[32] Paul Cohen, Robert Schrag, Eric Jones, Adam Pease, Albert Lin, Barbara Starr, David Gunning, and Murray Burke. The DARPA high-performance knowledge bases project[C]. AI Magazine, pages 25–49, Winter 1998.

[33] Qianli Jin, Jun Zhao, Bo Xu. NLPR at TREC2003 - Novelty and Robust Track[C]. Text Retrieval Conference (TREC-12), NIST, Maryland, USA, 2003.

[34] Qianli Jin, Jun Zhao, Bo Xu. Window-based Method for Information Retrieval[C]. The First International Joint Conference on Natural Language Processing. (IJCNLP-04), Hainan Island, China, 2004.

[35] Ravichandran, D. and E.H. Hovy. Learning Surface Text Patterns for a Question Answering System[C]. In 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002) Conference, Philadelphia, PA, July 2002.

[36] Sanda Harabagiu, Dan Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan Bunescu, Roxana Girju, Vasile Rus and Paul Morarescu. FALCON: Boosting Knowledge for Answer Engines[C]. Proceedings of the Text Retrieval Conference (TREC-9). Gaithersburg, MD. November 13-16, 2000.

[37] Sujian Li, Jian Zhang, Xiong Huang and Shuo Bai. Semantic Computation in Chinese Question Answering System[J]. Journal of Computer Science and Technology, 2002.

[38] Susan Dumais, 2002, Web Question Answering: Is More Always Better? [C], In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002), August, 2002.

[39] TAKAHASHI Tetsuro, NAWATA Kozo, KOUDA Shinya and INUI Kentaro. Seeking Answers by Structural Matching and Paraphrasing[C]. NTCIR Workshop3 Meeting, Question Answering Task. 2002.

[40] Terry Winograd. Five Lectures on Artificial Intelligence[J]. Linguistic Structures Processing, volume 5 of Fundamental Studies in Computer Science, pages 399- 520, North Holland, 1977.

[41] U. Hermjakob. Parsing and Question Classification for Question Answering[C]. In Proceedings of the ACL Workshop on Open-Domain Question Answering, Toulouse, France, 2001.

[42] W. A.Woods. Lunar rocks in natural english: Explorations in natural language question answering[J]. Lingusitic Structures Processing, volume 5 of Fundamental Studies in Computer Science, pages 521–569, North Holland, 1977.

[43] Xiaoyan Li, W. Bruce Croft, Evaluating Question-Answering Techniques in Chinese[C]. Computer Science Department University of Massachusetts, Amherst, MA , 2001.

[44] Yi Zhang, DongMo Zhang. Enabling answer validation by logic form reasoning in Chinese question answering[C]. In Proceeding of 2003 International Conference on Natural Language Processing and Knowledge Engineering. pages 275- 280, Beijing, 2003.

[45] Youzheng Wu, Jun Zhao, Bo Xu. Chinese Named Entity Recognition Combining Statistical Model with Human Knowledge[C]. In: The Workshop attached with 41st ACL for Multilingual and Mix-language Named Entity Recognition: Combining Statistical and Symbolic Models, pp.65-72, Sappora, Japan, 2003

[46] 崔恒, 蔡东风, 苗雪雷. 基于网络的中文问答系统及信息抽取算法研究[J]. 中文信息学报, 2004, 3.

[47] 张刚, 刘挺, 郑实福, 车万翔, 秦兵, 李生. 开放域中文问答系统的研究与实现[C]. 中国中文信息学会二十周年学术会议, 2001,11

[48] 郑实福, 刘挺, 秦兵, 李生. 自动问答综述[J]. 中文信息学报, 2002, 6.

[49] 吴友政, 赵军, 段湘煜, 徐波. 构建中文问答系统评测平台. 第一届全国信息检索与内容安全学术会议, 上海, 2004,11.