

第八章

问答系统



提纲

- 引言
- 问答系统发展历程
- 问答系统的种类
- 基于常问问题集的问答系统实现
- 基于大规模文档集的问答系统实现
- **TREC**问答评测任务与性能度量指标

引言

——什么是问答系统？

- 是一种不同于传统搜索引擎的**信息检索系统**
- 应用形式
 - 用户以自然语言问句的形式提供自己的信息需求，
 - 系统根据用户的输入，给用户返回能回答用户问题的准确、简洁的答案
- 例如
 - 用户输入：“中国的首都是哪个城市？”
 - 系统返回答案：“北京”
- 英文翻译：Question Answering System
 - Question Answering: 问答技术
 - 一般指学术研究者围绕该类系统所研究的各种相关技术
 - 有人也翻译为问答系统

引言

——为什么需要问答系统

- 传统的搜索引擎

- 应用模式

- 用户输入表达信息需求的关键词
 - 系统返回可能包含用户需要的文档列表

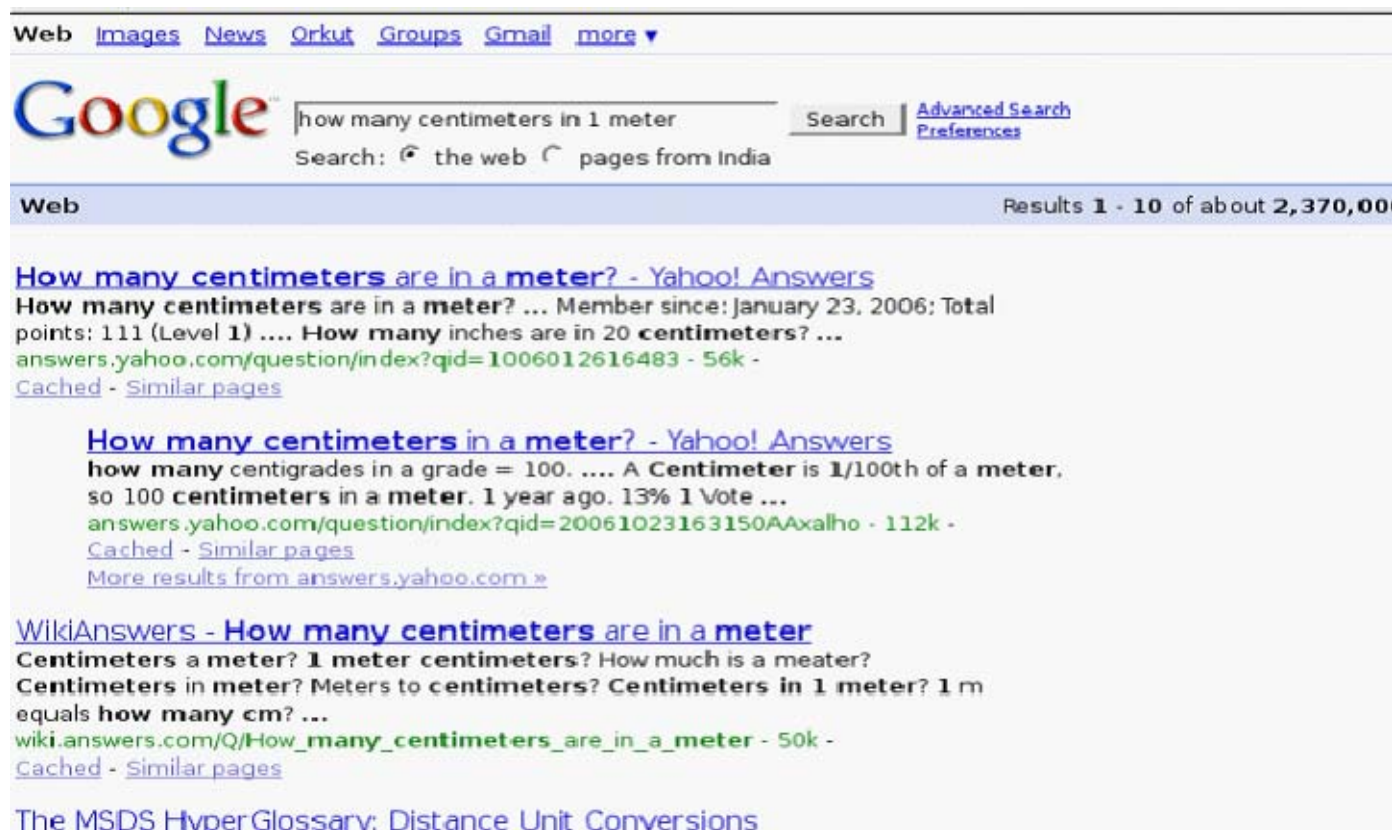
- 缺点

- 关键词构造的好坏非常影响搜索结果，很多时候对用户的应用水平要求较高
 - 用户真正的需求“埋”在返回结果中，需要用户自己浏览定位想要的内容

引言

——为什么需要问答系统（续）

■ Google检索



The screenshot shows a Google search interface with the query "how many centimeters in 1 meter". The search results are displayed under the "Web" tab, showing approximately 2,370,000 results. The first three results are from Yahoo! Answers, WikiAnswers, and The MSDS HyperGlossary, all providing the answer that there are 100 centimeters in a meter.

Web [Images](#) [News](#) [Orkut](#) [Groups](#) [Gmail](#) [more](#) ▼

Google [Advanced Search](#)
Search: ☒ the web ☐ pages from India

Web Results **1 - 10** of about **2,370,000**

[How many centimeters are in a meter? - Yahoo! Answers](#)
How many centimeters are in a meter? ... Member since: January 23, 2006; Total points: 111 (Level 1) How many inches are in 20 centimeters? ...
[answers.yahoo.com/question/index?qid=1006012616483](#) - 56k -
[Cached](#) - [Similar pages](#)

[How many centimeters in a meter? - Yahoo! Answers](#)
how many centigrades in a grade = 100. A Centimeter is 1/100th of a meter, so 100 centimeters in a meter. 1 year ago. 13% 1 Vote ...
[answers.yahoo.com/question/index?qid=20061023163150AAxalho](#) - 112k -
[Cached](#) - [Similar pages](#)
[More results from answers.yahoo.com »](#)

[WikiAnswers - How many centimeters are in a meter](#)
Centimeters a meter? 1 meter centimeters? How much is a meater? Centimeters in meter? Meters to centimeters? Centimeters in 1 meter? 1 m equals how many cm? ...
[wiki.answers.com/Q/How_many_centimeters_are_in_a_meter](#) - 50k -
[Cached](#) - [Similar pages](#)

[The MSDS HyperGlossary: Distance Unit Conversions](#)

引言

——为什么需要问答系统（续）

- 问答系统START返回的结果

START's reply

==> how many centimeters in 1 meter?

There are 100 centimeters in a meter.

Source: START KB

-
- [Go back to the START dialog window.](#)

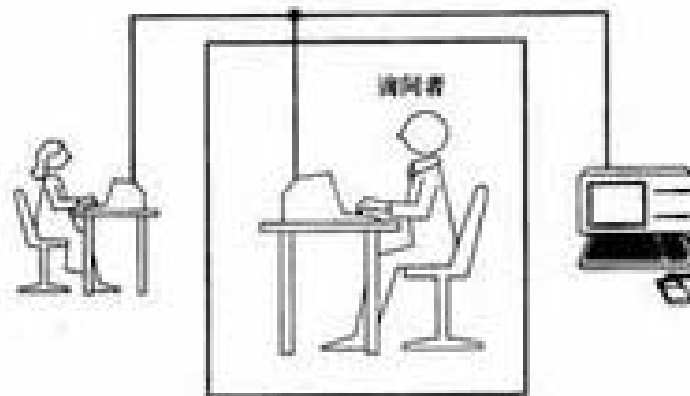


提纲

- 引言
- 问答系统发展历程
- 问答系统的种类
- 基于常问问题集的问答系统实现
- 基于大规模文档集的问答系统实现
- TREC问答评测任务与性能度量指标

问答系统发展历程

——问答系统研究



- 不是从20世纪90年代才开始，而是有很长的历史
- 图灵测试：图灵在论文《Computing Machinery and Intelligence》中提到测试机器是否有智能的问题（1950年）
 - “机器能思考吗？”
 - 判断机器能否思考的方法——图灵测试
 - 由测试人与另一房间中的两个对象E和F对话
 - 如果经过一段时间的对话之后，测试人如果不能断定E和F中谁是人、谁是计算机，则认为计算机已经具备了人的智能



问答系统发展历程

——历史上经典的问答系统

- **BASEBALL（1961年），Green等开发**

- **功能：**

- 可回答美国一个季度棒球比赛的时间、地点、成绩等自然语言问题

- **ELIZA（1966年），Weizenbaum等开发**

- **功能：**

- 系统扮演一个心理学专家的角色
 - 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，达到对病人进行心理治疗的目的

- **特点**

- 能给用户有和它“对话”的感觉



问答系统发展历程

——历史上经典的问答系统

- SHRDLU（1972年），Terry Winograd开发
 - 功能：
 - 呈现一个图形界面的虚拟世界
 - 系统根据和用户的对话，按照用户的要求来把不同形状和大小的彩色木块进行移动
- LUNAR（1973年），Woods等开发
 - 功能：
 - 帮助地质学家了解、评估阿波罗登月计划积累的月球土壤和岩石的各种化学分析数据
 - 在1971年的第二届年度月球科学会议上得到成功的展示
 - 性能：
 - 有关月球岩石数据的111个问题，能正确回答其中的78%



问答系统发展历程

——历史上经典的问答系统

- **SAM（1977年）**，耶鲁大学人工智能实验室开发
 - 功能：理解一篇文章
 - 由使用者针对文章提出各种问题
 - 系统根据对文章的理解，给出文章中出现的、可能包含答案的句子
- **LILOG项目**（由IBM在1985年开始开发，1991年产生演示版本）
 - 功能
 - 可回答关于德国城市旅游领域的问题
- **Deep Read（1999年）**，MITRE公司开发
 - 功能
 - 和SAM类似



问答系统发展历程

——历史上经典的问答系统

- 90年代后，产生了各种在线问答系统
 - 如Ask.com, START, AnswerBus等等
- 问答技术评测
 - TREC在1999年开始，设立开放域问答技术评测任务
 - 极大激发和促进了问答相关研究
 - 2004年的最好系统能正确回答77%的基于事实的问题



提纲

- 引言
- 问答系统发展历程
- 问答系统的种类
- 基于常问问题集的问答系统实现
- 基于大规模文档集的问答系统实现
- TREC问答评测任务与性能度量指标



问答系统的种类

——分类方法

- 应用不同，需要的问答系统形式不同，采用的语料和技术也会有不同
- 可以从不同角度去分类问答系统
 - 应用的领域
 - 使用的语料规模、语料的格式等
 - 实现的技术
- 按应用领域进行分类
 - 限定域问答系统
 - 系统所能处理的问题只能限定于某个领域或者某个内容范围
 - 开放域问答系统。
 - 系统可回答的问题不应当限定于某个特定领域



问答系统的种类

——按实现技术分类

- 自然语言的数据库问答系统
- 对话式问答系统
- 阅读理解系统
- 基于常用问题集的问答系统
- 基于知识库的问答系统
- 基于大规模文档集的问答系统



问答系统的种类

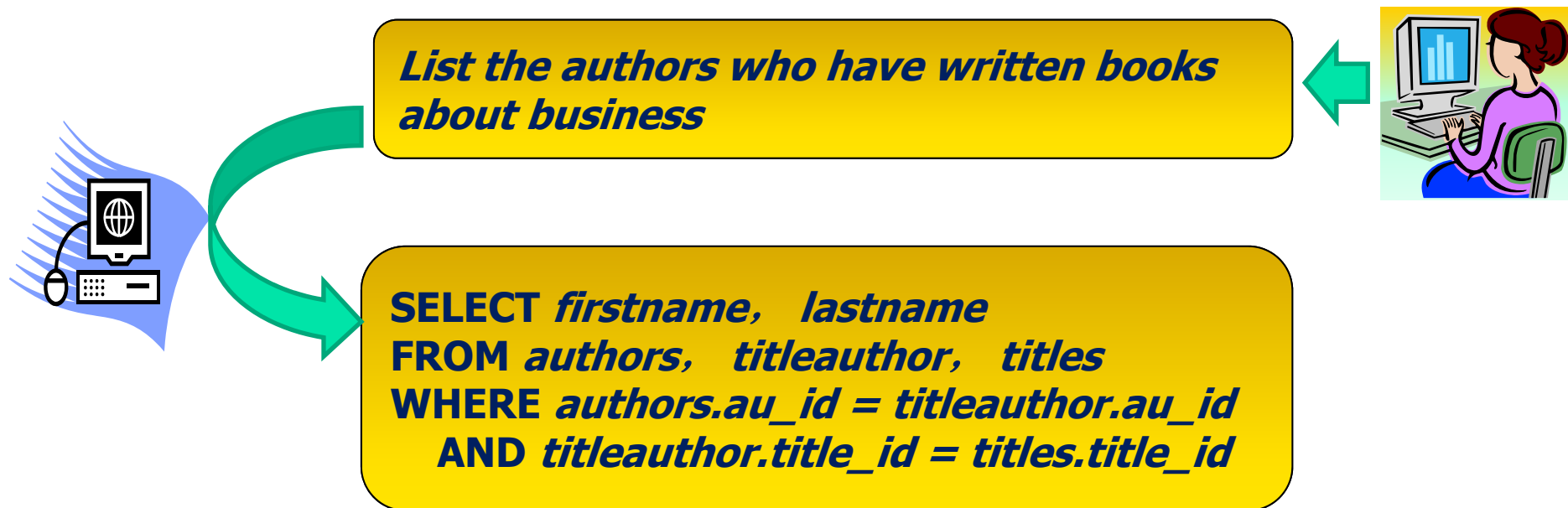
——自然语言的数据库问答系统

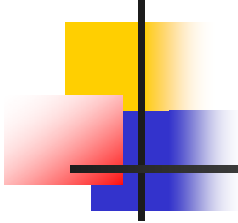
- 代表性的系统
 - 如前述的BASEBALL、LUNAR
- 基本原理
 - 将问句进行形式化处理，生成数据库系统能接受的查询语句（如SQL语句）
 - 将生成的查询语句发送给数据库系统，获取答案

问答系统的种类

——自然语言的数据库问答系统

■ 举例





问答系统的种类

——对话式问答系统

- 重要特征

- 能以对话的形式和使用者进行交流，问题和相应答案的描述可以在一个上下文环境中

- 实例

- 前述的ELIZA、SHRDLU
- 聊天机器人（ChatBot），如ALICE等

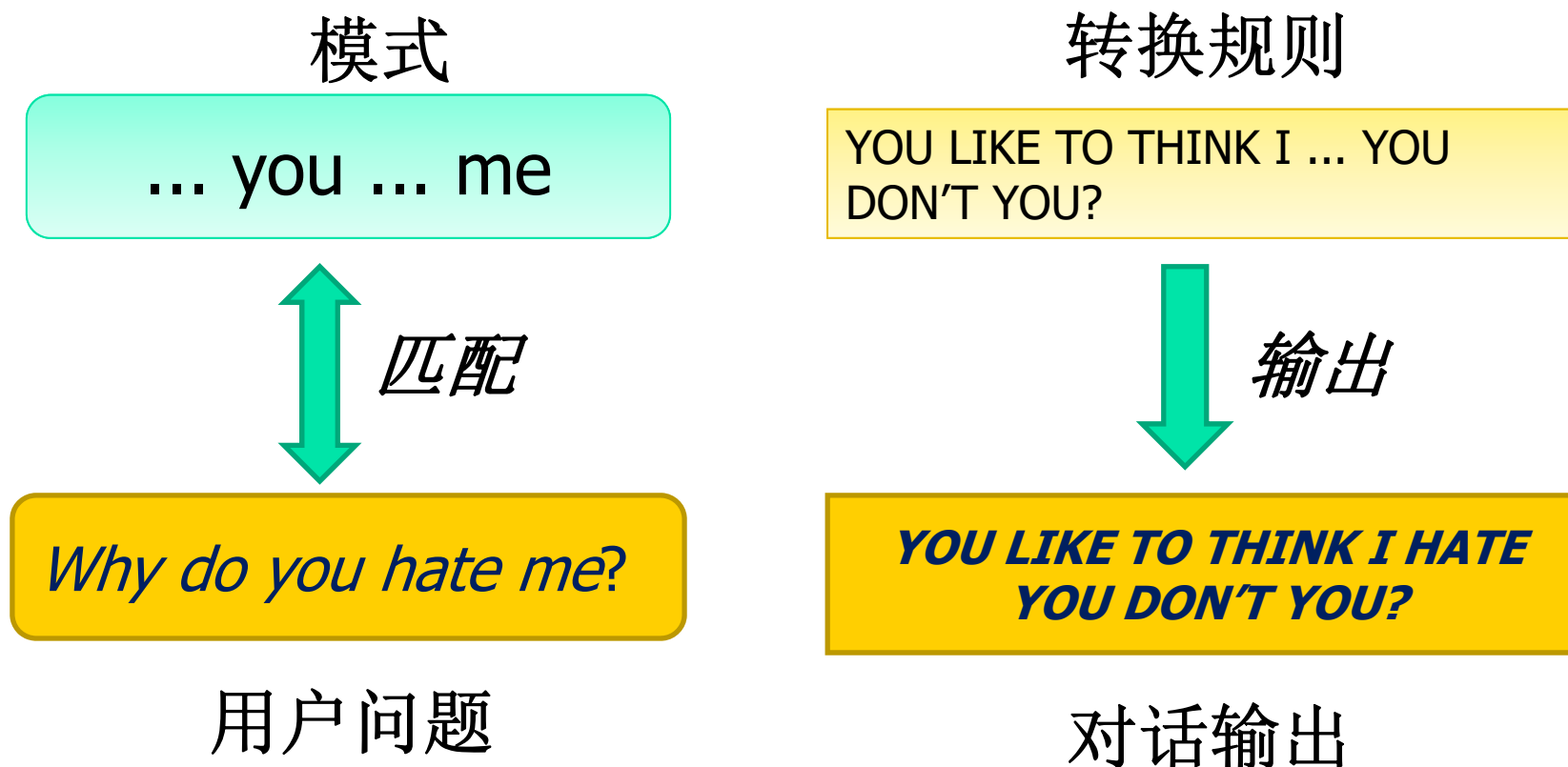
- ALICE

- 通过模式匹配和字符串的子串替换方法，来对用户的提问进行处理，并返回应答

问答系统的种类

——对话式问答系统

ALICE模式匹配

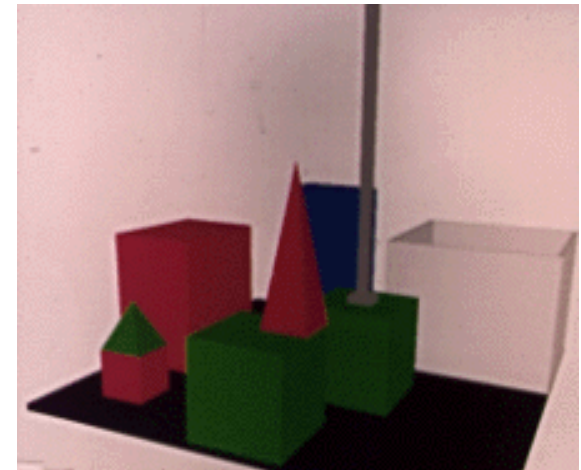


问答系统的种类

——对话式问答系统

■ SHRDLU

- 用户可以用对话的方式，请求系统在虚拟环境中控制一个机器人手臂，做各种操作
- 虚拟环境由一张桌子、垒在桌子上的一些不同形状大小的彩色积木、一个可放积木的盒子组成



问答系统的种类

——阅读理解系统

- 和早期的对话系统类似，早期的阅读理解系统也一般是由人工智能研究人员来开发
- **SAM**：也就是“**Script Applier Mechanism**”的简称
 - 要求有一个和文章内容相关的场景脚本
 - 在用户提出问题时，系统根据文章内容、场景脚本来回答问题
 - 脚本不存在时，系统无法正常工作



问答系统的种类

——基于知识库的问答系统

- 实例

- CYC、NKI（US）、NKI等

- 特点

- 优点：回答准确，可以进行一定的推理计算
 - 缺点是：需要建立大规模知识库，消耗大量的人力物力



问答系统的种类

——基于常问问题集的问答系统

- 常问问题集（Frequently-Asked Question，简称FAQ）
 - 是指被经常询问的问题
- 业务或者商业相关的用户经常会咨询某项业务、产品等的各种问题
- 如果能将经常问的问题及其答案整理出来，在用户提出和现有记录相同或者相似的问题时，可直接给出现成的答案
 - 提高效率，减少人工回答时的重复劳动

问答系统的种类

——基于大规模文档集的问答系统

- 通过对用户所提出的问题进行分析，根据问题从大规模文本集中检索到与之相关的文本
- 从检索到的文本中抽取出对问题的答案
- 实例：
 - Ask、AnswerBus、START 等
 - 都以海量的互联网页作为文档集



提纲

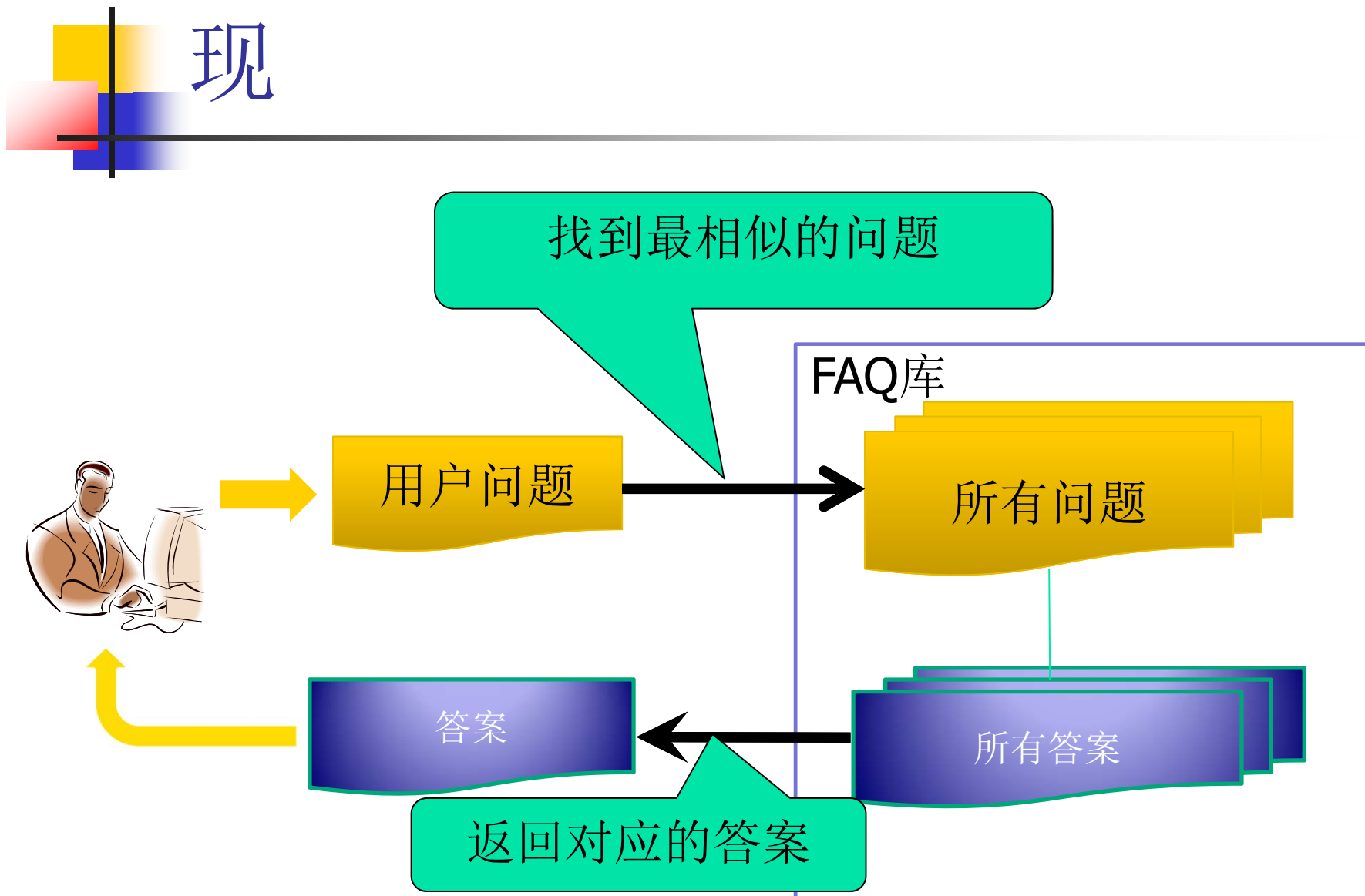
- 引言
- 问答系统发展历程
- 问答系统的种类
- **基于常问问题集的问答系统实现**
- 基于大规模文档集的问答系统实现
- TREC问答评测任务与性能度量指标



基于常问问题集的问答系统实现

- 常问问题集（Frequently Asked Questions，简称FAQ）
 - 把用户经常提问的问题和答案积累、保存起来，从而得到的一个大的“问答对”库
- 系统的基本实现过程
 - 接受用户输入的新问题
 - 计算新问题和问题库中所有问题之间的相似程度，找出和新问题最相似的若干个问题
 - 把最相似的这些问题及其答案返回给用户，用户从中获取新问题的答案

基于常问问题集的问答系统实现





基于常问问题集的问答系统实现

- 在系统实现中，需要解决三个问题
 - 候选问题集的建立
 - 问句相似度计算
 - 常问问题集的更新
- 候选问题集的建立
 - 问题库规模庞大时，查找和新问题（可称之为目标问句）相似的现存问题
 - 如果对库中每一个现存问题都作比较，则会降低系统的效率
 - 一种解决方法：按照一定策略，从中选择出有可能和新问题相似的一个较小的候选问题集
 - 在候选集中利用相似度计算方法选择最相似的现存问题，则减少了计算的时间，提高了速度

基于常问问题集的问答系统实现

■ 候选问题集的建立

■ 建立方法

- 设用户输入的目标问句中中共有 n 个词 (W_1, W_2, \dots, W_n)
- FAQ 库中共有 m 个问句，第 i 个问句含有 n_i 个词 (Q_1, Q_2, \dots, Q_{n_i})
- 第 i 个问句和目标问句之间重叠的词个数记为

$$Num_i = \left| \{W_1, W_2, \dots, W_n\} \cap \{Q_1, Q_2, \dots, Q_{n_i}\} \right|$$

- 则可以从 Num_i 最大的问题中，选择一定比例的问题组成候选问题集
- 可以设计合理的数据结构来保存FAQ库，以方便快速地计算 Num_i

基于常问问题集的问答系统实现

- 问句相似度计算

- 基于向量空间模型的相似度计算

- 将FAQ库中的问题表示成一个 n 维的向量

$$T = \langle T_1, T_2, \dots, T_n \rangle$$

- T_i 的计算方法

- 设 n 为 w_i 在这个问句中出现的个数， m 为FAQ中含有 w_i 的问句的个数， M 为FAQ中间句的总数，则

$$T_i = n \times \log(M / m)$$

- 将用户的新问题（目标问句）也表示为一个类似的 n 维向量 $T' = \langle T'_1, T'_2, \dots, T'_n \rangle$

基于常问问题集的问答系统实现

■ 问句相似度计算

■ 基于向量空间模型的相似度计算

■ 目标问句和FAQ库问题的相似度计算

$$\text{Similarity}(T, T') = \frac{\sum_{i=1}^n T_i \times T'_i}{\sqrt{\sum_{i=1}^n T_i^2 \sum_{i=1}^n T'^2_i}}$$

■ 缺点

- 一种统计的方法，只有当句子包含的词数足够多时，相关的词才会重复出现，该方法的效果才能体现
- 只考虑了词在上下文中的统计特性，而没有考虑词本身的语义信息

基于常问问题集的问答系统实现

■ 问句相似度计算

■ 基于语义的相似度计算方法

- 决定问题语义的不仅仅是词，也有词之间的关系
 - 两个问题包含的词汇相同，语义可能不同
 - 两个问题包含的词汇不同，但语义可能相同
- 可以借助《知网》等知识资源
- 实际是研究中的难点



基于常问问题集的问答系统实现

■ FAQ库的更新

- 用户新问题和库中所有问题的相似度均小于一定阈值
 - 可以认为库中没有对应用户新问题的答案
 - 可将新问题保存
- 对在库中没有类似问题的新问题，可以用人工等方式增加该问题的答案，并将“新问题-答案”添加到FAQ库中



提纲

- 引言
- 问答系统发展历程
- 问答系统的种类
- 基于常问问题集的问答系统实现
- 基于大规模文档集的问答系统实现
- TREC问答评测任务与性能度量指标



基于大规模文档集的问答系统实现 ——问题的种类

■ 背景

- 在开放域的应用中，用户的问题各种各样
- 技术发展的限制，要求这类系统的实现先简后难，循序渐进
- 需要分清简单问题、复杂问题

■ 根据预期答案形式的不同，将系统要解决的问题进行种类划分

- 事实型问题
- 定义型问题
- 复杂型问题



基于大规模文档集的问答系统实现 ——问题的种类

- 事实型的问题
 - 针对某事件，提问发生时间、地点、涉及的人物等
 - 或者针对某个实体，提问关于实体的属性
 - 举例：“参加**2008**年奥运会的有多少个国家？”
- 定义型的问题
 - 提问对某个目标的定义。举例：“什么是**SARS**？”
- 复杂问题
 - 观点型的问题
 - 提问某个人或者组织、机构对某事件或者实体的看法和认识
 - 举例：“欧盟对于朝鲜核试验有什么样的看法？”
 - 过程描述型的问题
 - 提问对某个动作的叙述、某件事情的原因、解释等
 - 举例：“怎么做鱼香肉丝啊？”
 - 其它问题



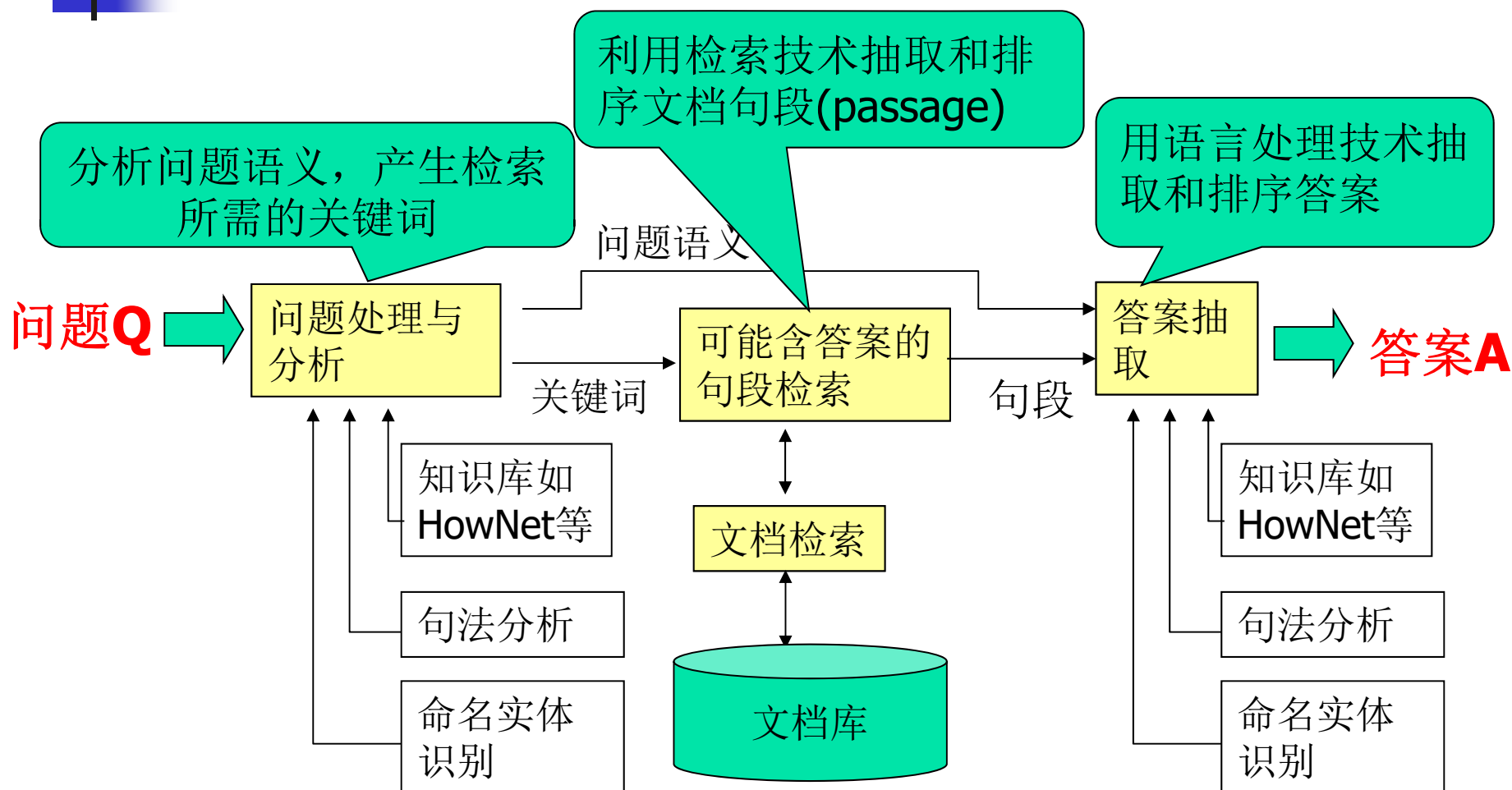
基于大规模文档集的问答系统实现

——问题的种类

- 不同种类问题的回答难度不同
 - 事实型问题相对最容易
 - 观点型问题和过程描述型等复杂问题较困难
- 本章只讨论事实型问题和定义型问题的回答
- 事实型问题又分两种
 - 简单事实型问题：答案是一个命名实体或者短语
 - 如：“世界上人口最少的国家是哪个？”
 - 罗列事实型问题：答案是多个命名实体或者短语
 - 如：“奥运会曾经在哪些国家举办过？”

基于大规模文档集的问答系统实现

——一般的体系结构





基于大规模文档集的问答系统实现 ——问题分析与处理

■ 功能

- 产生对预期答案的约束条件，使得答案抽取模块只从满足这些约束条件的候选中抽取答案来

- 构造在检索阶段需要的关键词

■ 约束条件分析

- 判别问题所期望的答案语义类型，称之为“问题分类”
- 识别所期望的答案和问题中其它词之间的语法和语义关系



基于大规模文档集的问答系统实现 ——问题答案语义类型分类

- 举例
 - 问题：哪个国家人口最多？
 - 期望的答案类型：“国家”
 - 问题：毛泽东生于哪一年？
 - 期望的答案类型：“日期”
 - 问题：中国人口有多少？
 - 期望的答案类型：“数量”
- 抽取答案时只需判断属于该语义类型的词或者短语，以提高答案抽取的准确率
- 显然，首先要求有一个类别体系



基于大规模文档集的问答系统实现 ——问题答案语义类型分类

- 哈尔滨工业信息检索研究室“中文问题分类体系”
 - 首先是粗类，每个粗类下又分为若干细类
 - 粗类别数：7类，细类别数：85类
 - 粗类：人物、地点、数字、时间、实体(包括抽象实体)、描述、未知
- 问题的答案形式种类与答案语义类型的区分
 - 前者是描述问题答案的简单或者复杂程度
 - 答案是表达事实的短语？还是一段复杂描述？
 - 后者是描述答案的语义类型
 - 答案是“国家名称”？还是“数量”？还是不能确定具体类型的“描述”？



基于大规模文档集的问答系统实现 ——问题答案语义类型分类

■ 基于规则的分类方法

- 对每个类别设计大量的规则，一旦问题和一个规则相匹配，则问题就属于该规则对应的类别
- 举例
 - 规则：“*＜哪一个国家＞？” ， 对应类别：“国家”
 - 问题：“世界上耕地面积最大的是哪一个国家”？
 - 问题和规则匹配，则问题的类别为“国家”
- 优点：简单，准确率高
- 缺点：耗费大量人力去设计规则



基于大规模文档集的问答系统实现 ——问题答案语义类型分类

- 基于统计机器学习的分类方法
 - 形式和文本分类类似
 - 思路
 - 用人工方式对一批问题的类别进行标注
 - 利用已有机器学习方法，在已标注类别的训练问题集合上训练分类模型
 - 用训练得到的模型对测试问题进行自动分类
 - 可以使用的机器学习方法
 - 朴素贝叶斯、支持向量机、最大熵模型等
 - 优点：得到的分类模型灵活性好
 - 缺点：需要训练问题集合进行标注



基于大规模文档集的问答系统实现 ——问题答案语义类型分类

- 基于统计机器学习的分类方法
 - 选择合适的问题特征，是保证有较高分类性能的关键
- 特征选择
 - 有些问题的疑问词就可以确定问题的类型
 - 如“哈工大的校长是谁？”，疑问词“谁”就确定类型是“人”
 - 有些问题的疑问词和问题焦点词共同确定了问题的类别
 - 问题焦点词是问题中指示期望答案语义类型的词
 - 如“哪个国家人口最多？”
 - 疑问词“哪个”、焦点词“国家”确定了该答案类型是“国家”



基于大规模文档集的问答系统实现 ——文档检索

- 答案抽取需要利用语言处理（如词性标注、语法分析等）和信息抽取（如名实体识别）等
 - 这些处理的计算复杂度较高
- 问答系统中的文档检索模块
 - 要在答案抽取之前，从海量文档集中检索出有可能包含答案的文档
 - 可节省处理的时间，提高系统的效率
 - 不是去找到问题的答案，而是查找有可能包含答案的文档，一般被称为“预取”



基于大规模文档集的问答系统实现 ——文档检索

- 大规模文档集的构造
 - 可以根据应用需要收集，也可以从Web上下载抓取
- 检索实现过程
 - 文档集预处理
 - 抽取答案时所需的耗时的处理用离线方式预先完成
 - 如句子的语法和语义分析、命名实体识别等
 - 根据问题构造查询关键词
 - 相关文档检索
 - 涉及的内容：检索模型、查询构造方法与查询扩展方法
 - 可以使用传统的文档检索模型和工具
 - 模型：向量空间模型、统计语言模型等
 - 工具：START、Lucene、Lemur等



基于大规模文档集的问答系统实现 ——文档检索

- 根据问题构造查询关键词
 - 简单的构造方法
 - 将问题中主要的词汇作为查询关键词
 - 举例：
 - 问题： *在中国的大学中，哪个学校的校园面积最大？*
 - 关键词： *中国 大学 学校 校园 面积 最大*
 - 缺点
 - 有些不包含这些关键词的文档，却包含问题的答案
 - 如何构造关键词，能使得检索模块能检索到包含答案的文档？
 - 仍然是需要继续深入研究的问题



基于大规模文档集的问答系统实现 ——文档检索

- 根据问题构造查询关键词时的扩展
 - 扩展的必要性
 - 答案在文档中往往以不同于问题的形式来表述
 - 关键词构造要让系统检索到和问题表达方式相近的文档
 - 关键词构造也要能让系统检索到和问题表达方式不同的相关文档
 - 扩展方法
 - 也是问答系统研究中需要解决的主要问题之一



基于大规模文档集的问答系统实现 ——文档检索

- 根据问题构造查询关键词时的扩展
 - 基于知识库的扩展
 - 可以利用问题中的词在 *WordNet*、*HowNet*、《同义词词林》等知识资源中的同义词、上位词和同义词来扩展
 - 如按《同义词词林》中的同义词扩展
 - 原始关键词
 - 中国 大学 学校 校园 面积 最大
 - 扩展后的关键词
 - 中国 (大学∨高校∨高等学校) (学校∨全校∨学府∨院所∨院校) 校园 (面积∨总面积) (最∨最为) (大∨广大)



基于大规模文档集的问答系统实现 ——句段检索

- 句段检索：Passage Retrieval
 - 有些也翻译为：片段检索
 - 把文档检索部分获得的文档拆分成文档片断或句子，从中选择最相关的部分
 - 更进一步减少答案抽取所需处理的内容长度
- 文档切分方法：不同的系统采用不同的方法
 - 以连续的 n 个句子作为一个文档句段
 - 根据经验来确定 n 的大小，如取1或者3
 - 以篇章的一个自然段(paragraph)为一个文档句段
 - 文档进行子话题(subtopic)分割，把一个子话题作为一个句段



基于大规模文档集的问答系统实现 ——句段检索

■ 检索方法

■ 最简单的方法

- 计算句段和问题之间匹配的词个数，将该数目作为句段的排序权值
- 计算问题和句段之间的余弦相似度，将该相似度作为句段的排序权值
- 对句段和问题间匹配词个数、词的idf权值、词间的密度综合加权，将该值作为句段的排序权值
 - 不仅考虑匹配词的个数及其它们在句段中的idf权值
 - 也考虑匹配词在句段中的相邻距离，即考虑匹配词在句段中的密度



基于大规模文档集的问答系统实现 ——简单事实型问题的答案抽取

■ 简单匹配的答案抽取方法

■ 基本思想

- 在检索到的句段中抽取和问题的预期答案类型相一致的命名实体，作为候选答案
- 对候选答案进行排序时，综合其所在文档句段的顺序和在所有文档句段中出现的次数作为排序分值
- 排序最高的候选答案，将被选择为最终的答案

■ 缺点之一

- 当句段中有多个满足答案语义类型的命名实体时，将难以判断



基于大规模文档集的问答系统实现 ——简单事实型问题的答案抽取

■ 基于表层模式匹配的方法

■ 基本思想

- 利用规则模式，从文档句段中抽取满足模式的答案
- 规则模式描述了问题的主要词汇与候选答案在句段中的出现形式
- 不需要太多深层的语言处理

■ 规则模式的构造

- 用手工方式构造或者自动学习得到



基于大规模文档集的问答系统实现 ——简单事实型问题的答案抽取

- 基于表层模式匹配的方法
 - 手工构造模式
 - 例如：对关于“某人出生年月日”的英文问题，构造的部分答案模式如下

```
1.0 <NAME> <ANSWER> 2
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>
```



基于大规模文档集的问答系统实现 ——简单事实型问题的答案抽取

- 基于表层模式匹配的方法

- 自动学习规则模式

- 需要有训练数据和测试数据
 - 数据包含：问题、包含答案的文档句段、问题的答案
 - 过程
 - 先从训练数据文档句段中，提取出包含问题词和答案的子串
 - 将子串中的问题词和答案词替换成变量，得到一个候选规则
 - 对候选规则进一步泛化
 - 在测试数据中用候选规则抽取答案，以判断候选规则的准确率
 - 将准确率高于一阈值的候选规则选为自动学习到的规则



基于大规模文档集的问答系统实现 ——简单事实型问题的答案抽取

■ 利用大规模文档集中数据冗余的方法

■ 基本思想

- 由于数据集的海量特点，总会存在一些和问题的描述方式相近的答案句子
- 系统可以不用复杂的语言处理技术，而使用简单的统计方法辅助浅层语言技术就可确定问题的答案

■ 举例：世界上最长的河流是哪个？

- 在检索到的前**150**文档句段中
 - 尼罗河：出现**20**次，亚马逊河：出现**10**次；伏尔加河：出现**9**次；雅鲁藏布江：出现**2**次
- 尼罗河为正确答案



基于大规模文档集的问答系统实现 ——简单事实型问题的答案抽取

■ 基于多特征的统计机器学习方法

■ 基本思想

- 在检索到的文档句段中，可能存在多个满足问题预期语义类型的命名实体或者短语
- 将这些命名实体或者短语作为候选答案
- 对这些候选答案进行排序或者选择时，利用特定类型的单一特征总会存在不足，一种方法是将多种特征结合起来
- 构造训练“问题-文档句段-答案”集合，用机器学习方法训练得到答案选择模型



基于大规模文档集的问答系统实现 ——简单事实型问题的答案抽取

- 基于多特征的统计机器学习方法

- 特征的选择

- 句子特征：问题和候选答案所在句段间匹配的词和依存弧的分值
 - 语言特征：候选答案是否是特定动词的主语或宾语等
 - 词汇模式特征：候选答案所在句段是否匹配某种词汇模式
 - 候选答案的冗余特征：候选答案在检索结果中出现的次数
 - 其它特征



基于大规模文档集的问答系统实现 ——罗列事实型问题的答案抽取

- 罗列（或者列举）事实型的答案是多个命名实体或者短语
- 答案抽取
 - 可以采用和简单事实型问题答案抽取相同的方法，选择多个候选答案
 - 可以设定一个阈值，将分值超过一定阈值的候选答案全部返回，则就是最终的答案



基于大规模文档集的问答系统实现 ——定义型问题的答案抽取

- 定义型问题提问对一个概念或者实体的定义
 - 举例
 - 问题：三角函数是什么？
 - 回答：由直角三角形边的比值得到的函数
 - 问题的答案在表述时存在一定的模式
 - 如“xxx是xxx”，“xxx指的是xxx”
- 可从大规模语料中总结出大量的抽取模式，利用这些模式抽取定义的答案句
 - 举例
 - <DEFTERM>代表要定义的术语
 - <DEFINITION>代表定义串



基于大规模文档集的问答系统实现 ——定义型问题的答案抽取

<DEFTERM>是<DEFINITION>
<DEFTERM>指的是<DEFINITION>
<DEFTERM>指的就是<DEFINITION>
<DEFTERM>的意思是说<DEFINITION>
<DEFTERM>的意思是<DEFINITION>
<DEFTERM>是一<DEFINITION>



提纲

- 引言
- 问答系统发展历程
- 问答系统的种类
- 基于常问问题集的问答系统实现
- 基于大规模文档集的问答系统实现
- **TREC问答评测任务与性能度量指标**

TREC问答评测的任务与性能度量指标

■ TREC-8

- 只设置简单事实型问题
- 答案形式：包含答案的文本块(snippet)，文本块的长度须在250字节之内
- 系统可以返回5个排好序的结果，评测的度量指标为*MRR*:

$$MRR = \frac{1}{N} \sum_{i=1}^N \frac{1}{rank(first_answer)}$$

- N 是问题总数
- $rank(first_answer)$ 是对于第*i*个问题的5个返回结果中，正确答案的顺序



TREC问答评测的任务与性能度量指标

■ TREC-9

- 使用了用户在搜索引擎中所提的问题，而不是评测的组织者根据语料专门设计的问题
- 与TREC-8相比，对答案的要求和度量指标没有变化

■ TREC-2001

- 任务中增加了罗列事实型问题
- 要求返回的答案串不能超过50个字节
- 度量指标仍然用MRR
- 文档集合中不一定包含有答案
 - 这类问题的标注答案则为空（NIL）



TREC问答评测的任务与性能度量指标

■ TREC-2002

- 问题形式和TREC-2001相同，包含简单事实型问题和罗列事实型问题
- 要求返回的结果是一个精确的回答，而不是包含答案的词串
- 度量指标
 - 系统能正确回答的问题数占全部测试问题总数的比例

TREC问答评测的任务与性能度量指标

■ TREC-2003

■ 问题形式

- 既包含简单事实型问题和罗列事实型问题，也包含有定义型问题
- 要求返回的答案必须为精确的答案

■ 度量指标

- 简单事实型问题：系统能正确回答的问题数占全部测试问题总数的比例，即准确率（accuracy）
- 罗列事实型问题：综合了准确率和召回率的F值

$$F = \frac{2 * IP * IR}{(IP + IR)} \quad IP = D / N \quad IR = D / S$$

其中， S 是已知答案的总个数， N 是系统返回的答案个数， D 是返回结果中正确答案的个数



TREC问答评测的任务与性能度量指标

■ TREC-2003

■ 定义型问题的评测方法

- 对定义型问题，评测组织者预先构造了标准答案
- 答案是由被标注为“**vital**”的一些信息块组成
- 对各系统返回的信息块，由评测者进行人工判断
 - 要么被标为“**vital**”：信息对目标的正确解释不可少
 - 要么被标为“**acceptable**”，即尽管不是必不可少的，但对问题目标的解释来说是可以接受的

TREC问答评测的任务与性能度量指标

■ TREC-2003

■ 定义型问题的度量指标

$$F(\beta = 5) = \frac{26 * precision * recall}{25 * precision + recall} \quad recall = r / R$$

$$precision = \begin{cases} 1, & \text{if } len < allowance \\ 1 - \frac{len - allowance}{len}, & \text{otherwise} \end{cases}$$

$$allowance = 100 * (r + a)$$

r: 被判为“**vital**”的信息块个数, *a*: 被判为“**acceptable**”的信息块个数, *len*: 系统返回结果中, 非空白的字符总个数

TREC问答评测的任务与性能度量指标

■ TREC 2004开始

■ 任务设置

- 针对某话题，提出若干个事实型和罗列型问题
- 最后是一个基本类似定义型的“other”问题
- 问题之间不独立

■ 问题举例

■ 度量指标

- 对于定义型问题，度量指标中 β 值被设为3
- 其它并无变化

3	Hale Bopp comet	
3.1	FACTOID	When was the comet discovered?
3.2	FACTOID	How often does it approach the earth?
3.3	LIST	In what countries was the comet visible on its last return?
3.4	OTHER	
21	Club Med	
21.1	FACTOID	How many Club Med vacation spots are there worldwide?
21.2	LIST	List the spots in the United States.
21.3	FACTOID	Where is an adults-only Club Med?
21.4	OTHER	