

信息检索 Information Retrieval



第十章 问答技术

引言

- 当前搜索引擎存在的问题

- 检索需求的表达不够准确

- 用户的检索需求往往是非常复杂而特殊的
 - 关键词的简单逻辑组合很难表达用户的检索需求

- 缺乏语义处理技术的支撑

- 传统信息检索以关键词为基础的索引、匹配算法尽管简单易行，但仅停留在语言的表层，没有涉及语义

引言

● 当前搜索引擎存在的问题

■ 检索结果不够简洁

➤ 信息检索的理想目标是提供用户精确的查询信息

□ 但无论是传统文档信息检索还是Web检索都只提供和用户查询相关的一批文档集合

➤ 文档检索系统返回的相关信息太多，用户很难快速准确地定位到所需的信息

□ 例如，在Google上输入“中国 首都”

– 它有可能返回成千上万个网页（约有205,000,000项符合“中国 首都”的查询结果）

引言

Google

中国 首都

全部 图片 地图 新闻 视频 更多 设置 工具

找到约 205,000,000 条结果 (用时 0.79 秒)

中华人民共和国 / 首都




北京市

北京：意为“北方的首都”，是目前中国的首都，也是明清两个王朝的首都。南京：意为“南方的首都”。明朝初期首都，也是中华民国初期首都。西安：古称“长安”，秦朝、西汉与唐朝的首都。

[中国历代首都之行- 来自维基导游的旅行指南 - 維基導遊 - Wikivoyage](https://zh.wikivoyage.org/zh-hans/中國歷代首都之行)
<https://zh.wikivoyage.org/zh-hans/中國歷代首都之行>

用户还搜索了

还有10+项

						
上海市	香港	慕田峪长城	故宫博物院	东京都	首尔	深圳市

引言

- 当前搜索引擎存在的问题
 - 用搜索引擎查找答案，有时候能成功
 - 例如: *Who was the third prime minister of Australia?*
 - 答案: *Chris Watson* (克里斯·沃森)

引言

Who was the third prime minister of Australia



全部 新闻 图片 视频 地图 更多 设置 工具

找到约 151,000,000 条结果 (用时 0.65 秒)

澳大利亚总理 (3)

克里斯·沃森



用户还搜索了

还有3+项



安德鲁·费希尔 阿瑟·法登 乔治·里德 比利·休斯 埃德蒙·巴顿 约瑟夫·库克 阿尔弗雷德·迪肯

 更多有关“克里斯·沃森”的信息

反馈

third prime minister of australia



全部 图片 新闻 地图 视频 更多 设置 工具

找到约 90,100,000 条结果 (用时 0.86 秒)

澳大利亚总理 (3)

克里斯·沃森



用户还搜索了

还有3+项



安德鲁·费希尔 阿瑟·法登 乔治·里德 比利·休斯 埃德蒙·巴顿 约瑟夫·库克 阿尔弗雷德·迪肯

 更多有关“克里斯·沃森”的信息

反馈

引言

australia's third prime minister

全部

新闻

图片

视频

地图

更多

设置

工具

找到约 85,100,000 条结果 (用时 0.56 秒)

Chris Watson - Wikipedia

https://en.wikipedia.org/wiki/Chris_Watson 翻译此页

John Christian Watson commonly known as Chris Watson, was an Australian politician who served as the third Prime Minister of Australia. He was the first Prime ...

Born:

John Christian Tanck; c. 9 April 1867; Va...

Died:

18 November 1941 (aged 74); Double B...

Political party:

Labor (to 1916); National Labor ...

Children:

1

Early life

Prime Minister in 1904

Later life

References

Prime Minister of Australia - Wikipedia

https://en.wikipedia.org/wiki/Prime_Minister_of_Australia 翻译此页

The Prime Minister of Australia is the head of government of Australia. The individual who holds Percy Spender (Minister for the Army) seated third from the right. The youngest person to become prime minister was Chris Watson – 37, who ...

Deputy Prime Minister

Spouse of the Prime Minister ...

List of Prime Ministers

Chris Watson — Prime Ministers — Australian Prime Ministers

<https://primeministers.moadoph.gov.au/prime-ministers/chris-watson> 翻译此页

Chris Watson became Australia's third prime minister following the resignation of Prime Minister Alfred Deakin, leader of the Protectionist

australia's 3rd prime minister

全部

新闻

图片

地图

视频

更多

设置

工具

找到约 47,200,000 条结果 (用时 0.69 秒)

Prime Minister of Australia - Wikipedia

https://en.wikipedia.org/wiki/Prime_Minister_of_Australia 翻译此页

The Prime Minister of Australia is the head of government of Australia. The individual who holds ... 1 Constitutional basis and appointment; 2 Powers and role; 3 Privileges of office. 3.1 Salary; 3.2 Allowances; 3.3 After office. 4 Acting and interim ...

Salary:

\$538,460 (AUD)

Seat:

Canberra

Deputy:

Michael McCormack

Member of:

Cabinet; National Security Commit...

Acting and interim Prime ...

Former Prime Ministers

List and timeline

Chris Watson - Wikipedia

https://en.wikipedia.org/wiki/Chris_Watson 翻译此页

John Christian Watson commonly known as Chris Watson, was an Australian politician who served as the third Prime Minister of Australia. He was the first Prime Minister from the Australian Labour Party, and led the ... 3rd Prime Minister of Australia. In office 27 April 1904 – 18 August 1904. Monarch, Edward VII. Governor- ...

Born:

John Christian Tanck; c. 9 April 1867; Va...

Died:

18 November 1941 (aged 74); Double B...

Political party:

Labor (to 1916); National Labor ...

Children:

1

Early life

Prime Minister in 1904

Later life

References

克里斯·沃森

前澳大利亚总理，亦是1904年至1904年澳大利亚工党议会议员

生于：1867年

逝世于：1941年

配偶：无

子女数：无

教育：无

曾经担任：无

1910年

引言

● 当前搜索引擎存在的问题

■ 用搜索引擎查找答案，有时候能成功

➤ 例如： *Who was the third prime minister of Australia?*

□ 答案： *Chris Watson*（克里斯·沃森）

■ 用搜索引擎查找答案，经常找不到

➤ 例如：王安石变法和商鞅变法的区别是什么？

□ 答案：社会制度发生了变革

➤ 例如：王安石是怎么死的？

□ 答案：积郁成疾

引言

王安石变法和商鞅变法的区别



全部 图片 新闻 视频 更多

设置 工具

找到约 113,000 条结果 (用时 0.33 秒)

有谁知道商鞅变法和王安石变法的区别?_百度知道

<https://zhidao.baidu.com/question/15584410.html>

2017年7月17日 - 商鞅变法是指战国时(公元前475~公元前221年)商鞅在秦国进行的两次政治改革。商鞅姓公孙,卫国贵族,又称卫鞅或公孙鞅。秦孝公六年(即公元前356年)任用商鞅...

商鞅变法与王安石变法的相同点 ... 2018年5月9日

商鞅变法与王安石变法的相同点是_百度知道 2018年1月2日

王安石变法与商鞅边法异同_百度知道 2017年7月18日

商鞅变法与王安石变法一成一败的原因_百度知道 2009年1月23日

zhidao.baidu.com站内的其它相关信息

商鞅变法和王安石变法的本质区别和再目的上的区别_百度知道

zhidao.baidu.com > 教育/科学 > 人文学科 > 历史学

商鞅变法是站在新兴的地主阶级立场上,代表的是先进生产力的变革,但是他触及了旧势力者即奴隶主阶级的利益,必然为部分人所反对,但本质上强大了秦国。

试比较一下商鞅变法和王安石变法的区别,以及各自的历史进步性和历史

...

<https://www.pin-cong.com/p/13774/?s=18463&c=18688>

所以商鞅新政简直当作救命稻草一样,虽然有副作用,也坚定的执行下去;.宋朝并没有大...儒生的力量非常强大,王安石的变法最后失败了,守旧的力量取得了胜利。

王安石怎么死的



全部 新闻 图片 视频 地图 更多

设置 工具

找到约 2,620,000 条结果 (用时 0.29 秒)

王安石是怎么死的? 北宋大政治家王安石之死 - 趣历史

www.qulishi.com/news/201403/11696.html

2014年3月28日 - 王安石是怎么死的? 元祐元年(1086年)四月六日,66岁的王安石在江宁府(南京)的半山园去世。死亡是一道黑色门槛。王安石死了,这个王朝再也没有...

王安石简介_王安石怎么死的_王安石变法-趣历史网

www.qulishi.com/renwu/wanganshi/

王安石(1021年12月18日-1086年5月21日),字介甫,晚号半山,谥号“文”,世称王文公,自号临川先生,晚年封荆国公,世称临川先生又称王荆公,江西临川盐阜岭(今...

王安石怎么死的,北宋王安石死因揭秘(积郁成疾) - 90后励志网

<https://www.cy1990.com/lizhigushi/mingren/22794.htm>

2018年7月30日 - 在北宋历史上,王安石是一个悲剧型的人物,推行了变法,最终却被罢免。那么,王安石怎么死的,你知道吗?是被奸人陷害而死,还是积郁成疾而死...

主要内容

- 问答技术的发展历史
- 问答技术的分类
- 问答技术的基本体系框架
- 问答技术的评测和实例

问答技术的发展历史

● 问答技术的定义

- 问答式信息检索是一种允许用户 以自然语言方式询问，系统从单语或多语文档集中查找并返回 确切答案或者蕴含答案文本片断 的新型的信息检索方式

- 问答系统允许用户以自然语言的形式查询信息

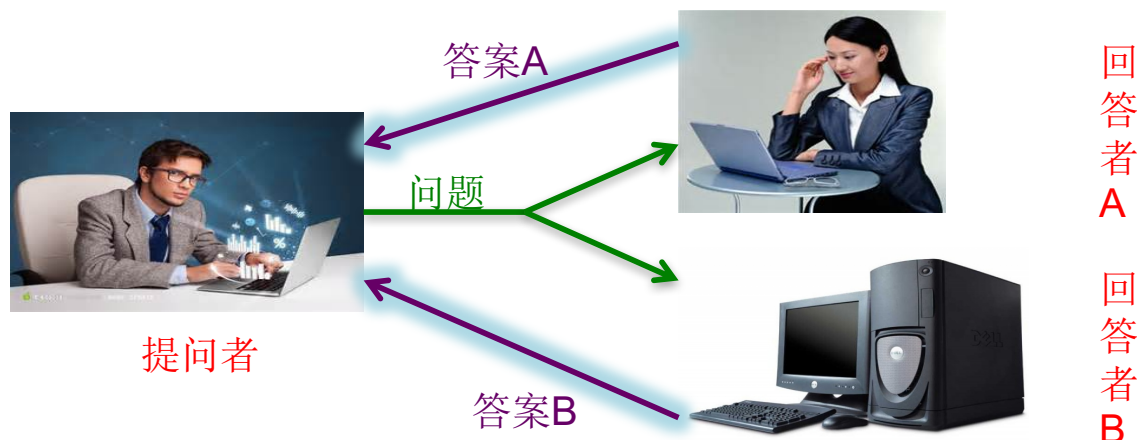
- 例如：世界上最大的宫殿是什么宫殿？

- 系统则直接提供用户准确、简洁的答案

- 例如：紫禁城/故宫

问答技术的发展历史

- “图灵测试” 可以看作是问答系统的蓝图
 - 1950年，英国著名数学家*A.M.Turing*在其论文《Computing Machinery and Intelligence》中提到测试机器是否具有智能的问题：机器能思考吗？
 - 并提出了判定机器能否思考的方法—**图灵测试**



问答技术的发展历史

- 最早的聊天机器人 *ELIZA*

- 1966年，麻省理工学院（MIT）的约瑟夫魏泽鲍姆 (*Joseph Weizenbaum*) 开发
- 临床治疗中模仿心理医生
- 主要的实现技术
 - 关键词匹配寻找存在的知识，并找到相应的答案
 - 人工编写的回复规则

虽然人工智能可能成为可能，但我们永远不应该让计算机做出重要的决定，因为计算机总是缺乏人的品质，如同情心和智慧。

— 《计算机能力与人类理性》，1976

问答技术的发展历史

● 早期两个比较著名的问答系统

■ BASEBALL (1961年)

- 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

■ LUNAR (1973)

- 帮助地质学家方便地了解、比较和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

■ 后台用数据库，保存系统可提供的各种数据

- 用户提问时，系统把用户的问题转换成SQL查询语句，从数据库中查询到数据提供给用户

问答技术的发展历史

● 问答式检索系统

■ *Ask Jeeves*, *AnswerBus*, *START*等

- *Ask Jeeves*虽然接受自然语言提问，但返回的结果还是和提问相关的文章
- *AnswerBus*是一个句子级的多语言的问答系统，对于法语、西班牙语、德语、意大利语或葡萄牙语表述的用户提问，系统返回可能包含答案的8个句子
- *START*直接向用户的自然语言提问提供简洁答案
 - 输入提问: How many people in China ?
 - 系统返回: 1,286,975,468 (July 2003 est.)

问答技术的发展历史

- WolframAlpha (<https://www.wolframalpha.com>)
 - 2009年5月18日正式发布
 - 第一版用约1500万行 $Mathematica$ 代码编写的，在10000个CPU上运行
 - $Mathematica$ ：囊括了计算机代数、符号、数值运算、可视化和统计功能的计算平台和工具包
 - 直接向用户返回答案，而不是像传统搜索引擎一样提供一系列可能包含用户所需答案的相关网页



how many people in china



Assuming "china" is a country | Use as a given

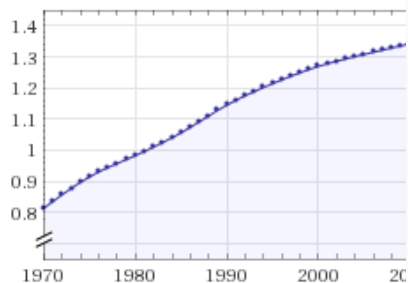
Input interpretation:

China population

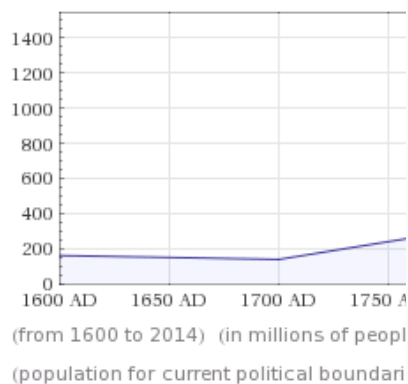
Result:

1.36 billion people (world rank: 1st) (2014 estimate)

Recent population history:



Long-term population history:



Demographics:

Show rates

Show distribution

Show non-metric

population	1.36 billion people (world rank: 1 st) (2014 estimate)
population density	146 people/km ² (people per square kilometer) (world rank: 84 th) (2014 estimate)
population growth	0.426 %/yr (world rank: 167 th) (2013 estimate)
life expectancy	75.8 years (world rank: 114 th) (2013 estimate)
median age	38 years (world rank: 114 th) (2013 estimate)

Age distribution:



Largest cities:

More

city	population
Shanghai	14.01 million people
Beijing	12.46 million people
Zhoukou, Henan	12.07 million people
Nanyang, Henan	11.68 million people
Baoding, Hebei	11.55 million people

(2009 estimates)

Comparisons:

$\approx (0.19 \approx 1/5) \times$ world population (7.13001 billion people)

$\approx 1.1 \times$ current population of India (1.292 billion people)

$\approx 1.2 \times$ current population of Africa (1.10583 billion people)

Sources

Download page

POWERED BY THE WOLFRAM LANGUAGE

Related Queries:

= populations of Shanghai Cooperation Or...

= currency of China

= religion fractions China

= population of China vs Kazakhstan

= full boundary length of China

问答技术的发展历史

● Watson（沃森）

■ 2011年，参加综艺节目《危险边缘》来测试它的能力，第一次人与机器对决

- 2月14日~2月16日的3集节目中，沃森前两轮与对手打平，最后一集中，打败了对手获得了冠军
- 是一个集自然语言处理、信息检索、知识表示、自动推理、机器学习等开放域问答技术的应用
- 硬件

□ 90台IBM Power 750服务器组成的集群服务器，共计2880颗Power7处理器核心及16TB内存（300万美元），每秒可处理500GB的数据（相当于100万本书）



问答技术的发展



● 汪仔机器人

■ 2017年2月6日，搜狗

➤ 江苏卫视《一站到底》

■ 基于深度学习技术

➤ 汪仔的大脑由深度神经网络技术构建

▣ 卷积神经网络、多层循环网络.....

➤ 云端连接上万个服务器进行复杂的运算

➤ 结合知识图谱技术，从检索到的信息中提炼出精准的答案

主要内容

- 问答技术的发展历史
- 问答技术的分类
- 问答技术的基本体系框架
- 问答技术的评测和实例

问答技术的分类

- 根据答案来源的数据类型分类
 - 基于大规模真实文本的问答技术
 - 基于网络文本的问答技术
 - 基于知识库的问答技术
 - 阅读理解式的问答技术

问答技术的分类

● 根据答案来源的数据类型分类

■ 基于大规模真实文本的问答技术

➤ 从预先建立的大规模真实文本语料库中查找答案

□ 类似于 *TREC QA Track*

□ **优点**：能够提供一个优良的算法评测平台，适合对不同问答技术的比较研究

□ **缺点**：不能涵盖用户所有问题的答案

■ 基于网络文本的问答技术

■ 基于知识库的问答技术

■ 阅读理解式的问答技术

问答技术的分类

● 根据答案来源的数据类型分类

- 基于大规模真实文本的问答技术

- 基于网络文本的问答技术

- 从网络文本中查找问题的答案

- **优点**：网络是最大规模的“语料库”，基本涵盖所有提问的答案

- **缺点**：网络是一个动态变化的“语料库”，不适合评价各种问答技术的优劣

- 基于知识库的问答技术

- 阅读理解式的问答技术

问答技术的分类

● 根据答案来源的数据类型分类

- 基于大规模真实文本的问答技术
- 基于网络文本的问答技术
- 基于知识库的问答技术

➤ 从预先建立好的结构化知识库中查找答案

- 优点：在结构化知识库的基础上可以设计出具有较强推理能力的问答技术
- 缺点：如何建立大规模的知识库？基于知识库的问答系统只能限定在特定领域

- 阅读理解式的问答技术

问答技术的分类

● 根据答案来源的数据类型分类

- 基于大规模真实文本的问答技术
- 基于网络文本的问答技术
- 基于知识库的问答技术
- 阅读理解式的问答技术

- 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

问答技术的分类

- 根据问答技术的应用领域分类
 - 基于常问问题集的问答系统 (Frequently Asked Question)
 - 限定域问答系统
 - 开放域问答系统
 - 聊天机器人

问答技术的分类

- 根据问答技术的应用领域分类
 - 基于常问问题集的问答系统 (Frequently Asked Question)
 - 系统在已有的“问题-答案”对的集合中查找与用户问题相匹配的问题
 - 将其对应的答案直接返回给用户
 - 限定域问答系统
 - 开放域问答系统
 - 聊天机器人

问答技术的分类

- 根据问答技术的应用领域分类
 - 基于常问问题集的问答系统 (Frequently Asked Question)
 - 限定域问答系统
 - 用户给出的问题只能限定在有一个特定领域
 - 如：体育、法律、医疗.....
 - 开放域问答系统
 - 聊天机器人

问答技术的分类

- 根据问答技术的应用领域分类
 - 基于常问问题集的问答系统 (Frequently Asked Question)
 - 限定域问答系统
 - 开放域问答系统
 - 用户给出的问题不受任何领域限制
 - 聊天机器人

问答技术的分类

● 根据问答技术的应用领域分类

- 基于常问问题集的问答系统 (Frequently Asked Question)
- 限定域问答系统
- 开放域问答系统
- 聊天机器人
 - 与用户闲聊，通常是人发起话题，机器人根据人说话的内容，进行回复
 - 对话的轮数反映了聊天机器人的性能
 - 不能解决实际问题，更像是一个玩具

主要内容

- 问答技术的发展历史
- 问答技术的分类
- 问答技术的基本体系框架
- 问答技术的评测和实例

问答技术的基本体系框架

● 基于模式匹配的问答技术

苏轼

网页 资讯 视频

百度为您找到相关结果约

苏轼_百度百科





苏东坡生于1037年,卒于1101年,本名苏轼,字子瞻,号“东坡居士”,眉州眉山即今四川眉山人,是北宋著名文学家、书画家、散文家、诗人、词人。苏东坡一生曾两任杭州...

tv.cntv.cn/videoset/C2... - 百度快照



2018年11月29日 - 北宋,1037年,苏轼生于四川眉山的书香之家,19岁考取进士,笔墨了得,与父亲苏洵、弟弟苏辙并列“唐宋八大家”;为政亲民,修西湖、救弃婴,广为世人道;...

news.163.com/18/1129/1... - 百度快照

简介: 苏轼 (1037年1月8日—1101年8月24日), 字子瞻, 又字和仲, 号铁冠道人、东坡居士, 世称苏东坡、苏仙, 汉族, 眉州眉山 (今属四川省眉山市) 人, 祖籍河北栾城, 北宋著名文学家、书法家、画家。嘉祐二年 (1057年), 苏轼进士及第。宋神宗时曾...

[人物生平](#) [主要成就](#) [人物评价](#) [轶事典故](#) [亲属成员](#) [更多>>](#)

baike.baidu.com/ ▾

问答技术的基本体系框架

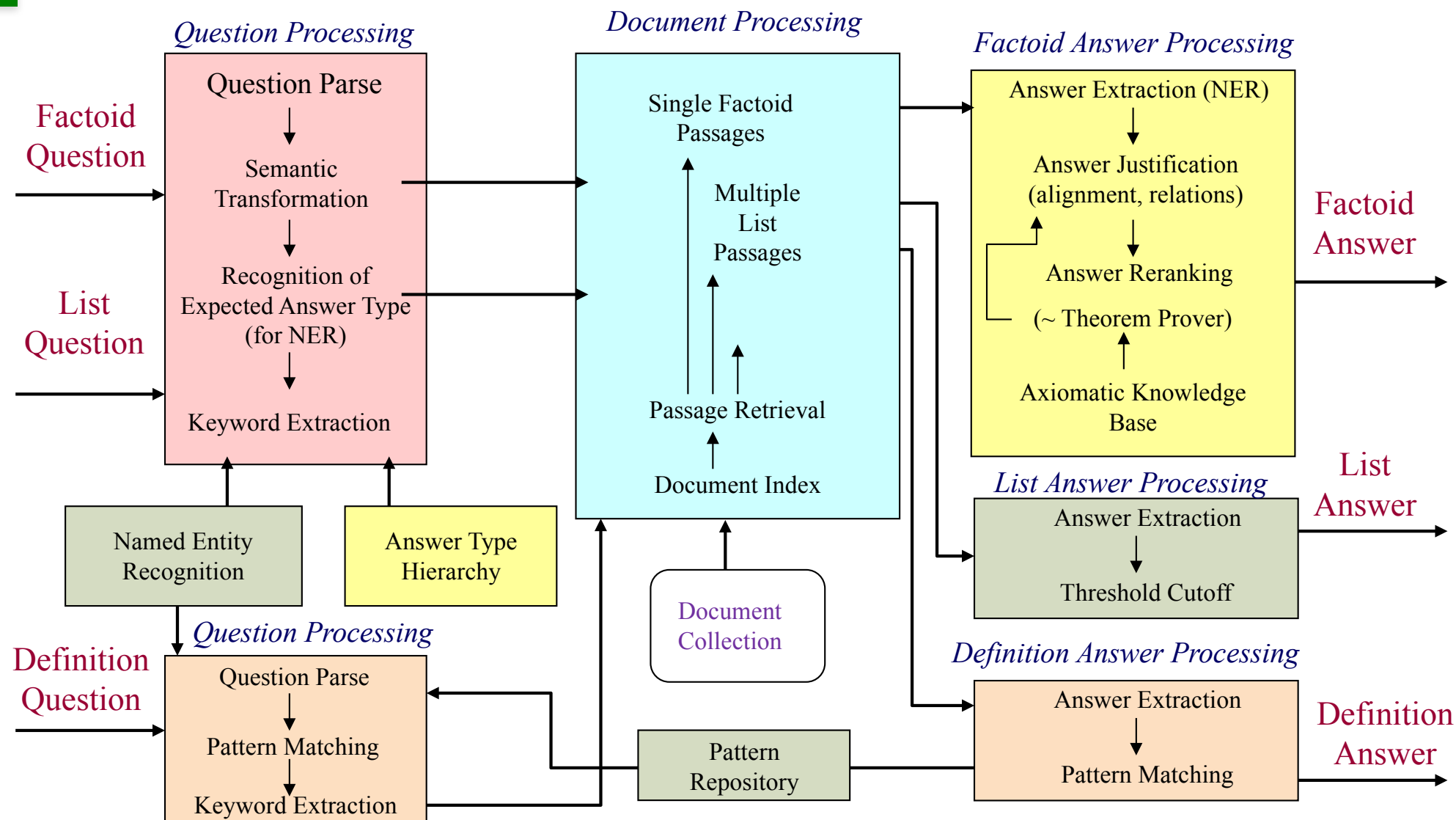
● 基于模式匹配的问答技术

- 离线获取各类问题答案的模式，使用这类问题的所有模式来对抽取的候选答案进行验证
- 基于模式匹配的问答技术中关键问题
 - 如何自动获取某些类型问题尽可能多的答案模式？
 - 例如：询问“某人生日年月日”，部分答案模式：
 - 1.0 <name> (<answer> -)
 - 0.85 <name> 生于 (<answer> -)
 - 0.6 <answer><name>出生于
 -

问答技术的基本体系框架

- 基于自然语言处理的问答技术
 - 运用语言分析技术
 - 问题理解
 - 词法分析、句法分析等
 - 复述、指代及省略识别等
 - 答案抽取
 - 运用信息检索技术
 - 候选答案句检索
 - 候选答案句排序

问答技术的基本体系框架



问答技术的基本体系框架

- 基于自然语言处理的问答技术
 - 基于*NLP*的问答技术主要包括三个部分
 - 问题理解
 - 对用户的问题进行语义层次的分析 and 理解
 - 问题分类
 - 问题的扩展
 - 问题的形式化转换
 -
 - 段落检索
 - 答案抽取

问答技术的基本体系框架

● 基于自然语言处理的问答技术—问题理解

■ 问题分类

➤ 确定问题的答案类型（人物、地点、时间……）

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

问答技术的基本体系框架

● 基于自然语言处理的问答技术—问题理解

■ 问题分类

➤ 中文问题分类体系

大类(Coarse)	小类(Fine)
人物(HUM)	特定人物 团体机构 人物描述 人物列举 人物其他
地点(LOC)	星球 城市 大陆 国家 省 河流 湖泊 山脉 大洋 岛屿 地点 列举 地址 地点其他
数字(NUM)	号码 数量 价格 百分比 距离 重量 温度 年龄 面积 频率 速度 范围 顺序 数字列举 数字其他
时间(TIME)	年 月 日 时间 时间范围 时间列举 时间其他
实体(OBJ)	动物 植物 食物 颜色 货币 语言文字 物质 机械 交通工具 宗教 娱乐 实体列举 实体其它
描述(DES)	简写 意义 方法 原因 定义 描述其它
未知(Unknown)	未知

问答技术的基本体系框架

● 基于自然语言处理的问答技术—问题理解

■ 问题分类

➤ 基于规则的问题分类方法

- 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
- 例如，对于“国家(country)”类别
 - 可以设计规则 “*<哪>一个国家？”
 - 当问题的后部包含“哪一个国家”时，该问题就被分到“国家”类别
- 应用规则的方法比较简单，要耗费大量人力去设计规则，另外对有些问题很难设计规则去覆盖

问答技术的基本体系框架

● 基于自然语言处理的问答技术—问题理解

■ 问题分类

➤ 基于统计的问题分类方法

- 人工方式对一批问题的类别进行标注（训练语料库）
- 设计机器学习算法，对已标注的这个训练集进行分类模型的自动训练
 - 贝叶斯分类、决策树、*SVM*、*CNN*
 - 可以使用词、词性、句法、疑问词等特征
 - 使用训练得到的模型对测试问题进行自动分类

问答技术的基本体系框架

● 基于自然语言处理的问答技术—问题理解

■ 问题分类

➤ 问题分类的特点

- 对于文本，问题一般很短，其中包含的词很少，因此可以利用的特征少
- 在问题所包含的词中，决定问题类别的特征具有明显的倾向性，即只有某些词才是真正确定问题类别的主要特征

问答技术的基本体系框架

● 基于自然语言处理的问答技术—问题理解

■ 问题扩展

- 对问题中的关键词进行扩展，提高返回结果的召回率
 - “充电宝”和“移动电源”互为同义词
- 通过同义词词典或复述词典对问题进行扩展
 - 同义词：上下文语境无关
 - 复述词：上下文语境相关

问答技术的基本体系框架

- 基于自然语言处理的问答技术—问题理解

- 问题的形式化转换

- 示例

- “When was the paper clip invented?”

- “The paper clip was invented”

- “When did Nixon visit China?”

- “Nixon visited China” “Nixon did visit China”

- 转化后的查询质量直接决定了相关文档返回的质量

- 转化过程中还需要对关键词进行提取（搜索引擎利用关键词进行文档检索）

- 关键词的质量也会影响结果的质量

问答技术的基本体系框架

- 基于自然语言处理的问答技术—段落检索
 - 相关文档检索
 - 检索那些文档中可能包含答案
 - 方法与前面介绍的信息检索方法相同
 - 片段检索
 - 需要将检索获得的文档拆分成片段或者句子
 - 减少答案抽取所需处理的内容长度

问答技术的基本体系框架

● 基于自然语言处理的问答技术—答案抽取

■ 以词或者短语作为答案

➤ 处理起来相对简单一些。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

□ 例如：“中华人民共和国是什么时候成立的？”

□ “自从 1 9 4 9 年 1 0 月 1 日中华人民共和国成立以来至 1 9 9 4 年底止，我国已经同世界上的约 1 6 0 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。”

➤ 所需要的答案只是这句话中的一小部分，如果把整句话作为答案都提交给用户，显然冗余信息太多

问答技术的基本体系框架

● 基于自然语言处理的问答技术—答案抽取

■ 答案类型的统计

➤ 从搜索网站的日志中共提取5400多个问题

▣ 其中很多问题省略了疑问词、表达模糊、要求回答的是完成某件事的程序而非简短答案等等

- 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？

➤ 从中可以看出，绝大多数是简述型问题，超过90%

问答技术的基本体系框架

● 基于自然语言处理的问答技术—答案抽取

■ 以文摘作为答案

➤ 对于有些问题，通过简短的一个短语或者一句话很难给出答案

□ 例如：“9.11事件是怎么回事？”

➤ 关于这个问题，在互联网上有许多相关的报道，如果能把这些相关报道做成一个简短的**文摘**，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

问答技术的基本体系框架

● 基于自然语言处理的问答技术

■ 其他相关技术

- 短语结构分析或依存分析
- 词汇链
- 复述（paraphrase）
- 指代消解
- 文本蕴含
- 复杂问句分解
-

问答技术的基本体系框架

● 基于自然语言处理的问答技术

■ 其他相关技术—短语结构分析或依存分析

- 得到句子的短语结构句法树或依存句法树
- 在句子排序或答案抽取阶段，使用句法信息

□ 问题: 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.

- 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。
- 引入句法信息，系统只会返回答案 **Jack Ruby**

问答技术的基本体系框架

● 基于自然语言处理的问答技术

■ 其他相关技术—词汇链

- 很多情况下，问题的关键词和文本关键词不一致，但表达的意思相同
- 利用 *WordNet* 构建词汇链，连接问题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*的宾语是*die*的主语
 - $kill(e, x1, x2) \rightarrow cause(e1, x1, e2) \ \& \ die(e2, x2)$

问答技术的基本体系框架

● 基于自然语言处理的问答技术

■ 其他相关技术—复述 (paraphrase)

- 指用不同的词汇-句法结构表达同样的意思
- 可以解决因问题和答案的表述不同给问答系统的设计带来的麻烦
- *When did Colorado become a state?*
 - Colorado became a state in 1876.
 - Colorado was admitted to the Union in 1876.
- *Who killed Abraham Lincoln?*
 - John Wilkes Booth killed Abraham Lincoln.
 - John Wilkes Booth ended Abraham Lincoln's life with a bullet.

词汇	词汇复述结果	词汇	词汇复述结果
开掘	挖掘、发掘、挖	少许	少量、一点点
伙食团	食堂	翼子板	叶子板
从军	当兵、参军、服役、入伍	急挫	狂跌、大跌
脸面	面子、颜面	动漫	动画、漫画、动画片
百货店	百货商店、商场、杂货店、量贩店、百货公司	牲畜	畜禽、禽畜、家畜、牲口、畜
声势	气势、威望、声望	情节	桥段
谋杀案	凶杀案、杀人案	此次	本次、这次
皇帝	帝王、帝、皇上	展馆	展厅、展览厅、展区
啜泣	抽泣、哭泣、呜咽	年代久远	古老、年深日久
疾风暴雨	暴风骤雨、急风暴雨、狂风暴雨	禁不住	忍不住、情不自禁、不禁、不由得

短语	短语复述结果
铁窗生涯	牢狱生涯、牢狱生活、监狱生涯
季后赛揭幕战	季后赛首战
出生日期	出生时间、诞辰、生辰、生日、出生年月日、出生年月、诞生日期
人大教育科学文化卫生委员会	人大教科文卫委员会、人大教科文卫
仍屹立不倒	仍然巍然耸立、至今依然屹立、仍然屹立不倒、依然屹立不倒
悍然攻打	悍然进攻、悍然打击
西方国家政客	西方政客、西方国家政客、西方国家政要、西方政要、西方政治家、西方政治人物、西方国家政界人士
薪酬结构	工资结构、薪金结构、薪资结构
迈阿密热火队	热火队、迈阿密热队

问答技术的基本1

- 基于自然语言处理
 - 其他相关技术一指代
 - 指代消解

*he*指代*Ken Cuccinelli*

*them*指代
*Terry McAuliffe*和*Ken Cuccinelli*

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q₁: What are the candidates **running** for?

A₁: Governor

R₁: The Virginia governor's race

Q₂: **Where**?

A₂: Virginia

R₂: The Virginia governor's race

Q₃: Who is the democratic candidate?

A₃: **Terry McAuliffe**

R₃: Democrat Terry McAuliffe

Q₄: Who is **his** opponent?

A₄: **Ken Cuccinelli**

R₄: Republican Ken Cuccinelli

Q₅: What party does **he** belong to?

A₅: Republican

R₅: Republican Ken Cuccinelli

Q₆: Which of **them** is winning?

A₆: Terry McAuliffe

R₆: Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May

问答技术的基本体系框架

● 基于自然语言处理的问答技术

■ 其他相关技术—文本蕴含

➤ 自然语言文本中的单向推理关系

➤ 多源答案验证

▣ 答案的来源多种多样（知识库、普通文本、常问问题集）

▣ 以统一的标准对不同来源的结果给出置信度

✧ T1: 丁磊1997年5月创立网易公司。

✧ H1: 丁磊是网易公司的创办人。（Positive Textual Entailment）

✧ T2: 丁磊1997年5月创立网易公司。

✧ H2: 丁磊不是网易公司的创办人。（Contradiction）

✧ T3: 丁磊1997年5月创立网易公司。

✧ H3: 丁磊是个中国人。（Unknown Entailment, Neutral）

问答技术的基本体系框架

- 基于自然语言处理的问答技术

- 其他相关技术—复杂问句分解

- 通常用户的问题形式比较复杂，里面会包含多个子问题

- 姚明哪年出生的？身高多少？（并列式）

- 二次大战期间美国总统是谁？（嵌套式）

- 包括问句的分解和子问句的生成

主要内容

- 问答技术的发展历史
- 问答技术的分类
- 问答技术的基本体系框架
- 问答技术的评测和实例

问答技术评测

● TREC问题样例

1. Who is the author of the book, "The Iron Lady: A Biography of Margaret Thatcher"?
2. What was the monetary value of the Nobel Peace Prize in 1989?
3. What does the Peugeot company manufacture?
4. How much did Mercury spend on advertising in 1993?
5. What is the name of the managing director of Apricot Computer?
6. Why did David Koresh ask the FBI for a word processor?
7. What debts did Quintex group leave?
8. What is the name of the rare neurological disease with symptoms such as: involuntary movements (tics), swearing, and incoherent vocalizations (grunts, shouts, etc.)?

问答技术评测

● TREC评测

TREC-8 ~ TREC-12 QA Track发展情况

	TREC-8（1999）	TREC-9（2000）	TREC-10（2001）	TREC-11（2002）	TREC-12（2003）
任务类型	Main Task	Main Task	Main/List/ Context Task	Main/List Task	Main/Passage Task
测试提问数	200	693	500/50/10 Series	500/25	500
测试集来源	FAQ Finder日志 QA参赛者 TREC Team	Encarta/ Excited日志	MSNSearch/ Ask Jeeves日志		---
测试集真实度	部分提问采用逆 构造法产生，故缺 乏一定的真实度	测试集都是从自然语言检索系统的日志中提取出来的，只进行了适当的修正			
提问类型	每个提问都能从语料库中找到答案		Main Task中有的提问没有答案，此时系统应返回NIL		
答案个数	每个提问给出按概率大小排列的5个答案			每个提问只给出一个答案	
答案长度	≤50字节			准确答案	
	≤250字节				
评测指标	Main Task: MRR	List/Context Task: Accuracy		CWS	
最佳系统性能	66.0%	58%(50byte) 76%(250byte)	77%	85.6%	---

问答技术评测

● TREC评测

■ Main Task（主任务）

- 主要测试系统对**基于事实、有简短答案**的问题的处理能力
 - Where is Belize located?
 - What type of bridge is the Golden Gate Bridge?
- 那些需要总结、概括的问题不在测试之列
 - 如何办理出国手续?
 - 如何制作网页?
 - 如何赚钱?

问答技术评测

● TREC评测

■ List Task（列举任务）

- 要求系统列出满足条件的几个答案
- 在 *TREC2003* 之前，*Track* 要求被测试系统给出 **不少于给定数目** 的实例
 - Name 22 cities that have a subway system
- TREC2003 要求系统要给出 **尽可能多实例**
 - List the names of chewing gums

问答技术评测

● TREC评测

■ Context Task（语境任务）

➤ 测试系统对上下文的理解和把握，问题 i 的回答基于对问题 j ($i > j$) 的理解和把握

- 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?
- 问题A对问题B和C进行答案检索很重要

问答技术评测

- TREC评测

- Passage Task（语块任务）

- TREC2003 QA Track提出的新任务

- 与其他任务不同：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串

- a small chunk of text that contains an answer

问答技术评测

- TREC QA的评价指标

- TREC QA Track的评测指标主要包括

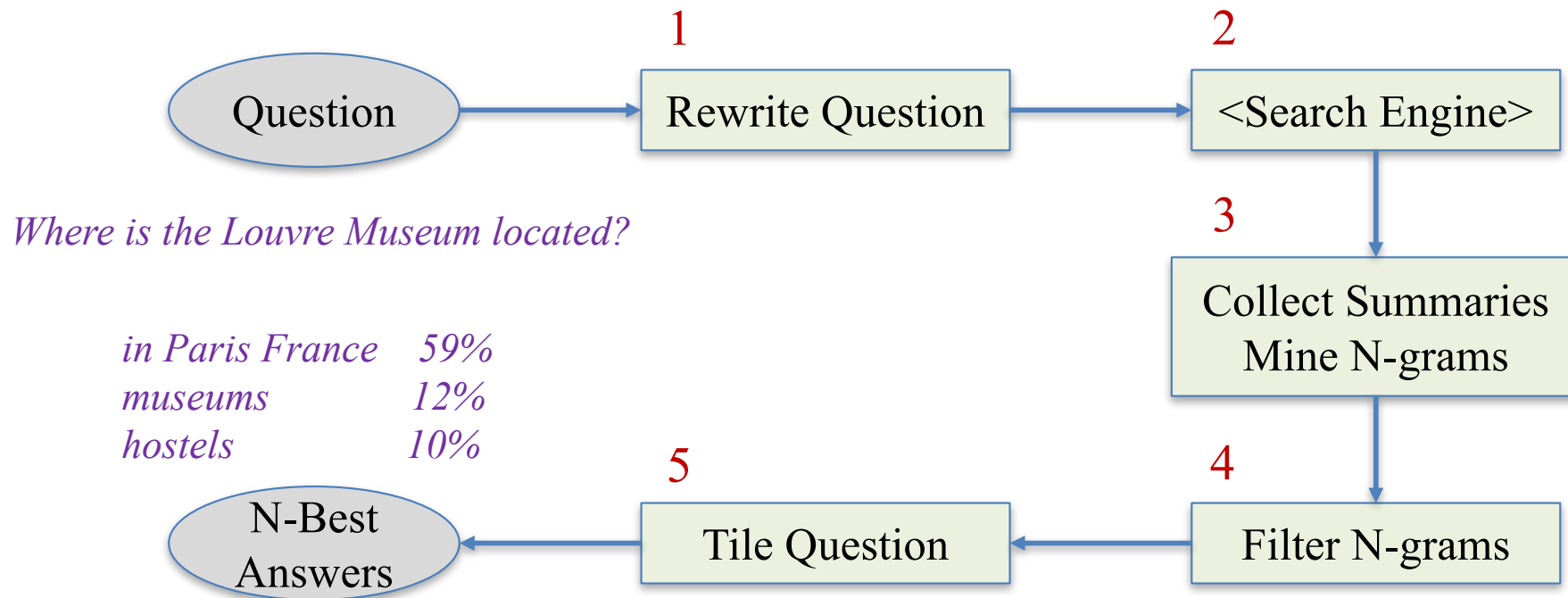
- 平均排序倒数 (Mean Reciprocal Rank, MRR)
 - 准确率 (Precision)
 - 置信度 (Confidence Weighted Score)

$$CWS = \frac{1}{N} \sum_{i=1}^N \frac{\text{前}i\text{个问题中被正确回答的问题数}}{i}$$

- 公式中的N表示测试集中的提问个数

问答系统实例

● AskMSR



问答系统实例

● AskMSR

■ Step 1. 问题改写 (Rewrite Question)

➤ 用户的问题常常和包含答案的句子在句法上很相似

- Where is the Louvre Museum located?
- The Louvre Museum is located in **Paris**
- Who created the character of Scrooge?
- **Charles Dickens** created the character of Scrooge.

➤ 将问题分类7类

- **Who is/was/are/were...?**
- **When is/did/will/are/were ...?**
- **Where is/are/were ...?**

问答系统实例

● AskMSR

■ Step 1. 问题改写 (Rewrite Question)

➤ 根据类别制定问题改写规则

□ 例如：针对地点类问题，将 ‘is’ 移动到所有可能的位置

- “Where **is** the Louvre Museum located”
- → “**is** the Louvre Museum located”
- → “the **is** Louvre Museum located”
- → “the Louvre **is** Museum located”
- → “the Louvre Museum **is** located”
- → “the Louvre Museum located **is**”

➤ 根据预期的答案类型来确定问题的类别

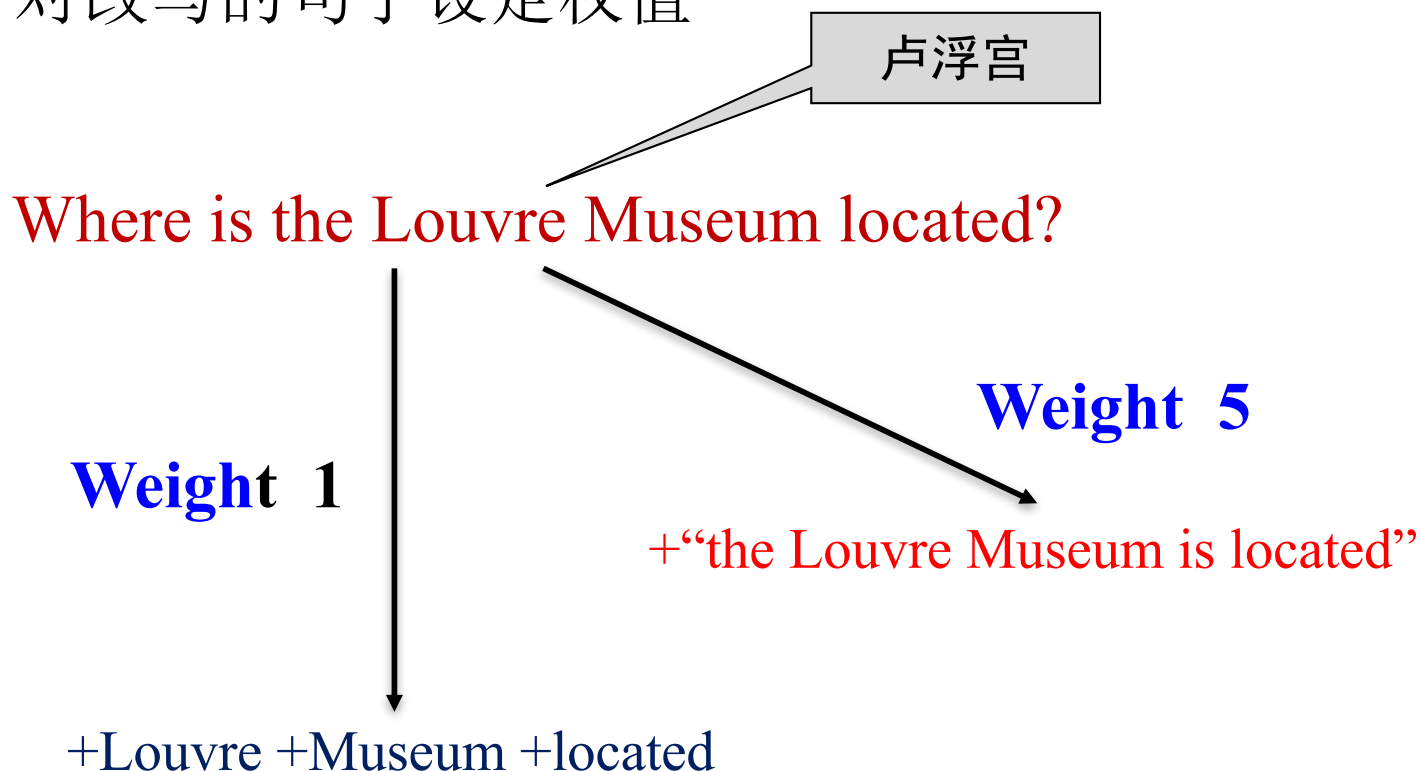
□ Date, Person, Location, ...

问答系统实例

● AskMSR

■ Step 1. 问题改写 (Rewrite Question)

➤ 对改写的句子设定权值



问答系统实例

● AskMSR

■ Step 2. 调用搜索引擎

- 将所有改写后的问题提交给搜索引擎
- 找到top N答案（100？）
- 为了提高速度，可以只依赖“snippets”而不是全文

Louvre Museum - Facts and tips - Come to Paris

<https://www.cometoparis.com/paris-guide/paris.../louvre-museum-s927> ▼ 翻译此页

Louvre Museum is one of the largest museums in the world, **located in** the center of Paris, in the 1st district. Nearly 35,000 objects from prehistory to the 21st ...

Le Musee du Louvre - The Louvre Museum - DiscoverFrance.net

www.discoverfrance.net/France/Paris/Museums-Paris/Louvre.shtml ▼ 翻译此页

Louvre {looov'-ruh} — a French palace and the national art **museum** of France. **Located in** Paris, **the Louvre** is one of the largest palaces in the world and, as a ...

问答系统实例

● AskMSR

■ Step 3. 挖掘 *N*-grams

- *N*-gram: 在一个序列中 *N* 个相邻的词汇
- 例: “*Web Question Answering: Is More Always Better*”

□ Unigrams

- *Web, Question, Answering, Is, More, Always, Better*

□ Bigrams

- *Web Question, Question Answering, Answering Is, Is More, More Always, Always Better*

□ Trigrams

- *Web Question Answering, Question Answering Is, Answering Is More, Is More Always, More Always Betters*

问答系统实例

● AskMSR

■ Step 3. 挖掘*N-grams*

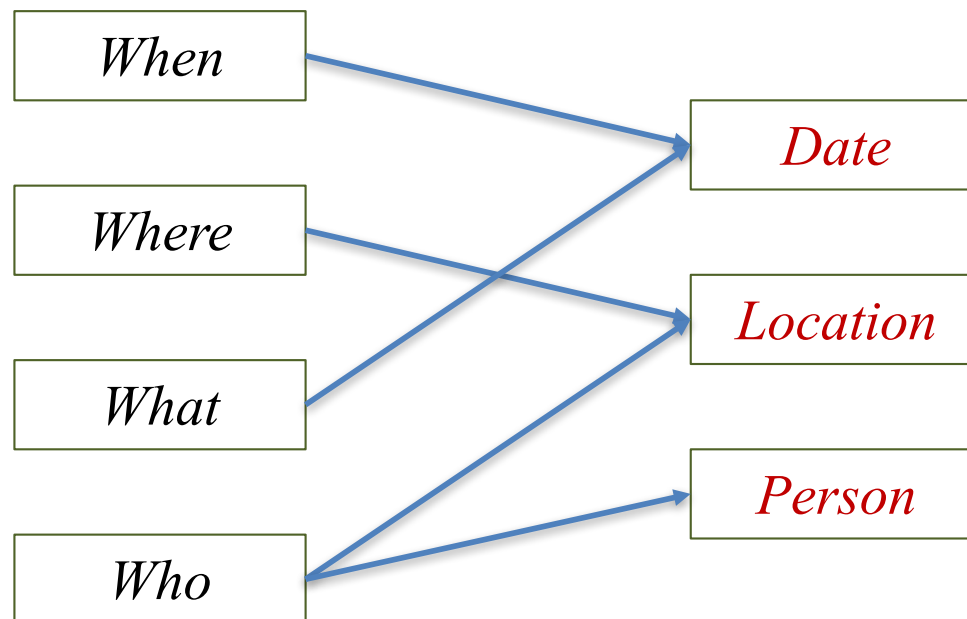
- 列举出全部检索到的*snippets*中的*N-grams* (N=1, 2, 3)
- 一个*N-gram*的权重取决于出现的次数
- 例: “Who created the character of Scrooge?”
 - *Dickens* – 117
 - *Christmas Carol* – 78
 - *Charles Dickens* – 75
 - *Disney* – 72
 - *Carl Banks* – 54
 - *A Christmas* – 41
 - *Mr. Charles* - 10

问答系统实例

- AskMSR

- Step 4. 过滤*N-grams*

- 每一个问题类型和一个或多个类别相联系



问答系统实例

- AskMSR

- Step 5. 获得答案

分值

75

Charles Dickens

117

Dickens

10

Mr. Charles

合并,



分值 202

Mr Charles Dickens

问答系统实例

● AskMSR

■ 评测结果

➤ 标准TREC 测试语料

□ ~1M 个文档; 900 个问题

➤ 在超过20个参加机构中排名第9

□ $MRR = 0.262$ (相当于在第4-5个候选答案中命中)

➤ 使用 Web文档, 而不是TREC's 1M documents

□ $MRR = 0.42$ (相当于在第2-3个候选答案中命中)

本章小结

- 了解问答技术的发展历史
- 了解问答技术的分类
- 掌握基于自然语言处理的问答技术基本流程及方法
- 了解问答技术的相关评测，根据给定的实例对问答技术有更好的理解