深度问答技术初探

李超 cli@mobvoi.com

提纲

- 简介深度问答
- 深度问答技术框架
- 如何快速搭建一个深度问答系统
- 大公司和创业公司技术选型

Deep

DeepQA deeply analyzes natural language input to better find, synthesize, deliver and organize relevant answers and their justifications from the wealth of knowledge available in a combination of existing natural language text and databases.

Not deep learning

什么是深度问答

目标：

对用户输入自然语言query进行理解，并给出问题的精准答案

依赖技术：

- Query理解（分析、改写、推理）
- 文本分类
- 实体识别技术
- 语义相似性计算
- 信息检索技术
- Learning to rank
- ...

利用资源：

- 知识图谱
- 问答对
- 海量自然语言文本
-

深度问答 vs 搜索引擎

	DeepQA	Search engine
输入	自然语言问句	关键词组合
输出	精准答案	问答列表
query需求	需求明确	可以是泛需求
应用范围	相对更窄	广泛

答案三种来源

- 从问题库中获取答案
  - 手表蓝牙断连怎么办
- 从结构化知识图谱获取答案
  - 北京大学校长是谁
- 从无结构化文本中抽取答案
  - 杀了兄弟夺取皇位的皇帝

中文名	北京大学
英文名	Peking University [6]
简　称	北大、PKU [6]
创办时间	1898年（戊戌年）7月3日 [7]
类　别	公立大学、全国重点大学
学校类型	综合类
属　性	九校联盟 [8]
	985工程
	211工程
	2011计划 [4]
所属地区	中国-北京
现任校长	林建华
知名校友	李克强、邓稼先、朱自清、屠呦呦、李彦宏等

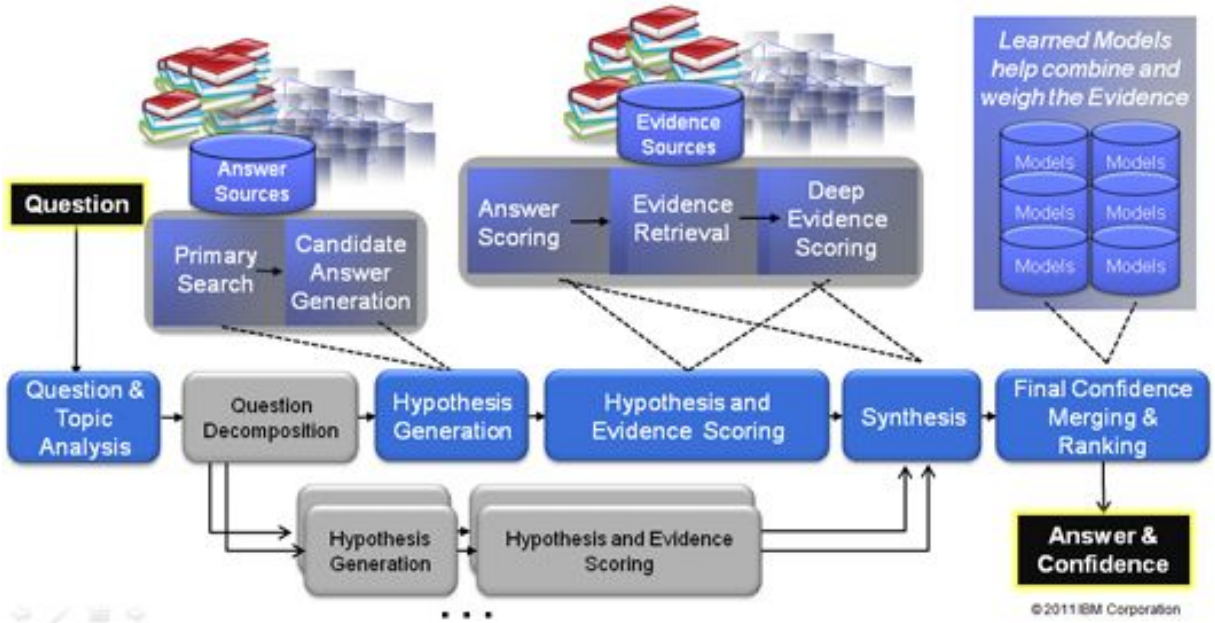
2013年7月19日 - 而他们之所以不顾手足之情、争得你死我活,则完全是为了夺取帝王之位。例如,唐代初年,李世民在玄武门之变中杀掉长兄李建成与四弟李元吉,成为皇位继承人...

李世民杀兄弟当皇帝是真的吗

答：当然是真的，历史上著名的玄武门之变 "玄武门之变"是唐高祖武德九年六月初四（公元626年7月7日）由当时的天策上将、唐高祖李渊的次子秦王李世民在唐王朝的首都长安城（今陕西省西安市）大内皇宫的北宫门——玄武门附近发动的一次流血政变。 在李家...

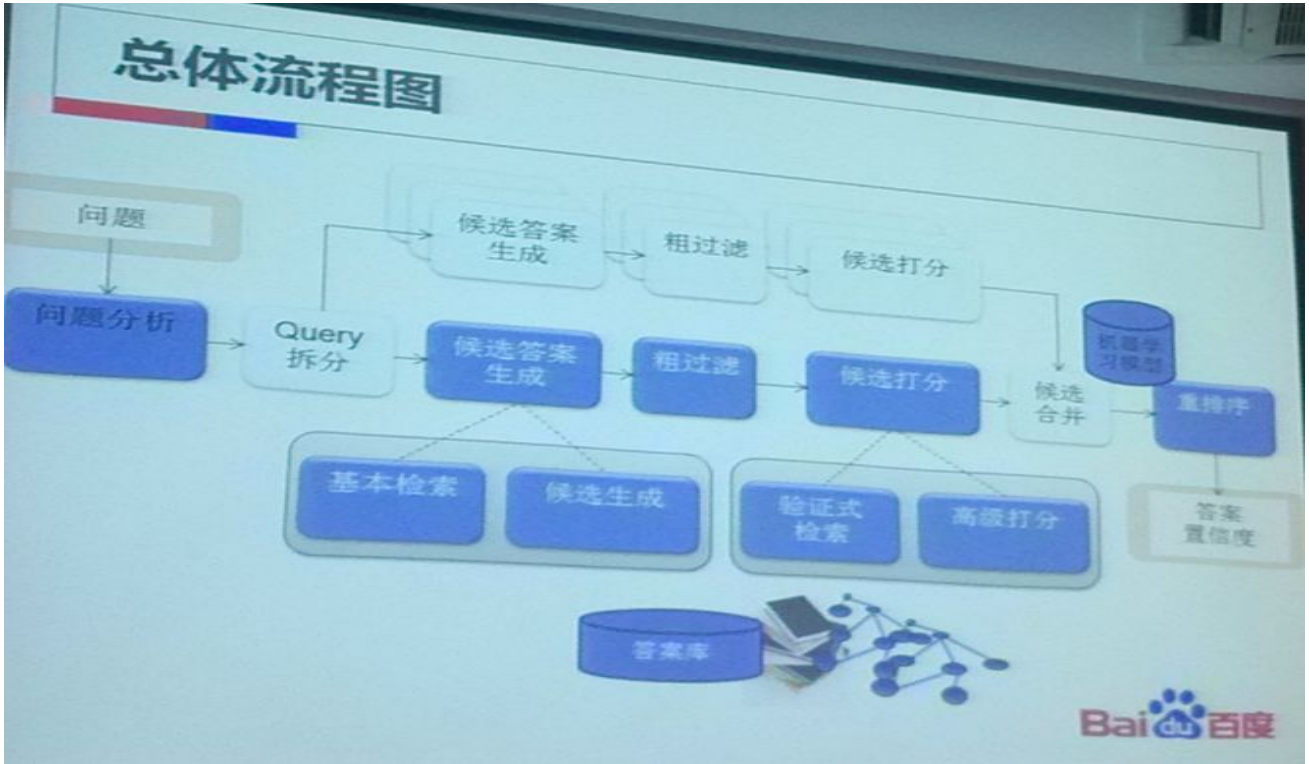
深度问答技术框架

IBM waston 技术框架



深度问答技术框架

百度DeepQA技术框架

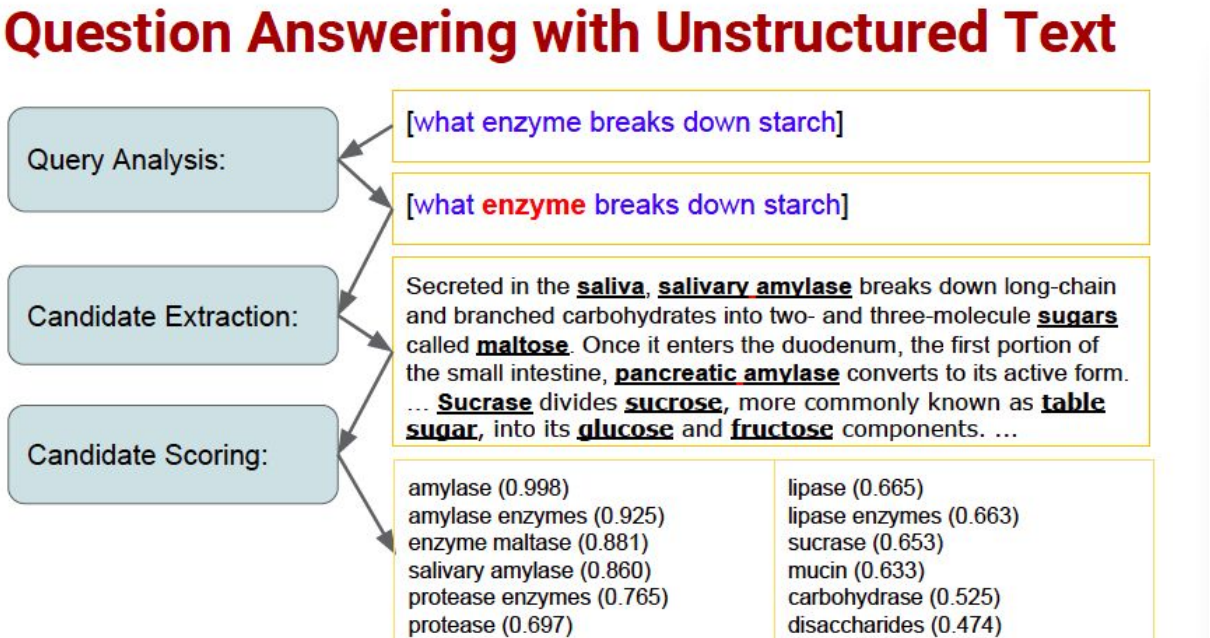


深度问答技术框架

DeKang Lin技术框架

Use IR to identify top documents/passages

Extract answers from the retrieved text



Query分析

- Query理解（分析、改写、推理）
- 问题类型
  - 问题角度：what/when、how/why、whether、
  - 答案角度：实体、数字、短观点
- Lexcial answer type
  - 第一个发现美洲大陆的航海家
- Keywords and weights
  - 第一个发现美洲大陆的航海家
- Keywords and relations
  - 第一个->航海家、美洲->大陆、发现->大陆

候选答案抽取

- 搜索技术
- NER
- 实体链接
- Pattern抽取

答案排序

features:

- LAT和候选答案一致性
- 候选答案和query keyword的距离
- 候选答案的出现次数
- 候选答案所在的条目位置
- 候选答案所在的条目 质量
- 答案和问题的相关性

问问问答系统

- 问题对
  - 手表相关、热点事件运营
- 结构化知识图谱
  - Kgbase QA
- 无结构化文档
  - 自然语言问答

技术问题

- 问题分析
  - Query理解（分析、改写、推理）
  - 问题分类类型
  - Lexical answer type
  - Keywords and their weights
  - Keywords and their relations
- 候选答案获取
  - 通用搜索（全网索引）
  - 通用领域的NER
  - 实体链接
- 答案排序
  - LAT和候选答案一致性
  - 候选答案和query keyword的距离
  - 候选答案的出现次数
  - 答案和问题的相关性

技术问题

- 问题分析
  - Query理解（分析、改写、推理）
  - 问题分类类型
  - Lexical answer type
  - Keywords and their weights
  - Keywords and their relations
- 候选答案获取
  - 通用搜索（全网索引）
  - 通用领域的NER
  - 实体链接
- 答案排序
  - LAT和候选答案一致性
  - 候选答案和query keyword的距离
  - 候选答案的出现次数
  - 答案和问题的相关性

办法总比问题多

利用搜索引擎

技术问题

- 问题分析
  - Query理解（分析、改写、推理）
  - 问题分类类型
  - Lexical answer type
  - Keywords and their weights(全部是1)
  - Keywords and their relations(全部相同)
- 候选答案获取
  - 通用搜索（借助搜索引擎全网索引）
  - 通用领域的NER(词表匹配)
  - 实体链接（暂时不用）
- 答案排序
  - LAT和候选答案一致性(词表匹配)
  - 候选答案和query keyword的距离
  - 候选答案的出现次数
  - 答案和问题的相关性（暂时不用）

效果展示

- 测试集准确率
  - 87%
- 真实query log结果统计
  - entityQA 85.1%
  - ticwatch召回率4‰左右

效果展示

数据资源

- 支持更多的Answer Type
- 热点事件运营
- 手表相关问题
- 车载相关问题

改进

- 传统方法
  - 问句相似性计算优化LAT识别
  - LAT和答案一致性计算
  - 问题答案相关性计算
  - 候选答案排序改成LTR
  - 答案归一化
- Deep Learning

创业公司和大公司技术选型

- 数据
  - 海量 vs 稀缺
- 技术储备
  - 丰富 vs 缺乏
- 人员
  - 团队 vs 一到两人
- 时间
  - 半年以上 vs 一个季度之内

创业公司和大公司技术选型

- 各模块选最简单的方法
- 快速搭建出base line版本
- 扩大产品影响力
- 争取资源、迭代完善

一点总结

学术是学术，应用是应用

想办法利用资源

取法其众

如无必要 勿增实体

QA about DQA

谢谢