问答系统中的知识图谱

Knowledge Graphs for Question Answering

沈李斌
joint work with colleagues at Mobvoi
2014-10-17

概要

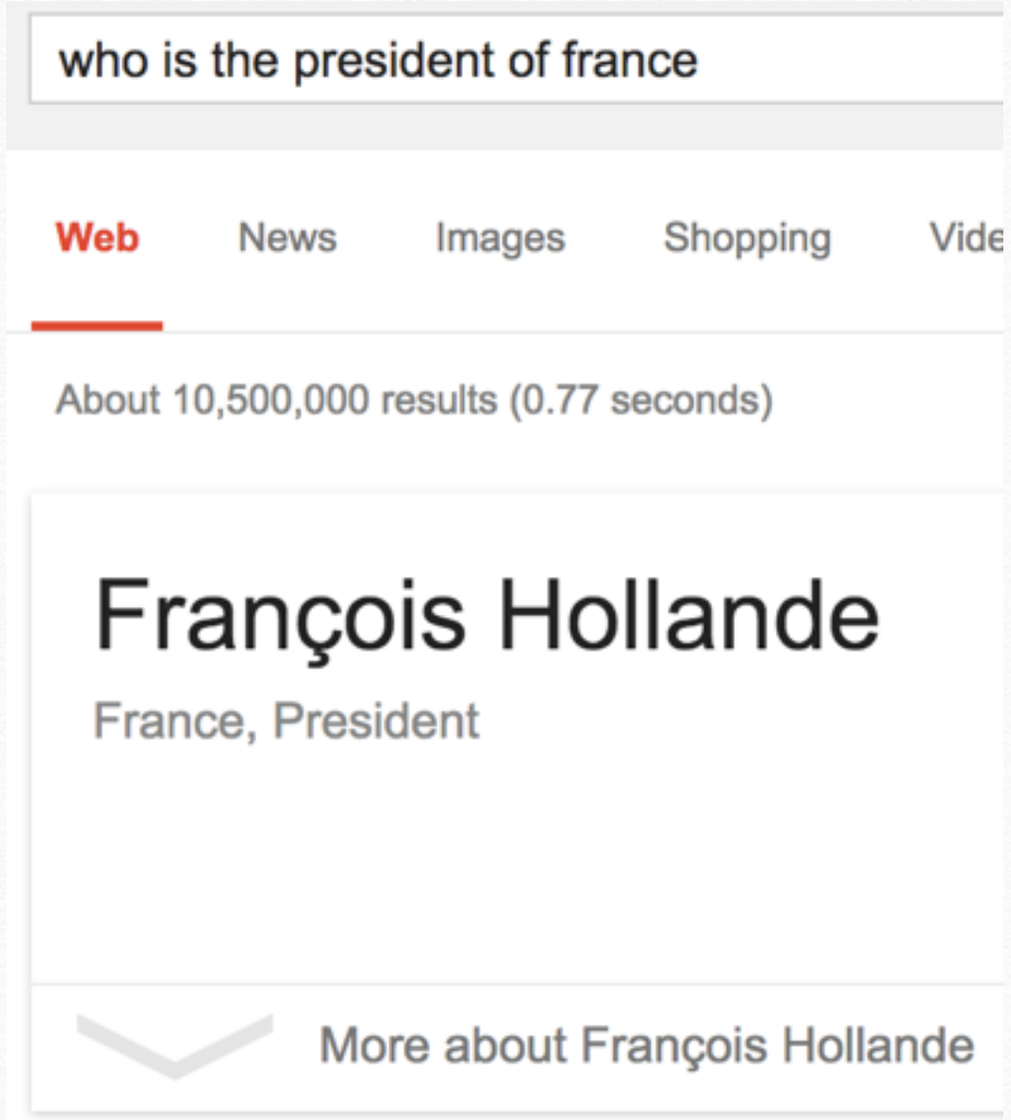
- 问答系统简介
- Watson系统中的知识图谱及使用
- 出门问问的知识图谱及使用

问答系统

- IBM Watson in Jeopardy!
  - 受限的百科问答
  - 知识主要依赖wikipedia
  - 非结构化文本分析 + 结构化领域知识
  - 约70个子模型
  - IR+NLP+ML

问答系统

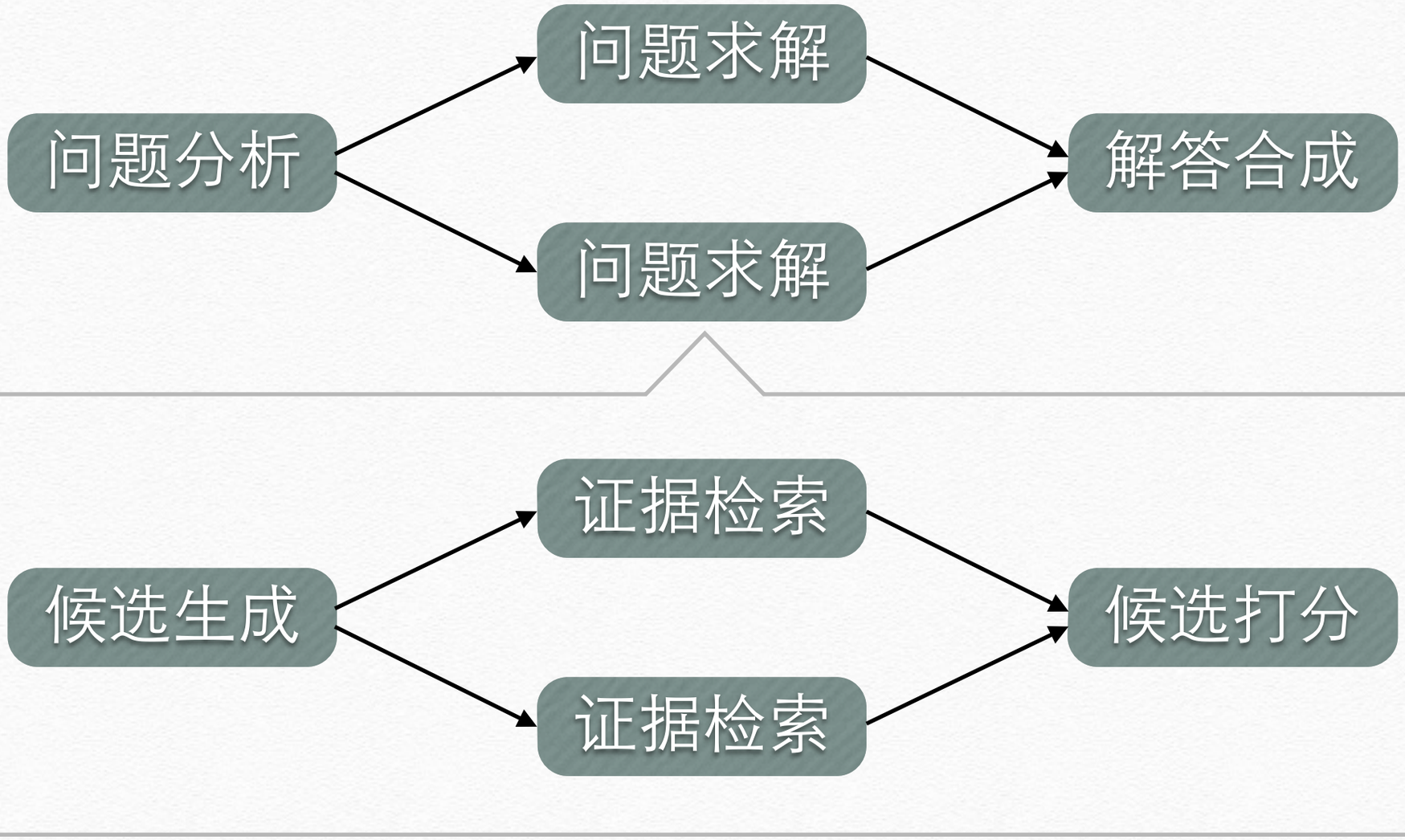
- Google Now
  - 58%的搜索结果使用了结构化的输出
  - 结构化输出准确率达到约88%
  - 依赖Google知识图谱
  - 5亿objects，35亿facts和relations（2012/05）



问答系统

- 出门问问
  - 智能移动语音搜索
  - 手机、Android wear、Glass、车载、微信
  - POI、导航、娱乐、查询等60多个生活领域
  - ASR + NLP + IR+ML

问答系统

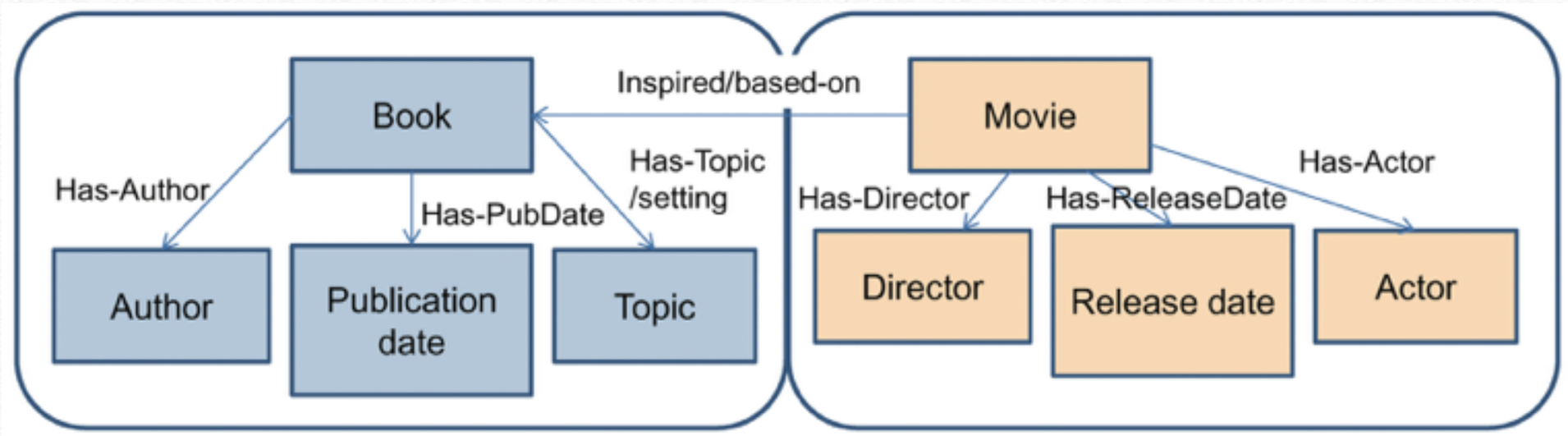


Watson

- 结构化知识的用途
  - 生成候选解答
  - 提供证据支持
- 结构化知识的种类
  - 大规模实体关系库，例如DBPedia
  - 大规模（自动）领域知识库，例如事件时间关系
  - 小规模（手工）领域知识库，例如电影知识

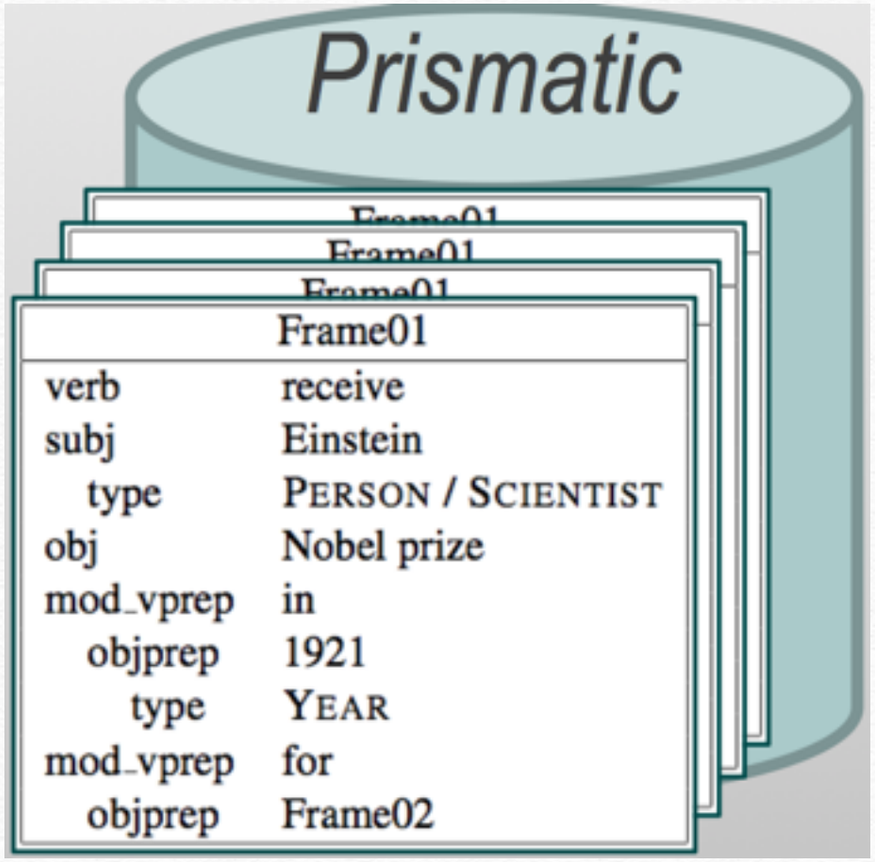
Watson

- 根据对Jeopardy!的分析，手工建立一些领域的frames
  - 例如电影、小说、国家、州、总统等
- 数据从相关的结构化或半结构化数据中提取
  - 例如Wikipedia表单、DBPedia



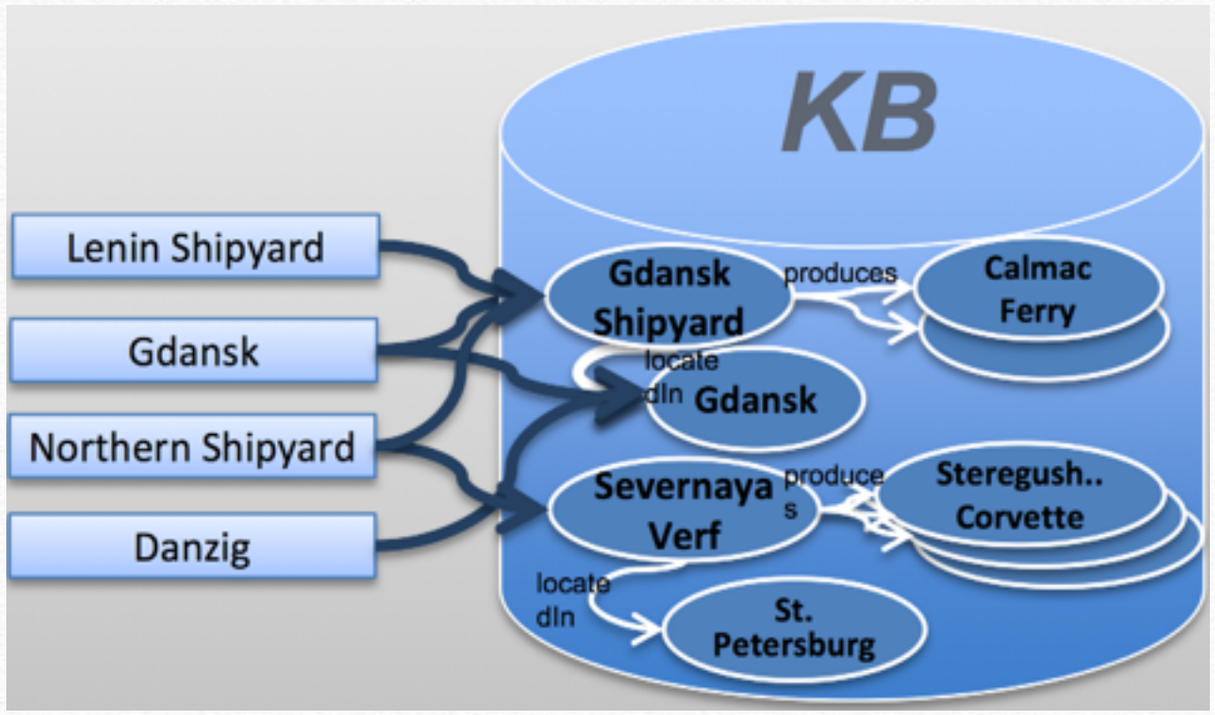
Watson - Prismatic

- Linguistic Frame Extraction
  - 10亿多Frames
  - 从2TB网页中挖掘
  - SVO/is-a等关系



Watson - KAFE

- Knowledge From Extracted Content
  - 实体消岐
  - 实体类型
  - 类型消岐



Watson中KG的使用

- 检测时间关系的吻合度
  - Query：*A 1992 movie starring Anthony Hopkins was based on a book by this author*
  - temporal_relation(Event A, Event B)
  - Accuracy提高1%

Watson中KG的使用

- 检测空间关系的吻合度
  - Query： *This state which lies to the NE of Nebraska*
  - spatial_relation(Entity A, Entity B)
  - Accuracy提高1.5%

Watson中KG的使用

- 比较候选答案的类型和问题的LAT
  - EDM：将实体映射到WordNet
  - PDM：将LAT映射到WordNet
  - 计算WordNet类型间的距离
    - Equivalent/subclass，Disjoint，Sibling，Superclass，Statistical，LCA
  - Accuracy提高3%

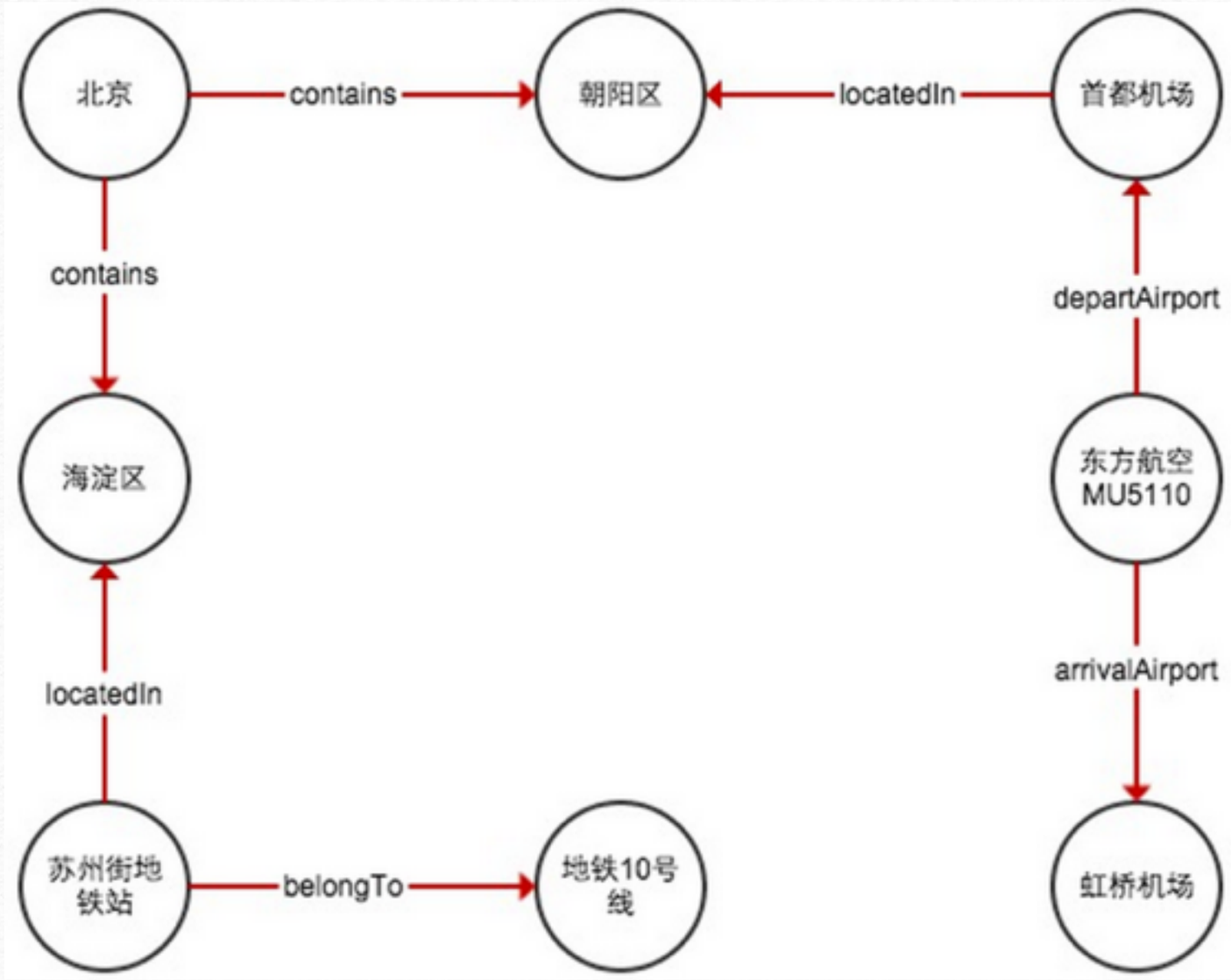
Watson中KG的使用

- Semantic frame用于候选生成
  - *LANGUAGE: The lead singer of the band Dengue Fever is from this country & often sings in Khmer. (Answer: Cambodia)*
  - 在Jeopardy!中，一个国家与一种语言共现，则该语言是该国家的通用语言
- 为7%的query提供至少一个候选

出门问问

- 知识图谱类别
  - 行政区划
  - POI，如机构、餐厅、宾馆、公交等
  - 人物
  - 歌曲、影视、小说、手机App

出门问问


北京
contains
朝阳区
locatedIn
首都机场
contains
departAirport
海淀区
东方航空
MU5110
locatedIn
arrivalAirport
苏州街地铁站
belongTo
地铁10号线
虹桥机场

出门问问

- 知识图谱的数据来源
  - 合作网站API
  - 网站抓取
- 知识图谱的数据整合
  - 实体聚合：不同来源的相同实体聚合
  - 属性聚合：不同来源的相同属性的聚合
  - 数据规范化：映射到我们的schema

出门问问

- 知识图谱的数据存储
  - 离线数据处理：Graph database
  - 在线服务：定制的便于检索的数据结构
- 知识图谱的研发重点
  - 三元组的抽取
  - 无监督/半监督的数据获取

出门问问

- 知识图谱的使用
- 自然语言处理
- Query分析
- 垂直搜索

出门问问

- 用于自然语言处理
- Query：东莞庄路富力院士庭广场
  - 庄路？
  - 院士？
  - 东莞庄路
  - 富力院士庭广场

出门问问

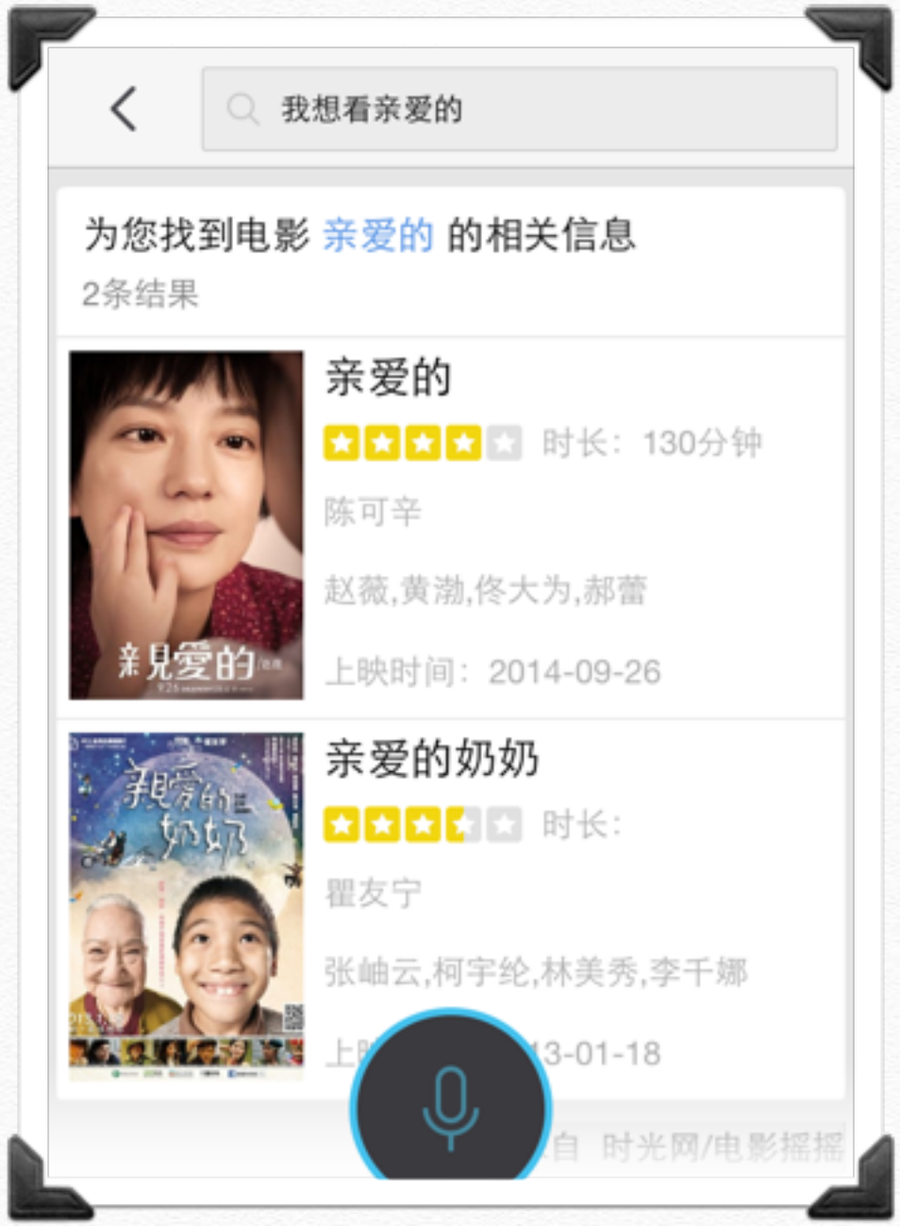
- 用于自然语言处理
- Query：广州同德围附近的光大银行
  - 同德／NN 围／VV
- 同德围／NN

出门问问

- 用于Query分析
- Query：睡在我上铺的兄弟
  - 整个Query作为歌名

出门问问

- 用于Query分析
- Query：我想看亲爱的
  - “亲爱的”也是一部电影的名字



出门问问

- 用于垂直搜索
- Query：交通大学怎么走
  - “交通大学”被解释成“北方交通大学”，而不是其它更常见的
- 实体地理位置的因素

出门问问

- 用于垂直搜索
- Query：甜蜜蜜
  - 歌曲
  - 电影
  - 电视
  - 蛋糕店
  - 婚庆公司

结语

- 问答系统是知识图谱开发和应用的前沿
- 问答系统依赖于大规模的知识图谱，也需要手工建设的领域知识图谱
- 知识图谱的使用渗透在各个子模块中
  - 自然语言处理、Query分析、垂直搜索、候选打分、专有领域问题的解答