科大讯飞 iFLYTEK

从应用角度来看知识图谱的价值和挑战

科大讯飞研究院　胡国平

2014年10月17日 武汉

第二届全国中文知识图谱研讨会

- 本次知识图谱研讨会
  - 着重探讨中文知识图谱的构建的资源、技术、方案、策略以及待研究问题和挑战
  - 促进研究单位之间以及研究界和产业界之间的学术交流
  - 探索今后大规模中文知识图谱构建的研讨与合作机制

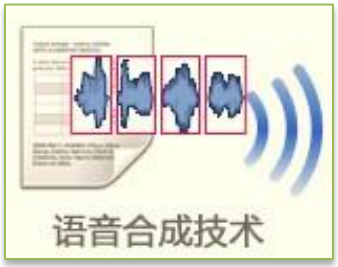
讯飞研究院简介

- 科大讯飞公司（中文信息领域的三个上市公司之一）下设的专门从事语音及语言核心技术研究的部门

- 指导思想：用正确的方法，做有用的研究

讯飞发展史上“有用的研究”

引擎厂商



语音合成

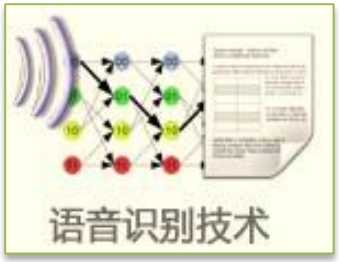
1999~

教育学习

语音评测

2002~

运营商&安全



语音识别

2005~

移动互联

语音交互

2011~

交流内容

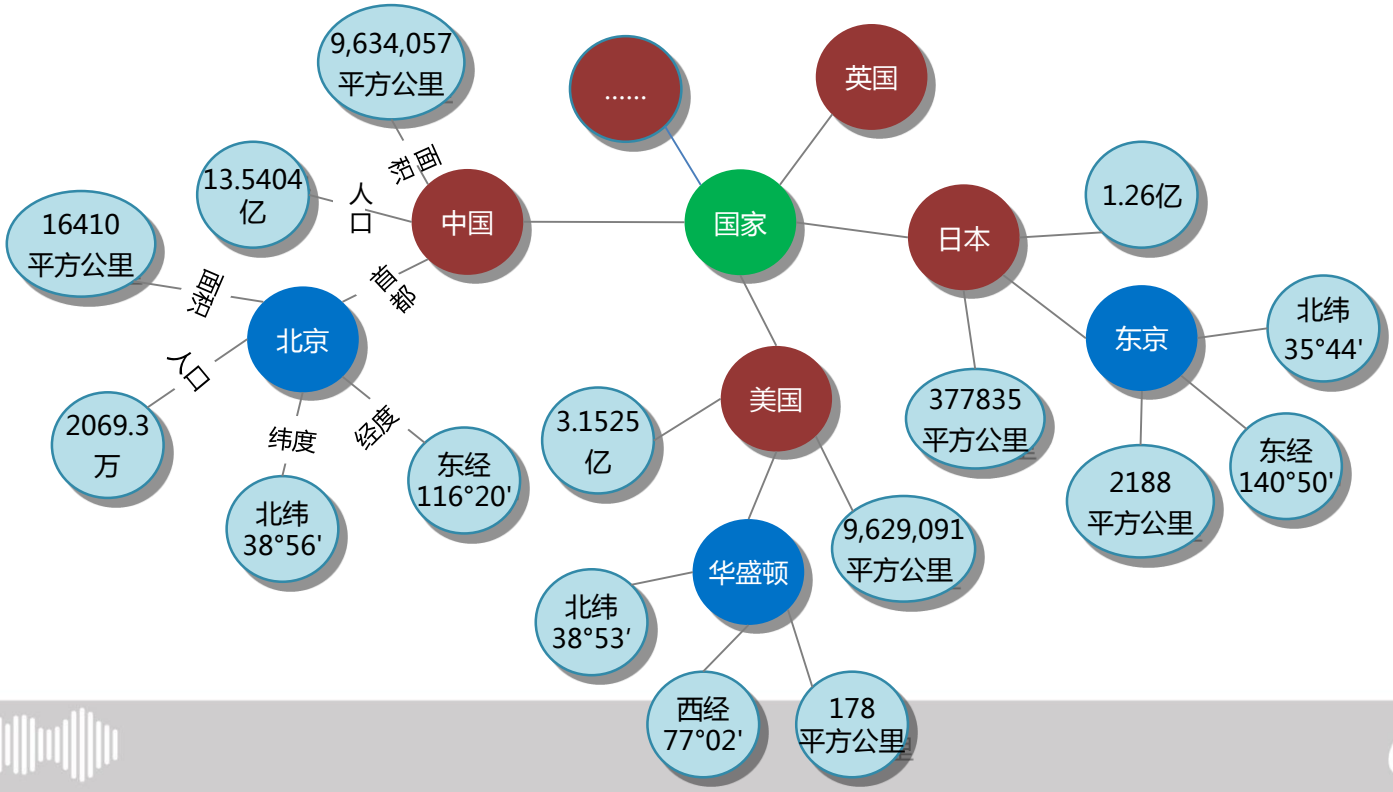
从应用角度来看知识图谱的价值和挑战

一、讯飞在知识图谱领域的一些尝试

二、从应用看知识图谱的价值和挑战

什么是知识图谱(Knowledge Graph)?

- 知识图谱旨在描述真实世界中存在的各种实体或概念及其关系，一般用三元组表示
- 知识图谱亦可被看作是一张巨大的图，节点表示实体或概念，边则由属性或关系构成

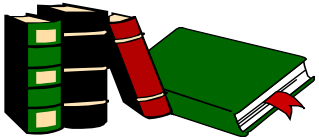


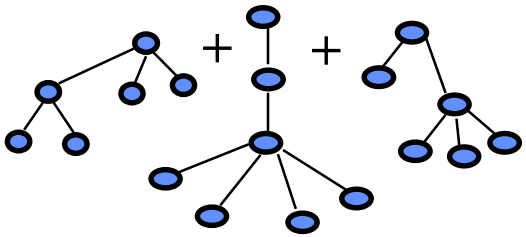
知识图谱构建后台的技术族

让世界聆听我们的声音

Web

字典，术语表，
百科，书本



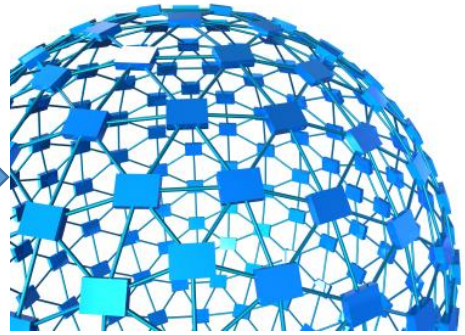


已有知识库

知识获取

知识融合

知识验证



知识图谱

科大讯飞 iFLYTEK

现有的大规模知识

名称	规模	
Yago	1千万实体，35万类别， 1.8亿事实，100种属性， 100语言	
Dbpedia	4千万实体，250类别， 5亿事实，6000种属性，	
Freebase	2千5百万实体，2000主题， 1亿事实，4000种属性	
谷歌知识图谱	5亿实体名字， 35亿条事实	
NELL	3百万实体名字，300类别 500属性，100万事实 1千5百万学习规则	

中文知识库

- 传统的语言类知识库，基于人工编写方式，构建了一系列的中小规模中文知识库
  - 知网(HowNet)[董振东 和 董强,1999]
  - 《同义词词林》[梅家驹等,1996]
  - 概念层次网络(HNC)[黄曾阳,1997]

- 大规模事实类知识库，针对自身业务需要建立
  - 百度知心，优化搜索
  - 搜狗知立方，优化搜索
  - 阿里巴巴知识库(商品知识库）

讯飞的一些初步尝试

让世界聆听我们的声音

音乐领域

- 重点探索了如何基于互联网构建高质量的大规模音乐知识库

10086客服领域

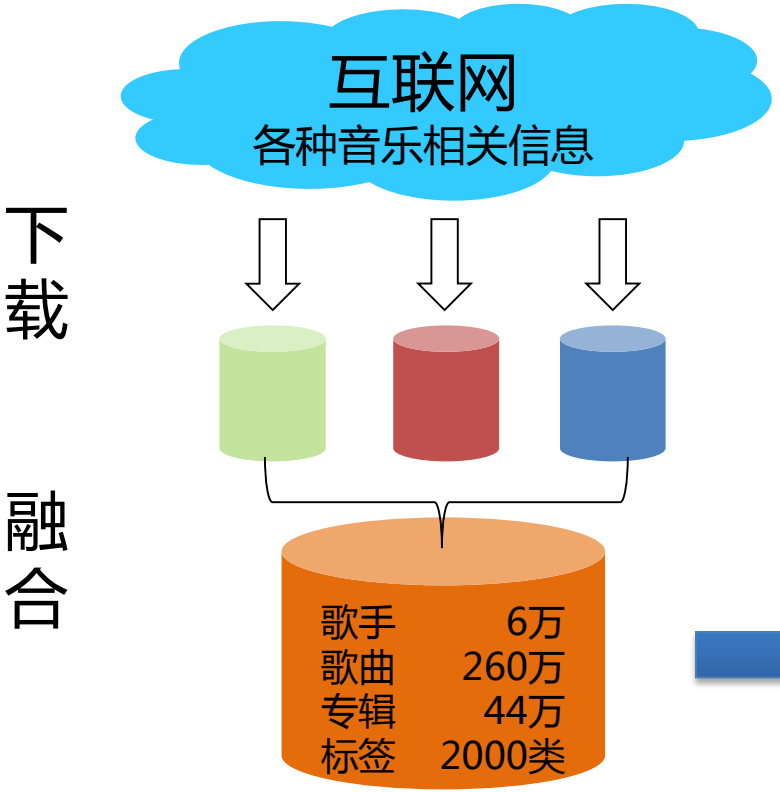
- 重点探索了如何基于知识库改进自动客服系统

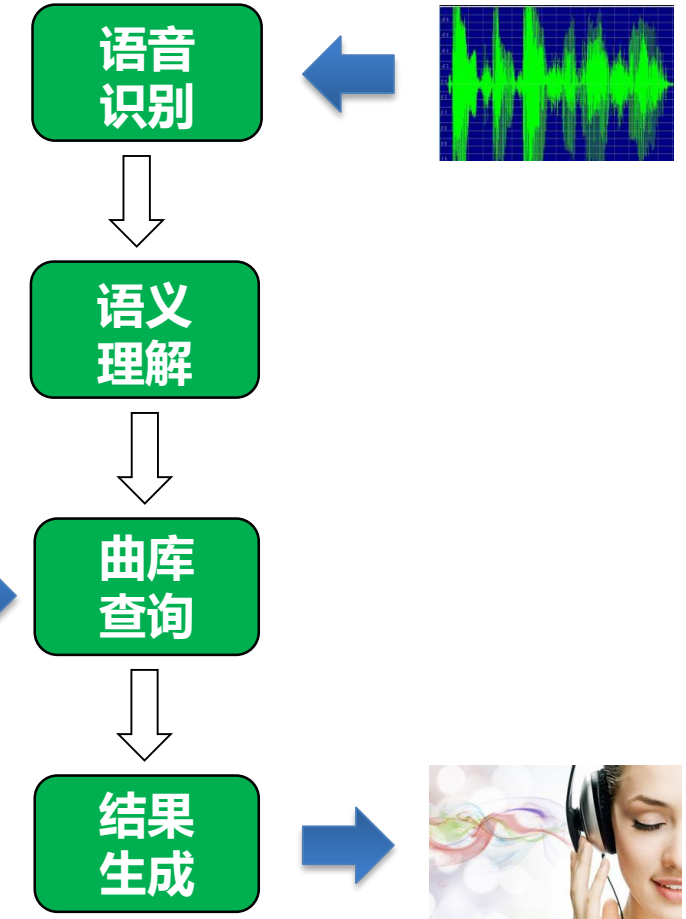
音乐领域

- 任务定义：想听什么，想问什么，直接说
  - 放一首王力宏的歌（王力宏_歌手）
  - 我想听江南style神曲(江南style_歌曲)
  - 来一首他一定很爱你，阿杜的（阿杜_歌手_他一定很爱你_歌曲）
  - 来首儿歌（儿歌_标签）
  - 我想听一首歌吻吻的幸福 (稳稳的幸福_歌曲 ， 需容错)
  - 播放周杰伦的新歌 （周杰伦_歌手，歌曲属性为“新” ）
  - 周董和费玉清合作的那首歌（周杰伦_歌手& 费玉清_歌手）
  - 周杰伦和哪些台湾女歌手合作过（知识问答）
  - 神话是孙楠和谁合唱的（知识问答）

语音识别的突破，大大激发了语音理解的需求

系统实现





基于互联网的音乐曲库构建 + 基于人工规则的语义理解

基于互联网的音乐曲库构建

- 遇到了一系列的问题
  - 同一个歌曲有多个歌手演唱过的版本，甚至多个完全不同的歌曲，网上信息比较混乱

基于互联网的音乐曲库构建

让世界聆听我们的声音

- 遇到了一系列的问题
  - 同一个歌曲有多个歌手演唱过的版本，，甚至多个完全不同的歌曲，网上信息比较乱
  - 歌曲的标签体系比较乱，需要融合
    - “蓝调” “布鲁斯”（对应的英文都是Blues）
    - “慢摇” “R&B” “HIP&HOP”
    - “戏曲” “京剧” “昆曲” “黄梅戏”（相互平行）
    - “纯音乐” “钢琴曲” “古筝” “萨克斯”
  - 歌曲名、歌手名很随意
    - 我想给我的*朋友*点首*我爱你*中国*这首歌*

科大讯飞 iFLYTEK

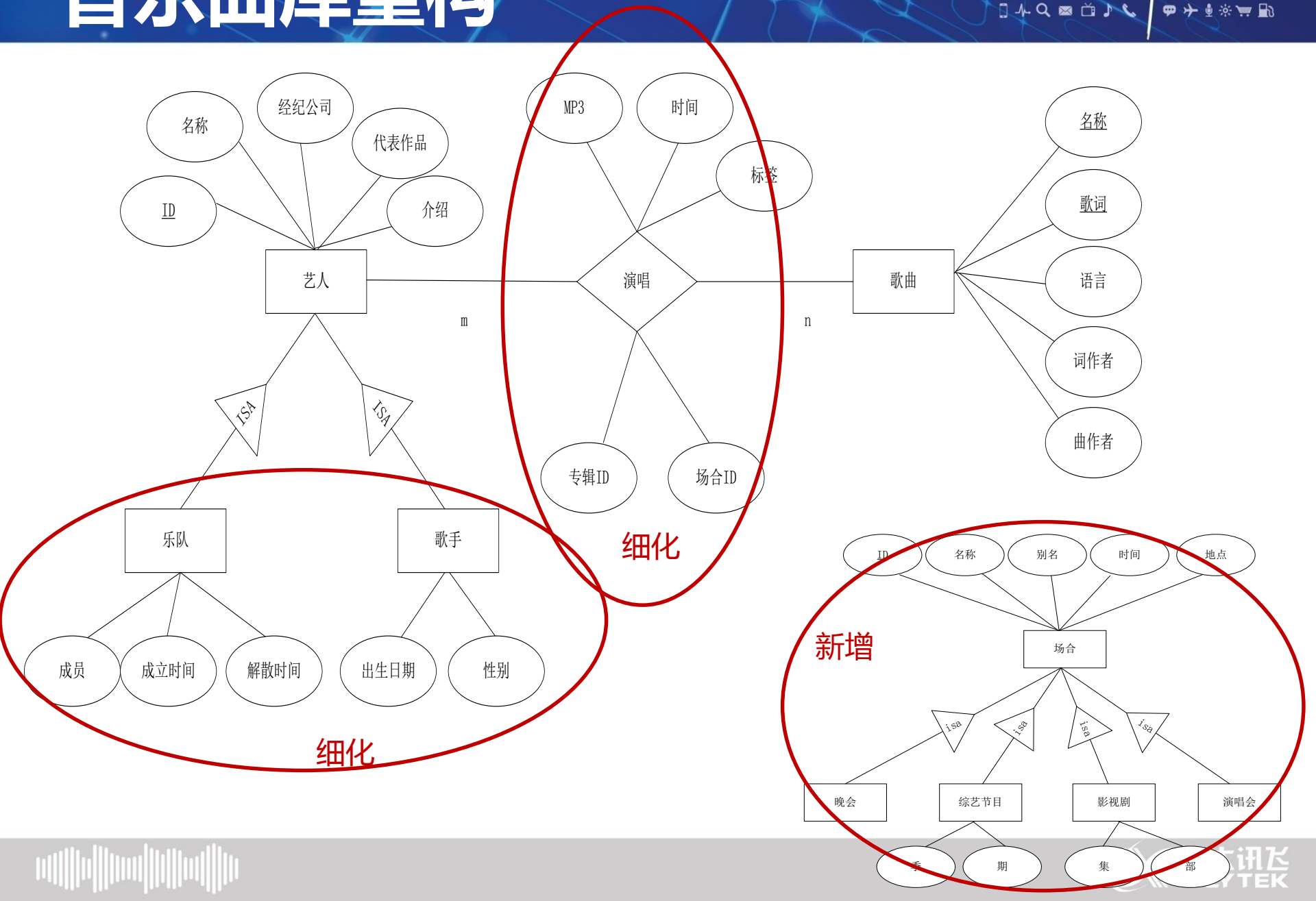
音乐曲库重构

- 音乐曲库重构
  - 以事件本体为依据重构音乐本体库，替换三元组方案

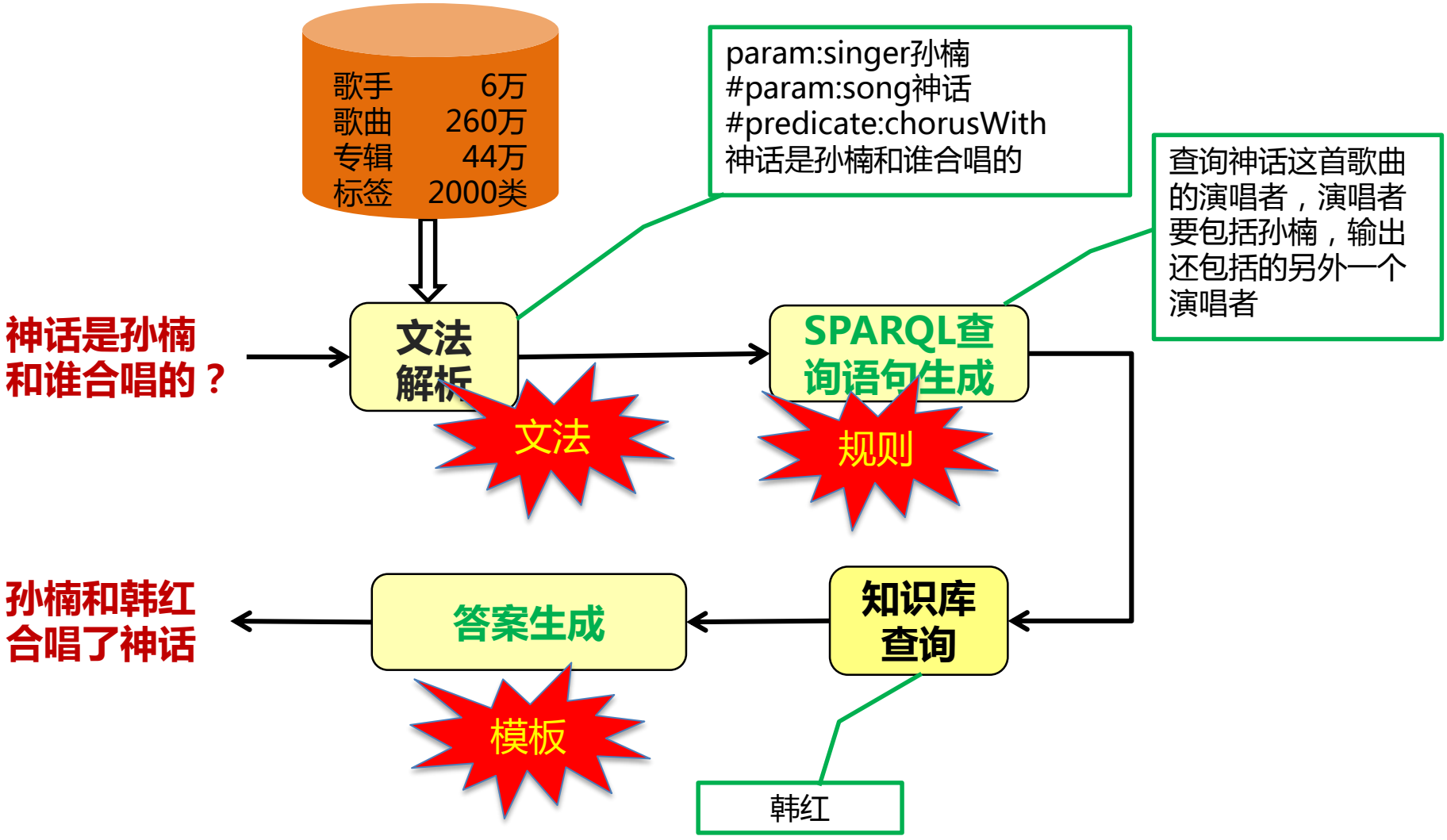
三元组：（那英_Subject，演唱_Predicate，春暖花开_Object）

事件本体：（那英_主体，演唱_动作，春暖花开_客体，蛇年春晚_场合，2013年2月9日_时间）

音乐曲库重构


名称
经纪公司
代表作品
ID
介绍
艺人
m
演唱
n
歌曲
MP3
时间
标签
专辑ID
场合ID
名称
歌词
语言
词作者
曲作者
细化
ISA
ISA
乐队
歌手
成员
成立时间
解散时间
出生日期
性别
细化
新增
ID
名称
别名
时间
地点
场合
isa
isa
isa
isa
晚会
综艺节目
影视剧
演唱会
季
期
集
部

基于人工规则的语义理解



基于曲库资源的理解和纠错

识别结果：我是你的安东尼演唱的
文法规则： {歌曲名}{歌手名}演唱的
容错结果：倪安东（歌手名）我是你的（歌曲名）

识别结果：汪峰的北京青年
文法规则：{歌手名}的{歌曲名}
容错结果：汪峰（歌手名）北京北京（歌曲名）

讯飞的一些初步尝试

- 音乐领域
- 10086客服领域

智能客服领域

智能客服：改变传统的按键式自助服务，用户使用自然语音与系统交互，实现菜单扁平化，提升用户满意度，减轻人工服务压力，降低运营成本。

10086智能客服

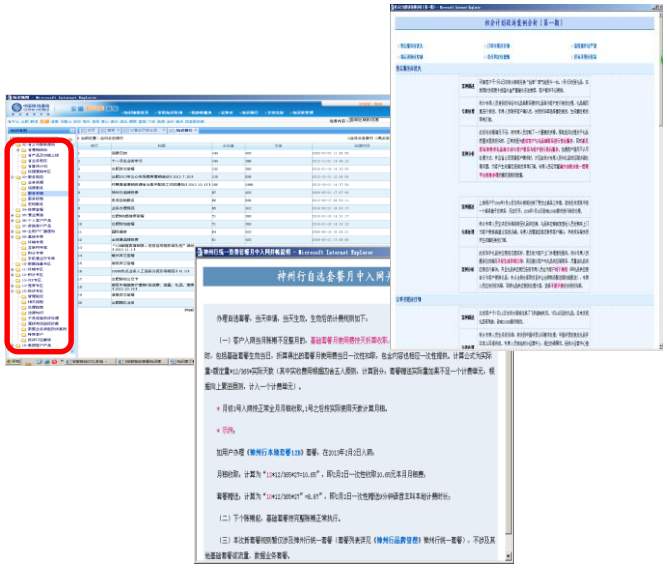
业务办理类：

1.帮我查询话费。
2.办理彩铃业务。
3.开通5元30M套餐。
4.我要交100元话费

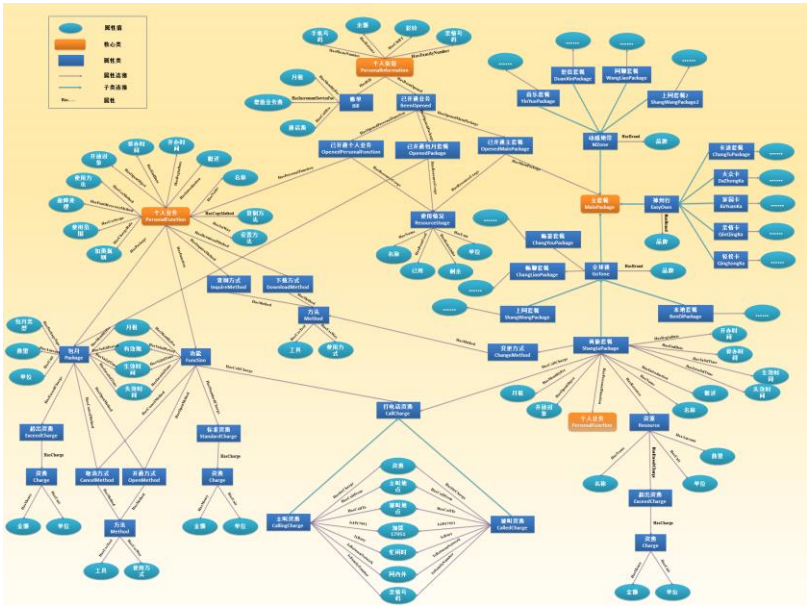
业务咨询类：

1.怎样退订国内移动数据5元？
2.办两城一家后给省内打电话,是怎么收费的？
3.怎么办理10元100M流量包？
4.我要到新加坡去旅游，需要办什么手续？

总体思路



移动作业手册

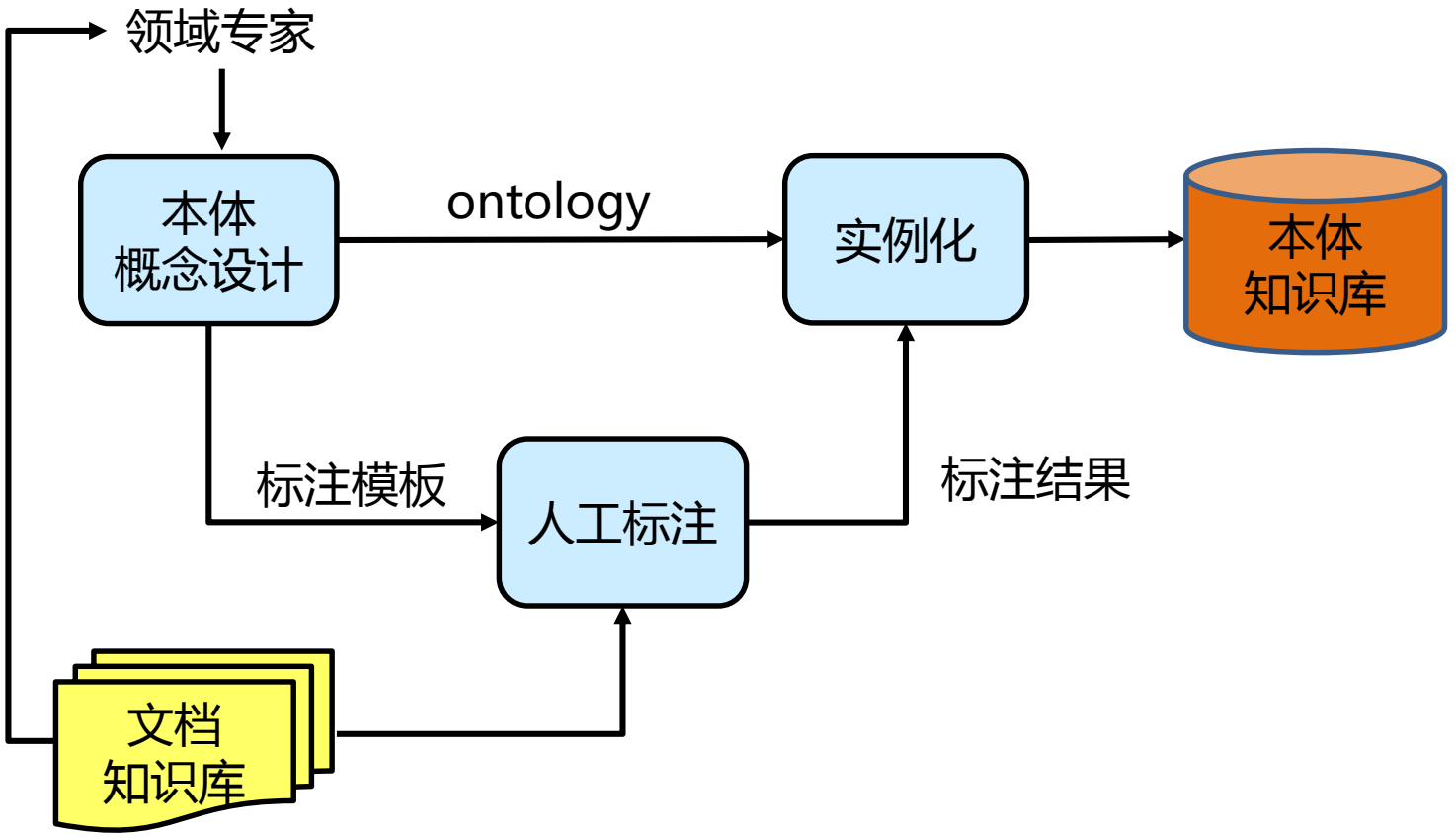


结构化知识库

原始文档知识库

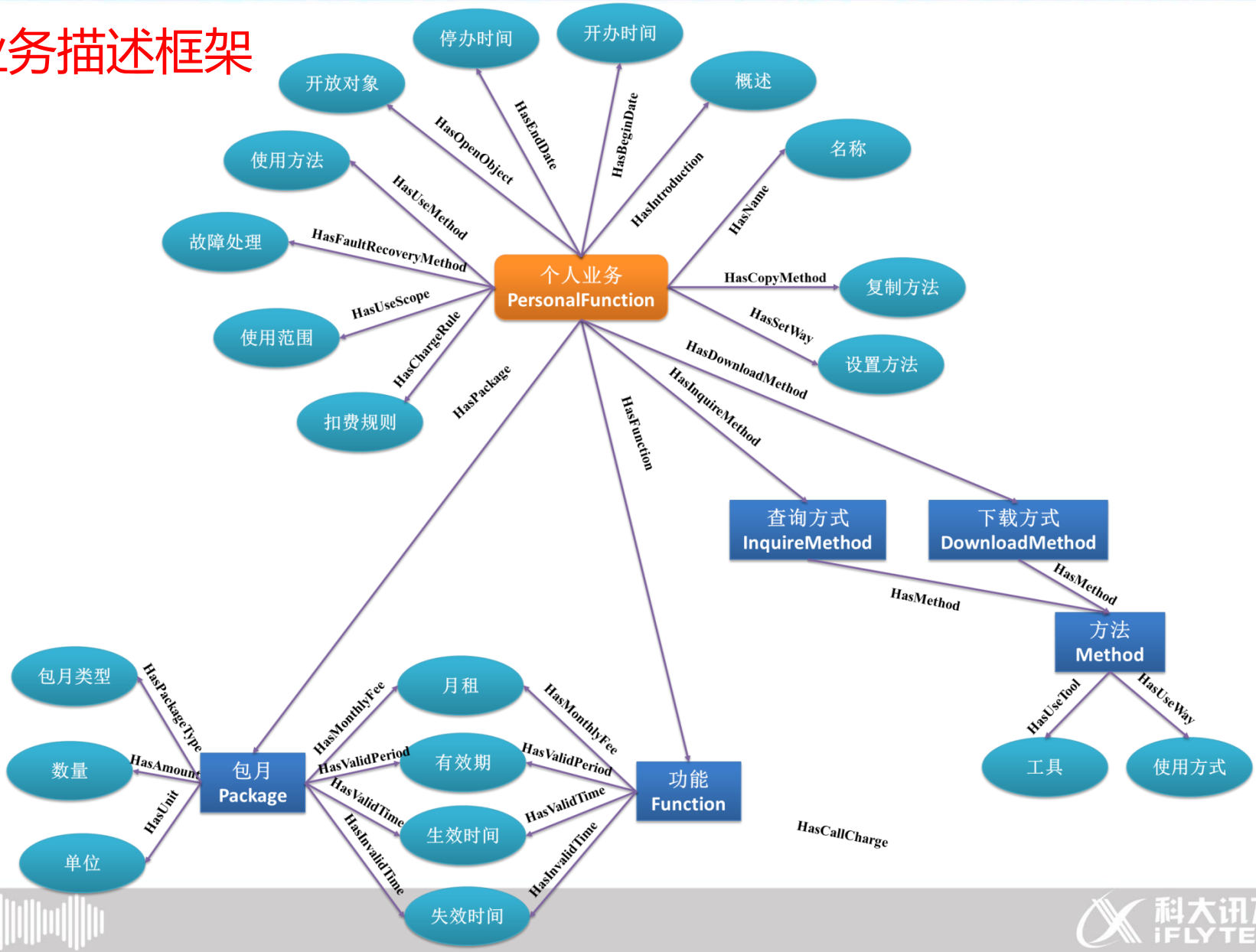
安徽移动Web版知识库

人工参与的本体知识库构建



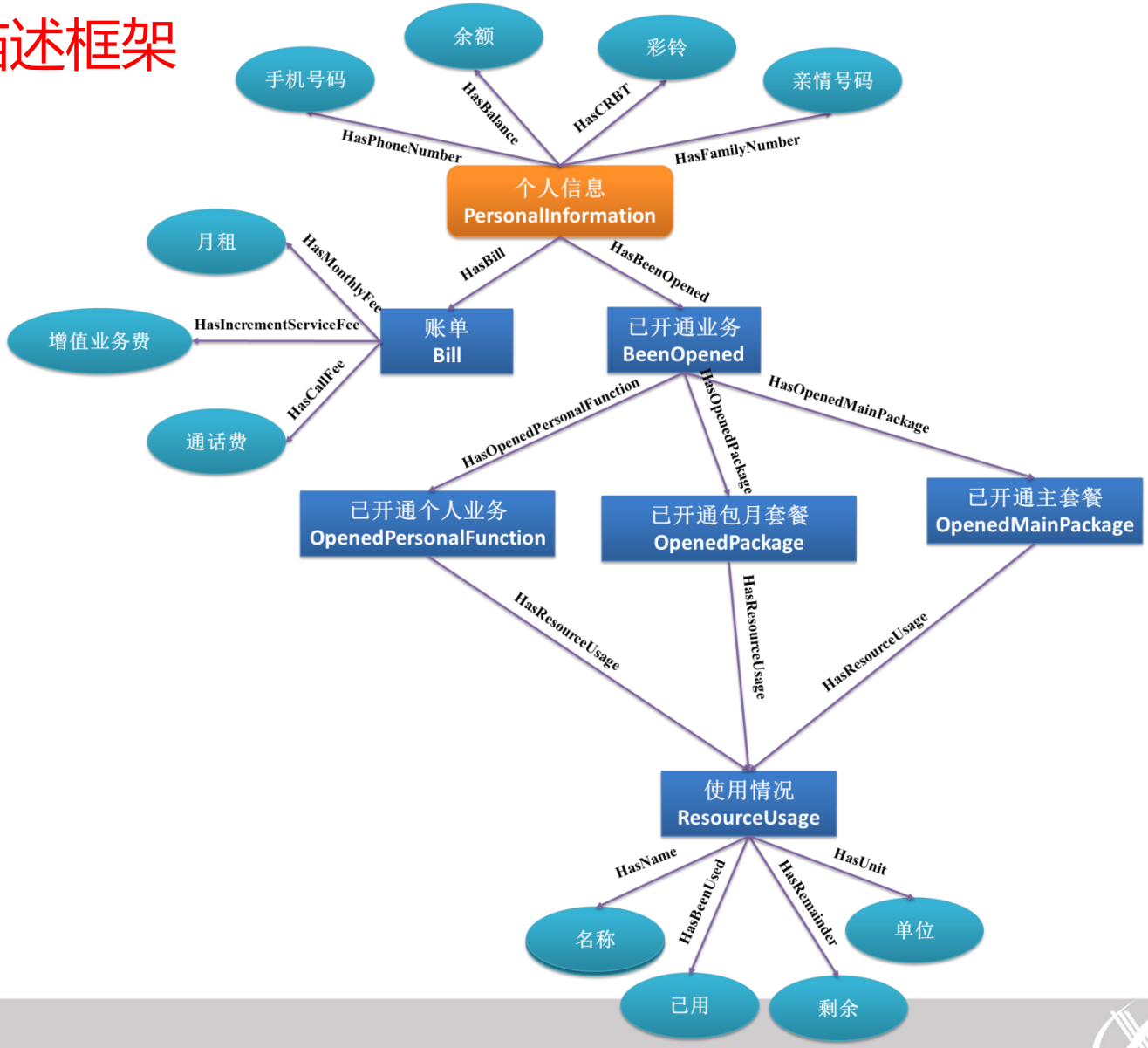
知识库结构

个人业务描述框架



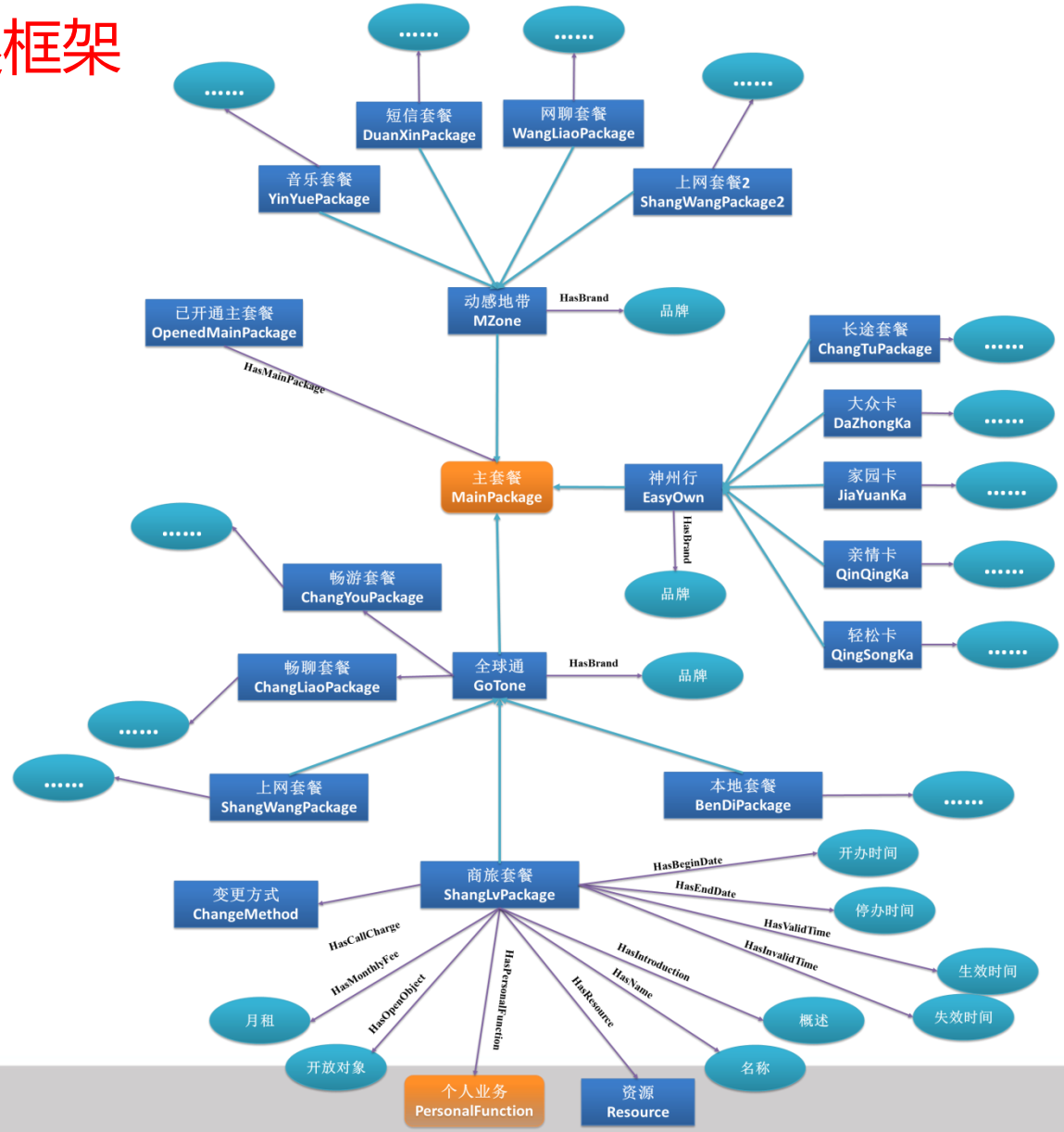
知识库结构

个人信息描述框架



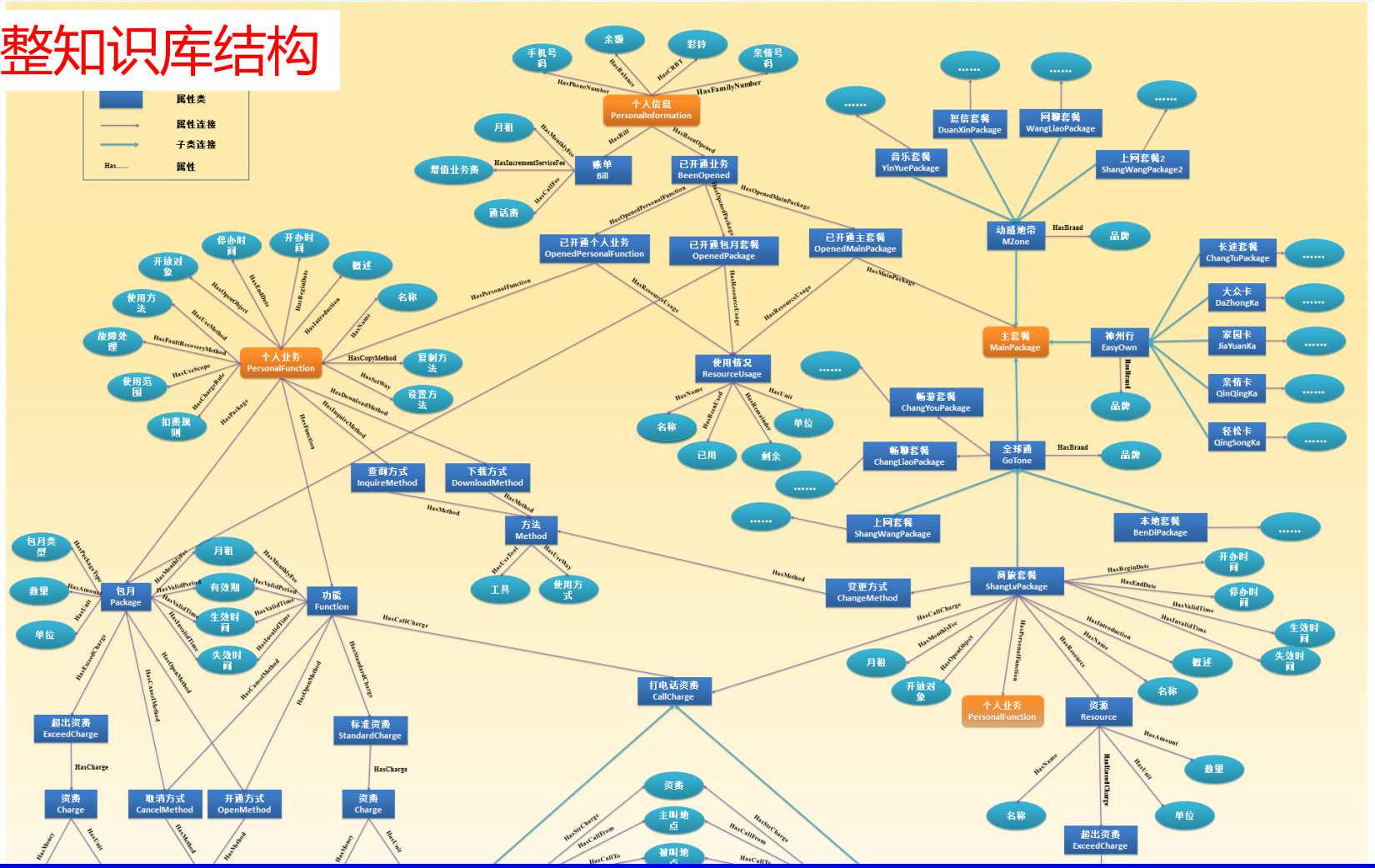
知识库结构

主套餐描述框架



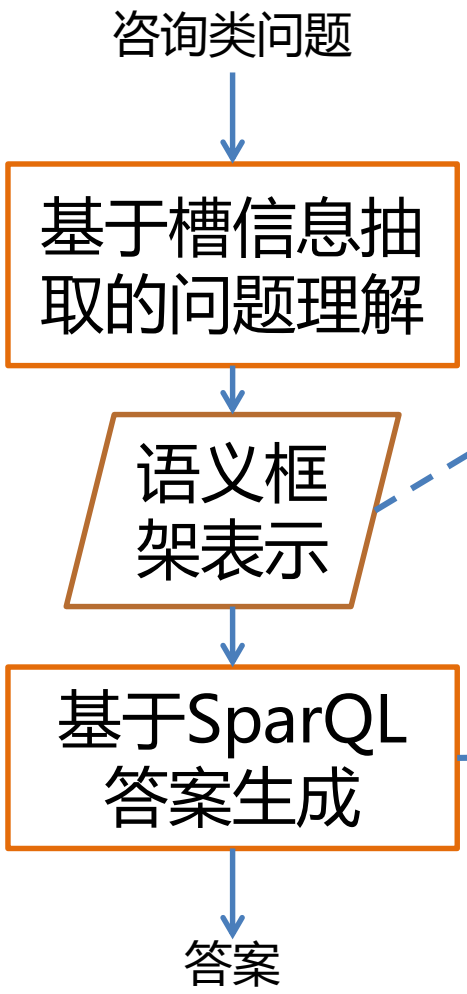
知识库结构

完整知识库结构



当前规模：106属性，100业务，3102个三元组

基于本体库的咨询问答服务



问题：怎么办理10元100M流量包？

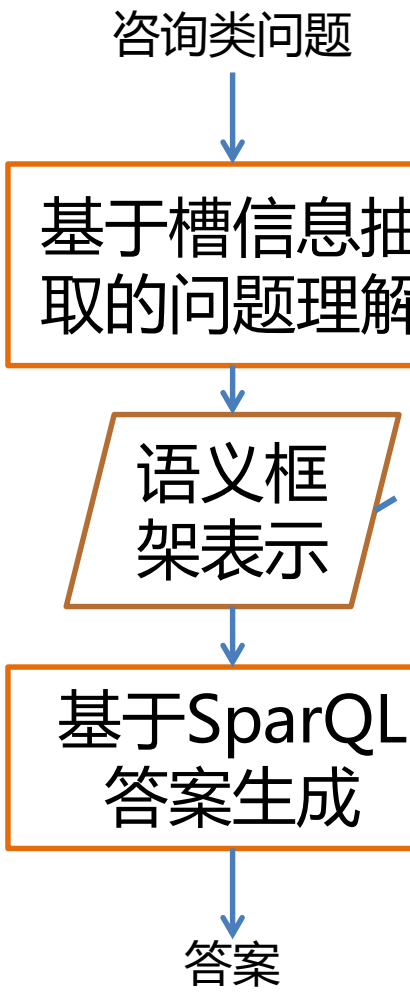
句子级 问题焦点 开通方式

词汇级
- 问题类型 怎么
- 操作 开通
- 业务 流量
- 金额 10元
- 数量 100M

```
Select ?money ?amount ?open_way{ ?busi
has_name “流量”. ?busi
contain_pack ?pack. ?pack
has_money ?money.?pack
has_amount ?amount. ?pack
has_open_way ?open_way}
```

答案：对不起，没有找到10元100M的流量套餐，但是有以下套餐:10元70M;5元30M;50元500M.

基于本体库的咨询问答服务



问句：办两城一家后给省内打电话,是怎么收费的?

业务名称：两城一家省内套餐
关键动词：是、拨打、收费
参　　数：地区-省内
关键名词：两城一家、电话
问题类型：how
问题焦点：资费

Select ?money ?amount ?open_way{ ?busi
has_name "两城一家" .....

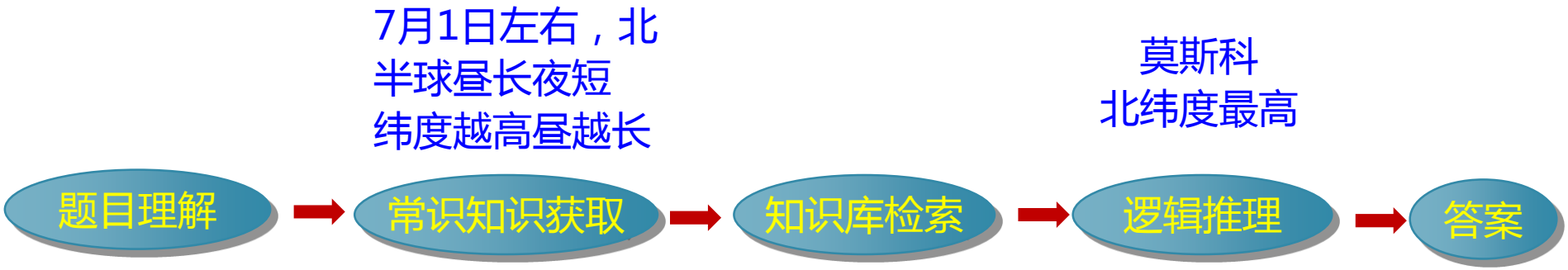
答案：开通后省内定点市打电话每分钟0.29元

863类人答题任务难度分析

题目（改编自2012年海南省高考文综地理试卷）

7月1日下列城市白昼最长的是？

A. 悉尼 B. 上海 C. 雅加达 D. 莫斯科



类人答题系统的研发需要多种知识库的支撑！

让世界聆听我们的声音

从应用角度来看知识图谱的价值和挑战

一、讯飞在知识图谱领域的一些尝试

二、从应用看知识图谱的价值和挑战

人工智能的三个阶段

计算智能

能存会算

感知智能

能听会说、能看会认

认知智能

能理解会思考

一、知识图谱研发需区分并针对应用

- 和NLP一样，不同应用场景，对知识图谱的需求点和难点差别很大

应用场景

- 搜索：展示及优化搜索结果
- 音乐：难在构建曲库
- 客服：对自动客服的帮助有限
- 答题：需要各种常识和学科知识库

- 潜在应用价值的发现，能拉动知识图谱的研究水平和投入的提升

二、积极探索知识图谱对理解的价值

知识图谱对语义理解本身的帮助有限

- 列表型资源的利用
- 基于知识图谱的纠错

记忆超出任何人，但理解不及5岁小孩

- 给你一个苹果，再给你一个苹果，你有几个苹果了
- 给你一个苹果，你吃了一个苹果，你有几个苹果了

语义理解正是人类知识积累的渠道

- 乌龙球：踢进自家球门的球
- 高铁：就是很快的火车

三、直面当前知识图谱技术方案的问题

研究范围

- 语言知识库
  - 何谓“坐”
- 事实类知识库
- 专业领域知识库

表示能力

- 三元组的方式，表达能力不足
- 知识图谱表示是离散型的，没有连续、模糊的概念
  - 收获了七八斤苹果

应用模式

- 串行的知识图谱如何追赶人类是高度并行
  - “跑”比“走”快
  - 从北京到天津不用坐飞机
- 知识图谱上的推理能力，也还太弱

“深”比“广”难

一个哲学问题

- 仅用文字描述的世界，是不是 “假”的世界？
  - 没有视觉、没有听觉、没有触觉
  - 没有情感、没有信仰

格桑花

一种生长在高原上的花朵，杆细瓣小，花为复叶羽状，五瓣花，颜色鲜艳而美丽，随着季节变化而逐渐变换颜色。花径长约1.5～3cm，为落叶乔木。

总结和建议

“用正确的方法，做有用的研究”，知识图谱的研究应该与实际应用相结合，应该与语义理解、知识学习相结合

讯飞公司针对音乐和客服两个领域做了一些知识图谱相关的尝试，目前还都不太成功

“深”比“广”难，现有知识图谱的解决方案和应用价值都面临诸多挑战，是一个很值得深入和广泛合作研究的领域

谢谢

让世界聆听我们的声音

第二届全国中文知识图谱研讨会

主办单位：中国中文信息学会

承办单位：华中师范大学, 武汉, 2014年10月17日

Home　会议简介　演讲嘉宾　会议日程　主办　English