AiCon
全球人工智能与机器学习技术大会

知识图谱在小米的落地与挑战

刘作鹏

小米人工智能实验室知识图谱总监

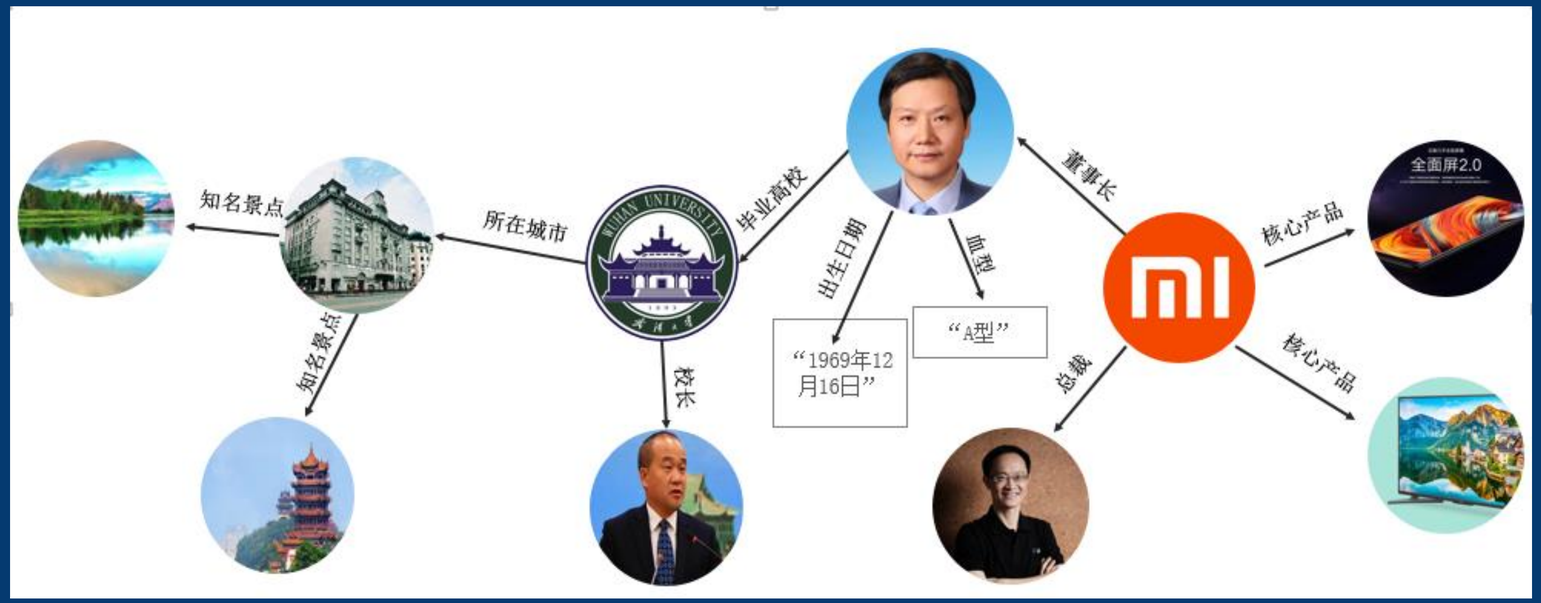
主办方 Geekbang 极客邦科技 InfoQ

TABLE OF

CONTENTS 大纲

- 知识图谱概述
- 知识图谱构建
- 基于图谱的问答
- 图谱的其他典型应用

知识图谱简介



- 本质上，知识图谱是大规模的语义网络（Semantic Web）；
- 语义网络是知识表示的重要方式之一，富含实体、概念和多种语义关系；

语义网络

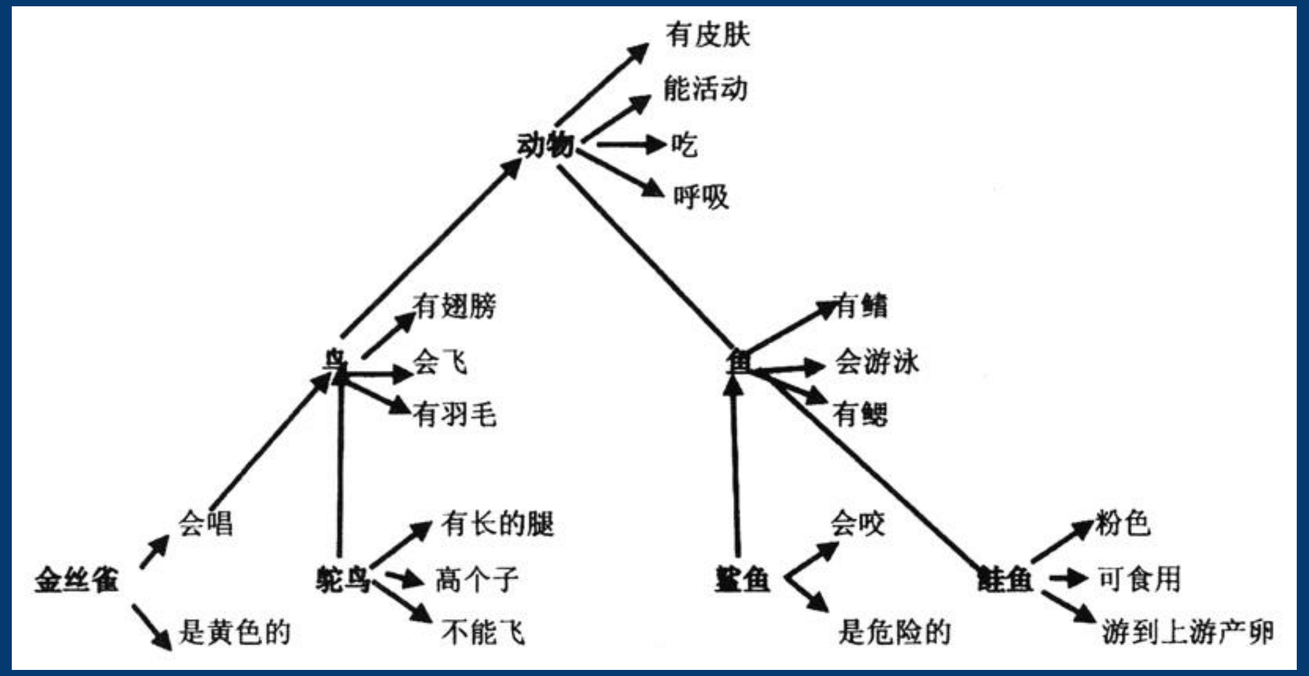
语义网络

知识表示

知识工程

人工智能

语义网络的位置



典型的语义网络

图谱的价值

规模巨大

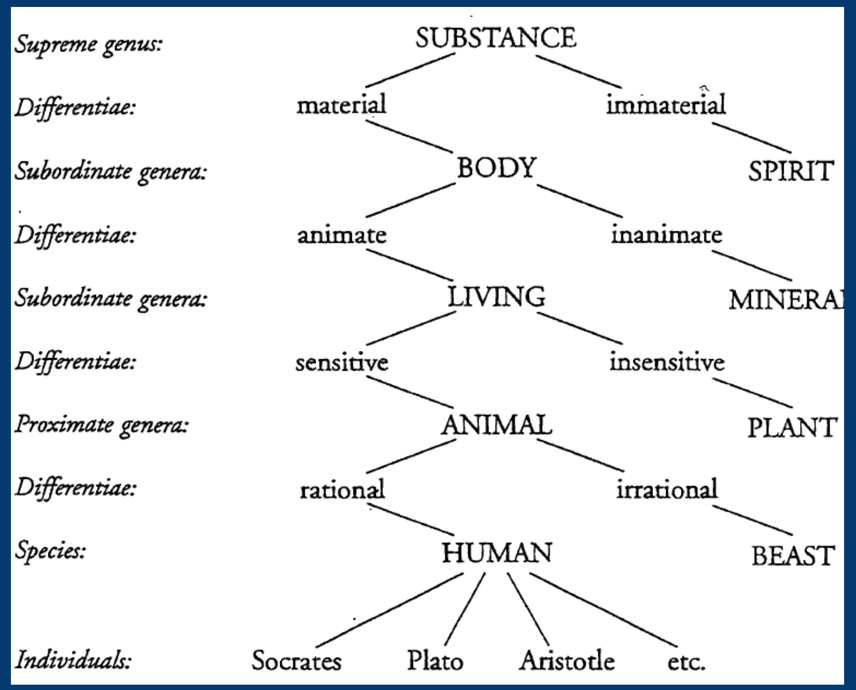
结构精良

语义丰富

质量上乘

- 知识图谱是让机器具备认知能力的关键技术；
- 同深度学习相比，它在推理、理解、解释等问题的处理上，有独特的优势；

简单回顾

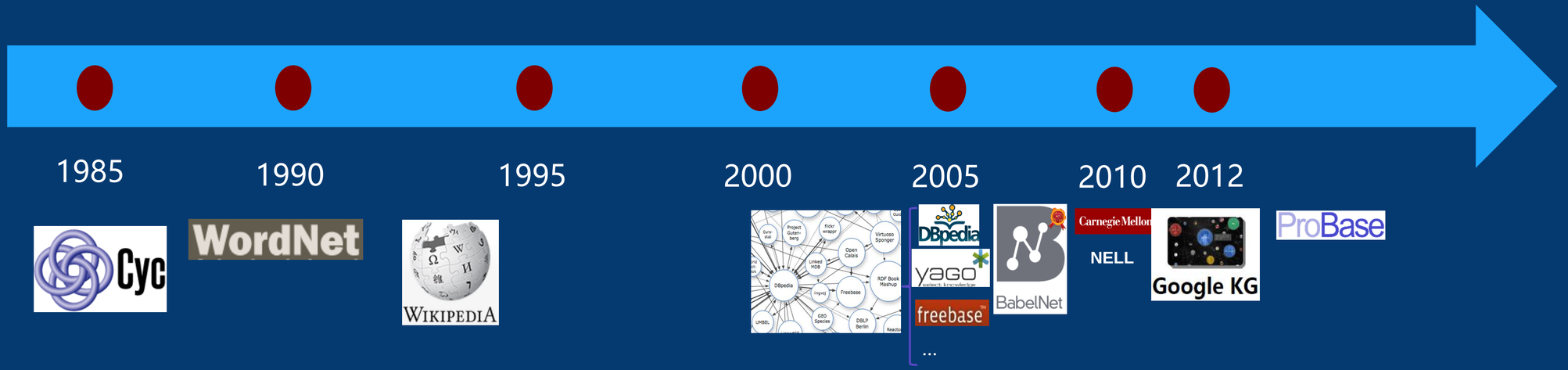


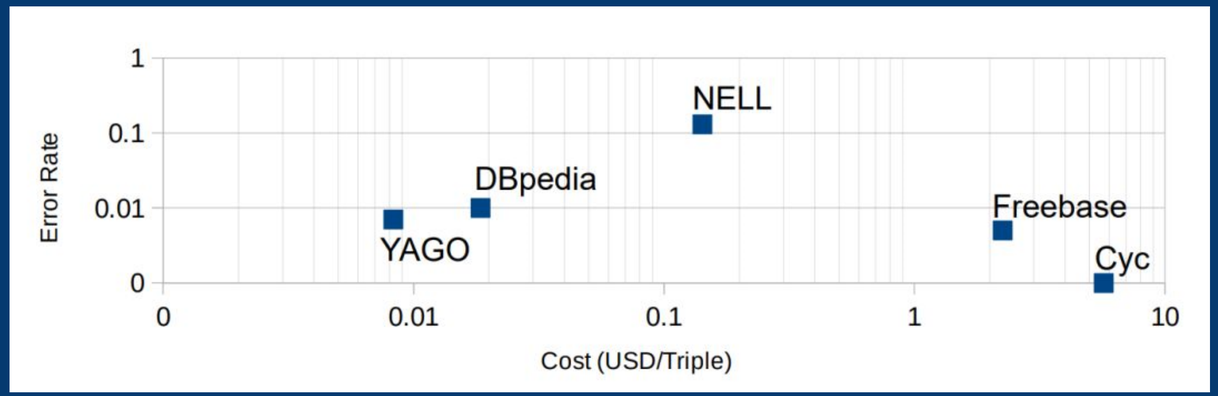
亚里士多德(“百科全书之父”)：三段论 & 形而上学

莱布尼茨：齿轮计算器 & 符号逻辑

波菲利之树(公元270年)

近代发展





- Cyc 最早通用知识图谱之一，耗时1000人年，成本1.2 亿美金；
- Freebase 有30 亿条英文关系，基于众包，总成本约为 67.5 亿美元；

OpenBase项目

知识的抽取与编辑

知识的挖掘与融合

知识更新

知识众包

联合发起机构

OpenKG　小米集团　清华大学　浙江大学　东南大学　海知智能　狗尾草

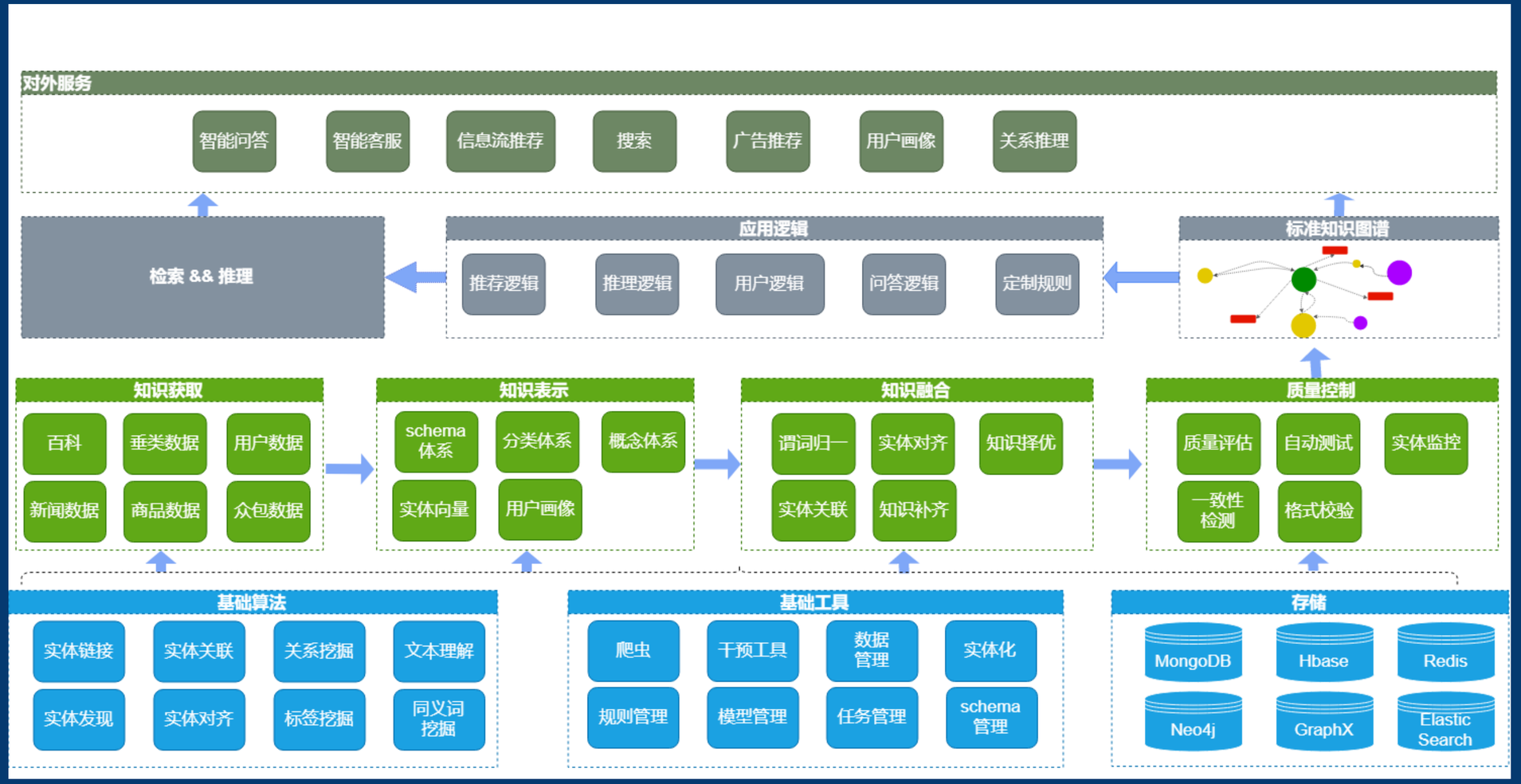
https://openbase.ai.xiaomi.com/

TABLE OF

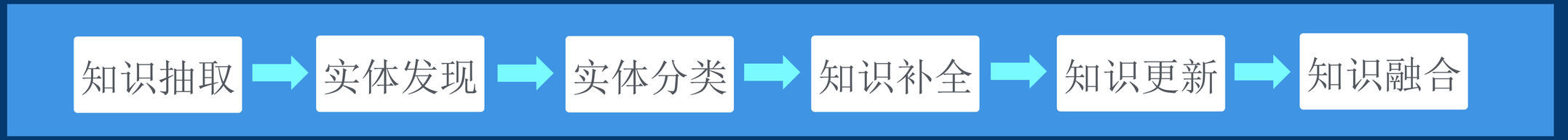
CONTENTS 大纲

- 知识图谱概述
- 知识图谱构建
- 基于图谱的问答
- 图谱的其他典型应用

关键构建技术

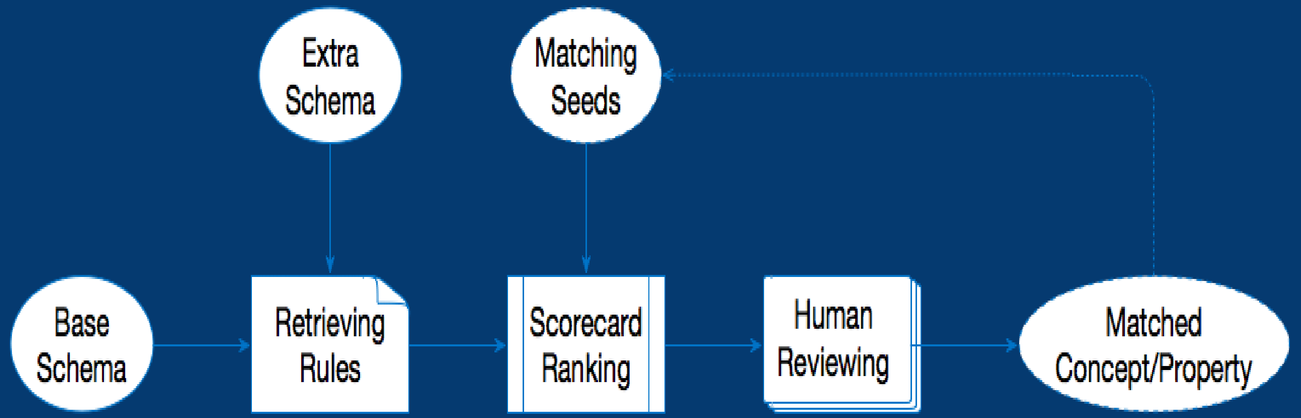

对外服务
智能问答
智能客服
信息流推荐
搜索
广告推荐
用户画像
关系推理
检索 && 推理
应用逻辑
推荐逻辑
推理逻辑
用户逻辑
问答逻辑
定制规则
标准知识图谱
知识获取
百科
垂类数据
用户数据
新闻数据
商品数据
众包数据
知识表示
schema 体系
分类体系
概念体系
实体向量
用户画像
知识融合
谓词归一
实体对齐
知识择优
实体关联
知识补齐
质量控制
质量评估
自动测试
实体监控
一致性检测
格式校验
基础算法
实体链接
实体关联
关系挖掘
文本理解
实体发现
实体对齐
标签挖掘
同义词挖掘
基础工具
爬虫
干预工具
数据管理
实体化
规则管理
模型管理
任务管理
schema 管理
存储
MongoDB
Hbase
Redis
Neo4j
GraphX
Elastic Search

百科知识图谱的构建



- 从多个百科数据源开始的、自下而上的图谱构建方式；
- 缺乏完整、一致的本体知识；
- 大规模百科图谱的对齐是一个难题；

百科图谱的对齐：Schema层对齐



- 以cnSchema为基础，半自动地抽取等价概念和上下位关系；
- 利用等价概念发现等价谓词，必要时利用人工干预提升结果质量；
- 等价谓词又可以帮助发现等价概念，然后迭代循环；

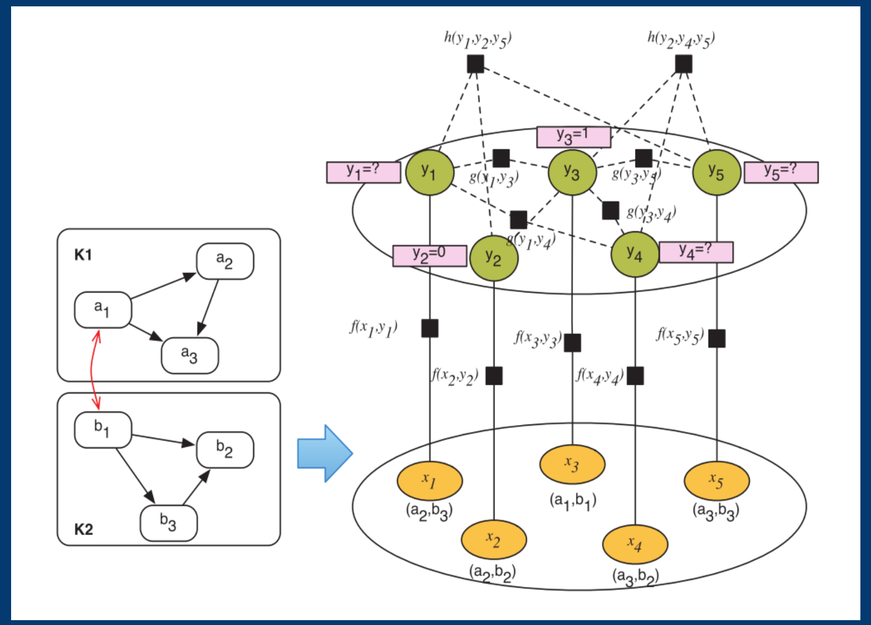
百科图谱的对齐：实例层对齐

- 方法一：使用基于EM算法的半监督学习策略寻找等价实例：
  - 等价实例和属性做为种子集合，并从种子集合中挖掘匹配规则；
  - 根据匹配规则寻找置信度高的等价实例；
  - 将新的等价实例加入到种子集合中，然后迭代循环；

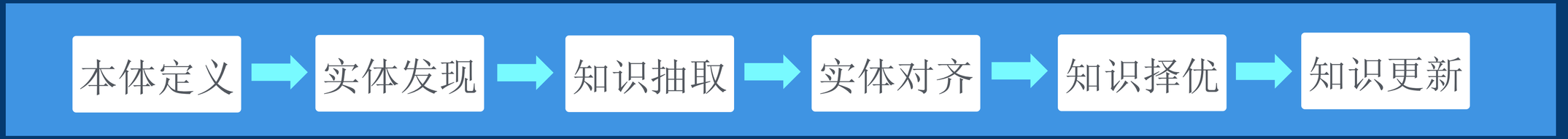
方法二： Linkage Factor Graph Model在跨语言链接上，有着很好的表现

$$p(Y) = \prod_i f(y_i, x_i) g(y_i, G(y_i)) h(y_i, H(y_i))$$

《Cross-lingual Knowledge Linking Across Wiki Knowledge Bases》
(Zhichun Wang… 2012)

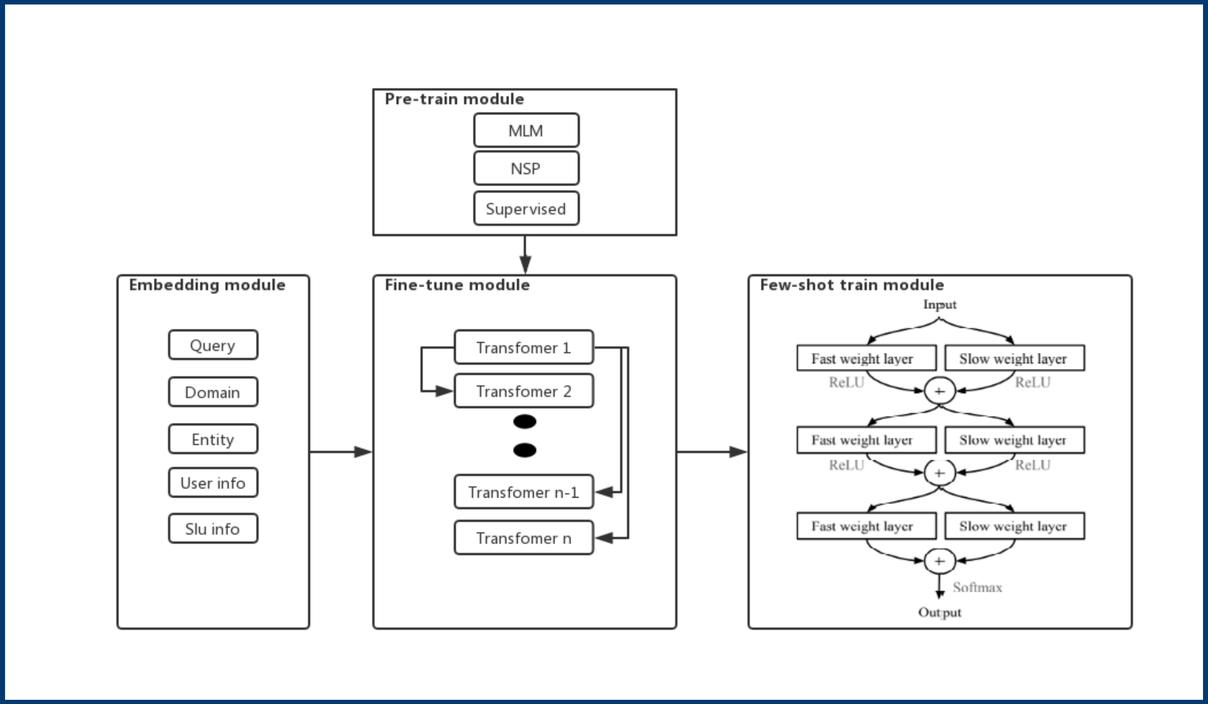


垂域知识图谱的构建



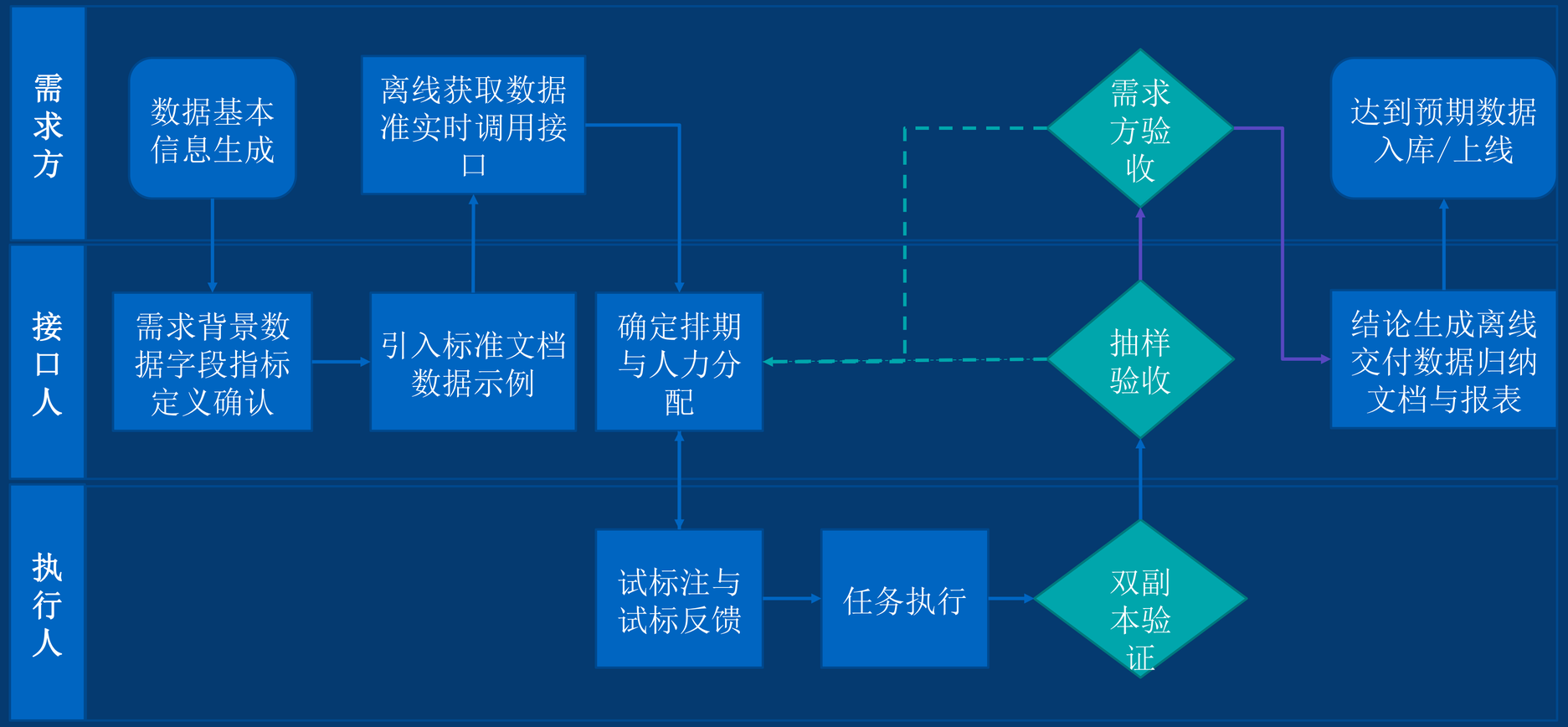
- 垂域图谱的构建从本体定义开始，采取自上而下的方式；
- 数据源的质量不同，知识择优异常重要；
- 在构建全新的垂域图谱时，基于小样本的模型学习非常重要；

小样本条件下的文本分类模型

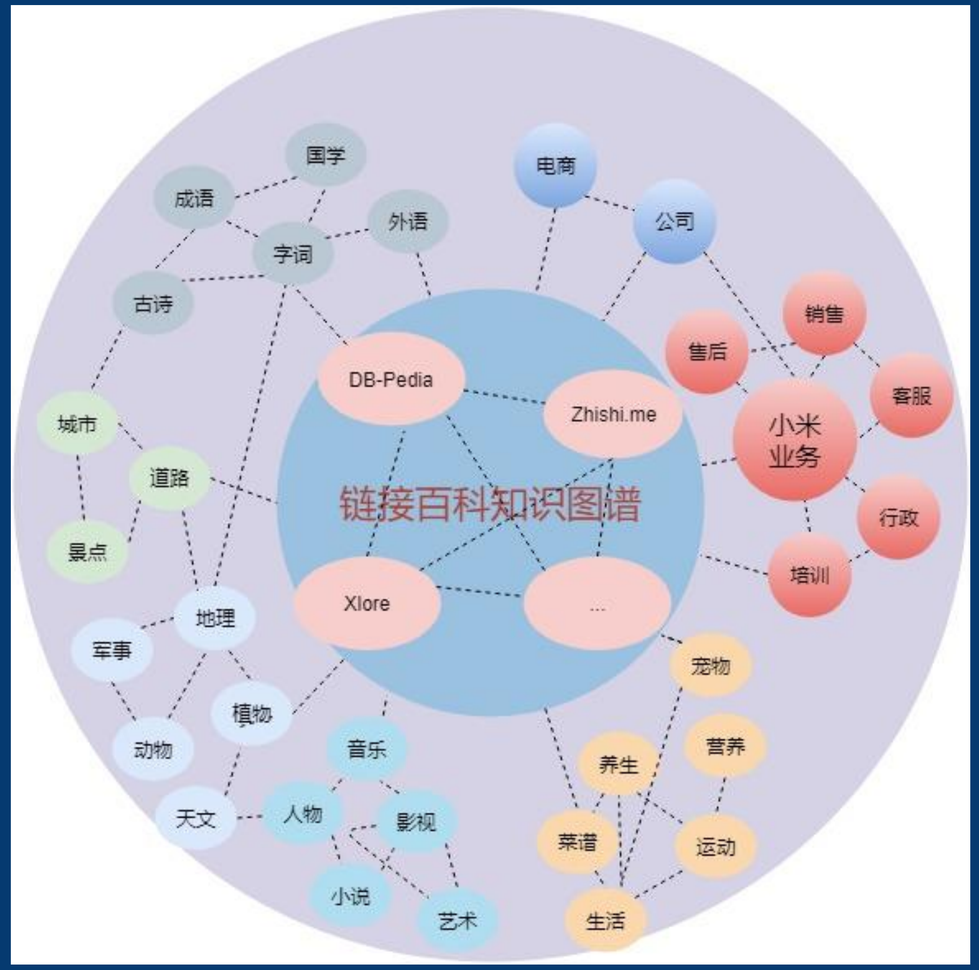


- 联合BERT和Meta Network的网络模型；
- Transformer层之间用残差网络连接，提升了模型训练速度和分类效果

图谱的质检流程

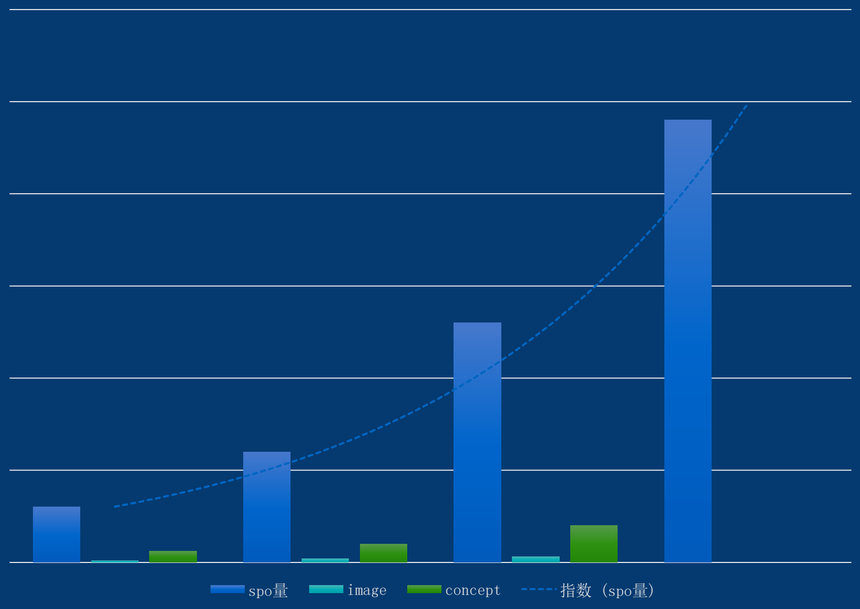

需求方
接口人
执行人
数据基本信息生成
离线获取数据准实时调用接口
需求背景数据字段指标定义确认
引入标准文档数据示例
确定排期与人力分配
试标注与试标反馈
任务执行
双副本验证
抽样验收
需求方验收
结论生成离线交付数据归纳文档与报表
达到预期数据入库/上线

小米图谱



- 以百科图谱为中心，链接了各垂类图谱；
- 在小米智能客服和行政助手场景下，小米业务图谱显著地提升了问答准确率；

小米图谱规模



- 截止到19年Q3，SPO数超过百亿；
- 高质量关系数目每季度翻倍增长；

TABLE OF

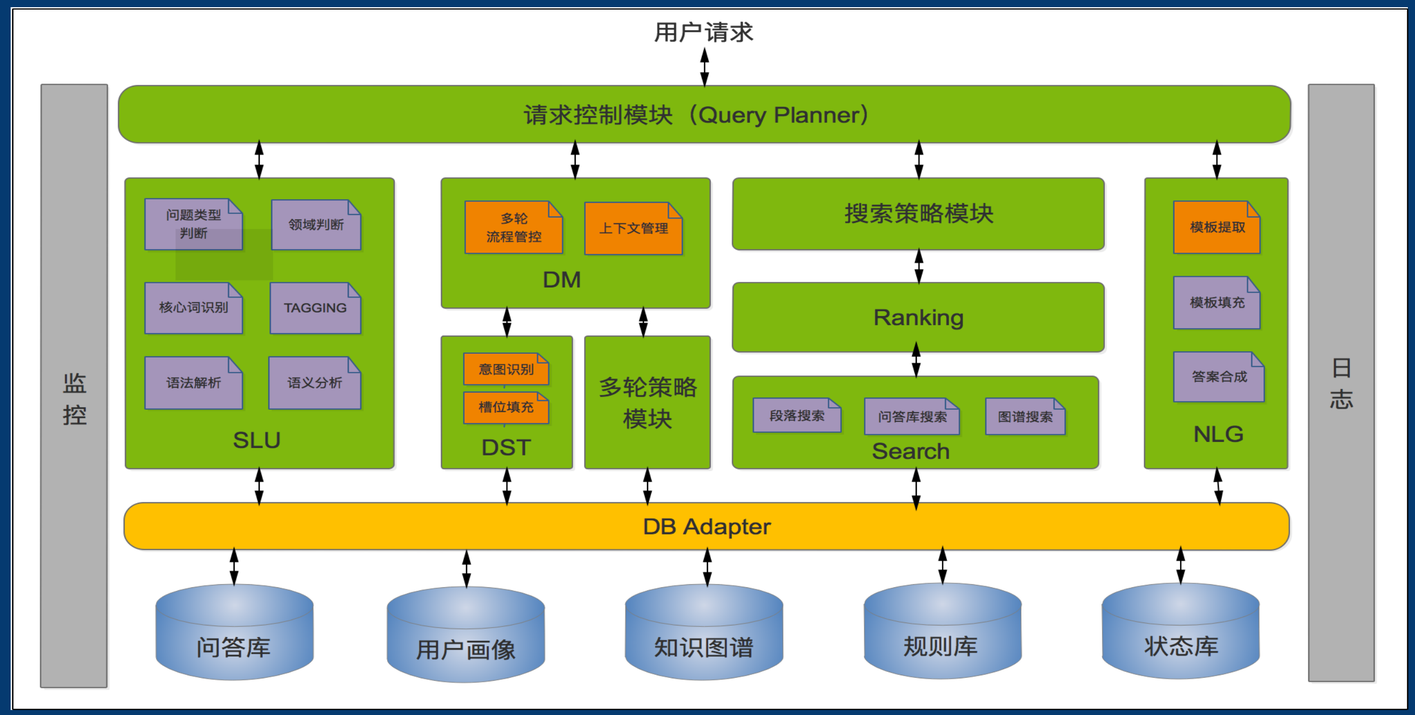
CONTENTS 大纲

- 知识图谱概述
- 知识图谱构建
- 基于图谱的问答
- 图谱的其他典型应用

小爱同学生态架构

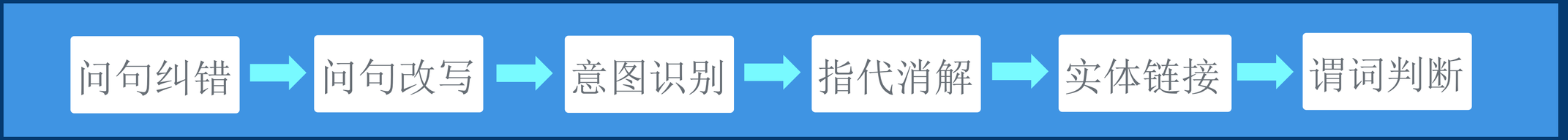

生态场景
手机
音箱
TV
电视
生态链
其他
开放平台
技能开发平台
语音服务平台
交互形式
语音交互
多模交互
内容服务
内容类
工具类
知识类
控制类
娱乐类
生活服务类
帮助类
APP类
闲聊类
...
已接入9大类别52个垂域

小爱开放域问答系统



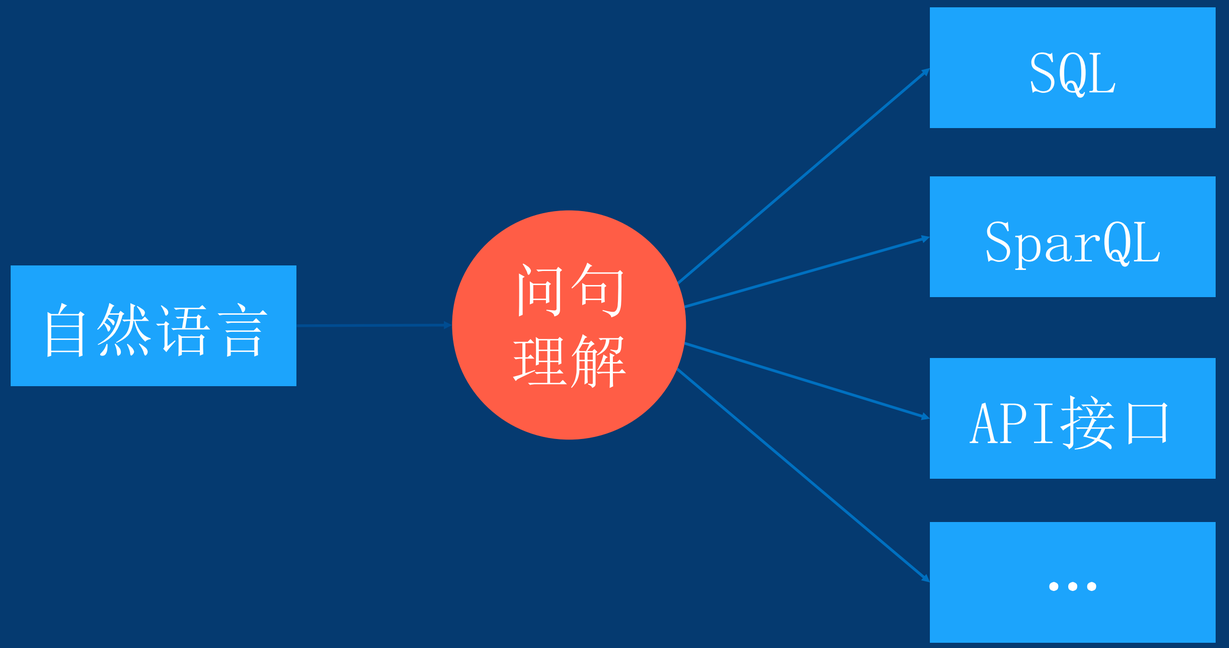
- KBQA是小爱开放域问答系统的重要组成部分；
- 从小爱的线上表现来看，KBQA的准确率，要显著高于基于FAQ的问答和基于阅读理解的问答子系统；

问句理解是当前的难点



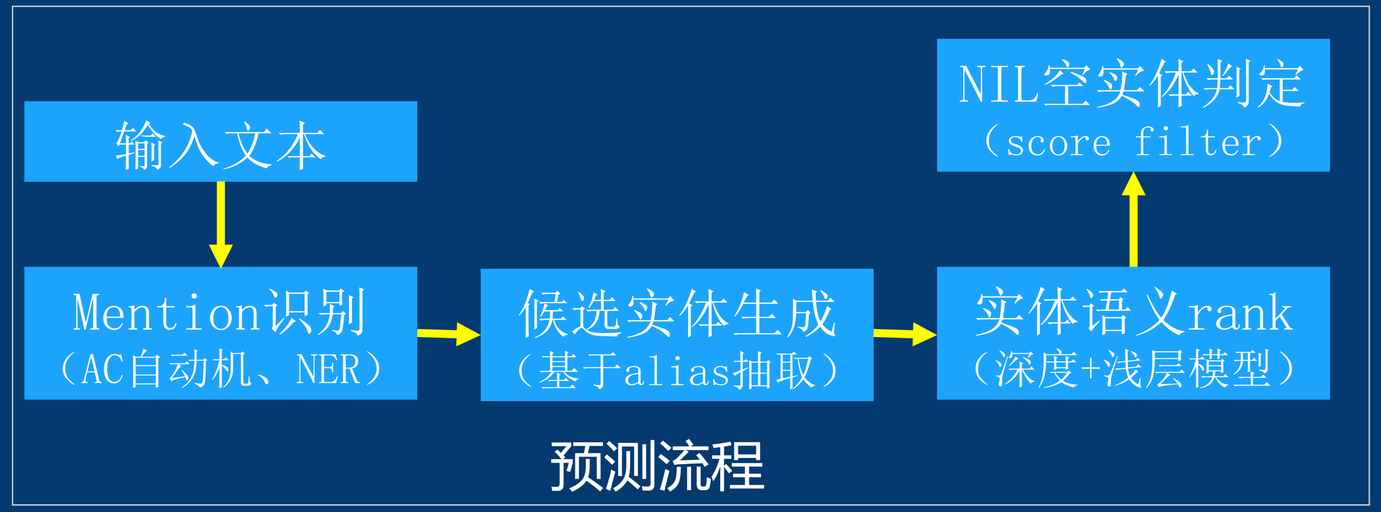
- 小爱的KBQA分成许多步，部分借鉴了搜索和NLP的技术
- 需要在80毫秒内完成全部处理，每天处理6千万次以上的请求

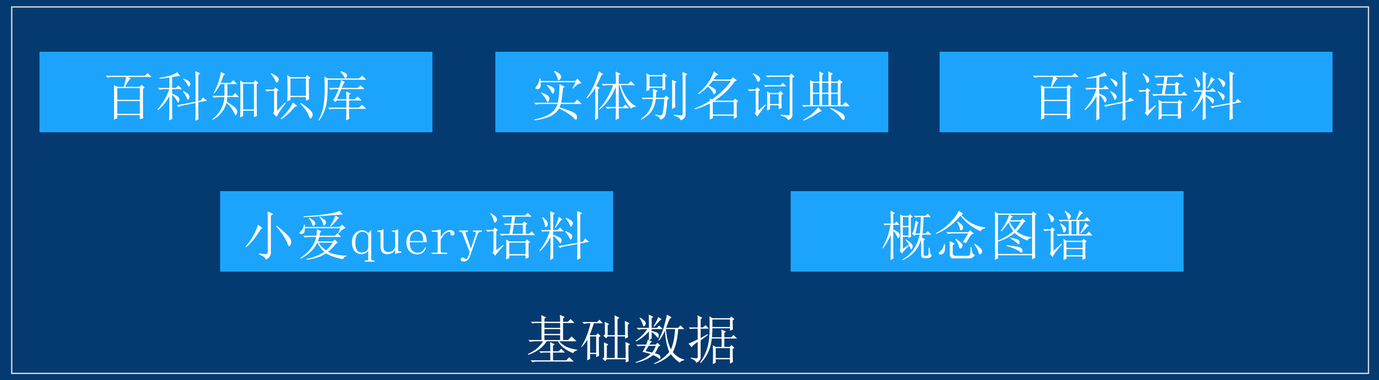
从翻译的角度看用户问句理解

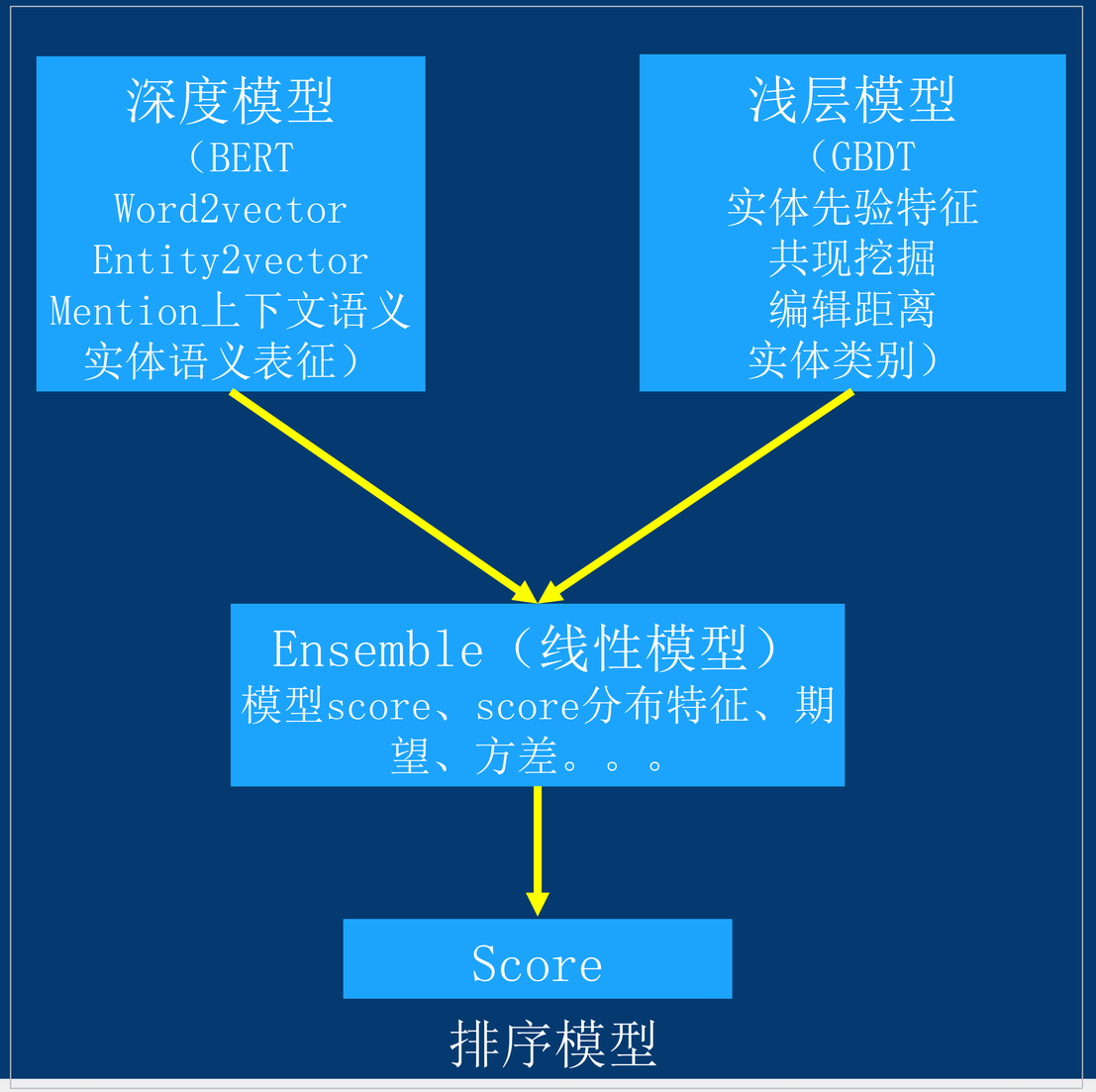


将自然语言翻译成SQL/SparQL等机器语言，方便后续的查询，如：
“你知道张三的电话是多少么？”=>SELECT Phone_number FROM Employees WHERE Name =“张三”

KBQA关键技术：实体链接





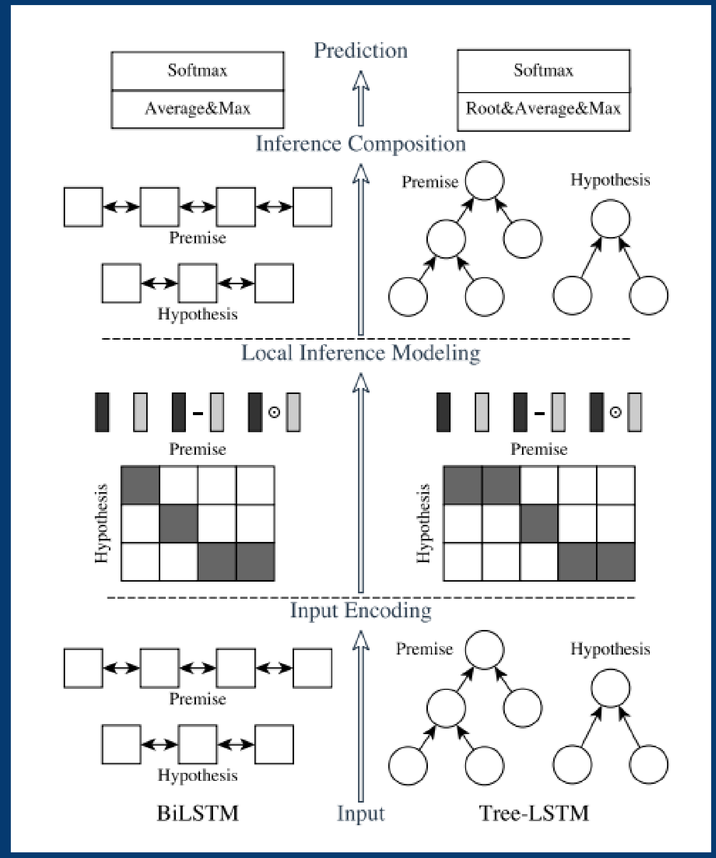


KBQA关键技术：谓词判断

- 头部问题用模板和规则的方式处理，模板的挖掘，可以采用BootStraping方法 ；

- 结合CFG-CKY Parser来解析成Lambda 算式，
例：“唐家三少的家乡有什么机场”=>
lambda_x_y.Birthplace(唐家三少,x) And Airport(x,y)；

- 长尾问题采用模型来处理，如利用ESIM模型：
将谓词关联到口语短语，利用短语与用户问句做匹配；



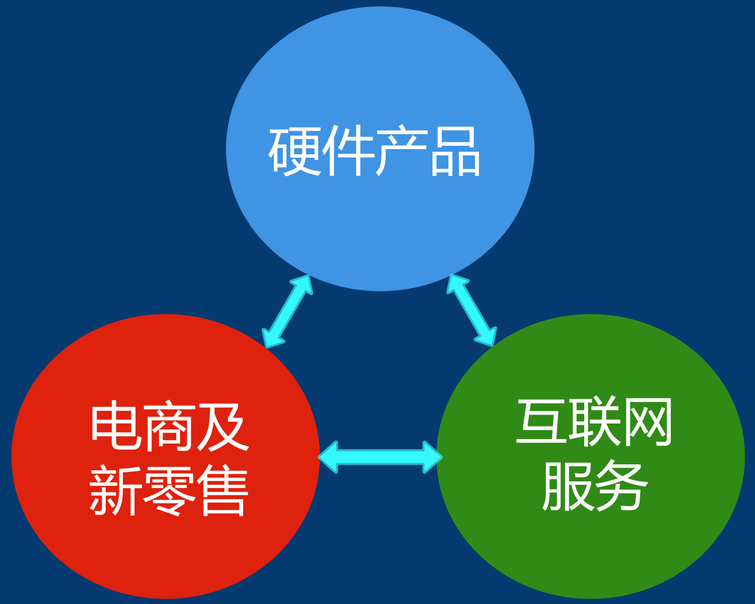
ESIM论文：《Enhanced LSTM for Natural Language Inference》

TABLE OF

CONTENTS 大纲

- 知识图谱概述
- 知识图谱构建
- 基于图谱的问答
- 图谱的其他典型应用

众多的应用场景



铁人三项

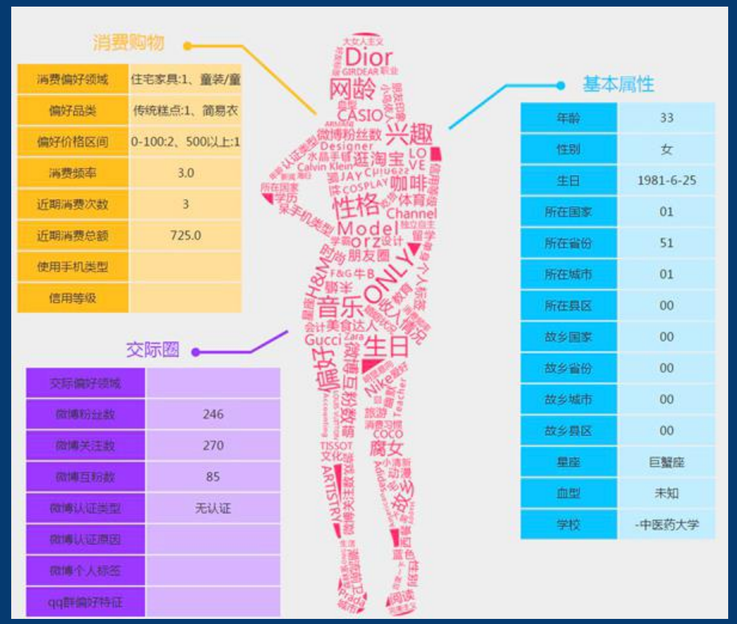
典型场景（一）

有品电商：构建电商图谱，提升推荐精度

新闻推荐：新闻去重

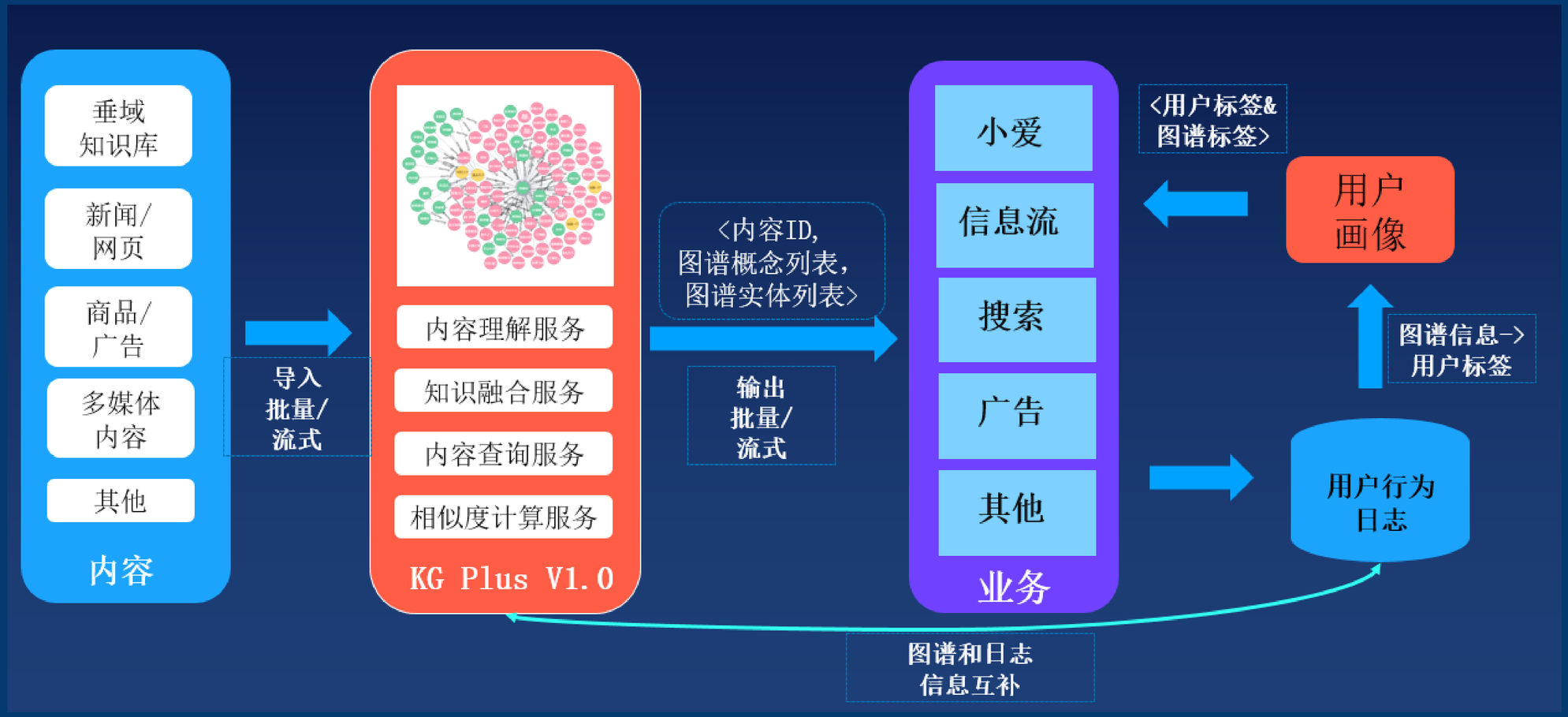
典型场景（二）

全局搜索：补充结构化内容



用户画像：补充标签数量

从KG到KG Plus


垂域
知识库
新闻/
网页
商品/
广告
多媒体
内容
其他
内容
导入
批量/
流式
内容理解服务
知识融合服务
内容查询服务
相似度计算服务
KG Plus V1.0
<内容ID,
图谱概念列表,
图谱实体列表>
输出
批量/
流式
小爱
信息流
搜索
广告
其他
业务
<用户标签&
图谱标签>
用户
画像
图谱信息->
用户标签
用户行为
日志
图谱和日志
信息互补

KG Plus

- 将结构化的图谱数据与非结构化的文本数据和多媒体数据相关联；
- 通过KG Plus平台，统一为各上游产品提供服务；
- KG Plus 平台提供的核心能力如下：

1、KG Plus查询：根据内容名称或者内容ID，从KG Plus里查询内容；

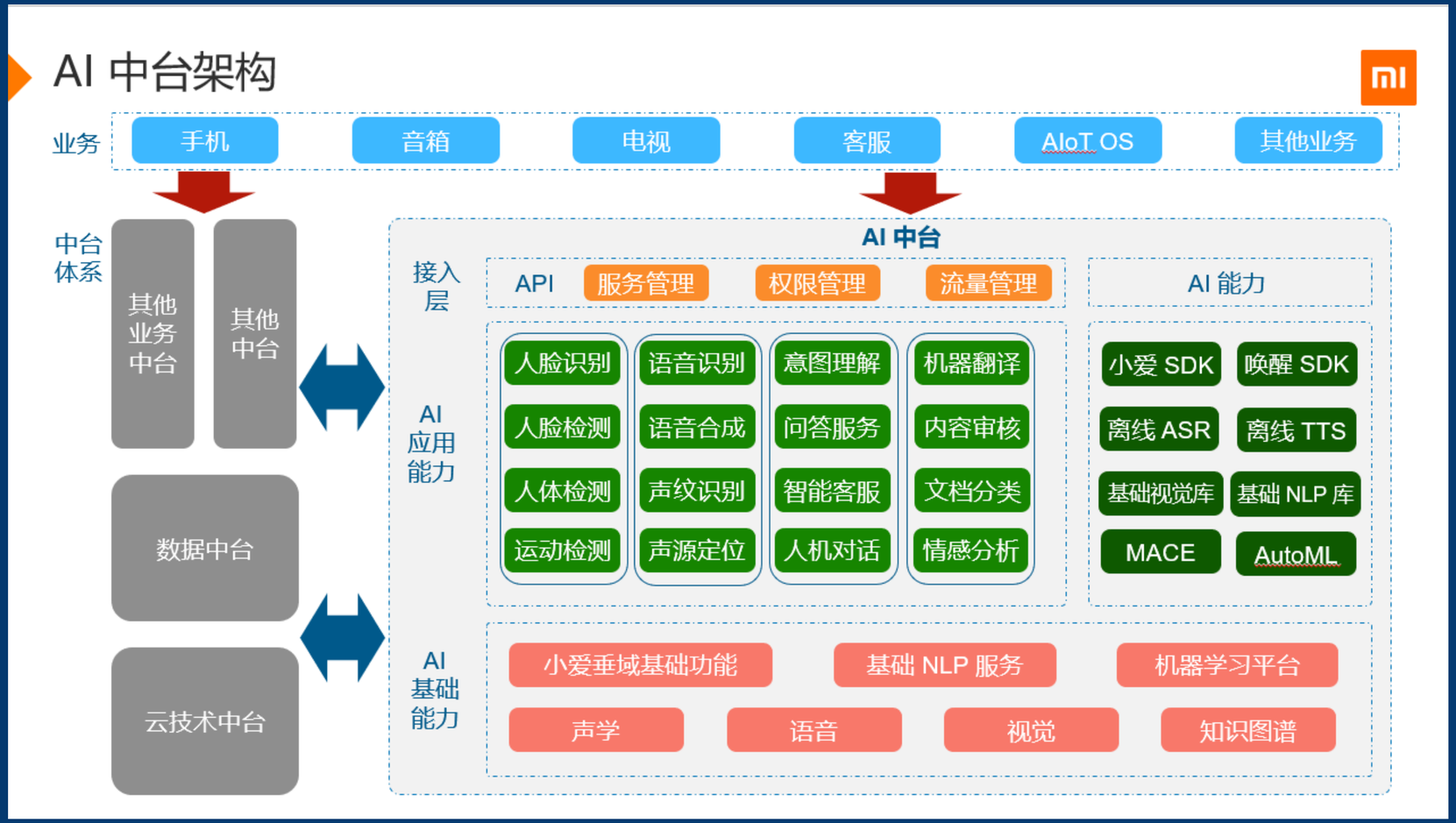
2、理解服务：根据内容名称或者内容ID，查询补充了各类信息的内容数据；

3、融合服务：将结构化的和非结构化的数据融合到KG Plus中；

4、距离计算：查询实体之间或者概念之间的图谱距离 ；

5、实体推荐：根据给定实体，推荐关联实体；

KG Plus在小米AI中台的位置



主办方 Geekbang. 极客邦科技 InfoQ

THANKS

AiCon
全球人工智能与机器学习技术大会