

# 电商领域的智能私人助理

## --阿里小蜜技术实践与演进

陈海青(海青)

阿里巴巴客户体验事业群-智能创新中心-高级技术专家



促进软件开发领域知识与创新的传播



关注InfoQ官方微信  
及时获取ArchSummit  
大会演讲视频信息



全球软件开发大会 [北京站]

2017年4月16-18日 北京·国家会议中心

咨询热线: 010-64738142



全球架构师峰会 2016 [深圳站]

2017年7月7-8日 深圳·华侨城洲际酒店

咨询热线: 010-89880682

# 大纲

第一部分 阿里小蜜平台介绍

第二部分 智能人机交互构建技术实践与演进

第三部分 挑战与未来

# 大纲

第一部分 阿里小蜜平台介绍

第二部分 智能人机交互构建技术实践与演进

第三部分 挑战与未来

# 电商领域的私人助理实践-阿里小蜜



基于阿里海量消费数据，结合线上、线下的生活场景需求，以智能+人工的模式提供智能导购、服务、助理的业务体验

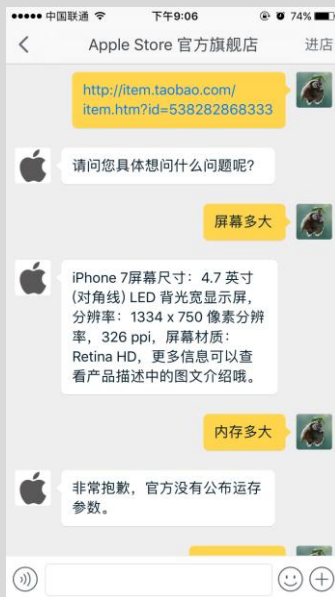
电商领域的平台化开放

- 商家开放-千牛平台
- 企业开放-钉钉平台

# 阿里小蜜及平台输出展示



阿里小蜜

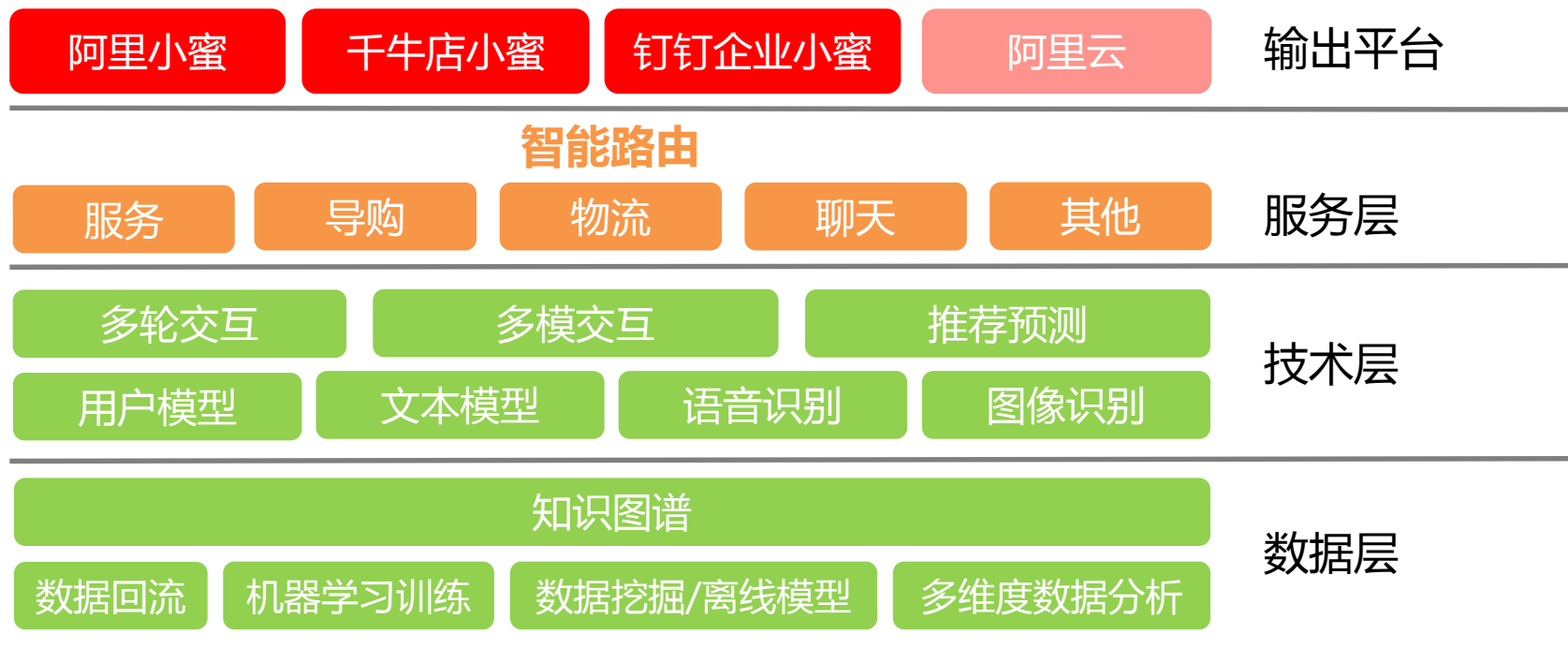


店小蜜



企业钉小蜜

# 阿里小蜜平台



# 大纲

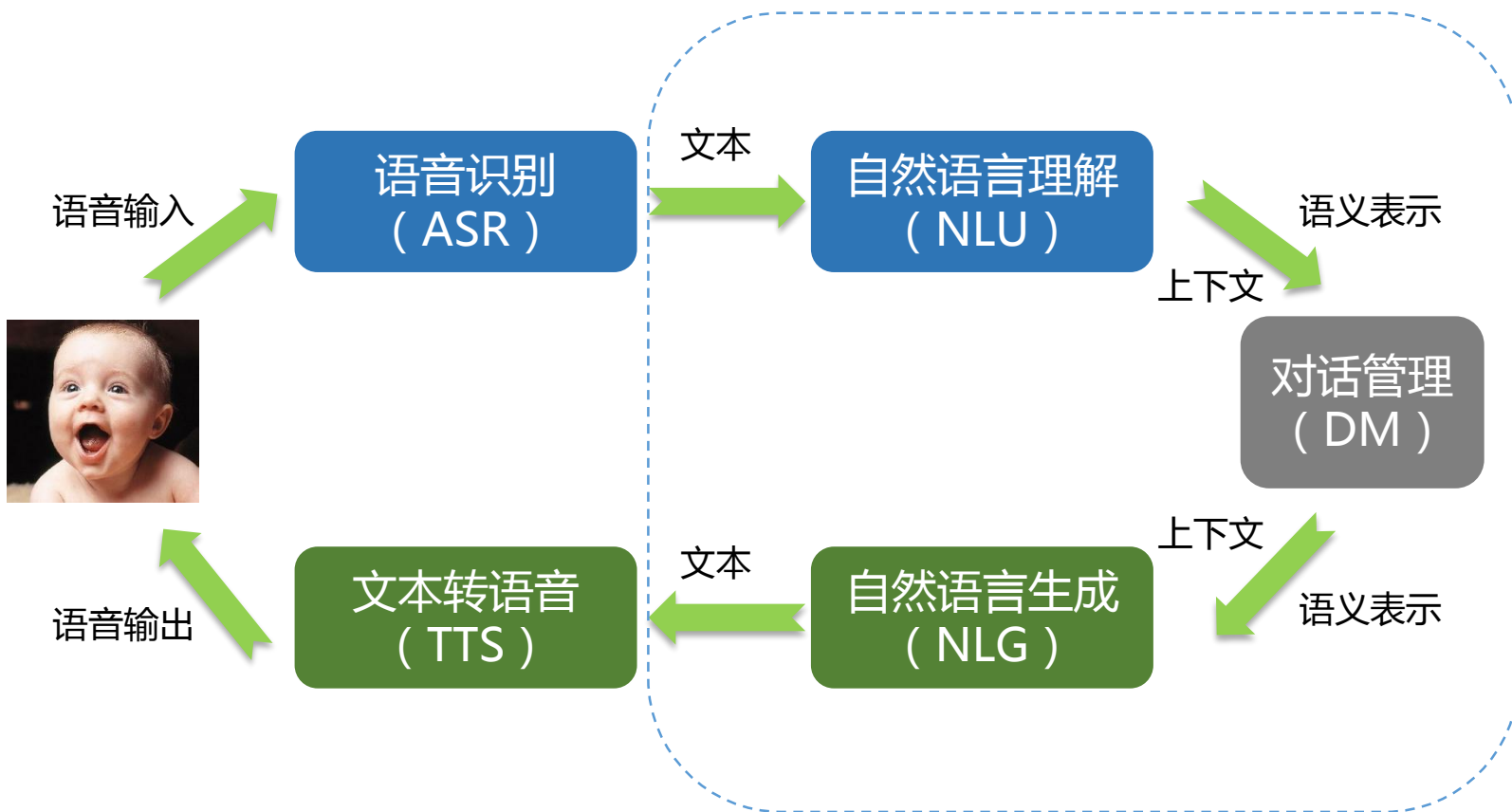
第一部分 阿里小蜜平台介绍

第二部分 智能人机交互构建技术实践与演进

第三部分 挑战与未来



# 智能人机交互的基本技术流程



# 现今四种主流的问答匹配技术

- 1 基于模板式匹配(Rule-Based)
- 2 基于检索的模型(Retrieval model)
- 3 基于统计机器翻译模型(SMT)
- 4 基于深度学习模型(Deep Learning)

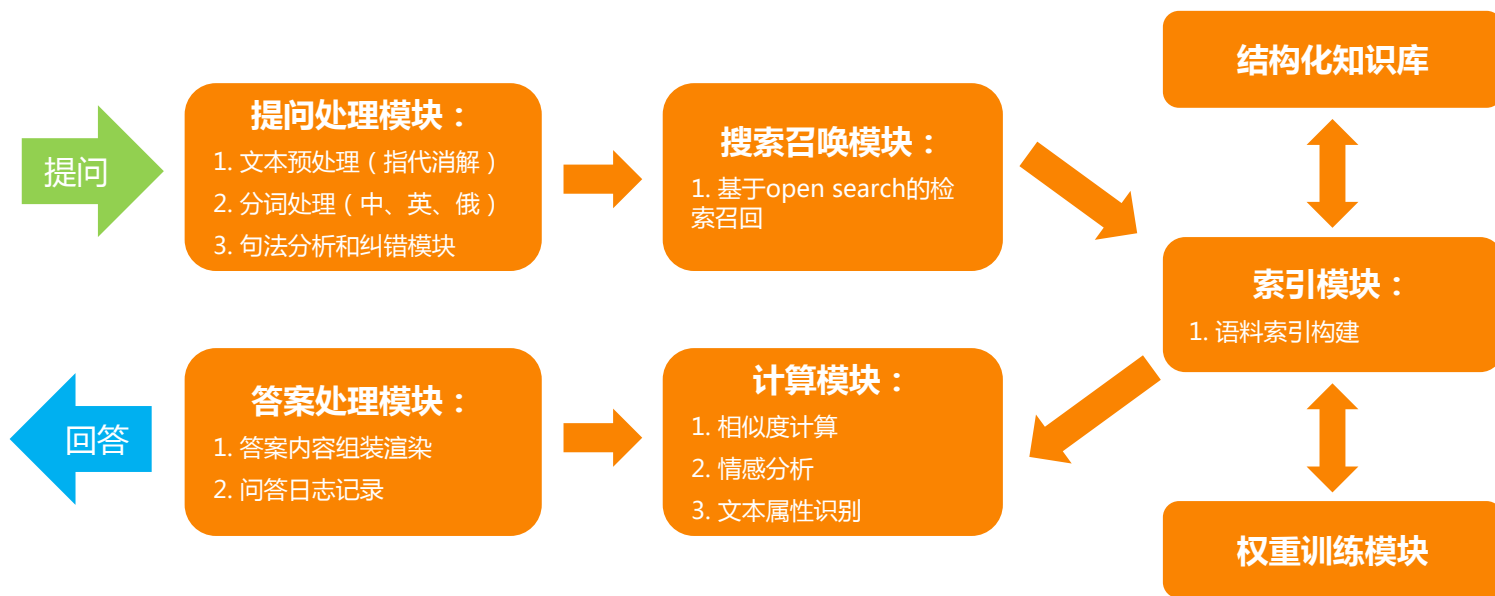
# 技术演进之路

- 一切从搜索出发
- 意图识别与匹配的分离
- 以知识为核心的挖掘与优化
- 深度学习技术的应用

# 一切从搜索出发

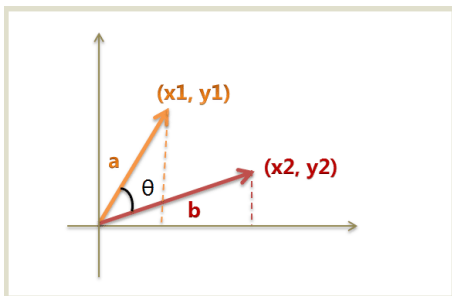
## | 基于检索模型的问答系统构建

# 检索模型(Retrieval Model)构建体系



# 检索模型的构建

- 基本思路：
- 基于传统的检索作为基础，以TF\*IDF做为向量，通过文本距离计算进行计算，获取Recall@1
- 距离算法举例：余弦相似度算法



$$\begin{aligned}\cos\theta &= \frac{\sum_{i=1}^n (A_i \times B_i)}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} \\ &= \frac{A \cdot B}{|A| \times |B|}\end{aligned}$$

- 优点与缺点：
- 优点：具备问答体系的平台化与统一性，可平台扩展，并且计算特征可不断添加提升
- 缺点：意图和匹配混淆，在复杂问答场景无法细分

# 意图识别与匹配分离

- | 分层策略
- | 意图识别

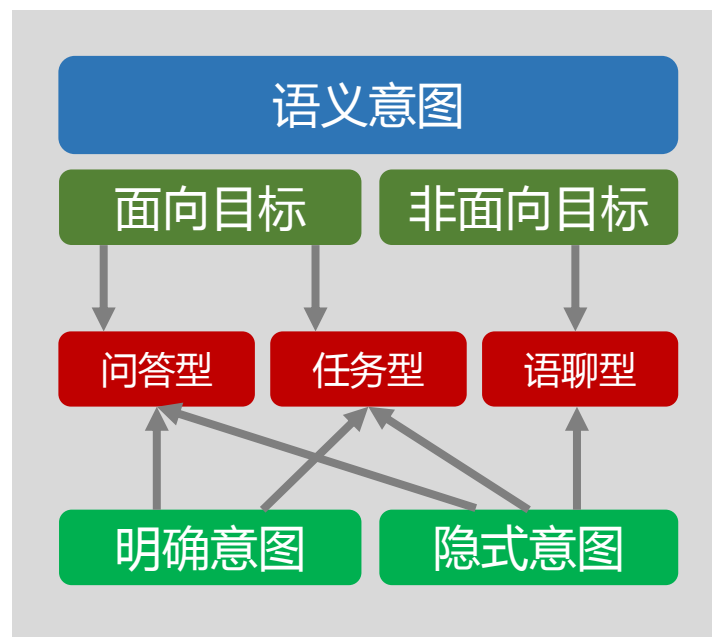
# 智能人机交互分层策略

## 语义意图识别：

- 目标和非目标任务的细分
- 意图的明确与推理匹配

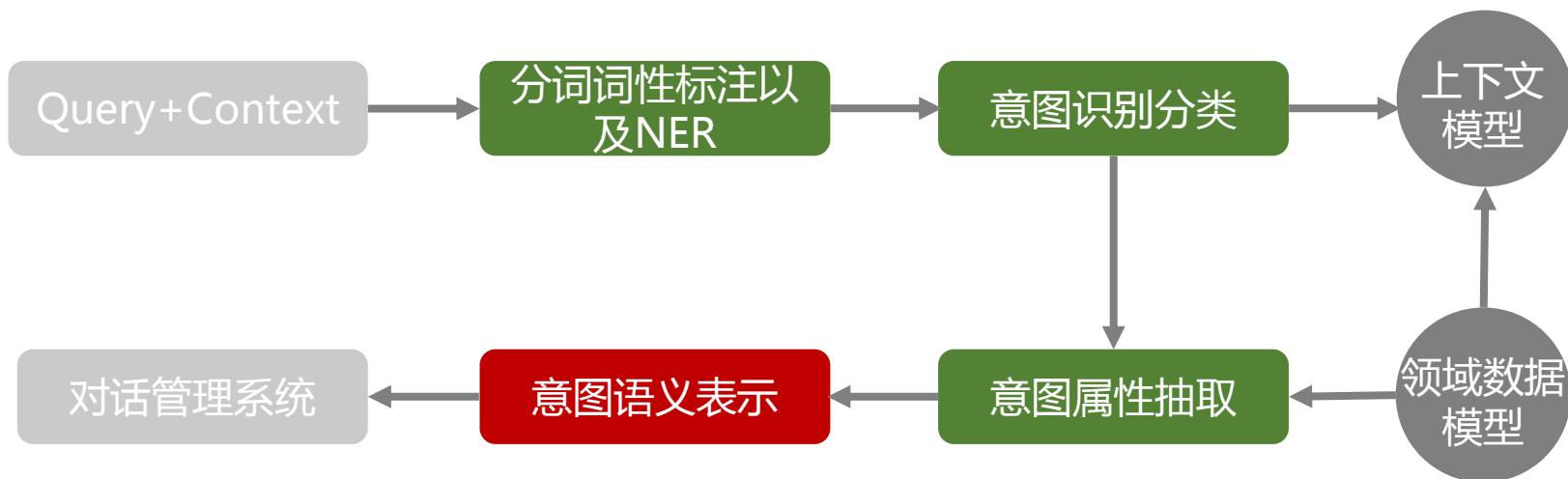
## 面向3类业务划分并建立技术体系：

- 问答型：“密码忘记怎么办？”
- 任务型：“我想订一张明天从北京到杭州的机票”
- 语聊型：“我心情不好”





# 语义意图识别的基本技术流程



# 意图识别构建

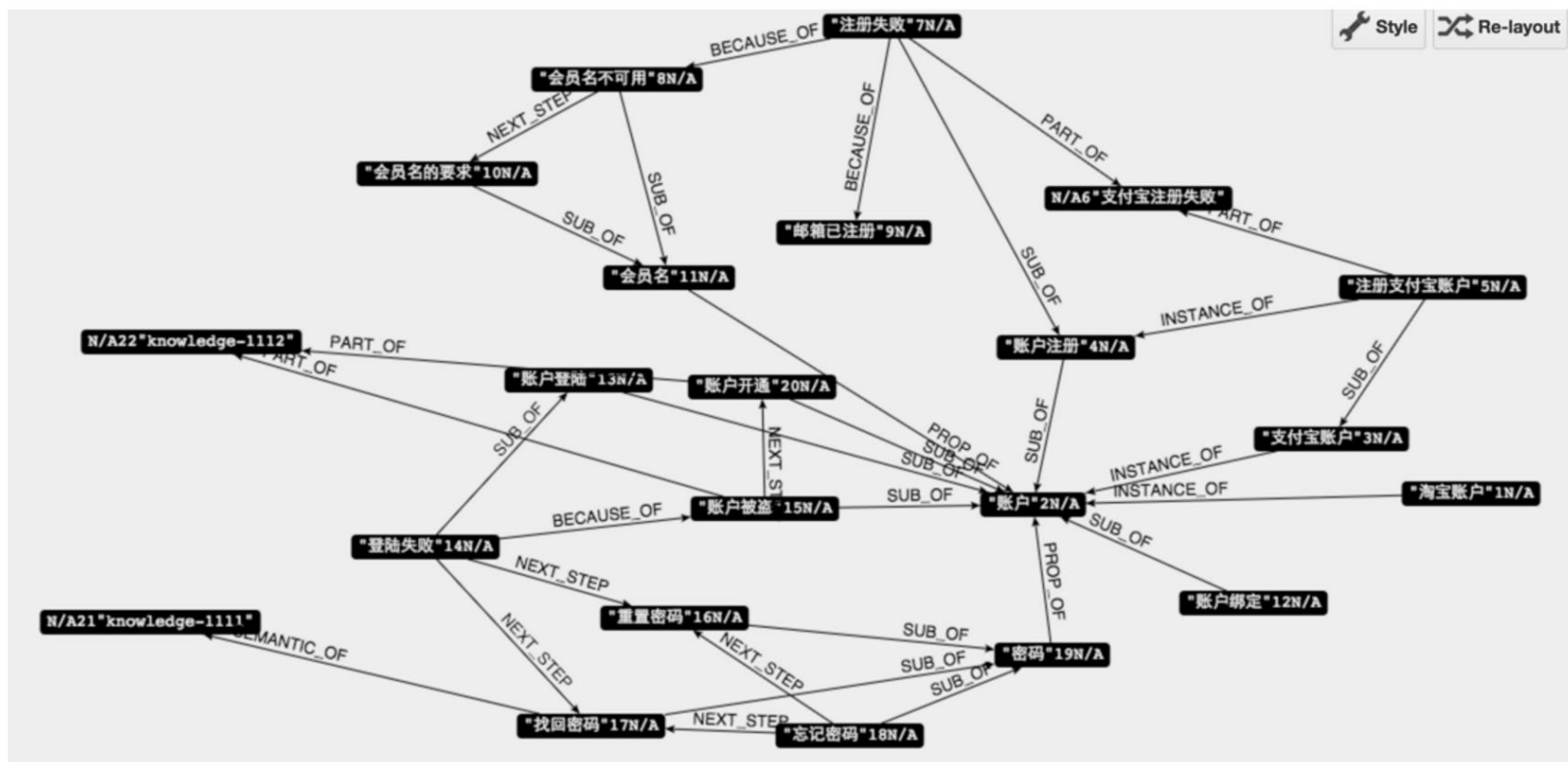
- **基本思路：**
- 抽象成为分类问题，构建分类模型对意图进行预测
- **多分类和二分类选型对比**
- 多分类模型(有监督的分类算法，依据具体场景进行选型  
Bayes\Knn\最大熵\...)：适用于相对简单场景且分类数稳定领域
- 二分类模型(按照意图领域做成多个二分类模型 SVM\...)：适用于领域分类相对独立，并且经常需要新增修改的场景，能做到相互独立

# 以知识为核心的挖掘与优化

- | 知识构建的两种形式：图与本体
- | 知识的挖掘与构建

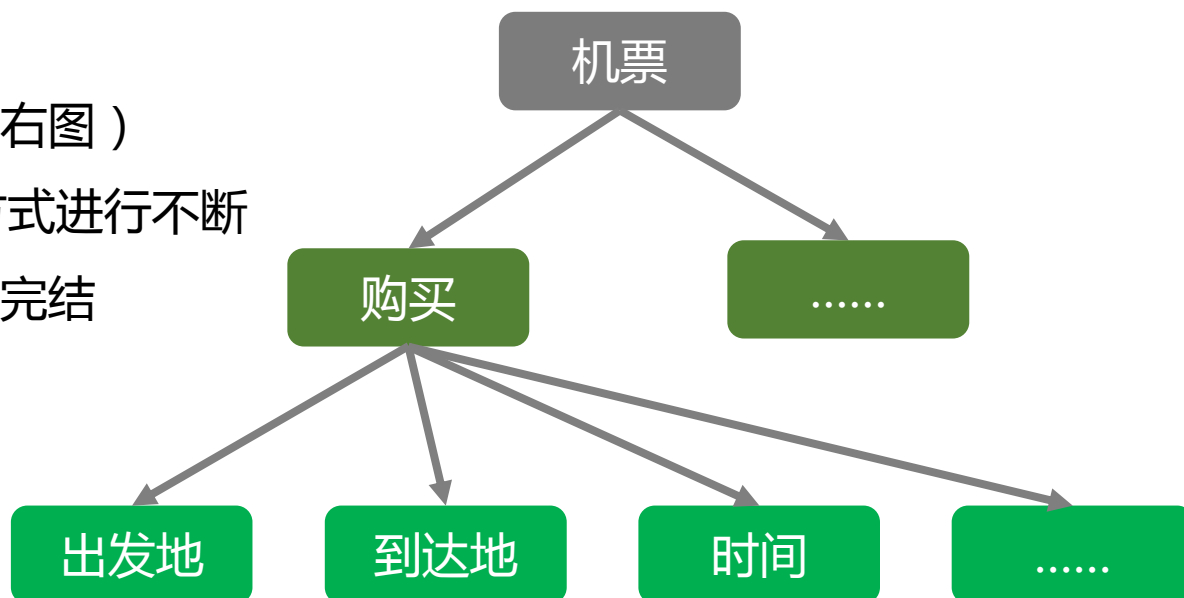


# 知识图谱的构建



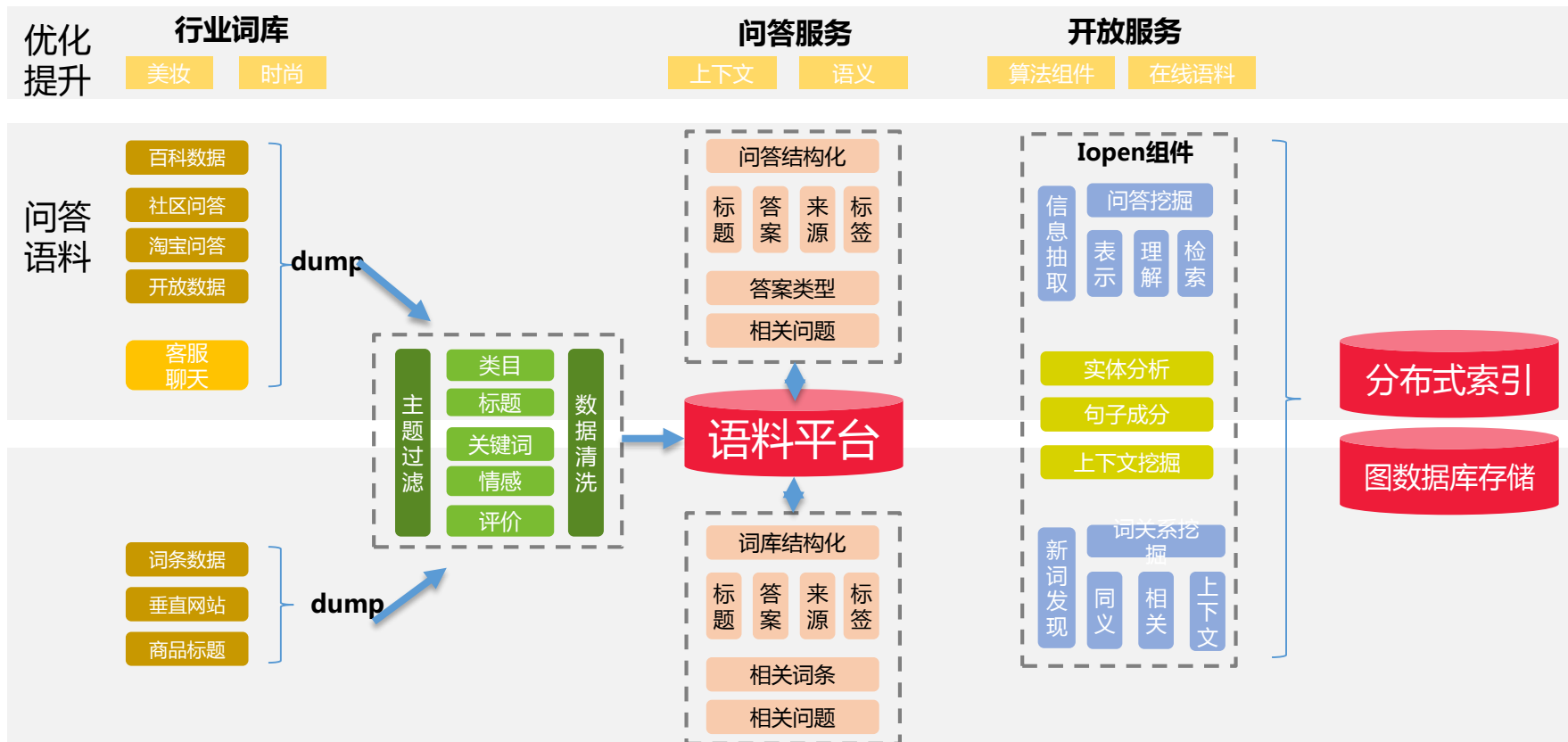
# 本体知识的构建

- 构建领域意图树（如右图）
- 通过slots filling的方式进行不断判断、填充、转移和完结



# 知识构建与挖掘

通过多数据源完成实体和结构化短句的挖掘与积累，并最终生成知识图谱或可用语料



# 语义挖掘

- 同义语义挖掘、相似词挖掘、构造模式 ( pattern )
- 相似词挖掘主要是选取标签中的业务词作为种子，例如：通过 Word2Vec 扩展相近词汇，用于词语的归一化
- 同义语义挖掘旨在挖掘同义问法





# 同义语义挖掘方法

## | 文本相似度：

通过相似度算法，结合了同义词替换和同音词替换等预处理

# 潜在语义空间构造流程

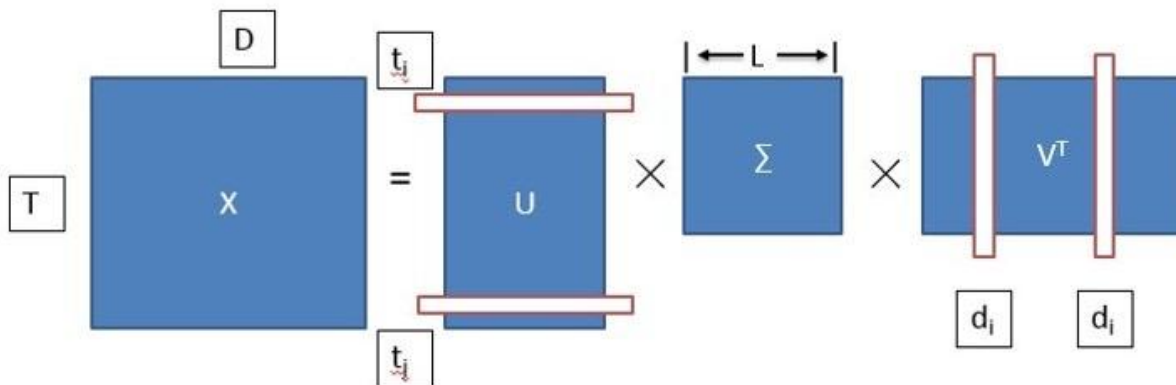
- 1 分析文档集合，建立Term-Document矩阵
- 2 对Term-Document矩阵进行奇异值分解(SVD)
- 3 对SVD分解后的矩阵进行降维，低阶近似
- 4 构建潜在语义空间，用相似度或者聚类

# 潜在语义分析

**潜在语义分析(LSA)**：文本从稀疏的高维词汇空间映射到一个低维的向量空间，相比传统向量空间，维度更小，语义关系更明确

**潜在语义空间的构造方法**：SVD奇异值分解

•核心在于降维。将原始的Term-Document矩阵，转换为三个矩阵乘积，其中 $\Sigma$ 即为潜在语义空间（假设有L维），词（Term）和文档（Document）可映射为L维的向量（下图中： $t_i, t_j, d_i, d_j$ ），再做后续基于向量的计算。



# 深度学习技术的应用

- | 结合用户行为特征的深度意图识别
- | 检索模型与深度学习模型相结合的语聊模型构建

# 结合用户行为特征的深度意图识别

- **基本思路：**
- 由于文本缺失、不明确或者不完整的情况下，增加用户行为特征进行意图分类预测
- 深度学习分类模型(CNN\DNN\LSTM\...)：大数据量下的提升

# 深度学习模型意图数据积累

## 保证相关性

初步建模 - 保证相关性

- 用户问句和知识点标题的相似度
- 用户问句和历史问句的相似度

## 保证多样性

预留位置展示 - 保证多样性

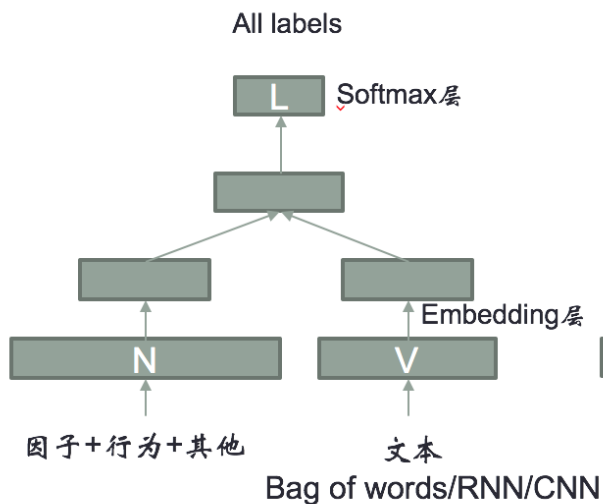
- 随机展示（均匀采样）
- 按历史的知识点使用频次来推荐

# 结合用户行为特征的深度学习意图预测模型

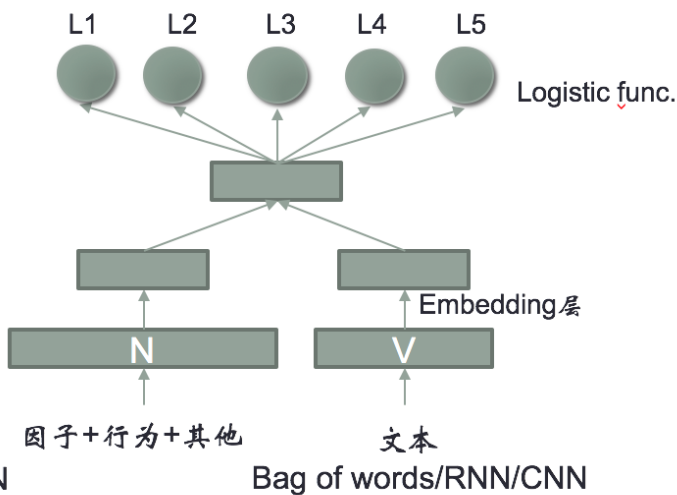


# 深度学习模型意图识别方案

- DNN 2-channel inputs



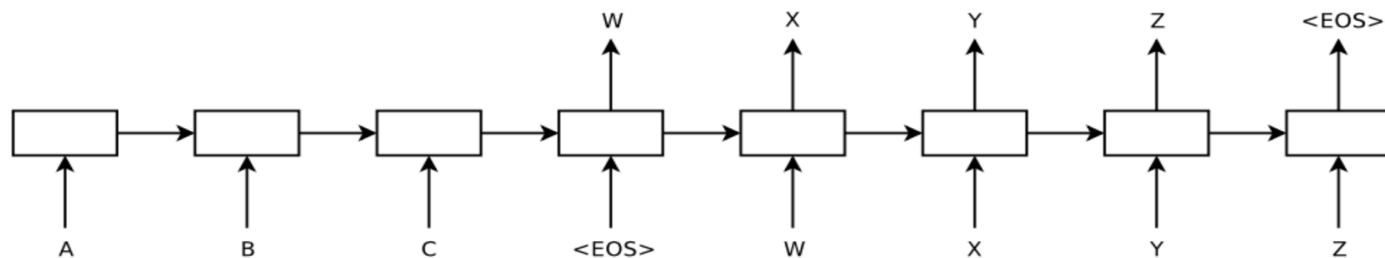
- DNN 2-channel inputs + multiple labels





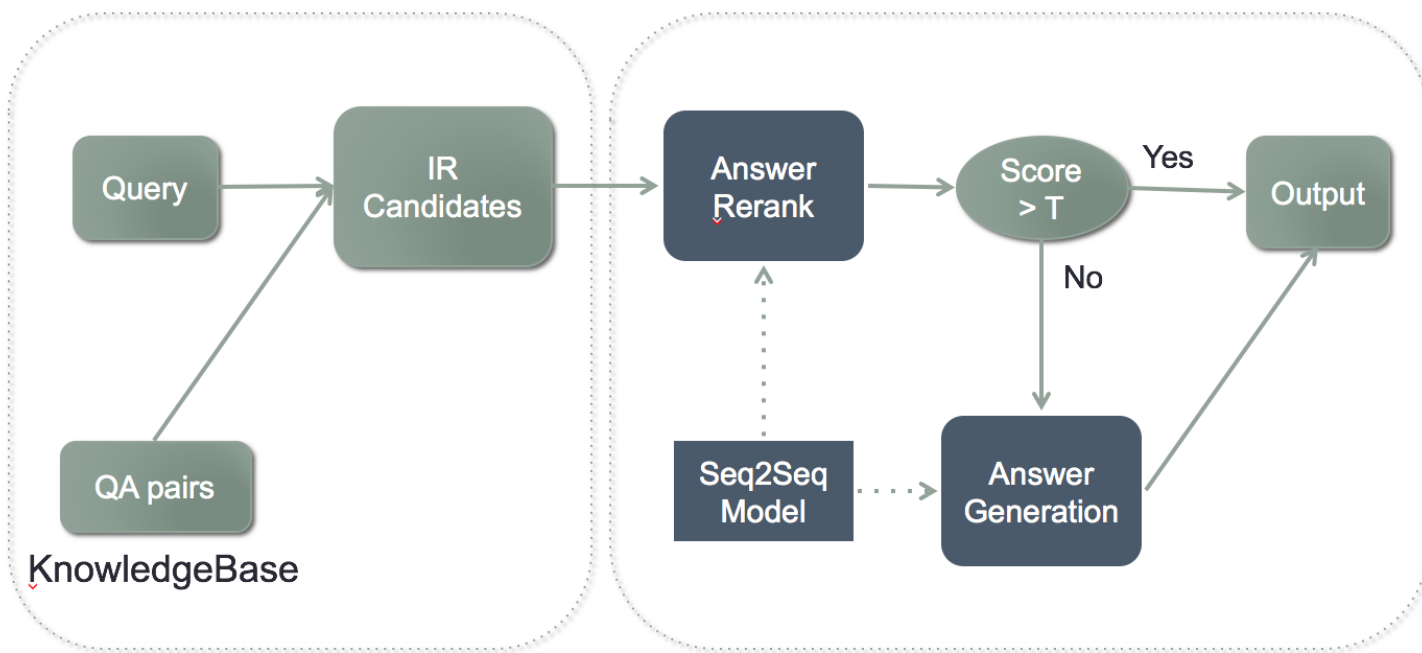
# 检索模型与深度学习模型相结合的语聊模型构建

- **检索模型与生成模型：**
- 检索模型：具有平台化扩展和答案语法合理性的优势，但是对结构化数据依赖大受限于语料库的完备性，在语义递进等聊天领域效果不好
- 生成模型：不受语料库范围，具有语义层面递进与识别优势，但是难以保证语法一致性与答案合理性



- **基本思路：**
- 结合检索模型和深度学习生成模型的优势进行混合模型构建

# 检索模型与深度学习生成模型相结合的方案

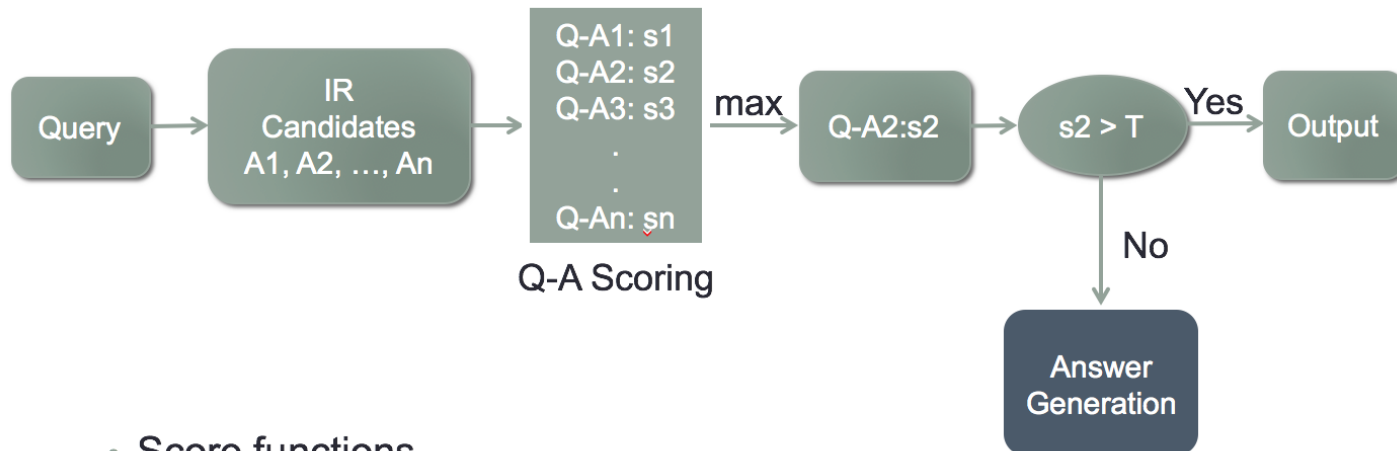


Information Retrieval

基于Seq2Seq的Rerank和Generation融合方案

# Seq2Seq Rerank Module

- Answer reranking



- Score functions
  - Q for encoder, A for decoder
  - Cross-ent, Averaged prob. > Geometric average

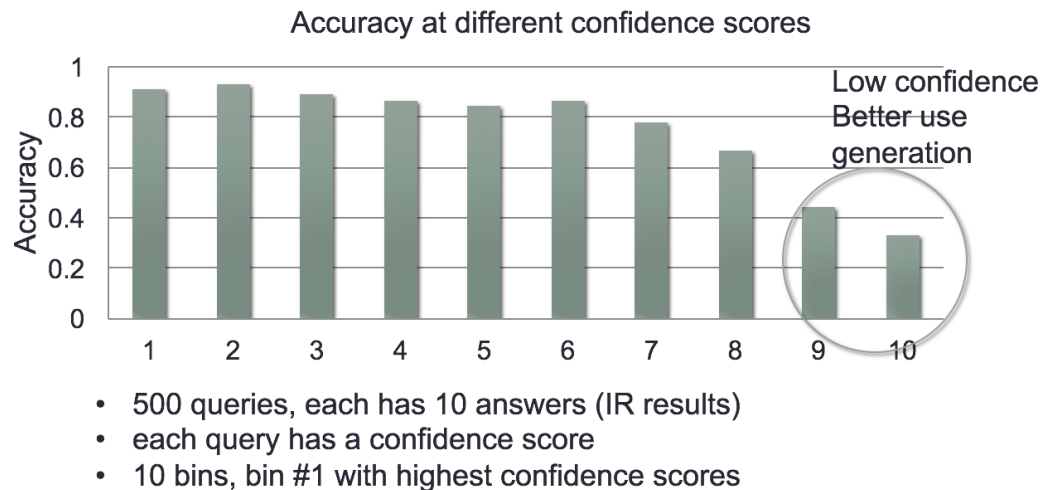
# Rerank and Generation

## Generation:

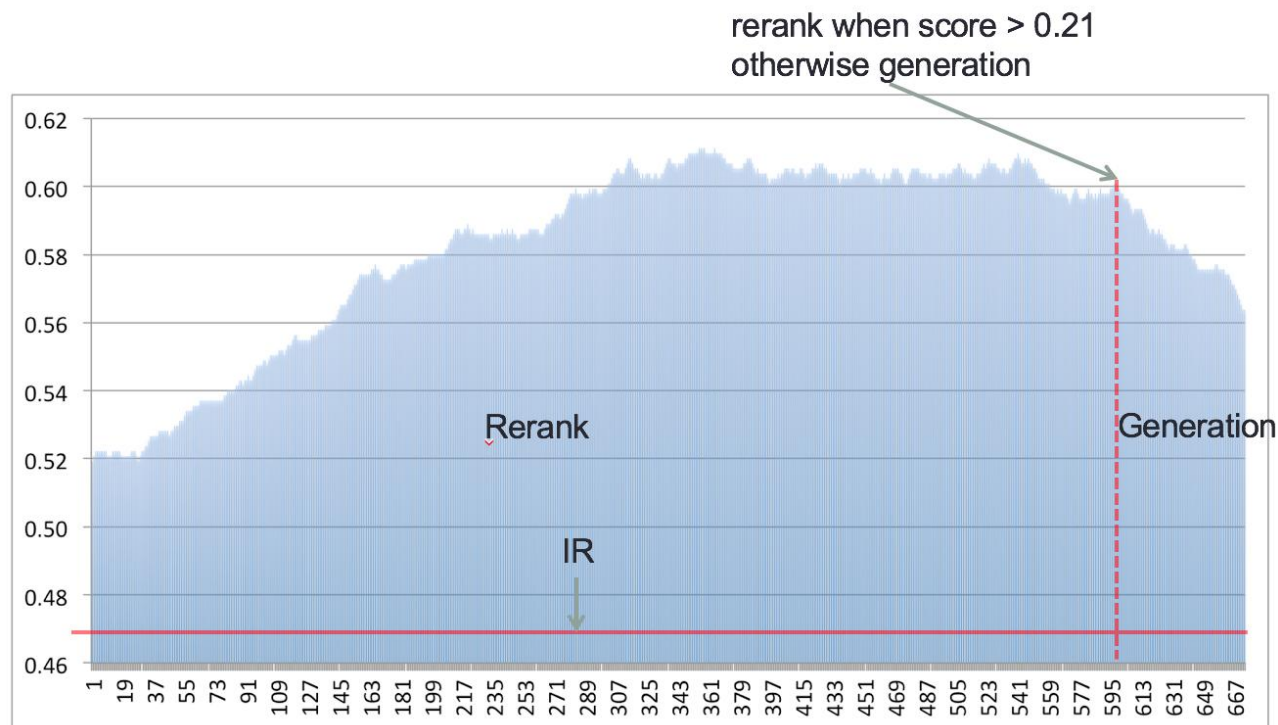
- 不受语料库限制，难以保证一致性和合理回答

## Reranking:

- 虽有固定语料库的局限性，但是能rank出更合理的答案



# Rerank Generation 离线数据结果



Rerank和Generation的融合结果 (Top 1 Acc)

# 大纲

第一部分 阿里小蜜平台介绍

第二部分 智能人机交互构建技术实践与演进

第三部分 挑战与未来

# 挑战与未来

## 挑战

- 目前智能人机交互机器人的智能程度还比较低，还有很长一段路要走
- 在工业领域由于涉及的领域以及复杂度情况很多，需要进行不断的细分并通过不同的方案来解决

# 挑战与未来

未来

- 随着计算机计算能力的大幅度提升，数据将发挥更加重要的作用，通过增强学习的模式构建智能数据闭环，循环不断提升智能程度
- 技术的不断突破，特别是Deep Learning、Transfer Learning在NLP领域的不断突破与应用将成为未来需要大家一同探索的方向



# Thanks

