行业知识图谱关键技术及应用

唐呈光(蒽竹)

阿里巴巴智能服务-云小蜜对话智能技术-算法专家

2020/11/15

Contents
目录

云小蜜技术整体概览


电信运营商
数字政府
金融客户
生态客户
三方ISV实施
客户自研
PaaS接入层（提供机器人平台输出能力）
平台产品能力
对话工厂
知识云
AIBoost
策略中心
智能路由
智能决策辅助
智能调度&预测
智能质检
权限管理
数据报表丨蜂眼
开发者及合作伙伴生态
认证课程
培训材料
工具包
SDK
算法模型能力
多轮对话
FAQs
KBQA
机器阅读
闲聊
预训练语言模型
营销模型
Ranker
情感识别
多语言
阿里云 – 公有云/私有云

行业KBQA面临的挑战

冷启动成本高

新场景上线KBQA要2周~1个月，时间花费：

1.图谱Schema构建；

2.三元组抽取与答案校验；

3.语料收集与标注；

4.模型训练与调优；

复杂句语义理解难

Q：给53岁的老人投终身防癌险，保20年的话，到时可以领的保险金是多少?

A：请问您是否想要咨询以下问题?
1.投保电话；2.如何选择保险产品；3.投保流程

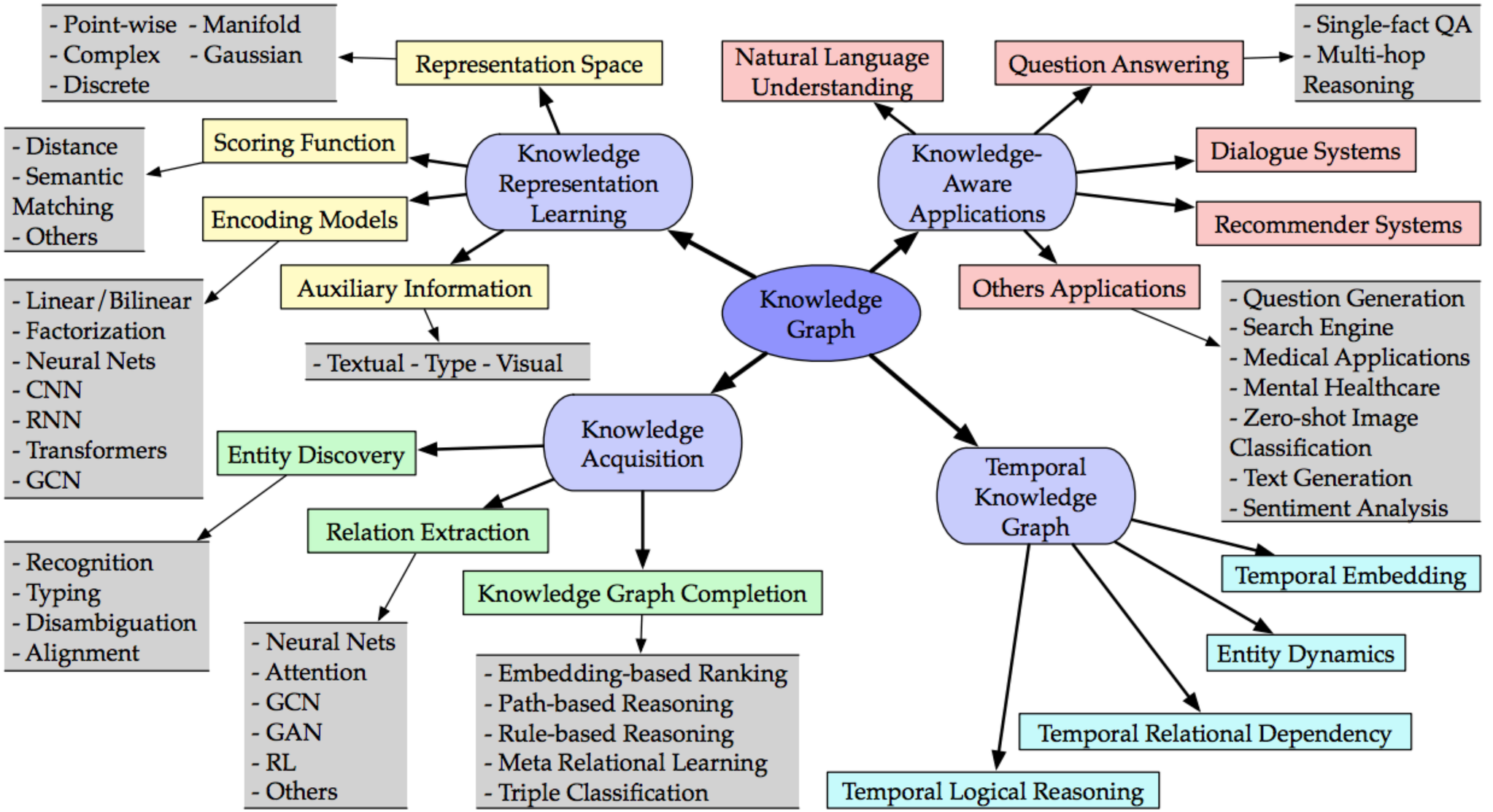
存在问题：系统无法给出精准答案

运营成本高

图谱是动态的，新增属性是常态，而数据标注成本高，效果难以持续提升

以税务为例，运营第1个月新增15个属性，平均每个属性7.6条样本，属性识别准确率70%

学术界图谱研究热点

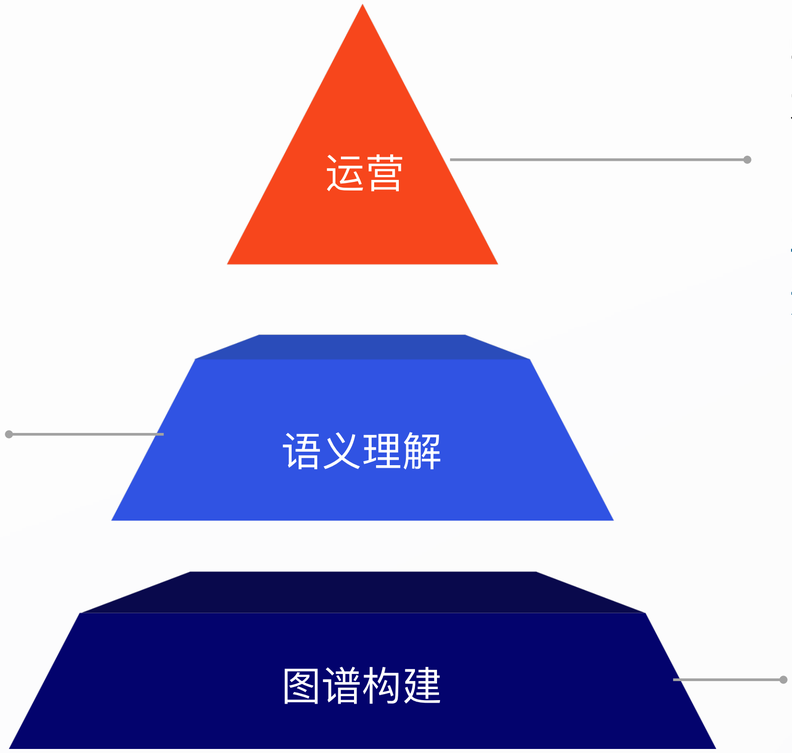


A Survey on Knowledge Graphs: Representation, Acquisition and Applications[J]. arXiv preprint arXiv:2002.00388, 2020.

学术界与工业界的差别

学术界：基于模板和改写构造数据集，语义丰富度不足

工业界：线上用户表述多样，语义理解难度大



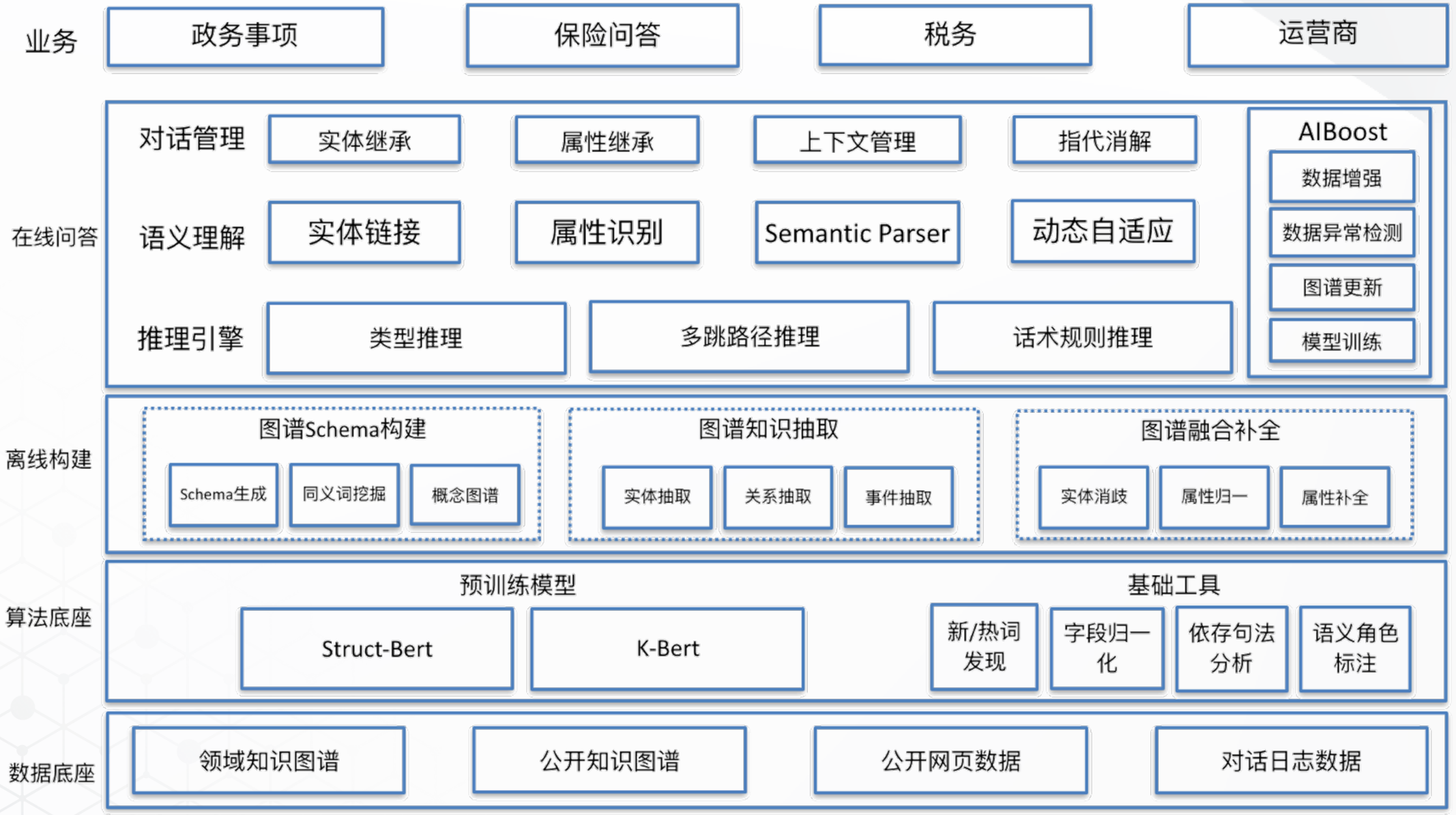
学术界：关注算法本身创新，不需要考虑运营问题

工业界：是否具备可干预、可运营能力，会影响业务落地

学术界：大多数实体、关系抽取方法聚焦在通用知识图谱领域

工业界：行业知识图谱，降低人工构建成本是关键，需要的是完整的人机协同构建链路

我们的整体方案


业务
政务事项
保险问答
税务
运营商
在线问答
对话管理
实体继承
属性继承
上下文管理
指代消解
语义理解
实体链接
属性识别
Semantic Parser
动态自适应
推理引擎
类型推理
多跳路径推理
话术规则推理
AIBoost
数据增强
数据异常检测
图谱更新
模型训练
离线构建
图谱Schema构建
Schema生成
同义词挖掘
概念图谱
图谱知识抽取
实体抽取
关系抽取
事件抽取
图谱融合补全
实体消歧
属性归一
属性补全
算法底座
预训练模型
Struct-Bert
K-Bert
基础工具
新/热词发现
字段归一化
依存句法分析
语义角色标注
数据底座
领域知识图谱
公开知识图谱
公开网页数据
对话日志数据

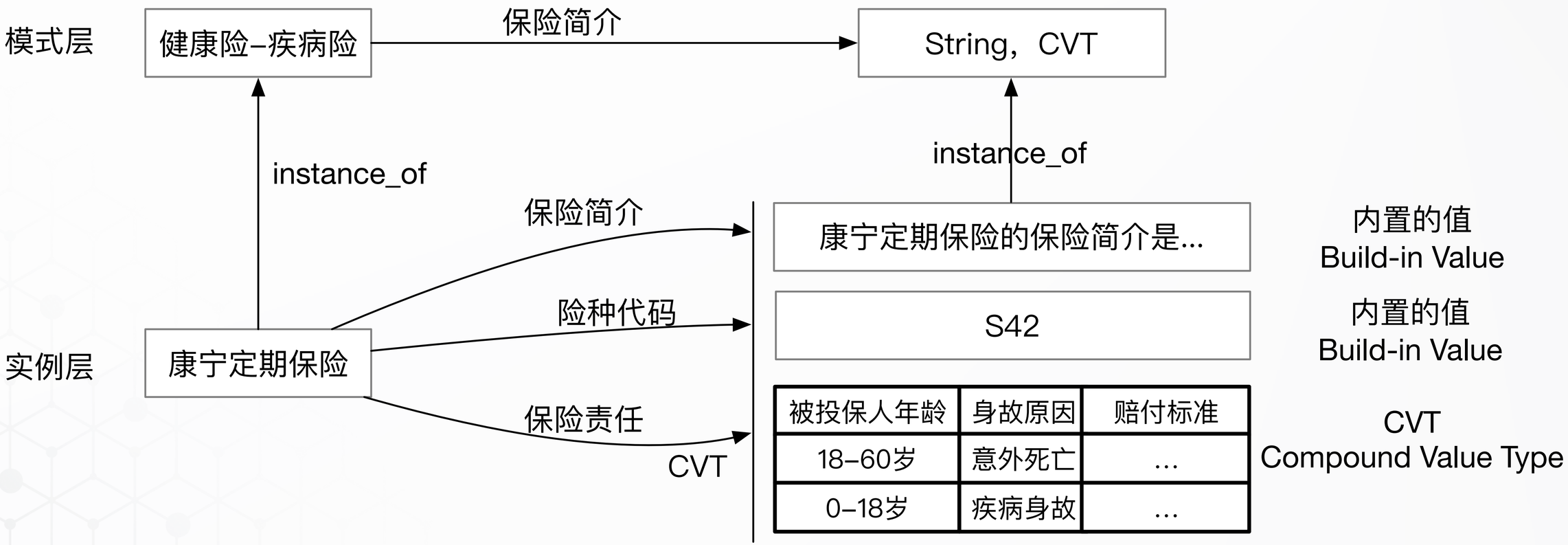
行业知识图谱构建

02

行业知识图谱建模

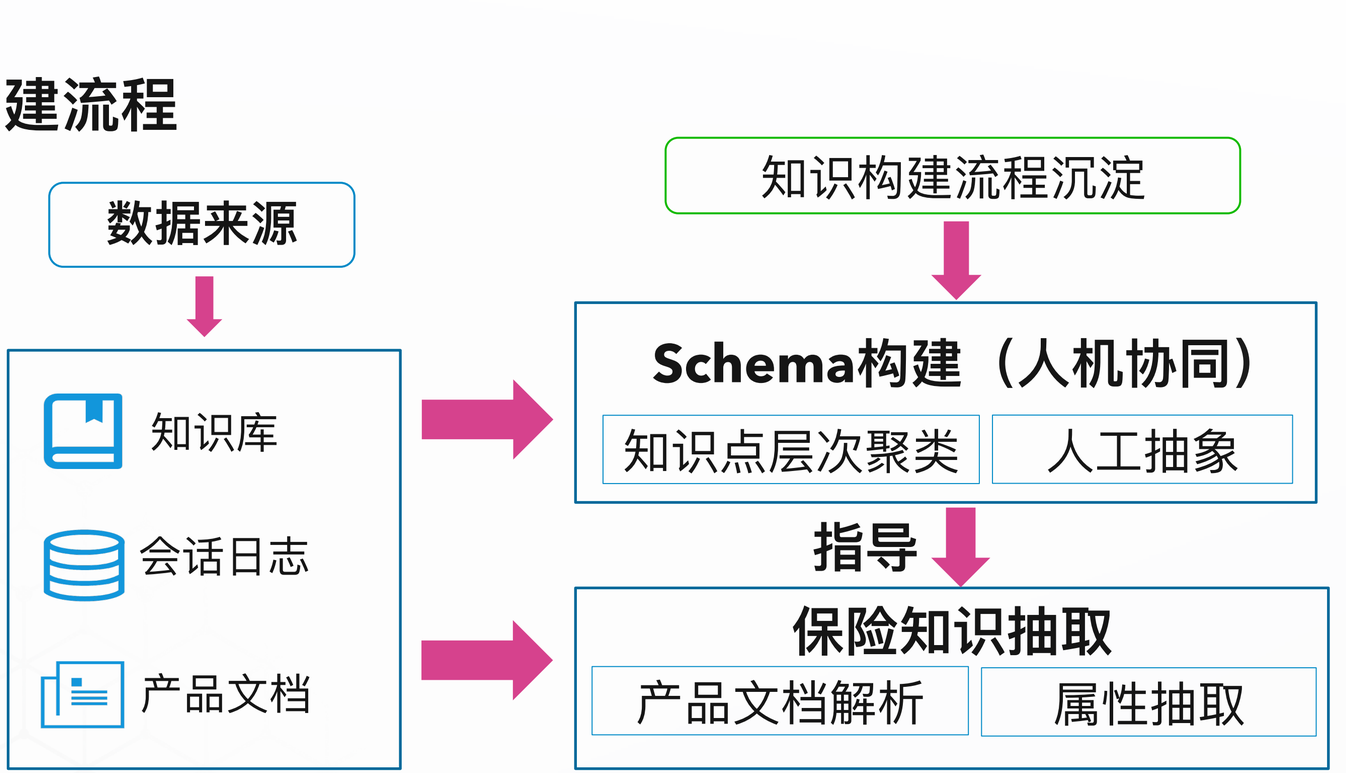
三元组结构：Class-Property-Value （CPV），其中Value分为2类：

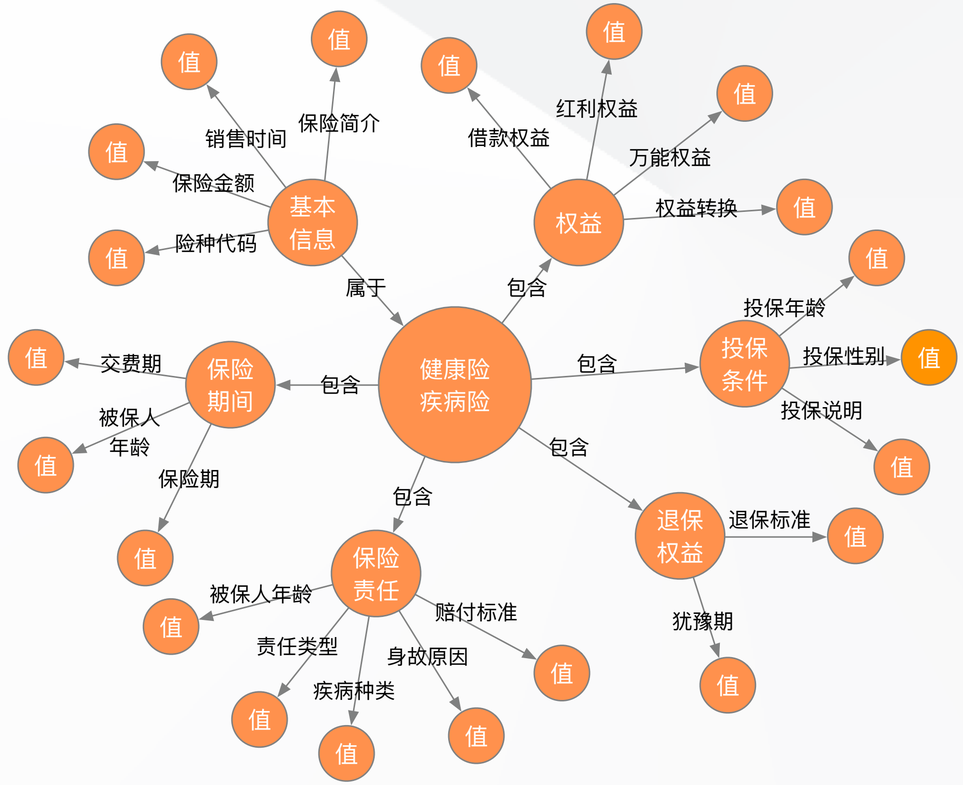
- 单一值：使用内置值类型（String、Int等），如保险简介的值类型
- 复合值：采用Compound Value Type（CVT），有条件依赖的值



保险知识图谱构建

构建流程





保险产品知识图谱

- Schema：5个类型和150+属性
- 国寿：1002个实体
- 全国：28450实体

成果

- 图谱：完成保险产品知识图谱构建
- 降本：构建工具为业务节省**60%**人力；国寿单渠道知识点缩容**42%**
- 标准：经业务确认，图谱Schema适用全国**2.8万**保险产品条款

基于文档的知识抽取

挑战：

十六、咨询途径

电话咨询：0572-5022949

十七、监督投诉渠道

电话投诉：0572-5123449 或 12345 投诉热线

网上投诉：http://zxts.zjzwfw.gov.cn

十八、办公地址和时间

受理地点：安吉县行政服务中心（湖州市安吉县昌硕街道天荒坪南路 62 号安吉商会大厦

浙江政务办事指南文档

规则模板无法推广

咨询监督

咨询方式

咨询电话：020-87933393
咨询窗口地址：从化区城郊街河滨北路128号一楼公安业务大厅
微信号：广州从化公安
信函地址：从化区从化区城郊街河滨北路128号公安业务大厅

监督投诉方式

投诉电话：020-12345
投诉窗口地址：广州市公安局从化区分局从城大道233号监督室
投诉网址：http://wsbs.gz.gov.cn/gz/hotline/
信函地址：广州市公安局从化区分局从城大道233号监督室

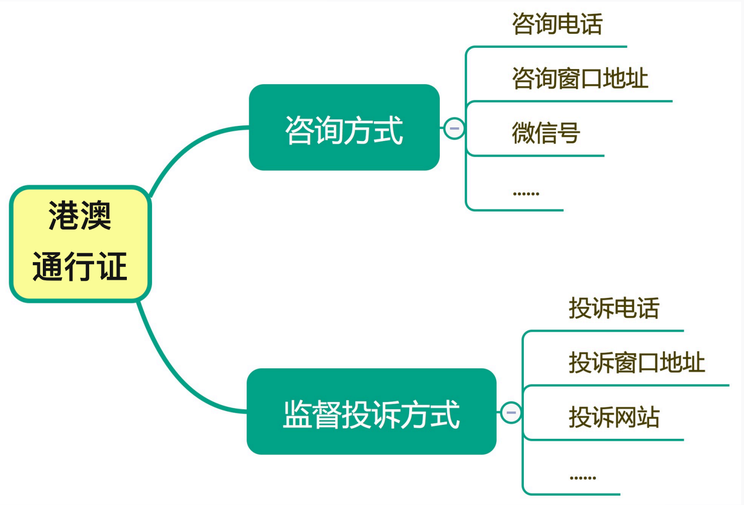
广东政务办事指南文档

方案：

文档解析
1.文档内容提取
2.保留结构、字体信息

小样本SPO抽取模型

融入文本结构信息的预训练语言模型



知识+模型沉淀，降低冷启动成本

行业/场景沉淀阶段

企业客户使用阶段

政务图谱

保险图谱

税务图谱

众包标注

带标语料

直接复用

行业Schema

实体实例

知识增强

行业KBQA模型

3天工作量

新增少量属性

录入属性值(答案)

1天工作量

模型finetune

平台：80%工作量

Baseline系统出厂

客户：20%工作量

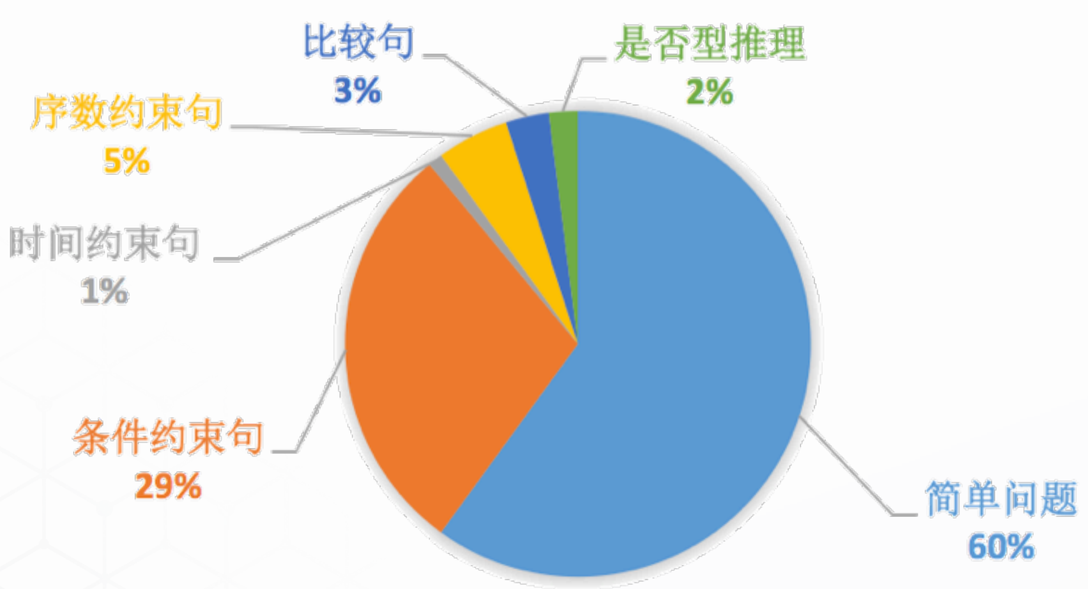
知识图谱问答(KBQA)

03

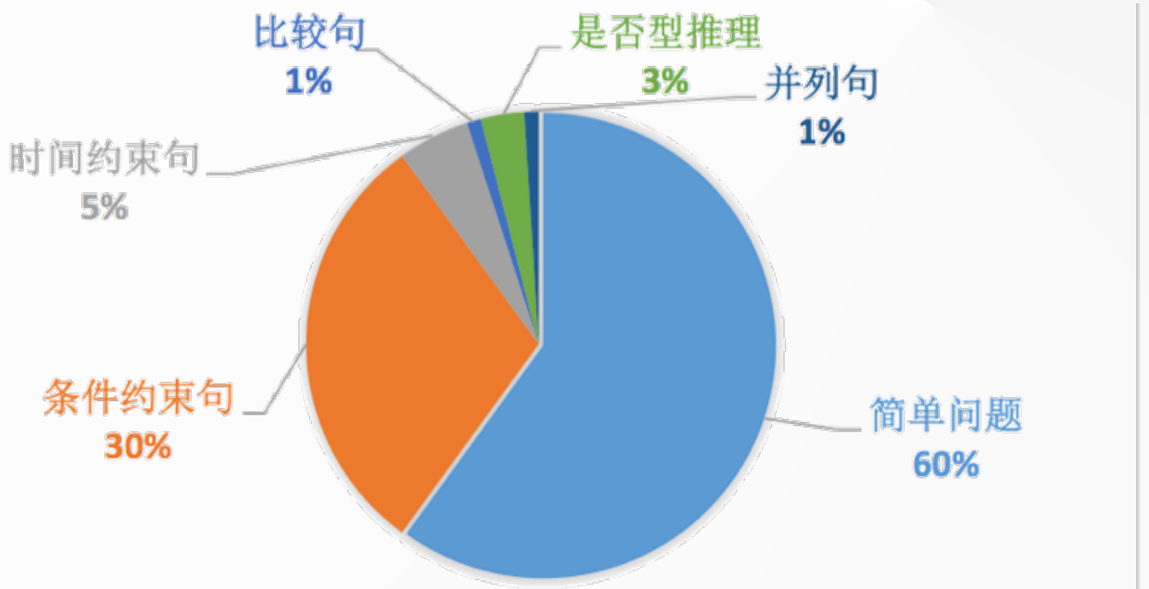
云小蜜业务中的问句类型

	问题分类	例句
单跳问题	简单问题句	1. 为什么不能办理欢享包 2. 税金延期缴纳
多跳问题	多跳句	1. 非诚勿扰在海南拍摄地的酒店的名字
约束问题	条件约束句	1. **售房**的个人所得税 2. **纳税人所得为外币**的怎么办理个人所得税的退税和补税? 3. 原先扣减个税费使用的是**房贷**，现在可以改成**房租**可以扣吗
	时间约束句	1. **2019年后**发放的年终奖个税怎么计算? 2. 怎么查询不到**2019**的个税
推理问题	序数约束句	1. 价格**最低**的套餐多少钱 2. 意外险**最多**能赔多少保险金
	比较句	1. **增值税普通发票**和**增值税专用发票**有什么区别 2. **企业所得税**和**个人所得税**有什么区别
	是否型问句	1. 考核奖要缴纳个税吗 2. 我今年38岁了，可以投保重疾险吗
	并列句	1.少儿意外保险最多能赔多少钱?最高保额是多少? 2.平安保险我想退保,.要带些什么证件,和办理的手续的位置在哪里?

复杂句占比达40%

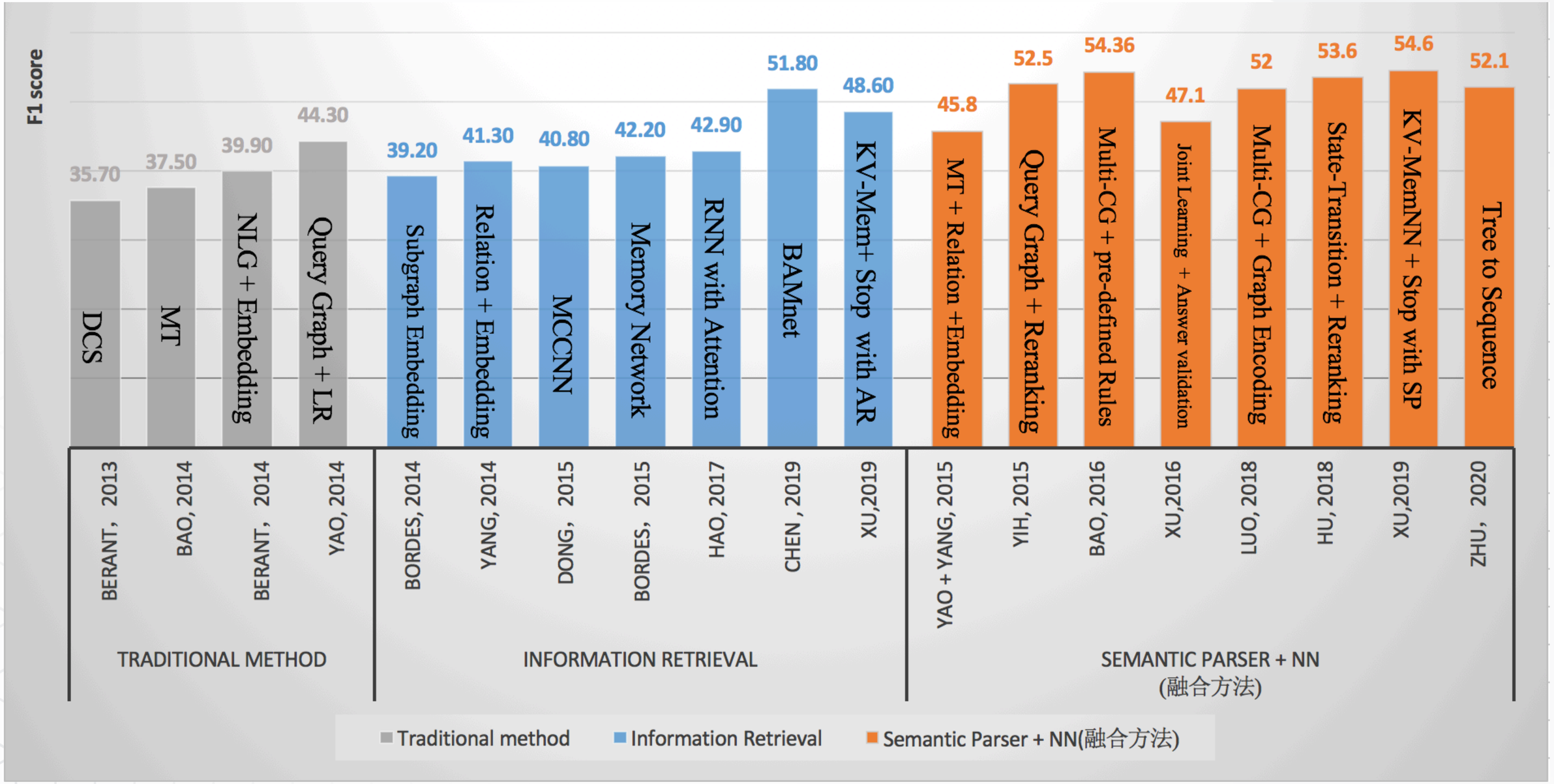


电信运营商场景



税务场景

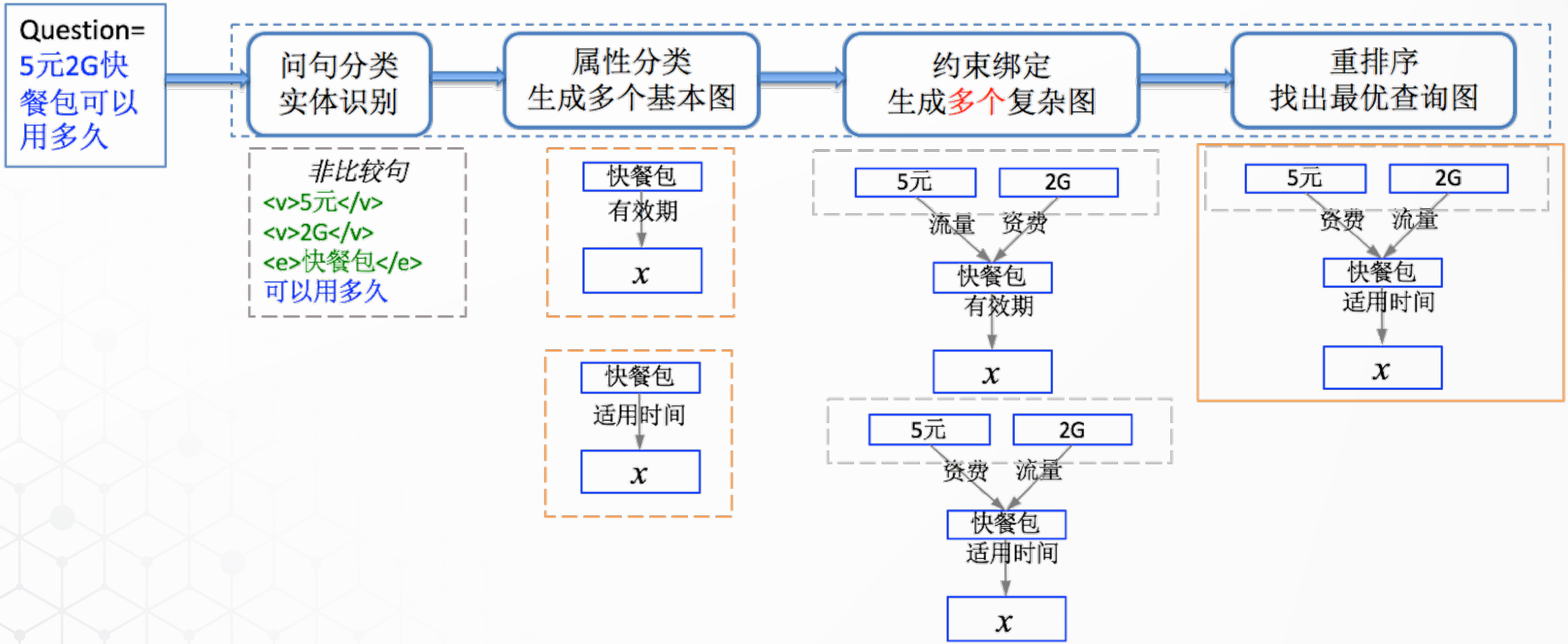
WebQuestions Benchmark



KBQA算法1.0：MultiCG Pipeline

MultiCG方案：

- 双11大促无人机器人，服务商家客户
- 中移动在线IM：湖北、福建项目



KBQA算法1.0存在的不足

问题&不足：

- 约束识别泛化能力弱，依赖手工规则
- 用户Query中省略实体现象
  - 子女教育扣除标准 (省略个税实体)

解法：

Hierarchical KB-Attention Model (HKBAM)

- 不具备消歧能力
  - 你们有没有新出的5G套餐
- 依存分析能力依靠规则
  - 有没有大于5元小于10元的套餐

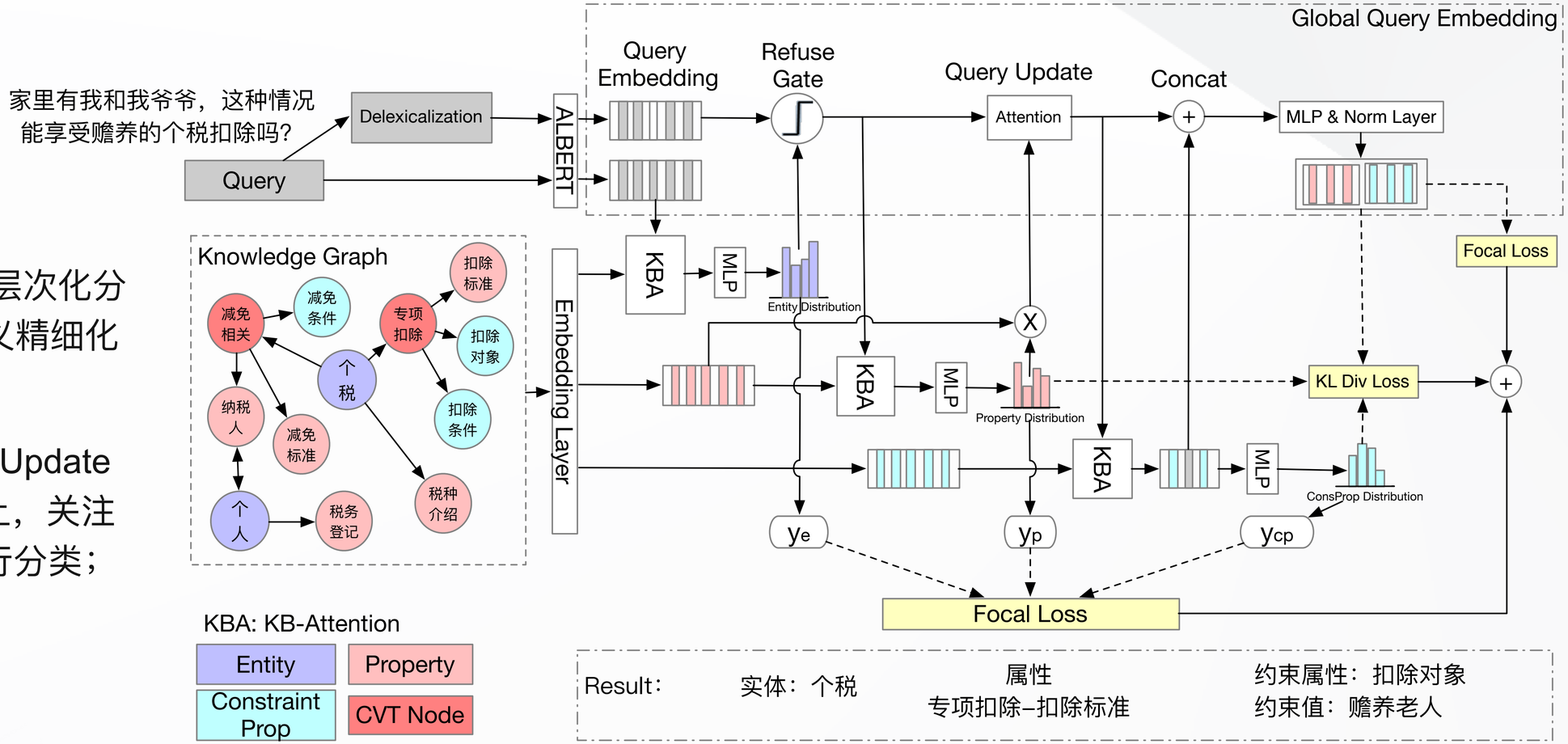
with KAMR parser

HKBAM：Hierarchical KB Attention Model

方案要点：

1.利用KG信息对Query进行层次化分步(multi-hop)解析，实现语义精细化理解；

2.引入Refuse Gate + QueryUpdate机制，使得模型在不同hop上，关注Query中不同词的信息来进行分类；



模型	预训练模型	Acc	提升
MultiCG		46.31%	
HKBAM原生模型		84.60%	
HKBAM原生模型	政务StructBert	86.15%	1.55%
HKBAM + Refuse Gate + QueryUpdate	政务StructBert	**88.70%**	**2.55%**

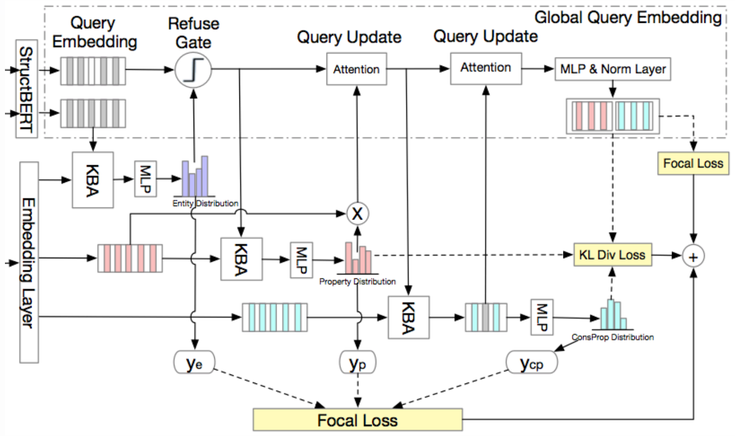
KBQA算法2.0：泛化与消歧

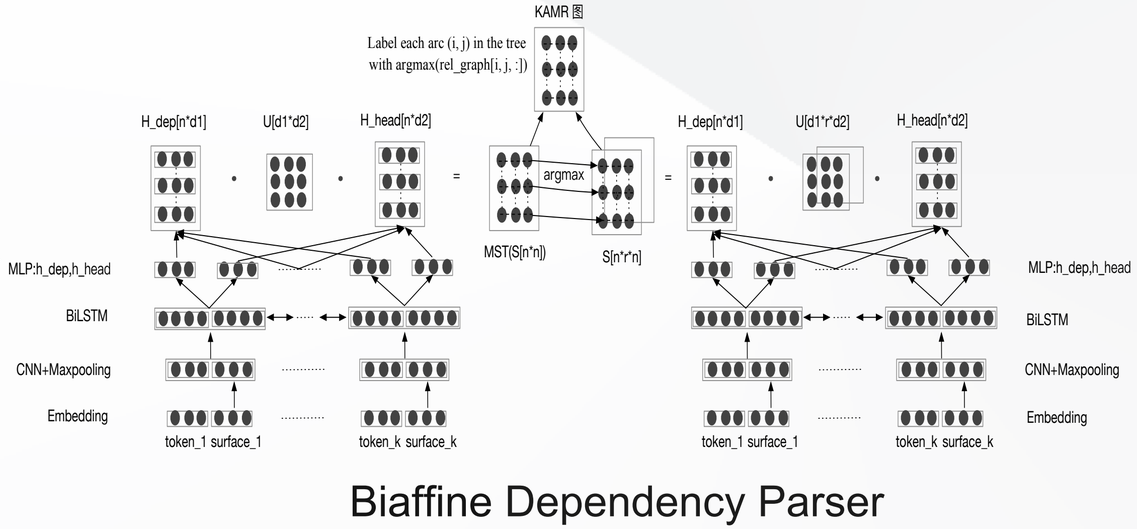
实体识别

规则匹配

BILSTM–CRF

HKBAM





Biaffine Dependency Parser

移动4G的，大流量38元套餐有么

实体识别 → 属性识别 → 约束识别 → Semantic Parser → KAMR TO MultCG → 答案查询

主体套餐

大流量套餐

业务简介

Null, 套餐, 介绍

开通条件

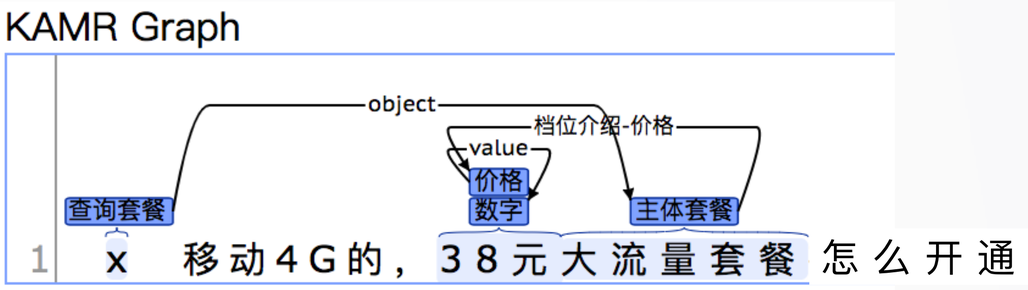
开通, 套餐, 祈使

开通方式

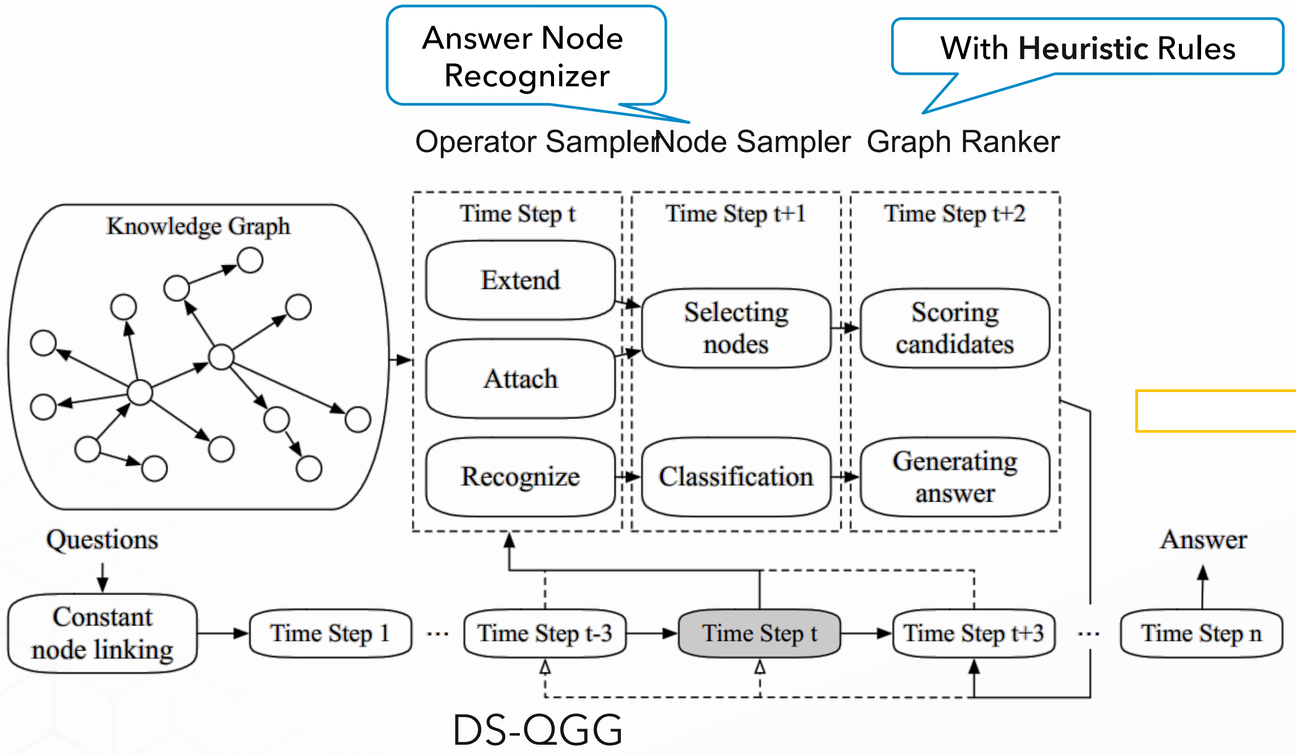
开通, 套餐, 方式

ID	Content	Surface_str	Score
1	开通方式	&	1
2	x	&	1
3	Other	移动	1
4	流量	4G	0.33
5	数字	4G	0.33
6	Other	的，	1
7	价格	38元	1
8	数字	38元	1
9	主体套餐	大流量套餐	0.98
10	Other	怎么开通	1

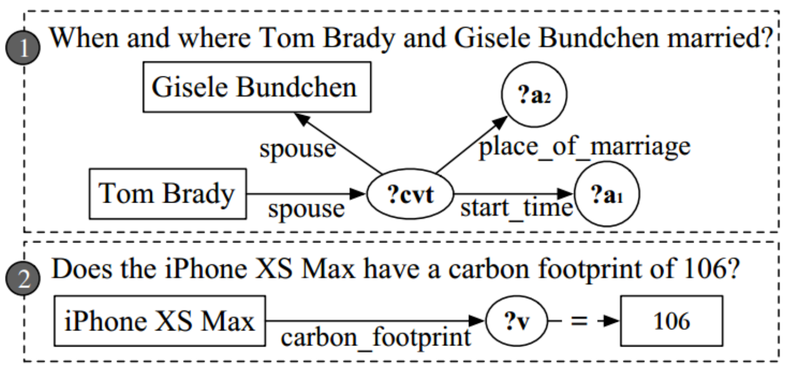
KAMR Graph



KBQA算法3.0：统一多跳和条件推理



- 拓展查询图结构，增强表示能力
  - 单一答案节点限制 → 答案节点可任意选择
  - 理解多意图句和是否类问句
- 强化学习训练加速
  - 启发式规则提供先验
  - 课程学习由易到难逐步优化参数



成果

- ComplexWebQuestions 超越Google Pullnet **4.1%**，官方LeaderBoard**第一名**
- LcQuAD 2.0 效果超越ACL 2020 SOTA方法
- 具备多跳和条件推理的问答能力

Table 3: Performance (%) on WebQSP and CWQ.

Method	WebQSP Avg F_1	CWQ Hits@1/Avg F_1
STAGG [46] (2015)	66.8	-
KV-MemNN [33] (2016)	38.6	-
HR-BiLSTM [48] (2017)	62.3	33.3/31.2
GRAFT-Net [38] (2018)	62.8	30.1/26.0
KBQA-GST [26] (2019)	67.9	39.3/36.5
TEXTRAY [9] (2019)	60.3	40.8/33.9
PullNet [37] (2019)	68.1	45.9/-
QGG [25] (2020)	74.0	44.1/40.4
DS-QGG	**79.6**	**50.0/48.3**
w/o rules	79.1	45.4/41.7

Table 4: Performance (%) on the test set of LCQ.

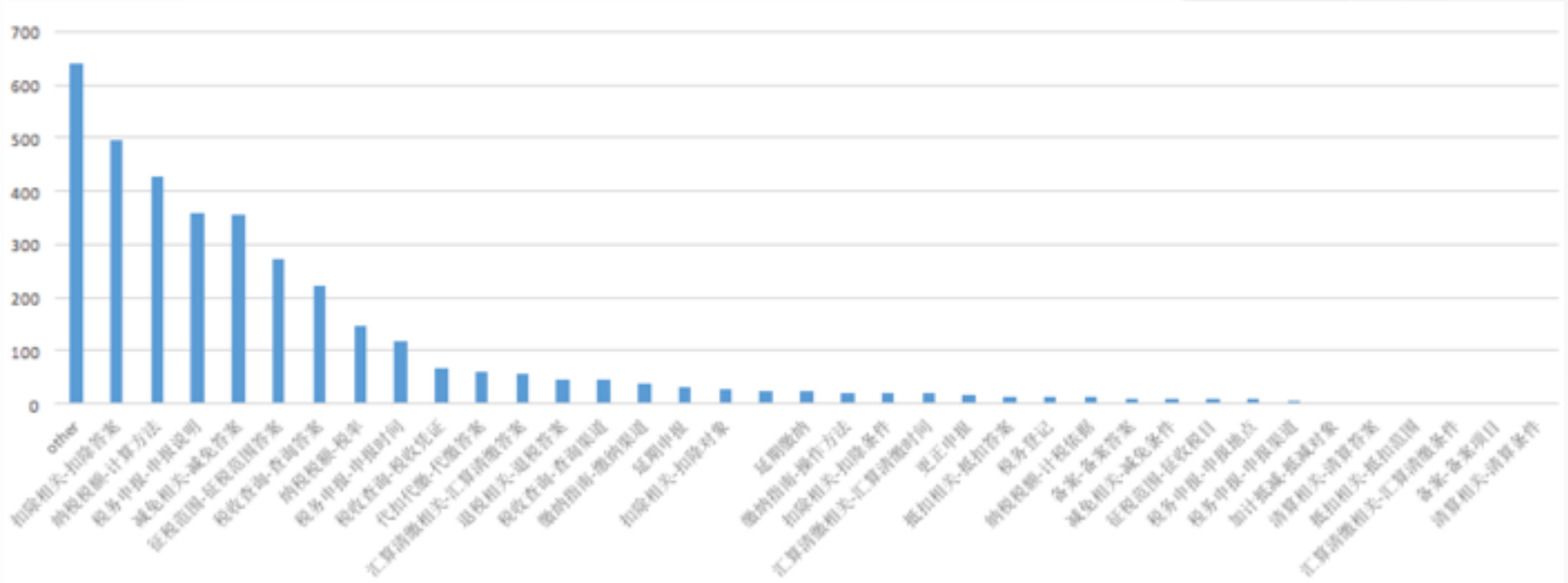
Method	Precision	Recall	Avg F_1
QGG	22.0	39.8	28.3
DS-QGG	**23.5**	**44.0**	**30.6**
w/o rules	23.1	42.8	30.0

顺滑的多轮问答体验：

- 实体澄清与上下文继承
- 精准理解：条件约束句
- 精准理解：是否型问句

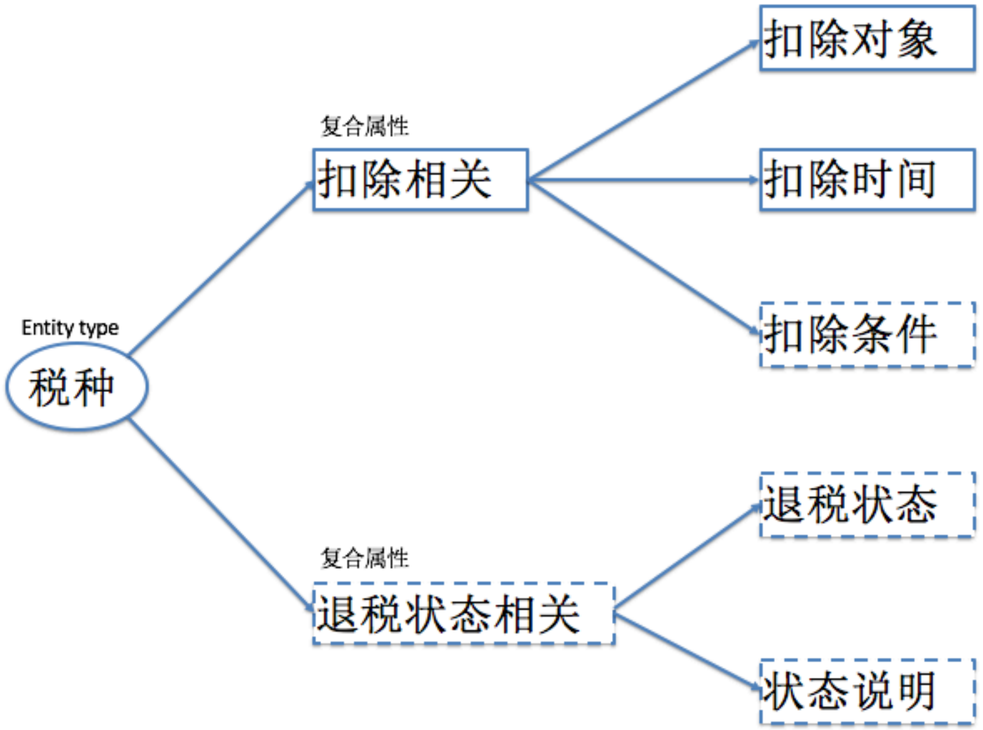
运营成本高，效果难以持续提升

痛点1：
新增属性是常态，很多长尾Query需要识别，但样本收集困难



痛点2：
有监督模型训练链路长，使用成本高，且重训后效果不一定正向

以税务场景为例



- Query1：子女教育专项附加扣除需要满足什么条件?
- Unseen Property：扣除条件

- Query2：4.10就提交了退税申请，为什么一直在审核，这是什么意思?
- Unseen Property：状态说明
- Unseen Constraint：退税状态=“退税审核中”

技术方案：基于KGE的动态自适应

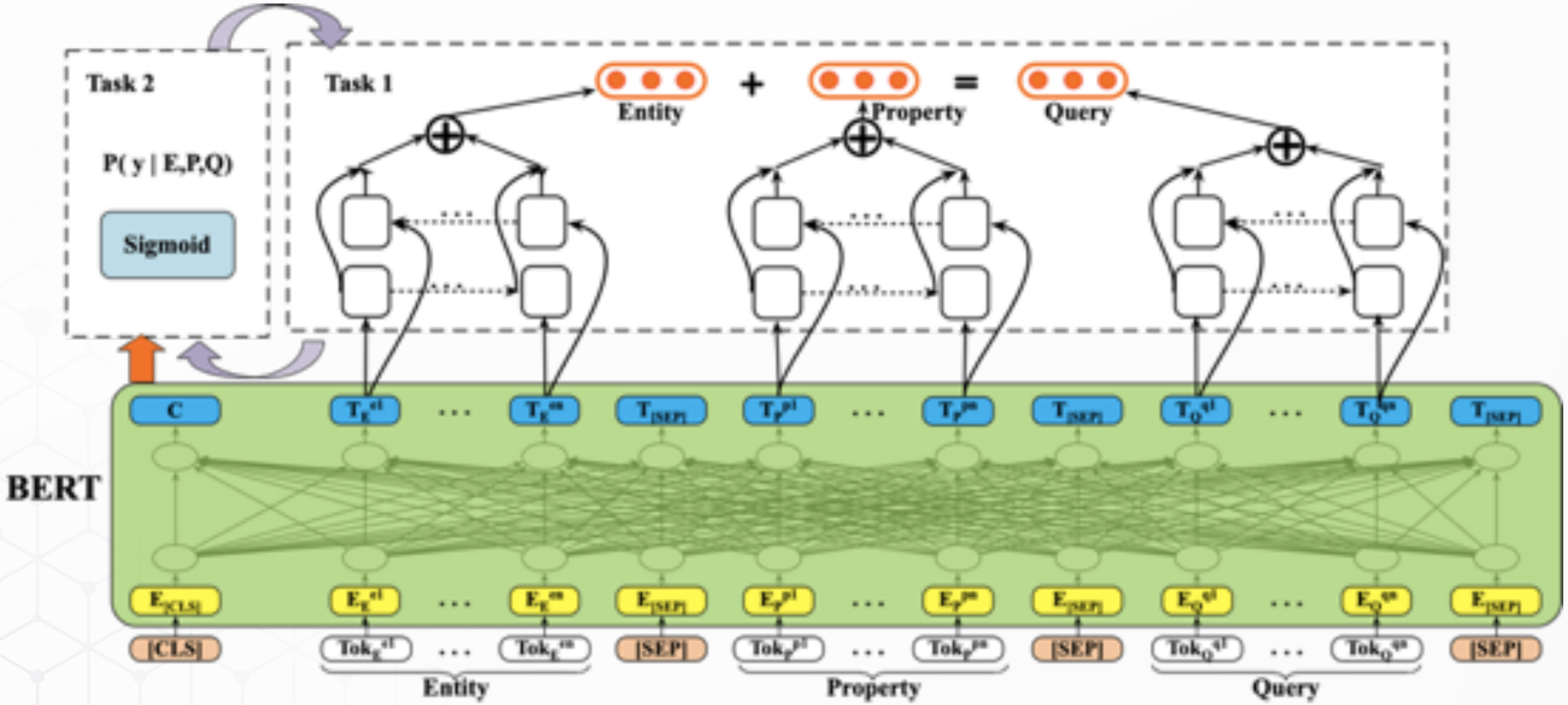
方案：

1.引入KGE任务辅助提升unseen property KBQA效果（利用KGE任务finetune底层BERT）

2.Head + Relation = Tail --> Head + Property = Query

效果：

1.新增属性，无需重训立即生效；在少样本(5条)情形下，属性识别准确率达75%+



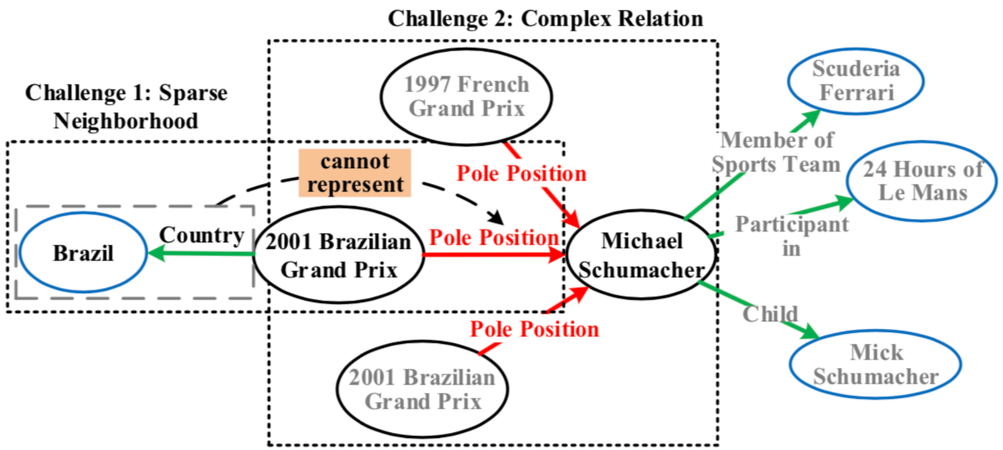
稀疏图谱表示学习：小样本KGE

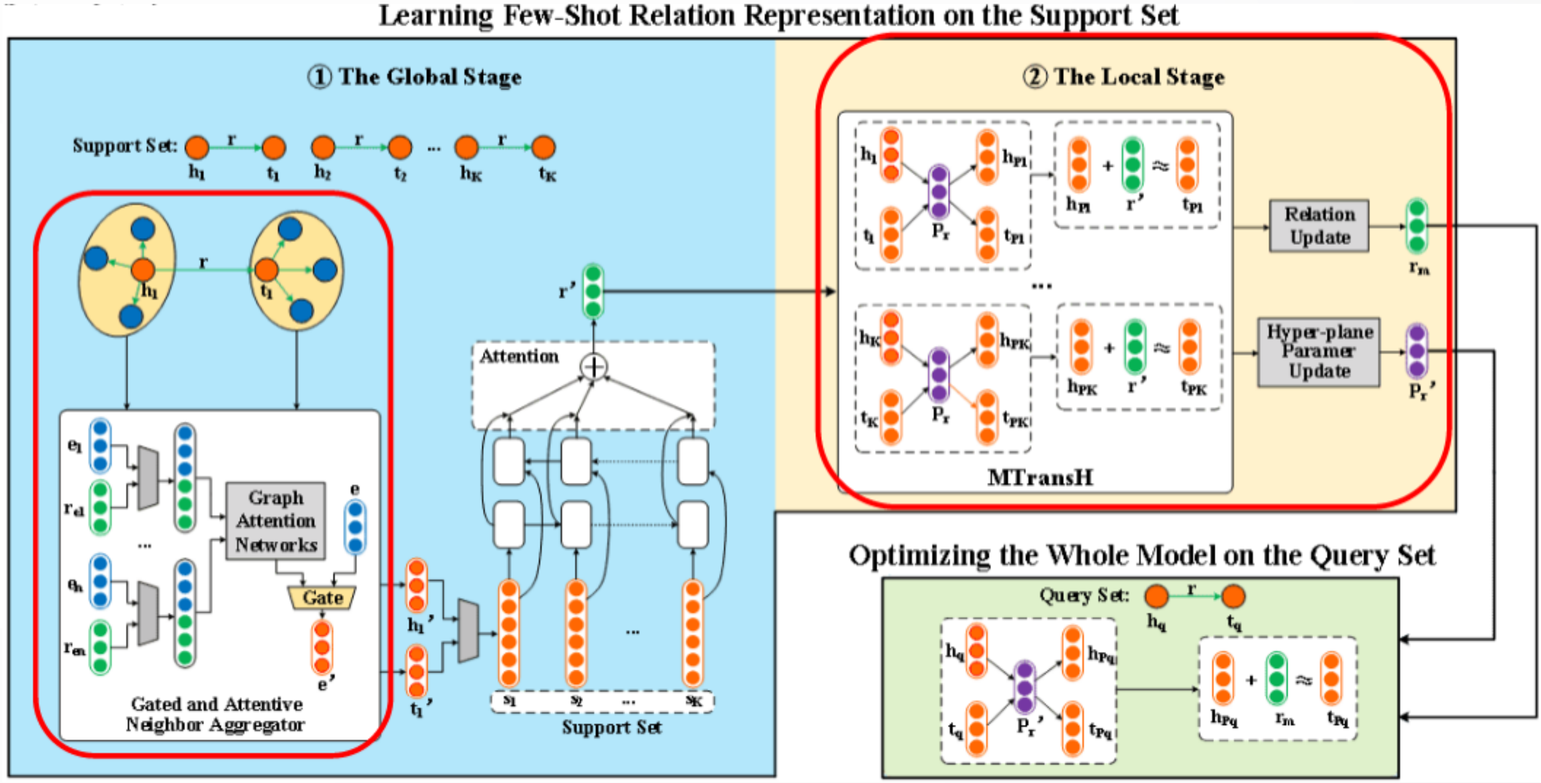
挑战：

- 面临图谱稀疏问题
- 小样本下复杂关系难以表示

方案：

- 门控加注意力的邻居聚合器：滤除噪声邻域信息
- Meta Based TransH：小样本情况下学习Relation的超平面参数





总结与展望

04

行业KBQA面临的挑战与解法

挑战：	冷启动成本高	复杂句语义理解难	运营成本高
解法：	➢ 知识图谱构建链路 ➢ 知识+模型沉淀	➢ 提升模型泛化与消歧能力 ➢ 统一多跳和条件推理 ➢ 理解多意图句与是否型问句	➢ 图谱动态自适应能力 ➢ 配套干预、回流机制

行业知识图谱应用的实践心得

1 图谱概念简化，降低认知门槛

1. 图谱概念对一般客户而言，有比较高的理解和认知成本
2. 分步骤、带案例的讲解，效果会比较好

2 明确业务范围，合理控制预期

1. KBQA适合以实体为中心，问句条件依赖有组合性的场景
2. 依据问句占比，明确KBQA覆盖业务范围

3 利用工具，提升客户自运营能力

1. 半自动化工具辅助构建和知识抽取，帮客户提效
2. 建设KBQA的可干预能力，支持快速fix线上badcase

未来展望

更强大的知识赋能

隐式赋能：融入行业知识的语言模型

显示赋能：基于KGE的KBQA

更轻量的业务运营

实体、属性自动发现（热点事件）

图谱自动融合与补全

更高效的知识生产

建设文档结构化知识平台，

提升图谱知识生产效率

更丰富的应用场景：推荐、导购

基于结构化知识的话术生成，

用于产品推荐、导购等场景

感谢聆听！