小米小爱智能问答探索与实践

代文

小米AI Lab

目录

介绍—小爱同学

手机
+
AIoT

入口

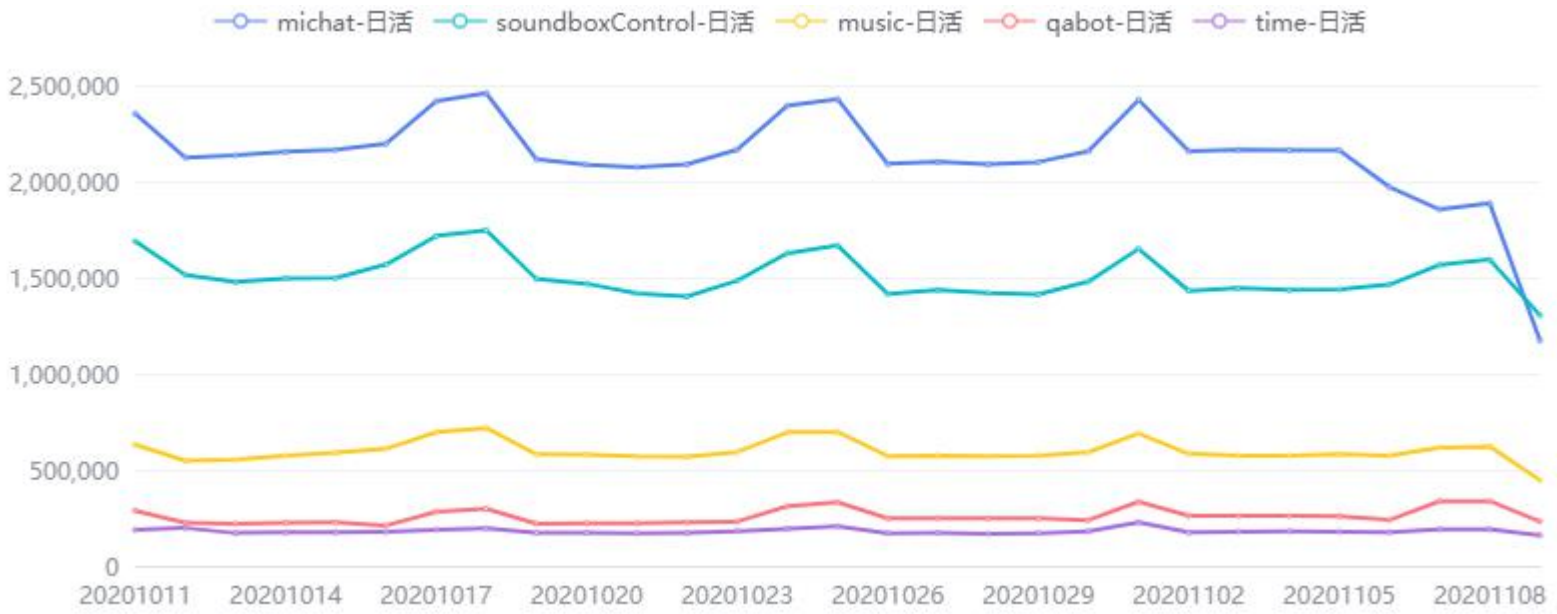
介绍—应用场景

智能场景

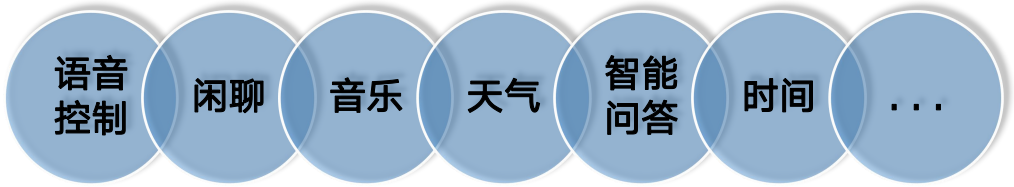
35个设备，751款智能设备*，全方位覆盖智能家庭生活的方方面面

*数据截至2019年11月

介绍—应用场景



小爱同学



智能问答：

- 有客观事实性答案
- 满足用户信息查询类需求

智能问答

天空为什么是蓝色的

为你找到以下信息

"天空为什么是蓝色的"

天空之所以是蓝颜色的，是因为太阳光是七种光组成。阳光进入大气时，波长较长的色光，如红光，透射力大，能透过大气射向地面；而波长短的紫、蓝、青色光，碰到大气分子、冰晶、水滴等时，就很容易发生散射现象。被散射了的紫、蓝、青色光布满天空，但是大气层中紫色几乎看不到，因为紫色光在被散射的同时大部分也被吸收，而且人类的眼睛对紫色也并不敏感，就使天空呈现出一片蔚蓝了。

小爱这里的天气：今夜到明天上午有点想你，预计下午转为持续想你，哈哈，你那的天气呢？

本次回答满意吗？ 赞 踩

搜索 玫瑰花的花语 春天里的

诗词歌赋

静夜思是哪位诗人写的

作者：李白

静夜思

【唐】李白

00:01 | 00:37

chuáng qián míng yuè guāng
床 前 明 月 光 ，

yí shì dì shàng shuāng
疑 是 地 上 霜 。

本次回答满意吗？ 赞 踩

搜索 蝶恋花 小荷才露尖尖

人物·视频

姚明的老婆身高是多少

190 cm，姚明的妻子是叶莉，叶莉的高度是190 cm

暂无数据

190 cm

简介:姚明的妻子是叶莉，叶莉的高度是190 cm

精打细算靠小爱，准确无误严谨派。
计算、换算小爱都可以！

换个话题 | 感兴趣

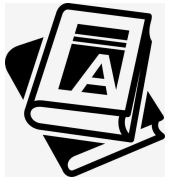
本次回答满意吗？ 赞 踩

目录

基于知识图谱的问答

三元组spo量：近百亿

实体量：数十亿+


书籍


本地生活


字词、古诗


房产


商品


百科


音乐


景点

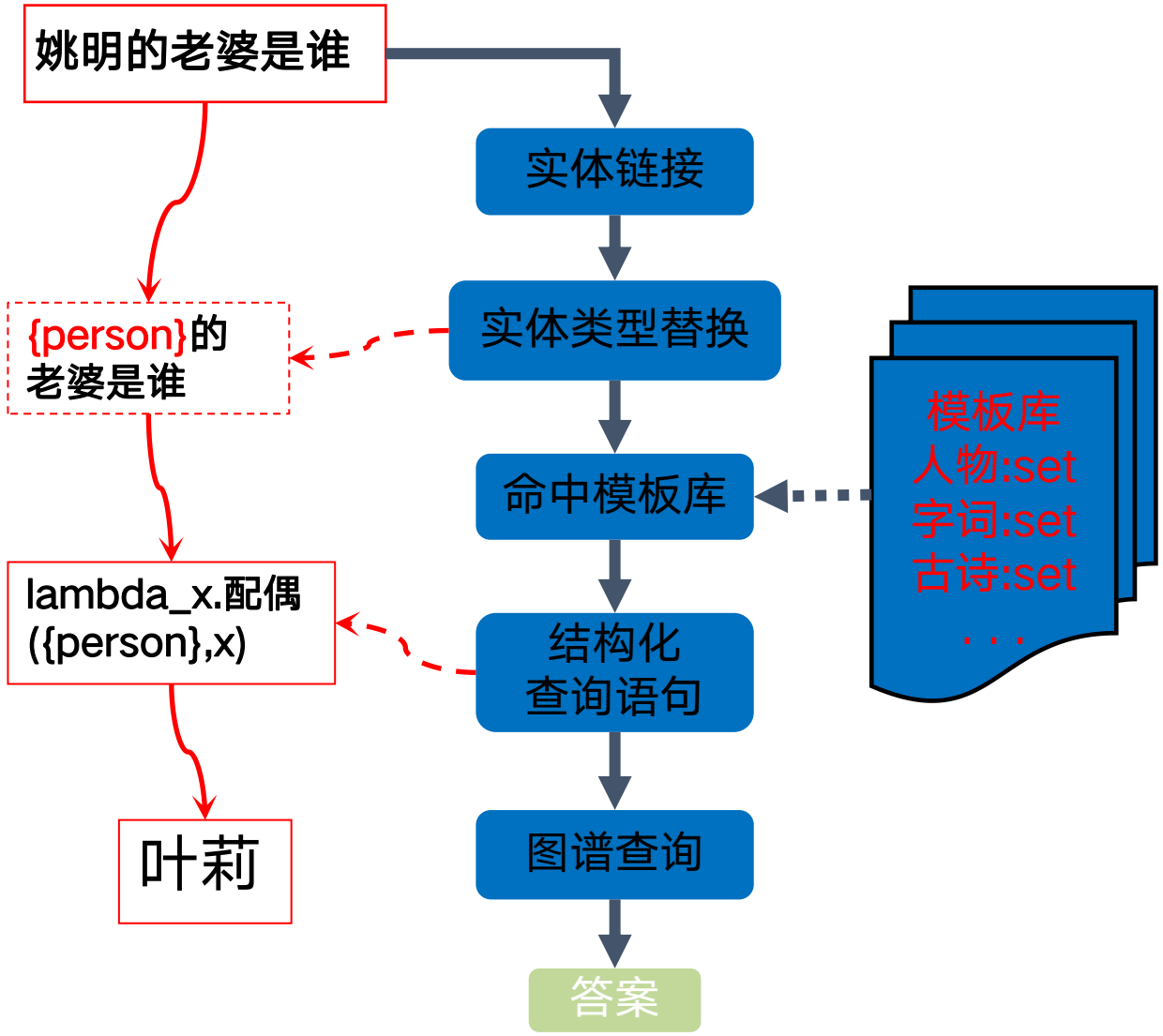

人物、作品

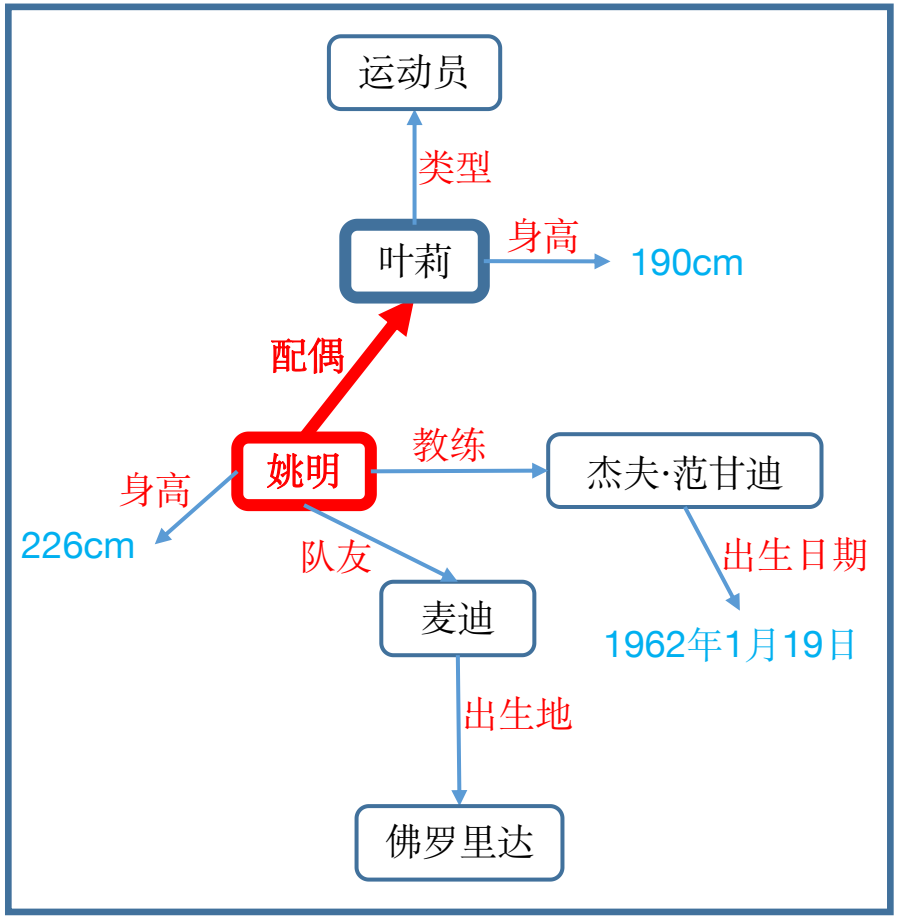
……

基于知识图谱的问答

- 基于模板的方法
- 基于槽填充-意图识别联合模型的方法
- 基于深度学习的方法

基于模板的方法

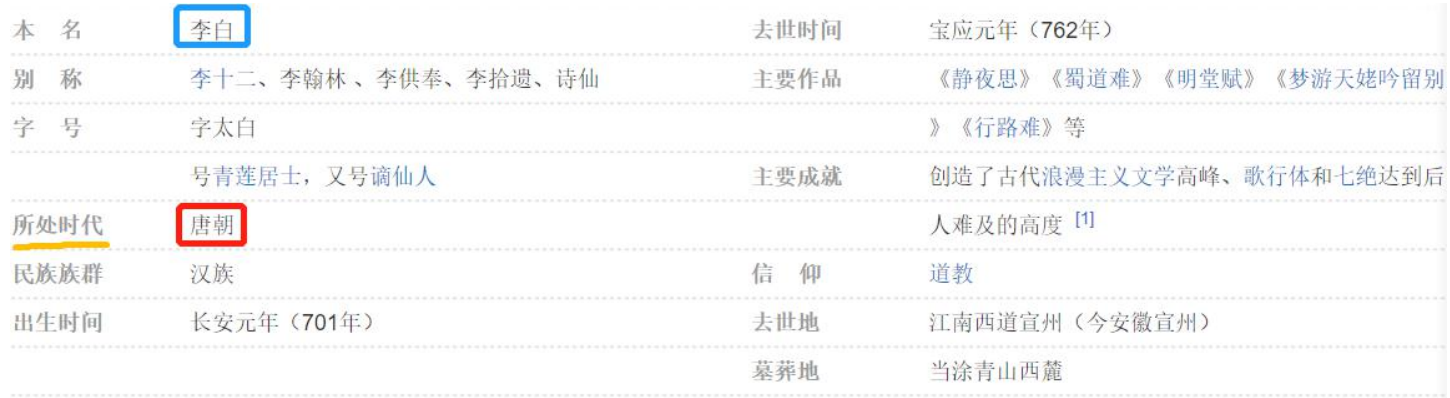




模板的获取是一个困难的问题
- ✓ **从线上高频query扩展补充**
- ✓ **自动化获取**

基于模板的方法-- 模板挖掘方法

结构化词条



本　名	李白	去世时间	宝应元年（762年）
别　称	李十二、李翰林、李供奉、李拾遗、诗仙	主要作品	《静夜思》《蜀道难》《明堂赋》《梦游天姥吟留别
字　号	字太白		》《行路难》等
	号青莲居士，又号谪仙人	主要成就	创造了古代浪漫主义文学高峰、歌行体和七绝达到后
所处时代	唐朝		人难及的高度 [1]
民族族群	汉族	信　仰	道教
出生时间	长安元年（701年）	去世地	江南西道宣州（今安徽宣州）
		墓葬地	当涂青山西麓

李白（实体）-所处时代（属性）-唐朝（属性值）

问答论坛

结构化解析

问法：{person}是什么朝代的？

lambda_x.所处时代({person},x)

一些清洗规则：

1.答案中出现多个属性值的问答对要过滤掉

2.出现频次低的问法要过滤掉

…

基于模板的方法

优点

- 准确率很高：95%
- 线上性能好
- 适合做体验精品化

缺点

- 模板扩充耗时耗力
- 泛化性差，召回有限

作文

菜谱

人物

字词

基于槽填充-意图识别联合模型的方法

目的：提升泛化性，提高召回

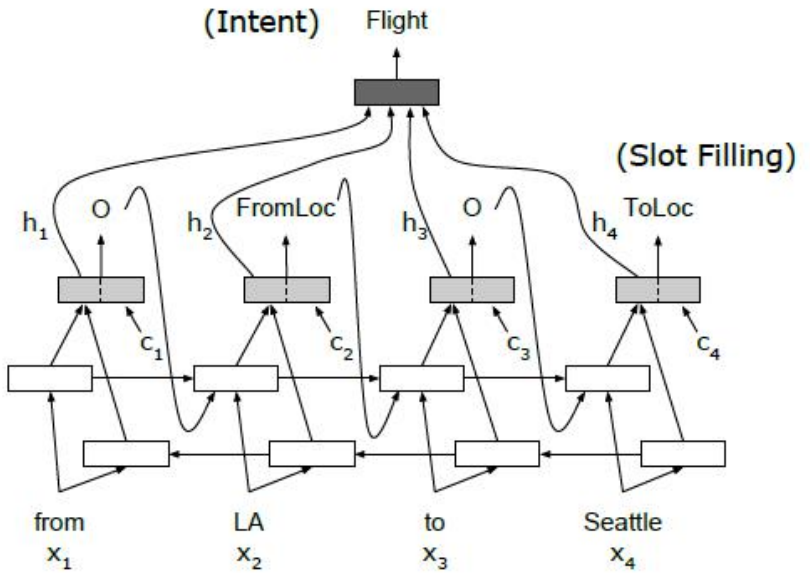
模型：

- 槽填充：通过NER方式提取槽位
- 意图识别：按文本分类方式识别query意图
- 多任务学习：将二者联合学习

策略：

- 引入CRF层，提升槽位抽取准确率
- 根据领域词表对识别槽位进行纠错
- 实现意图、槽位、纠错一步到位

菜谱、古诗垂域：对欠召回badcase扩召约30%，意图准确率87.2%，槽位抽取准确率82.0%

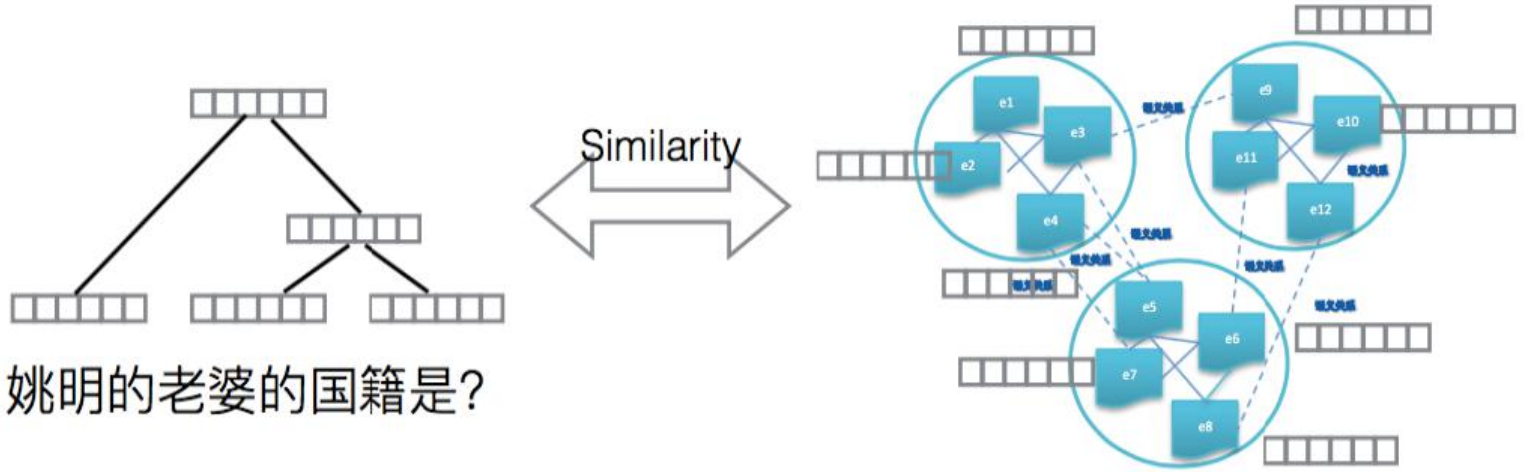


Liu B, Lane I. Attention-based recurrent neural network models for joint intent detection and slot filling[J]. arXiv preprint arXiv:1609.01454, 2016.

手机助手召回率	上线前	上线后
菜谱	80.41% +5%	85.76%
古诗	84.59%	87.79%

基于深度学习的方法

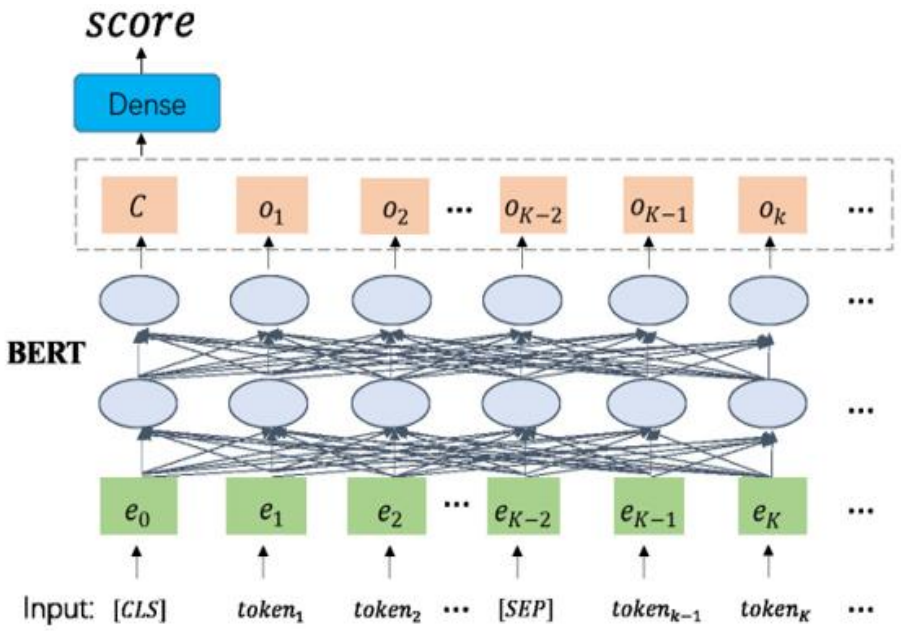
- 实体链接
- 子图检索
  - ➢子图模板
- 子图匹配
  - ➢语义相似度计算
  - ➢Rank排序



基于深度学习的方法—实体链接

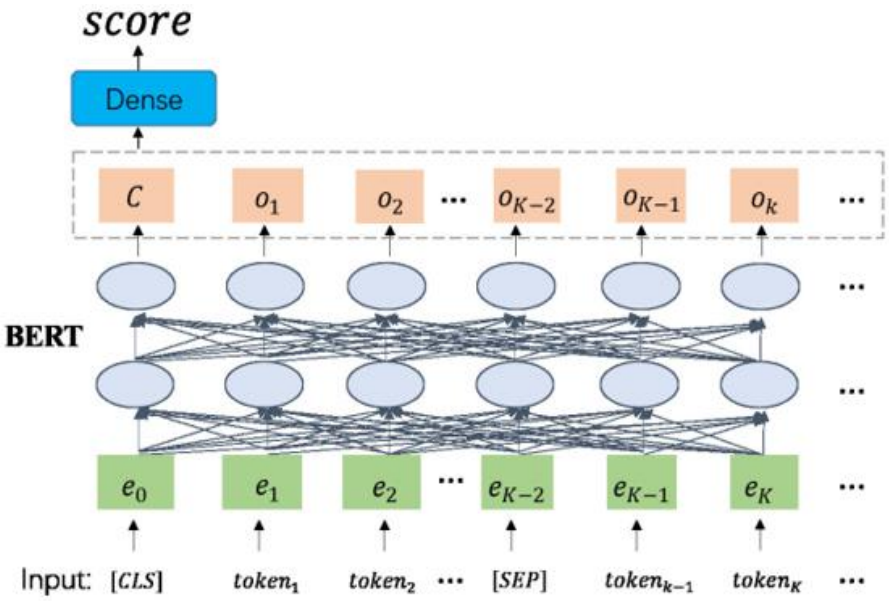
实体链接：

Step 1: Entity Recognition



Text_a: [标记符]龙卷风[标记符]的英文名是什么？ Text_b: None

Step 2: Entity Disambiguation

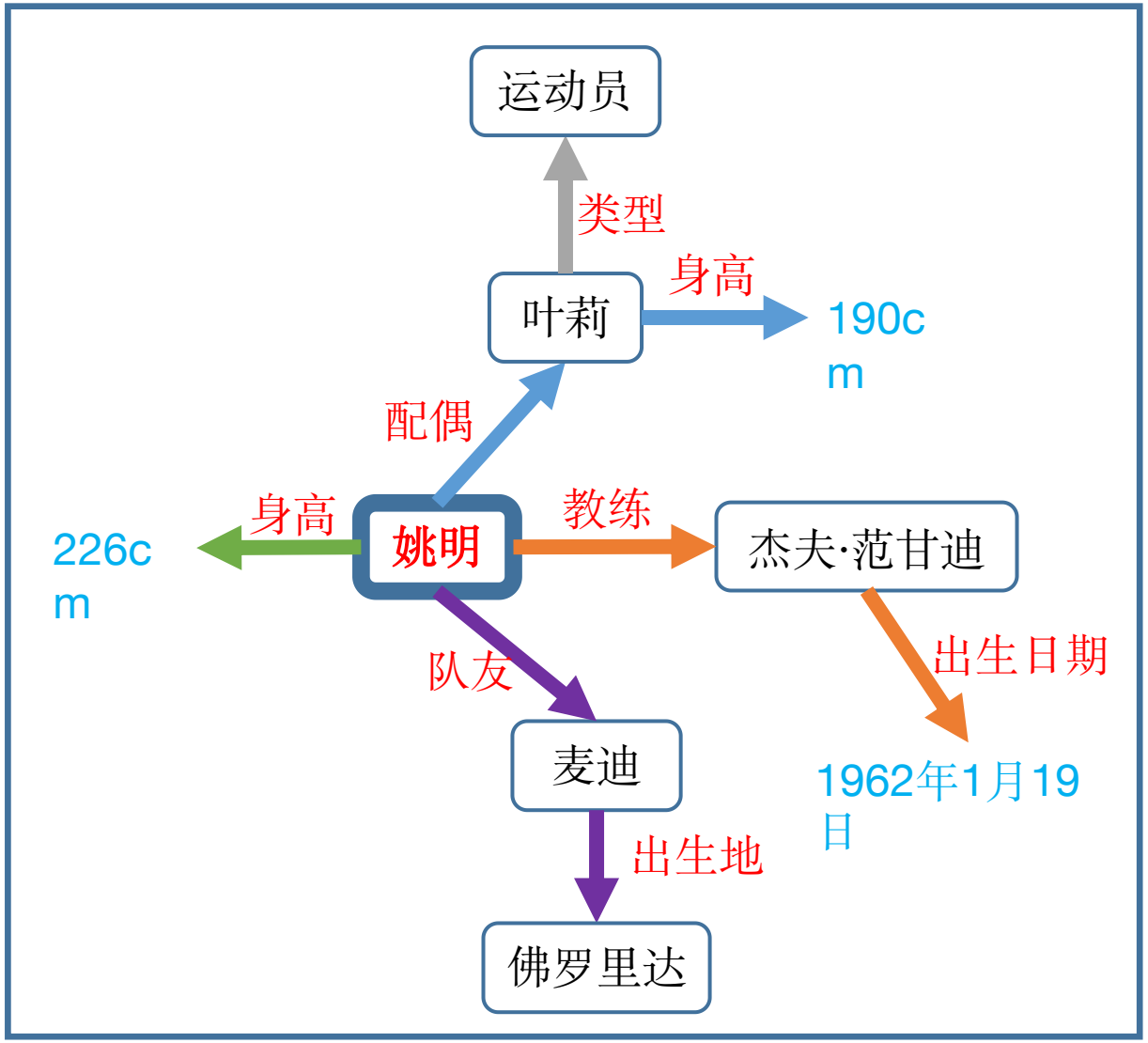


Text_a: [标记符]龙卷风[标记符]的英文名是什么？ Text_b: 一种自然天气现象

基于深度学习的方法—子图检索

子图检索：

- 以单实体/多实体作为起点
- 按照路径模板挖掘候选子图



预定义路径模板

类型	路径模板
单实体单跳	Subject-Predicate-Answer
单实体单跳	Answer-Predicate-Object
单实体两跳	Subject-Predicate-Intermediate-Predicate-Answer
单实体两跳	Intermediate-Predicate-Object-Intermediate-Predicate-Answer
单实体两跳	Intermediate-Predicate-Object-Answer-Predicate-Intermediate
两个实体两跳	Two-Entity-Path
两个实体三跳	Two-Entity-Path-1-hop
三个实体三跳	Three-Entity-Path
三个实体四跳	Three-Entity-Path-1-hop

基于深度学习的方法—子图匹配

姚明的老婆的身高是多少

候选路径

姚明配偶类型^
姚明配偶身高^
姚明身高^
姚明教练出生日期^
姚明队友出生地^
....

匹配流程:

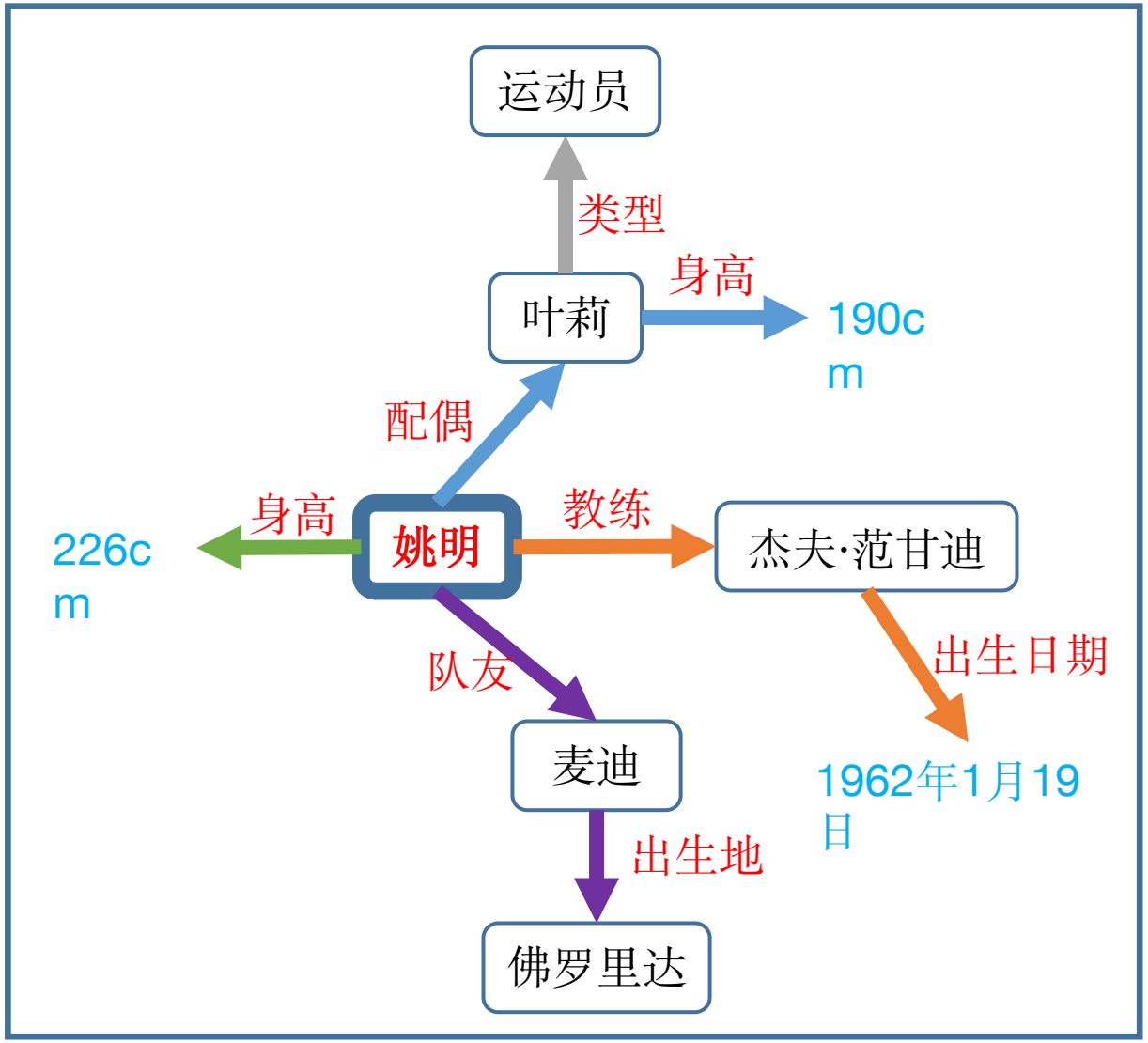
RoBERTa 语义相似度

人工特征

Rank模型

Top 1

在CCKS2020中文知识图谱问答评测中，排名3/430

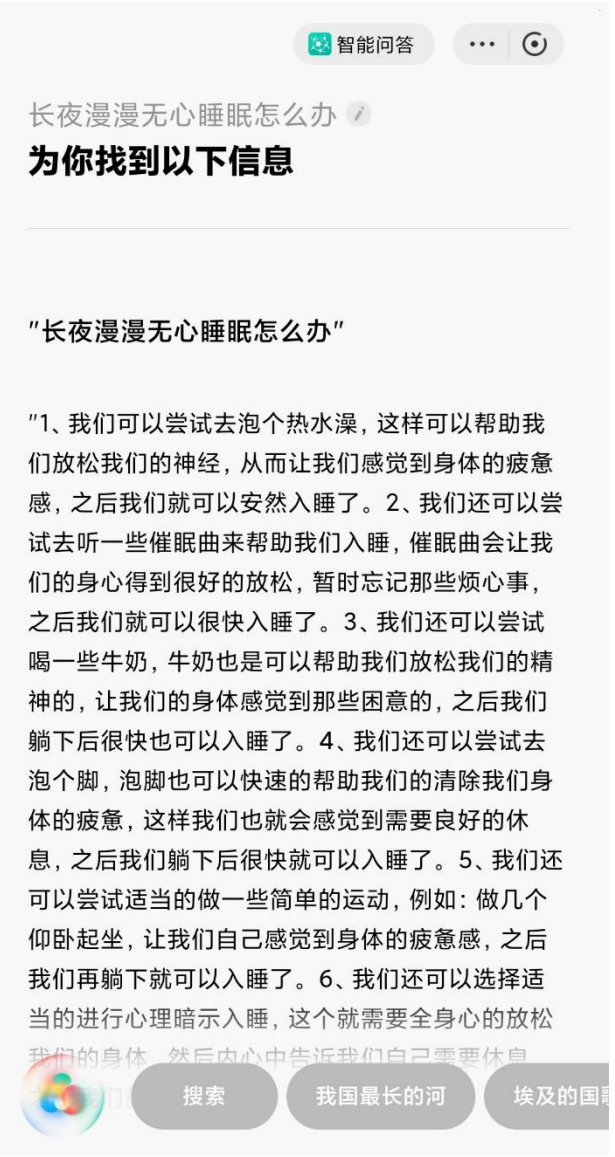


目录

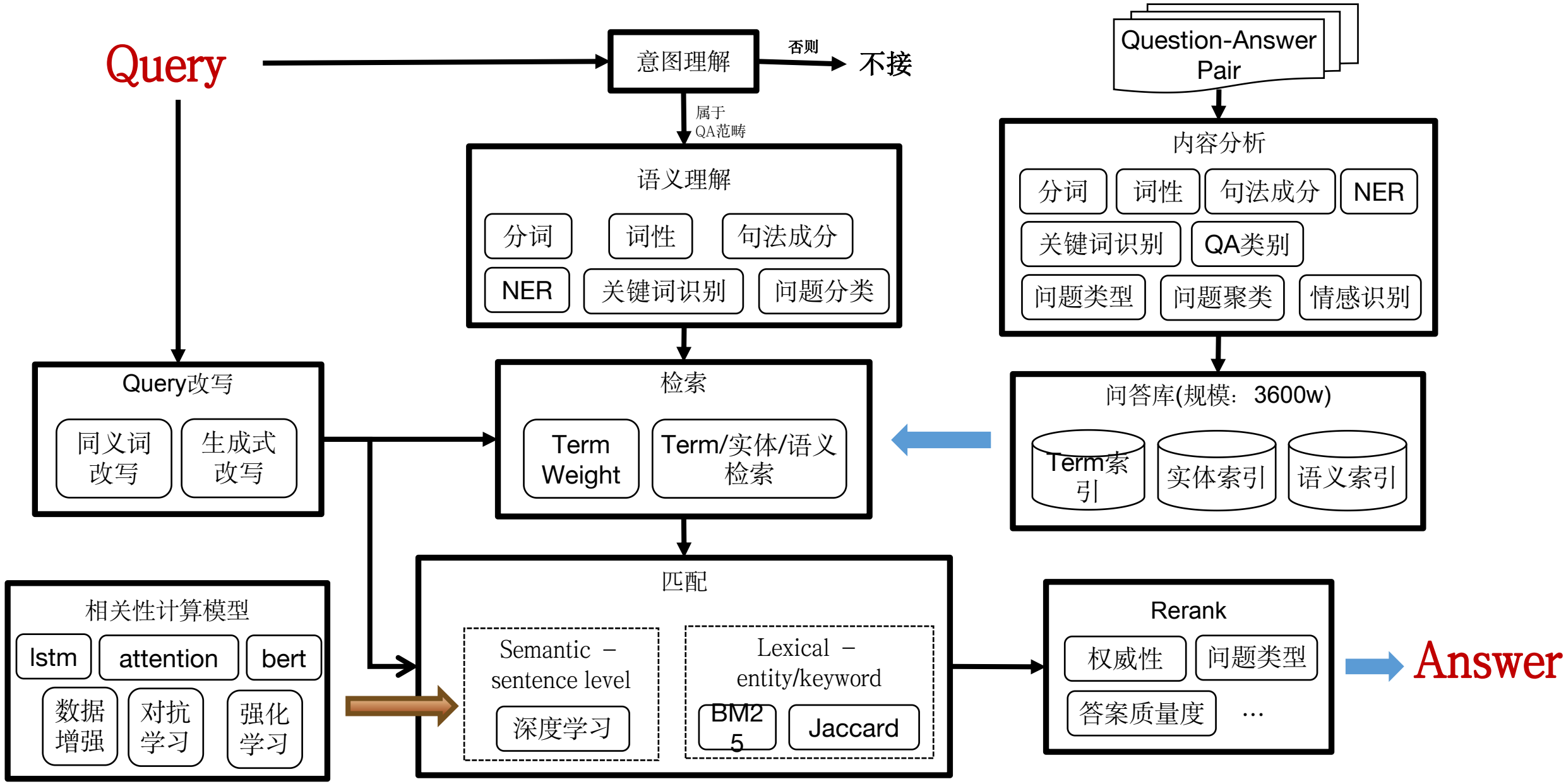
基于检索的FAQ问答

非结构化query:

WHY	HOW	WHETHER
• 天空为什么是蓝色的 • 为什么海水是咸的	• 怎样控制猫咪食量 • 长夜漫漫无心睡眠怎么办	• 青蛙王子是不是安徒生童话 • 柠檬是酸性食物吗



FAQ问答框架


Query
意图理解
否则
不接
属于
QA范畴
语义理解
分词
词性
句法成分
NER
关键词识别
问题分类
Question-Answer Pair
内容分析
分词
词性
句法成分
NER
关键词识别
QA类别
问题类型
问题聚类
情感识别
Query改写
同义词改写
生成式改写
检索
Term Weight
Term/实体/语义检索
问答库(规模：3600w)
Term索引
实体索引
语义索引
匹配
相关性计算模型
lstm
attention
bert
数据增强
对抗学习
强化学习
Semantic – sentence level
深度学习
Lexical – entity/keyword
BM25
Jaccard
Rerank
权威性
问题类型
答案质量度
…
Answer

基于检索的FAQ问答

检索

- term检索
- 实体检索
- 语义检索

匹配

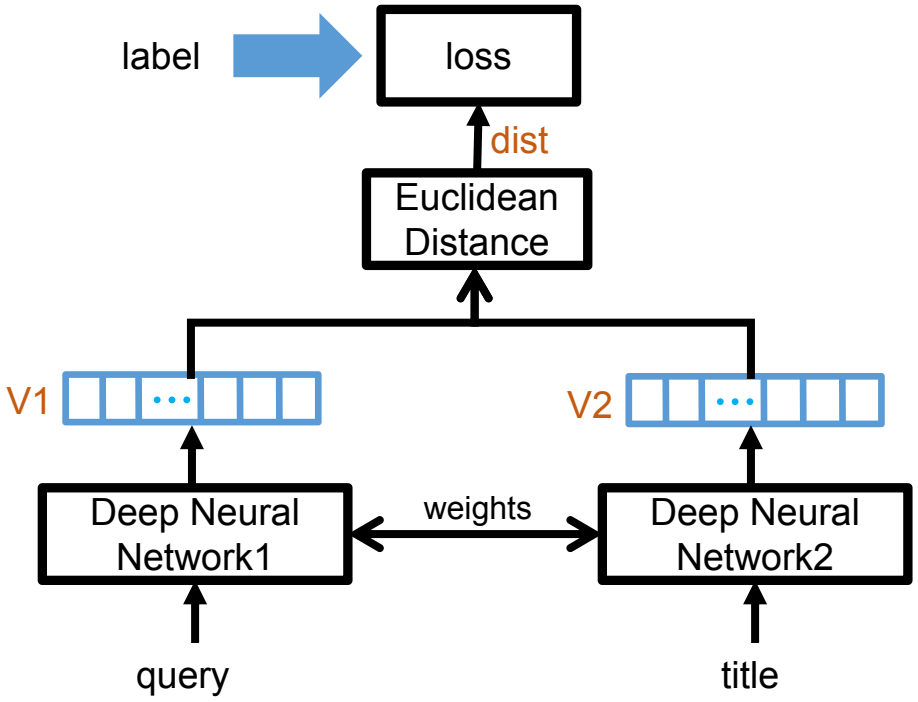
- representation-based vs interaction-based
- 数据增强
- 知识蒸馏

基于检索的FAQ问答--检索

- term检索
- 实体检索
  - query:中国餐馆的主角有谁; doc:中国餐馆在播放抗日神剧
  - 依靠实体链接技术
- 语义检索
  - 为每一个问题计算句子的语义向量
  - 通过乘积量化技术优化线上查询效率

语义检索

- Step 1: 学习得到每个doc的语义向量



$$\text{dist}(V_1, V_2) = \|V_1 - V_2\|^2$$

$$\text{loss} = \sum (y_+ \cdot dist(V_1, V_2) + y_- \cdot L(dist(V_1, V_2)) + \lambda \cdot \text{regularization})$$

$$\text{其中，} L(x) = \begin{cases} t - x, & x \le t \\ t, & x > t \end{cases}$$

语义检索

类型	Pearson相关系数	F1	Inference(cpu)
双向lstm	0.5039	72.63	
双向lstm + attention	0.4947	71.94	
stack lstm	0.4877	71.47	
Transformer-6layers	0.7171	85.68	7ms
Transformer-5layers	0.7129	85.25	7ms
Transformer-4layers	0.7012	84.81	5.5ms
Transformer-3layers	0.7008	84.52	3.5ms

- Step 2：搭建faiss语义检索服务，以docid作为key，语义向量作为value
- Step 3：寻找最近邻，通过docid查询倒排索引，获取doc结果

基于检索的FAQ问答

检索难题：

长尾query

表达冗余。例如“孙子兵法智慧的现代意义”，在这个语境下，“智慧” 是一个无关紧要的词。如果强制去召回“智慧”的话，反而召回不出真正想要的结果

语义鸿沟。比如 “谁发明了新中国”，其中“发明” 这个词较少用来形容国家的建立，使得检索到的结果很少

表达冗余： termweight

孕妇的正常体温是多少度

检索失败

中工国际股什么时间发行的

狗狗的正常体温应该是多少度呢

人体正常体温是多少度范围

什么时间发行的火影忍者

检索失败

中工国际股什么时间发行的

王力宏的最新专集什么时间发行啊

中国福娃邮票的发行时间是什么时候

问答场景query的特点：

- ✓ 一般是一个语义完整的句子
- ✓ term数量比较多

termweight → 召回不佳 → 候选队列没有正确答案 → 回答出错 →  影响用户体验

表达冗余： termweight

- tf-idf
- 点击数据：根据Q=abc中a/b/c三个term在点击结果中的出现次数来计算；为了解决从未出现过的query没有点击数据的问题，把点击细化到ngram的粒度；
- 提取特征训练xgb模型

特征类型	特征名称
静态特征	term词性、长度信息、term数目、位置信息、ner结果、句法依存tag、是否数字、是否英文、是否停用词、是否专名实体、是否重要行业词、idf
动态特征上下文特征	embedding模长、删词差异度、前后词pmi、左右邻熵、删除当前词之后ppl打分

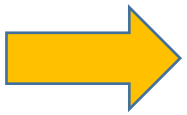
- 根据语境动态自适应的termweight。训练基于embedding的lstm网络，来动态计算每个term的词权重。

语义鸿沟：同义词挖掘

- 模板挖掘：
  - Bootstrapping
    1. 初始化种子数据(如：刘德华，华仔)
    2. 获取包含种子的句子集合(如：刘德华也被叫作华仔)
    3. 生成pattern
       a) pattern生成(如：XXX也被叫作XXX)
       b) pattern按频次排序，清洗过滤
    4. 基于pattern集合获取更多的SPO数据(如：姚明也被叫作大姚)
    5. 将4的结果回灌1中，迭代整个流程
  - 依存句法分析

巨型 |水虎鱼| 又 | 叫做 |非洲 | 虎鱼

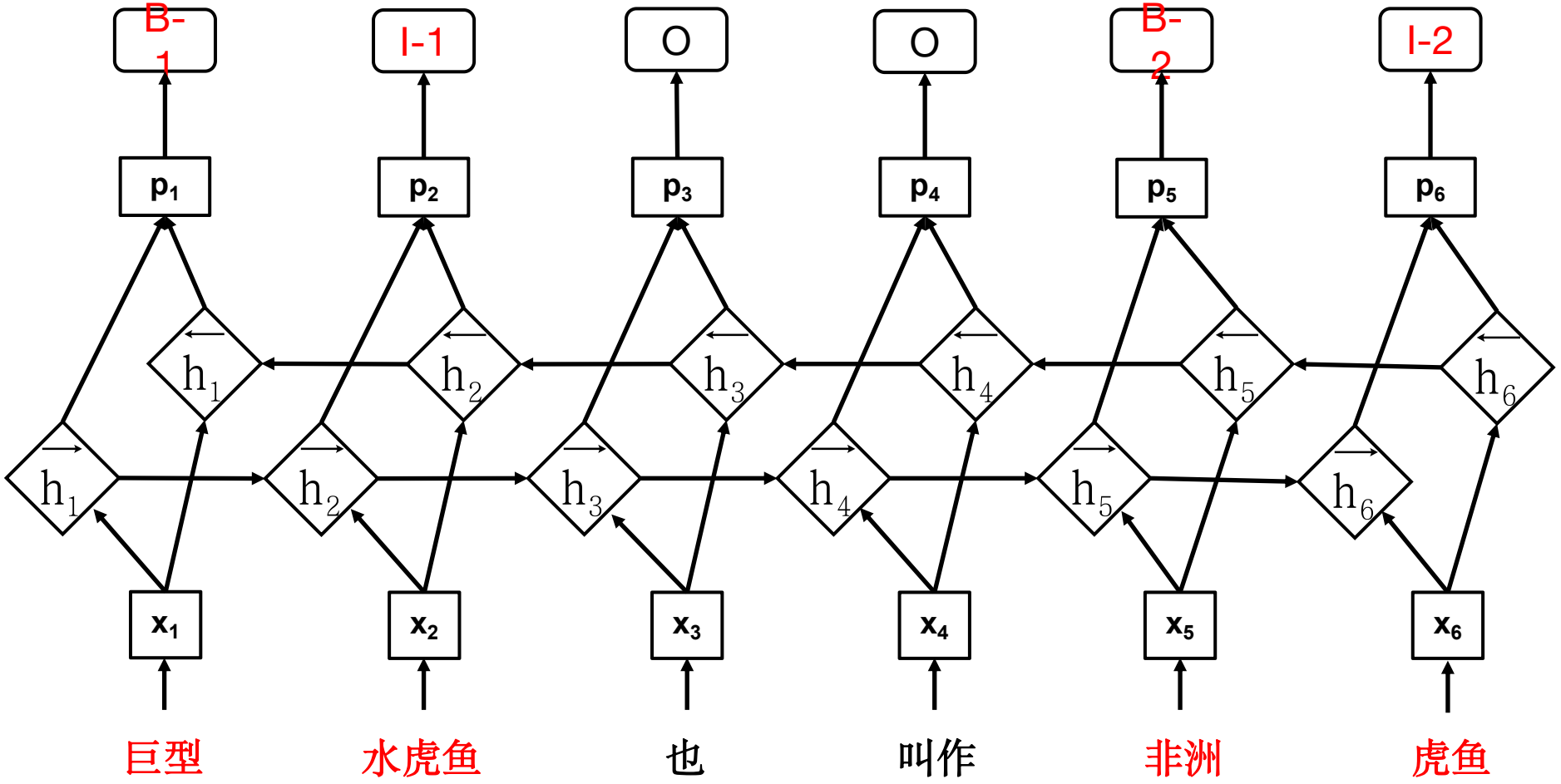
2:ATT| 4:SBV |4:ADV|0:HED|6:ATT|
4:VOB



巨型水虎鱼 同义词 非洲虎鱼

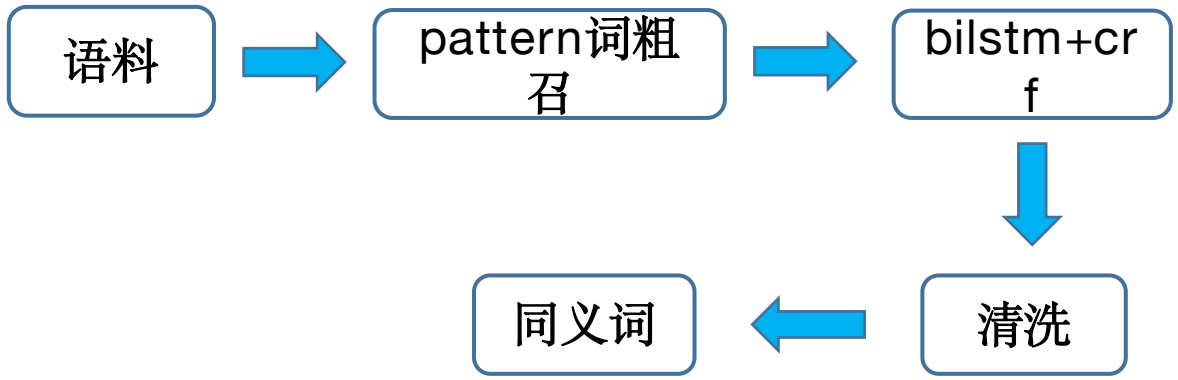
语义鸿沟：同义词挖掘

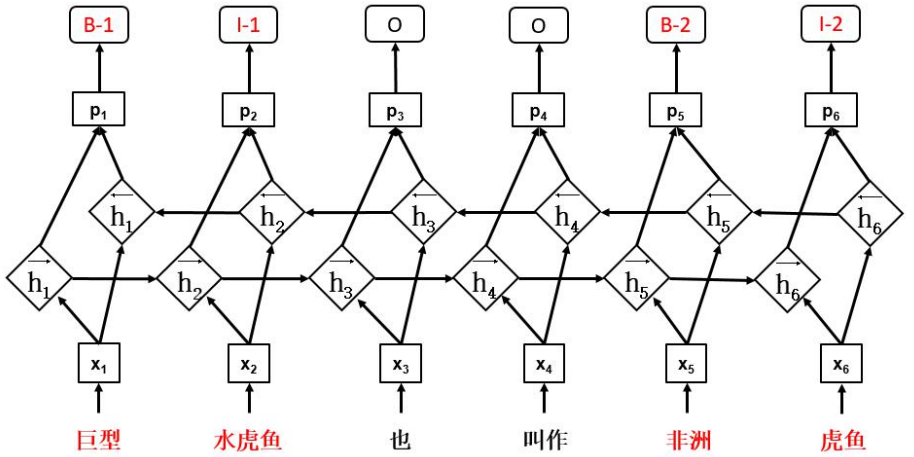
bilstm-crf模型



语义鸿沟：同义词挖掘

bilstm-crf模型





从线上业务效果出发，优先准确率

	百度百科	互动百科summary	互动百科content	搜狗问问
准确率	87%	89%	72%	77%
召回率	38.6%	43.3%	30.6%	26.2%

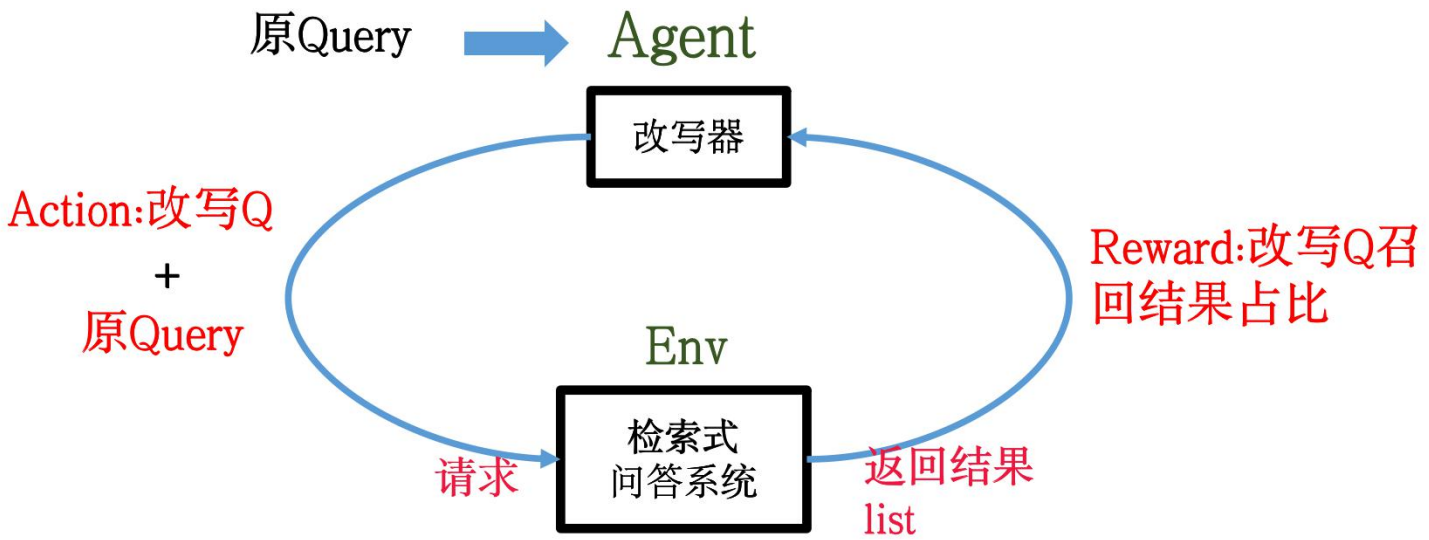
口语化问题

O	O	O	O	O	B-1	E-1	O	O	B-2	I-2	I-2	I-2	E-2	O
酒	精	依	赖	,	医	学	也	叫	慢	性	酒	中	毒	。

不含同义词的语句

O	O	O	O	O	O	O	O	O	B-1	E-1	O	O	O	B-2	E-2	O
与	杆	件	轴	线	相	重	合	的	内	力	,	称	为	轴	力	。

语义鸿沟：生成式改写



改写器：

- ✓ 利用人工标注数据预训练
- ✓ 利用线上未召回query进行强化学习训练

	query	检索结果
改写前	历史上最短的朝代是哪个朝	1.历史上中国最长的是哪个朝代 2.赵飞燕是历史上哪个朝代的 3.历史上禅让只在哪个朝代 4.唐朝是不是历史上最强大的朝代 5.中国第一个朝代是哪个朝
改写后	哪个朝代历史最短	1.中国历史上哪个朝代时间最长，哪个朝代时间最短 2.中国历史上哪个朝代存在的时间最长哪个最短 3.中国时间最短的朝代是哪个 4.青帝是哪个朝代的历史人物 5.裴松之是哪个朝代的历史人物

基于检索的FAQ问答

检索

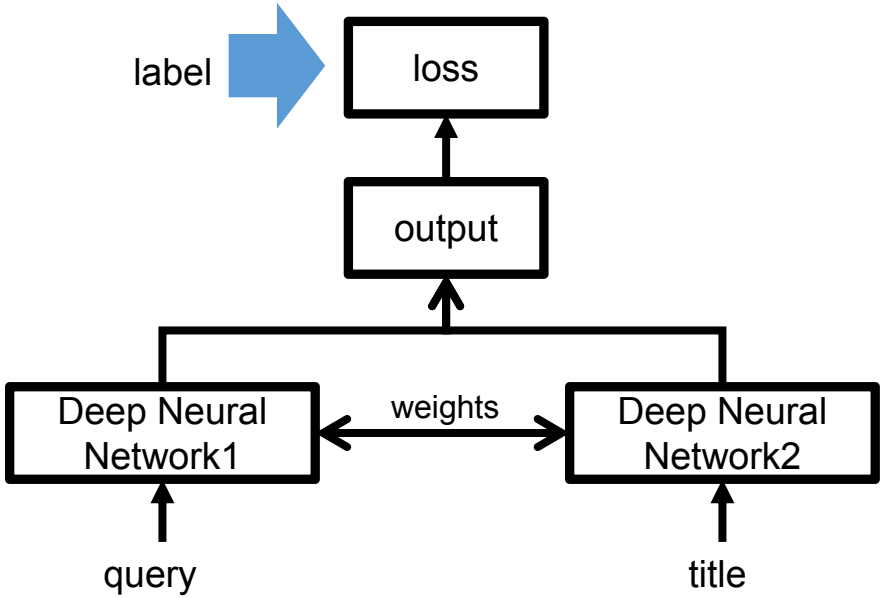
- term检索
- 实体检索
- 语义检索

匹配

- representation-based vs interaction-based
- 数据增强

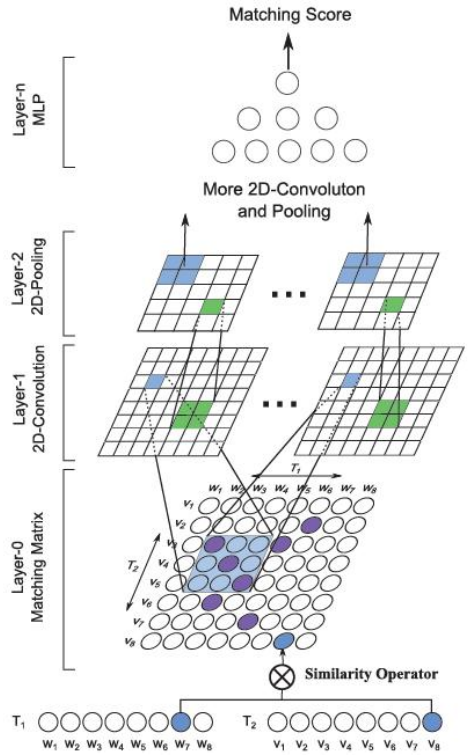
基于检索的FAQ问答—匹配

Representation-based Methods



- 双塔结构
- DSSM、CNN-DSSM、ARC-I等

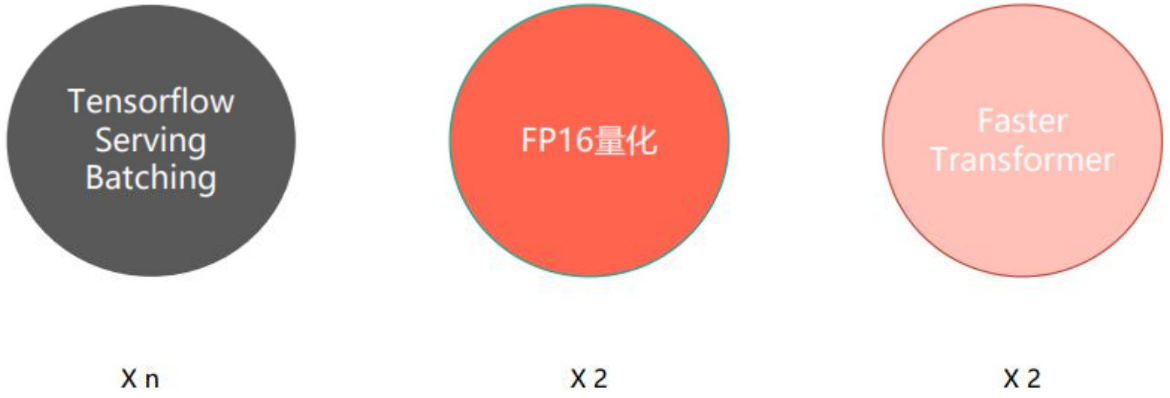
Interaction-based Methods



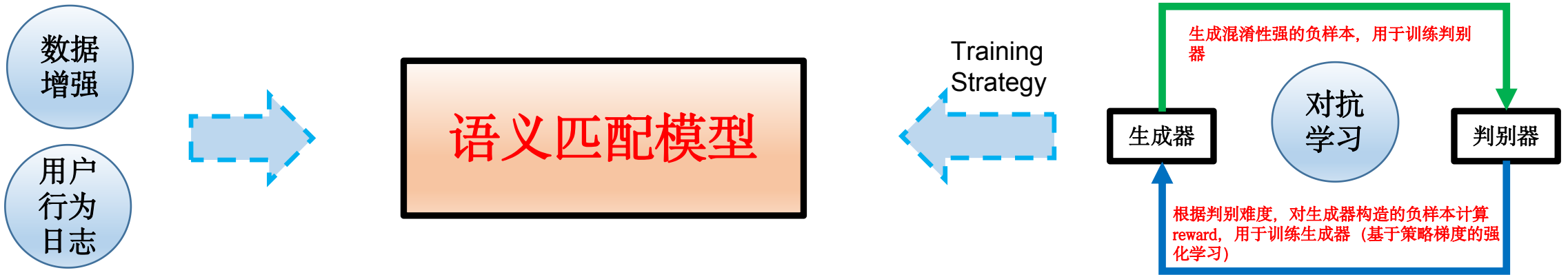
- ESIM、MatchPyramid、K-NRM、ABCNN、BiMPM、DIIN等

基于检索的FAQ问答—匹配

模型	字面匹配	word-level bilstm	Char&word-level bilstm	ESIM	BERT
改进点	base版本	对语义相似的pattern进行学习	解决未登录词的问题，扩充语义维度	对核心词对相关性的影响进行学习	效果最好
F1	71.5%	76.1%	81.8%	83.8%	86.6%



基于检索的FAQ问答—匹配



- 首先，基于用户行为日志的海量数据做粗训练，这部分海量数据质量较低，噪音偏多。
- 然后，用高质的人工标注数据+数据增强做进一步的fine-tuning。
- F1-score：86.6% => 88.0%

数据增强

- 正样本：Q1-Q2
  - Q1-Q1’ > 0.7
  - Q2-Q2’ > 0.7
  - 增强结果：Q1’-Q2’
- 负样本：Q1-Q2
  - Q1-Q1’ < 0.3
  - Q2-Q2’ < 0.3
  - 增强结果：Q1’-Q2’

基于检索的FAQ问答

匹配难题：语义焦点

Case 1:“古代如意是用来干什么的” – “古代石斧是用来干什么的”（语义焦点）

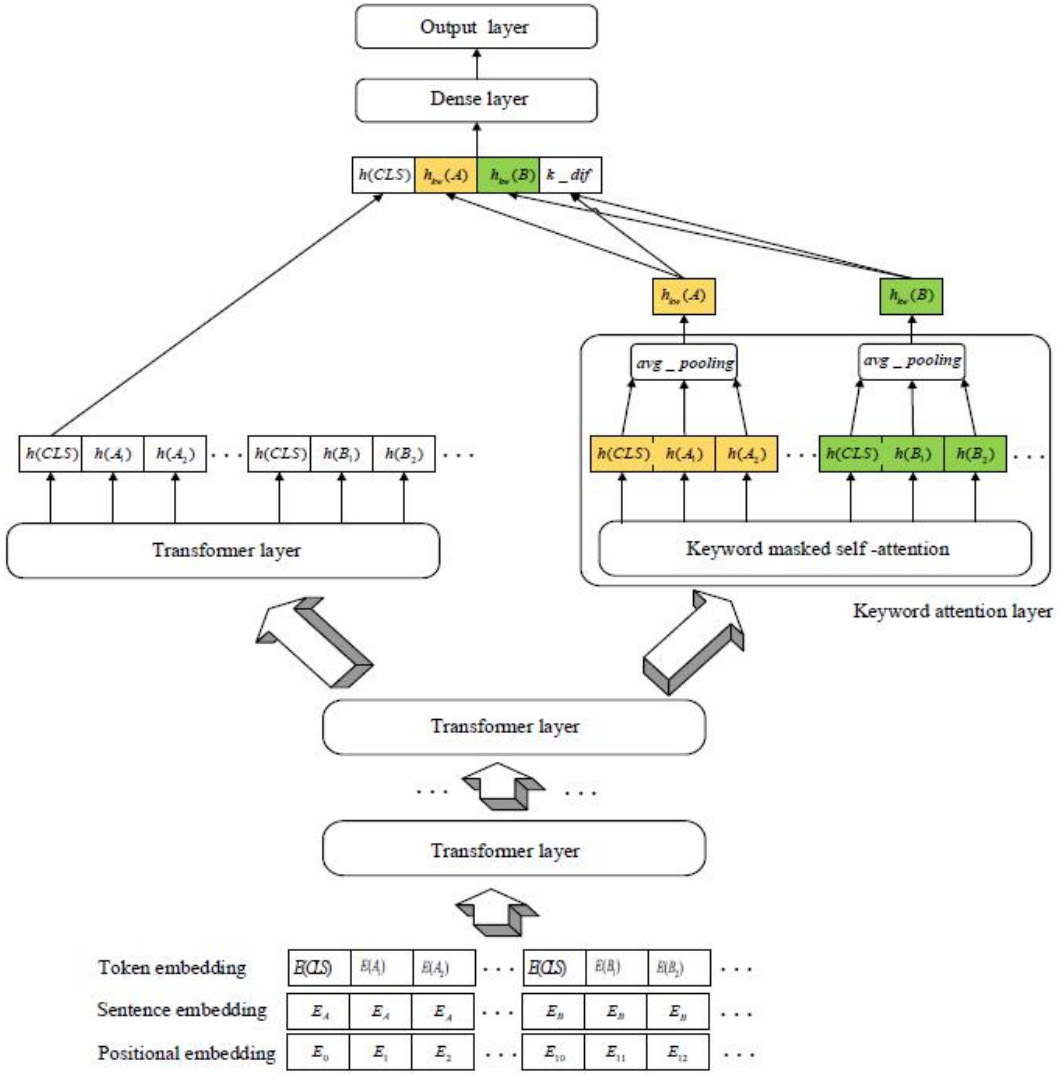
Case 2:西方情人节在国内相当于什么节” – “西方情人节在农历相当于什么节”（非语义焦点）

语义焦点：额外特征

构造样本：

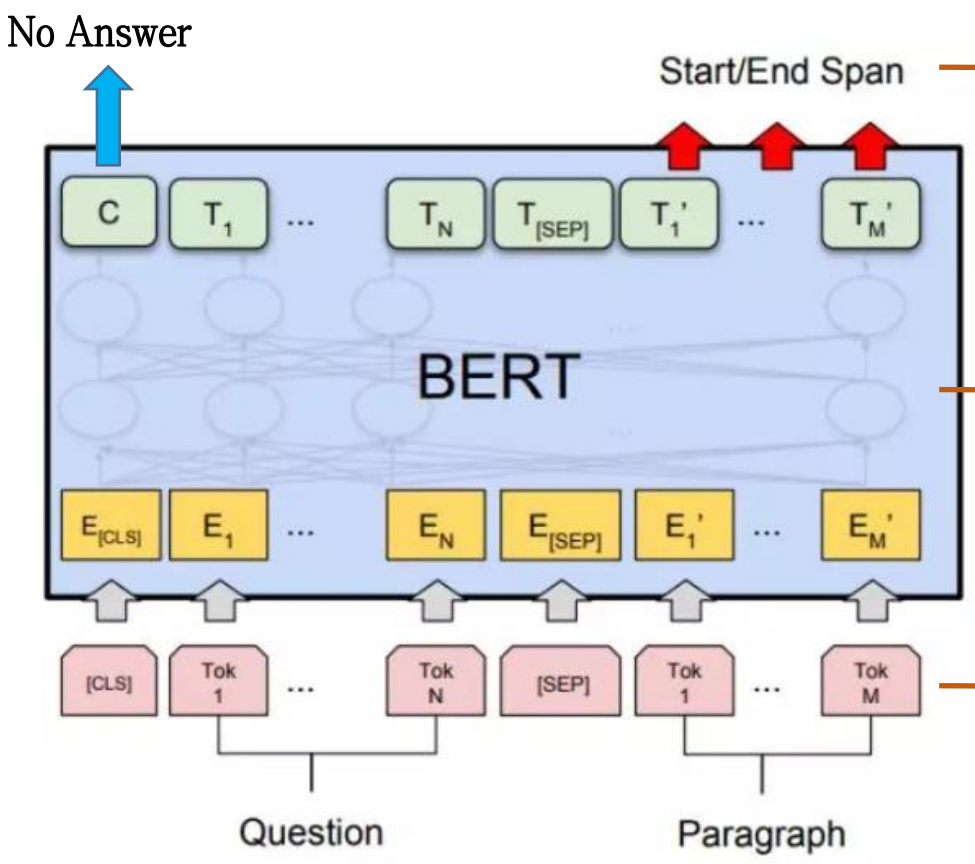
query → term-weight &实体链接 → 图谱相关实体替换

F1-score	without keyword	keyword-BERT
头部评测集	84.5%	86.7%
长尾评测集	84.0%	85.2%



目录

基于BERT的MRC模型



输出层：答案起点、终点位置预测

表示层：信息抽取

输入层：问题和篇章

主要问题

- 难点：
  - 开放场景下，候选passage太长，检索难度大
  - 误召比较严重，准确率难以适应产品应用
- 限定场景下的阅读理解
  - 只针对给定passage，去掉检索环节
  - 从passage中提取答案片段，给用户一个更直接的答案展现
  - 线下实验效果：recall/precision=81.9%/47.1%

目录

总结

KBQA

- 优点：准确率高，体验好
- 缺点：召回较低，泛化性差
- 定位：覆盖头部，打造精品

FAQ问答

- 优点：覆盖广，泛化性好
- 缺点：数据质量参差不齐，体验一般
- 定位：兜底模块，照顾长尾

阅读理解

- 优点：用户体验好
- 缺点：准确率低
- 定位：扩展问答能力

智能问答系统

欢迎大家使用小爱同学

谢谢!