阿里健康 | DataFun.

交互式BERT在医药电商搜索相关性中的探索

杨蕾(琼羽) 阿里健康 算法工程师

目录
CONTENT

阿里健康 | DataFun.

01 健康搜索业务和技术简介

健康电商搜索简介

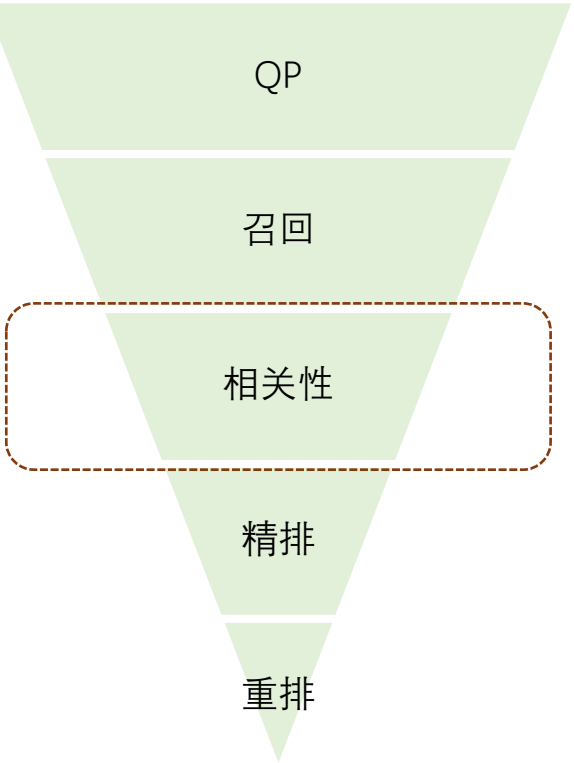
淘宝健康行业搜

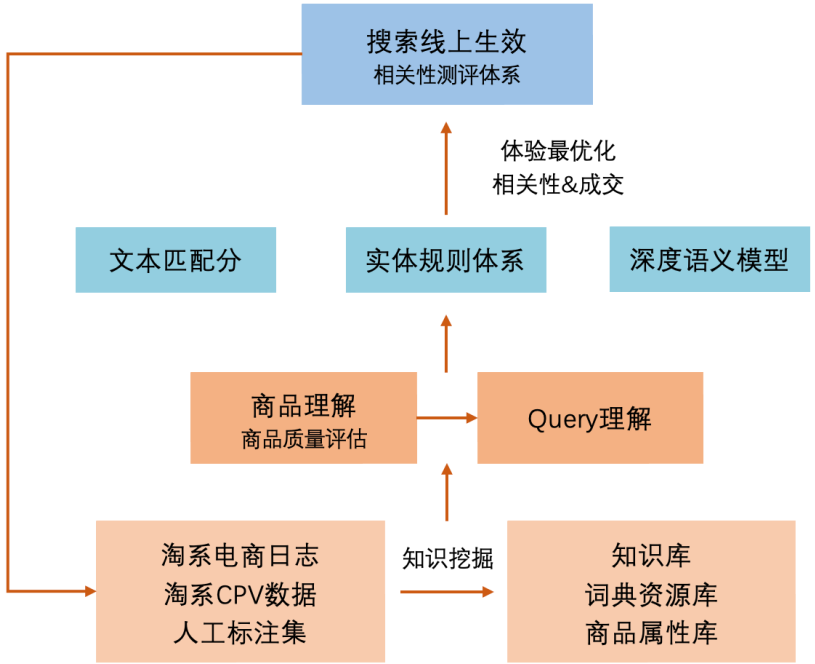
天猫好药

阿里健康大药房

阿里健康 | DataFun.

健康电商搜索的主要技术





AliHealth 阿里健康 | DataFun.

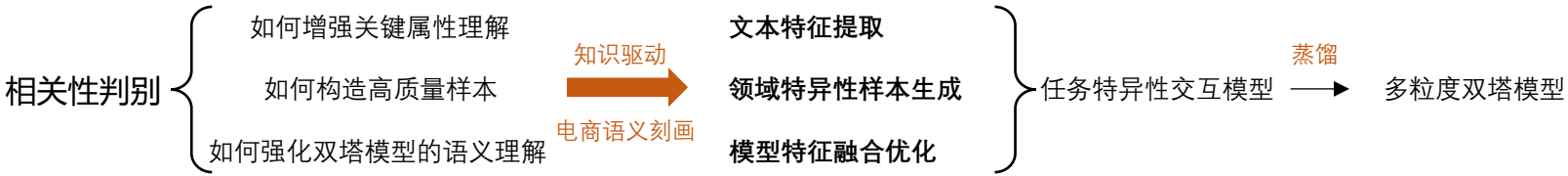
02 交互式BERT算法探索

语义模型设计背景

- 背景
  - 文本实体相关性强依赖知识和规则，成本高且难以保证覆盖率
  - query和商品的文本描述存在语义鸿沟
- 问题与挑战
  - 垂直搜索特点强，常规模型对关键属性的捕捉能力不足
  - 标注成本昂贵，缺少高质量训练样本
  - 上线RT要求高，双塔模型的表征能力有限

query	title	label
感冒清热颗粒	999感冒清热颗粒18袋感冒咳嗽咽干流清涕风寒感冒	good
	999感冒灵颗粒9袋感冒颗粒清热解毒鼻塞头疼	bad-药品名不符
	999板蓝根颗粒清热解毒咽肿痛喉咙发炎风热冲剂	bad-药品名不符
	欧意感冒清热软胶囊非颗粒24粒头疼发热咽痛咳嗽	bad-剂型不符

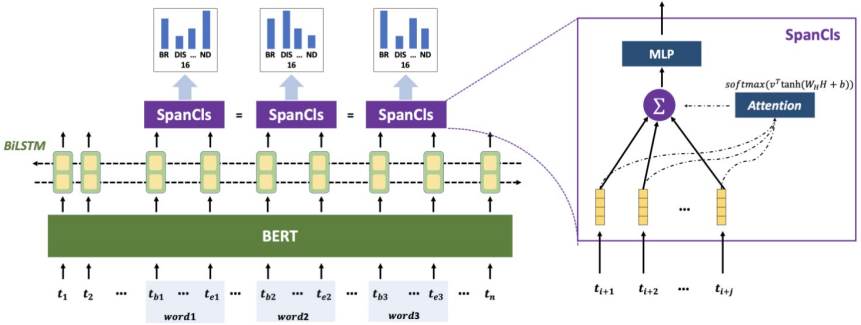
- 优化思路

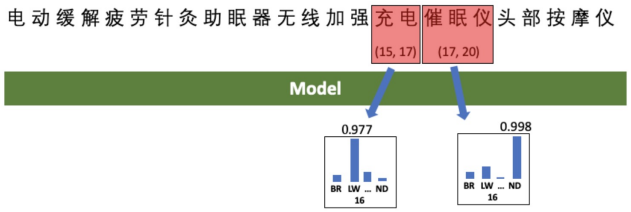


文本特征提取

query特征提取

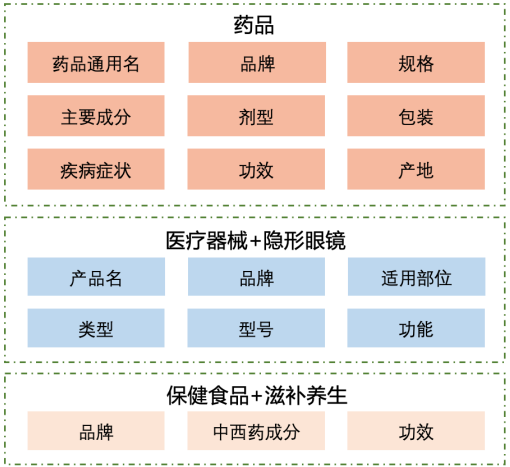
- Step1：分类目圈定重要实体类型，建立实体词典、同义/归一化词典
- Step2：基于词典和贪心算法，计算query多重分词打标
- query = 阿莫西林胶囊（药品通用名）白云山（品牌）
  - 阿莫西林胶囊：主要成分 + 剂型

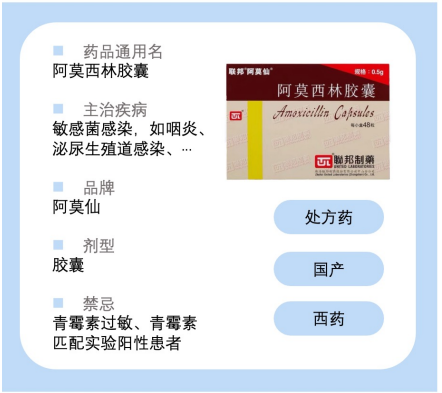




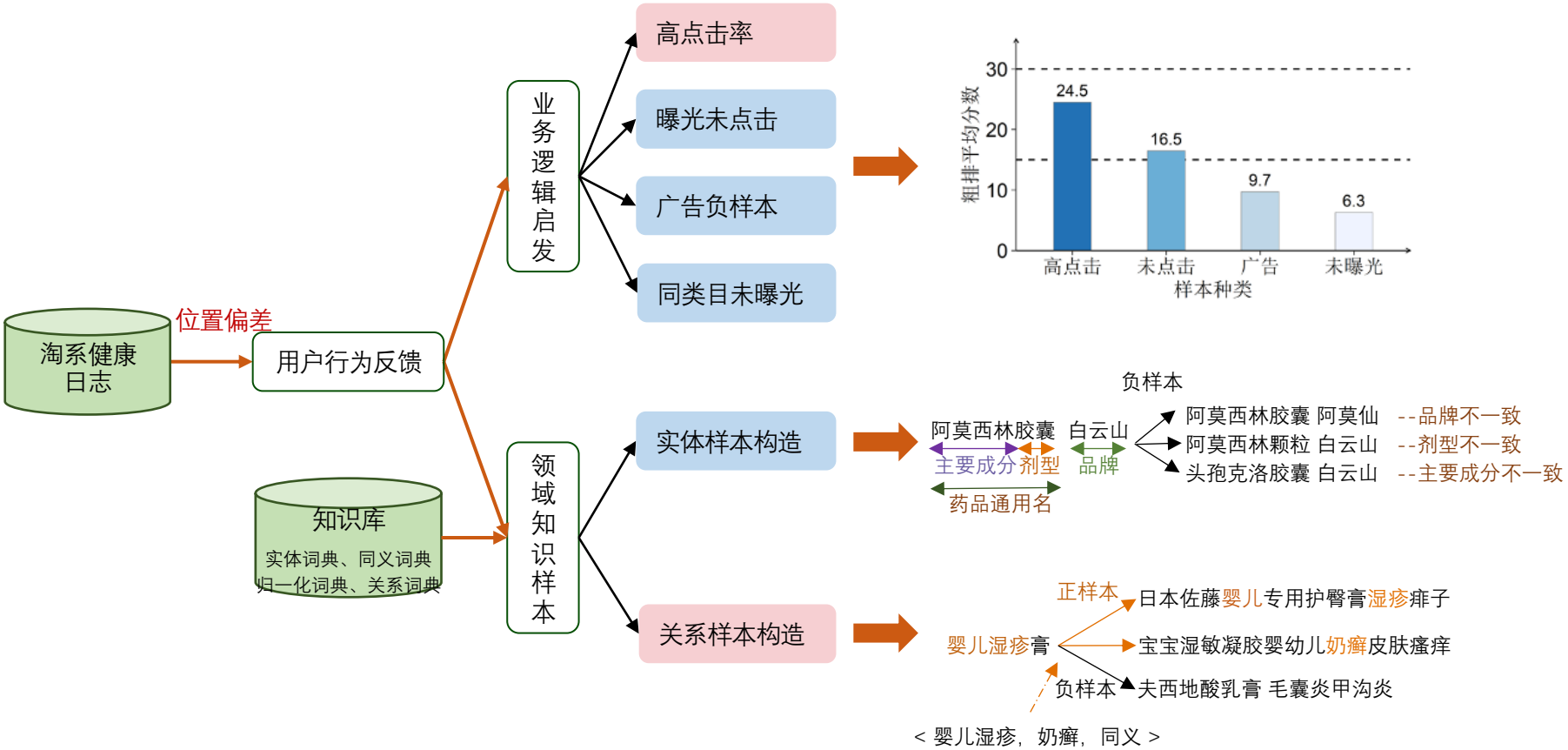
商品特征提取

- Step1：商品结构化理解，依赖淘系CPV数据
- Step2：商品属性补充和核实，结合商品标题、商品详情页、国家药监局等内外部信息
- 难点：商品信息真实性，标题堆砌、属性乱填





领域特异性样本生成



阿里健康 | DataFun.

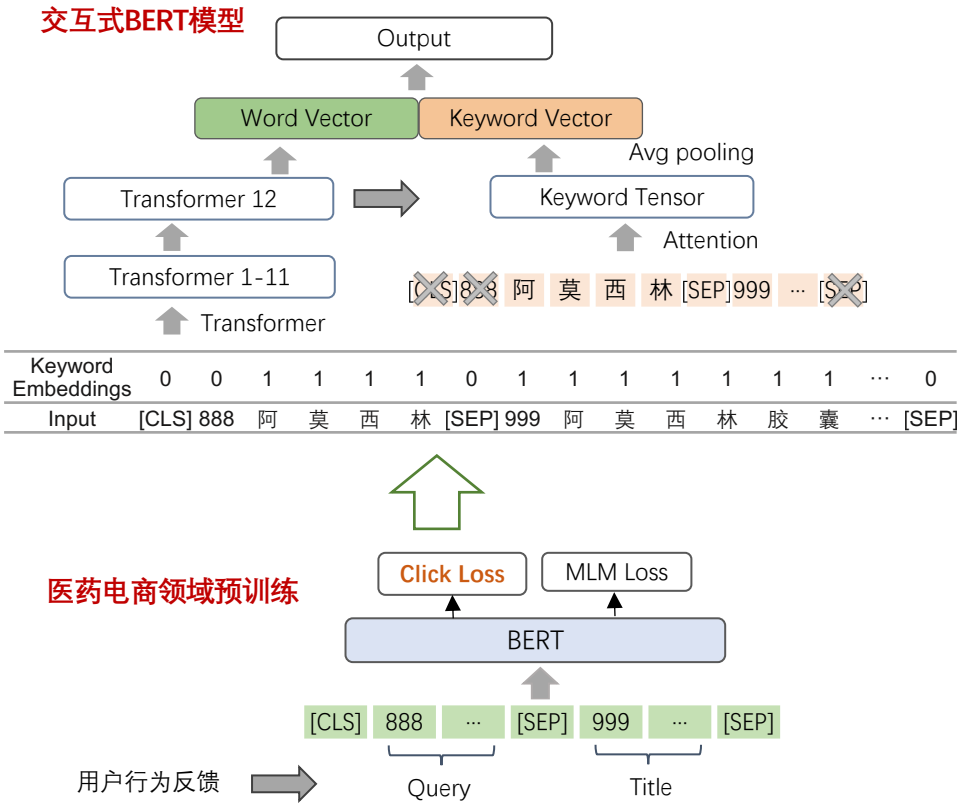
交互式模型结构

- 交互式BERT模型
  - query/title/query特征/商品特征分别输入
  - 新增keyword embeddings记录是否为实体特征
  - keyword transformer强化关键特征交互
- 医药电商领域预训练
  - 2亿用户行为样本continue-train
  - 训练目标：Click Loss + MLM Loss

$$Click\ loss = -\sum_{i=1}^{N}(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i))$$

- 离线实验

Models	AUC
DSSM	0.885
BERT(sample-enhanced)	0.908
$BERT_{kw}$(keyword-enhanced)	0.919
$BERT_{kw}$(continue-train)	0.927



阿里健康 | DataFun.

03 模型应用实践

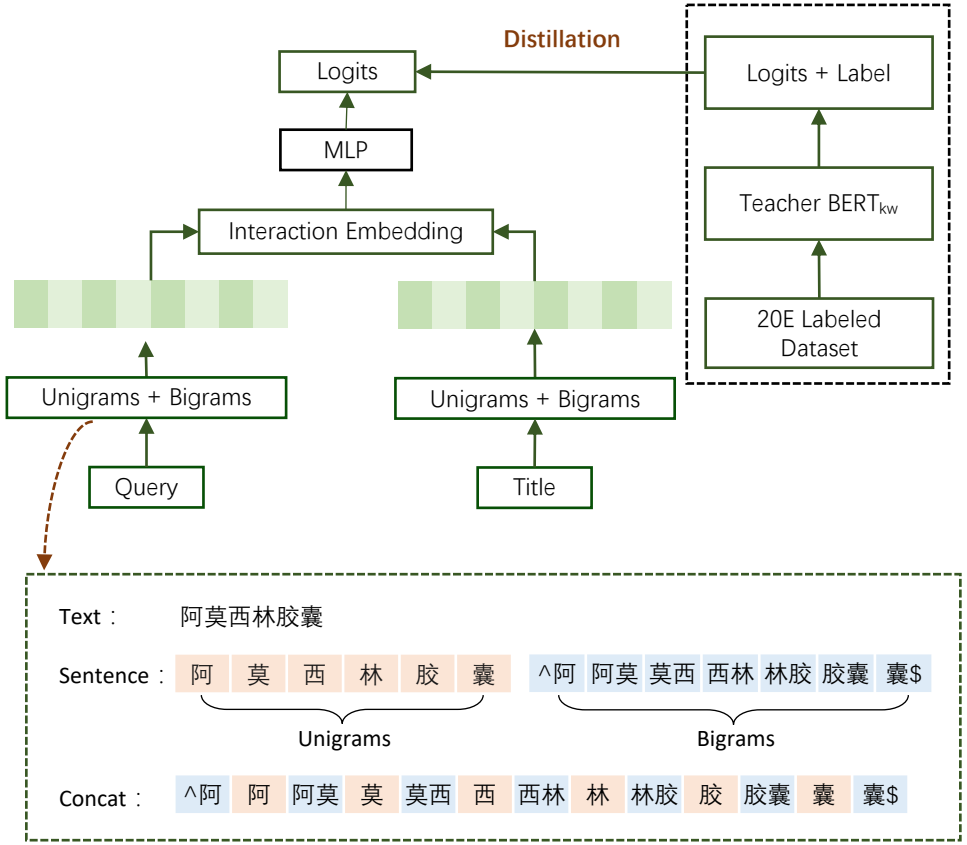
双塔模型蒸馏

- 多粒度双塔模型
  - query/title文本分别输入，单字+双字组合
  - Avg-pooling和Max-pooling输出编码
  - query和商品向量算分
- 任务导向的知识蒸馏
  - 20亿领域特异性带标签数据
  - teacher soft label+真实hard label

$$loss = -\sum_{i=1}^{N} \lambda\left(y_i^{soft} \log(\hat{y}_i) + \left(1 - y_i^{soft}\right) \log(1 - \hat{y}_i)\right)$$
$$+(1-\lambda)\left(y_i^{hard} \log(\hat{y}_i) + (1 - y_i^{hard}) \log(1 - \hat{y}_i)\right)$$

- 离线实验

Models	AUC
$BERT_{kw}$[teacher]	0.927
$BERT_{kw}$2DNN[student]	**0.921**
$BERT_{kw}$2DNN(soft label only)	-1.7pt
$BERT_{kw}$2DNN(MSE loss)	-1.1pt
$BERT_{kw}$2DNN(Brigrams only)	-0.9pt
$BERT_{kw}$2DNN(Trigrams)	-0.3pt



双塔模型效果分析

- 上线方式
  - 内容表存储商品，在线产出向量并计算
  - 文本相关性+实体相关性+语义相关性 融合计算
  - 结合成本样本集，确定权重分数和阈值
- 线上效果
  - 相关性指标：曝光pv为权重随机采样，人工标注
  - 成交效率：在线AB

场景	人工测评 good 率	成交订单	成交金额
阿里健康大药房	+2.64pt	+3.72%	+3.16%
天猫好药	+3.94pt	+1.94%	+0.90%

- case展示

  语义模型上线后，query=“测压器” 的商品由3个涨至19个。

非常感谢您的观看

阿里健康 | DataFun.