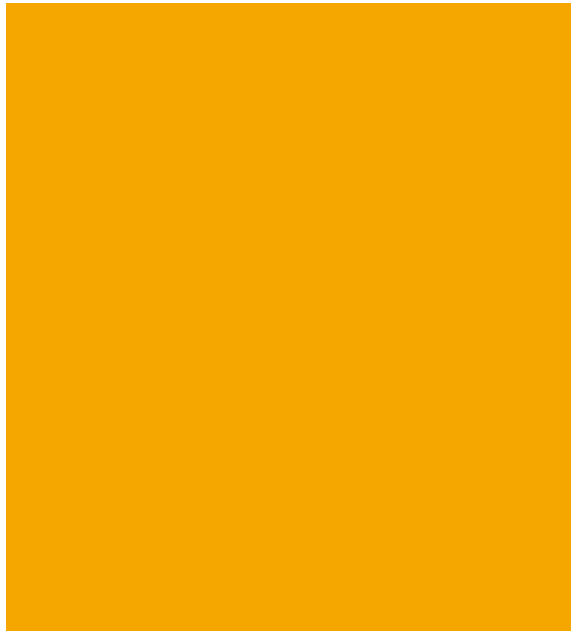
阿里实习介绍

张颖而

目录

- 交互式搜索Tag选择任务
  - 项目背景
  - 目标
  - 方案
  - 效果评估及应用场景

交互式搜索Tag选择



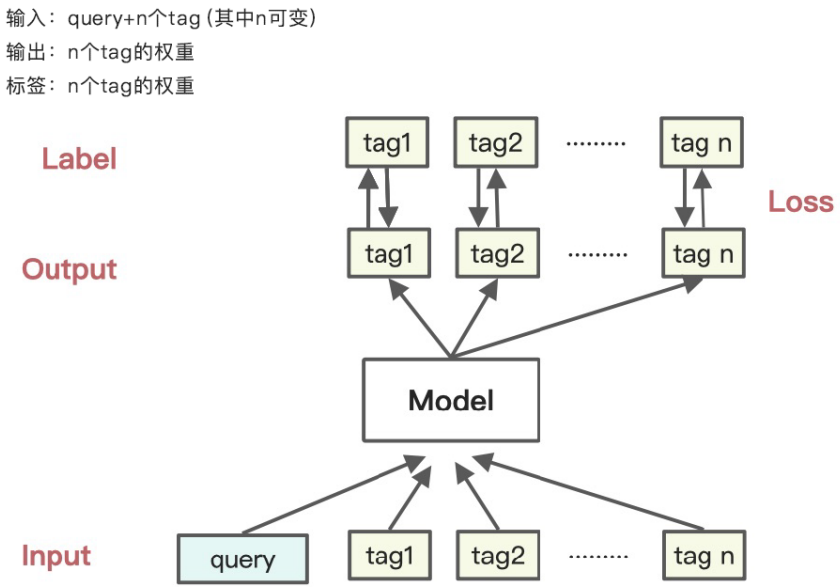
业务背景

- 医疗交互式搜索场景下，用户输入query=“嗓子痛说话没声 ”提供一些tag让用户进行选择补充query的症状
- 当前搜索方案：query+tags 组成新 query 重新发起搜索

  Q = 嗓子痛说话没声 讲话多 过度用声 一周 咽痛
- 当前搜索方案缺点：
  - **这些tag的来源是根据医生问诊的术语，和搜索结果并不直接相关的**
  - **tag 过多，太长 （目前选前3个）**
  - **直接拼接非自然语言**

改进目标：如何选tag？如何组tag？评估标准？

目标: 整体架构



- 给一个qurey 和一堆tag, 选出最好的tags
  - 方法一：query+n tag 组合输出一个结果，但是组合数非常多且耗时
  - **方法二：给每个tag 打分，输出n个tag 的分数 (query只和单个tag有关)**
- 最佳搜索效果的标准(label)：

  搜索结果确保原始query不偏移且新添加tag的信息

  Q=烫完头头皮屑特别多 tag=头皮瘙痒

  搜索结果： 烫完头头皮痒而且头皮屑多

  Q=酒后喝牛奶吐 tag=喷射性

  搜索结果：喝牛奶出现喷射性呕吐（query语义丢失）

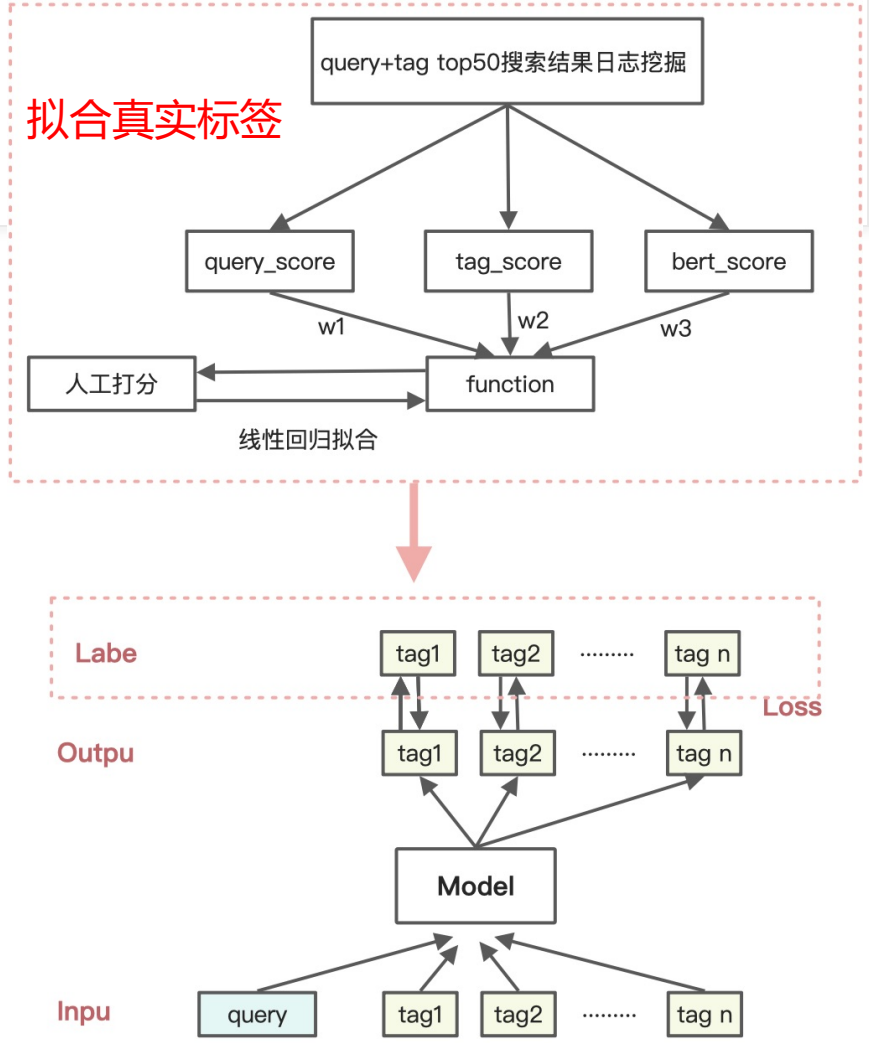
  喝酒吐了能不能喝牛奶（query 偏离且tag丢失）

方案—标签设计

- QPP (Query Predicting Performance)
  - 需要客观指标 -人工标注存在数量的局限性和质量的主观性
  - 从搜索结果计算(post-retrieval) tag 的label能反映整个搜索系统的效果
  - Step1:使用query+单个tag发起搜索，在top50搜索结果中统计三个指标：**query_score，tag_score，bert_score**
    (NER实体匹配，文本飘红，同义词替换，语义重复)
  - Step2:使用线性模型融合三个指标。具体的参数需要拟合人工打分标签。

- 为何线上不直接计算搜索结果，而需要模型预测？
  因为发起一次搜索耗时

- **同query同tag下正序比例：0.82。设计的label基本满足实际需求**



方案—模型设计：DeepCT

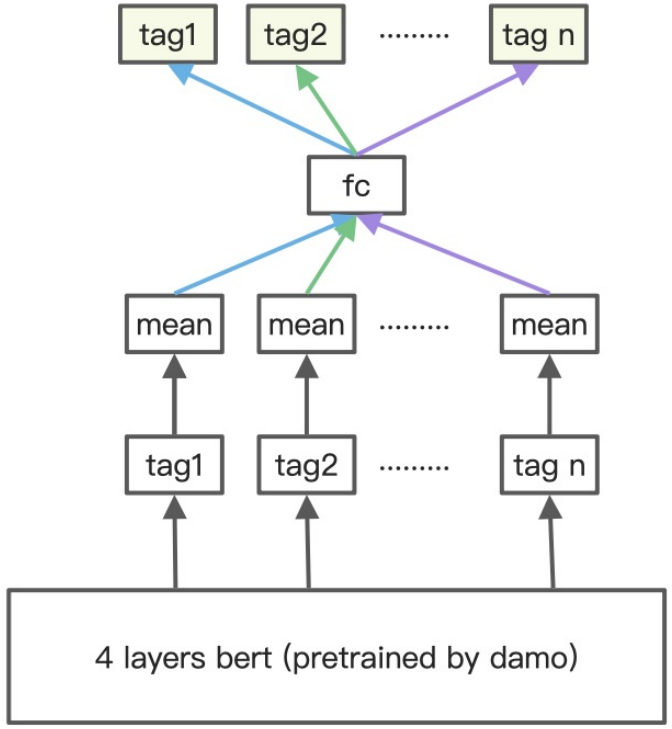
解决Tag数量不确定 : mask

得到预测的tag

每个tag单独输入全连接层

对tag中所有字的向量求平均

根据 tag_mask确定每个tag
在bert中输出的位置



attention_mask: 标识有没有字

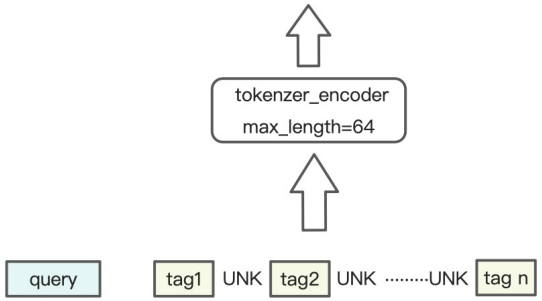
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,

input_ids: 编码

[101, 1391, 1776, 691, 6205, 2861, 5496, 2094, 102, 4635, 1921, 100, 671, 1453, 102, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,

token_type_ids：标识text1/text2

[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]



tag_mask: shape (batch_size, max_tag_len, max_seq_len) 对每个样本
(dim=0) 中的每个tag(dim=1)记录是否哪些位置上有效tag

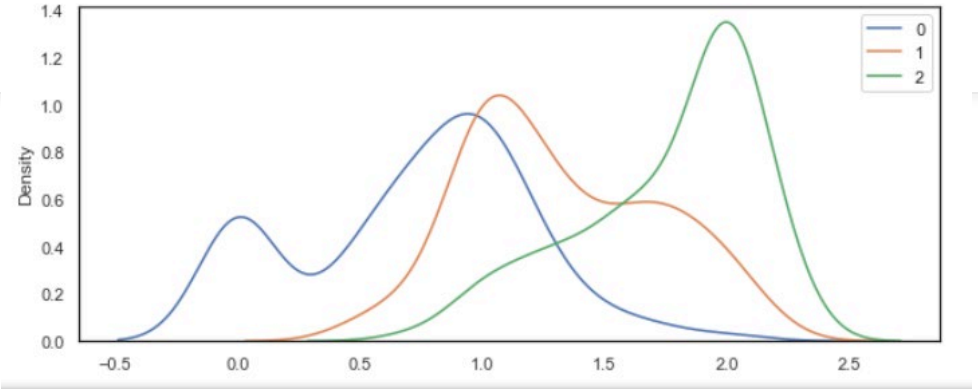
[[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],.....
[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]]

labels :标识每个tag的tern weight (补全到10位)

[0.5, 0.3, −1,−1,−1,−1,−1,−1,−1,−1]

Query+5个tag为一组去训练，打乱顺序的扩充样本

方案—模型训练及结果



- 数据量
  - 训练集 41730
  - 测试集 10433
- Loss
  - MSE loss
  - Pair loss:当两个tag之间label差值大于阈值/小于负阈值时，分类为1或0，将差值做sigmoid后，按照二分类计算loss
- 指标
  - 模型正逆序 7.99 （评价模型学习能力）
  - NDCG 0.9571（deepct预测结果和真实人工测评500份）
    - 例： 人工 [2, 2, 1, 2, 0] 预测：[1.7683, 1.5864, 1.4471, 1.3742, -0.0204]
  - 相比于直接相关性模型NDCG 0.7970有所提升
  - 实际应用场景中，选择TOP3 VS原来选择最新的三个，三个标签发生变化比例40% Diff 率40%，Diff中新老版本人工打分G:S:B=47:40:13

方案—效果

- Case

query=腹胀
原始tag: '多食产气食物', '过量饮用碳酸饮料', '运动少'. 评测：2,2,0
deepct后tag:'多食产气食物', '暴饮暴食', '过量饮用碳酸饮料',评测 2,2,2

query=下车后头晕
原始tag='乘坐交通工具' '其它条件''呕吐'. 评测：0,0,2
deepct后tag= '恶心', '呕吐', '头痛'. 评测：2,2,2

应用场景

- 出卡片过程中截断阈值以下tag
- 用户点选后选择TOP3 Tag
- 展示界面给出tag的排序，提升用户满意度