CCKS 2020：基于标题的大规模商品实体检索

任务：本评测任务为基于标题的大规模商品实体检索。即对于给定的一个商品标题，参赛系统需要匹配到该标题在给定商品库中的对应商品实体。针对检索任务采用了召回 -> 粗排序 -> 精排序的大思路。

数据处理：

原始数据比较乱，需要做一些预处理方便后续模型的训练，预处理如下：

1. 训练集中存在 text_id 不唯一的情况，我们对 text_id 重新赋值保证其唯一性，方便后面的处理。
2. 训练集存在大量噪音数据，存在相同文本对应多个实体 id 的情况，直接删除这部分数据。
3. 知识库存在图书类别相关实体很多，但是训练集中图书类别实体占比较小，为了提高性能和整体效果删除图书类别实体。
4. 知识库中存在一些相近，实体名字相同产地相同，只有部分小的差别，这种数据模型很难学习，去除没有在训练集中出现的那个。
5. 实体描述文本构建，是将谓语和宾语通过‘为’相连，例如‘产地为新加坡’，谓语和宾语相连后再通过以下谓语顺序连成一条文本， ['产地', '功能', '症状', '主要成分', '生产企业', '规格']。

对于情况 2,4，这样的数据在最后的排序阶段会影响模型的收敛，导致最后的性能下降，故直接去除这部分数据。

召回模型：

召回模型采用 Triplet BERT ，模型图如图 1 所示，Ancho 为商品标题，Positive 为对应的实体描述文本，也就是正样本，Negative 为随机选取的负样本。训练时将上述三个文本输入到 BERT 模型中，选取 BERT 模型的所有 Token 向量平均作为输出，将三个输入向量进过 TripletMarginLoss 得到损失值完成模型的训练。部分训练细节如下：

1. online triplet mining 负样本选择，即动态负采样，在训练中的每个批次（batch）中，都得对三元组进行动态地采样样本。
2. 学习率：BERT 层采用 3e-5，其他层采用 4e-5，为了更好的收敛采用了指数衰减，衰减指数为 0.5.

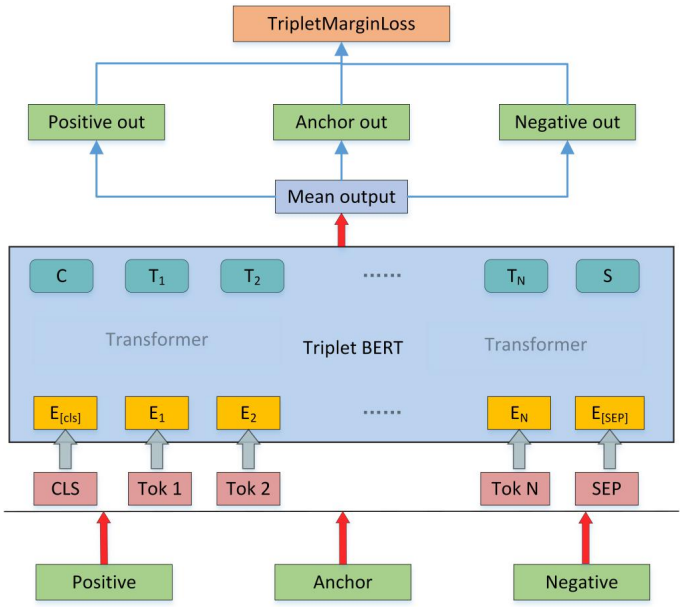

图 1

模型预测阶段则将所有标题文本和所有实体描述文本都经过模型得到向量。然后针对某一个标题文本的向量和所有实体描述文本的向量进行距离计算，选择距离最近的 Top100，得到了该文本的前 100 个召回实体。

我们采用了交叉验证对训练集进行预测，得到训练集每个标题的前 100 个召回实体。对于测试集则采用了概率求平均的进行模型的融合。其中 BERT 模型采用了两种预训练，分别是百度的 ernie-1.0 和 roberta-wwm ，对于两个模型预测的结果也是采用了取平均的方式。

排序模型：

排序分为两个阶段，分别是有 top100 排序得到 top10，再有 top10 排序得到 top1。每个标题的对应实体的 top100 有前面召回得到，在 top100 的基础上构建粗排序模型。由粗排序模型得到 top10，然后在 top10 的基础上构建精排序模型。

排序模型如图 2 所示，排序模型采用了二分类的方法，对每一个候选实体进行预测，然后对预测的概率进行排序。模型输入有标题文本和实体描述文本构成，如：

标题文本：虎镖肩颈舒

实体描述文本：虎标颈肩舒产地为新加坡，症状为舒压按摩，缓解肌肉紧绷，僵硬，酸痛等，主要成分为薄荷脑，水杨酸甲酯

将上述两段文本连在一起为：

[CLS]虎镖肩颈舒[SEP]虎标颈肩舒产地为新加坡，症状为舒压按摩，缓解肌肉紧绷，僵硬，酸痛等，主要成分为薄荷脑，水杨酸甲酯[SEP]

将上述文本输入到 BERT 模型进行二分类，得到该标题与该实体的概率。

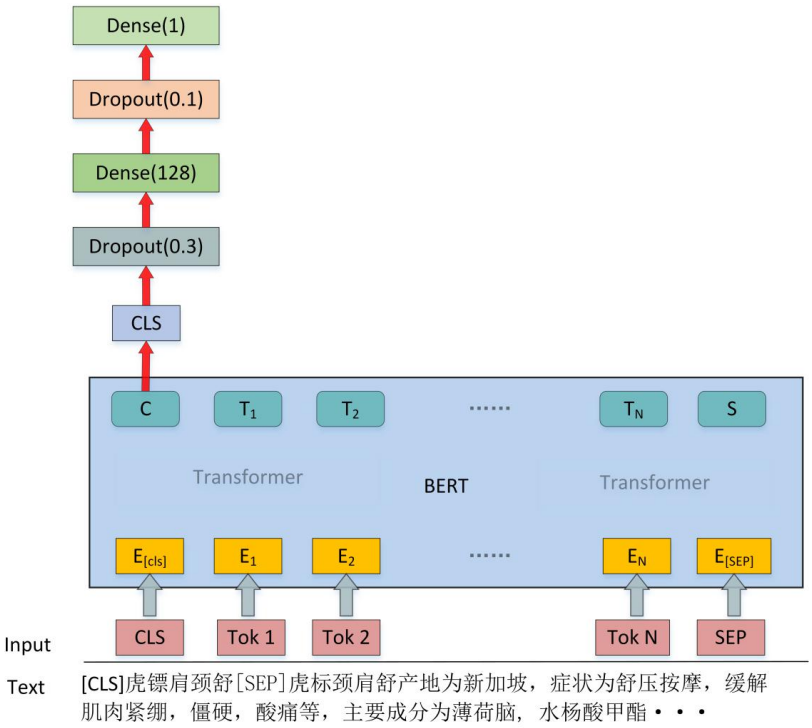

图 2

top100 -> top10，模型细节

1. 负样本个数为 3
2. 采用 ernie-1.0，roberta 两个预训练模型，对预测结果求平均

top10 -> top1，模型细节

1. 负样本个数为 2
2. 采用 ernie-1.0，roberta-wwm，bert-wwt 两个预训练模型，对预测结果求平均