心法利器[13] | 任务方案思考：句子相似度和匹配

原创 机智的叉烧 CS的陋室 2020-12-19

收录于话题

#心法利器 15 #自然语言处理 14 #搜索 10 #对话 9

Auf Und Auf Voll Lebenslust

Franzl Lang - Aus Freude am Leben

【前沿重器】

全新栏目，本栏目主要和大家一起讨论近期自己学习的心得和体会，与大家一起成长。具体介绍：

仓颉专项：飞机大炮我都会，利器心法我还有。

往期回顾

- 心法利器[8] | 模型热更新小记
- 心法利器[9] | 算法项目从0到1孵化过程
- 心法利器[10] | 算法项目从1到N的进化
- 心法利器[11] | 任务方案思考：文本分类篇
- 心法利器[12] | 任务方案思考：序列标注（NER）篇

0 小系列初衷

自己接触的项目大都是初创，没开始多久的项目，从0到1的不少，2020年快结束，感觉这个具有一定个人特色的技术经验可以在和大家分享一下。

计划篇章：

- （已完成）文本分类篇。针对NLP文本分类任务。
- （已完成）序列标注（NER）篇。针对命名实体识别、序列标注任务。
- 文本匹配篇。针对语义相似度计算、向量匹配等问题。
- 人工特征学习篇。针对多特征的机器、深度学习方案。

开始我把这个标题叫做语义匹配，后来感觉还是不能叫这个名字，应该把问题放大为句子相似度和匹配问题。

1 语义匹配的场景

语义匹配的核心其实是评价两个query之间的相似度，可以看看现在常用的场景：

- 搜索领域，语义向量召回是一个比较新潮的召回方式，灵活性更高，下游的精排部分也可以通过语义相似度来进行排序。
- 智能客服，之前的阿里小蜜的文章也提过，对于长尾的结果，可以通过向量召回的方式来进行处理。
- 对话领域，可以说是智能客服的眼神，闲聊类的，可以通过语义匹配完成闲聊的回复，当然多轮也有多轮的玩法。

可以看到，各种领域，其实语义匹配的舞台非常大，了解这方面的方案对NLP技术栈的了解非常有用。

2 方法选型

2.1 文本层面的相似

最简单的方法往往就是最浅层的方案，所以还是文本层面的相似，方法逐步升级是这样的：

- 编辑距离，这应该是最严格的一种相似了。
- cqr，分子是句子1和句子2词汇的交集词汇量，分母是句子1和句子2的并集词汇量。
- 加权的cqr，可以做一个简单的词权重，然后做加权的cqr。
- BM25。传统搜索的常用方法。

文本层面的方法，在搜索领域已经非常成熟，BM25已经具有很高的准度，结合上游常用的一些改写，其实已经能够达到很好的效果，这也是经典搜索最常用的一套范式。

2.2 向量表征作召回

向量召回是当前比较流行的一种新的搜索技术，这里以来两个关键技术点，向量索引和句子表征技术。

向量索引的是指就是一种向量最近邻的搜索方案，最常用的场景就是KNN，而在我们的场景中，就是把句子表征成一个向量，构建索引，新来一个句子，用同样的放哪个还是构建一个向量，就可以完成相似度召回，常用的构建索引方式推荐两种，这两种都已经有开源工具支持。

- annoy，一种基于树的构造方法。
- hnsw，一种基于图的构造方法，这应该是目前我已知速度最快的方法了。

说完了向量索引，就要说向量表征了，只有足够好的向量表征，上面说的向量召回，召回的东西才会足够好，为什么我说好呢，就是因为这里涉及的好的维度多：

- 准确率足够高，召回的内容真的是和句子足够接近。
- 有比较强的泛化能力，这也是语义向量召回相比传统搜索的相似召回最突出的优势，只要语义足够接近，“查询”和“查看”就可能匹配到，“幂幂”和“杨幂”也能打中，这样能降低我们挖掘数据带来的成本。
- 好的相似度匹配能识别关键词，只需要模型端到端处理，不需要单独抽关键词。

那么，这个语义表征，一般都是什么方法呢，这里也是提几个：

- word2vector预训练。如果语料不足甚至没有语料，我们其实可以用开源的预训练好的w2v词向量作为基线，取均值就能拿到句向量。
- 如果有一些平行样本，可以开始考虑用一些平行预料（sentence1,setence2,label）进行finetuning，说白了就是两个向量分别去词向量后均值，最终用余弦或者欧氏距离计算相似度就行。
- 数据量足够后，就可以开始在上面搭积木了，CNN、LSTM之类的都可以尝试，当然经验之谈，self-attention可以尝试。
- 数据量再多点，我们就可以上bert之类的大家伙了。

现在的语义相似度，更多是通过优化交互特征来提升相似度计算的效果，但是在向量召回这里，由于目前只能支持简单的相似度召回，两个query只有在计算相似度的最后一步才能够见面，因此query之间的交互特征是无法提取的，所以很多现在流行的方法是用不了的。

2.3 语义相似度

如果语义相似度要被用在后续的精排，无论是搜索、对话甚至是推荐，在经历初筛之后，我们往往有更多时间和经历来比对剩余的结果和用户query之间的相似程度，此时我们就可以使用交互特征逐一匹配，完成最后的精排，这些方案往往在大量比赛中就有提到，以DSSM为基，升级很多方案，包括很多人知道的EISM等，当然比赛的经验也告诉我们，模型本身还可以加入更多的文本特征来协助衡量语义相似度，因此在用语义相似度模型的同时，可以加入一些人工特征来协助优化，这也是推荐系统的wide&deep中所提到的深浅层特征均用的思想。

这里给一篇蚂蚁金服比赛的文章吧，大家可以根据这个思路去参考优化：
https://blog.csdn.net/u014732537/article/details/81038260

3 优化手段

当然，上面的方式是让大家用最快的速度去完成一个demo或者说baseline，然后我们需要一系列的手段进行优化，在这里也给大家介绍一些有用的方案。

- 如果你的场景里需要一些英文，可以加入一些英文文本去finetuning，开放域的。
- 针对问答场景，由于用户的问题都有明显意图，因此做一些词权重、attention的操作有利于效果提升，包括提槽，当然在浅层模型的情况下，词的归一化也有好处。

- 通过传统的搜索，用ES召回之类而方式，可以召回很多文本接近但是语义遥远的case，通过人工标注的样本对效果的提升很有好处。
- 同样是hard case挖掘，用自己的语义模型做召回，召回在阈值附近的case，做一下人工的复核，这样做样本也对效果提升有好处，这其实用的是主动学习的思想。

4 小结

做完搜索，后来又开始做向量表征和召回，感觉就很奇妙，能够理解传统搜索和相对新潮的向量表征召回之间的关系，这两者之间的关系还是挺微妙地，互相借鉴的过程中能够产生一些火花，例如向量检索之前可以召回一些相似的、标准的query然后来检索，这样能大幅提升准确率，也一定程度降低了对模型深度的要求。（隐约感觉是时候写一篇有关模型和规则特征之间关系的文章了？）

我是叉烧，欢迎关注我！

叉烧，OPPO搜索算法工程师，主做Query理解，NLP方向。19届北科技统计学硕士（保研），17届北京科技大学信息与计算科学、金融工程双学位毕业，论文7篇，学生一作3篇，参与国家级及以上学术会议4次，优秀论文一次，国奖金。曾任去哪儿网大住宿事业部产品数据，美团点评出行事业部算法工程师。

微信个人公众号
CS的陋室

微信 zgr950123
邮箱 chashaozgr@163.com
知乎 机智的叉烧