基于“三阶段”多粒度集成的篇章级检索方法

华东师范大学　　锆德教育资讯

华東師範大學
EAST CHINA NORMAL UNIVERSITY

Global Education Alliance
We Navigate

汇报人: 刘曙

目录

1. 团队介绍

团队成员

刘曙
华东师范大学非全研究生，来自兰曼教授CubeNLP实验室[1], 目前在上海犀语担任算法工程师，主要研究金融领域的信息抽取与文档智能。

邓顺子
广州大学本科毕业，目前在锆德教育资讯担任算法工程师，主要研究留学教育领域的信息抽取与智能对话。

---

[1] http://www.cubenlp.cn

2. 问题分析

赛题任务

基于民航领域相关语料集合S，结合用户问句Q，采用信息检索相关模型与方法，返回与问句Q较相关或检索模型得分较高的N篇文章（Si, Sj, Sk等）。
选手预测结果对应格式为{question, content_key_list}。

```
{
 "question": "机场转机需要多久时间",
 "answer": [
  "cd0b3ef10ce295a292a9376e",
  "cd0b3ef10ce295a292a9376e",
  ]
}
```

title	content	label_content	key
支付方式	网上银行支付	{"h1_0":[{"title":"支付方式","texts"	7a50d6e2ccbbd66817105da5342dd7fb
国内客票使用条件	»（一）	{"h1_0":[{"title":"","texts":[{"tex	04188a33147e665b172a48183df2984a
行李规定	一、随身携带物品	{"h2_0":[{"title":"一、随身携带物品"	4c20b81a35d0e2668ff5425c386b04be
旅客乘机健康指南	感谢您选择中国南方航空！	{"h1_0":[{"title":"旅客乘机健康指南"	ef1c1574c4025bf25040fafd9d71f810
锂电池安全运输提示	当您携带手机、手提电脑、摄像机等含锂电池电子设备	{"h1_0":[{"title":"锂电池安全运输提	cad141ea827438157d8ca1f3b0e52a6a
美国境内航班延误应对	美国境内航班停机坪长时间延误应对计划	{"h4_0":[{"title":"美国境内航班停机	724be0fa8a2ab4dc2f3b692c6f878362
行程单规定	一、行程单（报销凭证）的领取说明	{"h1_0":[{"title":"行程单规定","text	20d17dfab06979fa1d30bc5574f6221f
旅行证件提醒	有效的乘机证件：	{"h1_0":[{"title":"旅行证件提醒","te	ea77dc8107aafc7c12cae74f08f1e3ff
南航提前选座产品购买须知	第一条产品概念	{"h1_0":[{"title":"南航提前选座产品	b3376b4821fbceefe5d20f6dcc79eb32
额外行李产品购买须知	第一条产品概念	{"h1_0":[{"title":"额外行李产品购买	bf0b06af8fb5e514a05a3d4d2f6eede3
空地联运产品使用须知	空地联运产品使用须	{"h3_0":[{"title":"","texts":[{"tex	271a84486069759bd600273b492b1588
国内临时团保证金规则	一、适用对象	{"h1_0":[{"title":"国内临时团保证金	1e3e442cd3a440039a9a078643d9d1f8
国内团队销售管理规定	一、团队旅客人数的计算规定	{"h1_0":[{"title":"国内团队销售管理	e54c36502777d4fab1c414ea7cdac95c
国际团队销售管理规定（暂行）	一、适用范围	{"h2_0":[{"title":"一、适用范围","te	0f593bf1b8c39ce53fcedf18f34c7fe4
机上延误应急预案	1前言	{"h2_0":[{"title":"1前言","texts":[	29a81bf7f69bc2c606960568ad97d515
南航北京大兴产品权益说明	●优惠内容	{"h1_0":[{"title":"南航北京大兴产品	6c30f7a91ed2729e1e4bdc2bfe00a879
隐私通知	发布生效日期：2018年5月24日	{"h2_0":[{"title":"一、南航的业务是	185ffba3da1db07db22b02aeecedf5d6
cookie政策	最新更新截至【2018年5月24日】	{"h1_0":[{"title":"cookie政策","tex	662871f24fe58d09c0812e8d0344le47
网上订票服务协议	网上订票服务协议	{"h3_0":[{"title":"","texts":[],"h4	2ca6335f900d29e8639d64d891e93c18
南航网上座位预选协议	南航网上座位预选协议	{"h3_0":[{"title":"","texts":[],"h4	035f1371ca2ceeb0c083cd1ca821cdab
自动值机服务须知与免责声明	自动值机服务须知与免责声明	{"h2_0":[{"title":"自动值机服务须知	13914d06414e5076f561a01a1d6cfa9d
非南航航班使用条件	厦航国内航空运输票价适用条件	{"h4_0":[{"title":"厦航国内航空运输	f3adc30a1614c051f562dc223cfcd469
休闲旅游产品服务条款	“机票加酒店”休闲旅游产品服务条款及规则	{"h1_0":[{"title":"","texts":[]}],"[	738ed158623150782afa292e24e07f39

输入： 测试问题

输出： {测试问题，答案所在文章的key列表}

难点分析

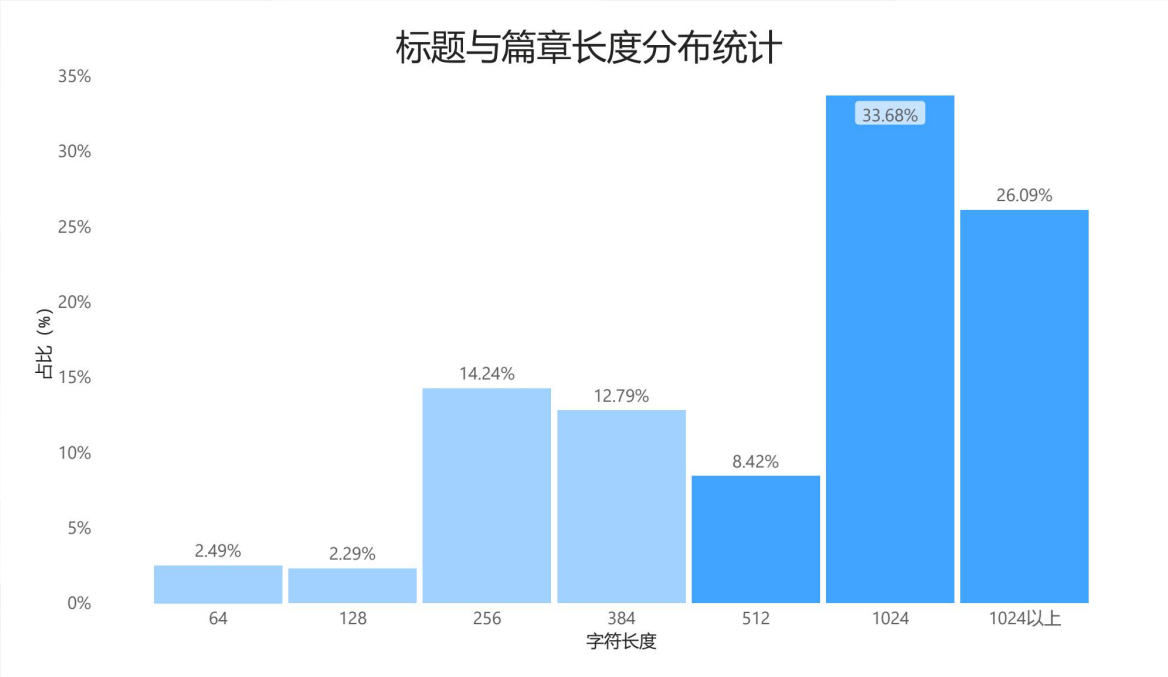
- **长文本篇章**

总数	平均长度	最小长度	最大长度
3319	12	2	55
25%	50%	75%	90%
8	10	13	20

query字符长度分布[1]

篇章总数	平均长度	最小长度	最大长度
962	916	8	15501
25%	50%	75%	90%
306	589	1076	1858

标题与篇章长度分布[2]



注1：数据来源于train.txt与valid.txt,已剔除无法回答的query。
注2：数据来源于section.xlsx，包含少量detail信息遗漏（约占1%）。

难点分析

- **细粒度排序**

输入：机场转机需要多长时间

输出：

一票到底行李免提南航经乌鲁木齐中转一票到底业务范围：1）当日国内中转旅客2）隔日国内中转旅客，在始发站一次性办理疆内转疆外目的地登机牌和行李牌，在乌鲁木齐无需提取行李。.....5.衔接时间机转机服务南航中转为转机时间大于30分钟小于50分钟的国内旅客提供“机转机”服务，即舱门口接机、免二次安检、全程引导。6、行李委托无忧急转对于因前序航班晚到，行李来不及转机的旅客，如旅客授权，南航中转可提供行李委托服务，即旅客签订《行李委托书》后可乘坐原定南航航班先成行，托运行李由南航中转协助转运后续航班，旅客后续自行前往机场提取行李。7、中转信息全天守候提供24小时中转业务咨询服务。热线电话：0991-3800215

query对应ans候选数量	1	2	3	4	5
每类所占数量	561	832	161	276	54
每类所占数量(%)	30%	44%	9%	15%	3%

文章包含的detail数量分布[1]

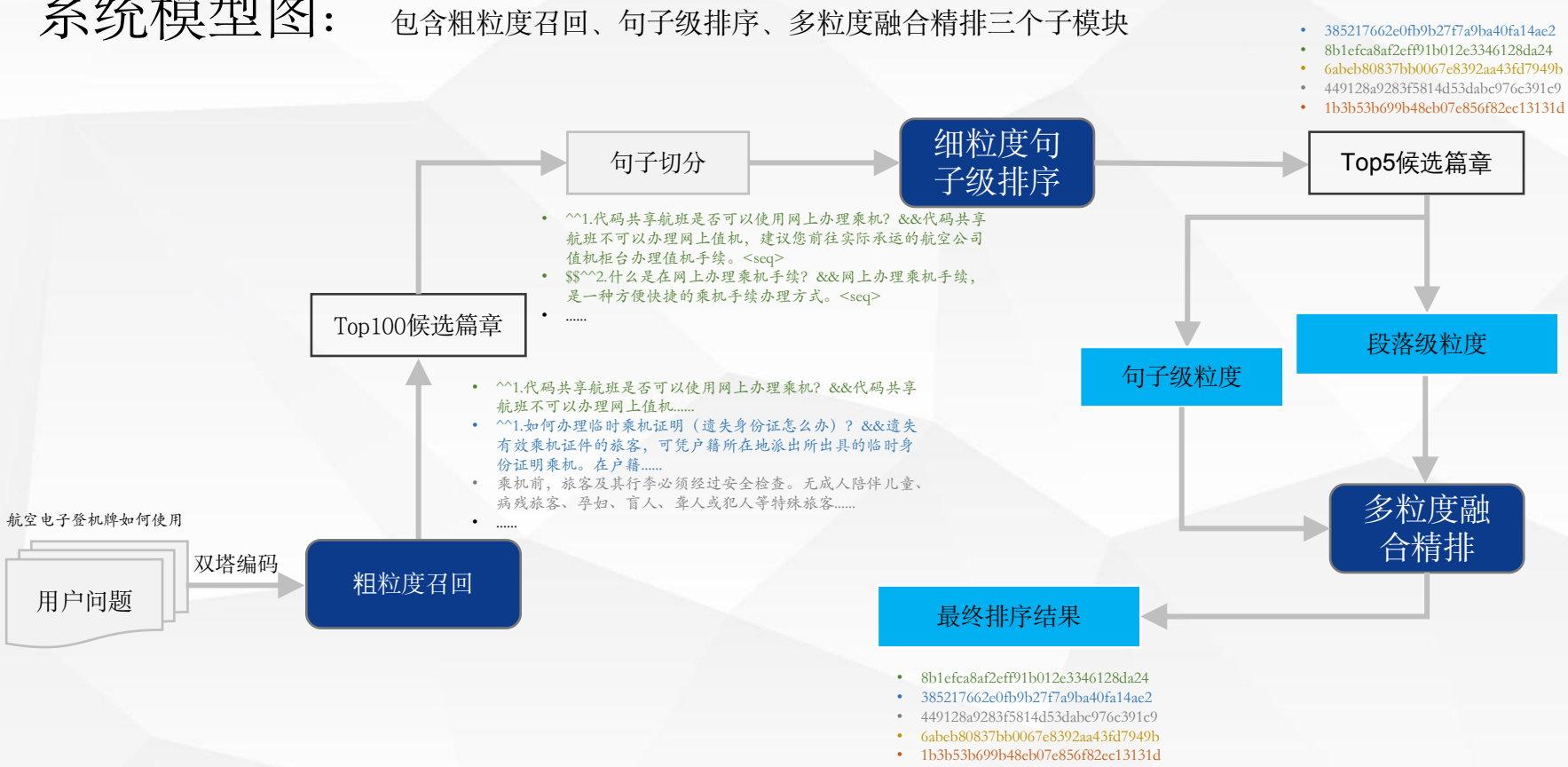
注1：数据来源于train.txt与valid.txt,已剔除无法回答的query。

3. 算法模型

3.1模型概述

系统模型图：

包含粗粒度召回、句子级排序、多粒度融合精排三个子模块



3.2 粗粒度召回

粗粒度召回

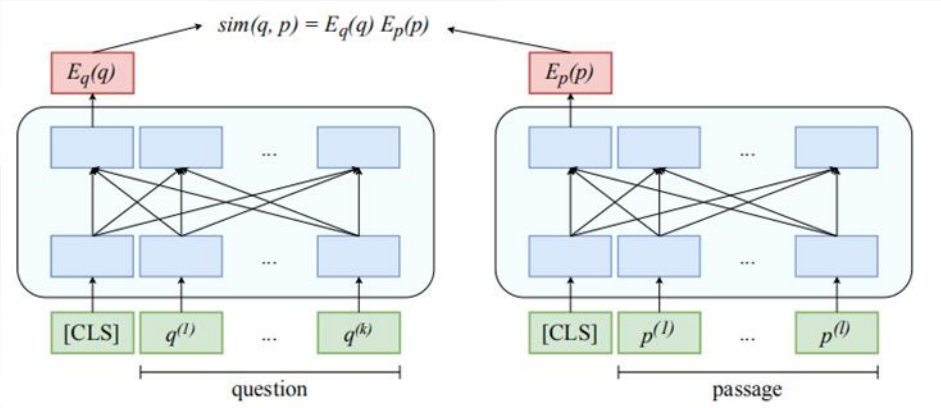
检索问题，分为两个阶段，召回-精排
召回是为了快速检索与问题相关的候选答案，缩小选择范围
本模块采用了经典的双塔模型，召回问题top100候选篇章

解决:
双塔模型
采用百度RocketQA的Dureader模型

问题：航空电子登机牌如何使用

输出：


```
{
  "question": "航空电子登机牌如何使用",
  "answer": [
    "8b1efca8af2eff91b012e3346128da24",
    "cad9cb637782266f54df1e3c77257acb",
    "449128a9283f5814d53dabc976c391c9",
    "385217662e0fb9b27f7a9ba40fa14ae2",
    "895e6c32d5d589c894f0f3d188b97a25",
    "6abeb80837bb0067e8392aa43fd7949b",
    "8f4f560a6087c88394afc18db6b04448",
    "1d0f544547fb6fc33ce298befca10040",
    "6516568c25442323e08a333ac3eaecf7",
    ......
  ]
}
```

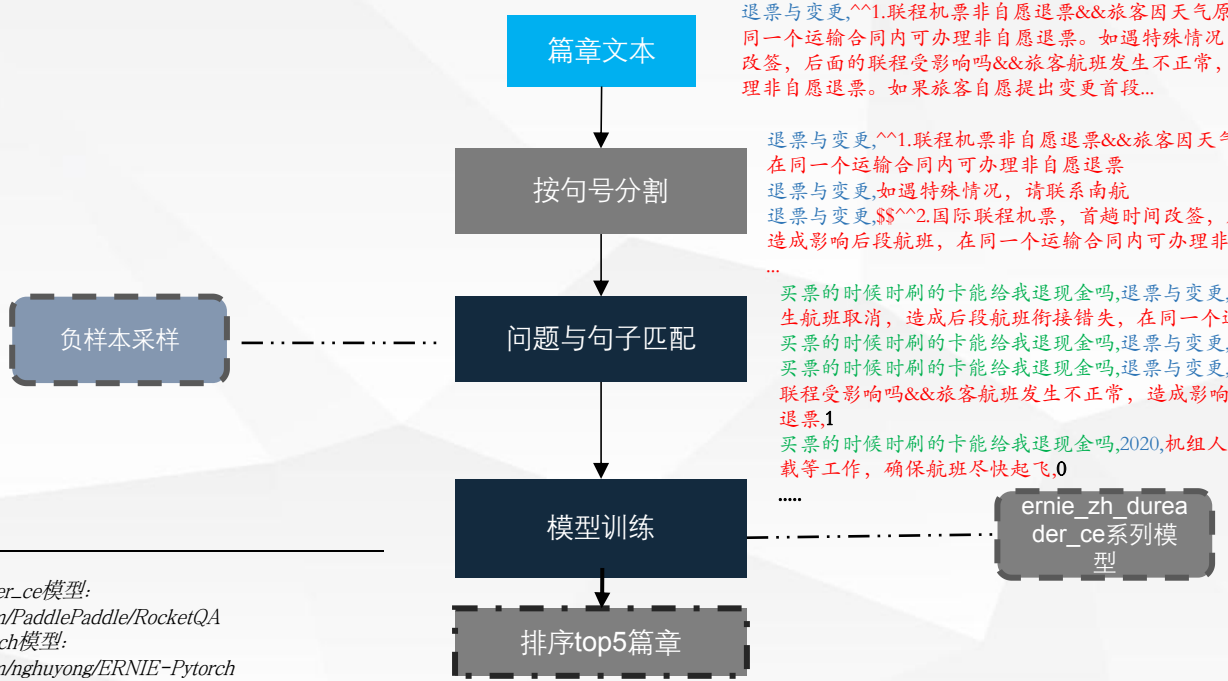
Dual-model

双塔模型训练代码：https://github.com/PaddlePaddle/RocketQA/tree/main/rocketqa

3.3 细粒度句子级排序

细粒度句子级排序

将篇章按照句号进行分割并基于交互模型训练，获取篇章的细粒度信息，对测试question的候选top100篇章进行排序，精排模型保证了top5的召回范围。



ernie_zh_dureader_ce模型:
https://github.com/PaddlePaddle/RocketQA
paddle模型转torch模型:
https://github.com/nghuyong/ERNIE-Pytorch

负样本采样

思路：

在实际检索场景中，question对应的答案候选项占比较少，因此为了保证训练与预测的一致性，同时拉近quesiton与正样本的距离，拉远question与负样本的距离，需要采取负样本采样技术。

方法：

- Random：从所有篇章语料中随机抽样
- BM25：使用BM25检索相似，但不包含答案的篇章
- Gold：训练集中其他问题出现过的答案篇章

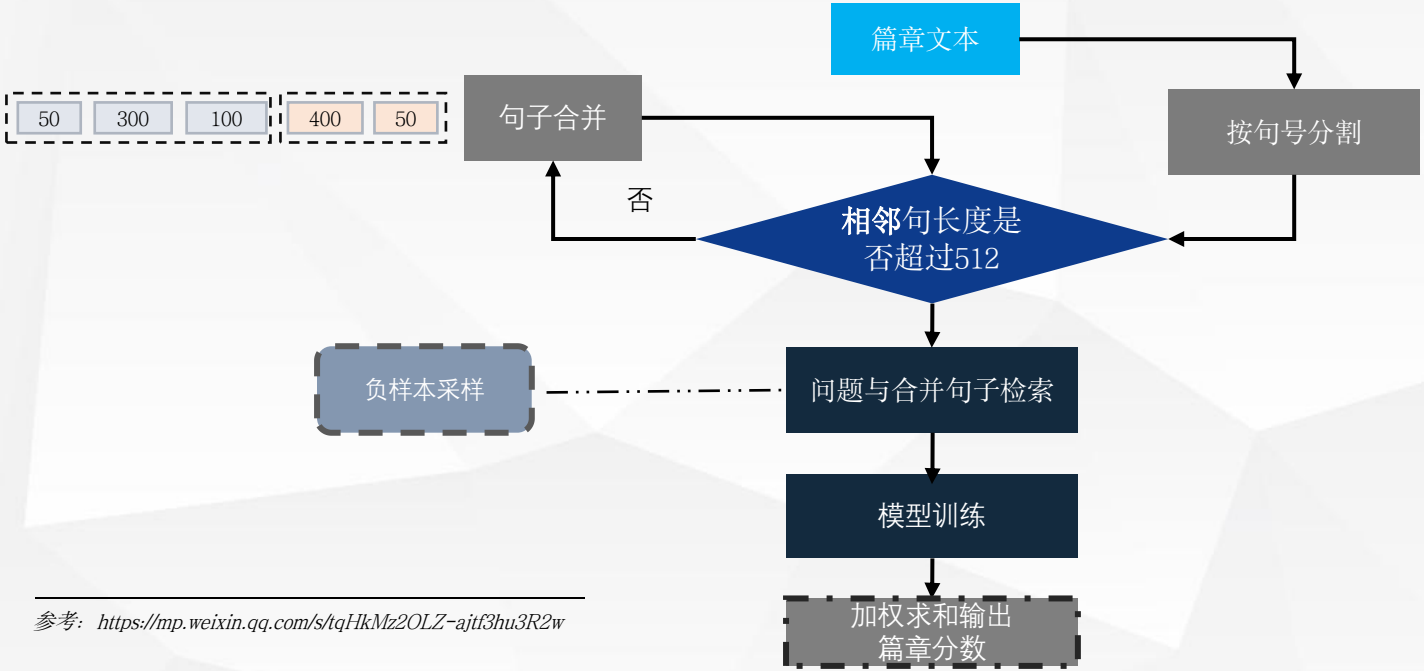
p@1	score	p@3	p@5	方法
0.6772	0.8893	0.905	0.9647	Random：正负样本比例1：5
0.7832	0.9239	0.9354	0.9732	Random：正负样本比例1：100
0.8331	**0.9346**	0.9379	0.9732	Gold+Random：正负样本比例1：100
0.8514	**0.9404**	0.9452	0.9732	Gold+Bm25：正负样本比例1：100

论文参考：https://arxiv.org/abs/2004.04906

3.4 多粒度融合精排

多粒度融合精排

将篇章按照句号分割后，分句后将句子聚合形成不超过512字的段落，减少离散的语义信息。
获取篇章的段落信息，对测试question的候选top5进行排序。

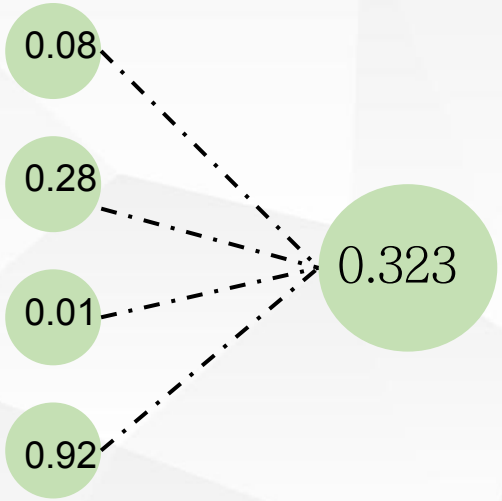


参考：https://mp.weixin.qq.com/s/tqHkMz2OLZ-ajtf3hu3R2w

篇章分数计算

方法：

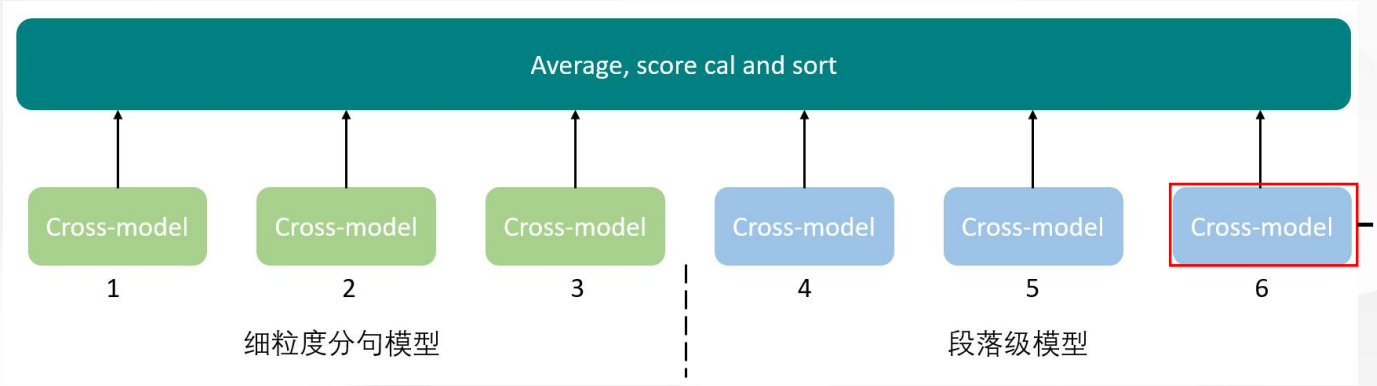
篇章被分割成段落，question会与每个段落分别计算获取概率分数，为了得到整体篇章的分数，因此需要将所有段落得分进行加权平均计算，获取整体篇章分数。

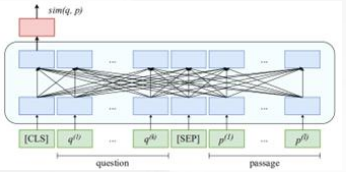


论文参考：https://arxiv.org/format/2104.08253

多粒度融合ensemble

将训练好的三个细粒度句子级模型和三个段落级模型进行ensemble，logits求和平均，不同粒度篇章信息的聚合





Cross-model

p@1	score	p@3	p@5	方法
0.8514	**0.9404**	0.9452	0.9732	未ensemble
0.866	**0.9456**	0.9525	0.9732	多粒度ensemble

3.5 工程落地

线上推理

- docker封装，开箱即用
- 将候选篇章缓存成向量矩阵，加速召回阶段检索
- 接口采用fastapi，即可以采用requests请求进行测试，也可使用内置的swagger网页进行交换式测试

效果：

- 推理资源，GPU占用2G，CPU内存占用1G
- 单条数据平均推理速度：7s

网页测试界面示意图

4. 思考与展望

思考与展望

- 相对与段落级别的查询，篇章级正样本分句后的存在伪正样本。
- 目前检索模型，训练与预测存在一定gap，训练使用二分类模型，hard分类。预测则仅输出logits大小，是soft排序。
- 采用多粒度融合方式，导致推理速度较慢。

THANKS