基于“三阶段”细粒度精排的段落级检索方法

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：邓顺子

目录

1. 团队介绍

团队成员

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。

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。

---

[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篇段落（Si, Sj, Sk等）。
选手预测结果对应格式为{问题，文章key,文章所对应的detail段落}。

```
{
 "question": "买票的时候时刷的卡能给我退现金吗",
 "answer": [{
   "content-key": "57997c7ea0308171ca3be0cba3282df6",
   "detail": ["h1_0", 0, "h2_7", "text"]
 }, {
   "content-key": "e68660af5d233eaa37ce2f2241f6b042",
   "detail": ["h1_0", 0, "h2_11", "text"]
 }]
}
```

content-key	detail	title	text
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'title']	支付方式	支付方式
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'texts', 0, 'text']	支付方式	网上银行支付
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'texts', 1, 'text']	支付方式	您需拥有银行卡，并
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'texts', 2, 'text']	支付方式	开通网上银行
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'texts', 3, 'text']	支付方式	支付平台
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'texts', 4, 'text']	支付方式	南航官网支持银联在线支付、支付宝、财付通、汇付天下、易宝信用
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'texts', 5, 'text']	支付方式	无需开通网上银行
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'texts', 6, 'text']	支付方式	银联电话支付
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'texts', 7, 'text']	支付方式	银联电话支付是一种安全快捷的电话支付方式，您无需开通网上银行
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'texts', 8, 'text']	支付方式	国际卡支付
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'texts', 9, 'text']	支付方式	南航英文官方网站和各海外站点为您提供各种国际支付方式，支持中
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'texts', 10, 'text']	支付方式	提示：
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'texts', 11, 'text']	支付方式	1、为了保障您的用卡安全，我们可能会需要您进行支付卡验证。如
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'texts', 12, 'text']	支付方式	2、南航不会向您收取任何支付手续费或汇率费用，但您的发卡机构
7a50d6e2ccbbd66817105da5342dd7fb	['h1_0', 0, 'texts', 13, 'text']	支付方式	3、为了提供安全的网上交易环境，本网站针对Visa/Mastercard付
04188a33147e665b172a48183df2984a	['h1_0', 0, 'texts', 0, 'text']	国内客票使用条件	»（一）南航北京大兴机场进出港国内客票使用条件规定（适用于乘
04188a33147e665b172a48183df2984a	['h1_0', 0, 'texts', 1, 'text']	国内客票使用条件	南航北京大兴机场进出港国内客票使用条件规定（适用于乘机日期为
04188a33147e665b172a48183df2984a	['h1_0', 0, 'texts', 2, 'text']	国内客票使用条件	»（二）南航国内运输散客票价使用条件规定（适用于2021年12月20
04188a33147e665b172a48183df2984a	['h1_0', 0, 'texts', 3, 'text']	国内客票使用条件	南航国内运输散客票价使用条件规定（适用于2021年12月20日（含）
04188a33147e665b172a48183df2984a	['h1_0', 0, 'texts', 4, 'text']	国内客票使用条件	»（三）国内运输散客公布票价使用条件总则（适用于2020年10月25

输入： 测试问题

输出： {测试问题，答案所在文章的key列表，以及文章中对应的detail段落}

难点分析

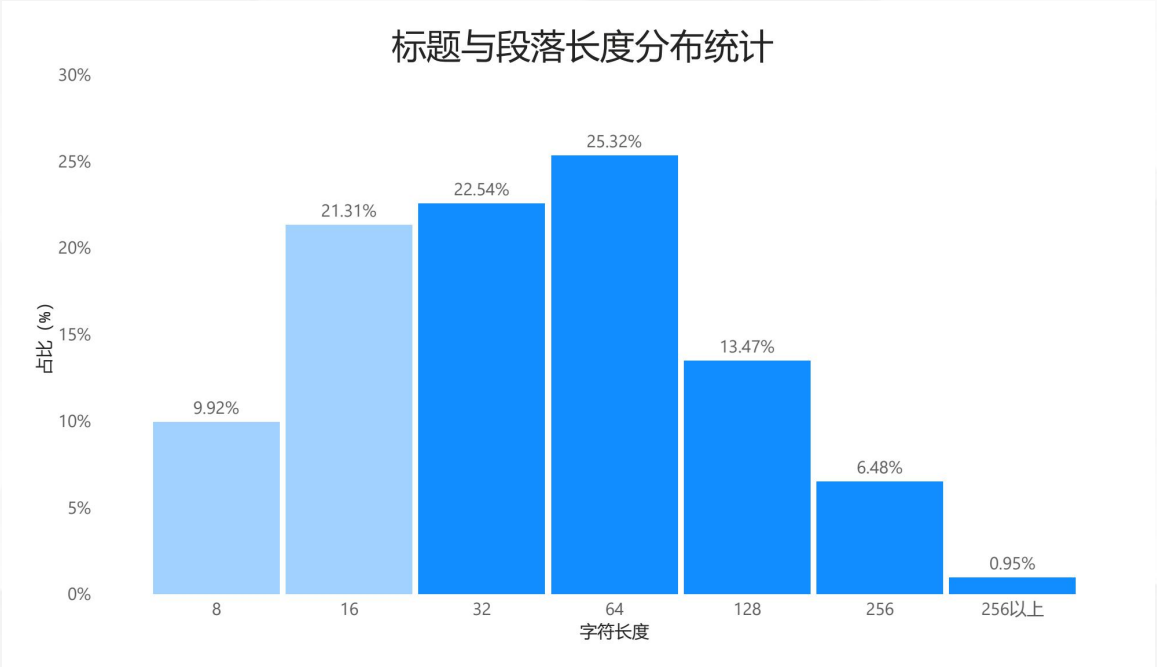
- **细粒度段落文本**

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]

段落总数	平均长度	最小长度	最大长度
19952	48	3	2519
25%	50%	75%	90%
14	29	56	110

标题与段落长度分布[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

- **细粒度多段落排序**

段落总数	平均个数	最小个数	最大个数
962	21	1	315
25%	50%	75%	90%
9	15	22	37

文章包含的detail数量分布[1]

段落总数	平均个数	最小个数	最大个数
3319	8	1	81
25%	50%	75%	90%
2	3	8	15

query包含的answer数量分布[2]

输入：猫咪随身携带上飞机

输出：
回复1：小动物
回复2：（1）作为行李运输的活体小动物是指家庭饲养的狗、猫、鸟或者其他玩赏宠物；
回复3：（2）如动物的体形过小（例如：乌龟、鼠类、观赏鱼等）或体形过大（笼体包装超过最大体积限制）或对运输安全可能造成危害，以及野生动物和具有形体怪异或者易于伤人等特性的动物如蛇等，不属于活体动物范围，不能作为行李运输；
回复4：（3）不收运短鼻猫、短鼻犬类动物作为行李运输；
回复5：短鼻猫：缅甸猫、异国短毛猫（加菲猫）、波斯猫、喜马拉雅猫、英国短毛猫等。
回复6：短鼻犬：猴面犬、波士顿梗犬、拳师犬（所有品种）、布鲁塞尔葛里芬犬、斗牛犬/牛头犬（所有品种）、斗牛梗犬、巴哥犬/哈巴犬、卡斯罗犬、松狮犬、波尔多犬、英国玩具猎狐犬、日本狮子犬、骑士查理王猎犬、拉萨犬、獒犬（所有品种）、北京犬、西班牙加纳利犬、沙皮犬、西施犬、西藏猎犬、恶霸犬、洋基犬、巴基斯坦牛头梗、斯塔福德梗等。
……

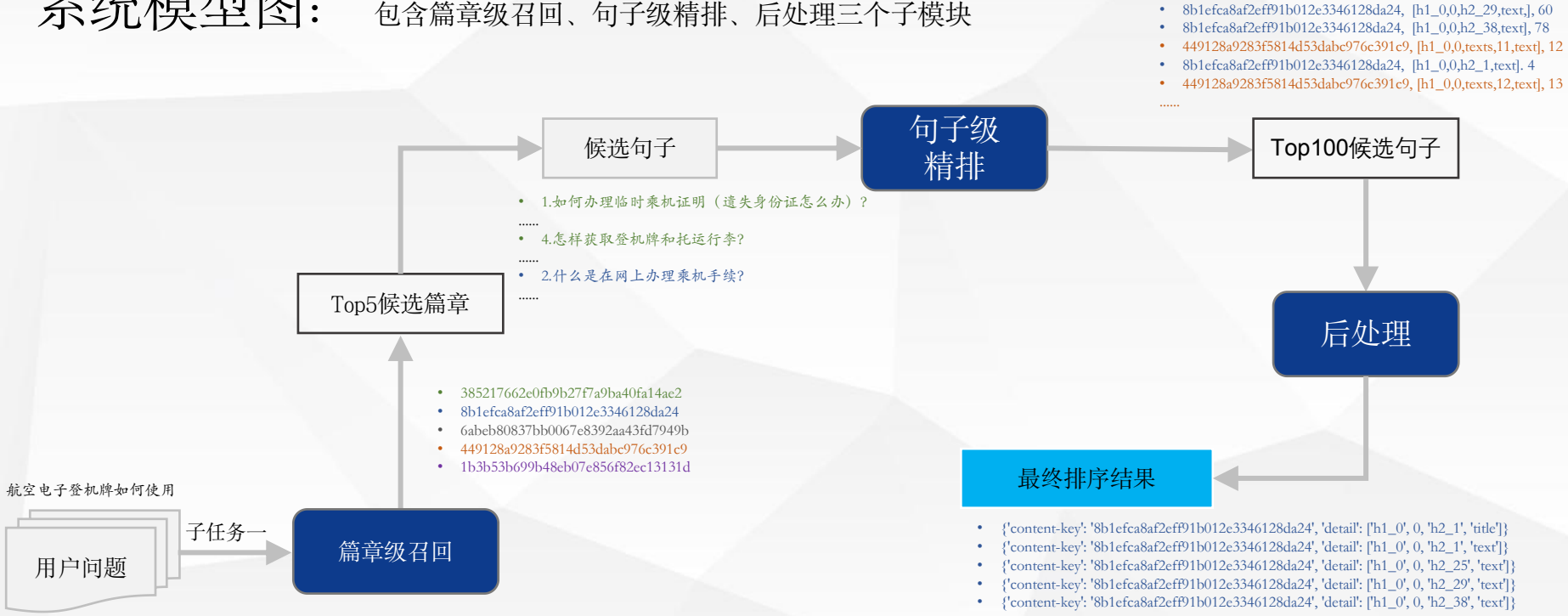
注1：数据来源于section.xlsx，包含少量detail信息遗漏（约占1%）。
注2：数据来源于train.txt与valid.txt,已剔除无法回答的query。

3. 算法模型

3.1模型概述

系统模型图：

包含篇章级召回、句子级精排、后处理三个子模块



3.2 篇章级召回

篇章级召回

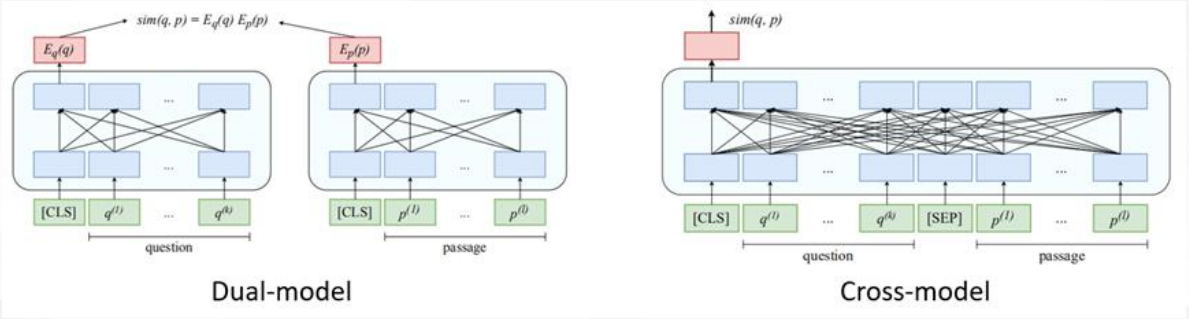
基于子任务一进行篇章级召回，缩小候选段落范围
本模块基于召回-精排的pipeline方式，采用经典的双塔模型和交互模型
基于篇章多粒度信息对整体打分排序，获取top5的结果

解决:
召回-精排pipline方式
融入多粒度信息对篇章整体打分

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"
  ]
}
```

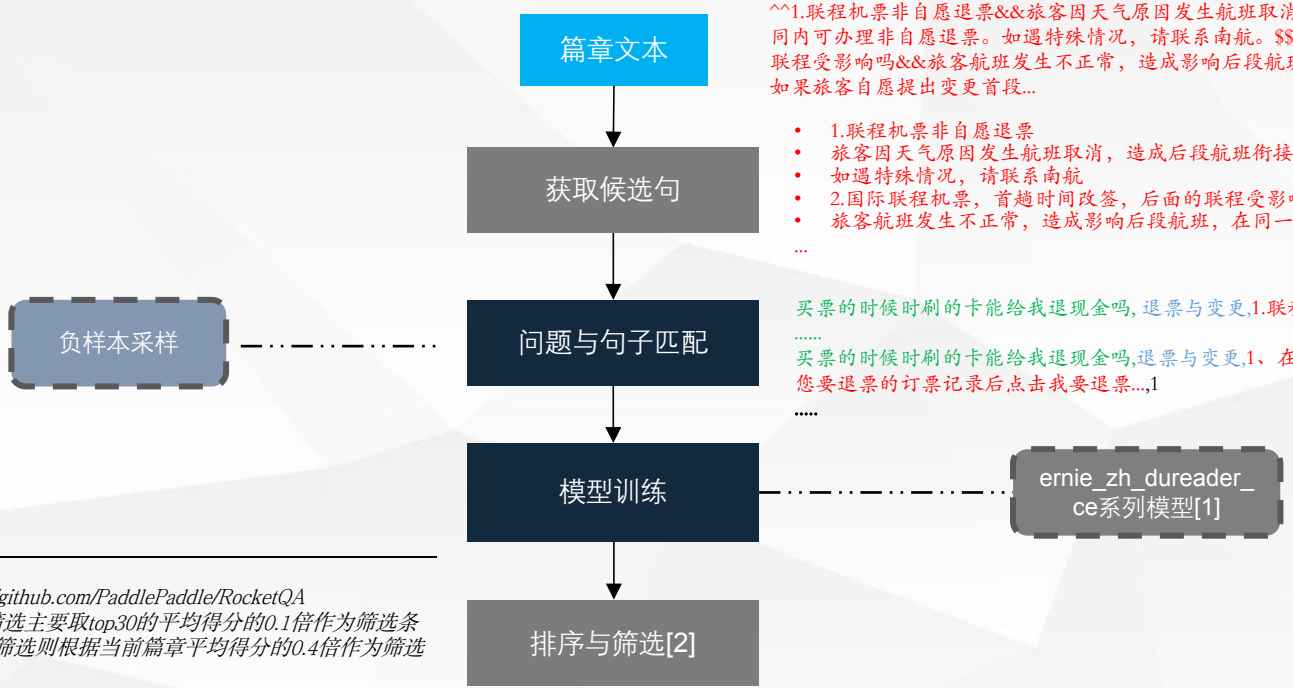



模型训练代码：https://github.com/PaddlePaddle/RocketQA/tree/main/rocketqa

3.3 细粒度句子级精排

句子级排序与二次筛选

根据篇章key获取其关联的句子信息，对测试question的top100候选句子进行排序，经过全局筛选与篇章内筛选，使得模型可以输出较高质量的句子。



注1：https://github.com/PaddlePaddle/RocketQA
注2：全局筛选主要取top30的平均得分的0.1倍作为筛选条件，篇章内筛选则根据当前篇章平均得分的0.4倍作为筛选条件。

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，缩小question与负样本的距离，需要采取负样本采样技术。

方法：

- Random：从所有篇章包含的句子语料中随机抽样
- Gold[1]：正样本中所有出现过的篇章，扣除当前query的篇章后，从中随机抽取50条。
- Corr：若样本的answer之间存在交集，则answer构成的并集，里面的篇章互为相关。

p@1	score	p@3	p@5	方法
0.5918	0.6284	0.5966	0.6618	Random(1:10)：句子正负比1:200
0.7954	0.834	0.8368	0.8477	Corr(1/4[2])：句子正负比：1:44
0.8051	**0.8715**	0.8782	0.894	Corr(1/4)+Gold(1/4)：句子正负比1:97
0.7881	0.8532	0.8611	0.8745	Corr(1/2)+Gold(1/2)：句子正负比1:196

注1：论文参考：https://arxiv.org/abs/2004.04906
注2：1/4表示每个篇章的句子随机抽取25%。

3.4 后处理

篇章重排与句子排序

根据原篇章排序对篇章内的句子得分进行降分处理（每个排序降低0.05），同时结合篇章句子个数，篇章内句子总分，篇章内句子字符长度因素对篇章进行重排。之后结合篇章内句子索引以及句子得分输出最终句子。

$$D_{score} = (1 - 0.05 * D_i) * (S_{mean} + \frac{S_{sum}}{200} - \frac{SeqLen}{20000})$$

D_{score}：篇章分数。

D_i：原篇章顺序，i为索引。

S_{mean}：篇章内句子平均分。

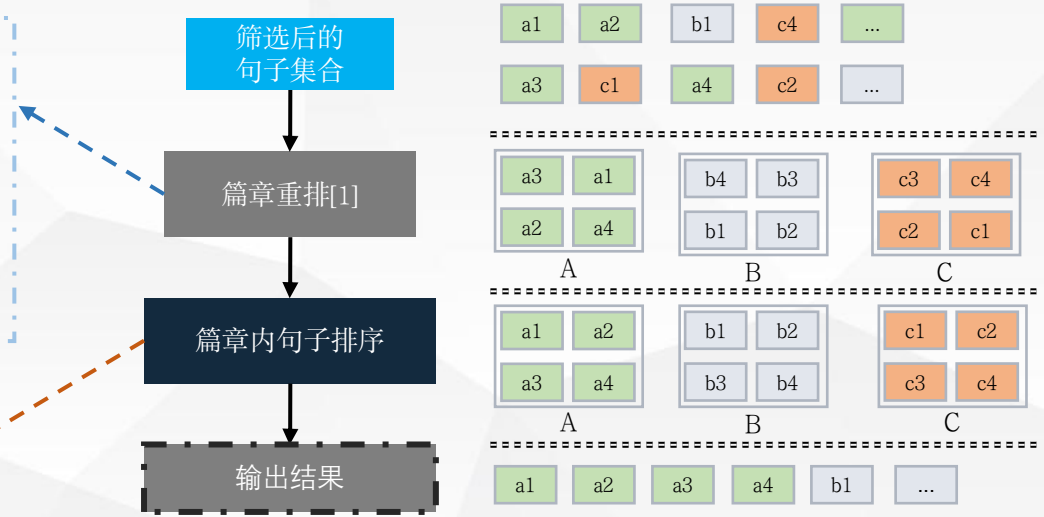
S_{sum}：篇章内句子总分。

$SeqLen$：篇章中句子字符长度之和。

$$Score = str(int(10 * (1 - Score)) + str(index(4))$$

Score:句子分数(倒序)。

index(4):句子在篇章中索引，按4位长度向左补零。



注：篇章重排用于防止任务一带来的误差累计至任务二，造成误差放大。

3.5 工程落地

线上推理

- docker封装，开箱即用

- 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3.7G，CPU内存占用2.9G
- 单条数据平均推理速度：0.37s

网页测试界面示意图

4. 思考与展望

思考与展望

- 目前模型输入高度依赖于任务一，导致任务一的预测结果与预测效率会对任务二造成较大影响，能否脱离任务一单独建模?

- 拆分后的句子短文本较多，分开训练无法学习到整体篇章信息，导致相对较短的文本预测效果较差，能否利用篇章信息融入短文本提升模型效果?

THANKS