ES+MRC方案

总体思路：给定了问题，context和答案，我们可以训练MRC（机器阅读理解）模型。该模型可以从context定位出答案的起始和结束位置。
预测时，由于未给出context，我们利用ES（elastic search）根据问题对context排序选择出top1的context。并利用训练的MRC从中抽取出答案片段。

训练

数据预处理

首先利用 src/convert_jsonl.py将训练数据转换为json格式，data/train.jsonl.

{"qid": "47a41a03966431739257ef215cdc1caa", "context": "工业和信息化部组织开展负压救护车重点生产企业督导检查 2020年2月4日，为做好新型冠状病毒感染的肺炎疫情防控物资保障工作，加强负压救护车生产质量检查工作，工业和信息化部装备工业一司会同国家卫健委、国家药监局相关司局赴北京北铃专用汽车有限公司进行督导检查，重点了解企业生产及检测过程、产品质量和生产一致性保障能力、安全生产工作等情况。 与此同时，工业和信息化部装备工业一司委托河南、江苏、山东、天津等省（市）工业和信息化主管部门分别对辖区内生产负压救护车、负压设备等关键零部件的企业开展督导检查，了解并协调解决企业生产过程中遇到的困难和问题，确保产品质量并按时交付，为疫情防控工作做出积极贡献。", "query": "工业和信息化部到哪家企业进行督导检查？", "answer": {"text": "北京北铃专用汽车有限公司"}}

利用src/preprocess.py从训练数据中计算出答案的起始结束位置，存入到 data/train_answer.jsonl 文件。

{"qid": "47a41a03966431739257ef215cdc1caa", "context": "工业和信息化部组织开展负压救护车重点生产企业督导检查 2020年2月4日，为做好新型冠状病毒感染的肺炎疫情防控物资保障工作，加强负压救护车生产质量检查工作，工业和信息化部装备工业一司会同国家卫健委、国家药监局相关司局赴北京北铃专用汽车有限公司进行督导检查，重点了解企业生产及检测过程、产品质量和生产一致性保障能力、安全生产工作等情况。 与此同时，工业和信息化部装备工业一司委托河南、江苏、山东、天津等省（市）工业和信息化主管部门分别对辖区内生产负压救护车、负压设备等关键零部件的企业开展督导检查，了解并协调解决企业生产过程中遇到的困难和问题，确保产品质量并按时交付，为疫情防控工作做出积极贡献。", "query": "工业和信息化部到哪家企业进行督导检查？", "answer": {"text": "北京北铃专用汽车有限公司", "span": [109, 120]}}

MRC训练

我们利用transformers中的bert模型和中文预训练模型参数。为了复用squad数据的预处理，我们把 data/train_answer.jsonl 文件转换为squad数据格式data/train_squad.json。
训练参数如下

```
python src/run_squad.py --model_type bert --model_name_or_path bert-base-chinese
--do_train --train_file data/train_squad.json --output_dir debug_squad_v1/
```

预测

ES检索

对context文件建索引，利用测试集问题检索context，存储在data/query_docids_v1.csv

数据预处理

类似于训练数据处理同样把测试数据转化为json合适，再转化为squad数据格式。
执行以下测试命令

```
python src/run_squad.py --model_type bert --model_name_or_path debug_squad_v1/checkpoint-6000/ --do_eval --predict_file data/test_squad.json --output_dir debug_squad_v1/
```

生成的测试文件利用src/format_submission.py转换成可提交格式。

方案总结分析

该方案分数大约在0.37左右
性能可能的改进点在于：

- 改进context的检索，目前只是基于BM25分数，引入排序模型会更好。
- 改进MRC训练，目前训练BERT模型的参数未经调整，此处可以提升。
- 可以将利用MRC工多个context抽取答案，然后对候选答案再次排序。