2019中国法研杯
阅读理解赛道参赛报告

队伍：like cxk

成员：丘德来，鲍亮　　指导老师：刘康，廖祥文

所属单位：中科院自动化所模式识别国家重点实验室

福州大学信息检索课题组

云知声智能科技股份有限公司

赛题分析

经审理查明,原、被告于2010年11月5日登记结婚,婚生子王凯翔(现改名为那8)于2012年2月10日出生2013年1月14日,经本院主持调解,双方当事人就抚养权自愿达成如下协议:婚生子王凯翔由葛x1抚养,王0从2013年1月起每月承担抚养费10000元至王翔凯独立生活之日止2012年9月22日,原告王0与他人孕育一男孩离婚后原、被告双方均已重组家庭,现原、被告双方都有稳定的工资收入,原告王0之妻没有固定收入来源,孕育两个男孩,居住于原告王0父母房屋,被告葛x1之夫有收入来源,带有一女,一家×××住×××,现就读于准格尔旗民族幼儿园以上事实由原、被告陈述及原告出示的工资收入及存款证明、葛x1、那8常驻人口登记卡、出生医学证明在案予以证实

问题一：双方约定王0每月支付多少抚养费	**回答一**：10000元
问题二：被告葛x1的继女目前就读于何处?	**回答二**：准格尔旗民族幼儿园
问题三：婚生子王凯翔出生日期为何时?	**回答三**：2012年2月10日
问题四：原、被告双方是否均已重新组成家庭?	**回答四**：YES
问题五：王翔凯是否有意愿同原告王0一起生活?	**回答五**：UNK

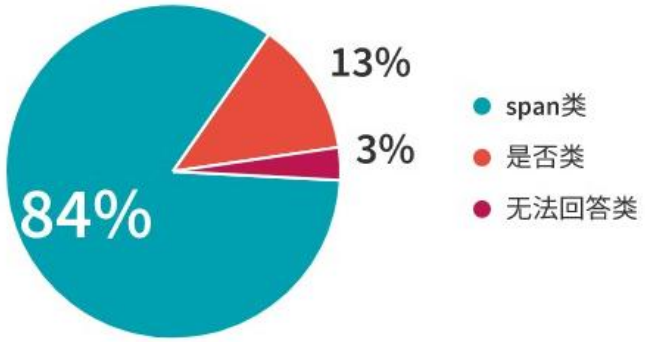
https://github.com/china-ai-law-challenge/CAIL2019/

三个难点：

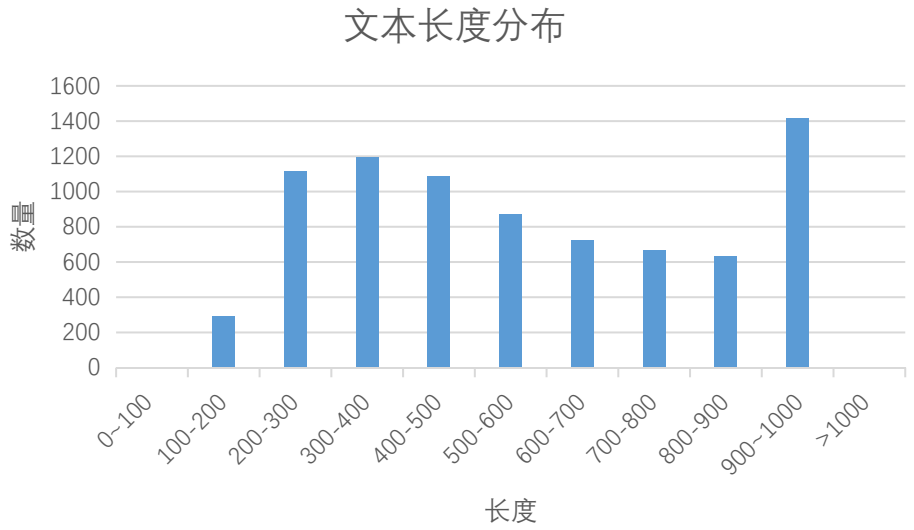
- 混合两种不同类型的法律文书
- 参考SQuAD2.0，加入了无法回答的问题
- 参考CoQA数据集，包含了是否类（YES/NO）问题

数据集分析

两种文书的数量



三种答案类型的问题占比



分析：

- 训练集的问题组成以片段抽取为主，同时包含了5000+的YES/NO类问题，以及1000+的无法回答类型的问题，需要一个合理的方案处理不同类型的问题。
- 文书长度较长，超过50%的文书的长度在500以上，模型设计上要考虑长文本的相关问题。

解决方案

- **不同类型问题的联合学习**

设计一个**答案片段抽取+YES/NO分类+无法回答分类**三个任务多任务学习的端到端模型，对不同类型的问题进行统一学习。

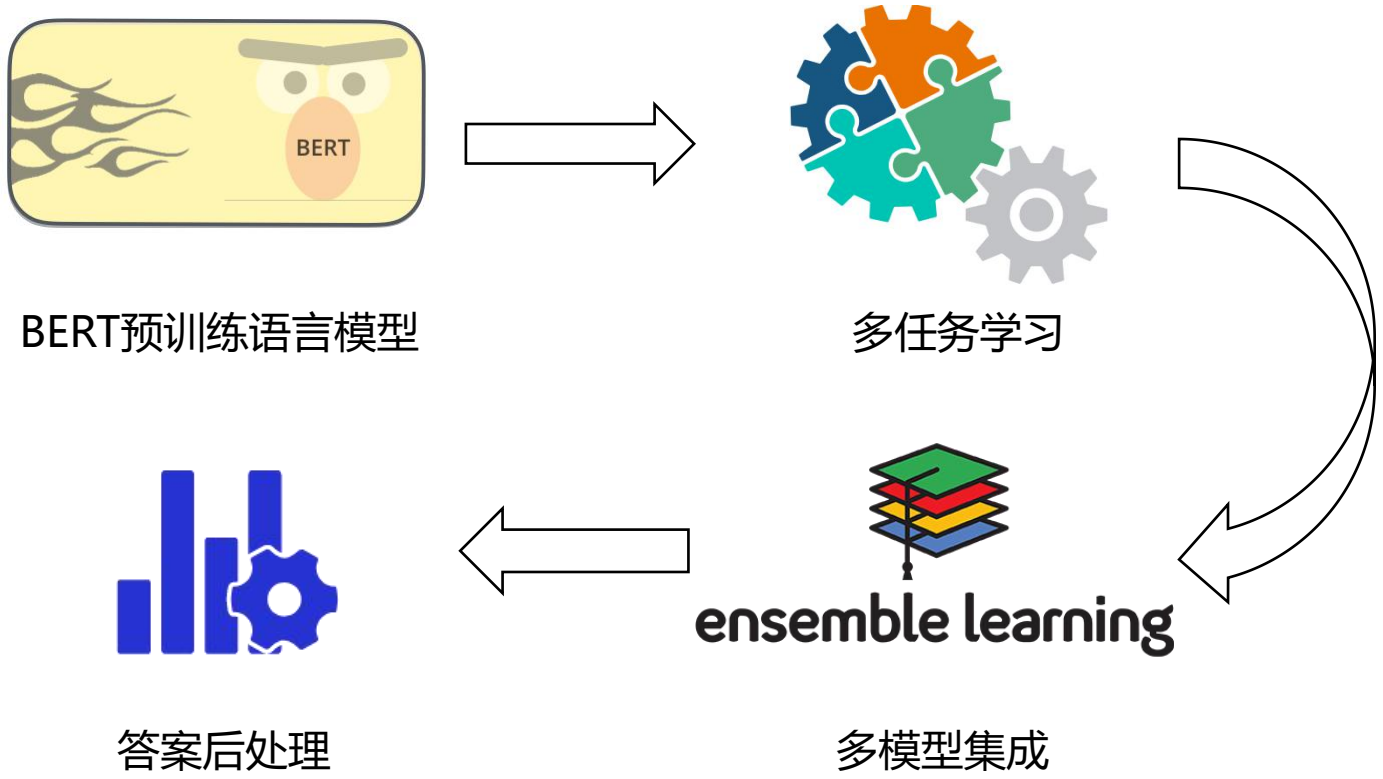
- **长文本问题**

借鉴BERT官方放出的用于SQuAD数据集的fine-tune方案中预处理的思路，即在数据预处理时利用滑动窗口方法将长文本切割成多个doc_span，并且对于出现在多个span中的单词在后续计算得分时以该单词具有“最大上下文”的doc_span为准。

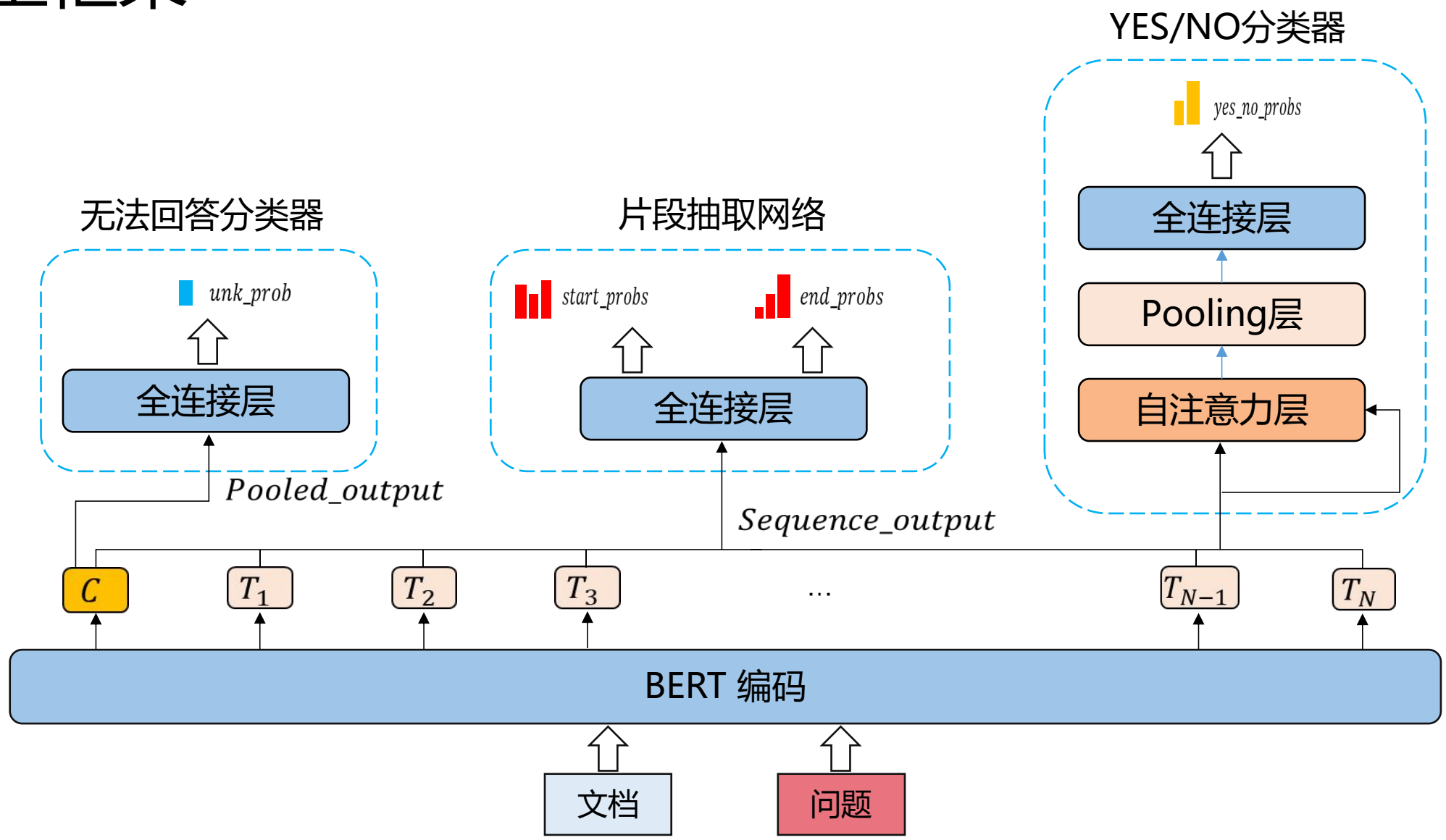
- **后处理**

深入分析数据集，我们发现一部分问题存在一定的规律或是模型预测的答案可以进一步修正，所以在第二阶段我们在整体模型结构中加入了后处理模块，以进一步提升性能。

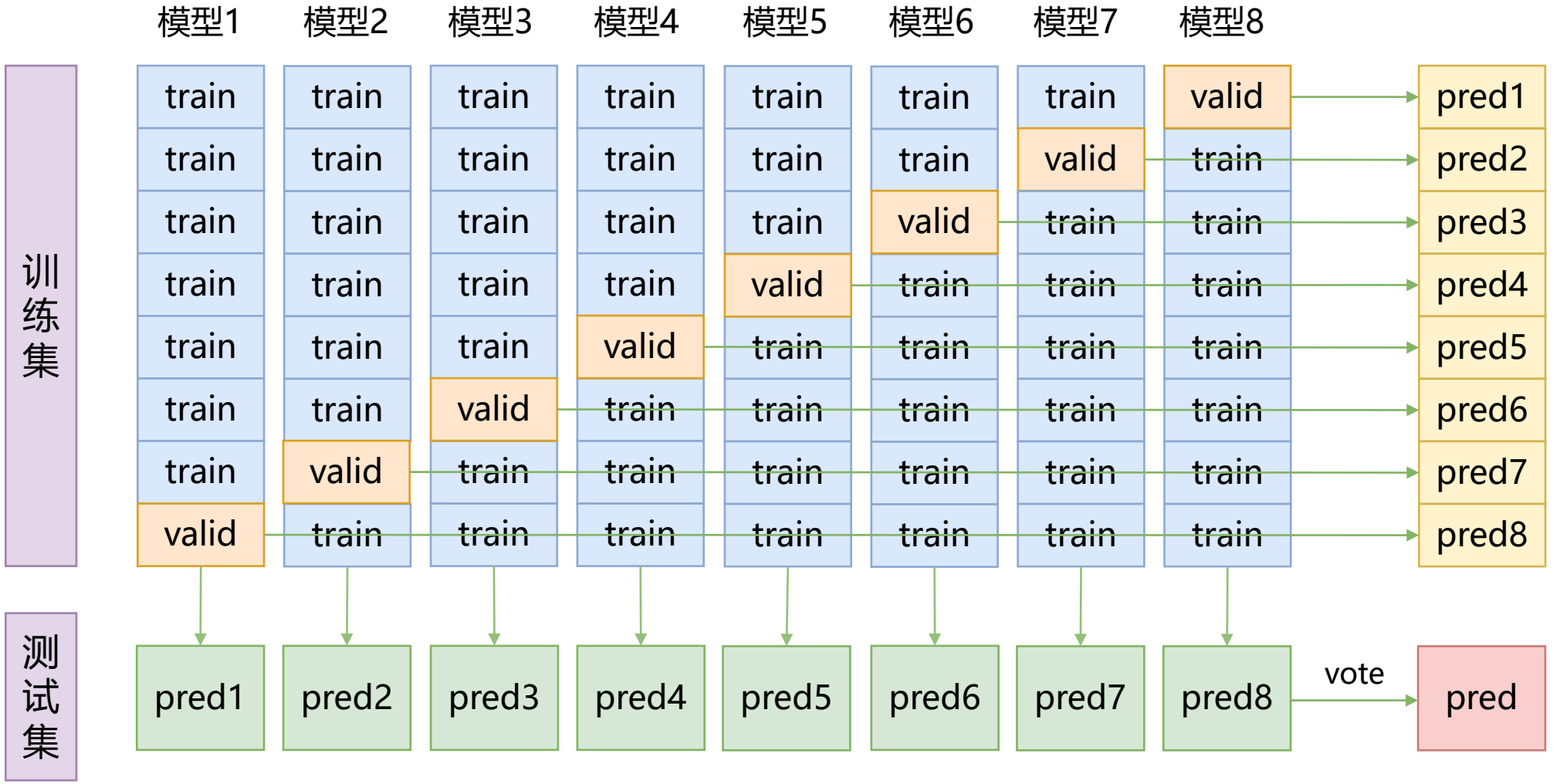
整体方案



模型框架


YES/NO分类器
yes_no_probs
全连接层
Pooling层
自注意力层
无法回答分类器
unk_prob
全连接层
片段抽取网络
start_probs
end_probs
全连接层
Pooled_output
Sequence_output
C
T_1
T_2
T_3
...
T_{N-1}
T_N
BERT 编码
文档
问题

集成方案



采用vote的方式，即对多个基模型的预测进行投票，选择得票最多的预测作为最终答案

后处理

分析模型预测的答案，我们发现以下两个问题：

1. 在刑事类问题上，模型性能较低，这与刑事类问题较为复杂，答案片段往往很长的特点有关。
2. 某些特定问题，模型预测的答案边界跟ground truth比有些偏差，可以通过一定的规则修正。（标注不一致）

针对上述问题，我们采用一些规则对模型预测的答案进行后处理，以得到更合理的输出。

问题类型	问题样例	规则
投保人类	“投保的人是谁？”，“投保人是谁？"	“A作为被保险人”，“投保人为A”，“A向B投保了C险”
目标投保公司类	“向什么公司投保？"	从文档中抽取“xx公司”实体名与模型预测答案进行校验
时间类	“投保人所投保险的有效时间是多久？”，“原告购买的保险的有效时间是多久？”	对模型答案不符合“x年x月x日”形式的进行修正
案件经过类	“案件发生经过是怎样的？ “，“抢劫案件发生经过是怎么样的？”	从文中匹配符合“xx案件发生经过…”，“抢劫罪…”等模式的span

比赛结果

第二阶段	第三阶段	总成绩
81.815	81.595	81.661

更多细节

1. BERT预训练模型的选择：bert-wwm(HIT)>google>ERNIE>THU。
2. 根据模型大小要求（3G），采用了8折交叉的集成方案。集成模型是提升性能的一个关键点，我们尝试了多种集成方案，保留了效果最好的方法。
   - 方案一：从多个base模型的预测中选择得分最大的作为最终预测。
   - 方案二：对yes/no，unk和span分别处理，使用得分最大+vote的方式：若半数base模型预测为yes/no或无法回答，依据vote得到相应答案；否则根据最大得分得到最终预测。
   - **方案三：完全使用vote的方式，即如果一个span被多个base模型所预测，则选它作为答案。**
   - 方案四：所有base模型预测的概率序列求平均，作为最终的预测。
3. 对模型的预测，特别是回答效果不够好的刑事类问题进行后处理验证，实验证明有效，线上的F1值略有提升。不过时间原因，没有对错误样例进行更细致的分析和后处理修正。
4. 尝试了几种在BERT基础上改进的模型，如答案抽取网络中增加self-match层，增加全连接层的层数，或是将是否分类器和无法回答分类器两部分合并为三分类器的结构，然而均无提升。

进一步提升的设想

- **模型：**

1. 这是一个法律文本的任务，可以从法律文本的特点出发设计模型。
2. 引入外部法律专业知识和常识助力模型推理。

- **数据：**

1. 标注不准确：标注不准确的问题如：答案span在文中多次出现，但ground truth总是取该span第一次出现的位置；可以使用答案句检索的方式重新标注答案位置。
2. 数据扩充：通过回译的方法扩充数据集也是一个突破性能瓶颈的良好途径。

感谢您的聆听，欢迎交流！

联系方式

delaiqiu@163.com

caldreaming821618@gmail.com

代码开源

https://github.com/NoneWait/cail2019