2020法研杯

阅读理解赛道参赛报告

队伍：Cola
成员：吴云朝、吴康康
所属单位：南京擎盾信息科技有限公司

赛题分析

- 对于每个问题，需要结合案情描述内容，给出回答，回答为Span（内容的一个片段）、YES/NO、Unknown中的一种，并且给出答案依据，即所有参与推理的句子编号。评价包括两部分：1）Answer-F1，即预测答案会与标准答案作比较，计算F1；2）SupFact-F1，即预测句子编号序列会与标准句子编号序列作比较，计算F1。最终为这两部分F1的联合F1宏平均。

分析：

1. 需要预测答案的类型，分为4个类型（Span、YES、NO、Unknown）。
2. 如果时Span类型需要预测答案片段的开始和结束位置。
3. 预测答案依据。

数据集

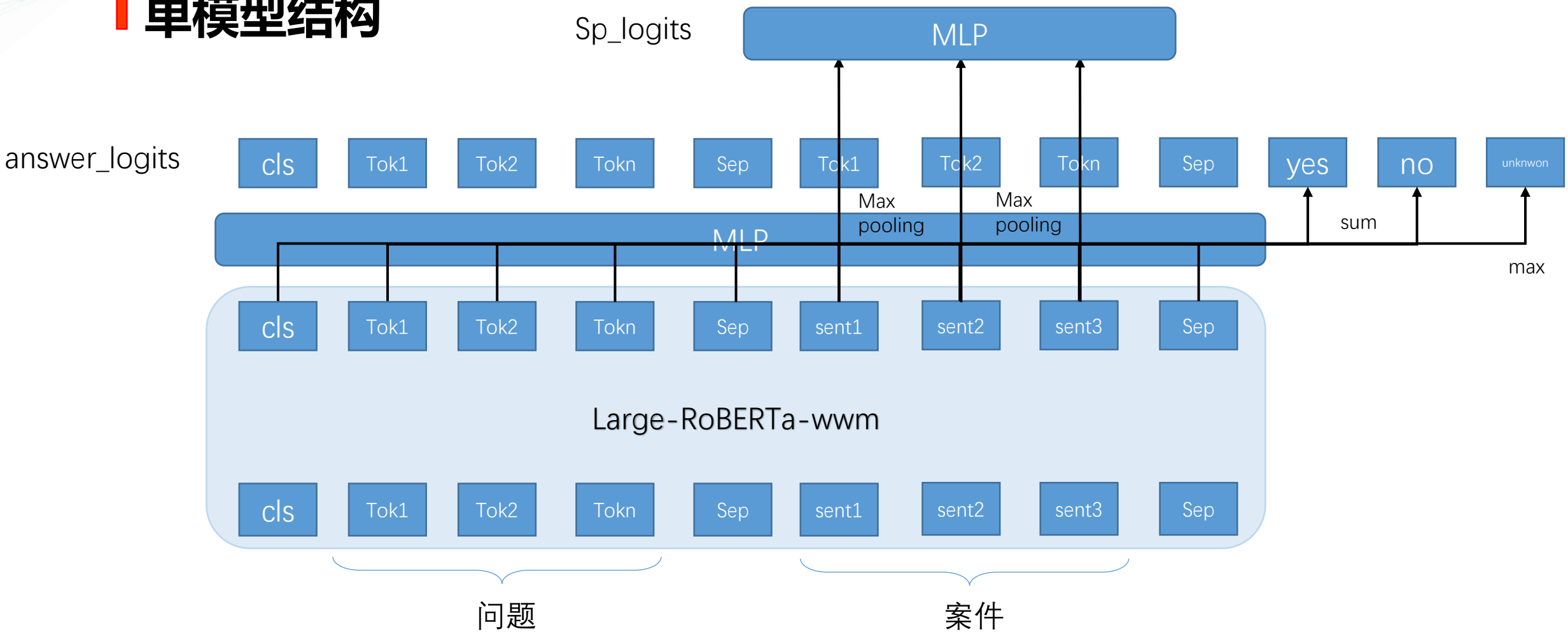
CAIL2019阅读理解赛道数据集（4000民事+4000刑事+2000官方测试验证集，只标注了答案，缺少证据标注）

CAIL2020阅读理解赛道数据集（民事、刑事、行政各约1700）

单模型探索

1. 迁移学习：为了利用CAIL2019的数据，将模型answer模块的网络结构在CAIL2019的数据上预先进行训练。在CAIL2020的模型中加载answer模块的参数。
2. 预测sp的时候尝试了用mean_pooling、max_pooling、 mean_pooling max_pooling拼接以及预先经过transformer_encoder模块等方法。
3. 尝试了很多预训练模型，包括brightmart和讯飞的Roberta，wwm，Albert以及清华的民事、刑事预训练BERT，最终发现讯飞的RoBERTa-wwm-ext-large效果最好。
4. 案件过长，采用了滑动窗口进行优化。
5. Answer任务和sp任务联合训练似乎并没有起到1+1>2的效果？
6. 超参数调整。

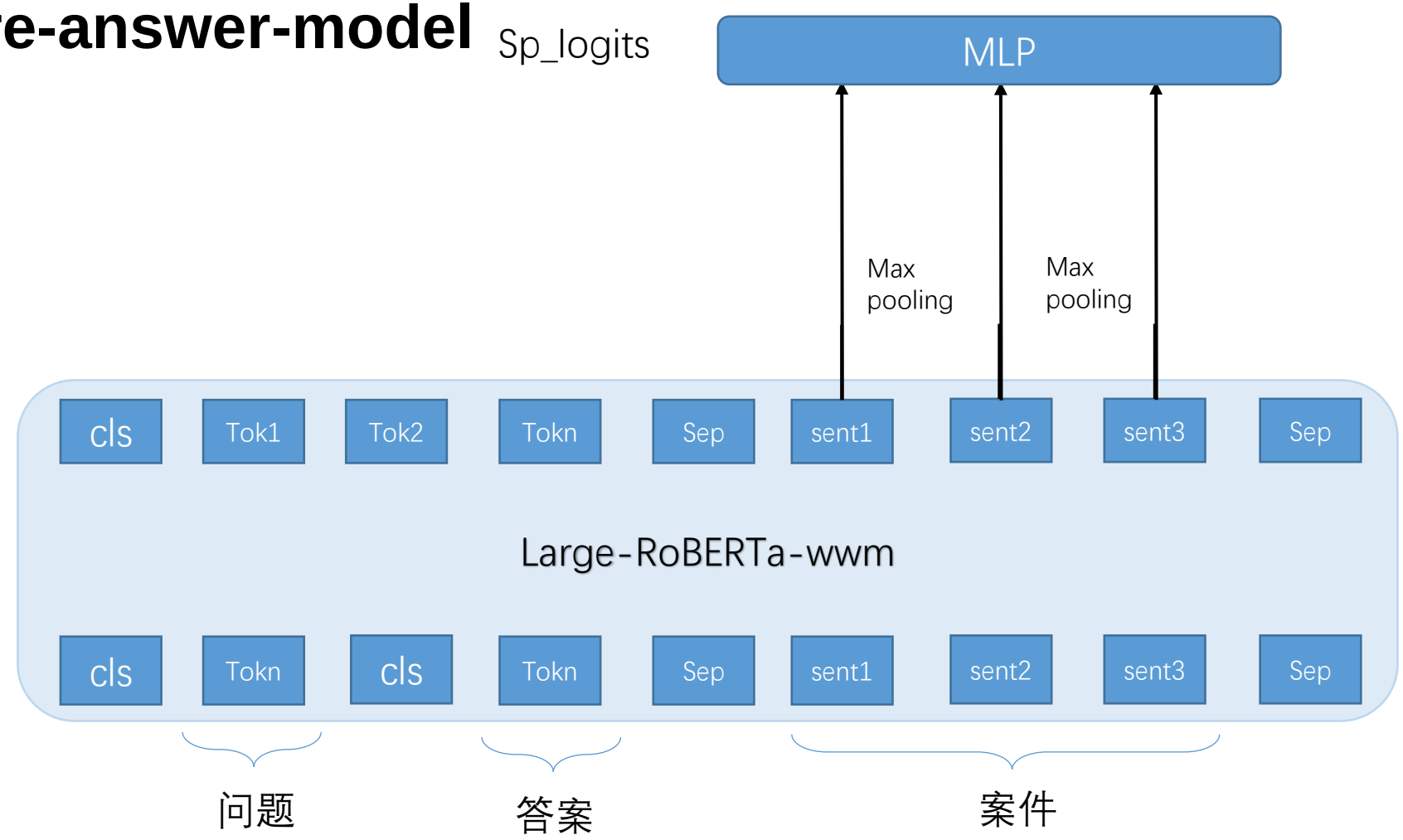
单模型结构


Sp_logits
MLP
answer_logits
cls
Tok1
Tok2
Tokn
Sep
Tok1
Tok2
Tokn
Sep
yes
no
unknwon
Max pooling
Max pooling
sum
MLP
max
cls
Tok1
Tok2
Tokn
Sep
sent1
sent2
sent3
Sep
Large-RoBERTa-wwm
cls
Tok1
Tok2
Tokn
Sep
sent1
sent2
sent3
Sep
问题
案件

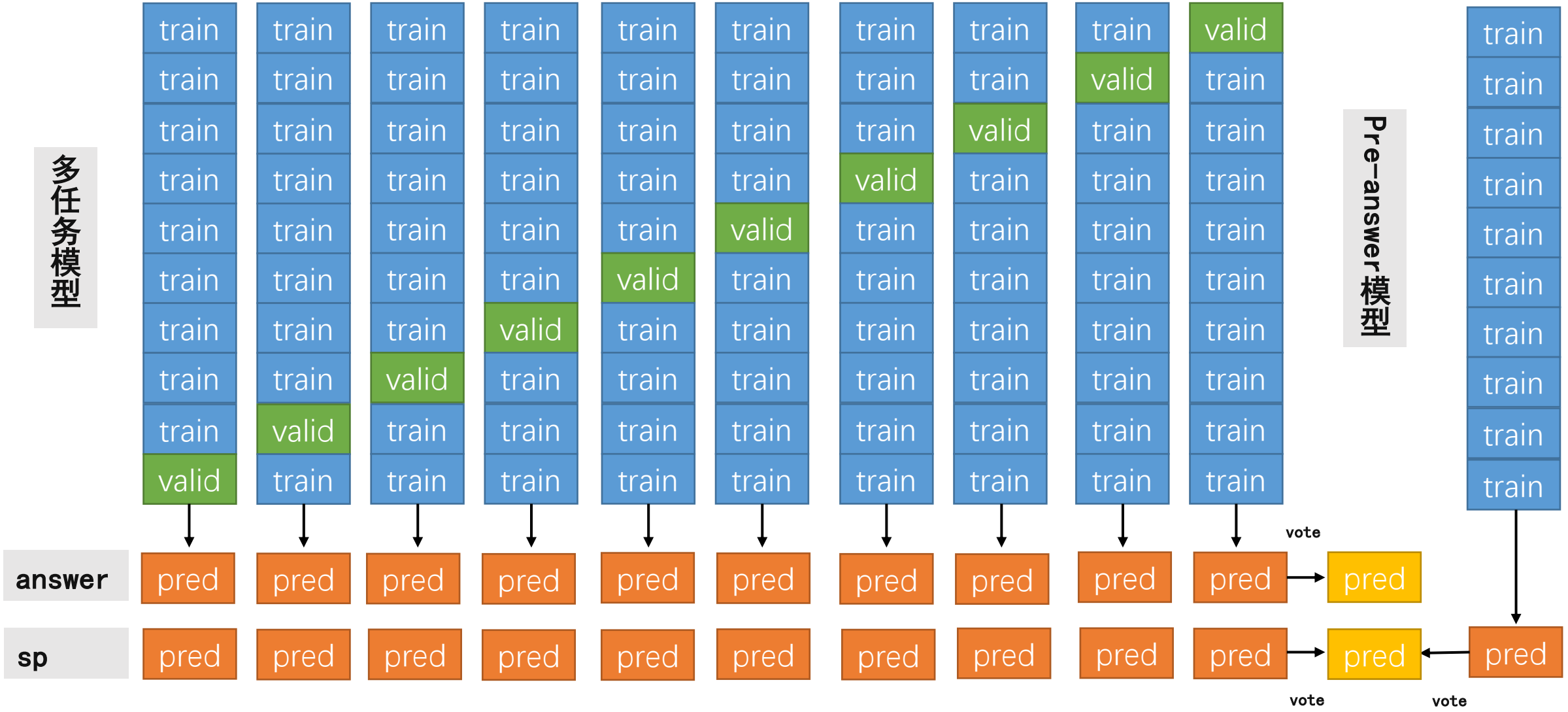
多模型融合

1. 经过实验发现模型在预先知道答案的情况下可以显著提高sp的预测准确性。(preanswer_sp：将答案字段拼接在问题后面输入模型)
2. 在10折和5折的情况下，Vote方法比logits平均在线上有更好的效果。
3. 考虑提交模型大小上限是8G，半精度压缩后单模型大小为750M（再经过zip压缩后为715M），采用10折交叉+单个模型（preanswer_sp）进行加权vote。（尝试过取并集、交集，效果不佳）
4. 调低sp预测的阈值，并且同时调低vote的阈值。线上略微提高。

Pre-answer-model Sp_logits



多模型结构


多任务模型
Pre-answer模型
train
valid
answer
sp
pred
vote
vote
vote

进一步设想

- 预先用海量法律文书微调BERT
- 对数据进行增广
- 图神经网络

谢谢聆听，欢迎交流

队伍：Cola

成员：吴云朝(ericperfecttttt@gmail.com)
吴康康

单位：南京擎盾信息科技有限公司