第三届“讯飞杯”中文机器阅读理解评测
CMRC2019

容钰添 张鑫睿

目录

问题背景

问题背景

- **句子级填空型阅读理解**
  - 篇章中多个句子缺失
  - 候选答案中包含假答案
- **评价指标**
  - 问题准确率：QAC=答对问题数 / 总问题数
  - 篇章准确率：PAC=完全答对篇章数 / 总篇章数

	总篇章数	总问题数	最大选项数	假答案
Train data	9638	100009	15	无
Trial data	139	1504	15	无
Dev data	300	3053	15	有
Quality data	500	5081	15	有
Test data	-	-	-	-

数据预处理

数据预处理

- **数据清理**
  - 统一标点符号、去除拼音标注、繁转简…
- **增加假答案**
  - 从原文中随机选取一定数量的句子作为候选答案（假答案）参与训练
- **候选答案回填**
  - 篇章中的所有候选答案分别回填到篇章中的BLANK处
- **多[MASK]填充**
  - 截取的上下文中出现其他BLANK，用多个[MASK]填充

数据预处理-候选答案回填

原文： 一艘太空船在宇宙中慢慢漂浮着。“桔子是一个什么样的星球?**[BLANK1]**?”驾驶舱中,小米问雪儿。[BLANK2]……

候选答案：

1. “他见黄金鸟飞远了,就走进山洞”,
2. “小米第一个从舱里跳了下来”,

……

n. “小米来到一个山谷,里面很安静,到处都散射着耀眼的金光”

候选答案回填：

1.一艘太空船在宇宙中慢慢漂浮着。“桔子是一个什么样的星球?**他见黄金鸟飞远了,就走进山洞**?”驾驶舱中,小米问雪儿。[BLANK2]……

2.一艘太空船在宇宙中慢慢漂浮着。“桔子是一个什么样的星球?**小米第一个从舱里跳了下来**?”驾驶舱中,小米问雪儿。[BLANK2]……

……

n.一艘太空船在宇宙中慢慢漂浮着。“桔子是一个什么样的星球?**小米来到一个山谷,里面很安静,到处都散射着耀眼的金光**?”驾驶舱中,小米问雪儿。[BLANK2]……

依据： 受CoLA[1]和SWAG[2]启发，将任务转换为判断句子**合理性**、**连贯性**问题，所以采用候选答案回填的方式

1. Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. arXiv preprint arXiv: 1805.12471.
2. Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. Empirical Methods in Natural Language Processing (EMNLP), 2018.

数据预处理-多[MASK]填充

原文： 拉拉本领特别大，他会变好多种魔术哦！比如，**[BLANK1][BLANK2]**把红兔子的围巾变成裙子……

候选答案回填： 拉拉本领特别大，他会变好多种魔术哦！比如，**把白兔子的帽子变成一朵花；[BLANK2]**把红兔子的围巾变成裙子……

[MASK]填充[BLANK]： 拉拉本领特别大，他会变好多种魔术哦！比如，**把白兔子的帽子变成一朵花；[MASK]**把红兔子的围巾变成裙子……

多[MASK]填充[BLANK]： 拉拉本领特别大，他会变好多种魔术哦！比如，**把白兔子的帽子变成一朵花；[MASK][MASK][MASK][MASK][MASK][MASK]**把红兔子的围巾变成裙子……

- **与掩码语言模型保持输入一致性**，用[MASK]替换其他[BLANK]
- **一定程度还原候选答案与上下文的相对距离**，用多[MASK]替换单[MASK]（6个）

模型预训练

模型预训练-特定领域预训练[3]（SpanBERT-wwm-ext）

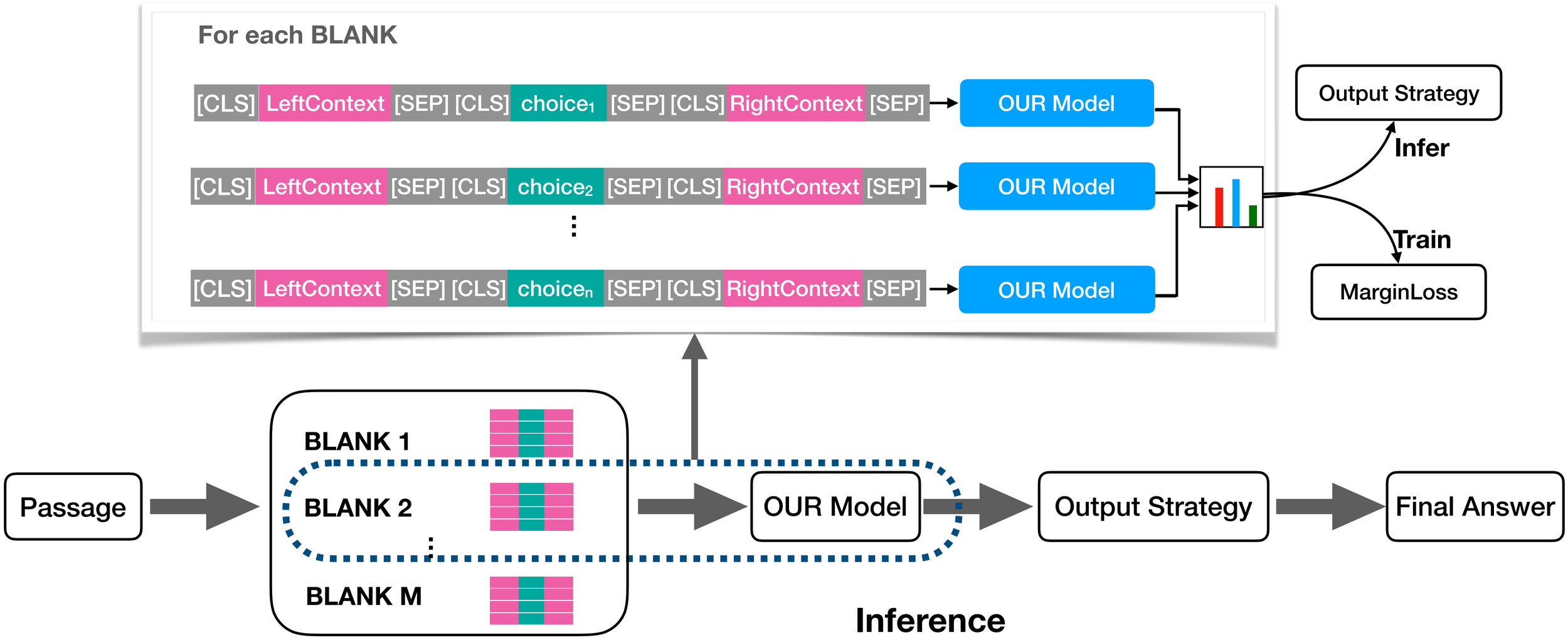
- 预训练初始模型：**bert-wwm-ext**[4]
- 特定领域预料：CMRC2017+CMRC2019 Dataset
- 动态MASK[5]
- Span MASK[5]
- 舍弃Next Sentence Prediction[5]
- 学习率：1e-4
- Epochs：3
- 批量大小：128
- 文本最大长度：128
- 优化器：AdamW + LookAhead[6]
- 其他参数：bert原始设置

3. Howard J , Ruder S . Universal Language Model Fine-tuning for Text Classification. Association for Computational Linguistics (ACL), 2019.
4. https://github.com/ymcui/Chinese-BERT-wwm
5. Joshi M, Chen D, Liu Y, et al. Spanbert: Improving pre-training by representing and predicting spans[J]. arXiv preprint arXiv:1907.10529, 2019.
6. Zhang M R, Lucas J, Hinton G, et al. Lookahead Optimizer: k steps forward, 1 step back[J]. arXiv preprint arXiv:1907.08610, 2019.

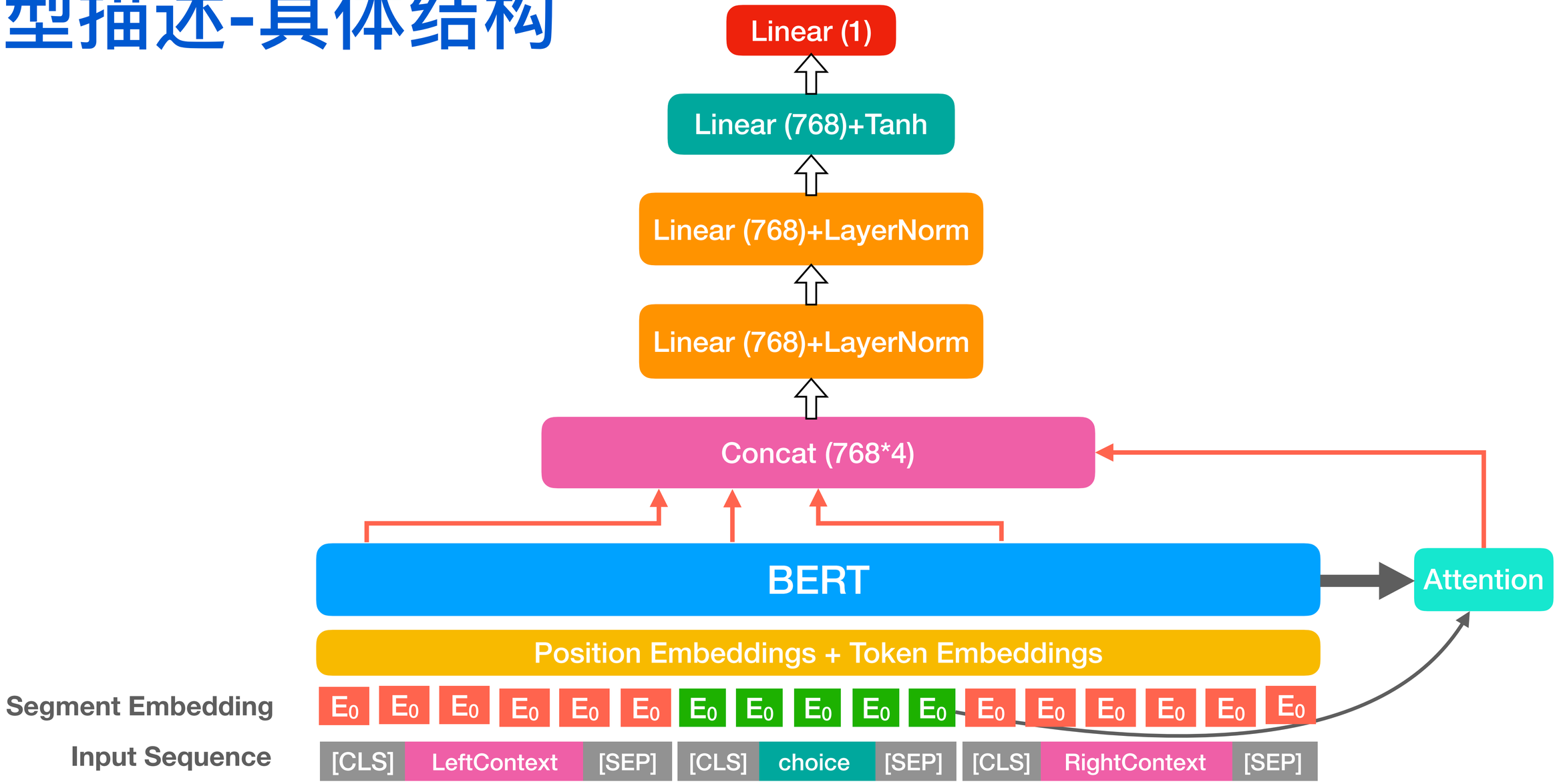
04

模型描述

模型描述-整体结构


For each BLANK
[CLS] LeftContext [SEP] [CLS] choice1 [SEP] [CLS] RightContext [SEP]
[CLS] LeftContext [SEP] [CLS] choice2 [SEP] [CLS] RightContext [SEP]
[CLS] LeftContext [SEP] [CLS] choicen [SEP] [CLS] RightContext [SEP]
OUR Model
OUR Model
OUR Model
Output Strategy
Infer
Train
MarginLoss
Passage
BLANK 1
BLANK 2
BLANK M
OUR Model
Output Strategy
Final Answer
Inference

模型描述-具体结构


Linear (1)
Linear (768)+Tanh
Linear (768)+LayerNorm
Linear (768)+LayerNorm
Concat (768*4)
BERT
Attention
Position Embeddings + Token Embeddings
Segment Embedding
Input Sequence
E0 E0 E0 E0 E0 E0 E0 E0 E0 E0 E0 E0 E0 E0 E0 E0 E0 E0
[CLS] LeftContext [SEP] [CLS] choice [SEP] [CLS] RightContext [SEP]

模型描述-Margin Loss

$$\ell_{\text{margin}} = \frac{1}{N-1} \sum_{c' \in C, c' \neq c} \max\left\{0, \xi - f(p,c) + f(p,c')\right\}$$

N为候选答案数量，f得分函数，P为上下文，c为正确选项，c'为错误选项，ξ为间隔参数

依据：候选答案选择视为一个**合理性排序问题**[7]，而不是分类问题

7. Li Z, Chen T, Van Durme B. Learning to Rank for Plausible Plausibility. Association for Computational Linguistics (ACL), 2019.

模型描述-知识蒸馏

student采用和teacher相同的网络结构（重生网络[8]），使用teacher annealing策略[9]，本系统使用的损失函数如下：

$$\lambda \ell_{\mathrm{KL}}\left(f(x_i,\theta_{\mathrm{teacher}}), f(x_i,\theta_{\mathrm{student}})\right)+\left(1-\lambda\right)\ell_{\mathrm{margin}}\left(y_i, f(x_i,\theta_{\mathrm{student}})\right)$$

ℓ_{KL}为KL散度损失，ℓ_{margin}为margin损失。在训练过程中，λ从1到0线性递减

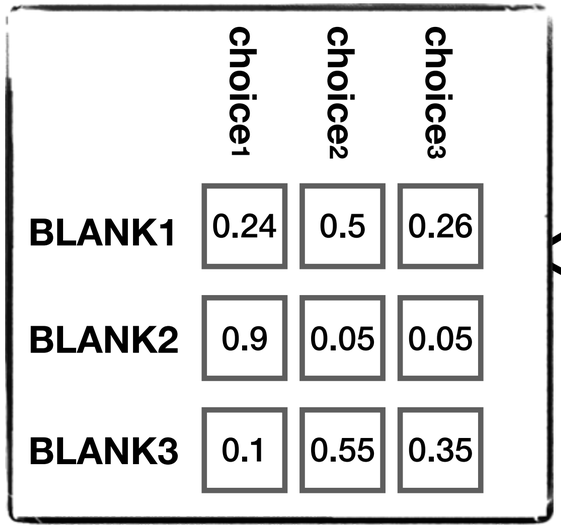
8. Furlanello T, Lipton Z C, Tschannen M, et al. Born again neural networks. International Conference on Machine Learning (ICML), 2018.
9. Clark K, Luong M T, Khandelwal U, et al. Bam! born-again multi-task networks for natural language understanding. Association for Computational Linguistics (ACL), 2019.

模型描述-训练策略

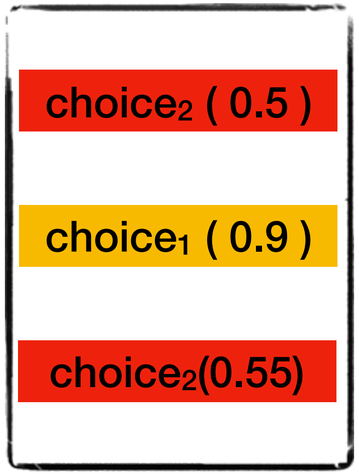
- **模型训练数据及参数**
  - 初始模型参数：spanbert-wwm-ext
  - 训练数据：CMRC2017+CMRC2019 Dataset
  - 学习率：3.5e-5（2e-5、3.5e-5、4e-5）
  - epochs：3
  - 批量大小：16
  - 文本最大长度：128（64、128）
  - 损失函数：margin loss（ξ=0.36） （margin loss、交叉熵）
  - 优化器：AdamW + LookAhead
  - 其他参数：bert原始设置

预测策略

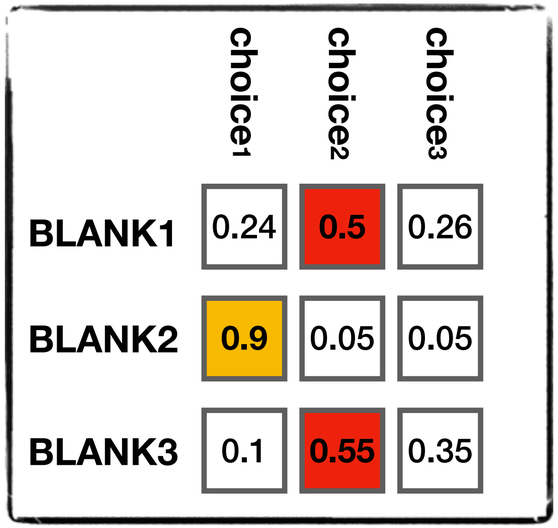
预测策略

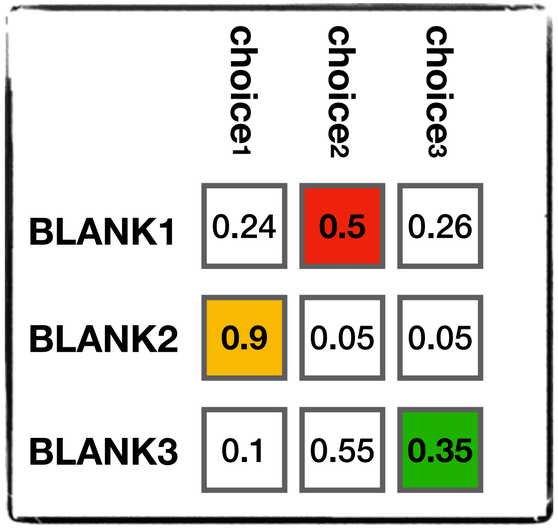


方案A(最高得分)



方案B(差值排序选择):
0.5 - 0.26(BLANK1) >
0.55-0.35(BLANK2)





将篇章中所有BLANK对应的候选答案得分向量拼接在一起，构成BLANK-choice的得分矩阵M，在M上采用差值排序策略，得到最终结果

06

实验结果

实验结果

验证集（dev）消融实验结果对比

	QAC/PAC			
	方案A(最高得分)	△	方案B(差值排序)	△
Complete Model with MarginLoss	**79.30 / 16.89**	—	**86.15 / 42.33**	—
- answer back	77.20 / 17.00	-2.10 / +0.11	83.13 / 36.33	-3.02 / -6.00
- multi [MASK]	78.10 / 15.67	-1.20 / -1.22	84.91 / 38.89	-1.24 / -3.44
- fake answer	78.67 / 16.11	-0.63 / -0.78	84.93 / 38.78	-1.22 / -3.55
- attention	79.08 / 16.73	-0.22 / -0.16	85.88 / 40.67	-0.37 / -1.66
Complete Model with CrossEntropyLoss	79.00 / 15.67	-0.30 / -1.22	84.60 / 38.00	-0.85 / -4.33

实验结果

验证集（dev）不同预训练模型实验结果对比

	QAC/PAC			
	方案A(最高得分)	△	方案B(差值排序)	△
SpanBERT-wwm-ext	**79.30 / 16.89**	—	**86.15 / 42.33**	—
BERT-wwm-ext	77.76 / 13.67	-1.54 / -3.22	84.80 / 38.33	-1.35 / -4.00
BERT-wwm	75.43 / 12.67	-3.87 / -4.22	82.38 / 30.67	-3.77 / -11.66
BERT	75.20 / 13.33	-4.10 / -3.56	81.46 / 32.67	-4.69 / -9.66
RoBERTa-wwm-ext	78.58 / 14.67	-0.72 / 2.22	84.93 / 39.33	-1.22 / -3.00

实验结果

验证集（dev）模型集成实验结果对比

	QAC/PAC			
	方案A(最高得分)	△	方案B(差值排序)	△
Complete Model	79.30 / 16.89	__	86.15 / 42.33	__
Complete Model+Distillation	79.20 / 16.83	-0.10 / -0.06	86.77 / 42.00	+0.62 / -0.33
Ensemble	**80.97 / 19.67**	+1.67/ +2.78	**88.21 / 48.00**	+2.06 / + 5.67

Ensemble模型：包含7个模型，分为3个类型，分别如下

1. +distill+attention（2个）
2. +distill-attention（3个）
3. -distill+attention（2个）

实验结果

排名	队伍	开发集PAC	开发集QAC	资格集PAC	资格集QAC	测试集PAC	测试集QAC
1	**bert_scp_spm (ensemble)** PINGAN-GammaLab	60.0	90.927	58.2	90.789	57.6	90.055
2	**mojito system (ensemble)** SFTech	48.0	88.208	43.4	86.459	41.8	85.991
3	**DA-BERT (ensemble)** 百度	34.3	86.341	29.2	84.905	27.6	84.447
4	**CMRC2019 MULTIPLE BERT (ensemble)** Six Estates https://www.6estates.com	38.7	82.968	35.6	83.507	32.2	82.591
5	**nkuzhangyi_cmrc_v2 (ensemble)** CICC	29.7	80.937	26.0	80.319	26.6	79.562
6	**MRC-ZZ SYSTEM (single model)** 哈工大&汉仪字库	29.0	80.380	25.8	78.292	26.6	78.781
7	**MB-Reader (ensemble)** ECUST	18.7	78.218	17.8	76.422	15.6	76.319

模型最终实验结果

实验结果-错误分析

1. 对于一些需要推理的[BLANK]，无法给出正确答案，缺乏一定的知识推理能力。如：

走过来说：＂今天你迟到了，按照公司的规定，每月迟到满三次者，扣发当月奖金。＂小芸暗叫一声倒霉。下班的时候，**[BLANK5]**，却看见很多同事都不约而同地挤向电梯。难道，已经修好了吗？她走过去看了看，果然，电梯运转正常。

正确答案：她下意识地往楼梯口的方向走

预测答案：她在一楼的电梯口停下

2. 存在无法区分的候选答案，即答案都合理。如：

年二十七岁，何不学苏老泉呢？＂学苏老泉，发愤图强。[BLANK4]。待到三十岁的时候，他的画就画得很不错了，**[BLANK5]**。不仅学画，学写诗，齐白石还对书法、篆刻很有兴趣。有一次，在请他作画的人家里，遇到了一位篆刻＂

正确答案：在家乡一带开始有了名气

预测答案：赢得了中国和世界人民的崇高评价

3. 方案B（差值排序）会把方案A（最高得分）预测正确的答案看作错误答案。如：

天动地的命令：＂免除老百姓三年赋税，王公贵族的牲畜分一半给没有牲畜的人。＂全国人民欢天喜地，而王公贵族却扬言：＂如果奥塔娜不取消这命令，一定要把她活活杀死！＂**[BLANK9]**，赶紧往回走，刚动身就被阿拉齐汗捉去了。

正确答案：阿吾兰齐汗闻讯后非常惊慌

预测答案（方案B）：奥塔娜把这一一告诉父亲

预测答案（方案A）：阿吾兰齐汗闻讯后非常惊慌

实验结果-错误分析

4. 关键上下文缺失（多出现在多个BLANK连续情况下），已有信息无法得到真正的答案，可以通过方案B部分解决。如：

说：＂送你两个嫩葫芦吃吧！＂小白兔咧开三瓣嘴笑了：兔爷家的葫芦多得成灾了，得！我帮帮他吧。[BLANK7]。**[BLANK8]**。[BLANK9]。[BLANK10]。[BLANK11]。兔爷看看天说：＂暴风雨就要来了！＂风是雨的头，狂风过后，天上下起了瓢泼大雨，一下

正确答案：兔爷接过花慢吞吞地走了

预测答案：这时,洪水快要淹到脖子了

＂送你两个嫩葫芦吃吧！＂小白兔咧开三瓣嘴笑了：兔爷家的葫芦多得成灾了，得！我帮帮他吧。[BLANK7]。[BLANK8]。**[BLANK9]**。[BLANK10]。[BLANK11]。兔爷看看天说：＂暴风雨就要来了！＂风是雨的头，狂风过后，天上下起了瓢泼大雨，一下就是

正确答案：小白兔到城里卖花,赚了不少钱

预测答案：小白兔采下鲜花,扎成一束一束的放在车上,他要到城里去卖花

实验结果-错误分析

5. 顺序不明确，候选正确答案无明显的语序差别。如：

了，希拉克的所有家产就是130万欧元。这在法国，实在算不上什么有钱。[BLANK5]，不过事实就是这样。[BLANK6]**[BLANK7]**[BLANK8]乡村别墅50万元;[BLANK9]，就是这些。一笔一笔都清楚明白，任何人都可以查询，可以提出疑问，甚至可

正确答案： 20万元的家具和艺术品；

预测答案： 夫人股市资金42万元；

，希拉克的所有家产就是130万欧元。这在法国，实在算不上什么有钱。[BLANK5]，不过事实就是这样。[BLANK6][BLANK7]**[BLANK8]**乡村别墅50万元;[BLANK9]，就是这些。一笔一笔都清楚明白，任何人都可以查询，可以提出疑问，甚至可以指

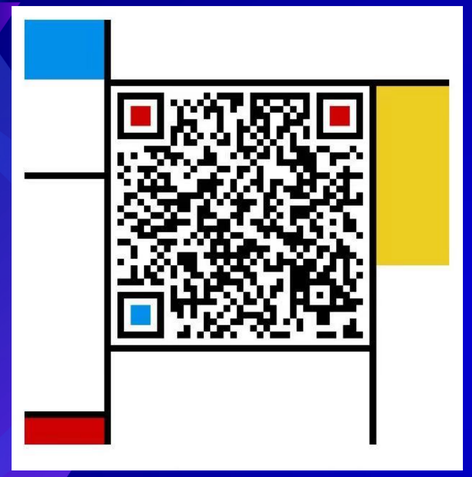
正确答案： 夫人股市资金42万元；

预测答案： 20万元的家具和艺术品；

实验结果-赛后实验

去掉在原文中出现的假答案实验结果对比（dev）

	QAC/PAC			
	方案A(最高得分)	△	方案B(差值排序)	△
Sigle Model	79.30 / 16.89	—	86.15 / 42.33	—
Sigle Model +后处理时去掉假答案	**80.18 / 18.00**	+0.88 / +1.11	**88.60 / 53.67**	+2.45 / 11.34
Ensemble Model	80.97 / 19.67	—	88.21 / 48.00	—
Ensemble Model+后处理时去掉假答案	**81.26 / 19.67**	+0.29 / 0.00	**90.21 / 57.67**	+ 2.00 / 9.67



容钰添

张鑫睿

TECHNOLOGY 顺丰科技

Thanks