第五届“讯飞杯”中文机器阅读理解评测总结

Overview of the 5th Workshop on Chinese Machine Reading Comprehension (CMRC 2022)

CMRC 2022 评测委员会

江西·南昌

2022年10月30日

评测概况

INTRODUCTION

大赛历史

- **中文机器阅读理解评测（CMRC）**
  - CCL大会从2017年开始举办技术评测，CMRC 2017是CCL 2017首个评测
  - CMRC也是国际范围内机器阅读理解技术方向上最早的评测研讨会
  - 至今已举办5届技术评测：CMRC 2017/2018/2019/2020/2022

大赛历史

- **第一届“讯飞杯”中文机器阅读理解评测（CMRC 2017）**
  - 任务：填空型阅读理解

- **第二届“讯飞杯”中文机器阅读理解评测（CMRC 2018）**
  - 任务：抽取型阅读理解

大赛历史

- **第三届“讯飞杯”中文机器阅读理解评测（CMRC 2019）**
  - 任务：句子级填空型阅读理解

- **第四届“讯飞杯”中文机器阅读理解评测（CMRC 2020）**
  - 与中国法研杯司法人工智能挑战赛（CAIL 2020）联合举办，聚焦司法阅读理解任务

本届评测概况

- **第五届“讯飞杯”中文机器阅读理解评测（CMRC 2022）**
  - 由认知智能国家重点实验室、哈尔滨工业大学社会计算与信息检索研究中心（HIT-SCIR）、科大讯飞（北京）有限公司联合举办
  - 由CCL大会颁发获奖证书，由科大讯飞股份有限公司提供奖金
- **CMRC 2022 评测委员会**
  - 主席：崔一鸣（科大讯飞）、车万翔（哈尔滨工业大学）
  - 委员：杨子清（科大讯飞）、潘雨晨（科大讯飞）

本届评测概况

- **奖项设置**
  - 每个赛道分别评选出冠军、亚军、季军各一名，颁发奖金和荣誉证书
  - 入围决赛队伍会获赠由电子工业出版社出版的图书《自然语言处理：基于预训练模型的方法》

奖项	数量	奖励
冠军	一名	奖金10,000元 + 荣誉证书
亚军	一名	奖金5,000元 + 荣誉证书
季军	一名	奖金3,000元 + 荣誉证书

本届评测概况

- **时间安排**
  - 本届评测从5月份开始报名，9月完成整个比赛环节

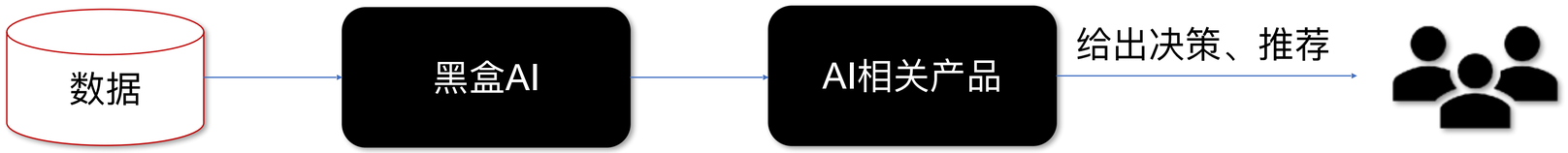
阶段	时间
报名阶段	2022年5月18日 - 2022年7月31日
资格赛阶段	2022年6月1日 - 2022年8月1日
资格审查：提交	2022年8月1日 - 2022年8月7日
资格审查：公布结果	2022年8月15日
决赛阶段	2022年8月16日 - 2022年8月31日
公布决赛结果	2022年9月中旬
撰写评测报告	2022年9月下旬
评测研讨会	2022年10月

评测任务介绍

Task Description

可解释人工智能

- **黑盒AI和可解释AI**
  - **黑盒AI**：为何作出这种决策，依据是？什么时候会导致方法失效？如何调整和修改决策过程？



  - **可解释AI**：同时给出决策和相应依据；可以知道AI方法在什么情况下有效；决策结果可控可反馈；用户信赖度更高

任务介绍

- **CMRC 2022 评测任务：可解释性阅读理解**
  - 给定一个篇章以及问题，选手需要设计一个可以同时抽取答案和佐证依据的机器阅读理解系统
  - 根据阅读理解类型分为以下两个赛道：
    - 抽取型阅读理解赛道：答案和佐证依据均是篇章中的某个连续片段
    - 选择型阅读理解赛道：答案从若干个候选选项中选出，佐证依据是篇章中的某个连续片段
  - 评测难点：不提供带标注的训练集合，需要设计无/弱监督系统来完成答案和佐证依据的抽取

Subset	Passage	Question & Answer
抽取型	...钩盲蛇（学名："Ramphotyphlops braminus"）是蛇亚目盲蛇科下的一种无毒蛇种，主要分布在非洲及亚洲，不过现在钩盲蛇的分布已推广至世界各地。钩盲蛇是栖息于地洞的蛇种，由于体型细小，加上善于掘洞...	**Q**: 钩盲蛇一般生活在什么地形中？ **A**: 地洞
选择型	...大学生活是走上社会的预演，可以说，大学里的处世态度和人际关系的成功与否，直接决定着将来在社会上的成败。人是社会性的动物，生活中的每个人都离不开别人的帮助，同时也在帮助着别人。不管是学习、生活、工作，都要求自己要有良好的处理人际关系的能力。一个人要想有良好的人际关系，就要遵循以下几个原则：一是"主动"。要主动和别人交往，主动帮助别人。二是"诚信"。...	**Q**: 说话人认为什么因素决定在社会上的成败？ **A**: 工作的态度 **B**: 朋友的数量 **C**: 大学里的学习成绩 **D**: 大学里的人际关系

[Cui et al., 2022] ExpMRC: Explainability Evaluation for Machine Reading Comprehension. In ***Heliyon Vol.8***

任务介绍

- **评测数据统计信息**
  - 开发集提供给参赛选手，测试集由评测委员会保留（不对外公开）

赛道	抽取型		选择型	
集合	开发集	测试集	开发集	测试集
答案类型	篇章片段	篇章片段	选择题	选择题
领域	维基百科	维基百科	考试	考试
篇章数	369	399	273	244
问题数	515	500	505	500
最大参考答案数量	3	3	1	1
最大参考证据数量	2	2	4	4

任务介绍：数据获取和使用要求

- **可以使用的数据**
  - 任何公开数据集的训练集部分，例如CMRC 2018、DRCD、C3等的训练集
  - 任何无标注的自由文本数据
  - 可以对上述两类数据进行自动加工形成伪训练数据，例如抽取答案所在句作为证据文本，以此形成弱监督的训练数据
- **不可以使用的数据**
  - 任何未公开的数据，例如私自人工标注的数据
  - 任何公开数据集中的开发集和测试集
  - 本届评测提供的开发集，即ExpMRC开发集

任务介绍：评价方法

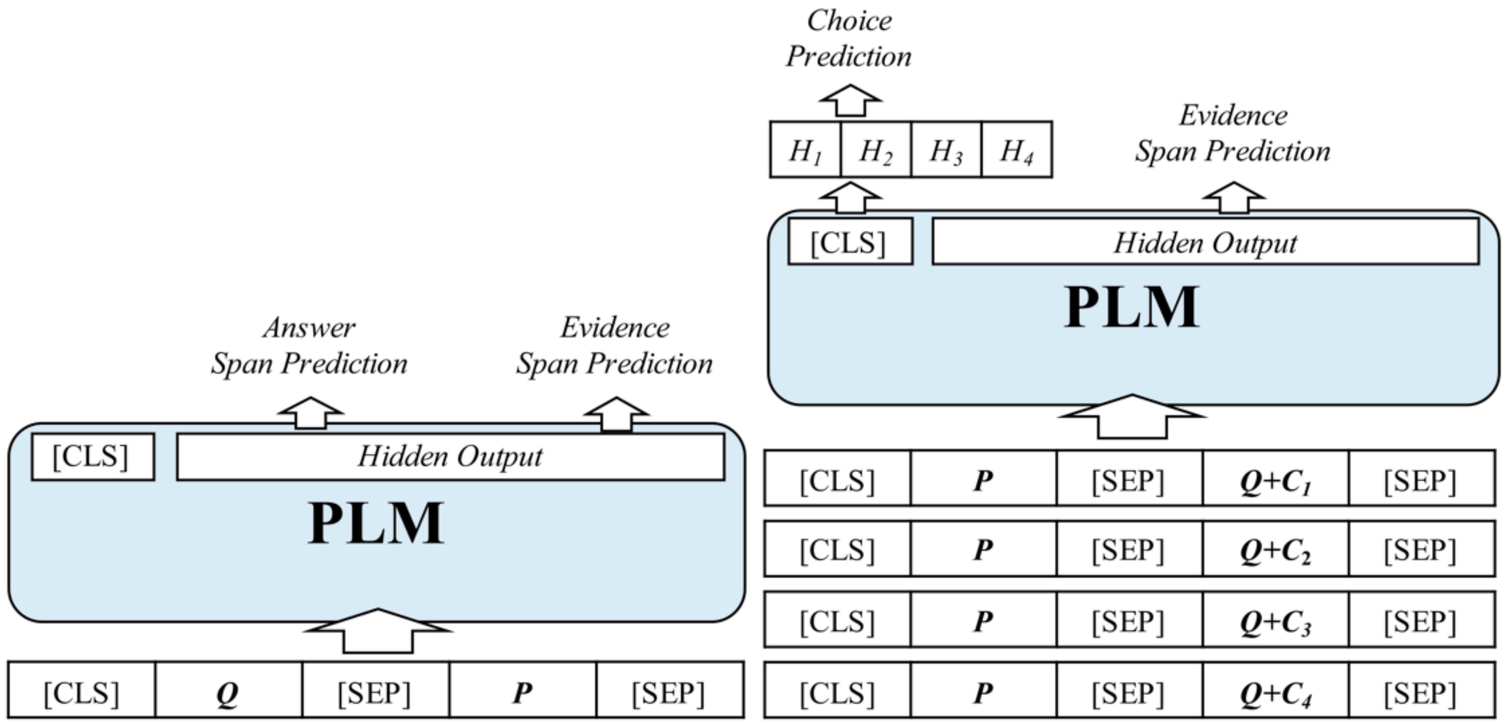
- **评价方法**
  - 提供三种评价方法，分别对答案、证据以及两者综合进行评价
    - 抽取型阅读理解赛道：答案F1、证据F1、综合F1
    - 选择型阅读理解赛道：答案准确率、证据F1、综合F1
  - F1指标用来计算正确答案（证据）和预测答案（证据）之间的重叠程度

$$F1(\text{综合}) = F1(\text{答案}) \times F1(\text{证据})$$

  - 最终排名将通过测试集上的“综合F1”指标降序排列得出

任务介绍：基线系统

- **基于伪数据的基线系统**
  - 通过启发式方法构建伪数据，构建相应的神经网络模型加以训练，以抽取答案和证据

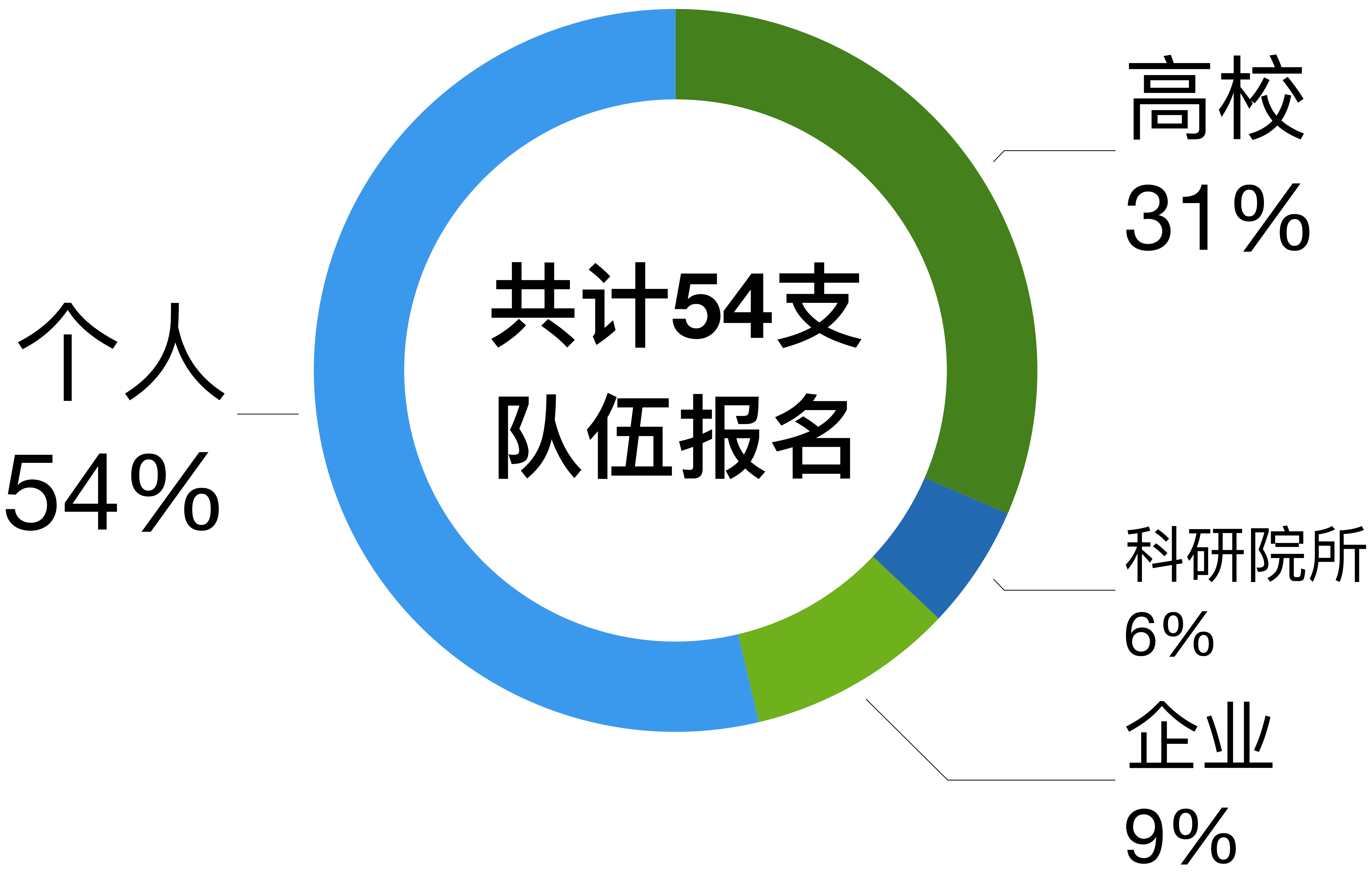


[Cui et al., 2022] ExpMRC: Explainability Evaluation for Machine Reading Comprehension. In ***Heliyon Vol.8***

参赛情况及评测结果

PARTICIPATION & EVALUATION RESULTS

报名情况


共计54支
队伍报名
高校
31%
科研院所
6%
企业
9%
个人
54%


山西大学
北京理工大学
青岛科技大学信息科学技术学院
哈尔滨工业大学
Cpic
Knt
Mediago
Nothing
个人
百度
南理工
阿里巴巴
北京
腾讯科技
苏州大学
自由职业
南开大学
中国矿业大学
Dataark
平安银行
有限公司
长沙理工大学
兰州大学
快手
Ob
Ecust
武汉大学
北京邮电大学
商汤科技
Nvidia
字节跳动
北京大学
大数据学院
华南理工大学
上海交通大学
北京语言大学
北京师范大学
中国科学院大学

比赛阶段

- **资格赛阶段（6月1日-8月1日）**
  - 参赛队伍将通过CodaLab平台自助提交开发集结果，取前10名进入到资格审查环节
- **资格审查阶段（8月1日-8月15日）**
  - 入围队伍需要提交代码和模型以完成资格审查
  - 同时需要提供训练数据来源的详细说明，不符合参赛规则的队伍将失去决赛资格
- **决赛阶段（8月16日-8月31日）**
  - 资格审查通过的队伍将提交两组系统（资格审查阶段提交的模型和决赛阶段提交的模型）
  - 由组委会给出测试集结果，取两组系统测试集“综合F1”指标高的结果作为该队伍的最终评测结果，参与最终的排名

参赛系统概况

- **数据方面**
  - 除了ExpMRC提供的pseudo-train之外，采用了额外弱监督标注方法生成evidence
  - 利用DuReader、CMRC 2018、DRCD、CAIL 2021等进行数据增广
  - 使用EDA、U3E、对抗训练等方法
- **模型方面**
  - 多数采用RoBERTa、MacBERT（包含MRC专版）等常用预训练模型
  - 也有队伍使用ERNIE3等其他预训练模型

资格赛成绩

抽取型阅读理解赛道

Results

#	User	Entries	Date of Last Entry	Team Name	Answer F1 ▲	Evidence F1 ▲	Overall F1 ▲
1	Two	11	07/29/22	Gamma_kg	91.89093 (4)	90.76024 (1)	83.73154 (1)
2	jerome	16	07/28/22	jerome	94.86694 (2)	85.58790 (2)	81.38900 (2)
3	Smile	2	07/30/22	不才	96.59644 (1)	81.95546 (4)	79.26139 (3)
4	yinpei_su	7	08/01/22		94.57919 (3)	81.18179 (5)	76.89272 (4)
5	Supreme	13	07/25/22	SXU_NLP	87.35584 (6)	84.15848 (3)	75.02840 (5)
6	yunxiaomr	3	07/30/22	YunxiaoMr	88.19959 (5)	80.91901 (6)	72.25283 (6)
7	yucheng-zeng	2	06/30/22	ITNLP_CMRC	81.48322 (7)	78.38410 (7)	64.82126 (7)
8	zgjiangtoby	3	07/31/22	Qust_AI	75.06074 (8)	74.63994 (8)	58.59530 (8)

选择型阅读理解赛道

Results

#	User	Entries	Date of Last Entry	Team Name	Answer F1 ▲	Evidence F1 ▲	Overall F1 ▲
1	yunxiaomr	10	08/01/22	YunxiaoMr	80.59406 (1)	75.82941 (2)	67.74704 (1)
2	jerome	6	07/28/22	jerome	79.00990 (2)	79.57246 (1)	66.95658 (2)
3	hsz779	5	07/24/22	suzhe	72.87129 (5)	72.52306 (3)	58.25260 (3)
4	Two	4	07/31/22	Gamma_kg	74.85149 (4)	68.53325 (4)	56.46481 (4)
5	Smile	1	07/30/22	不才	76.83168 (3)	66.57672 (5)	55.68041 (5)
6	Leishu	1	07/31/22	BNU_icip	70.29703 (6)	60.70876 (6)	45.02872 (6)

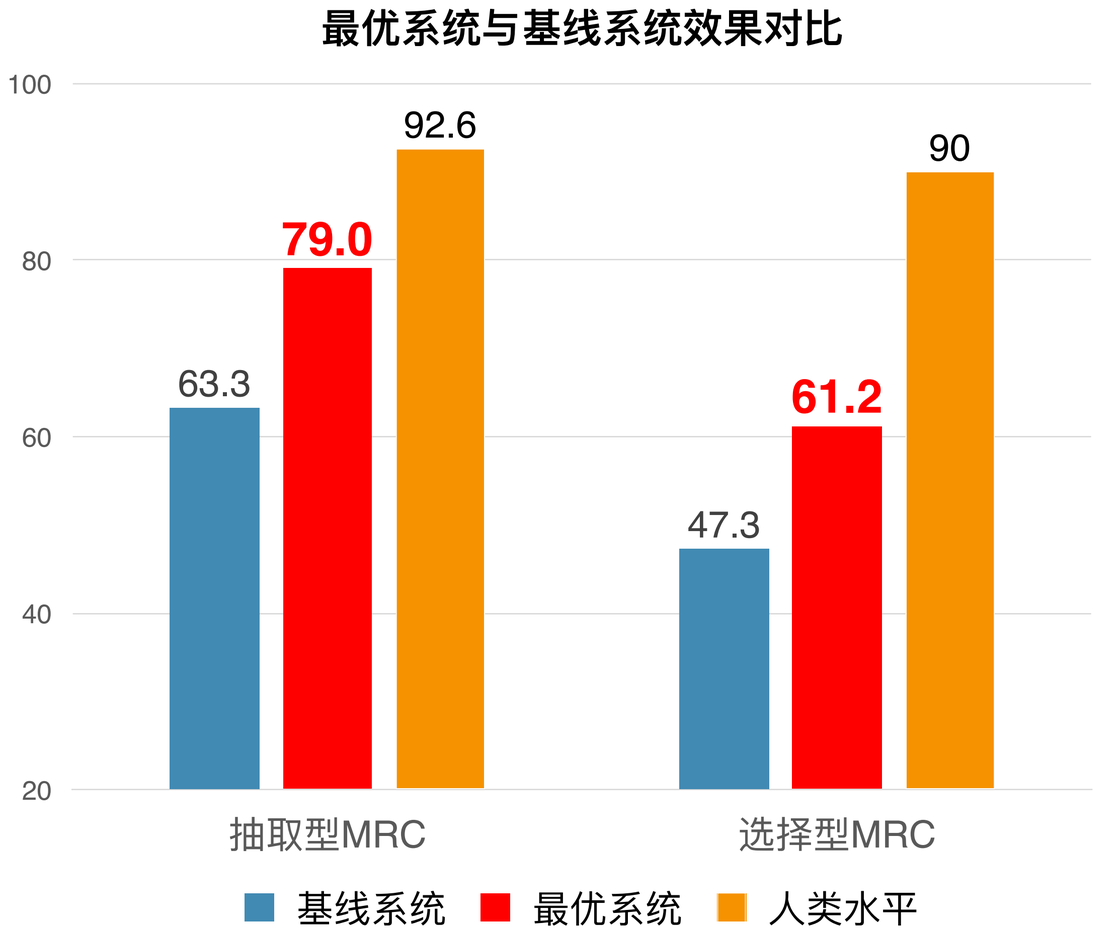
决赛成绩

抽取型阅读理解赛道

排名	队伍名	答案F1	证据F1	综合F1
1	Two	91.979	85.191	79.028
2	Supreme	92.804	79.404	73.934
3	jerome	95.428	75.568	72.294
4	Smile	94.576	72.552	69.104
5	yunxiaomr	91.786	71.715	65.944

选择型阅读理解赛道

排名	队伍名	答案Acc	证据F1	综合F1
1	yunxiaomr	81.0	70.902	61.196
2	jerome	77.0	71.241	60.133
3	hsz779	75.4	69.857	57.784
4	Smile	74.0	61.526	50.952
5	Two	69.2	61.605	47.833
6	leishu	66.4	58.646	41.458



获奖单位（抽取型阅读理解赛道）

Two（深圳壹帐通智能科技有限公司）

Supreme（山西大学）

jerome（阳光出行）

获奖单位（选择型阅读理解赛道）

yunxiaomr（山西大学）

jerome（阳光出行）

hsz779（北京理工大学）

总结展望

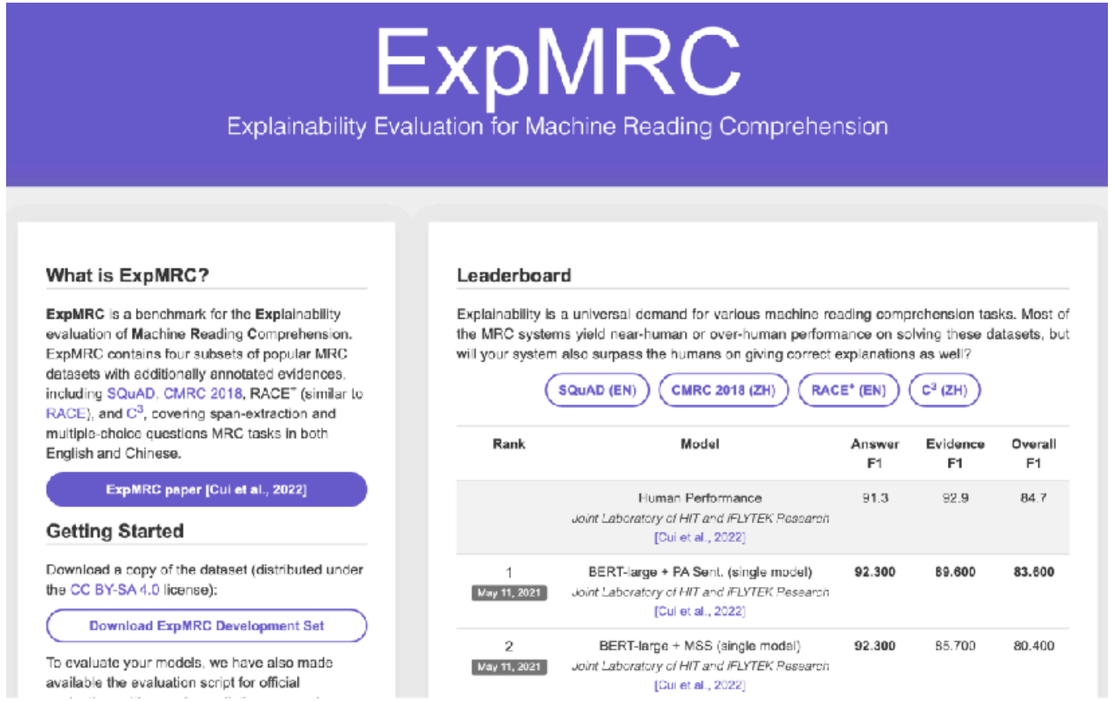
SUMMARY & PROSPECT

总结展望

- **总结**
  - 本届评测聚焦可解释性阅读理解，在给出答案的同时给出佐证依据
  - 由于没有有标注训练集，多数参赛系统应用了各种弱监督方法生成数据并结合了预训练模型
  - 最优系统相比baseline有较大性能提升，但距离人类水平还有不小差距
- **展望**
  - 机器是否真正的“理解”了文章？预训练模型到底学习到了什么？
  - 解答过程是否可信？如何进一步增加模型的可解释性？
- **欢迎持续关注CMRC系列评测，共同推动中文机器阅读理解技术研究**

开放式挑战

- **开放式挑战赛ExpMRC**
  - 开放式挑战需要通过CodaLab worksheet完成，并在隐藏的测试集上得到最终效果
  - 参赛者可任意选择四个子任务中的一个或多个来进行提交
  - 更详细的提交方法请参考评测网站：http://expmrc.hfl-rc.com/



相关资源

GitHub

- 预训练模型：BERT、RoBERTa、RBT(L)、XLNet、ELECTRA、MacBERT、PERT、少数民族语言模型CINO、**LERT**
- 开源工具包：TextBrewer、TextPruner
- 中文数据集：阅读理解、文本分类、文本纠错等
- 访问HFL典藏集了解更多：http://anthology.hfl-rc.com

Model Hub

- 提供了9大类，共计47个优质预训练模型
- 搭配🤗transformers库，快速加载各类预训练模型
- 模型库地址：http://huggingface.co/hfl

哈工大讯飞联合实验室
微信公众号

再次感谢各参赛单位的大力支持！

https://cmrc2022.hfl-rc.com