第二届“讯飞杯”中文机器阅读理解评测总结

Overview of the Second Evaluation Workshop on Chinese Machine Reading Comprehension

崔一鸣

哈工大讯飞联合实验室(HFL)，科大讯飞

2018年10月19日

概况介绍

- 第二届“讯飞杯”中文机器阅读理解评测 (The Second Evaluation Workshop on Chinese Machine Reading Comprehension, CMRC 2018)
  - 主办方：中国中文信息学会计算语言学专委会 (CIPS-CL)
  - 承办方：哈工大讯飞联合实验室 (HFL)
  - 冠名方：科大讯飞股份有限公司
- 旨在促进中文阅读理解研究及发展，并且为相关领域学者提供一个良好的交流平台

评测委员会

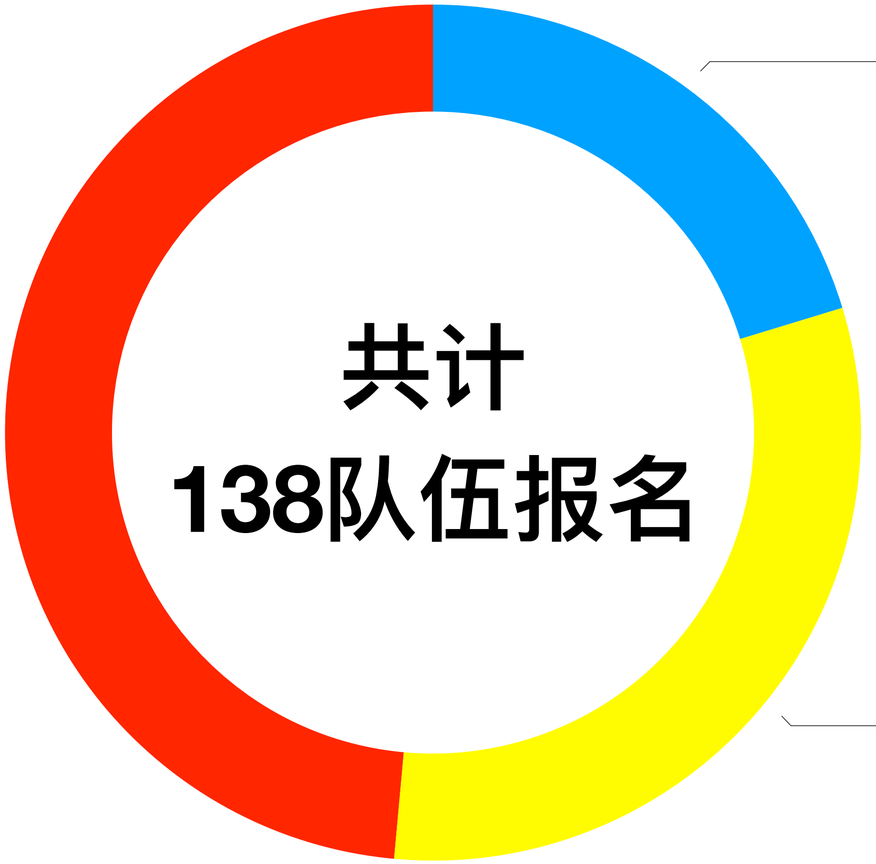
- **CCL 2018评测联合主席**
  - 刘　挺（哈尔滨工业大学）
  - 宋　巍（首都师范大学）
- **CMRC 2018评测主席**
  - 刘　挺（哈尔滨工业大学）
  - 崔一鸣（科大讯飞）

时间安排

事件	时间
发布试验集	2018.3.5
预报名	2018.4.22
报名信息确认	2018.4.23 ~ 2018.4.27
发布训练集和开发集	2018.5.7
系统搭建及调整	2018.5.7 ~ 2018.8.10
提交系统验证开发集	2018.6.7 ~ 2018.8.10
提交最终评测系统	2018.8.13 ~ 2018.8.18
撰写系统描述摘要	2018.9.30
召开评测大会	2018.10.19

报名情况

学生组队
49%



学术界研究人员
20%

工业界研究人员
31%

任务介绍

- 本届阅读理解评测的主要任务是“基于篇章片段抽取的阅读理解” (Span-Extraction Machine Reading Comprehension)
  - 作为去年填空型阅读理解评测的进一步扩展，更具有挑战性，更接近实际应用场景
  - 参赛者需要对给定的篇章进行建模并回答与篇章相关的问题
  - 答案需要从篇章中抽取出一个连续片段

[Document]
静电感应是物体内的电荷因受外界电荷的影响而重新分布。这个现象由英国科学家约翰·坎通和瑞典科学家**分别在1753年和1762年发现**。静电发电机，例如威姆斯赫斯特电机、范德格拉夫起电机和起电盘，都使用这个原理。正常的物质都带有等量的正电荷和负电荷，因此总的来说是不带电的。如果把带电的物体靠近不带电的导体，例如一片金属，则导体上的电荷将会重新分布。例如，如果把带正电的物体靠近一块金属（参见右面的图），则金属上的负电荷将会被吸引过去，而正电荷则会被排斥。这样便导致金属的靠近外界电荷的部分带有负电荷，而远离外界电荷的部分则带有正电荷。

.....

在绝缘体中，电子被原子束缚着，不能在物体中自由移动；但是在原子内可以移动一点点。如果把带正电的物体靠近绝缘体，则每一个原子中的电子都会被吸引而稍微移动一点，而原子核则会被排斥，而往相反的方向移动一点。这种现象称为极化。由于这时物体中的负电荷离外面的带电物体较近，而正电荷则距离较远，将导致吸引力比排斥力大一点点。这个现象是微观的，但因为有那么多的原子，加起来效果就很明显了，足以使较轻的物体（如小纸片）被吸引。

[Question]
静电感应是什么时候发现的?
[Answer 1]
1753年和1762年
[Answer 2]
分别在1753年和1762年发现
[Answer 3]
分别在1753年和1762年发现

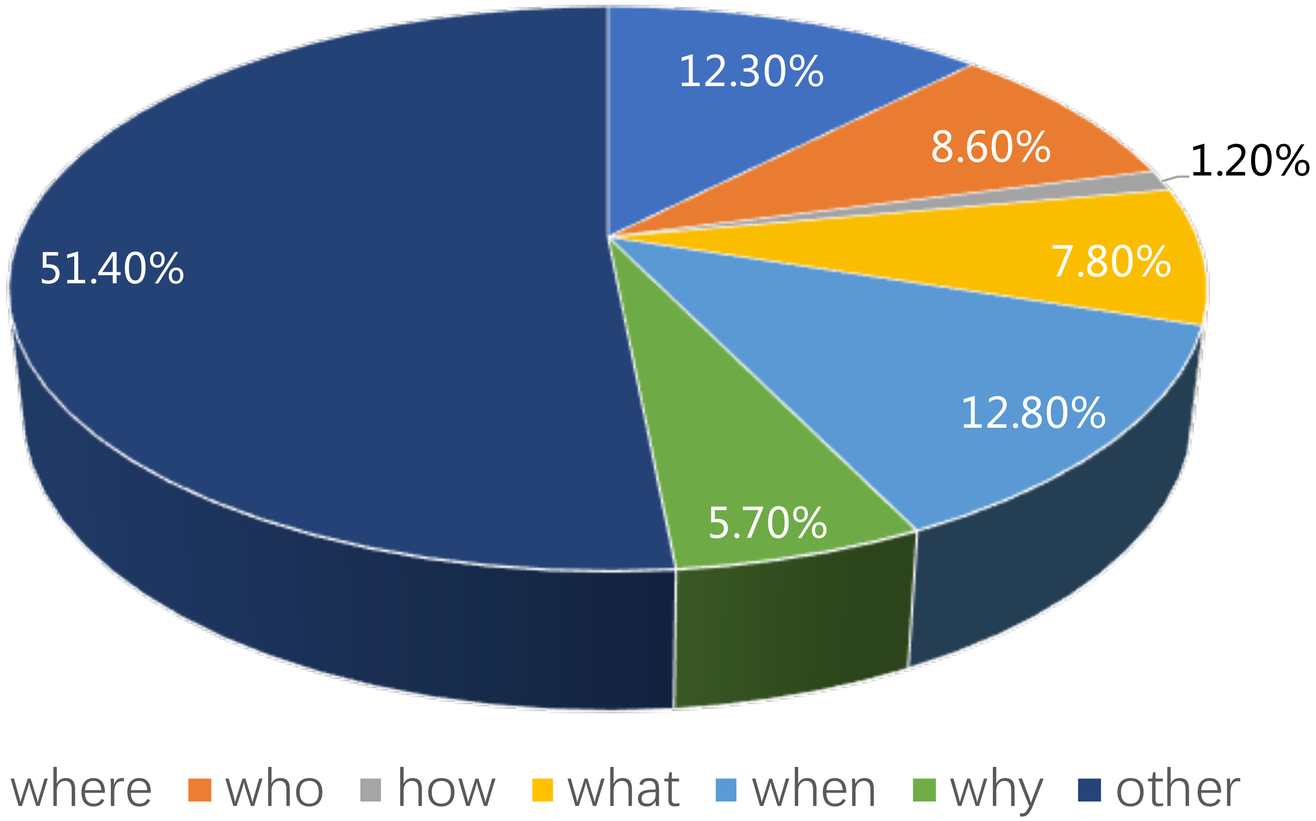
评测数据

- 本次评测数据的领域是维基百科类文本
  - 数据来源：中文维基百科dump数据（时间戳：2018年1月22日）
  - 篇章筛选：根据一定的规则进行筛选，保证篇章能被大多数人理解
  - 问题标注：由人工进行编写

	Train	Dev	Test	Challenge
Question #	10,321	3,351	4,895	504
Answer # per query	1	**3**	**3**	**3**
Max doc tokens	962	961	980	916
Max question tokens	89	56	50	47
Max answer tokens	100	85	92	77
Average doc tokens	452	469	472	464
Average question tokens	15	15	15	18
Average answer tokens	17	9	9	19

评测数据

- 问题类型分布


12.30%
8.60%
1.20%
7.80%
12.80%
5.70%
51.40%
where
who
how
what
when
why
other

评测数据

- 除了已公开的数据外，我们还保留了一部分数据作为**挑战集**
- **特点**
  - 答案不能仅从一个句子中简单推断出来
  - 如果答案是实体，则该类实体在篇章中的数量必须大于2（起到混淆效果）
  - 问题类型主要集中在why/how类型

[Document]
《黄色脸孔》是柯南·道尔所著的福尔摩斯探案的56个短篇故事之一，收录于《福尔摩斯回忆录》。孟罗先生素来与妻子恩爱，但自从最近邻居新入伙后，孟罗太太则变得很奇怪，曾经凌晨时份外出，又藉丈夫不在家时偷偷走到邻居家中。于是孟罗先生向福尔摩斯求助，福尔摩斯听毕孟罗先生的故事后，认为孟罗太太被来自美国的前夫勒索，所以不敢向孟罗先生说出真相，所以吩咐孟罗先生，如果太太再次走到邻居家时，即时联络他，他会第一时间赶到。孟罗太太又走到邻居家，福尔摩斯陪同孟罗先生冲入，却发现邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿，所以不敢说出真相。

[Question]
孟罗太太为什么在邻居新入伙后变得很奇怪?

[Answer 1]
邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿

[Answer 2]
邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿，所以不敢说出真相。

[Answer 3]
邻居家中的人是孟罗太太与前夫生的女儿，因为孟罗太太的前夫是黑人，她怕孟罗先生嫌弃混血儿，所以不敢说出真相。

评测数据

- 目前公开集合已可以在CodaLab以及Github平台下载
  - CodaLab：https://worksheets.codalab.org/worksheets/0x92a80d2fab4b4f79a2b4064f7ddca9ce/
  - Github: https://github.com/ymcui/cmrc2018
- 若后续在研究中使用本届评测的数据或对比相关结果，请引用以下文章

A Span-Extraction Dataset for Chinese Machine Reading Comprehension

Yiming Cui[†‡], Ting Liu[‡],
Li Xiao[†], Zhipeng Chen[†], Wentao Ma[†], Wanxiang Che[‡], Shijin Wang[†], Guoping Hu[†]
[†]Joint Laboratory of HIT and iFLYTEK (HFL), iFLYTEK Research, Beijing, China
[‡]Research Center for Social Computing and Information Retrieval (SCIR),
Harbin Institute of Technology, Harbin, China
[†]{ymcui,lixiao3,zpchen,wtma,sjwang3,gphu}@iflytek.com
[‡]{ymcui,tliu,car}@ir.hit.edu.cn

系统评价

- **评价指标1：精准匹配率（Exact Match，EM）**
  - 计算预测结果与标准答案是否**完全一致**
  - 一致=1分，不一致=0分
- **评价指标2：模糊匹配率（F1-score）**
  - 计算预测结果与标准答案之间的**字级别**(character-level)匹配程度
- **排名方式：单模型、多模型混合排名**
  - 按照测试集EM和F1的平均值倒序排列
  - 要求提交多模型融合结果时必须同时提交单模型结果

系统提交

- **系统提交方式：CodaLab线上提交**
  - SQuAD官方提交平台
  - 保证结果真实、有效、可复现

- **本次评测过程中测试集仅内部可见，参赛者不可见**
  - 避免参赛者针对测试集进行优化，进一步保证了评测的公平性
  - 限制参赛队伍一周只能提交一次结果

评测结果 & 获奖单位

最终评测结果

	Development		Test		Challenge	
	EM	F1	EM	F1	EM	F1
Estimated Human Performance	*91.083*	*97.348*	*92.400*	*97.914*	*90.382*	*95.248*
Z-Reader (single model)	**79.776**	**92.696**	**74.178**	**88.145**	13.889	**37.422**
MCA-Reader (ensemble)	66.698	85.538	71.175	88.090	15.476	37.104
RCEN (ensemble)	76.328	91.370	68.662	85.753	15.278	34.479
MCA-Reader (single model)	63.902	82.618	68.335	85.707	13.690	33.964
OmegaOne (ensemble)	66.977	84.955	66.272	82.788	12.103	30.859
RCEN (single model)	73.253	89.750	64.576	83.136	10.516	30.994
GM-Reader (ensemble)	58.931	80.069	64.045	83.046	**15.675**	37.315
OmegaOne (single model)	64.430	82.699	64.188	81.539	10.119	29.716
GM-Reader (single model)	56.322	77.412	60.470	80.035	13.690	33.990
R-NET (single model)	45.418	69.825	50.112	73.353	9.921	29.324
SXU-Reader (ensemble)	40.292	66.451	46.210	70.482	N/A	N/A
SXU-Reader (single model)	37.310	66.121	44.270	70.673	6.548	28.116
T-Reader (single model)	39.422	62.414	44.883	66.859	7.341	22.317
Unnamed Sys by usst (single model)	34.490	59.539	37.916	63.502	5.159	18.687
Unnamed Sys by whu (single model)	18.577	42.560	22.288	46.774	2.183	21.587
Unnamed Sys by LittleBai (single model)	7.021	31.657	10.848	37.231	0.397	9.498
Unnamed Sys by jspi (single model)	13.793	39.720	0.449	34.224	2.579	20.048

最终评测结果

	Development		Test		Challenge	
	EM	F1	EM	F1	EM	F1
Estimated Human Performance	*91.083*	*97.348*	*92.400*	*97.914*	*90.382*	*95.248*
Z-Reader (single model)	**79.776**	**92.696**	**74.178**	**88.145**	13.889	**37.422**
MCA-Reader (ensemble)	66.698	85.538	71.175	88.090	15.476	37.104
RCEN (ensemble)	76.328	91.370	68.662	85.753	15.278	34.479
MCA-Reader (single model)	63.902	82.618	68.335	85.707	13.690	33.964
OmegaOne (ensemble)	66.977	84.955	66.272	82.788	12.103	30.859
RCEN (single model)	73.253	89.750	64.576	83.136	10.516	30.994
GM-Reader (ensemble)	58.931	80.069	64.045	83.046	**15.675**	37.315
OmegaOne (single model)	64.430	82.699	64.188	81.539	10.119	29.716
GM-Reader (single model)	56.322	77.412	60.470	80.035	13.690	33.990
R-NET (single model)	45.418	69.825	50.112	73.353	9.921	29.324
SXU-Reader (ensemble)	40.292	66.451	46.210	70.482	N/A	N/A
SXU-Reader (single model)	37.310	66.121	44.270	70.673	6.548	28.116
T-Reader (single model)	39.422	62.414	44.883	66.859	7.341	22.317
Unnamed Sys by usst (single model)	34.490	59.539	37.916	63.502	5.159	18.687
Unnamed Sys by whu (single model)	18.577	42.560	22.288	46.774	2.183	21.587
Unnamed Sys by LittleBai (single model)	7.021	31.657	10.848	37.231	0.397	9.498
Unnamed Sys by jspi (single model)	13.793	39.720	0.449	34.224	2.579	20.048

人工评价结果估算方法

- 以#1标注者编写的答案为输出，#2、#3为标准答案
- 以此类推可以计算出3组评测结果
- 对三组评测结果进行平均

人工评价结果观察

- 测试集效果略高于开发集效果
- 挑战集效果低于开发集、测试集效果，但相差不大

最终评测结果

	Development		Test		Challenge	
	EM	F1	EM	F1	EM	F1
Estimated Human Performance	*91.083*	*97.348*	*92.400*	*97.914*	*90.382*	*95.248*
Z-Reader (single model)	**79.776**	**92.696**	**74.178**	**88.145**	13.889	**37.422**
MCA-Reader (ensemble)	66.698	85.538	71.175	88.090	15.476	37.104
RCEN (ensemble)	76.328	91.370	68.662	85.753	15.278	34.479
MCA-Reader (single model)	63.902	82.618	68.335	85.707	13.690	33.964
OmegaOne (ensemble)	66.977	84.955	66.272	82.788	12.103	30.859
RCEN (single model)	73.253	89.750	64.576	83.136	10.516	30.994
GM-Reader (ensemble)	58.931	80.069	64.045	83.046	**15.675**	37.315
OmegaOne (single model)	64.430	82.699	64.188	81.539	10.119	29.716
GM-Reader (single model)	56.322	77.412	60.470	80.035	13.690	33.990
R-NET (single model)	45.418	69.825	50.112	73.353	9.921	29.324
SXU-Reader (ensemble)	40.292	66.451	46.210	70.482	N/A	N/A
SXU-Reader (single model)	37.310	66.121	44.270	70.673	6.548	28.116
T-Reader (single model)	39.422	62.414	44.883	66.859	7.341	22.317
Unnamed Sys by usst (single model)	34.490	59.539	37.916	63.502	5.159	18.687
Unnamed Sys by whu (single model)	18.577	42.560	22.288	46.774	2.183	21.587
Unnamed Sys by LittleBai (single model)	7.021	31.657	10.848	37.231	0.397	9.498
Unnamed Sys by jspi (single model)	13.793	39.720	0.449	34.224	2.579	20.048

评测系统结果观察

- Z-Reader在EM显著高于MCA-Reader，但F1相差较小，可能在输出边界处理上更精细
- 在挑战集上，所有系统效果均有大幅度下降，且与人工评测结果有较大差距
- 开发集、测试集上表现较好的系统并不一定在挑战集上达到很好的效果

获奖单位

- **我们非常荣幸的宣布以下参赛单位获奖**
  - **冠军：**深圳追一科技有限公司
  - **亚军：**北京信息科技大学智能信息处理实验室
  - **季军：**6ESTATES PTE LTD

排名	时间	系统名称	Pre-Test	Test		
			Average	EM	F1	Average↓
1	2018/9/17	**Z-Reader (single)** ZhuiYi	**81.608**	**74.178**	**88.145**	**81.161**
2	2018/9/17	**MCA-Reader (ensemble)** 北京信息科技大学智能信息处理实验室	79.147	71.175	88.090	79.632
3	2018/9/17	**RCEN (ensemble)** 6ESTATES PTE LTD	77.978	68.662	85.743	77.203

冠军

深圳追一科技有限公司

杨雪峰，巨颖，徐爽，孙宁远

亚军

北京信息科技大学智能信息处理实验室

毛腾，张禹尧，郑佳，李小龙，郭泽晨，蒋玉茹

季军

6ESTATES PTE LTD

王超，李若愚，方帆

最佳单系统奖

- 评测委员会根据以下几个层面评选出最佳单系统奖
  - 系统描述报告的完备性
  - 预测试集、测试集的效果
  - 挑战集上的效果
- 征集系统描述报告阶段，共收集到4份报告
  - 其中3份来自于评测前三名
  - 另一份由于参赛队伍无法参加评测研讨会而放弃评奖

奖项	奖励
冠军	¥ 20,000 + 荣誉证书
亚军	¥ 10,000 + 荣誉证书
季军	¥ 5,000 + 荣誉证书
最佳单系统奖	¥ 10,000 + 荣誉证书

最佳单系统奖

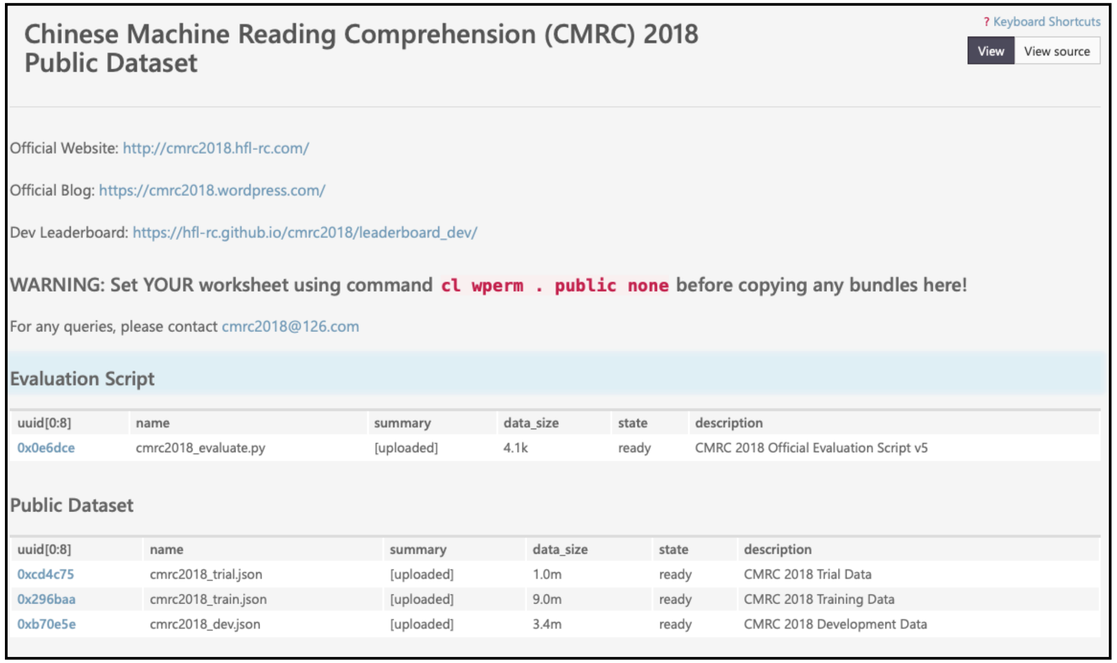
北京信息科技大学智能信息处理实验室

毛腾，张禹尧，郑佳，李小龙，郭泽晨，蒋玉茹

开放式评测

开放式评测

- 本届评测的隐藏测试集、挑战集均不对外公开
  - 以开放式评测的形式来继续评测系统效果
  - 由于CodaLab上需要提交程序代码，所得出的结果可复现，更加具有说服力
- 开放式评测流程
  - 提交方在官方开发集上跑通系统
  - 评测方在隐藏的测试集、挑战集上给出结果
- 本届参赛队伍的所有结果将复制到开放式评测榜单中
- 所有评测结果将持续更新在CMRC 2018网站



Chinese Machine Reading Comprehension (CMRC) 2018
Public Dataset

? Keyboard Shortcuts

View | View source

Official Website: http://cmrc2018.hfl-rc.com/

Official Blog: https://cmrc2018.wordpress.com/

Dev Leaderboard: https://hfl-rc.github.io/cmrc2018/leaderboard_dev/

WARNING: Set YOUR worksheet using command `cl wperm . public none` before copying any bundles here!

For any queries, please contact cmrc2018@126.com

Evaluation Script

uuid[0:8]	name	summary	data_size	state	description
0x0e6dce	cmrc2018_evaluate.py	[uploaded]	4.1k	ready	CMRC 2018 Official Evaluation Script v5

Public Dataset

uuid[0:8]	name	summary	data_size	state	description
0xcd4c75	cmrc2018_trial.json	[uploaded]	1.0m	ready	CMRC 2018 Trial Data
0x296baa	cmrc2018_train.json	[uploaded]	9.0m	ready	CMRC 2018 Training Data
0xb70e5e	cmrc2018_dev.json	[uploaded]	3.4m	ready	CMRC 2018 Development Data

总结

- **总结**
  - 在该任务中，EM指标普遍偏低，表明如何更好的决定输出答案的边界是中文阅读理解任务中需要考虑的问题
  - 从结果可以看到，虽然在测试集的结果上最佳系统与人类平均水平差距仅为10个点左右，但在挑战集合上的效果仍然有较大差距(>60个点)
  - 希望通过开放式评测看到更多的系统刷新挑战集的结果
  - 期待接下来各参赛队伍精彩的技术报告！

致谢

- **主办方**
  - 中国中文信息学会计算语言学专委会
- **承办方**
  - 哈工大讯飞联合实验室
- **冠名方**
  - 科大讯飞股份有限公司
- 感谢长沙理工大学协调安排办会场地
- 感谢科大讯飞提供数据支持

资源

- 评测官方网站
  - http://cmrc2018.hfl-rc.com/
  - https://hfl-rc.github.io/cmrc2018/
- 评测数据下载
  - https://github.com/ymcui/cmrc2018
- 评测报告下载
  - shttps://arxiv.org/abs/1810.07366

再次感谢各参赛单位的大力支持！