第三届“讯飞杯”中文机器阅读理解评测总结

Overview of the 3rd Evaluation Workshop on Chinese Machine Reading Comprehension

CMRC 2019评测委员会

2019年10月19日，云南昆明

往届评测回顾

概况介绍

- **“讯飞杯”中文机器阅读理解评测 (Evaluation Workshop on Chinese Machine Reading Comprehension)**
  - 始于2017年，CCL大会的创始评测活动
  - 国际范围内最早的机器阅读理解评测研讨会
  - 已成功举办两届：CMRC 2017, CMRC 2018

CMRC

- **第一届“讯飞杯”中文机器阅读理解评测**
  - CCL大会首次举办评测活动
  - 任务：填空型阅读理解
  - 该数据集已发表在LREC 2018
- **第二届“讯飞杯”中文机器阅读理解评测**
  - 任务：篇章片段抽取型阅读理解
  - 该数据集已发表在EMNLP 2019

Dataset for the First Evaluation on Chinese Machine Reading Comprehension

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

A Span-Extraction Dataset for Chinese Machine Reading Comprehension

Yiming Cui[†‡], Ting Liu[†], Wanxiang Che[†],
Li Xiao[†], Zhipeng Chen[†], Wentao Ma[†§], Shijin Wang[†§], Guoping Hu[†]
[†]Research Center for Social Computing and Information Retrieval (SCIR),
Harbin Institute of Technology, Harbin, China
[‡]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
[§]iFLYTEK AI Research (Hebei), Langfang, China

CMRC

- **CMRC 2018 开放式评测挑战**
  - 以SQuAD为代表的英文阅读理解数据集上，机器已超过人类平均水平
  - 但中文系统上，机器距离人类平均水平还有一定的距离
  - CMRC2018评测结束后，评测委员会继续接收测试系统，以测试模型在**隐藏的“挑战集”**上的效果

CMRC 2018 Open Challenge

Season: System Evaluation for CMRC 2018 Open Challenge

Leaderboard

CMRC 2018 challenge set requires comprehensive reasoning over **multiple clues** in the passage, while keeping the original *span-extraction* format, which is far more challenging than the test set. Will your system surpass the humans on this task?

Rank	Date	System	TST-EM	TST-F1	CHL-EM	CHL-F1
		Human Performance CMRC 2018 Officials	*92.400*	*97.914*	*90.382*	*95.248*
1	2019/8/19	**XLNet-mid (single model)** Joint Laboratory of HIT and iFLYTEK Research https://github.com/ymcui/Chinese-PreTrained-XLNet	69.300	89.200	**29.100**	**55.800**
2	2019/5/1	**Dual BERT (w/ SQuAD) (single model)** Joint Laboratory of HIT and iFLYTEK Research https://arxiv.org/abs/1909.00361	73.800	**90.200**	27.800	55.200
3	2019/9/10	**RoBERTa-wwm-ext (single model)** Joint Laboratory of HIT and iFLYTEK Research https://github.com/ymcui/Chinese-BERT-wwm	72.600	89.400	26.200	51.000
4	2019/5/1	**Dual BERT (single model)** Joint Laboratory of HIT and iFLYTEK Research https://arxiv.org/abs/1909.00361	70.400	88.100	23.800	47.900
5	2019/7/30	**BERT-wwm-ext (single model)** Joint Laboratory of HIT and iFLYTEK Research https://github.com/ymcui/Chinese-BERT-wwm	71.400	87.700	24.000	47.300
6	2019/6/20	**BERT-wwm (single model)** Joint Laboratory of HIT and iFLYTEK Research https://arxiv.org/abs/1906.08101	70.500	87.400	21.000	47.000
7	2019/3/27	**P-Reader (single model)** swjtu_PF	66.189	84.386	16.079	39.683

本届评测概况

CMRC 2019

- **第三届“讯飞杯”中文机器阅读理解评测 (The Third Evaluation Workshop on Chinese Machine Reading Comprehension, CMRC 2019)**
  - 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)
  - 赞助商：科大讯飞股份有限公司
- 旨在促进中文阅读理解研究，并且为相关领域学者提供一个良好的交流平台

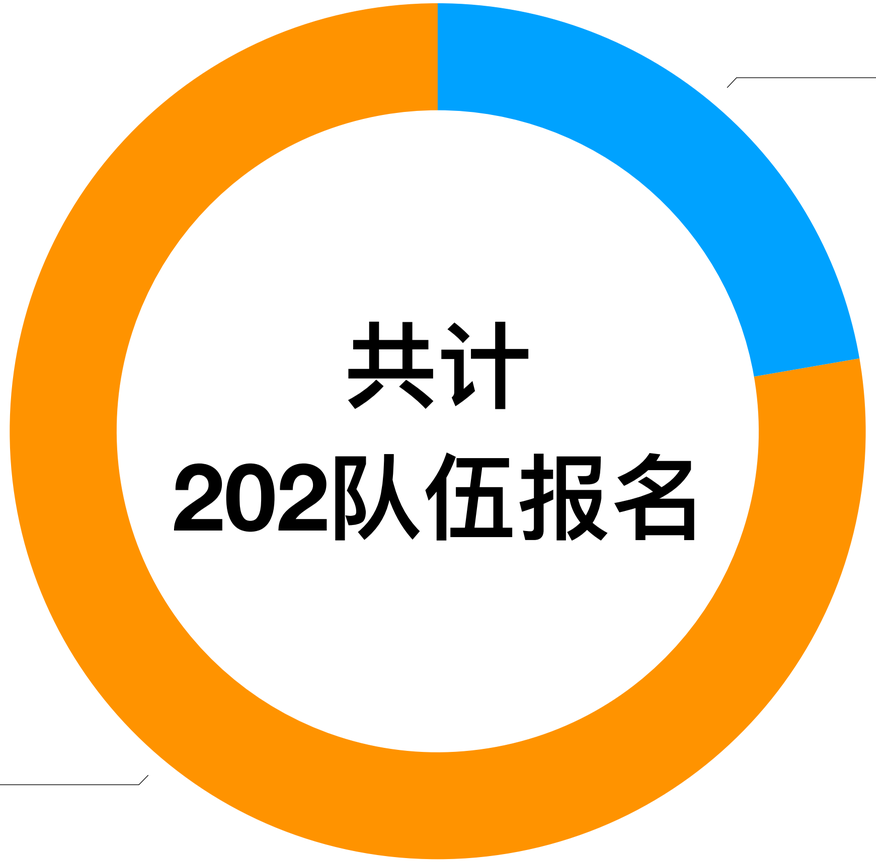
评测委员会

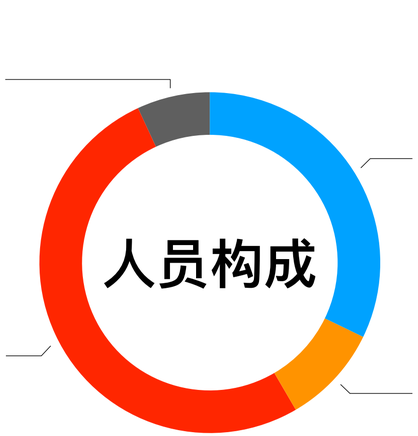
- **CCL 2019 评测主席**
  - 刘　挺（哈尔滨工业大学）
  - 徐睿峰（哈尔滨工业大学）
  - 宋　巍（首都师范大学）
- **CMRC 2019 评测主席**
  - 刘　挺（哈尔滨工业大学）
  - 崔一鸣（科大讯飞股份有限公司）

答辩委员会

- **主席**
  - 孙茂松（清华大学）
- **委员**
  - 林鸿飞（大连理工大学）
  - 车万翔（哈尔滨工业大学）
  - 邱锡鹏（复旦大学）
  - 崔一鸣（科大讯飞股份有限公司）

报名情况





奖项设置

- **CMRC 2019 评测奖项设置**
  - 根据**客观评测**以及**专家评审**结果综合评选出获奖单位
  - 中国中文信息学会计算语言学专委会（CIPS-CL）提供荣誉证书
  - 科大讯飞股份有限公司提供奖金

奖项	数量	奖励
冠军	一名	¥20,000 + 荣誉证书
亚军	一名	¥10,000 + 荣誉证书
季军	三名	¥ 5,000 + 荣誉证书

时间安排

阶段	事件	时间
赛前	参赛报名	2019年5月23日 ~ 2019年6月30日
资格赛	发布训练集、试验集	2019年5月23日
	发布开发集	2019年6月10日
	系统搭建及调整	2019年5月23日 ~ 2019年7月31日
	发布资格集，获取决赛资格	2019年8月1日 ~ 2019年8月7日
决赛	提交最终评测系统	2019年8月14日 ~ 2019年8月21日
赛后	公布客观结果排名	2019年9月中旬
	撰写系统描述报告	2019年9月下旬
	召开CMRC 2019评测研讨会	2019年10月19日

比赛阶段

- **CMRC 2019 客观评测共设置3个比赛环节**
  - **开发阶段（5月23日~7月31日）**
    - 选手可以任意提交开发集上的实验结果，不计入成绩
  - **资格赛（8月1日~8月7日）**
    - 每个队伍可至多提交10次结果，取前10名进入到决赛环节
  - **决赛（8月14日~8月21日）**
    - 每个队伍提交最终评测系统，得到最终的客观评测结果

评测任务

任务介绍

- 通常，阅读理解任务包含如下几个要素<P, Q, C, A>
  - **P**assage：需要机器阅读的篇章
    - 根据篇章数量，分为单文档阅读理解、多文档阅读理解等
  - **Q**uestion：根据篇章内容所提出的问题
    - 根据问题类型，分为填空型、用户提问型等
  - **C**andidate：候选答案
    - 根据任务类型，会有一些候选答案，如选择型阅读理解等
  - **A**nswer：答案
    - 根据任务类型，可能是单个词、篇章片段、生成的句子等

RACE

Passage

Is it important to have breakfast every day? A short time ago, a test was given in the United States. People of different ages, from 12 to 83, were asked to have a test. During the test, these people were given all kinds of breakfast, and sometimes they got no breakfast at all. ...

Question

What do the results show?

Candidates

A ***They show that breakfast has affected on work and study.***
B Breakfast has little to do with a person's work.
C A person will work better if he only has fruit and milk.
D They show that girl students should have less for breakfast.

任务介绍

- **本届评测任务：句子级填空型阅读理解**（Sentence Cloze-style Machine Reading Comprehension, SC-MRC）
  - 篇章：带有空缺位置的文本
  - 问题：空缺位置（需填入对应的句子）
  - 候选：若干个候选句子
    - 真：该句子可填入某一空缺位置中
    - 假：该句子并不属于任何一个空缺
  - 答案：将候选句子正确填入，给出句子顺序

篇章	森林里有一棵大树，树上有一个鸟窝。[BLANK1]，还从来没有看到过鸟宝宝长什么样。 小松鼠说：“我爬到树上去看过，鸟宝宝光溜溜的，身上一根羽毛也没有。”“我不相信，”小白兔说，“所有的鸟都是有羽毛的。” “鸟宝宝没有羽毛。”小松鼠说，“你不信自己去看。” 小白兔不会爬树，它没有办法去看。小白兔说：“我请蓝狐狸去看一看，我相信蓝狐狸的话。”小松鼠说：“蓝狐狸跟你一样，也不会爬树。” 蓝狐狸说：“我有魔法树叶，我能变成一只狐狸鸟。”[BLANK2]，一下子飞到了树顶上。 “蓝狐狸，你看到了吗？”小白兔在树下大声喊。 “我看到了，鸟窝里有四只小鸟，他们真是光溜溜的，一根羽毛也没有。”蓝狐狸说。 就在这时候，鸟妈妈和鸟爸爸回来了， [BLANK3]，立刻大喊大叫：“抓强盗啊！抓强盗啊！强盗闯进了我们家里，想偷我们的孩子！” [BLANK4]，全都飞了过来。他们扇着翅膀，朝蓝狐狸冲过来，用尖尖的嘴啄他，用爪子抓他。 蓝狐狸扑扇翅膀，赶紧飞。 鸟儿们排着队伍，紧紧追上来。[BLANK5]，它飞得不高，也飞得不快。“救命啊，救命！”蓝狐狸说，“我不是强盗，我是蓝狐狸！” 小白兔在草丛说：“你不是鸟，你飞不过他们，你赶快变回狐狸吧！”蓝狐狸急忙落到地上，变回了狐狸，躲进深深的草丛里。 鸟儿们找不到蓝狐狸，只得飞走了。 蓝狐狸对小白兔说：“谢谢你。
候选	0: 狐狸是第一次变成狐狸鸟 1: 森林里所有的鸟听到喊声 2: 他们看到鸟窝里蹲着一只蓝色的大鸟 3: 蓝狐狸真的变成了一只蓝色的大鸟 4: 小动物们只看到过鸟妈妈和鸟爸爸在鸟窝里飞进飞出 5: 小松鼠变成了一只蓝色的大鸟（假选项）
答案	[4, 3, 2, 1, 0]

评测数据

数据	篇章数	问题数	是否包含假选项	是否公开
试验集	139	1,504	否	是
训练集	9,638	100,009	否	是
开发集	300	3,053	是	是
资格集	500	5,081	是	仅篇章问题
测试集	500	5,118	是	否

评价标准

- **主评价指标：问题级准确率（Question-level ACcuracy, QAC）**
  - 计算问题级别的准确率

$$\mathbf{QAC} = \frac{\text{答对的问题数}}{\text{总问题数}}$$

- **次评价指标：篇章级准确率（Passage-level ACcuracy, PAC）**
  - 计算篇章级别的准确率（一个篇章的所有问题答对）

$$\mathbf{PAC} = \frac{\text{完全答对的篇章数}}{\text{总篇章数}}$$

- 客观评测结果以QAC为主要指标，相同时则比较PAC

基线系统

- **基于BERT的基线系统**
  - 使用中文BERT搭建了简单的基线系统供参赛者参考
  - 决赛阶段，提供了Codalab代码提交的样例

数据	问题级准确率 QAC	篇章级准确率 PAC
开发集	70.586%	13.333%
资格集	70.006%	8.200%
测试集	69.969%	9.400%

系统提交

- **开发&资格赛阶段**
  - 使用的是CodaLab Competition平台
  - 选手只需提交结果输出文件，无需提交代码
  - 排行榜实时更新
- **决赛阶段**
  - 使用的是CodaLab Worksheet平台
  - 选手必须提交可执行程序，并在开发集上跑通
  - 由评测委员会运行相关代码，得到测试集上的最终结果

客观评测结果

客观评测结果

- 资格赛成绩

Results						
#	User	Entries	Date of Last Entry	Team Name	QAC ▲	PAC ▲
1	ewrfcas	9	08/07/19	Gammalab	85.67211 (1)	43.60000 (1)
2	sixEstates	7	08/07/19	6Estates	83.84176 (2)	36.60000 (2)
3	RichardRui	5	08/07/19	匡扶汉室	83.42846 (3)	36.40000 (3)
4	Decalogue	4	08/07/19	ZZ	78.07518 (4)	24.60000 (4)
5	huifei	2	08/07/19	飞天神器	76.69750 (5)	13.60000 (8)
6	wyxlzsq	9	08/07/19	雨沐车车	75.89057 (6)	16.60000 (5)
7	ndsc01	6	08/05/19		75.33950 (7)	14.40000 (7)
8	X-X	9	08/07/19		75.10333 (8)	15.40000 (6)
9	niji123	10	08/07/19	Wall.E	74.41449 (9)	12.40000 (10)
10	nkuzhangyi	9	08/07/19	nkuzhangyi	73.01712 (10)	13.40000 (9)
11	HelloNil	5	08/07/19	33lab	72.87935 (11)	11.00000 (12)
12	berton	9	08/07/19	baseline	71.40327 (12)	11.60000 (11)
13	luckyboys	5	08/05/19	luckyboys	71.26550 (13)	10.80000 (13)
14	CMRC_2019	2	08/01/19	CMRC 2019 Evaluation Committee	70.00590 (14)	8.20000 (15)
15	HKP	7	08/07/19	Messi	68.90376 (15)	9.80000 (14)
16	renxingkai	2	08/06/19	RedCat7	68.45109 (16)	8.00000 (16)

前十名进入决赛

客观评测结果

- **决赛成绩：7个队伍的成绩有效**

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	90.9269570914	58.2	90.7892147215	57.6	90.0547088707
2	**mojito system (ensemble)** SFTech	48.0	88.2083196856	43.4	86.459358394	41.8	85.9906213365
3	**DA-BERT (ensemble)** 百度	34.3	86.3413036358	29.2	84.9045463491	27.6	84.4470496288
4	**CMRC2019 MULTIPLE BERT (ensemble)** Six Estates https://www.6estates.com	38.7	82.9675728791	35.6	83.507183625	32.2	82.5908558030407
5	**nkuzhangyi_cmrc_v2 (ensemble)** CICC	29.7	80.9367834916	26.0	80.318834875	26.6	79.5623290348
6	**MRC-ZZ SYSTEM (single model)** 哈工大&汉仪字库	29.0	80.3799541435	25.8	78.2916748672	26.6	78.7807737397
7	**MB-Reader (ensemble)** ECUST	18.7	78.2181460858	17.8	76.4219641803	15.6	76.3188745604

现场答辩环节

评分标准

- **答辩评分标准（每一项20分）**
  - **任务理解**：对本次评测的任务理解是否充分，对相关工作是否有详细调研
  - **系统设计**：从系统整体上考虑设计是否合理，是否运用了科学的方法解决问题
  - **技术创新**：除了现有的技术手段之外，是否有模型创新的部分，创新是否合理，是否具有先进性
  - **实验结果**：对实验结果的量化分析，消融实验等来证明系统的有效性
  - **语言学分析**：是否有错误样例分析等语言学相关分析，从阅读理解任务的角度进行分析

打分方法

- **答辩环节得分：所有专家评分的平均值*20%**

打分范围	对应5分制	评判标准
17~20	**5 (strong accept)**	**优秀**：技术完备，分析透彻且深入，具有创新性、先进性
13~16	**4 (accept)**	**良好**：技术方案或分析具备一定的先进性，超过平均水平
9~12	**3 (borderline)**	**中等**：达到了基本要求，方案尚可，但无明显亮点
5~8	**2 (reject)**	**略差**：存在明显技术漏洞或分析不充分
0~4	**1 (strong reject)**	**差**：基本未涉及到相关打分项

答辩队伍

- **答辩顺序确定方法**
  - 客观评测折算后的成绩*1000%2019，按降序排列

出场顺序	队伍	客观评测（80%）	伪随机结果
1	6ESTATES PTE LTD	66.073	1465
2	PINGAN-GammaLab	72.044	1379
3	CICC	63.650	1061
4	哈工大&汉仪字库	63.025	436
5	SFTech	68.792	146

最终成绩

名次	队伍	客观评测（80%）	专家评审（20%）	最终成绩
1	PINGAN-GammaLab	72.044	17.36	**89.434**
2	SFTech	68.792	15.68	**84.472**
3	6ESTATES PTE LTD	66.073	16.2	**82.273**
4	CICC	63.650	14.64	**78.290**
5	哈工大&汉仪字库	63.025	13.4	**76.425**

获奖单位

平安金融壹账通

顺丰科技有限公司

6ESTATES PTE LTD

中金公司（CICC）

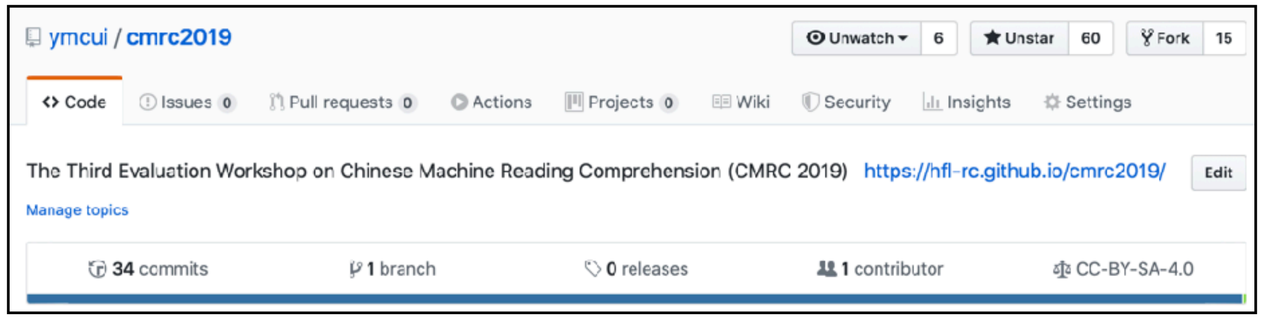
哈工大&汉仪字库

总结

- **总结**
  - 中文机器阅读理解在近两年得到了更多关注
  - 机器客观指标不断刷新记录，但
    - 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？如何增加模型的可解释性？
    - 如何解决不同语种间阅读理解系统性能的差异？
- 希望大家持续关注CMRC系列评测，共同推进中文信息处理技术前进

资源

- 评测官方网站
  - http://cmrc2019.hfl-rc.com/
  - https://hfl-rc.github.io/cmrc2019/
- 评测数据下载
  - https://github.com/ymcui/cmrc2019



资源

- **中文预训练系列模型**
  - BERT、RoBERTa系列：https://github.com/ymcui/Chinese-BERT-wwm
  - XLNet系列：https://github.com/ymcui/Chinese-PreTrained-XLNet

BERT/RoBERTa

XLNet

致谢

- **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）
- **赞助商**：科大讯飞股份有限公司
- 感谢答辩委员会的各位评审专家
- 感谢昆明理工大学协调安排办会场地
- 感谢参加本届评测研讨会的所有老师和同学

再次感谢各参赛单位的大力支持！

预祝评测研讨会圆满成功！