科大讯飞 iFLYTEK HIT-SCIR

文档密级：外部公开

第一届“讯飞杯”中文机器阅读理解评测总结

崔一鸣

哈工大讯飞联合实验室，科大讯飞

2017年10月14日

概况介绍

- 第一届“讯飞杯”中文机器阅读理解评测 (The First Evaluation Workshop on Chinese Machine Reading Comprehension)
  - 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)
  - 冠名商：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沟通平台

评测委员会

- **评测大会联合主席（Evaluation co-Chairs）**
  - 刘挺（哈尔滨工业大学）
  - 王士进（科大讯飞）
- **评测委员（Evaluation Members）**
  - 崔一鸣（科大讯飞）
  - 刘铭（哈尔滨工业大学）

时间安排

事件	时间
预报名	2017.4.5 ~ 2017.4.17
正式报名	2017.4.19 ~ 2017.4.25
发布训练集和开发集	2017.5.3
系统搭建及调整	2017.5.3 ~ 2017.8.13
提交系统验证开发集	2017.7.13 ~ 2017.8.13
提交系统验证测试集	2017.8.14 ~ 2017.8.18
撰写系统描述摘要	2017.8.31
召开评测大会	2017.10.14

任务介绍

- 本届阅读理解评测的主要任务是“填空型阅读理解” (Cloze-style Reading Comprehension)
  - 即限定答案是**篇章中出现的单个词**
- 参赛者需要对给定的篇章进行分析并回答与篇章相关的问题
- 根据问题的形式，分为两个评测任务
  - 填空型任务
  - 用户提问型任务

任务介绍

- **Track 1: 填空类问题**
  - 篇章：带有单词空缺的篇章
  - 问题：空缺词所在的句子
  - 答案：单个词（主要以名词和名实体为主）
  - 特点：训练、开发、测试集的形式完全一致

Cloze Track
1 ||| 为了 让 森林 变 得 更加 茂盛 ， 大伙 都 在 努力 地 工作 着 。 2 ||| **XXXXX** 每天 天 不 亮 就 起来 ， 在 这 棵 树 上 啄啄 ， 在 那 棵 树上 敲敲 ， 他 的 尖利 的 长嘴 ， 使 害虫 没有 藏身 之 地 。 3 ||| 小 猴 呢 ， 整天 从 这 棵 树 跳 到 那 棵 树 ， 他 手脚 不 停 ， 把 缠 在 树干 上 的 细藤 扯 下来 ， 让 树木 长 得 更 茂盛 一点 。 4 ||| 小 松鼠 更 忙 了 ， 他 用 自己 的 尖牙 把 一颗颗 松果 咬 下来 ， 然后 再 用 爪子 把 土刨 开 ， 把 松球 埋 下去 —— 他 在 植树 呢 ! 5 ||| 惟有 大象 没 活 干 ， 他 整天 游荡 ， 大家 问 他 为什么 不 干活 ， 他 说 ： “ 我 没有 啄木鸟 的 长嘴 ， 也 没有 猴子 的 巧手 和 松鼠 的 尖牙利爪 ， 我 能 干 什么 呢 ? 6 ||| ” 有 一 天 ， 大象 被 地上 的 枯树 干绊 了 一跤 ， 他 气 极了 ， 就 用 鼻子 卷 起 枯树干 ， 把 它 扔 得 远远 的 。 7 ||| 就 在 这 一刹那 ， 大象 发现 了 自己 有 鼻子 ， 有 细长 而 有力 的 鼻子 。 8 ||| 他 高兴 地 告诉 啄木鸟 、 猴子 、 松鼠 ： “ 我 发现 我 有 鼻子 …… 9 ||| ” 大伙 不 明白 ， 都 笑 了 起来 ： “ 你 本来 就 有 鼻子 ! 10 ||| ” “ 对 ， 但 我 忘 了 我 有 鼻子 ， 忘 了 我 的 鼻子 也 能 干活 ! 11 ||| ” 大象 高兴 地 说 。 12 ||| 于是 ， 他 用 鼻子 卷 走 了 森林 里 横七竖八 的 枯树干 ， 给 小 松鼠 腾 出 了 更 多 的 播种 的 地方 。 13 ||| 不久 ， 在 原先 堆 着 枯树干 的 地方 ， 长 出 一 支 支 小 绿苗 。 14 ||| 大伙 夸奖 大象 有 一 只 多么 能干 的 鼻子 。
XXXXX 每天 天 不 亮 就 起来 ， 在 这 棵 树 上 啄啄 ， 在 那 棵 树上 敲敲 ， 他 的 尖利 的 长嘴 ， 使 害虫 没有 藏身 之 地 。
啄木鸟

任务介绍

- **Track 2: 用户提问类问题**
  - 篇章：完整篇章
  - 问题：人工标注的问题
  - 答案：单个词（主要以名词和名实体为主）
  - 特点：训练集与最终的开发/测试集不一样，参赛者需要考虑类型迁移的问题

User Query Track
1 ||| 一天 ， 阿凡提 在 头 上 缠 了 一 顶 非常 华丽 而 庄重 的 色兰 进 了 王宫 。 2 ||| 他 心想 国王 肯定 会 羡慕 他 这 顶色兰 ， 没准儿 还 会 出 大 价钱 买 下 这个 色兰 。 3 ||| “ 阿凡提 你 花 了 多少 钱 买 的 这 顶 色兰 ？ 4 ||| ” 国王 果然 很 感 兴趣 地 问道 。 5 ||| “ 陛下 ， 我 用 了 一千 枚 金币 买 下 了 这 顶 高贵 的 色兰 。 6 ||| ” 阿凡提 回答 。 7 ||| 国王 的 宰相 看出 了 阿凡提 的 企图 ， 悄声 对 国王 说 ： “ 只有 傻瓜 才 会 用 一千 枚 金币 买 这 顶色兰 。 8 ||| ” “ 阿凡提 ， 我 从未 听说 过 一 顶 色 兰值 一千 枚 金币 ， 你 为什么 出 这么 大笔 钱 买 它 呢 ？ 9 ||| ” “ 尊贵 的 陛下 ， 因为 我 知道 全世界 只有 一 位 国王 知道 它 的 价值 ， 并且 会 把 它 买 下来 。 10 ||| ” 阿凡提 神秘 地 说道 。 11 ||| 国王 为 显示 自己 高贵 ， 果然 花 了 一千 枚 金币 买 下 这 顶色兰 ， 并 沉醉 于 赞美 和 恭维 之中 。 12 ||| 阿凡提 拿 上 这 一千 枚 金币 后 ， 来到 那位 宰相 跟前 ， 悄悄 地 对 他 说道 ： “ 您 确实 知道 这 顶色兰 的 价值 ， 但 我 却 懂得 国王 的 弱点 。 ”
阿凡提 在 头 上 缠 了 一 顶 非常 华丽 而 庄重 的 什么 进 了 王宫 ？
色兰

评测数据

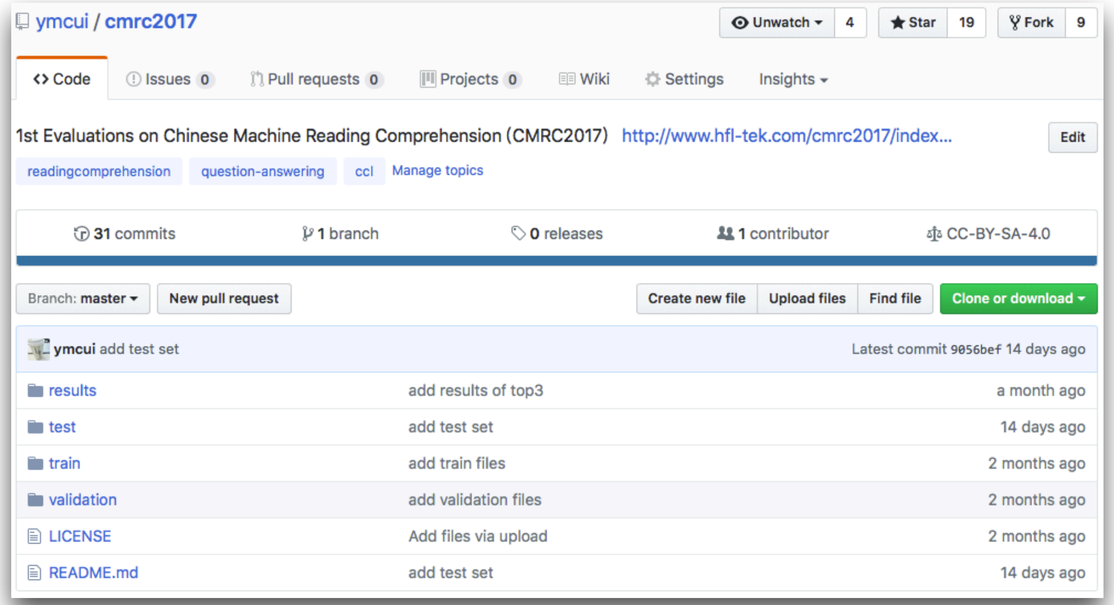
- 本次评测数据的领域是儿童读物
  - 数据来源：内部收集的20,000个儿童读物篇章
- 数据处理过程（自动生成问题）
  - 利用LTP(Che et al., 2010)对篇章进行分句、分词、POS以及依存分析
  - 抽取COO/SBV/VOB等依存关系，并只保留有依存的部分且这些词不能是代词或动词
  - 筛选出的词频必须大于等于2

评测数据

- 本次评测使用的所有开发集和测试集均经过人工筛选或标注
  - 填空型：机器自动生成，人工筛选
  - 用户提问型：完全由人工提问
- 其他一些限制（部分）
  - 篇章或者问题是否合适
  - 避免简单的语言模型选出答案
  - 每个篇章只允许[筛选出]/[提问]有限个问题（填空类1，用户提问5）

评测数据

- 目前所有数据已公开在Github上
  - https://github.com/ymcui/cmrc2017
- LICENCE
  - CC-BY-SA-4.0
- International Standard Language Resource Number (ISLRN)
  - 451-824-550-408-2



- BIBTEX

```
@article{cmrc2017-dataset,
  title={Dataset for the First Evaluation on Chinese Machine Reading Comprehension},
  author={Cui, Yiming and Liu, Ting and Chen, Zhipeng and Ma, Wentao and Wang, Shijin and Hu, Guoping},
  journal={arXiv preprint arXiv:1709.08299},
  year={2017}
}
```

系统评价

- 评价指标：答案准确率（Exact Match）

准确率 = 预测正确的样本数 / 总样本数

- 排名方式：单模型、多模型混合排名
  - 要求提交多模型融合结果时必须同时提交单模型结果

系统评价

- **系统提交方式**
  - Codalab线上提交（快捷）
  - 线下提交给评测委员会测试（较慢）
- **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
  - 赛后已将评测数据公开，进一步保证评测结果透明化

–	填空类问题	用户提问类问题
训练集类型	填空型	填空型
开发集类型	填空型	用户提问型
测试集类型	填空型	用户提问型
侧重点	系统本身的优化	类型迁移

参赛者不可见

报名概况

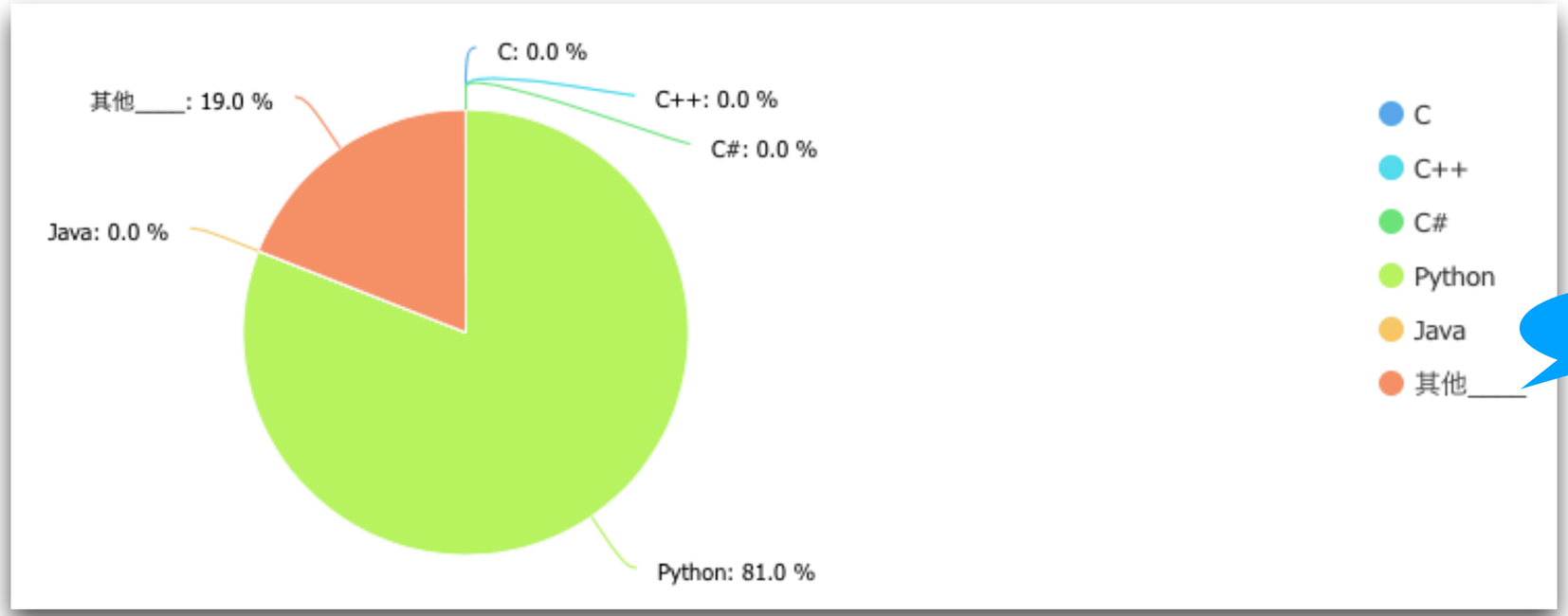
- 2017年4月份发布了本次评测的通知，共收到35份报名
  - 由下表可见，填空类任务相对来说上手难度较低，最终留存率也较高；相对较难的用户提问类最终只有3家单位提交结果

任务	正式报名阶段	最终提交阶段
填空类	报名：30 留存率：N/A	报名：14 留存率：14/30=46.7%
用户提问类	报名：26 留存率：N/A	报名：3 留存率：3/26=11.5%

- 留存率 = 本阶段人数 ÷ 上一阶段人数

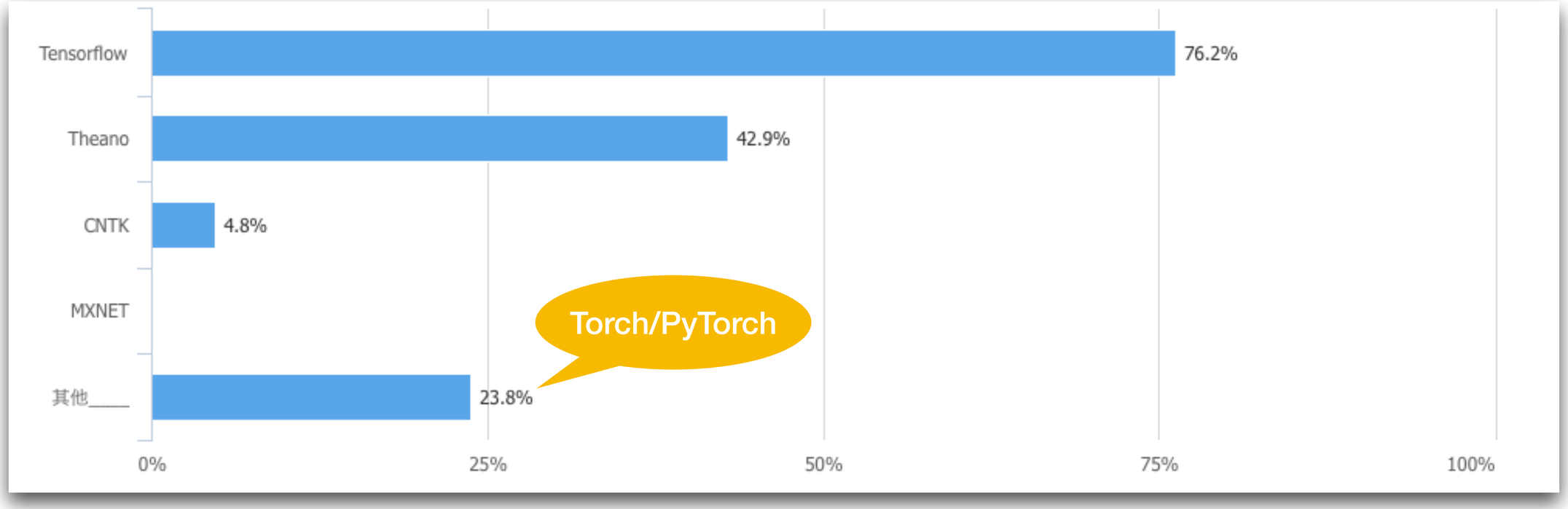
报名概况

- 对参赛者的系统概况进行了调查，部分统计结果如下（样本数：21）
  - 首选编程语言：Python压倒性的胜利，其次是Lua



报名概况

- 对参赛者的系统概况进行了调查，部分统计结果如下（样本数：21）
  - 首选深度学习库：Tensorflow > Theano > PyTorch



参赛系统分析

- 根据参赛队伍提交的报告，得出如下汇总信息（填空类任务）

	6ESTATES	上海交大	云思创智	武汉大学
系统结构	神经网络	神经网络	神经网络	神经网络
词向量	Word-level word2vec Char-level	Word & Char-level	Word-level word2Vec	Word-level GloVe Char-level
篇章建模	BiGRU	BiGRU	BiGRU	BiRNN
Attention计算	Attention-over-Attention Gated Attention	Gated Attention	Gated Attention	Gated Self-Matching
答案预测	Pointer Network	Pointer Network	Pointer Network	Pointer Network
模型融合	Probability Accumulation	-	Bagging	-

参赛系统分析

- 根据参赛队伍提交的报告，得出如下汇总信息（用户提问类任务）

	华东师大	山西大学3队	郑州大学
系统结构	神经网络	**传统NLP**	神经网络
词向量	word/POS/NER Alignment feature	计算相似度	Word-level
篇章建模	BiGRU	-	BiGRU
Attention计算	Gated Attention	-	Attention-over-Attention Gated Attention
答案预测	Pointer Network	基于依存句法分析抽取	Pointer Network
解决方案	Pre-training → Fine-tuning	通过传统NLP对篇章分析	Pre-training → Fine-tuning

参赛系统分析

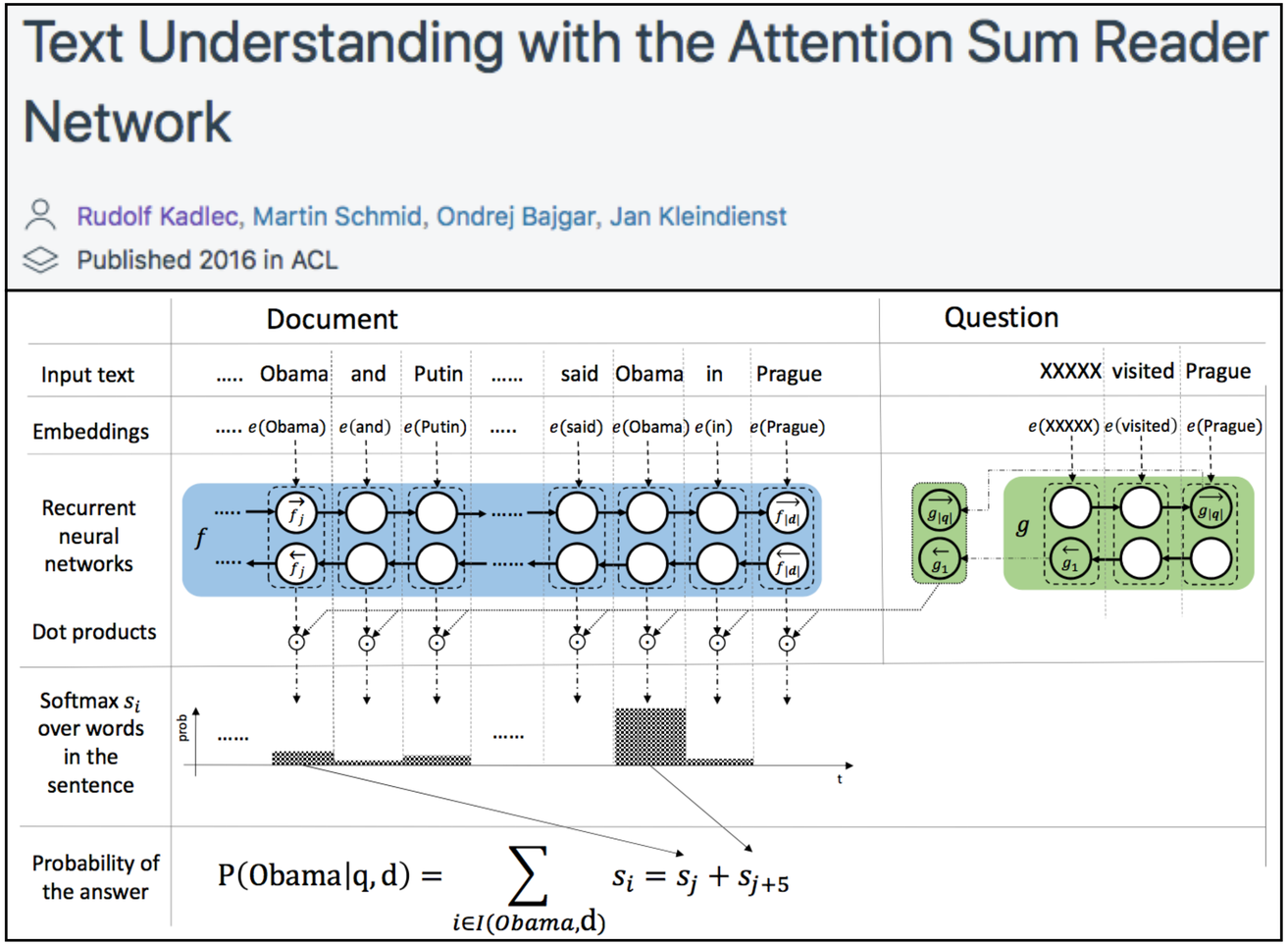
- **填空类任务**
  - 几乎所有参赛队伍均使用了神经网络系统
  - 在Attention结构上，多数借鉴了Gated-Attention以及Attention-over-Attention结构
- **用户提问类任务**
  - 少数参赛队伍考虑将传统NLP特征加入到神经网络系统中，甚至采用传统NLP为主的系统结构
  - 均考虑到了问题的迁移问题，多数采用Pre-training → Adaptation的方案

基线系统

- 我们一共提供了4个基线系统，其中包括两个简单系统和两个神经网络系统
- **简单系统**
  - Random Guess: 从篇章中随机选择一个词作为答案
  - Top Frequency: 从篇章中选择频次最高的词作为答案
- **神经网络系统**
  - Attention Sum Reader (Kadlec et al., 2016)
  - Attention-over-Attention Reader (Cui et al., 2017)

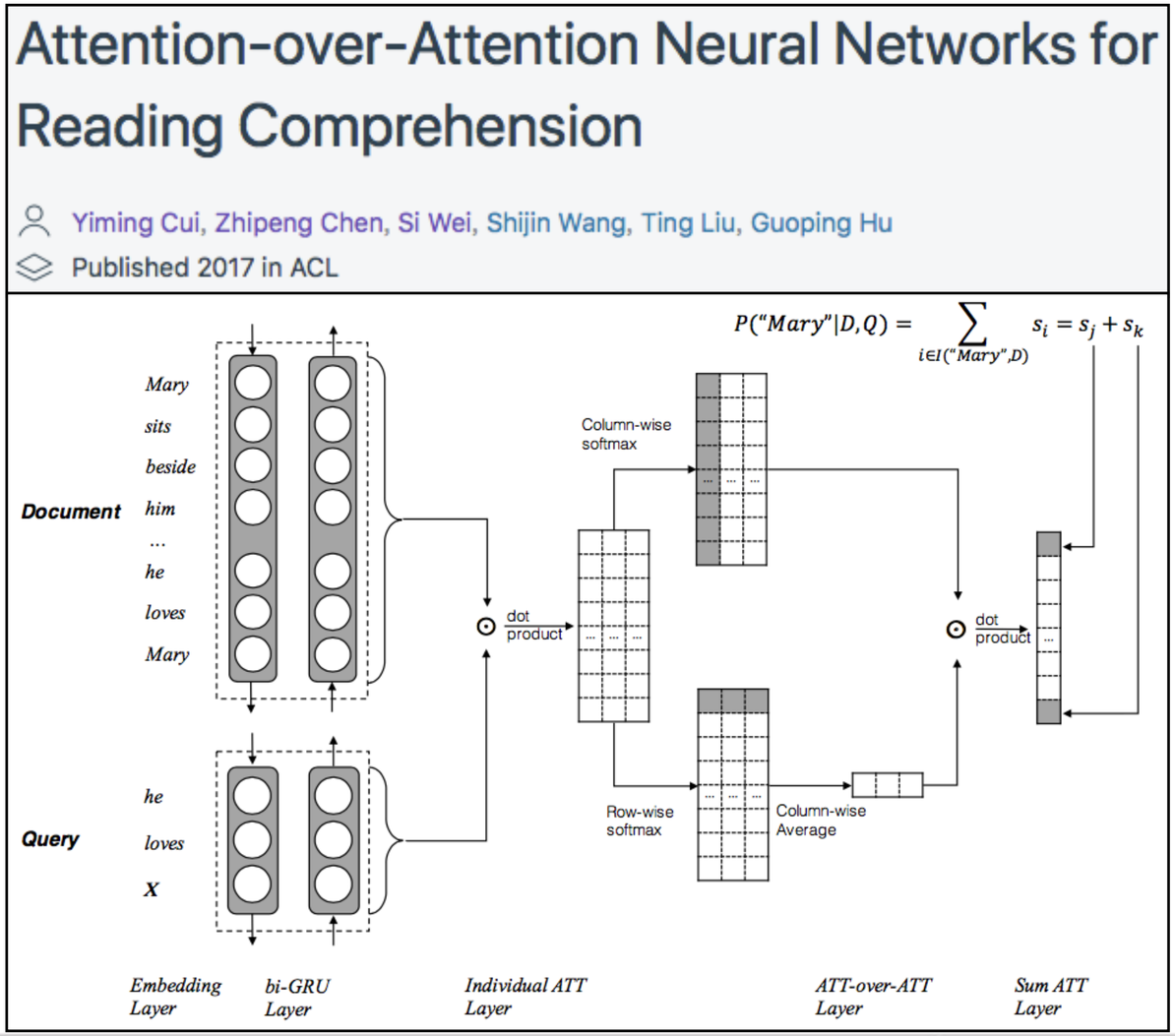
AS READER

- **AS Reader (Kadlec et al., 2016)**
  - 首次将Pointer Network应用在阅读理解任务上，巧妙利用了完形填空的特点
  - 模型结构相对简单但效果显著
  - 另外提出了Sum Attention的机制将候选词在篇章中出现的位置相加作为最终的概率



AoA Reader

- **AoA Reader (Cui et al., 2017)**
  - 业界首次提出层叠式注意力机制（Attention-over-Attention），在阅读理解任务中效果显著
  - 作为CAS Reader的进一步优化，Attention-over-Attention机制取代了原有的启发式规则，动态调整Query中的每个词对答案预测的贡献
  - 同时我们引入了Re-ranking机制，使用其他特征（例如语言模型、词频等）进一步提升答题效果



基线系统

- 基线系统效果（单模型）
  - 神经网络基线系统表现优异
  - Top Frequency表现较差，表明在制作数据时并没有大量选择高频词作为答案

系统（填空类问题）	开发集	测试集
Random Guess	1.65	1.67
Top Frequency	14.85	14.07
AS Reader (Kadlec et al., 2016)	76.05	77.67
AoA Reader (Cui et al., 2017)	77.20	78.63

系统（用户提问类问题）	开发集	测试集
Random Guess	1.50	1.47
Top Frequency	10.65	8.73
AS Reader (Kadlec et al., 2016)	-	49.03
AoA Reader (Cui et al., 2017)	-	51.53

最终评测结果

排名	系统（填空类问题，单系统）	开发集	测试集↓
1	上海交通大学仿脑计算与机器智能研究中心自然语言组	76.15	**77.73**
2	南京云思创智信息科技有限公司	77.15	77.53
3	华东师范大学	77.95	77.40
4	武汉大学语言与信息研究中心	**78.20**	76.53
5	哈尔滨工业大学（深圳）	76.05	75.93
6	广州火焰信息科技有限公司	73.55	75.77
7	鲁东大学	74.75	75.07
8	6ESTATES PTE LTD	75.85	74.73
9	武汉科技大学	73.80	74.53
10	北京信息科技大学	70.06	70.20
11	沈阳航空航天大学	63.15	65.80
12	山西大学二队	62.60	64.70
13	山西大学一队	64.85	64.67
14	郑州大学	52.80	54.53

最终评测结果

排名	系统（填空类问题，多系统）	开发集	测试集↓
1	6ESTATES PTE LTD	**81.85**	**81.90**
2	上海交通大学仿脑计算与机器智能研究中心自然语言组	78.35	80.67
3	南京云思创智信息科技有限公司	79.20	80.27
4	华东师范大学	79.45	79.70
5	鲁东大学	77.05	77.07
6	山西大学二队	62.60	64.70

最终评测结果

排名	系统（填空类问题，混合排名）	单/多模型	开发集	测试集↓
1	6ESTATES PTE LTD	**多模型**	**81.85**	**81.90**
2	上海交通大学仿脑计算与机器智能研究中心自然语言组	**多模型**	78.35	80.67
3	南京云思创智信息科技有限公司	**多模型**	79.20	80.27
4	华东师范大学	**多模型**	79.45	79.70
5	鲁东大学	**多模型**	77.05	77.07
6	武汉大学语言与信息研究中心	单模型	78.20	76.53
7	哈尔滨工业大学（深圳）	单模型	76.05	75.93
8	广州火焰信息科技有限公司	单模型	73.55	75.77
9	武汉科技大学	单模型	73.80	74.53
10	北京信息科技大学	单模型	70.06	70.20
11	山西大学二队	**多模型**	66.65	68.47
12	沈阳航空航天大学	单模型	63.15	65.80
13	山西大学一队	单模型	64.85	64.67
14	郑州大学	单模型	52.80	54.53

最终评测结果

- **用户提问类问题**
  - 3家单位提交最终结果
  - 其中山西大学使用了纯NLP的系统结构，其他两家使用了神经网络结构
  - 从单系统结果来看，准确率相差较大，表明选择正确的问题迁移方法对最终结果影响较大

用户提问类问题（User-Query Question）

最终排名	参赛单位	单/多系统	开发集准确率	测试集准确率↓
1	华东师范大学 East China Normal University (ECNU)	多系统	90.45%	69.53%
		单系统	85.55%	65.77%
2	山西大学三队 Shanxi University (SXU-3)	单系统	47.80%	49.07%
3	郑州大学 Zhengzhou University (ZZU)	单系统	31.10%	32.53%
–	Baseline – AS Reader CMRC2017 Official	单系统	*TBA*	47.77%
–	Baseline – Top Frequency CMRC2017 Official	单系统	10.65%	8.73%

获奖单位

- 我们很荣幸的宣布以下单位获奖
  - **填空类问题**

填空类问题（Cloze-style Question）

最终排名	参赛单位	单/多系统	开发集准确率	测试集准确率↓
1	6ESTATES PTE LTD	多系统	81.85%	81.90%
2	上海交通大学仿脑计算与机器智能研究中心自然语言组 Shanghai Jiao Tong University (SJTU BCMI-NLP)	多系统	78.35%	80.67%
3	南京云思创智信息科技有限公司	多系统	79.20%	80.27%

获奖单位

- 我们很荣幸的宣布以下单位获奖
  - **用户提问类问题**

用户提问类问题（User-Query Question）

最终排名	参赛单位	单/多系统	开发集准确率	测试集准确率↓
1	华东师范大学 East China Normal University (ECNU)	多系统	90.45%	69.53%
2	山西大学三队 Shanxi University (SXU-3)	单系统	47.80%	49.07%
3	郑州大学 Zhengzhou University (ZZU)	单系统	31.10%	32.53%

获奖单位

- 我们很荣幸的宣布以下单位获奖
  - **最佳单系统奖**

最佳单系统（Best Single System）

最终排名	参赛单位	单/多系统	开发集准确率	测试集准确率↓
1	上海交通大学仿脑计算与机器智能研究中心自然语言组 Shanghai Jiao Tong University (SJTU BCMI-NLP)	单系统	76.15%	77.73%

总结

- 填空类问题已被广泛研究，从各参赛单位提交的结果来看目前的神经网络技术能够很好的解决这类问题
- 对于问题类型迁移，本次评测中仅有少许队伍提交了结果，表明该任务存在一定的难度，值得进一步深入探究
- 期待接下来各参赛队伍的精彩技术报告！

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（HFL）
- 冠名商：科大讯飞股份有限公司
- 感谢南京师范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://www.hfl-tek.com/cmrc2017/index.html
- 评测数据：https://github.com/ymcui/cmrc2017
- 评测报告：https://arxiv.org/abs/1709.08299

再次感谢各参赛单位的大力支持！