百度阅读理解技术及应用

百度自然语言处理部 刘璟

2018年7月6日

智能设备迅速普及

智能问答

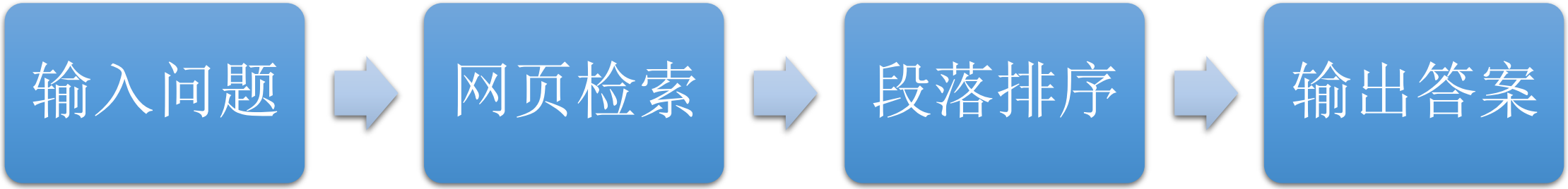
知识图谱

网页

网页

Baidu 百度

传统的检索式问答



“最后一公里”：答案精准定位

Q 儿童能汗蒸吗

P 如今，随着人们生活水平的不断提高，人们的健康意识逐渐增强，所以养生成为人们现在比较喜欢的休闲方式。而且，随着汗蒸的不断普及，人们对汗蒸的好处也越来越了解，并且很喜欢进行汗蒸这项活动。但是，汗蒸对所有人都使用吗?儿童能进行汗蒸吗?儿童能汗蒸吗?答案是否定的。现今比较流行的托玛琳汗蒸房汗蒸，可以产生有利于身体健康的远红外线，这些远红外线对于身体的保健作用是不容忽视的。但是，提醒广大热爱汗蒸的家长朋友们注意了，儿童是不适宜进行汗蒸的。儿童之所以不适宜进行汗蒸，是因为儿童的体温远比成年人的体温上升的要快，这就使得儿童的新陈代谢加快。而快速的体内回圈会增加心脏的负担，也很难通过排汗来调节人体温度，这样会损害儿童的身体健康，所以儿童最好是不要进行汗蒸。不过也有人认为，有些汗蒸房如某些韩式汗蒸馆的温度不会太高，效果也可以，儿童在此进行汗蒸也不会有什么严重的影响，利大于弊，所以也可以去试试。但最后要提醒大家的是，如果家长们很想带着儿童一块去汗蒸，可以先咨询一下小儿科医生，遵照医生的建议指导进行这项活动。

基于阅读理解技术的答案精准定位

定义

数据

阅读理解

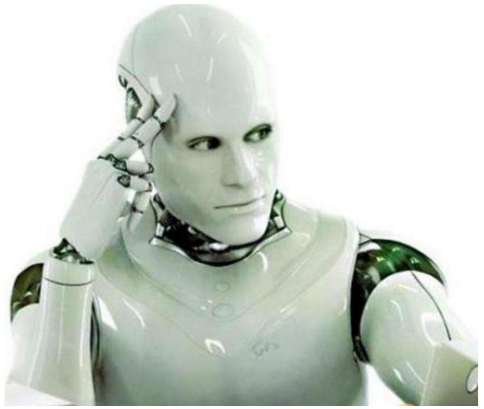
模型

应用

Baidu 百度

任务定义

机器阅读理解（Machine Reading Comprehension）是指让机器阅读文本，然后回答和文本内容相关的问题。



任务定义（续）

- 给定篇章P和问题Q， 从P中抽取答案A， A是P中的一个连续片段

段落（P） + 问题（Q） → 答案（A）

P

防水作为目前高端手机的标配,特别是苹果也支持防水之后,国产大多数高端旗舰手机都已经支持防水。虽然我们真的不会故意把手机放入水中,但是有了防水之后,用户心里会多一重安全感。那么近日最为火热的小米6防水吗?小米6的防水级别又是多少呢? 小编查询了很多资料发现,小米6确实是防水的,但是为了保持低调,同时为了不被别人说防水等级不够,很多资料都没有标注小米是否防水。根据评测资料显示,**小米6支持IP68级的防水**,是绝对能够满足日常生活中的防水需求的。

Q 小米6防水等级是多少？

A 小米6支持IP68级的防水

数据集

Neural network

2015之后 **>=100K 问题!**

SQuAD (Rajpurkar et al, 2016)
- **100K** 问题 (众包)
- 536 文档 (维基百科)

DuReader (He et al, 2017)
- **200K** 问题 (搜索日志)
- **1M** 文档 (网页和知道)

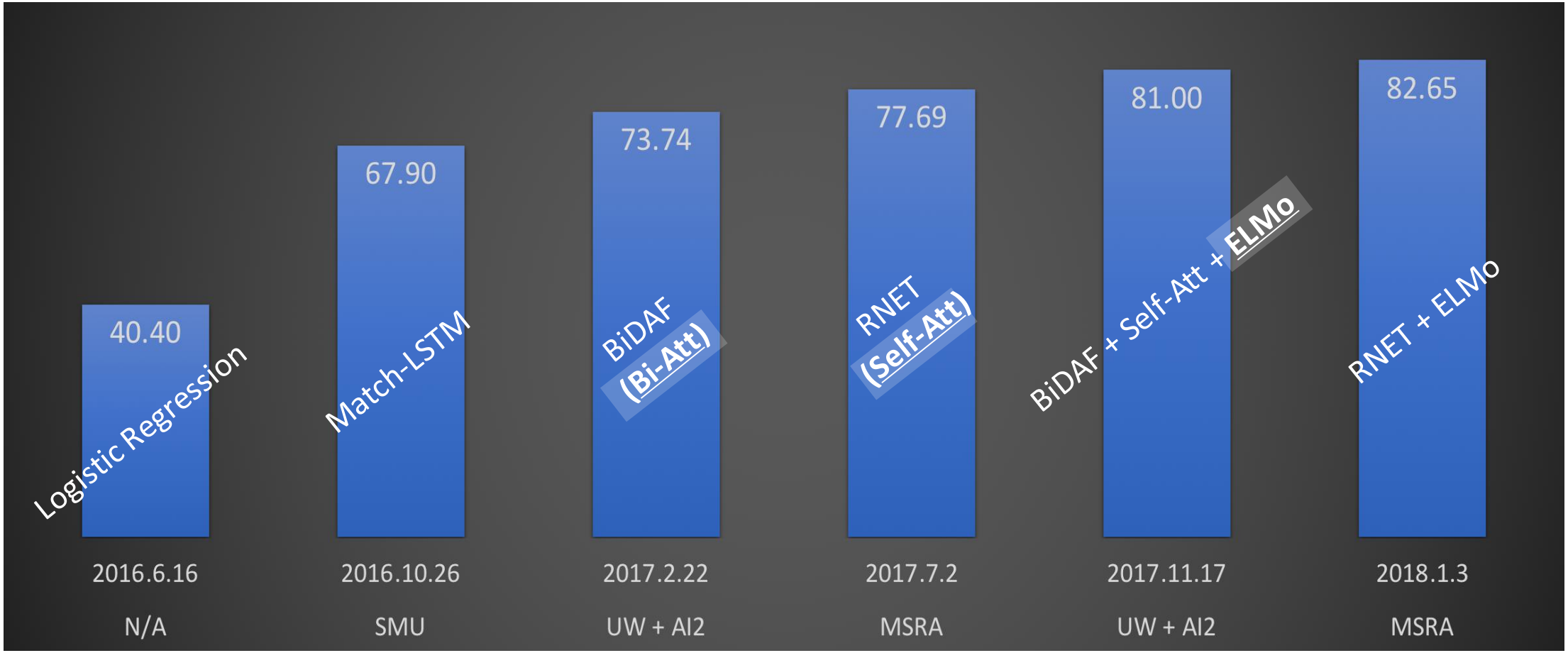
MCTest (Richardson et al, 2013)
- 2600 问题 (众包)
- 500 儿童故事

MSMARCO (Nguyen et al, 2016)
- **100K** 问题 (搜索日志)
- **200K** 文档 (网页)

2015之前

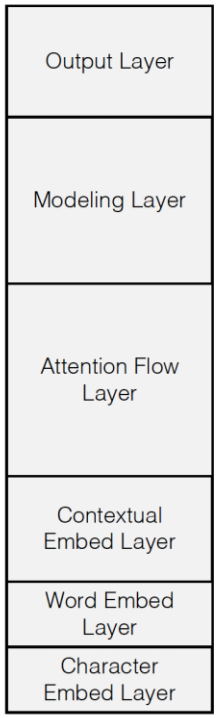
Logistic regression

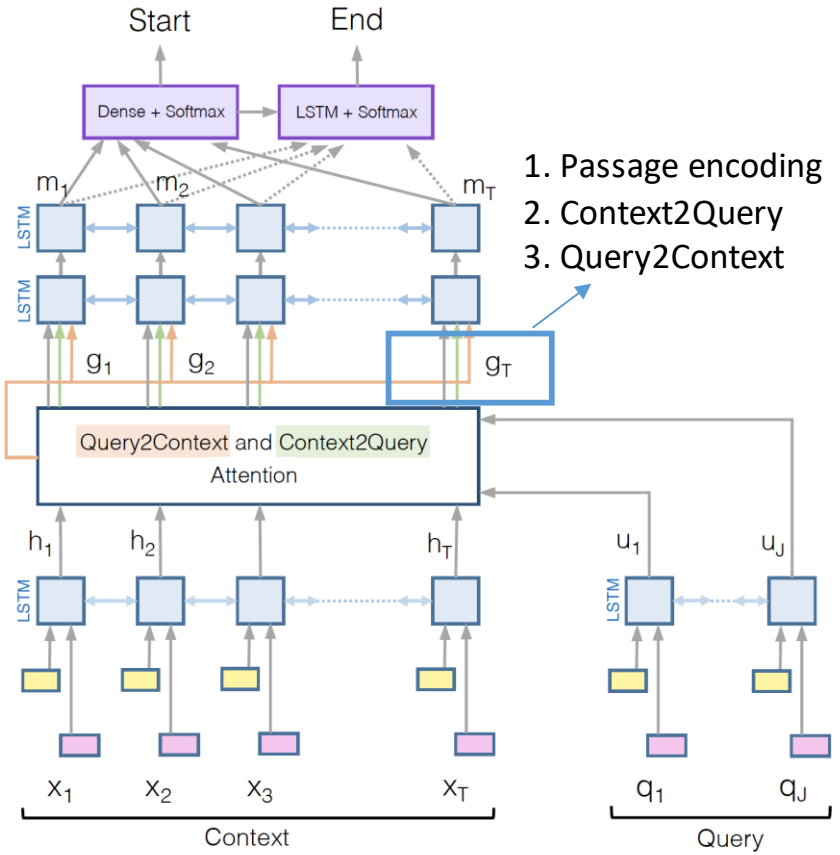
数据集驱动的技术进步（SQuAD）

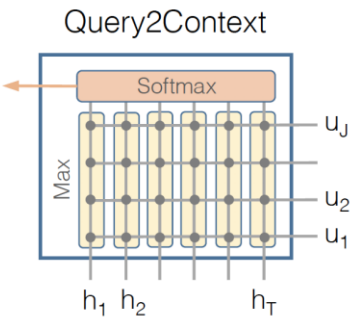


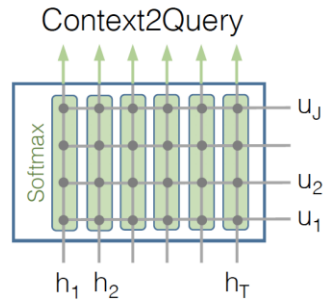
Bi-Directional Attention Flow Model (BiDAF)

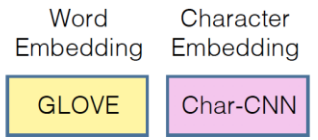
- 预测层：选择最佳答案片段
  - Pointer Network
- 匹配层：基于attention机制的问题和篇章匹配
  - Question-aware passage encoding
  - Passage-aware question encoding
- 编码层：问题和篇章编码
  - Word/Char, RNN/BiRNN











Seo, Minjoon, et al. Bidirectional attention flow for machine comprehension. ICLR 2017

样例分析（正确）

P

Established originally by the Massachusetts legislature and soon thereafter named for John Harvard (its first benefactor), Harvard is the United States' oldest institution of higher learning, and the Harvard Corporation (formally, the President and Fellows of Harvard College) is its first chartered corporation. Although never formally affiliated with any denomination, the early College primarily trained Congregationalist and Unitarian clergy. Its curriculum and student body were gradually secularized during the 18th century, and by the 19th century Harvard had emerged as the central cultural establishment among Boston elites.

Q What individual is the school named after?

Predicted John Harvard ✓

样例分析（错误）

P

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. **The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title.** The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

Q Which team won Super Bowl 50?

Predicted Carolina Panthers ✗

（Chen et al. 2017）

Baidu 百度

面向搜索场景的阅读理解

DuReader (He et al, 2017)
- **200K** 问题 (搜索日志)
- **1M** 文档 (网页和知道)

MSMARCO (Nguyen et al, 2016)
- 100K 问题 (搜索日志)
- 200K 文档 (网页)

SQuAD (Rajpurkar et al, 2016)
- 100K 问题 (众包)
- 536 文档 (维基百科)

多文档阅读理解的挑战

- 文化
- 细胞的培养

Question: What is the difference between a mixed and pure culture?

Passages:

预测

[1] **A culture is a society's total way of living and a society is a group that live in a defined territory and participate in common culture.** While the answer given is in essence true, societies originally form for the express purpose to enhance . . .

[2] . . . There has been resurgence in the economic system known as capitalism during the past two decades. 4. **The mixed economy is a balance between socialism and capitalism.** As a result, some institutions are owned and maintained by . . .

[3] **A pure culture is one in which only one kind of microbial species is found whereas in mixed culture two or more microbial species formed colonies.** Culture on the other hand, is the lifestyle that the people in the country . . .

[4] Best Answer: **A pure culture comprises a single species or strains. A mixed culture is taken from a source and may contain multiple strains or species.** A contaminated culture contains organisms that derived from some place . . .

[5] . . . It will be at that time when we can truly obtain a pure culture. **A pure culture is a culture consisting of only one strain.** You can obtain a pure culture by picking out a small portion of the mixed culture . . .

正确

[6] **A pure culture is one in which only one kind of microbial species is found whereas in mixed culture two or more microbial species formed colonies.** A pure culture is a culture consisting of only one strain. . . .

. . . . . .

Reference Answer: A pure culture is one in which only one kind of microbial species is found whereas in mixed culture two or more microbial species formed colonies.

端到端的多文档阅读理解模型V-NET

- 文化
- 细胞的培养

Question: What is the difference between a mixed and pure culture?

Passages:

预测 [1] **A culture is a society's total way of living and a society is a group that live in a defined territory and participate in common culture.** While the answer given is in essence true, societies originally form for the express purpose to enhance . . .

[2] . . . There has been resurgence in the economic system known as capitalism during the past two decades. 4. **The mixed economy is a balance between socialism and capitalism.** As a result, some institutions are owned and maintained by . . .

[3] **A pure culture is one in which only one kind of microbial species is found whereas in mixed culture two or more microbial species formed colonies.** Culture on the other hand, is the lifestyle that the people in the country . . .

证据 [4] Best Answer: **A pure culture comprises a single species or strains. A mixed culture is taken from a source and may contain multiple strains or species.** A contaminated culture contains organisms that derived from some place . . .

[5] . . . It will be at that time when we can truly obtain a pure culture. **A pure culture is a culture consisting of only one strain.** You can obtain a pure culture by picking out a small portion of the mixed culture . . .

正确 [6] **A pure culture is one in which only one kind of microbial species is found whereas in mixed culture two or more microbial species formed colonies.** A pure culture is a culture consisting of only one strain. . . .

. . . . . .

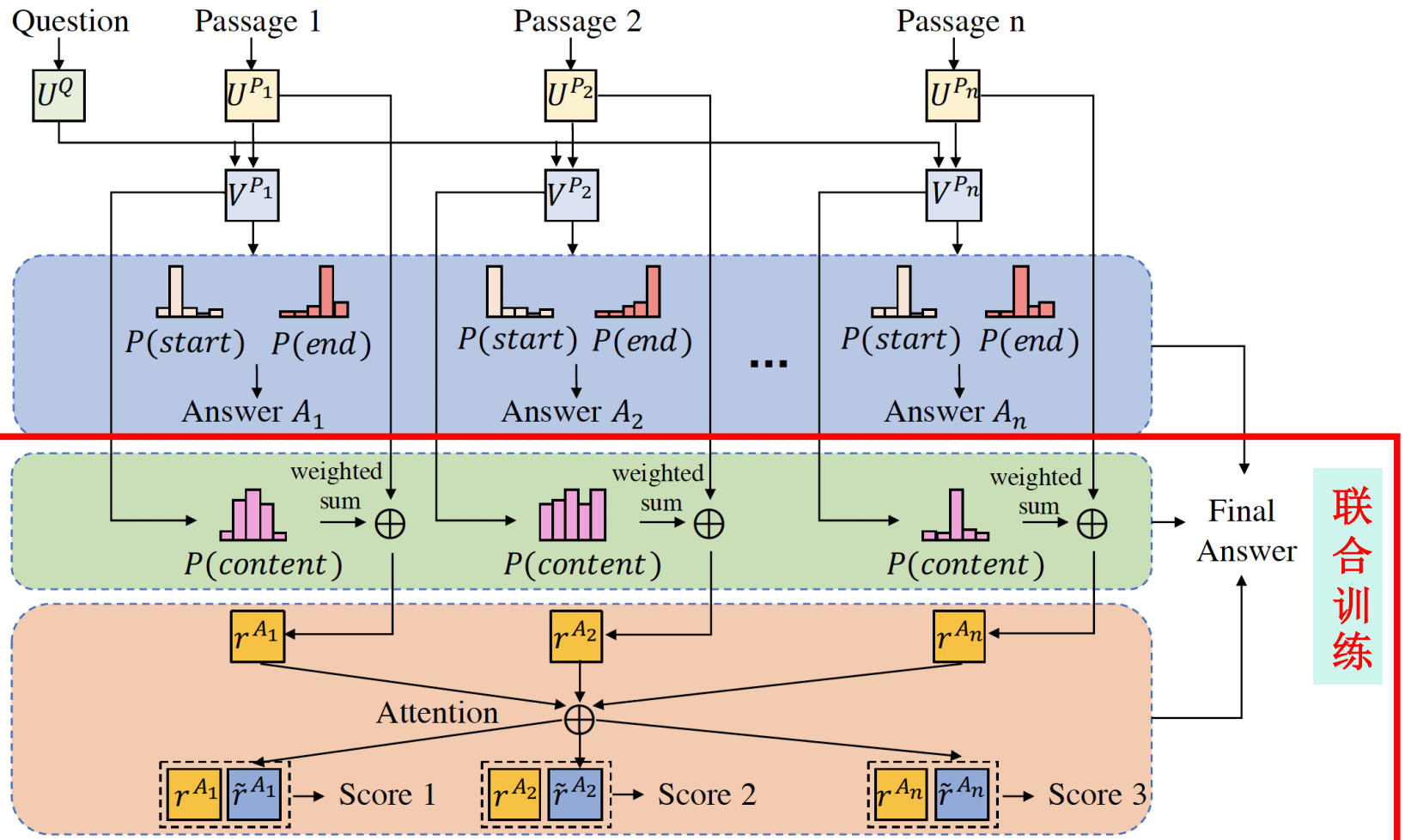
Reference Answer: A pure culture is one in which only one kind of microbial species is found whereas in mixed culture two or more microbial species formed colonies.

Wang, Yizhong, et al. Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification. ACL 2018

Baidu 百度

端到端的多文档阅读理解模型V-NET（续）

- 编码层：问题和篇章编码 **Encoding**
- 匹配层：基于attention机制的问题和篇章匹配 **Q-P Matching**
- 答案边界预测层：选择答案起始位置 **Answer Boundary Prediction**
- **答案内容预测层：预测答案词** **Answer Content Prediction**
- **答案验证层：答案互相验证** **Answer Verification**



Wang, Yizhong, et al. Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification. ACL 2018

端到端的多文档阅读理解模型V-NET（续）

模型对比实验（测试集结果）

模型	Rouge-L
ReasoNet (Shen et al. 2016)	38.81
R-Net (Wang et al. 2017)	42.89
S-Net (Tan et al. 2018)	45.23
V-Net (Wang et al. 2018)	**46.15**

模块有效性验证实验（开发集结果）

模型	Rouge-L	△
V-NET (Full Model)	45.65	
—Answer Verification	44.38	-1.27
—Content Model	44.27	-1.38
—Joint Training	44.12	-1.53

Question: What is the difference between a mixed and pure culture	**Scores**		
Answer Candidates:	**Boundary**	**Content**	**Verification**
[1] A culture is a society's total way of living and a society is a group ...	$\mathbf{1.0 \times 10^{-2}}$	$\mathbf{1.0 \times 10^{-1}}$	1.1×10^{-1}
[2] The mixed economy is a balance between socialism and capitalism.	1.0×10^{-4}	4.0×10^{-2}	3.2×10^{-2}
[3] A pure culture is one in which only one kind of microbial species is ...	5.5×10^{-3}	7.7×10^{-2}	1.2×10^{-1}
[4] A pure culture comprises a single species or strains. A mixed ...	2.7×10^{-3}	8.1×10^{-2}	1.3×10^{-1}
[5] A pure culture is a culture consisting of only one strain.	5.8×10^{-4}	7.9×10^{-2}	5.1×10^{-2}
[6] A pure culture is one in which only one kind of microbial species ...	5.8×10^{-3}	9.1×10^{-2}	$\mathbf{2.7 \times 10^{-1}}$
.....		.....	

Wang, Yizhong, et al. Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification. ACL 2018

面向搜索问答的阅读理解数据集DuReader

真实问题
搜索日志

真实文本
网页库和UGC文本

最大规模
30万问题，150万文档

丰富标注
66万人工标注答案、
问题类型、实体及观点

- 主流数据集仅覆盖部分实际问答需求

Answer type	Percentage	Example
Date	8.9%	19 October 1512
Other Numeric	10.9%	12
Person	12.9%	Thomas Coke
Location	4.4%	Germany
Other Entity	15.3%	ABC Sports
Common Noun Phrase	31.8%	property damage
Adjective Phrase	3.9%	second-largest
Verb Phrase	5.5%	returned to Earth
Clause	3.7%	to avoid trivialization
Other	2.7%	quietly

SQuAD数据集问题类型分布，答案平均长度3.09

- DuReader问题类型分布

答案类型	事实类（Fact）	观点类（Opinion）
实体类（Entity）	23.4% 例：iphone x哪天发布	8.5% 例：2017最好看的十部电影
描述类（Description）	34.6% 例：消防车为什么是红的	17.8% 例：丰田卡罗拉怎么样
是非类（YesNo）	8.2% 例：39.5度算高烧吗	7.5% 例：学围棋能开发智力吗

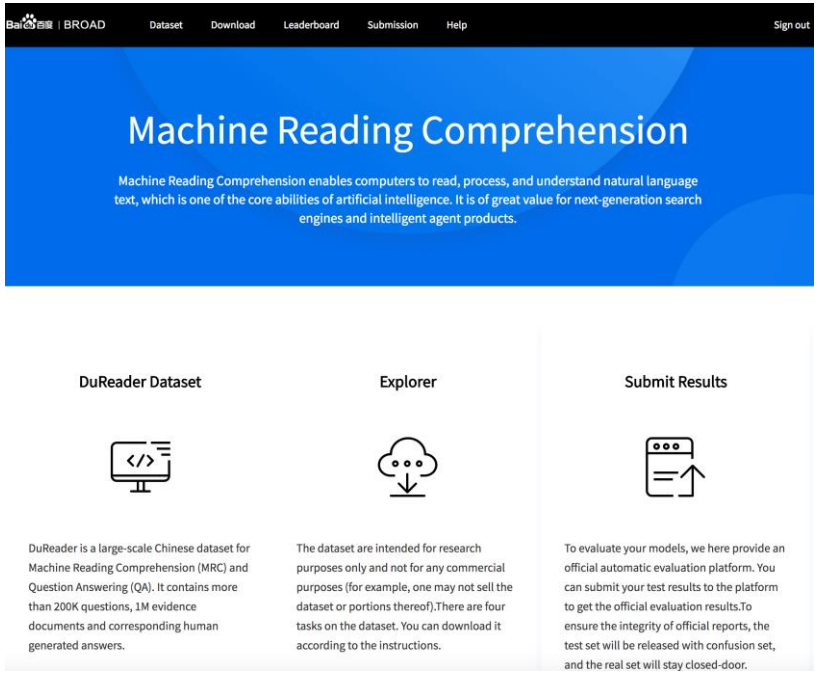
He, Wei, et al. DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications. MRQA workshop at ACL 2018

Baidu 百度

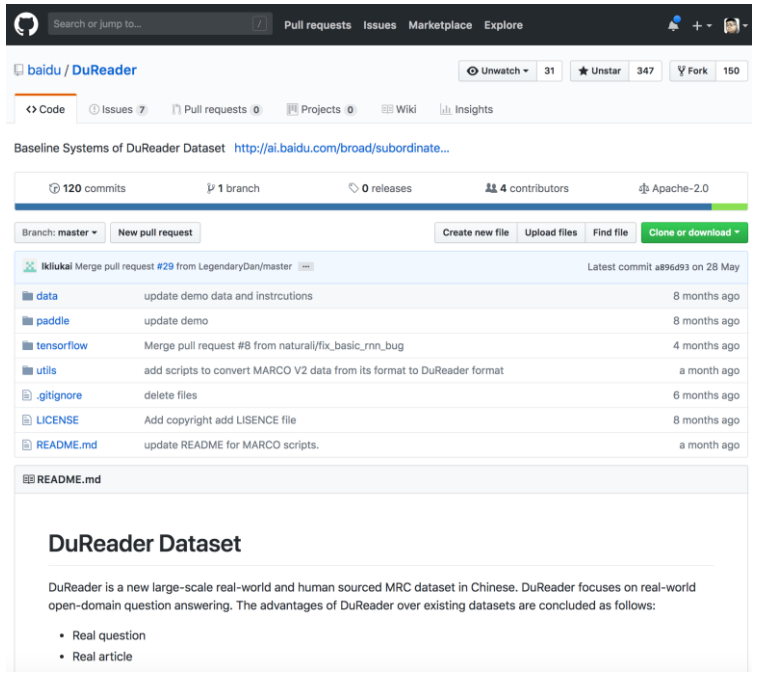
DuReader @ 百度开放数据平台

- http://ai.baidu.com/broad
  - 数据集总下载量1,7000+*

  * 统计截至2018.5.30



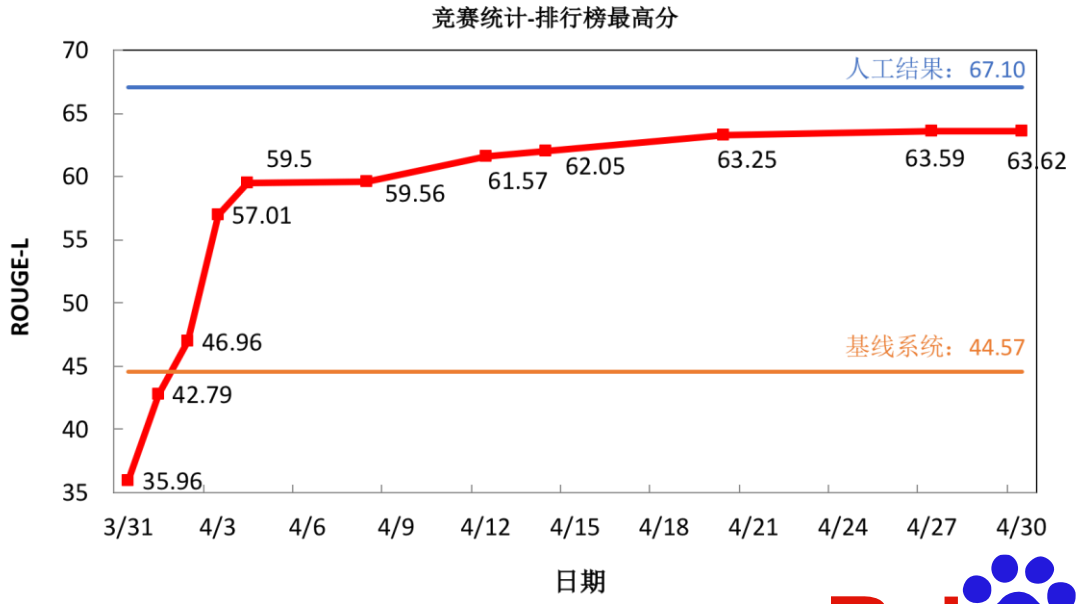
- 开源基线系统 @ Github
  - Paddlepaddle
  - Tensorflow



He, Wei, et al. DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications. MRQA workshop at ACL 2018

- 与中文信息学会和计算机学会联合举办技术竞赛
- 1062支国内外队伍踊跃注册
  - 466支团队来自高校和科研所
    - 国内“211”院校超过半数都参与了本次竞赛
  - 440支队伍来自企业
    - 腾讯、阿里、华为、微软、IBM...
  - 156支队伍是个人参赛

- 1489次结果提交，有效推动技术提升



答案精准定位例子

P

喝牛奶的时间对人体消化吸收有很大的影响，因此喝牛奶应该讲究时间安排。营养专家们认为，牛奶最好在傍晚或临睡之前半小时饮用。晚上喝牛奶有利于休息和睡眠。牛奶中含有一种能使人产生疲倦欲睡的生化物L色氨酸，还有微量吗啡类物质，这些物质都有一定的镇静催眠作用，特别是L色氨酸它是大脑合成羟色的主要原料，五羟色胺对大脑睡眠起着关键的作用，它能使大脑思维活动暂时受到抑制，从而使人想睡眠，并且无任何副作用，而且牛奶粘在胃壁上吸收也好，牛奶中的钙还能清除紧张情绪。睡前喝牛奶有利于钙的吸收利用。晚餐摄入的钙，睡前大部分被人体吸收利用。睡后特别是晚上零点以后血液中钙的水平会逐渐降低，血钙的下降，促进了甲状旁腺分泌亢进，激素作用于骨组织，使骨组织中的一部分钙盐，溶解入血液中，以维持血钙的稳定平衡。此种溶解作用是人体的自我调节功能，时间长了，会成为骨质疏松症的原因之一。晚上睡前喝牛奶，牛奶中的钙可缓慢的被血液吸收，整个晚上血钙都得到了补充、维持平衡，不必再溶解骨中的钙，防止了骨流失、骨质疏松症，所以睡前喝牛奶好。

Q

牛奶什么时候喝最好

A

牛奶最好在傍晚或临睡之前半小时饮用。

答案精准定位例子（续）

P

孕妇吃金针菜，可以补充营养，强身健体，促进胎儿的健康发育，但是孕妇吃金针菜前要先将金针菜做熟，不能吃生的金针菜，因为金针菜含有秋水仙碱，遇热才能破坏掉，孕妇如果生吃金针菜就会摄入秋水仙碱，会导致中毒。金针菜炒鸡蛋原料：适量鸡蛋，金针菜，油，盐，葱。做法：1、将葱洗干净，切成段，将金针菜洗干净，先浸泡一会儿，将鸡蛋打散，搅拌均匀。2、往锅内倒入适量油，然后下葱段，炒出爆香味。3、然后加入鸡蛋，炒得差不多熟的时候就加入金针菜，翻炒至熟。4、加入适量盐，搅拌均匀即可。功效：孕妇怀孕，身体疲劳而胃口不佳，吃金针菜炒鸡蛋，可以改善胃口，还可补充蛋白质，提高免疫力，增强抵抗力，还可促进胎儿健康发育。

Q

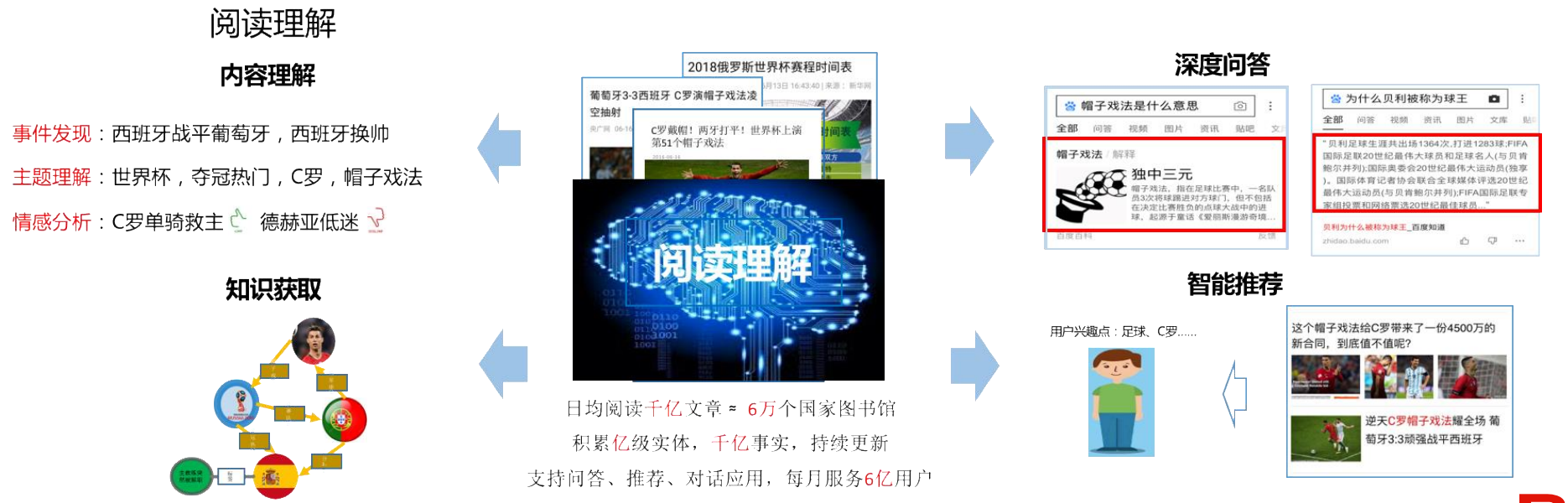
孕妇吃金针菜的做法

A

1、将葱洗干净，切成段……搅拌均匀。
2、往锅内倒入适量油……炒出爆香味。
3、然后加入鸡蛋……翻炒至熟。
4、加入适量盐，搅拌均匀即可。

总结

- 新时代下的阅读理解技术得到了快速发展
  - 数据规模的增大使得端到端的学习变的可能
  - 神经网络在匹配方面展现出很大的优势
- 未来工作



谢谢

Q&A