斯坦福SQuAD挑战赛的中国亮丽榜单

崔一鸣[1] 刘 挺[2] 王士进[1] 等
[1]哈工大讯飞联合实验室 科大讯飞
[2]哈尔滨工业大学

关键词：阅读理解评测 斯坦福阅读理解数据集 (SQuAD)

由斯坦福大学发起的大规模阅读理解挑战赛SQuAD[1](Stanford Question Answering Dataset) 正处于如火如荼的角逐阶段。SQuAD 是目前竞争最为激烈且影响力最大的阅读理解任务，被誉为阅读理解领域的 ImageNet。在挑战赛中，你会看到一份亮丽榜单。它来自中国——哈工大讯飞联合实验室 (HFL) 经过一段时间的摸索，改进了早期在填空型阅读理解问题上使用的基于层叠式注意力模型 (Attention-over-Attention Reader) [2]，提出了基于交互式层叠注意力模型 (Interactive Attention-over-Attention Reader)，在该任务上获得性能突破，两个评测指标 (模糊准确率、精准匹配准确率) 均达到世界第一的水平。同时，该模型也是目前首个模糊准确率 (F1-score) 超过 85% 的系统。

阅读理解评测进展

让机器能听会说，能理解会思考是人工智能的终极目标。如何让机器能够阅读并且理解自然语言，也是认知智能领域研究的主要问题。近年来，在自然语言处理领域掀起了一股研究阅读理解的热潮，吸引了大量研究机构加入其中。

谷歌 DeepMind 的研究员赫尔曼 (Hermann) 等人在 2015 年发表文章，提出使用基于注意力的神经网络模型 Attentive Reader[3] 来解决阅读理解问题，同时发布了一个大规模的填空型阅读理解数据集 CNN/Daily Mail[1]。Facebook 的研究员希尔 (Hill) 等人也在同一年发表文章，提出利用记忆网络 (Memory Network)[2] 解决阅读理解问题，并同步公开了儿童读物填空型阅读理解数据集 (Children's Book Test) [4]。这两个数据集是在 2016 年阅读理解领域中使用最多的数据集。2016 年 7 月，哈工大讯飞联合实验室发布中文阅读理解数据集 PD&CFT[5]，进一步推动了中文阅读理解的研究。

在 2016 年自然语言处理实证方法会议 (EMNLP 2016) 上，斯坦福大学 Rajpurkar 等人发布了 SQuAD[6]，使阅读理解研究上了一个新台阶。图 1 给出了 SQuAD 中的一个例子。

SQuAD 有几个特点：(1) 答案不再只是词级别，其粒度扩展到短语甚至是句子。(2) 问题不再是由机器自动生成，而是采用人工标注，保证了问题的可用比例。(3) 该数据集规模更大 (10 万个问题)，可

[1] 由Hermann等人提出的阅读理解数据集。该数据集从美国有线电视新闻网(CNN)和《每日邮报》上摘取了大量语料，然后从每篇文章对应的摘要中去掉某个实体构成问题，要求机器能够根据文章的内容自动找到对应的实体。

[2] 由Weston等人提出的一种机器学习的框架，包含输入特征映射、泛化层、输出特征层、反馈层。

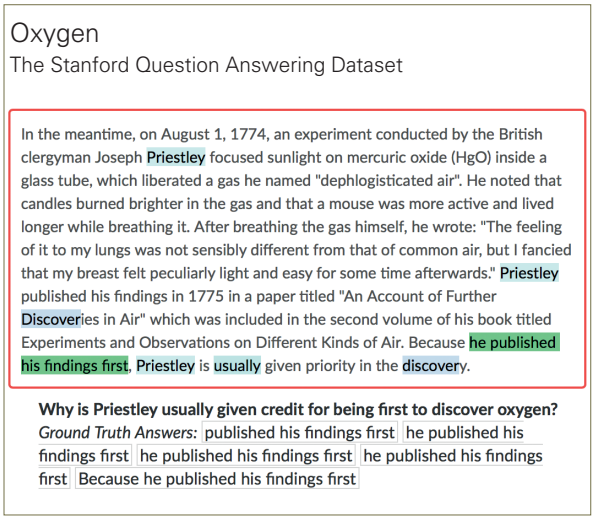

图1 SQuAD阅读理解数据集问题实例

以更好地利用神经网络模型来解决阅读理解问题。

与以往不同的是，斯坦福大学仅公开了训练集和开发集而保留了测试集，并提供一个开放平台供参赛者提交自己的算法，由 SQuAD 官方利用隐藏的测试集对参赛系统进行评测，将相关结果更新到 SQuAD 官网上。

SQuAD结果及模型分析

SQuAD 挑战赛是业内公认的机器阅读理解标准水平测试，也是该领域的顶级赛事。参赛者来自全球学术界和产业界的研究团队，包括微软亚洲研究院、艾伦研究院 (AI2)、IBM、Salesforce、Facebook、谷歌等知名企业以及卡内基梅隆大学、斯坦福大学等高校。2017 年 8 月初，SQuAD 挑战赛榜单再次更新，将每个参赛队伍的最好成绩进行排名，结果如表 1 所示。

表1 斯坦福SQuAD榜单（截至2017年8月初）

系统名（参赛队伍）	Exact Match	F1
Interactive AoA Reader（哈工大讯飞联合实验室）	77.845	85.297
r-net（微软亚洲研究院）	77.688	84.666
smarnet（Eigen & 浙江大学）	75.989	83.475
DCN+（Salesforce）	74.866	82.806
ReasoNet（微软雷德蒙德研究院）	75.034	82.552
Mnemonic Reader（国防科技大学&复旦大学）	74.268	82.371
SEDT（卡内基梅隆大学）	74.090	81.761
SSAE（清华大学）	74.080	81.665
Human Performance（斯坦福大学）	*82.304*	*91.221*

可以看出，中国本土机构在 SQuAD 任务中相当活跃，表明目前机器阅读理解研究在国内已成为热门的研究方向。榜单中的大多数模型公开了模型结构，我们根据已有的文献对这些系统进行了深入分析，得出了如下结论。

基于注意力机制 基于注意力机制 (Attention-based) 的神经网络方法 [7] 在自然语言处理领域获得了较大成功，从最初的神经网络机器翻译到目前的阅读理解系统，大多数采用了这种机制。应用注意力机制，主要是利用问题在篇章中寻找与问题最相关的部分，得到一个篇章级别的注意力分布，从而使机器能够更好地识别，从这种软机制上缩小解答的范围。目前基于注意力机制的方法是阅读理解任务上标配的结构之一。

基于字与词的词向量表示 人们在阅读文章时，如果篇章中存在大量的生词，就很难理解其表达的具体含义，机器也是如此。训练神经网络时，考虑到时间和空间复杂度，通常无法使用全词表，需要从整个词表中抽取出频次较高的 *N* 个词作为训练用的词表 (shortlist)，这样难免会将一些低频词过滤掉，但这些信息有助于机器对篇章的深层次理解。在传统的基于词的向量表示基础上，引入基于字的向量表示有助于减少未登录词所带来的影响，对于一些词缀的学习也有一定帮助 [8]。目前引入基于词和字的向量表示的系统均获得了一定的性能提升。

基于指针网络的答案预测 早在填空型阅读理解的研究中，IBM Watson 研究员卡德莱克 (Kadlec) 等人就尝试使用改进的指针网络 (Pointer Network)[9, 10] 来直接从篇章中抽取出单个词的答案，这也成为了填空型阅读理解模型的标配。随后 Shuohang Wang（王烁航，音译）等人提出了

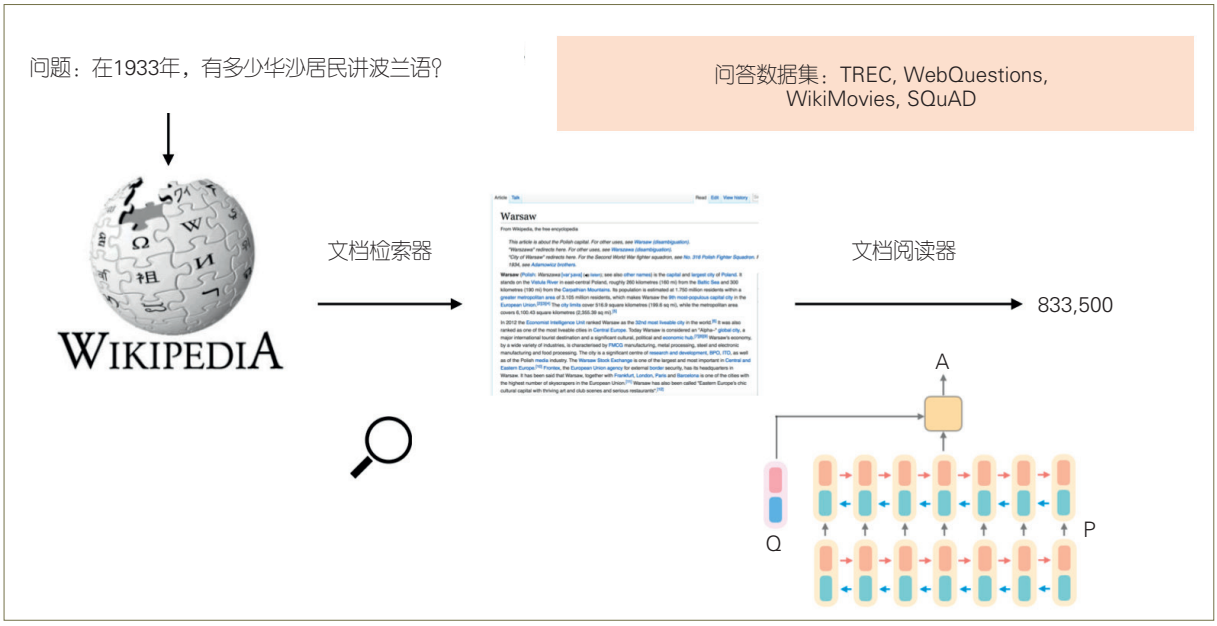

图2 斯坦福大学提出的开放域阅读理解模型

Match-LSTM 模型，用于解答 SQuAD 中的任务 [11]。由于 SQuAD 任务的答案是篇章中的连续片段，所以只须利用指针网络预测答案在篇章中的起始位置和结束位置即可。这种方法较答案生成模型有显著的性能优势，且与 SQuAD 任务更加契合。

阅读理解研究趋势及展望

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. Scientists wanted to see how well their bodies worked after eating different kinds of breakfast.

The results show that if a person eats a right breakfast, he or she will work better than if he or she has no breakfast. If a student has fruit, eggs, bread and milk before going to school, he or she will learn more quickly and listen more carefully in class. Some people think it will help you lose weight if you have no breakfast. But the result is opposite to what they think. This is because people become so hungry at noon that they eat too much for lunch. They will gain weight instead of losing it.

Question: What do the results show?

A) They show that breakfast has affected on work and studies.

B) The results show that breakfast has little to do with a person's work.

C) The results show that a person will work better if he only has fruit and milk.

D) They show that girl students should have less for breakfast.

图3 RACE数据集中的问题实例

SQuAD 任务是一个长期的挑战赛，从去年在 EMNLP 2016 亮相至今一年的时间，排行榜不断被刷新，与之对应的深度学习模型层出不穷，极大地推动了机器阅读理解的研究进程。但机器阅读理解的发展绝不会止步于此，未来值得探索的方向有：

基于多文档的阅读理解 多文档阅读理解技术是原有技术在内容上的自然拓展。需要从大量的文档中挑选出与问题最相关的若干文档，缩小机器阅读的范围，再利用传统的基于单文档的阅读理解模型从候选篇章中抽取出精准的答案。

面向开放域的阅读理解 在回答开放域的问题时，需要从海量的篇章中抽取出与问题最相关的一些篇章，然后筛选出正确答案。斯坦福大学陈丹琪博士在 ACL 2017 上发表文章，提出了面向开放域的阅读理解技术方案（见图 2），用来解答维基百科上相关的问题 [12]。

更大规模及更高难度的阅读理解数据集 更大规模的数据集可以让我们挖掘到更多信息。在 EMNLP 2017 上发表的一篇论文，提供了一个新的阅读理解数据集 RACE[13]（见图 3）。该数据集包含了我国中考和高考中使用的英文阅读理解问题，其形式与早期的 MCTest[3] 数据集非常相似，但不论从难度上还是规模上，都比 MCTest 有了大幅度提高。

面向实用场景的阅读理解 在实际生活中，人们查阅资料时往往需要阅读长篇甚至大量的文档才能找到答案。阅读理解技术可以将人们从繁杂的阅读中解放出来。例如，人们想知道“汽车的天窗怎么打开”时，只需提出问题，机器就可以从长篇的资料中抽取出答案和相关图片并展示出来，方便人们的生活。

随着机器阅读理解数据集的不断迭代更新，相关技术不断进步，有效地推动了机器阅读理解研究的发展。现阶段的机器阅读理解模型是否真正地“理解”了自然语言，仍是需要进一步深入研究的问题。相信总有一天，机器会像人类一样对语言进行归纳、总结、推理，进而真正地理解自然语言。 ■

崔一鸣

CCF专业会员。科大讯飞AI研究院高级研究员，哈尔滨工业大学计算机专业硕士。主要研究方向为阅读理解、机器翻译、自然语言处理等。

ymcui@iflytek.com

刘 挺

CCF理事、CCF哈尔滨主席、*CCCF*前译文栏目主编。哈尔滨工业大学教授、社会计算与信息检索研究中心主任。曾任顶级国际会议ACL、EMNLP领域主席。主要研究方向为自然语言处理和社会计算。tliu72@foxmail.com

王士进

CCF专业会员。科大讯飞AI研究院副院长，中国科学院自动化所博士。主要研究方向为语音识别、口语评测、机器翻译、阅读理解等。

sjwang3@iflytek.com

其他作者：陈致鹏 马文涛 胡国平

参考文献

[1] The Stanford Question Answering Dataset[OL]. http://stanford-qa.com

[2] Cui Y, Chen Z, Wei S, et al. Attention-over-Attention Neural Networks for Reading Comprehension[C]// *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017, Volume 1: Long Papers)*, 2017: 593-602.

[3] Hermann K M, Kocisky T, Grefenstette E, et al. Teaching Machines to Read and Comprehend[C]// *Advances in Neural Information Processing Systems*. MIT Press, 2015: 1693-1701.

[4] Hill F, Bordes A, Chopra S, et al. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations[OL]. (2015).arXiv preprint arXiv:1511.02301.

[3] MCTest是Richardson等人提出的一个阅读理解数据集，其形式类似于英语考试中的阅读理解单项选择题。MCTest 严格限制了文档本身是儿童能够理解的故事内容，从而保证阅读理解的过程仅仅依赖于篇章的内容，而不需要过多的外部知识。

[5] Cui Y, Chen Z, Wei S, et al. Consensus Attention-based Neural Networks for Chinese Reading Comprehension[C]// *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016: 1777-1786.

[6] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100,000+ Questions for Machine Comprehension of Text[C]// *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*(EMNLP 2016), 2016: 2383-2392.

[7] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[OL]. (2014). arXiv preprint arXiv:1409.0473.

[8] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional Attention Flow for Machine Comprehension[OL]. (2016). arXiv preprint arXiv:1611.01603.

[9] Kadlec R, Schmid M, Bajgar O, et al. Text Understanding with the Attention Sum Reader Network[C]//*Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL2016, Volume 1: Long Papers)*.2016: 908-918.

[10] Vinyals O, Fortunato M, Jaitly N. Pointer networks[C]// Advances in Neural Information Processing Systems, 2015: 2692-2700.

[11] Wang S, Jiang J. Machine Comprehension Using Match-LSTM and Answer Pointer[OL]. (2016).arXiv preprint arXiv:1608.07905.

[12] Chen D, Fisch A, Weston J, et al. Reading Wikipedia to Answer Open-Domain Questions[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017, Volume 1: Long Papers), 2017: 1870-1879.

[13] Lai G, Xie Q, Liu H, et al. RACE: Large-scale ReAding Comprehension Dataset from Examinations[OL]. arXiv preprint arXiv:1704.04683, 2017.