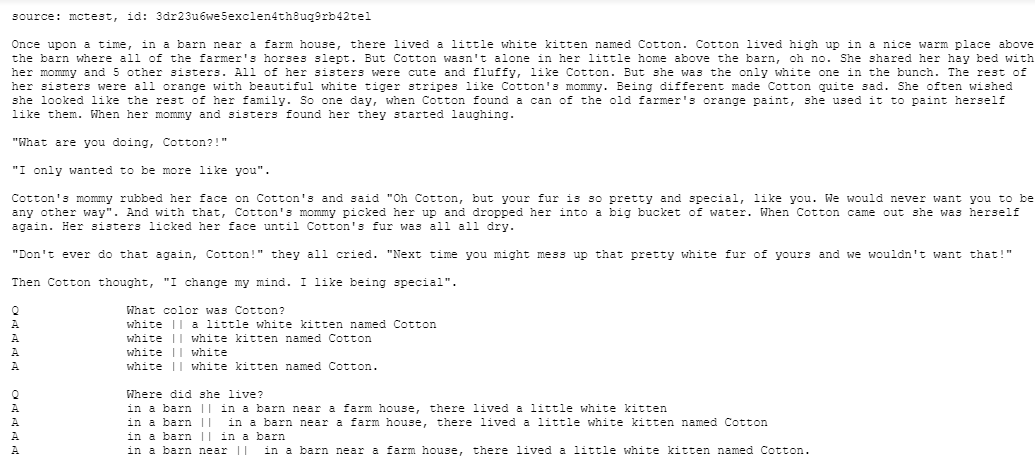
2019年7月24日：SQuAD，CoQA和QuAC三个阅读理解数据集的对比（2019NAACL）

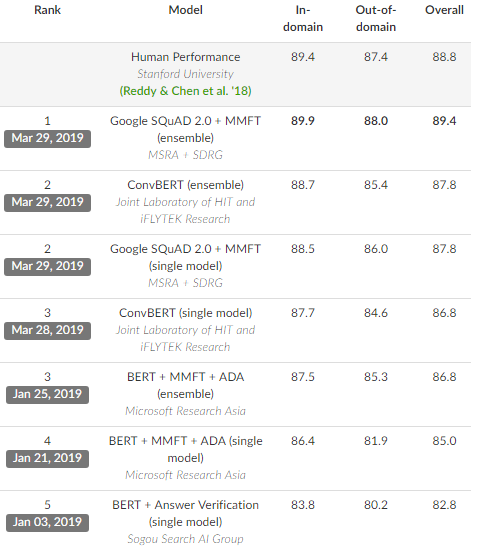
这两天在看图网络相关的内容，在艰难理解中……

SQuAD,CoQA和QuAC三个数据集的比较及内容\_19年NAACL（Yatskar M. A qualitative comparison of coqa, squad 2.0 and quac[J]. arXiv preprint arXiv:1809.10735, 2018.），三个数据集都是阅读理解形式的数据集，分析如下：

1. CoQA:该数据集提出时的挑战目标是衡量机器理解文本段落的能力，该数据集包含从12.7万多个问题，并从8000多个对话中收集问题。其独特之处有：1）问题是对话形式的（类似阅读理解题中的问答）2）答案可以是自由格式的文本3）每个答案中还附有段落中突出的证据子序列4）一共有7个不同领域。该数据集中还包含对于常识的推理，其大致内容如下：



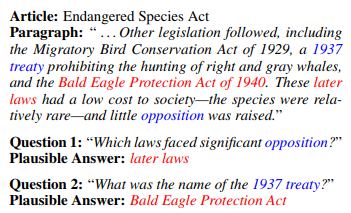
该数据集目前的研究状况为：



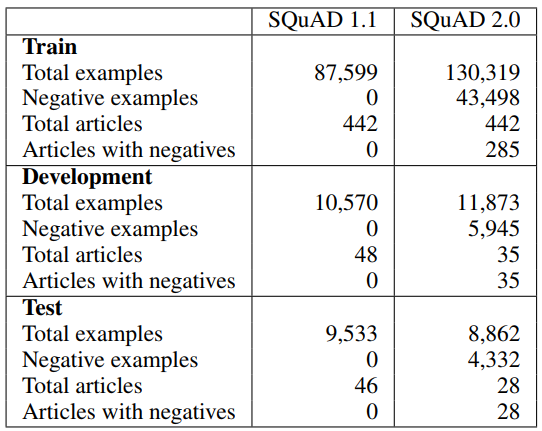
该数据集领域，现有算法已经超过了人类平均水平。

1. SQuAD2.0：发布该数据集的相关文章为：Rajpurkar P, Jia R, Liang P. Know What You Don't Know: Unanswerable Questions for SQuAD[J]. arXiv preprint arXiv:1806.03822, 2018.

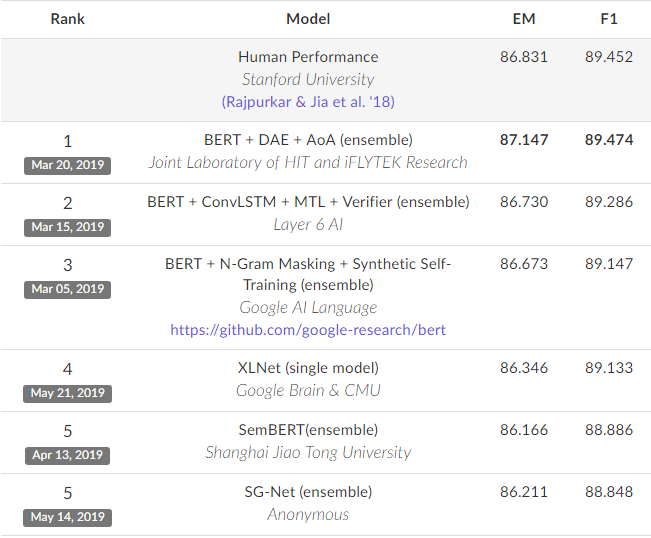
该数据集取自维基百科，包含了十万条问题，并且包含了一些难以直接从文章中抽取信息回答的问题，如：



下表是该数据集2.0版本与1.1版本的对比：



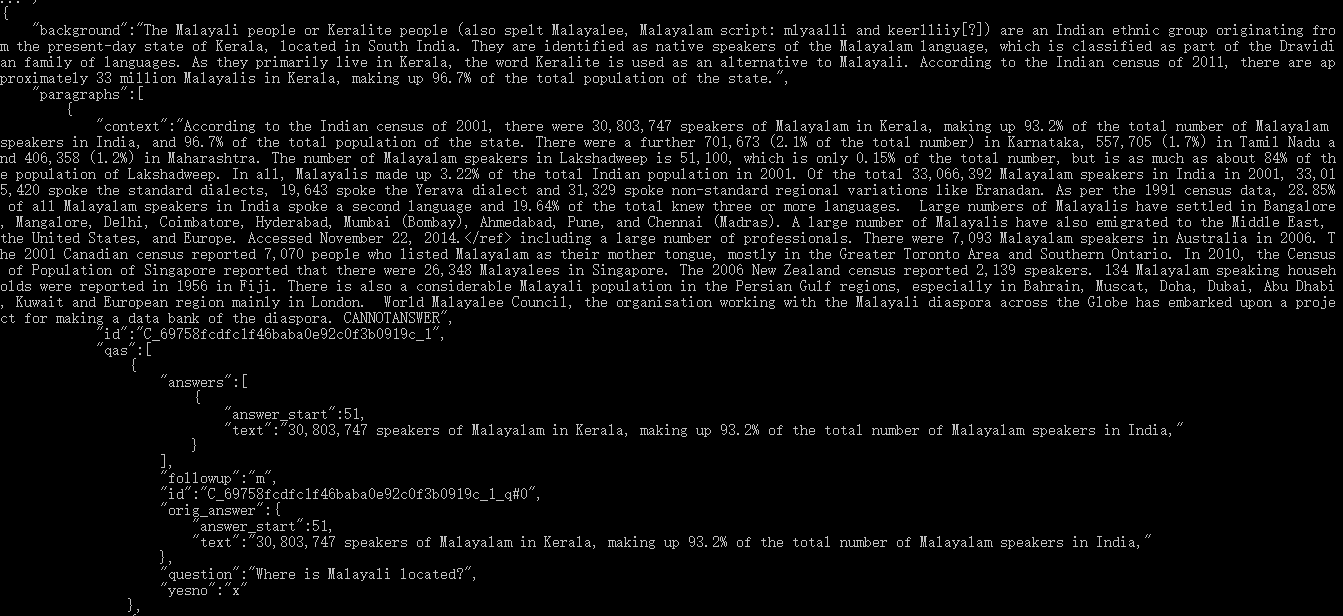
该数据集目前的刷榜状况为：



该数据集从测试表现上来说已经超过了人类的平均水平。

1. QuAC数据集： Choi E, He H, Iyyer M, et al. Quac: Question answering in context[J]. arXiv preprint arXiv:1808.07036, 2018.

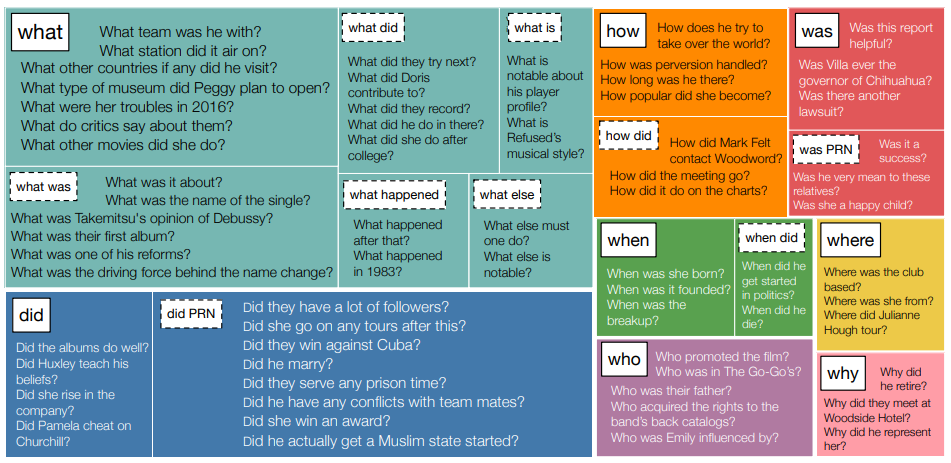
该数据集包含的是学生和教师之间的对话文本，其引入了一个之前数据集没有的新挑战：其问题更加开放，难以回答，或是只在当前上下文中有意义。其内容格式如下：



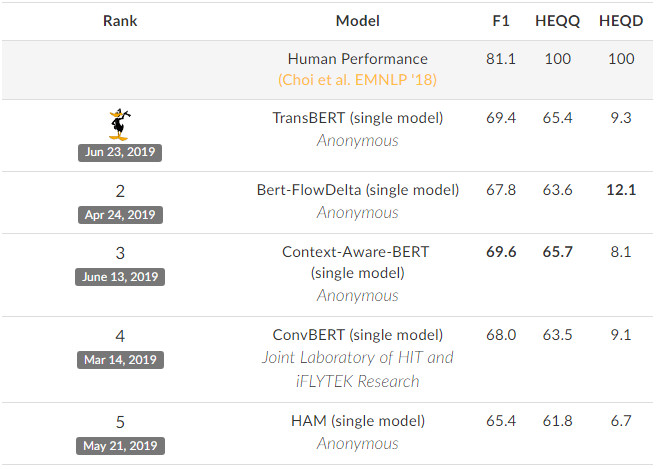


一组数据包含背景，文章内容，一系列问题和回答。其中，问题的回答来自于文中内容的概括。

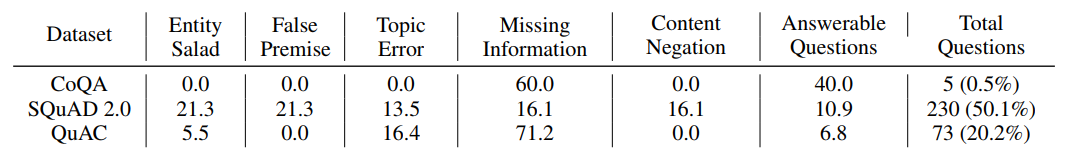
下图中的矩阵图表示了关于该数据集的几个最常出现的“Wh“式问题：



相比于其他数据集，该数据集的提升空间还很大：

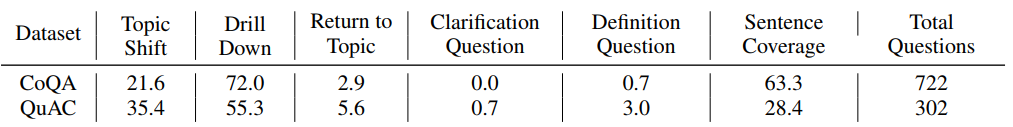


1. 今天看的文章综合分析了如上的3个数据集。首先，文章从随机抽取的50句上下文中查看各种数据的特殊情况。



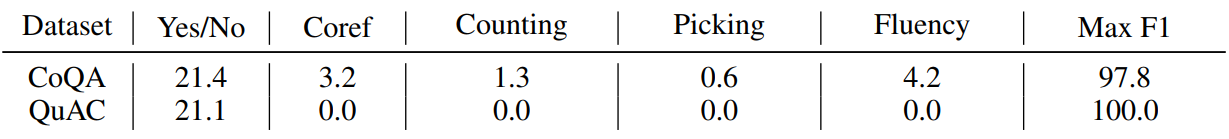
文章分析，SQUAD2.0包含了多种会导致问题无法回答的情况，而QuAC专注于信息推测（丢失信息项目数据很多），CoQA则没有显著地包含无法回答的问题。

同时，文章对比了CoQA和QuAC数据集中问题特征的比较，比较内容如下表：



文章分析，CoQA包含的多是深入主题细节的问题，而QuAC对话中切换主题的频率更高。

接下来，文章分析了CoQA和QuAC两个数据集中摘要式问题（对错问题，共指问题，数数问题，选择问题，“流畅度问题”——如‘Q: how did he get away? A: by foot’）类型的分布，从表格可以看出QuAC的摘要式问题类型更为单一。



从各数据集目前的SOTA来看，QuAC数据集还有很大的提升空间，但其对于文章细节和推理的要求较高；另外两个数据集目前的SOTA都很好，但是普遍使用了基于BERT或XLNET的结构，因此也许在非预训练结构上还有进步的空间