汇报 0525

朱睿涵　李宸亦

项目三

2022 年 5 月 25 日

目录

CoQA 简介

CoQA A Conversational Question Answering Challenge
CoQA：一个用于建立对话式问题回答（Conversational Question Answering）系统的新型数据集。数据集包含 12.7 万个带答案的问题，这些问题来自 7 个不同领域的 8 千条文本段落的对话，后续实验中五个用于域内评估，两个用于域外评估。问题是对话式的，而答案是自由格式的文本，并在段落中突出了相应的证据。

CoQA 开发目标

- 还原人类对话的性质：人们在日常对话中很少像阅读理解一样，基于材料生搬硬套出一个答案，要还原对话的这一本质，就需要解决传统阅读理解问题的问题-文章依赖性，以及实现基于对话历史的问答
- 确保对话中答案的自然度：以往的阅读理解会依赖材料截取答案，导致答案不够自然，不够口语化。因此要通过 CoQA 训练出形式较为自由的抽象答案，而不是简单的信息提取。
- 建立跨领域表现稳健的 QA 系统：以往的 QA 数据集来源于单一领域而 CoQA 的数据来源于七个领域

目标任务

给定一篇文章和一段对话，回答对话中的下一个问题。对话中的每一轮由问题 (Q)，答案 (A)，依据 (R) 组成，答案往往比依据简洁很多。回答问题时，需要考虑对话中的历史信息，比如回答 Q_2 时，要基于对话历史 Q_1, A_1 以及答案依据 R_2，可表示为：

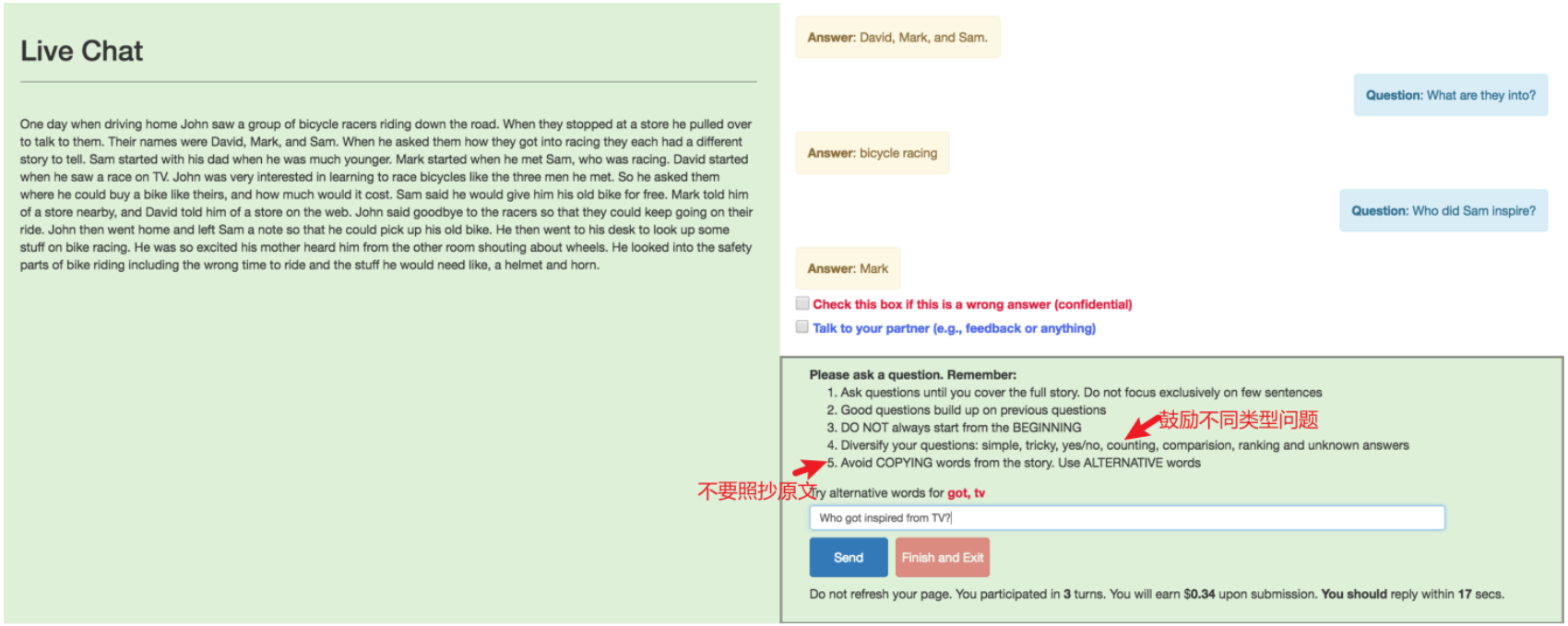
$$A_2 = f(Q_1, A_1, Q_2, R_2)$$

$$A_n = f(Q_1, A_1, ..., Q_{n-1}, A_{n-1}, R_n, Q_n)$$

对于无法回答的问题，给出“unknown”的回答，不标注任何依据（R）

数据收集

提出新的问题者，希望避免使用段落中的确切词汇，以增加词汇的多样性。当他们输入一个已经出现在段落中的词时，提醒他们尽可能地转述问题。界面如图：



数据收集

回答问题者，希望回答者坚持使用段落中的词汇，以限制可能的答案数量。界面如图：

数据集划分

选取儿童故事、文学、初中和高中英语考试、新闻、维基百科、Reddit 和科学 7 个方面。

Domain	#Passages	#Q/A pairs	Passage length	#Turns per passage
In-domain				
Children's Sto.	750	10.5k	211	14.0
Literature	1,815	25.5k	284	15.6
Mid/High Sch.	1,911	28.6k	306	15.0
News	1,902	28.7k	268	15.1
Wikipedia	1,821	28.0k	245	15.4
Out-of-domain				
Reddit	100	1.7k	361	16.6
Science	100	1.5k	251	15.3
Total	8,399	127k	271	15.2

Table 2: Distribution of domains in CoQA.

数据集划分：

对于领域 1-5：开发集 100 篇文章，测试集 100 篇文章，其余文章作为训练集

对于领域 6-7：测试集 100 篇文章，其余文章作为训练集

模型

模型

给定段落 p
对话历史 $q_1, a_1, \ldots, q_{i-1}, a_{i-1}$
黄金答案 $a_1, a_2, \ldots, a_{i-1}$ 被用来预测 a_i
输入：问题 q_i
输出：答案 a_i
模型：PGNet，DrQA，PGNet+DrQA

组合模型中，阅读理解模型 DrQA 首先指出文本中的答案证据，而对话模型 PGNet 则将证据归化为答案。根据经验对 DrQA 和 PGNet 做了一些改变。对于 DrQA，如果答案是理由的一个子串则直接预测答案，否则就预测理由。对于 PGNet，提供当前问题和 DrQA 的跨度预测作为编码器的输入，解码器的目的是预测最终的答案。

结果

	In-domain					Out-of-dom.		In-domain	Out-of-dom.	Overall
	Child.	Liter.	Mid-High.	News	Wiki.	Reddit	Science	Overall	Overall	
Development data										
Seq2seq	30.6	26.7	28.3	26.3	26.1	N/A	N/A	27.5	N/A	27.5
PGNet	49.7	42.4	44.8	45.5	45.0	N/A	N/A	45.4	N/A	45.4
DrQA	52.4	52.6	51.4	56.8	60.3	N/A	N/A	54.7	N/A	54.7
Augmt. DrQA	**67.0**	**63.2**	**63.9**	**69.8**	72.0	N/A	N/A	**67.2**	N/A	**67.2**
DrQA+PGNet	64.5	62.0	63.8	68.0	**72.6**	N/A	N/A	66.2	N/A	66.2
Human	90.7	88.3	89.1	89.9	90.9	N/A	N/A	89.8	N/A	89.8
Test data										
Seq2seq	32.8	25.6	28.0	27.0	25.3	25.6	20.1	27.7	23.0	26.3
PGNet	49.0	43.3	47.5	47.5	45.1	38.6	38.1	46.4	38.3	44.1
DrQA	46.7	53.9	54.1	57.8	59.4	45.0	51.0	54.5	47.9	52.6
Augmt. DrQA	**66.0**	63.3	66.2	**71.0**	71.3	57.7	63.0	**67.6**	60.2	**65.4**
DrQA+PGNet	64.2	**63.7**	**67.1**	68.3	**71.4**	**57.8**	**63.1**	67.0	**60.4**	65.1
Human	90.2	88.4	89.8	88.6	89.9	86.7	88.1	89.4	87.4	88.8

Table 7: Models and human performance (F1 score) on the development and the test data.

1. seq2seq 模型的表现最差
2. 组合模型优于两者的单一模型，可以与增强的 DrQA 相比较
3. 最好的模型比人类表现差

分析点

History size	Seq2seq	PGNet	DrQA	Augmt. DrQA	DrQA+ PGNet
0	24.0	41.3	50.4	62.7	61.5
1	**27.5**	43.9	**54.7**	66.8	**66.2**
2	21.4	44.6	54.6	**67.2**	66.0
all	21.0	**45.4**	52.3	64.5	64.3

所有的模型都成功地利用了历史，但超过一个以前的回合，收益就很少了。随着我们增加历史记录的大小，性能会下降。在人的实验中，前一回合在理解当前问题中起着重要作用；对话中的大多数问题在两个回合的范围内有有限的依赖性。

目录

百度千言数据集

百度千言针对阅读理解任务有三个数据集，其中 Dureader checklist 和 DuReader robust 是单篇章、抽取式阅读理解数据集，DuReader yesno 是观点型的。

百度千言数据集

DuReader yesno的数据集：


```
{
"documents": [
    {
    "title": "香蕉能放冰箱吗 香蕉剥皮冷冻保存_健康贴士
    "paragraphs": [
        "本文导读：..............."
    ]
    }
],
"yesno_answer": "No",
"question": "香蕉能放冰箱吗",
"answer": "香蕉不能放冰箱，香蕉如果放冰箱里，
        会更容易变坏，会发黑腐烂。",
"id": 293
}
```

百度千言数据集

checklist 和 robust 的数据集：


```
{
  "data": [
    {
      "paragraphs": [
        {
          "context": "【皋】字读音既可读gāo,又可读háo。读作gāo时,字义有三种意思,水边的高地或岸;沼泽,湖泊;姓氏。
                      读作háo时,有号呼;呼告的意思。皋读作háo时 ...  全文",
          "qas": [
            {
              "question": "皋怎么读",
              "type": "in-domain",
              "id": "e3ffa587bba2478191e357cd9a56d10b",
              "answers": [
                {
                  "text": "既可读gāo,又可读háo",
                  "answer_start": 6
                }
              ],
              "is_impossible": false
            }
          ],
          "title": "皋怎么读 - 懂得"
        }
      ]
    }
  ]
}
```

百度千言数据集

checklist 和 robust 和之前用过的数据集类型差不多意义不是很大，也许可以作为训练和测试的补充。DuReader yesno 这种观点型的数据集和之前的类型不同，把 DuReader yesno 装进模型训练的意义更大一些。

目录

baseline

上周对于苏剑林 baseline 进行了简单的 pytorch 改写, 为保证之后基于此 baseline 的微调工作能够顺利进行, 本周继续完成对于 baseline 的改写工作.
本周对于我们改写过的 baseline 模型又进行了结构化调整, 消除了之前存在的一些潜在问题, 目前 baseline 从 keras 到 pytorch 的移植工作基本完成, 如果下周计算资源申请到位后可以进行模型微调和数据集的加入.

THANKS!