**基于大模型的军事装备领域问答生成技术**

**评测任务**

近年来，大型语言模型（简称大模型）的高速发展，推动了知识获取方式、信息服务形态的颠覆性变革。由于大模型的技术特点，其赋能军事应用的效果评估、能力提升等工作，均极大依赖于专业的问答数据集。而由于军事装备领域数据的细分性、封闭性难获取性等特点，依赖领域专家完全人工构建数据集的代价过于高昂而难以有效实施。基于优势大模型自动生成领域问题与答案，辅助领域专家人工审核修正，是垂直领域专业问答评测数据集构建的可行途径。为此，任务组织方在2020-2023年连续4年组织军事装备无人系统等测评任务的基础上，协调相关领域专家，针对军事装备领域的互联网公开数据，人工标注了问答对生成样例数据集，包括填空题、单选题、判断题、复杂问答题等四种类型问题及答案。在此基础上组织本次技术评测，以期提升基于大模型的领域问答生成技术水平，更好的支持军事装备领域专业化数据集构建工作，同时遴选优秀团队，共同促进军事装备领域大模型应用技术生态建设。

**1、任务定义**

问答生成（Question and Answer Generation，QAG）是一项文本生成任务，旨在根据给定的文本内容，生成与上下文相关的问题和答案。在本次评测任务中，需要根据给定的文章，生成不同类型的问题答案对。

1.1 输入

文章&问题类型。

1.2 输出

指定类型的问题和答案。

1.3 问题类型

当前关注的问题类型有：填空题、单选题、判断题、复杂问答题。

其中的复杂问答题，指问题和答案有一定的复杂度，即答案需要深入理解文章主题、需要包含推理归纳等过程、需要通过阅读多个段落获取。

|  |  |  |
| --- | --- | --- |
| 问题类型编号 | 问题类型 | 问答对示例 |
| 1 | 填空题 | {  "question\_type\_id":1,  "question": "\_\_是歼-10的总设计师。",  "answer":"宋文骢"} |
| 2 | 单选题 | {  "question\_type\_id":2,  "question": "第十一届航空展在哪里举行的？"，"choices":["A 珠海","B 深圳","C 香港","成都"]  "answer":"A"} |
| 3 | 判断题 | {  "question\_type\_id":3,  "question": "当前，法国具有三位一体的核力量结构。"，  "answer":"错误"} |
| 4 | 复杂问答题 | {  "question\_type\_id":4,"question":"美军认为GBU-24/B的激光制导套件不适合快沉项目，原因是什么？"，  "answer":"原因包括两个方面，一是激光制导需要现场进行照射引导，难以实施；二是激光制导受气象条件影响大，不符合快沉项目的作战条件要求。"} |

**2、数据集描述**

样例数据包括军事装备领域有关装备技术、装备效能、装备运用、装备保障、装备战略等方面的40篇文章，每篇文章会有10个问答对示例样本，其中包含2道填空题、3道单选题、3道判断题以及2道复杂问答题。

样例数据分为两个文件：文章内容文件document.csv， 问答对文件qa.csv。

每个文件中包含的字段如下：

document.csv

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 字段名称 | 字段解释 | 字段内容示例 |
| 1 | doc\_id | 文章的唯一id | doc\_1 |
| 2 | content | 文章的文本内容 | 此次部署第13陆战队远征队将搭舰随行。\n\n地面作战单元为陆战4团2营，航空作战单元为..... |

qa.csv

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 字段名称 | 字段解释 | 字段内容示例 |
| 1 | qa\_id | 问答对的唯一id | qa\_1 |
| 2 | doc\_id | 问答对所属的文章id | doc\_1 |
| 3 | content | 问答对的内容，格式为json字典，参考1.3中的问答对示例 | {  "question\_type\_id":4,  "question": "美军认为GBU-24/B的激光制导套件不适合快沉项目，原因是什么？"，  "answer":"原因包括两个方面，一是激光制导需要现场进行照射引导，难以实施，二是激光制导受气象条件影响大，不符合快沉项目的作战条件要求。"} |

**3、评价指标**

本次评测会针对10篇文章来进行问答对生成，分以下五个维度对生成的问答与答案进行评估，每个维度的计分由大模型给出（届时将根据国内评测榜单选取具体的评分大模型，并提前告知参赛队；由于大模型生成结果具有变化性，会采取多次取平均的方式），每个维度评分所用的prompt见附录evaluation\_prompt.py文件：

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 评估指标 | 评估指标含义 | 分值范围 |
| S1 | 问题相关性 | 所生成的问题是否与文本相关 | 0-100 |
| S2 | 问答匹配度 | 所生成问题和答案的匹配程度 | 0-100 |
| S3 | 问题流畅性 | 问题文本的流畅度 | 0-100 |
| S4 | 答案流畅度 | 答案文本的流畅度（针对复杂问答题） | 0-100 |
| S5 | 问题复杂度 | 针对复杂问答题，判断其复杂程度。其他类型问题不涉及。 | 0-100 |

（1）对于每个问答对结果，令其得分为*S*q。

a) 对于填空题、判断题、单选题，其得分计算方法为：

*S*q = 0.4\**S*1+0.4\**S*2+0.2\**S*3

b) 对于复杂问答题，其得分计算方法为：

*S*q = 0.2\**S*1+0.2\**S*2+0.1\**S*3+0.1\**S*4+0.4\**S*5

（2）对于每类问题，令其得分为*TS*q。

该类问题的得分为该类问答对得分平均值，对于填空题和复杂问答题，K=2，其余情况下K=3。

*TS*q = $\frac{1}{K}\sum\_{k=1}^{K}Sq\_{k}$

（3）对于每篇文章，令其得分为*S*d。

*S*d = 0.5\*$TS\_{q复杂问答题}$+0.2\*$TS\_{q单选题}$+0.2\*$TS\_{q判断题}$+0.1\*$TS\_{q填空题}$

提交结果的最终评分S为：

S=$\frac{1}{10}\sum\_{i=1}^{10}Sd\_{i}$

**4、任务提交及评分**

4.1 实现方案要求

本次评测，要求参赛队调用**开源大模型**来实现问题与答案的生成，一方面需要构建有效的prompt，一方面需要考虑多个大模型的联合使用。

模型选择：开源大模型（便于测试以及复现）。

实现方式：可以只调用一个大模型，也可以同时使用多个大模型进行整合。

4.2 最终提交材料

参赛队最终提交如下材料：

a) 方法描述文档

b) 方法的实现代码，代码的具体格式请参见附录中的代码样例demo.py

4.3 依托平台

本次评测的任务提交，将依托红山开源平台（https://www.osredm.com/competition/zstp2024/）开展，参赛队在该平台完成注册、数据下载、结果提交等工作。

4.4 评估方式

组织方在示例的40篇文章之外，专门遴选军事装备领域的10篇文章作为测试数据集。参赛队提交材料后，组织方将运行参赛队提交的代码，针对测试数据集生成问答对，每篇文章生成10道题目，其中包含2道填空题、3道单选题、3道判断题以及2道复杂问答题，而后对生成的结果进行打分。本测评任务在比赛期间不公开测试数据集，比赛结束后将向所有提交结果的参赛队公示测试数据集。

**5、报名方式**

参赛队需首先在红山开源平台完成注册（https://www.osredm.com/competition/zstp2024），并在“参赛报名”处填写相关信息，鉴于本次评测专家标注样例数据集的价值，参赛队需提交队长所在单位证明（相关要求见报名页面）。参赛队通过审批环节后，可在红山开源平台完成样例数据集下载与结果提交。

**6、奖励设置**

CCKS组委会为本次任务提供3万元奖金，奖励3支获奖团队，奖金设置如下：

第一名：12,000元

第二名：10,000元

第三名： 8,000元

**7、时间安排**

时间安排初定如下，后续如果有调整，以CCKS发布信息为准。

* 评测任务发布：5月1日
* 报名时间：5月15日—7月21日
* 样例数据集发布：5月20日
* 结果提交：8月8日
* 结果及测试数据集公示：8月15日
* CCKS会议日期(评测报告及颁奖)：9月19日—22日

**8、组织者信息**

**任务组织者**：

**张 静，军事科学院系统工程研究院**

**任务联系人邮箱：**

**（可通过该邮箱联系沟通）**

**lrp\_ph@163.com**

# 附录

##  evaluation\_prompt.py

# 问题相关性：S1
prompt\_1 = '''
### 文章内容:
{text}

### 问题:
{question}

### 评分结果示例：
{
    "score":78
}

### 任务:
请分析上述问题与提供的文章内容的相关性，并给出一个0到100的分数来量化这种相关性。0表示完全不相关，100表示问题直接且紧密地关联于文章的核心信息。同时，请遵循评分结果示例的格式返回结果。
'''
# 问答匹配度:S2
prompt\_2 = '''
### 文章内容:
{text}

### 问题类型:
{type}

### 问题:
{question}

### 答案:
{answer}

### 评分结果示例：
{
    "score":78
}

### 任务:
请综合考虑文章内容、问题类型以及提供的问题与答案，评估它们之间的匹配度。特别注意，问题类型对评价标准有重要影响：例如，事实性问题的答案需直接且准确引用文章中的信息；解释性问题则要求答案能合理解析文章相关内容；推理性问题的解答应基于文章内容进行合理推断。请仅输出一个0到100的整数，用来量化问题与答案的匹配程度，其中0表示完全不匹配，100表示答案完美符合问题要求及类型标准。同时，请遵循评分结果示例的格式返回结果。
'''
# 问题流畅度、答案流畅度:S3,S4
prompt\_3 = '''
### 问题类型:
{type}

### 问题内容:
{text}

### 评分结果示例：
{
    "score":78
}

### 任务:
请依据问题类型及内容，评估该问题表述的清晰度、语法正确性、逻辑连贯性以及对目标受众的适宜性。其中0分表示问题表述极不流畅，存在严重理解障碍；100分表示问题表述完美，逻辑清晰，易于理解。
同时，请遵循评分结果示例的格式返回结果
'''
# 问题复杂度：针对复杂问答题，S5
prompt\_4 = '''
### 文章内容:
{text}

### 问答对:
\*\*问题:\*\* {question}
\*\*回答:\*\* {answer}

### 评分结果示例：
{
    "score":78
}

### 复杂度评估任务:
请根据文章内容及上述问答对，评估该问答对的复杂程度。复杂度考量因素包括但不限于：
- \*\*问题深度:\*\* 问题是否要求深入理解文章主题或涉及高级概念？
- \*\*逻辑结构:\*\* 问题与答案的逻辑是否复杂，是否包含多层次推理？
- \*\*语言表达:\*\* 是否使用了复杂的词汇、术语或句式结构？
- \*\*跨段落:\*\* 问题的答案是否需要通过阅读多个段落获得？

请给出一个0-100的分数来量化其复杂程度，其中0表示非常简单，100表示极其复杂。同时，请遵循评分结果示例的格式返回结果：
'''

##  [demo.py](https://demo.py)

#!/usr/bin/env python
# -\*- coding: utf-8 -\*-

def qag(text,question\_type):
 '''
    :text: 输入的文本内容
    :question\_type: 期望生成的问题类型
    '''
 # TODO: question and answer generation
 # 需要选手自行完成
 # result:根据问题类型，返回对应格式的问答对结果
    result = {
 "question\_type\_id":question\_type,
 "question": "第十一届航空展在哪里举行的？",
 # 对于非选择题，choices字段值为空就可以
 "choices":["A 珠海","B 深圳","C 香港","D 成都"]
 "answer":"A"
 }
 return result