高人 CCF-基于文心 NLP 大模型的阅读理解可解释评测

网址：https://www.datafountain.cn/competitions/589/datasets

注意事项

参赛选手须使用深度学习平台飞桨进行模型的设计、训练和预测，不得使用其他相关平台、框架及任何飞桨中未包含的学习方法参赛。

本赛题以 BDCI 大赛官方竞赛平台评测结果为准，除 BDCI 官方竞赛平台以外，百度飞桨 AI Studio 作为官方指定的竞赛日常训练平台，可为参赛选手练习机会，需要的参赛选手可在百度飞桨 AI Studio 平台注册报名进行练习。

AI Studio 平台报名地址：

https://aistudio.baidu.com/aistudio/competition/detail/394/0/introduction?ad-from=bs-CCF1

点击查看：

飞桨官方基线：

https://aistudio.baidu.com/aistudio/projectdetail/4282843?ad-from=bs-DF1

文心大模型官网： https://wenxin.baidu.com/?fr=bs

飞桨官网： https://www.paddlepaddle.org.cn

赛题任务

本次比赛聚焦于阅读理解任务的可解释评测。选手需使用飞桨（PaddlePaddle）深度学习框架，根据给定的一段文本 T 及与其相关的问题 Q，从文本 T 中抽取问题 Q 对应的的答案，同时给出模型预测答案所依赖的证据。如下方示例所示，“答案”和“证据”字段对应模型输出，其中答案和证据皆来自输入文本 T。

- **示例**

文本(T)：一公里等于两里，可以通过公里和里之间的关系来进行换算，就可以得出最终的结果。一公里等于多少里。

问题(Q)：一公里等于多少里

文本分词：[“一”，“公”，“里”，“等”，“于”，“两”，“里”，“，”，“可”，“以”，“通”，“过”，“公”，“里”，“和”，“里”，“之”，“间”，“的”，“关”，“系”，“来”，“进”，“行”，“换”，“算”，“，”，“就”，“可”，“以”，“得”，“出”，“最”，“终”，“的”，“结”，“果”，“。”，“一”，“公”，“里”，“等”，“于”，“多”，“少”，“里”]

答案：两里

证据：[“0”，“1”，“2”，“3”，“4”，“5”，“6”]

基于参赛系统给出的预测依赖证据，也就是输入文本中对预测影响较大的若干词（我们的基线系统提供了基于注意力权重、梯度等多种证据抽取方法，可供大家参考及使用），我们给出模型的可解释性评测结果。

数据简介

我们仅提供评测数据，其包含 4366 条数据（对于模型训练数据，推荐大家使用 DuReader_Checklist 数据集)。我们基于 DuReader_chechlist 的测试集构建我们的评测数据，针对每一条评测数据都人工标注了证据。同时，针对每一条数据，我们从抗干扰性、敏感性、泛化性等维度标注了扰动数据。我们的评测数据集中包含 2042 原始数据和 2324 条扰动数据。

数据说明

数据格式：JSON;

- **输入数据字段及说明：**

Id：数据的编号，作为该条数据识别 key;

context：原文本数据;

question：问题文本数据;

sent_token：原文本数据的标准分词，注意：golden 证据是基于该分词的，预测证据也需要与该分词对应；输入数据示例：

{

“id”: 452,

“context”：“一公里等于两里，可以通过公里和里之间的关系来进行换算，就可以得出最终的结果。一公里等于多少里”，

“question”：“一公里等于多少里”，

“sent_token”: [“一”，“公”，“里”，“等”，“于”，“两”，“里”，“，”，“可”，“以”，“通”，“过”，“公”，“里”，“和”，“里”，“之”，“间”，“的”，“关”，“系”，“来”，“进”，“行”，“换”，“算”，“，”，“就”，“可”，“以”，“得”，“出”，“最”，“终”，“的”，“结”，“果”，“。”，“一”，“公”，“里”，“等”，“于”，“多”，“少”，“里”]

}

- **评测数据字段及说明：**

sent_id：数据的编号，作为该条数据识别 key;

sent_text：原文本数据;

question：问题文本数据;

sent_token：原文本数据的标准分词，注意：golden 证据是基于该分词的，预测证据也需要与该分词对应；

rationale_tokens：人工标注证据的标准分词；

rationales：人工标注证据对应的标准分词 ID；

sample_type：数据类型（原始数据还是扰动数据）；

rel_ids：原始数据对应的扰动数据 ID 的列表；

ans：问题对应的答案；评测数据示例：

{

“sent_id”: 452,

“question”：“一公里等于多少里”，

“sent_text”：“一公里等于两里，可以通过公里和里之间的关系来进行换算，就可以得出最终的结果。一公里等于多少里”，

“sent_token”: [“一”，“公”，“里”，“等”，“于”，“两”，“里”，“，”，“可”，“以”，“通”，“过”，“公”，“里”，“和”，“里”，“之”，“间”，“的”，“关”，“系”，“来”，“进”，“行”，“换”，“算”，“，”，“就”，“可”，“以”，“得”，“出”，“最”，“终”，“的”，“结”，“果”，“。”，“一”，“公”，“里”，“等”，“于”，“多”，“少”，“里”],

“rationale_tokens”: [“一”，“公”，“里”，“等”，“于”，“两”，“里”],

“rationales”: [“0”，“1”，“2”，“3”，“4”，“5”，“6”],

“sample_type”：“ori”，

“rel_ids”: [2776],

"ans ": “两里”

}

提交要求

选手提交命名为 mrc_rationale.txt 的文件，压缩文件名为 mrc_rationale.txt，文件内部格式为：

每行为一个输入文本的预测结果。
每行包含 3 列内容，分别为 sent_id（必需字段）、predicted answer（必需字段）、rationale list（必需字段），以 table 键隔开。其中，sent_id（输入的编号）来自测试集文件，predicted answer 是模型对于问题 Q 预测的答案，rationale list 是给出的证据（按重要度顺序给出 token id 序列，按逗号隔开）

评测标准

我们分别基于模型预测答案、证据来评估模型本身表现和其可解释性。

对于模型本身表现，我们采用 F1-score 指标（见公式一）来评估。对于可解释性，我们从合理性、忠诚性 2 个维度来评估。合理性评估模型给出的证据与人工标注证据的拟合程度，我们使用 Macro-F1（见公式二）作为评估指标。忠诚性评估模型预测实际对证据的依赖程度，我们使用扰动下证据的一致性来评估忠诚性，并采用 MAP（见公式三）作为评估指标。更多内容见我们开源项目 TrustAI 中的 可信评测 部分。

公式一：

$$F1-\text{scor}e=\frac{1}{N}\sum_{i=1}^{N}\left(2\times\frac{P_i\times R_i}{P_i+R_i}\right)$$

公式二：

$$F1-\text{score}=\frac{1}{N}\sum_{i=1}^{N}\left(2\times\frac{P_i\times R_i}{P_i+R_i}\right)$$

$$\text{where}\quad P_i=\frac{|s_i^p\cap s_i^g|}{|s_i^p|}\text{ and }R_i=\frac{|s_i^p\cap s_i^g|}{|s_i^g|}$$

公式三：

$$MAP=\frac{\sum_{i=1}^{|X^p|}\left(\sum_{j=1}^{i}G\left(x_j^p,X_{1:i}^o\right)\right)/i}{|X^p|}$$

公平竞技

参赛团队需共同维护竞赛环境的公平公正，禁止在指定考核技术能力的范围外，利用规则漏洞或技术漏洞等不良途径提高成绩与排名，禁止在比赛中抄袭他人作品、交换答案、使用多个小号，一经发现将取消比赛成绩并严肃处理。

DataFountain 基于自动化反作弊系统、结合人工审核，赛中动态反违规、反作弊，若收到团队封禁通知，可在指定页面申诉。