https://www.biendata.com/competition/2019diac/

**比赛背景**

虽然近年来智能对话系统取得了长足的进展，但是针对专业性较强的问答系统（如法律、政务等），如何准确的判别用户的输入是否为给定问题的语义等价问法仍然是智能问答系统的关键。举例而言，**“市政府管辖哪些部门？”**和**“哪些部门受到市政府的管辖？”**可以认为是语义上等价的问题，而**“市政府管辖哪些部门？”**和**“市长管辖哪些部门？”**则为不等价的问题。

针对问题等价性判别而言，除去系统的准确性外，系统的鲁棒性也是很重要、但常常被忽略的一点需求。举例而言，虽然深度神经网络模型在给定的训练集和测试集上常常可以达到满意的准确度，但是对测试集合的稍微改变（Adversarial Attack）就可能导致整体准确度的大幅度下降（一些相关文献综述可见<https://arxiv.org/pdf/1902.07285.pdf>和<https://www.mitpressjournals.org/doi/full/10.1162/tacl_a_00254>）。

如以下样例：

|  |  |
| --- | --- |
| **origin example** | **adversarial example** |
| 检察机关提起公益诉讼是什么意思 | 监察机关提起公益诉讼是什么意思 |
| 检察机关提起公益诉讼是什么意思 | 检察机关发起公益诉讼是什么意思 |
| 寻衅滋事一般会怎么处理 | 寻衅兹事一般会怎么处理 |
| 什么是公益诉讼 | 请问什么是公益诉讼 |

右列数据是左列数据出现一些错别字或者添加一些无意义的干扰词产生的，并不影响原句的意思，但是这样的微小改动可能会导致完全不同结果。从用户的角度而言，这意味着稍有不同的输入就可能得到完全不一样的结果，从而严重降低用户的产品使用体验。

**任务提交**

参赛选手需要在验证集和测试集（测试集将于比赛结束前一天发布）上提交预测结果。提交文件格式参考sample\_submission.csv。

注意：**请勿改变提交文件的行的顺序**。文件每行末尾不要有多余的空格，包括header line和每一行各列之间；文件末尾加一个空行；文件要用无BOM的utf8编码。

**比赛讨论区**

添加请备注：姓名，机构，2019DIAC

**规则**

1. 所有参赛选手都必须在管理系统中注册；

2. 参赛选手在管理系统中组队，每支队伍最多不超过5名队员，一旦加入一支队伍，不允许退出，队伍也不允许解散；

3. 每支队伍需指定一名队长，队伍名称不超过15个字符；

4. 每名选手只能参加一支队伍，一旦发现某选手以注册多个账号的方式参加多支队伍，将取消相关队伍的参赛资格；

5. 参赛队伍可以合并，也可以邀请队友，但不能拆分，且合并后的队伍不得超过总人数上限（5人）；

6. 允许使用开源代码或工具，但不允许使用任何未公开发布，也不能使用需要授权的代码或工具；

7. 不允许使用主办方提供的数据集之外的任何外部标注数据，可以使用数据增强后的数据集或从其他渠道获得的无等价性标注的数据集。参赛选手不得在未经主办方授权情况下将比赛数据作为其他用途使用；

8. 参赛队伍可在参赛期间随时上传验证集的预测结果，一天不能超过2次；请注意，组队时，新成立的队伍的所有队员的历史提交数不能超过比赛开始天数乘以每日提交上限次数。例如，如果三名成员希望在比赛开始的第五天组队，他们三人过去总提交数不能超过10天。所以，请尽早组队；

9. 获奖队伍必须提交可复现结果的参赛代码，运行说明和算法说明书以备审核，参赛选手需要配合主办方对比赛作品的有效性与真实性进行验证；

 10. 获奖模型必须遵守GPL v3协议，并发布在组织者指定的网站上；

11. 参赛作品必须保证原创性，不违反任何中华人民共和国的有关法律，不侵犯任何第三方知识产权或者其他权利；一经发现或经权利人提出并查证，主办方将取消其比赛成绩并进行严肃处理；

12. 比赛作品的知识产权归参赛作者所有。主办方有权将参赛作品、作品相关、参赛团队信息用于宣传品、相关出版物、指定及授权媒体发布、官方网站浏览及下载、展览（含巡展）等活动项目。

### 数据集

|  |  |  |  |
| --- | --- | --- | --- |
| 名称 | 格式 | 链接 | MD5 |
| baifendian\_data.zip | zip 992.00KB | [百度](https://pan.baidu.com/s/1KY_b7lfpIABb54Ov33A2OQ)( 提取码：ayvx ) ， [Dropbox](https://www.dropbox.com/s/bbz0giitt8fjdby/baifendian_data.zip?dl=0" \t "_blank) | 9faef2b6c07c7a8680b3cf713b0e4b2b |

**数据**

**1. 数据详情**

本次大赛提供的是一个法律领域的问句等价性数据集，该数据集为我们在实际项目中开发系统所使用的数据集。除去该数据集外，参赛选手可以使用数据增强后的数据集或从其他渠道获得的无等价性标注的数据集。**参赛选手不得使用人工进行标注的数据。**

训练集根据在实际项目中的数据情况，以问题组的形式提供，每组问句又分为等价部分和不等价部分，等价问句之间互相组合可以生成正样本，等价问句和不等价问句之间互相组合可以生成负样本。我们提供6000组问句的训练集，每组平均有三个等价问句和3个不等价问句。验证集和测试集则以问句对的格式提供，其中验证集有5000条数据。测试集中除了人工标注的样本外，还会有大量adversarial example。

训练集用于模型的学习，比赛期间选手可以提交验证集的预测结果，但最终成绩由测试集的预测结果决定，由于测试集中会有大量adversarial example，因此最终在测试集上的成绩可能会与验证集上不同。（测试集将于比赛结束前24小时发布，具体可参考时间轴页面。）

**2. 数据格式**

**train\_set.xml**

训练集以XML文件提供，用于训练模型。XML文件中内容格式如下：

|  |
| --- |
| <?xml version="1.0" encoding="utf8"?>  <TrainCorpus>  <Questions number="0">  <EquivalenceQuestions>  <question>什么是公益诉讼？</question>  <question>公益诉讼的定义？</question>  <question>公益诉讼的概念</question>  <question>公益诉讼的定义是什么？</question>  </EquivalenceQuestions>  <NotEquivalenceQuestions>  <question>环境公益诉讼的原告是什么意思？</question>  <question>什么样的鉴定依据算是民事公益诉讼</question>  <question>检察机关提起公益诉讼是什么意思</question>  <question>什么是行政诉讼？</question>  </NotEquivalenceQuestions>  </Questions>  <Questions number="1">  <EquivalenceQuestions>  <question>检察机关提起公益诉讼的目的是什么？</question>  <question>检察机关为什么要提起公益诉讼？</question>  <question>检察机关提起公益诉讼的目的？</question>  </EquivalenceQuestions>  <NotEquivalenceQuestions>  <question>十八大提出由检察机关提起公益诉讼有利于</question>  <question>如何认识检察机关在公益诉讼中的地位</question>  <question>检察机关提起公益诉讼是什么意思</question>  <question>检察机关提起公益诉讼有哪些优势</question>  </NotEquivalenceQuestions>  </Questions>  </TrainCorpus> |

每一个Questions标签中为一组数据，其中EquivalenceQuestions标签内的问句之间互为等价关系，NotEquivalenceQuestions标签内的问句与EquivalenceQuestions为不等价关系。EquivalenceQuestions之间的问句互相组合可以生成正样本（label为1），EquivalenceQuestions和NotEquivalenceQuestions之间的问句互相组合可以生成负样本（label为0），具体需要生成多少正样本多少负样本由参赛选手自行决定。

**dev\_set.csv**

此数据集用于测试，数据格式如下：

|  |  |  |
| --- | --- | --- |
| **qid** | **question1** | **question2** |
| 1 | 醉酒驾驶，保险公司赔偿吗 | 当事人醉酒驾驶，保险公司会不会赔偿 |
| 2 | 酒驾会吊销驾照吗 | 醉驾会被吊销驾照吗 |
| 3 | 被他人的摩托车撞到，导致骨折，对方应赔多少 | 被摩托车撞到骨折，该赔多少 |

**test\_dev.csv**

此数据集用于最终结果的评定，其格式和dev\_set一致，不同的是该数据集中会包含大量adversarial example。

**sample\_submission.csv**

提交样例，格式如下：

|  |  |
| --- | --- |
| **qid** | **label** |
| 1 | 1 |
| 2 | 0 |
| 3 | 1 |
| ... | ... |

其中，qid是问句对id，label是该id对应的问句对的等价关系，int类型，取值范围为{0,1}。qid和label之间使用tab分隔符('\t')分隔。**提交文件不需要添加header, 请参考sample submission。**