2022 全球算法精英大赛初赛参赛手册

题目： 知识驱动口语对话

1. 简介

当前生成式对话技术存在回复无意义、信息量少等问题，因此希望对话系统能充分利用外部知识，生成更有意义、内容更丰富的回复。知识涉及多个领域，各领域的样本量并不一定均衡；并且有时需要对图谱知识进行复杂的查询或推理，才能得到问题的答案。此外，现实的对话往往有口语化表达和 ASR 识别错误等问题，因此希望对话模型鲁棒性更强，能够容忍口语表述不规范、不精确以及 ASR 中的词法或语法错误问题，并生成正确的标准的回复。

2. 赛题说明

本题目将为选手提供多轮对话数据和知识图谱数据，参赛选手基于给定的数据构建知识对话模型。赛题任务如下：

2.1 知识选择

- 目标：理解用户问题，选择与用户问题相关的知识三元组；
- 输入：对话历史、知识库
- 输出：知识三元组
- 评估指标：精准率、召回率、F1 值

2.2 对话生成

- 目标：结合对话模型和知识三元组，生成回复语句；
- 输入：对话历史、知识库、知识三元组
- 输出：自然、流畅、合理的回复语句
- 评估指标：自动化评估指标（基于字的 BLEU-1/2，Distinct-1/2，generation_F1 值）和人工评估（丰富度 0-2，连贯性 0-2，知识准确率 0-2）

数据划分成训练集/验证集/测试集，对每支队伍的各个自动评估指标计算加权得分，先根据加权后最终得分筛选出 top15 队伍。然后评审人员再对 top15 模型进行人工评估，决出 top7 队伍，最终比赛排名以人工评审为准。

3. 数据说明

数据包括对话数据和图谱数据，对话数据划分为训练集、验证集和测试集。

3.1. 训练集包含多个对话样例，对话中标注了话语 message 涉及的知识三元组 attrs，一个语句可能涉及多个知识三元组，每个知识三元组来自图谱文件，包含了 attrname，attrvalue，name 三个字段信息，数据样例如下：

- "messages":[
- {
  - "message":"对话语句"

  },
- {
  - "message":"对话语句"
  - "attrs":[
    - "attrname":"实体属性名",
    - "attrvalue":"实体属性值",
    - "name":"实体"]

  },

  …

]

- "name":"对话开始的实体"

3.2. 验证集的数据格式和训练集类似，不同的是，验证集中部分语句带有 ASR 错误，参赛者需要正确理解带有 ASR 错误的语句，选择正确的知识并给出正确的回复语句。

3.3. 测试集中给出多个对话样例，每个对话样例带有 id 和对话历史，参赛者根据对话历史，选择知识（如果语句涉及知识），给出回复，数据样例如下：

- "样例 id":[
  - {"message":"对话语句"},
    …
  ]

3.4. 知识图谱文件

知识图谱文件包含多个实体，实体包含多个属性三元组，属性三元组的格式为（实体，属性，属性值）。

- "实体":[
  - ["实体",
  - "属性",
  - "属性值"],

  …
  ]

4. 评估方式

对每支队伍的各个自动评估指标计算加权得分，根据分数筛选 top15 队伍。评审人员对 top15 模型进行人工评审，决出 top7 队伍，最终比赛排名以人工评审为准。

4.1、知识选择采用精确率（Precision），召回率（Recall），F1 值进行评估：

$$Precision = \frac{Count(correct\ predicted\ knowledge\ triples)}{Count(predicted\ knowledge\ triples)}$$

$$Recall = \frac{Count(correct\ predicted\ knowledge\ triples)}{Count(ground-truth\ knowledge\ triples)}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Count(correct predicted knowledge triples) 表示预测正确的知识三元组数量，Count(predicted knowledge triples)表示当前样例预测的知识三元组数量，Count(ground − truth knowledge triples)表示当前样例真实的知识三元组数量。

4.2、文本生成采用基于字粒度计算 BLEU-N（N=1,2）、DISTINCT-N（N=1,2）和 generation_F1 值进行评估。计算公式如下：

$$BLEU - N = BP \cdot exp(\sum_{n=1}^{N} w_n\ logP_n)$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{1-r/c} & if\ c \le r \end{cases}$$

其中w_n表示权重，取 1/N；N 取值为[1,2]；P_n表示 n-gram 准确率；c 表示预测回复文本长度，r 表示标准回复文本长度。

$$DISTINCT - N = \frac{Count(unique\ ngram)}{Count(prediction\_response\ ngram)}$$

$Count(unique\ ngram)$ 表示回复中不重复的 ngram 数量，$Count(prediction_response\ ngram)$表示预测回复中 ngram 的总数量，N 取值为[1,2]，DISTINCT-N 越大表示生成的多样性越高。

$$p = \frac{Count(common\ word)}{Count(prediction\ response\ word)}$$

$$r = \frac{Count(common\ word)}{Count(\text{ground} - truth\ response\ word)}$$

$$generation\_F1 = \frac{2 * p * r}{(p + r)}$$

$Count(common\ word)$ 表示预测回复和标准回复文本中共同出现的字，$Count(prediction\ response\ word)$ 表示预测回复文本长度，$Count(\text{ground} - \text{truth}\ response\ word)$表示标准回复文本长度。

4.3、对每支队伍的各个指标进行加权得分，公式如下：

$$score = 0.3 * (Precision + Recall + F1) + 0.7 * (BLEU - 1 + BLEU - 2 + generation\_F1)$$

4.4、根据总分数筛选 top15 队伍，评审人员对 top15 模型进行人工评估，决出 top7 队伍，最终比赛排名以人工评审为准。人工评估指标包括丰富度、连贯性和知识准确率，具体描述如下：

丰富度（0-2）：评价回复句子本身的信息丰富程度。

连贯性（0-2）：评价回复句子回复输入上文的合适程度，是否话题契合、逻辑正确等。

知识准确率（0-2）：评价回复句子所用知识的准确率。

5. 提交方式

选手提交结果为一个 result.json 文件，编码采用无 BOM 的 UTF-8。需要给出对应样例的 id，以及其中涉及的 attrs，和回复文本 message。如果样例不涉及知识，只需给出对应的 message，数据格式如下所示：

- "样例 id":{
  - "message":"对话语句"
  - "attrs":[
    - "attrname":"实体属性名"
    - "attrvalue":"实体属性值"
    - "name":"实体"]

  },
- "样例 id":{
  - "message":"对话语句"}