Manual for 2022 Huawei GLOBAL AI CHALLENGE (Preliminary)
Proposal: Knowledge-driven spoken dialogue

1. Introduction

Chatbots are often let down by their passiveness, which often leads to meaningless responses to user requests and limited information. That is why a dialogue system that is capable of interpreting messages and being informative is needed. However, required knowledge may come from different domains, and the samples are not evenly distributed across these domains. Worse still, a dialogue system may need to make complex queries or inferences to fetch the correct information from the knowledge graph. Another challenge is when users verbally make requests, which are converted to text using the automatic speech recognition (ASR) technology, the converted text often contains errors. Therefore, a dialogue model is expected to be more robust when encountering errors caused by colloquial expressions, as well as word and syntax errors in text converted using ASR, thus giving users standard and expected responses.

2. Proposal Description

The data of multi-round dialogues and knowledge graphs will be provided for you to set up knowledge-driven dialogue models. The tasks are as follows:

2.1. Knowledge selection

- Objective: Interpret user questions and select the correct knowledge triples to answer the questions.
- Input: dialogue history and knowledge base
- Output: knowledge triples
- Scoring indicators: precision, recall, and F1 score

2.2. Response generation

- Objective: Generate responses based on the dialogue model and selected knowledge triples.
- Input: dialogue history, knowledge base, and knowledge triples
- Output: natural, smooth, and reasonable responses
- Scoring indicators: BLEU-1/2, DISTINCT-1/2, and generation_F1, all of which are calculated based on Chinese characters, for automatic scoring; informativeness (0–2), coherence (0–2), and factual accuracy (0–2) for evaluation by judges

We will provide you with a training set, a validation set, and test sets. A team's rank is determined by the sum of their weighted scores in indicators for automatic scoring. The 15 models with the highest total scores will then be evaluated by judges, who will select 7 teams out of the 15 ones. Evaluation by judges assumes a major role in the scoring process.

3. Data Description

You will be provided with dialogue data and knowledge graph data. There are three dialogue data sets: training, validation, and test sets.

3.1. The training set contains multiple dialogue samples. In each dialogue sample, the knowledge triples (defined as **attrs**) involved in the utterance text (defined as **message**) are annotated. Knowledge triples are fetched from the knowledge graph,

and each triple consists of **attrname**, **attrvalue**, and **name**. Here is a data example.

- "messages":[
- {
  - "message":"Utterance text"

  },
- {
  - "message":"Utterance text"
  - "attrs":[
    - "attrname":"Entity attribute name"
    - "attrvalue":"Entity attribute value"
    - "name":"Entity"]

  },

  ...

  ]
- "name":"Entity mentioned at the beginning of the dialogue"

3.2. The data format of the validation set is similar to that of the training set. The difference is that, some utterance text in the validation set contains ASR errors. Your model needs to correctly interpret erroneous text, select the right knowledge points, and give the expected response.

3.3. A test set contains multiple dialogue samples, each of which is assigned a sample ID and contains historical dialogue data. Your model needs to select knowledge triples (if involved in the utterance text) from the knowledge graph and generate a response. Here is a data example.

- "Sample ID":[
  - {"message":"Utterance text"},

    ...

  ]

3.4. Knowledge graph file

The knowledge graph file contains multiple entities. Each entity has an entity name, attribute, and attribute value, all of which compose an attribute triple.

- "Entity":[
  - ["Entity",
  - "Attribute",
  - "Attribute value"],

  ...

  ]

4. How We Score

A team's rank is determined by the sum of their weighted scores in indicators for automatic scoring. The 15 models with the highest total scores will then be evaluated by judges, who will select 7 teams out of the 15 ones. Evaluation by judges assumes a

major role in the scoring process.

4.1. Scoring indicators for knowledge selection are **Precision**, **Recall**, and **F1**. They are calculated as follows:

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

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

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

In the preceding formulas, $\text{Count(correct predicted knowledge triples)}$ indicates the number of correct knowledge triples predicted by your model for a sample; $\text{Count(predicted knowledge triples)}$ indicates the number of all knowledge triples predicted by your model for the sample; and $\text{Count(ground} - \text{truth knowledge triples)}$ indicates the number of correct knowledge triples in the sample.

4.2. Scoring indicators for text generation are **BLEU-***N* (*N* = 1 or 2), **DISTINCT-***N* (*N* = 1 or 2), and **generation_F1**, which are calculated based on Chinese characters as follows:

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

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

In the preceding formula, w_n indicates the weight and equals 1/*N*, where the value of *N* ranges from 1 to 2; P_n indicates the accuracy of ***n*gram**, **c** indicates the text length of the response predicted by your model, and **r** indicates the text length of the standard response.

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

In the preceding formula, $Count(unique\ ngram)$ indicates the number of ***n*gram** strings that appear only once in the response; $Count(prediction\_response\ ngram)$ indicates the number of all ***n*gram** strings in the response predicted by your model. The value of *N* ranges from 1 to 2, and a greater **DISTINCT-***N* indicates higher diversity of the generated response.

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

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

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

In the preceding formulas, $Count(common\ word)$ indicates the number of words that appear both in the response predicted by your model and the standard response; $Count(prediction\ response\ word)$ indicates the text length of the response predicted by your model; and $Count(\text{ground} - \text{truth}\ response\ word)$ indicates the text length of

the standard response.

4.3. A team's rank is determined by the sum of their weighted scores in indicators for automatic scoring. The formula is as follows:

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

4.4. The 15 models with the highest total weighted scores will then be evaluated by judges, who will select 7 teams out of the 15 ones. Evaluation by judges assumes a major role in the scoring process. Scoring indicators for this phase are as follows:
Informativeness (0–2): how informative the response generated by your team's model is.
Coherence (0–2): how cohesive the response generated by your team's model is with the dialogue context, which is determined by topic relevancy, logic, and other factors.
Factual accuracy (0–2): how accurate the knowledge triples fetched by your team's model to answer a user question is.

5. How to Submit

Submit a **result.json** file encoded in UTF-8 without BOM. The file shall contain **id** (ID of the corresponding dialogue sample; mandatory), **attrs** (knowledge triples; mandatory if required for generating a response to the utterance text in the dialogue sample), and **message** (response generated by your model; mandatory). Here is a data format example.

- "Sample ID":{
  - "message":"Utterance text"
  - "attrs":[
    - "attrname":"Entity attribute name"
    - "attrvalue":"Entity attribute value"
    - "name":"Entity"]

  },
- "Sample ID":{
  - "message":"Utterance text"}