Multimodal Dialogue Understanding and Generation

1. Background

The multimodal dialogue understanding and generation task can be divided into two phases: multimodal context understanding and response generation. Specifically, the former includes dialogue session identification (i.e., determining whether the dialogue content has changed) and dialogue scene identification (i.e., determining whether the video context has changed). The ultimate goal is to generate a response that is coherent to the dialogue context and relevant to the video context.

2. Task Overview

This task includes three tracks:

- Track 1. Dialogue scene identification: predict the boundaries of different dialogue scenes given a set of continuous dialogue utterances and a related video.
- Track 2. Dialogue session identification: predict the boundaries of different dialogue sessions given a set of continuous dialogue utterances and a related video (which is identical to Track 1).
- Track 3. Dialogue response generation: generate a response based on scene and session predictions, while coherently catching up with the conversation.

Track 1

Input: A dialogue session and the corresponding video.

Output**:** A binary value indicating whether or not each utterance in the session begins a new dialogue scene.

Track 2

Input**:** A dialogue session and the corresponding video.

Output**:** A binary value indicating whether or not each utterance in the session begins a new dialogue topic.

Track 3

Input**:** A dialogue session and the corresponding video, predicted session identification results.

***Output*:** If the session's last utterance is not the end of the current dialogue topic or dialogue scene, the output is an utterance related to the current dialogue topic or dialogue scene.

3. Dataset

The videos and dialogues for this task are crawled from online TV series, with the dialogues being the subtitles (in Chinese and English) of the associated video snippets. The transition of topics and scenes is manually labeled through crowdsourcing. The dataset is split into a training set, a validation set, and a test set. During the competition, the test set is not available to the public. Each example includes a dialogue session as well as the associated video clip, which is a sequence of frames. The frames from the videos have been downsampled to 3fps. Each frame has two labels, which indicate whether the scene and topic have changed.

An example is displayed as follows:


```
{
    "vid": "Designated.Survivor_S02E01_clip_010",
    "dialogues":
        [{
            "turn_id": 0,
            "en_text": "It is two pages!",
            "ch_text": "一共只有两页",
            "start": 0.0,
            "end": 1.26,
            "scene": 0,
            "topic": 0},
        {
            "turn_id": 1,
            "en_text": "Chicken korma from Rasika, and you're not gonna touch it?",
            "ch_text": "拉西卡的腰果滑汁鸡　你都不尝一下吗",
            "start": 8.44,
            "end": 11.85,
            "scene": 1,
            "topic": 1},
        {
            "turn_id": 2,
            "en_text": "Sorry, I'm just not hungry.",
            "ch_text": "抱歉　我只是不饿",
            "start": 12.97,
            "end": 14.58,
            "scene": 0,
            "topic": 0},
        …
```

where "vid" denotes this example's identification and "turn id" is the index of the utterance in the dialogue session. "start" and "end" refer to the beginning and end of the session in the original video, respectively. "scene" and "topic" are binary, with 1 indicating that the current utterance belongs to a new dialogue scene or topic and 0 indicating that the scene or topic remains unchanged. More details about the dataset as well as the download links can be found in https://github.com/patrick-tssn/NLPCC-2022-Shared-Task-4

4. Evaluation Metrics

We evaluate the dialogue scene identification accuracy (i.e., ACC_s) in track 1 and dialogue topic identification accuracy (i.e., ACC_t) in track 2. In track 3, we evaluate the BLEU, ROUGE, METEOR, and CIDER scores of the generated response.

Let S_i and $\hat{S}_i$ denote the predicted and golden labels of dialogue scene identification for the i^{th} utterance respectively, the ACC_s is computed as follows:

$$ACC_s = \frac{1}{I}\sum_{i=1}^{I} I_{\{S_i=\hat{S}_i\}}$$

Similarly, ACC_t is defined as follows:

$$ACC_t = \frac{1}{I}\sum_{i=1}^{I} I_{\{T_i=\hat{T}_i\}},$$

where T_i and $\hat{T}_i$ denote the predicted and golden labels of dialogue topic identification for the the i^{th} utterance respectively. BLEU, ROUGE, METEOR, and CIDER scores are calculated through https://github.com/tylin/coco-caption