Standard Project

Description

Emotion Discovery and Reasoning its Flip in Conversation (EDiReF), SemEval 2024 Task 10.

We only consider SUBTASK iii (English dialogues only)

The challenge provides the **training dataset** for research purposes.

The data contains 4000 short English dialogues: from 5 to 17 utterances

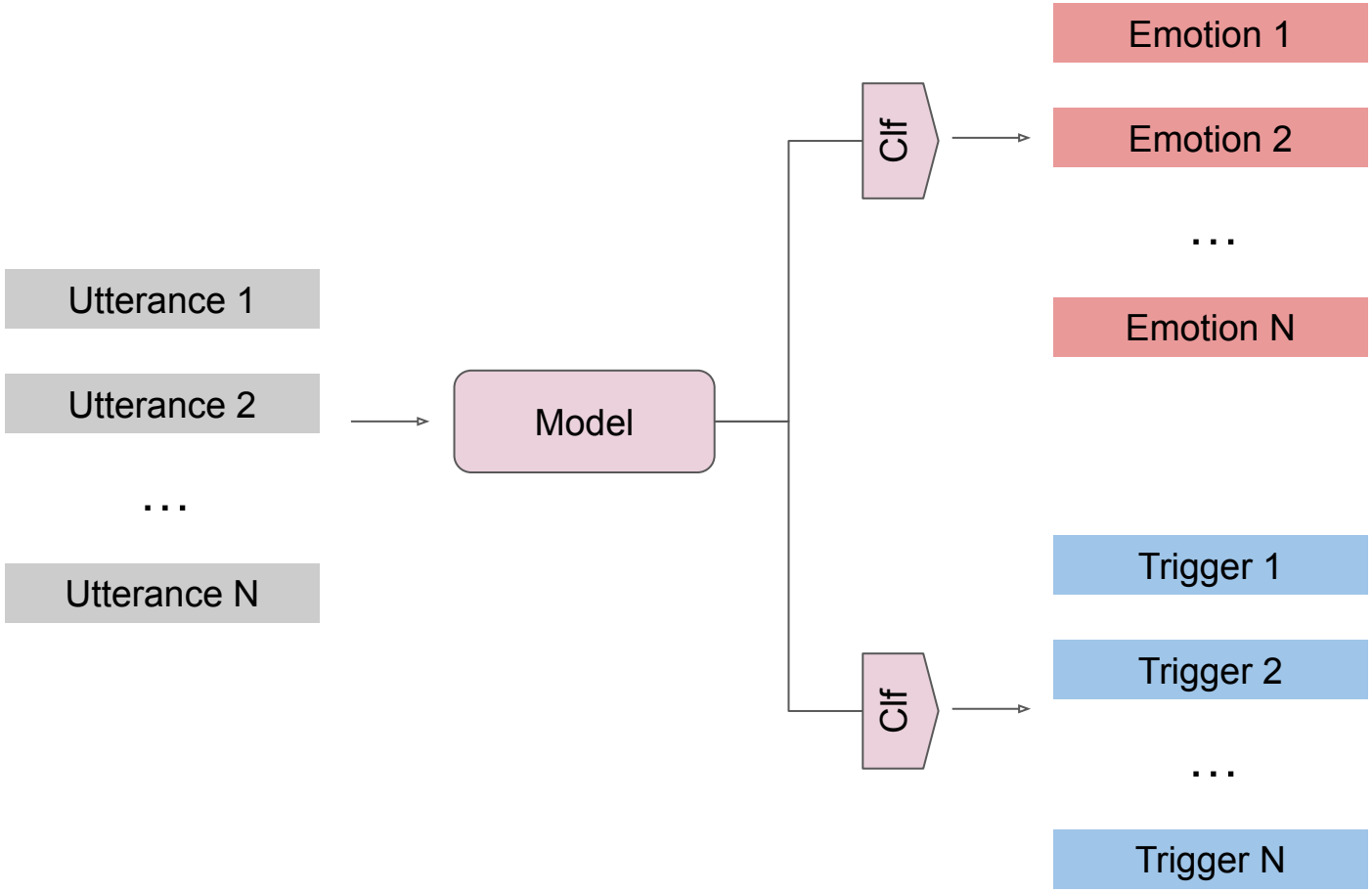
If you like, you can also **join the official challenge**.

However, we do not manage this and is not required in any way to pass the NLP exam.

Description


```
"emotions":
[
   "neutral",
   "surprise",
   "neutral",
   "surprise",
   "neutral",
   "anger"
],
"utterances":
[
   "Yes that's right.",
   "Why?",
   "I tired attacking two women, did not work.",
   "What?!",
   "No, I mean it's okay, I mean, they're-they're my friends.",
   "In fact, I-I-I was married to one of them."
],
"triggers":
[
   0.0,
   0.0,
   1.0,
   0.0,
   1.0,
   1.0
]
```

Task


Utterance 1
Utterance 2
...
Utterance N
Model
Clf
Clf
Emotion 1
Emotion 2
...
Emotion N
Trigger 1
Trigger 2
...
Trigger N

Recommendations

Some of the trigger labels are NaN

The way the dataset is provided, some trigger labels are not correctly formatted and appear as NaN.

You should convert them to zero to avoid errors.

Recommendations

No default data splits are provided.

You should perform an 80/10/10 train/val/test split.

Ensure you perform the split at the dialogue level: utterances from the same dialogue are in the same split.

Given the dataset format, this property should be trivially guaranteed.

The baselines presented in this document are trained and evaluated using the above-mentioned splitting strategy.

Recommendations

You should train and evaluate a BERT baseline on two different settings:

- **Freezed**: we freeze the BERT embedding layer weights and fine-tune the classifier heads on top

- **Full**: we fine-tune the whole model architecture. Make sure you set a small enough batch size. We recommend 1.

We recommend `bert-base-uncased`.

In addition, you should evaluate a random and a majority classifier for emotions and triggers.

You should report a comparison with your selected model(s) and the provided baselines.

Recommendations

You should report the following metrics for model evaluation.

Sequence F1: compute the f1-score for each dialogue and report the average score.

Unrolled Sequence F1: flatten all utterances and compute the f1-score.

Compute the above metrics for emotions and triggers labels.

Recommendations

To assess model robustness and report a sound evaluation, **you should train and evaluate your model(s) on 5 different seeds.**

You can pick the seeds as you like, but make sure you fix the seed at the beginning of your script(s).

You should report all metrics' average and standard deviation computed over the 5 seeds.

Any Questions?