PTG Model's Evaluation

Modified from Itamar Bar-Yossef

Recap of Prompts

- Prompt ChatGPT whenever there is a user/instructor utterance or no one said anything for the past 10 seconds.
- Every prompt asks ChatGPT these questions:
  1. What is their dialog intention? Choose among Question, Answer, Confirmation, Hesitation, Self Description, and Other
  2. Which step is the user at? For example Step 3
  3. Should you say anything? Yes or no
     3.1. If yes, what would you say?
     3.2. If yes, choose your dialog intention among Instruction, Confirmation, Question, Answer, or Other
     3.3. If your dialog intention is Instruction, is it about current step, next step, details, or mistake correction
  4. Did the user make a mistake? Yes or No? If yes, choose among wrong object, wrong state, wrong action

Overview

- Each response to a prompt looks like this (ish):

```
##### FRAME: 9167/14191
2. Step 11
3. Yes
3.1. "We should only be using a total of 5 toothpicks with one inch apart"
3.2. Instruction
3.3. Mistake Correction
4. Yes
wrong state
```

```
##### FRAME: 1356/14191
1. Self Description
2. The user is at step 1.
3. No.
4. Yes, wrong step.
```

```
##### FRAME: 10142/14191
1. Confirmation
2. Step 7
3. Yes
3.1. Repeat step 8: "We need to cut the pinwheels. Let's trim the ends of the tortilla
roll with the butter knife, leaving 1/2 inch margin between the last toothpick and the
end of the roll. Discard ends."
3.2. Instruction
3.3. current step
4. No
```

- We loop through all the responses in a trial, and for each response we evaluate the answer to each question and aggregate results across a trial and across all trials to compute micro/macro F1 scores
- The responses could be messy, takes a bit of filtering.

Evaluate Question 1 (Dialog intention of last user utterance)

- Compare GPT response to ground truth - dialog intent annotations can be found in the Dropbox.
- Keep a count of the number of correct answers as well as the number of total answers (some prompts don’t include question 1).
- Aggregate the counts across all trials

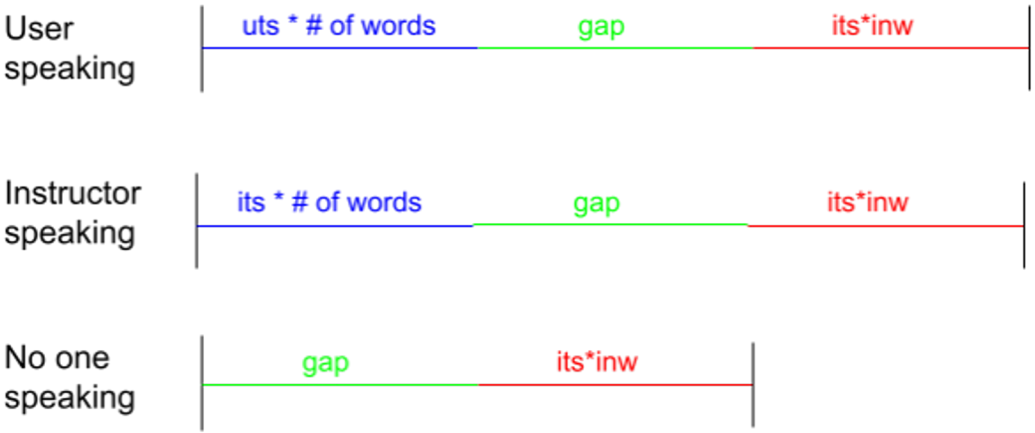
Evaluate Question 2 (Which step the user is at)

- Again, compare GPT response to ground truth step annotations.
- Keep a count of the number of correct answers by GPT and aggregate across all trials.

Evaluate Question 3 (Talk or Not?)

- To evaluate this question we first need to come up with a way to find the the ground truth of whether an instructor should talk or not in a given moment.
- We can calculate a time interval, that starts at the current time and ends by the time the instructor should have started talking.
- There are three cases:

Evaluate Question 3 (Talk or Not?)



According to Figure 2:

uts = 522 ms/word
its = 398 ms/word
inw = 6 words/utterance

uts - median user talking speed
its - median instructor talking speed
inw - median number of words per instructor comment
gap - 3 seconds long (eyeballed)

Evaluate Question 3 (Talk or Not?)

- If there is an instructor utterance within the interval we assume the instructor should talk at the current time, otherwise they shouldn’t.
- We use this as ground truth to evaluate ChatGPT’s response.
- If GPT correctly guessed it should talk, we also evaluate if it can guess the correct intention for Q 3.1 ~ Q 3.3.

Evaluate Question 4 (Did the user make a mistake?)

- This evaluation is slightly tricky because there isn't a direct annotation of the mistake per query point. The ground truth comes from Instructor dialog act annotation
- Assume that whenever a user makes a mistake, it elicits an utterance from instructor, with a mistake correction dialog intent.
- For example, user made a "wrong object" mistake in Frame A, Frame B is prompted due to an instructor utterance with intent "Instruction wrong object"