Annotator Guidelines

Clarifying Question Generation

Annotation Time: 60 minutes

Task Description

The text-generation task chosen is clarifying question generation. Given a story, 1-2 history turns and an ambiguous question, the model produces a clarifying question for the user. Following are the definitions of each of these terms:

- **Clarifying question:** A question that seeks to gather additional information or clarification about something that is not clear or ambiguous. In the context of this task, the model is expected to generate a question that helps to clarify the meaning or intention behind an ambiguous question provided by the user.

- **Story:** A piece of text or narrative that provides context or background information for the task. It typically contains information about events, characters, and situations that the model can use to generate relevant clarifying questions.

- **History turns:** Refers to previous interactions or statements in a conversation. These turns provide a history or context for the current question or response. In the context of this task, the history turns help the model understand the conversation context and generate appropriate clarifying questions based on that context.

- **Ambiguous question:** A question that is unclear or can have multiple interpretations or meanings. The model's task is to generate a clarifying question to seek further details or clarification from the user about the ambiguous question.

Note that for a question to be a clarifying question it is not mandatory for it to start with “Do you mean”. Questions which are able to clarify the ambiguous question or highlight one of the many possibilities that an ambiguous question might be referring to are also considered as good clarifying questions.

Evaluation

Each model being evaluated generates a pair of clarifying questions given a single story and a set of history turns. The following are the evaluation criteria for the clarifying questions generated:

Grammatical Correctness: Judges whether a question generated is grammatical correct according to the syntactic structure of a language. Even if the question is non-relevant or non-diverse but correct grammatically it should be given a high score.

Diversity: Judges the extent of diversity between two questions generated by a model. If the two questions touch on different aspects of the story they should be given a high diversity score. **This rating should be the same for both the clarifying questions.**

ClariScore: Metric to judge how capable the clarifying question is in resolving the ambiguity of the previously asked question. This metric judges if the clarifying question asked is helpful in resolving the ambiguity of the user asked question. A technique to score on this front is to compare the ambiguous question and then to ask yourself the clarifying question to judge if any ambiguity is able to be resolved.

Relevance: Judges if the clarifying question asked is relevant to the story, history and the ambiguous question asked. A technique to score on this front is to see if the clarifying question contains any information outside of the passage provided. If it does, rate it low.

Scoring

Each annotator has been given about 30 passages in a reference.txt file. For each passage, there are a total of **7 question pairs** generated by 7 models. Please match the ID in the annotator.csv file and the reference.txt file before scoring to avoid mismatch.

Examples

The following are two complete examples provided outside of the dataset.

Story
The story of the day I lost my best friend to a car accident. The day a precious life was taken from us way too soon. \n\nIt was a bright and Sunny day in November.

Thanksgiving had been celebrated only two days before. Since it was a holiday weekend I had been on the phone with Greg the night before many times. His dad didn't want him to come over because of the holiday. I guess he finally wore him down and he called and said, \"I can stay\". So, my mom, brother, and I went to pick him up. He was always smiling. The complete opposite of my shy self, Greg was always the life of the party. \n\nWe got two large pizzas that Friday night. I've never known anyone in my entire life who loved to eat more than Greg. That's the way he was though. He was just enjoying life. And if it meant gaining weight or whatever, so be it. He would sit back and put his hands on his belly and just laugh. We (Greg, David, and I) did so many funny things together and had such great times. Things we should have done and things we shouldn't have done, I'll \"Never\" forget. \n\nOn Saturday morning Dad took us out for breakfast. We all finished eating and followed my Dad up to the cashier. Greg asked Dad if he could have a candy bar. I looked at Greg shaking my head. He just laughed. After breakfast, Father took us to my Mom's house. \n\nWhen we got out at Mom's house there was no one home. So, one of us grabbed a big wheel and rode it down the steep driveway into the street. Just boys being boys. Greg and I did it several times until the last time. The car hit him on the head, knocking him around 75-- 100 yards. My brother and I both ran screaming just yelling for help and crying. One of the neighbors called 911. I was in shock. That day was forever etched into our memories. \n\nIt still hurts to think about it. Wishing we could have grown old together. Wondering how it would have been. I'm sure It WOULD HAVE BEEN GREAT.

History
User: Did anybody actually see the accident happen?
Bot: yes.
User: Who saw it?
Bot: My brother and I

User Ambiguous Question
What was everyone doing?

Ground Truth
Do you mean before the accident?

Clarification Questions
Who was Greg in shock when he saw the car crash?
How did Greg and Greg miss each other?

Evaluation
Clarifying Question 1 - [3, 5, 3, 4]

Clarifying Question 2 - [3, 5, 2, 2]

Grammatical Correctness - “Who was Greg in shock” is grammatically incorrect. The correct grammatical structure should have been “Who was in shock when he saw the car crash?”. Similarly the second clarifying question is grammatically incorrect as well. The correct question should have been “How did my brother and I miss Greg?”.

Diversity - Both the questions are diverse and touch on different aspects of the story.
ClariScore - The user wants to know what everyone was doing in the passage. The ground truth clarifying question points out the time which the user is referring to. Both clarifying questions do not try to resolve this ambiguity.

Relevance - The first clarifying question is relevant since the passage refers to people who were in shock when the accident happened. The second clarifying question, even though it captures the implicit meaning from the last part of the paragraph, is not relevant since the passage does not mention anyone “missing” anyone’s presence explicitly.