

Guidelines for Evaluating the Quality of response candidate in a given dialogue

Description of the annotation task	1
Sub-task 1: Evaluating Response Candidate	1
Sub-task 2: Providing an explanation for the selected choice	2
Examples	2
Examples of Appropriateness	2
Examples of Contextualization	3

Description of the annotation task

In this task, you are asked to evaluate the quality of possible response candidates, given a dialogue history. The dialogue is between two participants who are chit-chatting.

The conversations are about ordinary topics, such as relationships, work, and ordinary life events. The objective of this task is to evaluate the quality of the proposed response to continue the dialogue, given the dialogue history.

The task is composed of two sub-tasks:

Sub-task 1: Evaluating Response Candidate

In the first sub-task, you are asked to evaluate each proposed response candidate using 3 criteria: *Appropriateness*, *Contextualization* and *Correctness*.

- A)** In the first evaluation criterion, ***Appropriateness***, you are asked to evaluate if a response candidate is appropriate for a given dialogue history or not. That is if the response candidate is coherent and it makes sense in the current dialogue history. For this evaluation, you can choose from the following options:
- **Appropriate**, the response candidate makes sense and it can be the natural continuation considering the dialogue history.
 - **Not Appropriate**, the response candidate does not make sense in the current dialogue history.
 - **I don't know**, the response candidate has some parts that make sense with respect to the dialogue history, but some that do not.
- B)** For what concerns the second evaluation criterion, ***Contextualization***, you are asked to evaluate if the response candidate contains references (implicit or explicit) to the shown dialogue history. In this evaluation you can choose from the following options:

- **Contextualized**, the response candidate contains implicit or explicit references to the dialogue history (e.g. events, people, objects or emotions).
- **Not Contextualized**, the response candidate does not contain any reference to the dialogue history.
- **I don't know**, the response candidate contains some references to the dialogue history, but it contains other references that are not clear or relevant.

C) For what concerns the fourth and last criterion, **Correctness**, you are asked to evaluate either if the response candidate contains grammatical and/or structural errors or if the response candidate is correct. For the evaluation, this evaluation you can choose from the following options:

- **Correct**, the response candidate does not contain any type of grammatical or structural error, repetitions, misspelling or any other type of error.
- **Not Correct**, the response candidate contains some errors, for example, repetitions, misspellings, punctuation or any other type of error.
- **I don't know**, it is hard to identify if the response candidate contains errors or not.

ALL proposed TEXTS are in LOWERCASE characters, as a consequence, **the absence of uppercase is not to be considered an error**.

Sub-task 2: Providing an explanation for the selected choice

After this subtask, in some cases, you are asked to select an explanation that best describes the motivation behind your evaluation from a predefined list of items. If none of the options suits your scenario, please select “other”.

Examples

Examples of Appropriateness

Dialogue History		
<p>A: did you watch the news yesterday ?</p> <p>B: yes , i was quite shocked to see the destruction from the flood .</p> <p>A: yeah , it really did a number on me too . did you know anyone from there ?</p>		
Response candidate	Evaluation	Explanation
B: a friend of mine lived there . fortunately nothing bad happened to them .	Appropriate	The response candidate is a natural response because there is a clear logical thread between the

		candidate and the dialogue history.
B: of course i love ice cream , who doesn't ?	Not Appropriate	The response candidate does not make sense with respect to the dialogue history.

Dialogue History		
<p>A: sir , are you interested in buying some hot dogs ?</p> <p>B: yes please, how much for two ?</p> <p>A: they are 2 dollars each .</p>		
Response candidate	Evaluation	Explanation
B: you got yourself a deal . can i also get some mustard ?	Appropriate	The response candidate is a natural response because there is a clear logical thread.
B: what are you selling ?	Not Appropriate	The response candidate is not coherent with what has been said before in the dialogue history.

Examples of Contextualization

Dialogue History		
<p>A: did you watch the news yesterday ?</p> <p>B: yes , i was quite shocked to see the destruction from the flood .</p> <p>A: yeah , it really did a number on me too . did you know anyone from there ?</p>		
Response candidate	Evaluation	Explanation
B: a friend of mine lived there . fortunately nothing bad happened to them .	Contextualized	In the response candidate there is a reference to the dialogue history.
B: no, but have you also seen the latest football	Contextualized	In the response candidate



match ?		there is a reference to the dialogue history.
B: yes .	Not Contextualized	The response candidate is generic and does not mention any information from the dialogue history.

Dialogue History		
A: sir , are you interested in buying some hot dogs ?		
B: yes please, how much for two ?		
A: they are 2 dollars each .		
Response candidate	Evaluation	Explanation
B: here 's 4 dollars . can i also get some mustard ?	Contextualized	In the response candidate there is a reference to the dialogue history.
B: then give me three basketballs .	Not Contextualized	The response candidate contains some elements (i.e. basketballs) not mentioned in the dialogue history.