UNIVERSITÀ DI TRENTO | Department of Information Engineering and Computer Science

SiS LAB
Signals & Interactive Systems

Guidelines for Evaluating the Quality of response candidate for question answering

Description of the annotation task

In this task, you are asked to evaluate the quality of several response candidates, given a **context** and a **question**.

The questions are based on the provided context. The objective of this task is to evaluate the quality of the proposed response as a possible answer to the given question based on the provided context.

The task is composed of two sub-tasks:

Sub-task 1: Evaluating the Response Candidates

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

A) For what concerns the first evaluation criterion, ***Contextualization**,* you are asked to evaluate if the response candidate contains (implicit or explicit) <u>references to the context</u>. In this evaluation you can choose from the following options:
   - **Contextualized**, the response candidate contains implicit or explicit references to the context (e.g. events, people, objects, or emotions).
   - **Not Contextualized**, the response candidate does not contain any reference to the context.
   - **I don't know**, the response candidate contains some references to the context, but it contains other references that are not clear or relevant.

B) For what concerns the second evaluation criterion, ***Validity***, you are asked to evaluate if the response candidate includes the right information to adequately answer the proposed question based on the context. In this evaluation you can choose from the following options:

- **Valid**, the response candidate includes the right information from the context to adequately answer the proposed question.
- **Not Valid**, the response candidate does not include the right information from the context to adequately answer the proposed question.
- **I don't know**, the response candidate includes some information that is adequate but some that is not.

C) For what concerns the third 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 (grammatically and structurally) 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 Contextualization

Context + Question
Context: Rome then became the focus of hopes of Italian reunification after the rest of Italy was united as the Kingdom of Italy in 1861 with the temporary capital in Florence. That year Rome was declared the capital of Italy even though it was still under the Pope's control. Today, Rome is still the capital of Italy. **Question:** Which is the capital of Italy?

Response candidate	Evaluation	Explanation
Answer: "rome"	Contextualized	The response candidate contains

		elements mentioned in the context.
Answer: "florence"	Contextualized	Even if the answer is wrong, the response candidate contains elements mentioned in the context.
Answer: "london"	Not Contextualized	The response candidate does not contain elements mentioned in the context.

Context + Question
Context: In 2025, the following eclipses are expected in Denmark: a partial lunar eclipse on the 14th of March, a partial solar eclipse on the 29th of March, and a total lunar eclipse on the 7th of September. **Question:** How many eclipses will happen in Denmark in 2025?

Response candidate	Evaluation	Explanation
Answer: "three: one on the 14th of March, one on the 29th of March, and one on the 7th of September"	Contextualized	The response candidate contains elements mentioned in the context.
Answer: "three"	Contextualized	The response candidate implicitly refers to some elements mentioned in the context.
Answer: "i don't know"	Not Contextualized	The response candidate does not contain elements mentioned in the context.

Examples of Validity

Context + Question
Context: Rome then became the focus of hopes of Italian reunification after the rest of Italy was united as the Kingdom of Italy in 1861 with the temporary capital in Florence. That year Rome was declared the capital of Italy even though it was still under the Pope's control. Today, Rome is still the capital of Italy. **Question:** Which is the capital of Italy?

Response candidate	Evaluation	Explanation
Answer: "rome"	Valid	The response candidate answers the proposed question with the correct information.
Answer: "london"	Not Valid	The response candidate does not answer the proposed question with the correct information.
Answer: "florence"	Not Valid	The response candidate does not answer the proposed question with the correct information.

Context + Question
Context: Rome then became the focus of hopes of Italian reunification after the rest of Italy was united as the Kingdom of Italy in 1861 with the temporary capital in Florence. That year Rome was declared the capital of Italy even though it was still under the Pope's control. Today, Rome is still the capital of Italy. **Question:** What year was rome declared the capital of Italy?

Response candidate	Evaluation	Explanation
Answer: "1861"	Valid	The response candidate answers the proposed question with the correct information.
Answer: "i don't know"	Not Valid	The response candidate does not answer the proposed question with the correct information.
Answer: "2012"	Not Valid	The response candidate does not answer the proposed question with the correct information.