

Annotation Guideline

Original draft by Dana Ruiters

Modified by David Adelani

Task Description

Together with this annotation protocol, you have received one excel Sheets. Your task is to annotate the data contained within the sheet based on the annotation rules presented in this annotation protocol. The annotations consist of **2 tasks** which are distributed across the three files. Each file contains around **100 data points**. Please commit to a single Excel Sheet at once.

Data

In each Excel Sheet, you will be confronted with **2 data columns**: **reference** and **hypothesis** sentences. The reference is human translated while the hypothesis is the model translation. This is the data you will be asked to annotate. Next to the data columns, you will find **2 task columns**. Content preservation and Grammaticality.

Tasks

In the following, you will find the instructions that are related to each of the task columns. Be precise in your assignments and try to avoid overusing any one of the category labels. We have included tests within the data to check whether annotations have been performed in a consistent manner.

The three tasks are defined as such:

Content Preservation

- The Excel Sheet will contain a content preservation task. In column 1 and 2 ("reference", "hypothesis") you will see a sentence each. Your job is to evaluate how much of the meaning of the reference was preserved in the model output. How many named entities and contents are preserved? Words replaced by their synonyms are fine.
- To do this task successfully, picture what is being described and contrast exactly what is conveyed by one statement versus what is being conveyed by the other. Do the statements refer to the exact same person, action, event, idea, or thing? Is the MEANING of the sentence kept in the translation?
- **Ignore grammatical errors** and awkward wordings within the statements as long as they do not obscure what a statement is supposed to convey.
 - To rate the adequacy of a sentence pair, please think of the following statement:
"The sentences in content A and B are identical in their underlying meaning."
And rate your agreement on a scale of...

A five-point Likert scale, where 1 is “Not at all”, 2 is “Barely”, 3 is “Partly”, 4 is “Mostly” and 5 is “Fully”.

- **1:** Not at all: The two sentences are on different topics.
- **2:** Barely: The two sentences do not carry the meaning of the reference, but are on the same topic.
- **3:** Partly: only a portion of the sentence carry some meaning of the reference
- **4:** Mostly: The meaning of the two sentences are mostly equivalent, but some details differ.
- **5:** Fully: The two sentences are completely equivalent, as they mean the same thing.
- The rating (1,2,3,4 or 5) should be written into **column 3** (“Content Preservation”) into the same row as the sentence pair being rated.
- Some examples:
 - A: *The Governor has been impeached*
B: *Don't worry*
→ 1: Completely carry different meanings but on the same topic
 - A: *Trump's a liar*
B: *Obama speaks*
→ 2: Different meaning, but the general topic (i.e. *politicians doing/being something*) is the same.
 - A: *The Mayor of London was her on Thursday*
B: *The Mayor of Canada visited Aso Rock*
→ 3: These are not equivalent, but share some details (i.e. *A governor is visiting or was at a place*).
 - A: *Nigeria confirms 86 new cases with COVID-19*
B: *On the 20th of April, Nigeria confirms 86 new cases with coronavirus*
→ 4: The two sentences carry the same meaning, but B is missing information (*On the 20th of April was missing in the first sentence*). The two could be more similar. I included this information, e.g. by adding “On the 20th of April”.
 - A: *Nigeria confirms 86 new cases with COVID-19*
B: *Nigeria registers 86 new cases with coronavirus*
→ 5: Identical in meaning (despite the replacement of “confirm” with “register”).

Fluency

- The Excel sheet will ask you to annotate the fluency of a sentence.
- In **column 2** (“**prediction**”) you will see a sentence. Your task is to rate how **fluent** the sentence is. Ignore the sentence in column 1 (“original”).
- To rate the fluency of a sentence, please think of the following statement:
“*This is a fluent sentence written by a native speaker.*”
And rate your agreement on a scale of...

- **1:** Strongly disagree: Incomprehensible
- **2:** Disagree: Disfluent English
- **3:** Neither agree nor disagree: Non-Native English
- **4:** Agree: Good English
- **5:** Strongly agree: Flawless English
- The rating should be written into column 4 (“Fluency”) into the same row as the sentence being rated.
- Please note that we are not asking for **grammaticality** per se. Please focus on whether a sentence is fluent based on the **typical language use** of a **native speaker**.
- Some examples:
 - *He the around you: go.*
→ 1: This is incomprehensible and not a typical (fluent) language use, even in social media.
 - *it found, never anyone.*
→ 2: This is not a typical (fluent) language use.
 - *I was very comfort the night after*
→ 3: This could be a typical, fluent comment in social media, but contains an untypical grammatical mistake that a native speaker would not make (*comfort* instead of *comfortable*).
 - *I love the food ... however fries here are delicious.*
→ 4: This is a very fluent sentence that could have been written by a native speaker. But the usage of *however* is slightly out of place here (using *and* instead might have been more appropriate).
 - *The Mayo of London visited Texas*
→ 5: This is a very fluent sentence. Despite being ungrammatical (“Mayo” instead of “Mayor”), this can be assumed to be written by a native speaker.