

<Triplet set>-<Text> Pair Validation:

Guidelines & Tips

Greetings!

Thank you for accepting my offer.

The data I am dealing with consists of Triplet-Text pairs, where a set of triplets is verbalized into a text. Your task will be to validate whether the <Triplet set>-<Text> pairs correctly encompass the mutual context for a total of 30 examples per file.

This work requires a clear understanding of the triplet's structure, making it somewhat challenging. Therefore, I would like to provide you with a few guidelines and examples, along with detailed explanations.

First and foremost, it is crucial to understand the structure of a triplet.

[Triplet Composition]

Format: (<S> subject | <P> predicate | <O> object)

A triplet is composed of three elements: subject, predicate, and object. Each of these elements is delineated by specific symbols with capital letters such as <S>, <P>, and <O>. These symbols help to clearly identify the role of each element within the triplet.

- <S> stands for the subject, which represents the entity that is performing an action or being described.
- <P> stands for the predicate, which illustrates the relationship or action that connects the subject to the object.
- <O> stands for the object, which is the entity that is receiving the action or being described in relation to the subject.

For example, “(<S> Alice | <P> Occupation | <O> Writer)” can be paired with “Alice is a writer.” Or “Alice was a former writer.” If there are multiple triplets such as “(<S> Alice | <P> Occupation | <O> Writer), (<S> Alice | <P> Nationality | <O> U.S)”, then “Alice is a writer from United States of America.” Or “Alice who was a former writer is from U.S.” can be proper match.

[Triplet-Text Pair Validation]

Now, let's explain how to validate Triplet-Text pairs. As mentioned earlier, a triplet includes a subject, predicate, and object. The AI's task involves generating sentences from multiple triplets, making it essential to understand the meaning of each triplet when validating the Triplet-Text pairs.

We will now explain how to determine whether a Triplet-Text pair is correctly or incorrectly aligned, along with some examples.

[Examples]

When you open up a file named as “**number.docx**”, you can see 30 tables per each file. One example is followed.

<Triplet Set>
1. (<S> Glyphostoma epicasta <P> Family <O> Clathurellidae) 2. (<S> Glyphostoma epicasta <P> Class <O> Marine gastropod mollusk) 3. (<S> Glyphostoma epicasta <P> Type <O> Sea snail)
<Reference Text>
Glyphostoma epicasta is a species of sea snail, a marine gastropod mollusk in the family Clathurellidae.
<Error>
[Unused triplets]: x [Unguessable text]: x

Each table contains three distinct fields, described as follows:

<Triplet Set>

This field includes multiple triplets, ranging from 1 to more than 10 triplets. Each triplet is a structured representation of a subject, predicate, and object. Single or multiple triplets together describe a single sentence

<Reference Text>

This field contains the text that corresponds to the <Triplet Set>. The text is generated by understanding the structure and meaning of the triplets in the <Triplet Set>. It should clearly and accurately contain “**Only and Every**” information from <Triplet Set>.

<Error>

This field is used to document instances where the relationship between the Triplet Set and the Reference Text is not inferable or clearly understood. If there are discrepancies or ambiguities between the triplets and the generated text, such issues are recorded here.

The methods for error annotation will be followed.

1. Triplet-Text Pair Miss-Aligned

In case of error, there are two types of error.

1.1. Unused Triplet

<Triplet Set>
1. (<S> Muijala <P> Is part of <O> Lohja, finland) 2. (<S> Muijala <P> Country <O> Finland) 3. (<S> Lohja, finland <P> Country <O> Finland)
<Reference Text>
Muijala is a district in Lohja, Finland.
<Error>
[Unused triplets]: 3 [Unguessable text]: x

In this case, the reference text “Muijala is a district in Lohja, Finland.” can be verbalized with only 1,2 triplets. Reference text contains “Only” information from <Triplet set>, but not “Every” information from <Triplet set>.

In this case, please write all IDs of unused triplets next to “[Unused triplets]:” in the <Error> filed.

* If there are duplicate triplets, one of them is considered an unused triplet. For example, in case of [1.(<S> Alice | <P> date of birth | <O> 1997) 2.(<S> Alice | <P> date of birth | <O> 12DEC97)] and “Alice was born in December 12th in 1997” are paired, the text can be guessed with only triplet 2. Therefore 1 can be determined as unused triplet.

1.2. Unguessable Text

<Triplet Set>
1. (<S> Flora Drummond <P> sex or gender <O> female)
<Reference Text>
Nicknamed 'The General' for Flora Drummond's habit of leading Women's Rights marches wearing a military style uniform 'with an officers cap and epaulettes' and riding on a large horse, Drummond was an organiser for the Women's Social and Political Union (WSPU) and was imprisoned nine times for her activism in the Women's Suffrage movement.
<Error>
[Unused triplets]: x [Unguessable text]: Nicknamed 'The General' for, habit of leading, Rights marches wearing a military style uniform 'with an officers cap and epaulettes' and riding on a large horse, Drummond was an organiser for the Women's Social and Political Union (WSPU) and was imprisoned nine times for her activism in the, Suffrage movement.

In this case, “Flora Drummond” is mentioned in the **<Reference Text>**, but many other words are unguessable with the information provided from the **<Triplet set>** alone. Reference text contains “Every” information from **<Triplet set>**, but not “Only” information from **<Triplet set>**.

In this case, please copy and paste words from **<Reference Text>** excluding the words that is from **<Triplet Set>** next to the “[Unguessable text]:” in the **<Error>** filed.

Blue-Colored Words are examples that is excluded in the error because they are from **<Triplet Set>**. To enhance clarity, shading has been used to distinguish the splits containing error annotations.

* Gray and Yellow were used to distinguish splits, without any additional meaning beyond simple differentiation.

* In practice, you do not need to distinguish splits by shading. Copy and Paste is enough 😊

* If Some subjects or objects are absent in **<Reference Text>**, that could be considered as [Unguessable text].

For Example, (<S> Alice | <P> Occupation | <O> Writer) is paired with “She is the writer”, then “She” became unguessable part in this case.

*** If the information in the <Reference Text> is based on facts or common knowledge but cannot be inferred from the <Triplet Set>, it is considered [Unguessable text].**

2. Triplet-Text Pair Well-Aligned

<Triplet Set>
1. (<S> Michel alberganti <P> Nationality <O> French) 2. (<S> Michel alberganti <P> Occupation <O> Radio producer) 3. (<S> Michel alberganti <P> Occupation <O> Writer) 4. (<S> Michel alberganti <P> Occupation <O> Journalist)
<Reference Text>
Michel Alberganti was a French radio producer, writer, and journalist.
<Error>
[Unused triplets]: x [Unguessable text]: x

In this case, every triplet is properly used to predict reference text. Also reference text include 'Every and Only' information from given **<Triplet Set>**. You can just write x mark on **[Unused triplets]** and **[Unguessable text]** in the **<Error>** filed.