Guidelines for error annotation of semantic error types in data-to-text systems

Table of Contents

1 Introduction

The participants will annotate semantic error types (addition, omission and repetition) on 30 pairs of triple(s) and verbalisations.

Triple(s) is/are the input to the data-to-text generation system and *verbalisation* is the system output from the data-to-text generation system.

In this context, *triple(s)* consist(s) of the elements 'Subject | Predicate | Object,' and *verbalisation* is the natural language text generated from the input triple(s).

For example, 'Joe_Biden | president | United_States' is a *triple* with the following elements:

- Subject: Joe_Biden
- Predicate: president
- Object: United_States

One possible *verbalisation* is 'Joe Biden is the president of the United States.' Note that the verbalisation can be seen as being composed of components that correspond to the subject, predicate and object of the triple, i.e., Joe Biden as a subject, the president as a predicate and the United States as an object.

Another example, 'Morelos | leader | Adrián_Rivera_Pérez' from Figure 1 is a *triple* with the following elements:

- Subject: Morelos
- Predicate: leader
- Object: Adrián_Rivera_Pérez

‘The leader of Morelos is Adrián Rivera Pérez.’ from Figure 1 is the *verbalisation.* It is composed of components that correspond to the subject, predicate and object of the triple, i.e., ‘Morelos’ as a subject, ‘leader’ as a predicate and ‘Adrián Rivera Pérez’ as an object.

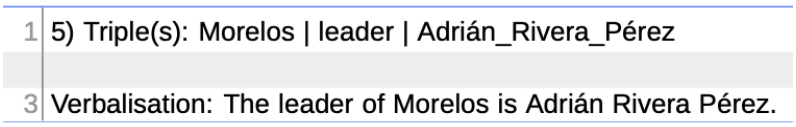

Figure 1. Example of triple and verbalisation as given on the brat annotation tool.

Your task is to label parts of triple(s)/verbalisation pairs with semantic error types, if there are any, as per the annotation steps provided in **Section 2.B**. Please note that there can be zero or more semantic error types in any triple(s)/verbalisation pair. Definitions are provided for all error types in **Section 2.A**.

Instructions on how to use the brat annotation tool is provided in **Section 3**.

2 Instructions for error annotation:

A. Definitions

Three types of errors need to be identified and labelled: omissions, additions and repetitions. These are defined as follows. In this context, ‘input’ is the set of triples and ‘output’ is the verbalisation (text).

Definition of OMISSION:
Some content that is present in the input and should be rendered in the output is not present in the output. Moreover, there are no word span(s) in the output that are intended to render it, but do so wrongly. *i.e. this type of error can be fixed by adding something to the output.*

Definition of ADDITION: The output text contains word span(s) for which there is no corresponding part of the input that they render. In other words, some content that is **not** present in the input and **should not** be rendered in the output is nevertheless rendered by some word span(s) in the output. Moreover, there is no content in the input that the word span(s) are intended to render, but render wrongly. *i.e. this type of error can be fixed by removing something from the output.*

Definition of REPETITION: Some content is repeated verbatim in the output, but there is no corresponding repetition in the input.

B. Annotation Steps

In the first step, you should examine whether each element in the input triples is verbalised or not.

If an element is not expressed in the verbalisation, mark the element as an omission error type in the triple.

If the whole triple is not expressed in the verbalisation, mark each element as an omission error type in the triple.
For example, if the triple 'ENAIRE | city | Madrid' is not expressed in the verbalisation, then mark 'ENAIRE' as an omission, 'city' as an omission and 'Madrid' as an omission.

In the brat annotation tool, select the entity type as omission.

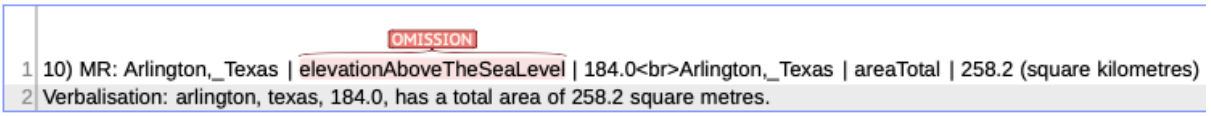

Figure 2. Example of omission error annotation in the brat annotation tool interface.

If each element in the input triples is verbalised which means there is no omission error, then proceed to the **second step**.

In the second step, examine whether all the content words and phrases in the verbalisation render a corresponding element(s) in the triples.

If a content word phrase does not render a corresponding element in the input triples, mark it as an **addition** error type.

In the brat annotation tool, select the entity type as addition:

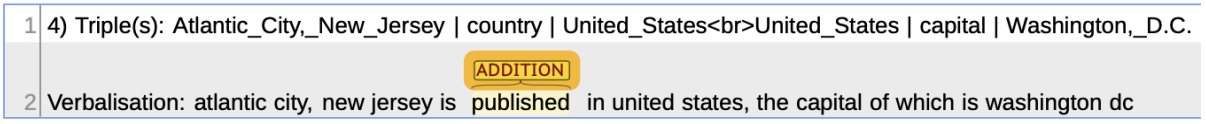

Figure 3. Example of addition error annotation in the brat annotation tool interface.

If all the content phrases in the verbalisation render a corresponding element in the input triples this means there is no addition error, so proceed to the **third step**.

In the third step, check if any part of the output is repeated, including close paraphrases. This is the case e.g. if an element in the triples is rendered more than once. If there is a content phrase that is repeated in this sense, mark it as a **repetition** error type.

In the brat annotation tool, select the entity type as repetition.

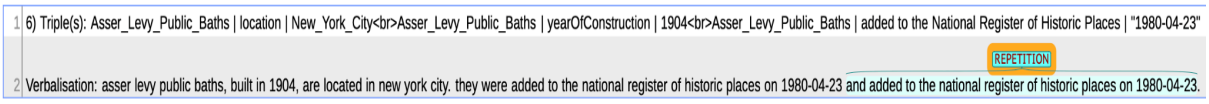

Figure 4. Example of repetition error annotation in the brat annotation tool interface.

If all the content phrases in verbalisation include all the elements in the triples without an extra in the verbalisation that has no relation in the input triples, which means there is no repetition error, then proceed to the next pair of triple(s) and verbalisation.

Please note that:

- If there is more than one triple in the input, triples are enclosed within single quotes (‘ ’) and separated by commas. **For example,** ‘Joe_Biden | president | United_States’**,** ‘Joe_Biden | birthPlace | Pennsylvania’.
- Please be careful while selecting the word span when marking an error. You should select complete tokens, i.e., words in the text, that are delimited by whitespace.

  For example, the selection for ‘president’ in ‘Joe Biden is the president of the United States.’ is correct, but selecting just ‘pres’ is not correct, as in ‘Joe Biden is the president of the United States.’ Similarly, ‘Joe_Biden | president | United_States’ is correct, but ‘Joe_Biden | president | United_States’ is not.
- You should consider the inferred verbs and tenses correct in verbalisations as long as they are implied by the information in the input triple(s).

  For example, consider the input triple “Alessio_Romagnoli | youthclub | A.S._Roma” and the corresponding verbalisation “alessio romagnoli plays for the a . s . roma youth team.” Here ‘plays for’ can be inferred from the presence of ‘youthclub’ in the input triple. This is considered valid/correct and should not be marked as an error.
- However, cases such as, ‘youthclub’ being verbalised as ‘youthteam’ (‘youthclub’ is not rendered in the output and ‘youthteam’ is added in the output) or ‘AC_Hotel_Bella_Sky_Copenhagen’ verbalised as ‘hotel bella sky copenhagen’ (‘AC_Hotel_Bella_Sky_Copenhagen’ should be marked as an omission and ‘hotel bella sky copenhagen’ as addition) should be marked as errors.
- You should take extra care with units, dates and other numerical values and their conversions. For example, if ‘1234 m’ is verbalised as ‘1.234 km’ then it should not be considered an error. If ‘2006-12-31’ is verbalised as ‘31st July 2016’ then it should be marked as an omission (‘2006-12-31’ is not rendered in the output), and addition (‘31st July 2016’ is added in the output). If ‘610.0’ is verbalised as ‘610 metres’ then it should be considered an error where ‘metres’ will be an addition error.

3 Instructions for using the brat annotation tool:

1. The annotators will be provided with the link to the brat annotation tool in the email sent to them. The link will look like this: https://39f5-80-233-42-86.ngrok-free.app/index.xhtml#/main/.[FIRSTNAME-LASTNAME]

2. Please copy-and-paste the link into your browser (preferably, Chrome or Safari). This will either take the annotator to a page as seen in (a) Figure 5 or (b) Figure 6.
    a. If it takes the annotator to a page that looks like Figure 5, please click on "**Visit Site**".
    b. Or if it takes the annotator to a page that looks like Figure 6, please click "**OK**".

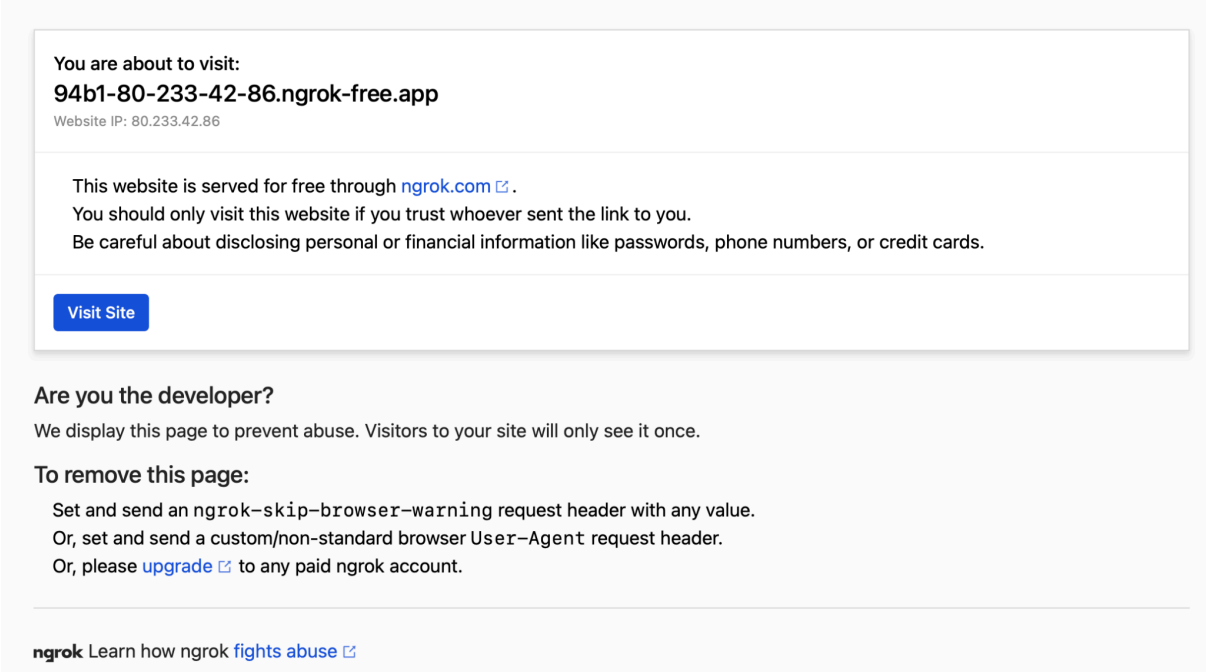

Figure 5. Possible first page (a) that comes up after the link is opened on the browser.

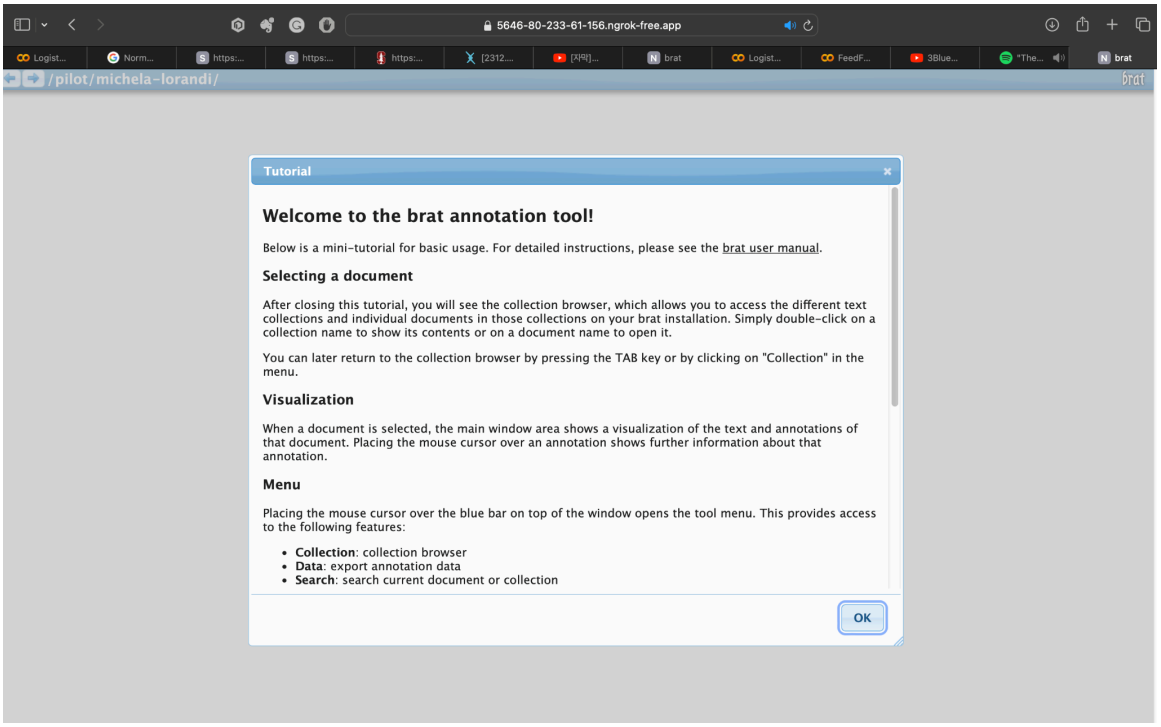

Figure 6. Possible first page (b) that comes up after the link is opened on the browser.

3. This will take the annotator to the folder that contains the files that have to be annotated as shown below. For this experiment, there will be 30 individual files in total. Figure 7 shows an example of how the brat annotation tool interface will look after Step 2. Select a file (for example, "**1**") and click "**OK**". Each file contains one input triple(s) and the verbalisation pair. Please see Figure 8 for an example. The annotator can navigate the files by using the left & right arrow on the top-right (as seen in Figure 8) OR use the left & right buttons on your keyboard.

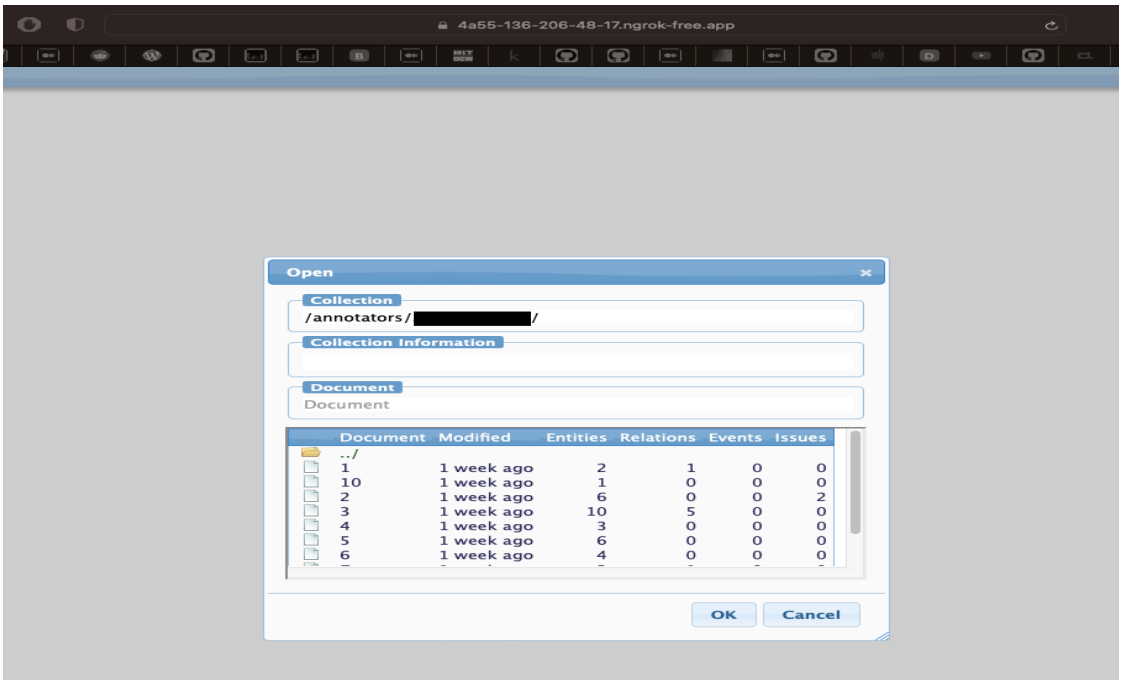

Figure 7. The first instance of brat annotation tool interface that the annotator will see.

Figure 8. Example of input triples and the verbalisation found in each file when opened.

4. The next step for the annotator is to log in to the brat interface so that the annotations will be recorded. This step is *crucial* for the annotator.
    a. Click on the **brat logo** on the top-right and click "**Login**". See Figure 9.
    b. This will take the annotator to the login interface. See Figure 10.
    c. The login 'username' and 'password' will be shared with the annotator along with the link to the brat interface in the email.
    d. Once the annotator logs in and begins their annotation, the annotations are saved automatically.

Figure 9. The login button's visibility after clicking on the brat logo on the top-right.

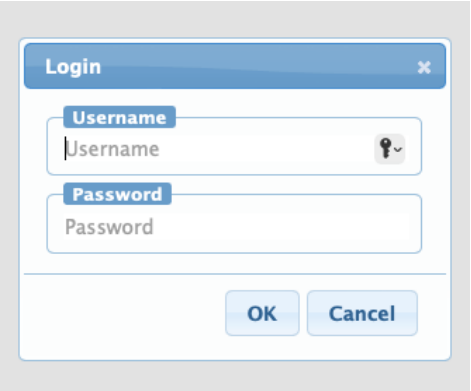

Figure 10. Login interface in the brat annotation tool.

5. To annotate the errors, the annotator should select the word span to be marked as an error. This will give you the pop-up window that contains the list of our error types under the 'entity type' label in the interface. See Figure 11.
   The annotator can select the appropriate error types according to the instructions for error annotation and the definitions of error types provided in **Section 2**. The annotator has to select the text span which corresponds to the error and mark it.
   If the annotator believes that there is no error, please go to the next file by either navigating by using the left & right arrow on the top-right OR using the left & right buttons on your keyboard.

6. To add to the previous point, if the annotator made a mistake in the selection of word span and/or in selecting a different error type and wish to correct their mistake, the annotator should select the word span and then select ‘delete’ from the pop-up window. See Figure 11.

7. After the annotator finishes the task, please logout (go to top-right → click on the brat logo (as you did in order to login) → you will see the “**Logout**” button, click on it) and send an email to rudali.huidrom(at)adaptcentre(dot)ie informing about completion of your task.

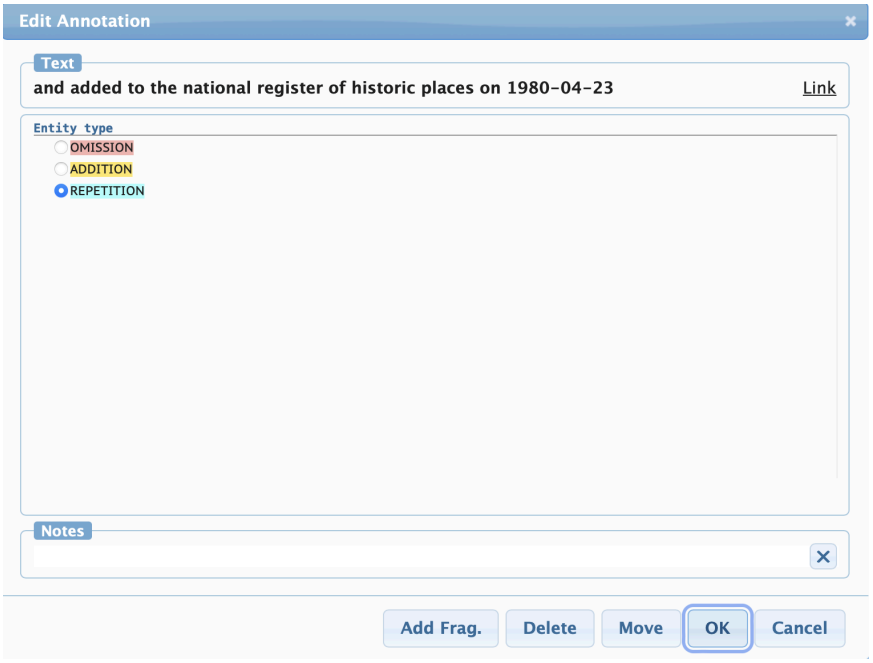

Figure 11. The pop-up window contains the list of our error types under the ‘entity type’ label in the brat interface.

For any queries, please don’t hesitate to contact:

Rudali HUIDROM
rudali.huidrom2@mail.dcu.ie, rudali.huidrom@adaptcentre.ie .