Manual Annotation Projection

Given a sentence and a target in a source language, the task is to manually select the same target in a translated target sentence.

APP TUTORIAL:

Requirements:

The annotation projection app uses tkinter and tqdm, install them with the following commands:

```
sudo apt-get install python-tk
pip3 install tqdm
```

Run the app

```
python3 app.py \
--source_dataset examples/en-test.tsv \
--target_sentences examples/en2es-test.txt \
--output_dataset examples/output_sample.txt
```

source_dataset: Path to the source dataset in TSV format and IOB2 annotation format.
target_sentences: Path to the translated sentences, one per line
output_dataset: Path where the manually annotated target sentences will be stored
font_size (optional): If you want to increase or reduce the font size of the app you can set a new font size with the arg "--font_size 12" the default font size is 12

App interface

The app consists of two text boxes.
The grey one contains the source sentence. The white one contains the translated target sentence.
The target in the source sentence that we must project will appear between brackets: [Place] and highlighted in orange
pyt
The text box of the target sentence is editable: **DO NOT WRITE ANYTHING IN IT OR MODIFY THE TARGET SENTENCE**
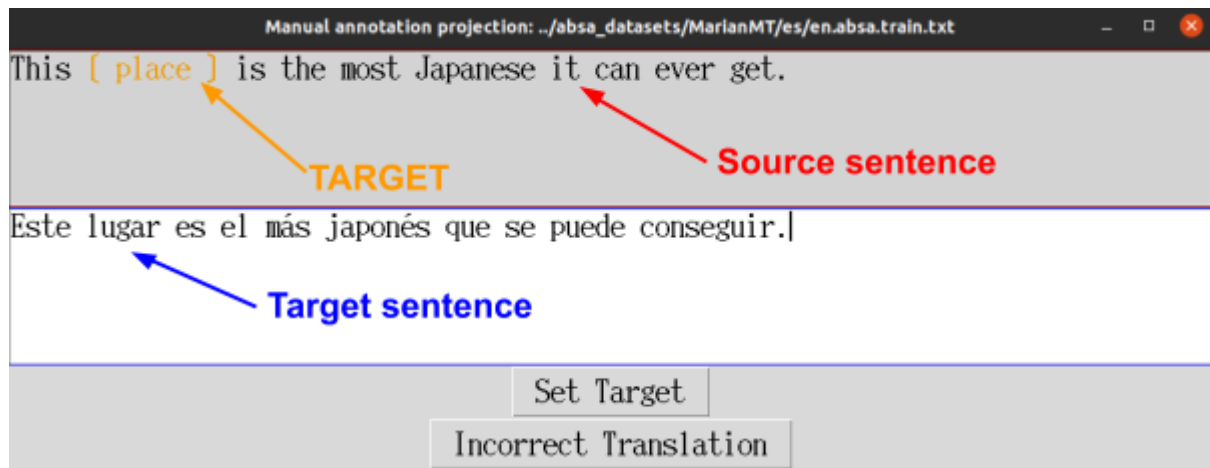
To project the target in the source sentence to the target sentence we select the text that contains the target in the target sentence with the mouse. **You can select multiple words.**

There is no need to select the whole words or avoid selecting spaces. The app will correctly handle the selection as long as you select at least one character of the word. The following selections are all equal "té verde creeme" will be stored as the projected target. However, to avoid possible miss projections, try to correctly select the target.
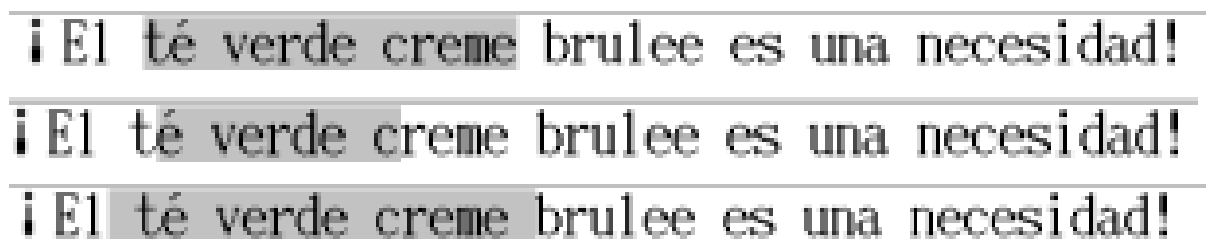
After you have selected the target push the "Set Target" or the "Incorrect Translation" button (See guidelines to know when you should push each one).

If the target is missing in the target sentence, don't select anything in the target sentence, just push the "Set target" button. A confirmation box will appear, push "YES" to skip the target.
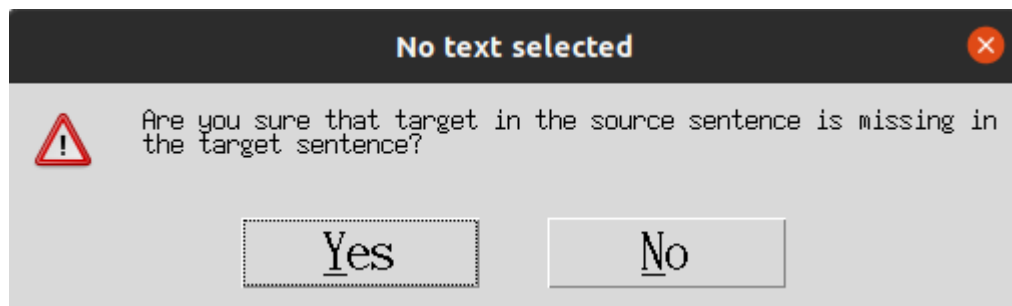
After projecting the target the next target in the source sentence will be highlighted or the next source and target sentences will appear if there are no more target in the source sentence.

IMPORTANT: Targets in the target sentence CANNOT overlap, that is, a word that cannot be part of two different targets. If you select two times the same word an error will appear in the terminal where you launched the app and the app will get stuck. If this happens close the app and launch it again to reattempt projecting the same example.

```
  0%|                                      | 0/654 [00:00<?, ?it/s]Exception in Tkinter callback
Traceback (most recent call last):
  File "/home/iker/miniconda3/lib/python3.8/tkinter/__init__.py", line 1883, in __call__
    return self.func(*args)
  File "app.py", line 85, in OnButton
    source_sentence_txt, target_sentence_txt = self.sentence_reader.step(
  File "/home/iker/Documents/Manual-Annotation-Projection/sentence_reader.py", line 318, in step
    self.get_next_tag()
  File "/home/iker/Documents/Manual-Annotation-Projection/sentence_reader.py", line 173, in get_next_tag
    self.write_sentence()
  File "/home/iker/Documents/Manual-Annotation-Projection/sentence_reader.py", line 210, in write_sentence
    raise ValueError(
ValueError: Collision between tags, entity overlap not allowed. A word cannot belong to more than one entity. target_sentence_tags: [[14], [14]]
```
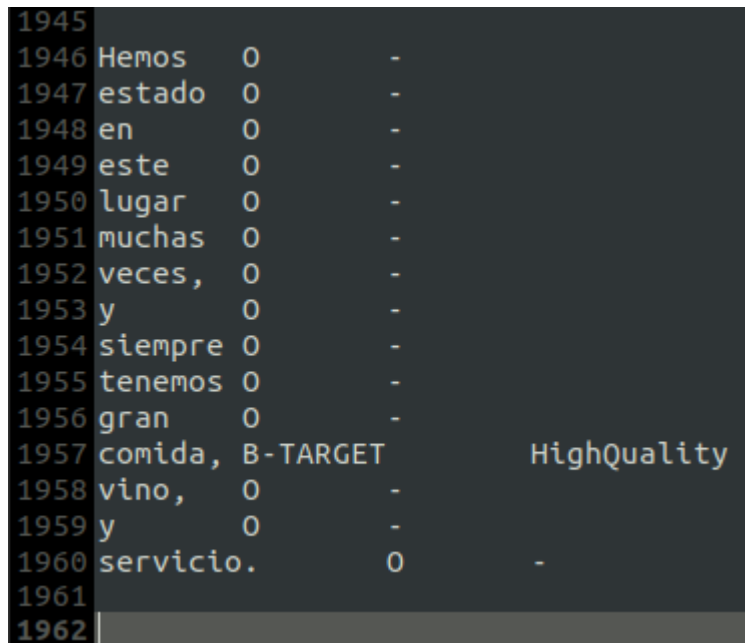
Save/Restore progress

After you have projected all the target of a source sentence the target sentence and the manually annotated targets will be immediately stored in the output file. You can close the app without losing any progress.

When you relaunch the app, use the same --output_dataset, the app will read the sentences from the output_path that you have already projected and it will resume from the last projected sentence.

For example, if you project 5 sentences, you close the app, and you launch it again with the same output_path you will continue the projection from the 6º sentence.

If you think you made a mistake you can go to the --output_dataset file to manually edit it. You can also remove the last n sentences to project them again (I don't recommend removing anything to avoid any formatting error, just edit the tags). Do not remove a sentence from the middle of the file or the app will restore the progress from that sentence ignoring the next ones in the file. You can only remove the last ones. To avoid formatting errors if you remove a sentence remember that after the last sentence in the file you need to add two new lines.

```
1945
1946 Hemos     O          -
1947 estado    O          -
1948 en        O          -
1949 este      O          -
1950 lugar     O          -
1951 muchas    O          -
1952 veces,    O          -
1953 y         O          -
1954 siempre   O          -
1955 tenemos   O          -
1956 gran      O          -
1957 comida,   B-TARGET        HighQuality
1958 vino,     O          -
1959 y         O          -
1960 servicio.        O       -
1961
1962 |
```

If you are going the manually edit the output file please close the app first.

See your progress

In the terminal where you have launched the app, you can see a progress bar

Estimated time that it will take you project the remaining sentences.

`2%|▭▭▭        | 14/654 [00:22<17:37,  1.65s/it]`

Percentage of sentences projected

Number of sentences you have projected.

Total number of sentences that you need to project

Total time you have expenend projecting sentences
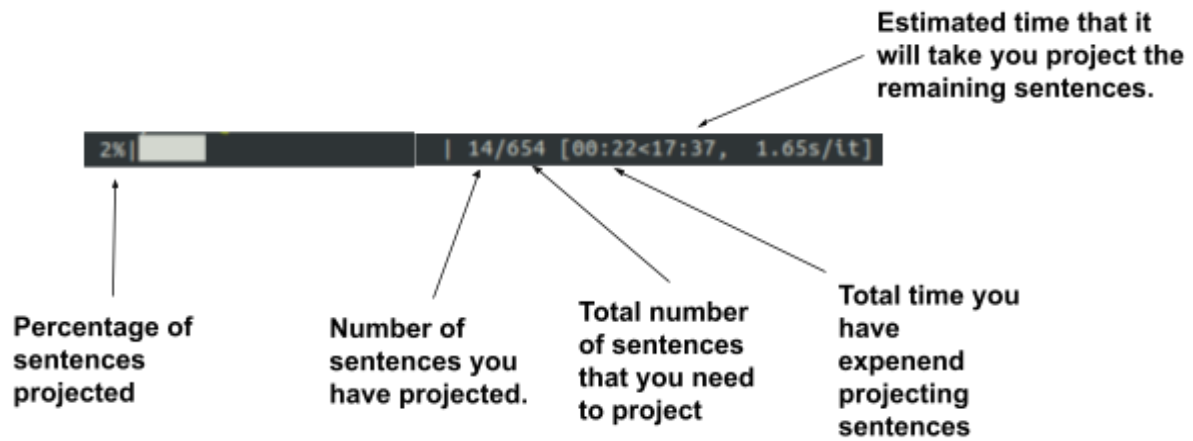
Close the app

Push the "X" button on the top of the app

ANNOTATION GUIDELINES

Simple projection

The target has been correctly translated. **For these projections, we use the**

Set Target

button.

| Judging from previous posts this used to be a good [place], but not any longer. |
| A jungar por publicaciones anteriores, esto solía ser un buen **lugar**, pero ya no. |

| The [seafood] is amazing, there's a good [wine list] , and the ever-changing menu always offers some great surprises. |
| El **marisco** es increíble, hay una buena **carta de vinos**, y el menú siempre cambiante siempre ofrece grandes sorpresas. |

| Every time in New York I make it a point to visit [Restaurant Saul] on Smith Street |
| Cada vez que estoy en Nueva York me gusta visitar el **restaurante Saul** en la calle Smith. |

| In the evening, this place attracted a well dressed, with it, NY [crowd] . |
| Por la noche, este lugar atrajo a un **público** de Nueva York bien vestido, con él. |

Considerations:

- We don't select as target prepositions in the target sentence if they are not selected as a target in the source sentence. For example, if the target is [food] we select as target [comida] we DO NOT select as target [la comida].
- We don't select punctuation in the target sentence if it doesn't belong to the target word even if the punctuation is part of the target in the source sentence. For example, if the target in the source sentence is [place!] but in the target sentence

the word and the exclamation mark are written separately ("lugar !") we only select as target "[lugar] ! "we do not select the exclamation mark "[lugar !]"

Contrary, if in the source sentence the punctuation is not part of the target but in the target sentence they appear as a unique word we select it ignoring the punctuation. For example, if the target in the source sentence is "[place] !" but in the target, the sentence is written as "lugar!" we select the target as usual "[lugar!]"

Mistranslations

Sometimes the target is mistranslated but it still appears in the target sentence. **In the case, we use the** Incorrect Translation **button and we follow the same guidelines as for "Simple projections"**

Ive asked a [cart attendant] for a lotus leaf wrapper rice and she replied back rice and just walked away.

Le he pedido a un **cartero** una hoja de loto envuelta en arroz y ella respondió arroz y simplemente se fue.
[1] A better translation of the target would be "Encargado de carritos / dependiente", "cartero" doesn't have the same meaning as "cart attendant".

[Chow fun] was dry; [pork shu mai] was more than usually greasy and had to share a table with loud and rude family.

Chow divertido estaba seco; **shu mai cerdo** era más que generalmente grasiento y tenía que compartir una mesa con la familia ruidosa y grosera.
[2] Chow fun is a type of beef, in Spanish is also known as chow fun, there is no translation for this name, it should not be translated
[4] "shu mai cerdo era …" is grammatically incorrect, it should have been translated as ""el shu mai de cerdo era..".

Went on a 3 day oyster binge, with [Fish] bringing up the closing, and I am so glad this was the place it 0 trip ended, because it was so great!

Fue en un atracón de ostras de 3 días, con **Fish** sacando el cierre, y me alegro de que este fuera el lugar donde O viaje terminó, porque era tan grande!

> *[3] "Fish" is not a proper name in this sentence, it should have been translated as "pescado" in the target sentence*

> I had never had [Edamame pureed] before but I thought it was innovative and tasty (could've used a bit more salt).

> Nunca había tenido **Edamame puro** antes, pero pensé que era innovador y sabroso (podría haber utilizado un poco más de sal).
> *[1] "Edamame pureed" should have been translated as "pure de Edamame" or "Edamame en pure". "Edamame puro" doesn't have the same meaning as "Edamame pureed".*

Considerations:

- We consider a target mistranslated if the target still appears in the target sentence but the translation is not correct, that is:
  - The target in the source sentence and the target in the target sentence do not have the same meaning (even if by contact we can still understand the sentence) [1]
  - Targets that should not be translated (i.e proper names) but they have been translated [2]
  - Targets that should be translated (i.e proper names) but they have not been translated [3]
  - Grammatically incorrect translation [4]
- We don't consider a target mistranslated if there is a mistranslation in the surroundings of the target but the target is correctly translated.
- The same considerations that for "Simple translations"

Missing targets

Sometimes the translator doesn't translate the whole sentence, missing parts of it or generates a translation that doesn't contain the target. **In this case, we push the**

Set Target **button without selecting a target. A confirmation message should appear to confirm that the target does not appear in the target sentence.**

> The food was average to above-average; the French Onion soup filling yet not overly impressive, and the [desserts] not brilliant in any way.

La comida era media por encima de la media; el relleno de sopa de cebolla francesa no es demasiado impresionante, no son brillantes de ninguna manera.
The target "desserts" does not appear in the target sentence"

Considerations:
- We only consider the target missing if it does not appear in the target sentence and there is no possible target. If a possible target appears even if the meaning has been changed by the translator we consider it a "**mistranslation**".
- If a multi-word target is missing some words, but at least one appears in the target sentence and it can be considered as the target we consider this a "**mistranslation"**.

Split targets

Sometimes the translator splits a multi-word target across a sentence, that is, extra words that do not belong to the target appear in the target sentence between the target words. In this case, we consider the target a missing target.

My wife had the [fried shrimp] which are huge and loved it

Mi esposa tenía los **camarones** que son enormes y le encantaron **fritos**.
We consider this a missing target