Named Entity Annotation Guide

Group Number: 35

Group Members: YoungJun Cho (Unikey: ycho6245)

The initial annotation guide was based on "An Annotated Dataset of Literary Entities," David Bamman, Sejal Popat and Sheng Shen, NAACL 2019.

We consider six types of entities: Person (PER), Facility (FAC), Geo-political entities (GPE), Location (LOC), Vehicle (VEH), Organisation (ORG). Note, we will not annotate ‘Other’ or ‘Misc’ entities, a practise that is used in some Named Entity annotations.

1.Examples

Person (PER)

Here are some examples from group35-stage1.txt that were classified as PER:

Barbican

the travellers

Captain

scientists

astronomers

Facility (FAC)

Here are some examples from group35-stage1.txt that were classified as FAC:

the Cambridge Observatory

~~faculties~~

his own observatory

Geo-political entities (GPE)

Here are some examples from group35-stage1.txt that were classified as GPE:

Parsontown

Long's Peak

Stony Hill

New York

a _Mediterranean_

Location (LOC)

Here are some examples from group35-stage1.txt that were classified as LOC:

the Moon

the Earth

the summit

the mountains

The northern hemisphere

Vehicle (VEH)

Here are some examples from group35-stage1.txt that were classified as VEH:

An express train

the Projectile

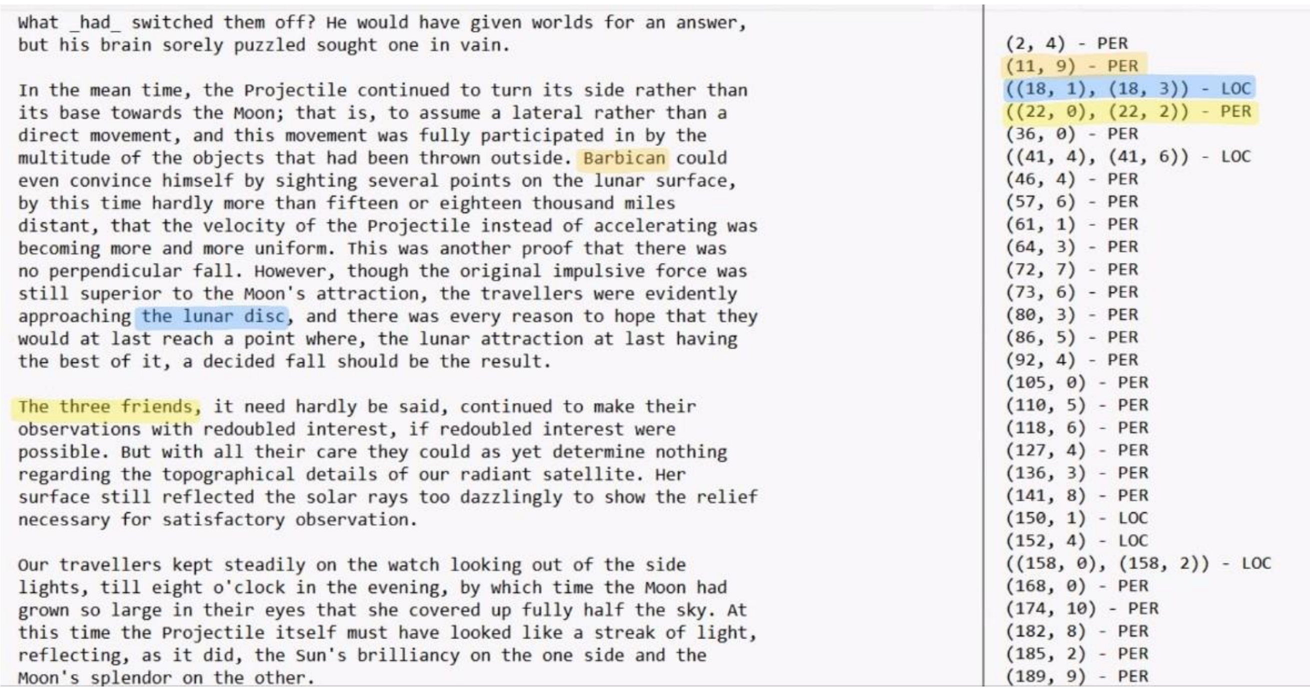
Organisation (ORG)

Here is an example from group35-stage1.txt that was classified as ORG:

~~Representatives~~

2.Explanation

Annotation



```
What _had_ switched them off? He would have given worlds for an answer,
but his brain sorely puzzled sought one in vain.

In the mean time, the Projectile continued to turn its side rather than
its base towards the Moon; that is, to assume a lateral rather than a
direct movement, and this movement was fully participated in by the
multitude of the objects that had been thrown outside. Barbican could
even convince himself by sighting several points on the lunar surface,
by this time hardly more than fifteen or eighteen thousand miles
distant, that the velocity of the Projectile instead of accelerating was
becoming more and more uniform. This was another proof that there was
no perpendicular fall. However, though the original impulsive force was
still superior to the Moon's attraction, the travellers were evidently
approaching the lunar disc, and there was every reason to hope that they
would at last reach a point where, the lunar attraction at last having
the best of it, a decided fall should be the result.

The three friends, it need hardly be said, continued to make their
observations with redoubled interest, if redoubled interest were
possible. But with all their care they could as yet determine nothing
regarding the topographical details of our radiant satellite. Her
surface still reflected the solar rays too dazzlingly to show the relief
necessary for satisfactory observation.

Our travellers kept steadily on the watch looking out of the side
lights, till eight o'clock in the evening, by which time the Moon had
grown so large in their eyes that she covered up fully half the sky. At
this time the Projectile itself must have looked like a streak of light,
reflecting, as it did, the Sun's brilliancy on the one side and the
Moon's splendor on the other.
```

```
(2, 4) - PER
(11, 9) - PER
((18, 1), (18, 3)) - LOC
((22, 0), (22, 2)) - PER
(36, 0) - PER
((41, 4), (41, 6)) - LOC
(46, 4) - PER
(57, 6) - PER
(61, 1) - PER
(64, 3) - PER
(72, 7) - PER
(73, 6) - PER
(80, 3) - PER
(86, 5) - PER
(92, 4) - PER
(105, 0) - PER
(110, 5) - PER
(118, 6) - PER
(127, 4) - PER
(136, 3) - PER
(141, 8) - PER
(150, 1) - LOC
(152, 4) - LOC
((158, 0), (158, 2)) - LOC
(168, 0) - PER
(174, 10) - PER
(182, 8) - PER
(185, 2) - PER
(189, 9) - PER
```

Annotation was proceeded like the picture above.

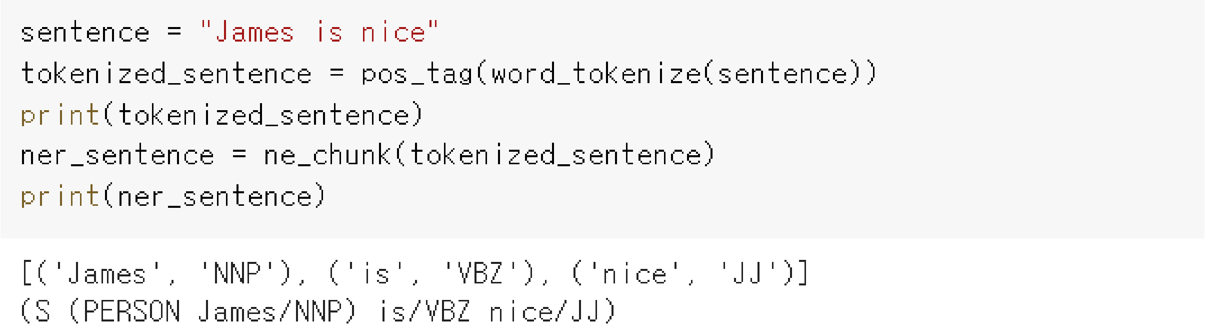
For example, In the text, the word "Barbican" appears in the 11th line and is located in the 9th position based on the whitespace. Therefore, we annotated it as a named entity of type PER with the coordinates (11,9).

Also, I identified a nested annotation in the text, specifically the phrase "the lunar disc", which appears in the first three words of the 18th line. Therefore, we annotated it as a named entity of type LOC with the coordinates ((18,1), (18,3)).

Unusual Cases

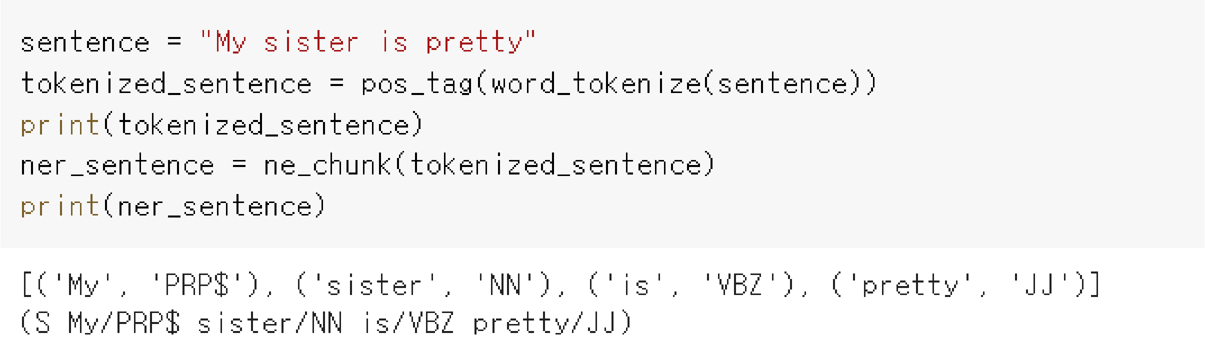
1. Personal and Objective pronouns

   In order to determine whether personal pronouns and objective pronouns should be classified as PER in Named Entity Recognition, we utilised the NLTK package.



```
sentence = "James is nice"
tokenized_sentence = pos_tag(word_tokenize(sentence))
print(tokenized_sentence)
ner_sentence = ne_chunk(tokenized_sentence)
print(ner_sentence)
```

```
[('James', 'NNP'), ('is', 'VBZ'), ('nice', 'JJ')]
(S (PERSON James/NNP) is/VBZ nice/JJ)
```



```
sentence = "My sister is pretty"
tokenized_sentence = pos_tag(word_tokenize(sentence))
print(tokenized_sentence)
ner_sentence = ne_chunk(tokenized_sentence)
print(ner_sentence)
```

```
[('My', 'PRP$'), ('sister', 'NN'), ('is', 'VBZ'), ('pretty', 'JJ')]
(S My/PRP$ sister/NN is/VBZ pretty/JJ)
```

Hence, I decided Personal pronouns such as "I", "you", "he", "she", "it", "we", and "they" and Objective pronouns such as "his", "her", and "them" are not annotated as PERSON in Named Entity Recognition (NER) tasks.

This is because PERSON is a category that is reserved for specific named entities and those pronouns are not specific enough to identify a named entity on their own.

However, as an exception, when a person comes after an objective pronoun, we decided to tag them together as PER. One example for this format would be ‘My sister’.

2. ‘The southern’

   In the context of the sentence, this word refers to 'the southern hemisphere' with the word 'hemisphere' being omitted. Therefore, I classified it as a named entity of type LOC.

~~3. ‘Representatives’~~

~~"Representatives" in this sentence refers to a group of individuals who represent Earth and humanity. While "Representatives" could be interpreted as referring to specific people, the more appropriate interpretation in this context is that "Representatives" refers to a collective entity.~~

~~Therefore, I decided to annotate it as ORG.~~

4. ‘the track’

   The word used in the context does not refer to an actual track, but rather to being 'off the track'. Therefore, we did not tag it as a location but rather interpreted it to mean 'deviating from the desired path'.

5. ‘Latin’

   The word 'Latin' was used to refer to the Latin language, and therefore I did not annotate it as a named entity of type PER.

6. ‘somewhere’

   Although the word indicates a location, its physicality is unclear. Therefore, I excluded it from the LOC category.

7. Words with ‘--’

   There were cases where words had '--' attached either at the beginning or end of the word. For instance, in the phrase 'a Polish astronomer--more', the words 'astronomer' and 'more' were connected by '--'. Since 'a Polish astronomer' needed to be annotated as a person, we classified the word 'more' as a named entity of type PER even though it was attached to the word 'astronomer' with '--'.

Updated explanations – stage2

1. In the case of annotated words with articles (a, the)

   Although the annotated words were the same between the guide I created and the one provided by the school, there was a difference in whether or not articles (a, the) were included before the annotated words. To maintain consistency, I decided to include the articles when annotating the words.

   <Example>

   My previous annotation: seas

   Provided annotation: the seas

   Final chosen annotation: the seas

2. In cases where there were phrases or clauses that modify the word to be annotated

   In this case, my guide only annotated the word itself, excluding the modifying phrases or clauses. However, the guide provided by the school included the modifying phrases or clauses in the annotation process. Therefore, if there were parts that modify the word to be tagged according to the school's guide, such as prepositional phrases, I included those phrases in the annotation process.

   <Example>

   My previous annotation: observers

   Provided annotation: the former observers of the moon

   Final chosen annotation: the former observers of the moon

3. Wrong labels assigned for each annotated word

   When reviewing the guide provided by the school, we found that there were many cases where the label assigned to the annotated word was deemed inappropriate. In such cases, I excluded those words from our annotation process. Here is an example of such cases.

   <Example>

   Provided annotation: therefore - PER

   Final chosen annotation: eliminated

   Since, therefore is inappropriate to be annotated to PER label, these cases are eliminated.

4. Same annotated words, different labelling

   There were cases where the same word was annotated but the labels I chose and the labels provided by the school were different. In such cases, I referred to the explanations of the labels provided by the school and chose a more descriptive and appropriate label.

   <Example>

   My previous annotation: Caspian Sea - GPE

   Provided annotation: Caspian Sea - LOC

   Final chosen annotation: Caspian Sea – GPE

   *As the annotation guide provided says ‘known geographical locations within the real world’ is included in GPE, I referred the explanations from the guide, and made a decision.

5. Personal and Objective pronouns

   The annotation guide provided by the school annotated Personal and Objective pronouns as PER. However, I decided to exclude Personal and Objective pronouns from the annotation, following the previous annotation guide. I made this decision based on the analysis using the NLTK package.

6. When performing Named Entity Recognition (NER) annotation, the goal is typically to identify named entities that actually exist, such as people, places, companies, dates, and times. Normally, imaginary or fictional entities such as characters and place names in literature are generally excluded from NER annotation. However, depending on the task or project, they can be included. For this annotation task, which is for analysing a novel, I consider characters and place names in the novel as valid targets for NER annotation and include them in the annotation process.