Annotation Guidelines

Version 1

Stephen Mayhew, November 2022, based on NorNe guidelines (https://github.com/ltgoslo/norne/blob/master/NorNe%20Annotation%20Guidelines.pdf)

1. What is an annotation?

An annotation is “a markup of a text span in a specific format that indicates a feature or features of the text within the span.” (MUC-7 (https://www-nlpir.nist.gov/related_projects/muc/index.html))

In addition to a span, all annotations have a type: e.g. person, location, organization.

2. What should be annotated?

In the Universal Dependencies NER project we annotate named entities. All names of people (fictional and real), organizations, and locations that are found in the texts should be annotated.

The main rules of thumb for annotating something as a name are:

- The word is or contains a proper noun
    - ‘Michael Jordan’ is composed of two consecutive proper noun tokens with a single reference
    - Names may include words which are not proper nouns, e.g. prepositions, as in ‘John of Burgundy’
    - ‘Pearl Harbor’ is a name of a location, yet it consists of two common nouns, “pearl” and “harbor”
- The word has a unique reference, i.e. it points to a single thing in the world
- The reference is constant over time
    - ‘Tottenham goalkeeper’ is not a Person name, as its reference will change over time
    - ‘Al Yankovic’ is a name, as its reference does not change

Some examples which are regarded names:

- Names: E.g., ‘Charlie Chaplin’, Chaplin’, ‘Charlie’
- Initials: E.g. ‘C.C.’ or ‘CC’
- Spelling mistakes: E.g. ‘Charie Chaplin’
- All orthographic variations E.g. ‘charlie chapline’, ‘Charlie CHAPLIN’
- Names of fictional characters: ‘Bilbo Baggins’, ‘Hercule Poirot’, ‘Ryan North’

Some examples which are regarded non-names:

- Pronouns (he, her, they, them)
- Expressions of time
- Currencies (Euro, Kroner, Dollar, etc.)
- Language names and nationalities (English, Spanish, German, Chinese)

2.1 Entity types

In this tagset, there are 3 entity types:

- Person (PER)
- Organization (ORG)
- Location (LOC)

location (LOC)

The same surface form can belong to different types due to different contexts (metonymy). Thus, the same name can refer to both the organization and the place where it is:

- She worked for Theater Cafe$_{(ORG)}$, but they met at the Cafe Love$_{(LOC)}$.

“The White House” is a location when something happens there, and an organization when they decide something.

2.1.1 Entity type: Person

The person name category includes names of real people and fictional characters.

- 10 Best Sonnets by William Shakespeare$_{(PER)}$
- You’re a mean one, Mr. Grinch$_{(PER)}$
- She reminds me of Red Riding Hood$_{(PER)}$’s grandma.

Family names should be annotated as Person, even if they refer to several people.

- We’ve been invited to the Baums’$_{(PER)}$ house tomorrow.
- And, uninvited, the Brothers Grimm$_{(PER)}$

Gods are also annotated as Persons, when capitalized and having a single reference

- God$_{(PER)}$ plays a big role in life
- Thor’s$_{(PER)}$ hammer is too heavy for me

When not having a single reference, it is not tagged:

- God or not, Zlatan$_{(PER)}$ is the best.
- After a long run, I feel like a god.

Do not tag animal names as persons:

- The famous TV-dog Lassie
- Bo Obama is a pet dog of the Obama$_{(PER)}$ family

Usernames should be annotated as Persons if they are used as names:

- lolgrapeuk$_{(PER)}$ told me that thunderbiscuit$_{(PER)}$ is joining later

2.1.2 Entity type: Organization

Include any named collection of people, such as firms, institutions, organizations, pop groups, sports teams, unions, political parties etc.

- CBS$_{(ORG)}$ bought BBC$_{(ORG)}$
- Pixar$_{(ORG)}$ created Monsters Inc.
- The Department of Defense$_{(ORG)}$ is a top group in the US$_{(LOC)}$ government.
- Having left the Republican Party$_{(ORG)}$, John Smith$_{(PER)}$ will head up United Airlines$_{(ORG)}$.

Organization also includes names of places when they act as administrative units or sports teams:

- Baltimore$_{(ORG)}$ lost to Indianapolis$_{(ORG)}$ last weekend
- Boston$_{(ORG)}$ announced yesterday that mask restrictions would be lifted

Documents from the “various” corpus in UNER can be tricky. The establishments they mention are very often a location

Documents from the "reviews" corpus in UNER can be tricky. The establishments they mention are very often a location, but are treated like an organization. Use your judgment in these cases.

- I'm never going back to Dusty's(LOC)! The food was disgusting!
- Even though I've never tried Dusty's(ORG), I've heard they have disgusting food.

Corporate designators like Co. and Ltd. are to be included as part of the name. The term "& co" in "Obama & co" should not be annotated because it is a designator for unnamed persons, and not an organization or a company:

- Obama(PER) & co cleaned the table

In contrast to e.g.:

- Law firm Johnson & Co(ORG) represented Hansen(PER)

Only tag brands if referring to the organization itself, not as a brand label.

- Nike(ORG) has confirmed the new shoe sizes.
- I think I got injured because I wore Nike(OTH).

2.1.3 Entity type: Location

Includes geographical places, buildings and facilities. Examples are airports, churches, restaurants, hotels, tourist attractions, hospitals, shops, street addresses, roads, oceans, fjords, mountains, planets, parks and also fictional locations.

Examples:

- Germany(LOC) is very close to Sweden(LOC)
- I have reservations tomorrow morning at The Breakfast Club & Grill(LOC)
- Church of the Ascension(LOC) will have Easter services at 7am and 10am
- The Gulf of Mexico(LOC) is a beautiful blue in the morning.
- I cannot recommend the restaurant on Carson Street(LOC).
- It takes only 15 minutes, if you go on I-70(LOC)
- If I ever traveled from Earth(LOC) to Mars(LOC), I would want to be asleep.

When two locations are consecutive, we annotate separately:

- i have like 40 friends who live in philly(LOC) pennsylvania(LOC).

Do not include locational descriptors unless they are part of the name:

- She lives in northern California(LOC).
- My two favorite places are West Virginia(LOC) and South Sudan(LOC).

Postal addresses are an exception to the separate annotation rule: they should be annotated as a whole span, as they constitute a unique reference. Examples:

- 5900 Penn Avenue, Pittsburgh(LOC)
- Target Headquarters, 1000 Nicollet Mall TPS-3165 Minneapolis MN 55403, United States(LOC).
- Apple Inc. One Apple Park Way. Cupertino, CA, 95014. United States(LOC).

2.1.4 Entity type: OTHER

In the course of annotation, there may be cases in which a span feels like an entity, but doesn't fit into one of the above categories. For these cases, use the OTHER tag. These annotations will most likely not be included in the final set, but

categories. For these cases, use the OTHER tag. These annotations will most likely not be included in the final set, but can be used to check disagreements between annotators, and possibly stored as a separate annotation layer.

Tag nationalities and languages as OTHER.

- The Argentinian(OTH) diplomat spoke Greek(OTH)

Tag brands as OTHER:

- I refuse to buy an Apple(OTH) Watch.
- My dad was adamant that Saab(OTH) was better than Porsche(OTH)

2.1.5 Relationship to to Universal Dependencies PROPN

We expect that most occurrences of named entities will coincide with the PROPN (https://universaldependencies.org/u/pos/PROPN.html) tag in Universal Dependencies. However, slight differences remain: UD will at times give parts of speech to individual constituents, and may include a broader category of terms.

2.2 Ambiguity

Ambiguity is a frequent source of doubt when annotating. Solve ambiguity in general as follows:

- Choose the entity type based on the local context
- We assume that every entity has a base, or literal, meaning
- When there is ambiguity, either because of lack of context or genuine ambiguity, always choose the literal meaning of the word(s)
- If the context doesn't help, and the surface form is ambiguous, choose the most common usage
- If you had to link this entity to a Wikipedia page, what type would the page be?

Examples:

- Baltimore is top of the table.
    - Choose ORG by the context, sports
- I can't wait to see Indianapolis
    - Could be either ORG or LOC, choose LOC as the literal meaning
- Lawrence is my favorite.
    - Could be either PER or LOC, choose PER as the more common usage

2.3 Examples

2.3.1 Nested names

Always annotate the whole name, never nested parts.

Annotate like this:

- University of Washington St. Louis(ORG)

And not like this:

- *University of Washington(LOC) St. Louis(LOC)

2.3.2 Possession/genitive

Annotate the genitive marker as part of the name

- Charlie Chaplin's$_{(PER)}$ office.

Note that the possessor and possessed named entity substring should be tagged separately:

- Morten's$_{(PER)}$ Sweden$_{(LOC)}$.

2.3.3 Titles

We never annotate titles as a name or part of a name:

- Your Majesty
- President Lincoln$_{(PER)}$
- Honorable King Harald$_{(PER)}$
- Mullah Omar$_{(PER)}$

When the same words refer to the person/occupation, they are not considered names:

- The governor drove into the ditch.

2.3.4 Names connected by conjunctions

Annotate names connected by conjunctions as multiple distinct entities:

- Eli$_{(PER)}$ and Carl I. Hagen$_{(PER)}$

2.3.5 Names that include numbers

Include numbers when they are part of the entity name:

- 10 Downing St$_{(LOC)}$
- 21 Pilots$_{(ORG)}$

2.3.6 Quotations around a name

Exclude quotations and other punctuation characters as part of the name:

- "Becoming John Malkovich$_{(PER)}$" starring John Malkovich$_{(PER)}$.
- Oslo$_{(LOC)}$-Bergen$_{(LOC)}$ is quite far…

2.3.7 Words and Phrases derived from names

Phrases that contain names should not be annotated if they are not in use as a proper noun:

- The Bergen wave is a term for popular musical styles from theBergen$_{(LOC)}$ area
- The Pittsburgh left causes 10 million traffic accidents a year inPittsburgh$_{(LOC)}$.
- I have a Matisse next door, but I think it was painted byPicasso$_{(PER)}$.

2.3.8 Proper names conjoined to other tokens

If a name includes a hyphen, include the non-name tokens in the annotation:

- New York-based$_{(LOC)}$ ACLU$_{(ORG)}$

2.3.9 Combinations of Names

Names that are combined using various conventions, should be given type by the combination of the names:

- He was born in Champaign-Urbana(LOC)

In some cases, two different types of named entity will be combined. Tag first for context, and if the context is ambiguous, or there is no context, tag with the first entity.

- Steve Jones@EPC(PER) told me the prices would drop
- The firm/person I would contact is EPC/Steve(ORG) Jones.

2.3.10 Emails and Usernames

Email addresses and Usernames should not be tagged unless they are used in the place of a name. Examples:

- bob@motels.com(PER) said I should contact him
- I need to email help@homedepot.com(ORG) to get a refund
- You can contact us at hello@dollargeneral.com
- anybody know why @slumpdog(PER) never comments any more?
- thanks for the tip! ~SlamDunk84

2.3 Questions

In cases where the locative meaning applies for an organizational area, there is sometimes doubt as to whether the appropriate annotation should be LOC, or ORG.

- This year, the scheme will be introduced in the EEA
- Police attorney XX at Helgeland police district says one person has been arrested and charged
- So far this year, there have been three arrests in the Hordaland police district, and two in the Hedmark and Rogaland police districts

There is often doubt related to whether a sentence-initial word should be annotated, given that many words can both be used as a proper name and a generic term.

- The Governor of Svalbard has the highest official authority in the area
- The church has its own bodies that can take over the tasks that the King (government) currently handles
- The government has not yet commented on the incident
- The police say that there may be a connection between the misdeeds Sometimes the name is related to the context
- Hacker News sent me a question last week

Sometimes nicknames can contain a proper name. The question in these cases is whether the whole construction should be annotated, or just the proper name.

- The question is who wins the 11 electoral votes in Arizona(LOC), the Grand Canyon state

3. Previous work, basis of the guidelines

These guidelines are heavily based on NorNe (https://github.com/ltgoslo/norne/blob/master/NorNe%20Annotation%20Guidelines.pdf) annotation guidelines, with the biggest changes being the removal of Event, Product, and GPE named entities.

The guidelines are also partially based on:

- ACE (Automatic Content Extraction) English Annotation Guidelines for Entities Version 6.6 2008.06.13
    - https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf (https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf)

- MUC-6 Guidelines
    - https://cs.nyu.edu/~grishman/muc6.html (https://cs.nyu.edu/~grishman/muc6.html)
    - https://cs.nyu.edu/~grishman/NEtask20.book_1.html (https://cs.nyu.edu/~grishman/NEtask20.book_1.html)
- MUC-7 Guidelines
    - https://www-nlpir.nist.gov/related_projects/muc/index.html (https://www-nlpir.nist.gov/related_projects/muc/index.html)
- CONLL NER shared task:
    - http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt (http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt)
- Ontonotes 5
    - https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf (https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf)

4. Acknowledgements

Thanks to Dan Zeman, Amir Zeldes, Constantine Lignos for comments and thoughtful feedback.

5. Changelog

Note: these guidelines originally started at v1, and incremented to v1.1, but as part of UNER v1 release, we have retroactively decremented the versions, so that the guideline version matches the release version.

v0

Initial

v1 (Release Version)

- Added wording about consecutive locations, and tagging of planets