Annotation guidelines for Nested Named Entities for English

Barbara Plank and Sif Dam Sonniks
IT University of Copenhagen
Department of Computer Science
Rued Langgaards Vej 7, 2300 København S
bplank@gmail.com

Abstract

This technical report describes the annotation guidelines for our Nested Named Entities corpus for English, developed for the English Web Text corpus (EWT-NNER). Nested named entities here are treated as a two layer annotation scheme, where the outermost span embraces the longer span and is the most prominent entity reading, and the inner span contains secondary or sub-entity readings. Our guidelines were adopted from the German NoSta-D guidelines (Benikova et al., 2014), which were also the basis for the Danish DaN+ NNER guidelines (Plank et al., 2020), which is a predecessor of the guidelines presented here.

Basics Named entities are nominal phrases that determine specific people, organizations, locations or miscellaneous specific objects like film titles or products. Given the following example:

[Leila] bought [the house]

There are two nominal phrases. Only one of them is a named entity (*Leila*), the second nominal is a common noun.

Only full nominal phrases are potential full NEs. Pronouns and all other phrases should be ignored. For mediums such as social media we do mark usernames, hashtags and emails as potential NEs. It should be noted that capitalization is not a reliable indicator of NE status.

- Full NEs are annotated as LOC (location), ORG (organization), PER (person) or MISC (miscellaneous other)

The data is annotated with the BIO scheme in tabular CoNLL-style format. An example:

```
Texas        B-ORG   B-LOC
Ultimate     I-ORG   O
SteakHouse   I-ORG   O
```

Note that the inner layer is annotated in the second annotation slot (in the example, the LOC), and that empty slots are marked with O.

We next describe each of the four entity types. At the end of the document are summarizing tables, inspired by the tables in the German NoSta-D guidelines (Benikova et al., 2014), to show examples for each category and non-NEs.

Person The tag PER is used for personal names, as well as names of fictional characters and pets. We also use the tag for personal emails and twitter handles (non-personal ones are tagged MISC). Names can include numbers, as in the Danish queen *Margareth II*. For the sake of uniformity, we follow the German NoSta-D guidelines (Benikova et al., 2014) and do not tag *God* and *Satan*.

Organization The ORG tag is used for organizations. Apart from organizations and corporations, this also includes shops, restaurants, newspapers, airports, and more. It is also used for groups (what might also be called "tribes"), though note that if the name is derived from a location, it is LOCderiv and not ORG. See also the section on "Geopolitical units".

Location The LOC tag is used for places. This includes countries, continents, cities, sightseeing spots, rivers, mountains, stadiums, and more. Some locations have city, town, county, etc. included in their official names - others do not. We only tag it as part of the NE if it part of the official name ([*New York City*]LOC vs [*Chicago*]LOC *City*).

Miscellaneous The MISC category covers a lot of ground. It covers titles of books, songs, movies, currency, blog post, as well as publications that are not an institution themselves (in which case the proper tag is ORG). It also covers specific agreements (such as *Indo-Sri Lanka Peace Accord*) but

not general agreement types (*Employment Agreement*). Some MISC NEs might be very generally named, but there should still be differentiation between *the Metro*, which is not an NE, *Metro 4 line*, which is the name of a specific metro line and therefore *does* qualify, and *RER B*, which is the name of a metro system, and *also* qualifies. MISC further covers languages whose names are not derived from locations, as well as product names.

Nested NEs We annotate nested NEs up to two layers. Subsequent layers are noted down separately. A nested NE is an instance where a token contains both a primary and a secondary or subentity readings. This usually occurs if an entity is named after another named entity, or because several instances of one NE share a name (an organization and a physical event, a place and an organization, etc.). For example:

- [[Oregon]LOC University]ORG
- [[Smith]PER Street]LOC
- [[Sweden]LOC]ORG (see "Geopolitical units")
- [Greater [Nadu]LOC]LOC
- [[UN]ORG Security Council]ORG

A NE is only nested if the inner NE is not just another name for the same entity. If we take the NEs *United States of America* and *Republic of Iraq*, both are names of countries. But in the first, *America* is also a nested LOC, as America is an independent area from the USA. *Iraq* is identical with *Republic of Iraq*, and thus is not a nested NE.

The city name is not considered a nested NE in place names prefaced by *New*. For instance:

- [New York]LOC
- [New Delhi] LOC

We also do not annotate nested NEs in NEderivs and NEparts:

- [North American]LOCderiv
- [Oxford-educated]ORGpart

Though note that LOCderivs can be on the inner layer:

- [[British]LOCderiv [Columbia]MISC]LOC

NEpart is not used when NEs appear as part of emails, website names, or file names, as these are only tagged MISC.

Geopolitical units Some entities are in the grey zone between locations and organizations. These are most often countries, but all sorts of area types can fit in this category. As we do not have a specialized tag for these geopolitical units, we tag either as LOC or ORG depending on the context.

A Geopolitical unit is tagged as LOC if it is referred to as a place. If someone is *from* that place, something is happening *in*, or in the context of an attack on, invasion of, strike against, occupation ofthe place (and so on), it is tagged as LOC. Likewise if a statement is given on the country, or if it is commented on in-text.

ORG is used when the entity has the ability to act, make decisions, and respond. In these cases the entity is always tagged ORG on the outer layer and LOC on the inner layer, i.e., [[Sweden]LOC]ORG or in CoNLL-format:

```
Sweden  B-ORG   B-LOC
```

When judging entities in genitive, there is also a distinction to be made. In cases such as *India's Manipur province*, *India* is a LOC, but in cases such as *India's ambassador*, it is considered an ORG (LOC on inner layer). Edge cases such as *USA's power grid* are harder to judge. Here the rule of thumb is: if the ownership can be read as meaning location (Here: the power grid of/in the USA), it is tagged as LOC.

What is not included as NE? We do not include determiners or titles in front of NEs, irregardless of capitalization:

- the [United States]LOC
- The [Beatles]ORG
- Queen [Elizabeth II]PER

The exception is determiners in non-English names, most prominently the Arabic determiner *al*, which is included in names of individuals, groups, newspapers, etc. The determiner is also included in titles of songs, books, movies, etc.

If a NE is introduced (e.g. *US Supreme Court*) and later referred to with a shortened form that is the general noun (e.g. *the Court*), the shortened form is not tagged as NE, despite potential capitalization. If the shortened form does not correspond to the general noun, it is still tagged:

US Supreme Court → the Court → Surpreme Court is a Court → Not NE
House of Lords → the Lords/the House → House

of Lords is neither lords or a house → NE

Note that some NEs have names that are also regular nouns (e.g. *The Senate*) and are tagged as NE.

Internal company departments such as *Finance and Banking* or *HR Department* are not NE. Likewise, numbered codes used internally in a company for other companies and departments are not treated as NE.

Abbreviated forms are annotated, but are not tagged as NEpart if part of a bigger abbreviation:

- [LTTE]ORG (Liberation Tigers of Tamil Eelam - TE not tagged as LOCpart)
- [USD]MISC (United States Dollar - US not tagged as LOCpart)

Be aware that the distinction between NEs and identical words not used as NEs: (*[[Chinatown]LOCpart]LOC* as a location name vs *a [chinatown]LOCpart* used generally)

Issues with tokenization, punctuation, etc. The way a text is written or pre-tokenized for annotation might cause difficulties for annotation. This section covers what to do in these cases.
NEs might share a word part in coordination, leaving them as incomplete names. Following the German NoSta-D guidelines (Benikova et al., 2014), we tag both parts separately as if they were the full names:

```
North   B-LOC   O
and     O       O
South   B-LOC   O
Korea   I-LOC   B-LOC
```

Parts of an NE might be added on with punctuation, or a part might be added in the middle of the NE. The full span, including punctuation, is included in the tagging:

```
(   B-MISC O
Second   I-MISC O
World   I-MISC O
)   I-MISC O
War   I-MISC O

Joe   B-PER O
"   I-PER   O
The   I-PER   O
Killer   I-PER   O
"   I-PER   O
Smith   I-PER   O
```

Sometimes words might be split up in the data. Regardless of this, it is tagged as one semantic unit:

```
Lora  B-PER   O
Sullivan@ENRON I-PER O

Cologne   B-LOCpart   O
-  I-LOCpart   O
based I-LOCpart O

Jim B-PER   O
's   I-PER   O
```

NEderiv and NEpart Derivations of NEs, i.e., words which are derived through morphological derivation processes, are marked with NEderiv (e.g., *Danish*). Note that this includes zero-derivation, where the word form remains unchanged (e.g. Google(NE) → Google(verb)). NEderivs do not need to be nominal phrases. Declination (e.g., possessives and plural forms) are not considered derivations and are directly annotated as NEs. The most common NEderivs are languages and nationalities, derived from places. Examples:

- PERderiv - Saddamites, Shakespearean, Orwellian
- ORGderiv - Republican, Soviet, Democrats, Google(verb)
- LOCderiv - Iranian, Mexicans, caucasian
- MISCderiv - eSpeakers(from computer program eSpeak), Cajunish

NEpart is used when a NE is incorporated into another word. Unlike with a NEderiv, the rest of the word must function on its own without the NEpart. Note that the NEpart is not tagged if the NE is part of an email address, website name or filename. Examples:

- PERpart - TomiPilates
- ORGpart - Enron-sponsored
- LOCpart - Chinatown, Pro-India
- MISCpart - Swahili-speaking

If html-tags are attached to a NE (e.g. *mailto:email@email.com*), it is not tagged as NEpart.

It is possible for a word to be both a NEpart and NEderiv. In these cases it is tagged NEpart on the outer layer and NEderiv on the inner, as the word was derived before it was used as a NEpart.

If more than one NE is part of a given word, it is annotated only once for each NEpart-type:

- [[Israeli-Palestinian]LOCderiv]LOCpart
- [[anti-Italian]LOCderiv]LOCpart

Zero derivation and meaning extension In cases where the word form remains unchanged, it can be hard to discern between NEs and derivations. To illustrate the reasoning one might employ, *ISDA Master Agreement* will be used as an example. The *ISDA Master Agreement* is not in itself a NE, as it is a contract type, not a specific agreement. *ISDA* is, however, an ORG (*ISDA = International Swaps and Derivatives Association*). Which gives us this annotation:

```
the       O    O
ISDA      B-ORG    O
Master    O    O
Agreement     O    O
```

This agreement is referred to in different ways in the text, including *ISDA doc* (which has a similar annotation to the full title, as it is still a compound noun) and *the ISDA*, meaning the ISDA master agreement and not the organization. When used like this, it counts as a derivation:

```
send O    O
them O    O
the O    O
ISDA B-ORGderiv     O
```

Similarly with the chicken breed *Sussex chicken*, annotation is straightforward at first (noting that animal breeds are not tagged as NE):

```
Sussex    B-LOC      O
Chicken O    O
```

Often the chicken is referred to simply as a *sussex*, in which case it also counts as a derivation:

```
he    O    O
bought    O    O
a    O    O
Sussex LOCderiv O
```

Thus we are included meaning extension under the umbrella of derivation, as it would be incorrect to mark a document as an organization, and a chicken as a location. Artworks referred to by the artists name, such as *a Picasso* or *a Monet* are PERderiv.

Noun adjuncts used adjectively Another cause of difficulty is the somewhat fuzzy line between compound nouns and noun adjuncts which are used adjectively. To keep the guidelines as clear and uniform as possible, we only annotate known adjectives (e.g. *US, UK, Texan*) as NEderiv. This means that the NEs in noun phrases such as *The NY Times, Iraq oil fields, Dehli police officer* are tagged as pure NE, regardless of function.

The only exception is when a non adjunct is coordinated with a true adjective, in which case both are tagged as NEderiv:

```
the O O
English B-LOCderiv O
and O O
Singapore B-LOCderiv O
financial O O
transactions O O
```

Summary of Difficult Cases

- Geopolitical units
  - [Hungary]LOC (area)
  - [[Hungary]LOC]ORG (when country has agency)
- Coordinated words with shared word part are tagged separately: *[North]LOC and [South [America]LOC]LOC*
- Abbreviations are tagged as NE but without inner layers: *[USD]MISC, [UCPH]ORG, [USA]LOC, [EU]ORG*
- Punctuation is included if in the middle of NEs: *[Thomas "Tom" Smith]PER, [(Second World) War]MISC*
- Html-tags appended to words are ignored in the tagging: *[mailto:Jim@email.com]PER, [<p>Mis]PER is my cat*
- Disregard tokenization borders if it splits semantic units: *[EU's]ORG, [[French-speaking]B-LOCderiv]B-LOCpart*
- Zero derivation - sometimes word form does not change with derivation: *[Google]ORGderiv (verb)*
- Meaning extension - NE's used metonymically are tagged as NEderiv: *[Picasso]PERderiv (painting), [ISDA]ORGderiv (document), [Sussex]LOCderiv (Chicken)*

- NES acting as attributive nouns are ONLY tagged as NEderiv if they are established adjectives: *[UK]LOCderiv politics, [United Kingdom]LOC politics*, OR if coordinated with a true adjective *[UK]LOCderiv and [France]LOCderiv politics*,

Examples of categories The following tables provide a quick reference for examples of the different categories.

Entity	Examples
Names	Anna Hansen, [Hansen, Anna K.], Smith, Margareth II
Pet names	Fido, Mis
Pseudonyms and artist names	Lady Gaga
Characters	Donald Duck, Spiderman
Initials	PS, JB
Nicknames and handles	Humanpixel, CoolGuy88
Personal email addresses	Anna.Hansen@email.com

Table 1: Entity type PER - Person

Entity	Examples
Organizations	UN, EU
Companies	Chanel, Enron
Clubs	FC Barcelona, Dallas Cowboys
Bands	Beatles
Peoples, Indigenous nations	Comanche Nation, Sitka Tribe
Airports	LAX, Flughafen Frankfurt am Main
Universities, Institutes	Manipur University, Institute of Health Sciences
Restaurants, Hotels	Hyatt Hotel, Holiday Inn
Military	US Army, Corps of Engineers
Hospitals	St. George's Hospital
Sports events, Festivals	Bundesliga, Copenhell
Political parties	Democratic Party, GOP, Democrats
Libraries	Royal Library

Table 2: Entity type ORG - Organization

Entity	Examples
Planets, Moons	Mars, Moon
Streets, Plazas, Squares	Crowle Road, A614, Red Square
Sightseeing spots, Churches	Taj Mahal, Sacre Coeur
Shopping centers	Crescent Shopping Center
Mountains, Lakes, Rivers, Oceans	Mt. Fuji, Lake Victoria, Ganges, Atlantic Ocean
Cities, Districts	New York, Chicago Loop
Countries, Continents, Regions	Slovenia, Africa, Subsaharan Africa, South India

Table 3: Entity type LOC - Location

Entity	Examples
Valuta	Eur, €, USD $
Book, Movie, Song & Chapter titles etc.	War and peace, Inception, Symphony No. 4
Wars	Second World War
Political events	9/11
Market indexes	Dow Jones, Bondad pool
Languages	Swahili
Games	Guild Wars, Chess
Artworks	Guernica
Blogs, magazines that are not institutions unto themselves	Hidden Nook
Prizes and medals	Varnegie Medal of Bravery
Websites, forum channels	www.google.com, r/denmark
Project names	War on Drugs
Computer software	Firefox, Photoshop
Natural and other disasters	Hurricane Katrina, Chernobyl

Table 4: Entity type MISC - Miscellaneous

References

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. Nosta-d named entity annotation for german: Guidelines and dataset. In *LREC*, pages 2524–2531.

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. DaN+: Danish nested named entities and lexical normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662, Barcelona,

Entity	Examples
Food dishes	Sushi, empanadas
Animal Species	Zebra, Oscar(Fish)
National holidays, Religious events	Christmas, Ramadan, Valentine's day
Religions, Political and artistical schools of thought	Christianity, Nazism, Conservatism, Beat, Modernism
General Nouns	Government, the Court
Internal departments	HR, North America (still taged as LOC), Business & Finance
Articles	a, the

Table 5: NOT NE

Spain (Online). International Committee on Computational Linguistics.