MaiNLP NER Annotation Guidelines

Siyao Peng Zihang Sun Huangyan Shan Marie Kolm
Verena Blaschke Ekaterina Artemova Barbara Plank
LMU Munich, Germany
{siyao.peng, b.plank}@lmu.de

Version: March 19, 2024

Introduction

This is an exemplified annotation guideline for Named Entity Annotations across projects within the MaiNLP group. The guideline targets varied genres – such as Wikipedia, news, academic, and social media – and different languages and dialects, such as English, German, and Bavarian German.

We base our criteria on the datasets and guidelines of the CONLL2003 Shared Task: Language-Independent Named Entity Recognition (Tjong Kim Sang & De Meulder, 2003).[1] In the meantime, we also refer to specific annotation details discussed in the following guidelines:

- German NER guidelines from the NoSta-D corpus (Benikova et al., 2014)[2]
- Danish NER guidelines for the DaN+ corpus (Plank et al., 2020)[3]
- English NER guidelines for the EWT corpus (Plank & Sonniks, 2021)[4]

Since the CONLL2003 NER annotation schema has been broadly adapted and the dataset widely used in NER tagging experiments, we decided to follow such a schema to make our human annotations more compatible with state-of-the-art corpora and parsers.
Most essentially, this implies:

- Only annotating named entities, but not common nouns or pronouns
- Not annotating nested entities
- Focusing on PERSON, LOCATION, and ORGANIZATION entities but also experimenting with sub-categorizing the MISCELLANEOUS category (see NameYourType – an experiment with the MISC category)

What is a Named Entity (NE)

Named Entities (NE) are nominal phrases that determine specific objects in the real world (Plank et al., 2020). Person (PER), location (LOC), and organization (ORG) are the most frequently annotated entities in NE corpora.

Here are some examples:

- **[Germany]_LOC** 's representative to the **[European Union]_ORG** 's veterinary committee **[Werner Zwingmann]_PER** said on Wednesday consumers should buy sheepmeat from countries other than **[Britain]_LOC** until the scientific advice was clearer .
- Der Opernsänger , spätere Intendant des **[Darmstädter Hoftheaters]_ORG** und Meister heimatlicher Erzählkunst , **[Ernst Pasqué]_PER** , zeichnete sie literarisch nach .

[1] https://www.cnts.ua.ac.be/conll2003/ner/annotation.txt
[2] https://drive.google.com/drive/folders/1kC0I2UGl2ltrluI9NqDjaQJGw5iliw_J
[3] https://aclanthology.org/2020.coling-main.583.pdf
[4] https://github.com/bplank/nested-ner/blob/master/techreport/EWT_NNER_annotation_guidelines.pdf

NE Encoding

Editing and storing named entity annotations in a tab-separated document (*.tsv) is a common practice in the field. In the TSV file, sentences are separated by an empty line, and within one sentence, each token and its part-of-speech, named entity tags reside in the same line with tabs (\t) as separators.

BIO encoding is used for NE tags – B for Beginning, I for Inside, and O for Outside. If a token is annotated as B-X, it is the beginning of an X-entity. If an entity spans multiple tokens, the second to last tokens are tagged I-X. Elsewhere, if a token does not belong to any entity, it is labeled O (Lample et al., 2016).

The following example shows a transliteration between TSV BIO encoding and plain text sentence with NE bracketing:

Sentence examples with NE bracketing:
Only [France]_LOC and [Britain]_LOC backed [Fischler]_PER 's proposal .
The [EU]_ORG 's scientific veterinary and multidisciplinary committees are due …

TSV format BIO-encoded annotations:
Only O
France B-LOC
and O
Britain B-LOC
backed O
Fischler B-PER
's O
proposal O
. O

The O
EU B-ORG
's O
scientific O
veterinary O
and O
multidisciplinary O
committees O
are O
due O
to O
……

NE Spans

This project follows the CoNLL 2003 schema to mark the longest Named Entity. This implies that we do not annotate non-named and nested entities.

In this section, we illustrate a few strategies for determining the span of NEs, namely, an entity's start and end tokens.

Constraints by Part-of-speech (POS)

We can exclude specific tokens from an NE based on their grammatical categories, i.e., part-of-speeches. The following grammatical items are not included if they occur at the edge of an entity.

Demonstratives and determiners

Demonstrative and determiners such as *the* in English, *der, das, die* in German, or *de* in Bavarian are not included in a named entity.

English example:

- the [United States]_LOC

German examples:

- [Niederlanden]_LOC
- In den [USA]_LOC

Bavarian example:

- Vo de [indogermanischn]_LANGderiv

Salutations

Titles and other terms that address peoples' names are not included in annotations, for example, *Mr. Ms. Mrs., King, Queen, Prince, Princess* in English, and *Frau, Herr* in German.

English examples:

- Mr. [Adel Ibrahim]_PER
- King [Hussein]_PER
- Princess [Diana]_PER

German example:

- Frau [Blaul]_PER

Bavarian example:

- Da König [Ludwig]_PER

Numeral

Numbers are not included, for example, year, house number of a street, etc.

English example:

- 2008

German example:

- der [Eschersheimer Landstraße]_LOC 247

Bavarian example:

- In da [Altebergenstroßn]_LOC 3

Adjectives

Adjectival modifiers of nouns are not part of the named entity unless they occur as a fixed expression.

Bavarian example:

- [deitschn]_LANGderiv [Duden]_WOA

Prepositional phrases

Modifiers such as prepositional phrases are not included in entities.

English example:

- state of [South Dakota]_LOC

Bavarian example:

- im [boarischn]_LOCderiv [Schwoam]_LOC vo links in de [Doana]_LOC

Only named tokens are included in the span

It is important to remember that we only annotate the named portion of a noun phrase.

English example:

- [Partizan]_ORG and [Red Star]_ORG of [Yugoslavia]_LOC , [Alba]_ORG of [Germany]_LOC , and [Benetton]_ORG of [Italy]_LOC

German examples:

- der **[osmanischen]_ORGderiv** Annexion des [Jemen]_LOC
- Haltestelle [Rathaus]_LOC
- Compare: [stuttgarter]_LOCderiv Flughafen VS. [Flughafen Stuttgart]_LOC

Bavarian example:

- [europäischen]_LOCderiv [Hunnen]_ORG

Longest if nested

Since we do not annotate nested entities for practical reasons, only the longest entity is annotated for nominal compositional phrases.

English examples:

- [San Francisco Giants]_ORG with no annotation for *San Francisco*
- [Oklahoma State University]_ORG with no annotation for *Oklahoma*

German examples:

- [Bayerische Vereinigung]_ORG with no annotation for Bayerische
- [Deutschen Oper Düsseldorf-Duisburg]_ORG with no annotation for *Deutschen* or *Düsseldorf-Duisburg*
- das „ [Forschungszentrum für öffentliche Meinung und Massenkommunikation]_ORG “ der [Fakultät der Sozialwissenschaften der Universität Ljubljana]_ORG

Bavarian examples:

- fum [Untarichtssminisderium]_ORG (fria [Bundesminisdearium fia Büdung , Wissnschoft und Kuitua]_ORG , seid 2007 [Bundesminisdearium fia Untaricht , Kunst und Kuitua]_ORG) fum [Estareichischn Bundesfalog]_ORG ([ÖBV]_ORG)

Split entities, if syntactically separable

Repeated NEs

If an NE is repeated both as a full name and as an abbreviation, we annotate the two occurrences separately.

English example:

- the [China National Nonferrous Metals Import and Export Corp]_ORG ([CNIEC]_ORG)

German example:

- [Kommunalen Ausländer- und AusländerinnenVertretung]_ORG ([KAV]_ORG)

Bavarian example:

- In [Dinkelsbühl]_LOC ([Wernitzstrandbod]_LOC) , [Wassatrüdinga]_LOC ([Wernitzbod]_LOC) und [Oettingen]_LOC ([Wernitzfreibod]_LOC)

Coordinated NEs

If conjunction coordinates two entities, we annotate them separately.

English examples:

- [North]_LOC and [South America]_LOC
- the [Irish Navy]_ORG and [Air Corps]_ORG

German example:

- [Arbeitsgemeinschaft sozialdemokratischer Frauen]_ORG , [Grünen-Frauen]_ORG und [Eschborner Frauenforum]_ORG

Bavarian examples:

- [Nieder]_LOC , oder [Oberbayern]_LOC
- [Deitsche Museum]_LOC , oda d' [oide]_LOC , d' [neie]_LOC und d' [Pinakothek vo da Modeane]_LOC

Interrupted named entities

If other tokens intervene in a multi-word named entity, the multi-word named entity gets canceled. However, we still mark smaller NER spans that remain uninterrupted.

German example:

- die [Slowenische]_LOCderiv [Evangelische]_RELIGIONderiv ([lutherische]_RELIGIONderiv) Kirche

NE Types

In this section, we introduce the three most common types of entities – PERSON, LOCATION, and ORGANIZATION – referring to examples from CONLL2003 and other NE annotations.
We also experiment with expanding the MISCELLANEOUS label into fine-grained sub-types, and hopefully, we can observe some interesting phenomena regarding the diversity and agreement among innovative entity types.

PERSON

The following sub-types are classified as PER NEs:

First, middle, and last names of people

English examples:

- [John Lloyd Jones]_PER
- [Hendrix]_PER

German examples:

- [Detlev Engel]_PER
- [Biwer]_PER

Bavarian example:

- [Zankl Albert]_PER

Stage, fiction names, nicknames

English examples:

- [Batman]_PER
- [Snow White]_PER
- [Don Quixote]_PER

German examples:

- [Hauptmann Deutschland]_PER
- [Wilhelm Meister]_PER

LOCATION

The following sub-types are classified as LOC NEs:

Geo-political locations

These include country, province, state, city, town, village, continent, street, etc.

English examples:

- [Munich]_LOC
- [Wall Street]_LOC
- [China]_LOC

German examples:

- [Deutschland]_LOC
- [San Martin]_LOC
- [Cali]_LOC
- [Vereinigten Staaten]_LOC
- [Korffstraße]_LOC

Bavarian examples:

- [Minga]_LOC
- [Daetschland]_LOC

Natural locations

These include rivers, mountains, oceans, beaches, national parks, etc.

English examples:

- [Lake Maracaibo]_LOC
- [Island Beach State Park]_LOC

German examples:

- [Ratzeburger See]_LOC
- [Nationalpark Triglav]_LOC
- [Adriaküste]_LOC

Bavarian example:

- [Starnberga See]_LOC

Human-built constructions

These include bridges, tunnels, towers, buildings, etc.

English examples:

- [Trent Bridge]_LOC
- [Eurotunnel]_LOC

German examples:

- [Friedensbrücke]_LOC
- [Johanniterturm]_LOC

Bavarian example:

- [Eisenbaunwossaturm]_LOC

Public places[5]

These include museums, airports, stations, galleries, libraries, etc.

English example:

- [Metropolitan Museum]_LOC

German examples:

- [Museum Großauheim]_LOC
- [Alte Josefshaus]_LOC

Bavarian example:

- [Passauer Glasmuseum]_LOC

Metaphorical and artificial locations

English example:

- He searched for ngrams on [Google]_LOC.

German examples:

- Ankündete auf [Facebook]_LOC
- Zwengs dem wead [Bassau]_LOC aa eftas as " [Venedig vo Bayan]_LOC " ghoaßn . (Passau metaphorically as the "Venice of Bavaria")

[5] This differs from the EWT and NoSta-D guidelines in which museums, airports, and hotels are tagged as ORG.

ORGANIZATION

The following sub-types are classified as ORG NEs:

Countries as a political unit

When countries function as a political unit, it is an ORG.

German examples:

- [Republik Slowenien]_ORG
- [Slowenien]_ORG trat der [EU]_ORG und der [NATO]_ORG
- [Herzogtum Estarreich]_ORG

Bavarian example:

- [Bundesrepublik Deitschland]_ORG

Companies

These include media (press, radio, etc.), bank, stock, manufacturer, brand, etc.

English examples:

- [Chase Securities Inc.]_ORG
- [Kathmandu Post]_ORG
- [BBC radio]_ORG

German examples:

- [Edeka Nordbayern]_ORG
- [Bundespost]_ORG
- [Volkswagen]_ORG

Public facilities

These include hospitals, nurseries, schools, universities, etc.

English examples:

- [Royal Free Hospital School of Medicine]_ORG
- the [Military College]_ORG
- the [University of the Witwatersrand]_ORG

German examples:

- [Freien Universität]_ORG
- [Hilfe für krebskranke Kinder]_ORG
- Der [Michael-Grzimek-Schule]_ORG

Political and public service organizations

These include governments, non-government organizations (NGOs), commonwealth societies, etc.

English examples:

- the [Government of National Unity and Reconciliation]_ORG
- [Chesapeake Police Department]_ORG
- [World Bank]_ORG
- the [Transportation Ministry]_ORG

German examples:

- [Frankfurter Verband für Alten- und Behindertenhilfe]_ORG
- Die [Bundesvereinigung der Deutschen Arbeitgeberverbände]_ORG
- [Weltbank]_ORG

Bavarian examples:

- [Regiarungsbeziak Owabayern]_ORG
- [Boarische Kuitusministerium]_ORG

Recreational groups

These include sport clubs and associations, bands, orchestras, etc.

English examples:

- [Philharmonia Orchestra]_ORG
- [English Football League]_ORG

German examples:

- [FC Bayern]_ORG
- [Berliner Philharmoniker]_ORG

Bavarian example:

- [Fédération Internationale de Tir aux Armes Sportives de Chasse]_ORG

Groups of People

An organized group of people could be labeled as ORG, e.g. tribes.

German examples:

- [Vikings]_ORG
- Den [Slawen]_ORG
- Die [Franken]_ORG

However, we only label established or historically well-known families, but not families of ordinary individuals. For example, [Kardashian family] would be a named entity but not the “Tom family”.

Empires and established families

German examples:

- [Herzogtum Krain]_ORG
- [Heiliges Römisches Reich]_ORG

WOA (Work-of-art)

This sub-type includes human-made products (i.e., work-of-art), for example, songs, films, TV programs, art pieces, books, laws, video games, etc.

English examples:

- [The Shining]_WOA
- [Knockin' on Heaven's Door]_WOA
- [Independent Spirit Awards 1988]_WOA

German examples:

- [Moonwalker]_WOA
- [Bequeme Monatsraten]_WOA
- [Verfassung der Republik Slowenien]_WOA
- [Wörterbuch der deutschen Sprache in Österreich]_WOA=

Bavarian examples:

- [NS-Gwoitherrschaft]_WOA
- Din [16er TV]_WOA
- Der [Nein Mittn vo Bossa]_WOA

EVENT

This sub-type includes historical events, conferences, festivals, sports competitions, and temporal terms such as periods or epochs.

English examples:

- [Second World War]_EVENT
- the [Conference on Disarmament]_EVENT
- the [Olympic Games]_MISC

German examples:

- [Ersten Weltkrieg]_EVENT
- [Schrimpegassefest]_EVENT
- [Salzburger Festspielen]_EVENT
- [Internationale Konflikte der Nachfolgestaaten Jugoslawiens]_EVENT
- [die Zeit des Vormärz]_EVENT
- [Edo-Periode]_EVENT
- [Eiszeit]_EVENT

Bavarian examples:

- [Politische Aschamiedwoch]_EVENT
- [Maya-Klassik]_EVENT

LANG (Languages)

This sub-type includes various languages.

English example:

- [English]_LANG

German examples:

- [Deutsch]_LANG
- [standarddeutsch]_LANG

Bavarian example:

- [neihochdeitsch]_LANG

RELIGION

German example:

- [Satanismus]_RELIGION

Bavarian example:

- [Protestantismus]_RELIGION

MISC (Miscellaneous)

MISC is a last-resort category.
Many corpora label NEs that do not fit in any of the previous categories as MISC.

German examples:

- Award: Einen [Grammy]_MISC
- Project: [Pap Smear]_MISC, [Truth About Seafood]_MISC
- [Eurobarometers]_MISC
- [Grünen Band Europas]_MISC, [Blauen Herzen Europas]_MISC (a natural conservation network)
- [Hussiten]_MISC (a religious movement)
- Ludwig-Thoma-Bahn (a railway system)
- [Triglav-Nationalpark]_LOC ([WDPA]_MISC 2517) – WDPA is a catalog
- [Goldenen Schallplatte]_MISC (an award)

Bavarian examples:

- [Dochau-Sid]_LOCpart
- Konzeption vo da Bundeswehr
- [Naše cesta kolem světa]_MISC

Suffixes -deriv and -part

We follow the GermEval and subsequent NER datasets to include subtypes -deriv and -part on top of each preceding entity type.

The subtype -deriv is mainly used for adjectival derivations of nominal named entities, e.g., “[chinesisches]_ORGderiv Protektorat” is derived from the country [China]_ORG.
To determine the entity type, we examine alternative usages of the entity in the nominal base form. If a nominal paraphrase is a PER, then the derived word is PERderiv.
Here are more examples:

English example:

- My [French]_LOCderiv boyfriend

German examples:

- die [Drau]_LOC ([slowenisch]_LANGderiv [Drava]_LOC)
- der [Pole]_LOCderiv [Georg Franz Kolschitzky]_PER
- weiterhin lebten damals in [Slowenien]_LOC 1,98 % [Serben] _LOCderiv , 1,81 % [Kroaten] _LOCderiv , 1,1 % [Bosniaken] _LOCderiv

The subtype -part is used for words that only partly contain named entities,
This is especially frequent in German languages due to the long compounding structure.
For example, [deutschlandweit]_LOCpart.

German examples:

- [deutschsprachige]_LANGpart Restgruppe
- [Renaissance-Arkadenhof]_EVENTpart
- [Natura-2000-Vogelschutz-]_ORGpart und [FFH-Gebiet]_ORGpart , als [UNESCO-Biosphärenreservat]_ORGpart

Bavarian examples:

- [Chat-Arabischn]_LANGpart
- [spanischn]_LANGderiv
- [Mingara]_LOCderiv