DEFT ERE Annotation Guidelines: Entities V1.8

Linguistic Data Consortium
May 23, 2014

Changes from V1.7

- 2.3: Added “here” and “there” to pronominal list for ERE purposes.
- 3.1, 3.5: Added a note about generic PER mentions that reflect LOC names (“the west favors him”). Default to LOC entity type instead of PER if the entity is otherwise not taggable.
- 4.4: Added three more examples to show how tokens with multiple entities should be handled: [he/she], the [Brogan/Tremolo] wedding, and [USPS].
- 4.5: Added mention level subscript to entities in modifier position examples
- 5.3: Revised “general description” examples to remove attributive PER mentions that would be tagged as TTLs.
- 5.3: Expanded referentiality explanation to include “Japanese Americans” example
- 7.1: Updated unattributed discussion forum quotes section and added post author nicknames section.
- 7.2: Notes on allowable XML tagging in hyperlinks (between <a href=”URL”> and </a> markers).

Changes from V1.6

- 3.5: Specified that rooms or wings within a building are the lowest level location that we will annotate. Places or objects within a room are off limits.
- 3.1, 3.4: Added note about generic PER mentions that reflect GPE names (“Americans love fast food”) and added previous demonym example to this section (“The French enjoy wine”). Default to GPE entity type instead of PER if the entity is otherwise not taggable.
- 3.1, 3.3: Added note about generic PER mentions that reflect ORG names (“The Democrats are all the same”). Default to ORG entity type instead of PER if the entity is otherwise not taggable.
- 3.1: Added note about generic religious group/ethnicity mentions and explained not to tag or default to another entity type.
- 3.1: List of PER examples now includes “[65] deaths”
- 3.4: Added new examples to the demonym note.
- 4.4: Added an example with excessive coordinating conjunctions and how to split entities.
- 7.1: Expanded section to include information about quotes and end-post XML
- 2.3: Removed negative pronouns

Changes from V1.5

- 5.3: Further expanded section to illustrate the difference between asserted entity mentions and general descriptions.

Changes from V1.4

- Removed all language regarding GPE mentions being tagged as ORGs, with the exception of sports teams.

- Reorganized sections overall.
- 5.3: Added tree diagram.
- 4.4: Added rule about cases of multiple entities in a single, continuous string.

Changes from V1.3
- Added additional examples to many sections. Removed confusing examples throughout.

Changes from V1.2
- 2.6: Altered section to pertain to discussion forum data.
- 2.6.1: Added section to explain how to handle post metadata.

Changes from V1.1
- 2.2.3: Removed generic pronouns from “Pronominal Entity Mentions” list (anyone, anything, anybody)
- 2.3.3: Replaced the last example under ORG vs. GPE metonym note with a less ambiguous ORG mention (“the White House”)
- 2.4.4: Included examples of cases where only single entity tags can be annotated from phrases involving coordination.

Changes from V1.0
- 2.2.2: Noted that appositives and certain other NAM+NOM combinations expressing identity or categorization (including some usages of “of” which function to express identity or of which general class something is a particular example) should be annotated by tagging the NAM and NOM mentions autonomously.
- 2.2.2, 2.2.3: Specified that pronominal phrases will be tagged NOM.
- 2.2.3: Made pronoun list under “Pronominal Entity Mentions” comprehensive.
- 2.2.4, etc.: Adjusted “Tag for Meaning” rule to “Tag for Usage” to better reflect its intent, and emphasized that some entity types may not reflect their surface forms, and annotators must use their judgment to assign the appropriate entity type.
- 2.3.2: Specified that TTL entities will consist of personal titles and honorifics, official rank or status, and specific professional positions or occupations.
- 2.3.2: Specified that TTL strings consist only of the title, position, or profession itself, independent of its organizational circumstances, and excluding articles and modifiers.
- 2.3.3, 2.3.4: Updated ORG and GPE sections and examples to reflect decision to tag mentions of nation-states or other metonyms which clearly refer to a GPE’s government as ORGs, and to use GPE as the default when in doubt.
- 2.3.4: Noted language names should not be tagged as GPE mentions.
- 2.3.5: Included more examples of informal and ad hoc LOC mentions.
- 2.4.2: Adjusted language for tagging as LOC mentions of place names in common use but which do not refer to a region corresponding to a formal GPE.

- Former 2.4.6 “Compound and Multi-Word Titles” moved under 2.3.2 “Titles”
- 2.4.9: Included examples of text with all entities completely tagged.
- 2.5.3: Added positive examples of specific mentions for contrast with general/non-specific examples.
- 3.4: Added section to specify that mentions of earlier and later versions of an organization will be coreferenced of as the same ORG entity.

Table of Contents

1. Introduction

The purpose of this annotation project is to mark up texts for entities, coreference, events and relations. The primary purpose is for the annotations to describe the meaning of the text, as opposed to its syntactic or lexical aspects. The annotation is carried out level by level. This document describes the level of entity annotation and coreference.

An **entity** is a unique object or set of objects in the world – for instance, a specific person, place, or organization. A **mention** is a single occurrence of a name, nominal phrase, or pronominal phrase that refers to, or describes, a single entity. The **mention extent** is a string of text that we annotate to indicate the occurrence of an entity mention. In a later task (coreference) we cluster together multiple mentions of the same entity.

Entities may be referenced in text at three different **mention levels –** a name, a common noun or noun phrase, or a pronoun. For example, the following are all mentions of the same entity occurring at different levels:

- Name Mention (NAM): Barack Obama
- Nominal Mention (NOM): the incumbent
- Pronoun Mentions (PRO): he, his

In this task, we will label five **entity types**:

- Person (PER) - Person entities are limited to humans. A PER entity may be a single person or a group.
- Title (TTL) – Used for titles, roles, professions, and honorifics.
- Organization (ORG) - Organization entities are corporations, agencies, and other groups of people defined by an established organizational structure. An ORG entity may be a single organization or a group. **NOTE:** A key feature of an ORG is that it can change members without changing identity.
- Geopolitical Entity (GPE) - GPE entities are composite entities, consisting of a physical location, a government, and a population. All three of these elements must be present for an entity to be tagged as a GPE. A GPE entity may be a single geopolitical entity or a group.
- Location (LOC) - Location entities are geographical entities such as geographical areas and landmasses, bodies of water, and geological formations as well as buildings and other permanent human-made structures. A LOC entity may be a single location or a group.
- Vehicle (VEH) -
- Weapon (WEA) –
- Medium (MED) –
- Value (VAL) -

For our purposes, a “taggable” entity is one that is explicitly mentioned in the text, regardless of part of speech, and falls into one of the five types above. For all entities we label the text string constituting the entity mention, and assign an entity type. For entity types other than Title, we also indicate the mention level (NAM, NOM or PRO), see section below for details.

2. Tagging the Entity Mention, Determining its Level, and Labeling its Extent

NOTE: Throughout this document the extent of each entity mention is marked with [square brackets]. Counter-examples are marked with a ~~strikethrough~~. Entity types are indicated by subscript ($_{\text{PER, ORG, GPE, LOC, TTL}}$). For most sections, examples will only mark the entity type relevant to that section. Examples with multiple entity types can be found in section 2.4 and onward.

Your first task is to find each taggable entity mention in the document, and label its extent – that is, the string of text that refers to the entity. Mention extents generally begin and end at word (token) boundaries. However, possessive endings ('s) and verbal contractions (‘m, ‘ve, ‘re) should be <u>excluded</u> from the mention extent. As a rule, you should also exclude punctuation characters like commas, periods, and quotation marks unless the same entity mention continues after the punctuation mark. Special rules for entity extents apply depending on whether they are named, nominal, or pronominal mentions, so we will consider each mention level in turn.

Once you have determined the entity mention extent), you must label the entity mention level (NP type in the current tool) and entity type (section 2.3).

2.1 Tagging Named Entity Mentions

A named entity mention (NAM) is a mention that uniquely refers to an entity by its proper name, acronym, nickname, alias, abbreviations, or another alternate name. For our purposes, the extent of a name is the entire string representing the name, <u>excluding</u> the preceding definite article (‘the’) and any other pre-posed or post-posed modifiers. These are excluded because they are not part of the entity’s actual name (e.g. Bill Clinton’s name is ‘Bill Clinton’ not ‘former president Bill Clinton’).

NOTE: Mentions of entities with names referred to as “so-called” may also be tagged NAM.

Some examples of named entity mentions follow:

- [Bob Austin]$_{\text{PER}}$
- former president [George W. Bush]$_{\text{PER}}$
- the [Eiffel Tower]$_{\text{LOC}}$
- [IBM]$_{\text{ORG}}$
- the [Yankees]$_{\text{ORG}}$ (sports team)

- [Coca-Cola Bottling Co.]$_{\text{ORG}}$
- [Uganda]$_{\text{GPE}}$
- [Bowdon]$_{\text{GPE}}$, [Georgia]$_{\text{GPE}}$
- [Mt. Fuji]$_{\text{GPE}}$
- the [Kremlin]$_{\text{ORG}}$
- the [Kennedys]$_{\text{PER}}$
- [Bill]$_{\text{PER}}$ and [Hillary Clinton]$_{\text{PER}}$
- [Sean “Puffy” Combs]$_{\text{PER}}$, aka [Puff Daddy]$_{\text{PER}}$ or [P. Diddy]$_{\text{PER}}$
- the [North]$_{\text{GPE}}$ (for ‘North Korea’)
- the so-called [[Northern Cyprus]$_{\text{GPE}}$ Chess Federation]$_{\text{ORG}}$
- [Land of the Rising Sun]$_{\text{GPE}}$ (for ‘Japan’)
- the famous [Lincoln Memorial]$_{\text{LOC}}$
- the incomparable [Steven Spielberg]$_{\text{PER}}$
- the [US]$_{\text{GPE}}$ [State Department]$_{\text{ORG}}$
- the [States] $_{\text{GPE}}$ (as a nickname for the US)
- The [Gambia]$_{\text{GPE}}$

2.2 Tagging Nominal Entity Mentions

A nominal entity mention (NOM) is an entity mention not including the entity’s proper name, referring to it by common noun phrase. For our purposes, the extent of a nominal mention is the full mention of the noun or noun phrase, including articles and all pre-posed and post-posed modifiers. This is because modifiers provide information about an entity that could later be used by systems to identify they entity by name.

NOTE: Noun phrases beginning with pronominals (see section 2.2.3 below), like “this group”, “the other party”, “few of the attendees”, will be tagged as nominals.

NOTE: A good rule for identifying the extent of a nominal mention is that it is the extent of the text that would be replaced by a pronoun (e.g. ‘[the war-torn country] elected a new president’ the GPE mention extent can be replaced by a pronoun ‘[it] elected a new president’. Replacing part of the mention extent would not make sense ‘the war-torn ~~[it]~~$_{\text{GPE}}$’ elected a new president’).

Some examples of possible nominal mentions are given below:

- [a monument]$_{\text{LOC}}$
- [a few well-known monuments]$_{\text{LOC}}$
- [some teams]$_{\text{ORG}}$
- [the building]$_{\text{LOC}}$
- [that city]$_{\text{GPE}}$
- [her country]$_{\text{GPE}}$
- [the director of the Oscar winning film *Lincoln*]$_{\text{PER}}$

- [the family]$_{\text{PER}}$
- [another large company whose investors revolted]$_{\text{ORG}}$
- [the presidential hopeful from Chicago]$_{\text{PER}}$

NOTE: Appositives and certain other NAM+NOM combinations expressing identity or categorization should be tagged with care. Some usages of "of" which function to express identity or of which general class something is a particular example. Where possible, do not include nominal mention extents with named mention extents; tag them autonomously:

- [Reuters]$_{\text{NAM}}$ [international news agency]$_{\text{NOM}}$
- [his loudest critic]$_{\text{NOM}}$, [Jon Stewart]$_{\text{NAM}}$
- [my brother]$_{\text{NOM}}$ [John]$_{\text{NAM}}$
- the [Financial Accounting Standards Board]$_{\text{NAM}}$, [the private-sector body based in Norwalk, Conn., that sets the nation's accounting standards]$_{\text{NOM}}$
- [the informant]$_{\text{NOM}}$ called [Deep Throat)$_{\text{NAM}}$
- [the London borough]$_{\text{NOM}}$ of [Greenwich]$_{\text{NAM}}$
- [the city]$_{\text{NOM}}$ of [Denver]$_{\text{NAM}}$
- [Google employee]$_{\text{NOM}}$ [John Doe]$_{\text{NAM}}$

When it is not possible to tag NAM+NOM combinations autonomously (such as in cases where a named mention is embedded within a coreferential nominal mention), phrases are still tagged exhaustively:

- [the [Tamil Tigers]$_{\text{NAM}}$ separatist organization]$_{\text{NOM}}$
- [the now-defunct [G17 Plus]$_{\text{NAM}}$ political party]$_{\text{NOM}}$

2.3 Tagging Pronominal Entity Mentions

A pronominal entity mention (PRO) is a referring expression that corresponds to a nominal or a named entity. The extent of a pronominal mention is just the single referring unit. Below is a list of pronominal entity mentions (The referring expressions on this list will be tagged as PRO in this task.):

- all
- another
- any
- both
- each
- each other
- either
- everybody
- everyone
- everything

- few
- he
- her
- here
- hers
- herself
- him
- himself
- his
- I
- it
- its
- itself
- little
- many
- me
- mine
- more
- most
- much
- my
- myself
- one
- one another
- other
- others
- our
- ours
- ourselves
- several
- she
- some
- somebody
- someone
- something
- that
- their
- theirs
- them
- themselves
- there
- these
- they
- this

- those
- us
- we
- what
- whatever
- where
- wherever
- which
- whichever
- who
- whoever
- whom
- whomever
- whose
- you
- your
- yours
- yourself
- yourselves

Reflexive pronouns are marked in the same way as other pronouns, e.g.,

- John hurt [himself]$_{\text{PER}}$

NOTE: Noun phrases beginning with a pronoun listed above, like “this group”, “the other party”, “few of the attendees”, will be tagged as nominals.

Relative pronouns should only be tagged as entities if they are **not** part of a nominal entity mention. For example:

- [John Smith]$_{\text{PER}}$, [who]$_{\text{PER}}$ is a friend of mine, arrived late.
- I work at [the Black Cat]$_{\text{LOC}}$, [which]$_{\text{LOC}}$ is a small restaurant downtown.
- He saw [the students ~~[that]$_{\text{PER}}$~~ he would be teaching]$_{\text{PER}}$.

The possessive ending (‘*s*) and verbal endings (‘m, ‘re, ‘ve, etc.) should be excluded from the extent of named, nominal or pronominal mentions. For example:

- [David]$_{\text{PER}}$’s house
- the [buildings]$_{\text{LOC}}$’s entrance
- [I]$_{\text{PER}}$’ve never been there
- [She]$_{\text{PER}}$’s a smart girl

2.4 General Rules

The following general rules apply at all times:

1. We always tag an expression according to its usage in the context being evaluated. In other words, the annotation of an expression depends on how it is being used. We will call this rule **Tag for Usage**. For example, if we have the sentence '[Kansas] beat [Georgetown] last night' we tag 'Kansas' and 'Georgetown' as ORGs since they are referring to sports teams, even though superficially the strings appear to be referring to a GPE or LOC. See sec. 2.4.3.
2. Entities are tagged regardless of which syntactic function they fulfill. For example, entity mentions may be adjectives ("[Korean] cars"), possessive determiners ("[her] three convictions"), or prepositional complements ("at [the beach]").
3. We allow overlapping or embedded annotations in cases where a modifier of an entity itself refers to a taggable entity. We will call this the **Modifier Referent Rule**. For example, the expression 'the Pakistani people' would have two entity mentions tagged: '[the [Pakistani]$_{\text{GPE}}$ people]$_{\text{PER}}$' and 'his army' would be tagged as '[[his]$_{\text{PER}}$ army]$_{\text{PER}}$'. See sec. 2.4.5 for more.

3. Labeling the Entity Type

Once you have determined and input the entity mention extent, in addition to tagging the entity mention level, you must label the entity type. In this task, we will label X entity types: person (PER), titles (TTL), organizations (ORG), geo-political entities (GPE), locations (LOC), vehicles (VEH), weapons (WEA), medium (MED), and value (VAL). A description of each type follows.

3.1 Person Entities (PER)

NOTE: For examples in this section, only PER entities are labeled with [square brackets].

Person entities are limited to individual humans or groups of humans identified by a simple referring expression (PER.NOM), a name/nickname/alias (PER.NAM), or pronoun (PER.PRO).

If a group of people meets the definition of an ORG or GPE it should be tagged as such. Otherwise, the group should be tagged as PER. By this standard, family names and ethnic and religious groups that lack formal organizational backing are tagged as PER entities.

NOTE: For entities such as movements (e.g. 'Occupy Wall Street', 'the Tea Party', 'rebel movements') which encompass gray areas regarding existence of formal name and structure, use your best judgment as to whether to tag them as ORG or a PER-group. We should usually default to making fewer assumptions and use the less-specific, more conservative entity type (PER).

NOTE: Generic PER mentions that reflect GPE names (such as “Americans” in “Americans love fast food”) should not be annotated as the PER entity type. They should be tagged as GPE entities. We should default to GPE when a PER mention would be generic and otherwise not tagged.

NOTE: Generic PER mentions that reflect ORG names (such as “Democrats” in “The Democrats are all the same”) should not be annotated as the PER entity type. They should be tagged as ORG entities. We should default to ORG when a PER mention would be generic and otherwise not tagged.

NOTE: Generic PER mentions that reflect LOC names (such as “the west” in “the west still favors him” or “the world” in “the world just never accepted her for who she was”) should not be annotated as the PER entity type. They should be tagged as LOC entities. We should default to LOC when a PER mention would be generic and otherwise not tagged.

Fictional characters, religious deities, and non-human characters should not be tagged as PER entities. However, deceased people may be tagged as PER entities (though phrases such as “corpse” or “dead bodies” are not tagged as PER entities). Some examples of PER entities are given below. Recall that counter-examples are given in strikethrough.

- [Bill Clinton]
- [Analysts] told the Guardian that…
- [Judy Garland]
- [Adam West]’s costume from the [~~Batman~~] TV series
- [~~Harry Potter~~]
- [GOP hopeful]
- the [Cartwrights]
- [the squad of [Marines]]
- [the family]
- [the house painters]
- [the Christians]
- [the linguists under the table]
- [the Arabs]
- [[Her] friend] was [[a provincial doctor]’s wife].
- [I]’ve read [~~Jane Eyre~~] 7 times.
- [~~Seinfeld~~]was [my] favorite show.
- [We] have eradicated terrorism from [[our] society].
- He has reported on [65] deaths in the last eight months.

NOTE: If a document refers to a religious group or ethnicity in a generic fashion, we do not tag an entity, nor do we default to an ORG or GPE entity type.

- [~~Catholics~~] talk about the Bible like it's the best book ever written.
- They kept talking about how [~~Arabs~~] all eat really well.

You may occasionally encounter an ordinal suffix like 'Jr.', 'Sr.', or 'IV'. These are considered part of a person name and should be included within the mention extent. However, Titles (including honorifics) should NOT be included in a Person entity mention extent for instance:

- Pope [Benedict XVI]
- Mr. [Albert Franklin, Jr.] was on [the research team]

3.2 Title Entities (TTL)

NOTE: For examples in this section, many entity types are labeled with [square brackets] to illustrate the relationship among the types.

Personal titles and honorifics, official rank or status, and specific employed occupations or professional positions will be labeled as TTLs when they are used to establish a particular relationship between a PER mention and the title, honorific, position, or occupation, both in the same sentence. Annotators will need to use best judgment when determining whether a position or occupation is specific enough to be tagged as a title (e.g., general types such as worker, official, member, employee, will not be tagged TTL).

Because TTLs occur in conjunction with other entities and refer to other entities, they have some special rules. Unlike other entity types, titles do not have a mention level (NAM, NOM or PRO). Title elements are often but not always adjacent to a PER entity, which can be of mention level NAM, NOM or PRO. The most frequent constructions TTLs occur in are title+name, appositives, and copula constructions.

Just as with appositives, all titles, positions, and honorifics will be marked separately from the individual's name. For instance, in the following sentences, there are two separate entities marked:

- [Vice President]$_{\text{TTL}}$ [Biden]$_{\text{PER}}$
- [spokesperson]$_{\text{TTL}}$ [Mary Gillette] $_{\text{PER}}$
- [Michelle Obama] $_{\text{PER}}$, [First Lady] $_{\text{TTL}}$, spoke at the event.
- [Michelle Obama] $_{\text{PER}}$, [who]$_{\text{PER}}$ is the [First Lady] $_{\text{TTL}}$, spoke at the event.
- [She]$_{\text{PER}}$ is the [First Lady] $_{\text{TTL}}$.
- The famous Greek [philosopher]$_{\text{TTL}}$ [Zeno]$_{\text{PER}}$ was born in 490 BC.

Sometimes the name of the person is split into two pieces by the title. In these cases, we will annotate a nested title.

- [Alfred [Lord]$_{\text{TTL}}$ Tennyson]$_{\text{PER}}$

For our purposes, the extent of a TTL is limited to the string of text that describes the position or rank itself, independent of its organizational circumstances or relationships, and excluding a preceding definite article ('the') and any other pre-posed or post-posed modifiers. TTLs are often mentioned in tandem with the organization (ORG) or geo-political (GPE) entity to which they pertain, e.g.: "Mayor of Boston", "Vice President of Marketing", "the Dover Rotary Club's treasurer", "the commissioner of the National Basketball Association", etc. Thus, if the phrase containing the position title itself also contains another taggable entity type, be sure to tag each entity separately (TTL-ORG relationships will be tagged in the Relations level of annotation). Examples:

- [Treasury]$_{\text{ORG}}$ [Secretary]$_{\text{TTL}}$ [Jackson]$_{\text{PER}}$
- [Justice]$_{\text{ORG}}$ [Minister]$_{\text{TTL}}$ [Giovannia Maria Flick]$_{\text{PER}}$
- [Mission Control]$_{\text{ORG}}$ [Chief]$_{\text{TTL}}$ [Vladimir Solovyov]$_{\text{PER}}$
- [[US]$_{\text{GPE}}$ Army]$_{\text{ORG}}$ [negotiator]$_{\text{TTL}}$ [Harold Norman]$_{\text{PER}}$
- [US]$_{\text{GPE}}$ [Secretary]$_{\text{TTL}}$ of [State]$_{\text{ORG}}$ [Hillary Clinton]$_{\text{PER}}$

Compound and multi-word titles may be tagged so that their mention extents include all words essential to the understanding of the title or position itself, excluding an adjacent organization or GPE mention, e.g.:

- As [Assistant Deputy Sheriff]$_{\text{TTL}}$, it is Bob's duty to inquire into certain deaths.
- Japan's parliament elected him [prime minister]$_{\text{TTL}}$.
- Anne became the [heir apparent]$_{\text{TTL}}$ for the remainder of William's reign.
- White House [Chief of Staff]$_{\text{TTL}}$ Jack Lew resigned last year.
- former [President]$_{\text{TTL}}$ Bill Clinton

The only exception to excluding an ORG or GPE mention from a title mention extent is when one is nested within it:

- [Deputy [Foreign]$_{\text{ORG}}$ Minister]$_{\text{TTL}}$.

In some cases, more than one title will be presented for a single person. In case all of these titles are adjacent to the person name, we will tag each of the titles separately as TTL:

- [Karachi]$_{\text{GPE}}$ [Mayor]$_{\text{TTL}}$ and [GlobalCom]$_{\text{ORG}}$ [Chairman]$_{\text{TTL}}$ [Mr]$_{\text{TTL}}$ [Anwar Ayub]$_{\text{PER}}$

NOTE: When a string that may serve as a title or position is being used as a nominal to refer to a person who is not named in relation to it within the sentence, the entity mention should be tagged as PER and not as TTL. For instance, consider these two contrasting cases:

- [Joe Biden]$_{\text{PER}}$ is [Vice President]$_{\text{TTL}}$
- [The strongest supporter]$_{\text{PER}}$ was [the vice president]$_{\text{PER}}$.

NOTE: Phrases that establish ad hoc, informal, or general roles—as opposed to more specific, formal positions, occupation, titles, and honorifics—should not be tagged as titles:

- Suzie, the ~~[brain]$_{\text{TTL}}$~~ in [the family]$_{\text{PER}}$., explained how the machine worked.
- ~~[Class clown]$_{\text{TTL}}$~~ John made a joke.
- Amy was always the ~~[philosopher]$_{\text{TTL}}$~~ in our discussions.

3.3 Organization Entities (ORG)

NOTE: For examples in this section, only ORG entities are labeled with [square brackets].

Organization entities are groups of people defined by an established organizational structure, identified by a simple referring expression (ORG.NOM), a named expression (ORG.NAM), or a pronoun (ORG.PRO).

NOTE: Sets of people who are not formally organized into a unit should be treated as a PER entity rather than an ORG entity. This distinction can sometimes be difficult. If in doubt, label the group as PER instead of ORG. Some examples of entities that should be treated as PER entities instead of ORG entities are:

- [the delegation]$_{\text{PER}}$
- [Occupy Wall Street]$_{\text{PER}}$
- [Police]$_{\text{PER}}$ arrested [the group of rebels]$_{\text{PER}}$

NOTE: Organizations that share their name with a publication (whether printed or digital) should only be tagged as ORGs when it's clear that the organization is being referred to, not the publication. Publications are not, themselves, considered organizations. For instance:

- The [New York Times]$_{\text{ORG}}$ announced that it has named a new CEO.
- Bob enjoys reading the ~~[New York Times]$_{\text{ORG}}$~~ on Sunday.
- [Facebook]$_{\text{ORG}}$ is headquartered in Menlo Park, CA.
- i saw on ~~[facebook]$_{\text{ORG}}$~~ there was something on the bbc saying the earth had exploded

NOTE: Generic PER mentions that reflect ORG names should not be annotated as the

PER entity type. They should be tagged as GPE entities. We should default to ORG when a PER mention would be generic and otherwise not tagged.

- The [Democrats]$_{ORG}$ are all the same.
- The [Republicans]$_{ORG}$ supported Ryan.

Organizations include the following subtypes: Governmental; Commercial, Educational, Scientific, Medical; Media; Religious, Social, Advocacy; and Sports. Though we will not be labeling these subtypes explicitly, it is useful to consider examples of them:

Governmental (includes Political, Quasi-Governmental, Military, and Para-Military Groups)

- [Republican Party]
- [Labour Party]
- the [Socialist People's Party]
- [Republican National Committee]
- [ACLU]
- The [Cato Institute]
- [NATO]
- The [World Bank]
- three of the [U.N.] workers stationed in East Timor
- [International Monetary Fund] aid
- [Hizbollah]
- [Islamic Resistance]
- [Rally for Congolese Democracy]
- [Institutional Revolutionary Party]
- the [KKK]
- [Al Aqsa Martyr's Brigade]
- [Tamil Tigers]
- the [Caravan of Death], [a military party that killed 73 political prisoners]
- the leading deputy of the [Rally for Congolese Democracy], [one of the biggest rebel movements supported by Uganda]
- [The Salzburg prosecutor's office] is investigating the disaster to determine if criminal charges could be filed.
- Putin, a former [KGB] agent, defended [the court that convicted Pope and [the security services]].
- The [Financial Accounting Standards Board] will take no conclusive action on [its] current project on business combinations until [Congress] has reconvened in 2001.
- [The US navy] now says the USS Cole was being refueled when an explosion ripped through it in Yemen last week, killing 17.

Commercial

- the [Roundabout Theater Company] is calling [its] new facility in Times
- [Pixar], [the award-winning animation company]
- the [American Airlines Theater]
- Pope, who owns [TechSource Marine Industries] in State College
- Like [the famous Irish group] the [Chieftains], [Solas] frequently headlines in Celtic festivals.

Educational, Scientific, Medical

- [George Washington University]
- [Overseas Chinese Physics Institute]
- [Gulf Coast Research Laboratory]
- [A coalition of medical and health groups from [Massachusetts General Hospital]]
- Pope had worked for the [Applied Research Laboratory] at [Pennsylvania State University].
- [NDSU] and [University of Minnesota] weeds specialist Alan Dexter says 98% of the plants survived.

Media

- [Agence France Presse]
- [abc news]
- [Associated Press]
- [Chicago Sun-Times]
- [National Geographic]

Religious, Social, Advocacy

- [German Bishops Conference]
- [Rock the Vote]
- [American Medical Association]
- [American Council on Education]
- [National Rifle Association]
- [American Diabetic Association]
- [NAACP]
- [American Bar Association]
- [National Center for Public Policy and High Education]
- The [Red Cross] said about 15 people managed to escape…

Sports

- [Taekwondo Association]
- [Philippines Olympic Committee]
- [national hockey league]
- [San Francisco 49ers]
- A group of survivors belonging to [a German ski club in Vilseck, Germany]

3.4 Geopolitical Entities (GPE)

NOTE: For examples in this section, unless specified, all entity types labeled with [square brackets] are GPE.

Geo-Political Entities are nations or subordinate types of politically-defined territory such as provinces, states, counties, cities, etc.). For something to be taggable as a GPE, it must consist of three elements: political organization, population, and physical territory. Note that sometimes a GPE mention may appear to refer more strictly to the physical location, but in such cases we still tag it as a GPE—for example:

- We went to [France]$_{\text{GPE}}$ for our vacation.
- They delivered the supplies to [Pakistan]$_{\text{GPE}}$

Sometimes the context makes it appear that the mention of the geo-political unit, the capital, or government location is referring specifically to the government itself. In these cases we still tag the mention as a GPE. For instance:

- [Iraq]$_{\text{GPE}}$ signed a treaty with [Kuwait]$_{\text{GPE}}$.
- [Washington]$_{\text{GPE}}$ discussed economic policies with [Moscow]$_{\text{GPE}}$ at the summit.
- [The government of [France]$_{\text{GPE}}$]$_{\text{ORG}}$ welcomed the agreement.
- [India]$_{\text{GPE}}$ is interested in strengthening economic ties with the [US]$_{\text{GPE}}$.
- The Premier said [China]$_{\text{GPE}}$ would continue on a path of economic liberalization.
- [Turkey]$_{\text{GPE}}$ regards [Northern Cyprus]$_{\text{GPE}}$ as a sovereign country.

GPE entities can be single GPEs or groups of GPEs, for example:

- I visited [Britain]$_{\text{GPE}}$, [France]$_{\text{GPE}}$, and [Germany]$_{\text{GPE}}$ last summer. I had a great time visiting [these countries]$_{\text{GPE}}$.

NOTE: When a GPE name or demonym or adjectival GPE name is used to refer to the *people* of a GPE (and the mention is specific), it should be tagged as a PER entity. For example:

- The [Swiss]$_{\text{PER}}$ have joined us on the bus tour.

- Luckily, the [Australians]$_{PER}$ made it to the barbeque on time.

NOTE: Generic PER mentions that reflect GPE names should not be annotated as the PER entity type. They should be tagged as named GPE entities. We should default to GPE when a PER mention would be generic and otherwise not tagged. For example:

- [Americans]$_{GPE}$ love fast food.
- The [French]$_{GPE}$ enjoy wine.
- [Canadians]$_{GPE}$ appreciate hockey.

NOTE: Use caution with languages. Generally names of languages are not taggable as GPE mentions:

- ~~French~~ is spoken in much of Africa.
- All nations of the former [Yugoslavia]$_{GPE}$ have ~~Serbian~~-speaking regions.
- ~~Arabic~~ is a major international language.
- The most widespread [Indian]$_{GPE}$ languages are ~~Hindi, Marathi, Tamil, Urdu, Bengali~~, and ~~Telugu~~.

NOTE: Sometimes the names of GPE entities may be used to refer to other things associated with a region besides the government, people, or aggregate contents of the region. The most common examples are sports teams:

- [New York]$_{ORG}$ defeated [Boston]$_{ORG}$ 99-97 in overtime.

As always, we Tag for Usage. So in the example above, both 'New York' and 'Boston' are ORG entities. Note however, that GPE names nested within sports team names should still be tagged as GPEs:

- The [[Philadelphia]$_{GPE}$ Eagles]$_{ORG}$

Additional examples of GPEs include the following. In the examples below, only GPE entities are enclosed in [square brackets].

- Hospital officials said all eight survivors were [German]$_{GPE}$.
- the conversion to Christianity of the [Roman]$_{GPE}$ emperor Constantine
- [Salzburg] governor Schausberger said…
- Recounts are only just beginning in [Palm Beach]$_{GPE}$ and [Volusia counties]$_{GPE}$.
- The economic boom is providing new opportunities for women in [New Delhi]$_{GPE}$.
- …said Norbert Karlsboeck, mayor of [Kaprun]$_{GPE}$, [a town some 50 miles south of [Salzburg]$_{GPE}$ in the central [Austrian]$_{GPE}$ Alps]$_{GPE}$
- [France]$_{GPE}$'s greatest national treasure

- [France]$_{\text{GPE}}$ produces better wine than [New Jersey]$_{\text{GPE}}$.
- [Israeli]$_{\text{GPE}}$ troops
- The [Palestinian]$_{\text{GPE}}$ Authority has banned rallies that are pro-[Iraq]$_{\text{GPE}}$

NOTE: Countries of countries, such as ‘the [European Union]$_{\text{GPE}}$’ and ‘the [United Kingdom]$_{\text{GPE}}$’ will be annotated as GPEs, since they have all three GPE components (i.e., a population, a government, and a location).

The same formula applies to contested areas like Taiwan; see also sec. 2.4.8 below.

3.5 Location Entities (LOC)

NOTE: For examples in this section, only LOC entities are labeled with [square brackets].

Location entities are geographically or astronomically defined places that do not have a political component, natural structures like bodies of water and mountains, and human-made structures like airports, factories, and streets. Locations are identified by a simple referring expression (LOC.NOM), a named expression (LOC.NAM), or a pronoun (LOC.PRO).

Examples of place-related strings that are tagged as LOC include heavenly bodies, continents, non-politically-defined regions, airports, highways, street names, factories, cafes, manufacturing plants, street addresses, oceans, seas, straits, bays, channels, sounds, rivers, islands, lakes, national parks, mountains, and monumental structures, such as the Eiffel Tower and Washington Monument. Fictional or mythical locations should not be tagged. Also, rooms or wings within a building are the lowest level of granularity that we will annotate. Objects or places within a room should not be tagged. For instance:

- repairs began on the [Alaskan Pipeline]
- [his driveway] was being refinished
- ~~[The wall]~~ and ~~[coffee table]~~ in [the living room]
- Vice President Cheney visited [the site].
- In Armenia, the three of them will join other, similar delegations from around [the world]...
- The droids landed on ~~[Tatooine]$_{\text{LOC}}$~~
- ... eclipse fans are being warned not to look directly at [the sun] …
- the [Missouri River]
- the collapse of the newly-constructed [Teton Dam]
- [the region where the movement has found most success recently]
- Police are asking everyone to avoid [the affected area].
- Many people in [North America] saw a partial solar eclipse yesterday.

NOTE: Generic PER mentions that reflect LOC names (such as “the west” in “the west still favors him”) should not be annotated as the PER entity type. They should be tagged as LOC entities. We should default to LOC when a PER mention would be generic and otherwise not tagged. Other examples include:

- [Europeans]$_{LOC}$ travel all summer.
- [the world]$_{LOC}$ just never accepted her for who she was

4. Difficult Cases and Interactions Among Entity Types

4.1 Determiners and Mention Span

The general rule is that determiners are included with nominal mention extents, but not with named mention extents. Determiners are included in the annotation of nominal entities that contain a named entity, as in the following example.

- [a [Gulshan Hotel]$_{LOC}$ driveway]$_{LOC}$
- [the [Smith]$_{PER}$’s house]$_{LOC}$

This nesting is particularly common when a NAM entity is adjacent to a NOM entity over which the article has scope.

4.2 The Extent of LOC and GPE mentions

There are several issues surrounding the expression of LOC and GPE entities and which parts of a string to tag.

LOC or GPE compound expressions in which place names are typically separated by a comma in English should be tagged as separate entities.

- [Kaohsiung]$_{GPE}$, [Taiwan]$_{GPE}$
- [Ford’s Theater]$_{LOC}$, [Washington D.C.]$_{GPE}$

When a "designator" is customarily used as a regular part of a place name, that word should also be included in the extent of the entity. For example, include in the tagged string the word ‘River’ in the name of a river, ‘Mountain’ in the name of a mountain, ‘City’ in the name of a city, etc., if such words are contained in the string.

- [Mississippi River]$_{LOC}$
- the [Himalayan Mountains]$_{LOC}$
- [New York City]$_{GPE}$

Often times place names are modified by words like ‘Southern’, ‘Lower’, ‘West’, ‘the former’ and so on. When these modifiers are part of a location's official name they should be tagged as part of the name. For instance:

- [Upper Volta]GPE
- [North Dakota]GPE

A place name in common use but which does not refer to a region corresponding to a formal GPE should be tagged as a Named location:

- the [Middle East]LOC
- the [West Bank]LOC
- [Eastern [Europe]LOC]LOC
- [Siberia]LOC

Place names may present difficulties. If you are not sure whether a modifier is part of an official name, you should include the modifier as part of the place name.

Names of regions within GPEs should be tagged as Nominal locations, and the GPE within them should be tagged as well

- the [western [United States]GPE]LOC
- [southwest [Germany]GPE]LOC

4.3 Entity Types and Tag for Usage

It often happens that the name of one entity is used to refer to another entity. You may also encounter multiple mentions of the same entity that invoke different entity types. Surface forms and meanings may belie actual usage for some entities, so you will need use your judgment in assigning the appropriate entity type—always Tag for Usage, as in the examples below.

- [Wouters]PER, 42, died an hour later at **[St. John Macomb Hospital]LOC**
- [The suspect]PER died later the same night, **[hospital]ORG** [spokeswoman]TTL [Rebecca O'Grady]PER said Thursday.
- **[America]ORG** brought home the gold. (sports team)
- Secretary of Defense William S. Cohen said today that he is satisfied **[Beijing]GPE** will not continue sales of anti-ship missiles
- **[The Guggenheim Museum]ORG** announced a new acquisition
- **[The Guggenheim Museum]LOC** was designed by [Wright]PER
- **[Deep Throat]PER** was the pseudonym given to [the secret informant to [Woodward]PER and [Bernstein]PER]PER.
- [He]PER flew into **[JFK]LOC** yesterday.

Notice we may need to ignore references to certain entity types within a mention in order to tag the string's basic usage in context. E.g., while in "The Armenians said…", "Armenians" means "persons who are citizens of the nation of Armenia", it will only be tagged as PER, and not GPE, because it is being used as a PER entity, and we wish to avoid multiple tags of one string.

4.4 Expressions that refer to multiple entities

Care is needed when dealing with coordination in entities. When a phrase refers to multiple, coordinated entities, mark each entity separately where possible. For instance:

- [China]$_{\text{GPE}}$ and [South Korea]$_{\text{GPE}}$ signed the agreement.
- [Jimmy]$_{\text{PER}}$ and [Rosalyn Carter]$_{\text{PER}}$
- [North]$_{\text{LOC}}$ and [South America]$_{\text{LOC}}$

But be careful not to split apart proper names that contain a conjunction. For instance:

- [Trinidad and Tobago]$_{\text{GPE}}$
- the [Fish and Wildlife Service]$_{\text{ORG}}$

The latter example is the name of one organization and should be tagged as a single named entity (it's not 'the Fish Service' and 'the Wildlife Service' as separate names).

When conjunctions are used excessively (mainly when referring to nominal entities), you may split apart the entities at the coordinating conjunctions. For example:

- Instead of tagging: [[my] stepkids and friends and family]
- We can tag: [[my] stepkids] and [friends]and [family]

Some cases of coordination may necessitate a phrase being tagged as a single entity, such as in cases where only a single noun is present but coordinated modifiers might suggest two distinct entities. For instance:

- [American and Canadian soldiers]$_{\text{PER}}$
- [the East and South of Iran]$_{\text{LOC}}$
- [the CEOs of Google and Youtube]$_{\text{PER}}$

Cases where multiple entities are joined together in a single, continuous string are also tagged as a single mention:

- they crossed the [Af-Pak] border
- [Brad&Angelina]
- [me+you]
- [he/she]
- the [Brogan/Tremolo] wedding

- [USPS] - we're unable to tag [[US]$_{\text{gpe}}$PS]$_{\text{org}}$ because we cannot split the token
- [Obamacare] – we'd tag this as a named PER because "Obamacare" wouldn't be an otherwise taggable entity. Obama's name is within the token and we should capture this information.

4.5 Entities in modifier position

If an entity mention contains another taggable mention nested within it, these nested entities should also be tagged. This applies both to named and nominal entity mentions, for example:

- [the [Clinton]$_{\text{PER.NAM}}$ government]$_{\text{ORG.NOM}}$
- [[Treasury]$_{\text{ORG.NAM}}$ employees]$_{\text{PER.NOM}}$
- [[U.S.]$_{\text{GPE.NAM}}$ exporters]$_{\text{ORG.NOM}}$
- [[Texas]$_{\text{GPE.NAM}}$ schools]$_{\text{ORG.NOM}}$
- [Kentucky]$_{\text{GPE.NAM}}$ Fried Chicken]$_{\text{ORG.NAM}}$
- [[government]$_{\text{ORG.NOM}}$ offices]$_{\text{ORG.NOM}}$
- [[Kurdistan]$_{\text{GPE.NAM}}$ Freedom Fighters]$_{\text{ORG.NAM}}$
- [the [Midwestern]$_{\text{LOC.NOM}}$ bank]$_{\text{ORG.NOM}}$
- [the [Russian]$_{\text{GPE.NAM}}$ foreign minister]$_{\text{PER.NOM}}$
- [the [American]$_{\text{GPE.NAM}}$ companies]$_{\text{ORG.NOM}}$
- [[Israeli]$_{\text{GPE.NAM}}$ troops]$_{\text{PER.NOM}}$
- [[Republican]$_{\text{ORG.NAM}}$ voters]$_{\text{PER.NOM}}$
- [[airline]$_{\text{ORG.NOM}}$ regulators]$_{\text{PER.NOM}}$
- [[Chrysler]$_{\text{ORG.NAM}}$ factories]$_{\text{LOC.NOM}}$
- [[union]$_{\text{ORG.NOM}}$ leaders]$_{\text{PER.NOM}}$
- The [[Chinese]$_{\text{GPE.NAM}}$ military]$_{\text{ORG.NOM}}$

4.6 Possessives

When you encounter a possessive construction, it may contain two taggable entity mentions. Note that when the construction is comprised of two named mentions, such as in the third example below, the two entities are tagged separately (i.e. the possessive entity is not embedded).

- [[Temple University]$_{\text{ORG.NAM}}$'s graduate school of business]$_{\text{ORG.NOM}}$
- [[Canada]$_{\text{GPE.NAM}}$'s parliament]$_{\text{ORG.NOM}}$
- [Singapore]$_{\text{GPE.NAM}}$'s [Central Narcotics Bureau]$_{\text{ORG.NAM}}$

4.7 Hyphenated pre-modifiers

Taggable entities that are part of a pre-modifying hyphenated construction should be tagged separately, for example:

- The [GOP]$_{\text{ORG}}$-backed candidates toured the area.

4.8 Places of contention

Places of contention can be tagged as GPEs long as they have all three components of a GPE (i.e. GPE = population + location + government). If a place of contention does not have all three of these components, it should be tagged as a LOC instead.

Using this rule, ‘Palestine’ is tagged as a GPE because it has all three GPE components, while ‘Gaza strip’ is tagged as a LOC, because though it has a population and a location, it doesn't have its own government.

4.9 Examples with entities completely annotated

- Videos circulated by [Osama bin Laden]$_{\text{PER.NAM}}$ have added to the evidence linking [him]$_{\text{PER.PRO}}$ and [the [al-Quaida]$_{\text{ORG.NAM}}$ network]$_{\text{PER.NOM}}$ to the Sept. 11 terrorist attacks in the [United States]$_{\text{GPE.NAM}}$, [the government]$_{\text{ORG.NOM}}$ said Wednesday in an updated dossier on the investigation.
- [Guzman]$_{\text{PER.NAM}}$ indicted [Pinochet]$_{\text{PER.NAM}}$, holding [him]$_{\text{PER.PRO}}$ responsible for the actions by the “[Caravan of Death]$_{\text{PER.NAM}}$”, [a military party that killed [73 political prisoners]$_{\text{PER.NOM}}$ shortly after the 1973 coup in which [Pinochet]$_{\text{PER.NAM}}$ ousted Marxist [President]$_{\text{TTL}}$ [Salvador Allende]$_{\text{PER.NAM}}$]$_{\text{PER.NOM}}$.
- Midway through the hearing, [Chief Justice]$_{\text{TTL}}$ [Renquist]$_{\text{PER.NAM}}$ seemed to scold [[his]$_{\text{PER.PRO}}$ colleagues]$_{\text{PER.NOM}}$ for being too talkative when [he]$_{\text{PER.PRO}}$ made an unusual offer to [the lawyer representing [[Florida]$_{\text{GPE.NAM}}$'s Attorney General]$_{\text{PER.NOM}}$]$_{\text{PER.NOM}}$.
- [Actors and singers also on the flight]$_{\text{PER.NOM}}$ held a benefit concert in [Baghdad]$_{\text{GPE.NAM}}$ Saturday evening, with most of the $13 cover charge to be donated to support the [Palestinian]$_{\text{GPE.NAM}}$ uprising.
- ...said [Archbishop]$_{\text{TTL}}$ [Khajag Barasamian]$_{\text{PER.NAM}}$, [head]$_{\text{TTL}}$ of the [Diocese of the [[Armenian]$_{\text{GPE.NAM}}$ Church]$_{\text{ORG.NAM}}$ in [America]$_{\text{GPE.NAM}}$]$_{\text{ORG.NAM}}$, [[whose]$_{\text{ORG.PRO}}$ headquarters]$_{\text{LOC.NOM}}$ are in [Manhattan]$_{\text{GPE.NAM}}$.

5. What NOT to tag

5.1 Events

Do not tag event names even if they refer to events that occur on a regular basis and are associated with institutional structures. However, the institutional structures themselves —steering committees, etc. —should be tagged.

- the ~~Pan-American Games~~
- the [Olympic Committee]$_{\text{ORG}}$

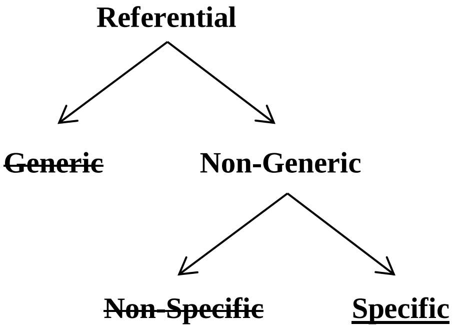
5.2 Artifacts and Products

Miscellaneous types of proper names that are not to be tagged as named entities include artifacts, other products, and plural names that do not identify a single, unique entity. For instance:

- the ~~Taurus~~ is the latest car model

5.3 Generic and Non-Specific Entities

Entity mentions that are generic or non-specific should also not be tagged. In ERE, only specific entities are annotated.



Do not tag mentions which do not refer to a particular, asserted entity which is understood to be real by the immediate writer, speaker or 'voice' in the text. Consider the following cases:

- I'd like to find ~~[a good dentist]~~$_{\text{PER}}$.
- I've got [a good dentist]$_{\text{PER}}$.
- Have you ever been to ~~[a trustworthy car dealership]~~$_{\text{ORG}}$?
- I once found [a trustworthy car dealership]$_{\text{ORG}}$, but it closed.
- ~~[More and more people]~~ are losing jobs to outsourcing.

Names used to refer to a generic category of things and do not refer to a specific entity should not be tagged

- the ~~Campbell Soups~~ of the world

Certain constructions can make more difficult the decision of whether or not to categorize an entity mention as specific. Sometimes a phrase may seem to be more a general description of an entity than an actual mention of an entity. Overly general descriptions are considered underspecified and are not tagged. As is always the case in ERE, tag for meaning/usage. If you think something is only a description and not a mention, don't tag it. Consider the following:

- [John Jacob Marshall]$_{\text{PER.NAM}}$, [the friend who drove me home last night]$_{\text{PER.NOM}}$
- [The guy you saw yesterday]$_{\text{PER.NOM}}$ is[Jerry]$_{\text{PER.NAM}}$
- [Dennis R. Beresford]$_{\text{PER.NAM}}$ is [that outspoken backgammon enthusiast]$_{\text{PER.NOM}}$
- [He]$_{\text{PER.PRO}}$ is ~~[a pain in the neck]~~
- [He]$_{\text{PER.NAM}}$ is ~~[my hero]~~
- [John]$_{\text{PER.NAM}}$ is a [plumber]$_{\text{TTL}}$
- [My friends]$_{\text{PER.NOM}}$ are ~~[great people]~~
- [She]$_{\text{PER.PRO}}$ is the [First Lady]$_{\text{TTL}}$
- [France]$_{\text{GPE.NAM}}$ is ~~[a nice place to vacation]~~

Similarly, phrases describing an entity's role (other than a Title) in some capacity are generally seen as descriptions, not coreferent entity mentions:

- [Atlanta]$_{\text{GPE.NAM}}$ is ~~[a major hub between the Northeast and the Caribbean]~~
- [Bob]$_{\text{PER.NAM}}$ is usually ~~[the one who picks up lunch for everyone]~~

The distinction between specific and generic/non-specific entities can be difficult to make, and you will need to use context to make this distinction. E.g.: 'analysts are interesting people' is a generic reference to a category of people in general, while 'analysts told the Guardian that ..' is a reference to specific people. Essentially, if you as a reasonable reader would interpret the entity as specific, go ahead and annotate it. If you are unsure as to whether something is specific or generic, leave it unannotated.

Fully annotated example with explanation:

- The [US]$_{\text{GPE.NAM}}$ confined [[Japanese]$_{\text{GPE.NAM}}$ Americans]$_{\text{PER.NAM}}$ during the war but did not do so with [Italian]$_{\text{GPE.NAM}}$ [Americans]$_{\text{GPE.NAM}}$ and [German]$_{\text{GPE.NAM}}$ [Americans]$_{\text{GPE.NAM}}$.

    - The mentions of "Italian Americans" and "German Americans" are *non-generic* in this example, but they're also *non-specific*. Because of this distinction, we do not tag the mentions as PERs. We default to GPE entities when the PER meaning is non-specific or generic. So, because "Italian America" and "German America" (arguably) aren't valid GPEs, we tag each word as separate GPE entities.

    - Conversely, "Japanese Americans" is *non-generic* as well as *specific*, because the speaker is referring a subset of Japanese Americans that were detained. In this case, we'd tag the string as a PER with Japanese nested within.

    - "Japanese Americans" would also be a Person ARG for a Justice.Arrest-Jail event. As a general rule, we don't change entity types just to grab a

Relation or Event argument. However, thinking of entities in terms of Relations and Events can help us decide how to annotate a mention.

6. Coreference

6.1 General Instructions for Coreference Tagging

Once all taggable entities have been labeled, the next task is entity coreference. During coreference labeling we cluster together multiple mentions of the same entity within the document. If two or more mentions refer to the same underlying entity, we must indicate this by coreferencing them, regardless of the entity level (NAM, NOM, PRO).

Coreference can only be done when mentions share the same entity type (PER, ORG, LOC, GPE or TTL). If you want change the type of an entity during coreference, you must go back to the Entity task and change its type before continuing with coreference.

In most cases annotating coreference is very straightforward. In a document about Osama bin Laden, we want all mentions of 'bin Laden' to be lumped together in the same entity of type PER. In the following passage, all the bracketed mentions should be coreferenced as one entity:

- Videos circulated by **[Osama bin Laden]** have added to the evidence linking **[him]** and the al-Qaida network to the Sept. 11 terrorist attacks in the United States, the government said Wednesday in an updated dossier on the investigation. The document, published by Prime Minister Tony Blair's office, said **[the Saudi dissident]** had come "closest to admitting responsibility" for the attacks in an "inflammatory video," allegedly made on Oct. 20 that was not released to the media but circulated to al-Qaida members. "The battle has been moved inside America, and we shall continue until we win this battle, or die in the cause and meet our maker," the document quotes **[bin Laden]** as saying.

The name mentions of 'Osama bin Laden' are easy to spot. However, it is important to annotate all mentions that refer to the entity that is 'bin Laden'. This will include nominal mentions bolded above, such as 'the Saudi dissident' and pronominal mentions such as 'him', which would all be coreferenced together.

NOTE: All of these coreferring mentions in the example above have the same entity type: PER.

By contrast, 'Saudi' is a GPE, which means it cannot be coreferential with the PER mentions of 'bin Laden':

- said [the [Saudi]$_{\text{GPE}}$ dissident]$_{\text{PER}}$ had come “closest to admitting…”

On the other hand, if we had a sentence like the following:

- The document said [the Saudi]$_{\text{PER}}$ had come “closest to admitting…”

we would label ‘the Saudi’ as a PER entity in accordance with the Tag for Usage rule, and this entity would be coreferred with other PER mentions of ‘bin Laden’.

6.2 Coreference for Titles

Corefer all title mentions of the same role within the same organization, regardless of whether the mentions are associated with different persons. In the example below, the title ‘President’ refers to the same specific role in both cases (President of the U.S.) so the two mentions are marked as coreferring:

- [President]$_i$ Clinton and [President]$_i$ Bush did not have the same policies.

The following example shows that if the title mentions refer to different roles (President of the US vs. President of France) they should not be marked as coreferring, even if their strings are identical:

- [President]$_k$ Obama greeted [President]$_j$ Hollande.

NOTE: Person entity mentions (PER) and title entity mentions (TTL), being different types of entities, will not be coreferenced – but since TTL mentions will only be tagged when establishing a relationship to a PER mention for which the TTL refers, the relationship will not be lost and will be annotated at the Relations level.

NOTE: If you are not sure whether two titles refer to the same position (e.g. if you are not sure whether two people are president of the same country), you can look these up online.

Generic titles that are used as part of a form of address, such as *Mr, Mrs, Ms, Dr, Professor* etc. should be marked as coreferring if the same string is used (i.e. we will <u>not</u> coreference these as positions, only as strings)

- [Mrs]$_j$ Jones and [Mrs]$_j$ Smith met for coffee.
- [Dr]$_k$ Mary Arroyo and [Dr]$_k$ Jim Hills published the paper.

6.3 Coreference in Questions

When an entity is being questioned coreference can be marked if context makes the identification clear, e.g.:

- Dialog:
    a. A: I went to see [the breeder]$_k$.
    b. B: [Who]$_k$'s [the breeder]$_k$? Is [that]$_k$ [the breeder that you saw yesterday]$_k$?
    c. A: Yes.

6.4 Coreferencing Organizations Over Time

When comparing mentions of earlier and later versions of an organization (e.g., "1950s IBM" VS "present-day IBM"; "the Blair government" VS "the Brown government"), we will still coreference them as the same ORG entity.

7. Discussion Forums

When annotating discussion forum documents, you should expect to find more colloquial language, including spelling errors, interruptions, unclear expressions and missing punctuation. Annotate each document to the best of your understanding, trying to focus on the author's presumed intent.[1]

7.1 Post Metadata

Post authors: Each discussion forum post begins with an XML heading similar to the following:

```
<post author="pollywog" datetime="2009-03-24T11:34:00" id="p3">
```

In ERE, this data is considered taggable. Therefore, in the above example, there is one taggable entity:

[pollywog]per.nam

Post author quotes: XML metadata also signifies the end of discussion forum posts and the boundaries of quotes. It's important to note shifts in post authors, because we will coreference speakers accurately. Take the following example:

```
<post author="Tsukasa" datetime="2011-11-09 id="p188">
<quote orig_author="Schrodinger's Cat">
not that I'm excusing it in any way
</quote>
good! youre starting to make sense to me
</post>
```

[1] Note that the policies set down regarding word tokenization in this section are different from Treebank policies on some items.

When a post author quotes another poster, XML displays <quote orig_author="X"> where X will be the name of whomever is being quoted. Additionally, the </quote> marker signals the end of quoted text. Similarly, </post> will mark the end of the post author's post.

So, in the above example, Tsukasa has written "good! youre starting to make sense to me," while also quoting Schrodinger's Cat, who previously stated, "not that I'm excusing it in any way."

Post author nicknames: Sometimes post authors will use nicknames that are based on screen names for other post authors. These nicknames should still be tagged as named Per mentions. Notice the usage in the following example:

```
<post author="Scrat" datetime="2003-11-26T23:01:00" id="p173">
<quote orig_author="Moot">
C'mon scratsy you gotta tell us the story
</quote>
Eh maybe when you're older Mootay.
</post>
```

Post author [Scrat] refers to post author [Moot] as "Mootay" after [Moot] refers to him as "scratsy". It's easy to miss these named PER mentions, especially when they aren't capitalized.

Unattributed quotes: Sometimes discussion forum quotes are unattributed, so <quote orig_author="X"> will not appear. Instead, <quote> will be the only indicator. In these cases, coreference can be more difficult, so be sure to create new equivalence classes for entities that aren't linked with previously mentioned entities. For instance, the unattributed quotes can be from other discussion forums.

7.2 URLs

Potential entity mentions embedded in URLs will generally not be tagged as entities, as it isn't usually clear in cases such as this that any real-world entity is truly being referenced (other than perhaps in a generic fashion). For instance, in:

```
<a href="http://www.lonelyplanet.com/usa/virginia>
```

neither "usa" nor "virginia" are tagged as entities. Similarly, in the following, "whitehouse" is not tagged as an entity:

whitehouse.gov

Hyperlinks: While we do not tag entities embedded in URLs, we **do** tag entities within XML hyperlink metadata. Generally appearing before or after a URL, hyperlink metadata surrounds text with <a href=”URL”> and </a> markers. For example, a document might contain the following text:

> <a href="http://www.flyers.nhl.com/club/schedule.htm">Flyers Captain Claude Giroux Scores with 5 Seconds Remaining to Win Stanley Cup</a>

> Within this hyperlink metadata, [Flyers]$_{\text{ORG}}$,[Captain]$_{\text{TTL}}$, and [Claude Giroux]$_{\text{PER}}$ are all taggable entities.

7.3 Misspellings and Incorrect Punctuation

Annotate misspellings according to the intended meaning, as far as that can be deciphered. In the example below the second “I” is a typo and we can assume that the author intended to write “a”. The second “I” should therefore not be marked as PER.PRO.

- I know **I** guy who can help us out

Similarly, incorrect punctuation should be ignored and the text marked according to the author’s presumed intent, e.g.,

- I bought two [book’s] at the store.

In the case of missing apostrophes, annotate the entire word, even if you would normally exclude the apostrophe from the mention span, e.g.,

- Call me when [your] in town.
- [Im] in town!

In the case of missing spaces, annotate the entire span even if it includes text that you would normally not annotate, e.g.,

- [Iwanna] get out of this town.
- [IDK] who that is.

7.4 Repetition and Fragments

In repeated text mark each mention separately, e.g., in the example below both mentions of “there” are marked as separate LOC entities, coreferring with each other:

- I wanna dive [there]LOC…. *drive [there]LOC I mean

Annotate fragments to the best of your interpretation, e.g., in the example below there are two fragments and one complete sentence mentioning "John". All three mentions should be annotated and marked as coreferring.

- dialogue:
    a. A: [John]
    b. B: [John] was
    c. A: I saw [John]

7.5 Coreference in Discussion Forums

Discussion forums contain dialogues between multiple participants. Care must be taken to mark coreference correctly, especially for first and second person pronouns, e.g.,

- dialogue:
    a. [I]$_k$ want to get rid of this one.
    b. Sure [I]$_j$'ll take it off [your]$_k$ hands.