DEFT Rich ERE Annotation Guidelines: Entities V2.4

Linguistic Data Consortium
May 21, 2015

Changes from V2.3

- 8.3: Added Hyperlinks section

Changes from V2.2

- Added Section 8.2: Twitter usernames
- Aligned NonSPC entity examples throughout the document
- 2.2: New note on referential pronoun tagging within nominal entities
- 2.2.1: New note on how to tag multiple heads in nominal entities
- 4.1: Added examples that show how to tag “who” next to PER.NAM entities

Contents

1 Introduction

The purpose of this annotation project is to mark up texts for entities, relations, and events. Relations and events are both created by using entities which are linked together in a process called coreference, as their building blocks. 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.

The Entity Detection task requires that annotators detect specific types of entities mentioned in the source data while disambiguating their intended meaning. Additionally, an annotator will select attributes for these entities which will be extracted and merged into a unified representation for each entity.

1.1 Basic Concepts

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.
- 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.
- Facility (FAC) – A facility is a functional, primarily man-made structure.

Facilities are artifacts falling under the domains of architecture and civil engineering.

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 10 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 type (NAM, NOM or PRO) and specificity level (SPC or NonSPC). See sections below for details.

1.2 General Rules

The following general rules apply at all times:

1. Entities are tagged regardless of which syntactic function they fulfill. For example, entity mentions may be adjectives ("[Korean]$_{GPE}$ cars"), possessive determiners ("[her]$_{PER}$ three convictions"), or prepositional complements ("at [the beach]$_{LOC}$").
2. 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]$_{GPE}$ people]$_{PER}$' and 'his army' would be tagged as '[[his]$_{PER}$ army]$_{PER}$'. See sec. 5.6 for more.

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 excluded 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 (section 2), entity class (section 3) and entity type (section 4).

2 Entity mention extents and mention levels

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 ($_{PER, ORG, GPE, LOC, FAC,}$). 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.

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, excluding 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]$_{PER}$
- former president [George W. Bush]$_{PER}$
- the [Eiffel Tower]$_{LOC}$
- [IBM]$_{ORG}$
- the [Yankees]$_{ORG}$ (sports team)
- [Coca-Cola Bottling Co.]$_{ORG}$
- [Uganda]$_{GPE}$
- [Bowdon]$_{GPE}$, [Georgia]$_{GPE}$
- [Mt. Fuji]$_{GPE}$
- the [Kremlin]$_{ORG}$
- the [Kennedys]$_{PER}$
- [Bill]$_{PER}$ and [Hillary Clinton]$_{PER}$
- [Sean "Puffy" Combs]$_{PER}$, aka [Puff Daddy]$_{PER}$ or [P. Diddy]$_{PER}$
- the [North]$_{GPE}$ (for 'North Korea')
- the so-called [[Northern Cyprus]$_{GPE}$ Chess Federation]$_{ORG}$
- [Land of the Rising Sun]$_{GPE}$ (for 'Japan')
- the famous [Lincoln Memorial]$_{LOC}$
- the incomparable [Steven Spielberg]$_{PER}$
- the [US]$_{GPE}$ [State Department]$_{ORG}$
- the [States] $_{GPE}$ (as a nickname for the US)
- The [Gambia]$_{GPE}$

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 the 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]~~$_{GPE}$' elected a new president').

Some examples of possible nominal mentions are given below:

- [a monument]$_{FAC}$
- [a few well-known monuments]$_{FAC}$
- [some teams]$_{ORG}$
- [the building]$_{FAC}$
- [that city]$_{GPE}$
- [her country]$_{GPE}$
- [the director of the Oscar winning film *Lincoln*]$_{PER}$
- [the family]$_{PER}$
- [another large company whose investors revolted]$_{ORG}$
- [the presidential hopeful from Chicago]$_{PER}$

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

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]$_{NAM}$ separatist organization]$_{NOM}$
- [the now-defunct [G17 Plus]$_{NAM}$ political party]$_{NOM}$
- [the president of [Ford] $_{NAM}$] $_{NOM}$

NOTE: When a nominal entity contains pronouns that refer to that same entity, do not annotate said pronouns. This is akin to double tagging the entity. Both "who" and "he" are not tagged, because they refer to "the man who said he would wash

your car". However, "your" is tagged because it refers to another entity:

- [Michael]$_{NAM}$, [the man ~~[who]~~ said ~~[he]~~ would wash [your]$_{PRO}$ car]$_{NOM}$, welcomed new business.

2.2.1 Tagging Entity Mention Heads in Nominal Entities

In addition to the extent for nominal mentions, the head of the nominal mentions must be marked as well. The **head** is the head noun of the nominal phrase. For example (the head is marked with an underline):

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

In general, NOM mentions are only allowed one word in the head extent. The main exceptions to this are LOC mentions that consist of GPE name and directional modifier.

- [the southwestern [United States]$_{GPE.NAM}$] $_{NOM.LOC}$

The entity, "United States," is originally a GPE.NAM, but the modifier causes the full nominal phrase to become a LOC.NOM. "United States" cannot be decomposed into a head and modifier, so it must be the entire head. Additionally, "United States" is tagged as a GPE.NAM, without marking anything as the head, as heads are only required for NOM mentions.

NOTE: When you annotate a nominal mention that contains more than one head, you will need to mark the heads separately. To do this, tag the full nominal extent twice if there are two heads, three times if there are three heads, etc. Each time you will mark a different head. For example:

- [nearly 1,000 men and women injured in Iraq and Afghanistan]$_{PER.NOM}$
- [nearly 1,000 men and women injured in Iraq and Afghanistan]$_{PER.NOM}$

2.2.2 "Headless" Mentions

Headless mentions are constructions in which the nominal head is not overtly expressed. Although these mentions are technically headless, we will assign as head

the right-most premodifier that falls directly before the spot where the head would be.

- It was [the toughest]$_{LOC}$ that she'd climbed without using a harness
- [more than 30]$_{PER}$ showed up at the park
- [many]$_{PER}$ on both sides disagreed
- [60%]$_{ORG}$ said they'd change distributors by the year's end
- [sixty percent]$_{PER}$ said they'd buy a new home in the next 5 years
- [35]$_{PER}$ were injured as a result of the crash

NOTE: Bare demonstratives followed by a relative clause (or modified in some way) should be tagged as:

- [Those who would disagree]$_{PER}$ won't be happy with the outcome.
- [Those present at the meeting]$_{PER}$ noticed how his demeanor changed.
- We must help [those in need]$_{PER}$

2.2.3 Partitive Constructions and Non-Partitives

Partitive constructions have two elements: the part and the whole, both of which need to be tagged. The extents of part-elements are included in the whole. Just as in Headless mentions, we will tag the right most premodifier of the first element as the head of the partitive construction. The whole element needs to be tagged separately.

- [some of [the lawyers]$_{PER}$] $_{PER}$
- [one of [the houses]$_{FAC}$]$_{FAC}$
- [half of [the team]$_{ORG}$] $_{PER}$
- [all of [them]$_{PER}$]$_{PER}$
- [sixty percent of [the participants]$_{PER}$]$_{PER}$

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	me	themselves
another	mine	these
any	more	they
both	most	this
each	much	those
each other	my	us
either	myself	we
everybody	one	what
everyone	one another	whatever

everything	other	where
few	others	wherever
he	our	which
her	ours	whichever
hers	ourselves	who
herself	several	whoever
him	she	whom
himself	some	whomever
his	somebody	whose
I	someone	you
it	something	your
its	that	yours
itself	their	yourself
little	theirs	
many	them	

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

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

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

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

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

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]$_{PER}$'s house
- the [buildings]$_{FAC}$'s entrance
- [I]$_{PER}$'ve never been there
- [She]$_{PER}$'s a smart girl

3 Mention Class

In addition to an entity's mention level – either NAM, NOM, or PRO – annotators will now decide on an entity's level of specificity, which we call mention class. A dropdown menu next to entity type and mention level options will direct annotators to choose between Specific or Non-Specific entity mentions. In Light ERE,

annotators only tagged specific entity mentions. However, in Rich ERE we will tag an entity every time one of the ten entity types appears in a document.

3.1 Specific (SPC)

Specific entities are asserted in a document (not hypothetical, generic, or other). An entity is SPC when the entity being referred to is a particular, unique object (or set of objects).The Light ERE understanding of taggable entity mentions is what Rich ERE now calls SPC.

- [[John] PER.NAM.SPC's lawyer] PER.NOM.SPC won the case.
- This afternoon, [a crowd of angry muslims] PER.NOM.SPC set fire to [a hotel] FAC.NOM.SPC.
- [Lee Hawk Seder] PER.NAM.SPC is Jerusalem Bureau Chief for the [Washington Post] ORG.NAM.SPC
- [Columbia University] ORG.NAM.SPC 's [Institute of War and Peace Studies] ORG.NAM.SPC
- [At least four people] PER.NOM.SPC were injured.

Sometimes a mention refers to a large number of entities (where the actual members of the set are not necessarily identifiable) and the number used is an estimate.

- [Over two hundred thousand people] PER.NOM.SPC participated in the riots.

In cases where the author mentions an entity whose identity would be difficult to locate, and then conflates it with multiple other fuzzy mentions, all mentions are tagged as SPC.

- [Sources] PER.NOM.SPC said…
- [Officials] PER.NOM.SPC reported…

3.2 Non-Specific (NonSPC)

Non-Specific entities are those which fall under the following categories: negated, generic, and irrealis.

Negated
A negated entity is one that has been quantified such that it refers to the empty set of the type of object mentioned.

- [No sensible lawyer] PER.NOM.NonSPC would take that case.
- [No one] PER.NOM.NonSPC has claimed responsibility.
- There are [no confirmed suspects] PER.NOM.NonSPC yet, but officials say several

Middle East groups are expected to be investigated.

NOTE: We do not tag nominals introduced by negated predicates. For example, in the following sentence, we would not annotate "lawyers": "They are not ~~lawyers~~."

Irrealis
Irrealis references include quantified nominal phrases in modal, future, conditional, hypothetical, uncertain, or question contexts (in all cases the entity/entities referenced cannot be verified, regardless of the amount of "effort").

- [Many people] PER.NOM.NonSPC will participate in the parade.
- I don't know [how many people] PER.NOM.NonSPC came.
- Do you know [how many people] PER.NOM.NonSPC came?
- We will elect [five new officials] PER.NOM.NonSPC.
- [You] PER.PRO.NonSPC know, I didn't even realize...

Generic
A mention is generic when the entity being referred to is not a particular, unique object (or set of objects). Instead GEN entities refer to a non-descript category of entities. Notice that the mentions in question are still understood to be referential in that they point to actual things in the world.

- [those dang Americans]PER.NOM.NonSPC love McDonald's even though it's gross food (we would also tag [Americans]GPE.NAM.SPC)
- I think [parks]FAC.NOM.NonSPC are my favorite places within a big city
- [Democrats]PER.NOM.NonSPC and [Republicans]PER.NOM.NonSPC are always at each other's throats
- [Lawyers] PER.NOM.NonSPC don't work for free.
- But the sense of urgency for this meeting matches the rage felt by both [Israelis] PER.NOM.NonSPC and [Palestinians]PER.NOM.NonSPC after yesterday's violence.
- …[extremist groups] PER.NOM.NonSPC have a lot of support these days and a lot of power.

4 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 5 entity types: person (PER), organizations (ORG), geo-political entities (GPE), locations (LOC), facilities (FAC. A description of each type follows.

4.1 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 be annotated as NonSPC nominal PER entities.

NOTE: Generic PER mentions that reflect ORG names should be annotated as NonSPC PER entities.

- [The Democrats]$_{\text{PER.NOM.NonSPC}}$ are all the same. (Separately, Democrats will be tagged as ORG.NAM.SPC.)
- [Democrats]$_{\text{PER.NOM.NonSPC}}$ are all the same.

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], [the president who took office in 1993], declined comment.
- [Bill Clinton], [who] declined comment, left the meeting.
- [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.

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]

4.2 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 be annotated as NonSPC PER entities.

- [The Democrats]$_{\text{PER.NOM.NonSPC}}$ are all the same. (Separately, Democrats will be tagged as ORG.NAM.SPC.)
- [Democrats]$_{\text{PER.NOM.NonSPC}}$ are all the same.

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]

4.3 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]$_{GPE}$ for our vacation.
- They delivered the supplies to [Pakistan]$_{GPE}$

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

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

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

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. We do not double-tag "Americans" in the example below as a PER and a GPE, as no definite or indefinite article (the/a/an) appears:

- [The Swiss]$_{PER.NOM.SPC}$ have joined us on the bus tour. ("Swiss" will also be tagged as a GPE.NAM.SPC)
- Luckily, [the Australians]$_{PER.NOM.SPC}$ made it to the barbeque on time. ("Australians" will also be tagged as a GPE.NAM.SPC)
- [Americans]$_{PER.NOM.NonSPC}$ love fast food.
- [The French]$_{PER.NOM.NonSPC}$ enjoy wine. ("French" will also be tagged as a GPE.NAM.SPC)

- [Canadians]$_{PER.NOM.NonSPC}$ 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]$_{GPE}$ produces better wine than [New Jersey]$_{GPE}$.
- [Israeli]$_{GPE}$ troops
- The [Palestinian]$_{GPE}$ Authority has banned rallies that are pro-[Iraq]$_{GPE}$

NOTE: Countries of countries, such as 'the [European Union]$_{GPE}$' and 'the [United Kingdom]$_{GPE}$' 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. 5.2 below.

NOTE: We will not annotate the adjectives "local" or "federal" as GPE or LOC entities, as these words evoke entity traits. "Local" and "federal" should not be tagged as entities by themselves. Not all adjectives are untaggable. GPEs are sometimes in adjective form, as in "American people", in which American is tagged as GPE.NAM.SPC

- He'll be tossed in [a [state]$_{\text{GPE.NOM.NonSPC}}$ or ~~[federal]~~$_{\text{~~GPE.NOM.NonSPC~~}}$ penitentiary]$_{\text{FAC.NOM.NonSPC}}$ next month.
- He'll be tossed in [a [state]$_{\text{GPE.NOM.NonSPC}}$ or federal penitentiary]$_{\text{FAC.NOM.NonSPC}}$ next month.

4.4 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 or natural structures like bodies of water and mountains. 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, street addresses, oceans, seas, straits, bays, channels, sounds, rivers, islands, lakes, national parks, and mountains. Fictional or mythical locations should <u>not</u> be tagged. For instance:

- Vice President Cheney visited [the <u>site</u>].
- In Armenia, the three of them will join other, similar delegations from around [the <u>world</u>]...
- The droids landed on ~~[Tatooine]~~$_{\text{~~LOC~~}}$
- ... eclipse fans are being warned not to look directly at [the <u>sun</u>] ...
- the [Missouri River]
- [the <u>region</u> where the movement has found most success recently]
- Police are asking everyone to avoid [the affected <u>area</u>].
- Many people in [North America] saw a partial solar eclipse yesterday.
- [~~[federal]~~ land] will be used for construction.
- [federal land] will be used for construction.

4.5 Facility Entity (FAC)

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

A facility is a functional, primarily **man-made** structure. These include buildings and similar facilities designed for human habitation, such as houses, factories, stadiums, office buildings, gymnasiums, prisons, museums, and space stations;

objects of similar size designed for storage, such as barns, parking garages and airplane hangars; elements of transportation infrastructure, including streets, highways, airports, ports, train stations, bridges, and tunnels. Roughly speaking, facilities are artifacts falling under the domains of architecture and civil engineering. Facility entities which do not fit into the types defined below will not be tagged.

NOTE: 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.

- ~~[The wall]~~ and ~~[coffee table]~~ over [there]

Types of Facilities

Airport
A facility whose primary use is as an airport.

- new york's [la guardia airport] has been a nightmare this year

Plant
One or more buildings that are used and/or designed solely for industrial purposes: manufacturing, power generation, etc.

- the train ran directly from [the oil refinery] to [the smelter]

Building-or-Grounds
Man-made/-maintained buildings, outdoor spaces, and other such facilities. This includes anything from a tent to a hotel to a ranch to Disneyland.

- We found ourselves at the [national archives].
- The [Berlin Wall]
- the parades at [Disneyland]

Subarea-Facility
Taggable portions of facilities. The threshold of taggability of subarea-facility is the ability of the area to contain a normally proportioned person comfortably. Individual rooms of buildings are considered subarea-facilitiy, but other portions of buildings, such as walls, windows, or doors, are not tagged.

- Two men who rented [an Aden apartment]

Path
A facility that allows fluids, energies, persons or vehicles to pass from one location to another. For example: streets, canals, and bridges.

- Undercover agents have been patrolling [Aden's streets].
- [Telephone lines] were knocked down…

5 Difficult Cases and Interactions Among Entity Types

5.1 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]$_{FAC}$ driveway]$_{FAC}$
- [the [Smith]$_{PER}$'s house]$_{FAC}$

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

5.2 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]$_{FAC}$, [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]$_{LOC.NAM}$
- the [West Bank]$_{LOC.NAM}$
- [Eastern [Europe]$_{LOC.NAM.SPC}$]$_{LOC.NAM}$
- [Siberia]$_{LOC.NAM}$

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]$_{GPE.NAM}$]$_{LOC.NOM}$
- [southwest [Germany]$_{GPE.NAM}$]$_{LOC.NOM}$

5.3 NAM vs. NOM

Some ambiguities can arise when trying to make a distinction between NAM and NOM entities. It may appear that a NOM is being used to name something, or that a NAM mention may be separated into a few NOMs.

A general property of NAMs is that they are defined to pick out one particular entity as a referent. They are unique identifiers, like "Vladimir Putin" or "United States."

NOMs, on the other hand, define an entire category. They can pick out a referent which belongs to that category, but only after disambiguating it from all other potential members of the category. If a nominal mention is used as an individual reference in a discourse, the head noun often has to be "individualized" via quantification and/or qualification with determiners, adjectives, relative clauses, etc.

[Vladimir Putin] sat at the table.	*[Vladimir Putin]* $_{NAM}$
[The man] sat at the table.	*[The man]* $_{NOM}$

One of the trickiest parts of distinguishing between NAMs and NOMs is NOM categories modified by NAMs such that they only have one referent, such as:

the Pakistani army
the Chinese embassy
the Egyptian supreme court
the University of Chicago payroll department

With the GPE/ORG modifying the categories, they pick out a specific referent in each NOM category. It is hard to decide whether the whole string should be treated as a NAM, or as a NOM mention with a modifying GPE/ORG named entity.

Some ORGs are unambiguously NAM, as they automatically pick out one specific entity, not a member of a set.

[Nazareth Academy] $_{ORG.NAM.\ SPC}$
the [Danger Danger Gallery] $_{ORG.NAM.SPC}$
the [United States Armed Forces] $_{ORG.NAM.SPC}$

Some ORGs are unambiguously NOM, as they could not be considered the name of an organization, only a type of organization.

[the [U.S.] military] *ORG.NOM.SPC*
[the [Chinese] embassy] *ORG.NOM.SPC or FAC.NOM.SPC*

Some are tricky and you probably need to rely on external resources such as Wikipedia for reference.

the [Pakistani army] *ORG. NAM.SPC*
[the Egyptian supreme court] *ORG.NOM.SPC*
[the [University of Chicago] *ORG.NAM.SPC* *payroll department]* *ORG. NOM.SPC*

5.4 Entity Types and Tag for Usage

Rule: We always tag an expression according to its usage in context. 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.

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

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

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

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

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 in nominal mentions, you should tag the full nominal mention extent multiple times with a different head marked each time. For example, we will tag "my" once, but the full extent will be annotated three times in the following:

- [[my]$_{PER.PRO}$ stepkids and friends and family]$_{PER.NOM}$
- [my stepkids and friends and family] $_{PER.NOM}$
- [my stepkids and friends and family] $_{PER.NOM}$

Also, if the modifier comes after the coordinated nouns, we would tag the full extent in the same fashion:

- [students and faculty at [Penn]$_{ORG.NAM}$] $_{PER.NOM}$
- [students and faculty at Penn] $_{PER.NOM}$

- [the East and South of [Iran]$_{GPE.NAM}$]$_{LOC.NOM}$
- [the East and South of Iran]$_{LOC.NOM}$

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]$_{GPE.NAM}$ and [Canadian] $_{GPE.NAM}$ soldiers]$_{PER.NOM}$
-
    -
    -
- [the CEOs of [Google]$_{ORG.NAM}$ and [Youtube] $_{ORG.NAM}$]$_{PER.NOM}$

Cases where multiple entities are joined together by punctuation marks in a single, continuous string can still be tagged separately:

- [Af]$_{GPE.NAM}$ --[Pak]$_{GPE.NAM}$
- [Brad]$_{PER.PRO}$&[Angelina]$_{PER.NAM}$
- [me]$_{PER.PRO}$+[you] $_{PER.PRO}$

However, if multiple entities are merged by neologism or slang, we tag only one entity:

- [Brangelina]$_{PER.NAM}$ (where Brad and Angelina are merged into one entity)

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

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

5.8 Hyphenated pre-modifiers

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

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

5.9 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.

5.10 Examples with entities completely annotated

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

6 What NOT to tag

6.1 Event Names

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]$_{ORG}$

7 Coreference

7.1 General Instructions for Coreference Tagging

The annotation tool requires you to make decisions of entity coreference each time you annotate an entity mention. The basic concept of coreference is that 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). In the tool, you drop all mentions referring to the same entity to the same entity bin.

Coreference can only be done when mentions share the same entity type (PER, ORG, GPE, LOC, or FAC) and same entity class.

In most cases annotating coreference is very straightforward. In a document about Osama bin Laden, we want all mentions of 'bin Laden' to be in the same entity bin whose entity type is PER and entity class is SPC. 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 and entity class SPC.

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

- said [the [Saudi]$_{GPE}$ dissident]$_{PER}$ had come "closest to admitting…"

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

- The document said [the Saudi]$_{PER.NOM}$ 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'.

7.2 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.

7.3 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.

7.4 Coreferencing NonSPC mentions

We do not coreference NonSPC mentions with SPC mentions. When encountering a NonSPC entity in the Coref tab, we should remember to base coreference decisions on specificity level. For example, the following entities are all NonSPC, so they should be coreferenced:

- [Lawyers]$_{NonSPC}$ don't always work for free, but [they]$_{NonSPC}$ do from time to time. [Lawyers]$_{NonSPC}$ also generally make a lot of money…

It is important to note that articles or determiners do not always appear in the source text. Contrast the previous example with the following excerpt:

- [The lawyers]$_{SPC}$ entered the court room a few minutes before the trial started. [Lawyers]$_{SPC}$ approached the bench to talk with the judge. Isn't it odd how [lawyers]$_{NonSPC}$ sometimes seem chummy with judges and other times [they]$_{NonSPC}$ seem like complete enemies?

In this example, the SPC mentions would be coreferenced in one equivalence class and the NonSPC would be coreferenced in a separate equivalence class.

When in doubt, do not coreference NonSPC mentions.

8 Informal Genres: Discussion Forums and Twitter data

When annotating discussion forum and Twitter 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]

8.1 Post Metadata

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 entity tag:

- [pollywog]per.nam

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="2011-11-09T 17:40:00" 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>

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."*

Sometimes discussion forum quotes are unattributed, so *<quote orig_author="X">* will not appear. Instead, *<quote>* will be the only indicator.

8.2 Twitter usernames

Usernames in Twitter data are denoted by the '@' symbol. Annotators will not

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

include the '@' symbol when tagging the usernames as PER.NAM or ORG.NAM entities.

- when's the next event @slyfoxbrewery ?
    - [slyfoxbrewery]annotated as an ORG.NAM

- @MayorNutter have you ice skated at Dilworth?
    - [MayorNutter] annotated as a PER.NAM

8.3 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:

- 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 and markers. For example, a document might contain the following text:

> Flyers Captain Claude Giroux Scores with 5 Seconds Remaining to Win Stanley Cup
>
> Within this hyperlink metadata, [Flyers]$_{ORG}$,[Captain]$_{TTL}$, and [Claude Giroux]$_{PER}$ are all taggable entities.

8.4 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.

8.5 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.

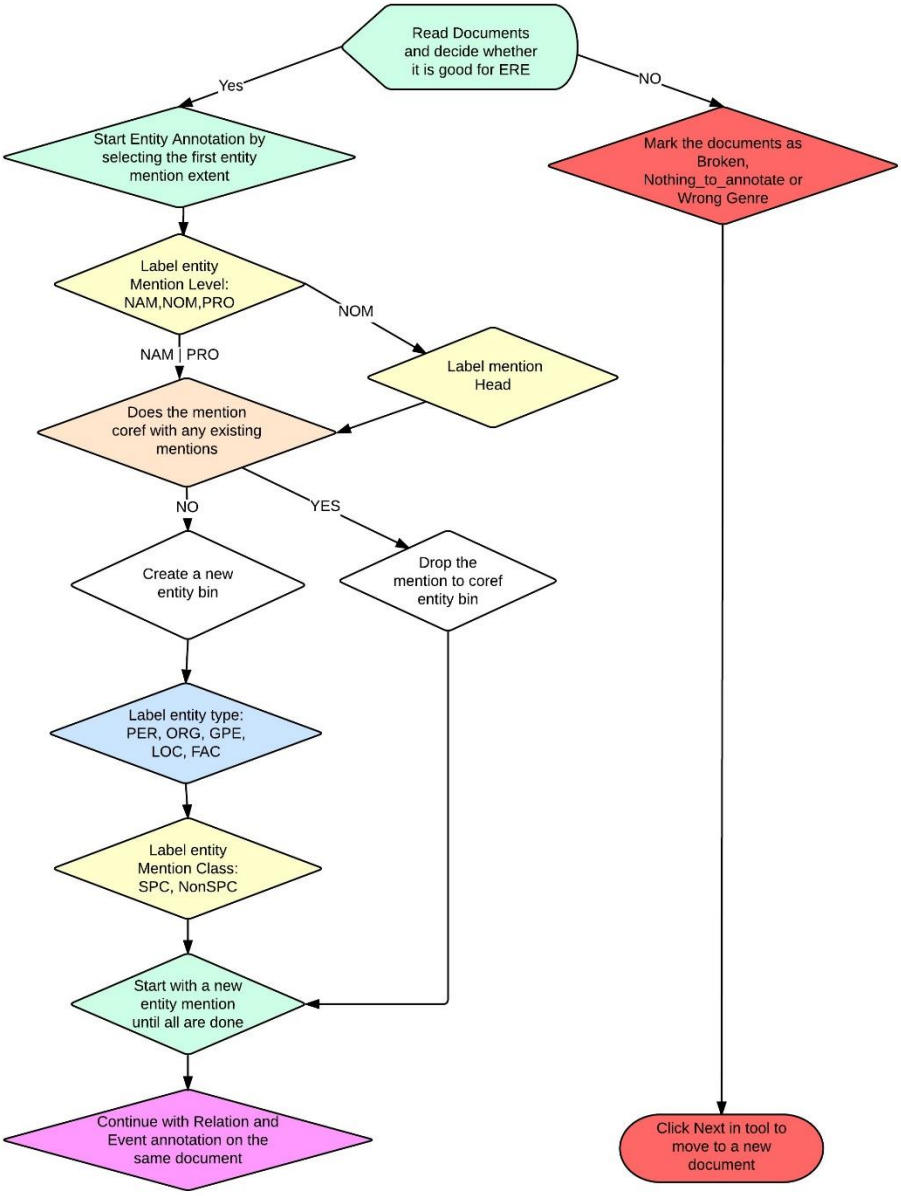
- dialogue:
    a. A: [John]
    b. B: [John] was
    c. A: I saw [John]

8.6 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.

Appendix: Annotation Decision tree


Read Documents and decide whether it is good for ERE
Yes
NO
Start Entity Annotation by selecting the first entity mention extent
Mark the documents as Broken, Nothing_to_annotate or Wrong Genre
Label entity Mention Level: NAM,NOM,PRO
NOM
NAM | PRO
Label mention Head
Does the mention coref with any existing mentions
NO
YES
Create a new entity bin
Drop the mention to coref entity bin
Label entity type: PER, ORG, GPE, LOC, FAC
Label entity Mention Class: SPC, NonSPC
Start with a new entity mention until all are done
Continue with Relation and Event annotation on the same document
Click Next in tool to move to a new document