

Simple Named Entity Guidelines V6.4

Developed by Stephanie Strassel — Linguistic Data Consortium
For the Low Density Languages Project
(Based largely on the MUC-7 NE Guidelines)

1 Introduction

An entity is some object in the world -- for instance, a place or a person. A named entity is a phrase that uniquely refers to that object by its proper name, acronym, nickname or abbreviation. Some examples of named entities follow:

Coca-Cola Bottling Co.
Bob Austin
the Eiffel Tower
IBM
the Yankees
Uganda
Bowdon, Georgia
Mt. Fuji
the Kremlin
the Kennedys

2 Entity Types

We will identify four types of named entities:

PERSON (PER): Person entities are limited to humans identified by name, nickname or alias.

TITLE/ROLE (TTL): Named personal titles or roles. These are restricted to titles that occur directly before or after the person name they describe.

ORGANIZATION (ORG): Organization entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure.

LOCATION (LOC): Location entities include names of politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments.

Other types of named entities like animals, inanimate objects and monetary units will not be annotated.

Within this document, named entities are indicated by [square brackets].

Note that annotations **may not** overlap or embed in the text. In other words, every annotation must end before another can begin.

2.1 Person Names

People may be specified by name, nickname or alias. Family names should also be tagged as PERSON. Names of deceased people, as well as fictional human characters appearing in movies, television, books and so on, should be tagged as PERSON entities. Religious deities should also be tagged as persons.

2.2 Titles, roles and appositives

Titles, roles and honorifics such as "Mr." and "President" are tagged as title entities and are separated from the individual's name. For instance, in the following sentence, there are two separate entities marked:

[Vice President] [Cheney] visited the site.

For this task we define titles and roles as occurring either directly before or directly after a person name. Therefore, titles and roles are only tagged when they occur directly next to the person name they modify. In the following example, for instance, the phrase "Vice President" is not considered a title and is not tagged:

The strongest supporter was the Vice President.

If a title contains within it a taggable entity, tag that entity separately¹. For instance:

[Microsoft] [Chairman] [Bill Gates] stated that...
 ORG TTL PER

You may occasionally encounter an appositive like "Jr.", "Sr.", and "III". These are considered part of a person name and should be marked as part of the name, for instance:

[Mr.] [Albert Franklin, Jr.] was part of the research team.
 TTL PER

Finally, sometimes the name of the person is split into two pieces by the title. In these cases, we will annotate the two pieces of the PERSON name as two separate PERSON entities:

¹ Notice that the prohibition against overlapping or nested annotations (we may **never** insert one Named Entity *inside* of another one) forces us to turn the string "Microsoft Chairman Bill Gates" into a sequence of Named Entities. This is a special case for PERSON entities. In general, we may not break Named Entities into smaller pieces.

[Alfred] [Lord] [Tennyson]
 PER TTL PER

Some more examples of names and titles:

[GlobalCorp] [Vice President] [John Smith]
 ORG TTL PER

[Treasury] [Secretary] [Jackson]
 ORG TTL PER

the [U.S.] [Vice President], [Dick Cheney]
 LOC TTL PER

[Justice] [Minister] [Giovanni Maria Flick]
 ORG TTL PER

[British] [Rashtrodut] [Anwar Coudhury]
 LOC TTL PER

[Mission Control] [Chief] [Vladimir Solovyov]
 ORG TTL PER

2.3 Organization Names

Tag all proper name mentions of groups with a defined organizational structure. These include

Businesses

[Bridgestone Sports Co.] profits

Stock exchanges

[NASDAQ] shares

Multinational organizations

[European Union] representatives

Political parties

[GOP] hopeful

Non-generic government entities

[the State Department]

Sports teams

[the Phillies]

Military groups

[the Tamil Tigers]

Many other kinds of entities refer to facilities or buildings that are primarily defined by their established organizational structure, and can do things like issue statements, make decisions, hire people, raise money and so on. A mention of such an entity should be tagged as an ORGANIZATION when it functions like an ORG in the document. These include things like:

Churches and other religious institutions

[Trinity Lutheran Church]

Hospitals

[Finger Lakes Area Hospital Corp.]

Hotels

[Four Seasons Hotel Group]

Museums

[the Guggenheim Museum]

Universities

[the University of Chicago]

Government offices

[the White House]

Note that definite and indefinite determiners 'the' and 'a' are included in the annotation, except for cases when they quantify something other than the tagged entity, as in the following examples:

A [Gulshan Hotel] spokesman

the [U.S.] Vice President

As in the above examples, this exception is particularly common when the tagged name is used in the pre-modifier (adjective) position.

2.3.1 General ORGANIZATION-like non-entities

General entity mentions such as "the police" and "the government" should not be tagged, since these are not unique proper name references to specific entities.

2.4 Location Names

Examples of place-related strings that are tagged as LOCATION include named heavenly bodies, continents, countries, provinces, counties, cities, regions, districts, towns, villages, neighborhoods, airports, highways, street names, factories, manufacturing plants, street addresses, oceans, seas, straits, bays, channels, sounds, rivers, islands, lakes, national parks,

mountains, fictional or mythical locations, and monumental structures, such as the Eiffel Tower and Washington Monument. For instance:

```

the collapse of the newly-constructed [Teton Dam]
                                     LOC
the dispute over votes in [Dade County]
                                     LOC
[The Walt Whitman Bridge] remained closed
LOC
repairs began on a 10-mile stretch of [the Alaskan Pipeline]
                                     LOC
[The Garden State] is known for its tomatoes.
LOC

```

2.4.1 Extent of Location Names

There are several issues surrounding the expression of location names and which parts of a string to tag.

2.4.1.1 Compound expressions

Compound expressions in which place names are separated by a comma in English should be tagged as separate instances of LOCATION.

```

[Kaohsiung], [Taiwan]
[Washington], [D.C.]

```

2.4.1.2 Designators

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 LOCATION 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]
[the Himalayan Mountains]

```

2.4.1.3 Location modifiers and "semi-official" place names

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 LOCATION name. For instance:

```

[Upper Volta]
[North Dakota]

```

Even if the place name does not have "official" status but has an agreed-upon definition and is in very frequent use, the string should be tagged as a LOCATION, as in:

[the Middle East]
 [the West Bank]
 [Eastern Europe]

When these modifiers are not the official name of a place, or when the definition of the place might vary from person to person, do not tag the modifier as part of the LOCATION entity name.

[Mississippi River] west bank
 former [Soviet Union]
 [Gaul], in present-day [France]
 lower [Manhattan]
 Northern [California]

These place names can sometimes be tricky. 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.

2.5 Deciding among entity types

There are some situations where deciding what entity type to assign can be somewhat tricky.

2.5.1 ORG referring to LOC, LOC referring to ORG

Many organizations have not only an organizational structure, but a physical location. For instance, museums are primarily organizations but are also housed in a specific building or facility. So while we normally tag museums as ORG entities, there are cases when a particular example might function more like a LOCATION. In cases like this, annotators should tag the named entity based on the way it functions in the sentence. For instance:

[The Guggenheim Museum] announced a new acquisition.
 ORG

[The Guggenheim Museum] was designed by [Wright].
 LOC PER

Thirty people were wounded in the bomb blast in front of the city's [Gulshan Hotel].
 LOC

A [Gulshan Hotel] spokesman called the incident a tragedy.
 ORG

Similarly, city, country, and other place names are frequently used to refer to organizations located in those places rather than the geographical places themselves. For instance:

[Washington] announced a new tax policy today.
 ORG

In this case, the name Washington is used to refer to the US Government, located in Washington. Because Washington is referring to an organization entity in this example, it should be tagged as ORG.

Also, when the name of a unique structure or building (normally a location) is used to refer to the government or other organization housed in that facility, the name should be tagged as an ORG:

[The Pentagon] issued a statement about the incident.
ORG

The same logic applies to place names referring to sports teams:

[Boston] beat [New York] last night in extra innings.
ORG ORG

In summary, for any cases where a place name is used to refer to an organization, you should tag the name based on function in the sentence:

ORG: used when the example primarily refers to the organizational structure, and is acting like an agent (issuing a statement, making a decision, hiring people, raising money, etc.)

LOC: used when the example primarily refers to the physical structure, rather than the people/groups who run it.

3 Difficult Cases

3.1 Expressions that refer to multiple entities

When a phrase refers to multiple named entities, mark each entity separately.

For instance, this sentence contains two entities:

[China] and [South Korea] signed the agreement.

Similarly,

[Jimmy] and [Rosalyn Carter]
[North] and [South America]

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

[the Fish and Wildlife Service]

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).

3.2 Nested Expressions

Recall that no nested expressions will be marked. When the name of one entity contains within it another entity name, do not pull out the name of the other entity and mark it separately². Only tag the larger entity. For instance

```
[Arthur Anderson Consulting] no markup for Arthur Anderson
                             alone
[Boston Chicken Corp.]      no markup for Boston alone
```

3.3 Entities as modifiers

If an entity name modifies another word (even if that word is not a taggable entity type), you should still tag the entity name.

```
[Bridgestone] profits
the [Clinton] government
[Treasury] bonds and securities
[U.S.] exporters
[Apple] computers
[Texas] intermediate crude oil
[China] film festival
```

Similarly, if the entity name occurs in the form of an adjective you should also tag it:

```
the [American] companies
[Cuban] citizens
[Chinese] food
```

3.4 Possessives

When you encounter a possessive construction, tag the two parts individually as two separate names. For instance:

```
[Temple University's] [Graduate School of Business]
[Canada's] [Parliament]
```

Keep in mind that annotation requires you to select whole words to tag as names, so you have to include the "s" even though it's not part of the name.

² PERSON entities that are adjacent to TTL and ORG pre-modifiers are a notable exception. Please see Section 2.2.

3.5 Other types of names

Aliases, acronyms, nicknames and abbreviations for proper names should be tagged as a name:

IBM	[abbreviation for International Business Machines]
Big Blue	[alias for International Business Machines]
Big Board	[alias for New York Stock Exchange]
Mr. Fix-It	[nickname for candidate for head of the CIA]
the Big Apple	[nickname for New York City]
Red Sox	[alias for the Boston Red Sox]
Sears	[alias for Sears Roebuck and Co.]

4 What NOT to tag

4.1 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	[no markup]
vs.	
[the Olympic Committee]	[Organization]

4.2 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 [no markup]

4.3 Generics

Also, generic names that do not refer to a specific entity should not be tagged

the Campbell Soups of the world [no markup]

5 Annotation Uncertainty

In some cases, you may encounter examples that you don't know how to handle. If so, you should proceed as follows:

- If it's an example not covered in the guidelines, note it in your copy of the guidelines and let your supervisor know about it.

- If it's an example where your language differs from the rules as written for English, note it in your copy of the guidelines and let your supervisor know about it.
- If it's a case where there's something wrong with the file you're working on, stop working on that file and let your supervisor know.

Your supervisor might not be available at the moment you have a question or issue, so it's important for you to write down the problem so it can be resolved later.

So that you can keep working even after you have a question about a particular example, we've created one more tag in the annotation tool: "No_Annotation". When you encounter a problem or have a question about a particular word or phrase and you can't get an immediate answer, label the item "No_Annotation". That will let us easily find it later when we try to resolve the problem.

You should also use the "No_Annotation" label in cases where there's some kind of problem with a single word or handful of words in the file – e.g., they're badly translated or the font isn't displaying properly.

If the whole file is problematic (i.e., poor translation, corrupted, font problems), stop working on it and let your supervisor know. If you are using AWS to receive file assignments, you may simply mark the file as "Broken" in the file-assignment interface. If you use this approach, please be sure to include a descriptive comment when AWS asks what is wrong with the file.