.

Annotation Guidelines for Named Entities

Version 1.1

Emi MAEKAWA

2018/06/12

1. About Our Tag Set

The tag set used here combines the 18 categories described in "OntoNotes Release 5.0" [1] and 4 newly added categories. The 18 OntoNote tags have been defined based on the tag definitions given in "OntoNotes Named Entity Guidelines Version 14.0" [2]. The 18 tags have been slightly modified for our purpose.

2. Tag Set

Tag Name (OntoNote Tag)	Brief Description	Comments	Example *Named entity mentions are marked within round brackets.
NE-PER (PERSON)	Person Name	Proper names of people including first names, last names, individual or family names, fictional names and unique nicknames. Generational markers such as Jr. and IV, and royal titles such as Queen and Sir are included. Unmarked: honorific titles other than the above such as Mr., Mrs., Miss, Ms, and Dr.	(Ray Parker Jr.) Dr. (Brown) / (St. Marks) (Queen Elizabeth II) (Princess Diana) (Charlie Brown) the President (Kennedy)
NE-PER-N (newly added)	Post / Job Title	Names of posts and job titles. Unmarked: honorific titles such as Mr. and Dr., modifiers such as "former" and "previous", and any common nouns.	(American President) Bush (Microsoft CEO) former (First Lady) (Judge) Brown
NE-CHA (newly added)	Non-human Creature Name	Names of creatures other than humans including animals and fictional non-human characters. Names of gods and goddesses are marked with NE-PER, NOT NE-CHA.	(Snoopy) (Flipper) the dolphin
NE-NRP (NORP)	Nationality / Religion / Political affiliation	Adjectival forms of GPE and non-GPE place names (such as American), named religions, heritage, and political affiliation. Regardless of context, adjectival forms such as "French" are NRPs while nominal forms such as "France" are GPEs.	(European) euphoria an (American) publishing group He is (Jewish). The (Hindu) newspaper Three (Democrats) the (Muslims) France's (Socialist) government
NE-FAC	Facility	Names of man-made structures, including	the (Brooklyn Bridge)

(FACILITY)	Name	buildings, airports, stations, infrastructures (bridges and streets), monuments, oil fields, golf courses, hospitals, zoos, shopping centers, etc. Facility names can be ORG depending on context.	(Yankee Stadium) (Madison Avenue) (42nd Street) / the (Eiffel Tower) (Statue of Liberty) We got out of the (City Library). (**NE-ORG** City Library) said that
NE-ADD (newly added)	Contact Information	Addresses, phone numbers, URLs, email addresses. The expression "[place name], [higher-level municipality's name]" such as "Boston, MA" and "Paris, France" is NOT marked with NE-ADD.	(10025 5th Ave., NY, NY 10003) c.f. on (NE-FAC 5th Ave), (NE-GPE NY).
NE-ORG (ORGANIZA-TION)	Organization Name	Names of companies, government agencies, political parties, educational institutions, sport teams, hospitals, museums, libraries etc. Government and political facilities such as White House and Pentagon are marked with NE-ORG. Organization names can be NE-FAC depending on context. Especially, hotels, museums, hospitals, libraries, churches and temples, commercial facilities, stock exchanges, etc. are often marked with NE-FAC. Adjectival forms of organization names are also marked with NE-ORG. Names of newspapers, magazines, and websites are always marked with NE-ORG regardless of whether they refer to the artifact or the organization. City names referring to a sport team are marked with NE-ORG while country names are always marked with NE-GPE.	(IBM) / (Capitol Hill) (White House) (Democratic Party) (Hilton Hotel) A (Democratic) candidate (New York Stock Exchange) (LIFE) / The (New York Times) a (Boston) shortstop (YouTube) / a (Wikipedia) article the (Japanese Government) the (Japanese government) (Obama administration) c.f. outside the (**NE-FAC** NYSE)
NE-GPE (GPE)	Country / Municipality Name	Names of geographical administrative entities including countries, villages, cities, states, provinces, prefectures, and other forms of municipalities. Island names are also marked with NE-GPE if they are administrative units. When you do not know whether a geographical entity is an	a (US) company / (Paris) southern (California) the (United Kingdom)

		administrative unit or not, mark it with NE-GPE.	
NE-LOC (LOCATION)	Non-GPE Location Name	Names of locations other than GPEs including celestial bodies, stars, continents, mountains, oceans, coasts, rivers, lakes, borders, etc. Named regions, areas, and neighborhood such as "Middle East", "Europe", and "East Village" are included in this category.	the (Hudson River) / (Europe) (Brighton Beach) (Long Island) / (Latin America) (Silicon Valley) (Earth) / the (Sun) (Greenwich Village)
NE-PRO (PRODUCT)	Product Name	Name of any product including non-commercial vehicles (automobiles, rockets, aircraft, ships). Although references including manufacturer and product names are marked as two different entities, products using only their manufacturer's name (as in "I bought a Ford") are marked with NE-PRO. Unlike the OntoNotes definition, financial products and services are also included in this category.	((NE-ORG Ford) (NE-PRO Taurus)) Two (NE-PRO Fords) (Space Shuttle Discovery) (Lotto 6)
NE-EVT (EVENT)	Event Name	Named events and phenomena including natural disasters, hurricanes, revolutions, battles, wars, demonstrations, concerts, sports events, etc.	(Hurricane Hugo) the (Vietnam War) the (Mexican Revolution) the (2016 World Cup) / (F1)
NE-ART (WORK OF ART)	Title	Titles of books, songs, films, plays and other creations such as awards, stock price indexes, and social security systems including health insurance systems or pension plans. Newspaper headlines are marked with NE-ART only when they are referential. Headlines used as "headlines" should not be marked. Series names, as in the "Harry Potter series", are also marked.	(Gone with the Wind) (Oscar) / (Nobel Prize) (Academy Award for Best Actor) I read a (Harry Potter) book. (Dow Jones Industrial Average) the (Union Flag)
NE-PJT (newly added)	Project Name	Names of projects, systems, ideas, policies, plans, movements, doctrines, religions, thoughts, systems, etc. Can be NE-ORG depending on context.	the (New Deal) project (Medicare)
NE-LAW	Name of	Named legal documents including laws,	(Bill of Rights)

(LAW)	Law	treaties, sections, and chapters.	the (Johnson Act) the (Warsaw Pact) (Article II of the Constitution)
NE-LAN (LANGUAGE)	Language	Any named language including programming languages.	a (Japanese) dictionary written in (English)
NE-DT (DATE)	Date	Date or period of 24 hours or more, including day, week, month, certain named period, season, year, etc. Age is also included in this category whether it is a noun, adjective, or adverb phrase. Definite articles should be included if they are needed to denote specific time. Definite articles should not be included when time-denoting words are premodified either by last, next or ordinal number, while definite articles should be included when premodified by other adjectives. Note, however, that definite articles should be included when a numerical expression is inserted between a time-denoting word and one of the premodifiers last/next/ordinal (e.g. last five years). Unmarked: "now", "past", "future", time-denoting adjectives and adverbs such as "recent", "recently", and "previously".	(Dec. 31 2001) / (Monday) (several years) / (today) (every year) / (last month) Mark, (aged 35), is … at (the age of 65) / (50 years old) (the previous quarter) / (Spring) (a year ago) / (1990-1995) (these days) / (the weekend) (23, 24, and 28 January) (the following/previous year) the (last/next/second year) (the past/next/previous 3 days) (the recent months) / (the 60's) the (summer months) the (21st century) / the (first day)
NE-TM (TIME)	Time	Times of day and time duration less than 24 hours. The expression "24 hours" is marked with NE-DT, NOT NE-TM.	(4 p.m.) / (this morning) (3 hours) / (the first 5 minutes) the (morning/afternoon/night) the (second quarter) the (42nd minute) / the (first half)
NE-PC (PERCENT)	Percentage.	Percent symbol or the word "percent" should be included.	(60 %)
NE-MO (MONEY)	Monetary Value.	Monetary units must be explicitly written. Rate expressions such as "per share" in "$5 per share" should not be included.	(50 yen) / ($ 1000.00) ($3) per share (3 dollars) per share
NE-QT (QUANTITY)	Quantity	Measurements including length, distance, area, weight, heat, velocity, temperature,	(about 5 miles) the (100-m) race

		byte size, etc. Units of measurement must be explicitly written.	(more than 4 acres) (60 miles an hour) (40 megabytes)
NE-OD (ORDINAL)	Ordinal Number	All ordinal numbers including adverbials.	(first) / (second) / (secondly)
NE-CD (CARDINAL)	Cardinal Number	Any numerical expression not categorized in any of the above categories. Numbers appearing as list item numbers or numbers used in addresses are also marked with NE-CD. Quantificational modifiers such as "only" and "just" are also included in the extent.	(half) / (hundreds) / (one-third) (100) people / (7) goals The Yankees won (4-1). (10 times) more Unmarked: many times
NE-UC (newly added)	Unclassified	Any named entity not categorized in any of the above categories.	(D-Calif.) / (R-NY) (Billboard Hot 100 Chart) the case (Brown vs. New York)

3. Extents

■Articles

Articles should not be included in the extent (exceptions: time and numerical expressions and NE-ART).

the Philippines => the (NE-GPE Philippines)

Definite articles before organizational names should not be included even when they begin with the upper-case letter "T".

He was interviewed by The New York Times. => … by The (NE-ORG New York Times).

When an article is part of a title, the article should be included.

the film "The Thing" => the film "(NE-ART The Thing)"

When not including articles does not conform to the Penn Treebank's bracketing guidelines, articles are included in the extent.

the United States of America
(NP (NP the United States)
(PP of
(NP America)))

In the phrase above, the article "the" and the noun phrase "United States" are first bracketed and therefore cannot be separated.

=> (NE-GPE the United States of America)

■Punctuations and Prepositions

Phrases separated by punctuations or prepositions should be marked as separate entities (exceptions: NE-DT, NE-TM, and names officially including punctuations or prepositions such as "United States of America").

It happened in Boston, MA. => It happened in (NE-GPE Boston), (NE-GPE MA).

Time expressions to denote certain times or durations should not be separated into their component parts. They should be marked in their entirety.

at two p.m. of the third day => at (NE-TM two p.m. of the third day)

Note, however, that if marking a phrase in such way does not conform to the Penn Treebank structure, the phrase should be separated.

The accident occurred at 3 p.m. on Monday. => … at (NE-TM 3 p.m.) on (NE-DT Monday).
* Should not be marked as (NE-TM 3 p.m. on Monday).

Prepositions used in range expressions such as "from", "to", and "between" are included in the extent.

from the 26th to 28th of April => (NE-DT from the 26th to 28th of April)

■Bracketed Phrases: Apposition

When a bracketed phrase is a named entity and apposition of the immediately preceding phrase, they should be separately marked.

International Business Machine [IBM]
=> (NE-ORG International Business Machine) [(NE-ORG IBM)]

3550 pounds [16100 kg] => (NE-QT 3550 pounds) [(NE-QT 16100 kg)]

Even when a bracketed phrase is inserted in the middle of a named entity or numerical/temporal expression, the whole phrase should NOT be marked as one NE.

next Sunday [November 25] morning
=> (NE-DT next Sunday) [(NE-DT November 25)] (NE-TM morning)

* Wrong marking: (NE-TM next Sunday [November 25] morning)

■Bracketed Phrases: Supplementary Information

When a numerical and temporal expression is followed by supplementary information within parentheses, the bracketed part should NOT be included in the extent.

at two p.m. (UTC) => at (NE-TM two p.m.) (UTC)

■Lowercase Letters

When a word(s) in a named entity happens to begin with a lowercase letter, that word(s) should also be included in the extent.

Central Intelligence agency => (NE-ORG Central Intelligence agency)
U.S. navy => (NE-ORG U.S. navy)

When such words are not included in the original name, they should not be included.

KGB security agency => (NE-ORG KGB) security agency

Words beginning with a lowercase letter in GPE/non-GPE location names and facility names should be included in the extent.

the (NE-LOC Hudson river) / (NE-FAC Waterman street) / (NE-GPE Normandy province)

4. Newly Added Categories

■NE-PER-N: Post and Job Title

Names of posts and job titles.
Basically, each word in a NE-PER-N should begin with a capital letter, but in some cases, words beginning with a lowercase letter are also considered NE-PER-Ns as in some of the examples below. Honorific titles such as Mr. and Dr. are not marked.

Examples (NE-PER-Ns are marked within round brackets) :
ambassador to: (ambassador to France)
Baron + Person name: (Baron) Brown
Coach + Person name: (Coach) Chip Kelly
district court judge: (Western Australian district court judge)
governor: (California governor)
government attorney: (government attorney)
FA chief executive: (FA chief executive)
FIFA general secretary: (FIFA general secretary)
FIFA President: (FIFA President), (FIFA president)
Foreign minister: (Foreign minister), (foreign minister)
leader: (Conservative Party leader)
Member of Parliament: (Member of Parliament)
President: (President) Bush, (US President) Bush, (US president) Bush
President of: (President of the United States), (president of the United States)
prince: (Moroccan Prince)
Professor: Professor (NE-PER Brown)
Senator: (NE-GPE Arizona) (NE-PER-N Senator)
spokesman: (White House spokesman), (Utah Department of Transportation spokesman),
(Pentagon spokesman)
Lowercase lettered "spokesmen/women/persons" are marked with NE-PER-N only when they are working for government agencies. Spokespersons working for other types of organizations should not be marked: (ORG IBM) spokesman
White House adviser: (White House adviser)

■CHA : Name of Non-human Creature

Names of creatures other than humans including animals and fictional non-human characters.

Names of gods and goddesses or names of human-like fairies are marked with NE-PER, NOT NE-CHA.

■ADD: Contact Information

Addresses, phone numbers, email addresses, URLs, etc.
Unlike OntoNote annotation, where addresses are broken down into several parts, addresses should be marked with NE-ADD as one entity. Each element in an address should not be marked with NE-ADD when it is used singly.

23 Spring Street, New York, New York => (NE-ADD 23 Spring Street, New York, New York)

It is located on Spring Street. => It is located on (NE-FAC Spring Street).

New York is a big city. => (NE-GPE New York) is a big city.

Combinations of a number and location name (street name) are marked with NE-ADD.

It is 23 Spring Street. => It is (NE-ADD 23 Spring Street).

When URLs are used as website names, they should be marked with NE-ORG, NOT NE-ADD.

■PJT: Project

Names of projects, plans, movements, doctrines, (new) religions, thoughts, systems, etc. Five major religion names (Buddhism, Christianity, Hinduism, Islam, and Judaism) are NOT marked.

■UC: Unclassified

Any named entity not categorized in any of the rest of the categories.

5. Problematic Cases

■Head Sharing

Example: Scottish and Irish FAs

When two or more entities share the same head, the entire phrase is marked as one named entity so that the treebank structure can be maintained, not like the OntoNotes guidelines where only the named entity that is closest to the head word is marked.

(NE-ORG Scottish and Irish FAs)

NE-DTs and NE-TMs are also marked the same way.

June 7 and 8 => (NE-DT June 7 and 8) c.f. (NE-DT June 7) and (NE-DT June 8)

■Named Places: NE-LOC or NE-GPE

When you do not know whether a geographical entity is an administrative unit or not, mark it with NE-GPE.

■NE-DT and NE-TM

Adverbs and adjectives to denote specific time are included, but prepositions, such as "for" in "for two days", are NOT included.

two years ago => (NE-DT (ADVP (NP two years) ago))
a two-year-old girl => a (NE-DT two-year-old) girl
Tom Brown, aged 32, is … => Tom Brown, (NE-DT (ADJP aged 32)), is …

in two hours => in (NE-TM two hours) *Wrong marking: (NE-TM (PP in two hours))

6. References

[1] Ralph Weischedel et al. 2012, *OntoNotes Release 5.0*", accessed 1 June 2017, <https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf>

[2] Raytheon BBN Technologies 2004, *OntoNotes Named Entity Guidelines Version 14.0,* accessed 1 July 2017, <https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxjbHRsYW5ub3RhdGlvbnN8Z3g6MzVlOGFkMjQ4ZWMxM2Y5MA>