Raytheon
BBN Technologies

OntoNotes Named Entity Guidelines

VERSION 14.0

Contents

1 Overview

The purpose of this OntoNotes task is to tag “Named Entities”, or proper name 'mentions', in text. Texts annotated in this way will help the computer learn to correctly identify multiple mentions of the same entity.

OntoNotes annotates 18 types of named entities. This list of types combines lists of categories originally used for Question Answering and in the ACE evaluation, including the EDT types Person, Organization, Location, GPE, and Facility, with some modifications. The list also includes the MUC categories Money, Percent, Time and Date, again with some modifications. The remaining categories were determined in part based on categories found in literature on the question answering task and in part based on BBN’s own examination of the data. These mentions are annotated over the Treebank tokens (and so do not necessarily align with nodes in the Treebank parse tree—especially in case of flat NP structures).

Each of the 18 categories is described in the next section. The guidelines are mostly language independent, and we provide English and Chinese examples throughout, although there is also a supplement covering Chinese language-specific issues at the end of the document.

2 Named Entities

2.1 Person Name (PERSON)

These are proper names of people, including fictional people, first names, last names, individual or family names, unique nicknames. Generational markers (Jr., IV) are included in the extent, while honorifics (Ms., Dr.) and occupational titles (President, Secretary) are NOT included. Following are some examples. The named entity mentions are marked within square brackets.

Examples

Dr. [Bob Smith, Sr.]
[Henry IV]
Secretary [Rice]
The President (no markup)
总统 [布什]
[刘江永]教授
[杨]先生
[布什]政府
[柯林頓]當局

2.2 Nationality, Other, Religion, Political (NORP)

This type represents adjectival forms of GPE and Location names, as well as adjectival forms of named religions, heritage and political affiliation. Also marked are head words which refer to people using the name of an entity with which they are affiliated, often a GPE or Organization. The distinction between NORP and other types is morphological. "American" and "Americans" are adjectival nationalities, while "America" and "US" are GPEs, regardless of context.

Examples

[American] forces
[Eastern European] cuisine
the [Democratic] candidate
the three [Democrats]
a [Chinese-American] dentist
a [Jewish] doctor
the [Muslims]
[俄罗斯]总统
[亚洲]国家
日本的[防卫厅]长官
[大陆]方面
[中国]人
[台湾]同胞
[台北]市長

2.3 Facility (FAC)

Names of man-made structures: infrastructure (streets, bridges), buildings, monuments, etc. belong to this type. Buildings that are referred to using the name of the company or organization that uses them should be marked as FAC when they refer to the physical structure of the building itself, usually in a locative way: "I'm reporting live from right outside [Massachusetts General Hospital]"

Examples

[5th Avenue]
[Logan Airport]
[Tobin Bridge]
the [Lincoln Memorial]
[I-95]
[靖国神社]
[虹桥机场]
第一 时间赶到[北京医院]
[宾西法尼亚大街]
[士林 夜市]

2.4 Organization (ORG)

These are names of companies, government agencies, educational institutions, sport teams. Names of hospitals, museums, and libraries should be marked, unless the mentions are clearly referring to the building in a locative way. Adjectival forms of organization names are included, as are metonymic mentions of the buildings or locations associated with these organizations.
A group, team, force, etc. must be officially sanctioned in some way to be classified as an Organization. Organized crime groups, such as the Mafia, are not marked. Terrorist groups such as Al-Qaeda, however, should be marked.

Examples

[Congress]
the [Senate]
the [Supreme Court]
the [University of Michigan]
[Bank of America]
the [New York Times]
the [Museum of Fine Arts]
the [White House]
the [Pentagon]
the [Kremlin]
[Capitol Hill]
[白宫]
[世界贸易组织]
[中央电视台]
[中文国际频道]
[ＡＰＥＣ]会议
[黑旗军]
[联合舰队]
[国会山]

2.5 Geographical/Social/Political Entity (GPE)

Names of countries, cities, states, provinces, municipalities. In cases where the GPE name is modified, such as "southern California," [California] is marked as a GPE name and there is NO other markup.

Examples

the south of [Baghdad]
[韩国釜山]
[莫斯科]
[台湾]

2.6 Location (LOC)

Names of geographical locations other than GPEs. These include mountain ranges, coasts, borders, planets, geo-coordinates, bodies of water. Also included in this category are named regions such as the Middle East, areas, neighborhoods, continents and regions of continents. Do NOT mark deictics or other non-proper nouns: here, there, everywhere, etc. As with GPEs, directional modifiers such as "southern" are only marked when they are part of the location name itself.

Examples

[South Boston]
[Eastern Europe]

north [Chinatown]
[朝鲜半岛]
[亚洲]
[华北]
[澎湖列岛]
[台湾岛]
在[国会山]上
[火星]
[地球]
[信义 商圈]

2.7 Product (PRODUCT)

This can be name of any product, generally a model name or model name and number. Named foods are also included. Credit cards, checking accounts, CDs, and credit plans are NOT marked. References that include manufacturer and product should be marked as two separate named entities, ORG + PRODUCT: [Apple] [iPod], [Dell] [Inspiron], [Ford] [Mustang].

Examples

[iPod]
[Inspiron 1700]
[Mustang GT]
[Velveeta]

2.8 Date (DATE)

Used to classify a reference to a date or period, etc. Age also falls under this category, even when it’s a noun phrase referring to a person: the 5-year-old, 5 years old, Jane Doe, 5, etc. Extent should include modifiers & prepositions that denote specific time, such as [2 days ago], [the past two days], but not those that mark duration, such as "for [2 days]." Do not separate mentions into their component parts: [November 2, 2001] and [from the fall of 1962 through the spring of 1967] should be marked in their entirety, without separate markups for "November 2," "2001, "the fall," "1962," "the spring," and "1967." Dates that are part of rate expressions such as "once per **day**" or "twice a **year**" should NOT be marked.

Examples

[November 2 2001]
[January]
[Monday]
[seventies] fashion
[the 1940’s]
[several years ago]
for [two months]
since [last week]
[tomorrow]

in [spring]
[this past summer]
[the fall of 2008]
[winter]
[our fourth quarter]
[昨天]
[今年]
[今年的十月十七号]
上任[四年]
[四十周年]
[日韩友好年]
[今年一到八月份]
[春节]
[从二零零一年到二零零五年]
[周末]
[台湾光复节]
[第四年]
[５１岁]
經過[四天]
卑劣 当选 后 的 [前 七 个 半 月]
[数 年] 后

2.9 Time (TIME)

Any time ending with "a.m." or "p.m." If the "a.m." or "p.m." is explicit, it must be tagged along with the numbers. Other times of day (units smaller than a day) and time durations less than 24 hours are also marked: morning, noon, night, 3 hours, a minute. Do not separate mentions into their component parts: [the evening of July 4th] and [5:00am, April 5, 2008] should be marked in their entirety, without separate markups for "evening," "July 4th," "5:00am," and "April 5, 2008"

Examples

[1:00 a.m.]
[yesterday morning]
[noon]
[this evening]
[night]
[three hours]
[a minute]
[昨天晚上]
[明天早晨]
[半个小时]
每[五分鐘]
２００１年 ９月 的 [一 个 早晨]

2.10 Percent (PERCENT)

Any percentage. A percent symbol or the actual word percent must be explicit and included in the extent. If the percent is implicit, the number should be marked CARDINAL.

Examples

[50%]
[a hundred and twenty percent]
[百分之三点二]

2.11 Money (MONEY)

Any monetary value including all monetary denominations. The monetary unit must be explicit and included in the extent. If the monetary unit is implicit, the number should be marked CARDINAL. Only values should be tagged—generic references to money should not. For example, in "money invested overseas," there is no markup for "money." In rate expressions such as "$ per unit," the unit should not be included in the extent. For example, in "$3 per share," the extent is [$3].

Examples

[50 yen]
[one million dollars]
[17,000 British pounds]
[$10.20]
[ten cents] apiece
[将近一千亿美元]
[五百万 两 ， 银两]

2.12 Quantity (QUANTITY)

Used to classify measurements with standardized units. If the unit of measurement is implicit or non-standard (3 handfuls, 2 football fields, 10 points), the number should be marked CARDINAL. One exception to this rule is formulaic references to the age, height, and weight of a person: Joe Smith, 44, five ten, two twenty. In this instance, [five ten] and [two twenty] should be marked QUANTITY. (While [44] should be marked DATE)

Examples

[4 miles]
[4 grams]
[4 degrees]
[4 pounds]
[4 ounces]
[九十多公里]
[四五级]的风
[零下十一度]

2.13 Ordinal (ORDINAL)

All ordinal numbers, including adverbials.

Examples

in the [first] place
[third] in line
[fourth] place
[secondly]
[第五]次会晤
[第五]个人
[第一]任

2.14 Cardinal (CARDINAL)

Numerals, including whole numbers, fractions, and decimals, that provide a count or quantity and do not fall under a unit of measurement, money, percent, date or time. For "Nasdaq composite fill [1.39] to [451.37]." the numbers are marked CARDINAL because there is no monetary unit. Headless numerical phrases are also covered in this category: "reducing employment from [18,000] to [16,000]." Numbers identifying list items should also be included. Pronominal mentions of "one" should not be tagged.

Examples

[about half]
[hundreds] and [hundreds]
[one-third]
[four]
[exactly 4534.5]
[两]位领导人
[两]岸
外交[三]原则
[一]根支柱
[二]到八度
[三分之一]
[半数]
[六十五万]人
[二十]次
[成千上万]
[三打]
[二十多] 枚隕石
[幾千]倍輻射

2.15 Event (EVENT)

Named hurricanes, battles, wars, sports events, attacks. **Metonymic** mentions (marked with a ~) of the date or location of an event, or of the organization(s) involved, are included:

Examples

"the impact of [nine-eleven]"
    ~ the events of September 11, 2001
"Lincoln's speech after [Gettysburg]"
    ~ the battle of Gettysburg
"[Enron] has made us all suspicious"
    ~ the Enron scandal
[WWII]
[Hurricane Katrina]
the [New York City Marathon]
the [Oklahoma City bombing]
last year's [Oscars]
the [2007 World Series]
[ＡＰＥＣ 第十三 次 领导人 非正式 会议]
[冷战]
[甲午战争]
[辛亥革命]
热带 风暴 [“ 尤特” （ ＵＴＯＲ ）]

2.16 Work of Art (WORK_OF_ART)

Titles of books, songs, television programs and other creations. Also includes awards. These are usually surrounded by quotation marks in the article (though the quotations are not included in the annotation).

Newspaper headlines should only be marked if they are referential. In other words the headline of the article being annotated should not be marked but if in the body of the text here is a reference to an article, then it is markable as a work of art.

Examples

[The Empire Strikes Back]
the [Bible]
[Blue Moon]
[Larry King Live]
[Nobel Peace Prize]
her [Emmy]
an [Oscar] nomination
[今日关注]节目
[《 眼镜蛇 ２ 》]这 本 书

2.17 Law (LAW)

Any document that has been made into a law, including named treaties and sections and chapters of named legal documents.

Examples

the [Bill of Rights]
[IRS code 4]
The [1988 trade act]
the [Johnson Act]
[Article II of the Constitution]
the [Warsaw Pact]
the so-called special [301 provision] of the act
[马关 条约]
[宪法]

2.18 Language (LANGUAGE)

Any named language.

Examples

[Latin]
[Arabic]
[Filipino]
[西语]裔
[中] [英文]版

Supplement A: Guideline CheatSheet

Markables/Extents

- Mark the full name for name categories.
- Only mark proper names, NOT nominals.
- If a common noun is modified by a proper PreMod, mark ONLY the PreMod, and categorize it based on its own meaning
  - the [Clinton] administration – [Clinton] is marked as PER
  - a [Supreme Court] ruling – [Supreme Court] is marked as ORG
- For categories dealing with numbers, mark the full expression, including modifiers
  - [Just 30%] of those polled favor the proposal
  - In the winter the average temperature in Baghdad is [about 10°C]
  - [Almost 700] people attended the conference
  - Many of the talks, which numbered [over 70], were focused on outsourcing

Unmarkable

- Do NOT include determiners or articles in the extent
  - For "the White House" and "the US," mark only [White House] and [US]
- Pronouns and pronominal elements like anaphoric "one," "someone," "everyone," "others," etc.
- Names embedded in atomic names
  - Mark [Bank of America] as ORG, but do not mark "America"
- Occupational titles and honorifics should NOT be included in extent of PERSON entities
  - For "President Bush" and "Mr. Bush," mark only [Bush] as PER
- Contact Information
  - Nothing should be marked in "cnn.com," "1600 Pennsylvania Ave," or "1-800-555-1234"

Commonly Confused Entities

- **GPE v. LOC**
  - Most named places are GPEs (Geo-Political Entities)
  - Locations lack governmental structure (Geo, but not Political)
    - A country is a GPE, but a continent or a region is a LOC
    - A city is a GPE, but a neighborhood is a LOC

- **NORP v. GPE/ORG**
  - NORPs are generally adjectival
  - Countries are GPEs, but Nationalities are NORPs
  - Religions, political parties, and other named groups are ORGs, but their members are NORP
    - "The [Democrats] have yet to choose a nominee" = NORP

    - "The [Democratic Party] has yet to choose a nominee" = ORG
  - This can be especially ambiguous with groups named "The _______s"
    - "He always dreamed of joining the [Marines]" = ORG
    - "The [Marines] completed their mission and returned to their base" = NORP

- **NORP v. LANGUAGE**
  - Often, the adjectival form of a nationality is identical to the name of that country or people's language. Annotators should use context to determine which tag is appropriate
    - the [English] people = NORP
    - the [English] language = LANGUAGE

- **DATE v. TIME**
  - Mentions longer than or equal to 24 hours should be marked DATE

- **ORG v. FAC**
  - If a building houses an organization of the same name, mentions should be marked ORG, unless clearly referring to the physical building alone in a locative way

- **FAC v. LOC**
  - Facilities are man-made, Locations are natural
    - A canal is a FAC, but a river is a LOC
    - A building is a FAC, but a cave is a LOC
    - A road is a FAC, but a mountain pass is a LOC
    - A farm is a FAC, but a forest is a LOC

- **ORG v. WORK OF ART**
  - Television broadcast companies are ORGs
  - Television programs can be ORGs or WORK OF ART, depending on context
    - "[60 Minutes] tried to contact him for an interview" = ORG
    - "I love watching [60 Minutes]" = WORK OF ART
  - Specific episodes or segments of television programs are WORK OF ART

- **ORG v. PRODUCT**
  - Makes, models, and versions are PRODUCTS; the company that produces them is ORG
    - Mustang is a PRODUCT, but Ford is an ORG
    - Pepsi is a PRODUCT, but Pepsi Co. is an ORG
  - If an organization produces a product of the same name, annotators should use context to determine whether a mention should be marked ORG or PRODUCT

- **EVENT v. DATE/LOC/FAC/ORG**
  - It is possible to refer to an Event using the Date when it occurred, the Location, or
  - Facility where it occurred, or the Organization(s) involved. These mentions should be marked EVENT if the context makes it clear that the intended referent is the Event itself.
    - "After [Columbine], many schools installed metal detectors" = EVENT
    - "[Columbine High School] was the site of a deadly shooting" = FAC

- **CARDINAL v. QUANTITY/MONEY/PERCENT**
    - Quantites, Money, and Percents MUST have explicitly-mentioned units, otherwise they should be marked CARDINAL (even the implicit unit seems obvious)
        - She wanted [a hundred percent] <PERCENT>, but he only gave her [**fifty**] <CARDINAL>
        - She wanted [a hundred dollars] <MONEY>, but he only gave her [**fifty**] <CARDINAL>
        - She wanted [a hundred gallons] <QUANTITY>, but he only gave her [**fifty**] <CARDINAL>

- **CARDINAL v. DATE/TIME**
    - Dates and Times MUST have explicitly-mentioned units of time, otherwise they should be marked CARDINAL (even if the implicit unit seems obvious)
        - She wanted to stay for [two weeks] <DATE>, but had to leave after [one] <CARDINAL>
        - She wanted to stay for [two hours] <TIME>, but had to leave after [one] <CARDINAL>

Supplement B: Chinese Addendum

How to Determine NORPs

NORPs are only used to indicate Nationality (including race, clan, or tribe), Religious, or Political affiliation. NORPs should only be modifiers of people: almost every NORP will modify the character for "person" or "people," (i.e., 人 or 族), although there are some exceptions.

1. Is the entity in question a GPE, a Location or an Organization? If not, then it cannot be a NORP.

    a. [China] person [中國]人 = NORP
    b. [Europe] person [歐洲]人 = NORP
    c. [Democrat party] people [民主黨]人 = NORP
    d. [Han] race [漢]族= NORP
    e. [World Cup] air [世界盃]空氣= EVENT

2. If the entity is an Organization, is it a religious or political one? If not, then it cannot be a NORP

    a. [Christian religion] person [基督教]徒= NORP
    b. [Communist Party] person [共產黨]人= NORP
    c. [BBN] person [BBN]人員 = ORG
    d. [FBI] person [FBI] 人員 = ORG

3. If the entity is a GPE, Location, or Religious/Political Organization, is it modifying "person" or "people"? If so, then it is a NORP. If not, see #4.

    a. [Mexico] people [墨西哥]人= NORP
    b. [Mexico] cuisine [墨西哥]菜= GPE
    c. [Mexico] history [墨西哥]歷史= GPE
    d. [Mexico] dance [墨西哥]舞蹈= GPE
    e. [Islam religion] people [伊斯蘭教]徒= NORP
    f. [Islam] culture [伊斯蘭]文化= ORG
    g. [Islam] society [伊斯蘭]]社會= ORG
    h. [African descendant American] person [非洲裔美國]人= NORP

4. If the entity is modifying a job title that is directly related to the entity itself, then it cannot be a NORP. If the entity is modifying a job title, but the title is unrelated to the entity itself, then it is a NORP.

    a. [Republican party] person [共和黨]員 = NORP

b. [Republican party] chairman [共和黨]主席= ORG
c. [America country] doctor [美國]醫生= NORP
d. [America country] President [美國]總統= GPE
e. [Jewish] doctor [猶太]醫生= NORP
f. [Jewish religion] rabbi [猶太教]祭司= ORG
g. [Foreign Affairs Ministry] chief [外交部]長= ORG
h. [FBI] agent [FBI]特工= ORG
i. [Washington] resident [華府]居民= NORP
j. [Washington] insider [華府]圈內人 = GPE

DATE/TIME

DO mark things like "in recent years" [近幾年], "the next few days" [接下來幾天] and "in [those days]"

DO NOT mark "recently" 最近 or "soon" 不久 or "at that time"當時

However, on [that night] [当晚] should be tagged as a Time.

Names of Dynasties should be considered periods of time:
[Qing Dynasty] [清朝]= DATE
the [Qing] government [清]政府= DATE

Named years should be marked:
[year of the dog] [狗年]= DATE
[Beijing Olympics year] [北京奥林匹克年]= DATE

Durations should be marked inclusively:
[from 1995 to 2000][從1995年到2000年]

Approximations should be included in tag:
[less than one year] [不到一年]

Otherwise, prepositions should not be included:
beginning in [1995]自[1995年]

If a mention includes a time, day, month, and year, all these should be included in a single tag:
"the morning of Sept 11, 2001" = TIME

However, these mentions should not be tagged as a single entity, since the day is missing:
[one morning] in [September, 2001] = TIME + DATE

If the time description includes a time zone, include it in the mention:
[Eastern time 8:30am]

However, if the time description includes a city name, mark the city as a GPE:
[Beijing] time [8:30am] = GPE + TIME

ORDINAL

DO NOT mark words used to begin or end a list if they are not numeric in nature.
For example, do not mark "finally,"最終 "last,"最後一次 or "head"首次

CARDINAL

DO NOT mark "both" or "pair," 雙方even though they are understood to mean two.
DO mark Arabic numerals
DO mark numbers that are part of an idiomatic or fixed expression. 不管[三七二十一]

ONLY mark "one" if it is used to indicate the number
DO NOT mark "one" when it is used as a pronoun:
"One should always do one's homework"
DO NOT mark "one" when it is used as an article:
"One day it was raining, and the next, it was sunny"

Shortened or Elided Forms

Only tag mentions that include proper names, even if the proper name is implicit.

[China country] [中國]= GPE "country" 國家= no tag

[Communist Party] [共產黨]= ORG "party" 黨= no tag

[Hong Kong Special Administrative District International Consultant Committee][香港 特区行政长官特国际顾问委员会]= ORG
[International Consultant Committee] [国际顾问委员会]= ORG
“the committee” 该委员会or 委员会"committee" = no tag

[Beijing Anti-Drug Volunteer Team] [北京市禁毒志愿者总队] = ORG
[Anti-Drug Team] [禁毒志愿者队伍]= ORG
"team" 总队= no tag

[Central Committee of the CPC] [中共黨中央]= ORG
[Party Central] [黨中央]= ORG (unless context is outside mainland China)
[Central] [中央]= ORG (unless context is outside mainland China)

[Treasury Department] = ORG
[Treasury] = ORG
"department" = no tag

Please note that “Mainland” is considered a proper name.
[Mainland] government [大陸]政府= GPE

ORG

Organization names that contain GPEs should be marked as one ORG:

[Taiwan Ministry of Foreign Affairs] [台灣的外交部]
[US White House] [美國白宮]
(if the meaning is not the building itself but the government)
[Israel Army] [以色列軍隊]
(if the meaning is the entire military unit, not a subset of soldiers)
[American Embassy] [美國大使館]
(if the meaning is not the building itself but the administrative unit)

If the GPE is NOT part of the Organization name, but rather serves to clarify the location of the Organization, it should be marked separately:

[British] [Cambridge University] [英國][劍橋大學]= GPE + ORG

[Chinese] branch of [Voice of America] [美国 之 音] [中文]部 = Lang + ORG

For industries that are frequently mentioned using the name of the GPE, LOC, or FAC where they are located, the mention should be tagged as ORG unless it is clear the speaker meant to refer ONLY to the physical location:

[Broadway] musical [百老匯]音樂劇= ORG
[Wall Street] banker [華爾街]銀行家= ORG
[Hollywood] actor [好萊塢]演員= ORG

Newspapers and magazines should ALWAYS be tagged as ORG:

today's [New York Times] 今天的[紐約時報]= ORG
an issue of [Newsweek] 一期[新聞週刊]= ORG

HEAD-SHARING MENTIONS

When two mentions share the same head word, it is impossible to mark them as two separate entities. Therefore, we can only mark the mention that is closest to the head word. For example, "South and North Korea" 南北韓share the head word "Korea,"韓 so we can tag [North Korea] [北韓] as a GPE, but we cannot tag "South."南

LONG MENTIONS

Long names of administrative units that include several smaller place names should be tagged as one long mention:

[Liaoning Province Fuxin City Pingan Coal Mine] [辽宁省阜新市平安煤矿] = FAC

Names of cities should be tagged separately from names of countries:

For "Norway capital Oslo挪威首都奧斯路, mark [Norway] [挪威] as a GPE, and [Oslo] [奧斯路] as another GPE.

For "China Beijing"中國北京, mark [China]中國 as a GPE, and [Beijing]北京 as another GPE.

[Indian-occupied Kashmir] [印控克斯米尔] should be tagged as a GPE, but [Kashmir] area [克斯米尔] 地區 is a Location.

the [West Bank of Jordan River] [约旦河 西岸] or [West Bank] [西岸] is a Loc.

For "old Bush" 老布什and "little Bush" 小布什, just tag [Bush] [布什]

For address, break it down into several mentions. For example, 2899 Xietu Road, Room 207, Shanghai City"上海市斜土路二八九九 号二零七室", tag [Shanghai City] [上海市] as a GPE, [Xietu Road] [斜土路] as a Facility, [2899] as a Cardinal, and [207] as another Cardinal.

Don't include quotation marks in mentions.