DEFT Chinese Rich ERE Annotation Guidelines: Entities V1.0

Linguistic Data Consortium

May 4, 2015

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): 普京, 刘敬民
- Nominal Mention (NOM): 有些城市
- Pronoun Mentions (PRO): 他, 他们

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]$_{\text{GPE}}$ cars"), possessive determiners ("[her]$_{\text{PER}}$ three convictions"), or prepositional complements ("at [the beach]$_{\text{LOC}}$").
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 '巴基斯坦人民' would have two entity mentions tagged: '[[巴基斯坦]$_{\text{GPE}}$ 人民]$_{\text{PER}}$' and '他的部下' would be tagged as '[[他]$_{\text{PER}}$ 的部下]$_{\text{PER}}$'. 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 ($_{\text{PER, ORG, GPE, LOC, FAC,}}$). 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.

The 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 ending such as “的” as in “中国的” should be excluded from the mention extent of “中国”. 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.

Mentions are frequently nested; that is, they will contain mentions of other entities. For example, the phrase [[百度] $_{\text{ORG.NAM}}$的总裁] $_{\text{PER.NOM}}$is a mention of an entity of type Person, and contains the name "*百度*", a mention of an entity of type Organization. It is even possible for a noun phrase to contain an embedded mention of the same entity. For instance, the phrase [那个教会[自己] $_{\text{PER.PRO}}$编程的历史学家] $_{\text{PER.NOM}}$ evokes a Person entity with two mentions: the entire phrase, and the words "*自己*".

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 (such as “该”, “这个”) 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. 胡锦涛’s name is “胡锦涛” not “前国家主席胡锦涛”).

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

Some examples of named entity mentions follow:

[贾斯汀·比伯] $_{\text{PER}}$

[四川] $_{\text{GPE}}$

[北京]$_{\text{GPE}}$

[红十字会]$_{\text{ORG}}$

[中国] $_{\text{GPE}}$

[基督教协会] ORG

奥运会将在 [北京]GPE举行

Mentions of entities with names referred to as alias (别号, 绰号) may also be tagged NAM.

[胡 core]PER *as alias of 胡锦 涛*
[习大大]PER *as alias of 习近平*
[温宝宝]PER *as alias of 温家宝*
[枫叶国]GPE的自驾生活 *as alias of* 加拿大
[韩棒子]GPE凭什么与[中国]GPE一战 *as alias of* 韩国
[狗狗] ORG 股票飙升 *as alias of* 谷歌公司

There are also cases where one proper name is nested in another proper name. In such cases, we will annotate all taggable named entities. Note that the extent of each name should include all texts that are required for such name. For example:

[[中国] NAM.ORG社会科学院] NAM.ORG has two names: 中国社会科学院 and 中国
[[北京] NAM.ORG大学] NAM.ORGhas two names: 北京大学 and 北京

A proper name is always taggable as long, even when it is part of an atomic mention of an entity, regardless of whether or not the entity is taggable under the current specifications. For example:

[意大利] GPE 面	Italian pasta
[法国] GPE 红酒	French wine
[诺贝尔] PER 奖	Noble Prize

If one proper name is modifying another proper name, then they need to be tagged separately. Note that the extent of each name should only include string of texts of such name.

[[北] NAM.GPE大] NAM.ORG [中文系] NAM.ORG has two names: 北京大学, 北京 and 中文系
[ ] GPE[ ] GPE has two names: 辽宁and 阜新
[美国] GPE [佛罗里达州] GPE has two names: 美国and 弗罗里达州
[中国] GPE [红十字会] ORG has two names: 中国 and 红十字会

When a person proper name is followed by a title or a proper name is preceded by a position, the title is not part of the extent of the proper name. Refer to section 3.2 for detail of TITLE entity.

[李克强]NAM [总理]TTL
[张]NAM [教授] TTL

中国[总理] TTL [李克强] NAM

Sometimes, names may be conjoined and part of the name is shared, as in 北京和清华大学，南京和武汉大桥. We will annotate mentions like this as follows:

[北京] NAM.ORG和[清华大学] NAM.ORG
[南京] NAM.LOC和[[武汉] NAM.GPE大桥] NAM.LOC

Note in the second example, 武汉 is tagged as GPE, but 南京 is tagged as LOC, as we 南京 here refers to 南京大桥. Since we can't double tag the same mention as two different entities, there is no way to capture 南京 as 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.

Some examples of possible nominal mentions are given below:

[这些中国人] PER

[有的国家]GPE

[作家协会成员]PER

[有些城市]GPE

[一名中学生]PER

[媒体发言人]PER

[口腔癌患者]PER

[禽流感定点医院]ORG

[一些媒体记者]PER

[一群恐怖分子]PER

[那座楼]LOC

[这个地方]LOC

[这些地区]GPE

[一些组织]ORG

NOTE: Noun phrases beginning with a pronominals (see section 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’).

NOTE: Appositives and certain other NAM+NOM combinations expressing identity or categorization should be tagged with care. Do not include nominal mention extents with named mention extents; tag them autonomously:

［药家鑫］$_{NAM}$ ［这个人］$_{NOM}$
［诺贝尔奖获得者］$_{NOM}$ ［莫言］$_{NAM}$
［美国首都］$_{NOM}$ ［华盛顿］$_{NAM}$
［香港］$_{NAM}$ ［这个城市］$_{NOM}$
［正在崛起的新兴城市］$_{NOM}$［宁波］$_{NAM}$

A lot of cases, a nominal mention may contain another mention either of the same entity or different entity. All those mentions need to be tagged, but their extents need to follow the rules of each mention type. For example,

[[我]$_{PRO}$的房子］$_{NOM}$
[这座位于[[中国]$_{NAM}$大西北]$_{NOM}$的城市]$_{NOM}$
［美丽的[[北］$_{NAM}$大]$_{NAM}$校园］$_{NOM}$
[代表[俄罗斯]$_{NAM}$的运动员］$_{NOM}$现在进场

2.2.1 Tagging Entity Mention Heads in Nominal Mentions

In addition to the extent of the nominal phrase, the head of the phrase must be marked. In

The hurricane destroyed the new glass-clad skyscraper.

the full extent of the mention is

The new glass-clad skyscraper

and the head is *skyscraper*. In the examples below, the head will be marked by underscoring.

Determining the head of a common noun (NOM) in Chinese is intrinsically hard and there are no agreed rules on word segmentation. Some basic rules of thumb are:

1. Never tag suffixes alone as the head (suffixes such as 家 in 作家, 者 in 记者, 人 in 候选人,残疾人, 室 in 研究室 etc

2. Bi-syllabicity is always preferred by the Chinese speaker. If possible, treat a bi-syllabic string as one word except when the syntactic or morphological properties favor the other way around, for example:

    [各地] [各国] [各党]
    [本校] [全球] [全国] [当地] [每人]
    but [本单位]

3. For a tri-syllabic string "ABC", do similarly as in 3:

    [饺子馆] [料理店] ［石油化工厂];

4. If a measure word can be inserted between the number without changing the meaning, separate the number from the head. Otherwise, treat it as a single head, for example:

    [三人] [数人] [三国] [四省] [两国] [多国]
    [三排] [一方] [三者] [一行]

5. When the string is a pronoun and a noun, if both the pronoun and noun are monosyllabic and the noun is bound, treat it as a single head; otherwise, treat the noun only as a head, for example:

    [我校] [我国] [我单位] [我公司] [我们国家]

6. For the pattern X + N, where X modifies the N, treat X + N as a single head if X is a prefix. If x is a non-predicative adjective, treat X + N as a single head if both X and N are monosyllabic when the meaning of X + N is noncomposistional. A simple test of compositionality is to insert 的 between X and N to see whether the meaning changes. If it doesn't, treat N alone as head. For unclear cases, if both X and N are monosyllabic, treat X + N as a single head. For example:

    [阿爸] [女人][强队] [小媳妇] [大海] [小国] [大国] [前总统] 原在华老挝难民]

7. For localizer + N and N + localizer, if both localizer and N are monosyllabic, treat the string together as a head. For example:

    [前院] [境外] [国内] [海外] [后花园]

8. Ask.

2.2.2 "Headless" Mentions

A headless mention is an NP without an overt syntactic head, but the head (in Chinese) is recoverable syntactically from the context even though the version that includes the head may not be preferred for the context by the native speaker.

Although these mentions are technically headless, we will mark the rightmost character or word of the headless mention as the head in annotation.

A: 你喜欢红色的小轿车吗？

B: 我喜欢白色的。

Here [白色的] is a headless mention and we tag the last word 的 as the head.

Other cases include “number + classifier” (三个没来), “bare quantifier” (很多没来), etc. Note that Chinese does not have an equivalent to the partitive constructions described in the English guidelines.

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.):

我	我们	咱	咱们
你	您	你们	您们
他	他们	她	她们
它	它们	自己	自身
这	那	这些	其它/他
其余	本人	那些	这儿
那儿	这里	那里	大家
前者	后者	其间	俺
人家	俺们	本席	敝人

全新的[我们] PRO
那时的[他] PRO
80 后的[俺们] PRO

The reflexive morpheme 自己 (“self”) can be used in two ways, as a reflexive pronoun or as an adverbial to serve to contrast with oneself with others. Regardless of what function it serves in a sentence, it’s always marked as pronominal if the entity it refers to falls into one of the seven types.
However, the morpheme 自己or 本人can appear immediately after a person pronoun or nominal (such as **他自己**,**面瘫患者自己**, 我本人) For simplicity, we mark the two mentions separately, even though they refer to the same entity.

[面瘫患者] NOM

[自己] PRO
[我] PRO [本人] PRO
[他] PRO [自己] PRO

Noun phrases beginning with a pronoun listed above, like “这组”, “别国”, “少数与会者”, will be tagged as nominals**.**

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 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...

  对于长时间耽误请求外援的作法，[人们]尤其感到愤怒。

  [您]现在收听的是美国之音的《时事经纬》节目。

  有[人]猜测当初的爆炸发生以后，船上可能有[人]没有死亡。

  他们 要把 [多数人]的利益 , [多数人] 的愿望, [多数人]的意见作为我们制定政策
  的出发点和归宿.

  瑞士政府暂时将不向伊拉克派驻[大使],但不久将派[两名外交官]赴 伊重开□伊使□.

Generic
A mention is generic when the entity being referred to is not a particular, unique object (or set of objects). Instead generic 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]$_{\text{PER.NOM.NonSPC}}$ and [Republicans]$_{\text{PER.NOM.NonSPC}}$ are always at each other’s throats
- [Lawyers] $_{\text{PER.NOM.NonSPC}}$ don’t work for free.
- But the sense of urgency for this meeting matches the rage felt by both [Israelis] $_{\text{PER.NOM.NonSPC}}$ and [Palestinians]$_{\text{PER.NOM.NonSPC}}$ after yesterday’s violence.
- …[extremist groups] $_{\text{PER.NOM.NonSPC}}$ have a lot of support these days and a lot of power.
- 建立 [一支与打赢未来战 争相适应的人才指指 部队].
- [高科技 部队] 如果没有 [高素质人才] 支撑, 再先进的装备也是一堆废铁 .

Personal pronouns – in particular, first and second person pronouns – may have a generic reading even when they are not used anaphorically. In Chinese the third person pronouns can only have a generic reading when they refer back to a generic NP in the discourse.

一个共□党□无所畏惧。哪里有□□[他]就会出□在哪里。

Some hints for distinguishing between generic and non-generic:
1. If the NP is in the form of “a(n) + N”, can you replace the NP with a bare plural? If so, it’s generic.
2. For a bare plural NP, can you precede the NP with a demonstrative in subsequent references? If not, it’s generic.
Note that a universal quantifier does not necessarily trigger a generic reading even though the quantified NP refers to the entire set of entities.
Everyone likes fish. (generic)
Everyone went there. (specific)

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. ‘茶党, ‘反政府团体’) 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, unless it is used as an alias to refer to an entity. 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.

[~~林黛玉~~]与[~~贾宝玉~~]是两个家喻户晓的小说人物
［林黛玉］旷了那么多天课总算今天来上学了。
[一个学生]$_{\text{NOM.PER}}$
[访华代表团]$_{\text{NOM.PER}}$
［华人］$_{\text{NOM.PER}}$
［阿拉伯人］$_{\text{NOM.PER}}$
［基督教徒］$_{\text{NOM.PER}}$
［巴勒斯坦人］$_{\text{NOM.PER}}$
［维吾尔族人］$_{\text{NOM.PER}}$
［阿尔巴尼亚族］$_{\text{NOM.PER}}$
[刘诗诗]$_{\text{NAM.PER}}$祝[小狮子们]$_{\text{NOM.PER}}$新年快乐！
[齐国熊猫]$_{\text{NAM.PER}}$送给[老大]$_{\text{NOM.PER}}$的礼物

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:

[老王] *NAM.PER*
[小王] *NAM.PER*
[奥利弗三世] *NAM.PER*
[小布什] *NAM.PER* 先生

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:

[代表团]PER
[警察]PER 逮捕了 [这群反政府人员]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:

- [纽约时报]ORG 宣布它任命了新的执行总裁.
- 他每天早上都 看~~[纽约时报]ORG~~.
- [脸书]ORG 的总部在加州.
- 我在 ~~[脸书]ORG~~ 上看到这个惊人的消息

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

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

[共产党]
[劳工党]
[共和党全国委员会]
[北约]
[世界银行]
三个[联合国]工作人员遭绑架
[基地组织]
韩国[统一部] 长官朴在归12号率领韩国代表团乘专机抵达平壤
中国[国务院]的工作规则首次公布， 领导一般不得为地方题词。
美国[军队]
中国[人民解放军]
[泰米尔猛虎组织]
[尼泊尔游击队]“封锁”首都陆路交通
[联合国][维和部队]
[卡托研究所]
[东方人文思想研究所]
中国[民间保钓协会]

Commercial

中国大陆也将逐步对[诺基亚]（[Nokia]）或[易利信]([Erisson])等[外国厂商]开放。
从传言到现实：[盛大]联姻[新浪]，震惊业界。
视频领域的老大[优酷]和[土豆]终于正是加入客厅屏幕战斗。
[苹果]和[中国移动通信]达成协议

Educational, Scientific, Medical

[台湾大学][国家发展研究所]教授洪镰德指出......
这台类人形机器人与[国防科技大学]1990年研制的我国首台两足步型机器人相比
[北京大学][中国经济研究中心]成立五周年庆典

Media

[人民日报]
[新华社]
[美国之音]
[中国地产杂志社]
[香港中国通讯社]

[中国建筑工业出版社]

[中国新闻社]

Religious, Social, Advocacy

[罗马教廷]

中国[基督教协会]

中国[天主教爱国会]

中国[基督教“三自”爱国运动委员会]

[中国道教学院]

[红十字会]

Sports

费城[76人队]

[中国队]

中国[奥林匹克委员会]

[中华全国体育总会]

中国[残疾人体育协会]

中国[羽毛球协会]

武汉[红金龙]击败四川[全兴]

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:

代表团计划前往[缅甸]$_{\text{GPE}}$。

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

在[[[黑龙江]、[陕西]、[河北]、[浙江]、[四川]、[上海]等省]市]纪念抗美援朝50周年的展览也相继开幕

该委员会认为，[香港]经济复苏的情况比 [[亚洲]其他国家]都要好

[哈尔滨市]一些参加过抗[美]援[朝]战争的老战士和民间收藏爱好者.....

作为[[中国]最大的经济中心城市][上海]，由于[它]独特的区位优势和管理、人文优势，也必然会成为[跨国公司地区总部和外资研发中心的首选地]。

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.

[中]GPE[美] GPE达成“密约”要往死里收拾安倍。
[美] GPE [俄] GPE [韩] GPE 要求继续进行[朝鲜] GPE 核谈判。
[平壤] GPE 警告[澳] GPE 勿参与拦截否则遭核打击。
[朝鲜] GPE 拒绝与[日本] GPE 就横田惠假遗骨案进行会谈的要求.

NOTE: When a GPE name +人 is used to refer to the people of a GPE, the whole string should be tagged as a PER entity and the GPE is tagged separately as GPE. For example:

[[中国]_ NAM.GPE人民] NOM.PER热爱和平.
[[日本] NAM.GPE人] NOM.PER一直没有为上个世纪的战争道歉。
当前，[一些[中国] NAM.GPE人] NOM.PER热衷于地产投资。

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

非洲很多地方说~~法语~~.
~~阿拉伯语~~是国际性语言。.
很多[印度]语言很普遍，~~包括印地语，泰米尔语，乌尔都语，旁遮普语~~等。

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:

[费城]ORG 在加时赛击败了 [波士顿]ORG

Note however, that GPE names nested within sports team names should still be tagged as GPEs:

[[曼切斯特]GPE.NAM联队]ORG.NAM

GPE names may also modify sports team names:

[费城]GPE [老鹰队]ORG

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

2008年奥运会将在[中国]举行.

[苏州]距离[上海]比[杭州]更近
新华社记者[纽约]报道
[中国]近年来的蓬勃发展让世界嘱目
[中国][北京]
[辽宁]副省长
[辽宁][阜新][孙家湾]煤矿
[美国]可能会在今年6月就对[伊朗]实施军事打击
维护和平、促进发展，是[中国]政府外交方面的一贯立场

NOTE: Countries of countries, such as '[欧盟]$_{\text{GPE}}$', '[英联邦]$_{\text{GPE}}$' will be annotated as GPEs, since they have all three GPE components (i.e., a population, a government, and a location). The same formula applies to contested areas like "大陆", "台湾"

[[中国]$_{\text{NAM.GPE.}}$大陆]$_{\text{NOM.GPE}}$
[海峡两岸] $_{\text{NOM.GPE}}$

Cluster of countries that do not have all three GPE components can't be considered as GPEs, but rather as Location; see also section 2.3.5 below.

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, airports, highways, street names, factories, cafes, manufacturing plants, street addresses, oceans, seas, straits, bays, channels, sounds, rivers, islands, lakes, national parks, mountains, and monumental structures, such as the Eiffel Tower and Washington Monument. For instance:

Continents and non-politically-defined regions:

[东欧] $_{\text{NAM}}$
[欧] $_{\text{NAM}}$ [美] $_{\text{NAM}}$国家
[南亚] $_{\text{NAM}}$
[[非洲] $_{\text{NAM}}$南部] $_{\text{NOM}}$
[中东] $_{\text{NAM}}$
[台湾地区] $_{\text{NOM}}$
[北京城区] $_{\text{NOM}}$
[[华中] $_{\text{NAM}}$地区] $_{\text{NOM}}$
[德国南部] $_{\text{NOM}}$

Natural structures

［山东半岛］$_{\text{NAM}}$
［钓鱼岛］$_{\text{NAM}}$
［这座山］$_{\text{NOM}}$

Human-made structures and facilities

[大亚湾核电站]
按国际四星级标准设计建造了多功能现代化酒店--[广州大厦]
[这个位于加基武吉的工人宿舍]，占地共2万平方公尺，楼面近3万4000平方公尺，可容纳6000名住客
自１０月１２日美国一艘驱逐舰在[也门亚丁港]发生爆炸以来，没有一艘美国船只驶过苏伊士运河。
[柏林墙]
[位于直落古楼3楼的 房间] 内......
设施包括[餐厅]、[医药中心]（备有X-光设施）、[迷你市场]、[行政大厦]等
四川[成绵高速公路] 是四川省[高等级公路]主骨架的重要组成部分
载满滑雪者的列车向[[基茨坦霍恩山]顶]爬升途中, 在进入[3公里长的隧道]行驶了600米后, 突然起火
[国际空间站]
[济宁高速]
[京广铁路]

Fictional or mythical locations should <u>not</u> be tagged.

在~~[多塔大陆]~~，近卫军团与天灾军团从诞生以来就是敌对的双方
~~[艾泽拉斯]~~是一块被无边海洋包围的巨大陆地

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

该系统将会在［115 个美国飞机场］和 14 个海港安装

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]

［大亚湾湾核电站]

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]

按国际四星级标准设计建造了多功能现代化酒店--[广州大厦]

[这个位于加基武吉的工人宿舍]， 占地共 2 万平方公尺， 楼面近 3 万 4000 平方公尺， 可容纳 6000 名住客

自１０月１２日美国一艘驱逐舰在[也门亚丁港]发生爆炸以来， 没有一艘美国船只驶过苏伊士运河。

[柏林墙]

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]

［位于直落鼓楼 3 楼的房间］内

设施包括［餐厅］，［医药中心（备有 X-广设施）］［迷你市场］，［行政大厦］等

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…
[四川城高速高速公路] 是四川省高等级
公路主骨架的重要组成部分 载满滑雪者的列车向基茨坦霍恩山顶爬升途中,
在进入[3 公里长的隧道]行驶 了 600 米后,突然起火

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.

那个[王进喜]NAM 简直要把人气死 vs [那个人]NOM 简直要把人气死
这个刚创建的［建新科技公司］NAM vs ［这个刚创建的科技公司］NOM

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.

[中国]GPE, [北京]GPE
[北京] GPE [国家大剧院] LOC/ORG

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.

[西伯利亚地区] NAM.LOC
[密西西比河] NAM.LOC
[加沙地带] NAM.LOC
[纽约市] NAM.GPE
[[济宁] NAM.GPE 高速] NAM.LOC
[京广铁路] NAM.LOC

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:

前[苏联] NAM.GPE
[北韩] NAM.GPE

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. 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:

[中东] NAM.LOC
[东欧] NAM.LOC
[华南] NAM.LOC

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

[[德国] NAM.GPE南部] NOM,LOC

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.

[一名 66 岁的精神病患者]PER 在[[哈尔滨] GPE 第一医院]LOC 死亡

[[哈尔滨]GPE 第一医院] ORG 拒绝公布患者的病历。

[美国]ORG 摘取首枚冬奥会金牌

[华盛顿] GPE 如此纵容[东京] GPE 也不怕到时落得农夫和蛇的下场

[国家大剧院]ORG 将于其它世界顶级乐团合作

去［国家大剧院］$_{\text{LOC}}$ 参观还要买门票
“[深喉］$_{\text{PER}}$”的真实身份一直是个谜
[奥巴马]$_{\text{PER}}$抵达[北京]$_{\text{GPE}}$

In Chinese, cases like’巴基斯坦人’ indicate a person’s citizenship or origin, which includes both a GPE or LOC and a PER. We will not consider the whole string as a NAM, but rather treat the GPE as a NAM and the PER as a NOM. For example,

[[巴基斯坦] $_{\text{NAM.GPE}}$人] $_{\text{NOM.PER}}$
[[东欧] $_{\textit{NAM.LOC}}$ 人] $_{\textit{NOM.PER}}$

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:

［南］$_{\text{GPE}}$［北韩］$_{\text{GPE}}$
［江西］$_{\text{GPE}}$和［湖北省］$_{\text{GPE}}$
[[布莱恩] $_{\text{NAM.PER}}$ 兄弟] $_{\text{NOM.PER}}$ 2011[休斯顿] $_{\text{GPE}}$ 夺冠后跳泳池庆祝

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

[特立尼达和多巴哥共和国]$_{\text{NAM.GPE}}$
[工业和信息化部] $_{\text{NAM.ORG}}$

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

When conjunctions are used excessively 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]$_{\text{PER.PRO}}$ stepkids and friends and family]$_{\text{PER.NOM}}$
- [my stepkids and friends and family] $_{\text{PER.NOM}}$
- [my stepkids and friends and family] $_{\text{PER.NOM}}$

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:

[[布什]NAM.PER政府]NOM.ORG
[[俄罗斯] NAM.GPE总理] NOM.PER
[[美国] NAM.GPE选民] NOM.PER
[[新华社] NAM.ORG记者] NOM.PER
[[美国] NAM.GPE [国防部] NAM.ORG发言人] NOM.PER
[[费城] NAM.GPE市政府] NOM.ORG
[[[哈市] NAM.GPE医院] NOM.ORG患者] NOM.PER
[[中国] NAM.GPE人民银行] NAM.ORG
[[北京] NAM.GPE大学] NAM.ORG

5.7 Possessives

When you encounter a possessive construction, it may contain two taggable entity mentions, as in:

> [[[[北]$_{\text{GPE}}$ 大]$_{\text{ORG}}$ 的各个院]系]$_{\text{ORG}}$
> [[加拿大]$_{\text{GPE}}$ 的 议院]$_{\text{ORG}}$
> [[皇马足球队]$_{\text{ORG}}$队员]$_{\text{PER}}$

5.8 Hyphenated pre-modifiers

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

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

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, '巴勒斯坦' is tagged as a GPE because it has all three GPE components, while '加沙地带' 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]$_{\text{PER.NAM.SPC}}$ have added to the evidence linking [him]$_{\text{PER.PRO.SPC}}$ and [the [al-Quaida]$_{\text{ORG.NAM.SPC}}$ network]$_{\text{PER.NOM.SPC}}$ to the Sept. 11 [terrorist]$_{\text{PER.NOM.NonSPC}}$ attacks in the [United States]$_{\text{GPE.NAM.SPC}}$, [the government]$_{\text{ORG.NOM.SPC}}$ said Wednesday in an updated dossier on the investigation.
- [Guzman]$_{\text{PER.NAM.SPC}}$ indicted [Pinochet]$_{\text{PER.NAM.SPC}}$, holding [him]$_{\text{PER.PRO.SPC}}$ responsible for the actions by the "[Caravan of Death]$_{\text{PER.NAM.SPC}}$", [a military party that killed [73 political prisoners]$_{\text{PER.NOM.SPC}}$ shortly after the 1973 coup in which [Pinochet]$_{\text{PER.NAM.SPC}}$ ousted Marxist President [Salvador Allende]$_{\text{PER.NAM.SPC}}$]$_{\text{PER.NOM.SPC}}$.
- Midway through the hearing, Chief Justice [Renquist]$_{\text{PER.NAM.SPC}}$ seemed to scold [[his]$_{\text{PER.PRO.SPC}}$ colleagues]$_{\text{PER.NOM.SPC}}$ for being too talkative when [he]$_{\text{PER.PRO.SPC}}$ made an unusual offer to [the lawyer representing [[Florida]$_{\text{GPE.NAM.SPC}}$'s Attorney General]$_{\text{PER.NOM.SPC}}$]$_{\text{PER.NOM.SPC}}$.
- [Actors and singers also on the flight]$_{\text{PER.NOM.SPC}}$ held a benefit concert in [Baghdad]$_{\text{GPE.NAM.SPC}}$ Saturday evening, with most of the $13 cover charge to be donated to support the [Palestinian]$_{\text{GPE.NAM.SPC}}$ uprising.
- ...said Archbishop [Khajag Barasamian]$_{\text{PER.NAM.SPC}}$, head of the [Diocese of the

[[Armenian]$_{\text{GPE.NAM.SPC}}$ Church]$_{\text{ORG.NAM.SPC}}$ in [America]$_{\text{GPE.NAM.SPC}}$]$_{\text{ORG.NAM.SPC}}$, [[whose]$_{\text{ORG.PRO.SPC}}$ headquarters]$_{\text{FAC.NOM.SPC}}$ are in [Manhattan]$_{\text{GPE.NAM.SPC}}$.

十二届全国人大第一次会议3月14日选举[李源潮]$_{\text{PER.NAM}}$为[国家副主席]$_{\text{TTL}}$。引人注意的是，今次以[[中央政治局]$_{\text{ORG.NAM}}$委员]$_{\text{TTL}}$身份成为[国家领导人]$_{\text{TITLE}}$的[李源潮]$_{\text{PER.NAM}}$，打破了[胡锦涛]$_{\text{PER.NAM}}$从1998年3月出任[国家副主席]$_{\text{TTL}}$开始形成的惯例，[[国家副主席]$_{\text{TTL}}$职务担任者]$_{\text{PER.NOM}}$15年来首次不“入[常]$_{\text{ORG.NAM}}$”。

[一名不愿透露姓名的警察]$_{\text{PER.NOM}}$告诉[[新华社]$_{\text{ORG.NAM}}$记者]$_{\text{PER.NOM}}$,第一起爆炸发生在位于[[萨那市] $_{\text{GPE.NAM}}$中心]$_{\text{LOC.NOM}}$的[也门]$_{\text{GPE.NAM}}$[国防部] $_{\text{LOC.NAM}}$附近， 导致[附近[老城] $_{\text{LOC.NOM}}$的城墙] $_{\text{LOC.NOM}}$受损， 但没有造成人员伤亡。第二起爆炸地点距 [也门]$_{\text{GPE.NAM}}$前[总统]]$_{\text{TTL}}$[[萨利赫]$_{\text{PER.NAM}}$官邸]$_{\text{LOC.NOM}}$和[[法国]$_{\text{GPE.NAM}}$驻[也门]$_{\text{GPE.NAM}}$使馆]]$_{\text{LOC.NOM}}$仅几十米远， 至少[两名市民]$_{\text{PER.NOM}}$受伤。

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.

~~全运会~~
the [Olympic Committee]$_{\text{ORG}}$
[国际奥林匹克委员会]$_{\text{ORG}}$

6.2Artifacts 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:

~~最新款的路虎~~很霸气

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 <u>dissident</u>]** 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.
- 在参加北京科博会时，[诺基亚新任中国公司负责人][谷思华]表示，[他]对WP 手机在中国的销售情况感到鼓舞。“[我]在诺基亚工作的时间已经近八年，[我]先在诺基亚芬兰公司，然后又到诺基亚英国公司，现在又到诺基亚中国公司工作，”一见到媒体，[谷思华]如此介绍了[他][自己]

The name mentions of ‘谷思华’ are easy to spot. However, it is important to annotate all mentions that refer to the entity that is ‘谷思华’。This will include nominal mentions bolded above, such as ‘诺基亚信任中国公司负责人’and pronominal mentions such as ‘him’ 他’，’ 我’ ‘自己’ 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, ‘诺基亚’ is an ORG, which means it cannot be coreferential with the PER mentions of ‘谷思华’:

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: 我昨天在街上碰见了［那个人]k.
  - B: ［哪个人］$_k$？是王姐给你介绍的［那个人］$_k$ 吗？
  - A: 是的，就是[他] k

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]$_{\text{NonSPC}}$ don't always work for free, but [they]$_{\text{NonSPC}}$ do from time to time. [Lawyers]$_{\text{NonSPC}}$ 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]$_{\text{SPC}}$ entered the court room a few minutes before the trial started. [Lawyers]$_{\text{SPC}}$ approached the bench to talk with the judge. Isn't it odd how [lawyers]$_{\text{NonSPC}}$ sometimes seem chummy with judges and other times [they]$_{\text{NonSPC}}$ 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:

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

- [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 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

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.

[我]就知道[他][我个人]会帮[我]的

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

[王，岐山]打老虎

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.

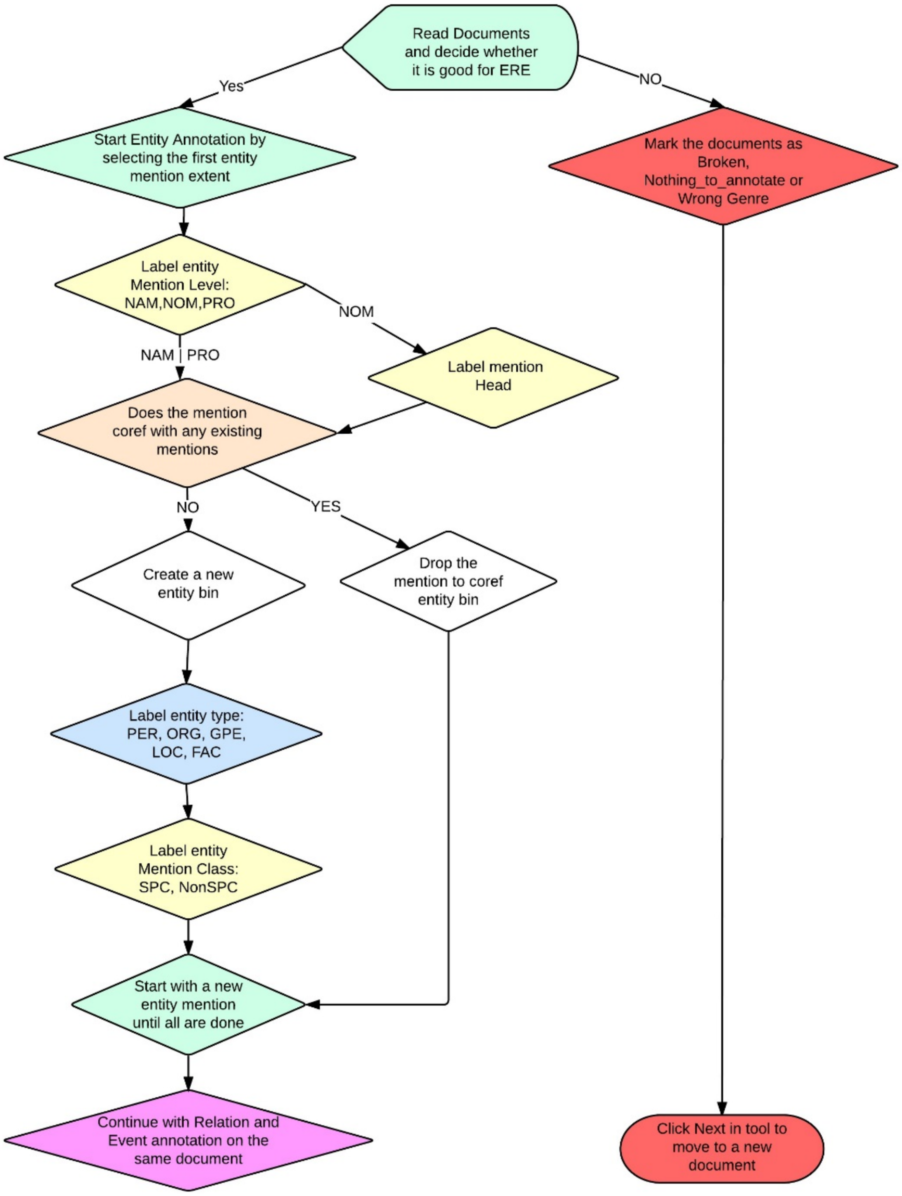
- dialogue:
    a. A: [John]
    b. B: [John] was
    c. A: I saw [John]

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