Entity Detection and Tracking – Phase 1

EDT and Metonymy Annotation Guidelines
(adapted from English version 2.5 by
Shudong Huang
for use in Chinese ACE annotation)

Version 2.5 20030530

1 Intro

The objective of the ACE program is to develop automatic content extraction technology to support automatic processing of source language data. This includes classification, filtering, and selection based on the language content of the source data, i.e., based on the meaning conveyed by the data. Thus the ACE program requires the development of technologies that automatically detect and characterize this meaning.

Ultimately, ACE applications will maintain a database of what is happening in the world. Ideally, this will be in terms of who is doing what, where, and when. As information from source language data is accumulated over time, the database will be updated and maintained. In this way the database becomes a vehicle for tracking the information we are interested in. The database should also maintain pointers into the source data so as to ensure more detailed examination of the information represented in the database.

The ACE research objectives are viewed as the detection and characterization of Entities, Relations, and Events. ACE Phase 1 begins the technology R&D effort by focusing on entity detection. This task is being defined so as to support applications as well as to provide a basis for further development in extracting relations and events.

The Entity Detection task requires that selected types of entities mentioned in the source data be detected, their sense disambiguated, and that selected attributes of these entities be extracted and merged into a unified representation for each entity. Tracking of entities across document boundaries will be deferred until after the initial phase.

This document outlines the ACE Phase 1 annotation tasks (Entity Detection and Tracking, Metonymy Annotation, and Generic/Specific Classification). It is intended to integrate section 6 of the ACE Pilot Study Task Definition v 2.2, EDT Metonymy Annotation Guidelines v 2.4, and various addenda to both documents into up-to-date annotation guidelines. Please refer to NIST's ACE website (www.itl.nist.gov/iaui/894.01/tests/ace/index.htm) for the ACE task definition and evaluation plan.

2 Basic Concepts

An entity is an object or set of objects in the world. A mention is a reference to an entity. Entities may be referenced by their name, indicated by a common noun or noun phrase, or represented by a pronoun. For example, the following are several mentions of a single entity:

Name Mention: *江泽民*

Nominal Mention: *现担任中央军委主席的人*

Pronoun Mentions: *他的儿子曾留学美国；自己*曾留学苏联*

(*for reflexives, please refer to section 4.2.5.3, "Pronominals", for details)

For Phase 1 of ACE, entities are limited to the following five types:

- Person - Person entities are limited to humans. A person may be a single individual or a group.
- Organization - Organization entities are limited to corporations, agencies, and other groups of people defined by an established organizational structure.
- Facility - Facility entities are limited to buildings and other permanent man-made structures and real estate improvements.
- Location - Location entities are limited to geographical entities such as geographical areas and landmasses, bodies of water, and geological formations.
- GPE (Geo-political Entity) - GPE entities are geographical regions defined by political and/or social groups. A GPE entity subsumes and does not distinguish between a nation, its region, its government, or its people.

We do not identify mentions of animals or most inanimate objects at this time.

For each entity, the annotation records the type of the entity (PER, ORG, GPE, LOC, or FAC), its class (Generic/Specific), all of the mentions of the entity from the text (Name, nominal, Pronoun), and the role of those mentions if applicable (see section 4.1.5.3 GPE Mention Roles).

It's important to note that a given linguistic expression without any context does not predict whether or not it's markable, or what type of entity it refers to. Its markability is decided by the entity it refers to in the context of use. For example,

3 Text to Annotate

Only material between <TEXT> and </TEXT> tags is to be annotated. In newswire documents, material in headlines and slug sections is not to be tagged. In broadcast news, only the transcribed speech is to be tagged; added information, such as that within <TURN> tags or speaker identification tags, is not to be tagged.

4 Entities and Mentions

4.1 Entity Types

4.1.1 Persons

Each distinct person or set of people mentioned in a document refers to an entity of type person. People may be specified by name ("John Smith"), occupation ("the butcher"), family relation ("dad"), pronoun ("he"), etc., or by some combination of these. Dead people and human remains are to be recorded as entities of type person. So are fictional human characters appearing in movies, TV, books, plays, etc.

There are a number of words that are ambiguous as to their referent. For example, nouns that normally refer to animals or non-humans can be used to describe people. If it is clear to the annotator that the noun refers to a person entity in a given context, it should be marked as so.

He is [a real turkey]
[The political cat of the year]
He was [one of the dark horses]
[The film star]
She's known as [the brain of the family]
[Californian transplants]
He is [a harmonic force]
[大腕]
他是[害群之马]
他是个[狐狸]

4.1.1.1 Saints and other religious figures

Religious titles such as saint, prophet, imam or archangel are to be treated as titles.

St. Christopher, the patron of transportation

References to "God" will be taken to be the name of this entity for tagging purposes. If it is used as a descriptor rather than a name, it will be considered a nominal mention. Note that capitalization information may not be available in speech transcripts.

If you believe in god, you must… name mention
如果你相信上帝…
He felt like he was [a god]. nominal mention
他觉得自己是个上帝.

Note the last example entails that a proper name does not necessarily imply that the mention is a named mention.

4.1.1.2 Fictional characters, names of animals, and names of fictional animals

Names of fictional characters are to be tagged; however, character names used as TV show titles will not be tagged when they refer to the show rather than the character name. Compare the following examples:

Batman has become a popular icon
Adam West's costume from Batman the TV series

Names of animals are not to be tagged, as they do not refer to person entities. The same is true for fictional animals and non-human characters. These two examples do not yield mentions.

Morris the cat

Snuggle, the fabric softener bear

4.1.1.3 Groups of people

Groups of people are to be considered an entity of type Person unless the group meets the requirements of an organization or a GPE described below.

The family

The house painters

The linguists under the table

4.1.1.3.1 Ethnic, Religious, and Political Groups

Ethnic groups, religious groups and political groups are often referenced by the name of the ethnicity, religion and political party, for example:

African-Americans

Catholics

Democrats

华人(vs.中国人)

华裔

汉人

藏人

天主教徒

Those groups that have an organizing body are name mentions of the organization. If a mention refers to the members of an organization in general, we consider the mention to refer to the organization.

Democrats support social programs.

Catholics celebrate Lent every year.

Democrats is an organization name because it is used in a context describing the beliefs of the greater organization of the Democratic Party. In other words, DO NOT tag it as “generic PER”, but as “specific ORG”.

When a mention of this kind refers to an individual person, as in

Mike is a Democrat

or to a small group of individuals, as in

Mike and Bob are both Democrats

the mention is a person nominal (NOT named) and is a mention of the same entity as the person to whom the phrase is attributed.

It is also possible that such mentions refer to the entire members of the group, but as type PER, not as ORG. Here’s such an example:

Catholics have been systematically persecuted and jailed in China.

In this context, *Catholics* should be tagged as “nominal generic PER” since a religious organization itself can’t really be persecuted (?).

Ethnic groups do not generally have a formal organization associated with them. As a result, we mark these mentions as names of a person entity if the entity is not specific.

> ***{[PER-name]** Cuban Catholics}* *are expecting the Pontiff to preach about the value of religious freedom, something they're just beginning to experience.*

Note that here “Cuban Catholics” is one unit, i.e. Cuban is not a modifier of Catholic and doesn’t evoke mention of another entity.

When ethnic designation is given to an individual person or a small group of individuals, the mention is marked as a nominal mention of that person entity.

> *Joe is **{[PER-nominal]** a Cuban Catholic**}**.*

In this example, the mentions “Joe” and “a Cuban Catholic” refer to the same entity.

Note in the examples above, “Cuban” is meant to mean “Cuban American”, not “Cuban Cuban”, which would be type GPE. In other words, in many other contexts, “Cuban” may refer to a GPE entity. Therefore it is important that the annotator pays particular attention to the context to decide how a particular linguistic expression is being used. 不要只看字面意义。

4.1.1.3.2 Family Names

Family names are to be tagged as Person.

> *The Kennedys*
>
> *The Kennedy family*

Please note that the second example contains two mentions of the same entity: one name mention and one nominal mention. That is, “Kennedy” is considered a modifier of “family”.

A monosyllabic family name (mostly Han, or Korean/Vietnamese/etc.) in Chinese rarely occurs alone to refer to a person, though a multi-syllabic family name (a few Han but mostly non-Han/foreign) can be used directly to refer to a person. There’s no Chinese counterpart of “the Kennedys”, but like in English, a family name in Chinese can be used with “家” to refer to a family as a social unit with a man and woman and their offspring. A monosyllabic family can be followed by “氏” and then “家族” to mean a group of people sharing a common ancestry – in the case of multi-syllabic family names (or full names), the morpheme “氏” is not needed. Further, a first name can always be followed by “家” to mean that person’s family (as a social unit). We distinguish the following cases:

- Monosyllabic family name +家: entire expression as a mention of person entity:

> *{张家}*
>
> *{李家}*

- Monosyllabic family name +氏 +家族 or multi-syllabic family name +家族: two mentions of the same person entity:

{{张氏}家族}
{{肯尼迪}家族}

- Multi-syllabic family name +家 or given name +家 or full name +家 or family name with title/honorific +家: two mentions of two distinct person entities:

{{张三}家}
{{布什}家}
{{书东}家}
{{黄书东}家}
{{李先生}家}
{{李部长}家}
{{老王}家}

The rationale behind treating monosyllabic + 家 and other cases + 家 is conjunction test.

张三和李四家 OK to mean both families

**张李家 not OK to mean both families; instead must use:*

张家和李家

4.1.2 Organizations

Each organization or set of organizations mentioned in a document gives rise to an entity of type organization. An organization must have some formally established association and a persistent, established existence. Typical examples are businesses (Microsoft, SUN), government units (the Bush administration, the State Department), sports teams, and formally organized music groups. Industrial sectors are also treated as organizations (often as a group).

Sets of people who are not formally organized into a unit are to be treated as a person entity rather than an organization entity. It is often difficult to tell the difference between organization entities and collections of individuals tagged as person entities. Example organization-like nouns which are *not* organizations are “family,” “employees,” and “crew.” In the latter two cases, although the members of a company or crew may work together in an organized and even hierarchical fashion, the groups are not organizations by themselves.

Some words like “team,” “delegation” and “police” achieve organization status only in certain contexts. “[The home *team*] flies to Connecticut to meet the Huskies in Hartford” clearly refers to a named sports team and is thus taggable as an organization. However, the “[U.N. weapons inspection *team*]” is less permanent and cohesive, and is thus a person entity rather than an organization. The noun “police” is a person entity in contexts like “[*police*] outnumbered [*demonstrators*]” but an organization entity in “[*police* in East Timor] have arrested [two *men*].”

The base type of hotels and restaurants is considered as an ORG type, a kind of business organization. However, a hotel or restaurant name may also be used to refer to a facility entity or location entity. When this kind of metonymy happens, we must mark two entity types – one literal (base-type), and the other intended. Compare the following examples:

> *{[ORG]狗不理包子店}将在北美设立{[ORG]特许连锁店}.*
>
> *{[lit:ORG][int:FAC]狗不理包子店}的橱窗里没有狗不理包子！*

For details on metonymy, please refer to Section 5.

Note the base type of factories/plants (工厂) is FAC. Please refer to 4.1.4 below.

Note also that government units often have a preceding government name modifier. Please refer to ? for details.

?An organization name may sometimes be used to refer to the members of the organization in aggregate (“SRI defeated BBN in softball”) or the buildings housing that organization (“SRI was destroyed by the 2003 earthquake.”) These concepts are subsumed by the organization entity. Thus, in each of these examples “SRI” should be considered a mention of (the same) entity of type organization. ?

4.1.2.1 Organization Entities used in Person Contexts

Whenever an organization takes an action, there are people within or in charge of the organization that one presumes actually made the decision and then carried it out. Thus many organization mentions could be thought of as metonymically referring to people within the organization. However, there seems to be little to be gained in the usual case by thus “reaching inside the organization” to posit a PER metonymy. It seems better to adopt the view that organizations can be agentive, and take action on their own. Only when something in the context draws particular attention to the people within the organization should a separate mention of a PER entity be marked.

4.1.2.2 First Person Pronouns Referring to Organizations

First person plural pronouns are often used by representatives of an organization to refer to that organization. Pronouns are often used in this way by reporters representing a broadcasting station and spokespeople representing organizations. For example, in *our top story*, *our* refers to the broadcasting organization. In these cases, annotators should mark first person plural pronouns as ORG mentions, and not as PER mentions – unless the story is about the reporters/anchors themselves ☺

4.1.3 Locations

Locations defined on a geographical or astronomical basis which are mentioned in a document and do not constitute a political entity give rise to location entities.

These include, for example, the solar system, Mars, the continents, the Hudson River, Mt. Everest, and Death Valley.

In general, terrestrial locations must have some two-dimensional extent. Abstract coordinates ("31° S, 22° W") and positions relative to a GPE or location ("30 miles east of Mount Fuji") are not themselves entities. Borders, considered as (one-dimensional) boundaries between two regions, are not entities. Positions distinguished *only* by the occurrence of an event at that position ("the scene of the murder", "the site of the rocket launching", “抗洪战场”) are not entities.

4.1.3.1 Sub-parts of Locations and GPEs

Portions of GPE entities or location entities, such as "*the center of the city*", "*the outskirts of the city*", or "*the southern half of New Jersey*" constitute location entities in their own right. When general locative phrases like “top,” “bottom,” “edge,” “periphery,” “center,” and “middle” are used to pinpoint a portion of a markable location, they are markable locations.

> *“They tend to live not in [the **center** of [the country]] but at [its **periphery**]”*

Note that location entities may also refer to the population of a region, or other aggregates within that region:

> *[The Deep **South]** voted for Bush.*
>
> *[Southern **France]** drinks more wine than Boston.*

4.1.3.2 Non-Locations

It is easy to start interpreting all objects as locations. Every physical object implies a location because the space that each physical object occupies is the “location” of that object. In addition, our language is full of location modifiers (which are often prepositional phrases) that pinpoint objects and activities, and even abstract concepts:

> *“Your coat is under the dog.”*
>
> *“The rabbit is hiding behind that rock.”*
>
> *“I have an idea in my head.”*

Viewed from a certain angle, “the dog,” “that rock” “my head” become locations. Very “location-ish” nouns make such an interpretation even more tempting:

> *“He dropped the logs on the ground.”*
>
> *“He put the lamp back in its place.”*

However, none of these are taggable location expressions. They do not fall within any of the classes defined above for taggable locations. The annotator must be careful not to fall down this slippery slope.

Do not tag compass points when they serve as adjectives or refer to directions, as in “the ants are heading north” and “they are found as far north as Maine.” Compass points should only be tagged when they refer to sections of a region, as in “the far west.”

4.1.3.3 Chinese Locative Phrases (*New*)

Chinese has a special class of affix-like words known as locatives (方位词). A typical locative phrase consists of "在" – which sometimes may be optional - followed by an NP and then by the locative. A locative specifies a spatial relationship between the entity referred to by the preceding NP and another entity or an event in the sentence (or discourse). The following table shows a list of such words according to Li & Thompson (1981):

	+ suffix 边	+ suffix 面	+ suffix 头	Gloss
上	上边	上边	上头	On top of
下（底下）	下边	下面	下头	Under
	左边	左面		Left of
	右边	右面		Right of
前	前边	前面	前头	In front of
后	后边	后面	后头	In back of
里（内）	里边	里面	里头	Inside of
外	外边	外面	外头	Outside of
旁	旁边			Beside
中间（当中）				Middle of
	东边	东面		East of
	西边	西面		West of
	南边	南面		South of
	北边	北面		North of
这儿（这里）	这边	这面		This side of
那儿（那里）	那边	那面		That side of

Due to the disyllabic tendency in modern Chinese, often, monosyllabic locatives are used with monosyllabic noun. The suffixes 边, 面, and 头 are used with monosyllabic locative to form disyllabic locatives so that they can be used with multi-syllabic nouns. The affixes themselves have lost their original meaning in the composition and in fact all carry neutral tone. Their choice may vary across dialects and personal preferences.

The above locatives only refer to some space or location in relation to the entity referred to by the preceding NP but the location itself is not specific and hence doesn't constitute a markable location entity, even though the English translation may be markable. In general, the locatives themselves are not markable – this is particular true of multi-syllabic locatives. Therefore, we only mark the preceding N if the entity is a markable entity (usually type LOC or GPE.LOC.)

However, many "N + locative" compositions are highly lexicalized, particularly those mono-syllabic noun + mono-syllabic locative. For example, "室内" (indoor), "室外" (outdoor), "海上" (at sea, on the sea), "国内" (domestic, internal), "国外"

(foreign, external), “乡下” (countryside), “” and so on and so forth. These lexicalized compositions, or rather words, are tagged iff the entity it refers to falls into one of the entity types. For example, “省里” usually means “provincial government” and thus should be tagged as a mention of an ORG entity.

Certain words such as “国内外” (meaning “domestic and foreign”, i.e. the whole world) are marked by decree. We’ll have a list of such words as we move along with annotation.

Note that Li & Thompson also lists the following as locatives:

东部	East part of
西部	West part of
南部	South part of
北部	North part of

However, the second morpheme “部” is not really a suffix like 边, 面, and 头. Furthermore, these words differ from those in the first table in that these locatives divide the entity that the preceding NP refers to into portions rather specifying some kind of space which could be external. We do not consider “loc + 部” as locatives but as common nouns. Thus if the preceding N is markable, the entire expression is mostly like markable and it usually invoke two entities. For example “中国东部”, meaning “the eastern part of China”, has two mentions: 中国 as a GPE.LOC and 中国东部 as a LOC with 东部 as its head.

4.1.4 Facilities

A facility is a large, functional, primarily man-made structure. These include buildings, and similar facilities designed for human habitation, such as houses, factories, oilfields, 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.

Individual rooms of buildings are facilities, but other portions of buildings, such as walls, windows, closets, or doors, are not facilities.

Rivers have LOC as the base type. But some rivers may be used for transportation. If the context specifically mentions them as a transportation route (as 交通枢纽), then the intended type should be FAC.

?By contrast, man-made canals should have the base type FAC?

4.1.4.1 Facility Entities used in Organization Contexts

In some cases, a facility name is used to refer to an organization (which, typically, operates the facility) or a set of people (the people employed by that organization).

1. *The museum is located on Fifth Avenue.*
2. *I walked into the museum.*
3. *Mary works for the museum.*
4. *The museum insisted that the exhibition was not obscene.*
5. *The museum received a gift of $100,000.*

Examples 1 and 2 clearly refer to the museum building. Examples 3, 4, and 5 refer to the organization housed in or operating the museum facility. In cases like this, the annotation will reflect both the facility and organization entities. Please see the Metonymy section below for more information.

Likewise, 工厂 (factory) in Chinese, whose base type is considered as FAC following English, is often used as a mention of an ORG entity.

4.1.5 Geographical/Social/Political Entities (GPE)

Geo-Political Entities are composite entities comprised of a population, a government, a physical location, and a nation (or province, state, county, city, etc.). In the political system, the last aspect includes all the following entities: 国 (“country/nation”), 省 (“province”), 地区/专区 (“prefecture”, a political entity between a province and a county, e.g. 扬州地区), 市 (“municipality/city”, including cities at all levels, e.g. 北京市，上海市，扬州市), 县 (“county”), 区 (“district (within a city)”, e.g. 海淀区), 镇 (“town”), 乡 (“township”), 村 (“village”). But 街道办事处 and 居民委员会/居委会 (“neighborhood committee”) will not be annotated as GPE, but simply ORG instead.

All mentions of the four aspects of a GPEs will be marked GPE and coreferenced. In this sentence,

The people of France welcomed the agreement.

there are two mentions

[The people of France]	*GPE*
[France]	*GPE*

The mention of the population of France is marked GPE, rather than PER. These mentions would be coreference as they refer to different aspects of a single GPE. A site note for Chinese, the word “人民” usually translates into “the people” in English. Thus, xx 国人民 or xx 市人民 will be usually marked as GPE. In other words, 人民 is usually associated with a GPE entity. On the hand, xx 国人 can be tricky. Unless the mention refers to the population as a whole, it should be marked as PER, indicating the person’s ethnicity/nationality. Sometimes, it can be ambiguous. For example,

中国人喜欢喝茶。

> *朝鲜战争中国人和美国人打了个平手。*

Explicit references to the government of a country (state, city, etc.) are to be treated as references to the same entity evoked by the name of the country. Thus "*the United States*" and "*the United States Government*" are mentions of the same entity. On the other hand, references to a portion of the government ("*the Administration*", "*the Clinton Administration*") are to be treated as a separate entity (of type organization), even if it may be used in some cases interchangeably with references to the entire government (compare "*the Clinton Administration signed a treaty*" and "*the United States signed a treaty*"). However, one must be careful with other political systems. The distinction between "government" and "administration" is less clear in China's political structure. This is also reflected in the Chinese language. Often, one might see "美国政府" and "布什政府" used interchangeably. On the other hand, China's "国务院" should be ORG, instead of GPE.ORG.

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:

> *<u>New York</u> defeated <u>Boston</u> 99-97 in overtime.*

These are to be recorded as distinct entities, not as mentions of the GPE entity. Thus, in this example, both "*New York*" and "*Boston*" would evoke organization entities.

4.1.5.1 GPE Clusters to be treated as GPEs

Like GPEs, clusters of GPEs consist of a populace, a well-defined physical territory, and in some cases (like Europe), have an organizing body (the European Union) associated with it. Because of their similarities to GPEs, these entities appear in contexts similar to those of GPEs. For example:

> *President-elect Kim Dae Jung today blamed much of* ***Asia****'s devastating financial crisis on governments that "lie" to their people and "authoritarian" leaders who place economic growth ahead of democratic freedoms. [9801.404]*

> *Many of the leaders of* ***Asian*** *society have been saying that military dictatorship was the way and democracy was not good for their nations," Kim said. [9801.404]*

> *They concentrated only on economic development," he said, without singling out any nations but referring to "****Asian****-style democracy," in which governments are built around a strong leader who controls economic policy. [9801.404]*

For this annotation task, named geographical entities that are commonly referred to by those names will be considered GPEs rather than Locations. Following is a non-exhaustive list of entities that were Locations in the Pilot Study, but should be GPEs for this task.

> *Asia, Europe, Eastern Europe, Western Europe, EU, the Middle East, Palestine, Southeast Asia, New England, South Africa, all continents.*

> *华东(地区), 华南(地区), 华北(地区), 华中(地区), 东北(地区), 西北(地区), 西南(地区), 西北(地区) (GPE clusters specific to Chinese regions)*

Other, more incidental clusters of GPEs are still considered Locations. For example, *the southern United States* is a Location. And similarly {中国东部}.

On the other hand coalitions of governments, as well as the UN, are organizational bodies and should be marked Organization.

How about “the world”? In most cases, it probably invokes a LOC mention. But it may also invoke a cluster of GPE entities. For example:

> *Other than the US, UK and a couple of other countries, the rest of the word is opposed to the war.*
>
> *The third word countries resist such globalization.*

4.1.5.2 Nested Region Names

A series of nested region names, such as "*Provo, Utah*" evokes one entity for each region. Thus "*Provo, Utah*" evokes one entity for the city (with mention "*Provo, Utah*") and a second one for the state (with mention "*Utah*").

4.1.5.3 GPE Mention Roles

Since an GPE entity has four aspects as specified above, we distinguish four roles associated with an GPE mention. Annotators need to decide for each entity mention in the text which role (Person, Organization, Location, GPE) the context of that mention invokes. This judgment typically depends on the relations that the entity enters into.

***France** likes to eat cheese.*	Person Role
***France** signed a treaty with Germany last week.*	Organization Role
*The world leaders met in **France** yesterday.*	Location Role

In the examples above, the name “France” refers to a range of concepts. Annotators must select the Role which matches the function of the GPE mention.

The GPE role may be used in contexts that highlight the nation (or state or province or city, etc.) aspect of the GPE entity, as distinct from the government, populace, and location, but that it may also be used in contexts referring to an indistinct amalgam of more than one of the aspects of a GPE (government, population, location, and nation).

***France** produces better wine than New Jersey.*	GPE Role (whole nation)
***France**’s greatest national treasure*	GPE Role (indistinct referent)

Even if more than one aspect of the entity is invoked by the context, only one role should be assigned. This usually occurs in the case of conjoined predicates. For example,

> ***Washington** is preparing for potentially massive demonstrations against the World Bank and the International Monetary Fund as ministers from those organizations arrive for Sunday's opening session.*

In the above example, it is the government of Washington (ORG) that is preparing for the demonstrations, but ministers will arrive at the location Washington. In these cases, the annotator should assign a role based on the closest local predicate. In this example, only the ORG role should be assigned to

Washington because “preparing…” is the local predicate and invokes an ORG reading.

The following sections give particular guidelines for frequently encountered cases, with examples.

GPEs Modifying People and Artifacts

Pre-modifiers are inherently vague and difficult to decompose. For this reason, all GPE pre-modifiers of people and artifacts will be assigned the role GPE.GPE. For the sake of consistency, the corresponding post-modifiers should also be marked GPE.GPE. For example, ***[[GPE.GPE] French}*** *president* should be marked in the same way as *president of* ***{[GPE.GPE] France}***. More examples of GPEs modifying people include:

> ***{[GPE.GPE]*** *Israeli}* *troops*
>
> ***{[GPE.GPE]*** *New York}* *policemen*
>
> *Prime Minister of* ***{[GPE.GPE]*** *Britain}*
>
> *Joe Smith of* ***{[GPE.GPE]*** *the United States}*
>
> ***{[GPE.GPE]*** *New York}* *attorney*
>
> ***{[GPE.GPE]*** *U.S.}* *Commander-in-Chief*

GPEs modifying artifacts should also be marked GPE.GPE. Common artifacts modified by GPEs include but are not limited to vehicles, weapons, and flags. Some examples follow:

> ***{[GPE.GPE]*** *U.S.}* *surveillance aircraft*
>
> ***{[GPE.GPE]*** *Iraqi}* *flag*

Note that Chinese doesn’t usually use a GPE name as pre-modifier before a person name. Instead it usually uses a relative clause with the verb “来自” or “从…来”, or to some lesser degree, the DE (的) construction (which can also be used with nominal mentions). Given the following sub-section from the original English guidelines,

> #### *Political associations - representatives*
>
> Political associations hold between people and GPEs. So in *Hillary Clinton (D-NY)*, *NY* is marked GPE.GPE.
>
> > *“This is going to be a brutal fight," said Rep. Thomas C. Sawyer (D-* ***{[GPE.GPE]*** *Ohio}), who has been closely involved in the census and is among those who believes the ongoing debate played a role in Riche's departure.*

we decide that for any person name that involves a GPE modifier, the annotator should decide if that GPE entity has the role of GPE or LOC. If the person entity represents the GPE entity as in sports or other kinds of competition that involve multiple GPE entities, politics, etc., tag the GPE entity as GPE.GPE. Otherwise, tag it as GEP.LOC.

> *来自{[GEP.LOC]江苏}的小李在学校里表现最好.*

> *来自{[GPE.LOC]美国}的旅游者…*
>
> *来自{[GPE.GPE]山东}的代表在会上…*
>
> *{[GPE.GPE]美国}（的）选手…*

Activities Associated with GPEs

Certain activities are associated with GPEs and therefore invoke a GPE role. For example, in *a pro-Iraq rally, Iraq* is assigned a GPE.GPE annotation. A rally is generally concerned with a nation, rather than exclusively a location or government.

> *The Palestinian Authority has banned pro-**{[GPE.GPE**] Iraq**}** rallies, but that ban has been widely ignored.*

Military Activity

Similarly, military activities like invasions, military strikes, bombings, etc. are considered to be acts carried out by and directed at entire nations (not distinguishable from the government, people and location of that nation) and therefore are associated with GPEs. Both the aggressors and the victims in these cases are marked GPE.GPE.

> *The city could have used some special protection in nineteen seventy-nine when **{[GPE.GPE]** the Soviet Union**}** invaded **{[GPE.GPE]** Afghanistan**}**.*

Political Communication and Decision-making

On the other hand, ORGs are responsible for decisions to take military actions. ORGs are also responsible for political communication events such as announcements, agreements, statements, denials, expressions of approval and disapproval, etc. So, if *China* agrees to something, *China* is a GPE.ORG.

> Ritter*'s return is seen as something of a test of that agreement, under which **{[GPE.ORG]** Iraq**}** agreed to give inspectors full access to eight of Saddam Hussein's presidential palaces.*

Embedding

GPE names embedded in mentions of the government have a GPE role. For example, in *the British government*, *British* is a GPE.GPE. This annotation conveys the relationship between nation and government. Similarly, in cases in which the embedded GPE conveys a political relationship with the location, the GPE is assigned a GPE role, as in the **{[GPE.GPE]**Israeli settlement**}**.

However, in cases in which there is only a locative relationship between the GPE and the LOC, the GPE is assigned a LOC role. For example, in *the heartland of America*, *America* is a GPE.LOC because a locative relation is conveyed.

> *Meanwhile, secretary of state Madeleine Albright, Berger and defense secretary William Cohen announced plans to travel to **{[GPE.LOC]** an unnamed city in **{ [GPE.LOC]** the **{[GPE.LOC]** US**}** heartland**}** **}** next week, to explain to the American people just why military force will be necessary if diplomacy fails.*
>
> ***{[LOC]** the**{[GPE.LOC]** Washington**}** area**}***

Athletes, Sports Teams, and GPEs

Athletes and teams are associated with GPE.GPEs as in *Picabo Street of the United States* below. Please note that *Picabo Street* is a person who was a member of the United States Olympic team.

> *Six days into the Nagano Games, one Alpine event _ the women's super-G won on Wednesday by Picabo Street of the* ***{[GPE.GPE]*** *United States}_ has been completed.*

However, when a GPE name is used as a team name (as in *Boston beat Philly*), the entity is marked as a metonymy, with the Literal mention being the city and the Intended mention being the team.

> ***{[GPE.GPE-Lit] [ORG-Int]*** *New York}* *had a shot to win but Chris Childs missed a three.*

In addition, because all GPEs are assigned a role, the Literal GPE mention is assigned a GPE role.

GPEs modifying non-government organizations

In cases where GPEs modify non-government organizations, the organizations are considered to be located in that GPE. Those GPEs should be marked GPE.LOC. So, in *New York corporation, New York* gets a GPE.LOC markup.

> *The* ***{[GPE.LOC]*** *California}* *company also asked that CAI be ordered to pay restitution to CSC "in an amount to be determined at trial."*

Governments

While the entity type for governments is GPE, the role for governments should always be GPE.ORG, while the modifying GPE name should be GPE.GPE if any.

> *But* ***{[GPE.ORG]*** *the Russian government}* *and many politicians will be stridently critical of the United States if they believe they are being ignored.*

(In that particular example, *Russian* would also be marked, so that the full annotation for that phrase would be *{[GPE.ORG] the {[GPE.GPE] Russian} government}*, and the two GPE mentions would be coreferential.)

GPEs and Government Organizations

GPEs modifying government organizations, like *New York police department* and *Kentucky state fire marshall's office,* reflect a relationship between the organizations and the governmental aspect of the GPE, so they are assigned a GPE.ORG markup.

> *The department said Sonabend can appeal to* ***{[GPE.ORG]*** *Switzerland}'s supreme court.*

GPEs and Populations

As stated above, populations of a GPE are treated as GPE.PER. However, it is sometimes difficult to determine whether a reference to people is a reference to the population.

*The **Japanese** have a considerable responsibility for the wars of the first half of the century*

In this example, the phrase *the Japanese* may be interpreted as the population of Japan, or the government of Japan, or the Japanese military, or even some part of the Japanese population. If the annotator believes that the phrase in question refers to the population of the GPE, or most of the population of a GPE, then the annotation should be GPE.PER and the mention is a name mention. However, if the annotator believes the phrase refers to a group of people, then PER is the assigned annotation and the mention is nominal because it does not refer to the name of a person. Examples:

***{[GPE.PER - name]** Cubans**}** have been waiting for this day for a long time.*

***{[GPE.PER - nom]** A majority of **{[GPE.PER - name]** Americans**} }** believe the allegations against Mr. Clinton are true.*

*You and th- **{[GPE.PER - nom]** the **{[GPE.GPE - name]** American**}** people**}** have a right to- to get answers.*

***{[PER - nom]** A majority of **{[PER - nom]** Americans surveyed**} }** believes allegations Mr. Clinton had an affair while he was President are not relevant.*

*Yet another cutting edge development by **{[GPE.PER - name]** the French**}** in their ongoing dealings with their enormous pet population.*

*Butler says those sanctions could end soon if **{[GPE - name]** the Iraqis**}** allow the inspectors to do their job.*

*The Missouri will come to rest near the memorial for the USS Arizona, which was sunk by **{[GPE - name]** the Japanese**}** during the attack on Pearl Harbor.*

{[GPE.PER – nom]** The rest of **{[GPE.PER - name]** America**} }

{[PER - nom]** idealistic Europeans**}

{[PER - nom]** Americans who want to come and, and learn, uh, from the communities how to live in a community, how to take decisions among the community**}

*I do think there is a danger that **{[PER – nom]** some Chinese**}** may underestimate American will on the Taiwan issue.*

Additional examples in Chinese that should be tagged as GPE by decree:

国内,国外,境内,境外,全国,全省…

Summary table of GPE modifiers:

4.2 Mentions

A mention is a linguistic expression – a string of words – used in text or speech to refer to or describe an entity. For each entity, we record and coreference all mentions of the entity. For ACE annotation, a mention is usually an NP (headed by a proper name, a common noun, or a pronoun) and can be nested within another mention. For example, the phrase

The president of Ford

is a mention of an entity of type person, and contains the name "*Ford*", a mention of an entity of type organization. It is also possible for a noun phrase to contain an embedded mention of the same entity. For instance, the phrase

The historian who taught herself COBOL

evokes a person entity with two mentions, the entire phrase and the word "*herself*".

Modifiers such as English proper adjectives (e.g. “American” as in “American flag”), English possessive pronouns (e.g. “your country”) and Chinese pre-nominal proper nouns (e.g. 美国国旗), etc. are also annotated. In other words, mentions in ACE are not limited to syntactic phrases only, but also include modifiers that are nouns or other types of words that have nominal counterparts.

For each mention, we record its head (usually the syntactic head except for proper names which are atomic) and its full extent (head and modifiers if any). But before we continue, we’ll first have an introduction to the Chinese NP.

4.2.1 Noun Phrases in Chinese

Functionally, an NP labels something such as a person, a thing, a class of things, an activity, an event, or an abstract quality or concept and usually appears as a subject or topic of a sentence or an object of a verb or a preposition. A typical NP in Chinese may consist of just a noun (common noun or proper noun) or a pronoun. Both common nouns and proper nouns may take modifiers. The Chinese word order is such that all modifiers – including attributive adjectives, nouns, relative clauses, demonstrative and quantifying determiners, numerals, classifiers/measure words, etc. must all occur BEFORE the noun. Here’s a typical example of complex NP’s with such modifying elements:

我的 昨天买的 那 三 件 蓝 衬衫

“those three blue shirts of mine that I bought yesterday”

Note: Chinese does not have articles like English “a(n)” and “the”. Thus a bare common noun can be used an NP and the (in)definiteness is determined by the context.

4.2.1.1 Singularity/Plurality

Chinese does not have inflection for plurality. There is only one morpheme that can be attached to a noun for plurality, namely 们. However, its usage is rather limited – it only follows a noun that refers to people, most often a multisyllabic noun. Even then it can often omitted. Thus plurality in Chinese is very implicit unless either 们 is used or a plural quantifier/determiner is used.

4.2.1.2. Classifier/Measure Phrases

In Chinese, when demonstratives, numbers, and some quantifiers are used with nouns, they must be followed by a classifier or a measure word (量词) before the noun. The choice of the classifier or measure word will depend the type of entity the noun refers to. And the demonstrative/number/quantifier + classifier/measure word sequence is sometimes refer to as “classifier/measure phrase”.

In Chinese ACE annotation, the classifier/measure phrase is generally not markable. However, the head noun may be omitted if it’s clear from the context/discourse (for example, the entity has just been mentioned in a previous clause/sentence), or in some particular syntactic constructions (e.g. 五个人当中的三个). When such an “elision” happens, the classifier/measure word is marked as the head of the mention and the mention is considered as pronominal in consistence with the English treatment of headless NP’s.

4.2.1.3 The Ubiquitous 的 Marker

The morpheme (or particle) 的 (-de) is everywhere in Chinese. As far as NP is concerned, it has two primary functions: (a) it appears between two an NP – which can be a common noun phrase, a proper name or a pronoun – and a noun to indicate some kind of genitive/possessive/associative relationship between them; and (b) it appears at the end of a clause and nominalize the clause such that the clause can function as an NP by itself, as a relative clause modifying another noun, or as a complement to a following abstract noun head. Note that the –de phrase and the noun head it modifies may not be adjacent to each other: other modifying elements may be inserted in between. Here are some examples:

(a)

我的朋友	*my friend*
公司的同事	*colleagues from (my) company*
宾大的教授	*professor from Penn*

(b)

我不喜欢你喜欢的 *I don’t like what you like*

他去的国家我没去过。我去过的他没去过。 *I’ve never visited the countries he’s been to and he’s never visited the countries I’ve been to.*

我不喜欢你喜欢的电影 *I don’t like the films you like*

他昨天买的苹果我吃了五个 *I ate five of the apples he bought yesterday*

多数美国人赞成政府攻打伊拉克的决定 *Most Americans support the government’s decision to invade Iraq*

As the first example in (a) shows, Chinese does not really have possessive pronouns. The concept of “possession” in Chinese is expressed in the same way for both pronouns and other noun phrases via the –de marker.

The first two examples in (b) are more interesting. In some sense, there’s something missing after –de. This “something” is usually understood by the reader/listener from the context. For example, the second example clearly shows that the missing element refers to a set of countries. In ACE annotation, we consider such NP’s as headless. If the NP is a mention of a markable entity, we treat the particle –de as the head of that mention. To be consistent with the English treatment of headless NP’s, we annotate the markable –de construction as pronominal, though in the future, we may consider a fourth category – headless mentions.

An adjective is a class of words modifying a noun by limiting, qualifying, or specifying. However, since Chinese adjectives are not morphologically distinguished and since nouns can also modify nouns, it's often not clear whether a particular nominal modifier is a noun or adjective. For example, "American Army" – where "American" is the adjectival form of "America" – is simply translated to "美国军队". In such cases, there's no need to consider "美国" as a *proper adjective* – we should just treat it as a regular proper name modifier.

4.2.1.4 Pronouns

Please refer to section 4.2.5.3.1.

4.2.1.5 Proper Names

For ACE annotation, we define a proper name as an NP used to name a unique person, organization, facility, place, or a geo-political entity. A proper name may consist of single proper noun (e.g. "Madonna") or a mixture of common nouns with or without proper nouns (e.g. "Microsoft Cooperation", "Linguistic Data Consortium). As a non-alphabetic language, Chinese has no capitalization.

4.2.1.6 Summary of NP's in Chinese

Simple NP's

- Bare Proper Name
- Bare Common Noun
- Bare Pronoun

Complex NP's

- Proper Name with modifier
- Common Noun with modifier
- (Rare) Pronoun with modifier

Compound NP's

- Two or more head nouns conjoined with conjunctions 和, 跟, 与, etc. with/without modifiers

4.2.2 Mention Extent

The extent of a mention consists of the entire nominal phrase. In case of structures where there is some irresolvable ambiguity as to the attachment of modifiers, the extent annotated should be the maximal extent. In the case of a discontinuous constituent, the extent goes to the end of the constituent, even if that means including tokens that are not part of the constituent. Thus, in

I met some people yesterday who love chess.

the extent of the mention is the entire phrase

{some <u>people</u> yesterday who love chess}

even though “yesterday” modifiers the main VP and bears no syntactic relations with “people who love chess”.

The extent of the mention includes all the modifiers of a nominal phrase, including prepositional phrases, relative clauses, appositional phrases, etc. Thus the phrase

Fred Smith, the noted general

constitutes two mentions of one entity.

{Fred Smith, the noted general}

{the noted general}

Similarly,

Fred Smith, who is a noted general

constitutes two mentions.

{Fred Smith, who is a noted general}

{a noted general}

Punctuations are treated as separate characters. As a rule, we do not include punctuation marks such as commas, periods, and quotation marks in the extent of a mention unless words included within the extent continue on after the punctuation mark. Extent boundaries on both sides must not split a word, though we have yet to define what constitutes a word in Chinese for ACE annotation.

4.2.3 Mention Head

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 mention is “the new glass-clad skyscraper” and the head is “skyscraper” (underlined). Except for proper nouns and proper adjectives, the head is always a single token (or a word as defined in ACE). If the syntactic head of the phrase is a multi-token item, the last token is marked. If the head is a proper name, however, then the entire string sequence of the atomic name is considered to be the head. In the following examples, the mention is enclosed in brackets and the head is underlined:

{Fred Smith} became {the new prime minister}

The job fell to {Abraham Abercrombie III}.

Note, however, a proper name may have modifiers. For example:

{That stupid Fred Smith} became {the new prime minister}.

{聪明的张三}不喜欢{愚蠢的李四}

If the phrase is "headless", as in the case of a partitive construction, the last token/word of the mention is to be marked as the head:

A course in linguistics for {the young} and {the restless}

尊{老}爱{幼}

He was introduced to {five of the analysts}

Note that in the last example, there is a second entity, whose full mention is "*the analysts"* and whose head is "*analysts"*. Please refer to 4.2.4.3.5 for more details on headless mentions.

4.2.4 Markability

This section is devoted to specific NP patterns in need of special attention in ACE annotation and may be frequently updated.

4.2.4.1 Plurals

An entity can be a set with two or more members (though in Chinese, the plurality of the mention head is often implicit):

The injured passengers

受伤的乘客

十名受伤的乘客

Two distinct sets produce separate entities, regardless of whether they have elements in common; so, for example,

Ten passengers were injured, six seriously

evokes two entities, one for the ten passengers, one for the six. Distinct sets produce separate entities, even if they have the same string, so

Five people like vanilla, five people like chocolate

evokes two entities (the five people who like vanilla and the five who like chocolate). Furthermore, a set is a distinct entity from each of its members;

Fred Smith married Harriet Hope; they lived happily for 6 weeks.

evokes three entities, one for Fred Smith, one for Harriet Hope, and one for the set with members Fred and Harriet. The only mention of the set is the pronoun "they".

4.2.4.2 Conjunctions

In conjoined expressions, there should always be one and only one Nominal Entity per head noun. Thus, conjoined noun phrases with no elision of the head noun are to be tagged separately. If a pre-nominal modifier is present it gets included only with the initial noun phrase of the conjunct, and if a post-nominal modifier is present, it gets included only with the final noun phrase of the conjunct.

{Muslims} and {Croats}

{many streams} and {rivers}

{很多学生}和{老师}

{almost every Serb}, {Croat} and {Muslim in Bosnia}

{bus stations}, {train stations}, and {shopping areas throughout the country}

Note that the task of combining such conjoined expressions into "super-entities" is left for higher levels of processing. For example, one could imagine a preprocess for co-reference analysis in which additional entities are derived from conjoined Nominal or Named Entities:

> *{{many streams} and {rivers}} are overflowing their banks.*
>
> *{Jimmy and Rosalyn Carter} donate their time to Habitat for Humanity.*

4.2.4.3 “Word-Internal” Shortened Forms

Chinese does not have contractions such as “let’s” for “let us” or “it’s” for “it is”, nor does it have acronyms such as “U.S.” or “U.N.” (except for borrowed words where roman letters are used). However, Chinese word-formation often shortens long words or phrases into disyllables (which can be above the word level). One will frequently find that a single syllable/character is used in place of a full country or region name such as “一中一台” (“one China, one Taiwan”), “美英两国” (“the U.S. and the U.K.”), and so on and so forth. Take “华” as an example. It’s a bound morpheme such that it cannot usually be used alone as a syntactic word. But it can be used as an object of a coverb/preposition and together they modify other phrases as in “对华政策” (“policy on China” or simply “China policy”), “驻华大使” (“ambassador to China”). For ACE annotation, such shortened forms are annotated just as their full-fledged forms even if they are considered bound morphemes. This marks an important departure from the Penn Chinese Treebank system (CTB) and other Chinese syntactic approaches. It also serves as one of the major reasons that Chinese source texts in ACE are not preprocessed for word segmentation – without word segmentation, the annotator has the “liberty” to look into the “word-internal” structure and annotate mentions that are otherwise not taggable (e.g. “日” as in “日货”). Here are some examples:

> *{{中}{美}两国}继续就{北韩}问题进行磋商*
>
> *{美}式武器*
>
> *抵制{日}货*
>
> *{{美}驻{华}人员}*

4.2.4.4 Elision

Where elision of the head noun occurs in a conjunction (explicit or implicit), a single entity is delineated (these could also be viewed as conjoined modifier phrases):

> *{the rain-soaked mid-Atlantic and new England **states**}*
>
> *{the successful and socially-responsible **manufacturers**}*
>
> *[GEP.GOV]{[GPE.GPE]{British} and [GPE.GPE]{Irish} **governments**}*
>
> *[GPE.GEP(?)]{[GPE.GPE]{英}[GPE.GPE]{美}两国}*
>
> *[GPE.ORG]{[GPE.GPE]{[GPE.GPE{英}[GPE.GPE]{美}两国}**政府**}*

Note that in the last example, four entities are invoked.

4.2.4.5 Range Expressions and Elision

Components of range expressions are tagged separately if there is no elision of any head noun:

from [the foothills] to [the prairie]

from [the downtown area] to [the suburbs]

从{农村}到{城市}

从{北京}到{上海}

However, in examples like the following there is only a single head noun. In these cases we will treat the range expression as a pre-modifier, so that it gets included in the maximum extent of the entity:

ranging from {five to six companies} per day

每天有{五到六家公司}

4.2.4.6 Predicate complements

Mentions should include nominal predicate complements that are affirmatively asserted of a reportable entity, since they describe the entity. Thus

Fred is a real linguist.

evokes an entity of type person with two mentions, "*Fred*" and "*a real linguist*". (Thus, the question of whether the usage is “generic”, as discussed below, does not arise in this context.) On the other hand,

Fred is not a real linguist.

evokes two entities: one of type person with only one mention, "*Fred*" and one of type person that is generic with only one mention “*a real linguist*”. Similarly,

Fred is studying to be a real linguist.

evokes a specific entity of type person with only one mention, "*Fred*" and a generic entity of type person with one mention, “a real linguist”, because the text does not assert that Fred has been, is, or will be a real linguist.

Note that the way we handle predicate complements (and likewise appositional predicates) may change in the future.

4.2.4.7 Apposition

According to NOAH, apposition is a “construction in which a noun or noun phrase is placed with another as an explanatory equivalent, both having the same syntactic relation to the other elements in the sentence.” For example,

The painter Copley was born in Boston.

To some extent, appositives are “syntactically equal” and therefore it shouldn’t matter which one extends to the entire structure. For ACE annotation, the default rule is that the first mention extends to the entire structure, for both English and Chinese, like so:

{The painter {Copley$_2$}}$_1$ was born in Boston.

{ Copley, the {painter}$_2$}$_1$, was born in Boston.

{中国北方最大的沿海城市 {天津市}$_2$}$_1$

In the simplest case where both appositives are single mentions as above, the two mentions are coreferenced. So the relation table for the last example looks like follows:

	Named mention	Nominal mention
1	天津市	中国北方最大的沿海城市天津市[城市]

However, complications arise when apposition appears with conjunction with the following possible patterns: single mention + conjoined mentions; conjoined mentions + single mention; and conjoined mentions + conjoined mentions.

Our proposed approach is as follows: if one and only one of the appositives is a single mention, maximize the extent of the single mention; otherwise, maximize the extent of the first mention. Coreference is allowed iff there is a one-to-one reference relationship between the two mentions.

Thus the three other possibilities are as follows:

single mention + conjoined mentions:

{中国沿海城市 {天津}$_2$ 和{珠海}$_3$}$_1$

	Named mention	Nominal mention
3		中国沿海城市天津和珠海[城市]
4	天津	
5	珠海	

conjoined mentions + single mention

{{中国贸易中心}$_2$ 和{最大城市}$_3$ 上海 $_3$}$_1$

	Named mention	Nominal mention
3	中国贸易中心和最大城市上海[上海]	中国贸易中心[中心] 最大城市[城市]

But

{{John}$_2$ and {Marry}$_3$, the happiest couple in the world}$_1$

	Named mention	Nominal mention
3		John and Mary, the happiest couple in the world [couple]
4	John	

5	Mary	

conjoined mentions + conjoined mentions

{沿海城市}$_1$ 和{贸易中心{上海}$_3$ 和{天津}$_4$}$_2$

	Named mention	Nominal mention
4		沿海城市 贸易中心上海和天津[中心]
5	上海	
6	天津	

In 4.2.4.1.1 we distinguish between positions and titles when they are used with names. Titles are treated as modifiers to the name head whereas position + name structures are treated like appositions. It should now be easier to see how we should annotate the 4 types of person entities involving positions and names given the above description of appositions

{副总理{钱其琛}}

{副总理{钱其琛}和{温家宝}}

{{副总理}兼{外交部长}钱其琛}

{作家}兼{记者{张三}和{李四}}

Expressions such as 中俄朝三国 are not considered as appositions. Rather, 中俄朝 are considered as modifiers. Thus this phrase has four mentions as follows:

{{中}$_2$ {俄}$_3$ {朝}$_4$ 三国}$_1$

Also distinguish from: {上海} {这颗东方明珠}; {西安} {这座历史悠久的城市} – these are not considered as appositions.

4.2.4.8 Proper adjectives

Does not apply to Chinese.

4.2.4.9 Quantified and partitive phrases

In English, a partitive construction of the form

quantifier of *ENP*

gives rise to two mentions: one for the entire phrase, and one for the embedded noun phrase *ENP* that is the object of "of". If the entire phrase represents a subset of *ENP*, these will be mentions of distinct entities. Thus in

three of the women

evokes two entities, for "*the women*" and "*three of the women*". Similarly,

some of the women

evokes two entities. On the other hand,

all of the women

has two mentions of *one* entity: "*the women*" and "*all of the women*" (the same set). This is also the case with the partitive-like phrase

a team of five experts

since the team is identical to the set of five experts.

Similar constructions in Chinese are easier to annotate since they must use the “container” word (当)中 with the marker 的 as NP_2 (当中)的 NP_1, where NP_1 (extending to the entire structure) and NP_2 evoke two distinct entities as in:

这些妇女当中的三人

这些妇女当中的一些人

Such expressions are not very common in Chinese and there’s no equivalent expression of “all of the woman” in Chinese. However, the pattern of 其中+NP is very frequent, where 其 is tagged as a pronominal mention. This happens when the larger set referred to by 其 was previously mentioned in the discourse.

4.2.5 Types of Mentions

We distinguish between mentions with a named head (name-mentions), those with a common noun head (nominal or nom-mentions) and those with a pronominal head (pro-mentions). Mentions with empty heads ("*five of the analysts*") are classified as pro-mentions.

4.2.5.1 Names

The terms “proper noun” and “proper name” are often used interchangeably. Although we will not attempt to distinguish them theoretically, it’s important to remember that a proper name may be formed entirely by common nouns (e.g. “Linguistic Data Consortium”) or by a mixture of common and proper nouns (e.g. “Microsoft Corporation”).

A named mention is a mention headed by a proper name. Often the proper name head is also the full extent of the mention, though proper names can have modifiers.

Names are atomic. This means that entity names wholly contained within another name are not annotated and that the proper name head is in general also the full extent of the mention. For example, in the following phrase only one entity is referenced.

The New York Times

This phrase references the organization of the newspaper. It does not evoke a separate entity for the city of "New York".

Likewise, :”北京” in the phrase “北京日报” does not evoke a city mention of “北京” because it’s an integral part of the entity name.

Proper names may have other modifiers that are not part of the name but should be included in the extent of the mention, for example, 贪婪的微软公司. The modifier can even be another name mention like “美国” in “美国微软公司”.

4.2.5.1.1 Head and Extent of Names

The following are head and extent rules that are specific to Name mentions.

Titles (and honorifics) and Positions

With a few exceptions, most position expressions in Chinese can also be used as titles. There is, however, a syntactic difference between the two structures. Titles in Chinese must appear after the person’s name and cannot take any modifiers. By contrast, positions appear before the person’s name and can take additional modifiers. Compare the following expressions:

> *{江泽民主席}*
>
> *{{中国} 国家主席 {江泽民}}*

where 主席 follows the name in the first example whereas it precedes the name in the second example. Thus the word order marks an important distinction between the two structures. Furthermore, pre-name positions can be con-joined and multiple names can appear after a position expression if that expression can imply multiple positions. Therefore, it’s OK to say

> *国务院副总理钱其琛和温家宝*
>
> *记者张三李四*

However, 钱其琛和温家宝副总理 doesn’t mean that 钱其琛 is also a vice premier.

There are also a few cases where expressions of position and title use different words. For example, 老师 and 教师 both mean “teacher”. Yet, 教师 is always used as a position whereas 老师 is usually used as a title, although it can be used as a position, particularly when it has a modifier such as 我们学校的老师张三.

Another difference is that a title can just be preceded by the person’s surname, but a position expression cannot be followed by the surname only. 国家主席江 looks rather awkward to the native speaker of Chinese.

In English, it appears that the following expressions

> *US President George Bush*
>
> *President Bush*

look identical. Indeed in earlier ACE annotation, they were treated the same: both were considered as title + name where the title is not part of the name head. The first example was thus different from

> *the US President George Bush*

where “the US President” and “George Bush” were tagged as appositions.

However, upon further scrutiny, the two English example bare some similarities to their Chinese counterparts. For example, it's OK to say

Former US Presidents Jimmy Carter and Bill Clinton

But it's not OK to say

**Presidents Bush and Jiang Zemin met last year.*

It appears that although English does not have the word order distinction as Chinese, it bears some similarities with Chinese. In English, it's also OK to pause between "US President" and "George Bush", but not OK between "President" and "George Bush". Furthermore, honorifics can precede a title, but not a position. Compare:

Mr. President

**Mr. US President*

In ACE annotation, we decide to distinguish the two structures as follows:

- A bare title + name in English or a name + a bare title in Chinese has only one mention where the name is head and the title is included in the extent of the mention as a modifier.

{江泽民主席}

- Complex titles are treated as positions and the position expression and the name(s) are appositions.

{{中国} 国家主席 {江泽民}}

For how to annotate appositions, refer to 4.2.4.7.

Organization name with GPE modifier

Proper names are considered atomic in ACE annotation, which means if a proper name contains another name as its integral part, the other name is not taggable. This can sometimes be troublesome when it comes to certain organization names as the annotator may not be able to determine if an embedded name of another entity is just a modifier or part of the name.

For government organizations, our approach is NOT to include their GPE "pre-modifier" in the head of the mention like so:

{{中国}$_2$外交部}$_1$

even though the official letterhead name is "中华人民共和国外交部". This is more or less consistent with the English approach.

Note that words such as "外交部", "国防部", "外事局" etc are NOT necessarily named mentions, e.g. "两国外交部", "各市外事局". Therefore, the annotator must determine whether such mentions are named mentions or nominal mentions from the context.

Likewise, 山东省政府 evokes two mentions of the same entity: {{[GPE.GPE]山东省}$_2$[GEP.ORG]政府}$_1$.

For non-government organizations, the annotator will have to rely on their world knowledge. For example, 美国 as in “美国微软公司” is clearly a mention of another entity and not part of Microsoft’s name whereas 南京 as in “南京石油化工公司” is more likely part of the name and hence should not be tagged as another entity mention.

4.2.5.1.2 Markable Names

The following are markability rules that apply specifically to name mentions.

Aliases: Shortened Forms and Nicknames

Generally, aliases for entities are to be tagged. Taggable aliases will include the following forms of entity names:

Truncated/shortened names, provided that the resulting form is clearly a proper noun referring to a specific entity, for example in:

Red Sox	alias for the Boston Red Sox
Sears	*alias for Sears Roebuck and Co.*
微软	微软公司
北大	北京大学

Nicknames and other aliases are tagged as names when they are established alternate ways of referring to an entity; if the annotator does not recognize the status of the nickname, it may be possible to determine from context whether the nickname is “established” or not.

The Big Apple	nickname for New York City
The garden state	nickname for New Jersey
山城	nickname for 重庆

Entity Names that Modify Persons/Positions

Entity names modifying a person or their positions are to be tagged.

Microsoft founder Bill Gates

The U.S. Vice-President

{{美国}国务卿}

Each of the examples above gives us two mentions.

Please note that nominal mentions of entities, which modify a person or their position, are not to be tagged.

company chairman James Smith

This example yields only one mention: “company” is not tagged.

4.2.5.2 Nominals

For the purposes of the ACE project, a nominal is a noun phrase headed by a common noun.

4.2.5.2.1 Nominal Modifiers

Common nouns directly modifying other nouns are not markable mentions.

Markable:

> *I love {French} food.*

Not Markable:

> *I love {prison} food.*
> *I love French fries.*

In the last example, “French” is part of the compound noun “French fries”, not a modifier.

Markable:

> *这个公司的经理*

Here “这个公司” is an NP, “公司” does not directly modify “经理”.

Not Markable:

> *明天召开全市公司经理会议.*

4.2.5.3 Pronominals

A pronominal is a word used as a substitute for a noun phrase. Pronominals refer to persons or things that are previously specified or understood from the context. Pronominals are marked whenever they reference a salient entity. The following are some additional rules that apply to pronominal mentions in Chinese.

4.2.5.3.1 Person Pronouns

The following table lists person pronouns in Chinese

	Singular	Plural
1st Person	我/咱	我们/咱们
2nd Person	你/您	你们
3rd Person	他/她/它	他们/她们/它们

As shown above, the plural form simply consists of a singular pronoun and the plural morpheme 们. Also, there are a few dialectal variations such as 俺.

Person pronouns in Chinese do not usually refer to animals and inanimate objects, not even the third person pronouns. The compensation for this is that Chinese is a PRO-drop language.

With a few exceptions such as 可怜的我 (“Poor me!”), Chinese person pronouns cannot usually be modified.

Lack of possessive pronouns in Chinese is actually a plus instead of a minus for ACE annotation. Consider:

你的孩子来了，我的没来.

Your child has come. Mine hasn't.

The English possessive pronoun "mine" here actually invokes two entities, "me – the speaker" and "the speaker's child". But there's no easy way to mark the two entity mentions in English. For Chinese, however, we can simply mark "我的" as one mention with "的" as head (see "headless mentions" below) and "我" as another mention: $\{\{我\}_2 \underline{的}\}_1$.

4.2.5.3.2 Reflexives

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 functions it serves in a sentence, it's always marked as pronominal if the entity it refers to falls into one of the five types.

However, the morpheme 自己 can appear immediately after a person pronoun (such as 他自己, 他们自己) and it's sometimes not clear which function it serves. For simplicity, we decide that if the sequence is in an object position (verbal or prepositional), mark it as a single pronominal mention. Otherwise, mark the pronoun and the reflexive as two pronominal mentions (of the same entity).

4.2.5.3.3 Demonstratives

Demonstratives (这, 那, 这些) are marked as pronominal when appear alone as NP's and stand for previously occurred NP's (or their reference can be determined from the context). They are not markable when they server as determiners of NP's.

4.2.5.3.4 Other pronouns and pronoun-like words/phrases:

- 其: third person pronoun used in written/formal language. Mark 其 as a pronominal mention if the entity is one of the five types. Additional expressions consisting of 其:
    - 其中: "lexicalized" phrase meaning "among them; of them; in it". Mark 其 as a pronominal mention if the entity it refers to is one of the five types.
    - 其间: "lexicalized" phrase meaning "between them; in it". As with 其中 above, mark 其 as a pronominal mention if the entity it refers to is one of the five types.
    - 其他/其它: highly lexicalized expression meaning "the rest of them; the others". Syntactically, it behaves a like a determiner. It's usually used a context where a subset of a set has previously been mentioned and this expression, with optional noun head and/or –de, is used to refer to the other subset. As with demonstratives, stand-alone 其他/其它 is marked as headless pronominal.

    - 其余: Similar to 其他/其它.

- 彼此: meaning “each other”. Treat the entire unit as pronominal. However, 相互 and 互相 is not taggable because they are adverbials.
- 本: this morpheme, originally meaning “the root of a plant”, can function as a “determiner” meaning “this” – often related to the speaker. Two words with this morpheme draws special attention.
    - 本人: can mean “I (myself, me)” or “oneself/in person”. In either case, treat it as pronominal.
    - 本身: meaning “itself/in itself”. Treat it as pronominal.

  If 本 is only used as a determiner, do not treat it as pronominal.

  *缅甸贸易部长吞基和泰国外交部长格森*格森西分别代表{本国政府}在协定上签字。(Chtb_027)*
- 自身: There are many words with 自 as their component morpheme. Among them, we consider 自身 as pronominal similar to 自己. We DO NOT mark 自 that’s part of a compound/complex verb (for example, 自重, 自爱, 自筹, like self-determination, self-esteem in English), where 自 is non-specific anyway.
- 前者/后者: meaning “the former/the latter”. Treat them as pronominal.
- Locatives with demonstrative determiners: 这里, 那里, 这儿, 那儿

  Strictly speaking, these expressions are full-fledged noun phrases and used as locative phrases. Morpho-syntactically, they’re different from “here” and “there” in English. But for ACE, we treat them just like “here” and “there” and tag them as pronominal mentions.

4.2.5.3.5 Headless Mentions

Headless mentions in ACE refer to those mentions without an explicit syntactic or semantic head either due to syntactic elision or as required by the syntactic structure. Except for non-final members of conjunctions, headless mentions are classified as pro-mentions, though this may change in the future.

> ***five** of the analysts*

Please note that this example also includes the nominal mention [the analysts].

The following shows some typical headless mentions:

Classifier/Measure Phrases Without Noun Head

In Chinese ACE annotation, the classifier/measure phrase is generally not markable. However, the head noun may be omitted if it’s clear from the context/discourse (for example, the entity has just been mentioned in a previous clause/sentence), or in some particular syntactic constructions (e.g. 五个人当中的三个). When such an “elision” happens, the classifier/measure word is marked as the head of the mention and the headless mention is considered as pronominal in line with the English treatment of headless NP’s.

–De (-的) Phrases Without Noun Head

Regardless of what function the morpheme –de serves, if it is not followed by a noun head, the mention is headless pronominal and –de is tagged as the head, like so:

> *他去的国家我没去过。{我去过的}他没去过。*
>
> *张三的父母退休了。{李四的}还没有。*

4.2.5.3.6 Pronouns Referring to GPEs

Like nominal mentions, pronouns that refer to GPEs, marked as mentions of the same GPE entit, may not be assigned the same role as , which may not be the same role of the antecedent.

> *Composite Example: The president flew to **{[GPE.LOC]** Israel**}** to meet with **{[GPE.GPE]** its**}** Prime Minister.*

Similarly, in the case of classic metonymies (where two entities are created), pronoun annotation is determined in part by the link to the antecedent and in part by the context in which the pronoun appears. If the antecedent is a classic metonymy, the pronoun will be a mention of the same entity as either the literal mention or the intended mention of the antecedent.

> *Metonymy Example: Thousands of parochial school and college students are joining this year's demonstration, including 1,500 high school students from across the country who spent last night at **{[ORG-Literal][FAC-Intended]** Catholic University**}**. **{[FAC]** It**}**’s in Georgetown.*

In some cases, the antecedent is not a metonymy but the context of the pronoun invokes an entity with a type that is different from that of the antecedent. In such cases, in addition to the mention of the new entity, the annotator should also mark the pronoun as a literal mention of the antecedent entity. (This allows us to maintain the connection between the pronoun and the antecedent.)

> *Metonymy Example: **{[FAC]** The museum**}** is located on 45th Street. **{[FAC-Literal] [ORG-Intended]** They**}** just hired a new guard.*

Since pronouns are rarely used in Chinese to refer to non-human entities, the annotator may find very few such cases in Chinese texts.

4.2.6 Coreference of Mentions

If two mentions refer to the same underlying entity, we must indicate this by coreferencing them. In most cases, this is very straightforward. In an article about Osama bin Laden, we want all mentions of Mr. bin Laden to be lumped together in the same entity and marked with the base type PER. So, if the following sentences appeared in the same article, we would want to include all the bold mentions in the Osama bin Laden entity.

> *Videos circulated by* ***Osama bin Laden*** *have added to the evidence linking* ***him*** *and the al-Qaida network to the Sept. 11 terrorist attacks in the United States, the government said Wednesday in an updated dossier on the investigation. The document, published by Prime Minister Tony Blair's office, said* ***the Saudi dissident*** *had come "closest to admitting responsibility" for the attacks in an "inflammatory video," allegedly made on Oct. 20, that was not released to the media but circulated to al-Qaida members. "The battle has been moved inside America, and we shall continue until we win this battle, or die in the cause and meet our maker," the document quotes* ***bin Laden*** *as saying.*

The name mentions of Osama bin Laden are easy to spot. Please note, however, that we must coreference all mentions that refer to the entity that is Mr bin Laden. This will include nominal mentions such as the Saudi dissident and pronominal mentions such as him.

5 Metonymy

Metonymy occurs when a speaker uses a reference to one entity to refer to another entity (or entities) related to it. For example, in the sentence below *Beijing* is a capital city name that is used as a reference to the Chinese government:

> ***Beijing*** *will not continue sales of anti-ship missiles to Iran.*

Classic metonymies make reference to two entities, one explicit and one indirect reference. Common examples are cases of capital city names standing for national governments, as shown above.

For ACE, the notion of metonymy has been extended to include the case where an entity has an established base type and but is used for a different type in the context. Common examples involve facilities and organizations, which are closely related in that organizations typically have facilities, and facilities are typically owned and administered by organizations. Thus when a facility is mentioned, the organization is sometimes also referenced. So, in *the museum announced its new exhibit*, the entity *museum* is a facility that houses artwork, but in this context it is the organization running the museum that is doing the announcing.

In cases like this, where both entities are expressed by the same phrase, two entity mentions – literal and intended – should be marked, one for each of the corresponding references. In the above example, the annotator would first mark mentions of a FAC entity (literal) and then an ORG entity (intended) for *the museum*.

If, elsewhere in the document, a mention of "the museum" occurred in the context like "New windows were ordered for the museum", that mention would be marked as an additional mention of the same FAC entity referred to above, but not as an additional mention of the ORG entity.

Note that we tag metonymies iff both entity types fall into one of the five (possibly more in the future) categories. If one of them is not a taggable entity type under the current ACE specifications, we do not mark two mentions. So 他是个[狐狸], 狐狸 only has one mention of a PER entity and is tagged as intended, even though its literal meaning is an animal. That's because the current ACE definition does not include animals.

The remainder of this section outlines specific annotation guidelines for metonymy in different contexts.

5.1 Capital City for Governmental GPE

Cases in which the capital city is used to refer to the nation's government are marked as true metonyms. (Because two separate GPEs are involved, this is not an exception to the general rule that GPEs are marked as one entity with a role rather than as two entities.)

> *Secretary of Defense William S. Cohen said today that he is satisfied **{[GPE.GPE-literal][GPE.ORG- intended]** Beijing**}** will not continue sales of anti-ship missiles to Iran as he wrapped up a four-day visit here that underscored improving Sino American military ties.*

In this example there are two mentions covering the word Beijing. The GPE.GPE is a mention of the city Beijing and the GPE.ORG is a mention of China. The GPE.ORG mention is a mention of the same China entity that would be referred to by other GPE mentions of "China" that might be found elsewhere in the document. Also if there were a later mention of the city of Beijing (for example, *Cohen left the city this morning*), it would be a GPE.LOC mention of the same Beijing entity referred to by the GPE.GPE mention in the above example.

5.2 Metonymies Involving ORG Base Entities

There is a table (see the Pilot Study task definition, Section 6.2.5) that specifies a "base" type for various kinds of entities. Mentions of entities with ORG base types like schools, restaurants, or churches are sometimes used to refer to the organization itself, and sometimes used to refer to the facility that houses that organization. Every mention of such an entity is to be marked (at least) as a mention of an entity of its base type. A second mention of a different type should also be marked if the context invokes a metonymic entity. Thus a mention whose base type is ORG but that is used in a FAC context will have mentions of both of those two entities associated with it.

Below are some examples of ORGs that refer either to a single base type entity, or else to both a base type and metonymic type entity.

Example 1

Universities have an ORG base type so both mentions of the university in 1A and 1B invoke an ORG entity. But 1B also invokes a FAC entity because it refers to the site.

> *A. Lee Jung Hoon, a political science professor at **{[ORG]** Yonsei University**}**…*

B. *Thousands of parochial school and college students are joining this year's demonstration, including 1,500 high school students from across the country who spent* last night at **{[ORG-literal] [FAC-intended]** Catholic University**}.**

Example 2

Embassies have an ORG base type so both 2A and 2B invoke an ORG entity. But 2A also invokes a FAC entity because FACs, not ORGs have gates.

A. *…a few hundred ethnic Albanians laid a black wreath at the gate of* ***{[ORG-literal] [FAC-intended]*** *Yugoslavian <u>embassy</u>**}**.*

B. *"Our Ministry of Defense is working very hard with* ***{[ORG]*** *the <u>U.S. Embassy</u> in Bogota**}** to get the information together," Cano said.*

5.3 Metonymies Involving FAC Base Entities

The same approach used for ORG entity mentions that refer to an associated FAC should also be used when a FAC entity mention refers to an associated ORG.

Here are two examples from the same document:

A. *Competing self-images of victim hood have long prevented Israelis and Arabs from acknowledging the full weight of each other's historical tragedies, and many Arab leaders have resisted efforts to lure them to* ***{[FAC]*** *the museum**}** and the similar Yad Vashem memorial in Jerusalem.*

B. *Lerman, reached at his New Jersey home, said the subject of Arafat and Israel's talks with the Palestinian Authority still profoundly divided U.S. and world Jewry and "we believe* ***{[FAC-literal] [ORG-intended]*** *the museum**}** should not get involved in a political dispute where half of the people are for something and half are against it."*

Since museums have a FAC base type, both examples A and B invoke a FAC entity. But example B also invokes an ORG entity because it is the organization that should not get involved in the dispute.

Note in the above examples that the two FAC mentions refer to the same FAC entity, as shown in the following table of entities and mentions:

Entity 1: ***{[FAC]*** *the museum**}**,* ***{[FAC]*** *the museum**}***

Entity 2: ***{[ORG]*** *the museum**}***

Another common class of FAC metonymies is found when named buildings are used to refer to the organizations based there:

It is unlikely ***{[FAC-literal] [ORG-intended]*** *the White House**}** would nominate a successor who did not support sampling, and equally unlikely Republican leaders would look favorably on such a candidate.*

5.4 Special Rule for Offices and Branches

Because the term "office" in English is often used to refer to an organization (branch), as in "the Office of the Attorney General," the base type for offices will be ORG in English. When the context suggests a reference to the physical entity, the entity should be marked both ORG and FAC. Examples that are ambiguous as to whether a facility or an organization is intended should be marked metonymically, with both an ORG and a FAC mention. Thus in the following example the office is marked both ORG and FAC because it is unclear whether the context suggests that the investigators are from the physical office or from the organization.

> *Investigators from* ***{[ORG-9] [FAC-10]*** *the Kentucky state fire marshal's <u>office</u>****}****.*

(In that particular example, *Kentucky* would also be marked, so that the full annotation for that phrase would be ***{[ORG-9] [FAC-10]*** *the* ***{[GPE.ORG]*** *<u>Kentucky</u>****}*** *state fire marshal's <u>office</u>****}***.)

The same general guidelines apply to other facility terms like "branches" (as in the local branch of a bank).

However, in Chinese, the base type for 办公室 is FAC. Its usage as an organization (e.g. 这个办公室的主任) is rather limited in Chinese (and perhaps more in translations). By contrast, the base type for 办公厅 is ORG and is rarely used as FAC. In Chinese, departments within a large organization have more specific words as such as xx 科, xx 处，xx 局，etc. The base type for such words is ORG and it's possible to refer to them as FAC.

5.5 Metonymies Involving LOC Base Entities

Entities whose base type is LOC can also be used in metonymic senses. In the following example, "the world" has literal type LOC but intended type PER, and thus is annotated with two separate mentions:

> ***{[LOC-literal] [PER-intended]*** *The whole <u>world</u>****}*** *was watching.*

6 Entity Class (Generic/Specific)

An entity is generic when it does not refer to a particular object or particular set of objects in the world. One might also say that an NP denoting a generic entity is used non-referentially. In general, only nominal and certain pronominal mentions can have generic interpretations.

Every entity must be designated as either generic or specific. In some cases this distinction is difficult to make.

This section will outline several tests that will help differentiate between the two classes. To help the annotator better understand the issue, and also due to lack of literature in Chinese linguistics on the subject, we're keeping most of the original materials from the English version.

6.1 Definition of Generic and Specific

A given common noun (*girl*, *motorcycle*, *bookmark*, *semantic theory*, etc.) denotes a *set of objects*, each of which is an example of the noun in question. In such a system, "boy" would refer to the set BOY whose membership would be precisely *all the boys in the world* (or perhaps: *in the Universe*).

The manner in which NPs refer can be easily explained relative to this backdrop:

1. Some NPs are used to refer to *a particular object in the world*. The set X (the common noun's referents) from which that object is drawn has little significance to the audience, other than to help in the selection of the (particular) object in question.

These NPs say something like: *there is a specific example of X, one that I have in mind, that ...* and are considered to be *non-generic*.

(Note that we will use non-generic and specific interchangeably in the present set of documents. The former is arguably more appropriate, since the annotation conventions adopted here tag the feature GENERIC as either *true* or *false*, but we will let the latter serve as form of shorthand notation.)

2. Other NPs are used to refer to *underspecified objects that may be an example of the set (X) in question, but need not be particular*. Here the set X has a greater degree of significance, since the only constraint on the entity in question is that it be drawn from that set.

These NPs say something like:

> *"Any member of the set X ..."; or*
>
> *"Each member of the set X ..."*

and are considered to be *generic*.

In short, a generic mention is used to refer to *any member of the set in question* rather than *some particular, identifiable member of that set* (which would be picked out by a *non-generic* mention) and a formal definition seems altogether impossible. As shall soon become clear, we can do little better in providing this notion with a precise definition.

We have therefore allowed the above informal (*folk*) definition --- together with the following discussion of the phenomena; the subsequent taxonomy of common generic-denoting mentions; and the concluding short list of (non-deterministic) tests for the applicability of generic status to a given mention --- to serve as the basis of our tagging decisions with regard to the attribution of generic status.

The (un-)reliability of syntactic or contextual tests here will become clear as the discussion proceeds --- it is helpful to correspondingly consider each of the examples which follow as having a (frequently secondary) role in illustrating this fact, whether or not this expository role is explicitly stated.

6.2 Classes of Mentions Frequently Associated with Generic Entities

We can make some loose generalizations about the classes of NPs, which are likely to refer to generic entities, but it is important to bear in mind the source of our reluctance to offer such categorical (or syntactic) criteria for the assignment of generic status to a given NP.

Typically, generic entities include types of entity, suggested attributes of entities, hypothetical entities, and generalizations across a set or sets of entities.

6.2.1 A Type of Entity (种，类)

This is the most typical case of generic mentions. Think of this as a kind/type entity, or better yet, think of the NP as a proper name for a class of entities.

{Mammals} are live bearers.

{Good students} do all the reading.

{Typical firemen} work hard all their lives in dangerous conditions.

{鲸鱼}是哺乳类动物。

{鲸鱼}是一种海洋动物。

{好学生}爱读书。

牛只吃草。

6.2.2 A Suggested Attribute of an Entity

John seems to be {a nice person}.

(cf. John is a nice person)

{Misfits} are sometimes {the best employees}.

张三看上去是个{好人}。

(比较：张三是个好人)

Note that for historical reasons, NP’s in affirmative predicate complement (e.g. “John is a nice person”) and appositive predicate (e.g “John, a real linguist, …”) are considered co-referential to the subject NP. While this may change in the future, it’s important for now to make the distinctin.

6.2.3 A Hypothetical Entity (假象, 假设)

If {a person} steps over the line, {they} must be punished.

Aides say he's plotting a political comeback, even considering a run for president} in two thousand.

只要{你}努力，就一定能成功。

如果{有人}打电话, 请帮我接一下。

6.2.4 A Generalization across a Set of Entities

{Outsiders} think that New Jersey is a different country.

{Purple houses} are really ugly.

{紫房子}难看。

Even if the *property* or the *set* underlying the entity in question is extremely constrained (i.e. such that there are very few possible members), that entity should still be considered *generic*.

{People who drive at night in red cars} are likely to get tickets.

The police are looking for {a man who wears green suits and carries a purple briefcase}.

The first of these examples falls into the *Type of Entity* category. The second is a *Hypothetical Entity*. The man in the second example may or may not exist (even though the police are looking for him).

Note that this mention would not be generic if the context went on to say specific things about the man wearing green suits. We have seen several examples of this case above. This is only *generic* if it is unclear if such a person actually exists.

6.3 Tests for Generic-hood

6.3.1 Words that are commonly generic

'anyone', 'any X', 'most Xs', 'more Xs' tend to be generic, even if the author has someone in mind.

{Anyone who carries a gun} is dangerous.

{Most doctors} are just in it for the money.

{More investigators} are needed for this case.

Similar quantifiers in Chinese include "任何 x", "(绝)大多(数)", "少数" etc.

One problem this class of noun phrases is the scope/domain in which the entities are referred to. The domain for the above example is the entire universe. But how big should the domain be? In other words, how are going to handle the case where the quantified NP has a domain modifier such "anyone from the earth/country/state/city/university/class/group"?

6.3.2 Determiners and Chinese classifiers/measure words

Generic noun phrases of the type "a" + singular noun or bare plurals can be distinguished using tests such as:

1. These noun phrases in negated contexts are *generic*:

I didn't see {gorillas} here. [generic]

I saw gorillas {a gorilla} here. [specific]

2. These noun phrases in modal contexts (such as *belief*, *desire*, ...) are *generic*:

I want to see {gorillas}.

I thought I heard {a gorilla}.

3. These noun phrases in questions are *generic*:

Have you seen {a gorilla} walking by?

Have you seen {gorillas} wearing hats?

Bare plurals with individual-level predicates are *generic*. *Individual-level predicates* mark characteristics of individual members of a set, e.g., "birds have wings" means that each bird has wings. In contrast, stage-level predicates ("Gorillas are wrecking my garden", "Gorillas are available") can be either *generic* or *non-generic*, depending on context.

Thus the subjects are *generic* in the following sentences:

{Gorillas} are intelligent

{Linguists} know French.

{Birds} have wings.

Occasionally noun phrases with "the" are generic, even though this is not typically the case. We find this when "the" plus a singular noun is used to represent a set, e.g.,

Turing invented {the computer}. [generic]

I wrote this on {the computer in my office}. [specific]

{The dodo} is extinct. [generic]

{The dodo} is dead. [specific]

Chinese does not have determiners similar to “a” and “the”, nor does the NP inflect for number. Recall, however, that Chinese often need a classifier or measure word when the head noun in question is preceded by a demonstrative or numeral. Other than the number “一”, other numerals and demonstratives should always indicate specific entities. “一” + classifier + noun is subject to similar interpretations as the English determiner “a” shown above.

6.3.3 Positive Assertion Test

This test applies to predications such as "*X* is *Y*" (as in the subsequent example). If *X* is specific, then *Y* will be as well, because *Y* is positively asserted of *X*. *Y* is assumed to be coreferential with *X* and therefore specific.

{Joe} is {a nice guy}.

If *X* is *generic* and *Y* is positively asserted of *X*, then *Y* is also *generic*.

{Firemen} are {nice guys}.

This test is less effective when someone other than the author of the story makes the positive assertion. This is just an instance of the case in which a modal context forces a generic reading (as in II-2 above).

Mary says that {Joe} is a {a nice guy}.

This sort of statement falls into the pattern

person Z says/said/thought/*etc. that* X *is* Y

This only counts as a positive assertion if *Y* is not an attribute and person Z is a trustworthy source of information. This case, however, is the exception rather than the rule. Most modal contexts are entirely opaque, and the assertions found inside will not generally hold "in the real world." This means that even the entities at play in such assertions cannot be reliably anchored in "reality;" that there is probably not a specific entity in the world to which the *beliefs*/*desires*/*assertions* of the speaker are linked (via the embedded proposition within which the mention intimating such an entity is located). In the case of:

> *John believes that a gorilla stole his lunch.*

We must assume that "any gorilla will do" (or, at least, that "it could be the case that any gorilla will do").

6.3.4 Negation Tests

1. Common nouns with "no" as a determiner are *generic*.

> *I saw no people in the room.*

2. Negated pronouns are *generic*.

> *I saw no one.*
>
> *I saw nobody.*

3. Negated full NPs can be *specific*.

> *Who would do that? Not {Joe}.*
>
> *Neither {Joe}, nor {Mary} said anything.*

4. Common nouns modified by "neither" and partitives with "neither" can be *specific* (depending on coreference) because the negative properties of "neither" have scope over more than just the NP.

> *{Neither person} left the room.*
>
> *{Neither of {them}} likes to talk much.*

6.3.5 Boiler Plate Test

These are NPs that have a legal-like hypothetical setting. We sometimes call them "*empty shell*" mentions.

> *Each year, we elect {one chairman} and {ten board members}.*
>
> *There can be only one {Miss America} for any given year.*

Given an actual instance of the hypothetical setting, these NPs would be "filled in" by actual entities. All these to-be-instantiated NPs should be marked *generic*.

Notice that this test is not exclusively forward-looking. We also see this phenomenon for classes of *previous* (or *iterative*) "empty shell mentions" serving as the generic entity in question. For example:

> *{Former U.S. presidents} have a hard time finding jobs.*
>
> *{The host} rarely steals the show on Saturday Night Live.*

The first example refers to a *generic entity* for which the entire membership is well defined. Any competent historian of the U.S. government can easily provide an exhaustive list of the members of **FORMER_US_PRESIDENT** --- a trick that does nothing to avert the assignment of *generic* status to the entity picked out by the relevant mention. Rather, we are still compelled to assign generic status by the observation that "former U.S. presidents" is used here to refer to any of a set of objects (**FORMER_US_PRESIDENT**), rather than someone in particular (e.g. Jimmy Carter).

The second example is an iterative case that includes as members both the membership of a (well-defined) set (**FORMER_SNL_HOSTS**) and the membership of a (presently undefined/unpopulated) set (**FUTURE_SNL_HOSTS**). Again, we are not torn by the (partial, extensional) definition of the set. We can see right away that the mention "The host" is being used to pick out *any of* a set of entities (without being particular). By our working definition, the mention is therefore **generic**.

It seems that the **Boiler Plate Test** has been poorly defined above (Test IV). We really intend to distinguish between ***the position itself*** and ***the (current) occupant of that position*** --- where the former is **generic** and the latter **specific**.

6.3.6 Verb-Object Compounds and Common Noun Modifiers

The object component of a verb-object compound is always non-referential if nothing separates the constituents. To be distinguished from a verb-object phrase, a verb-object compound must meet at least one of the following conditions:

- One or both of the constituents being bound morphemes (e.g. 革命, 照相)
- Idiomaticity of the meaning of the entire unit (e.g. 伤风)
- Inseparability or limited separability of the constituents (e.g. 革命)

The first question, though, is whether we should tag such NP's at all. Since idiomaticity is a matter of degree, to be on the safer side we annotate them as long as the entity belongs to one of the five categories. So in

他们修{桥}。

桥 is tagged as a mention of generic FAC entity.

The observation may be extended to the verb-object phrase if the object is a bare common noun and nothing else other aspect markers such as "过", "了" separates the constituents. For example,

他们造{铁路}。

他们造过{铁路}。

A common noun directly modifying another noun is not taggable in ACE. The major reason is that such nouns are non-referential and hence not of particular interest to the task.

6.3.6.1 Verb-Object-Complement Construction

We consider the object NP in such constructions as generic:

他们修建了{铁路}三条。

这些城市建立了{友好城市}五十多条。

Note that the complement NP's in the above examples are pronominal mentions because they are headless.

6.4 Summary

Deciding on the generic-hood of NP's can be very difficult for Chinese as the language does not have overt syntactic markings such as number inflections, determiners, and tense that are often used as texts in English and other languages. Furthermore, an NP may have both generic and specific readings and disambiguation will depend on the discourse. For example, the sentence

我喝牛奶

may have a generic reading for "牛奶" if it's uttered in a context where I'm asked what I usually drink for breakfast. But "牛奶" can also have a specific reading if I'm asked to make a choice between a glass of milk and a glass of orange juice on a particular occasion. Equivalent situations would be less ambiguous since in the context, "milk" must be preceded by the definite determiner "the". In general, the discourse has a far more important role to play in determining generic-hood.

Appendix

Appendix A: Chinese Word Segmentation – the ACE approach

In ACE annotation, we DO NOT segment Chinese texts into words with white spaces in between. But the notion of word still plays a role. The basic unit of ACE annotation is

Appendix B:

Co-reference with aliases which refer to more than one entity

Kobe Bryant is the next Michael Jordan.

Bill Clinton will go down in history as the Jon Bon Jovi of US presidents.