TAC KBP 2016 –Entity Discovery and Linking (EDL) Guidelines

Version 1.1

August 16, 2016

Linguistic Data Consortium

Created by: Joe Ellis, Jeremy Getman, Neil Kuster, & Dana Fore

With contributions from: Kira Griffitt, Xuansong Li, Alonso Indacochea, & Justin Mott

http://projects.ldc.upenn.edu/kbp/

Changes from V1.0

- Added reference to “Annotation Guidelines for Individuality of Specific Entities: DEFT Entities, Relations, Events (ERE)” under Individuality section (3.2)
- Revised instructions under 3.7.2 (ORGs) and 3.7.3 (GPEs) to match ERE on nominal top-level government mentions, to tag them as separate, distinct ORG mentions of the governments.

Changes from 2015:

- Section 3.2: Expanded language and examples for determining individual mentions for all entity types.
- Section 3.1: Nominal heads (NOM) to be tagged across all languages and entity types. Changed section name to “Mention Types: Name and Nominal” and created subsections for Named and Nominal mention types. The former absorbs the earlier subsection 3.8.1.1 on post authors (as named mentions), and latter absorbs the earlier subsection 3.8.1.2 on PER nominal mentions.
- Changes to reflect that all five entity types (FAC, GPE, LOC, ORG, PER) will be annotated across all three languages.
- Removed title entity type (TTL) section, retained comments on differences with NOM
- Removed distinctions between “full” ED and simplified ED task for Cold Start, as the latter is not active in 2016
- Removed Section 3.7 Embedded (intra-token) Mentions, not valid for 2016
- Revised Sections 3.2 and 3.3 on individuality and specificity of mentions to reflect consistency with 2016 Rich ERE annotation
- Wording changed to reflect that the 2016 Entity Discovery stage acts as an additional pass over ERE-generated Entity annotation
- Additional NOM examples

Table of Contents

1 Introduction 1

2 General Directives 1

2.1 Accuracy 1

2.2 Tag for Meaning 1

2.3 Online Searching 2

3 Entity Discovery 2

3.1 Mention Types 2

3.1.1 Named Mentions (NAM) 2

3.1.2 Nominal Mentions (NOM) 3

3.2 Individual Entities 5

3.3 Specific, Actual, Non-Fictional Entities 6

3.4 Ignore Mentions between Quotation Tags (Discussion Forums) 6

3.5 Complete Mentions 7

3.6 Nested Mentions 7

3.7 Entity Types 8

3.7.1 Person Entities (PER) 8

3.7.2 Organization Entities (ORG) 9

3.7.3 Geo-political Entities (GPE) 9

3.7.4 Locations (LOC) 10

3.7.5 Facilities (FAC) 10

3.8 Entity Morphs in Chinese Discussion Forums 11

4 Entity Linking 11

1 Introduction

Text Analysis Conference (TAC) is a series of workshops organized by the National Institute of Standards and Technology (NIST). TAC was developed to encourage research in natural language processing (NLP) and related applications by providing a large test collection, common evaluation procedures, and a forum for researchers to share their results. The Knowledge Base Population (KBP) track of TAC aims to advance the state of the art in systems that can determine whether or not specific entities appear in a knowledge base, extract information about those entities from natural text, and update the knowledge base with the extracted information. TAC KBP tests these capabilities of developing systems through multiple tasks, including Entity Discovery and Tri-Lingual Entity Discovery and Linking.

The 2016 TAC KBP Entity Discovery and Linking (**EDL**) evaluation track, systems are challenged to extract all valid entity mentions from a document collection and create cross-document, entity equivalence classes by either linking each mention to a knowledge base or directly clustering them. 2016 EDL will be performed in three languages, Chinese (Mandarin), English, and Spanish. These guidelines will be used to conduct annotators through data creation efforts for EDL.

There are two stages in EDL– Entity Discovery and Entity Linking. In Entity Discovery, annotators find and annotate ***mentions*** for certain kinds of entities that appear in a document. Mentions are occurrences of strings of text (words or other character strings) which refer to entities. For 2016, initial Entity Discovery (including coreference) will be completed during DEFT Rich ERE (Entities, Relations, and Events) annotation passes, with an additional pass done in EDL annotation to confirm or correct the ERE-generated entity mentions and equivalence class clusters. In the second stage, Entity Linking, annotators search through a knowledge base (KB) to determine whether it includes an entry for each entity annotated during Entity Discovery and, if so, link the entity cluster to the KB entry.

2 General Directives

2.1 Accuracy

Annotating entity mentions in the annotation tool requires annotators to be accurate in selecting the exact string of text (which could include letters, numbers, punctuation, Chinese characters, etc.) that represents the mention of an entity. Accuracy also extends to other aspects of annotation, including choosing entity types, grouping entity mentions within the correct cluster, providing translation when required, etc., as detailed below in the following sections. Annotators are encouraged to rely on the context within each document to assist in the annotation process.

2.2 Tag for Meaning

A fundamental rule of thumb which applies to both Entity Discovery and Linking is to tag for meaning: annotators must always ask themselves, “Given the context, who or what is the intended referent of this string of text?” This rule should be the first consideration in annotators' minds when finding, annotating, and linking entities. For instance, the same string of text may refer to two different entities, and annotators will need to use contextual clues to identify which entity is intended:

`[Philadelphia] is a diverse city.`
(Geo-Political Entity – GPE type)

`[Voráček] provided two assists to help [Philadelphia] beat [LA].`
(Both “Philadelphia” and “LA” here refer to the professional hockey organizations – ORG type)

2.3 Online Searching

You may use quick online searching to disambiguate the intended meaning of a string of text in a source document. This may be necessary in determining the proper reference or type for a name which may refer to more than one unique entity, for instance, common Geo-Political Entity (GPE) names. For example, if you are attempting to link a GPE named “Springfield”, a simple search of the KB for “Springfield” will return multiple KB entries for towns with this name, and the source document may not directly state which Springfield is intended. However, if the source document mentions contextual information about the town, such as “… the National Museum of Surveying opened in Springfield in 2007 ...”, you could perform an online search for The National Museum of Surveying, which would reveal that the museum is located in Springfield, IL, and disambiguate the GPE you are attempting to link.

3 Entity Discovery

Entity Discovery – the first stage in EDL – consists of annotating all valid entity mentions that occur in the assigned document. Entity Discovery (including coreference) will have already been completed during DEFT Rich ERE annotation in 2016. Thus, the Entity Discovery pass for 2016 EDL annotation will confirm or correct the entity mentions and coreferencing generated beforehand, adding or deleting mentions or clusters as necessary. ERE-tagged entity mentions will appear in the annotation tool as underlined in the document text and within coreferenced entity clusters, and EDL annotators will begin annotation by performing another pass to verify that all mentions and clusters are correct.

3.1 Mention Types

Mention types are the categories of nouns and nominal phrases which refer to entities. Only named mention types (NAM) and nominal mention types (NOM) are of interest in EDL annotation; we currently do not annotate pronouns/pronominals (e.g., he, they, it, etc.). All named and nominal entity mentions will be annotated for valid entity types (currently FAC, GPE, LOC, ORG, and PER; see section 3.7 below.

Annotators will need to be careful to distinguish between named/NAM mentions and nominal/NOM mentions. Nominals (NOM) are not proper names themselves but common nouns that refer to an entity by replacing its name.

3.1.1 Named Mentions (NAM)

A named (NAM) entity mention is a mention that uniquely refers to an entity by its proper name, acronym, nickname, alias, abbreviation, or other alternate name.

For our purposes, the extent of a named mention is the entire string representing the name, excluding the preceding any definite article (e.g., “the”) or any other pre-posed or post-posed modifiers. These are excluded unless they are part of the entity’s formal name. For example, Bill Clinton’s name is “Bill Clinton”, not “former president Bill Clinton”, while e.g., The Hague or Al Jazeera include articles which are formally part of their names.

It may be difficult to distinguish between named and nominal mentions -- especially for organizations or in cases where common nouns are pre-posed by proper noun modifiers. Some examples:

```
[Vietnamese Navy]NAM [spokesman]NOM
(the) [Vietnamese]NAM [navy]NOM
[Apple]NAM [headquarters]NOM
(an) [Ozark]NAM [forest reserve]NOM
the [University of Pennsylvania]NAM ’s [payroll department]NOM
```

While names are typically capitalized in English and Spanish, capitalization is not always reliable for indicating whether a string is actually a name or nominal, especially in informal data sources such as discussion forum (DF) documents. Annotation must rely on context and/or quick reference to external sources in such cases.

```
the [national security agency]NAM is just one of america’s
     intelligence organizations
China’s principal [national security agency]NOM is called the
     [National Security Bureau]NAM ([NSB]NAM)
```

Named mentions also include post author names found in the metadata of discussion forum threads and web documents (see section 3.1.1.1 for more information about post authors).

3.1.1.1 Post Authors

Discussion forum (DF) documents contain many instances of post authors in xml metadata, which are considered named PER mentions for the purposes of Entity Discovery & Linking. Below are different ways in which post author names occur in the source data, and how they are to be handled.

There are two kinds of metadata headings in which post authors occur in discussion forum documents:

```
<post author="[Ernie S.]" datetime="2011-04-
29T22:48:00" id="p10">
```

The above is an example of an individual post heading, in which there is one annotatable name: [Ernie S.]

```
<quote orig_author="~~Zona~~">
```

However, for cases like the second example, we do not annotate the name “Zona”, because it is considered to be within the boundaries of a DF quoted text region (see Sec 3.4 above).

3.1.2 Nominal Mentions (NOM)

When annotating documents for EDL, you will also be annotating nominal mentions (NOM) of valid entity types (FAC, GPE, LOC, ORG, and PER). A nominal mention consists of a common noun which refers to an entity in place of a name (proper names, aliases, shortened forms, etc., all count as “named mentions” – see Sec. 3.1.1 above).

For our purposes, only the head noun of nominal mentions will be annotated. For example (heads are marked with square brackets):

```
his loudest [critic]
my [brother]
the [informant]
the [president] of Ford
a Google [employee] told me…
the tech [company]
the diverse [city]
the coffee [shop] on the corner
```

It is important not to confuse nominal PER mentions with titles. A title is an official or unofficial name of an employment or membership position that has been held by some person. This includes personal titles and honorifics, official rank or status, and specific employed occupations or professional positions. Other personal honorifics that do not point to organizational positions, such as Mr., Mrs., Ms., Dr., Ph.D., M.D., Esq., etc., should not be tagged as nominals either.

Only a title being used as a nominal reference replacing a specific, real-world person will be tagged as a nominal PER. A title that does not in itself refer to a person by replacing that person's name -- instead usually occurring immediately before or after a name (capitalized in English), and acting as an honorific prefix/suffix or referring only to the position itself -- will not be annotated at all. Thus:

...[President] Carter ...

“President” here is not a taggable entity, because it only functions as a title

...the [president] signed a bill today...

“president” here is tagged as a nominal PER

As with other entity types, annotation of nominal mentions is limited to mentions of singular entities and to those referring to specific, real-world entities. For instance, consider the following sentence:

Some of the duties of a typical Kmart ~~employee~~ are inventory management, helping ~~customers~~, and merchandising.

In the above sentence, ‘employee’ would not be annotated, because it is generic – it does not refer to an actual, specific person in this context. The mention ‘customers’ would not be annotated either, not only because it is generic, but also because it does not refer to any individual entity. See the following sections 3.2 and 3.3 for more details on individual, specific, real-world mentions.

3.2 Individual Entities

Each entity mention annotated for EDL can only refer to a single, individual entity. For some entity types, especially PERs, making this distinction is clear. Consider for example the following sentence:

```
A [shooter] stormed a school outside Los Angeles on
Friday, claiming the lives of multiple ~~victims~~.
```

In the sentence above, you would annotate the singular nominal mention ‘shooter’, but not the mention of multiple ‘victims’. Accordingly, strings of text that conjoin more than one entity token, such as “Ford and Chrysler”, must also be divided and annotated according to each individual mention string:

~~[Miami and Tampa]~~ → [Miami] and [Tampa]

the ~~[senators and representatives]~~ (no singular mentions present)

.

Distinguishing between individual and group entity mentions is not always so transparent. Consider the following:

```
The winning [team] accepted its trophy.
```
(individual ORG (not multiple PER))

```
The massive [complex] is made up of 10 buildings.
```
(individual FAC)

Locations (LOC) can be particularly difficult, especially when presented as nominal and/or syntactically plural mentions. Consider the following valid EDL annotations:

```
the [Rocky Mountains]
the [keys] (the Florida Keys)
the [Great Lakes]
```

Note that degree of structure can sometimes be the deciding factor between a valid ORG entity and an invalid PER group. For instance:

...a ~~military~~ strike by the ~~coalition~~...
(“military” represents a group of ORGs; “coalition” is a group of GPEs)

the ~~delegation~~ arrived yesterday
(”delegation” refers to a PER group)

Distinguishing between singular and plural mentions may be difficult especially in Chinese, where a plural marker is not always present. For all languages, the determination of individuality must be made within the context of the document, using the tag-for-meaning rule, as with all EDL annotation. Context will help determine whether the entity is individual or not in many cases, as you may encounter mistypes or ambiguous cases which require you to rely on usage within context to distinguish whether a mention is singular or plural.

For LOC and FAC entities, there are two decisions that can help in determining individuality – made in this order:

- First, does the entity have a name, or is it likely or possible that it would have a name, in the real world? (This is really about whether the entity could have a proper name in the real world – it does not matter if that name actually appears in the document or not.)
- Second, is the entity a unique, single location or facility in the real world? (You may quickly check online if necessary to find this out.)

Meaning depends on context and usage, so exercise care in your judgment and think critically about what exactly is being referred to in each case. Consider the following text extent:

```
Even if all enemies are vanquished, the progressive
wars can never be won. The [Democrats]ORG will always
find new hostile territory to invade, always creating
a New Frontier.
```

"Democrats" in this case, although ostensibly a plural mention, is most likely being used as shorthand to refer to the Democratic Party (ORG) rather than some subset of Democratic politicians, making it a valid annotation following the "tag for meaning" principle.

For a more extended discussion and further examples for determining the individuality of mentions, please see LDC's "Annotation Guidelines for Individuality of Specific Entities: DEFT Entities, Relations, Events (ERE)" Version 2.6.

3.3 Specific, Actual, Non-Fictional Entities

Fictional or supernatural entities of any type (e.g., "Batman", "Mordor", "The Justice League", etc.) are invalid entities for Entity Discovery. Use caution when applying this rule as some entities known as fictional may have real-life counterparts (e.g., "Utopia" and "Paradise" can refer to real GPEs). Also, figures for which there is historical or archaeological evidence may be annotated; e.g., Mohammed, Jesus (Christ) of Nazareth, the historical Buddha, Confucius, Robin Hood – but not Satan, Maitreya Buddha, Paul Bunyan, etc.

Additionally, mentions that refers to generic, hypothetical, conditional, or negated entities will not be annotated – mentions must refer to specific entities. For instance:

> everyone is a ~~victim~~
> if I'm the next ~~victim~~
> ~~neither~~ of the sisters
> a typical ~~voter~~

3.4 Ignore Mentions between Quotation Tags (Discussion Forums)

NOTE: This rule does *not* refer to text in quotation marks from normally quoted or cited sources, but text set off by special computer coding for quoted posts in discussion forums.

When annotating discussion forum threads, do not annotate any mentions within sections of documents that are tagged with xml as quoted material. In discussion forum threads, these mark off quoted text from previous posts, and are displayed between the xml tags <quote> and </quote>.

3.5 Complete Mentions

The complete mention of an entity must be selected for annotation – mentions that only include part of a complete named-entity string are not adequate for annotation. For example, in the following text excerpt:

> [John Smith, Jr.] lives and works in beautiful
> [Philadelphia].

neither of the words “John” or “Smith” by themselves, nor the string “John Smith” alone, would be accurate mentions. This is because they each constitute substrings of a full mention – “John Smith, Jr.” However, if the text continued with the following sentence:

> [Smith] was born in the city, at which time his parents
> named him “[John]”.

both of the strings “Smith” and “John” should be selected as they appear in the text as separate and complete named mentions.

Sometimes a single mention may be interrupted by another phrase, punctuation, or other text characters:

> [F. Scott Fitzgerald]
> [Macaulay (is this the right spelling?) Culkin]
> [Mischa(sp?) Barton]
> [Dwayne “The Rock” Johnson]

In cases like the above, annotate the entire string as a single uninterrupted mention.

For nominals, keep in mind that only the head is to be tagged – the complete mention does not include any articles or modifying strings. E.g.:

> an additional suspected American [perpetrator] of the attack
> and not
> ~~an additional suspected American perpetrator of the attack~~

3.6 Nested Mentions

If an entity mention contains another valid mention nested within it, these nested entities should also be tagged. Some examples of overlapping mentions:

> [[Kentucky] Fried Chicken]
> [[Kurdistan] Freedom Fighters]
> [[Philadelphia] Eagles]
> [[American] Airlines]

However, we never “double-tag” a single entity mention string (i.e., tag the exact same span of text more than once). For instance, “Chicago” in “Chicago won the Stanley Cup....” (referring to the Chicago Bulls) would be tagged only once. Since we tag for meaning to the best of our ability, “Chicago” in this case would be tagged as a mention of the organization (ORG) known as the Chicago Bulls (and not also as the city/GPE Chicago, Illinois).

Also, names within fuller mention strings that refer to the same entity should not be annotated separately as a nested mention:

[United States of America]

“America” should not be annotated.

[Federal Republic of Nigeria]

“Nigeria” should not be annotated.

NOTE: Unlike in 2015, embedded mentions (intra-token mentions, relevant to English and Spanish only) will not be annotated. Some examples of mentions found embedded within tokens:

~~Obama~~care

[ShellOilNigeria]

(a post author name)

[~~[ShellOil][Nigeria]~~]

3.7 Entity Types

Five entity types are to be annotated, for all languages in 2016 EDL:

- persons (PER)
- organizations (ORG)
- geo-political entities (GPE)
- locations (LOC)
- facilities (FAC)

NOTE: In the following examples under each individual entity type, only strings referring to an entity of that type will be annotated (with brackets or strikethroughs); entities of other types not directly under discussion will be temporarily ignored.

3.7.1 Person Entities (PER)

PER is limited to individual humans. Groups of people (including e.g., families, sisters, married couples, etc.) are not valid person entities for EDL.

[Hillary Clinton] announced her candidacy.

The ~~Clintons~~ held a charity gala

3.7.2 Organization Entities (ORG)

ORGs are corporations, agencies, and other groups of people defined by an established organizational structure. Note that musical groups are considered to be organizations but individual artists (e.g., Britney Spears) are considered persons. Programs or projects should not be considered organizations and different iterations of the same organization (e.g., the Obama and Bush Administrations, or the 111th U.S. Congress and the 112th U.S. Congress) should not be considered as distinct entities. Media publications and productions (newspapers, magazines, TV shows, films, etc.) are not themselves considered organizations, though the entities that produce such works are often organizations.

> `In this week’s edition of` ~~`Time Magazine`~~`, ...`
> (No ORG mention)
>
> `Last night on [ABC] News, I saw...`
> (A parent company can be tagged as ORG, if it appears within a media outlet or publication’s name)
>
> `In an interview yesterday, the President of [Starbucks] said that...`
> (In this extent, “Starbucks” would be tagged as an ORG)
>
> `I met my [friend] at the` ~~`[Starbucks]`~~ `yesterday...`
> (In this extent, “Starbucks” would not be tagged as an ORG, but would be tagged as a facility (FAC) instead)

Nominal (NOM) mentions of the top-level government of a GPE (i.e. at the federal level) will be tagged as ORGs – and tagged, coreferenced, and linked separately from the corresponding GPE mentions.

```
The [French]GPE [government]ORG issued a release on…
A [government]ORG spokesperson today said…
The [government]ORG of [Vanuatu]GPE announced…
```

NOTE: Named (NAM) mentions of GPEs' top-level governments will be considered with the GPE itself; for discussion on top-level governments of GPEs, see the GPE section below.

3.7.3 Geo-political Entities (GPE)

Generally speaking, GPEs are composite entities comprised of a government, a physical location, and a population, with common types including countries, states, provinces, counties, cities, and towns.

Sometimes the context makes it appear that the mention of the geo-political unit (GPE) itself, the capital city, or other government location is referring specifically to the government itself. In these cases, we will still tag the mention of the GPE as a GPE and coref them. Ex.:

```
[Iraq] GPE signed a treaty with [Kuwait]GPE.
[Turkey]GPE regards [Northern Cyprus]GPE as a sovereign country.
[Washington]GPE discussed economic policies with [Moscow]GPE …
```

Note that in the last example above, even though the capital city names (Washington, Moscow) are being used as metonymic references to the governments of their respective nations, they should still be tagged as GPEs, clustered with other mentions of the parent GPEs (US and Russia, respectively), and linked with nodes for those GPEs in the KB.

Regions like “southeast US” are not GPEs because, while they have the physical location and population qualities, they do not have a single, corresponding government. In the text string “southeast Texas”, only [Texas] could be annotated as GPE, as southeast Texas has neither its own government nor a defined location – the entire string “southeast Texas” would be tagged as a Location (LOC) entity (see the next section).

While adjectival mentions of GPEs are tagged as named mentions of GPEs (for instance, [Canadian] from the string [[Canadian] Hockey League]), demonyms are **not** considered named mentions of their respective GPEs. For instance, “Americans” in “The ~~Americans~~ said...” is not considered a valid mention of the United States.

NOTE: The European Union (EU) and United Kingdom (Great Britain) are considered valid, individual GPEs.

3.7.4 Locations (LOC)

Location entities are geographically or astronomically defined places that do not have a political component or natural structures like bodies of water and mountains. Examples of place-related strings that are tagged as LOC include heavenly bodies, continents, non-politically-defined regions, oceans, seas, straits, bays, channels, sounds, rivers, islands, lakes, parks, and mountains.

```
[Cape Hatteras National Seashore] spans over 70 miles.
This [park] is famous for its fishing.

The [Rittenhouse Square] farmer’s market each Saturday

The [Midwest] was pummeled by severe storms. The
[region] now has several flood watches in progress.
```

3.7.5 Facilities (FAC)

A facility is a functional, primarily man-made structure. This includes buildings and similar facilities designed for human habitation, such as houses, factories, stadiums, office buildings, gymnasiums, prisons, museums, and space stations; objects of similar size designed for storage, such as barns, parking garages and airplane hangars; elements of transportation infrastructure, including streets, highways, airports, ports, train stations, bridges, and tunnels. Roughly speaking, facilities are artifacts falling under the domains of architecture and civil engineering.

```
…visited the [White House] last weekend

The [building] is located at [36th] and [Market]
```

3.8 Entity Morphs in Chinese Discussion Forums

The information in traditional formal genres such as newswire is usually explicitly expressed. However, in some certain conditions users need to create new ways to communicate sensitive subjects. For example, certain entity "morphs" widely exist in Chinese Twitter and discussion forums. These morphs are a special case of alias to hide the original objects (e.g. sensitive entities and events) for different purposes, including avoiding censorship, expressing strong sentiment, emotion or sarcasm, and making descriptions more vivid. Here is an example post using morphs: "由于瓜爹的事情，方便面与天线摊牌. (Because of Gua Dad's issue, Instant Noodles faces down with Antenna.)", where

- "瓜爹(Gua Dad)" refers to "薄熙来(Bo Xilai)" because the latter shares one character, "瓜(Gua)" with "薄瓜瓜(Bo Guagua)" who is the son of "薄熙来(Bo Xilai)";
- "方便面(Instant Noodles)" refers to "周永康(Zhou Yongkang)" because the latter shares one character "康 (kang)" with the well-known instant noodles brand "康师傅(Master Kang)";
- "天线(Antenna)" refers to "温家宝(Wen Jiabao)" because the latter shares one character "宝(baby)" with the famous children's television series "天线宝宝(Teletubbies)"

4 Entity Linking

In Entity Linking, the second stage of **EDL**, you will indicate whether or not the entities annotated and clustered together in the Entity Discovery stage are included in the ***Knowledge Base*** (KB). This is done by searching the KB interface in the tool for an entity, determining the proper corresponding **entry**, and then selecting one of the following labels:

- **Linking** them to KB entries in which they are the central topic
- Marking them as **NIL** (i.e., not included in the KB)
- Marking them as **Unknown** (i.e., impossible to determine whether the mention is for any particular entity in the KB)

Try your best to determine the exact equivalent KB entry for any entity, not just the closest one. Note that entities must be the exact **main topic** of an entry in the KB in order to be linked; links cannot be made when an entity is just mentioned within an entry on a different subject. For example, "George Lucas" could not be linked to a KB entry on the *Star Wars* movie franchise just because his name was mentioned within the entry. Likewise, an entry entitled "The Kingdom of England" would not be the correct entry to link to a KB entry on modern England, one should find the KB entry for "England" which refers to the modern state in Great Britain. Correspondingly, be sure the entry refers to the correct, intended entity – e.g., for the nation (GPE) of Australia, do not choose the entry for the continent (LOC) of Australia.

In general, you should not link to a KB entry for an entity which is not equivalent to the entity you're looking for. For example, the KB entry for the geographical location (LOC) "Europe" cannot be linked to mentions of "EU [European Union]" (ORG). If you determine that the actual intended entity to which an entity mention cluster refers does not have an equivalent KB entry, mark the entity as NIL. For instance, if there were no KB entry corresponding to

“Special Counsel to the Mayor of Albuquerque”, we could not link any mentions of this entity to (for instance) “Mayor of Albuquerque” or “Albuquerque”, etc.

If you are unable to determine if an entity has a KB entry or not – typically because it’s not possible to determine to which specific, actual entity the mention actually refers – mark the entity as Unknown, as opposed to NIL. For instance, post author names are considered named entity mentions and are thus annotated as PER entities. However, it is extremely unlikely that a post author would provide enough information about him or herself such that you could determine with certainty that the post author did or did not have a KB entry (while this is theoretically possible, it effectively never happens). Post authors are therefore almost always marked Unknown. Similarly, post authors can make references to entities without providing any disambiguating information about them (e.g. “my friend John”, where “John” would be an annotatable named mention). Cases such as this are also marked Unknown.

Sometimes the author of a document or discussion forum post will supply the reader with inaccurate or misleading information. In these situations, you should link an entity to the correct real-world entity, not some other entity which is potentially indicated incorrectly. For instance, if a document mentioned “Reno, NJ”, and then went on to discuss this city in enough detail that it was clear the author was referring to Reno, NV (where “NJ” was simply a typo), you should link the entity mention [Reno] to the KB entry for Reno, Nevada (and not, alternatively, mark the entity NIL since there is no Reno, New Jersey in the real world). Note that you would also need to annotate the string “NJ” in this mention of “Reno, NJ” as a mention of Nevada and not as a mention of New Jersey!