TAC KBP 2015 –Entity Discovery and Linking (ED&L) Guidelines

Version 1.2

July 10, 2015

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 2014:

- This document was completely overhauled to account for the differences between 2014 English monolingual ED&L and the two TAC KBP evaluations supported by these data in 2015 – Tri-lingual Entity Discovery & Linking and Cold Start – Entity Discovery.
- Section 2.6 – Embedded mentions – added
- Section 2.6 – Data Selection – removed as it is now a separate task
- Namestring Annotation renamed to Entity Discovery
- KB Linking renamed to Entity Linking
- The abbreviation for TITLE changed to TTL for consistency with ERE
- Examples added for all entity types.
- Section 3.1.1: Added Location (LOC), Facility (FAC) and Title (TITLE) entity types.
- Section 3.1.4: Added guidelines for the annotation of nominal PER mentions.
- Section 2.5: Removed restrictions on overlapping mentions; added explicit language disallowing double-tagging
- Section 1.1: Added General Directives section,
- Removed references to “web documents”, as these will not be annotated in 2015

Table of Contents

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.

In two of the 2015 TAC KBP evaluation tracks – Cold Start - Entity Discovery (**ED**) and Tri-Lingual Entity Discovery and Linking (**TED&L**) – 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. These guidelines will be used to conduct annotators through data creation efforts for both by two variations on the same basic annotation task – Entity Discovery & Linking (**ED&L**). Data created via full ED&L will be used to support the TED&L evaluation while data created via simplified ED&L (which includes only a subset of entity types, includes no nominal mentions, and is only conducted on English data) will support the ED evaluation.

There are two stages to both full ED&L and its simplified variation – 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. 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 full and simplified 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 online searching to disambiguate the intended meaning of a string of text in a source document. This is helpful 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 both simplified and full ED&L – consists of annotating all valid entity mentions that occur in the assigned document.

3.1 Named Mentions

All named entity mentions for valid entity types will be annotated in both full and simplified ED&L (see section 3.8 for more details about valid entity types for the two task variations). A named entity mention is a mention that uniquely refers to an entity by its proper name, acronym, nickname, alias, abbreviation, or other alternate name. Be careful to distinguish “named” mentions from common nouns or noun phrases that refer to an entity but which are not actually names (these are called ***nominal*** mentions, nominal mentions are also annotated for person entities in English documents for full ED&L; see section 3.8.1.2 for details). Note that it can sometimes be difficult to distinguish between a named and nominal mention, especially for organizations. Note as well that named mentions also include post author names found in the metadata of discussion forum threads and web documents (see section 3.8.1.1 for more information about post authors).

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”) and any other pre-posed or post-posed modifiers. These are excluded because they are not part of the entity’s actual name. For example, Bill Clinton’s name is “Bill Clinton”, not “former president Bill Clinton”.

3.2 Individual Entities

Entity mentions annotated as ED&L queries should refer only to a single entity. Strings of text that refer to more than one entity (e.g., [Ford and Chrysler], [Miami and Tampa]) are invalid.

While following this rule is seemingly simple to follow, it can be difficult to adhere to when dealing with nominal mentions (which currently only apply to English documents in full ED&L).

Consider the following text extent:

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

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

3.3 Actual, Non-Fictional Entities

Fictional or supernatural characters 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).

Additionally, mentions must refer to specific actual entities. For instance, "victim" from the generalized phrase "everyone is a victim" would ***not*** be annotated.

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, from the following text excerpt:

```
[John Smith] lives and works in beautiful
[Philadelphia].
```

we would not select either of the words "John" or "Smith" by themselves as accurate mentions. This is because they each constitute substrings of a full mention – "John Smith". 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.

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.

3.7 Embedded Mentions

If a single token within an entity mention includes another mention of a different, distinct entity, such an 'embedded' entity should also be annotated. Some examples of embedded mentions:

```
[Obama]care

[[ShellOil][Nigeria]]
```

In the latter example, there are 3 entities to annotate: "ShellOilNigeria" is being used as a post-author name, while "ShellOil" and "Nigeria" should also be tagged as distinct entity mentions.

3.8 Entity Types

For simplified ED&L, only three entity types are to be annotated:

- persons (PER)
- organizations (ORG)
- geo-political entities (GPE)

For full ED&L, there are two or three additional valid entity types, depending on language:

- locations (LOC)
- facilities (FAC)
- titles (TTL) – valid for **English documents only**

NOTE: your annotation tool should restrict your choices of entity types appropriate to the version of the task you are completing. However, if you are ever uncertain about which types you should annotate, do not hesitate to check with your supervisor.

3.8.1 Person Entities (PER)

PER is limited to individual humans. Groups of people (including families) are not valid person entities.

```
[Hillary Clinton] announced her candidacy.
The ~~Clintons~~ held a charity gala
```

3.8.1.1 Post Authors

Discussion forum documents contain many instances of post authors in xml metadata, which are considered names for the purposes of Tri-Lingual 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 quoted text region (see Sec 3.4 above).

NOTE: Section 3.8.1.2 is for English documents in full ED&L only

3.8.1.2 Nominal PER Annotation

When annotating English documents for full ED&L, you will also be annotating **nominal** mentions of PER entities. A nominal mention uses a common noun or noun phrase that refers to an entity in place of a **name** (proper names, aliases, shortened forms, etc., all count as “named mentions” – see Sec. 3.1above).

As with other entity types, annotation of nominal PER entities is limited to mentions of singular persons only. For instance, consider the following sentence:

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

From the above, you would annotate the singular nominal mention ‘shooter’ but not the plural mention ‘victims’.

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

> his loudest [critic]
> my [brother]
> the [informant]
> a Google [employee]
> the [president] of Ford

Only PER mentions referring to specific, real-world entities will be annotated. For instance, consider the below sentence:

> Some of the duties of a typical Kmart ~~employee~~ are inventory management and merchandising.

From the above sentence, ‘employee’ would **not** be annotated, because it is generic – it does not refer to an actual, specific person in this context.

It is important not to confuse nominal PER mentions with titles (TTL) (see section 3.8.6 for more details on TTLs). A mention of a title that refers to the position itself will be tagged as a title, whereas a title used as a reference to a specific person (such as in the second example) will be tagged as a nominal PER:

> ...[President] Clinton ...
> “President” here is a title TTL

> ...the [president] signed a bill today...
> “president” here is a nominal PER

3.8.2 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 for simplified ED&L but would be tagged as a facility (FAC) entity for full ED&L

3.8.3 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. Note, however, that for the purposes of Entity Discovery annotation (and all TAC KBP tasks), all top-level governments of GPEs should also be categorized as GPEs, not as ORGs.

Regions like "the southeast US" are not GPEs because, while they have the physical location and population qualities, they do not have their own 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 (in full ED&L "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, in "The ~~Americans~~ said..." is not considered a valid mention of the United States.

NOTE: The following three entity types are only annotated for full ED&L.

3.8.4 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.

The [Rittenhouse Square] farmer's market each Saturday

The [Midwest] was pummeled by severe storms.

3.8.5 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]

NOTE: Section 3.8.6 is for English documents in full ED&L only

3.8.6 Titles (TTL)

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. Honorifics that do not point to organizational positions, such as Mr., Mrs., Ms., Dr., Ph.D., M.D. should be excluded.

[Prof.] Mahmoud ElSawey

~~Dr.~~ Breanne Johnston, ~~M.D.~~

It is important not to confuse nominal PER mentions with titles. For instance, in the phrase “President Clinton”, “President” would indeed be tagged as a title. However, in the phrase “the president signed a bill today”, “president” would be tagged as a nominal PER. A mention of a title that refers to the position itself will be tagged as a title, whereas a title being used as a reference to a specific, real-world person (such as in the second example) will be tagged as a nominal PER.

3.9 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 full and simplified **ED&L**, 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 for the 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)

Note that entities must be the **central 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.

If you determine the actual intended entity to which an entity mention cluster refers, but are unable to determine if an entity has a KB entry or not, mark the entity as NIL.

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.