TAC KBP 2014 – Entity Discovery and Linking Query Development Guidelines

Version 1.5

Linguistic Data Consortium

Created by: Joe Ellis & Jeremy Getman

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

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

Changes from Version 1.4:

- Added brief discussion to section 3.2 regarding misleading or inaccurate document context.

Changes from Version 1.3:

- Added section 3.1.5 explaining how to annotate post authors in discussion forum and web documents.
- Added language to section 3.1.1 noting that demonyms are not considered named mentions of GPEs.

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 Linking.

In Entity Discovery and Linking (EDL), systems are challenged to extract all entity mentions in a document collection and to determine whether or not those entities are included in a Knowledge Base, currently a 2008 snapshot of Wikipedia. As such, the first step in evaluating EDL system performance is for annotators to perform full document entity annotation.

There are three stages to complete in the Entity Discovery and Linking task – Namestring Annotation, KB Linking, and NIL Coreference. During Namestring Annotation, you will search for and select named mentions of entities in text and then proceed to annotate all named mentions of entities that occur in that same text. In KB Linking, you will search the 2008 snapshot of Wikipedia and indicate whether or not it includes pages on the entities you annotated during Namestring Annotation. Lastly, for NIL Coreference, you will group together annotated entities that were not linked to the KB (i.e. NIL entities) into equivalence classes.

2 Entity Requirements for EDL

All entities annotated must fit the following requirements:

2.1 Named Mentions

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. Note that this includes post author names found in the metadata of discussion forum threads and web documents.

For our purposes, the extent of a name is the entire string representing the name, excluding the preceding definite article (i.e. “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”.

2.2 Individual Entities

A **namestring** is a piece of text selected from a source document that refers to an entity by name. Namestrings annotated as EDL 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 inappropriate.

2.3 Non-Fictional Entities

Fictional characters of any type (e.g., “Batman”, “Mordor”, “The Justice League”, etc.) are invalid entities for EDL. Use caution when applying this rule as some entities known for being fictional have real-life counterparts (e.g., “Utopia” and “Paradise” can refer to real GPEs).

2.4 Complete Mentions

Namestrings that are substrings of a complete named-entity mention are not allowed for any queries. For example, from the following text excerpt:

John Smith lives and works in beautiful Philadelphia.

you could not select either of the words “John” or “Smith” by themselves as Entity Linking queries. This is because they constitute a substring of the 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 are provided in the text as complete named mentions.

2.5 Overlapping Mentions

If an entity mention contains another taggable mention nested within it, these nested entities should also be tagged, assuming the overlapping mentions obey the following two rules:

1. One of the two mentions is completely nested within the other mention.
2. At least some part of the larger mention does not appear in any of its embedded/nested mention(s). In other words, the combined full extent of the nested mention(s) is not identical to the full extent of the larger mention.

Some examples of overlapping mentions which follow the rules above:

- [[Kentucky] Fried Chicken]
- [[Kurdistan] Freedom Fighters]
- [[Philadelphia] Eagles]

2.6 Appropriate Web Source Documents

Special care must be taken when annotating namestrings from web documents. Do not annotate namestrings from web documents that contain:

- Spam, pornography, or other offensive material

- Newswire. If one of the posts in a web or discussion forum document appears to contain part or all of a newswire article copied from an outside source, do not annotate the document.

3 Stages of Entity Discovery and Linking

3.1 Namestring Annotation

The first stage is Namestring Annotation. Since Entity Discovery and Linking challenges systems to identify entities in text and link them to the KB, a high level of confusability is the defining characteristic of a good EDL query.

3.1.1 Entity Types

All entities annotated in EDL must be one of three types: persons, organizations, or geo-

political entities.

- **Person Entities (PER)** – PERs are limited to individual humans. Groups of people (including families) are not valid person entities.

- **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 111th U.S. Congress and the 112th U.S. Congress) should not be considered as distinct entities.

- **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 EDL (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. Given the text “southeast Texas”, only “Texas” could be annotated as GPE, as southeast Texas has neither its own government nor a defined location.

  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” is not a mention of the United States.

3.1.2 Confusable Namestrings

A very desirable quality of EDL queries is that they have confusable namestrings, meaning that the namestring alone should not clearly indicate the entity to which it refers. For example, consider two different namestrings, “John” and “Barack Hussein Obama”. “John” is confusable because it is impossible to determine who or what the name refers to without additional information. “Barack Hussein Obama” is not confusable because one could probably guess correctly that the string refers to the US president elected in 2008. Note, however, that if you found a namestring "Barack Hussein Obama" that referred to someone other than the president, it would make an excellent EDL query.

Some other examples of confusable EDL namestrings include “John Smith”, “Democratic coalition”, “London”, and “Foreign Language Department”. Acronyms, such as “AA”, “NSP”, and “FBI” can also make strong EDL queries for ORG entities. Confusability across entity types is also an excellent quality for EDL queries (e.g., “George Washington” could refer to the president, the university, or the jazz musician). When selecting confusable EDL namestrings, keep in mind that all TAC KBP queries must be complete mentions (see section 2.3).

3.1.3 Measurable Standards

Since any namestring could feasibly refer to more than one entity, it is insufficient to base EDL query selection solely on an abstract notion of confusability. Consider the Obama example

above; while it is unlikely that the namestring in print refers to multiple individuals, such a possibility logically exists. For this reason, you should investigate whether potential EDL queries also meet two measurable standards of confusability: multi-entity namestrings and multi-namestring entities.

- **Multi-entity namestrings** are unique namestrings that refer to more than one entity *in the source corpus*. For example, if you have found and annotated the namestring "John" you should look for more occurrences of that exact same string in the corpus to ensure that a namestring refers to multiple entities.

- **Multi-namestring entities** are individual entities referred to by multiple confusable namestrings *in the source corpus*. Such entities result from misspellings (e.g., "Conneticut" and "Bagdad"), orthographic variations (e.g., "al-Assad" and "al-Asad" referring to the same person), aliases or nicknames (e.g., "Carlos the Jackal" and "Sanchez" referring to Ilich Ramírez Sánchez), or even differences in capitalization ("John" and "john"). Well-known nicknames (e.g. "Old Blue Eyes" or "The Biebs") are allowed.

3.1.4 Whole Document Annotation

After you have found a document containing a few good EDL queries based on the standards described above, you will then proceed to annotate all named entity mentions that occur in the document, with some exceptions.

First and foremost, all entities annotated must meet the requirements laid out in Section 2. Any entities that do not meet all the requirements in Section 2 are to be left unannotated.

Additionally, when annotating discussion forum threads or web documents, sections of documents that are tagged as quoted material are to be ignored. That is, no entity mentions within quotes are to be annotated. In discussion forum threads quotes are the text occurring between <quote> </quote> tags. In web documents, quotes are the text occurring between the quotation marks in <QUOTE PREVIOUSPOST= " ">

3.1.5 Post Authors

Discussion forum and web documents contain many instances of post authors ("posters" in web docs) in xml metadata, which are considered names 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 quoted text region.

Web documents contain a variety of formats for presenting poster names:

```
<POSTER> marcuswrobe...@hughes.net </POSTER>
```

The above contains one named mention: *marcuswrobe.* In cases like this, where the poster name given is an email address, we only annotate the portion prior to the '@' symbol as the extent of the name.

```
<POSTER> sf </POSTER>
```

There is one named mention in the above example: *sf.*

```
<POSTER> "poonam" <poonam8...@gmail.com> </POSTER>
```

Some web docs contain both a username and an email address in the poster tag. For cases like this, we annotate both the username and the first part of the email address. For instance, in the above there are two named mentions: *poonam* and *poonam8.*

3.2 KB Linking

In KB Linking, the second stage of EDL query development, you must indicate whether or not annotated EDL entities are included in the KB by either linking them to pages in which they are the central topic or marking them as NILs (i.e., not included in the KB). Note that entities must be the central topic of a page in the KB in order to be linked; they cannot just be mentioned in a page on a different subject. For example, "George Lucas" could not be linked to a KB page on the *Star Wars* movie franchise just because his name was mentioned in the text.

If you are unable to determine if an entity has a KB entry or not, mark the entity as NIL/Unknown, as opposed to NIL. For instance, post author names are considered named entity mentions and are thus annotated as PER entities in EDL. However, it is extremely uncommon for a post author to provide enough information about him or herself such that an EDL annotator could determine with certainty that the post author did or did not have a KB entry. Post authors are therefore almost always marked NIL/Unknown. Similarly, post authors often make references to entities without providing any disambiguating information about the entities (e.g. "my friend John", where "John" would be an annotatable named mention). Cases such as this are also marked NIL/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 (and "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).

3.3 Online Searching

You may use online searching to disambiguate the namestring in the source document. This

is helpful for disambiguating common GPE names. For example, if you are attempting to link "Springfield", which is labeled as a GPE, a simple search of the KB for "Springfield" will return KB entries for many towns with the name, and the source document may not directly state which Springfield is being referred to. 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.

3.4 NIL Coreference

The overall goal of the NIL coreference stage is to group together selected namestrings into equivalence classes when they refer to the same entity. For the purpose of this task, to **coreference** means to indicate that one or more mentions refer to the same thing by grouping them together as an entity. Note that a namestring does not necessarily have to be grouped with others to be coreferenced. If there is only one mention of a particular entity, it should be coreferenced by itself.

NIL coreference is performed on entities marked NIL/Unknown in addition to those marked NIL.