TAC KBP Entity Selection

Version 1.1 – May 21st, 2012

Linguistic Data Consortium

Created by: Joe Ellis (joellis@ldc.upenn.edu)

With contributions from: Kira Griffitt, Xuansong Li, Jeremy Getman, & Alonso Indacochea

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

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 primarily through two tasks, Entity Linking and Slot Filling.

In Entity Linking (EL) systems are challenged to determine whether or not specific entities mentioned in newswire or web documents are included in a 2008 snapshot of Wikipedia, which has served as the official knowledge base (KB) of TACK KBP since 2009. In Slot Filling (SF), however, systems must find and collect new information about entities in a large set of documents and add it to the KB. As entities are the basis of both EL and SF, the first step in evaluating system performance in the tasks is to select entities for use as queries.

There are three stages to complete in the Entity Selection annotation task – Namestring Annotation, KB Linking, and NIL Coreference. During Namestring Annotation, you will search for and select named mentions of entities in text for use in either EL or SF. In KB Linking, you will search the 2008 snapshot of Wikipedia and indicate whether or not it includes pages on the entities you selected during Namestring Annotation. Lastly, for NIL Coreference, you will group together selected entities that were not included in the KB (i.e., NIL entities) into equivalence classes.

A few terms related to Entity Selection tend to overlap in meaning and so some clarification is necessary. A **namestring** is simply a piece of text selected from a source document that refers to an entity by name. An **entity** is a specific person, organization, or geo-political entity. A **query** is the combination of a namestring and a source document that is used to test the ability of systems in either EL or SF.

2 Entity Requirements for EL and SF

All entities selected for use in both the Entity Linking and Slot Filling tasks must fit the following requirements:

2.1 Individual Entities

Selected namestrings 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 as queries for any TAC KBP task.

2.2 Non-Fictional Entities

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

2.3 Complete Mentions

Namestrings that are substrings of a complete named-entity mention are not allowed for any queries. For example, given the complete namestring “John Smith” in a document, you cannot select either the words “John” or “Smith” by themselves as they constitute a substring of a full mention. Additionally, you should not select “Springfield” from the sentence “She grew up in Springfield, OH”. These words could only be selected if they appeared separately in the document.

Note, however, that nested namestrings are allowable. For example, given the string “Governor of Arizona” the GPE namestring “Arizona” could be selected as its referent is something different than what is referred to by the full named mention.

2.4 Appropriate Web Source Documents

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

- Spam, pornography, or other offensive material
- Copyrighted material. Note that copyrighted material can often be identified by a “free reprint” statement such as:

  ****
  Please consider this free-reprint article written by:
  Sean Horton
  ****

2.5 Unique Namestring/Source Document Pairs

Every TAC KBP query should be a unique pairing of a namestring and source document. For example, if you select the namestring “John” from a document, you cannot select another “John” from the same source document, even if each namestring refers to a different entity.

3 Stages of Entity Selection

The first stage of Entity Selection – Namestring Annotation – involves different procedures depending on whether you are selecting entities for EL or SF as each task has different requirements. As a result, Namestring Annotation is discussed separately for EL and SF below. The processes used in KB linking and NIL coreference – the second and third Entity Selection stages – are the same for both EL and SF queries.

3.1 EL Namestring Annotation

Since Entity Linking challenges systems to identify entities in text, a high level of confusability is the defining characteristic of a good EL query. The task also allows for one additional type of entity.

Entity Types

All entities selected for use in all of the Entity Linking tasks 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. Brittany 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)** – GPEs are composite entities comprised of a government, a physical location, and a population. Common GPE types include countries, states, provinces, counties, cities, and towns. 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.

Confusable Namestrings

All EL queries should 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 EL query.

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

Measurable Standards

Since any namestring could feasibly refer to more than one entity, selecting EL queries by a language standard alone is insufficient. 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 EL 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 a “John” you should look for more occurrences of that exact same namestring in the corpus and annotate those that refer to different entities. Note, however, that confusable namestrings referring to only to a single entity are acceptable and need not be deleted.

- **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), and aliases or nicknames (e.g., “Carlos the Jackal" and “Sanchez” referring to Ilich Ramírez Sánchez). Be careful, however, not to select famous celebrity nicknames (e.g. "Old Blue Eyes" or “The Boss”) as such namestrings would not be confusable.

3.2 SF Namestring Annotation

In a sense, Slot Filling could be considered a second step after Entity Linking because an entity must first be linked to the KB (or identified as not having a KB page) before new information on it can be collected. Unlike EL queries, however, namestrings selected for use in the Slot Filling task should not be confusable and they should refer to entities that are richly documented in the corpus but not likely to be heavily documented in Wikipedia.

Entity Types

All entities selected for use in all of the Slot Filling tasks must be persons (PERs) or organizations (ORGs) meeting the standards described for each type above (see section 3.1). GPEs are not valid entities for Slot Filling.

Full Name Strings

All SF entities must be referred to by namestrings that are full enough to make the referent relatively unambiguous. Some examples of full namestrings are “Ilich Ramírez Sánchez”, “John Fahey” (unambiguous names containing shortened forms of names are acceptable), or “Philadelphia Classical Guitar Society”.

Not Famous

Since the primary focus of SF is collecting new information about entities and adding it to the KB, you generally want to avoid selecting famous entities for SF queries because they tend to have well-developed KB pages. As a result, when you have an entity in mind as a potential SF query, you should look up the person or organization in the KB to ensure that they do not already have a full entry.

Note that, because of some qualities of the KB, you should not apply the ‘not famous’ rule across the board for SF queries. Primarily, the KB was selected in 2008 and so entities that have only become well-known more recently might not be included in the snapshot. Additionally, some pages were removed from the KB as a result of formatting issues.

Measurable Standards

Since Slot Filling systems attempt to collect information about entities pertaining to the TAC KBP slots, the quality of SF queries is measured by the number and type of slots for which it can provide fillers (see *TAC KBP Slots* for a list and descriptions of all the types of information collected about entities in the Slot Filling task).

- **Multi-filler entities** are persons or organizations on which a great deal of information can be found in the source corpus pertaining to the TAC KBP slots. However, remember when looking for multi-filler entities that they cannot already have well-developed KB pages, which can pose a challenge.

- **Uncommon-filler entities** are persons or organizations on which uncommon information can be found in the source corpus pertaining to the TAC KBP slots. Some slots tend to appear less frequently in text (e.g., *PER: Cause of Death*, *PER: Religion*, *PER: Charges*, *ORG: Political/Religious Affiliations*, *ORG: Number of Employees/Members*, *ORG: Dissolved, ORG: Website*) and so entities that can provide valid fillers for these slots make challenging SF queries for systems.

Source Documents

Once you have ensured that a potential SF entity does not have a full KB page and can offer at least 2 valid slot fillers from the source corpus, look for a document that disambiguates your entity but, ideally, provides no data about it pertaining to the KBP slots. Once you have found such a document, annotate from it the fullest namestring referring to your entity.

Ideally, reference documents will contain no slot fillers for your entity. However, as such documents can be difficult to find, reference documents containing a few slot fillers for your assigned entities are permissible.

3.3 KB Linking

In KB Linking, the second stage of Entity Selection, you must indicate whether or not selected EL and SF 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). KB linking is the exact same process regardless of whether entities were selected for EL or SF. 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.

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 view 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.

4 The Entity Selection Tool

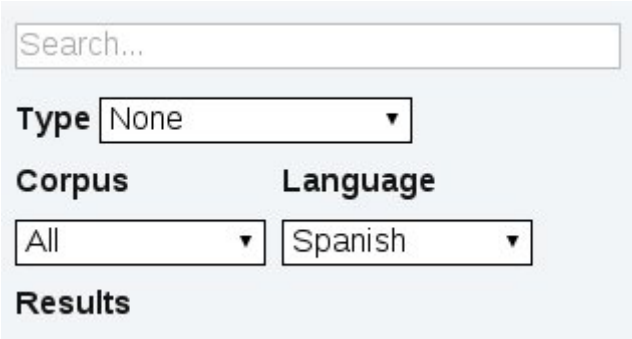
The following is a step-by-step walkthrough of the web-based Entity Selection tool, which can be accessed by logging in at https://webann.ldc.upenn.edu/ using your webann username and password.

4.1 Namestring Annotation

1. **Select Task** – It is **very** important that you begin Namestring Annotation by first selecting the correct tabs at the top of the right panel both for the language (Chinese, English, or Spanish) and the task (EL or SF) on which you are working:

   You should always be aware of which tab is selected (EL Entities or SF Entities) when making annotations. For example, If you are assigned to search for EL entities and happen to come across an entity that appears to instead be suitable for SF, you may annotate the entity for the Slot Filling task (despite being assigned to Entity Linking). To do so, however, you must switch to the 'SF Entities' tab, annotate the namestring, and then switch back to the 'EL Entities' tab to continue searching for EL entities.

2. **Search** – Following correct task selection, you can begin searching for entities that meet the requirements of your assigned task (either Entity Linking or Slot Filling). The search functions can be found at the top left of the Entity Selection view:



There are two methods by which to search:

To search by tagger output, select a corpus (Newswire or Web), language (Chinese, English or Spanish), and entity type (Person, Organization or GPE). After making these selections, place the cursor in the empty Search box and hit enter. This will produce a list of all documents in the selected corpus/corpora that meet the selected language and type criteria. Selecting a type is not technically required to produce results, but making this selection will highlight different strings in the displayed documents. For example, if Person is selected, all entity mentions in the corpus that have been tagged as Person will be highlighted, making the search for new entities much easier.

To search by namestring/keyword, a language must be selected but type and corpus can be left unspecified. Place the cursor in the search box, type the namestring that you want to search for, and hit enter. The search will result in a list of documents that contain the namestring that was searched for.

3. **Annotation** – Begin by clicking on a Doc-ID in the list of search results in the Namestring Annotation view. This will display a document in the center pane. After finding an entity that meets the criteria of your assigned task (Entity Linking or Slot Filling), use the mouse to highlight the mention of the entity you want to annotate, either by double-clicking or clicking and dragging, and then hit enter. This should create a new annotation in the right panel, with the annotated text filling the namestring bar at the top of the annotation, regardless of whether you are working under the Entity Linking tab or the Slot Filling tab.

   If you are searching for EL entities, select the annotated entity's type from the drop-down menu, and click the 'Finished' checkbox (if you later decide you need to make a change to the annotation, you can unclick 'Finished' and do so). You can then proceed to search for other entities that share the same (confusable) namestring as the first entity you annotated, making a new annotation for each confusable namestring you find (as long as it does not occur in the same document as another of these confusable entities).

   If you are searching for SF entities, select the annotated entity's type from the Type drop-down menu and then search for valid slot/fillers pairs involving the entity in corpus, following the guidelines in *TAC KBP Slots*. Once you've found a valid slot/filler

pair, select the slot from the ‘Slot’ drop-down menu that corresponds to the filler. To annotate the filler, make sure the field next to ‘Filler’ is selected, highlight the corresponding text string in the document, and hit the Enter key. Once you've found **at least** two different fillers for the entity, check the ‘Finished’ box. Additionally, remember that the entity's namestring should be annotated from a document that has as few fillers for the entity as possible (preferably zero).

4.2 KB Linking

1. **Select an Entity** – At the top left of the Entity Linking view, the type tabs (PER, ORG, GPE) will automatically sort the entities that were selected in the Namestring Annotation stage. Entities can be sorted alphanumerically by ID, Namestring, or Node columns directly below the type tabs.

   The information in the ID column will be carried over from the auto-generated namestring ID in the Entity Selection view. The namestring column contains the exact namestring that was selected from Namestring annotation, and the Node column lists whether or not a KB entry was found for each namestring (either KB-#, NIL, or Bad Entity) in this stage. Selecting an entity/namestring from the list in the left panel will display the entity's reference document in the bottom center pane.

2. **Search the Knowledge Base for the Entity** – The search box at the top of the right panel of the Entity Linking view allows you to search through the Knowledge Base entries for the namestring. Simply type a namestring in the search box and hit enter. The tool will only search for the titles of Knowledge Base entries, not content within an entry. However, if you type “body:” and then a search term or phrase, the tool will search for your string within the body of KB articles. Selecting a result from your search of the KB will display the KB entry in the top center pane.

3. **Determine the Entity's Status** – If the KB contains an entry for the entity, the entity and the KB entry should be linked. To link the two, click on the blue 'Link' button that separates the top and bottom center panes. When you click on the 'Link' button, the Knowledge Base's node number will appear in the 'Node' column on the left side of the tool (for instance, KB-52). If, after an exhaustive search, no KB entry is found for the entity, click the 'NIL' button, and 'NIL' appears in the 'Node' column. If the entity itself is determined to be bad (copyrighted material in the reference document; document contains offensive material; typed incorrectly (i.e. as a PER, ORG, or GPE); doesn't occur as a name by itself in the reference document; appears in the document as a part of a movie/book/TV show title; or, refers to more than one entity in the article), click the 'Bad' button to flag the entity/namestring.

4.3 NIL Coreference

1. **Selecting Entities** – In the NIL Coref view, the type tabs (PER, ORG, and GPE) above the 'UNDONE' column header will automatically sort the entities that were chosen in Namestring Annotation and marked as NIL in KB Linking. Clicking on a namestring brings the entity's reference document into the right pane. Clicking on an ID allows you to select the namestring you want to coreference. Only one reference document can be viewed at a time, but multiple IDs can be selected at once. A selected ID will be highlighted in red.

2. **Coreference** – After clicking the IDs of all namestrings you want to coreference, hit enter. This will move these entities into the center column (the 'DONE' column) under one header. In addition to simultaneously coreferencing multiple namestrings, you can also drag and drop a namestring from the 'UNDONE' column to an entity in the 'DONE' column to associate it with the already coreferenced entity.

   All of the namestrings from the 'UNDONE' column that refer to the same entity must be coreferenced together in the 'DONE' column. Once all of the namestrings for one entity are coreferenced together, repeat the above process for all other distinct entities. The coreference task is completed when there are no remaining namestrings in the 'UNDONE' column, each entity in the 'DONE' column is associated with namestrings that refer only to that specific entity, and no two groupings of namestrings refer to the same entity.