Task Description for Knowledge-Base Population at TAC 2009

Version of: 6/1/09

1. Overview

The main goal of the KBP track at TAC 2009 is to promote research in and to evaluate the ability of automated systems to discover information about named entities and to incorporate this information in a knowledge source. For the evaluation an initial (or reference) knowledge base will be provided along with a document collection that systems are to use to learn from. Attributes (a.k.a., "slots") derived from Wikipedia infoboxes will be used to create the reference knowledge base. There will be two related tasks: Entity Linking, where names must be aligned to entities in the KB, and Slot Filling, which involves mining information about entities from text. Slot Filling can be viewed as more traditional Information Extraction, or alternatively, as a Question Answering (QA) task, where the questions are static but the targets change. Groups may participate in either, or both tasks, though participation in both is encouraged.

The tasks will be structured by having participants process a list of target entities. The list will contain entity types of Person (PER), Organization (ORG), and Geo-Political Entity (GPE). As in the ACE evaluation, GPEs include inhabited locations with a government such as cities and countries.

2. Knowledge Base

Wikipedia infoboxes will be the basis for the reference knowledge base; however exact compliance with Wikipedia is not intended. The KB will be derived from the set of entities from Wikipedia that have infoboxes, and it will contain hundreds of thousands of nodes. Each entity in the knowledge base (sometimes called a node) will include the following:

- a name string
- an assigned entity type of PER, ORG, GPE, or UKN (unknown)
- a KB node ID (a unique identifier, like "E101")
- a set of 'raw' (Wikipedia) slot names and values
- some disambiguating text (i.e., text from the Wikipedia page)

For each of the three generic entity types there is a set of desired attributes ("slots"). Guidelines for each slot are available at: http://projects.ldc.upenn.edu/kbp/. The guidelines specify whether the slots are single-valued (e.g., *per:date_of_birth*) or list-valued (e.g., per*:employee_of*, per*:children*) and whether they should be linked to nodes in the KB. Official names for each slot are given in Table 1. [1]

[1] The earlier guidelines and the mapping files in the sample corpus mentioned a few slots that are not given in Table 1; those slots should be ignored.

Table 1. Slot names for the three generic entity types.

Person	**Organization**	**Geo-Political Entity**
per:alternate_names	org:alternate_names	gpe:alternate_names
per:date_of_birth	org:political/religious_affiliation	gpe:capital
per:age	org:top_members/employees	gpe:subsidiary_orgs
per:place_of_birth	org:number_of_employees/members	gpe:top_employees
per:origin	org:members	gpe:political_parties
per:date_of_death	org:member_of	gpe:established
per:place_of_death	org:subsidiaries	gpe:population
per:cause_of_death	org:parents	gpe:currency
per:residences	org:founded_by	
per:schools_attended	org:founded	
per:title	org:dissolved	
per:member_of	org:headquarters	
per:employee_of	org:shareholders	
per:religion	org:website	
per:spouse		
per:children		
per:parents		
per:siblings		
per:other_family		
per:charges		

The 'raw' slot names and the values in the reference KB are based on an October 2008 Wikipedia snapshot. Wikipedia infoboxes are not an ideal knowledge representation; one key disadvantage is a lack of inheritance (and therefore, consistency). Wikipedia infoboxes also tend to focus on presentation on a Wikipedia page instead of abstract representation. As an example, consider the table below. The infobox for each of these organizations contains a slot related to the date it was created; however, the name of the slot varies.

Table 2. Examples of Wikipedia infobox slot name and slot value variability.

Organization	**Infobox Class**	**Slot Name**	**Slot Value**
Bill/Melinda Gates Foundation	Non-Profit	founded_date	1994
Cornell	University	established	1865
FDA	Government_Agency	formed	1906
NASA	Government_Agency	formed	July 29, 1958
Washington Redskins	NFL_Team	founded	1932

To facilitate use of the reference KB a mapping from raw Wikipedia infobox slot-names to generic slots will be provided. For example, Wikipedia infobox slots named *established* or *creation_date* may contain a value for *org:founded*. These mappings should not be viewed as precise synonyms. Sometimes Wikipedia slots contain multiple pieces of information that belong in separate generic slots. For example, some Wikipedia infoboxes contain a slot named '*born'* that has values for both *per:date_of_birth* and

per:place_of_birth. Also, the 'semantics' of a Wikipedia infobox field will not always correspond with the generic slot guidelines. For example, the children slot in Wikipedia's Person infobox sometimes gives a number to indicate that the entity has that number of children; however, only names of children are acceptable for the generic slot *per:children*.

3. Entity Linking

The Entity Linking task is to determine for each query, which knowledge base entity is being referred to, or if the entity is not present in the reference KB. A query will consist of a name-string and a document-id. Each name-string will occur in the associated document in the test collection. The purpose of the associated document is to provide context that might be useful for disambiguating the name-string. Entities will generally occur in multiple queries using different name variants and/or different docids. It is also expected that some entities will share confusable names (e.g., *George Washington* could refer to the president, the university, or the jazz musician; *Washington* could refer to a city, state, or person).

Queries should be processed independently from one another. This is not meant to prohibit parallel computation on clusters, but only to require that systems may not leverage knowledge about the set of queries. Examination of multiple queries to make a better decision about a single query is expressly prohibited.

For each query a KB node-id must be returned. For entities that have no corresponding node in the reference KB a response of NIL should be returned.[2] System responses will be evaluated based on the correctness of linkages to KB nodes.

Query and Output Formats

Queries will be formatted in XML and a sample file is available on the website. Here is an example of the formatting:

```
<query id="EL1"><name>John Doe</name><docid>SUN-009</docid></query>
<query id="EL2"><name>John Doe</name><docid>NYT-005</docid></query>
<query id="EL3"><name>Johnny Doe</name><docid>CNN-001</docid></query>
```

System output files should have a single response per line, and consist of two space-separated columns. The first column is the query id from the input file and the second column is either a knowledge base id, or NIL. For example:

```
EL1 E101
EL2 NIL
EL3 E5871
```

[2] Note: this is a change from earlier guidelines which discussed clustering non-KB entities together and giving them a common identifier.

Thus, the name-string from query "EL1" is believed to refer to the entity in knowledge base node E101, and the entity in query "EL2" is believed to be absent from the reference KB.

Evaluation

Results for the entity linking task will be due before the target list for the slot-filling task is provided. This is because the slot filling queries will involve entities used in the entity linking task. The entity linking task is similar to cross-document co-reference; however, here the problem requires alignment to a knowledge base, not clustering of entities. The official evaluation measure will be micro-averaged accuracy. An example is given below.

Query	System Assignment	Correct Assignment	Correct?
EL1	E101	E101	Yes
EL2	NIL	E101	No
EL3	E5871	E5871	Yes
EL4	E101	E101	Yes
EL5	NIL	NIL	Yes

Here 4 out of 5 responses are correct, so the micro-averaged accuracy is 0.80. As this is a first-year track, other evaluation scores may be computed, but micro-averaged accuracy is the sole official measure. A script which computes micro-averaged accuracy, along with macro-averages across entities, is available from the track website.

4. Slot Filling

The Slot Filling task involves learning a pre-defined set of relationships and attributes for target entities based on the documents in the test collection. A query in the Slot Filling task will contain a name-string, docid, entity-type, node-id, and a list of slots to ignore. For example: [Paul Newman, ABC-20080611-9372, PER, E2317, per:date_of_birth] might be a query for actor Paul Newman. The node id that is provided will refer to a node representing the entity in the KB. For targets for which no node exists in the KB, the node-id will begin with "NIL", e.g., "NIL102". As in the entity linking task the provided docid is intended to give context for the entity. The list of slots to ignore will indicate that no response should be returned for these slots. This might be because the slot is single-valued and the reference KB has an existing value, or because the slot isn't appropriate for the specific target entity (e.g., *gpe:currency* would not be appropriate for the state of Florida).

Systems must process the target entities (i.e., each query) independently from one another. For each slot value returned, systems must also return a single docid from the test collection that supports the value returned for the given entity and slot.

Slots can be of one of two types: single-valued slots that admit only a single value (e.g., per*:date_of_birth*) and list-valued slots that can accept more than one value (e.g., per*:employee_of*). In some cases multiple correct and supportable values may exist in the corpus for a single-valued slot. For example, there may be distinct values for *per:age,*

per:religion. or *org:website.* In such cases, any correct and supported response is sufficient.

Systems are not expected to correct or modify values from the reference KB, but only to add information. Therefore no information is sought for single-valued slots that already have a value in the KB node. Normally such slots will be included in the list of slots to ignore, but if not, then NIL would be the correct response for this situation.

Redundant information should not be returned; only novel information is of interest. However, if an attribute has a value in the initial KB, the slot should not necessarily be ignored entirely. For example, if the Wikipedia alma-mater slot for investor Warren Buffet looked like:

<fact name="alma-mater">University of Nebraska</fact>

but a document is found which notes that he graduated from Columbia University, then this should be returned as a value for *per:schools_attended* (because *per:schools_attended* allows multiple values). However, University of Nebraska should not be returned, because it is redundant with what is already in the initial KB. Similarly, if multiple equivalent values occur in the test collection, the value should be returned only once, with any one of the supporting docids (i.e., Columbia University should be returned only once even if there are multiple documents that support it as a value for *per:schools_attended)*.

Correctness, Support, Exactness, & Redundancy

Each slot value that is returned will be be marked as Correct, Inexact, Redundant, or Wrong. In order to be Correct, the slot value must be supported by the associated docid, cannot already be in the reference KB, and must be expressed exactly in the returned string. Wrong includes both incorrect and not supported by the cited docid; this is a fusion of Wrong and Unsupported in the TREC Question Answering track.

Correctness

Assessors will be instructed to follow the slot-specific guidelines, however, correctness is ultimately decided based on the opinion of the assessor. Even with slot guidelines, there are times when determination of correctness will be subjective and in these cases assessors will rely on their judgment. As the task is filling in information into a knowledge base, a general rule of thumb is that the slot value should be reasonable for inclusion in a knowledge base or Wikipedia infobox.

NIL is the appropriate response when no novel information is available in the corpus to fill in an entity's slot. If a submission returns a NIL response, then the submission may not contain any additional value for the same slot of the target entity.

Unlike recent evaluations in TREC QA, it is explicitly not required to find the most recent information. Thus an older website (*org:website*) or a younger age (*per:age*) are acceptable.

Support
Assessors may use the entire document to make a decision about whether a slot fill is supported. Absolute logical soundness is not required. For example, if a document refers to "John Doe's first wife, Ruth" then "Ruth Doe" is a plausible response for *per:spouse* for John Doe. Her married name may have been Ruth Smith. However, absent any other (i.e., contradictory) information in the document, a reader of the document might reasonably decide to add Ruth Doe into the knowledge base.

Exactness
A slot value is marked Inexact if it is supported by the associated document, but the *string* returned for the value is incomplete or includes extraneous text. Determination of inexactness will be left up to the assessors. Generally speaking, inclusion of determiners or nominal pre-modifiers would not be considered inexact (e.g., "the Department of State", "coach Joe Gibbs", "city of Baltimore"). But "Seattle and Houston" would be inexact for *per:place_of_birth* and use of a given name alone would be inexact if the document provided a surname, even outside the immediate context where a relation was apparent. In other words, given only a description of "his aunt Emily", the given name alone is sufficient, but not if the document is unambiguous that the person's name is "Emily Williams". Note that it is not required that a slot value that is supported by a document be the most complete response from the collection; so if a name is just given as "Emily" in a document it does not matter that another document is more specific.

It is not required that slot values be a contiguous span of text from the supporting document. However, values directly extracted from the text or responses that remain as close as possible to the source text are preferred. Returning "10/31/1951" when a document states "Oct. 31, 1957" is reasonable (although such normalization is not required), but returning "Halloween 1951" or "the fifth Wednesday in Oct. 1951" from the same document is not likely to be accepted because these would be very unusual forms for that information in a knowledge base. Timestamps on documents can be used in determining dates; thus, if a document refers to someone dying "July 31st last year", a year could be returned in *per:date_of_death.*

Redundancy
Responses should not be redundant with information already provided in the <facts> field in the reference knowledge base. A slot value is marked Redundant if it is supported by the associated document and is exact, but the value is already in the reference KB. Additionally, responses for list-valued slots must be distinct; assessors will group correct slot fills into sets that are conceptually equivalent, and when a system returns multiple responses for a list slot that are considered equivalent, only one instance will be considered correct for scoring purposes.

Slot Value Linking

The Slot Filling task contains an entity linking component. For example, if "Cleveland" is returned as a response for *org:headquarters*, then that slot value should be linked to the appropriate node for the corresponding GPE, if it exists in the KB. If it is not present in

the knowledge base, NIL is the correct response. Teams that do not wish to attempt to link slot values within the KB may simply return links of NIL for all slot values.

Correctness of slot fills and within-KB linkage of slot values will be measured separately. However, if a system fails to find a correct answer then it isn't possible to assign a correct link (i.e., there is a cascading errors effect).

Query and Output Formats

Slot filling queries will be formatted in XML and a sample file is available on the website. Here is an example of the formatting:

```
<query id="SF1">
  <name>John Doe</name>
  <docid>SUN-009</docid>
  <enttype>PER</enttype>
  <nodeid>E101</nodeid>
  <ignore>per:date_of_birth per:place_of_birth per:religion</ignore>
</query>

<query id="SF2">
  <name>ACME Widget Corp</name>
  <docid>NYT-006</docid>
  <enttype>ORG</enttype>
  <nodeid>NIL102</nodeid>
</query>
```

System output files should contain at least one *response* for each query-id/slot combination, except that no response should be returned for slots listed in the <ignore> field. A response consists of a single line, with a separate line for each slot value. Lines should have either four or six space-separated columns:

Column 1: query id
Column 2: the slot name
Column 3: a unique run id for the submission
Column 4: NIL, if the system believes no information is learnable for this slot. Or, a single docid which supports the slot value
Column 5: an entity id for the slot value in the reference KB (e.g., E101), or NIL
Column 6: a slot value

Except for the last column containing the slot value, the columns cannot contain whitespace characters. When no novel information is believed to be learnable for a slot, Column 4 should be NIL and Columns 5 & 6 should be left empty. If a run is not intending to link slot values to KB entities, then NIL is the appropriate value in Column 5.

For each query, the output file should contain exactly one line for each single-valued slot that is not included in the <ignore> field.

For list-valued slots, the output file should contain a separate line for each list member. A response like "Tropicana Products and Frito-Lay" would be considered inexact for *org:subsidiaries*.

The file should be sorted by slot filling query id and multiple responses for the same list slot should be contiguous. But it is not necessary to order the slots for individual queries. For example:

```
SF1 per:spouse uva1x NBC-3218 NIL Jane Smith
SF1 per:spouse uva1x CNN-387 E19837 Mary Doe
SF1 per:date_of_death uva1x ABC-007 NIL April 23, 2008
SF1 per:children uva1x CNN-387 NIL Bobby Doe
SF1 per:employee_of uva1x NIL
SF1 per:schools_attended uva1x SUN-3321 E18872 Cornell
SF1 per:schools_attended uva1x SUN-3321 NIL Harvard Law School
SF1 per:schools_attended uva1x SUN-3321 E431197 NYU Law School
...
SF2 org:headquarters uva1x NYT-001 E2134 Cleveland, Ohio
SF2 org:alternate_names uva1x NYT-701 NIL Widgets-R-Us
SF2 org:founded_by uva1x ABC-119 NIL John "Hammer" Smithson
SF2 org:website uva1x NIL
SF2 org:parents uva1x NIL
SF2 org:subsidiaries uva1x NIL
```

Evaluation

To score a submission for the slot filling task two official scores will be calculated, a *SF-value* score based solely on correctness and a *SF-linkage* score that also considers the accuracy of linking slot values within the KB. Submissions that do not attempt slot value linking will only receive one score.

SF-value will be a composite score, computed based on the accuracy of single-valued and list-valued slots. SF-value = (single-slot-score + list-slot-score) / 2.0.

The single-slot-score will be calculated over all single-valued slots from all targets:

single-slot-score = # correct responses / # of single slots

List–valued slots will be individually scored based on instance precision and instance recall. Instance precision (IP) for a given target and slot is the (number of correct list responses / number of returned list responses). Instance recall (IR) for a given target and slot is the (number of correct list responses / number of desired list instances). The list-slot-score will be based on an F-measure that considers precision more important than recall.[3] The computation of the list-F-measure for a given target and slot is:

list-F-measure = (5 * IP * IR) / (4 * IR + IP)

[3] The reason for weighing precision more than recall is because the utility of a knowledge base is likely to degrade quickly as the proportion of incorrect information increases.

The list-slot-score is a micro-average of the list-F-measure across all list-valued slots from all targets.

The SF-linkage score will be the product of the SF-value score and micro-averaged linking Precision. Link Precision will be computed only for the links of slot values that are non-NIL and deemed correct.

There may be limitations on the use of human judgments for post hoc experimentation, as was the case with TREC QA. In other words, in a future experiment a correct slot value may be discovered for an entity, but that value might never have been judged in the original evaluation.

Other evaluation measures may be computed, such as accuracies computed for individual generic slots; however, SF-value and SF-linkage will be the official evaluation measures.

5. Data

A new document collection of mainly newswire will be released for the track. It will contain on the order of a million articles.

The reference knowledge base includes hundreds of thousands of entities based on articles from an October 2008 dump of English Wikipedia. (900,000 is probably an upper bound on the number of nodes.)

The number of queries for the entity linking task will most probably be between 1,000 to 3,000 name-mention/docid pairs. Between 50 and 100 entities are expected for slot filling.

6. Other issues

Up to three runs may be submitted by each team for each of the two tasks. Submission will take place in two phases. Entity linking results will be due before slot filling queries.

Systems should not be modified once queries are downloaded. However, because of the time gap in the two tasks, it is permitted for teams to modify systems for slot filling and slot value linking after the entity linking queries have been downloaded. Nonetheless, teams may not make changes based on the specific name-strings in the entity linking query list.

Details about submission procedures will be communicated to the track mailing list. A script will be made available to ensure that submission files comply with the prescribed format.

7. Schedule

Sample data released: May 1

Reference KB and document collections available: June 10
Entity Linking target list released: July 2
Entity Linking results due: July 9
Slot Filling target list released: July 13
Slot Filling results due: July 20
Assessments available: September 22
TAC workshop: November 16-17

8. FAQ

1. I have a question about the task.
   Please post it to the list tac-kbp@nist.gov. Information about subscribing to the list is available at: http://apl.jhu.edu/~paulmac/kbp.html
2. Are there any constraints concerning the use of external resources such as internet search, WordNet, and Wikipedia?
   As a general rule, such resources can be used and it is expected that workshop papers include details about how systems make use of them. However, since the evaluation is specifically proposing to use Wikipedia information as an initial knowledge source, teams should avoid using Wikipedia infoboxes to directly fill slots. Additionally, participants are asked to refrain from editing Wikipedia pages for target entities, either during, or after the evaluation. There are many ways in which Wikipedia can be legitimately used, including as training data for attribute learning, or to compile lists of name variation based on hyperlinks and redirects.