Proposed Task Description for Knowledge-Base Population at TAC 2011

Version of: 05/15/2011

1 Introduction

The main goal of the Knowledge Base Population (KBP) track at TAC 2011 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. The overall task of populating a knowledge base is decomposed into 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. Compared to previous information extraction evaluations such as MUC and ACE, KBP involves the following new research topics:

- Extraction at large scale (over 1 million documents) ;
- Using a representative collection (not selected for relevance);
- Cross-document entity resolution (extending the limited effort in ACE);
- Linking the facts in text to knowledge base;
- Offering the possibility of distant (and noisy) supervision through Infoboxes;
- Rapid adaptation to new relations (in KBP 2010 Surprise Task).

Compared to the KBP evaluation at TAC 2010, we aim to achieve three new research goals:

- Support multi-lingual information fusion via cross-lingual KBP entity linking;
- Capture temporal information via temporal slot filling;
- Automate novel KB entry creation via NIL entity clustering.

The tasks will be structured by having participants process a list of queries over target entities. For the Entity Linking task 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. For the Slot Filling task the list will only contain PER and ORG entities.

2 Entity Linking

2.1 Mono-lingual Entity Linking

In the Entity Linking task, given a query that consists of a name string and a background document ID, the system is required to provide the ID of the KB entry to which the name refers, or a "NILxxxx" ID if there is no such KB entry. The entity linking system is required to cluster together queries referring to the same non-KB (NIL) entities and provide a unique ID for each cluster, in the form of NILxxxx (e.g., "NIL0021").

An example query from the KBP2010 Entity-Linking evaluation is

```
<query id="EL000304">
  <name>Barnhill</name>
  <docid>eng-NG-31-100578-11879229</docid>
</query>
```

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).

For the primary task, the system may consult the text from the Wikipedia pages associated with the KB nodes. There will be also an optional task in which the systems should do linking without reference to these texts. – using only the slot values; this corresponds to the task of updating a knowledge base with no 'backing' text.

2.2 Cross-lingual Entity Linking

In KBP2011 we will introduce a new cross-lingual entity linking task. The basic task setup follows mono-lingual entity linking which consists of two steps: (1) link Non-NIL queries to English KB entries; (2) cluster NIL queries. The cross-lingual aspect comes from the fact that the queries will include both English queries (Section 2.1 above) and Chinese queries. An example Chinese query is

```
<query id="EL001234">
  <name>强尼凯许</name>
  <docid>AFC20030913.0300.0024</docid>
</query>
```

2.3 Scoring Metric

For a set of query names with source documents, an entity linking system is required to: (1) judge whether each query can be linked to any KB node; (2) Cluster all queries with NIL KB entries into clusters. Ultimately the system output can be viewed as a collection of various clusters; some clusters are labeled as KB node IDs. At the same time the answer key can also be viewed as a different collection of clusters. Therefore we will apply a modified B-Cubed metric (called B-Cubed+) to evaluate these clusters. Let us use the following notation:

$L(e)$ and $C(e)$ the category and the cluster of an item e,

$SI(e)$ and $GI(e)$ the system and gold-standard KB identifier for an item e,

we can define the correctness of the relation between e and e' in the distribution as:

$$G(e,e') = \begin{cases} 1 \, iff \; L(e) = L(e') \wedge C(e) = C(e') \wedge GI(e) = SI(e) = GI(e') = SI(e') \\ 0 \, otherwise \end{cases}$$

That is, two items are correctly related when they share a category if and only if they appear in the same cluster and share the same KB identifier in the system and the gold-standard. B-cubed+ precision of an item is the proportion of correctly related items in its cluster (including itself). The overall B-Cubed+ precision is the averaged precision of all items in the distribution. Since the average is calculated over items, it is not necessary to apply any weighting according to the size of clusters or categories. The B-Cubed+ recall is analogous, replacing "cluster" with "category". Formally:

$$\text{Precision } B\text{-}Cubed+ = Avg_e[Avg_{e'.C(e)=C(e')}[G(e,e')]]$$

$$\text{Recall } B\text{-}Cubed+ = Avg_e[Avg_{e'.L(e)=L(e')}[G(e,e')]]$$

The scorer is available at: http://nlp.cs.qc.cuny.edu/kbp/2011/scoring.html.

3 Slot Filling

3.1 Mono-lingual Slot Filling

The goal of Slot Filling is to collect from the corpus information regarding certain attributes of an entity, which may be a person or some type of organization. Guidelines for each slot are available at: http://nlp.cs.qc.cuny.edu/kbp/2011/annotation.html. 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). Official names for each slot are given in Table 1.

Person	**Organization**
per:alternate_names	org:alternate_names
per:date_of_birth	org:political/religious_affiliation
per:age	org:top_members/employees
per:country_of_birth	org:number_of_employees/members
per:stateorprovince_of_birth	org:members
per:city_of_birth	org:member_of
per:origin	org:subsidiaries
per:date_of_death	org:parents
per:country_of_death	org:founded_by
per:stateorprovince_of_death	org:founded
per:city_of_death	org:dissolved
per:cause_of_death	org:country_of_headquarters
per:countries_of_residence	org:stateorprovince_of_headquarters
per:stateorprovinces_of_residence	org:city_of_headquarters
per:cities_of_residence	org:shareholders
per:schools_attended	org:website
per:title	
per:member_of	
per:employee_of	
per:religion	
per:spouse	
per:children	
per:parents	
per:siblings	
per:other_family	
per:charges	

Table 1. Slot Names for the Two Generic Entity Types

Each query in the Slot Filling task consists of the name of the entity, its type (person or organization), a document (from the corpus) in which the name appears (to disambiguate the query in case there are multiple entities with the same name), its node ID (if the entity appears in the knowledge base), and the attributes which need not be filled. Attributes are excluded if they

are already filled in the reference data base and can only take on a single value. An example query is

```
<query id="SF114">
   <name>Masi Oka</name>
   <docid>eng-WL-11-174592-12943233</docid>
   <enttype>PER</enttype>
   <nodeid>E0300113</nodeid>
   <ignore>per:date_of_birth per:age per:country_of_birth per:city_of_birth</ignore>
 </query>
```

Along with each slot fill, the system must provide the ID of a document which supports the correctness of this fill. If the corpus does not provide any information for a given attribute, the system should generate a NIL response (and no document ID). For each attribute we indicate the type of fill and whether the fill must be (at most) a single value or can be a list of values. Since the overall goal is to augment an existing KB, two types of redundancy in list-valued slots must be detected and avoided. First, two fills for the same entity and slot must refer to distinct individuals. Second, if the knowledge base already has one or more values for a slot, items in the system output must be distinct from those already in the knowledge base. In both cases, it is not sufficient that the strings be distinct; the fills must refer to distinct individuals. For example, if the knowledge base already has a slot fill "William Jefferson Clinton", the system should not generate a fill "Bill Clinton" for the same slot.

System output files should be in UTF-8 and 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 the following tab-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: a slot value

When no novel information is believed to be learnable for a slot, Column 4 should be NIL and Column 5 should be left empty.
For each query, the output file should contain exactly one line for each single-valued slot. For list-valued slots, the output file should contain a separate line for each list member.

3.2 Cross-lingual Slot Filling

For KBP 2012 we are planning to extend the slot filling task to the cross-lingual paradigm. Given a query in English and a large collection of English and Chinese documents, a system should extract slot answers about the query and present the answers in English. Because of the complexity of this task, we have prepared preliminary specifications for the task (included in this document) and will be distributing initial training data conforming to these specifications later this summer. This will allow initial experimentation on this task well in advance of an actual evaluation. The cross-lingual slot filling task in KBP2012 will only focus on the following slot types:

-per:age

-per:origin
-per:employee_of
-per:cities_of_residence
-per:spouse
-org:top_members/employees
-org:subsidiaries
-org:city_of_headquarters

The cross-lingual query format is the same as the mono-lingual format. An example query is as follows:

```
<query id="SFXXX">
   <name> Johnny Cash</name>
   <docid> AFC20030913.0300.0024</docid>
   <enttype>PER</enttype>
   <nodeid>EXXXXXXX</nodeid>
   <ignore>per:age</ignore>
 </query>
```

Each line of annotation in the training data is in the following format:
Query-ID | English Query | Slot Type | English Answer | DOCID | Offsets in source document (English or Chinese)

An example is as follows:
SF3 | Johnny Cash | per:cities_of_residence | Nashville | AFC20030913.0300.0024 | 1 4

The system responses take the same form as the mono-lingual task (section 3.1) with only English slot fills. The pooled system responses will be assessed by bi-lingual annotators. Each system is required to submit three runs: slots filled from English documents, slots filled from Chinese documents and slots filled from documents in both languages. Separate scores will be reported for these three cases.

3.3 Scoring Metric for Mono-lingual and Cross-lingual Slot Filling Tasks

We will pool the responses from all the systems and have human assessors judge the responses. To increase the chance of including answers which may be particularly difficult for a computer to find, LDC will prepare a manual key which will be included in the pooled responses.

Each response is rated as correct, inexact, redundant, or wrong. A response is inexact if it either includes part of the correct answer or includes the correct answer plus extraneous material. No credit is given for inexact answers. Two types of redundant answers are flagged for list-valued slots. First, a system response may be equivalent to an answer in the reference knowledge base; this is considered incorrect. Second, two system responses for the same attribute may be equivalent; in the latter case, only the first of a set of equivalent answers is marked correct. (This is implemented by assigning each correct answer to an *equivalence class*, and only giving credit for one member of each class.)

Given these judgments, we can count

Correct = total number of non-NIL system output slots judged correct
System = total number of non-NIL system output slots
Reference = number of single-valued slots with a correct non-NIL response +
number of equivalence classes for all list-valued slots
Recall = Correct / Reference

Precision = Correct / System
F = 2*Precision*Recall/ (Precision + Recall)

The F score is the primary metric for system evaluation.

3.4 Temporal Slot Filling

In KBP2011 we will also add a new task of temporal slot filling. The goal of this new task is to add limited temporal information to selected slots in the KBP slot-filling output. We will limit temporal information to the following slot types:

per:spouse
per:title
per:employee_of
per:member_of
per:cities_of_residence
per:stateorprovinces_of_residence
per:countries_of_residence
org:top_employees/members

There will be two versions of the task, the *full* temporal task and the *diagnostic* temporal task. For the full temporal task, the system will be given a query file just as for the regular slot filling task, and will be expected to generate a slot filling output augmented with temporal information as described below. For the diagnostic temporal task, the system will be given two files, a query file and a slot file. The slot file will have the same form as the output of a run for the regular slot filling task: each line will specify a query, a slot, a slot value, and a document supporting that slot value. The system should determine the temporal information for each specified slot value, based only on the information in the specified document.

The output for the full temporal task will be scored through system output pooling, like the regular slot filling task. The diagnostic temporal task will be based on a set of slot fills tagged through manual annotation, and will be scored automatically.

3.4.1. Representation of temporal information

Associated with each non-NIL slot value will be a 4-tuple of dates

[T1 T2 T3 T4]

indicating that the slot value is true for a period beginning at some time between T1 and T2 and ending at some time between T3 and T4. A hyphen in one of the positions implies a lack of a constraint. Thus [- 20110101 20110101 -] implies that the value was true starting on or before January 1, 2011 and ending on or after January 1, 2011; i.e., that it was true on January 1, 2011 and no further information is available from the texts. Similarly, [20100101 20101231 - -] implies that the value was true starting at some time in 2010.

The goal in selecting this representation was to be able to capture most of the temporal information conveyed in the text while still retaining the structured data base style of KBP slot filling. A pair of dates would not be sufficiently flexible – the texts often do not give specific start and end dates. On the other hand, a more general representation involving multiple temporal predicates would be a sharp departure from infobox style.

Some types of information expressed in the text cannot be captured by a 4-tuple. These include

- durations where neither endpoint is known ("he worked for IBM for 7 years")
- relations between slots ("she married Fred two years after moving to Seattle")
- slot values which are true over multiple disjoint intervals ("Cleveland was President from 1885 to 1889 and from 1893 to 1897")
- regularly recurring events ("each Friday")
- fuzzy relations ("lately", "recently") that are encoded with the SET type in TimeML.

Here are some examples of 4-tuple representation, assuming the publication date of the text is January 1, 2001:

Document text	**T1**	**T2**	**T3**	**T4**
Chairman Smith	-	20010101	20010101	-
Smith, who has been chairman for two years	-	19990101	20010101	-
Smith, who was named chairman two years ago	19990101	19990101	19990101	-
Smith, who resigned last October	-	20001001	20001001	20001031
Smith served as chairman for 7 years before leaving in 1991	19840101	19841231	19910101	19911231
Smith was named chairman in 1980	19800101	19801231	19800101	-

Table 2. 4-tuple Representation Example

Note that these values assume that durations are interpreted as being literally exact. For example, "two years ago" is interpreted as exactly two years ago, not (for example) as between 1½ and 2½ years ago. Though this is unrealistic, it simplified the task and the evaluation. In the case of a slot holding over multiple disjoint intervals, the best response will capture the period from the start of the first interval to the end of the last interval.

3.4.2 Output format

The format of the input for the temporal task will be exactly the same as for the regular slot-filling task. As for the regular slot-filling task, system output files should be in UTF-8 and 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. Lines should have the following tab-separated values:

Column 1: query id
Column 2: the slot name
Column 3: a unique run id for the submission
Column 4: one of the strings 'NIL', 'V', 'T1', 'T2', 'T3', or 'T4'
Column 5: an eight-digit date
Column 6: a document ID
Column 7: a slot value

If there is no fill for a particular slot, a single response line should be generated for the slot, with column 4 containing NIL and the remaining columns empty.

If there is a fill for a particular slot, but there are no temporal constraints on the slot, a single response line should be generated for the slot, with column 4 containing V, column 5 containing '-', column 6 containing a single docid which supports the slot value (as for the regular slot-filling task), and column 7 containing the slot value.

If there is a fill for a slot with some (1 to 4) temporal constraints on the slot, up to 4 response lines should be generated for the slot, with column 4 containing the type of constraint (T1, T2, T3, or T4), column 5 containing the date, column 6 containing the docid of a document supporting the constraint, and column 7 containing the slot value.

3.4.3. Scoring

The simplest scoring scheme would mark each constraint as correct or incorrect. However, because the time information provided by the texts may be only approximate, such all-or-nothing scoring is likely to lead to problems. Instead we propose to use a score measuring the similarity of each constraint in the key and system response. Let the date in the key be k_i and the date in the system response be r_i; let $d_i = |k_i - r_i|$, measured in years. Then the score for the set of temporal constraints on a slot is

$$S(slot) = \tfrac{1}{4} \bullet \sum_{i=1}^{4} \frac{c}{c + d_i}$$

$$c = \begin{cases} c_{overconstraining}, if\ (i \in \{1,3\} \wedge r_i > k_i) \vee (i \in \{2,4\} \wedge r_i < k_i) \\ c_{vagueness}, otherwise \end{cases}$$

where $c_{overconstraining}$ and $c_{vagueness}$ are two constants (tentatively both set to 1 year) such that errors of that amount get 50% credit. This yields a score between 0 and 1.

The absence of a constraint in T1 or T3 is treated as a value of -∞; the absence of a constraint in T2 or T4 is treated as a value of +∞.

Overall system scores are computed the same way as for regular slot filling (see section 3.3) *except* that, in computing the value of *correct*, we take the sum over all correct slot fills of *S(slot)*.

3.4.4. Training data

To facilitate system development, we will be providing annotations corresponding not only to the final 4-tuples for selected queries but also to intermediate local information regarding temporal constraints. Each instance of a slot value in the text will be annotated with temporal information. If the slot value is associated with a temporal expression representing a date or interval, the annotation will specify the temporal expression, the offset of the expression within the document, its normalized form (for example, for specific dates, its 8-digit yyyymmdd value), and the relation between the slot value and the temporal expression.

We will use a set of seven relations developed for temporal annotation by the DARPA Machine Reading program:

Relation	Role of temporal expression	Example
Beginning	the start time for the slot value	Rob joined GE in 1999
Ending	the end time for the slot value	Rob left GE in 1999
Beg_and_end	the slot value is true exactly for the specified time	Rob was named *linguist of the month* for June 1999.
Within	the slot value is true for at least some portion of the specified time	Rob worked for GE in 1999
Throughout	the slot value is true for all of the specified time	Rob commuted to work from his home in Denver for all of 1999
Before_start	a moment before the start time for the slot value	In 1999, before Rob joined GE, …
After_end	a moment after the end time for the slot value	By 1999 Rob had already left GE

Table 3. Temporal Relation Types

In addition, if the slot is currently true (as of the document date) this will be indicated in the intermediate file. Further details regarding these relations can be obtained from the *MR_KBP_Training_Guidelines*, which will be distributed along with the training data.

Each of these relations can be translated into a 4-tuple. The basic guidelines for doing so will be distributed as part of the training data. The information in the individual 4-tuples can then be aggregated across sentences and documents, in the simplest cases by taking the maximum of T1 and T3 values and the minimum of T2 and T4 values. Because – as noted above – the 4-tuples do not capture all the temporal information in the text, the procedure based on the aggregation of 4-tuples will not necessarily produce the most accurate corpus-wide 4-tuple.

4 Data

4.1 Knowledge Base and Source Collection

The reference knowledge base includes hundreds of thousands of entities based on articles from an October 2008 dump of English Wikipedia which includes 818,741 nodes.

Each entity in the KB 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)

The 'raw' slot names and the values in the reference KB are based on an October 2008 Wikipedia snapshot. To facilitate use of the reference KB a mapping from raw Wikipedia infobox slot-names to generic slots is provided in training corpora.

The following Table 4 presents the profile of the source collection.

Language	Genre	Approximate Size (documents)
English	Broadcast Conversation	17
	Broadcast News	665
	Conversational Telephone Speech	1
	Newswire	1,286,609
	Web Text	490,596
Chinese	Newswire	1 million from Chinese Gigaword

Table 4. # Documents in Source Collection

4.2 Training and Evaluation Corpus

The following Tables summarize the KBP2011 training and evaluation data which we aim to provide for participants. For all tasks we try to achieve a balance between different genres, and between the queries with and without KB entry linkages. The surprise slot filling task will not be conducted in 2011, but we are providing the KBP2010 surprise slot filling training and evaluation data for research interest.

Corpus	Genre/Source	Size (entity mentions)		
		Person	Organization	GPE
Training	2009 Eval	627	2710	567
	2010 Training Web data	500	500	500
	2010 Eval Newswire	500	500	500
	2010 Eval Web data	250	250	250
Evaluation	Newswire	500	500	500
	Web data	250	250	250

Table 5. Mono-lingual Entity Linking Data

Corpus	Source	Size (entities)	
		Person	Organization
Training	2009 Evaluation	17	31
	2010 Participants	25	25
	2010 Training	25	25
	2010 Training (Surprise SF task)	24	8
	2010 Evaluation	50	50
	2010 Evaluation (Surprise SF task)	30	10
Evaluation	2011 Evaluation	50	50

Table 6. Mono-lingual Slot Filling Data

Corpus	Size (entities)	
	Person	Organization
Training	40	10
Evaluation	80	20

Table 7. Temporal Slot Filling Data

Corpus	Genre/Source	Size (entity mentions)		
		Person	Organization	GPE
Pilot	English KBP 2010 source collection + subset of Chinese Gigaword (newswire)	250	250	250
Training		250	250	250
Evaluation		250	250	250

Table 8. Cross-lingual Entity Linking Data

Corpus	Genre/Source	Size (entities)	
		Person	Organization
Pilot	English KBP 2010 source collection + subset of Chinese Gigaword (newswire)	25	25
Training		25	25
Evaluation		50	50

Table 9. Cross-lingual Slot Filling Data

5 External Resource Restrictions and Sharing

5.1 External Resource Restrictions

As in KBP 2010, participants will be asked to make at least one run subject to certain resource constraints, primarily that the run be made as a 'closed' system … one which does not access the Web during the evaluation period. Sites may also submit an additional run with access the Web. This will provide a better understanding of the impact of external resources.

Further rules for both of the primary runs and additional runs are listed in Table 10.

Specific Rules	Specific Examples
Allowed	Using a Wikipedia derived resource to (manually or automatically) create training data
	Compiling lists of name variation based on hyperlinks and redirects before evaluation
	Using a Wikipedia derived resource before evaluation to create a KB of world knowledge which can be used to check the correctness of facts
	Preprocess/annotate a large text corpus before the evaluation to check the correctness of facts or aliases
Not Allowed	Using Wikipedia infoboxes to directly fill slots
	Editing Wikipedia pages for target entities, either during, or after the evaluation

Table 10. Rules of Using External Resources

5.2 Resource Sharing

In order to support groups which intend to focus on part of the tasks, the participants are encouraged to share the external resources that they prepared before the evaluation. The possible resources may include intermediate results, entity annotations, parsing/SRL/IE annotated Wikipedia corpus, topic model features for entity linking, patterns for slot filling, etc. The sharing process can be informal (among participants) or more formal (through a central repository built by the coordinators). Please email the coordinators in order to access the central site.

6 Submissions and Schedule

6.1 Submissions

In KBP 2011 participants will have one week after downloading the data to return their results for each task (refer to the detailed schedule in Table 10 below) Up to three alternative system runs may be submitted by each team for each task. Systems should not be modified once queries are downloaded. Details about submission procedures will be communicated to the track mailing list. The tools to validate formats are available at: http://nlp.cs.qc.cuny.edu/kbp/2011/tools.html

6.2 Schedule

An approximate schedule for KBP 2011 is presented in Table 11.

Date	Event
03/16	Preliminary task definition available
03/16	New Entity Linking Scorer available (deal with NIL clustering)
03/21	Registration site available
03/21	Release of English source collection
03/21	Release of Chinese source collection
04/28	Revised task definition available
05/03	Final slot filling annotation guidelines available
05/16	Cross-lingual entity linking and Temporal slot filling sample data sets available
05/26	Corrected KBP2010 English regular slot filling annotations available
05/26	Temporal slot filling training corpora available
06/10	Cross-lingual entity linking training corpora (including NIL clustering) available
06/10	Registration deadline
07/11-07/17	Regular Slot Filling Evaluation
07/18-07/24	Full Temporal Slot Filling Evaluation
08/01-08/07	Diagnostic Temporal Slot Filling Evaluation
08/08-08/14	Regular Entity Linking Evaluation
08/15-08/21	Cross-lingual Entity Linking Evaluation
09/19	Assessments for all tasks available
09/25	Deadline for TAC 2011 workshop presentation proposals
10/26	System description paper due
11/14-11/15	TAC 2011 workshop (NIST)

Table 11. KBP 2011 Schedule (Tentative)

7 Mailing List and Website

The KBP 2011 website is http://nlp.cs.qc.cuny.edu/kbp/2011. Please post any questions and comments to the list tac-kbp@nist.gov. Information about subscribing to the list is available at: http://nlp.cs.qc.cuny.edu/kbp/2011/mailing.html.