Proposed Task Description for Knowledge-Base Population at TAC 2012

Version 1.1 of June 3, 2012

1 Introduction

The main goal of the Knowledge Base Population (KBP) track at TAC 2012 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 source document collection from which systems are to learn. 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 (EL), where names must be aligned to entities in the reference KB and with other entities discovered in the collection; and Slot Filling (SF), which involves mining information about entities from text. (A third major task in KBP 2012, Cold Start KBP, is defined separately and integrates the Entity Linking and Slot Filling tasks to produce a new knowledge base independent of a reference KB.) 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. KBP 2012 also includes a diagnostic task for Slot Filling, Slot Filler Validation (SFV), in which SFV systems must determine the correctness of candidate fillers that are provided by systems from the full Slot Filling task.

Compared to the KBP evaluation at TAC 2011, we aim to achieve the following new research goals:

- Add confidence values to facts that are added to the knowledge base
- Provide a more focused justification for each slot filler in the form of a clause or sentence in the source document which provides justification for the relation
- support further multi-lingual information fusion by adding Spanish source documents to the entity linking and slot filling tasks; and
- integrate the various aspects of the track to construct and evaluate knowledge bases.

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

2 Entity Linking

2.1 Monolingual Entity Linking

In the Entity Linking task, given a query that consists of a name string, a background document ID, and a pair of UTF-8 character offsets indicating the start and end location of the name string in the document, 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>
  <startoffset>xxx</startoffset>
  <endoffset>yyy</endoffset>
</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

The cross-lingual entity linking tasks follow monolingual entity linking; the steps are: (1) link Non-NIL queries to English KB entries; and (2) cluster NIL queries. The cross-lingual aspect comes from the fact that the queries will include Chinese and Spanish queries. An example Chinese query is

```
<query id="EL001234">
  <name>强尼凯许</name>
  <docid>AFC20030913.0300.0024</docid>
  <startoffset>xxx</startoffset>
  <endoffset>yyy</endoffset>
</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 $GI(e)$ represent, respectively, the system (*i.e.,* participant-submitted) and gold-standard (ground truth) KB identifiers for an item e.

We 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:

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

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

The scorer is available at: http://www.nist.gov/tac/2012/KBP/tools/

3 Slot Filling

3.1 Monolingual 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 of the slots will be available at: http://www.nist.gov/tac/2012/KBP/task_guidelines.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 KBP 2012 slot are given in Table 1. Guidelines for KBP 2012 slots are based on guidelines for KBP 2011, with some small revisions. Modifications to the 2011 guidelines include changes to some of the official slot names; Table 2 shows the mapping from the KBP 2011 slot name to the KBP 2012 slot name for those slots whose names have changed.

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:date_founded
per:city_of_death	org:date_dissolved
per:cause_of_death	org:country_of_headquarters
per:countries_of_residence	org:stateorprovince_of_headquarters
per:statesorprovinces_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. KBP2012 Slot Names for the Two Generic Entity Types

2011 Slot Name	2012 Slot Name
per:stateorprovinces_of_residence	per:statesorprovinces_of_residence
org:founded	org:date_founded
org:dissolved	org:date_dissolved
org:political/religious_affiliation	org:political_religious_affiliation
org:top_members/employees	org:top_members_employees
org:number_of_employees/members	org:number_of_employees_members

Table 2. Mapping from KBP2011 to KBP2012 Slot Names

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), the start and end offsets of the name as it appears in the document, 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 database 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>
   <startoffset>xxx</startoffset>
   <endoffset>yyy</endoffset>
   <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 filler, the system must provide a confidence score and provenance for the filler. Provinence is in the form of the ID of a document that supports the correctness of this filler, start and end offsets for the slot filler as it appears in the document, and start and end offsets for the clause or sentence that justifies the relation between the query and the slot filler. If the corpus does not provide any information for a given attribute, the system should generate a NIL response (and no provenance or confidence score).

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 fillers 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 are distinct; the fillers must refer to distinct individuals. For example, if the knowledge base already has a slot filler "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: 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 that justifies the relation between the query entity and the slot filler
Column 5: a slot filler
Column 6: start offset of filler
Column 7: end offset of filler
Column 8: start offset of justification
Column 9: end offset of justification
Column 10: confidence score

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. When no novel information is believed to be learnable for a slot, Column 4 should be NIL and Columns 5-10 should be left empty. Column 5 (if present) contains the string representing the slot filler; the string should be extracted from the document in Column 4, except that dates should be normalized (year, year/month, or year/month/day as appropriate; e.g., 20070329 for March 29, 2007) and any embedded tabs or newline characters should be converted to a space character.

Provenance: Columns 6 and 7 must contain the location of the slot filler string in the document. Each document is represented as a UTF-8 character array and begins with the "<DOC>" tag, where the "<" character has index 0 for the document. The startoffset in Column 6 is the index of the first character of the string and the endoffset in Column 7 is the index of the last character of the string (therefore, the length of the string is endoffset-startoffset+1). If the slot filler in Column 5 has been normalized, the offsets in Columns 6-7 should be for the unnormalized filler. For example, a date in Column 5 that is computed from the document date and the string "yesterday" should provide the offset for "yesterday" in Columns 6-7.
Columns 8 and 9 must contain the location of the sentence or clause that provides justification for the relation. For example, for query per:spouse of "Michelle Obama" and the sentence "He is married to Michelle Obama" ("He" referring to Barack Obama mentioned earlier in the document), the filler in Column 5 should be "Barack Obama", the offsets for filler must point to "He" and the offsets for justification must point to "He is married to Michelle Obama". A human assessor will judge correctness of the (possibly normalized) slot filler string, and correctness of the offsets. However, the offsets will not be scored; rather, they are used only to allow the assessor to quickly see the points in the document where the fact is attested, and to provide further training data for systems attempting to learn contextual patterns for slots.

Confidence Scores: To promote research into probabilistic knowledge bases and confidence estimation, each non-NIL response must have an associated confidence score. Confidence scores will not be used for any official TAC 2012 measure. However, the scoring system may produce additional measures based on confidence scores. For these measures, confidence scores will be used to induce a total order over the responses being evaluated; when two scores are equal, the response appearing earlier in the submission file will be considered to have a higher confidence score for the purposes of ranking. A confidence score must be a positive real number between 0.0 (representing the lowest confidence) and 1.0 (inclusive, representing the highest confidence), and must include a decimal point (no commas, please) to clearly distinguish it from a document offset. In 2012, confidence scores may not be used to qualify two incompatible fills for a single slot; submitter systems must decide amongst such possibilities and submit only one. For example, if the system believes that Bart's only sibling is Lisa with confidence 0.7 and Milhouse with confidence 0.3, it should submit only one of these possibilities. If both are submitted, it will be interpreted as Bart having two siblings.

NIST reserves the right to assess and score only the top-ranked N non-NIL responses in each submission file, where N is determined by assessing resources and the total number of responses returned by all participants.

3.2 Cross-lingual Slot Filling

For KBP 2012 we extend the slot filling task to the cross-lingual paradigm. Given a query and a large collection of English and Spanish documents, a system should extract slot fillers for the query and present them as they appear in the source document. The queries may come from either English or Spanish documents, and the slots in the cross-lingual slot filling task will be the same as those in the English slot filling task.

The cross-lingual query format is the same as the monolingual format.

The system responses take the same form as the monolingual task (section 3.1) with slot fills in the language of the source document. No translation of slot fillers is expected, except for normalization of dates. As in the monolingual task, systems should not return more than one response for equivalent slot fillers; two slot fillers are considered equivalent if they refer to the same entity (in the case of named entities), have the same value (in the case of dates and quantities), or have the same English translation (in the case of strings, such as for per:title and per:charges). The correctness of system responses and the equivalence of slot fillers will be judged by bi-lingual annotators.

3.3 Scoring Metric for Monolingual 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 that 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 the slot filler in Column 5 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.

4 Slot Filler Validation

The Slot Filler Validation (SFV) task is motivated by a use case in which the SFV system is used as a component of a full SF system. SFV is a diagnostic task that eliminates the need for a team to have a full Slot Filling system and concentrates on the refinement of existing system(s) output. The input to the SFV system is a set of submission files from several Slot Filling runs (with the run ID anonymized appropriately). The output of the SFV system is a binary classification (Correct/Incorrect) of each candidate slot filler in each SF run.

The evaluation measures the effect of using the SFV output to filter the contributing SF runs. Each contributing SF run will be filtered and rescored in the same way as for the full SF task, and the results compared against the scores for the unfiltered SF run. SFV tries to increase Precision of the contributing SF runs without significantly reducing Recall, and the objective function is to maximize the mean F-score over each of the filtered contributing runs.

5 Data

5.1 Knowledge Base and Source Document 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 source document collection for the KBP 2012 Entity Linking and Slot Filling tasks are composed of English, Spanish, and Chinese documents from the following LDC packages:

1. LDC2010E12: TAC 2010 KBP Source Data V1.1
2. LDC2011T13: Chinese Gigaword Fifth Edition
3. LDC2011T07: English Gigaword Fifth Edition
4. LDC2011T12: Spanish Gigaword Third Edition
5. LDC2012E23: TAC 2012 KBP Source Corpus Additions Web Documents

Only a subset of the English, Spanish, and Chinese Gigaword collections will included in the official KBP 2012 tasks; the document IDs of those Gigaword documents that are part of the KBP 2012 tasks are listed in LDC catalog item LDC2012E22 (TAC 2012 KBP Source Corpus Additions Newswire Doc-ID Lists).

The following Table 3 presents the profile of the collection of source documents for the KBP 2012 entity-linking and slot-filling tasks.

Language	Source	Genre	Size (documents)
English	LDC2010E12	Newswire, Web Text, Miscellaneous	1,777,890
	LDC2011T07	Newswire	1,000,257 (list in LDC2012E22)
	LDC2012E23	Web Text	1,000,000
Chinese	LDC2011T13	Newswire	2,000,256 (list in LDC2012E22)
	LDC2012E23	Web Text	815,886
Spanish	LDC2011T12	Newswire	1,000,020 (list in LDC2012E22)

Table 3. Distribution of Documents in KBP 2012 Source Document Collection

5.2 Training and Evaluation Corpus

The following Tables summarize the KBP 2012 training and evaluation data that we aim to provide for participants. For all tasks we try to achieve a balance among genres, and between the queries with and without KB entry linkages.

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
	2011 Eval Newswire	500	491	500
	2011 Eval Web data	250	259	250
Evaluation (estimate)	2012 Newswire	444	444	444
	2012 Web data	222	222	222

Table 4. English Monolingual Entity Linking Data

Corpus	Genre/Source	Size (entity mentions)		
		Person	Organization	GPE
Training	2011 Training English/Chinese Newswire	250	250	250
	2011 Eval English/Chinese Newswire	250	250	250
	2012 Training Chinese Web	50	50	50
Evaluation (estimate)	2012 Eval Chinese Newswire	444	444	444
	2012 Eval Chinese Web	222	222	222

Table 5. Chinese Cross-lingual Entity Linking Data

Corpus	Genre/Source	Size (entity mentions)		
		Person	Organization	GPE
Training	2012 Training	615	615	615

(estimate)	Spanish Newswire			
Evaluation (estimate)	2012 Eval Spanish Newswire	667	667	667

Table 6. Spanish Cross-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
	2011 Evaluation	50	50
Evaluation	2012 Evaluation	40	40

Table 7. English Monolingual Slot Filling Data

Corpus	Genre/Source	Size (entities)	
		Person	Organization
Training	2012 English News/Web; 2012 Spanish News	25	25
Evaluation		40	40

Table 8. Spanish Cross-lingual Slot Filling Data

6 External Resource Restrictions and Sharing

6.1 External Resource Restrictions

As in KBP 2010, participants will be asked to make at least one run (the first 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 additional runs which 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 9.

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 and/or Freebase to directly fill slots or directly validate candidate slot fillers for the evaluation query
	Editing Wikipedia pages for target entities, either during, or after the evaluation

Table 9. Rules of Using External Resources

6.2 Resource Sharing

In order to support groups that 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.

7 Submissions and Schedule

7.1 Submissions

In KBP 2012 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 five 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 will be made available at: http://www.nist.gov/tac/2012/KBP/tools/ .

7.2 Schedule

An approximate schedule for KBP 2012 is presented in Table 10.

Date	Event
03/05	Call for Participation; Registration site available
04/23	Release of English, Chinese, and Spanish source document collection
05/03	Preliminary task definitions available
06/01	KBP 2012 slot guidelines and Slot Filling assessment guidelines available
06/01	Spanish Entity Linking sample data set available
06/04	Cold Start sample corpora available
06/11	Spanish Slot Filling sample data set available
06/15	Spanish Entity Linking training corpora available
06/15	Registration deadline
06/18	Spanish Slot Filling training corpora available
07/18-07/25	English Slot Filling Evaluation
07/25-08/01	Spanish Slot Filling Evaluation
08/01-08/15	Cold Start Evaluation
08/01-08/31	Chinese Entity Linking Evaluation (pick one week)
08/01-08/31	English Entity Linking Evaluation (pick one week)
08/01-08/31	Spanish Entity Linking Evaluation (pick one week)
08/01-08/31	English Slot Filler Validation Evaluation (pick one week)

09/12	Assessments for Entity Linking and Slot Filling tasks available
09/26	Assessments for Cold Start task available
09/26	Deadline for TAC 2012 workshop presentation proposals
mid October	System description paper due
11/05-11/06	TAC 2012 workshop (NIST)

Table 10. KBP 2012 Schedule (Tentative)

8 Mailing List and Website

The KBP 2012 website is http://www.nist.gov/tac/2012/KBP/. Please post any questions and comments to the list tac-kbp@nist.gov. Information about subscribing to the list is available at: http://www.nist.gov/tac/2012/KBP/registration.html.