Task Description for English Slot Filling at TAC-KBP 2014

Version 1.1 of May 6th, 2014

1. Changes

- 1.0 – Initial release
- 1.1 – Changed output format: added provenance for filler values (Column 6 in Table 2) to help LDC judge Correct vs. Inexact fillers. Changed input format: removed the <nodeid> and <ignore> fields.

2. Introduction

The main goal of the Knowledge Base Population (KBP) track at TAC 2014 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. This document focuses only on the English Slot Filling (SF) task, 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. For the other tasks part of KBP 2014, please visit the KBP web page: http://www.nist.gov/tac/2014/KBP/.

The slot filling task at TAC-KBP 2014 follows closely the 2013 definition. There are, however, three important changes that will be implemented this year:

1. This year, a percentage of the queries will contain entity names that are ambiguous across the document collection. For example, "Michael Jordan" may refer to the basketball player or the Berkeley machine learning professor. The goal of this exercise is to encourage participants to combine multiple KBP tasks, in this particular case, entity linking and slot filling.
2. This year we accept outputs created through inference, provided it is justified in the KBP document collection. For example, a system could correctly infer the filler "per:country_of_birth=France" from two texts (potentially appearing in two different documents): "He was born in Paris" and "Paris is the capital of France". To accommodate this change, the output format for SF changes this year. See Section 5 for details.
3. This year the evaluation queries don't include links to the reference KB or specify slots to ignore.

3. Description

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://surdeanu.info/kbp2014/def.php. 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 2014 slot are given in Table 1. The guidelines for KBP 2014 slots are identical to the guidelines from KBP 2013.

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:employee_or_member_of	
per:religion	
per:spouse	
per:children	
per:parents	
per:siblings	
per:other_family	
per:charges	

Table 1 KBP2014 Slot Names for the Two Generic Entity Types

4. Input Format

This year's input query format is close to the 2013 format, with two changes:

1. We removed the <nodeid> field , which links the input entity to the reference knowledge base. The reasoning behind this decision is to align the input formats between Slot Filling and Cold Start. Note that the entity can still be disambiguated using the provided docid and beginning/end offsets.
2. Because the link between the entity and the KB is no longer provided, the <ignore> field, which listed slots to be ignored during extraction because they were already populated in the KB, is also removed.

Thus, 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, and the start and end offsets of the name as it appears in the document (to disambiguate the query in case there are multiple entities with the same name). An example query is:

```
<query id="SF_002">
  <name>PhillyInquirer</name>
  <docid>eng-NG-31-141808-9966244</docid>
  <beg>757</beg>
  <end>770</end>
  <enttype>ORG</enttype>
</query>
```

Along with each slot filler, the system must provide a confidence score and justification for the extraction. See Section 5 for details on the output format. If the corpus does not provide any information for a given attribute, the system should generate a NIL response (and no filler value 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. For list-valued slots, fillers returned for the same entity and slot must refer to distinct individuals. It is not sufficient that the strings be distinct; for example, if the system finds both “William Jefferson Clinton” and “Bill Clinton” as fillers for the same entity and slot, it should return only one of those fillers (the other would be considered redundant and reduce system precision).

5. Output Format

The new SF output format is driven by two observations:

1. It is designed to allow justifications that aggregate information from multiple different documents. This was not supported by the 2013 SF output format. However, please note that the **LDC’s assessment guidelines do not change, other than accepting justifications coming from multiple documents**. This means that a slot filler is considered correct only if the justification unambiguously supports the extraction.
2. During the 2013 assessment process, LDC derived no benefit from having the entity and filler provenances (Columns 6 and 7 in the 2013 format). Thus, we would like to simplify the requirements for provenance. We will still require the provenance for the relation itself (formerly Column 8 in the 2013 format) and a simplified form of filler provenance (see below).

Similar to 2013, the 2014 format requires that system output files be in UTF-8 and contain at least one *response* for each query-id/slot combination. A response consists of a single line, with a separate line for each slot value. Lines should have the following six tab-separated columns:

Column 1	Query id (same as 2013)
Column 2	Slot name (same as 2013)
Column 3	A unique run id for the submission (same as 2013)

Column 4	NIL, if the system believes that no information is learnable for this slot, in which case Columns 5 through 7 are empty; or provenance for the relation between the query entity and slot filler, consisting of up to ***4 triples*** in the format: docid:startoffset-endoffset separated by comma. Each of these individual spans may be at most ***150 UTF-8 characters***. Similar to 2013, 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. Note that the beginning <DOC> tag varies slightly across the different document genres included in the source corpus: it can be spelled both with upper case and lower case letters, and it may include additional attributes such as "id" (e.g., <doc id="doc_id_string"> is valid document start tag). Thus, offsets are counted *before* XML tags are removed. In general, start offsets in these columns must be the index of the first character in the corresponding string, and end offsets must be the index of the last character of the string (therefore, the length of the corresponding mention string is endoffset – startoffset + 1).
Column 5	A slot filler (possibly normalized, e.g., for dates) (same as 2013)
Column 6	Provenance for the slot filler string. This is either a single span (docid:startoffset-endoffset) from the document where the **canonical** slot filler string was extracted, or (in the case when the slot filler string in Column 5 has been normalized) a set of up to two docid:startoffset-endoffset spans for the base strings that were used to generate the normalized slot filler string. Same as Column 4, multiple spans must be separated by commas. The documents used for the slot filler string provenance must be a subset of the documents in Column 4. LDC will judge Correct vs. Inexact with respect to the document(s) provided in the slot filler string provenance.
Column 7	Confidence score (same as Column 9 in 2013)

Table 2 Description of SF output format

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 information is believed to be learnable for a slot, Column 4 should be NIL and Columns 5 through 7 should be left empty.

Relation Provenance: The provenance stored in Column 4 must contain text that justifies the extracted relation. That is, it must include some mention of the subject and object entities and some text supporting the slot/predicate that connects them. For example, consider the query "per:country_of_birth" for the entity "Michele Obama" and the texts:

> *Michelle Obama started her career as a corporate lawyer specializing in marketing and intellectual property. She was born in Chicago.*
>
> *…*

> *Chicago is the third most populous city in the United States, after New York City and Los Angeles.*

Using this information, a system can correctly extract the filler "per:country_of_birth=United States" for the above query. The provenance for this filler must include elements of the last two sentences, at least: "She was born in Chicago" and "Chicago is the third most populous city in the United States" (which were necessary to perform the inference that generated this slot filler). Importantly, the provenance no longer has to include text that disambiguates ambiguous mentions of entity and filler (although systems will not be penalized if they do). In this particular example, the entity mention is ambiguous in the above provenance ("She"). LDC assessors will manually disambiguate such mentions by reading a few sentences surrounding the provided provenance (this was proved sufficient in the previous evaluations). The human assessor will judge the correctness of the (possibly normalized) slot filler string, and correctness of the provenance offsets. We will report two different scores for this task: (a) ignoring the provenance offsets, and (b) scoring the provenance offsets, i.e., a slot filler will be considered correct only if both its value and its justification are correct.
All in all, assuming the first block of text starts at offset 100 in document D1, and the second starts at offset 200 in document D2, a valid encoding for this provenance would be (without the quotes): "D1:209-232,D2:200-260".

Although inference is encouraged, NIST reserves the right to ignore some slots and all their submitted responses for the purposes of assessment and scoring, especially slots such as per:countries_of_residence and per:statesorprovinces_of_residence, which may be inferred easily from per:cities_of_residence. Note that this decision will be taken **after** all submissions are received, when NIST can understand which query + slot combination led to a large number of fillers.

Filler Values: Column 5 (if present) contains the canonical string representing the slot filler; the string should be extracted from the filler provenance in Column 6, except that any embedded tabs or newline characters should be converted to a space character and dates must be normalized. Systems have to normalize document text strings to standardized month, day, and/or year values, following the TIMEX2 format of yyyy-mm-dd (e.g., document text "New Year's Day 1985" would be normalized as "1985-01-01"). If a full date cannot be inferred using document text and metadata, partial date normalizations are allowed using "X" for the missing information. For example:

- May 4th" would be normalized as "XXXX-05-04";
- "1985" would be normalized as "1985-XX-XX";
- "the early 1900s" would be normalized as "19XX-XX-XX" (note that there is no aspect of the normalization that captures the "early" part of the filler).

See the assessment guidelines document (available here: http://surdeanu.info/kbp2014/def.php) for more details on the normalization requirements.

Filler Provenance: As mentioned in Table 2, the filler provenance must point to a canonical mention, rather than an arbitrary mention. For example, if the provenance document for the above per:country_of_birth example contains both "United States" and "US", the filler and the corresponding provenance must point to "United States".

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 2014 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 2014, 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.

6. Particular Cases

Slots that require special provenance are handled similarly to 2013.

per:alternate_names

The per:alternate_name slot needs separate treatment because systems may extract it without much contextual information (other than occurrence in the same document). While textual patterns may sometimes provide useful context for this slot (e.g., "Dr. Jekyll, *also known as* Mr. Hyde"), it is possible to extract instances of this slot without such information. For example, a system may decide that "IBM" is an alternate name for "International Business Machines" solely based on the fact that the former is an acronym for the latter and they appear in the same document. In these situations, the provenance must contain sentences that mention the alias(es) used to extract the acronym. For the above example, it should contain two sentences mentioning "International Business Machines" and "IBM", respectively.

per:title

The definition of the per:title slot follows the 2013 changes. The main difference from 2012 (and before) is that titles that represent positions at different organizations must be reported as distinct fillers. For example, "Mitt Romney" has held three different "CEO" positions:

CEO, Bain Capital (1984–2002)
CEO, Bain & Company (1991–92)
CEO, 2002 Winter Olympics Organizing Committee (1999–2002)

These positions must be reported as distinct titles by the systems. Note that this is different from the past evaluations. In the previous evaluations, these titles would be merged into a single instance, because the strings ("CEO") are similar. This year, we are considering these as three distinct, valid fillers since they each refer to a different position at different organizations.

Note that this change in specification does not apply to occupations that have no clear affiliation (e.g., "actor", "star") or to positions where the affiliation is missing. In such situations, the systems (and the human assessors) should revert to the matching criterion of the previous year, where the context for the title slot filler is ignored. One more complicated scenario involves multiple positions with affiliation present for only a few. For example, "M. Smith" may appear in a document as "professor at NYU", "professor at Berkeley" or simply as "professor". In such situations, the position without an affiliation must be reported as separate filler, distinct from the ones with explicit affiliation. In the above example, an ideal system would extract three "professor" fillers, one for the position at NYU, one for the position at Berkeley, and a final one for the unaffiliated position.

The provenance for per:title must contain the corresponding organization, if present, e.g., the sentence containing "professor at NYU" for the previous example.

Please read the slot annotation guidelines for more details.

7 Scoring

The scoring procedure is carried over from 2013. 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.

The slot filler (Column 5) in each non-Nil response is assessed as Correct, ineXact, Redundant, or Wrong:

1. A response that contains more than four provenance triples (Column 4) will be assessed as Wrong.
2. Otherwise, if the text spans defined by the offsets in Column 4 (+/- a few sentences on either side of each span) do not contain sufficient information to justify that the slot filler is correct, then the slot filler will also be assessed as Wrong.
3. Otherwise, if the text spans justify the slot filler but the slot filler in Column 5 either includes only part of the correct answer or includes the correct answer plus extraneous material, the slot filler will be assessed as ineXact. No credit is given for ineXact slot fillers, but the assessor will provide a diagnostic assessment of the correctness of the justification offsets for the response. Note: correct filler strings will be assessed using the information provided in Column 6 of the output format (see Table 2).

4. Otherwise, if the text spans justify the slot filler and the slot filler string in Column 5 is exact, the slot filler will be judged as Correct (if it is not in the live Wikipedia at the date of query development) or Redundant (if it exists in the live Wikipedia). The assessor will also provide a diagnostic assessment of the correctness of the justification offsets for the response.

Two types of redundant slot fillers are flagged for list-valued slots. First, two or more system responses for the same query entity and slot may have equivalent slot fillers; in this case, the system is given credit for only one response, and is penalized for all additional equivalent slot fillers. (This is implemented by assigning each correct response to an *equivalence class*, and giving credit for only one member of each class.) Second, a system response will be assessed as Redundant with the live Wikipedia; in KBP 2014, these Redundant responses are counted as Correct, but NIST will also report an additional score in which such Redundant responses are neither rewarded nor penalized (i.e., they do not contribute to the total counts of Correct, System, and Reference below).

Given these judgments, we can count:

Correct = total number of correct equivalence classes in system responses
System = total number of non-NIL system responses
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.

8 Data

The 2014 SF task will use the same knowledge base and source document collection as 2013. We detail these resources below.

Knowledge Base and Source Document Collection

The reference knowledge base includes nodes for 818,741 entities based on articles from an October 2008 dump of English Wikipedia.
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 partial mapping from raw Wikipedia infobox slot-names to generic slots is provided in training corpora. Note that this year the reference KB is used solely as a potential training resource. As discussed above, the assessment of Redundant filler is performed against the live Wikipedia.

The source documents for the KBP 2014 English Slot Filling tasks will be identical to 2013, and will include ~1M newswire documents from a subset of Gigaword (5th edition), ~1M web documents, and ~100K documents from discussion fora. This collection will be distributed by LDC to KBP participants as a single corpus, entitled “TAC 2014 KBP English Source Corpus”, with Catalog ID LDC2014E13. In addition of the source documents, this corpus contains the output of BBN’s SERIF NLP pipeline on these documents. We hope that this simplifies data management and system development for participants.

Training and Evaluation Corpus

The following table summarize the KBP 2014 training and evaluation data that we aim to provide for participants.

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
	2012 Evaluation	40	40
	2013 Evaluation	50	50
Evaluation	2014 Evaluation	50	50

Table 3 English Monolingual Slot Filling Data

9 External Resource Restrictions and Sharing

External Resource Restrictions

As in previous KBP evaluations, 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, i.e., one which does not access the Web during the evaluation period. Sites may also submit additional runs that 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 5.

Specific Rules	Specific Examples
Allowed	Using a Wikipedia derived resource such as Freebase 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. Note that manual annotations of this data are allowed for what is considered world-knowledge (e.g., gazetteers, lists of entities) but only automatically-generated annotations are accepted for KBs of relations that can be directly mapped to slots used in this evaluation.
	Preprocess/annotate a large text corpus before the evaluation to check the correctness of facts or aliases. Same as above, only automatically-generated annotations are accepted for KBs of relations that can be directly mapped to slots used in this evaluation.
Not Allowed	Using structured knowledge bases (e.g., Wikipedia infoboxes, DBPedia, 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 4 Rules on Using External Resources

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 (such as query reformulations), 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 to get information about the central sharing site.

10 Submissions and Schedule

Submissions

In KBP 2014 participants will have one week after the evaluation queries are released to return their results for each task. Up to five alternative system runs may be submitted by each team for each task. Submitted runs should be ranked according to their expected score (based on development data, for example). 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://surdeanu.info/kbp2014/software.php.

Schedule

Please visit the slot-filling website for an approximate schedule for the English Slot Filling tasks at KBP 2014: http://surdeanu.info/kbp2014/.

11 Mailing List and Website

The KBP 2014 website is http://www.nist.gov/tac/2014/KBP/. The website dedicated to the English slot filling tasks is http://surdeanu.info/kbp2014/. Please post any questions and comments to the list tac-kbp@nist.gov. You must be subscribed to the list in order to post to the list. Information about subscribing to the list is available at: http://www.nist.gov/tac/2014/KBP/registration.html.