Cold Start Knowledge Base Population at TAC 2016

Task Description[1]

Version 1.0 of July 11, 2016

[1] The TAC organizing committee welcomes comments on this Task Description, or on any aspect of the TAC evaluation. Please send comments to tac-kbp@nist.gov.

What's New

Cold Start 2016 has two task variants: A full end-to-end Knowledge Base Construction (CSKB) task, and a component Slot Filling (CSSF) task. A component Entity Discovery and Linking (EDL) task is organized completely under the EDL track, but has the same input documents as CSKB and CSSF, to allow the entity discovery component of CSKB systems to be evaluated on the same documents as standalone EDL systems. In order to enable teams with slot filling systems to also participate in the end-to-end KB construction task, two EDL evaluation windows are offered and staged such that teams constructing a KB are given the output of EDL systems participating in the first EDL evaluation window.

This document describes the 2016 Cold Start SF/KB Construction tasks. The detailed task description for EDL is at the EDL 2016 web site (http://nlp.cs.rpi.edu/kbp/2016/).

The 2016 Cold Start SF/KB Construction tasks are identical to the 2015 tasks, with the following changes:

1. The Cold Start tasks are ***cross-lingual***; in addition to English, the 2016 source corpus includes Chinese and Spanish documents. Cross-lingual SF/KB Construction systems may return entity mentions, slot fillers and provenance from any combination of English, Chinese, and Spanish documents. Additionally, 3 diagnostic monolingual versions of these tasks are offered (one for each language), in which entity mentions, slot fillers and provenance must come from only the single language.
2. In addition to person (PER), organization (ORG), and geopolitical entity (GPE) types, KB Construction systems must return mentions of **location (LOC)** and **facility (FAC)** entities (although the slot inventory will *not* be modified to include LOC and FAC entities).
3. In addition to named mentions, KB Construction systems must extract and link all ***nominal mentions*** of *specific individual* PER, ORG, GPE, LOC, and FAC entities.
4. Cold Start SF/KB Construction systems may return a nominal mention as a filler *if **no** name mention is available in the source corpus*.

Introduction

Since 2009, TAC has evaluated performance on two important aspects of knowledge base population: entity linking and slot filling. The goal of the Cold Start track is to exercise both of these areas, and evaluate the ability of a system to use these technologies to actually construct a knowledge base (KB) from the information provided in a text collection. Cold Start participants build a software system that processes a large text collection and creates a knowledge base that is consistent with and accurately represents the content of that collection. The knowledge base is then evaluated as a single connected resource, using queries that traverse entity nodes and relation (slot) links in the KB to determine if the KB contains correct relations between correct entities.

In 2016, Cold Start has two task variants.

1. In the *Knowledge Base* variant (CSKB), participants submit entire knowledge bases, without prior knowledge of the evaluation queries.
2. The *Slot Filling* variant (CSSF) supersedes the 2014 Slot Filling track and is designed to make it easy for sites with slot filling systems to participate in Cold Start. In this variant, the Cold Start evaluation queries are split into Cold Start Slot Filling queries, with one entry point per query, and are distributed at the start of the task evaluation window. Participants do not have to submit entire knowledge bases. Rather, they apply their slot filling system

twice, the first time on the entry point for each query, the second time on each of the results of the first round.

The Entity Discovery and Linking task and the Slot Filling task have done a good job of evaluating key components of knowledge base population. They do not, however, evaluate every aspect of an automatically generated knowledge base. Things one might like to know about such a knowledge base include:

- Are the entities in the knowledge base correctly tied to real-world mentions of those entities? TAC Entity Discovery and Linking (EDL) tasks have measured this.
- Are the facts and relations in the knowledge base accurate reflections of the facts and relations described in the source documents? The TAC Slot Filling tasks have measured this, as will TAC Cold Start SF.
- Are entity linking and slot filling correctly coordinated to produce a meaningful knowledge base? The TAC Cold Start KB task measures this.
- Can the knowledge base correctly perform inference over the extracted entities, such as temporal reasoning, confidence estimation, default reasoning, transitive closure, etc.? Cold Start is just beginning to measure this; it is designed to facilitate this kind of evaluation more thoroughly in future years.

We call the task *Cold Start* Knowledge Base Population to convey two features of the evaluation: it implies both that a knowledge base schema has been established at the start of the task, and that the knowledge base is initially unpopulated. Thus, we assume that a schema exists for the entities, facts, and relations that will compose the knowledge base; it is not part of the task to automatically identify and name facts and relationships present in the text collection. In 2016, we use a schema that combines the entity types from TAC KBP 2016 Entity Discovery and Linking, and the relation types from TAC KBP 2015 Cold Start Knowledge Base Population. Thus, the schema will include five entity types (person, organization, geopolitical entity, facility, and location) and forty-one relation types and their inverses. For relations whose fills are themselves entities (such as `per:siblings` or `org:subsidiaries`), CSKB systems will be required to link that slot to the node in the submitted KB representing the correct entity[2]. Slots whose fills are strings (such as `per:title` or `org:website`) will simply use strings to represent the information.

Cold Start also implies that the knowledge base is initially empty. To avoid solutions that rely on verifying content already present in Wikipedia or other large data sources about entities, the queries used in Cold Start will be dominated by entities that are not present in Wikipedia.

All participating systems will receive the following input:

1. a *document collection*;
2. a *knowledge base schema*

From these, systems participating in the Knowledge Base variant will produce a knowledge base. This KB will be submitted to NIST as a set of augmented triples. Participating KB systems must tie each entity mention in the document collection to a particular KB entity node; in this way, the knowledge base can be queried without first aligning it to a reference knowledge base. Knowledge bases will include `mention`, `nominal_mention`, `canonical_mention`, and `type` triples, as well as the full range of slot fills (all triples are described more fully below).

Systems participating in the Slot Filling variant will also receive:

[2] Because facility and location entities are not included in the slot definitions, only person, organization, and geopolitical entity nodes must be linked to the slots.

3. a set of Cold Start Slot Filling (CSSF) evaluation queries (each evaluation query is a sequence of one or two slot filling queries to be applied in series).

For both variants, the results will then be evaluated by NIST:

- Systems participating in the Slot Filling variant return slot fillers directly in response to the given CSSF evaluation queries, and the fillers are then assessed and scored for precision and recall.
- Evaluation of the Knowledge Base variant will start by applying the same CSSF evaluation queries to the submitted knowledge base. Each query will start at a named entity mention in a document (identified by the query's <beg> and <end> tags), identify the knowledge base entity that corresponds to that mention, follow a sequence of one or more relations within the knowledge base, and end in a slot fill. The resulting slot fills will be assessed and scored in the same way as in the Slot Filling variant. For example, a CSSF evaluation query might ask 'what are the ages of the siblings of the *Bart Simpson*[3] mentioned in Document 42?' A system that correctly identified descriptions of Bart's siblings in the document collection, linked them to the appropriate node in the KB, and also found evidence for and correctly represented the ages of those siblings would receive full credit.

[3] Many of the examples used to illustrate the Cold Start task are drawn from *The Simpsons* television show. Readers lacking a detailed working knowledge of genealogical relationships in the Bouvier/Simpson family need not agonize over what they have been doing with their lives for the past quarter century, but may simply visit http://simpsons.wikia.com/wiki/Simpson_Family.

Relation	Inverse(s)
per:children	per:parents
per:other_family	per:other_family
per:parents	per:children
per:siblings	per:siblings
per:spouse	per:spouse
per:employee_or_member_of	{org,gpe}:employees_or_members*
per:schools_attended	org:students*
per:city_of_birth	gpe:births_in_city*
per:stateorprovince_of_birth	gpe:births_in_stateorprovince*
per:country_of_birth	gpe:births_in_country*
per:cities_of_residence	gpe:residents_of_city*
per:statesorprovinces_of_residence	gpe:residents_of_stateorprovince
per:countries_of_residence	gpe:residents_of_country*
per:city_of_death	gpe:deaths_in_city*
per:stateorprovince_of_death	gpe:deaths_in_stateorprovince*
per:country_of_death	gpe:deaths_in_country*
org:shareholders	{per,org,gpe}:holds_shares_in*
org:founded_by	{per,org,gpe}:organizations_founded*
org:top_members_employees	per:top_member_employee_of*
{org,gpe}:member_of	org:members
org:members	{org,gpe}:member_of
org:parents	{org,gpe}:subsidiaries
org:subsidiaries	org:parents
org:city_of_headquarters	gpe:headquarters_in_city*
org:stateorprovince_of_headquarters	gpe:headquarters_in_stateorprovince*
org:country_of_headquarters	gpe:headquarters_in_country*

Table 1. Entity-valued slots. Slots with asterisks represent inverse relations that will need to be added by participants from previous years Slot Filling task (2014 and earlier). The type qualifier of each relation (`per`, `org` or `gpe`) is the type of its subject, while the type qualifier for its inverse is the type of its object. A set of types means that any of those types is acceptable for that slot. All submitted slot names must use only a single type specification.

per:alternate_names	org:alternate_names
per:date_of_birth	org:political_religious_affiliation
per:age	org:number_of_employees_members
per:origin	org:date_founded
per:date_of_death	org:date_dissolved
per:cause_of_death	org:website
per:title	
per:religion	
per:charges	

Table 2. String-valued slots.

Schema

The schema for Cold Start 2016 combines the entity and mention types from TAC KBP 2016 Entity Discovery and Linking, and the relation types from TAC KBP 2015 Cold Start Knowledge Base Population. Thus, the schema includes five entity types (person, organization, geopolitical entity, facility, and location) and forty-one relation types and their inverses. Annotation/assessment guidelines are available on the TAC web site (http://www.nist.gov/tac/2016/KBP/ColdStart/guidelines.html), and are more fully documented in the data packages that can be requested from the LDC upon completion of TAC KBP track registration.

Cold Start entities and entity mentions are defined by DEFT Rich ERE. Full annotation guidelines for DEFT Rich ERE entities are included in the DEFT Rich ERE annotation packages, available from the LDC, but a high-level summary of the five entity types and their mentions are available in *Rich ERE Annotation Guidelines Overview*. For Cold Start, the entity mention types that must be extracted are limited to ***named and nominal mentions***, and the entities must be ***specific individual entities*** (as described in *Annotation Guidelines for Individuality of Specific Entities*). A Cold Start *named entity mention* is the same as a named entity mention in Rich ERE; i.e., a Cold Start named entity mention is a mention that uniquely refers to an entity by its proper name, acronym, nickname, alias, abbreviation, or other alternate name, and includes post author names found in the metadata of discussion forum documents. The extent of the named entity mention is the entire string representing the name, excluding the preceding definite article and any other pre-posed or post-posed modifiers. A Cold Start *nominal entity mention* is the head of the nominal entity mention in Rich ERE; i.e., a Cold Start nominal entity mention is a mention not including the entity's proper name, referring to it by a common noun phrase (but for Cold Start, the nominal mention is only the head noun of the nominal phrase). Entity mentions are allowed to nest or overlap; for example, the string "Philadelphia Eagles" might be a mention of an ORG (the football team), while the first word might simultaneously be a mention of a GPE (the city of Philadelphia).

The Cold Start inventory of slots is described thoroughly in *TAC KBP 2015 Slot Descriptions* and *TAC KBP 2015 Assessment Guidelines* available on the TAC Web site. Forty-one slots and their inverses are used for the evaluation. Twenty-six of these have fills that are themselves entities, as shown in Table 1. The remaining fifteen slots have string fills, as shown in Table 2. Each entity-valued slot will

have an inverse.[4] All inverse relations must be explicitly identified in the submitted knowledge base. That is, if the KB asserts that relation R holds between entities A and B, then it must also assert that relation R^{-1} holds between B and A. As a convenience, the Cold Start KB validation script can be used to introduce missing inverses into a KB.

Document Collection

The Cold Start 2016 evaluation document collection will be the ***TAC KBP 2016 Evaluation Source Corpus***, which comprises approximately 90,000 documents, roughly equally distributed between English, Spanish, and Chinese, and balanced between newswire (NW) and multi-post discussion forum (MPDF) documents. These documents will be new (previously unreleased) documents that will be distributed by NIST via Web download at the beginning of the Cold Start evaluation window. There will be exactly one file per document, and all files will be parsable as XML. Each file will begin with the opening tag of the <DOC> element (<doc> for MPDF); [5] note that <DOC> can be spelled with either upper case or lower case letters, depending on the genre, and may optionally include additional attributes (such as "type" for some newswire data).

Newswire data will use the following markup framework:

```
<DOC id="{doc_id_string}" type="{doc_type_label}">
<HEADLINE>
...
</HEADLINE>
<DATELINE>
...
</DATELINE>
<TEXT>
<P>
...
</P>
...
</TEXT>
</DOC>
```

[4] Some slots, such as per:siblings, are symmetric. Others, such as per:parents, have inverses that were already in the 2014 English Slot Filling track (in this case, per:children). The remaining slots (e.g., org:founded_by) had no corresponding slot in the 2014 English Slot Filling track; Cold Start specifies new slot names for these inverses. All such slots are list-valued.

[5] In contrast to some of the KBP source corpora from previous years, the TAC KBP 2016 Source Corpus will *not* contain any files that begin with xml declarations such as <?xml version="1.0" encoding="utf-8"?>. This is to ensure that offsets align across the various KBP 2016 tracks that are using this same evaluation source corpus, regardless of whether offsets are counted from the beginning of the file, or the beginning of the <DOC> tag.

where the HEADLINE and DATELINE tags are optional (not always present), and the TEXT content may or may not include "<P> ... </P>" tags (depending on whether or not the "doc_type_label" is "story").

Multi-Post Discussion Forum files (MPDFs) are derived from Discussion Forum threads. They consist of a continuous run of posts from a thread but they are only approximately 800 words in length (excluding metadata and text within <quote> elements). When taken from a short thread, a MPDF may comprise the entire thread. However, when taken from longer threads, a MPDF is a truncated version of its source, though it will always start with the preliminary post. The MPDF files will use the following markup framework, in which there may also be arbitrarily deep nesting of quote elements, and other elements may be present (e.g. "<a...>..." anchor tags):

```
<doc id="{doc_id_string}">
<headline>
...
</headline>
<post ...>
...
<quote ...>
...
</quote>
...
</post>
...
</doc>
```

All provenance/justifications for all KBP 2016 tasks must be drawn from the documents in the TAC KBP 2016 Evaluation Source Corpus. 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. Thus, offsets for provenance are counted *before* XML tags are removed. Start offsets 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 string is endoffset – startoffset + 1).

All KBP 2016 systems should return extractions from anywhere in the document, including <quote> regions of MPDF documents. However, for the following KBP tasks, in which evaluation is by comparison with gold standard Rich ERE annotations (which will not include annotations of <quote> regions), the track coordinator will automatically filter out <quote> regions from submitted runs before scoring, so as to avoid penalizing runs that include <quote> regions:

(a) EDL
(b) Belief and Sentiment
(c) Event Nuggets
(d) Event Arguments

For the following KBP tasks, in which evaluation is by assessment, assessment and scoring will allow provenance and extractions from anywhere in the document, including <quote> regions:

(a) Cold Start SF
(b) Cold Start KB Construction

Evaluation Queries

CSKB and CSSF systems are evaluated by the same set of Cold Start evaluation queries. A Cold Start evaluation query begins with one or more mentions of the same entity, followed by a sequence of slots to be filled for the entity. Each mention in the query is called an *entry point* because it can be used to select (at most) one entity node in a KB that is being evaluated; multiple entry points are included for each Cold Start evaluation query in order to increase the chances that the KB will have a response to the query even if it misses one entry point. Each Cold Start evaluation query is split into multiple Cold Start Slot Filling (CSSF) queries, with one entry point per CSSF query (the CSSF queries will request the same slots, but each will have a different entry point).

Participants in the Slot Filling variant of Cold Start will receive the CSSF evaluation queries at the beginning of the CSSF evaluation window, and will apply a script to incrementally convert those queries to a form that looks similar to queries from the 2014 English Slot Filling task. Participants in the Knowledge Base variant will not receive the queries; rather, NIST will apply the evaluation queries to each submitted knowledge base and assess the results. An outline of the NIST assessment process for both Cold Start variants is given below.

All CSSF evaluation queries start with an *entry point* into the knowledge base being evaluated. The entry point is defined by a named entity mention (name, docid, begin offset, and end offset), and is followed by the entity type and either one or two slots to be extracted for the entity.

Evaluation queries could take one of two forms: single-hop or multiple-hop. For example, here is a sample single-hop CSSF evaluation query that asks "What is the age of the *June McCarthy* mentioned at offsets 16931-16943 in Document 42?":

```
<query id="CSSF16_ENG_00243754cd">
  <name>June McCarthy</name>
  <docid>42</docid>
  <beg>16931</beg>
  <end>16943</end>
  <enttype>PER</enttype>
  <slot>per:age</slot>
  <slot0>per:age</slot0>
</query>
```

This single-hop query looks very much like a query from the 2014 English Slot Filling task, except that each query in Cold Start asks for a specific slot, rather than all slots for which there is information in the document collection. [6]

A more complex "two-hop" query might ask, "What are the ages of the children of the *June McCarthy* mentioned at offsets 16931-16943 in Document 42":

```
<query id="CSSF16_ENG_002109743e">
  <name>June McCarthy</name>
  <docid>42</docid>
  <beg>16931</beg>
  <end>16943</end>
```

[6] Participants in the Slot Filling variant should treat all other slots as if they appear in the <ignore> field of a Slot Filling query from 2013 or earlier.

```
    <enttype>PER</enttype>
    <slot>per:children</slot>
    <slot0>per:children</slot0>
    <slot1>per:age</slot1>
  </query>
```

In general, two-hop queries will start from an entry point (selecting the corresponding KB entity of a CSKB submission), follow a single entity-valued relation (from Table 1), then ask for a single slot value (from either Table 1 or Table 2). [7] Such queries will verify that the knowledge base is well-formed in a way that goes beyond basic entity linking and slot filling, without allowing combinations of errors to drive scores to zero.

Because two-hop queries do not look like any slot filling queries from KBP 2009-2014, participants in the Cold Start Slot Filling variant must process the CSSF queries in two "rounds" using the `CS-GenerateCSQueries.pl` script from NIST, which adds the `<slot>` entry to the NIST-distributed CSSF queries. Participants in the Slot Filling variant must treat `<slot>` as the slot to be filled. During the first round, `<slot>` will be identical to `<slot0>`. The `CS-GenerateCSQueries.pl` script will then convert a first round output file to a second round query file. Second round queries generated by this script will bear `<slot>` entries equivalent to `<slot1>`. Though some of the CSSF queries will differ only in having different mentions (possibly for the same entity) as their entry points, participating CSSF systems are prohibited from using information about one query to inform the processing of another query.

For the Knowledge Base variant, the following rules are applied to map from a CSSF evaluation query to a knowledge base entry: First, form a candidate set of all KB node mentions that have at least one character in common with the evaluation query mention and that have the same type. If this set is empty, the submission does not contain any answers for the evaluation query. Otherwise, for each mention K in the candidate set, calculate:

- COMMON, the number of characters in K that are also in the query mention Q.
- K_ONLY, the number of characters in K that are not in Q.

Execute each the following eliminations until the candidate set is size one, and select that candidate as the KB node that matches the query:

- Eliminate any candidate that does not have the maximal value of COMMON
- Eliminate any candidate that does not have the minimal value of K_ONLY
- Eliminate all but the candidate that appears first in the submission file

The proper specification of `mention` relations in a KB is therefore important for scoring well; CSKB participants should therefore take care to ensure that every named entity mention in the evaluation collection serves as a `mention` relation for a node in the KB.

The NIST evaluation of a KB will proceed by finding all entries in the KB that fulfill an evaluation query. For example, if the evaluation query 'schools attended by the siblings of *Bart Simpson*' finds two siblings for the node specified by the entry point, and the KB indicates that those siblings attended two and one schools respectively, then three results would be assessed by NIST. These

[7] In principle, multiple-hop queries could include more than two relations, but we limit ourselves to two in Cold Start 2016.

results will be converted to the same form as the output for the Slot Filling variant. Results will be pooled across all CSKB and CSSF submissions, and assessors will judge the validity of each result. Finally, a scoring script will report a variety of statistics for each submitted run.

In creating evaluation queries, LDC will strive to balance even distribution across slot types with productivity of those slots. Single hop queries, which are of greater interest for slot filling, will in many cases ask for multiple slots for a given entity regardless of whether fillers for those slots are attested in the document collection. Multiple hop queries will favor entities and slot sequences that are attested in the document collection (although here too, availability of answers is not guaranteed at any hop level).

Task Output – Knowledge Base Variant

CSKB systems must produce a knowledge base as output. The first line of the output file must contain a unique run ID. The remainder of the KB is represented as a set of augmented triples. Assertions will appear, one-per-line, in tab-separated format. The output file will be automatically converted to RDF statements during evaluation. All output must be encoded in UTF-8.

Each triple appears in the output file in subject-predicate-object order. For example, to indicate that entity-4 has entity-7 as a sibling, the triple might be:

```
:e4    per:siblings  :e7
```

If entity-4 has siblings in addition to entity-7, these relations should be entered as separate triples.

Entities

Each entity specification begins with a colon, followed by a sequence of letters, digits and underscores. Examples of legal entity specifications include `:Entity42`, `:EE74_R29`, and `:there_were_two_muffins_in_the_oven`. No meaning is ascribed to this sequence by the evaluation software; it is used only as a unique identifier. Any subsequent use of the same colon-preceded sequence will be taken as a reference to the same entity.

Predicates

The legal predicates are those shown in Table 1 (including inverses) and Table 2, plus `type`, `mention`, `nominal_mention`, and `canonical_mention`.

Predicates found in Table 1 must have entity specifications in both the subject and object positions. Predicates found in Table 2 must have an entity specification in the subject position, and a double quote-delimited string in the object position; the string in the object position will exactly correspond with the slot fill for that relation in the Slot Filling task. A backslash character must precede any occurrence of a double quote or a backslash in such a string.[8] At least one instance of each unique subject-predicate-object triple will be evaluated. If more than one instance of a given triple appears in the output (with each triple having different provenance), LDC will assess the instance with the highest confidence value (see below), and will assess additional instances if resources allow. If more than one such triple shares the same confidence value, the triple that appears earlier in the output will be considered to have higher confidence.

[8] Each backslash used to quote the following character doesn't itself have to be preceded by another backslash.

`type`

Each entity will be the subject of exactly one `type` triple. The object of that triple will be either `PER`, `ORG`, `GPE`, `FAC` or `LOC` depending on the type of the entity. It is up to submitting systems to correctly identify and report the type of each entity.

`mention` and `nominal_mention`

Each entity will be the subject of one[9] or more `mention` or `nominal_mention` triples. Together with the provenance information (see below), these triples indicate how the knowledge base is tied to the document collection. Each named entity mention in the collection must be submitted as the object of a `mention` triple, while each nominal entity mention in the collection must be submitted as the object of a `nominal_mention` triple. For example, if an entity is mentioned by name five times in a document, five `mention` triples should be generated. The object of a `mention` or `nominal_mention` triple is the double-quoted mention string; document ID and offset appear under provenance information (see below). The definition of what constitutes a named or nominal entity mention for Cold Start is described in the Cold Start schema above.

`canonical_mention`

For each document that mentions an entity, one of the `mentions` or `nominal_mentions` must be identified as the *canonical mention* for that entity in that document; it is the string that will be seen by the assessor if that entity appears as a slot fill, supported by that document (in Slot Filling task terms, it is the content of Column 5 of a CSSF 2016 submission, and its provenance will serve as Column 7 of the CSSF submission).[10] Canonical mentions are expressed using a `canonical_mention` triple. The arguments for `canonical_mention` are the same as for `mention` and `nominal_mention`. Note that there is no requirement that submissions select a single, global canonical mention for an entity. While such a mention might be useful, here we require that a canonical mention be provided within each *document* for the assessor to use during assessment. Each `canonical_mention` is also a `mention or nominal_mention`. As a convenience, if a submitted KB does not contain a `mention` or `nominal_mention` triple for each `canonical_mention` triple, the missing relations will be inferred (perhaps incorrectly) as named `mentions` (albeit with a warning). This shortcut is provided to make submitted KBs easier to view, and does not relieve submitters from the requirement to provide each of the required `mentions`, `nominal_mentions`, and `canonical_mentions`.

---

[9] Unlike previous years, Cold Start 2016 requires both named and nominal entity mentions to be extracted and included in the KB.

[10] In the Slot Filling task of KBP 2009-2014 (and in the Slot Filling variant of Cold Start), all slot fills are strings. Assessors verify the validity of a slot fill by looking for that string in the specified document, using the provenance information provided in the system response. In a submitted KB, slots that are filled with entities hold not strings, but pointers to the KB structure for the appropriate entity. Thus, a canonical mention must be identified by the Cold Start KB for each entity in each document, so that the assessor can be presented with a string that represents the entity during assessment. A relation provenance (see below) entry may include more than one document, and at least one of those documents must contain a mention of the object of the relation; that document must therefore contain a canonical mention for the object. When selecting a canonical mention for presentation to the assessor, the first document appearing in the relation provenance that contains a mention of the object will be used for the canonical mention.

Task Output – Slot Filling Variant

Output for the Slot Filling variant will be in the form of a tab-separated file. The columns of the submitted file are as follows:

Column 1	Query ID. For the first round, this is taken directly from the <query> XML tag. For the second round, this is drawn from the <query> tag of the query generated from the first round output.
Column 2	The name of the slot being filled.
Column 3	A unique run ID for the submission.
Column 4	Provenance for the relation between the query entity and slot filler, consisting of up to 4 docid:startoffset-endoffset triples separated by commas. Individual spans may comprise at most 150 UTF-8 characters. Unlike the 2014 Slot Filling task, there is no requirement to generate NIL entries when no information about the target entity is available.
Column 5	A slot filler (possibly normalized, e.g., for dates). This is used both to populate the <name> entry of the next round query, and by the assessor to judge the slot fill. The string should be *extracted* from the filler provenance in Column 7, except that any embedded tabs or newline characters should be converted to a space character and dates must be normalized (therefore, slot fillers should *not* be translated across languages). If a nominal mention is returned as a slot filler, only the head word of the nominal phrase should be returned (consistent with the EDL definition of nominal mentions). For dates, systems must 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.
Column 6	A filler type, selected from {PER, ORG, GPE, STRING}. The STRING filler is used for string-valued slots shown in Table 2.
Column 7	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 comma-separated docid:startoffset-endoffset spans for the base strings that were used to generate the normalized slot filler string. The documents used for the slot filler string provenance must be a subset of the documents provided in Column 4. This column serves two purposes. First, LDC will judge Correct vs. Inexact with respect to the document(s) provided in the slot filler string provenance. Second, this column is used to fill the <docid>, <beg> and <end> entries in second round queries. If more than one provenance triple is provided here, the first one will be used to fill the second round query.

Column 8	Confidence score.

The process for constructing a Slot Filling variant submission is as follows:

- Download the following from the NIST Web site:
  - The Cold Start evaluation documents`CS-GenerateQueries.pl` script
  - `CS-PackageOutput.pl` script
  - `CS-ValidateSF.pl` script
- Send an email to tac-manager@nist.gov to request the following:
  - The CSSF evaluation queries
- Configure your system to produce results only from the Cold Start evaluation documents.
- Run the `CS-GenerateQueries.pl` script on the evaluation queries to produce the first round queries your system will run on. Note that the raw evaluation queries might differ from the format given above, so you should not assume that you can use them as input to your system without running this script.
- Run your system, producing a slot-filling submission for the first round queries.
- Run the `CS-ValidateSF.pl` script on your first round output to verify that it is formatted correctly.
- Run the `CS-GenerateQueries.pl` script on the evaluation queries and your first round output to produce the second round queries.
- Run your system on the second round queries to produce a second output file.
- Run the `CS-PackageOutput.pl` script on the two output files to produce your submission.
- Run the `CS-ValidateSF.pl` script on your submission to verify that it is formatted correctly.
- Upload the submission to NIST.

Task Output – All Variants

Provenance

Each triple in CSKB submissions and each output line in CSSF submissions will include a set of augmentations (again using tabs as separators). Except for the `type` predicate (which does not require explicit support from a document) the first augmentations will describe the provenance of the assertion. Provenance for submissions for the Slot Filling variant have already been described above; corresponding provenance for triples in KB variant submissions are detailed here:

For predicates for relations from Table 1 or Table 2, up to four comma-separated justifications will be allowed for each entry, at the submitter's discretion. Justifications do not need to be explicitly associated with subject, relation or object. Each justification will include a document ID, followed by a colon, followed by two dash-separated offsets (begin and end offsets). The offsets that show the provenance of an extracted relation are used to narrow the assessor's focus within the documents when assessing the correctness of that relation. Provenance for a single relation may be drawn from more than one document. For the KB variant, when selecting a canonical mention for presentation to the assessor, the first document appearing in the relation provenance that contains a named or nominal mention of the object will be used for the canonical mention. (At least one of the documents in the KB's relation provenance must contain a named or nominal mention of the object of the relation; that document must therefore contain a canonical mention for the object.) Therefore, participants should be careful to ensure that if some documents contain nominal canonical mentions, and some documents contain named canonical mentions, that the document

containing a named canonical mention appears as the first document in the provenance. String-valued slots (from Table 2) whose values do not represent entities, place an additional constraint on provenance for Knowledge Base variant participants: the first justification must represent the document ID and offsets of the string fill. (Slot Filling variant participants are already providing this information in Column 7 of their submissions.) This requirement will allow assessors to quickly see the text from which the string fill was extracted.

Unlike entries for Slot Filling relations, the `mention`, `nominal_mention`, and `canonical_mention` predicates will have only a single justification, representing the exact location of the mention in the text. The `type` predicate requires no provenance.

Confidence Measure

To promote research into probabilistic knowledge bases and confidence estimation, each triple or slot fill may have an associated confidence score. Confidence scores will not be used for any official TAC 2016 measure. However, the scoring system may produce additional measures if confidence scores are included. Confidence scores will be used to induce a total order over the facts being evaluated (ties are broken when two scores are equal by assuming that the assertion appearing earlier in the submission has a higher score). Any submitted confidence score must be a positive real number between 0.0 (exclusive, 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. Confidence scores, if present, will appear at the end of each output line, separated from the provenance information with a tab. 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.

Comments

Output files may contain comments, which begin at any occurrence of a pound sign (#) and continue through (but do not include) the end of the line. Comments and blank lines will be ignored. The first line of a KB variant output file must contain the unique run ID (i.e., it may not be blank). Submitters may like to add a comment to this line giving further details about the run.

Examples

The following five lines from a Knowledge Base variant submission[11] show examples of: one triple without any augmentations, two with only mention extent, one with only relation provenance, and one with both relation provenance and confidence.

```
:e4    type          PER
:e4    mention       “Bart Simpson”      Doc726:37-48
:e4    nominal_mention     “brother”     Doc726:15-21
:e4    per:siblings  :e7    Doc124:283-288,Doc885:173-179,Doc885:274-281
:e4    per:age       "10"   Doc124:180-181,Doc885:173-179     0.9
```

Here are example lines from a Slot Filling variant submission:

[11] The first three lines can readily be converted to form part of an EDL submission, which can be evaluated as in the EDL track.

```
Q4 org:city_of_headquarters myrun1 Doc42:3-8,Doc8:3-11 Baltimore GPE Doc8:3-11 1.0
Q5 per:siblings myrun1 Doc124:283-288,Doc885:173-179 Lisa PER Doc124:283-286  0.7
Q6 per:age myrun1 Doc124:180-181,Doc885:173-179 10 STRING Doc124:180-181 0.9
```

Differences between 2014 Slot Filling and the 2016 Cold Start Slot Filling Variant

Slot filling systems that participated in the 2014 Slot Filling task will need to handle the following differences to successfully participate in the 2016 CSSF task:

- Only the slot specified by the `<slot>` entry is to be filled; all other slots should be ignored. The `<slot>` entry is added to the queries received from NIST by running the `CS-GenerateQueries.pl` script.
- Participants will need to do one round of slot filling, run the `CS-GenerateQueries.pl` script to create the second round queries, then run slot filling again on the new queries. The results of rounds one and two are to be concatenated before submission using the `CS-PackageOutput.pl` script.
- CSSF requires that participants be able to fill all slots in both directions. For example, the 2014 Slot Filling task required detection of the `per:cities_of_residence` slot. CSSF also requires systems to be able to detect the inverse of that slot, `gpe:residents_of_city`.
- Each slot filler must be assigned a type, selected from {PER, ORG, GPE, STRING}. This field represents an additional output column not found in the 2014 Slot Filling or CSSF tasks.
- NIL entries, indicating that no information about a particular slot is available, are not required in CSSF.
- Nominal mentions of slot fillers may be return if no named entity mention is available in the document collection. (Returning nominal entity mentions is not required, but may improve system recall if done correctly.)
- In addition to English, slot fillers and provenance may also be returned from Chinese and Spanish documents (only if the team is participating in one of the language conditions that isn't mono-lingual English).

Evaluation

The primary evaluation for both Cold Start SF and Cold Start KB construction is the slot filling evaluation, based on assessment of slot fillers found in response to Cold Start evaluation queries. In addition, the entity discovery component of Cold Start KBs is secondarily evaluated using the same set of evaluation documents and annotations as in the EDL track.

Slot Filling Assessment

Cold Start 2015 assessment and scoring will proceed as follows: The responses for each evaluation query (from both task variants and from human-generated results) will be pooled, and each response will be assessed by a person. The result of following the first relation will be assessed as if it were a Slot Filling query (for Knowledge Base variant entries, the canonical mention of the object entity in the first supporting document that mentions that entity will be used for the slot fill). The second relation in the query will also be assessed as a Slot Filling query, but only if the fill for the first relation is correct. ***If the fill for the first relation is not correct, each fill for the second relation is automatically counted as Wrong.*** For example, if the query asks for the ages of the siblings of "Bart Simpson," and the submitted knowledge base gives "Lisa age 8" and "Milhouse age

10" as siblings, then only the reported age of Lisa will be assessed (Milhouse is not Bart's sibling), and the reported age of Millhouse will automatically be counted as Wrong.

Cold Start uses *pseudo-slot* scoring to evaluate multiple-hop queries, in which each evaluation query is treated as if it selects a single indivisible slot. For example, an evaluation query that asks for the children of the siblings of an entity will be scored as if it were a query about a virtual `per:nieces_and_nephews` slot.[12] The guidelines in *TAC KBP 2015 Slot Descriptions* specify whether each of the component slots of a pseudo-slot is single-valued (*e.g.,* `per:date_of_birth`) or list-valued (*e.g.,* `per:employee_of`, `per:children`). A pseudo slot is single-valued if each of its component slots is single-valued, and list-valued otherwise. In contrast to the Slot Filling task, Knowledge Base variant submissions may contain multiple fills for single-valued slots. If such are present in the submission, LDC will assess the slot fill with the highest confidence value, and will assess additional slot fills if resources allow. If more than one such slot fill shares the same confidence value, the slot fill that appears earlier in the output will be considered to have higher confidence.

Each CSSF slot filler response (or CSKB object of each component relation that makes up a single evaluation query response) is assessed as Correct, ineXact, or Wrong, following guidelines in *TAC KBP 2015 Assessment Guidelines*. For each query, all system responses in which the slot filler is assessed as Correct or ineXact will be partitioned into equivalence classes, where slot fillers in the same equivalence class represent the same entity or value (as in the case of dates). Each Correct or ineXact response will receive an annotation for filler mention type (either NAM or NOM), and each equivalence class will receive an annotation for equivalence class mention type (NAM if the assessor can find a named mention for the filler anywhere in the provenances in any of the responses; otherwise, NOM if only nominal mentions appear in the provenances of all responses).

Pseudo-slots will be scored just as slots in the Slot Filling task, with the additional constraint that both the slot fill and the path leading to that fill must be correct for the entirety to be judged correct. To receive credit for identifying Maggie Simpson as Patty Bouvier's niece, the knowledge base must not only include Maggie as the slot fill, but must also represent Maggie as Marge's child, and Marge as Patty's sibling:[13]

Evaluation query:	Nieces and nephews of Patty Bouvier (`per:siblings`, `per:children`)
Ground Truth:	`:PattyBouvier per:siblings :MargeSimpson`
	`:MargeSimpson per:children :MaggieSimpson`
Submission:	`:PattyBouvier per:siblings :MargeSimpson`
	`:MargeSimpson per:children :MaggieSimpson` ⇒ **correct**

A KB that indicated that Maggie was Patty's niece because she was Patty's sister Selma's child would be scored as incorrect:

Evaluation query:	Nieces and nephews of Patty Bouvier (`per:siblings`, `per:children`)
Ground Truth:	`:PattyBouvier per:siblings :MargeSimpson`
	`:MargeSimpson per:children :MaggieSimpson`
Submission:	`:PattyBouvier per:siblings :SelmaBouvier`
	`:SelmaBouvier per:children :MaggieSimpson` ⇒ **incorrect**

[12] A pseudo-slot is similar to the concept of a *role chain,* which is supported by some knowledge representation systems based on description logic, including OWL 2.

[13] In each of these examples, only the subject, predicate and object are shown, and only a subset of the relevant knowledge base is presented. Each entity is named after the mention that gave rise to it.

A response is inexact if it either includes only a part of the correct answer or includes the correct answer plus extraneous material. Inexact answers are counted as Wrong for the purposes of scoring:

Evaluation query:	Titles of parents of Bart Simpson (`per:parents`, `per:title`)
Ground Truth:	`:BartSimpson per:parents :HomerSimpson` `:HomerSimpson per:title "Attack-dog trainer"`
Submission:	`:BartSimpson per:parents :HomerSimpson` `:HomerSimpson per:title "dog trainer Pitiless Pup"` ⇒ **inexact**

In addition, the object of the *final* relation in a pseudo-slot may be rated as redundant if it is equivalent to another fill for the pseudo-slot. Redundant answers are counted as Wrong for the purposes of scoring:

Evaluation query:	Nieces and nephews of Patty Bouvier (`per:siblings`, `per:children`)

Ground Truth:	`:PattyBouvier per:siblings :MargeSimpson`
	`:MargeSimpson per:children :MaggieSimpson`
	`:MaggieSimpson per:alternate_names "Margaret Simpson"`

Submission:	`:PattyBouvier per:siblings :MargeSimpson` `:MargeSimpson per:children :MaggieSimpson` ⇒ **correct** `:MargeSimpson per:children :MargaretSimpson` ⇒ **redundant**

However, objects of relations other than the final relation will never be rated as redundant:

Evaluation query:	Nieces and nephews of Patty Bouvier (`per:siblings`, `per:children`)

Ground Truth:	`:PattyBouvier per:siblings :MargeSimpson`
	`:MargeSimpson per:children :LisaSimpson`
	`:MargeSimpson per:children :BartSimpson`
	`:MargeSimpson per:alternate_names "Marjorie Simpson"`

Submission:	`:PattyBouvier per:siblings :MargeSimpson`
	`:PattyBouvier per:siblings :MarjorieSimpson`
	`:MargeSimpson per:children :LisaSimpson` ⇒ **correct**
	`:MarjorieSimpson per:children :BartSimpson` ⇒ **correct**

Here, Marge Simpson and Marjorie Simpson represent the same person in the ground truth, but two distinct entities in the KB. However, because the query is about Marge's children and not about Marge herself, both responses to the evaluation query are assessed as correct.

Since in Cold Start the facts being evaluated come from sequences of triples, confidence scores would need to be combined if we wanted to generate confidence scores for a derived pseudo-relation. The proper way to combine scores of course depends on the meaning of those scores, and for now, Cold Start is not mandating any particular meaning. Three general score combination functions are min, max and product; we welcome comments from the community on which combination methods to report.

Slot Filling Scoring

Given the above approach to assessment, basic scoring for a given system proceeds as follows:

- Each response assessed as Wrong or ineXact, is counted as *Spurious*
- Each response for Round 2 whose Round 1 parent filler is assessed as Wrong or ineXact, is counted as *Spurious*
- Responses assessed as Correct are grouped into equivalence classes. For each equivalence class, at most one response from the system is counted as *Right*; all other responses are counted as *Spurious* (therefore, systems should not return redundant answers to the same

query). If the system has a NAM entity mention in the equivalence class, or if the system has only NOM entity mentions and the equivalence class is annotated as NOM, then the one response is counted as *Right*; otherwise, if the system has only NOM entity mentions in the equivalence class and the equivalence class is annotated as NAM, then the one response is counted as *Ignore* (i.e., treated as if it was never returned by the system). Thus, named entity mentions are preferred.

- **Reference** = number of single-valued pseudo-slots with a correct response + number of equivalence classes[14] for all list-valued pseudo-slots
- **Recall** = #Right / Reference
- **Precision** = #Right / (#Right + #Spurious)
- $\mathbf{F_1}$ = 2 * Precision * Recall / (Precision + Recall)

As in 2015, each Cold Start evaluation query in 2016 may have more than one entry point. Because the number of entry points may differ arbitrarily between Cold Start evaluation queries, we focus on two primary metrics for the 2016 Cold Start Knowledge Base Population system evaluation:

- MAX (micro-average): compute F1 for each entry point as outlined above to select a single "maximal" entry point for each evaluation query, where the selected entry point has a maximal F1 among all entry points for that query. The MAX micro-average Precision, Recall, and F1 for the system is computed by summing the counts across all queries, using only the selected maximal entry point for each query.
- MEAN (macro-average): compute F1 for each entry point as outlined above. The query-level score for a query is the mean of the F1 scores of each of its constituent entry points. The MEAN score for the system is the mean of its query-level scores. The MEAN metric gives equal weight to each query, and (within each query) equal weight to each of its entry points.

Entity Discovery Scoring

The scoring for the Entity Discovery component of submitted Cold Start KBs will be identical to scoring for the 2016 TAC Trilingual Entity Discovery and Linking task, with the exception that no linking to an existing knowledge base is required (that is, all mentions will be treated as NIL entries). Please see *TAC KBP2016 Entity Discovery and Linking Task Description* for complete details on scoring.

Submissions

A four-week window from Monday August 1 to Monday August 29 will be available for downloading the TAC KBP 2016 Evaluation Source Corpus, producing CSSF and CSKB system output, and submitting results. Systems should not be modified once the corpus has been downloaded. Starting Monday, August 15, participants in the CSSF task may email NIST to request the CSSF evaluation queries, but teams participating in both the CSSF and CSKB tasks must submit all CSKB runs before requesting the CSSF evaluation queries from NIST. On August 15, automatic EDL output from systems participating in the first EDL evaluation window, will also be made available as an optional resource to Cold Start participants.

[14] See *TAC KBP 2015 Slot Descriptions* and *TAC KBP 2015 Assessment Guidelines* for further information on how and when two slot fills are treated as equivalent.

For each of the Cold Start task variants (CSSF and CSKB), a team may submit up to 5 runs for each of the following 4 language conditions:

1. Monolingual English: entity mentions, slot fills and provenances are extracted only from English documents
2. Monolingual Spanish: entity mentions, slot fills and provenances are extracted only from Spanish documents
3. Monolingual Chinese: entity mentions, slot fills and provenances are extracted only from Chinese documents
4. Cross-lingual: entity mentions, slot fills and provenances are extracted from any combination of English, Spanish, and Chinese documents.

If a team submits a run involving more than one language under the Cross-lingual condition, it must also submit at least one run under the monolingual condition for each language involved (with a description of which monolingual run configurations were used for each cross-lingual run).

Submitted runs must be ranked (1-5) in order of evaluation preference. The number of runs actually evaluated will depend upon resources available to NIST; the 3 top-ranked runs from each team will be assessed for each task and language condition, and lower-ranked submissions may be assessed if resources allow. The run ID included in each team's submission file must be a concatenation of the team's TAC KBP 2016 team ID, the task (KB or SF), the language condition (ENG, CMN, SPA, or XLING), and a rank (1-5); thus "Acme_KB_XLING_1" would be the top-ranked run for the Acme team for the CSKB task variant under the cross-lingual condition.

The top-ranked submission must be made as a 'closed' system; in particular, it must not access the Web during the evaluation period. All submissions must obey the following external resource restrictions:

- Structured knowledge bases (e.g., Wikipedia infoboxes, DBPedia, Freebase) may not be used to directly fill slots or directly validate candidate slot fillers.
- Structured knowledge base entries for target entities may not be edited, either during, or after the evaluation.

In addition, because Cold Start focuses on the condition where the knowledge base is initially empty, we ask that each participating site submit at least one run that consults external entity knowledge bases only after entities and relations have been extracted from the document collection. Details about submission procedures will be communicated to the track mailing list. Tools to validate formats will be available on the TAC Web site (http://www.nist.gov/tac/2016/KBP/ColdStart/tools.html).

Change History

- Version 1.0
    - Original version, based on the 2015 specification
    - Added description of multi-lingual tasks
    - Aligned definition of entity types and mention types in the KB Construction task, with those in the 2016 Entity Discovery and Linking track
    - Added description of nominal entity mentions and slot fillers