

Cold Start Knowledge Base Population at TAC 2012

Task Description¹

Version 1.0 of May 3, 2012

Introduction

Since 2009, the TAC Knowledge Base Population Track has evaluated performance on two important aspects of knowledge base population: entity linking and slot filling. However, the ability of a system to use these technologies to actually construct a knowledge base (KB) from the information provided in a text collection has not been exercised. The Cold Start task is designed to evaluate a system's ability to do just that. Participants will 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 will then be evaluated as a single connected resource.

The Entity Linking and Slot Filling tasks 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? The TAC Entity Linking task has 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 task has measured this.
- Are entity linking and slot filling correctly coordinated to produce a meaningful knowledge base? Cold Start 2012 will measure 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 2012 will not measure this, but is designed to facilitate this kind of evaluation 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 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. We will use the schema that is implicitly specified by the TAC 2012 Slot Filling task. Thus, the schema will include three entity types (person, organization and geopolitical entity) and forty-two relation types. For relations whose fills are themselves entities (such as `per:siblings` or `org:subsidiaries`), systems will be required to link that slot to the KB node representing the correct entity. Slots whose fills are strings (such as `per:title` or `org:website`) will continue to use strings to represent the information.

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 "at" nist.gov.

Cold Start also implies that the knowledge base is initially empty; in particular, we assume that a Wikipedia dump is *not* the starting point for the knowledge base. To avoid solutions that rely on verifying content already present in Wikipedia or other large data sources about entities, the document collection used in Cold Start will be dominated by entities that are not present in Wikipedia. Examples of suitable collections include:

- news articles collected from the Metro section of local newspapers for a medium-sized city;
- Web news documents from college or university news sources; and
- a subset of the TAC document collection whose documents focus on entities that are not in Wikipedia.

Each of these possibilities is under consideration for the Cold Start 2012 document collection.

Participating systems will receive three inputs:

1. a *knowledge base schema*;
2. a *document collection*; and
3. a set of *entry points*, which are particular entity mentions in the document collection.

From these, systems will produce a knowledge base. In Cold Start 2012, this KB will be submitted to NIST as a set of augmented triples; in future years, the entire KB (including inference capability) may be submitted and evaluated. Participating systems must tie each of the provided entry points to a particular KB entity node; in this way, the knowledge base can be queried without first aligning it to a reference knowledge base.

The submitted knowledge base will then be evaluated by NIST. Evaluation will use a set of KB evaluation queries. Each query will enter the knowledge base at one of the entry points, follow a sequence zero or more relations within the knowledge base, and end in a slot fill (most or all of the evaluation queries will follow exactly one relation before selecting a slot). The resulting slot fills will be assessed and scored in much the same way as is now done in the Slot Filling task. For example, a KB evaluation query might ask ‘what are the ages of the siblings of the *Bart Simpson*² 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.

For evaluation purposes, NIST will translate the KB schema into an OWL (McGuinness et al. 2004) ontology, and will translate each submitted KB into a collection of RDF (Lassila & Swick 1998) assertions using that ontology. The OWL schema axioms and the RDF assertions will together represent the evaluation KB. NIST will then translate each Cold Start evaluation query into a SPARQL query (Prud'Hommeaux & Seaborne 2008) that it will run against the RDF versions of the submitted KBs. The output from the SPARQL queries will then be assessed and scored.

Schema

² 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 twenty-five years, but may simply visit http://simpsons.wikia.com/wiki/Simpson_Family.

Relation	Inverse
per:children	per:parents
per:other_family	per:other_family
per:parents	per:children
per:siblings	per:siblings
per:spouse	per:spouse
per:employee_of	org:employees*
per:member_of	org:membership*
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:holds_shares_in*
org:founded_by	per:organizations_founded*
org:top_members_employees	per:top_member_employee_of*
org:member_of	org:members
org:members	org:member_of
org:parents	org: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 relations that are newly defined for Cold Start because they are not part of the current Slot Filling task definition. 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.

The schema for Cold Start 2012 is derived directly from the Slot Filling task specification. Slot Filling defines forty-two slots. Twenty-seven of these have fills that are themselves entities, as shown in Table 1.³ The remaining fifteen slots have string fills, as shown in Table 2. Each entity-valued slot will have an inverse. Some slots, such as per:siblings, are symmetric. Others, such as per:parents, have inverses that are already Slot Filling task slots (in this case, per:children). The remaining slots (e.g., org:founded_by) have no corresponding slot in the Slot Filling task; Cold Start specifies new slot names for these inverses. All inverse relations must be explicitly identified in the

³ Some of the slots from previous years have been renamed to support their use in RDF and SPARQL. Either the old names or the new will be accepted, although the new names are preferred. A mapping from old names to new names is given in Appendix A.

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 .

Document Collection

Details on the document collection will be forthcoming. Documents will be formatted in the same way as the documents for the other Knowledge Base Population tasks. Because the task can begin as soon as the document collection is available, the collection will not be distributed until the start of the task window. We anticipate a collection of approximately 10,000 to 15,000 documents.

Entry Points

Entry points will be provided in a format that looks like Entity Linking queries. That is, each entry point will have a mention string, a document ID, and an offset. The list of entry points will likely be extensive. The list will be created by running a named entity recognizer over a significant subset of the document collection, lightly curating the results, and including each string identified as a person, organization or GPE mention. Thus, the list of entry points may well contain errors—strings that are not actually person, organization or GPE mentions. It is not the intention of the Cold Start evaluation to test whether systems can correctly distinguish valid entry points from invalid ones. No invalid entry point will be used in a KB evaluation query (as all such queries will be human-generated). Nonetheless, systems may refuse to attach a knowledge base node to a given entry point if they feel the entry point is invalid. No penalty will be paid for doing so if the entry point is not used in a KB evaluation query.

Not every legitimate named entity mention will be an entry point. Systems are responsible for identifying and asserting in the knowledge base mentioned entities that are not entry points, as they may appear as part of the answer to an evaluation query (see below).

Task Output

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.

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.

Predicates

The legal predicates are those shown in Table 1 (including inverses) and Table 2, plus `type`, `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 slot, 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.⁴

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

Each entity will be the subject of one⁵ or more `mention` triples. Together with the provenance information (see below), these triples indicate how the knowledge base is tied to the document collection. Each valid entry point must appear in a `mention` triple for exactly one entity. In addition, any further named mentions of the entity in the collection that were not specified as entry points should also appear. The object of a `mention` triple is the double-quoted mention string; document ID and offset appear under provenance information (see below).

In the Slot Filling task, all slot fills are strings. Assessors verify the validity of a slot fill by looking for that string in the specified document (and in 2012 using the addition of offset information). In a submitted KB, slots that are filled with entities hold not strings, but pointers to the KB structure for the appropriate entity. During assessment, the assessor must be presented with a string that represents such an entity. Thus, for each document that mentions an entity, one of the `mentions` must be identified as the *canonical mention* for that document; it is the string that will be seen by the assessor when that entity appears as a slot fill, supported by that document. Canonical mentions are expressed using a `canonical_mention` triple. The arguments for `canonical_mention` are the same as for `mention`. Note that there is no requirement that submissions select a single, global canonical mention for an entity. While such a name might be useful (and is a part of the current Entity Linking KB), here we require only that a name be provided within each document for the

⁴ Each backslash used to quote the following character doesn't itself have to be preceded by another backslash.

⁵ While unmentioned but inferred entities may play a role in future TAC evaluations, Cold Start 2012 will work only with entities that have named mentions.

assessor to use. Each `canonical_mention` is also a `mention`. As a convenience, if a submitted KB does not contain a `mention` triple for each `canonical_mention` triple, the missing relations will be inferred. 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` and `canonical_mentions`.

Provenance

Each triple will be followed by a set of augmentations (again using tabs as separators). Except for the `type` slot (which does not require explicit support from a document) the first annotations will describe the provenance of the assertion. In 2012, provenance includes not just the document ID for the document that attests to the fact, but also now appropriate offsets. The offsets themselves will not be assessed and scored; rather, they are used only to allow the assessor to quickly see the points in the document where the fact is attested. Each offset pair must be the character positions of the first and last characters of the word or words that best conveys the entity or predicate. The Cold Start task will follow exactly the Slot Filling task rules for how offset pairs are to be specified.

The Slot Filling task requires two offset pairs for each slot fill, one for the position in the document that attests to the relation (i.e., the predicate), and the other for the point in the document that holds the extracted value (i.e., the object). Because the Cold Start task may view a relation in either direction, it requires three offset pairs (subject, relation and object). Thus, provenance for a triple is expressed as seven tab-separated values: the document ID,⁶ the start and end offset for the subject, the start and end offset for the predicate, and the start and end offset for the object. These seven tab-separated values follow the subject-predicate-object values, separated from them by a tab. The `mention` and `canonical_mention` slots will have only a single offset pair, representing the start and end of the mention in the text.

Confidence Measure

To promote research into probabilistic knowledge bases and confidence estimation, each triple may 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 if confidence scores are included. For these measures, confidence scores will be used to induce a total order over the facts being evaluated (how ties are broken when two scores are equal has not yet been decided). Since in Cold Start the facts being evaluated come from sequences of triples, confidence scores will need to be combined. 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. Any submitted 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. Confidence scores, if present, will appear at the end of each output line, separated from the provenance information with a tab. 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.

⁶ TAC 2012 will operate only over relations that are fully attested within a single document.

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 an 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 three lines show examples of a triple without any annotations, one with only provenance annotation, and one with both provenance and confidence annotations.

```
:e4    type          PER
:e4    per:siblings  :e7    D00124  173  179  274  281  283  288
:e4    per:age       "10"   D00124  173  179  182  191  180  181  0.9
```

Evaluation

Evaluation Queries

Each submitted knowledge base will be evaluated by running a set of evaluation queries against it and assessing the results. Participants need not concern themselves with the details of how these queries are expressed and applied to their submitted knowledge bases, as all such work will be performed by NIST. An outline of the NIST process is given here. First, the augmented triples will be converted by NIST to a collection of RDF triples, which will be loaded into an RDF triple store along with the Cold Start OWL ontology. The resulting knowledge base will field the set of SPARQL queries. The results will be converted to a form similar to the results of the Slot Filling task. Results will be pooled across all submissions, and assessors will judge the validity of each result. Finally, a scoring script will report a variety of statistics for each submitted run.

The evaluation queries could take many forms. For example, a query that asked for slot fills for an entity mentioned in a particular document would look very much like the Slot Filling task, while a query that asked for all of the mentions of each of a set of entities would look like the Entity Linking task. For Cold Start, most of the SPARQL queries will start from an entry point, select the corresponding KB entity, follow a single entity-valued relation (from Table 1), then ask for a single slot value (from either Table 1 or Table 2). For example, a SPARQL query corresponding to the question ‘what are the ages of the siblings of the *Bart Simpson* mentioned in Document 42?’ would be of this form. Such “one-hop” queries will verify that the knowledge base is well-formed in a way that goes beyond basic entity linking and slot filling, without (we hope) allowing combinations of errors to drive scores to zero. Note that unlike the Slot Filling task, each SPARQL query will ask for a specific slot, not all slots for which there is information in the document collection.

Evaluation queries and the answers to them produced from the submitted knowledge bases will be made available to participants, but only after the evaluation is concluded.

Assessment

Cold Start 2012 assessment and scoring will be similar to Slot Filling assessment and scoring. The results for each assessment query will be pooled, and each response will be assessed by a person. For one-hop queries, the result of following the first relation will be assessed as if it were a Slot Filling query (where the canonical name of the object entity in the supporting document 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. 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).

TAC 2012 will use *pseudo-slot* scoring, 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.⁷ The Slot Filling guidelines 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.

As with the Slot Filling task, the object of each component relation that makes up a single evaluation query response is rated as correct, inexact, or wrong. 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:⁸

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**

A response is inexact if it either includes only a part of the correct answer or includes the correct answer plus extraneous material. No credit is given for inexact answers:

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. No credit is given for redundant answers:

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"`

⁷ 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.

⁸ 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.

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.

Scoring

Given the above approach to assessment, scoring for a given evaluation query proceeds as follows:

Correct = total number of system output pseudo-slots judged correct

System = total number of system output pseudo-slots

Reference = number of single-valued pseudo-slots with a correct response + number of equivalence classes⁹ for all list-valued pseudo-slots

Recall = Correct / Reference

Precision = Correct / System

F = $2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$

The F score is the primary metric for the 2012 Cold Start Knowledge Base Population system evaluation.

Submissions

A two-week window will be available for downloading the Cold Start document collection, building the knowledge base, and submitting results. Systems should not be modified once the corpus has been downloaded. Participants may submit up to five knowledge bases, ranked in order of evaluation preference. 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 external resource restrictions in place for the TAC 2012 Slot Filling task. The number of submissions actually judged will depend upon resources available to NIST. Details about submission procedures will be communicated to the track mailing list. Tools to validate formats will be made available prior to the start of the evaluation period.

⁹ See the Slot Filling Task Guidelines for further information on how and when two slot fills are treated as equivalent.

Sample Collection

A sample Cold Start collection will be available from the NIST Web site (<http://www.nist.gov/tac/2012/KBP/>). Note that this is not a training collection; it serves only to illustrate the various facets of the task and the evaluation. The sample will include:

- A file describing the collection (README.txt).
- A knowledge base schema.
- A document collection, comprising seventeen documents drawn from the domain of *The Simpsons* television show. Each <DOC> tag includes the original Web source, of which the text in the collection is a snippet.
- A set of entry points for the collection.
- A reference KB for the collection. Note that a reference KB will not be created for the actual Cold Start task.
- A sample participant submission, including a variety of errors.
- A set of evaluation queries with ground truth answers for those queries. Note that participants *will not receive* evaluation queries prior to the results submission deadline; they are used only internally at NIST for evaluation. They are provided here to give a clearer picture of how the evaluation will proceed.

References

D. L. McGuinness, Frank Van Harmelen et. al, *OWL Web ontology language overview*, W3C recommendation, 2004.

Ora Lassila and Ralph R. Swick, *Resource description framework (RDF) model and syntax*, World Wide Web Consortium, 1998.

E. Prud'Hommeaux and A. Seaborne, *SPARQL query language for RDF*, World Wide Web Consortium, January 2008.

Appendix A

The following table shows the mapping from old to new slot names for slots whose names have changed since KBP 2011. Either version will be accepted in KBP 2012 submissions, but the new names are preferred.

Previous slot name	New (preferred) 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