

TREC Entity 2010 guidelines

(v1.1, 2010-07-06)

Introduction

The overall aim of the new TREC Entity track is to create a test collection for the evaluation of entity-related searches on Web data. Many user information needs would be better answered by specific entities instead of just any type of documents. An entity is a person, product, or organization with a homepage. The entity's homepage is considered the representative of that entity on the web.

The first edition the Entity track, in 2009, has featured a **related entity finding (REF)** task. This year, REF runs as the main task of the track. A number of changes has been made to the previous edition, these are highlighted in a separate subsection. In addition, the track introduces a second challenge, **entity list completion (ELC)**, which will run as a pilot task.

Main task: Related Entity Finding

The problem of *related entity finding* (REF) is defined as follows:

Given an **input entity**, by its name and homepage, the **type of the target entity**, as well as the **nature of their relation**, described in free text, **find related entities** that are of target type, standing in the required relation to the input entity.

The track defines **entities** as "typed search results," or "things," **represented by their primary homepages** on the Web (see the definition of the primary homepage below). Searching for entities thus corresponds to ranking these homepages. The track thereby investigates a problem quite similar to the QA list task, where answers are known to be entities (and never attribute values) and the underlying information need is of a more informational nature – entity finding is situated in explorative search tasks, instead of the more factual list question task of seeking "the" list of answers to a question. The two problems are closely related, but compare for example a typical list question "What are the capitals of Europe?" to an entity finding task of "Which are the cities with a local government that encourages bike-usage?".

Input

For each request (query) the following information is provided:

- Input entity, defined by its name and homepage
- Type of the target entity (person, organization, product, or location)
- Narrative (describing the nature of the relation in free text)

The track limits the target entity types to four: people, organizations, products, and locations.

NOTE: the input entity does not need to be limited to these four types.

Example topic

An example information need, “*find organizations that currently use Boeing 747 planes*” is formulated as follows:

```
<query>
  <num>7</num>
  <entity_name>Boeing 747</entity_name>
  <entity_URL>clueweb09-en0005-75-02292</entity_URL>
  <target_entity>organization</target_entity>
  <narrative>Airlines that currently use Boeing 747 planes.</narrative>
</query>
```

Output

- For each query, participants may return **up to 100 answers** (related entities).
- For each answer entity a single homepage must be returned; optionally, the name of the entity may also be returned.
(There are a number of criterion for what is to count as the homepage of an entity, see below.)
- Each query must have at least one entity retrieved for it.

Submission format

Each answer record must have the following format:

```
topicID Q0 docno rank score runID entityName
```

where

- The first column is the topic number.
- The second column should always be Q0.
- The third column is either the official document number (“docno” field of the document).
- The fourth column is the rank the entity is retrieved, and the fifth column shows the score (integer or floating point) that generated the ranking. This score **MUST** be in descending (non-increasing) order and is important to include so that we can handle tied scores (for a given run) in a uniform fashion (the evaluation routines rank from these scores, not from your ranks). If you want the precise ranking you submit to be evaluated, the **SCORES** must reflect that ranking.
- The sixth column is called the “run tag” and should be a unique identifier for your group **AND** for the method used. That is, each run should have a different tag that identifies the group and the method that produced the run. Run tags must contain 12 or fewer letters and numbers, with **NO** punctuation.
- The seventh column is optional, in which the entity's name may be returned. Entity names need to be normalized as follows:
 - Only the following characters are allowed: [a..z], [A..Z], [0..9], _
 - Accented letters need to be mapped to their plain ASCII equivalents (e.g., “á” => “a”, “ü” => “u”)
 - Spaces need to be replaced with “_”

Example output

```
7 Q0 clueweb09-en0003-03-28260 1 0.98 exampleRun Asiana_Airlines
7 Q0 clueweb09-en0001-07-19878 2 0.94 exampleRun Cargolux
...
```

Runs

Participating teams may submit up to **four runs**, at least one of which will be judged.

We accept both automatic and manual runs. Automatic runs must not involve human intervention at any stage. The retrieval system should not be modified between the time queries are downloaded and the time the runs are submitted. Manual runs, which can involve anything from manual query formulation to interactive searching and judging, are strongly encouraged.

Document collection

The 2010 track uses the English portion of the [ClueWeb09](#) collection (about 500 million pages).

Evaluation and assessments

NIST coordinates topic development and relevance assessments. For the 2010 edition of the track **50 new REF topics** will be created and assessed. The test topic set will also include the **20 topics from 2009**; if time and resources allow, these topics will also be reassessed, but will not be taken into account for the official ranking of systems.

Entity homepages

Real-world entities can be represented by multiple homepages; a clearly preferred one cannot always be given. We differentiate between *primary* and *relevant* homepages of a given entity:

- a primary homepage is devoted to and in control of the entity
- a relevant homepage is devoted to the entity, but is not in control of the entity

What is not a homepage. Unlike last year, the Wikipedia page of a given entity is non-relevant by definition. Pages that only mention the entity (but are not about the entity) are also considered non-relevant. News articles and blog posts, even if exclusively about the entity, are not considered as entity homepages.

Homepages for products. By definition, a product is the most specific object that has a separate page under its manufacturer's site. Accordingly, primary homepages for entities of type product must be from their manufacturer's site.

Examples

Example answers to the "Boeing 747" query.

- **Name:**
British_Airways

- **Primary:**
<http://ba.com>
- **Relevant:**
<http://www.guardian.co.uk/business/britishairways>
- **Non-Relevant:**
<http://www.justtheflight.co.uk/news/18339089-aer-lingus-announces-british-airways-codeshare.html>
<http://www.time.com/time/world/article/0,8599,1983547,00.html>
- **Name:**
Korean_Air
- **Primary:**
<http://www.koreanair.com>
- **Relevant:**
<http://www.alternativeairlines.com/korean-air>
<http://www.airlinequality.com/Forum/korean.htm>
- **Non-Relevant:**
<http://www.defense-aerospace.com/article-view/verbatim/90868/vision-and-strategy-of-the-korean-aerospace-industry.html>

Further examples

These are not (good) answers to the "Boeing 747" query, but might help distinguish primary/relevant/non-relevant pages.

Let's consider a person: Australian prime minister *Kevin Rudd*.

- **Primary:**
<http://pm.gov.au>
- **Relevant** (not under KR's control, but "homepages" about him):
<http://www.aph.gov.au/house/members/member.asp?id=83T>
<http://www.crikey.com.au/topic/kevin-rudd>
- **Non-Relevant:**
http://en.wikipedia.org/wiki/Kevin_Rudd [by definition; Wikipedia page]

And a product: *Windows 7*.

- **Primary:**
<http://www.microsoft.com/windows/windows-7/default.aspx>
- **Relevant** (not under Microsoft's control, but still "homepages"):
<http://www.cnet.com/windows-7>
<http://lifehacker.com/tag/windows-7>
- **Non-Relevant:**
<http://www.engadget.com/2009/08/12/windows-7-review>

Assessment procedure

The assessment procedure consists of two stages. In phase one, homepages are judged as primary or relevant. For primary homepages, the name (returned along with the homepage) is judged whether it is correct or not. Then, in phase two, homepages belonging to the same entity are grouped together.

The output of the assessments will therefore include a set of homepages and a set of names, that all refer to one entity; one or more of these homepages identified as primary, a set of homepages identified as relevant, and one of names identified as correct.

NOTE: assessors will only look at the textual parts of homepages. In particular, flash content will not be looked at.

Evaluation metrics

The following measures will be used:

- NDCG@R, where a primary homepage gets gain 3 and a relevant homepage gets gain 1 (note that we reward primary homepages more than last year)
- P@R and MAP, computed for relevance level 1 (both relevant and primary accepted) and 2 (only primary accepted)

For all metrics, only previously unseen entities will be rewarded; i.e., if a primary/relevant homepage has already been returned at earlier ranks for the same entity, then it will count as non-relevant.

Official evaluation results will be based on the homepage field only; alternative rankings of systems will also take entity names into account. I.e., accept an entity (homepage) as primary/relevant only if a correct name is also provided.

External resources

It is allowed to use any external resource, but it needs to be indicated on the submission form. Any data that is not part of the ClueWeb09 collection is considered as an external resource.

Key changes

The key changes introduced to the previous edition of the REF task are as follows:

- English subset of ClueWeb cat A
- Single record submission format
- No supporting documents
- New entity type: location
- Revised definition of primary and relevant homepages
- Wikipedia pages are not accepted
- Primary homepages are rewarded more
- Names are judged only for primary pages; the judgment is binary

Pilot task: Entity List Completion

This year we introduce a pilot task, meant to reach out to research groups active in the area that is often referred to as "semantic search", more specifically, semantic data search: the retrieval of semantic data. The motivation of the proposed task is very close to that of the main task, but instead of finding entities represented by their homepages on the Web, the task here is to find these entities in the Semantic Web, or, in other words, to perform entity search in the Linked Open Data (LOD) cloud.

The problem of the *Entity list completion* (ELC) task is defined as follows:

Given a **list of input entities**, represented by their URIs, complete the list with additional entities from a specific collection of Linked Open Data.

Input

We will use (most of) the 20 topics developed in the 2009 pilot run of the track. For each of these topics, the answer entities identified in the 2009 Entity Track will serve as the list of examples. These will be mapped to LOD by track organizers.

Topic definitions follow the same format as for the REF task, with the addition of known relevant entities, referred to as *examples*. Example entities are identified by one or more URIs, and are listed between `<examples>..</examples>`:

```
<query>
  <num>7</num>
  <entity_name>Boeing 747</entity_name>
  <entity_URL>clueweb09-en0005-75-02292</entity_URL>
  <target_entity>organization</target_entity>
  <narrative>Airlines that currently use Boeing 747 planes.</narrative>
  <examples>
    <entity>
      <URI>http://dbpedia.org/resource/KLM</URI>
      <URI>http://www.linkedin.com/companies/klm</URI>
      <URI>http://www.reference.com/browse/KLM</URI>
    </entity>
    <entity>
      <URI>http://dbpedia.org/resource/Northwest_Airlines</URI>
    </entity>
    ...
  </examples>
</query>
```

Note: The above topic is only meant as an illustrative example, it was not checked whether the provided URIs exist in the LOD crawl used by the track.

Output

The output is a ranked list of additional entities, defined by a URI, that would complete the list.

- For each query, participants may return **up to 100 answers** (entities).
- For each answer entity a single URI must be returned; optionally, the name of the entity may also be returned.
- Each query must have at least one entity retrieved for it.

Submission format

The submission format is similar to that of the REF task, but instead of ClueWeb document IDs, URIs are to be returned. The name field (7th column) is optional, see the REF section for the rules of formatting names.

```
topicID Q0 URI rank score runID entityName
```

Example

```
7 Q0 http://dbpedia.org/resource/Air_China 1 0.98 exampleRun Air China  
7 Q0 http://dbpedia.org/resource/Cargolux 2 0.94 exampleRun Cargolux  
...
```

Data set

To ease collection building and at the same time simplify participation by the target community, the track will use the Billion Triple Challenge 2009 dataset (<http://vmlion25.deri.ie/>). The same data collection has been used for the Semantic Search challenge posed by the Semantic Search workshop held at WWW 2010, so should be easy to process for those researchers we specifically organize the pilot task for.

Evaluation and assessments

Assessment procedure

Judging will be done by participants.

Entity resolution (i.e. the same entity represented under different URIs) will be done during the assessment phase. Qrels therefore will consist of a set of entities, each identified by one or more URIs (which are considered equivalent).

Evaluation metrics

Evaluation measures will be P@R and MAP, both reported over the residual collection (known relevant—i.e., example—entities left out).

Relevant entities previously seen in the ranked list will be considered irrelevant.