Overview of the TREC 2011 Entity Track

Krisztian Balog
NTNU, Trondheim, Norway
krisztian.balog@idi.ntnu.no

Pavel Serdyukov
Yandex, Russia
pavser@yandex-team.ru

Arjen P. de Vries
CWI, The Netherlands
arjen@acm.org

1   Introduction

The TREC Entity track aims to build test collections to evaluate entity-oriented search on Web data. The track works with two corpora: the ClueWeb 2009 web corpus and the new Sindice-2011 dataset [1].

Motivated by observations from the 2010 Entity track, we made the following changes in the track setup for 2011. In the REF task, we focused on modifications to simplify the evaluation and improve cross-system comparison. (1) Only primary homepages are accepted, i.e., relevance is binary; (2) For each answer, a (single) supporting document is required; (3) Target type is not limited anymore; (4) Groups that generate results using Web Search Engines are required to submit an obligatory run, using the Lemur ClueWeb Online Query Service.

The main change regarding the LOD task is the use of the Sindice-2011 corpus, an improved and larger Semantic Web crawl, replacing the BTC-2009 collection used in 2010. (1) The target corpus is a larger and more representative LOD crawl; (2) Examples are not mapped manually to LOD, but given as ClueWeb document identifiers.

Finally, we introduced a new pilot task REF-LOD to explore the differences between ranking web data and ranking semantic web data, possibly enabling deeper investigations into connecting semantic web data with the 'real' web. We basically repeat the REF task, but request results identified by their LOD URIs instead of their homepages. One of the goals of this task is to gain more insights in entity representation, and specifically how often entities that are not represented on the web with their own homepage are represented as entities in LOD.

In the remainder of the paper we first detail the setup of each task. We present the results collected in this year's track participation, and summarize the approaches applied (preliminary). The paper concludes with a brief discussion of the problems faced by the track, and the way forward.

2   Tasks

The third edition of the Entity track featured two main tasks and a pilot task; all are variations of the related entity finding problem, but differ in how the input is formulated (i.e., whether example entities are available), in the data collections used, and in the means of entity identification. The tasks are summarized in Table 1. Changes to last years edition are discussed in the description of the corresponding task.

Task	Examples	Entity identification	Collection(s)	
			ClueWeb09 (EN)	Sindice-2011
REF	N	URL	Y	N
REF-LOD	N	URI	Y	Y
ELC	Y	URI (URL for examples)	Opt.	Y

Table 1: Tasks at Entity 2011.

2.1 Main task 1: Related Entity Finding

The Related Entity Finding (REF) task 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.

2.1.1 Input

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

- Input entity, defined by its name and homepage (ClueWeb ID)
- Type of the target entity
- Narrative (describing the nature of the relation in free text)

An example information need, "manufacturers of vehicles used by UPS" is formulated as follows:

```
<query>
    <num>80</num>
    <entity_name>United Parcel Service (UPS)</entity_name>
    <entity_URL>clueweb09-en0014-05-00600</entity_URL>
    <target_entity>manufacturer</target_entity>
    <narrative>manufacturers of vehicles used by UPS</narrative>
</query>
```

A key change to last year's setup is that target entity types are not limited anymore to the four high-level entity types (person, organization, location, product). The target type is extracted from the narrative and is always given in singular form.

2.1.2 Output

- For each query, participants may return up to 100 answers (related entities). Each query must have at least one entity retrieved for it.

- For each answer entity a single homepage and a single supporting document must be returned; optionally, the name of the entity may also be returned.

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

- Groups that generate results using Web Search Engines are required to submit an obligatory run, using the Lemur ClueWeb Online Query Service,[1] to ensure reproducibility.

New in 2011 that we require a single supporting document for each answer and that Web Search Engines may not be used unless submitting a corresponding run using a common ClueWeb API.

---

[1] http://lemurproject.org/clueweb09.php/index.php#Services

2.1.3 Data collection

The document collection is the English portion of ClueWeb, comprising of approximately 500 million pages.

2.1.4 Topics and assessments

Both topic development and relevance assessments were performed by NIST. For the 2011 edition of the track 50 new REF topics have been created.

The evaluation methodology differs from what was originally set out in the guidelines. In particular, runs are evaluated using standard trec_eval and supporting documents are not incorporated into the evaluation. The judgments come from two sources: answers found by the assessors during topic development and pooled results from participants (pooled down to depth 30). Although not used in the scoring, a judgment file with the assessments of supporting documents and name correctness has also been made available.

The main evaluation measure we use is Mean Average Precision (MAP). We also report on R-Precision (precision at rank R).

2.2 Related Entity Finding, LOD-variant

In this pilot we investigate using Linked Open Data (LOD) URIs instead of homepages for entity identification. The task and the topics are the same as for the main REF task. The LOD crawl used is the Sindice-2011 data set; see Section 2.3.3 for details.

2.3 Entity List Completion

Entity List Completion (ELC) addresses essentially the same task as REF does: finding entities that are engaged in a specific relation with an input entity. There are two main differences to REF:

- Entities are not represented by their homepages, but by a unique URI (from a specific collection, a sample of the Linked Open Data cloud).[2]

- A number of entity homepages (i.e., ClueWeb docIDs) are made available as part of the topic definition, as examples of known relevant answers.

The ELC task then is defined as follows:

> Given an *information need and a list of known relevant entity homepages*, return a list of relevant entity URIs from a specific collection of Linked Open Data.

2.3.1 Input

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

- Input entity, defined by its name, homepage (ClueWeb docID), and one or more LOD URIs

- Type of the target entity (defined using the DBpedia Ontology[3])

- Narrative (describing the nature of the relation in free text)

---

[2]We acknowledge that Web of Data would be a more appropriate name for what we refer to as LOD. However, in order to not to confuse participants, we keep LOD for now.

[3]http://wiki.dbpedia.org/Ontology

- A set of example entities, each defined by one or more ClueWeb09 docIDs (and the corresponding URLs) and optionally one or more names.

An example information need, "Organization of Petroleum Exporting Countries (OPEC)" is formulated as follows:

```
<query>
<num>22</num>
<entity_name>Organization of Petroleum Exporting Countries
  (OPEC)</entity_name>
<entity_homepage id="clueweb09-en0010-21-28880">
  http://www.opec.com/</entity_homepage>
<target_entity>location</target_entity>
<target_type_dbpedia>Country</target_type_dbpedia>
<narrative>Find countries that are members of OPEC
  (the Organization of Petroleum Exporting Countries).</narrative>
<examples>
  <entity>
    <homepage id="clueweb09-en0002-20-01948">
      http://english.mofa.gov.qa/index.cfm</homepage>
    <homepage id="clueweb09-en0002-74-29899">
      http://portal.www.gov.qa/wps/portal/</homepage>
    <name>qatar</name>
  </entity>
  <entity>
    <homepage id="clueweb09-en0127-57-06714">
      http://en.iran.ir/</homepage>
    <name>iran</name>
  </entity>
</examples>
</query>
```

Changed since last year that neither the input entity nor examples (known relevant) entities are mapped manually to LOD, only ClueWeb IDs and the corresponding URLs are provided.

2.3.2   Output

- For each query, participants may return up to 100 answers (related entities). Each query must have at least one entity retrieved for it.

- For each answer entity a single URI must be returned; optionally, the name of the entity may also be returned.

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

- Groups that generate results using Web Search Engines are required to submit an obligatory run, using the Lemur ClueWeb Online Query Service,[4] to ensure reproducibility.

2.3.3   Data collection

Last year we used the Billion Triple Challenge (BTC) collection as a sample of Linked Open Data (LOD), and found that it did not contain many of the entities targeted by the topics. This

---

[4]http://lemurproject.org/clueweb09.php/index.php#Services

Group	REF	REF-LOD	ELC
Beijing University of Posts and Telecommunications (PRIS)	Y	N	Y
Chinese Academy of Sciences (ICT)	N	N	Y
Digital Enterprise Research Institute	N	N	Y
Laboratoire d'informatique d'Avignon	N	N	Y
Peking University	Y	N	N
Shanghai TongKey Network Technology Co., Ltd	Y	N	N
Team COMMIT	N	N	Y
Univ. of Arkansas at Little Rock	N	Y	N
University of Indonesia	N	N	Y
Wuhan University	Y	N	Y

Table 2: Groups participated in Entity 2011.

year we introduced a new collection, Sindice-2011, created by the Sindice team from DERI, NUI Galway. The collection is derived from data collected by the Sindice semantic search engine and is designed specifically for supporting research in the domain of web entity retrieval. It also comes with a set of tools to help researchers work with the data set [1]. The collection is available at `http://data.sindice.com/trec2011/`.

2.3.4 Topics and assessments

We created the ELC 2011 topic set based on the REF 2010 topics, where known relevant answers serve as examples. Target types have manually been mapped to the DBpedia ontology. Note that the input entity and example entities are not mapped to LOD. Looking them up in the LOD crawl is now part of the task. URIs of example entities may also be returned as answers but these are worth less credit than finding new entities. Relevance is binary, but a distinction is made between returning examples and new entities; the main metric is NDCG.

Relevance assessments are collected using community judging, the assessment procedure is underway.

3 Results

Ten groups submitted a total of 37 runs. Table 2 lists the participating groups.

3.1 REF task

Four teams submitted a total of 12 runs for the REF task; the results are shown in Table 3.

3.2 REF-LOD task

Since only a single team has submitted runs, this task did not result in reusable assessments.

3.3 ELC task

Evaluation results are not yet available at the time of writing.

RunID	Type	MAP	R-Prec
PRISREF1	manual	0.2509	0.2908
PRISREF3	manual	0.2450	0.2750
PRISREF4	manual	0.2448	0.2823
PRISREF2	manual	0.2329	0.2620
TongKeyEN2	manual	0.1266	0.1984
TongKeyEN2	automatic	0.1209	0.1972
WhuRun1	manual	0.0063	0.0176
WhuRun2	manual	0.0050	0.0229
ICSTmaxSni	automatic	0.0004	0.0015
ICSTmaxAll	automatic	0.0000	0.0000
ICSTaveSni	automatic	0.0000	0.0000
ICSTaveAll	automatic	0.0000	0.0000

Table 3: Results for the REF task. Runs are ordered by MAP scores.

4   Approaches

In this section we present the summaries of approaches used by some representative systems participated in the Entity track. We are going to add a more comprehensive analysis of approaches for the proceedings version of this overview.

4.1   REF task

Wuhan University used a natural language processing based method to construct an efficient query out of the topic description. Then they analyzed the top 200 documents, returned by Lemur ClueWeb Online Query Service in response to this query. Finally, they analyzed the co-occurrence of the candidate entity with the given entity using the context around the candidate entity to rank the candidate entities.

4.2   REF-LOD (pilot) task

University of Arkansas at Little Rock followed a network mining approach, which started from finding documents relevant to the given entities. All documents originated from ClueWeb, but were found in two different ways: either by running a query on the collection index built by Indri, or by sending a query to the Bing search API (and leaving only ClueWeb documents in the returned rankings). They further identified entities in the found documents by using a combination of Stanford Named Entity Recognizer (NER), Apache sentence detector, a sentence parser and the DBpedia lookup service. A combination of entity-entity and entity-document relationship network was constructed from the identified entities and the corresponding documents. The proposed network mining algorithm assigned a score to each entity (a vertex in the network) based on their proximity to relevant documents.

4.3   ELC task

COMMIT/ILPS, University of Amsterdam addressed the ELC task using two techniques, one based on retrieval of entities based on the terms in the relation and the terms in the objects and predicates of the entities in the LOD cloud (text based approach). The other technique used the link overlap between the example entities and entities in the LOD cloud to produce a ranking (linked based approach). This technique first requires a mapping of the example entities to the

LOD cloud. They also tried to select a set of candidate entities from the homepages of the source entities in a web corpus. They extracted all entities from the top 5 pages that covered the most example entities. These candidates were mapped to URIs in the LOD cloud and ranked by a link-based ranking approach.

University of Indonesia approached the ELC task by resolving the linguistic relation of the given entity and the query description using a part-of-speech (POS) tagger. Subsequently, they retrieved the top-100 snippets for each query term, and passed each retrieved snippet into the DBPedia Spotlight web service to identify the candidate entities. Further, they ranked those candidates using their term frequency scores, where each term frequency score was calculated using the retrieved snippets. Finally, to find specific URIs in the final entity list, they performed phrase-based search in the Sindice dump collection.

5   Summary

The second edition of the Entity track featured two main tasks: Related Entity Finding (REF) and Entity List Completion (ELC). Entity 2011 also featured a pilot task, aiming to get more insight in the question whether semantic web resources would address the problem of entity representation better than our previous 'an entity has a homepage' assumption.

Unfortunately, instead of attaining a healthy growth in participation when compared to 2010, we attracted only a handful of participants; especially the REF task has been disappointing. Also, judging from an informal analysis of the results submitted we estimate that the pools are not of great quality. It is unclear why the track has been less successful than before—maybe teams thought that handling ClueWeb A was too challenging, or our requirement of a non-web run (for reasons of comparison) was perceived as too much effort; maybe the deadlines fell earlier than anticipated, or, researchers may simply have moved on to studying other problems.

As a consequence, we decided to put the track on hold for 2012. We are organizing ourselves to writing a new and improved track proposal for 2013, and welcome suggestions from the community. At this point, we seriously consider an event oriented focus, where we may include a mix of web and news data. We solicit input on the mailing list and linked-in discussion group, and plan to approach people directly to inquire what has been perceived as the main bottleneck in this year's track setup. Please let us know how you would like to see the track develop in the future!

References

[1] S. Campinas, D. Ceccarelli, T. E. Perry, R. Delbru, K. Balog, and G. Tummarello. The Sindice-2011 dataset for entity-oriented search in the web of data. In *1st International Workshop on Entity-Oriented Search (EOS)*, 26–32 2011.