Overview of the TREC 2010 Entity Track

Krisztian Balog
University of Amsterdam
k.balog@uva.nl

Pavel Serdyukov
TU Delft, The Netherlands
p.serdyukov@tudelft.nl

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

1 Introduction

The overall goal of the track is to perform entity-oriented search tasks on the World Wide Web. Many user information needs concern entities (people, organizations, locations, products, ...); these are better answered by returning specific objects instead of just any type of documents.

Defining entities on the Web is still an unsolved problem. We settled on representing entities by their homepages, under the assumption that any entity of interest would have at least one homepage. The homepage URL is used as unique identifier. In this scenario, entity ranking corresponds to the task of returning the homepages of entities of a given type, that are relevant to the users information need (represented as natural language text). We have to also consider that many entity queries could have very large answer sets (e.g., "actors playing in hollywood movies"); extra problematic with corpora the size of ClueWeb. In 2009, we decided therefore that finding associations between entities would be a more challenging one (in terms of modeling) and also a more manageable one (from a test collection building perspective) than finding associations between entities and topics, and defined the *Related Entity Finding (REF)* task.

Related entity finding requests a ranked list of entities (of a specified type) that engage in a given relationship with a given source entity. REF run as a pilot in 2009 and is the track's main task in this year; the document collection has been enlarged to the English subset of ClueWeb. We intend to repeat the REF task at least one more time in 2011.

One observation from the 2009 edition of the track is that many of the proposed approaches build heavily on Wikipedia and use it as a "semantic backbone": considering Wikipedia a large repository of entity names and types. Our goal is however not to evaluate entity retrieval over Wikipedia (this task has already been looked at in INEX, and a test collection exists), nor to limit ourselves to the (mostly popular) entities that are present in Wikipedia. As of this year, we are therefore not accepting Wikipedia pages as entity homepages.

The issue of combining (noisy) textual material (the Web) with semi-structured data (like Wikipedia or slightly more structure data sources like IMDB) is however an interesting line of research. As many data sources, and in particular those being constructed as so-called Linked Open Data (LOD), are naturally organized around entities, it would be reasonable to examine this problem in the context of entity retrieval. To foster research in this direction, we introduced the new *Entity List Completion (ELC)* pilot task. ELC is motivated by the same user scenario as REF, but with the main difference that entities are represented by their URIs in a semantic web crawl (the Billion Triple Collection). In addition, a small number of example entities (defined by their URIs) are made available as part of the topic definition. Our goal is to turn this pilot task to an "official" task in 2011.

GroupID	Name	REF	ELC
BIT	Beijing Institute of Technology	Y	N
CARD_UALR	Center for Advanced Research in Data Mining, UALR	Y	N
CMU_LIRA	Carnegie Mellon University	N	Y
FDWIM2010	Fudan university	Y	N
HPI	Hasso Plattner Institute/ SAP Research	Y	N
ICTNET	Institute of Computing Technology, Chinese Academy of Sc.	Y	N
LIA_UAPV	Universit d'Avignon	Y	N
NiCT	National Institute of Information and Communications Tech.	Y	N
PITTSIS	School of Information Sciences, University of Pittsburgh	Y	N
PRIS	Beijing University of Posts and Telecommunications	Y	N
Purdue_IR	Purdue University	Y	Y
SIEL_IIITH	Search & Information Extraction Lab (IIIT, Hyderabad)	Y	N
UAms	University of Amsterdam	Y	Y
UAmsterdam	University of Amsterdam (Kamps)	Y	N
UWaterlooEng	University of Waterloo	Y	N

Table 1: Groups participated in Entity 2010.

In the remainder of the paper we discuss the REF and ELC tasks in detail, in Sections 2 and 3, respectively. Since evaluation results are not available at the time of writing, we report only on the task setup and participants' approaches.

2 Related Entity Finding

2.1 Task

The Related Entity Finding (REF) task is formulated 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.

2.2 Topics and data collection

This task ran as a pilot task (with 20 topics) on ClueWeb CatB in 2009, and is running as the main task (with 50 topics) on the English portion of ClueWeb in 2010. The target entity types considered in 2009 were: *person*, *organization*, and *product*; these are complemented with *location* in 2010. Participants were requested to submit results for both the 2009 and 2010 topics.

2.3 Approaches

Fourteen groups submitted a total of 48 runs. Table 1 lists the participating groups. The following are descriptions of the approaches taken by the different groups. These paragraphs were contributed by participants and are meant to be a road map to their papers.

(**BIT** did not submit a summary as of this writing.)

CARD-UALR To find relevant entities and their homepages, first, we identified the entities and their types using Stanford Named Entity Recognizer. Due to its limitations, we could only identify PERSON, LOCATION and ORGANIZATION type entities. Next, an entity-entity co-occurrence graph was established. If two entities co-occurred in a webpage more than a specified threshold, the two entities were linked. Given the query entity, relevant entities are extracted based on a novel centrality measure (Cumulative Structural Similarity-CSS) using the intuition that an important entity will share many common neighbors with adjacent entities. Additionally, PageRank, HITS and Ensemble-based approaches are submitted.

FDWIM2010 The FDWIM group proposes a multiple-stage retrieval framework for the task of related entity finding. In the document retrieval stage, search engine is used to improving the retrieval accuracy. In the next stage, they extract entity with NER tools, Wikipedia and text pattern recognition. Then stoplist and other rules is employed to filtering entity. Deep mining of the authority pages is effective in this stage. In entity ranking stage, many factors including keywords from narrative, page rank, combined results of corpus-based association rules and search engine are considered. Finally, an improved feature-based algorithm is proposed for the entity homepage detection.

HPI The approach of the HPI-group studies in particular the exploitation of advanced features of different Web search engines to achieve high quality answers for the related entity finding task. Thus, the system preprocesses a topic using part-of-speech tagging and synonym dictionaries, and generates an enriched keyword query employing advanced features of the particular Web search engine. After retrieving a corpus of documents, the system constructs an extraction rule that consists of the source entity (and synonyms), the target entity type and words that should occur in the context of both (taken from the narrative relation description). After the extraction of potentially related entities, they are subjected to a deduplication mechanism and scored for each document with respect to the distance to the source entity. Finally, these scores are aggregated across the corpus by incorporating the rank position of a document. For homepage retrieval the HPI-system further employed advanced features of the used Web search engines - for instance to retrieve candidate URLs by queries such as "entity in anchor". Homepages are ranked by a weighted aggregation of feature vectors. The weight for each of the 17 used features was determined beforehand using a genetic learning algorithm. The submitted runs compare the performance of the three most popular search engines, that were employed by the system.

ICTNET The ICTNET group proposes a bipartite graph reinforcement model for entity ranking. Firstly, the candidate entities are extracted from related text snippets and are ranked based on a probabilistic model. Secondly, the lists which may contain several target entities are also extracted. Thirdly, a bipartite graph is constructed in which candidate entities and lists are considered as the two disjoint sets of graph vertices. Finally, the reinforcement algorithm is applied over the graph to get the final score for each candidate entity. For the homepage finding, google is used to search for top-K urls and some heuristic rules are used to identify the real homepage.

(**LIA_UAPV** did not submit a summary as of this writing.)

NiCT In 2010, the NiCT group mainly focused on improving target entity extraction and entity ranking, both of them play vital roles in the REF system. A Named Entity Recognition tool is first used to extract entities that match types of target entities such as organization, person, etc. Secondly, dependency tree-based patterns learnt automatically are employed to filter out the extracted entities that do not match fine-grained types of name entities

such as university, airline, author, etc. In ranking part, a dependency tree-based similarity approach is proposed, which is better than language model.

PITTSIS Our method is based on a two-layer probability model for integrating document retrieval and entity extraction together. The document retrieval layer finds highly relevant documents, and the entity extraction layer extracts the right entities. Our goal in this year TREC is to set up a frame work for evaluating and exploring each individual layer as well as the overall workflows. This method helps to reduce the overall retrieval complexity while keeping high accuracy in locating target entities.

PRIS The PRIS group proposes Document-Centered Model (DCM) and Entity-Centered Model (ECM) for the entity finding task. In DCM, documents are seen as a bridge. Both probabilities of a query and entity with respect to a document are estimated. In ECM, snippets extracted from documents are at the bottom to support entities. BM25 method is also introduced into ECM besides indri retrieval model. Another improvement aims to entity extraction. Special web page, NER tool and entity list generated by some rules are all taken into account.

(Purdue_IR did not submit a summary as of this writing.)

SIEL_IIITH We use external resources like Wikipedia and Web, as Clueweb Category A dataset is not available. We extract all entities from Wikipedia using pattern finding techniques and indexed them with their type. We searched query in this index to find target entities. We use web search to find target entities not present in Wikipedia index. We then combine both the results to get final ranking. We then used Clueweb's URL-DocId mapping to find urls of target entities present in Clueweb dataset and present corresponding DocID as final results. This approach give satisfactory results in the absence of Clueweb dataset.

UAms To address REF we look for homepages of entities of the target type that co-occur with the source entity in contexts of a certain size, emphasizing contexts that contain terms from the relation (the narrative provided with a topic). We experimented with context size by varying a window size parameter. To perform filtering based on type and homepage finding we use Freebase, which provides category labels and homepage URLs. To remove NER errors we restrict the candidate entities to those in Freebase. In addition to Freebase homepage URLs we submitted entity strings to a web search engine to find homepages.

UAmsterdam The University of Amsterdam, group of Jaap Kamps, participates only in the main related entity finding task, and uses Wikipedia as a pivot to search for entities. The approach is very similar to last year's approach. Wikipedia topic categories are manually assigned to the query topics, which are more specific as the given target categories. These more specific target categories are used to retrieve entities within Wikipedia. To search web entities the external links in Wikipedia are used, and an anchor text index is searched.

UWaterlooEng The University of Waterloo investigated whether related entity finding problem can be addressed by unsupervised approaches that rely primarily on statistical methods and common linguistic tools, such as named-entity taggers and syntactic parsers. An initial candidate list of entities is extracted from top ranked documents retrieved for the query, and then refined using a number of statistical and linguistic methods. One of the key components of their method consists of finding hyponyms of the category name specified in the narrative, representing candidate entities and hyponyms as vectors of grammatical dependency triples, and calculating similarity between them.

3 Entity List Completion

3.1 Task

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),
- A small number of known relevant entities are made available as part of the topic definition, as examples.

3.2 Topics and data collection

We use the Billion Triple Challenge (BTC) collection[1], a publicly available Semantic Web crawl; we consider this collection as a reasonable sample of Linked Open Data (LOD). Not all nodes in this Semantic Web graph are entities; identifying the nodes which refer to an entity is one of the challenges introduced by the task. Besides, the BTC collection appears to be noisy and incomplete. For instance, it contains far less Wikipedia entities than those which are the part of the ClueWeb B collection. This may be representative of the situation where entity classes are not that well covered by specialized entity repositories (as opposed to the coverage of the most popular entity classes in Wikipedia).

In order to help participants of 2009 use their previous approaches in the new setup, we use a subset of the 20 topics developed in the 2009 pilot run of the track. We had to exclude 6 topics from this set which had either too many additional entities as answers, or whose answer set from 2009 was complete, so could not be extended (for instance, all members of a band were found by participants of REF task in 2009). For each of the remaining 14 topics, the answer entities identified in the 2009 Entity track serve as the list of examples. These were then manually mapped to LOD by track organizers with the help of a baseline entity search system.

3.3 Approaches

For the ELC pilot task, three groups submitted a total of 5 runs. Below are the summaries of approaches, contributed by the participating teams (edited slightly for better presentation).

CMU_Lira The team from CMU (CMU_Lira) focused on Entity List Completion using Set Expansion techniques. Set expansion refers to expanding a partial set of "seed" objects into a more complete set. They propose a two stage retrieval process. The first stage takes the given query_entity and target_entity examples as seeds and does set expansion. In the second stage, candidates generated by first stage are type checked and ranked. The first stage of this approach focuses on recall while the second stage tries to improve precision of the intermediate result list. They have submitted two runs, by doing set expansion on the Web and on the Clueweb corpus.

(Purdue_IR did not submit a summary as of this writing.)

UAms To address ELC we look for entities similar to the given example entities. We find items that are linked to by example entities and consider other entities that link to those items to be candidate entities. For each entity we consider its links as well as the items to which it

[1] `http://vmlion25.deri.ie/`

links. The combination of a link and a linked item forms a link-item pair. Each entity has its set of associated link-item pairs. We rank entities by set overlap between their link-item pairs and the example entity link-item pairs. We then re-rank these intermediate results based on word overlap between the topic narrative and entity link-item pairs.

4 Summary

The second edition of the Entity track featured the Related Entity Finding (REF) as the main task: given an input entity, the type of the target entity (person, organization, product, or location), and the relation, described in free text, systems had to return homepages of related entities, and, optionally, the name of the entity.

For the second year of the track, 50 topics were created and assessed. In addition, participants were also requested to generate results for the 20 REF topics from 2009. We had slightly more submissions compared to the previous year (14 vs. 13 participants, 48 vs. 41 runs). This serves as a good motivation to run the task again next year. However, it becomes especially interesting if there are other applications within the same domain which have the potential to attract as many researchers as the REF task.

Entity 2010 also featured a pilot task: Entity List Completion (ELC). ELC is motivated by the same user scenario as REF, but entities are represented by their URIs in a semantic web crawl (the Billion Triple Collection), and a small number of example entities are made available as part of the topic definition. Our pilot run of the ELC task was not as popular as REF, probably due to the fact that participation needed a significant additional effort, because of the different nature of the dataset. We plan to run the task again in 2011, so that participants could have enough time to build their systems and process the data.