Overview of the TREC 2009 Entity Track

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

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

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

Paul Thomas
CSIRO, Canberra, Australia
paul.thomas@csiro.au

Thijs Westerveld
Teezir, Utrecht, The Netherlands
thijs.westerveld@teezir.com

1 Introduction

The goal of the entity track is to perform entity-oriented search tasks on the World Wide Web. Many user information needs would be better answered by specific entities instead of just any type of documents.

The track defines entities as “typed search results”, “things”, represented by their homepages on the web. Searching for entities thus corresponds to ranking these homepages. The track thereby investigates a problem quite similar to the QA list task. In this pilot year, we limited the track’s scope to searches for instances of the organizations, people and product entity types.

2 Related entity finding task

The first edition of the track featured one pilot task: *related entity finding.*

2.1 Data

The document collection is the “category B” subset of the ClueWeb09 data set[1]. The collection comprises about 50 million English-language pages.

2.2 Task

The first year of the track investigates the problem of related entity finding:

> 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.

This task shares similarities with both expert finding (in that we need to return not “just” documents) and homepage finding (since entities are uniquely identified by their homepage). However, approaches to address this task need to generalize to multiple types of entities (beyond

[1] ClueWeb09: `http://boston.lti.cs.cmu.edu/Data/clueweb09/`

just people) and return the homepages of multiple entities, not just one. Also, the topic defines a focal entity to which returned homepages should be related.

2.2.1 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, or product)
- Narrative (describing the nature of the relation in free text)

This year's track limits the target entity types to three: people, organizations, and products. (Note that the input entity does not need to be limited to these three types).

An example topic is shown below:

```
<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>
```

2.2.2 Output

For each query, participants could return up to 100 answers (related entities). Each answer record comprises the following fields:

- (HP1..HP3) Up to 3 homepages of the entity (excluding Wikipedia pages)
- (WP) Wikipedia page of the entity
- (NAME) A string answer that represents the entity concisely
- (SUPPORT) Up to 10 supporting documents

For each target entity (answer) at least one homepage (HP1) and at least one supporting document must be returned. The other two homepages (HP2 and HP3), the wikipedia page (WP), and the entitys name (NAME) are optional. Homepage fields (HP1..HP3) will be treated as a set, i.e., the order in which these are returned is indifferent. The same entry must not be retrieved for multiple entities in the same topic. This means that documents returned in the homepage (HP1..HP3) and wikipedia (WP) fields must not be retrieved for multiple entities in the same topic.

Returned entity names were required 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 replaces with "_"

2.3 Topics and assessments

Both topic development and relevance assessments were performed by NIST. For the first year of the track, 20 topics were created and assessed.

Entities are not so easily defined very precisely; instead of endulging in a long discussion about the exact semantics underlying the notion of entity, we simply adopt the following working definition: *A web entity is uniquely identifiable by one of its primary homepages.* Real-world entities can be represented by multiple homepages; a clearly preferred one cannot always be given. As a work-around, entity resolution is addressed at evaluation time.

2.3.1 Assessment procedure

In the first phase of the assessment procedure, all runs were pooled down to 10 records, and for each record entry, judgments were made for the homepage (HP and WP) and the name (NAME) fields.

Homepages were judged on a three-point relevance scale: non-relevant, relevant ("descriptive") or primary ("authoritative"). If a HP entry was the homepage for a correct entity, it was judged "primary" (2). Likewise, if a WP entry was a correct Wikipedia page for an entity. Pages that were related without being actual homepages for the entities were judged "relevant" (1). All other pages were judged non-relevant (0).

Each name returned in the record was also judged on a three-level scale: incorrect, inexact or correct. A name was judged inexact or correct if it matched up with something else in the record, even if the record was not either primary or relevant for the topic. A name was "inexact" (1) if it was correct but was not a complete form (had extra words or was ambiguous). Otherwise it was judged incorrect (0).

In a second phase of the assessment procedure the assessors matched primary pages to correct names, creating a set of equivalence classes for the right answers to each topic (i.e., addressing the resolution of entities).

2.3.2 Qrels

In the qrels file, the fields are:

```
topic-entry_type docid_or_name rel class
```

Where "topic-entry_type" is e.g. 1-HP1, the docid_or_name is a document ID or a name, rel is $\{0, 1, 2\}$ as described above, and class is an integer where lines with the same topic number and class correspond to the same entity.

2.3.3 Evaluation measures

The main evaluation measure we use is NDCG@R; that is, the normalized discounted cumulative gain at rank R (the number of primaries and relevants for that topic) where a record with a primary gets gain 2, and a record with a relevant gets gain 1. We also report on P@10, the fraction of records in the first ten ranks with a primary.

In the next section, we report the official evaluation results for the tasks. These are computed only on the basis of the homepage (HP) fields. In alternative evaluation scenarios, extra credit is to be given for finding Wikipedia homepages and names for the related entities; we leave this to further work.

Run	Group	Type	WP	Ext.	NDCG@R	P@10	#rel	#pri
KMR1PU	Purdue	auto	Y	Y	0.3061	0.2350	126	61
uogTrEpr	uogTr	auto	N	N	0.2662	0.1200	347	79
ICTZHRun1	CAS	auto	N	N	0.2103	0.2350	80	70
NiCTm3	NiCT	auto	Y	Y	0.1907	0.1550	99	64
UAmsER09Ab1	UAms (Amsterdam)	auto	N	N	0.1773	0.0450	198	19
tudpw	TUDelft	auto	Y	N	0.1351	0.0950	108	42
PRIS3	BUPTPRIS	manual	N	N	0.0892	0.0150	48	3
UALRCB09r4	UALR_CB	auto	N	N	0.0666	0.0200	15	4
UIauto	UIUC	auto	N	N	0.0575	0.0100	64	3
uwaterlooRun	Waterloo	auto	N	N	0.0531	0.0100	55	5
UdSmuTP	EceUdel	auto	N	N	0.0488	0.0000	102	10
BITDLDE09Run	BIT	manual	N	Y	0.0416	0.0200	81	9
ilpsEntBL	UAms (ISLA)	auto	Y	Y	0.0161	0.0000	30	1

Table 1: The top run from each group by NDCG@R. The columns of the table (from left to right) are: runID, group, type of the run (automatic/manual), whether the Wikipedia subcollection received a special treatment (Yes/No), whether any external resources were used (Yes/No), NDCG@R, P@10, number of relevant retrieved entities, and number of primary retrieved entities.

3 Runs and Results

Each group was allowed to submit up to four runs. Thirteen groups submitted a total of 41 runs; of those, 33 were automatic runs. Four groups submitted a total of 8 manual runs.

Table 1 shows the evaluation results for the top run from each group (ordered by NDCG@R). As we see from Table 1, performance varies significantly over the participants. Interestingly, result rankings would be quite different dependent on the performance measure chosen.

The differences between P@10 and NDCG@R results show that even though teams Purdue and CAS find the same number of primary entity homepages in their top 10 results, the Purdue strategy seems better at identifying more relevant (but not primary) homepages.

University of Glasgow retrieves by far the highest number of relevant entities, but other groups achieve better early precision. This could be merely a matter of re-ranking the initial results list, possibly helped by improved spam detection (but we did not investigate in detail yet).

The complete list of all submitted runs along with the evaluation results is presented in Table 2.

4 Approaches

The following are descriptions of the approach taken by different groups. These paragraphs were contributed by participants and are meant to be a road map to their papers.

Purdue We propose a hierarchical relevance retrieval model for entity finding. In this model, three levels of relevance are examined which are document, passage and entity, respectively. The final ranking score is a linear combination of the relevance scores from the three levels. To train the homepage classifiers, we use an incremental learning strategy motivated by active learning to alleviate the manual effort of labeling the scarce descriptive homepages.

uogTr The uogTr group extended the Voting Model for people search to the task of finding related entities of a particular type. Their approach builds semantic relationship support for the Voting Model, by considering the co-occurrences of query terms and entities in a document as a vote for the relationship between these entities. Additionally, on top of the Voting Model, they developed a novel graph-based technique to further enhance the initial vote estimations.

CAS In our approach, a novel probabilistic model was proposed to entities finding in a Web collection. This model consists of two parts. One is the probability indicating the relation between the source entity and the candidate entities. The other is the probability indicating the relevance between the candidate entities and the topic.

NiCT We aim to develop an effective method to rank entities via measuring "similarities" between input query and supporting snippets of entities. Three models are implemented to this end: The DLM calculates the probabilities of generating input query given supporting snippets of entities via language model; The RSVM ranks entities via a supervised Ranking SVM; The CSVM estimates the probabilities of input query belonging to "topics" represented by entities and their supporting snippets via SVM classifier.

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

TUDelft In three of four methods used to produce our runs we treated Wikipedia as the repository of entities to rank. We ranked either all Wikipedia articles, or those articles that are linked by the "primary" Wikipedia page for the query entity. Then we considered only entities that are mentioned at the given primary or at the top ranked non-Wikipedia pages from the entire collection. Additionally we filtered-out entities that belong to non-matching classes using DBPedia, Yago, and articles infoboxes.

BUPTPRIS In our work, an improved two-stage retrieval model is proposed according to the task. The first stage is document retrieval, in order to get the similarity of the query and documents. The second stage is to find the relationship between documents and entities. Final scores are computed by combining previous results. We also focus on entity extraction in the second stage and the final ranking.

UALR_CB We used Lemur tool kit version 4.10 to index the WARC format documents which were given on Red Hat Enterprise Linux machine. Then we used the queries to retrieve the named entities using Indri Query Language which was very related to the Inquery language. First we retrieved the pages related to the given queries of people or organizations and products and then we found the exact home pages for them using some keywords related to them.

UIUC The team from University of Illinois at Urbana-Champaign focused on studying the usefulness of information extraction techniques for improving the accuracy of entity finding task. The queries were formulated as a relation query between two entities such that one of the entities is known and the goal is to find the other entity that satisfies the relation. The two-step approach of relation retrieval followed by entity finding helped explore techniques to improve entity extraction using NLP resources and corpus-based reranking based on other relations that link the entities.

UWaterloo All terms in the entity name and narrative except stopwords constitute our query terms. We retrieve the query's top-100 passages and expanded them using a sliding window size of 100. We fetch their n-grams where n = 1..10. We consider only n-grams that is a Wikipedia title. Tf-idf weight was assigned to each term in the n-gram. We now compute the ranking score for each n-gram using the sum of their term weights.

EceUdel Our general goal for the Entity track is to study how we may apply language modeling approaches and natural language processing techniques to the task. Specifically, we proposed to find supporting information based on segment retrieval, extract entities using Stanford NER tagger, and rank entities based on a previously proposed probabilistic framework.

BIT Related Entity Finding by Beijing Institute of Technology employs Lemur toolkit to index and retrieve dataset stemmed by Krovetz stemmer and stopped using a standard list of 421 common terms; devised OpenEphyras Question Analyzer to construct weighted query strings; OpenEphyras NETagger to extract typed entities; OpenNLPs ME classifier to rank extracted entities homepages whose model is trained by TREC-supplied test topics; DBPedia (dump date 05/11/09) to extract product name list for identifying product entity names.

UAms (ISLA) We propose a probabilistic modeling approach to related entity finding. We estimate the probability of a candidate entity co-occurring with the input entity, in two ways: context-dependent and context-independent. The former uses statistical language models built from windows of text in which entities co-occur, while the latter is based on the number of documents associated with candidate and input entities. We also use Wikipedia for detecting entity name variants and type filtering.

5 Summary

The first year of the entity track featured a related entity finding task. Given an input entity, the type of the target entity (person, organization, or product), and the relation, described in free text, systems had to return homepages of related entities, and, optionally, the corresponding Wikipedia page and/or the name of the entity.

Topic development encountered difficulties because it turned out that for many candidate topics, Category B collection did not contain enough entity homepages. Assessment took place in two stages. First, the assessors judged the returned pages. Here, the hard parts of relevance assessment are to (a) identify a correct answer and (b) distinguish a homepage from a non-homepage. Assessors were then shown a list of all pages they had judged “primary” and all names that were judged “correct”. They could assign each to a pre-existing class, or create a new class.

Concerning submissions, a common take on the task was to first gather snippets for the input entity, then extract co-occurring entities from these snippets, using a named entity tagger (off-the-self or custom-made). Language modeling techniques were often employed by these approaches. Several submissions built heavily on Wikipedia; exploiting links outgoing from the entity’s Wikipedia page, using it to improve named entity recognition, making use of Wikipedia categories for entity type detection, just to name a few examples.

Run	Group	Type	WP	Ext.	NDCG@R	P@10	#rel	#pri
KMR1PU	Purdue	auto	Y	Y	**0.3061**	**0.2350**	126	61
KMR3PU	Purdue	manual	N	Y	0.3060	**0.2350**	126	61
KMR2PU	Purdue	auto	Y	Y	0.2916	**0.2350**	115	56
uogTrEpr	uogTr	auto	N	N	0.2662	0.1200	**347**	**79**
uogTrEc3	uogTr	auto	N	N	0.2604	0.1200	331	75
uogTrEbl	uogTr	auto	N	N	0.2510	0.1050	344	75
uogTrEdi	uogTr	auto	N	N	0.2502	0.1150	343	74
ICTZHRun1	CAS	auto	N	N	0.2103	**0.2350**	80	70
NiCTm3	NiCT	auto	Y	Y	0.1907	0.1550	99	64
NiCTm2	NiCT	auto	Y	Y	0.1862	0.1750	99	61
NiCTm1	NiCT	auto	Y	Y	0.1831	0.1450	98	63
UAmsER09Ab1	UAms (Amsterdam)	auto	N	N	0.1773	0.0450	198	19
tudpw	TUDelft	auto	Y	N	0.1351	0.0950	108	42
tudpwkntop	TUDelft	auto	Y	Y	0.1334	0.1150	108	41
NiCTm4	NiCT	auto	Y	Y	0.1280	0.0950	87	45
UAmsER09Co	UAms (Amsterdam)	auto	N	N	0.1265	0.0400	87	23
tudwtop	TUDelft	auto	Y	N	0.1244	0.0650	125	50
tudwebtop	TUDelft	auto	N	N	0.1218	0.0600	103	28
basewikirun	UAms (Amsterdam)	auto	Y	N	0.1043	0.0500	77	40
PRIS3	BUPTPRIS	manual	N	N	0.0892	0.0150	48	3
wikiruncats	UAms (Amsterdam)	auto	Y	N	0.0805	0.0550	77	40
PRIS1	BUPTPRIS	auto	N	N	0.0729	0.0100	40	2
PRIS2	BUPTPRIS	manual	N	N	0.0712	0.0050	61	1
UALRCB09r4	UALR_CB	auto	N	N	0.0666	0.0200	15	4
PRIS4	BUPTPRIS	manual	N	N	0.0642	0.0150	70	4
UIauto	UIUC	auto	N	N	0.0575	0.0100	64	3
uwaterlooRun	Waterloo	auto	N	N	0.0531	0.0100	55	5
UdSmuTP	EceUdel	auto	N	N	0.0488	0.0000	102	10
UALRCB09r3	UALR_CB	manual	N	N	0.0485	0.0100	9	2
UdSmuCM50	EceUdel	auto	N	N	0.0476	0.0100	96	8
UdSmuCM	EceUdel	auto	N	N	0.0446	0.0100	102	13
UdSmuTU	EceUdel	auto	N	N	0.0430	0.0000	98	13
BITDLDE09Run	BIT	manual	N	Y	0.0416	0.0200	81	9
UALRCB09r2	UALR_CB	auto	N	N	0.0399	0.0150	7	3
UALRCB09r1	UALR_CB	auto	N	N	0.0392	0.0050	8	1
UIqryForm	UIUC	manual	N	Y	0.0251	0.0000	4	0
UIqryForm3	UIUC	manual	N	Y	0.0189	0.0000	16	0
ilpsEntBL	UAms (ISLA)	auto	Y	Y	0.0161	0.0000	30	1
ilpsEntcr	UAms (ISLA)	auto	Y	Y	0.0161	0.0000	30	1
ilpsEntem	UAms (ISLA)	auto	Y	Y	0.0128	0.0000	17	0
ilpsEntcf	UAms (ISLA)	auto	Y	Y	0.0105	0.0000	25	0

Table 2: All submitted runs by NDCG@R. The columns of the table (from left to right) are: runID, group, type of the run (automatic/manual), whether the Wikipedia subcollection received a special treatment (Yes/No), whether any external resources were used (Yes/No), NDCG@R, P@10, number of relevant retrieved entities, and number of primary retrieved entities. Highest scores for each metric are in boldface.