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/

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Overview of the TREC 2009 Entity Track	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Amsterdam,Spui 21,1012 WX Amsterdam, The Netherlands,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009) held in Gaithersburg, Maryland, November 17-20, 2009. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**10**	

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 entity's name (NAME) are optional. Homepage fields (HP1..HP3) are treated as a set, i.e., the order in which these are returned is indifferent. The same entry (i.e., documents returned in the HP1..HP3 and WP fields) must not be retrieved for multiple entities in the same topic.

Returned entity names are 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 replaced with "_"

2.3 Topics and assessments

Both topic development and relevance assessments were performed by NIST. Topic development encountered difficulties because it turned out that for many candidate topics, the "Category B" collection did not contain enough entity homepages. Trivial topics, i.e., topics for which all the related entities are linked from input entity's homepage/website or from its Wikipedia page, were avoided. For the first year of the track, 20 topics were created and assessed.

Entities are not so easily defined very precisely; instead of engaging 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

The assessment procedure consisted of two stages. In phase one, judgments were made for HP, WP, and NAME fields, individually. Then, in phase two, HPs, WPs, and NAMEs belonging to the same entity were grouped together.

Phase one. 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: (0) non-relevant, (1) relevant ("descriptive") or (2) primary ("authoritative"). If a HP entry was the homepage for a correct entity, it was judged "primary." 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." All other pages were judged non-relevant.

Each name returned in the record was also judged on a three-level scale: (0) incorrect, (1) inexact or (2) 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" if it was correct but was not a complete form (had extra words or was ambiguous). Otherwise it was judged incorrect.

Phase two. Assessors matched primary pages (HP and WP) 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` denotes the topic ID (first half) and the field (second half), e.g., "1-HP" is the HP field for topic 1; `docid_or_name` is a document ID (for fields HP1..HP3 and WP) or a name (for field NAME); `rel` is $\{0, 1, 2\}$ as described above; and `class` is an integer value, 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

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, using the default evaluation setting (HP-only). 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 (fraction of records in the first 10 ranks with a primary homepage), number of relevant retrieved homepages, and number of primary retrieved homepages.

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.

Note that evaluation results are not computed using the standard `trec_eval` tool, but a script developed specifically for the 2009 edition of the Entity track[2].

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 addition, we report on alternative evaluation scenarios, where extra credit is given for finding Wikipedia homepages and names for the related entities (see Section 3.1).

3 Runs and Results

Each group was allowed to submit up to four runs. Thirteen groups submitted a total of 41 runs; of those, 34 were automatic runs. Four groups submitted a total of 7 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 this in detail yet).

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

[2]http://trec.nist.gov/data/entity/09/eval-entity.pl

3.1 Alternative evaluations

We consider different variations for computing the gain for each record.

HP-only (default) only homepage (HP1..3) fields are considered; a record with a primary homepage gets gain 2, with a relevant homepage gets gain 1. Names are not taken into account. (For each record the maximum gain is 2.)

HP+NAME in addition to the homepage (HP1..3) fields, NAME is also taken into account. An extra gain of 1 is awarded if an exact name is returned along with a primary homepage. (For each record the maximum gain is 3.)

WP-only only the Wikipedia (WP) field is considered; a record with a primary Wikipedia page gets gain 2, with a relevant Wikipedia page gets gain 1. Names are not taken into account. (For each record the maximum gain is 2.)

HP+WP HP1..3 and WP fields are all considered, names are not; a record with a primary page (either homepage or Wikipedia page) gets gain 2, with a relevant page gets gain 1. (For each record the maximum gain is 2.)

HP+WP+NAME all fields are considered. An extra gain of 1 is awarded if an exact name is returned along with a primary homepage or Wikipedia page. (For each record the maximum gain is 3.)

The results of these alternative evaluation scenarios are presented in Table 3.

3.2 The usefulness of Wikipedia

In order to study how far we can go with Wikipedia only when looking for entities, we analyzed the list of relevant entities and the list of their description pages. We found that 160 out of 198 relevant entities (≈80%) have a Wikipedia page among their primary pages, while only 108 of them have a primary web page (70 entities have both). However, not all primary Wikipedia pages could be returned by participants or judged, or not all Wikipedia pages could exist on the date when the ClueWeb collection was crawled (January/February 2009). So, we manually looked for primary Wikipedia pages for those 38 entites that had only primary web pages, using online Wikipedia (accessed in December 2009). As a result, we discovered primary Wikipedia pages for 22 entities. Those 16 entities that are not represented in Wikipedia are seemingly not notable enough, however they include all answers for 3 of 20 queries (looking for audio cds, phd students and journals).

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 ranking. 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. Furthermore, we exploit the structure of tables and lists to identify the target entities from them by making a joint decision on all the entities with the same attribute. To find entity homepages, we train logistic regression models for each type of entities. A set of templates and filtering rules are also used to identify target entities. (Fang et al., 2009)

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. (McCreadie et al., 2009)

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. (Zhai et al., 2009)

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. (Wu and Kashioka, 2009)

UAms (Amsterdam) For the entity ranking track, we explore the effectiveness of the anchor text representation, we look at the co-citation graph, and experiment with using Wikipedia as a pivot. Two of our official runs exploit information in Wikipedia. The first run ranks all Wikipedia pages according to their match to entity name and narrative. To find primary homepages, we follow links on Wikipedia pages. The other run reranks Wikipedia pages of the first run using category information. The other two runs use an anchor text index where the queries consist of the entity name and the narrative, and co-citations of the given entity url. (Kaptein et al., 2009)

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. (Serdyukov and de Vries, 2009)

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. (Wang et al., 2009)

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. (Pamarthi et al., 2009)

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. (Zheng et al., 2009)

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. (Yang et al., 2009)

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. (Bron et al., 2009)

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, the "Category B" collection did not contain enough entity homepages. For the first year of the track, 20 topics were created and assessed. 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.

References

M. Bron, K. Balog, and M. de Rijke. Related Entity Finding Based on Co-Occurance. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, Gaithersburg, MD, 2009.

Y. Fang, L. Si, Z. Yu, Y. Xian, and Y. Xu. Entity Retrieval with Hierarchical Relevance Model, Exploiting the Structure of Tables and Learning Homepage Classifiers. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, Gaithersburg, MD, 2009.

R. Kaptein, M. Koolen, and J. Kamps. Result Diversity and Entity Ranking Experiments: Anchors, Links, Text and Wikipedia, University of Amsterdam. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, Gaithersburg, MD, 2009.

R. McCreadie, C. Macdonald, I. Ounis, J. Peng, and R. L. T. Santos. University of Glasgow at TREC 2009: Experiments with Terrier. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, Gaithersburg, MD, 2009.

J. Pamarthi, G. Zhou, and C. Bayrak. A Journey in Entity Related Retrieval for TREC 2009. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, Gaithersburg, MD, 2009.

P. Serdyukov and A. de Vries. Delft University at the TREC 2009 Entity Track: Ranking Wikipedia Entities. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, Gaithersburg, MD, 2009.

Z. Wang, D. Liu., W. Xu, G. Chen, and J. Guo. BUPT at TREC 2009: Entity Track. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, Gaithersburg, MD, 2009.

Y. Wu and H. Kashioka. NiCT at TREC 2009: Employing Three Models for Entity Ranking Track. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, Gaithersburg, MD, 2009.

Q. Yang, P. Jiang, C. Zhang, and Z. Niu. Experiments on Related Entity Finding Track at TREC 2009. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, Gaithersburg, MD, 2009.

H. Zhai, X. Cheng, J. Guo, H. Xu, and Y. Liu. A Novel Framework for Related Entities Finding: ICTNET at TREC 2009 Entity Track. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, Gaithersburg, MD, 2009.

W. Zheng, S. Gottipati, J. Jiang, and H. Fang. UDEL/SMU at TREC 2009 Entity Track. In *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, Gaithersburg, MD, 2009.

Run	Group	Type	WP	Ext.	NDCG@R	P@10	#rel	#pri
KMR1PU	Purdue	auto	Y	Y	**0.3061**	**0.2350**	126	61
KMR3PU	Purdue	auto	Y	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, using the default evaluation setting (HP-only). 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 homepages, and number of primary retrieved homepages. Highest scores for each metric are in boldface.

Run	(1) HP-only					(2) WP-only				(3) HP+WP				
	NDCG	P@10	#rel	#pri	+NAME	NDCG	P@10	#rel	#pri	NDCG	P@10	#rel	#pri	+NAME
KMR1PU	**0.3061**	**0.2350**	126	61	**0.3244**	0.3365	0.**3950**	3	90	0.3044	**0.4850**	129	151	0.3325
KMR3PU	0.3060	**0.2350**	126	61	0.3243	**0.3372**	**0.3950**	4	90	**0.3048**	**0.4850**	130	151	**0.3328**
KMR2PU	0.2916	**0.2350**	115	56	0.3108	0.3236	0.3800	3	87	0.2877	0.4750	118	143	0.3156
uogTrEpr	0.2662	0.1200	**347**	**79**	0.2521	0.1821	0.2250	6	73	0.2438	0.2550	**353**	**152**	0.2367
uogTrEc3	0.2604	0.1200	331	75	0.2480	0.1847	0.1950	7	74	0.2421	0.2300	338	149	0.2352
uogTrEbl	0.2510	0.1050	344	75	0.2392	0.1874	0.1950	7	73	0.2323	0.2250	351	148	0.2268
uogTrEdi	0.2502	0.1150	343	74	0.2390	0.1877	0.2050	7	71	0.2320	0.2400	350	145	0.2270
ICTZHRun1	0.2103	**0.2350**	80	70	0.2213	0.2121	0.2550	4	63	0.1875	0.3450	84	133	0.1996
NiCTm3	0.1907	0.1550	99	64	0.1991	0.1742	0.1900	6	67	0.1866	0.2800	105	131	0.1739
NiCTm2	0.1862	0.1750	99	61	0.1922	0.1845	0.2100	6	66	0.1865	0.3100	105	127	0.1720
NiCTm1	0.1831	0.1450	98	63	0.1919	0.1766	0.2000	5	66	0.1814	0.2850	103	129	0.1688
UAmsER09Ab1	0.1773	0.0450	198	19	0.1477	0.1559	0.0300	**63**	20	0.1823	0.0700	261	39	0.1430
tudpw	0.1351	0.0950	108	42	0.1360	0.2836	0.2300	32	80	0.1767	0.2400	140	122	0.1820
tudpwkntop	0.1334	0.1150	108	41	0.1386	0.2826	0.2600	32	79	0.1778	0.2700	140	120	0.1877
NiCTm4	0.1280	0.0950	87	45	0.1263	0.1919	0.2200	8	79	0.1544	0.2550	95	124	0.1354
UAmsER09Co	0.1265	0.0400	87	23	0.1035	0.0487	0.0200	26	39	0.1401	0.0600	113	62	0.1098
tudwtop	0.1244	0.0650	125	50	0.1245	0.2551	0.2150	43	**94**	0.1672	0.2250	168	144	0.1749
tudwebtop	0.1218	0.0600	103	28	0.1081	0.0000	0.0000	0	0	0.1009	0.0600	103	28	0.0859
basewikirun	0.1043	0.0500	77	40	0.0987	0.1843	0.1000	51	54	0.1324	0.1200	128	94	0.1223
PRIS3	0.0892	0.0150	48	3	0.0807	0.0656	0.0350	7	14	0.1030	0.0500	55	17	0.0864
wikiruncats	0.0805	0.0550	77	40	0.0753	0.1740	0.1550	52	56	0.1208	0.1650	129	96	0.1153
PRIS1	0.0729	0.0100	40	2	0.0650	0.0779	0.0400	18	15	0.0971	0.0500	58	17	0.0793
PRIS2	0.0712	0.0050	61	1	0.0623	0.1199	0.0600	32	25	0.1116	0.0650	93	26	0.0907
UALRCB09r4	0.0666	0.0200	15	4	0.0523	0.0000	0.0000	0	0	0.0516	0.0200	15	4	0.0392
PRIS4	0.0642	0.0150	70	4	0.0589	0.0973	0.0550	21	19	0.0898	0.0700	91	23	0.0740
UIauto	0.0575	0.0100	64	3	0.0563	0.0324	0.0450	2	13	0.0559	0.0500	66	16	0.0568
uwaterlooRun	0.0531	0.0100	55	5	0.0453	0.0148	0.0050	1	9	0.0513	0.0150	56	14	0.0415
UdSmuTP	0.0488	0.0000	102	10	0.0458	0.0538	0.0300	18	45	0.0689	0.0300	120	55	0.0643
UALRCB09r3	0.0485	0.0100	9	2	0.0382	0.0000	0.0000	0	0	0.0380	0.0100	9	2	0.0289
UdSmuCM50	0.0476	0.0100	96	8	0.0423	0.0379	0.0500	20	39	0.0590	0.0550	116	47	0.0520
UdSmuCM	0.0446	0.0100	102	13	0.0412	0.0344	0.0200	17	42	0.0570	0.0300	119	55	0.0506
UdSmuTU	0.0430	0.0000	98	13	0.0392	0.0399	0.0150	20	39	0.0573	0.0150	118	52	0.0510
BITDLDE09Run	0.0416	0.0200	81	9	0.0379	0.0984	0.1250	6	47	0.0705	0.1250	87	56	0.0731
UALRCB09r2	0.0399	0.0150	7	3	0.0317	0.0000	0.0000	0	0	0.0316	0.0150	7	3	0.0243
UALRCB09r1	0.0392	0.0050	8	1	0.0316	0.0111	0.0050	1	1	0.0368	0.0100	9	2	0.0282
UIqryForm	0.0251	0.0000	4	0	0.0202	0.0000	0.0000	0	0	0.0224	0.0000	4	0	0.0172
UIqryForm3	0.0189	0.0000	16	0	0.0167	0.0204	0.0100	0	2	0.0221	0.0100	16	2	0.0216
ilpsEntBL	0.0161	0.0000	30	1	0.0140	0.0080	0.0200	0	9	0.0174	0.0200	30	10	0.0169
ilpsEntcr	0.0161	0.0000	30	1	0.0140	0.0080	0.0200	0	9	0.0174	0.0200	30	10	0.0169
ilpsEntem	0.0128	0.0000	17	0	0.0112	0.0100	0.0200	0	6	0.0160	0.0200	17	6	0.0156
ilpsEntcf	0.0105	0.0000	25	0	0.0091	0.0036	0.0000	0	3	0.0097	0.0000	25	3	0.0085

Table 3: Results of all submitted runs using alternative evaluation scenarios: (1) official qrels (for each record, only the HP1..3 fields are considered), (2) Wikipedia-only runs (for each record, only the WP field is considered), and (3) combined (HP1..3 and WP fields are all considered). To save space, we write NDCG@R as NDCG when HP/WP fields are considered; +NAME denotes NDCG@R when the NAME field is also taken into account. P10, #rel, and #pri are as before. The ordering of runs corresponds to those of Table 2. Highest scores for each metric are in boldface.