Overview of the TREC Tasks Track 2016

Manisha Verma[1], Evangelos Kanoulas[2], Emine Yilmaz[1], Rishabh Mehrotra[1], Ben Carterette[3], Nick Craswell[4], Peter Bailey[4]

University College London[1]
University of Amsterdam[2]
University of Delaware[3]
Microsoft[4]

1 Introduction

Research in Information Retrieval has traditionally focused on serving the best results for a single query, ignoring the reasons (or the task) that might have motivated the user to submit that query. Often times search engines are used to complete complex tasks (information needs); achieving these tasks with current search engines requires users to issue multiple queries. For example, booking travel to a location such as London could require the user to submit various queries such as flights to London, hotels in London, points of interest around London etc.

Standard evaluation mechanisms focus on evaluating the quality of a retrieval system in terms of the relevance of the results retrieved, completely ignoring the fact that user satisfaction mainly depends on the usefulness of the system in helping the user complete the actual task that led the user issue the query. Similar to Tasks Track 2015 [1], Tasks Track 2016 is an attempt investigate quality of retrieval systems in terms of (1) how well they can understand the underlying task that led the user submit a query, and (2) how useful they are for helping users complete their tasks.

In this overview, we first summarise the three categories of evaluation mechanisms used in the track and briefly describe the corpus, topics, and tasks that comprise the test collections. We then give an overview of the runs submitted to the Tasks Track and present evaluation results and analysis.

2 Evaluation Goals

This year we kept the same evaluation goals as 2015[1]. The three evaluation goals are: (1) Task understanding, (2), Task completion, and (3) Adhoc Retrieval. Participants were provided with a set of 50 queries, together with the Freebase ID [2] for each entity in these queries. Participants were asked to use these 50

[1] Ranking stability subtask was proposed, however, due to shortage of resources and time was not evaluated in 2016.

[2] https://developers.google.com/freebase/

queries and Clueweb12 document collection (A or B) for each of these tasks. The details of each task and metrics are described in following subsections.

2.1 Task Understanding

The goal of this task is to test whether systems can understand the possible tasks users might be trying to achieve given a query. For this task, participants were asked to submit key phrases that captured user's search task given this query.

For each query, the participants were asked to submit a ranked list of up to 1000 key phrases that represent the set of all tasks the user of query may be looking for. For example, for the query ”"hotels in London”, some relevant key phrases can be: “cheap hotels in London”, “reviews of hotels in London”, “hotels in London city centre”, etc. The goal of this task is to return a ranked list of key phrases that provide a complete coverage of tasks for each query, while avoiding redundancy.

Evaluating the coverage and relevance of the tasks submitted by the participants requires that a set of “gold standard” tasks that cover the set of all possible tasks are identified in advance. These gold standard tasks were constructed by the organizers, but were not be provided to the participants until the evaluation results were disclosed (i.e. post completion of track).

In order to guarantee higher coverage of tasks and be fair to all participants, tasks were developed based on information extracted from the logs of a commercial search engine, as well as by pooling the key phrases submitted by the participants. An example set of tasks for the query “hotels in London” may be

- hotels in in London [price]
- hotels in London [location]
- hotels [reviews] in London
- London [other accommodation]
- hotels [in locations around] London

Given the gold standard tasks, each key phrase submitted by the participants were judged by NIST assessors with respect to each of the gold standard tasks by using a three level judging scheme:

- **Highly relevant (2):** The key phrase completely describes the task and could be used as a query to a search engine to complete the task.
- **Relevant (1):** The key phrase somehow describes the task but not fully, it can be used as a query to achieve the task but there are better queries than that.
- **Non Relevant (0):** The key phrase is not relevant to the task and cannot be used to complete it.

In the aforementioned example, the key phrase “cheap hotels in London city centre” would be judged as relevant to both “hotels in London [price]” and “hotels in London [location]. Similar to Tasks Track 2015, the quality of each ranked list has been evaluated using diversity metrics such as ERR-IA [2] and α-NDCG [3].

2.2 Task Completion

The aim of this evaluation goal is to test the usefulness of a retrieval system in helping a user complete her search task.

Participants had to retrieve a ranked list of up to 1000 documents that could be relevant to any task a user may be trying to achieve given a query. The ranked lists provided by the participants were evaluated for diversity and relevance with respect to predefined list[3] of possible tasks given a query.

Each document submitted by the participants (up to a certain rank, 20 in 2016) has been assessed in terms of its *usefulness* to complete each possible 'gold standard" task using a three level judging scheme:

- **Key (2):** The document is essential towards the completion of the task. The document is enough on its own to complete the task.
- **Useful (1):** The document is useful towards the completion of the task. However, more documents need to be investigated in order to complete the task.
- **Not Useful (0):** The document is not useful towards completion of the task.

Similar to Tasks Track 2015, we also obtained *relevance* judgements for pooled list of documents from NIST assessors. Each document was labelled using four level judging scheme:

- **Highly Relevant (2):** The page contains significant amount of information about the task.
- **Relevant (1):** The content of this page provides some information on the task, which may be minimal.
- **Non Relevant (0):** The content of this page does not provide useful information about the task.
- **Spam (-2):** This page does not appear to be useful for any reasonable purpose; it may be spam or junk.

Given these judgements, similar to Task Understanding evaluation, the quality of each ranked list has been evaluated using diversity metrics: ERR-IA [2] and α-NDCG [3].

2.3 Adhoc Retrieval

For comparison purposes, we continued to have a traditional Web adhoc evaluation mechanism this year as well [3]. Participants were asked to submit a ranked list of up to 1000 documents for each topic. Participants were provided a short description of query along with query text and Freebase ID of entities in query.

Similar to Task track 2015, we used Task completion judgements for relevance to evaluate quality of the runs submitted by the participants. We ignored the usefulness category for adhoc evaluation. For the task completion, given

[3] These subtasks were manually designed by organizers in 2016.

a query, NIST assessors assigned document multiple relevance grades, each for possible tasks provided in ground truth. For Adhoc, we derived document relevance by using the *maximum* relevance label assigned for that document over all possible tasks.

Once these relevance judgements were obtained, ERR and NDCG were used as the primary metrics for evaluation, similar to previous years' Web Track [4].

3 Participants and Runs

Table 1 summarizes the participation in Tasks track. Overall we received 24 runs from four groups: 12 task understanding and 9 task completion and 3 adhoc runs. Number of submissions for each task from every group is given below:

- Webis Group (Webis): 3 adhoc, 3 completion and 3 understanding runs.
- University of Delaware (Udel-fang): 3 completion and 3 understanding runs.
- University of Delaware (Udel): 3 understanding and 3 completion runs.
- University of Stavanger (UiS) : 3 understanding runs.

Table 1: Tasks Track 2016 participation

Task	Understanding	Completion	Adhoc
Groups	4	3	1
Runs	12	9	3

Overall, this year participants submitted more runs as compared to 2015. In 2015, only 21 runs were submitted, of which 11 were task understanding, 6 were task completion and 4 were adhoc runs. This year we had similar number of groups participating in the track.

4 Evaluation Results

This year all 50 topics were judged for three tasks as compared to 2015 Tasks track. Detailed results per task are provided in following subsections.

4.1 Task Understanding

Task understanding runs were evaluated depth-20 pools of key phrases. Each key-phrase was labelled using judging guidelines described in Section 2.1.

This year 47689 keywords were evaluated across 50 queries, of which 34246 were 'not-relevant', 12003 were assigned 'relevant' and only 1440 were assigned 'highly-relevant' by the assessors.

We evaluate each task understanding submission with α-NDCG@20 and ERR-IA@20, where ERR-IA@20 is the primary metric. For each run, we report the average α-NDCG@20 and ERR-IA@20 for all topics. Participant results

Table 2: Task understanding results

Group	Run	ERR-IA@20	αNDCG@20
UiS	UiS_8	0.57	0.70
UiS	UiS_4	0.53	0.66
Webis	webis1	0.51	0.68
Webis	webis3	0.51	0.67
Webis	webis2	0.50	0.67
UiS	UiS_9	0.47	0.61
Udel	udelRun3	0.41	0.56
Udel	udelRun1	0.40	0.52
Udel-fang	udelRun4	0.40	0.52
Udel-fang	udelRun6	0.38	0.50
Udel-fang	udelRun5	0.36	0.46
Udel	udelRun2	0.35	0.45

for task understanding are shown in 2, where evaluation results are sorted on ERR-IA@20 in descending order.

This year 12 runs were submitted compared to 11 submitted last year [1]. The maximum ERR@20 and α-NDCG@20 in 2015 was 0.471 and 0.573 respectively. This year, however, participants from UiS have achieved much higher values, even though more topics were evaluated this year[4]. The minimum ERR@20 and α-NDCG@20 is also higher this year, they were 0.234 and 0.313 respectively in 2015.

4.2 Task completion results

Adhoc and task completion runs were evaluated using depth-20 pools of documents submitted by participants. Each document was labelled in terms of usefulness and relevance to each task, based on the judging schemes described in Section 2.2.

This year 33525 documents were labelled for 50 queries for both relevance and usefulness. For usefulness, 282 documents across 50 topics that were assigned 'Key' label. There were 3799 and 29444 documents marked 'useful' and 'not-useful' across 50 topics by NIST assessors this year. For relevance, 394 documents were marked 'highly-relevant', 4673 were labelled as 'relevant', 24184 were assigned 'non-relevant' and remaining were marked as 'spam' by assessors.

Given the judgements based on usefulness and relevance, both α-NDCG and ERR-IA metrics were computed at rank 10, focusing on ERR-IA at rank 10 computed using judgements based on usefulness as the primary metric. Table 4 shows the evaluation results for this category, sorted on ERR-IA in descending order. All participants used Category A collection of Clueweb documents.

This year some documents were not rendered properly at time of evaluation at NIST. Such documents were assigned -3 label by the assessors. Since, document relevance was not known, we ignored these documents from evaluating for relevance. However, these documents were marked 'not-useful' by assessors. Of 33525 documents, 3512 documents were not rendered properly and have been ignored for evaluations in Table 3.

[4] only 34 topics were evaluated in 2015

Table 3: Task Completion (Relevance) results

Group	Run	ERRIA@10	αNDCG@10
Udel-fang	udelRun5C	0.293	0.406
Udel-fang	udelRun4C	0.286	0.398
Udel	udelRun1C	0.284	0.395
Webis	webisC2	0.274	0.418
Udel	udelRun3C	0.267	0.392
Udel	udelRun2C	0.263	0.366
Webis	webisC1	0.259	0.396
Udel-fang	udelRun6C	0.257	0.372
Webis	webisC3	0.243	0.364

Table 4: Task Completion (Usefulness) results

Group	Run	ERRIA@10	αNDCG@10
Udel-fang	udelRun5C	0.243	0.347
Udel-fang	udelRun4C	0.231	0.334
Udel	udelRun2C	0.230	0.323
Udel	udelRun1C	0.229	0.330
Webis	webisC2	0.223	0.349
Udel	udelRun3C	0.222	0.339
Udel-fang	udelRun6C	0.215	0.320
Webis	webisC1	0.214	0.335
Webis	webisC3	0.199	0.305

Table 3 shows the evaluation results based on judgements based on *document relevance*. The ranking of systems when evaluation metrics are computed based on relevance versus usefulness differ in some positions. Pearson's ρ correlation between relevance and usefulness based evaluation is 0.910 and 0.922 for ERR-IA@10 and α-NDCG@10 respectively. Kendall tau Rank Correlation between relevance and usefulness based evaluation is 0.77 (p-val = 0.004) and 0.72 (p-val=0.009) for ERR-IA@10 and α-NDCG@10 respectively.

Figure 1 shows how the ranking of systems change when evaluation metrics are computed using usefulness judgements (x axis in the plots) versus using judgements in terms of relevance (y axis in the plots). As it can be seen in these plots, α-NDCG@10 has higher variance in scores than ERR-IA@10.

On comparison to Tasks track 2015, participants achieved lower ERR-IA@10 and α-NDCG@10 this year. In Tasks track 2015, maximum ERR-IA@10 and α-NDCG@10 for usefulness based evaluation was 0.442 and 0.518 respectively. However, this year participants the maximum maximum ERR-IA@10 and α-NDCG@10 for usefulness is 0.243 and 0.347 respectively.

4.3 Adhoc Retrieval results

In order to evaluate the quality of Adhoc Retrieval runs, the judgements obtained for Task Completion were used in the way described in Section 2.3. ERR and NDCG at rank 10 values were then computing, using ERR at rank 10 as the primary metric. Table 5 shows the evaluation results for the adhoc runs,

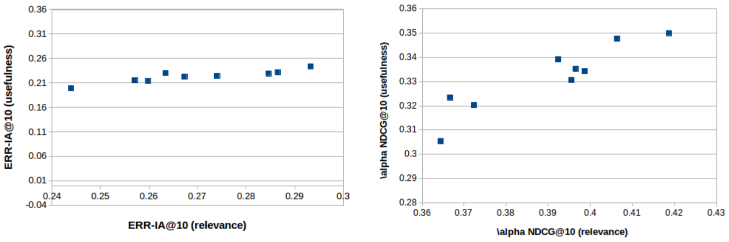

Figure 1: Comparison of evaluation results based on (left) ERR-IA, and (right) α-NDCG metrics when judgements based on usefulness versus relevance are used.

sorted in decreasing relevance in terms of the ERR scores.

The metric values for the adhoc runs are low but similar to what participants achieved in 2015 where maximum (minimum) ERR@10 and NDCCG@10 were 0.124 (0.001) and 0.455 (0.003) respectively. When the evaluation results for runs submitted by the same groups for Task Completion and Adhoc are compared, the evaluation results seem much higher for Task Completion.

Table 5: Adhoc results

Group	Run	ERR@10	NDCG@10
Webis	webisA2	0.12	0.44
Webis	webisA3	0.11	0.43
Webis	webisA1	0.11	0.42

5 Conclusion

The TREC 2016 Tasks Track ran for second year to build test collections to evaluate retrieval systems on relevance and usefulness of retrieved documents for a given user search task. We organized the track with three tasks: task understanding, completion and adhoc retrieval. We did not observe a significant rise in number of participants, however, the number of runs submitted this year were higher than 2015. Task understanding and task completion task received over 20 submissions from four participants. Submitted runs achieved higher accuracy for task understanding completion tasks. Tasks Track shall be organized again in 2017 with slight modifications to existing tasks.

References

[1] Emine Yilmaz, Evangelos Kanoulas, Manisha Verma, Ben Carterette, Nick Craswell, and Rishabh Mehrotra. Overview of the trec 2015 tasks track. 2015.

[2] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 621–630. ACM, 2009.

[3] Ian Soboroff, Iadh Ounis, Craig Macdonald, and Jimmy Lin. Overview of the trec-2012 microblog track. In *TREC*, volume 2012, page 20, 2012.

[4] Kevyn Collins-Thompson, Craig Macdonald, Paul Bennett, Fernando Diaz, and Ellen M Voorhees. Trec 2014 web track overview. Technical report, DTIC Document, 2015.