Overview of the TREC 2015 Tasks Track

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

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

November 4, 2015

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 usefullness of the system in helping the user complete the actual task that led the user issue the query. The TREC 2015 Tasks Track is an attempt in devising mechanisms for evaluating 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

The TREC 2015 Tasks Track consists of three evaluation goals: (1) Task understanding, (2), Task completion, and (3) Adhoc. Participants were provided with a set of 50 queries, together with the Freebase ID for each entity in these queries. The same queries were used for each of these tasks and Clueweb12 was used as the main corpus.

Details of each evaluation goal, as well as the metrics used for each goal are shown below.

2.1 Task Understanding

The aim of this evaluation goal is to test whether systems can understand the possible tasks users might be trying to achieve given a query. For this goal, the participants were asked to submit key phrases that represent the possible task the user may be trying to achieve 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 a user who submitted the 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 are out.

In order to guarantee the 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 London [price]
- hotels in London [location]
- hotels [reviews] in London
- [other accommodation] in London
- hotels [in locations around] London

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

- **Highly relevant:** The key phrase completely describes the task and could be used as a query submitted to the search engine to complete the task.
- **Relevant:** 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:** 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 judgedl be relevant to both "hotels in London [price]" and "hotels in London [location].

Given these judgments, the quality of each ranked list will then be evaluated using diversity metrics such as ERR-IA and α-NDCG [1].

2.2 Task Completion

The aim of this evaluation goal is to test the usefulness of a retrieval system in helping a user achieve a task. Participants were asked to retrieve a ranked list of documents that could be relevant to any task a user may be trying to achieve given a query.

For each query, the participants are expected to submit 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 in terms of the diversity and relevance of documents they have submitted with respect to the set of possible tasks the user may be trying to achieve given a query.

Each document submitted by the participants will then be assessed in terms of its *usefulness* to complete each possible "gold standard" task by using a three level judging scheme:

- **Key:** The document is essential towards the completion of the task. The document is enough on its own to complete the task.
- **Useful:** 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:** The document is not useful towards the completion of the task.

This is the firs time judgments in terms of usefulness have been used in evaluating quality of retrieval systems. Hence, for comparison purposes we also obtain relevance judgments by asking NIST assessors to label each document in terms of its relevance to the query. Fort this purpose, we used a four level judging scheme:

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

Given these judgments, similar to Task Understanding, the quality of each ranked list was then evaluated using diversity metrics such as ERR-IA and α-NDCG [1]. The primary judgments that were used in this evaluation category were based on usefulness.

Table 1: Tasks Track 2015 participation

Task	Understanding	Completion	Adhoc
Groups	5	3	2
Runs	11	6	4

2.3 Adhoc

For comparison purposes, we continued to have a traditional Web ad-hoc evaluation mechanism this year as well [1]. Participants were asked to submit a ranked list of up to 1000 documents.

For evaluating the quality of the runs submitted, the judgments that were obtained for the Task Completion Task were used, ignoring the usefulness category, and focusing on relevance. For the task completion category, a document is judged for each possible task, given a query. For Adhoc, each document is assigned a single relevance value, which is the maximum relevance label assigned for that document over all possible tasks.

Once these relevance judgments were obtained, ERR and NDCG were used as the primary metrics for evaluation, similar to previous years' Web Track [2].

3 Participants and Runs

Table 1 summarizes the participation in Tasks track. In total, we recieved 21 runs from five groups, consisting of 11 task understanding runs, 6 task completion runs, and 4 adhoc runs. In terms of the corpus, only one group (WHU) used Category B subset of Clueweb12 for all its runs. Remaining groups submitted runs using Category A corpus of Clueweb12.

The groups participated in the TREC 2015 Tasks Track, as well as the evaluation categories they participated in can be seen below:

- Microsoft Research (MSR): 1 task understanding run.
- Carnegie Mellon University (OAQA): 3 task understanding runs.
- University of Delaware (UDEL): 3 task understanding, 1 task completion runs.
- Webis Group (WEBIS): 1 task understanding, 3 task completion, and 3 adhoc runs.
- WHU group (WHU): 3 task understanding, 2 task completion and 1 adhoc runs.

4 Evaluation Results

4.1 Task Understanding Results

For the Task Understanding evaluation category, depth-20 pools of the key phrases submitted by the participants were constructed and each key phrase was labelled based on the judging scheme described in Section 2.1. Due to the limited judgment budget available at NIST, only 34 topics were judged for Task

Table 2: Task Understanding results

Group	Run	Category	ERR-IA@20	α-NDCG@20
Whu	TOPIC_RUN3	B	0.471	0.573
Udel	udelRun2	A	0.454	0.565
Webis	webis1	A	0.350	0.453
Udel	udelRun1	A	0.347	0.404
Whu	TOPIC_RUN2	B	0.336	0.413
Msr	MSRTasksQUrun3	A	0.303	0.359
Whu	NP_TU	B	0.299	0.375
Udel	udelTTTUAOL	A	0.269	0.327
CMU	rsf	A	0.260	0.335
CMU	lsf	A	0.249	0.324
CMU	lsfs	A	0.234	0.313

Table 3: Task Completion (Usefulness) results

Group	Run	Category	ERR-IA@10	α-NDCG@10
Udel	udelRun2CSpam	A	0.442	0.518
Webis	webisC2	A	0.254	0.300
Whu	TOPIC_RUN3_TC	B	0.177	0.210
Webis	webisC3	A	0.120	0.134
Webis	webisC1	A	0.096	0.108
Whu	TOPIC_RUN2_TC	B	0.014	0.021

Understanding. Hence, the evaluation results reported in this section mainly focus on these 34 topics.

Given the lables of key phrases, both α-NDCG and ERR-IA metrics were computed at rank 20, focusing on ERR-IA at rank 20 as the primary metric. Table 2 shows the evaluation results for this category, sorted in terms of decresing ERR-IA values.

4.2 Task Completion Results

For Task Completion, depth-10 pools of documents were constructed and each document was labelled in terms of *usefulness* and *relevance* to each task, based on the judging schemes described in Section 2.2. Similar to the Task Understanding evaluation category, only 35 topics were labelled for Task Completion due to limited judgment resources available at NIST.

Given the judgments 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 3 shows the evaluation results for this category, sorted in terms of decresing ERR-IA values.

Table 4 shows the evaluation results based on judgments in terms of relevance. The ranking of systems when evaluation metrics are computed based on relevance versus usefulness seem to be identical. Figuremetriccomparison shows how the ranking of systems change when evaluation metrics are computed using judgments in terms of usefulness (x axis in the plots) versus using judgments in

Table 4: Task Completion (Relevance) results

Group	Run	Category	ERR-IA@10	α-NDCG@10
Udel	udelRun2CSpam	A	0.469	0.554
Webis	webisC2	A	0.278	0.334
Whu	TOPIC_RUN3_TC	B	0.232	0.293
Webis	webisC3	A	0.126	0.149
Webis	webisC1	A	0.108	0.122
Whu	TOPIC_RUN2_TC	B	0.024	0.035

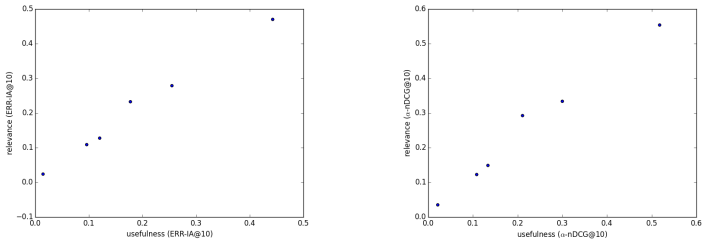

Figure 1: Comparison of evaluation results based on (left) ERR-IA, and (right) α-NDCG metrics when judgments based on usefulness versus relevance are used.

terms of relevance (y axis in the plots). As it can be seen in these plots, even though the values of evaluation metrics change when the different judgments are used, the ranking of the systems based on the two types of judgment seem identical.

4.3 Adhoc Retrieval Results

In order to evaluate the quality of Adhoc Retrieval runs, the judgments 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, sorted in decreasing relevance in terms of the ERR scores.

The metric values for the adhoc runs seem quite low. 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. When the documents retrieved are compared, the runs submitted for Task Completion do not have much overlap with the runs submitted for Adhoc. Hence, the low evaluation values for Adhoc seems to be due to the nature of the runs submitted for this evaluation category.

5 Conclusions

The TREC 2015 Tasks Track was the first attempt in building test collections for evaluating the usefulness of retrieval systems in terms helping people achieve their search tasks. Since this was the first year of the track, the number of participants was not very high. However, for the task understanding and the task

Table 5: Adhoc results

Group	Run	Category	ERR@10	NDCG@10
Whu	NORM_RUN1	B	0.124	0.455
Webis	webisA2	A	0.011	0.024
Webis	webisA3	A	0.006	0.019
Webis	webisA1	A	0.001	0.003

completion evaluation categories, the submitted systems seem to have achieved reasonable performance.

In 2016, the TREC Tasks Track will be running one more time. Building upon the findings of the 2015 track, in 2016 we plan to further continue our attempts in devising evaluation mechanisms for meausing the usefulness of a system in helping users achieve a task.

References

[1] Charles L. A. Clarke, Nick Craswell, and Ellen M. Voorhees. Overview of the TREC 2012 web track. In *Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6-9, 2012*, 2012.

[2] Kevyn Collins-Thompson, Paul N. Bennett, Fernando Diaz, Charlie Clarke, and Ellen M. Voorhees. TREC 2013 web track overview. In *Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19-22, 2013*, 2013.