TREC 2017 Tasks Track Overview

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

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

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; 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 topical 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. The TREC 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.

A number of application areas, beyond task-based retrieval, could benefit from such an evaluation framework and the constructed test collections, including, but not limited to, digital assistants tasks, query suggestions, session-based retrieval, exploratory search, entity-based search.

In this paper, we first summarize the three categories of evaluation mechanisms used in the track and briefly describe the corpus, topics, and tasks that comprise the test collection. We then give an overview of the runs submitted to the 2017 Tasks Track and present the evaluation results.

2 Evaluation Goals

This year we kept the same evaluation goals as in 2015 and 2016, namely: (1) Task understanding, (2) Task completion, and (3) Ad-hoc retrieval.

Participants were provided with a set of 50 queries, together with the Freebase ID[1] for each entity in these queries. For instance, the following query was one of the 50 provided to participants:

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

```
<task id = "1">
        <query>build a hydroponic garden</query>
        <freebase entity = "1">
                <id>/m/03nwn</id>
                <text>hydroponic</text>
        </freebase>
        <freebase entity = "2">
                <id>/m/0bl0l</id>
                <text>garden</text>
        </freebase>
</task>
```

The queries were manually constructed by the organizers of the track so that they have the following two properties: (a) they represent an action-oriented task, and (b) the task can be decomposed in a sequence of subtasks.

The corpus of the track remained the ClueWeb12 document collection. To alleviate the burden of indexing the collection on participants, the category B subset was indexed by Dr. Guido Zuccon and his team in Elastic 5.3 using stemming and stopping. The default retrieval algorithm was used, i.e. BM25. The Elastic Instance was deployed onto Microsoft Azure Platform (and is supported by a Microsoft Azure Research Award). Participants could query the API by sending CURL requests or writing a script. The spam scores from Waterloo Spam Rankings (http://www.mansci.uwaterloo.ca/~msmucker/cw12spam/) were also provided.

2.1 Task Understanding

The goal of *task understanding* is to test whether systems can understand the task and the possible subtasks users might be trying to complete given a query. For each query, participants were asked to submit a ranked list of up to 100 key phrases that represent the set of all subtasks the user needs to complete so that a complete coverage of potential subtasks is achieved, while avoiding redundancy.

Evaluating the coverage and relevance of the key-phrases submitted by the participants requires that a set of "gold standard" subtasks is identified in advance. These gold standard subtasks 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 achieve high coverage of tasks, tasks were developed by examining the corresponding to the query pages of WikiHow and the query suggestions provided by commercial search engines. Further, a task named "other" was added in the case participants submitted a relevant subtask not covered by the gold standard set. An example of a constructed topic can be seen below:

build a hydroponic garden [How do I construct a hydroponic garden and grow plants?]

- What are the different types of hydroponic systems to choose from, their pros and cons?
- What hydroponic system are used in commercial greenhouses?
- What material is needed to build a hydroponic system myself?
- Which plants can be grown in a hydroponics system?
- How long does it take to grow them?
- What nutrients to test the water for?
- How to prepare the seeds?
- How often to water the plants?
- What plant food to add and how often?
- What are the light and temperature requirements?

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.

Similar to 2015 and 2016, the quality of each ranked list was evaluated using diversity metrics such as ERR-IA [1] and α-NDCG [2].

2.2 Task Completion

The aim of this evaluation goal is to test the usefulness of a retrieval system in helping users complete their search task. Participants had to retrieve a ranked list of up to 1000 documents, to be evaluated by a three level "usefulness" judging scheme (key, useful, and not useful), and by four level "topical relevance" judging scheme (highly relevant, relevant, not relevant, and spam).

In 2017 no team participated in *task completion* hence no test collection was constructed.

2.3 Adhoc Retrieval

For comparison purposes, we continued to have a traditional Web ad-hoc evaluation mechanism this year as well [2]. 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.

In 2017 no team participated in *task completion* hence no test collection was constructed.

3 Participants and Runs

This year four runs were submitted in total from two participating groups: Beijing University of Technology (BJUT), and Siena College Institute for AI (SienaCLTeam). All the runs were task understanding runs, so only the key phrase qrels was constructed.

4 Evaluation Results

4.1 Task Understanding

Task understanding runs were evaluated depth-30 pools of key phrases. Each key-phrase was labeled using judging guidelines described in Section 2.1.

This year 3623 keywords were evaluated across 50 queries, of which 1787 were non-relevant, 1586 were relevant to at least one task, and 623 were highly relevant to at least one task 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. We choose depth-20 so that numbers are comparable (to some extent) with the previous two years of the Tasks track. For each run, we report the average α-NDCG@20 and ERR-IA@20 for all topics. Participant results for task understanding are shown in 1, where evaluation results are sorted on ERR-IA@20 in descending order.

Table 1: Task understanding results

Group	Run	ERR-IA@20	αNDCG@20
BJUT	BJUT_1	0.452	0.628
BJUT	BJUT_2	0.469	0.644
SienaCLTeam	SienaCLTeamRun1	0.317	0.502
SienaCLTeam	SienaCLTeamRun2	0.511	0.624

This year 4 runs were submitted compared to 12 submitted in 2016 [3] and to 11 submitted in 2015 [4]. The maximum ERR@20 and α-NDCG@20 in 2015 were 0.471 and 0.573 respectively, in 2016 they were 0.57 and 0.70, respectively, while this year they are 0.51 (SienaCLTeamRun1) and 0.64 (BJUT_1).

5 Analysis on Task Understanding from 2015 to 2017

In this section we present a few results that span the three years of the TREC Tasks Track, and in particular the Task Understanding task. Figure 1 displays the α-nDCG@20 scores for all the participating runs over the three years. To the best of our knowledge none of the participants froze their developed system to reuse it in consequent years, hence it is not easy to reach any conclusions from the presented results.

What has been apparent from the three years of the Task Understanding task is that it is hard for assessors to judge the relevance of a key-phrase to a task. Further, the collection of judged key-phrases may not even be a re-usable one given that in 2015 73% of the assessed key-phrases were unique, in 2016 30% and in 2017 50%, which indicates that different systems return by and large different key-phrases, hence future systems will probably return key-phrases for which no label is available. Hence, the one question that we attempted to answer was whether we can pivot on document labels to evaluate key-phrases.

To answer the afore-mentioned question we built an Indri index using Krovetz stemmer and removing stopwords, and employed the standard Indri Language Model with Dirichlet smoothing ($\mu = 5000$). Each produced key-phrase was run against the index and the top-k results where considered. If a returned document was relevant (or useful) for a task then the key-phrase was considered relevant (useful) for that task. Hence, we constructed a key-phrase qrel using the document qrel. Figure 2 demonstrates the correlation between system ERR-IA@20 performance using the assessors key-phrase qrels and the key-phrase qrels built on the document qrels on the basis of relevance (left) and usefulness (right). As one can observe there is a clear correlation between the scores, however in both cases this correlation, as measured by Kendall's τ is less than 0.8, which is considered a rather strong correlation in the field.

6 Conclusion

The TREC 2017 Tasks Track ran for a third 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

Figure 1: α-nDCG@20 on Task Understanding for all participating systems over 2015 (blue), 2016 (green) and 2017 (red).

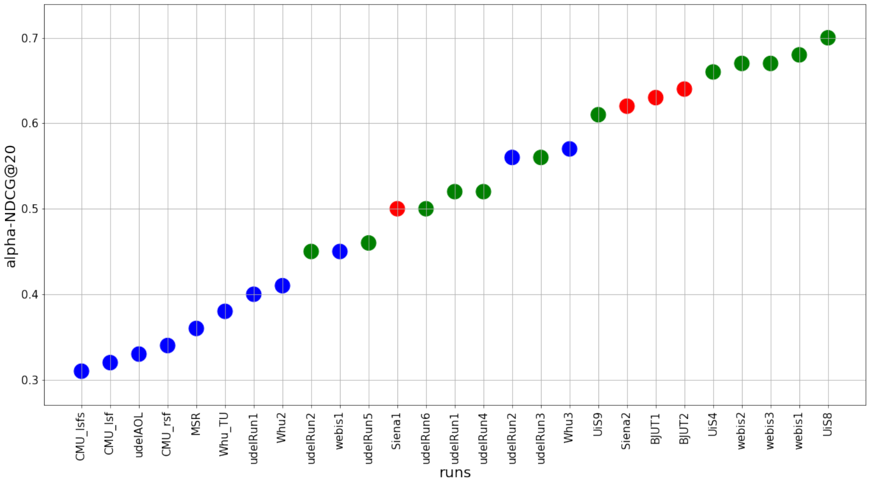
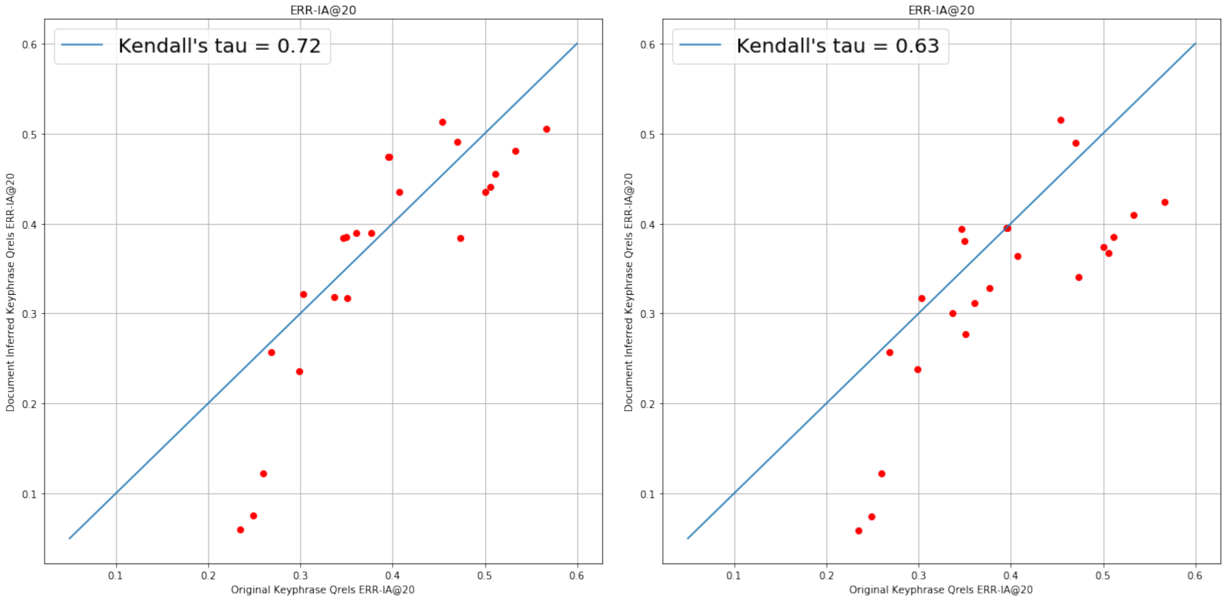


Figure 2: Correlation between system performance using the key-phrase qrels built on the document qrels on the basis of relevance (left) and usefulness (right).



understanding, completion and ad-hoc retrieval. We observe a significant drop in number of participants and the number of runs submitted this year, with 4 runs submitted in *task understanding*, while *task completion* and *ad-hoc retrieval* received no runs.

References

[1] 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.

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

[3] Manisha Verma, Evangelos Kanoulas, Emine Yilmaz, Rishabh Mehrotra, Ben Carterette, Nick Craswell, and Peter Bailey. Overview of the trec 2015 tasks track. 2016.

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