Overview of the CLEF dynamic search evaluation lab 2018

Kanoulas, E.; Azzopardi, L.; Yang, G.H.

DOI
10.1007/978-3-319-98932-7_31

Publication date
2018

Document Version
Final published version

Published in
Experimental IR Meets Multilinguality, Multimodality, and Interaction

Overview of the CLEF Dynamic Search Evaluation Lab 2018

Evangelos Kanoulas[1(✉)], Leif Azzopardi[2], and Grace Hui Yang[3]

[1] Informatics Institute, University of Amsterdam, Amsterdam, Netherlands
e.kanoulas@uva.nl
[2] Computer and Information Sciences, University of Strathclyde, Glasgow, UK
leif.azzopardi@strath.ac.uk
[3] Department of Computer Science, Georgetown University, Washington, D.C., USA
huiyang@cs.georgetown.edu

Abstract. In this paper we provide an overview of the CLEF 2018 Dynamic Search Lab. The lab ran for the first time in 2017 as a workshop. The outcomes of the workshop were used to define the tasks of this year's evaluation lab. The lab strives to answer one key question: how can we evaluate, and consequently build, dynamic search algorithms? Unlike static search algorithms, which consider user request's independently, and consequently do not adapt their ranking with respect to the user's sequence of interactions and the user's end goal, dynamic search algorithms try to infer the user's intentions based on their interactions and adapt their ranking accordingly. Session personalization, contextual search, conversational search, dialog systems are some examples of dynamic search. Herein, we describe the overall objectives of the CLEF 2018 Dynamic Search Lab, the resources created, and the evaluation methodology designed.

Keywords: Evaluation · Information retrieval · Dynamic search
Interactive search · Conversational search · User simulations
Query suggestion · Query generation · Meta-search · Result re-ranking

1 Introduction

Information Retrieval (IR) research has traditionally focused on serving the best results for a single query – so-called ad-hoc retrieval. However, users typically search iteratively, refining and reformulating their queries during a session. IR systems can still respond to each query in a session independently of the history of user interactions, or alternatively adopt their model of relevance in the context of these interactions. A key challenge in the study of algorithms and models that dynamically adapt their response to a user's query on the basis of prior interactions is the creation of suitable evaluation resources and the definition of suitable evaluation metrics to assess their effectiveness. Over the years various initiatives have been proposed which have tried to make progress on this long standing challenge.

P. Bellot et al. (Eds.): CLEF 2018, LNCS 11018, pp. 362–371, 2018.
https://doi.org/10.1007/978-3-319-98932-7_31

The TREC Interactive Track [12], which ran between 1994 and 2002, investigated the evaluation of interactive IR systems and resulted in an early standardization of the experimental design. However, it did not lead to a reusable test collection methodology. The TREC High Accuracy Retrieval of Documents (HARD) Track [1] followed the Interactive track, with the primary focus on single-cycle user-system interactions. These interactions were embodied in clarification forms which could be used by retrieval algorithms to elicit feedback from assessors. The track attempted to further standardize the retrieval of interactive algorithms, however it also did not lead to a reusable collection that supports adaptive and dynamic search algorithms. The TREC Session Track [2], which ran from 2010 through 2014, made some headway in this direction. The track produced test collections, where included with the topic description was the history of user interactions with a system, that could be used to improve the performance of a given query. While, this mean adaptive and dynamic algorithms could be evaluated for one iteration of the search process, the collection's are not suitable for assessing the quality of retrieval over an entire session. Further, algorithms that learn to optimize ranking over entire sessions are not feasible to be built. In 2015, the TREC Tasks Track [16,21] took a different direction, where the test collection provided queries for which all possible sub-tasks needed to be inferred, and the documents relevant to those sub-tasks identified. Even though the produced test collections could be used in testing whether a system could help the user to perform a task end-to-end, the focus was not on adapting and learning from the user's interactions as in the case of dynamic search algorithms. The Dynamic Domain Track [18], which ran in parallel to the Tasks Track, between 2015 and 2017, focused on domains of special interests, which usually produces complex and exploratory searches with multiple runs of user and search engine interactions. It was search in multiple runs of interactions where the participting systems were expected to adjust their systems dynamically based on the relevance judgments provided along the way. Figure 1 provides an overview of the task in this track. The user simulator, was practically feedback on the relevance of the returned documents and the passages in these document, to be used by retrieval systems in any way they could for the next iteration. Despite this over-simplification of what constitutes a user simulation the Dynamic Domain Track was the first benchmark collection that was designed to allow the development of dynamic retrieval systems in a controlled laboratory setting. The CLEF Dynamic Search Lab takes this effort one step forward, and instead of focusing on developing dynamic search algorithm, it focuses on developing effective user simulations.

In the related domain of dialogue systems, the advancement of deep and reinforcement learning methods has led to a new generation of data-driven dialog systems. Broadly-speaking, dialog systems can be categorized along two dimensions, (a) goal-driven vs. non-goal-driven, and (b) open-domain vs. closed domain dialog systems. Goal-driven open-domain dialog systems are in par with dynamic search engines: as they seek to provide assistance, advice and answers to a user over unrestricted and diverse topics, helping them complete their task, by not only taking into account the conversation history but optimizing the overall

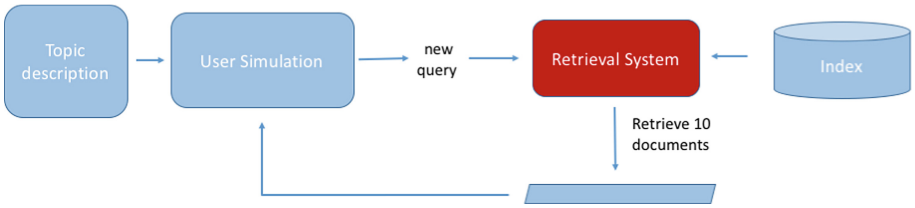

Fig. 1. TREC Dynamic Domain framework depicting the overall process and the task of ranking.

dialogue towards a specific user goal. While, a variety of corpora is available for training such dialog systems [15], when it comes to the evaluation, the existing corpora are inappropriate. This is because they only contain a static set of dialogues and any dialog that does not develop in a way similar to the static set cannot be evaluated. Often, the evaluation of goal-driven dialogue systems focuses on goal-related performance criteria, such as goal completion rate, dialogue length, and user satisfaction. Automatically determining whether a task has been solved however is an open problem, while task-completion is not the only quality criterion of interest in the development of dialog systems. Thus, simulated data is often generated by a simulated user [3,6,14]. Given a sufficiently accurate model of how user's converse, the interaction between the dialog system and the user can be simulated over a large space of possible topics. Using such data, it is then possible to deduce the desired metrics. This suggests that a similar approach could be taken in the context of interactive IR. However, while significant effort has been made to render the simulated data as realistic as possible [11,13], generating realistic user simulation models remains an open problem.

2 Lab Overview and Tasks

The focus of CLEF Dynamic Search is the evaluation and development of dynamic information retrieval algorithms that solve a user's complex task by continuously interacting with the user. When it comes to dynamic systems, the response of the system affects the user's next action. For instance, in dialog systems the response of the system highly affects how the user will continue the dialog, in search-based retrieval, the ranked results by the search engine highly affect the next query of the user. In these setups the evaluation of the systems becomes a really hard task, and it remains an open problem. Given the absence of a reliable evaluation framework in such a conversational/dynamic setup, the development of dynamic/conversational search engines also remains an open problem.

The CLEF 2018 Dynamic Search Lab focuses on the development of a dynamic search system evaluation framework, on the basis of the conclusions of the CLEF 2018 Dynamic Search Lab workshop [7,8]. The framework constitutes two agents – a question-agent and an answer-agent – which interact to solve

a user's task. The answer agent corresponds to the dynamic search system, or dynamic question answering system, while the question agent corresponds to the simulated user. In this 2018 edition, we focus on the development of a question agent, the goal of which is the production of effective queries given a verbose description of a user's information need (*query suggestion*). The question-agent will produce queries in a multi-round fashion; at every round a query is produced, submitted to the answer-agent, an Indri query language model over the New York Time corpus (more in Sect. 3), and the top-10 results of the query are fed back to the question-agent for the production of the next query. Potential participants are provided with a RESTful API to query the Indri index. The question-agent is running for 10 rounds, submitting in total 10 queries and obtaining 100 results retrieved by the answer agent. These 100 results, 10 per query, are then ranked in final ranking (*result composition*). Therefore, the lab offers two tasks to potential participants:

- Task 1: Query generation/suggestion
- Task 2: Results composition.

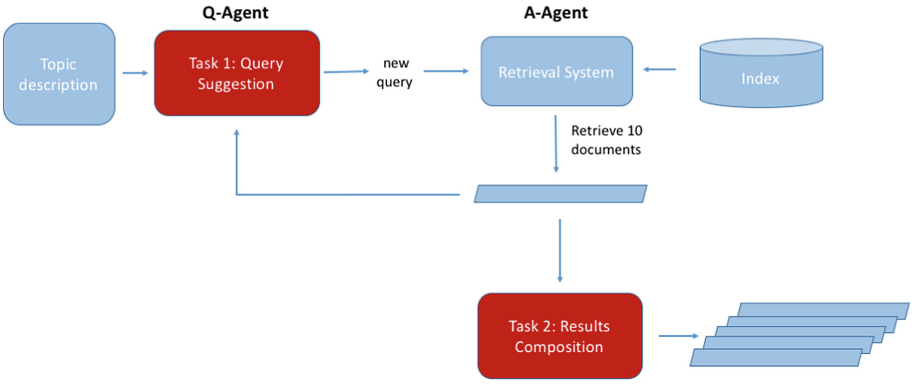

Fig. 2. CLEF 2018 Dynamic Search Lab framework depicting the overall process and the two tasks of query generation and result composition.

The two tasks are also depicted in red boxes in Fig. 2. In Task 1, participants were provided with a set of topics, split in a development and a test set. During the first search iteration, the participating runs in Task 1, were asked to generate a query to be submitted to a predefined, and provided by the participants retrieval system. In all follow up iterations, beyond the topic descriptions, the participating runs could use any information from the top-10 per-query returned documents resulted from all previous iterations. Participants were asked generate queries for 10 rounds of interactions. At the end of the 10 rounds, participants had to submit a run, with the following format:

```
TOPIC QUESTION DOCNO RANK SCORE RUN
```

TOPIC is the topic id and could be found in the released topics. QUESTION is the suggested by the participant query of this round. The question should be included within quotes, e.g. "london hotels". Each suggested query should be repeated over a maximum of 10 rows. DOCNO is the document id in the corpus. RANK is the rank of the document returned for this given round (in increasing order) SCORE is the score of the ranking/classification algorithm. RUN is an identifier/name for the system producing the run. Below is an example of a run:

```
dd17-51 "Katrina most costly hurricane" 1783276 10  ILPS-run1
dd17-51 "Katrina most costly hurricane" 1775816 9   ILPS-run1
dd17-51 "Katrina most costly hurricane" 1718269 8   ILPS-run1
dd17-51 "Katrina most costly hurricane" 1724162 7   ILPS-run1
dd17-51 "Katrina most costly hurricane" 1701311 6   ILPS-run1
dd17-51 "Katrina most costly hurricane" 1834929 5   ILPS-run1
dd17-51 "Katrina most costly hurricane" 1818307 4   ILPS-run1
dd17-51 "Katrina most costly hurricane" 1780634 3   ILPS-run1
dd17-51 "Katrina most costly hurricane" 1704548 2   ILPS-run1
dd17-51 "Katrina most costly hurricane" 1704526 1   ILPS-run1
dd17-51 "Hurricane Katrina's path of destruction"   1704322 10  ILPS-run1
... ... ... ...
dd17-60 "Tupac Amaru and Shining Path relationship" 0896459 1   ILPS-run1
```

In Task 2, participants were provided with all the top-10 results of the 10 queries submitted to the retrieval system, and they were asked to re-rank these 100 documents into a single ranking. The run to be submitted should have the following format:

```
TOPIC DUMMY DOCNO RANK SCORE RUN
```

TOPIC is the topic id and can be found in the released topics. DUMMY is a dummy column to be filled in with 0. DOCNO is the document number in the corpus. RANK is the rank of the document returned for this given round (in increasing order). SCORE is the score of the ranking/classification algorithm. RUN is an identifier/name for the system producing the run

Below is an example run:

```
dd17-51 0   1783276 100 ILPS-run1
dd17-51 0   1704322 99  ILPS-run1
dd17-51 0   1718269 98  ILPS-run1
dd17-51 0   1724162 97  ILPS-run1
dd17-51 0   1704548 96  ILPS-run1
dd17-51 0   1834929 95  ILPS-run1
dd17-51 0   1818307 94  ILPS-run1
dd17-51 0   1780634 93  ILPS-run1
dd17-51 0   1701311 92  ILPS-run1
dd17-51 0   1704526 91  ILPS-run1
dd17-51 0   1775816 90  ILPS-run1
... ... ... ...
dd17-60 0   0896459 1   ILPS-run1
```

3 Data Sets and Answer Agent

The collection that is used in the 2018 Dynamic Search Lab is the New York Times corpus[1]. The New York Times dataset consists of 1,855,658 articles published in New York Times from January 1, 1987 to June 19, 2007 with metadata provided by the New York Times Newsroom, the New York Times Indexing Service and the online production staff at nytimes.com. Most articles are manually summarized and tagged by professional staffs. The original form of this dataset is in News Industry Text Format (NITF).

The corpus was indexed by Indri and a Query Language Model with Dirichlet Smoothing has been implemented on the top of the Indri index, using Pyndri [4][2]. Potential participants are provided with a RESTful API to query the index[3].

The topics used are the topics developed by NIST assessors for the TREC 2017 Dynamic Domain Track [19]. A topic contains a title, which is like a query of few words, a more extended description of the user's information need, and a narrative that elaborates on what makes a document relevant and what not. It is the main search target for one complete run of dynamic search. An example of a topic provided to participants can be found below:

```
<topic name="Return of Klimt paintings to Maria Altmann"
       id="dd17-1">
    <description>Find information about the artwork by Austrian
    painter Gustav Klimt that was stolen by Nazis from its Austrian
    owners and subsequently returned to the rightful heir, Maria
    Altmann. </description>
    <narrative>This topic follows developments in the case of six
    specific paintings stolen by the Nazis during WWII.  The stolen
    paintings were given to a relative of the artist (Gustav Klimt),
    who in turn gave them to the Austrian government. ... Only the
    six paintings that comprised the Altmann case are relevant;
    work by other artists and other work by Klimt that had been
    confiscated by the Nazis is not relevant. </narrative>
</topic>
```

Each topic contains multiple subtopics, each of which addresses one aspect of the topic. The NIST assessors have tried produce a complete set of subtopics for each topic, and so they are treated as the complete set used in the evaluation. An example of a topic with subtopics is shown below:

[1] https://catalog.ldc.upenn.edu/ldc2008t19.
[2] https://github.com/cvangysel/pyndri.
[3] https://bitbucket.org/cvangysel/pyndri-flask.

```
<topic name="Hurricane Katrina's Effects" id="dd17-51"
       num_of_subtopics="4">
    <description>Hurricane Katrina, the most costly hurricane ever,
    effected millions of people ... </description>
    <narrative>Relevant documents report on how ... </narrative>
    <subtopic name="Katrina most costly hurricane" id="505"
              num_of_passages="10">
        <passage id="3628">
            <docno>1783276</docno>
            <rating>2</rating>
            <text><![CDATA[A federal judge in Mississippi ...]]>
            </text>
            <type>MANUAL</type>
        </passage>
        ...
```

Ten topics were released as a development set on April 15, 2018, and fifty topics as the test set on May 5, 2018. Subtopics for the test topics were not be released. The relevance judgments for these topics also followed the same sub-topic structure.

4 Evaluation

4.1 Task 1: Query Suggestion

The objective of task 1 is to generate a sequence of queries, in a sequential fashion, given a verbose description of a task (topic) and results of the answer agent for all previous queries. Each developed question agent is allowed to go over 10 rounds of query generations. At each round one query is submitted to the answer agent, and the top 10 results are collected. At the end of round 10, 100 search results will have been collected.

Therefore, in task 1 we focus on session based evaluation, with the quality of the A-Agent quantified by the Cube Test [10], sDCG [5], Expected Utility [20], and expected session nDCG [9]; other diagnostic measures such as precision and recall are to be reported. Cube Test is a search effectiveness measurement evaluating the speed of gaining relevant information (could be documents or passages) in a dynamic search process. It measures the amount of relevant information a system could gather and the time needed in the entire search process. The higher the Cube Test score, the better the IR system. sDCG extends the classic DCG to a search session which consists of multiple iterations. The relevance scores of results that are ranked lower or returned in later iterations get more discounts. The discounted cumulative relevance score is the final results of this metric. Expected Utility scores different runs by measuring the relevant information a system found and the length of documents. The relevance scores of documents are discounted based on ranking order and novelty. The document length is discounted only based on ranking position. The difference between the cumulative relevance score and the aggregated document length is the final score of each

run. Expected session DCG is an extension of the probabilistic model of the Expected Utility measure, that allows for modeling users that do not always see all the reformulations of a static set of them, i.e. allows for early abandonment.

4.2 Task 2: Results Composition

The objective of Task 2 is given the rankings obtained in Task 1 to merge them in a single composite ranking. At the end of round 10, 100 search results will have been collected. These 100 results coming from 10 queries should be re-ranked in a single optimal ranking. The evaluation of the quality of the composed ranking is done with traditional measures, such as nDCG, and diversity-based measures such as α-nDCG.

5 Lab Participation

Setting up the lab required more time than was originally anticipated. As a result, both the benchmark collection and the answer agent service were provided to participants only by mid-April, which did not allow the construction of a community around the lab. Hence, while 13 groups registered for the lab and 3 groups expressed very strong interest in participating, the lab received no submissions.

6 Conclusions

The CLEF Dynamic Search for Complex Tasks lab strives to answer one key question: how can we evaluate, and consequently build, dynamic search algorithms? The 2018 lab focused on how to devise an evaluation framework for dynamic search. Inspired by the Dynamic Domain framework, the lab sought for participants who would build user simulators, in terms of generated queries along multiple search iterations. The lab organizers decided to evaluate the effectiveness of the generated queries, rather than how close they are to actual user queries, and hence the main task of the lab (task 1) turned into a query suggestion task. This task was also followed by a result composition task (task 2) which focused on re-ranking the documents produced by a controlled retrieval system on the basis of the suggested queries.

Acknowledgements. This work was partially supported by the Google Faculty Research Award program. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

1. Allan, J.: HARD track overview in TREC 2003 high accuracy retrieval from documents. Technical report, DTIC Document (2005)
2. Carterette, B., Clough, P.D., Hall, M.M., Kanoulas, E., Sanderson, M.: Evaluating retrieval over sessions: the TREC session track 2011–2014. In: Perego, R., Sebastiani, F., Aslam, J.A., Ruthven, I., Zobel, J. (eds.) Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, 17–21 July 2016, pp. 685–688. ACM (2016). https://doi.org/10.1145/2911451.2914675
3. Georgila, K., Henderson, J., Lemon, O.: User simulation for spoken dialogue systems: learning and evaluation. In: Interspeech, pp. 1065–1068 (2006)
4. Van Gysel, C., Kanoulas, E., de Rijke, M.: Pyndri: a Python interface to the indri search engine. In: Jose, J.M., et al. (eds.) ECIR 2017. LNCS, vol. 10193, pp. 744–748. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56608-5_74
5. Järvelin, K., Price, S.L., Delcambre, L.M.L., Nielsen, M.L.: Discounted cumulated gain based evaluation of multiple-query IR sessions. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 4–15. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78646-7_4
6. Jung, S., Lee, C., Kim, K., Jeong, M., Lee, G.G.: Data-driven user simulation for automated evaluation of spoken dialog systems. Comput. Speech Lang. **23**(4), 479–509 (2009). https://doi.org/10.1016/j.csl.2009.03.002
7. Kanoulas, E., Azzopardi, L.: CLEF 2017 dynamic search evaluation lab overview. In: Jones, G.J.F., et al. (eds.) CLEF 2017. LNCS, vol. 10456, pp. 361–366. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65813-1_31
8. Kanoulas, E., Azzopardi, L.: CLEF 2017 dynamic search lab overview and evaluation. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, 11–14 September 2017. CEUR Workshop Proceedings, vol. 1866. CEUR-WS.org (2017). http://ceur-ws.org/Vol-1866/invited_paper_13.pdf
9. Kanoulas, E., Carterette, B., Clough, P.D., Sanderson, M.: Evaluating multi-query sessions. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, pp. 1053–1062. ACM, New York (2011). https://doi.org/10.1145/2009916.2010056
10. Luo, J., Wing, C., Yang, H., Hearst, M.A.: The water filling model and the cube test: multi-dimensional evaluation for professional search. In: 22nd ACM International Conference on Information and Knowledge Management, CIKM 2013, San Francisco, CA, USA, 27 October–1 November 2013, pp. 709–714 (2013). https://doi.org/10.1145/2505515.2523648
11. Maxwell, D., Azzopardi, L.: Agents, simulated users and humans: an analysis of performance and behaviour. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 731–740. CIKM 2016 (2016)
12. Over, P.: The TREC interactive track: an annotated bibliography. Inf. Process. Manag. **37**(3), 369–381 (2001)
13. Pääkkönen, T., Kekäläinen, J., Keskustalo, H., Azzopardi, L., Maxwell, D., Järvelin, K.: Validating simulated interaction for retrieval evaluation. Inf. Retr. J. **20**, 1–25 (2017)
14. Pietquin, O., Hastie, H.: A survey on metrics for the evaluation of user simulations. Knowl. Eng. Rev. **28**(01), 59–73 (2013)

15. Serban, I.V., Lowe, R., Henderson, P., Charlin, L., Pineau, J.: A survey of available corpora for building data-driven dialogue systems. CoRR abs/1512.05742 (2015). http://arxiv.org/abs/1512.05742
16. Verma, M., et al.: Overview of the TREC tasks track 2016. In: Voorhees and Ellis [17] (2016). http://trec.nist.gov/pubs/trec25/papers/Overview-T.pdf
17. Voorhees, E.M., Ellis, A. (eds.): Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, 15–18 November 2016, vol. Special Publication 500-321. National Institute of Standards and Technology (NIST) (2016). http://trec.nist.gov/pubs/trec25/trec2016.html
18. Yang, G.H., Soboroff, I.: TREC 2016 dynamic domain track overview. In: Voorhees and Ellis [17] (2016). http://trec.nist.gov/pubs/trec25/papers/Overview-DD.pdf
19. Yang, G.H., Soboroff, I.: TREC 2017 dynamic domain track overview. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, 15–18 November 2017. National Institute of Standards and Technology (NIST) (2017)
20. Yang, Y., Lad, A.: Modeling expected utility of multi-session information distillation. In: Azzopardi, L., et al. (eds.) ICTIR 2009. LNCS, vol. 5766, pp. 164–175. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04417-5_15
21. Yilmaz, E., Verma, M., Mehrotra, R., Kanoulas, E., Carterette, B., Craswell, N.: Overview of the TREC 2015 tasks track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, 17–20 November 2015, vol. Special Publication 500–319. National Institute of Standards and Technology (NIST) (2015). http://trec.nist.gov/pubs/trec24/papers/Overview-T.pdf