CLEF 2017 Task Overview: The IR Task at the eHealth Evaluation Lab

Evaluating Retrieval Methods for Consumer Health Search

Joao Palotti[1], Guido Zuccon[2], Jimmy[2], Pavel Pecina[3], Mihai Lupu[1], Lorraine Goeuriot[4], Liadh Kelly[5], and Allan Hanbury[1]

[1] Vienna University of Technology, Vienna, Austria,
[palotti,lupu,hanbury]@ifs.tuwien.ac.at
[2] Queensland University of Technology, Brisbane, Australia,
[g.zuccon, jimmy]@qut.edu.au
[3] Charles University, Prague, Czech Republic
pecina@ufal.mff.cuni.cz
[4] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
lorraine.goeuriot@imag.fr
[5] ADAPT Centre, Dublin City University, Ireland
liadh.kelly@dcu.ie

Abstract. This paper provides an overview of the information retrieval (IR) Task of the CLEF 2017 eHealth Evaluation Lab. This task investigates the effectiveness of web search engines in providing access to medical information for common people that have no or little medical knowledge (health consumers). The task aims to foster advances in the development of search technologies for consumer health search by providing resources and evaluation methods to test and validate search systems. The problem considered in this year's task was to retrieve web pages to support the information needs of health consumers that are faced with a medical condition and that want to seek relevant health information online through a search engine. The task re-used the 2016 topics, to deepen the assessment pool and create a more comprehensive and reusable collection. The task had four sub-tasks: ad-hoc search, personalized search, query variations, and multilingual ad-hoc search. Seven teams participated in the task; relevance assessment is underway and assessments along with the participants results will be released at the CLEF 2017 conference. Resources for this task, including topics, assessments, evaluation scripts and participant runs are available at the task's GitHub repository: https://github.com/CLEFeHealth/CLEFeHealth2017IRtask/

1 Introduction

This paper details the collection, systems and evaluation methods used in the IR Task of the CLEF 2017 eHealth Evaluation Lab (Task 3). This task is a continuation of the previous CLEF eHealth information retrieval (IR) tasks that ran between 2013 and 2016 [4,5,18,27] and embraces the TREC-style evaluation process, with a shared collection of documents and queries, the contribution of runs from participants and the subsequent formation of relevance assessments and evaluation of the participants submissions.

The task investigated the problem of retrieving web pages to support information needs of health consumers (including their next-of-kin) that are confronted with a health problem or medical condition and that use a search engine to seek better understanding about their health. This task has been developed within the CLEF 2017 eHealth Evaluation Lab, which aims to foster the development of approaches to support patients, their next-of-kin, and clinical staff in understanding, accessing and authoring health information [6].

The use of the Web as source of health-related information is a wide-spread practice among health consumers [13] and search engines are commonly used as a means to access health information available online [3].

Previous iterations of this task (i.e. the 2013 and 2014 CLEF eHealth Lab Task 3 [4,5]) aimed at evaluating the effectiveness of search engines to support people when searching for information about their conditions, e.g., to answer queries like "thrombocytopenia treatment corticosteroids length". These two evaluation exercises have provided valuable resources and an evaluation framework for developing and testing new and existing techniques. The fundamental contribution of these tasks to the improvement of search engine technology aimed at answering this type of health information need is demonstrated by the improvements in retrieval effectiveness provided by the best 2014 system [20] over the best 2013 system [23] (using different, but comparable, topic sets).

In 2015 the task took a different focus, specifically focusing on supporting consumers searching for self-diagnosis information [18], an important type of health information seeking activity [3]. Last year's task expanded on the 2015 task, by considering not only self-diagnosis information needs, but also needs related to treatment and management of health conditions [27]. Previous research has shown that exposing people with no or scarce medical knowledge to complex medical language may lead to erroneous self-diagnosis and self-treatment and that access to medical information on the Web can lead to the escalation of concerns about common symptoms (e.g., cyberchondria) [1,22]. Research has also shown that current commercial search engines are still far from being effective in answering such unclear and underspecified queries [26]. This year's task continues the growth path identified in past years and focuses on conducting assessments on deeper pooled sets than was possible in previous years of the task. The subtasks within this year's IR challenge are similar to 2016's: ad hoc search, query variation, and multilingual search. A new subtask is also introduced, aimed at exploring methods to personalize health search.

This paper is structured as follows: Section 2 details the four sub-tasks we considered this year; Section 3 describes the data collection, while Section 4 described the query set and the methodology used to create it; Section 5 lists the participants and their submissions; Section 6 details the methods used to create the assessment pools and relevance criteria; Section 7 lists the evaluation metrics used for this Task; finally, Section 8 concludes this overview paper.

2 Tasks

2.1 IRTask1: Ad-hoc Search

This is a standard ad-hoc search task, aiming at retrieving information relevant to people seeking health advice on the web. In this year's task, we re-used the 2016 topics, with the aim of improving the relevance assessment pool and the collection reusability (increase the pool depth). Because we re-used last year's topics, we asked participants to explicitly exclude from their search results for each query documents that have been already assessed in 2016 (a list of these documents was provided to participants, along with a script for checking submissions). Participants were highly encouraged to devise methods that explicitly explore relevance feedback, i.e. using the already assessed documents to improve their submissions.

2.2 IRTask2: Personalized Search

This task develops on top of the IRTask1. Here, we aimed to personalize the retrieved list of search results so as to match user expertise, measured by how likely the person is to be satisfied with the content of a document with respect to the expertise level of the health information.

Each topic in the collection has 6 query variations: the first 3 have been issued by people with no medical knowledge, while the second 3 have been issued by medical experts. When evaluating results for a query variation, we use a parameter alpha to capture user expertise. The parameter determines the shape of the gain curve, so that documents at the right understandability level obtain the highest gains, with decaying gains being assigned to documents that do not suit the understandability level of the modelled user. We use α=0.0 for query variation 1, α=0.2 for query variation 2, α=0.4 for query variation 3, α=0.6 for query variation 4, α=0.8 for query variation 5 and, finally, α=1.0 for query variation 6. This models increasing levels of expertise across query variations for one topic. The intuition in such evaluation is that a person with no specific health knowledge (represented by query variant 1) would not understand complex and technical health material, while an expert (represented by query variant 6) would have little or no interest in reading introductory/basic material. For more details about this evaluation measure, we refer the reader to Section 7.

Note that the 2016 collection includes assessments for understandability (for the same documents for which relevance was assessed), thus they could be used by teams for training. These understandability assessments are contained in the *qunder* files (similar to qrels, but for understandability) available at `https://github.com/CLEFeHealth/CLEFeHealth2016Task3`.

As for IRTask1, we asked participants to explicitly exclude from their search results for each query documents that have been already assessed in 2016.

2.3 IRTask3: Query Variations

IRTask1 and 2 treated query variations for a topic independently. IRTask3 instead explicitly explores the dependencies among query variations for the same information need. The task aims to foster research into building search systems that are robust to query variations. Different query variations were generated for the same forum entry (i.e. topic/information need), thus capturing the variability intrinsic in how people formulate queries when searching to answer the same information need.

For IRTask3 we asked participants to submit a single set of results for each topic (each topic has 6 query variations). Participants were informed of which query variations relate to the same topic, and should have taken these variations into account when building their systems.

2.4 IRTask4: Multilingual Search

The goal of this sub-task is to foster research in multilingual information retrieval, developing techniques to support users that can express their information need well in their native language and can read the results in English. This task, similar to the corresponding one in 2016, offers parallel queries in several languages (Czech, French, Hungarian, German, Polish, Spanish and Swedish).

3 Dataset

The 2013, 2014 and 2015 IR tasks in the CLEF eHealth Lab used the Khresmoi collection [8,7,4,5,18], a collection of about 1 million health web pages. Since last year we have set a new challenge to the participants by using the ClueWeb12-B13[6], a collection of more than 52 million web pages. As opposed to the Khresmoi collection, the crawl in ClueWeb12-B13 is not limited to certified Health On the Net websites and known health portals, but it is a higher-fidelity representation of a common Internet crawl, making the dataset more in line with the content current web search engines index and retrieve.

For participants who did not have access to the ClueWeb dataset, Carnegie Mellon University granted the organisers permission to make the dataset available through cloud computing instances[7] provided by Microsoft Azure. The Azure instances that were made available to participants for the IR challenge included (1) the Clueweb12-B13 dataset, (2) standard indexes built with the Terrier[8] [12], Indri[9] [21] and Elasticsearch[10] toolkits, (3) additional resources such

[6] http://lemurproject.org/clueweb12/

[7] The organisers are thankful to Carnegie Mellon University, and in particular to Jamie Callan and Christina Melucci, for their support in obtaining the permission to redistribute ClueWeb 12. The organisers are also thankful to Microsoft Azure who provided the Azure cloud computing infrastructure that was made available to participants through the Microsoft Azure for Research Award CRM:0518649.

[8] http://terrier.org/

[9] http://www.lemurproject.org/indri.php

[10] https://www.elastic.co/

as a spam list [2], Page Rank scores, anchor texts [9], urls, etc. made available through the ClueWeb12 website.

4 Query Set

Although a considerable number of documents (25,000) were assessed in the 2016 task, we decided to further deepen the assessment pools this year. Thus, the same topics developed in 2016 were kept for the 2017 task. For crafting the topics, we considered real health information needs expressed by the general public through posts published in public health web forums. Forum posts were extracted from the *AskDocs* section of Reddit[11]. This section allows users to post a description of a medical case or ask a medical question seeking medical information such as diagnosis, or details regarding treatments. Users can also interact through comments. We selected posts that were descriptive, clear and understandable. Posts with information regarding the author or patient (in case the post author sought help for another person), such as demographics (age, gender), medical history and current medical condition, were preferred.

The posts were manually selected by a student, and a total of 50 posts were used for query creation. Each of the selected forum posts were presented to 6 query creators with different medical expertise: these included 3 medical experts (final year medical students undertaking rotations in hospitals) and 3 lay users with no prior medical knowledge.

A total of 300 queries were created. Queries were numbered using the following convention: the first 3 digits of a query id identify a post number (information need), while the last 3 digits of a query id identify each individual query creator. Expert query creators used the identifier 1, 2 and 3 and laypeople query creators used the identifiers 4, 5 and 6. Queries and the posts used to generate them can be accessed at `https://github.com/CLEFeHealth/CLEFeHealth2017IRtask/tree/master/queries`.

For the query variations element of the task (sub-task 3), participants were told which queries were related to the same information need, to allow them to produce one set of results to be used as answer for all query variations of an information need.

For the multilingual element of the challenge (sub-task 4), Czech, French, Hungarian, German, Polish, Spanish and Swedish translations of the queries were provided. Queries were translated by medical experts hired through a professional translation company.

5 Participants Submission

Out of the 43 registered participating teams, 7 submitted runs. Table 1 lists the participating teams and the number of submitted runs for each one of the tasks. Teams were allowed to submit up to 7 runs and priority was sequentially given

[11] `https://www.reddit.com/r/AskDocs/`

Team Name	University	Country	Sub-Task			
			1	2	3	4
CUNI	Charles University in Prague	Czech Republic	3	-	-	28
IELAB	Queensland University of Technology	Australia	7	-	-	-
KISTI	Korean Institute of Science and Technology Information	Korea	3	-	-	-
SINAI	Universidad de Jaén	Spain	3	-	-	-
TUW	Vienna University of Technology	Austria	7	7	-	-
ub-botswana	University of Botswana	Botswana	5	-	-	-
UEvora	Universidade de Évora	Portugal	5	-	-	-
7 Teams	7 Institutions	7 Countries	33	7	0	28

Table 1: Participating teams and the number of submissions for each Sub-Task.

for assessment depending on the run number. Thus runs were sampled according to their priority: the priority of a run is expressed by the number that is assigned to the run by the participant, i.e., run 2 has a higher priority (and thus higher likelihood of inclusion in the assessment pool) than run 3. Run 1 (the baseline) has the highest priority; run 7 the lowest.

6 Assessments

Assessments are currently in progress. Similar to the 2016 pool, this year the pool was created using the RBP-based Method A (Summing contributions) by Moffat et al. [14], in which documents are weighted according to their overall contribution to the effectiveness evaluation as provided by the RBP formula (with p=0.8, following Park and Zhang [19]). This strategy was chosen because it was shown that it should be preferred over traditional fixed-depth or stratified pooling when deciding upon the pooling strategy to be used to evaluate systems under fixed assessment budget constraints [11], as it is the case for this task.

Following the suggestions of Palotti et al. [17], we adopted a two-stage approach to gather multi-dimensional relevance assessments. In the first stage of such a method, assessor time from highly-paid expert assessors is focused on assessing topical relevance and document trustworthiness (for relevant documents). In the second stage, understandability assessments are acquired employing less expert or less expensive assessors. The use of such a two-stage approach for collecting assessments has the potential of reducing the overall cost of evaluation, allowing assessment of more documents.

The relevance criteria created in 2016 were re-used this year. They were drafted considering the entirety of the forum posts used to create the queries, a link to the forum posts was also provided to the assessors. Relevance assessments were provided with respect to the grades *Highly relevant*, *Somewhat relevant* and *Not Relevant*. Readability/understandability and reliability/trustworthiness judgements were also collected for the documents in the assessment pool. These judgements were collected using a integer value between 0 and 100 (lower values meant harder to understand document / low reliability) provided by judges through a slider tool; these judgements were used to evaluate systems across different dimensions of relevance [25,24]. All assessments were collected through a purposely customised version of the Relevation toolkit [10].

7 Evaluation Metrics

System evaluation is conducted with both topical relevance centred metrics as well as understandability biased metrics in this task. Multiple evaluation metrics are used depending on the sub-task.

For IRTasks 1 and 4, evaluation is conducted with standard topical relevance centred metrics: Precision at 10 (P@10), Normalized Discounted Cumulative Gain at depth 10 (NDCG@10) and Rank Biased Precision with μ parameter set to 0.8 (RBP(0.8)), as done in previous years [27,18,5].

Submissions to IRTask3 are evaluated using the same measures as for IRTask1 but using the mean-variance evaluation framework (MVE) [28]. In this framework, evaluation results for each query variations for a topic are averaged and their variance also accounted for to compute a final system performance estimate. A script that implements the mean-variance evaluation framework is available at `https://github.com/CLEFeHealth/CLEFeHealth2017IRtask`.

Submissions to IRTask2 are evaluated by understandability biased metrics [24]. We consider the understandability-biased Rank Biased Precision, also with μ parameter set to 0.8 (uRBP(0.8)), and propose three new metrics for this subtask.

The first personalization-aware metric is a specialization of uRBP, α-uRBP, which uses an α parameter to model the kind of documents a user wants to read. We assume that α is a parameter that represents the understandability profile of an entity. A low α is assigned to items/documents/users that are experts, while a high α means the opposite. We assume that a user with a low α is interested in reading specialized documents as opposed to easy and introductory documents, while a high α represents users preferring the opposite, i.e. easy and introductory documents over specialized ones. We model in α-uRBP a penalty for the case in which a low α document is retrieved for a user that wants high α documents and vice versa. While we are still investigating which function is best to model this penalty, for now we assume penalty score drawn from a normal distribution. Figure 1 shows an example in which a user is seeking to read documents with α=20 and other values for α would have a penalty associated to them according to a Gaussian curve centered at 20 and with standard deviation of 30. We use the standard deviation of 30 in this evaluation campaign – methods to better estimate these parameters are left for future work.

The second and third personalization-aware metrics are simple modifications of Precision at depth X. For the relevant documents found up to rank X, we inspect how far the understandability label of each document is to the expected value required by a user. We could penalize the absolute difference linearly (LinUndP@X) or using the same Gaussian curve as in α-uRBP (GaussianUndP@10). Note that lower values are better for LinUndP@10, meaning that the distance from the required understandability value is small, and higher values are better for GaussianUndP@10, as a value of 100 is the best value one could reach.

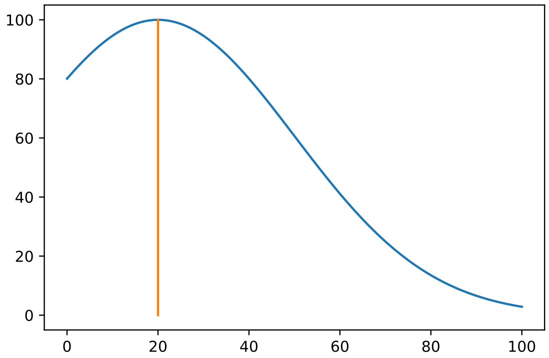

Fig. 1: A Gaussian model for penalty. This example of a normal distribution has its peak (mean) at 20 and standard deviation of 30. A document with α=60 would be worth only 41% of the score of a document with the desired α=20.

Scripts that implement the α-uRBP, LinUndP@X and GaussianUndP@10 are also available at `https://github.com/CLEFeHealth/CLEFeHealth2017IRtask`.

8 Conclusion

This document describes the settings and evaluation methods used in the IR Task of CLEF 2017 eHealth Evaluation Lab. The task considers the problem of retrieving web pages for people seeking health information regarding medical conditions, treatments and suggestions. The task was divided into 4 sub-tasks: ad-hoc search, personalized search, query variations, and multilingual ad-hoc search. Seven teams participated in the task; relevance assessment is underway and assessments along with the participants results will be released at the CLEF 2017 conference (and will be available at the task's GitHub repository).

Further development of the assessments made this year makes the collection stronger, providing to the research community a rich collection that goes beyond topical judgements. The understandability and trustworthiness assessments can be used to foster the development of retrieval methods for health information seeking on the web (e.g. [16,15]).

Baseline runs, participant runs and results, assessments, topics and query variations are available online at the GitHub repository for this Task: `https://github.com/CLEFeHealth/CLEFeHealth2017IRtask/`.

Acknowledgements We would like to thank Microsoft Azure grant (CRM:0518649), ESF for the financial support for relevance assessments, and the many assessor for their hard work. This work is supported in part by a Google Faculty Award.

References

1. M. Benigeri and P. Pluye. Shortcomings of health information on the internet. *Health promotion international*, 18(4):381–386, 2003.
2. G. V. Cormack, M. D. Smucker, and C. L. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval*, 14(5):441–465, 2011.
3. S. Fox. *Health topics: 80% of internet users look for health information online*. Pew Internet & American Life Project, 2011.
4. L. Goeuriot, G. J. Jones, L. Kelly, J. Leveling, A. Hanbury, H. Müller, S. Salantera, H. Suominen, and G. Zuccon. ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. *CLEF 2013 Online Working Notes*, 8138, 2013.
5. L. Goeuriot, L. Kelly, W. Lee, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, and H. M. Gareth J.F. Jones. ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. In *CLEF 2014 Evaluation Labs and Workshop: Online Working Notes*, Sheffield, UK, 2014.
6. L. Goeuriot, L. Kelly, H. Suominen, A. Névéol, A. Robert, E. Kanoulas, R. Spijker, J. Palotti, and G. Zuccon. Clef 2017 ehealth evaluation lab overview. In *Proceedings of CLEF 2017 - 8th Conference and Labs of the Evaluation Forum.* Lecture Notes in Computer Science (LNCS), Springer, September 2017.
7. L. Goeuriot, L. Kelly, G. Zuccon, and J. Palotti. Building evaluation datasets for consumer-oriented information retrieval. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).
8. A. Hanbury. Medical information retrieval: an instance of domain-specific search. In *Proceedings of SIGIR 2012*, pages 1191–1192, 2012.
9. D. Hiemstra and C. Hauff. Mirex: Mapreduce information retrieval experiments. *arXiv preprint arXiv:1004.4489*, 2010.
10. B. Koopman and G. Zuccon. Relevation! an open source system for information retrieval relevance assessment. *arXiv preprint*, 2013.
11. A. Lipani, G. Zuccon, L. Mihai, B. Koopman, and A. Hanbury. The impact of fixed-cost pooling strategies on test collection bias. In *Proceedings of the 2016 International Conference on The Theory of Information Retrieval*, ICTIR '16, New York, NY, USA, 2016. ACM.
12. C. Macdonald, R. McCreadie, R. L. Santos, and I. Ounis. From puppy to maturity: Experiences in developing terrier. *Proc. of OSIR at SIGIR*, pages 60–63, 2012.
13. D. McDaid and A. Park. Online health: Untangling the web. evidence from the bupa health pulse 2010 international healthcare survey. Technical report, 2011.
14. A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, Dec. 2008.
15. J. Palotti. Beyond topical relevance: Studying understandability and reliability in consumer health search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1167–1167. ACM, 2016.
16. J. Palotti, L. Goeuriot, G. Zuccon, and A. Hanbury. Ranking health web pages with relevance and understandability. In *Proceedings of the 39th international ACM SIGIR conference on Research and development in information retrieval*, 2016.
17. J. Palotti, G. Zuccon, J. Bernhardt, A. Hanbury, and L. Goeuriot. Assessors Agreement: A Case Study across Assessor Type, Payment Levels, Query Variations and

Relevance Dimensions. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association, CLEF'16 Proceedings*. Springer International Publishing, 2016.

18. J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanburyn, G. J. Jones, M. Lupu, and P. Pecina. CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving Information about Medical Symptoms. In *CLEF 2015 Online Working Notes*. CEUR-WS, 2015.
19. L. Park and Y. Zhang. On the distribution of user persistence for rank-biased precision. In *Proceedings of the 12th Australasian document computing symposium*, pages 17–24, 2007.
20. W. Shen, J.-Y. Nie, X. Liu, and X. Liui. An investigation of the effectiveness of concept-based approach in medical information retrieval GRIUM@ CLEF2014eHealthTask 3. In *Proceedings of the CLEF eHealth Evaluation Lab*, 2014.
21. T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, volume 2, pages 2–6. Amherst, MA, USA, 2005.
22. R. W. White and E. Horvitz. Cyberchondria: studies of the escalation of medical concerns in web search. *ACM TOIS*, 27(4):23, 2009.
23. D. Zhu, S. T.-I. Wu, J. J. Masanz, B. Carterette, and H. Liu. Using discharge summaries to improve information retrieval in clinical domain. In *Proceedings of the CLEF eHealth Evaluation Lab*, 2013.
24. G. Zuccon. Understandability biased evaluation for information retrieval. In *Proc. of ECIR*, 2016.
25. G. Zuccon and B. Koopman. Integrating understandability in the evaluation of consumer health search engines. In *Medical Information Retrieval Workshop at SIGIR 2014*, page 32, 2014.
26. G. Zuccon, B. Koopman, and J. Palotti. Diagnose this if you can: On the effectiveness of search engines in finding medical self-diagnosis information. In *Advances in Information Retrieval*, pages 562–567. Springer, 2015.
27. G. Zuccon, J. Palotti, L. Goeuriot, L. Kelly, M. Lupu, P. Pecina, H. Mueller, J. Budaher, and A. Deacon. The IR Task at the CLEF eHealth Evaluation Lab 2016: User-centred Health Information Retrieval. In *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*, September 2016.
28. G. Zuccon, J. Palotti, and A. Hanbury. Query variations and their effect on comparing information retrieval systems. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 691–700. ACM, 2016.