Overview of the CLEF 2018 Consumer Health Search Task

Jimmy[1,2], Guido Zuccon[1], Joao Palotti[3,4], Lorraine Goeuriot[5], and Liadh Kelly[6]

[1] Queensland University of Technology, Brisbane, Australia,
[jimmy,g.zuccon]@qut.edu.au
[2] University of Surabaya, Surabaya, Indonesia
jimmy@Staff.ubaya.ac.id
[3] Vienna University of Technology, Vienna, Austria,
palotti@ifs.tuwien.ac.at
[4] Qatar Computing Research Institute, Doha, Qatar,
jpalotti@hbku.edu.qa
[5] Université Grenoble Alpes, Grenoble, France
[firstname.lastname]@imag.fr
[6] Maynooth University, Maynooth, Ireland
liadh.kelly@mu.ie

Abstract. This paper details the collection, systems and evaluation methods used in the CLEF 2018 eHealth Evaluation Lab, Consumer Health Search (CHS) task (Task 3). This task investigates the effectiveness of search engines in providing access to medical information present on the Web for people that have no or little medical knowledge. 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. Built upon the the 2013-17 series of CLEF eHealth Information Retrieval tasks, the 2018 task considers both mono- and multilingual retrieval, embracing the *Text REtrieval Conference* (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.

For this year, the CHS task uses a new Web corpus and a new set of queries compared to the previous years. The new corpus consists of Web pages acquired from the CommonCrawl and the new set of queries consists of 50 queries issued by the general public to the Health on the Net (HON) search services. We then manually translated the 50 queries to French, German, and Czech; and obtained English query variations of the 50 original queries.

A total of 7 teams from 7 different countries participated in the 2018 CHS task: CUNI (Czech Republic), IMS Unipd (Italy), MIRACL (Tunisia), QUT (Australia), SINAI (Spain), UB-Botswana (Botswana), and UEvora (Portugal).

Keywords: Evaluation, Consumer Health Search, Medical Information Retrieval

1 Introduction

The use of the Web as a source of health-related information is a widespread practice among health consumers [19] and search engines are commonly used as a means to access health information available online [7]. However, there is on-going need for development of retrieval approaches and resources to support development in this domain as highlighted in for example [11].

This document reports on the CLEF 2018 eHealth Evaluation Lab information retrieval (IR) task (Task 3). 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 a medical condition and that use a search engine to seek better understanding about their health. This task has been developed within the CLEF 2018 eHealth Evaluation Lab, which aims to build efforts around the easing and support of patients, their next-of-kins, clinical staff, and health scientists in understanding, accessing, and authoring eHealth information in a multilingual setting [34].

The 2018 *Consumer Health Search* (CHS) continued the previous iterations of this task (i.e. the 2013, 2014, 2015, 2016, and 2017 CLEFeHealth Lab information retrieval task [9,12,27,39,29,10]) that aimed at evaluating the effectiveness of search engines to support people when searching for information about their medical conditions, e.g., to answer queries like "antiandrogen therapy for prostate cancer" with correct, trustworthy and understandable search results.

The 2013 and 2014 tasks focused on helping patients or their next-of-kin understanding information in their medical discharge summary. The 2015 task focused on supporting consumers searching for self-diagnosis information, an important type of health information seeking activity [7]. The 2016 task expanded the 2015 task by considering not only self-diagnosing information but also needs related to treatment and management of health conditions. Finally, the 2017 task used the corpus and topics of the 2016 task, with the focus of expanding the assessment pool and the number of relevance assessments.

The 2018 CHS task considered similar subtasks as in 2017: ad hoc search, query variation, methods to personalize health search, and multilingual search. A new subtask was also introduced: this required participants to classify queries with respect to their underlying query intent as detailed in [8]. For this task a new query set was introduced and a new document corpus, obtained from a subset of the CommonCrawl [7] data.

The remainder of this paper is structured as follows: Section 2 details the subtasks we considered this year; Section 3 describes the query set and the methodology used to create it; Section 4 describes the corpus used; Section 5 details the baselines created by the organisers as a benchmark for participants; Section 6 describes participants submissions; Section 7 details the methods used to create the assessment pools and relevance criteria; finally, Section 8 concludes this overview paper.

[7] http://commoncrawl.org/

2 The Subtasks of CLEF CHS

2.1 Subtask 1: Ad-hoc Search

This was a standard ad-hoc search task, aiming at retrieving information relevant to people seeking health advice on the Web. Queries for this task were generated by mining 50 queries issued by the general public to the HON search services, as detailed in Section 3. Every query was treated as independent and participants were asked to generate retrieval runs in answer to each query, as in a common ad-hoc search task.

Participants submitted a TREC result file containing a ranking of results in answer to each query. Each participant was allowed to submit up to 4 submissions which will be evaluated using normalized discounted cumulative gain [14] at 10 (NDCG@10), binary preference (BPref) and rank biased precision (RBP) [20], with a persistence parameter $p = 0.80$ (see [30]).

2.2 Subtask 2: Personalized Search

This task was developed on top of the Subtask 1 and followed on a similar task started in 2017. This subtask aimed to personalize the retrieved list of search results so as to match user expertise, measured by how likely the person is to understand the content of a document (with respect to the health information). To this end, submissions (i.e., in standard TREC run format and up to 4 submissions per participant) will be evaluated using the graded version of the uRBP evaluation measure [37], which uses both topical relevance and other dimensions of relevance, such as understandability and trustworthiness.

The parameters of this evaluation measure were further varied to evaluate personalization to different users. In addition to the 50 base queries as used in Subtask 1, each topic was released with 6 query variations issued by 6 research students at QUT; students had no medical knowledge. When evaluating results for a query variation, a parameter α will be used to capture user expertise. The parameter is used to determine 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. The alpha parameters used for each query variation are: $\alpha = 0.0$ for query variation 1 (the basis query), $\alpha = 0.2$ for query variation 2, $\alpha = 0.4$ for query variation 3, $\alpha = 0.5$ for query variation 4, $\alpha = 0.6$ for query variation 5, $\alpha = 0.8$ for query variation 6 and, finally, $\alpha = 1.0$ for query variation 7. The α variation models an increasing level 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 7) would have little or no interest in reading introductory/basic material.

2.3 Subtask 3: Query Variations

Subtask 3 aimed to foster research into building search systems that are robust to query variations. The task used the same 7 query variations set as in Subtask 2.

For this subtask, participants were asked to submit a single set of results for each topic in standard TREC run format. Each topic had 7 variations to be considered. Each participant was allowed to submit up to 4 submissions and submissions were evaluated using the same measures as for Subtask 1 but using the mean-variance evaluation framework (MVE) [40]. In this framework, evaluation results for each query variation for a topic were averaged, and their variance also accounted for to compute a final system performance estimate.

2.4 Subtask 4: Multilingual Ad-hoc Search

The multilingual task extended the ad-hoc subtask by providing a translation of the queries from English into Czech, French, and, German. The goal of this subtask was to support 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.

The queries for this subtask were manual translations of the Subtask 1 queries. Participant's submissions in standard TREC run format (up to 4 submissions per participant) were evaluated against Subtask 1 evaluation metrics.

2.5 Subtask 5: Query Intent Identification

This task, introduced this year, required participants to classify queries with respect to the underlying health intent. Health intents in this sub-task were clustered into 8 high-level intents: (1) Disease/illness/syndrome/pathological condition, (2) Drugs and medicinal substances, (3) Healthcare, (4) Test & procedures, (5) First aid, (6) Healthy lifestyle, (7) Human anatomy, (8) Organ systems. For each high-level intent, there was a maximum of 13 low-level intents. The health intent taxonomy is provided at `https://github.com/CLEFeHealth/CLEFeHealth2018IRtask/blob/master/clef2018_health_intent_taxonomy.csv`.

Given a query, participants needed to predict the correct intent underlying the query. Note that a query may have multiple intents. For each query, participants were asked to submit the top 3 intent predictions, in the form of the taxonomy ID corresponding to the intent. Each participant could submit up to 2 submissions in TREC format, where instead of a document ID, participants should list a low-level taxonomy ID, e.g., 1.1, 1.4, 2.1. Submissions were evaluated using mean reciprocal rank and NDCG@1,2,3. Matches at high-levels intents are differentiated from matches at low-levels intents.

3 Queries

This year CHS task used a new set of 50 queries, generated by the Khresmoi project[8] [2], which were issued by the general public to the HON[9] search service. These queries were manually selected by a domain expert from a sample of raw queries from the HON search engine collected over a period of 6 months to be representative of the type of queries posed to the search engine. Only non-capitalized queries were taken into account to remove possible influence by web crawlers using predetermined queries. Queries which seemed to be too "specialized" (for example, complex medical terms) and in languages other than English were excluded. Query intent was manually added to queries by the domain expert using the taxonomy provided at `https://github.com/CLEFeHealth/CLEFeHealth2018IRtask/blob/master/clef2018_health_intent_taxonomy.csv`. On generation of the query set it was found that all 50 queries contained less than or equal to two query terms, as detailed in [3]. As such a new set of 50 representative queries was generated using the same procedure, this time with the extra stipulation that queries must contain greater than two query terms. These 50 queries were considered as the base queries and used in our CLEF eHealth CHS Subtask 1. Queries were not preprocessed, for example any spelling mistakes that may be present has not been removed; systems submitted by challenge participants should have taken this into account.

For Subtasks 2 and 3, each base query was augmented with 6 query variations issued by 6 research students at QUT, with no medical knowledge. Each student was asked to formulate a query for each of the 50 queries narrative. No post-processing was applied to the formulated query variations: duplicates and spelling errors were kept. Subtask 4 used parallel queries in the following languages: French, German, and Czech. These queries are manual translations of Subtask 1's 50 base queries. Finally, Subtask 5 used the 50 base queries as in Subtask 1.

Queries were numbered using a 6 digits number with the following convention: the first 3 digits of a query ID identified a topic number (information need) ranged from 151 to 200. The last 3 digits of a query ID identified each individual query creator. The base queries used the identifier 1 and research students query variations used the identifier 2, 3, 4, 5, 6, and 7. Figure 1 shows the base query and the 6 query variations for topic 152.

For the multilingual queries, queries were placed within their respective language tags: cz (Czech), de (German), en (English), and fr (France). Figure 2 shows topic 152's queries in all 4 languages.

[8] `http://khresmoi.eu/`

[9] `https://hon.ch/en/`

```
<queries>
    ...
    <query>
        <id>152001</id>
        <en>emotional and mental disorders</en>
    </query>
    <query>
        <id>152002</id>
        <en>work colleague depression</en>
    </query>
    <query>
        <id>152003</id>
        <en>mental health problems, change in mood and withdrawn</en>
    </query>
    <query>
        <id>152004</id>
        <en>mental health cause withdrawn mood changes</en>
    </query>
    <query>
        <id>152005</id>
        <en>disease cause mental health behaviour change</en>
    </query>
    <query>
        <id>152006</id>
        <en>mood alterations causes</en>
    </query>
    <query>
        <id>152007</id>
        <en>uncommon mood change </en>
    </query>
     ...
</queries>
```

Fig. 1: Query variations for topic 152.

```
<queries>
    ...
    <query>
        <id> 152001 </id>
        <en> emotional and mental disorders </en>
        <fr> troubles mentaux et émotionnels </fr>
        <de> emotionale und psychische störungen </de>
        <cz> emocionální a duševní poruchy </cz>
    </query>
     ...
</queries>
```

Fig. 2: Multilingual queries for topic 152.

4 Dataset

The CHS task in 2016 and 2017 used the ClueWeb12-B13[10] collection, a corpus of more than 52 million Web pages. This year we introduced clefehealth2018 corpus. This was crated by compiling Web pages of selected domains acquired from the CommonCrawl [11]. An initial list of Websites was identified for acquisition. The list was built by submitting the CLEF 2018 base queries to the Microsoft Bing APIs (through the Azure Cognitive Services) repeatedly over a period of few weeks [12], and acquiring the URLs of the retrieved results. The domains of the URLs were then included in the list, except some domains that were excluded for decency reasons (e.g. pornhub.com). The list was further augmented by including a number of known reliable health Websites and other known unreliable health Websites, from lists previously compiled by health institutions and agencies.

The corpus was divided into folders, by domain name. Each folder contained a file for each Webpage from the domain available in the CommonCrawl dump. In total, 2,021 domains were requested from the CommonCrawl dump of 2018-09 [13]. Of the 2,021 domains in total, 1,903 were successfully acquired. The remaining domains were discarded due to errors, corrupted or incomplete data returned by the CommonCrawl API (a total of ten retries were attempted for each domain before giving up on a domain). Of the 1,903 crawled domains, 84 were not available in the CommonCrawl dump, and for these, a folder in the corpus exists and represents the domain that was requested; however, the folder is empty, meaning that it was not available in the dump. Note that .pdf documents were excluded from the data acquired from CommonCrawl. A complete list of domains and size of the crawl data for each domain is available at `https://github.com/CLEFeHealth/CLEFeHealth2018IRtask/blob/master/clef2018collection_listofdomains.txt`.

The full collection, clefehealth2018 [14], it contains 5,535,120 Web pages and its uncompressed size is about 480GB. In addition to the full collection, an alternative corpus named clefehealth2018_B [15] was created by manually removing a number of domains that were not strictly health-related (e.g., news Websites). This subset contains 1,653 domains and its size is about 294GB, uncompressed.

5 Baselines

We generated 21 runs, from which 19 were for Sub-Task 1, 3 for Sub-Task 2, and 2 for Sub-Task 5, based on common baseline models and simple approaches for fusing query variations. In this section we describe the baseline runs.

[10] `http://lemurproject.org/clueweb12/`
[11] `http://commoncrawl.org/`
[12] repeated submissions over time were performed because previous work has shown that Bing's API results vary sensibly over time, both in terms of results and effectiveness [15]
[13] http://commoncrawl.org/2018/03/february-2018-crawl-archive-now-available/
[14] clefehealth2018 is available at `https://goo.gl/uBJaNi`
[15] clefehealth2018_B is available at `https://goo.gl/uBJaNi`

5.1 Baselines for SubTask 1

A total of 14 standard baselines were generated using:

– Indri v5.9 with default parameters for models LMDirichlet, OKAPI, and TFIDF.
– Terrier v4.0 with default parameters for model BM25, DirichletLM and TFIDF.
– Elastic Search 5.1.1. with default parameters for model BM25F (i.e., b=0.75 and k=1.2) using two fields: title (weighted 1) and body (weighted 3). Tie-break was set to 1.0.

For these systems, we created the runs using and not using the default pseudo-relevance feedback (PRF) of each toolkit. When using PRF, we added to the original query the top 10 terms of the top 3 documents. All these baseline runs were created using the Terrier, Indri and ElasticSearch instances made available to participants in the Azure platform.

Additionally, we created a set of baseline runs that took into account the reliability and understandability of information.

Also based on the BM25 baseline run of Terrier, two understandability baselines were created using readability formulae. We created runs based on CLI (Coleman-Liau Index) and GFI (Gunning Fox Index) scores [4,13], which are a proxy for the number of years of the school required to read the text being evaluated. These two readability formulae were chosen because they showed to be robust across different methods for HTML preprocessing [28]. We followed one of the methods suggested in [28], in which the HTML documents are preprocessed using Justext[16][31], the main text was extracted, periods at the end of sentences were added whenever they were necessary (e.g., in presence of line breaks), and then readability scores were calculated. Given the initial score S for a document and its readability score R, the final score for each document is the combination of score obtained as $S \times 1.0/R$.

5.2 Baselines for SubTask 2

We explored three ways to combine query variations:

– Concatenation: we concatenated the text of each query variation into a single query.
– Reciprocal Ranking Fusion [5]: we fuse the ranks of each query variation using the reciprocal ranking fusion approach, i.e.,

$$RRFScore(d) = \sum_{r \in R} \frac{1}{k + r(d)},$$

where D is set the documents to be ranked, R is the set of document rankings retrieved for each query variation by the same retrieval model, $r(d)$ is the rank of document d, and k is a constant set to 60, as in [5].

[16] `https://pypi.python.org/pypi/jusText`

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

Team Name	University	Country	Sub-Task 1	2	3	4	5
CUNI	Charles University in Prague	Czech Republic	4	-	-	12	-
IELAB	Queensland University of Tech.	Australia	4	-	-	-	-
IMS	University of Padua	Italy	4	-	-	-	-
Miracl	University of Sfax	Tunisia	4	-	-	-	-
SINAI	Universidad de Jaén	Spain	1	-	-	-	-
UB-Botswana	University of Botswana	Botswana	-	-	4	-	-
UEvora	University of Évora	China	4	4	-	-	-
7 Teams	7 Institutions	7 Countries	21	4	4	12	-

– Rank Biased Precision Fusion: similarly to the reciprocal ranking fusion, we fuse the documents retrieved for each query variation with the Ranking Biased Precision (RBP) formula [20],

$$RBPScore(d) = \sum_{r \in R} (1 - p) \times (p)^{r(d)-1},$$

where p is the free parameter of the RBP model used to estimate user persistence. Here we set $p = 0.80$.

For each of the three methods described above, we created a run based on the BM25 baseline run of Terrier without pseudo-relevance feedback.

5.3 Baselines for SubTask 5

We generated two baselines for SubTask 5 with MetaMap [1] and QuickUMLS [33], both softwares were used to map the text in the queries to Unified Medical Language System (UMLS) concepts from which pre-defined semantic types can be extracted. A full list of UMLS semantic types can be found at `https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml`. Similar to previous work in the literature [16,25,24], a mapping was established between the UMLS semantic types and the search intents. This mapping is provided in the Github repository of this year's task.

6 Participant Submissions

This year, 7 participants from 7 countries submitted at least one run for any of the subtasks, as shown in Table 1. Each participant could submit up to 4 runs for Subtasks 1, 2, and 3, 4 runs for each language of Subtask 4, and 2 runs for Subtask 5.

We include below a self-described summary of the approach of each team (with minor editing by the task coordinators).

CUNI [32]: They submited runs for IRTask1 and IRTask4. For IRTask1, Run 1, they used the Terrier's index that is provided by the organisers without applying any data preprocessing. Terrier's implementation of Dirichlet smoothing language model is used as the retrieval model with its default parameters. Run 2, they used the same retrieval mode as in Run 1, while as an index, they used Terrier's index that uses Porter-stemming method and English stop-word list. Run 3, used Terrier's implementation of TF-IDF model, for the purpose of comparing between a vector-space model and an LM model (the one that is used in Run 1), they used the same index as in Run 1. Run 4, they used Terrier's implementation of Kullback-Leiber divergence (KLD) for query expansion, with number of top documents is set to 10 and number of terms for expansion is set to 3.
For IRTask 4, Run 1, they translated the queries in the source languages into English and get 1-best-list translations. Retrieval is conducted using Dirichlet model, and non-stemmed index. The same retrieval settings are used in the following runs. Run 2, they used hypotheses reranking approach, in which each query is translated into English and from the 15-best-list translations, 1-best-list (in terms of IR quality) translation is selected for the retrieval. Run 3, first they translated the queries into English and the 1-best-list that is produced by the SMT system is chosen as a base query, then this query is expanded by one term using term reranking approach. Run 4, this run is similar to Run1, the only difference is that Google Translate is used to translate the queries into English.

IELAB [17]: The IELAB team participated in Subtask 1. The team addressed the challenge by extending the Entity Query Feature Expansion model, a knowledge base (KB) query expansion method. To obtain the query expansion terms, first, we mapped entity mentions to KB entities by performing exact matching. After mapping, we used the Title of the mapped KB entities as the source for expansion terms. For our first three expanded query sets, we expanded the original queries sourcing expansion terms from each of Wikipedia, the UMLS, and the CHV. For our fourth expanded query set, we combined expansion terms from Wikipedia and CHV.

IMS [22]: They studied a query expansion approach that takes into account the relationships between the terms of the query and the Medical Subject Headings (MeSH) terms [2]; in addition, they evaluated different document scoring strategies given the multiple ranking list produced by the query expansions [1,2]. The methodology used for query expansion stick to the following sequence of steps: 1) they identified the MeSH terms that are present in the query by means of the MeshOnDemand1 API. 2) For each MeSH term found in the previous step, they used the MeSHRDF2 database to look for semantically related (MeSH) terms. 3) We choosed a subset of all the possible relations (predicates) between terms in the MeSHRDF database; then, we use this subset of predicates for query expansion in different ways (change width and depth in the graph of related terms). Given the list of MeSH terms, they combined these terms with the original query and create many variants with small differences. For each variant, they computed a ranked list of documents using the Elasticsearch engine. In order to merge the ranked lists, they compared many weighting strategies (average, sum of scores, normalized scores, etc.) to calculate a single score for each document. Then, they ordered the documents and produce the final ranked list.

Miracl [36]:The Miracl team submitted 4 runs in Subtask 1. The team submitted 2 baseline runs with different weighting models (TF-IDF and BM25) and the other 2 runs include query expansion using with two different ways, the MeSH ontology (via scopeNotes and via related terms) and the BM25 weighting model.

SINAI [6]: The SINAI team participated in Sub-task 1. They applied the query expansion technique using the most famous search engine at the moment: Google. Using the search engine, they added additional information not previously included in the query. They identified the medical concepts using cTAKES. This recognizer provides them with UMLS concepts in the expanded query. This way, they avoided introducing noise with words that are not related to the user query.

UB-Botswana [21]: They described the methods deployed in the different runs submitted for their participation to the CLEF eHealth2018 Task 3: Consumer Health Search Task, IRTask 3: Query Variations. In particular, they deployed data fusion techniques to merge search results generated by multiple query variants. As improvement, they attempted to alleviate the term mismatch between the queries and the relevant documents by deploying query expansion before merging the results. For their baseline system, they concatenated the multiple query variants for retrieval and then deploy query expansion.

UEvora [35]: The work of UEvora explored the use of learning to rank techniques as well as query expansion approaches. A number of field based features were used in training a learning to rank model. A medical concept model proposed in their previous work is re-employed with this year's new task. A word vectors model and UMLS are used as the query expansion sources.

7 Assessments

To assess submissions for Subtasks 1, 2, 3, and 4, first, we collected all unique topic-document pairs from all submitted runs. Then, using the RBP-based Method A (Summing contributions) by Moffat et al. [20], we weighted each document according to their overall contribution to the effectiveness evaluation as provided the RBP formula (with p=0.8, following Park and Zhang [30]). This strategy, named RBPA, 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 [18], as is the case for this task. For each topic, we selected the top 500 weighted documents to form an assessment pool of 25,000 topic-document pairs.

After the assessment pool was formed, we developed a number of relevance assessment tasks to be issued as HITs on Amazon Mechanical Turk. Workers, selected among those with a 90% acceptance rate and at least 1,000 tasks completed, were presented with the base query and the narrative of a topic with a link to the archived Webpage to be assessed. Currently, these assessments are in progress. Relevance assessments will be provided with respect to the grades *Highly relevant*, *Somewhat relevant* and *Not Relevant*. Readability/understandability and reliability/trustworthiness assessments will also be collected for the documents in the assessment pool. These assessments are collected using an integer value between 0 and 100 (lower values meant harder to understand document / low reliability) provided by assessors through a slider tool and will be used to evaluate systems across different dimensions of relevance [38,37].

As a pre-work task, workers have been asked to explicitly state what they understood on the topic and the narrative: workers that provided incorrect answers are not assigned HITs.

For Subtask 5, the 50 base queries were manually labelled with search intent(s) according to the search intents taxonomy (see Section 2.5. These manual labels were used for assessment.

8 Conclusions

This paper described methods and analysis of the CLEF 2018 eHealth Evaluation Lab Consumer Health Search Task. The task considered the problem of retrieving Web pages for people seeking health information regarding medical conditions, treatments and suggestions. The task was divided into 5 Sub-tasks including ad-hoc 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 2018 conference (and will be available at the task's GitHub repository).

As a by-product of this evaluation exercise, the task makes available to the research community a collection with associated assessments and evaluation framework (including readability and reliability evaluation) that can be used to evaluate the effectiveness of retrieval methods for health information seeking on the web (e.g. [23,26]).

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/CLEFeHealth2018IRtask`.

9 Acknowledgments

Jimmy is sponsored by the Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan / LPDP).

References

1. A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

2. N. Aswani, T. Beckers, E. Birngruber, C. Boyer, A. Burner, J. Bystron, K. Choukri, S. Cruchet, H. Cunningham, J. Dedek, L. Dolamic, R. Donner, S. Dungs, I. Eggel, A. Foncubierta, N. Fuhr, A. Funk, A. Garcia Seco de Herrera, A. Gaudinat, G. Georgiev, J. Gobeill, L. Goeuriot, P. Gomez, M. Greenwood, M. Gschwandtner, A. Hanbury, J. Hajic, J. Hlavácová, M. Holzer, G. Jones, B. Jordan, M. Jordan, K. Kaderk, F. Kainberger, L. Kelly, S. Kriewel, M. Kritz, G. Langs, N. Lawson, D. Markonis, I. Martinez, V. Momtchev, A. Masselot, H. Mazo, H. Müller, P. Pecina, K. Pentchev, D. Peychev, N. Pletneva, D. Pottecher, A. Roberts, P. Ruch, M. Samwald, P. Schneller, V. Stefanov, M. A. Tinte, Z. Urešová, A. Vargas, and D. Vishnyakova. Khresmoi - Multimodal Multilingual Medical Information Search. In *Proceedings of Medical Informatics Europe 2012 (MIE 2012), Village of the Future*, August 2012.
3. N. Aswani, L. Kelly, M. Greenwood, A. Roberts, M. Samwald, N. Pletneva, G. Jones, and L. Goeuriot. D1.3 - Report on results of the WP1 first evaluation phase. Technical report, Khresmoi Project, August 2012.
4. M. Coleman and T. L. Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284, 1975.
5. G. V. Cormack, C. L. A. Clarke, and S. Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 758–759, New York, NY, USA, 2009. ACM.
6. M. C. Díaz-Galiano, P. López-Úbeda, M.-T. Martín-Valdivia, and L. A. Ureña-López. Sinai at clef ehealth 2018 task 3. using ctakes to remove noise from expanding queries with google. In *CEUR Workshop Proceedings: Working Notes of CLEF 2018: Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings, 2018.
7. S. Fox. *Health topics: 80% of internet users look for health information online*. Pew Internet & American Life Project, 2011.
8. L. Goeuriot, A. Hanbury, B. Hegarty, J. Hodmon, L. Kelly, S. Kriewel, M. Lupu, D. Markonis, P. Pecina, and P. Schneller. D7.3 Meta-analysis of the second phase of empirical and user-centered evaluations. Technical report, Khresmoi Project, August 2014.
9. 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.
10. L. Goeuriot, G. J. Jones, L. Kelly, J. Leveling, M. Lupu, J. Palotti, and G. Zuccon. An Analysis of Evaluation Campaigns in ad-hoc Medical Information Retrieval: CLEF eHealth 2013 and 2014. *Springer Information Retrieval Journal*, 2018.
11. L. Goeuriot, G. J. Jones, L. Kelly, H. Mueller, and J. Zobel. Report on the SIGIR 2014 Workshop on Medical Information Retrieval (MedIR). *SIGIR Forum*, 48(2), 2014.
12. 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.
13. R. Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.
14. K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

15. Jimmy, G. Zuccon, and G. Demartini. On the volatility of commercial search engines and its impact on information retrieval research. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '18, 2018.
16. Jimmy, G. Zuccon, and B. Koopman. Choices in knowledge-base retrieval for consumer health search. In *European Conference on Information Retrieval*, pages 72–85. Springer, 2018.
17. Jimmy, G. Zuccon, and B. Koopman. Qut ielab at clef 2018 consumer health search task: Knowledge base retrieval for consumer health search. In *CEUR Workshop Proceedings: Working Notes of CLEF 2018: Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings, 2018.
18. A. Lipani, J. Palotti, M. Lupu, F. Piroi, G. Zuccon, and A. Hanbury. Fixed-cost pooling strategies based on ir evaluation measures. In *European Conference on Information Retrieval*, pages 357–368. Springer, 2017.
19. D. McDaid and A. Park. Online health: Untangling the web. evidence from the bupa health pulse 2010 international healthcare survey. Technical report, 2011.
20. 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.
21. N. Motlogelwa, E. Thuma, and T. Leburu-Dingalo. Merging search results generated by multiple query variants using data fusion. In *CEUR Workshop Proceedings: Working Notes of CLEF 2018: Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings, 2018.
22. G. M. D. Nunzio and A. Moldovan. A study on query expansion with mesh terms and elasticsearch. ims unipd at clef ehealth task 3. In *CEUR Workshop Proceedings: Working Notes of CLEF 2018: Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings, 2018.
23. 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.
24. J. Palotti and A. Hanbury. Tuw@ trec clinical decision support track 2015. Technical report, Vienna University of Technology Vienna Austria, 2015.
25. J. Palotti, A. Hanbury, and H. Müller. Exploiting health related features to infer user expertise in the medical domain. In *Web Search Click Data workshop at WSCM, New York City, NY, USA*, 2014.
26. 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.
27. 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.
28. J. Palotti, G. Zuccon, and A. Hanbury. The influence of pre-processing on the estimation of readability of web documents. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1763–1766, New York, NY, USA, 2015. ACM.
29. J. Palotti, G. Zuccon, Jimmy, P. Pecina, M. Lupu, L. Goeuriot, L. Kelly, and A. Hanbury. CLEF 2017 Task Overview: The IR Task at the eHealth Evaluation Lab. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*, CEUR Workshop Proceedings, 2017.

30. 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.
31. J. Pomikálek. *Removing boilerplate and duplicate content from web corpora*. PhD thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic, 2011.
32. S. Saleh and P. Pecina. Cuni team: Clef ehealth consumer health search task 2018. In *CEUR Workshop Proceedings: Working Notes of CLEF 2018: Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings, 2018.
33. L. Soldaini and N. Goharian. QuickUMLS: a fast, unsupervised approach for medical concept extraction. In *SIGIR MedIR Workshop*, Pisa, Italy, 2016.
34. H. Suominen, L. Kelly, L. Goeuriot, E. Kanoulas, L. Azzopardi, R. Spijker, D. Li, A. Névéol, L. Ramadier, A. Robert, J. Palotti, Jimmy, and G. Zuccon. Overview of the clef ehealth evaluation lab 2018. CLEF 2018 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS). Springer, September 2018.
35. H. Yang and T. Gonçalves. Improving personalized consumer health search:notebook for ehealth at clef 2018. In *CEUR Workshop Proceedings: Working Notes of CLEF 2018: Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings, 2018.
36. S. Zayani, N. Ksentini, M. Tmar, and F. Gargouri. Miracl at clef 2018 : Consumer health search task. In *CEUR Workshop Proceedings: Working Notes of CLEF 2018: Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings, 2018.
37. G. Zuccon. Understandability biased evaluation for information retrieval. In *Advances in Information Retrieval*, pages 280–292, 2016.
38. 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.
39. 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.
40. 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*, CIKM '16, pages 691–700, New York, NY, USA, 2016. ACM.