The IR Task at the CLEF eHealth Evaluation Lab 2016: User-centred Health Information Retrieval

Guido Zuccon[1], Joao Palotti[2], Lorraine Goeuriot[3], Liadh Kelly[4], Mihai Lupu[2], Pavel Pecina[5], Henning Müller[6], Julie Budaher[3], and Anthony Deacon[1]

[1] Queensland University of Technology, Brisbane, Australia,
[g.zuccon, aj.deacon]@qut.edu.au
[2] Vienna University of Technology, Vienna, Austria,
[palotti,lupu]@ifs.tuwien.ac.at
[3] Université Grenoble Alpes, France
[firstname.lastname]@imag.fr
[4] Trinity College Dublin, Irland
liadh.kelly@tcd.ie
[5] Charles University, Prague, Czech Republic
pecina@ufal.mff.cuni.cz
[6] University of Applied Sciences Western Switzerland, Switzerland
henning.mueller@hevs.ch

Abstract. This paper details the collection, systems and evaluation methods used in the IR Task of the CLEF 2016 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. 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 by a medical condition and that want to seek relevant health information online through a search engine. As part of the evaluation exercise, we gathered 300 queries users posed with respect to 50 search task scenarios. The scenarios were developed from real cases of people seeking health information through posting requests of help on a web forum. The presence of query variations for a single scenario helped us capturing the variable quality at which queries are posed. Queries were created in English and then translated into other languages. A total of 49 runs by 10 different teams were submitted for the English query topics; 2 teams submitted 29 runs for the multilingual topics.

Keywords: Evaluation, Health Search

1 Introduction

This document reports on the CLEF 2016 eHealth Evaluation Lab, 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 medical condition and that use a search engine to seek better understanding about their health. This task has been developed within the CLEF 2016 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 [15].

The use of the Web as source of health-related information is a wide-spread practice among health consumers [19] and search engines are commonly used as a means to access health information available online [7]. Previous iterations of this task (i.e. the 2013 and 2014 CLEFeHealth Lab Task 3 [8,9]) 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 [27] over the best 2013 system [30] (using different, but comparable, topic sets). The 2015 task has instead focused on supporting consumers searching for self-diagnosis information [23], an important type of health information seeking activity [7]. This year's task expands on the 2015 task, by considering not only self-diagnosis information needs, but also needs related to treatment and management of health conditions. 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) [3,29]. Research has also shown that current commercial search engines are still far from being effective in answering such unclear and underspecified queries [33].

The remainder of this paper is structured as follows: Section 2 details the sub-tasks we considered this year; Section 3 described the query set and the methodology used to create it; 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; Section 8 lists the evaluation metrics used for this Task; finally, Section 9 concludes this overview paper.

2 Tasks

2.1 Sub-Task 1: Ad-hoc Search

Queries for this task are generated by mining health web forums to identify example information needs, as detailed in section 3. Every query is treated as independent and participants are asked to generate retrieval runs in answer to such queries, as in a common ad-hoc search task. This task extends the evaluation framework used in 2015 (which considered, along with topical relevance, also

the readability of the retrieved documents) to consider further dimensions of relevance such as the reliability of the retrieved information.

2.2 Sub-Task 2: Query Variations

This task explores query variations for each single information need. Previous research has shown that different users tend to issue different queries for the same information need and that the use of query variations for evaluation of IR systems leads to as much variability as system variations [1,2,23]. This was the case also in this year's task. Note that we explored query variations also in the 2015 IR task [23], and we found that for the same image showing a health condition, different query creators issued very different queries: they differ not only in terms of the keywords contained in the query, but also with respect to their retrieval effectiveness.

Different query variations are generated for the same information need (extracted from a web forum entry, as explained in section 3), thus capturing the variability intrinsic in how people search when they have the same information need. Participants were asked to exploit query variations when building their systems: participants were told which queries related to the same information need and they were required to produce one set of results to be used as answer for all query variations of an information need. This task aims to foster research into building systems that are robust to query variations, for example, through considering the fusion of ranked lists produced in answer to each single query variation.

2.3 Sub-Task 3: Multilingual Ad-hoc Search

The multilingual task extends the Ad-hoc Search task by providing a translation of the queries from English into Czech, French, Hungarian, German, Polish, Spanish and Swedish. The goal of this sub-task is 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.

3 Query Set

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[7]. 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

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

post author sought help for another person), such as demographics (age, gender), medical history and current medical condition, were preferred.

In order to collect query variants that could be compared, we also selected posts where a main and single information need could be identified. These constraints guarantee as much as possible getting queries on the same aspects of the post.

The comments were also taken into account in the selection. Any user can add a comment to a post, and all users are labeled according to their medical expertise[8]. We mainly selected posts with comments, including some from labeled users. The post 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.

All queries were preprocessed to correct for spelling mistakes; this was done using the Linux program *aspell*. This was however manually supervised so that spelling correction was performed only when appropriate, as for example not to change drug names. We explicitly did not remove punctuation marks from the queries, e.g., participants could take advantage of the quotation marks used by the query creators to indicate proximity terms or other features.

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 creator used the identifiers 4, 5 and 6. In Figure 2 we show variants 1, 2 (both generated by laypeople) and 4 (generated by an expert) created for post number 103 (posts started from number 101), shown in Figure 1.

For the query variations element of the task (sub-task 2), 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 3), 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.

4 Dataset

Previous IR tasks in the CLEF eHealth Lab have used the Khresmoi collection [12,10], a collection of about 1 million health web pages. This year we set a new challenge to the participants by using the ClueWeb12-B13[9], a collection

[8] To be labeled as a medical expert, users have to send Reddit a proof such as a student ID, or a diploma.

[9] `http://lemurproject.org/clueweb12/`

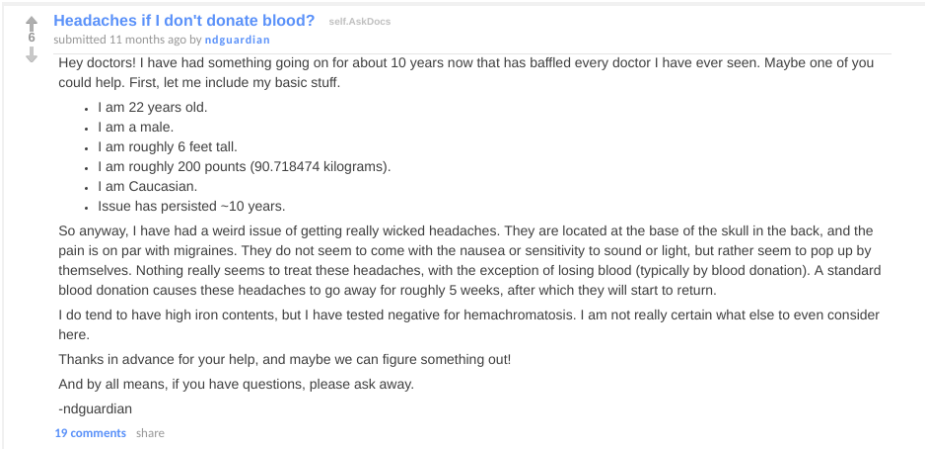

Fig. 1: Post from Reddit's Section AskDocs. It was used to generate queries ranging from 103001 to 103006

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[10] 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[11] [18] and the Indri[12] [28] toolkits, (3) additional resources such as a spam list [6], Page Rank scores, anchor texts [13], urls, etc. made available through the ClueWeb12 website.

5 Baselines

We generated 55 runs, from which 19 were for Sub-Task 1 and 36 for Sub-Task 2, based on common baseline models and simple approaches for fusing query variations. In this section we describe the baseline runs.

[10] 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.

[11] `http://terrier.org/`

[12] `http://www.lemurproject.org/indri.php`

```
<queries>
    ...
    <query>
        <id> 103001 </id>
        <title>headaches relieved by blood donation</title>
    </query>
    <query>
        <id> 103002 </id>
        <title>high iron headache</title>
    </query>
    ...
    <query>
        <id> 103004 </id>
        <title>headaches caused by too much blood or
                                "high blood pressure"</title>
    </query>
    ...
</queries>
```

Fig. 2: Extract from the official query set released.

5.1 Baselines for Sub-Task 1

A total of 12 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.

For both 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 and Indri instances made available to participants in the Azure platform.

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

Five reliability baselines were created based on the Spam rankings distributed with ClueWeb12[13][6]. For a given run, we removed all documents that had a spam score smaller than a given threshold th. We used the BM25 baseline run of Terrier, and 5 different values for th (50, 60, 70, 80 and 90).

Two understandability baselines were created using readability formulae. We created runs based on CLI (Coleman-Liau Index) and GFI (Gunning Fox Index) scores [4,11], 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

[13] http://www.mansci.uwaterloo.ca/~msmucker/cw12spam/

because they showed to be robust across different methods for HTML preprocessing [24]. We followed one of the methods suggested in [24], in which the HTML documents are preprocessed using Justext[14][26], the main text is extracted, periods at the end of sentences are added whenever they are necessary (e.g., in presence of line breaks), and then readability scores are 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 Sub-Task 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 variations 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].
- 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 each of three baselines for Terrier and Indri, with and without pseudo-relevance feedback. A total of 36 baseline runs were created for sub-task 2 (combination of 3 fusion approaches, 2 toolkits, 3 models, with and without PRF).

6 Participant Submissions

The number of registered participants for CLEF eHealth IR Task was 58; of these, 10 submitted at least one run for any of the sub-tasks, as shown in Table 1. Each team could submit up to 3 runs for Sub-Tasks 1 and 2, and up to 3 runs for each language of Sub-Task 3.

We include below a summary of the approach of each team, self described by them when their runs were submitted.

[14] `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
CUNI	Charles University in Prague	Czech Republic	2	2	21
ECNU	East China Normal University	China	3	3	8
GUIR	Georgetown University	United States	3	3	-
InfoLab	Universidade do Porto	Portugal	3	3	-
KDEIR	Toyohashi University of Technology	Japan	3	2	
KISTI	Korean Institute of Science and Technology Information	Korea	3	-	-
MayoNLPTeam	Mayo Clinic	United States	3	-	-
MRIM	Laboratoire d'Infomatique de Grenoble	France	3	-	-
ub-botswana	University of Botswana	Botswana	3	-	-
WHUIRgroup	Wuhan University	China	3	3	-
10 Teams	10 Institutions	8 Countries	29	16	29

CUNI: The CUNI team participated in all the subtasks but their main focus was put on the multilingual search in sub-task 3. The monolingual runs in sub-tasks 1 and 2 are mainly intended for comparison with the multilingual runs in sub-task 3. In sub-task 1 (Ad-Hoc Search), Run 1 employs the Terrier implementation of Dirichlet-smoothed language model with the μ parameter tuned on the data from previous CLEF eHealth tasks; Run 2 uses the Terrier vector space TF/IDF model. In sub-task 2 (Query variants), all query variants of one information need are searched for by the retrieval system (Dirichlet-smoothed language model in Run 1 and vector space TF/IDF model in Run2) and the resulting lists of ranked documents are merged and reranked by document scores to produce one ranked list of documents for each information need. In sub-task 3 (Multilingual search), all the non-English queries (including variants) are translated into English using their own statistical machine translation systems adapted to translate search queries in the medical domain. For each non-English query, 15 translation variants (hypotheses) are obtained. Their multilingual Run 1 employs the single best translation for each query as provided by the translation systems. In their multilingual Run 2, the top 15 translation hypotheses are reranked using a discriminative regression model employing a) features provided by the translation system and b) various kinds of features extracted from the document collection, external resources (UMLS - Unified Medical Language System, Wikipedia), or the translations themselves. Run 3 employs the same reranking method applied to the translation system features only.

ECNU: The ECNU team proposes a Web-based query expansion model and a combination method to better understand and satisfy the task. They use as baseline the Terrier implementation of the BM25 model. The other runs for sub-task 1, 2 and 3 explore Google search and MeSH to do query expansion. BM25, DFR_BM25, BB2, the PL2 models of Terrier and TF_IDF, the BM25 models of Indri were used and combined. For the sub-task 3 runs, Google Translator was used to translate the queries from Czech, French, Polish and Swedish to English before applying the same methods of sub-task 1.

GUIR: GUIR studies the use of medical terms for query reformulation. Synonyms and hypernyms from UMLS are used to generate reformulations of the queries; Terrier with Divergence from Randomness is used for retrieving and scoring documents. For sub-task 1, results obtained from the reformulated queries are used with the Borda rank aggregation algorithm. For sub-task 2, for each topic, results obtained are merged by any reformulated query in the topic using the Borda rank aggregation algorithm.

Infolab: Team InfoLab analyses the performance of several query expansion strategies using different methods to select the terms to be added to the original query. One of the methods uses the similarity between Wikipedia articles, found through an analysis of incoming and outgoing links, for term selection. The other method applies the Latent Dirichlet Allocation to Wikipedia articles to extract topics each containing a set of words that are used for term selection. In the end, readability metrics were used to re-rank the documents retrieved using the expanded queries.

KDEIR: KDEIR submitted runs for sub-tasks 1 and 2. In both sub-tasks the Waterloo spam score was used to filter out the spammiest documents, and the link structure present in the remaining documents was explored on top of their language model baseline.

KISTI: KISTI attempts two approaches using word vectors learned by Word2Vec based on medical Wikipedia pages. At first, initial documents are obtained using a search engine. Based on the documents, pseudo-relevance feedback (PRF) is applied with two different usage of the word vectors. In the first approach, PRF is performed with new relevance scores using the word vectors, while it is performed with a new query expanded using the word vectors in the second approach.

MayoNLPTeam: Mayo explores a Part-of-Speech (POS) based query term weighting approach which assigns different weights to the query terms according to their POS categories. The weights are learned by defining an objective function based on the mean average precision. They apply the proposed approach with the optimal weights obtained from the TREC 2011 and 2012 Medical Records Tracks into the Query Likelihood model (Run 2) and Markov Random Field (MRF) models (Run 3). The conventional Query Likelihood model was implemented as the baseline (Run 1).

MRIM: MRIM's objective is to investigate the effectiveness of the word embedding for query expansion on consumer health search, as well as the effect of the learning resource for learning on the results. Their system uses the Terrier index provided by the organizers. As a retrieval model the Dirichlet language model is used with default settings. Query expansion is applied on two training sets using word embedding sources. Word2vec is used for word embedding.

ub-botswana: In this participation, the effectiveness of three retrieval strategies is evaluated. In particular, PL2 is deployed with a Boolean Fallback score modifier as baseline system. If any of the retrieved documents contains all undecorated query terms (i.e. query terms without any operators), then documents are removed from the result set that do not contain all undecorated query terms With this score modifier. Otherwise, nothing is done. In another

approach, the collection enrichment approach is employed, where the original query is expanded with additional terms from an external collection (collection not being searched). To deliver an effective ranking, the first two rankers are combined using data fusion techniques.

WHUIRgroup: WHUIR uses Indri to conduct the experiments. CHV is used to expand queries and propose a learning-to-rank algorithm to re-rank the result.

7 Assessments

A Pool of 25,000 documents was created using the RBP-based Method A (Summing contributions) by Moffat et al. [20], 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 [25]). This stategy 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 [17]. A total of 100 runs were used (all baselines + all participant runs for Sub-Tasks 1 and 2) to form the assessment pools.

Assessment was performed by paid final year medical students who had access to queries, documents, and relevance criteria drafted by a junior medical doctor. The relevance criteria 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 judgments 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 [32,31]. All assessments were collected through a purpously costumised version of the Relevation toolkit [16].

8 Evaluation Metrics

System evaluation was conducted using precision at 10 (p@10) and normalised discounted cumulative gain [14] at 10 (nDCG@10) as the primary and secondary measures, respectively. Precision was computed using the binary relevance assessments by collapsing *Highly relevant* and *Somewhat relevant* assessments into the *Relevant* class, while nDCG was computed using the graded relevance assessments.

A separate evaluation was conducted using the multidimensional relevance assessments (topical relevance, understandability and trustworthiness) following the methods in [31]. For all runs, Rank biased precision (RBP) [20], with a

persistence parameter $p = 0.80$ (see [25]), was computed along with the multidimensional modifications of RBP, namely uRBP (using binary understandability assessments), uRBPgr (using graded understandability assessments), u+tRBP (using binary understandability and trustworthiness assessments).

Precision and nDCG were computed using **trec_eval**[15] along with RBP, while the multidimensional evaluation was performed using **ubire**[16] [31].

9 Conclusions

This paper describes methods, results and analysis of the CLEF 2016 eHealth Evaluation Lab, IR Task. 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 3 sub-tasks including ad-hoc search, query variations, and multilingual ad-hoc search. Ten teams participated in the task; relevance assessment is underway and assessments along with the participants results will be released at the CLEF 2016 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. [21,22]).

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/CLEFeHealth2016Task3`.

10 Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644753 (KConnect), and from the Austrian Science Fund (FWF) projects P25905-N23 (ADmIRE) and I1094-N23 (MUCKE). We also would like to thank Microsoft Azure grant (CRM:0518649), ESF for the support for financial relevance assessments and query creation, and the many assessor for their hard work.

References

1. L. Azzopardi. Query Side Evaluation: An Empirical Analysis of Effectiveness and Effort. In *Proc. of SIGIR*, 2009.
2. P. Bailey, A. Moffat, F. Scholer, and P. Thomas. User Variability and IR System Evaluation. In *Proc. of SIGIR*, 2015.

[15] `http://trec.nist.gov/trec_eval/trec_eval_latest.tar.gz`
[16] `https://github.com/ielab/ubire`

3. M. Benigeri and P. Pluye. Shortcomings of health information on the internet. *Health promotion international*, 18(4):381–386, 2003.
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. 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.
7. S. Fox. *Health topics: 80% of internet users look for health information online*. Pew Internet & American Life Project, 2011.
8. 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.
9. 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.
10. 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).
11. R. Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.
12. A. Hanbury. Medical information retrieval: an instance of domain-specific search. In *Proceedings of SIGIR 2012*, pages 1191–1192, 2012.
13. D. Hiemstra and C. Hauff. Mirex: Mapreduce information retrieval experiments. *arXiv preprint arXiv:1004.4489*, 2010.
14. K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.
15. L. Kelly, L. Goeuriot, H. Suominen, A. Neveol, J. Palotti, and G. Zuccon. Overview of the CLEF eHealth Evaluation Lab 2016. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*. Springer Berlin Heidelberg, 2016.
16. B. Koopman and G. Zuccon. Relevation! an open source system for information retrieval relevance assessment. *arXiv preprint*, 2013.
17. 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.
18. 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.
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. 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.

22. 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.
23. 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.
24. 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.
25. 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.
26. J. Pomikálek. *Removing boilerplate and duplicate content from web corpora*. PhD thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic, 2011.
27. 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.
28. 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. Citeseer, 2005.
29. R. W. White and E. Horvitz. Cyberchondria: studies of the escalation of medical concerns in web search. *ACM TOIS*, 27(4):23, 2009.
30. 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.
31. G. Zuccon. Understandability biased evaluation for information retrieval. In *Advances in Information Retrieval*, pages 280–292, 2016.
32. 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.
33. 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.