Building Evaluation Datasets for Consumer-Oriented Information Retrieval

Lorraine Goeuriot, Liadh Kelly, Guido Zuccon, Joao Palotti
Université Grenoble Alpes, Trinity College Dublin, Queensland University of Technology, Vienna University of Technology
France, Ireland, Australia, Austria
lorraine.goeuriot@imag.fr, liadh.kelly@tcd.ie, g.zuccon@qut.edu.au, palotti@ifs.tuwien.ac.at

Abstract

Common people often experience difficulties in accessing relevant, correct, accurate and understandable health information online. Developing search techniques that aid these information needs is challenging. In this paper we present the datasets created by CLEF eHealth Lab from 2013-2015 for evaluation of search solutions to support common people finding health information online. Specifically, the CLEF eHealth information retrieval (IR) task of this Lab has provided the research community with benchmarks for evaluating consumer-centered health information retrieval, thus fostering research and development aimed to address this challenging problem. Given consumer queries, the goal of the task is to retrieve relevant documents from the provided collection of web pages. The shared datasets provide a large health web crawl, queries representing people's real world information needs, and relevance assessment judgements for the queries.

Keywords: consumer health, evaluation campaign, information retrieval

1. Introduction

Information retrieval (IR) has evolved as a highly empirical discipline, where evaluation relies on carefully defined protocols and representative test collections. Shared tasks provide test collections to the community, in order to evaluate, compare and improve information retrieval systems. Shared evaluation test set generation is imperative for progression of the field of information retrieval. In generating test sets of this nature it is important to have robust and representative data.

CLEF eHealth is an evaluation campaign that informs the development of approaches to support patients, their next-of-kins, and clinical staff in understanding, accessing and generating health information. The campaign ran yearly since 2013, and each year organizes several tasks related to medical information extraction, management and retrieval.

The first CLEFeHealth lab (Suominen et al., 2013) contained three tasks: the first one on named entity recognition and/or normalization of disorders (Pradhan et al., 2013); the second one on acronyms/ abbreviations (Mowery et al., 2013) in clinical reports; and the third one on health-focused web information retrieval, supporting laypeople's information needs stemming from clinical reports (Goeuriot et al., 2013).

The second CLEFeHealth (Kelly et al., 2014) expanded our year-one efforts and again organized three tasks. Specifically, the first task aimed to help patients (or their next-of-kin) by addressing visualisation and readability issues related to their hospital discharge documents and related information search on the Internet (Suominen et al., 2014). The second task continued the IE work of the 2013 CLEFeHealth lab, specifically focusing on IE of disorder attributes from clinical text (Mowery et al., 2014). The third task further extended the 2013 IR task, with a cleaned version of the 2013 document collection being produced and the introduction of a new query generation method, as well as multilingual queries (Goeuriot et al., 2014).

The 2015 lab was split into two tasks focusing on information extraction and information retrieval. The IE task introduced two new challenges: a clinical speech recognition (SR) task of nursing shift changes (Suominen et al., 2015); and a named entity recognition in clinical reports in languages other than English, specifically French clinical reports (Névéol et al., 2015). The IR task focused on a new type of queries people issue to obtain information on the web (Palotti et al., 2015), with English queries, as well as their translations.

Figure 1 gives a summary of the CLEF eHealth tasks over the years.

The information retrieval task of the campaign has provided the research community with benchmarks for evaluating consumer-centered health information retrieval.Given queries issued by common people, the goal of the task is to retrieve relevant documents from the provided collection of web pages. In this paper we describe the datasets that were built for this evaluation task and the methodology followed to build them.

2. Related Work

Ours is not the first IR collection aimed at studying systems that support health information access and retrieval.

The OHSUMED collection contained around 350,000 abstracts from medical journals in the MEDLINE database. These covered a period of over five years and two sets of topics: a manually created one and one based on the controlled vocabulary thesaurus of the Medical Subject Headings[1] (MeSH). The collection was created for the TREC 2000 Filtering Track.

The Genomics Track (Roberts et al., 2009) ran an annual IR task on genomics data in biomedical papers and clinical reports from 2003–2007. The tasks ranged from ad-hoc IR to classification, passage IR, and entity-based question-answering.

The TREC Medical Records Track (Voorhees and Tong, 2011) ran an IR task in 2011 and 2012. The aim of the task was to develop IR techniques for finding patient cohorts relevant to inclusion criteria for clinical trial recruitment. Data

[1] http://www.ncbi.nlm.nih.gov/mesh

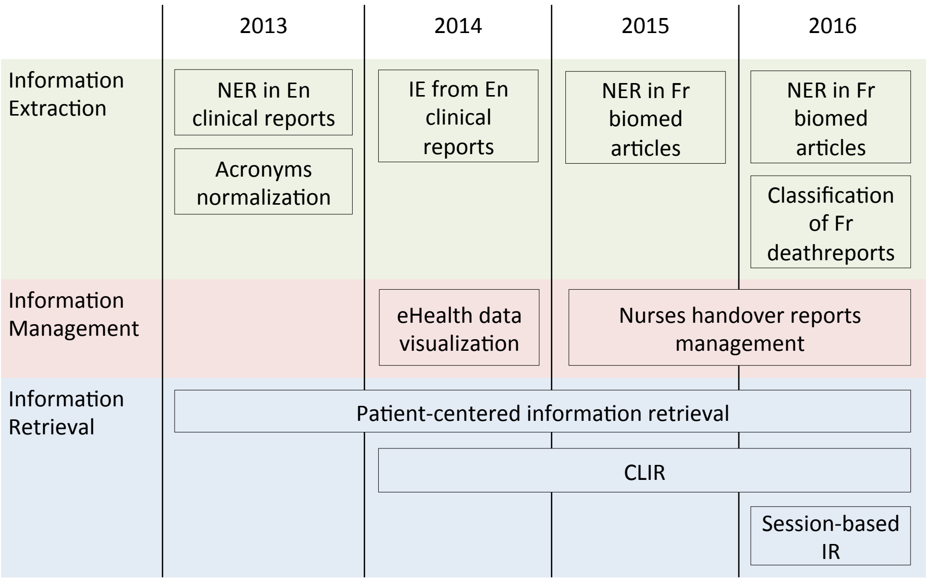
2013
2014
2015
2016
Information Extraction
NER in En clinical reports
Acronyms normalization
IE from En clinical reports
NER in Fr biomed articles
NER in Fr biomed articles
Classification of Fr deathreports
Information Management
eHealth data visualization
Nurses handover reports management
Information Retrieval
Patient-centered information retrieval
CLIR
Session-based IR

Figure 1: CLEF eHealth tasks since 2013

consisted of de-identified medical records, queries that resemble eligibility criteria, and associated relevance assessments.

The TREC clinical decision support track[2] investigates techniques for linking medical cases to information relevant for patient care. The document collection is the Open Access Subset of PubMed Central (PMC)[3], that contains around 650,000 biomedical articles. The topics are verbose medical case narratives. These are structured representations of medical records, containing information such as the patient's medical history, current symptoms, tests performed, etc. Each topic was labeled manually according to three common generic clinical questions (diagnosis, test, treatment). The goal of the task was to retrieve articles that would help a clinician answering the question.

In 2013, NTCIR (NII Test Collection for IR Systems) launched a new task called MedNLP, which aims to extract specific information from Japanese medical reports. For patient confidentiality reasons these are structured reports written by physicians about imaginary patients[4]. MedNLP includes two identification tasks: personal health information (e.g., name or gender, and complaints or diagnoses), and a "free task", where participants are invited to submit practical or creative solutions to other tasks.

In 2014 and 2015, the Question Answering evaluation lab at CLEF organized a new task called BioASQ (Balikas et al., 2014). The goal of this task was to address issues raised by large-scale datasets with an application to the biomedical context. BioASQ comprised two subtasks: a large-scale semantic indexing task; and a question-answering task. The former aimed at classifying documents from PubMed digital library[5] into MeSH. The latter focused on question-answering. Questions belong to one of the following categories: yes/no questions, factoid questions, list questions and summaries questions. Participants had to answer these questions with relevant concepts, articles, snippets and RDF triples.

The ImageCLEFmed Task (Kalpathy-Cramer et al., 2011; Müller et al., 2010) ran annually from 2005. Tasks focused on accessing biomedical images in papers and on the Internet. They targeted language-independent techniques for annotating images with concepts; multi-modal IR combining visual and textual features; and multilingual IR techniques.

The IR tasks in the CLEF eHealth evaluation campaign series(Suominen et al., 2013; Kelly et al., 2014; Goeuriot et al., 2015) that are described in this paper represent the first, and to our knowledge the only, evaluation effort focused on health information needs of common people.

3. The Datasets

An information retrieval evaluation test collection is typically composed of:

(1) A large set of documents from which documents relevant to the issued queries need to be identified

(2) A set of topics (composed of a query, contextual information, and any additional information such as images or metadata), representing the information need

(3) Assessments about the relevance of documents to queries

Along with the three parts listed above, the current CLEF eHealth collection also considers translations of queries to languages other than English, and assessments of the readability of documents.

We describe in this section each component of the dataset, and the methodology that led to their creation.

3.1. The Document Collection

A large web set of health resources was used as the document collection for the task. The set contains more than one million web page documents. These have been made

[2] http://www.trec-cds.org/

[3] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[4] http://mednlp.jp/medistj-en

[5] http://www.ncbi.nlm.nih.gov/pubmed

available to CLEF eHealth through the Khresmoi project[6]. This collection consists of web pages covering a broad range of health topics, targeted at both the general public and healthcare professionals. These domains consist predominantly of health and medicine websites that have been certified by the Health on the Net (HON) Foundation[7] as adhering to the HONcode principles[8] (approximately 60–70% of the collection), as well as other commonly used health and medicine websites such as Drugbank[9], Diagnosia[10] and Trip Answers[11]. The crawled documents are provided in the dataset in their raw HTML (Hyper Text Markup Language) format along with their uniform resource locators (URL). The dataset is made available for download on the web to registered participants on a secure, password-protected server. Apart from a few documents excluded because they had rendering problems (web documents that could not be displayed or with encoding problems) or raised copyright issues, the document collections distributed in 2013, 2014 and 2015 are similar.

3.2. The Query Collection

The queries were built from two different sources: medical reports, in 2013 and 2014; medical images in 2015. We report in this section the resources and methods used to create queries from them.

3.2.1. Diagnosed patients questions about their condition (2013 and 2014)

The 2013 and 2014 tasks aimed to study the information needs of people that have been diagnosed with specific conditions or given treatments. In order to obtain queries that are realistic for this information need, queries related to existing medical reports were created. The same set of medical reports was used for this task in both years. The medical reports originate from the de-identified MIMIC-II database[12] (Multiparameter Intelligent Monitoring in Intensive Care, Version 2.5) (Saeed et al., 2011). This corpus contains 200 reports belonging to four categories: discharge summaries (31%); electrocardiograms (27%), echocardiograms (21%) and radiology reports (21%).

Previous evaluation tasks in health IR have used MeSH[13] entries as queries. However, the queries considered by the task presented here are intended to be representative of real patients' information needs and statements. The topics aim to model queries used by laypeople (i.e., patients, their relatives or other representatives) to find out more about their disorders, once they have examined a discharge summary.

Topics to be used in this task were created by registered nurses and clinical documentation researchers involved in the CLEF eHealth consortium. This solution was chosen in place of recruiting patients because of the issues involved with recruitment and privacy. We believe that being in contact with patients on a daily basis and receiving their treatment and discharge summaries, nurses are familiar with patients' information needs and patients profiles, and able to state queries in a manner typical of patients.

A topic is generated for a given disorder and a discharge summary. Different strategies were used to create topics in 2013 and 2014. In both cases, a topic is built from a selected disorder in a given discharge summary.

- In 2013, a disorder was randomly selected from each discharge summary from among those already annotated. This selected disorder represents the main aspect of interest to a patient, e.g. a disorder mentioned in the discharge summary that a patient wants to find out more about.

- In 2014, instead of randomly selecting the disorder, we decided to create queries from the main one. This was done using the field "discharge diagnosis" or "main diagnosis" in the discharge summary. If several disorders were diagnosed, the medical professionals were free to pick one in the list. When this field did not appear in the report, we asked them to select a disorder that appeared to be the main one in the whole report.

An example of discharge summary used is given in Figure 2. Using the pair <*disorder - discharge summary*>, the experts developed a set of topics (and criteria for judging the relevance of documents to the queries, for use in the relevance assessment task described in the next section). Topics were created in a standard TREC format (see Figure 3 for an example), comprising a topic Title (text of the query), a Description (longer description of what the query means), a Narrative (expected content of the relevant documents) and a patient profile (relevant information on the patient identified in the discharge summary).

3.2.2. Self-diagnosis queries (2015)

Queries were manually built with the following process: images and videos related to medical symptoms were shown to users, who were then asked which queries they would issue to a web search engine if they, or their next-of-kin, were exhibiting such symptoms. Thus, these queries aimed to simulate the situation of health consumers seeking information to understand symptoms or conditions they may be affected by; this is achieved using image or video stimuli. This methodology for eliciting circumlocutory, self-diagnosis queries was shown to be effective by Stanton et al. (Stanton et al., 2014). Zuccon et al. (Zuccon et al., 2015) showed that current commercial search engines are yet far from being effective in answering such queries.

Following the methodology in (Zuccon et al., 2015; Stanton et al., 2014), 23 symptoms or conditions that manifest with visual or audible signs (e.g. ringworm or croup) were selected to be presented to users to collect queries. A cohort of 12 volunteer university students and researchers based in the organisers' institutions generated the queries. English was the mother-tongue for all volunteers and they had no

[6] http://khresmoi.eu/

[7] http://www.healthonnet.org

[8] http://www.hon.ch/HONcode/Patients-Conduct.html

[9] http://www.drugbank.ca/

[10] http://www.diagnosia.com/

[11] http://www.tripanswers.org/

[12] http://mimic.physionet.org

[13] (Medical Subject Headings), the NLM controlled vocabulary thesaurus used for indexing articles for PubMed – http://www.ncbi.nlm.nih.gov/mesh

```
Admission Date:  [**2014-03-28**]
Discharge Date:   [**2014-04-08**]
Date of Birth:  [**1930-09-21**]
Sex:   F
Service: CARDIOTHORACIC
Allergies:
Patient recorded as having No Known
Allergies to Drugs

Attending:[**Attending Info 565**]
Chief Complaint: Chest pain
Major Surgical or Invasive Procedure:
Coronary artery bypass graft 4.
History of Present Illness:
83 year-old woman, patient of Dr.
[**First Name4 (NamePattern1) **]
[**Last Name (NamePattern1) 5005**],
Dr. [**First Name (STitle) 5804**]
[**Name (STitle) 2275**], with
increased SOB with activity, left
shoulder blade/back pain at rest, +
MIBI, referred for cardiac cath.
This pleasant 83 year-old patient
notes becoming SOB when walking up
hills or inclines about one year ago.
This SOB has progressively worsened
and she is now SOB when walking
[**01-19**] city block (flat surface).
[]...]

Past Medical History:
arthritis; carpal tunnel; shingles
right arm 2000; needs right knee
replacement; left knee replacement
in [**2010**]; thyroidectomy 1978;
cholecystectomy [**1981**];
hysterectomy 2001; h/o LGIB 2000-2001
after taking baby ASA; 81 QOD
[...]
```

Figure 2: Extract of a discharge summary

particular prior knowledge about the symptoms or conditions, nor had they any specific medical background. This cohort was then somehow representative of the average user of web search engines seeking health advice (although they had a higher education level than average). Each volunteer was given 10 conditions and they were asked to generate up to 3 queries per condition (thus each condition/image pair was presented to more than one assessor[14]). An example of images and instructions provided to the volunteers is given in Figure 4[15].

A total of 266 possible unique queries were collected; of these, 67 queries (22 conditions with 3 queries and 1 condition with 1 query) were selected to be used in this year's task. Queries were selected by randomly picking one query per condition (we called this the *pivot* query), and then manually selecting the query that appeared most similar (called *most*) and the one that appeared least similar (called *least*) to the pivot query. Candidates for the *most* and *least* queries were identified independently by three organisers and then majority voting was used to establish which queries should be selected. This set of queries formed the *English query set* distributed to participants to collect runs. In addition, we developed translations of this query set into Arabic (AR), Czech (CS), German (DE), Farsi (FA), French (FR), Italian (IT) and Portuguese (PT); these formed the *multilingual query sets* which were made available to participants for submission of multilingual runs. Queries were translated by medical experts available at the organisers institutions.

3.3. Relevance Judgements

Relevance assessments were collected by pooling the runs submitted by participants along with baselines provided by the organizers. Assessment was performed by medical professionals[16] and researchers in clinical NLP or medical information retrieval[17]. In 2013 and 2014, relevance assessment was based on a four point scale, which were then also mapped into a binary scale:

- {0: non relevant, 1: on topic but unreliable} → non relevant
- {2: somewhat relevant, 3: relevant} → relevant

In 2015, the relevance assessment was done only on a three point scale, as the label *1: on topic but unreliable* was not used.

In 2013, the top 10 ranked documents from the participants baseline and top 2 priority runs were pooled. In 2014, the top 10 ranked documents from the baseline and top 4 priority runs were pooled. In 2015, the top 10 documents from the baseline and top 3 priority runs were pooled. The different amounts of pooling were due to different level of resources available for relevance assessments each year.

Table 1 gives details on the pool size and the distribution of relevance across the pools. As the table shows, the coverage of the pool is very limited. This is due to the limited resources organizers had.

In 2015, we also investigated the understandability of information provided by retrieved documents. An average user experiences difficulties in understanding a large number of results retrieved by current search engines (Benigeri and Pluye, 2003). The misunderstanding of medical information pose potential risks as people may dismiss serious symptoms or use inappropriate treatments. Thus, we foster research into developing better search engine technology that accounts for document understandability by collecting understandability assessments in 2015. The assessors were asked whether they believed a patient would understand the retrieved documents. Assessments were provided on a four point scale: 0, "It is very technical and difficult to read and understand"; 1, "It is somewhat technical and difficult to read and understand"; 2, "It is somewhat easy to read and understand"; 3, "It is very easy to read and understand".

[14] With exception of one condition, for which only one query could be generated.

[15] Note that additional instructions were given to volunteers at the start and end of the task, including training and de-briefing.

[16] In all three years.

[17] In 2013 and 2014.

	Pool size	Relevance			
		0	1	2	3
2013	6,391	4,316	197	1,439	439
2014	6,800	3,044	547	974	2,235
2015	8,713	6,741	-	1,515	457

Table 1: Pool statistics - number of documents, and distribution across relevance

	Pool size	Readability			
		0	1	2	3
2015	8,713	1,145	1,568	2,769	3,231

Table 2: Statistics for the readability assessment

Table 2 reports on the assessment distribution for each label.

4. Conclusion

In this paper, we described the creation process and the characteristics of three medical IR evaluation datasets, built within the CLEF eHealth evaluation campaign. More than thirty research teams have used these datasets to evaluate their search systems as part of the evaluation campaign. As these datasets have been made publicly available[18] the number of research teams using them for their research projects is growing.

5. Acknowledgements

This work has been supported in part by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°257528 (KHRESMOI), by Horizon 2020 program (H2020-ICT-2014-1) under grant agreement n° 644753 (KCONNECT), by the Austrian Science Fund (FWF) project n° I1094-N23 (MUCKE), and by the Guimuteic project funded by Fonds Européen de Développement Régional (FEDER) of région Auvergne Rhône-Alpes.

6. References

Balikas, G., Partalas, I., Ngomo, A.-C. N., Krithara, A., Gaussier, E., and Paliouras, G. (2014). Results of the bioasq track of the question answering lab at clef 2014. In *CLEF online working notes*, pages 1181–1193.

Benigeri, M. and Pluye, P. (2003). Shortcomings of health information on the internet. *Health Prom. Inter.*

Goeuriot, L., Jones, G., Kelly, L., Leveling, J., Hanbury, A., Müller, H., Salanterä, S., Suominen, H., and Zuccon, G. (2013). ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. In *Online Working Notes of CLEF*. CLEF.

Goeuriot, L., Kelly, L., Lee, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., and Gareth J.F. Jones, H. M. (2014). 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.

Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J., and Zuccon, G. (2015). Overview of the CLEF eHealth Evaluation Lab 2015. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 429–443. Springer.

Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., de Herrera, A., and Tsikrika, T. (2011). The CLEF 2011 medical image retrieval and classification tasks. In *Working Notes of CLEF 2011 (Cross Language Evaluation Forum)*.

Kelly, L., Goeuriot, L., Schreck, T., Leroy, G., Mowery, D. L., Velupillai, S., Chapman, W., Martinez, D., Zuccon, G., and Palotti, J. (2014). Overview of the ShARe/CLEF eHealth evaluation lab 2014. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 172–191. Springer.

Mowery, D., South, B., Christensen, L., Murtola, L., Salanterä, S., Suominen, H., Martinez, D., Elhadad, N., Pradhan, S., Savova, G., and Chapman, W. (2013). Task 2: ShARe/CLEF eHealth Evaluation Lab 2013. In *Online Working Notes of CLEF*. CLEF.

Mowery, D., Velupillai, S., South, B., Christensen, L., Martinez, D., Kelly, L., Goeuriot, L., Elhadad, N., Pradhan, S., Savova, G., and Chapman, W. (2014). Task 2 of the CLEF eHealth Evaluation Lab 2014: Information extraction from clinical text. In *CLEF 2014 Evaluation Labs and Workshop: Online Working Notes*, Sheffield, UK.

Henning Müller, et al., editors. (2010). *Experimental Evaluation in Visual Information Retrieval*, volume 32 of *The Information Retrieval Series*. Springer.

Névéol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., and Zweigenbaum, P. (2015). CLEF eHealth evaluation lab 2015 task 1b: clinical named entity recognition. In *CLEF 2015 Online Working Notes*. CEUR-WS.

Palotti, J., Zuccon, G., Goeuriot, L., Kelly, L., Hanburyn, A., Jones, G. J., Lupu, M., and Pecina, P. (2015). CLEF eHealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In *CLEF 2015 Online Working Notes*. CEUR-WS.

Pradhan, S., Elhadad, N., South, B., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W., and Savova, G. (2013). Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. In *Online Working Notes of CLEF*. CLEF.

Roberts, P. M., Cohen, A. M., and Hersh, W. R. (2009). Tasks, topics and relevance judging for the TREC genomics track: five years of experience evaluating biomedical text information retrieval systems. *Information Retrieval*, 12:81–97.

Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G. (2011). Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. *Critical care medicine*, 39(5):952–960.

Stanton, I., Ieong, S., and Mishra, N. (2014). Circumlocution in diagnostic medical queries. In *Proceedings of the 37th international ACM SIGIR conference on Research*

[18] See `https://sites.google.com/site/clefehealth/` for links to the datasets.

& development in information retrieval, pages 133–142. ACM.

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G., Elhadad, N., Pradhan, S., South, B. R., Mowery, D. L., Jones, G. J., Leveling, J., Kelly, L., Goeuriot, L., Martinez, D., and Zuccon, G. (2013). Overview of the ShARe/CLEF eHealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231. Springer.

Suominen, H., Schreck, T., Leroy, G., Hochheiser, H., Goeuriot, L., Kelly, L., Mowery, D., Nualart, J., Ferraro, G., and Keim, D. (2014). Task 1 of the CLEF eHealth Evaluation Lab 2014: visual-interactive search and exploration of eHealth data. In *CLEF 2014 Evaluation Labs and Workshop: Online Working Notes*, Sheffield, UK.

Suominen, H., Hanlen, L., Goeuriot, L., Kelly, L., and Jones, G. J. (2015). Task 1a of the CLEF eHealth evaluation lab 2015: Clinical speech recognition. In *CLEF 2015 Online Working Notes*. CEUR-WS.

Voorhees, E. M. and Tong, R. M. (2011). Overview of the TREC 2011 medical records track. In *Proceedings of TREC*. NIST.

Zuccon, G., Koopman, B., and Palotti, J. (2015). 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.

```
<query>
  <title> thrombocytopenia treatment corticosteroids length </title>
  <desc> How long should be the corticosteroids treatment
    to cure thrombocytopenia? </desc>
  <narr> Documents should contain information about
    treatments of thrombocytopenia, and especially
    corticosteroids. It should describe the treatment,
    its duration and how the disease is cured using it.
    <scenario> The patient has a short-term disease, or
      has been hospitalised after an accident (little to
      no knowledge of the disorder, short-term treatment)
    </scenario>
    <profile> Professional female </profile>
  </narr>
</query>
```

Figure 3: Example of a topic

```
Imagine you are experiencing the health problem shown below.
Please provide 3 search queries that you would issue to find out what is wrong.
Instructions:
* You must provide 3 distinct search queries.
* The search queries must relate to what you see below.
```

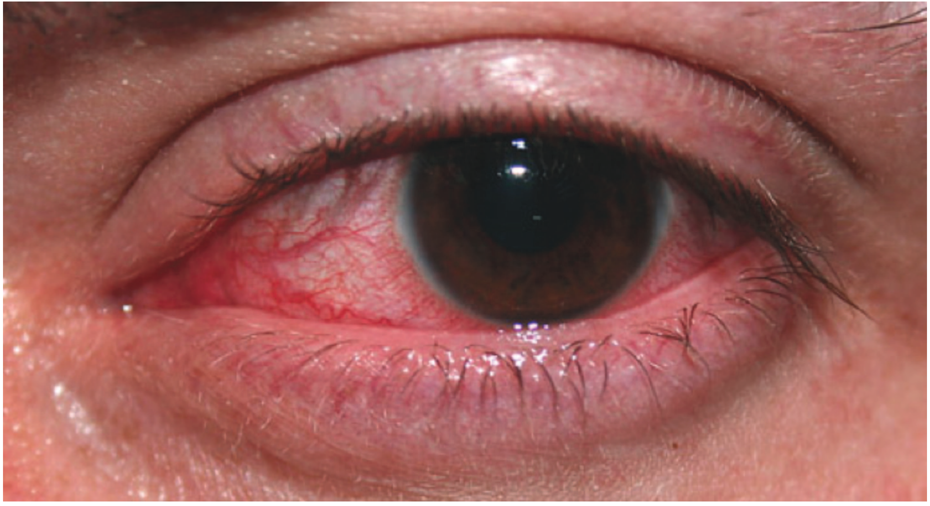

Figure 4: An example of instructions and images provided to volunteers for generating potential search queries.