# Test Collections for Medical Information Retrieval Evaluation

Lorraine Goeuriot, Liadh Kelly, Gareth J. F. Jones
Centre for Next Generation Localisation
School of Computing, Dublin City University, Dublin 9, Ireland
{lgoeuriot, lkelly, gjones}@computing.dcu.ie

## ABSTRACT

The web has rapidly become one of the main resources for medical information for many people: patients, clinicians, medical doctors, etc. Measuring the effectiveness with which information can be retrieved from web resources for these users is crucial: it brings better information to professionals for better diagnosis, treatment, patient care; and helps patients and relatives get informed on their condition. Several existing information retrieval (IR) evaluation campaigns have been developed to assess and improve medical IR methods, for example the TREC Medical Record Track [11] and TREC Genomics Track [10]. These campaigns only target certain type of users, mainly clinicians and some medical professionals: queries are mainly centered on cohorts of records describing a specific patient cases or on biomedical reports. Evaluating search effectiveness over the many heterogeneous online medical information sources now available, which are increasingly used by a diverse range of medical professionals and, very importantly, the general public, is vital to the understanding and development of medical IR. We describe the development of two benchmarks for medical IR evaluation from the Khresmoi project. The first of these has been developed using existing medical query logs for internal research within the Khresmoi project and targets both medical professionals and general public; the second has been created in the framework of a new CLEFeHealth evaluation campaign and is designed to evaluate patient search in context.

## 1. INTRODUCTION

The web is now used as one of the main resources for medical information by multiple user groups seeking to address many different classes of information need. Information Retrieval (IR) aims to provide results in response to user queries which address these information needs. Improving Medical IR constitutes a great challenge, as health is prevalent in everyone's life. Evaluating the effectiveness of IR for medical search tasks is key to developing effective systems and technologies. To date several IR evaluation campaigns have been developed in order to assist assess and improve medical IR methods, for example TREC Medical Record Track [11] or TREC genomics track [10]. However, these campaigns only target certain types of users, mainly clinicians and some medical professionals; and have only examined search of health records and evaluation search queries have mainly been centered on cohorts of records describing a specific case or on biomedical reports.

Some analysis of user query logs in the medical domain show that representative queries would be much shorter, whether they come from experts or non-experts [12, 1]. Thus existing benchmarks have not explored the type of online heterogeneous medical content typically searched by both professional and non professional searchers, they have done this using laboratory style queries which are not representative of the observed querying behaviour of real users.

We describe benchmark creation for medical IR evaluation within the Khresmoi project[1]. Khresmoi aims to develop a multilingual and multimodal search and access system for biomedical information and documents [5]. The project targets three user groups: general public, general practitioners and consultant radiologists. In this paper we focus on medical IR using text search over crawled resources, and hence on the first two user groups. In so doing, we describe two generated benchmarks: the first one has been created from existing query logs for internal research within the Khresmoi project and targets both medical professionals and general public; the second one has been created in the framework of the new CLEFeHealth evaluation campaign as part of the CLEF 2013 benchmark laboratories, and it targets patients only.

This paper is structured as follows: Section 2 presents a brief overview of past and present medical IR evaluation campaigns and the benchmarks used; Section 3 provides an overview of the Khresmoi project; the benchmarks created within the project are described in Section 4 and the future work using these benchmarks is briefly introduced in Section 5.

## 2. MEDICAL IR EVALUATION TO-DATE

In this section, we describe past and current medical IR evaluation campaigns and developed benchmarks. We show that, while these campaigns have been important in facilitating great progress in medical IR, they are very limited in the scope of the medical search tasks addressed and that the behaviour of end users has been overlooked.

### 2.1 Existing Benchmarks

OHSUMED, published in 1994, was the first collection containing medical data used for IR evaluation [6]. The collection contained around 350,000 abstracts from medical journals on the MEDLINE database over a period of five years and two sets of topics: a manually created one and one based on the controlled vo-


HSD 2013, August 1, 2013, Dublin, Ireland.

[1] http://khresmoi.eu/

cabulary thesaurus of the Medical Subject Headings[2] (MeSH). The collection was created for the TREC 2000 Filtering Track but also used for other research on health IR [2, 8].

The TREC Genomics Track, which ran between 2003 and 2007, investigated IR systems on biomedical genomics data [10]. This included tasks ranging from ad-hoc retrieval to document categorisation, passage retrieval, and entity-based question-answering. The test collection contained publications from medical journals and clinical reports related to genes and genomics.

Thus while these tasks were important in exploring search for scientific medical purposes they did not address the needs of less scientifically trained searchers.

The ImageCLEFmed Track on medical image retrieval, which ran between 2003 and 2013, provided several tasks supporting evaluation of medical image search [7, 9]. This included tasks on language-independent methods for the automatic annotation of images with concepts; multimodal IR based on the combination of visual and textual features; and multilingual image retrieval methods. The medical task in ImageCLEF concentrated on access to biomedical images in the literature and on the web. Several challenges of automatic image analysis were tackled in this benchmark by a sometimes large variety of participating research groups. While very important in areas where medical images form a vital part of the search data, these activities have again not addressed the more general medical search needs of many users.

The TREC Medical Records Track ran in 2011 and 2012 [11]. This task was based on a collection of de-identified medical records, queries that resembled eligibility criteria of clinical studies, and associated relevance judgements. Records were grouped into visits, corresponding to a patient admission in the hospital; visits ranged in length from a few hours to in excess of a year. The goal of the track was to find patient cohorts that are relevant to the criteria for recruitment as populations in comparative effectiveness studies. Again while an important search task, this activity did not address search over many of the useful and important resources now available.

Recently, NTCIR (NII Test Collection for IR Systems) launched a new campaign, called MedNLP, which aims to extract specific information from Japanese medical reports, written by physicians about imaginary patients[3]. This includes two identification tasks (i.e., 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. This is currently an exploratory activity, and while related to information access, this does not directly address search.

In summary, these previous campaigns have provided resources for evaluating various health IR techniques, aiming to support clinicians and other healthcare workers. Examples include identifying patient cohorts, searching medical images, and coding diagnoses. However, to date evaluation campaigns have not considered more general information needs such as patients and general practitioners information needs.

## 2.2 Towards Representing User Needs in Benchmarks

As shown in the previous section, existing medical IR evaluation benchmarks are highly oriented towards clinicians. Firstly, the datasets are very specialised: either health records, genomics articles, medline abstracts, etc. To our knowledge, existing benchmarks do not provide general health information that would meet the information needs of patients or GPs. Secondly, the queries themselves describe patient cases or are extracted from medical thesauri such as MeSH. As has been observed [12], medical queries tend to be much shorter than those used in existing benchmarks. The lack of resources representing patients and GPs information needs is motivated by several factors. First, it is much more difficult to target their diverse information needs than those of a community of practice such as clinicians due to differences in, for example, their health knowledge and computer skills. Second, they represent a much wider and more heterogeneous subject population than the populations focused on in other campaigns: patients and their relatives may have different interests, different abilities to interpret health information, and different health profiles. For example, diabetes patients may have more health knowledge on this chronic disease than patients with short-term diseases, and diabetic children will most likely wish to retrieve different types of information than their parents. However, finding documents that solve these information needs of laypeople is critical because of the effect incorrect information may have: cybercondria, self-medication, etc.

# 3. KHRESMOI PROJECT

As background to the development of our evaluation tasks, in this section we provide a brief overview of the Khresmoi project of which these form a part. Khresmoi aims to develop a multilingual and multimodal search and access system for biomedical information and documents [5].

The Khresmoi system is composed of multiple interacting component technologies that aim to help a user retrieve valuable medical information adapted to their requirements - preferred language, medical knowledge, etc. System components include machine translation, information retrieval, summarisation, semantic enrichment, spell checking, etc.

## 3.1 Use Cases

The Khresmoi project targets users who speak different languages, have different medical knowledge levels and different levels of knowledge of the language of the documents. Three use cases have been defined and studied in detail: two groups with general medical interests (general public and general practitioners); and one group of clinicians with specialized expertise (radiologists). Each of these groups have been studied within the project and their information needs and search behaviours have been classified through surveys and concrete scenarios [1].

## 3.2 Khresmoi System Evaluation

A major part of the Khresmoi project is the evaluation of Khresmoi technologies as used by our target user groups in order to assess the success and efficiency of Khresmoi project outcomes. Two types of evaluations are being carried out: user-centred evaluation, involving subjects performing predefined tasks on Khresmoi prototypes; and empirical evaluations, for automated assessment of system performance, both in terms of the effectiveness of individual components and the components in combination, and specifically how they interact in combination. Datasets are created to conduct all of these evaluations in a comprehensive and consistent manner.

## 3.3 CLEFeHealth

In order to extend our investigation of medical IR beyond the scope of the Khresmoi project itself, members of the Khresmoi project team are also participating in the organisation of a health

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

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

related evaluation workshop: CLEFeHealth[4] as part of the CLEF 2013 benchmarking laboratories. The goal of CLEFeHealth is to evaluate systems that support laypeople in searching for and understanding their health information. CLEF eHealth is comprised of three specific tasks related to information access.

The specific use case considered is as follows: Before leaving hospital, a patient receives a discharge summary. This describes the diagnosis and the treatment that they received in hospital.

The first task considered in the workshop aims at extracting names of disorders from the discharge summaries, while the second task requires normalisation and expansion of abbreviations and acronyms present in the discharge summaries. The use case then postulates that, given the discharge summaries and the diagnosed disorders, patients often have questions regarding their health condition. The goal of the third task is to provide valuable and relevant documents to patients, so as to satisfy their health-related information need. One of the features of this scenario is that we are able to identify the patient context in which the search is made from the contents of the discharge report. The role of Khresmoi within CLEF eHealth is as part of the team running the third task.

# 4. MEDICAL IR BENCHMARK CREATION

In this section we describe the benchmarks for medical IR evaluation, developed within the Khresmoi project and the CLEF eHealth workshop. A more detailed description of the benchmark developed for CLEF eHealth is described in [4]. These benchmarks are composed of a document collection, a set of queries and a list of relevant documents for each query. The document collection is shared across both benchmarks and described next. This is followed by details on the query set and relevant document set generation process used in the Khresmoi project test collection.

## 4.1 Khresmoi Document Collection

The Khresmoi document collection consists of a large web crawl of health resources, containing about 1.5 million documents. 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[5] as adhering to the HONcode principles[6] ( 60–70% of the collection), as well as other commonly used health and medicine websites such as Drugbank[7], Diagnosia[8] and Trip Answers[9].

## 4.2 Khresmoi Query Set

In order to perform some of the evaluations mentioned in Section 3.2, queries have been gathered for two use cases: general public and physicians. To obtain a set of queries representative of what our potential end-users would enter in a search system, we collected queries from existing query logs. For the general public, queries have been gathered from Health on the Net (HON) search engine. This query log contains queries issued in various languages, only the English ones were considered here. The physicians queries come from the Trip database[10] query logs. A set of 50 short general public (1-2 words in length), 50 long general public queries (>2 words in length), and 50 general practitioner (average 3 words in length) queries have been created for each use case. They have been manually selected by medical professionals to be representative of Khresmoi end-users. Moreover, they have been manually corrected (if they contained spelling errors) and translated into Czech, French and German. Classical IR dataset provide a description with each of the query in order to support the relevance assessment process. However, a description of the query can only be given by the author of the query when she is performing the search. As this information cannot be retrieved from query logs, it had to be generated by medical experts from selected queries, by estimating or inferring the likely search context based on their experience, and on the Khresmoi user requirements [1]. A category and a description are added manually to each query, as shown in the following example:

```
<query>
<title lang="en"> involuntary trembling or quivering
</title>
<title lang="fr">Tremblement et palpitation involontaires
</title>
<title lang="ge">unwillkürliches zittern oder zucken
</title>
<title lang="cz">neúmyslný třes a chvění</title>
<category>Symptoms</category>
<desc>results should provide possible health conditions
for which this symptom is known and also treatment
options</desc>
</query>
```

## 4.3 CLEFeHealth Query Set

The queries used in this task aim to model those used by patients to find out more about their disorders, once they have examined a discharge summary. The discharge summaries used for the task originate from the de-identified clinical free-text notes of the MIMIC II database, Version 2.5. Disorders have been identified within discharge summaries and linked to the matching UMLS (Unified Medical Language System) concept.

A query is generated for a given disorder and a discharge summary by nursing medical experts. Medical experts were used in this query generation process to overcome issues with patient privacy and recruitment. We believe that, being in daily contact with patients receiving treatments and discharge summaries, nurses are familiar with patients information needs and patient profiles.

65 disorders were randomly selected from a set of 1,006 disorders identified in CLEF eHealth Task 1. For each disorder, a discharge summary containing the disorder itself has been randomly selected. Using the pairs of disorder and associated discharge summary, the medical experts developed a set of patient queries (and criteria for judging the relevance of documents to the queries, for use in the relevance assessment task described in the next section). Queries are generated in the standard TREC format, consisting of a topic title (text of the query), description (longer description of what the query means), and a narrative (expected content of the relevant documents). A field describing the patient profile has also been added. The following example outlines a query:

```
  <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
```

[4] http://nicta.com.au/business/health/events/clefehealth_2013
[5] http://www.healthonnet.org
[6] http://www.hon.ch/HONcode/Patients-Conduct.html
[7] http://www.drugbank.ca
[8] http://www.diagnosia.com
[9] http://www.tripanswers.org
[10] http://www.tripdatabase.com/

```
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>
```

With this approach, five training and fifty test queries have been generated for use in the task. 65 disorders have been selected (i.e. more than the targeted number of queries) because some disorders/-queries may not be answerable using web pages from the document collection. During the query generation process, the experts manually removed disorders from the list of 65 that do not allow for realistic query generation. CLEF eHealth task participants were allowed to use the discharge summaries along with the query as contextual information.

### 4.4 Relevance Assessments

Relevance assessments for the Khresmoi query set were formed based on pooled sets generated using a combination of existing retrieval approaches. Documents in the pooled result sets have been rated as relevant or irrelevant to the queries by medical experts using details of document relevance given in the description field of each query topic. The relevance of each document was assessed by one expert.

Relevance assessments were conducted for the CLEF eHealth query set after task participants submitted their runs. Each participant was required to submit a baseline run that does not incorporate any advanced techniques (e.g., sophisticated annotation, query expansion, etc. techniques), and could submit up to three additional runs generated using the discharge summaries associated with the queries, and up to three runs using techniques of their choice which do not use the discharge summaries. To add diversity, while keeping the relevance assessment load as light as possible, pooled sets for relevance assessment were generated by merging the top 10 documents from participants baseline run, the best run using discharge summaries and the best run without using them, with duplicates removed. Relevance assessment was conducted on a 4-point scale (3: highly relevant, 2: somewhat relevant, 1: on topic but unreliable, 0: not relevant). Two qrel files were created: one which maintains this graded 4-point scale and one which maps this 4-point scale to a binary scale ($\{3, 2\} \rightarrow 1$: relevant, $\{1, 0\} \rightarrow 0$: not relevant).

## 5. FUTURE WORK

In this paper we described the creation of two new medical IR evaluation benchmarks. These benchmarks are rich resources representative of patients and general practitioners information needs. This benchmark generation also allowed us to investigate the creation of realistic query sets and useful contextual descriptions. This has been done either for existing queries, where the context has to be inferred, and made-up queries, where the context was set by real discharge summaries. While there are no other benchmarks covering such a context, their release represents great potential for improvement of medical IR.

In that sense, CLEF eHealth dataset has been released and 9 teams submitted runs to this campaign. Results were promising and their analysis is described in [3]. Participants results, outputs of the CLEF workshop and other fora will be used to improve the design of the task and the datasets for the 2014 lab.

Within Khresmoi, evaluation of the IR system will be conducted using the Khresmoi test collection described in this paper. Moreover, a set of global empirical evaluations will be performed using this same dataset, in order to evaluate the components interactions and the influence of their performances on each other.

## 6. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n$^{o}$ 257528 (KHRESMOI). The relevance judgements were funded by the ESF project ELIAS.

## 7. REFERENCES

[1] C. Boyer, M. Gschwandtner, A. Hanbury, M. Kritz, N. Pletneva, M. Samwald, and A. Vargas. Use case definition including concrete data requirements (D8.2). public deliverable, Khresmoi EU project, 2012.

[2] V. Claveau. Unsupervised and semi-supervised morphological analysis for information retrieval in the biomedical domain. In *Proceedings of COLING*, 2012.

[3] L. Goeuriot, G. J. F. Jones, L. Kelly, J. Leveling, A. Hanbury, H. MÃijller, S. SalanterÃd', H. Suominen, and G. Zuccon. Share/clef ehealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF)*, 2013.

[4] L. Goeuriot, L. Kelly, G. J. F. Jones, G. Zuccon, H. Suominen, A. Hanbury, H. Müller, and J. Leveling. Creation of a new evaluation benchmark for information retrieval targeting patient information needs. In *Proceedings of Evaluating Information Access Workshop (EVIA 2013), NTCIR-10 Conference*, 2013.

[5] A. Hanbury, C. Boyer, M. Gschwandtner, and H. Müller. Khresmoi: towards a multi-lingual search and access system for biomedical information. In *Med-e-Tel*, Luxembourg, 2011.

[6] W. R. Hersh, C. Buckley, T. J. Leone, and D. H. Hickam. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of SIGIR '94*, pages 192–201, 1994.

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

[8] B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley. An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In *Proceedings of CIKM 2012*, 2012.

[9] H. Müller, P. Clough, T. Deselaers, and B. Caputo, editors. *Experimental Evaluation in Visual Information Retrieval*, volume 32 of *The Information Retrieval Series*. Springer, 2010.

[10] P. M. Roberts, A. M. Cohen, and W. R. Hersh. 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, 2009.

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

[12] R. W. White, S. T. Dumais, and J. Teevan. How medical expertise influences web search interaction. In *SIGIR*, pages 791–792, 2008.