Creation of a New Evaluation Benchmark for Information Retrieval Targeting Patient Information Needs

Lorraine Goeuriot[1], Liadh Kelly[1], Gareth J. F. Jones[1], Guido Zuccon[2], Hanna Suominen[3], Allan Hanbury[4], Henning Müller[5], and Johannes Leveling[1]

[1]CNGL, School of Computing, Dublin City University, Ireland

[2]Australian E-Health Research Centre, CSIRO, Brisbane, Australia

[3]National ICT Australia and the Australian National University, Canberra, Australia

[4]Vienna University of Technology, Austria

[5]HES–SO, Sierre, Switzerland

{lgoeuriot,lkelly,gjones}@computing.dcu.ie

ABSTRACT

Searching for health advice on the web is becoming increasingly common. Because of the great importance of this activity for patients and clinicians and the effect that incorrect information may have on health outcomes, it is critical to present relevant and valuable information to a searcher. Previous evaluation campaigns on health information retrieval (IR) have provided benchmarks that have been widely used to improve health IR and record these improvements. However, in general these benchmarks have targeted the specialised information needs of physicians and other healthcare workers. In this paper, we describe the development of a new collection for evaluation of effectiveness in IR seeking to satisfy the health information needs of patients. Our methodology features a novel way to create statements of patients' information needs using realistic short queries associated with patient discharge summaries, which provide details of patient disorders. We adopt a scenario where the patient then creates a query to seek information relating to these disorders. Thus, discharge summaries provide us with a means to create contextually driven search statements, since they may include details on the stage of the disease, family history etc. The collection will be used for the first time as part of the ShARe/-CLEF 2013 eHealth Evaluation Lab, which focuses on natural language processing and IR for clinical care.

Keywords

Information Retrieval, Evaluation, Health Informatics, Test Set

1. INTRODUCTION

Searching for health advice is a common and important task performed by individuals on the web. Nearly seventy per cent of search engine users in the US have conducted a web search for information about a specific disease or health problem [3]. While health information retrieval (IR) is often considered as a domain-specific task [4], it is performed by a large variety of users, including various healthcare workers, but also, and increasingly commonly, by laypeople (e.g., patients and their relatives). This variety of potential information seekers, each characterised by different health knowledge, implies a broad range of information needs, and consequently a requirement for retrieval systems able to satisfy the health information needs of different categories of users.

The growing importance of health IR has provided the motivation for a number of evaluation campaigns focusing on health information. For example, the TREC Medical Records Tracks of the Text REtrieval Conference aim at identifying patient cohorts from medical reports to recruit for user studies [14]. In this task, topics include a particular disease/condition set and a particular treatment/intervention set; demographics or other characteristics may also be part of the topics (e.g., age group and hospitalisation status). Moreover, the ImageCLEFmed Tracks of the CLEF Initiative (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum) have created resources for the evaluation of image search in online resources or biomedical journal articles [7, 9]. However, while addressing different information needs (e.g., finding similar clinical cases vs. journal papers), these previous campaigns have targeted specific groups of users with expert health knowledge (e.g., clinicians and health researchers).

Previous research has considered the information needs of individuals seeking health advice on the web, but these studies mainly analysed query logs from large commercial search engines [15]. To the best of our knowledge, no evaluation campaign has considered the information needs that patients may have regarding their health conditions and provided resources for evaluating IR systems for this task. Such lack of attention to this task arises, at least partially, due to the complexity of assessing the information needs: laypeople that search for health information on the web have very varied profiles, and their queries and searching time tend to be much shorter than those considered in past health IR benchmarks [1, 13].

In this paper, we describe a new evaluation collection for IR and novel methods to generate contextualised statements of patient information needs. These are based on realistic short query state-

ments created in the context of patient discharge summaries. The discharge summaries can be considered as a description of the context in which the patient has been diagnosed with a given disorder and has written a query. The collection will be used for the first time for benchmarking as part of the ShARe/CLEF 2013 eHealth Evaluation Lab[1] (CLEFeHealth 2013), a three-task benchmarking activity coordinated by Shared Annotated Resources and CLEF that focuses on natural language processing (NLP) and IR for clinical care. The first two tasks of CLEFeHealth 2013aim at identifying disorders and expanding shorthands in anonymised discharge summaries (which are given to patients after hospitalisation). The third task focuses on IR methods with the aim to provide a patient with useful information related to their disorders as notified to them on a discharge summary provided to them after a clinical incident. A set of realistic patient queries is generated by healthcare professionals, and a collection of a broad range of medical documents is made available for search by the EU-FP7 Khresmoi project[2]. The discharge summaries originate from the de-identified MIMIC-II database[3] (Multiparameter Intelligent Monitoring in Intensive Care, Version 2.5).

This rest of this paper is organised as follows: Section 2 outlines the main evaluation campaigns on health IR. Section 3 describes the creation of the CLEFeHealth 2013dataset, that is, the document collection, query generation, and relevance assessment. Section 4 introduces the result sets and their evaluation. and Section 5 concludes the paper.

2. BACKGROUND AND RELATED WORK

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 vocabulary thesaurus of the Medical Subject Headings[4] (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.

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.

The TREC Medical Records Track ran in 2011 and 2012 [14]. 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.

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[5]. 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.

In summary, these previous campaigns have provided resources for evaluating various health IR techniques, aiming to support physicians and other healthcare workers. Examples include identifying patient cohorts, searching medical images, and coding diagnoses. However, to date evaluation campaigns have not considered the information needs that laypeople may have regarding their health conditions nor provided resources for evaluating IR systems which seek to meet these needs. The lack of these resources is motivated by several factors. First, it is much more difficult to target the versatile information needs of laypeople than those of a community of practice such as healthcare workers due to differences in, for example, their health knowledge and computer skills. Second, laypeople 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. Third, queries posed by patients are usually very short and often ambiguous or obscure [1], as opposed to, for example, the queries based on eligibility criteria considered by TREC Medical Records Track. This leads to ambiguous descriptions of information needs. However, finding documents that solve these information needs of laypeople is critical because of the effect incorrect information may have on health outcomes.

3. RESOURCE GENERATION

The goal of CLEFeHealth 2013is to evaluate systems that support laypeople in searching for and understanding their health information. CLEFeHealth 2013comprises three tasks. The specific use case that is 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 CLEFeHealth 2013aims 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. To evaluate systems that tackle this third task, we provide potential patient queries and a document collection containing various health and biomedical documents for task participants to create their search system. As is common in evaluation of IR, the test collection consists of documents, queries, and corresponding relevance judgements.

3.1 Document Set

[1] https://sites.google.com/site/shareclefehealth/
[2] http://www.khresmoi.eu
[3] http://mimic.physionet.org
[4] http://www.ncbi.nlm.nih.gov/mesh
[5] http://mednlp.jp/medistj-en

A large web crawl of health resources is used as the corpus for this task. The crawl contains about one million documents, which have been made available to CLEFeHealth 2013through the Khresmoi project [5]. 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[6] as adhering to the HONcode principles[7] (appr. 60–70% of the collection), as well as other commonly used health and medicine websites such as Drugbank[8], Diagnosia[9] and Trip Answers[10]. The crawled documents are provided in the dataset in their raw HTML 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.

3.2 Query Set

The queries used in the task aim to model those 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. 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. Previous evaluation tasks in health IR have used MeSH entries (the MeSH ontology is contained in the UMLS meta-ontology) as queries (see Section 2). However, the queries considered by the task presented here are intended to be representative of real patients' information needs and statements. Thus the possibility of issuing concept-queries is discarded. Layperson queries tend to be short, with an average length less than two words. However, different patients will have different information needs associated with the same query statement. For example, a patient that receives a cancer diagnosis for the first time would have a different information need than a patient at a terminal cancer stage. This type of contextual information related to the patient history is contained in the discharge summary. Thus, the discharge summaries can be used for contextually focused generation of queries. The information in a discharge summary can then be used to determine the relevance of retrieved information to this specific user.

Discharge summaries are semi-structured reports with the following appearance:

```
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
[...]
```

A query is generated for a given disorder and a discharge summary. To better structure the query generation process, patients' information needs have been grouped into three main scenarios:

1. the patient has a short-term disease, or has been hospitalised after an accident (little to no knowledge of the disorder, short-term treatment),
2. the patient has a chronic disease or a long-term disease that has *just* been diagnosed (little to no knowledge of the disorder, long-term treatment), and
3. the patient has a chronic or long-term disease, and this is the n-th diagnosis (potentially good knowledge of the disorder, long-term treatment).

Queries to be used in this task have been created by experts (each expert was a registered nurse and clinical documentation researcher) involved in the CLEFeHealth 2013consortium. This solution has been chosen in place of recruiting patients because of the issues involved with recruitment and privacy. We believe that, being on a daily basis in contact with patients receiving treatments and discharge summaries, nurses are familiar with patients information needs and patient profiles.

65 disorders have been randomly selected from the set of 1,006 disorders identified in the CLEFeHealth 2013Task 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 experts have 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 given following 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). 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
    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. A real log containing queries issued by

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

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

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

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

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

the general public on the HON website has also been used to exclude candidate queries which are unrealistic of the type of query that a patient would typically enter. For each query, an IR system that implements a standard BM25 weighting scheme [11] was used to retrieve a shallow pool of documents. This has been used to assess whether a standard retrieval system could match at least one relevant document to a candidate query. Queries with no relevant documents retrieved in the shallow pool have been removed.

4. RESULT SET AND EVALUATION

To allow task participants to develop effective systems, a set of five training queries with related relevance assessments has been distributed together with the corpus. Relevance assessments for these queries were formed based on pooled sets generated using the Vector Space Model [12] and Okapi BM25 [11]. Pooled sets were created generated by merging the top 30 ranked documents returned by the two retrieval models and removing duplicates.

Documents in the pooled result sets have been rated as relevant or irrelevant to the queries by the aforementioned experts using details of document relevance given in the narrative field of each query topic. The relevance of each document was assessed by one expert.

Relevance assessments on the test queries will be conducted after the task participants have submitted their runs. Each participant is required to submit a baseline run that does not incorporate any advanced techniques (e.g., sophisticated annotation, query expansion, etc. techniques), and can submit up to three additional runs generated using the discharge summaries associated with the queries, and up to three runs using a techniques of their choice.

Different methods to generate pooled result sets will be considered, depending on the number of submissions and the diversity of the retrieval results. Methods include, for example, pooling the baseline and top ranked runs submitted by task participants. Generated pooled result sets will be assessed by the experts. The standard *trec_eval*[11] tool will be used to determine the effectiveness of submitted runs. It is anticipated that the standard evaluation metrics of mean average precision (MAP) and precision at ten (P@10) will be used for this task. Precision at high rank is anticipated to be the measure of most interest to real users.

5. CONCLUDING REMARKS

This paper has described the creation of an evaluation collection for health IR which represents the first effort of producing a benchmarking resource for evaluating systems to support laypeople seeking health advice and information on the web in connection with medical records. The principles used to create a realistic set of queries were also described. These queries were created by healthcare professionals from a set of disorders and discharge summaries. Along with documents and queries, the collection provides a training set of relevance assessments, also developed by healthcare professionals.

This collection is part of the ShARe/CLEF eHealth Evaluation Lab, which is running for the first time in 2013 with three tasks. We anticipate that the outcomes of this task will provide insights on how the collection and evaluation methods can be improved for future evaluation campaigns. We hope that this benchmark will foster research in IR that targets patients' information needs.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 257528 (KHRESMOI). NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

[11] http://trec.nist.gov/trec_eval/

6. 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] S. Fox. Health topics: 80% of internet users look for health information online. Technical report, Pew Research Center, February 2011.

[4] A. Hanbury. Medical information retrieval: an instance of domain-specific search. In *Proceedings of SIGIR 2012*, pages 1191–1192, 2012.

[5] A. Hanbury and H. Müller. Khresmoi – multimodal multilingual medical information search. In *MIE village of the future*, 2012.

[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] S. E. Robertson and S. Jones. Simple, proven approaches to text retrieval. Technical Report 356, University of Cambridge, 1994.

[12] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[13] H. Suominen, editor. *The Proceedings of the CLEFeHealth2012 – the CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis*. NICTA, 2012.

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

[15] R. White and E. Horvitz. Cyberchondria: Studies of the escalation of medical concerns in web search. Technical report, Microsoft Research, 2008.