Building Realistic Potential Patient Queries for Medical Information Retrieval Evaluation

Lorraine Goeuriot[1], Wendy Chapman[2], Gareth J.F. Jones[1], Liadh Kelly[1], Johannes Leveling[1], Sanna Salanterä[3]

[1] Dublin City University, Ireland, [2] University of Utah, USA, [3] University of Turku, Finland

{lgoeuriot, gjones, lkelly, jleveling}@computing.dcu.ie, wendy.chapman@utah.edu, sansala@utu.fi

Abstract

To evaluate and improve medical information retrieval, benchmarking data sets need to be created. Few benchmarks have been focusing on patients' information needs. There is a need for additional benchmarks to enable research into effective retrieval methods. In this paper we describe the manual creation of patient queries and investigate their automatic generation. This work is conducted in the framework of a medical evaluation campaign, which aims to evaluate and improve technologies to help patients and laypeople access eHealth data. To this end, the campaign is composed of different tasks, including a medical information retrieval (IR) task. Within this IR task, a web crawl of medically related documents, as well as patient queries are provided to participants. The queries are built to represent the potential information needs patients may have while reading their medical report. We start by describing typical types of patients' information needs. We then describe how these queries have been manually generated from medical reports for the first two years of the eHealth campaign. We then explore techniques that would enable us to automate the query generation process. This process is particularly challenging, as it requires an understanding of the patients' information needs, and of the electronic health records. We describe various approaches to automatically generate potential patient queries from medical reports and describe our future development and evaluation phase.

Keywords: Query analysis, query generation, medical information retrieval

1. Introduction

Almost all over the world patient care in hospitals is documented carefully including admission information, documentation during the care episode and a discharge summary at the end of care. More and more often these documents are also shown or given to patients. They frequently include information that is not easy to understand by the patient, as they are written or dictated by a physician, nurse, therapist, specialist, or other clinician responsible for patient care. They describe the purpose of a hospital visit, completed procedures and investigations, the chosen treatment and care, a description of the recovery process, the status at discharge (discharge summary), and the future care plan. The primary purpose of discharge summaries is to support the care continuum as a handover note between clinicians, but they also serve legal, financial, and administrative purposes. The patient, her relatives, and other representatives are likely to have difficulties in understanding discharge text as in this simple example sentence from a US discharge: "AP: 72 yo f w/ ESRD on HD, CAD, HTN, asthma p/w significant hyperkalemia & associated arrythmias". These reports are written in specialised language, including abbreviations that are sometimes specific to individuals, as well as specialised vocabulary, that make them very hard to understand for patients (Allvin et al., 2011).

The overall, general objective of our research is to improve information access to health documents. Our research is conducted within within the ShARe/CLEF ehealth evaluation campaign (Suominen et al., 2013). The usage scenario of the campaign is to ease patients and their next-of-kin in understanding eHealth information. eHealth documents are much easier to understand after expanding shorthand, correcting the misspellings, normalizing all health conditions to standardized terminology, and linking the medical concepts to a patient-centric search on the Internet. The evaluation lab contains three tasks, the first one on visualization of eHealth data, the second one on information extraction from clinical data and the last one on patient-centered information retrieval. We are focusing here on the information retrieval task (Goeuriot et al., 2013a). Our main goal is to support patients in understanding their discharge summaries by understanding their information needs and automatically generating queries answering them. Although medical benchmarks for retrieving information related to queries exist, most of them focus on medical professionals' information needs rather than patients' needs. As commercial search engines query logs are generally not shared publicly, queries generated from patients' discharge summaries could provide essential material for patient-oriented novel applications, including generation of information to help patients understand their condition.[1] The automation of the query generation process would firstly greatly help the creation of new evaluation dataset. The more evaluation data there is, the more experiments there will be, and hopefully, the better medical IR system performances will be. Secondly, better IR results would also benefit medical professional: if the professionals could automatically generate queries and obtain relevant documents matching them from a medical report, this would greatly assist their patients in finding out about their condition.

In this paper we present a first step towards automatically creating these queries. After a description of related work (Section 2.), we identify the patients' information needs and detail them in Section 3. In Section 4. we describe the dataset (discharge summaries) and the annotations they

[1] Note that collections of medical documents typically contain sensitive (i.e. patient-related) information, which makes distribution even for research purposes difficult.

contain. In Section 5. we present the queries that have already been manually generated for the past evaluation campaign and the current one, and analyse them in the light of the identified patients' information needs. In the following section (Section 6.), we investigate possible techniques for automatic query creation and describe an evaluation framework for future experiments. Our conclusion and future work are presented in Section 7.

2. Related Work

McMullan (2006) provides a literature review on empirical studies on the use of the Internet for health information search by patients. The main outcome of that study is that the majority of health-related searches by patients target specific medical conditions. The search is carried out by the patient:

- before the clinical encounter to seek information to manage their own healthcare independently and/or to decide whether they need professional help;
- after the clinical encounter for re-assurance and deeper understanding or because of dissatisfaction with the amount of detailed information provided by the health professional during the encounter.

White and Horvitz (2013) study linking between patients' online behaviour and healthcare utilization. Their study is based on query logs obtained from a Microsoft browser toolbar and a survey. It shows a correlation between the query behaviour and the healthcare utilization. These findings fit our task scenario and confirm the patient need for information before and after being discharged from hospital.

To-date studies on patient queries have largely focused on query reformulation or query expansion. Their research is based on the following observations (Zeng et al., 2006):

- the patient queries are short;
- the queries often do not accurately reflect their information needs and are not effective for search.

Plovnick and Zeng (2004) conducted a pilot study where they reformulated consumer health queries with professional terms (UMLS preferred terms). While this method improved the results for queries containing acronyms and layperson terms, it seems to be mainly improving results from PubMed, which might not be the most searched and useful resource for patients. Zeng et al. (2006) describe the system HIQuA (Health Information Query Assistant), aiming at assisting users querying a search system to get health information. This system recommends additional or alternative query terms, and combines three sources: (1) usage patterns of consumers; (2) controlled medical vocabularies; (3) concept co-occurrence in medical literature. Their system recommendations resulted in statistically higher rates of successful queries, but had no statistically significant impact on user satisfaction or ability to accomplish predefined retrieval tasks.

Crain et al. (2010) skipped the interaction with the users, and instead created an information retrieval system based on Language Models and adapted to dialects. They mainly distinguish two dialects (common and technical language), but observed that the literature most often contains mixtures of dialects such as: common with some slang (on general discussion forums); common with technical (on laypeople health portals); and technical (on professional medical portals). Their system, called diATM is an extension of the polylingual topic models, that learn a language model in each language, for each topic. It shows improvement over state-of-the-art methods in providing relevant documents. While there has not been, to our knowledge, many studies on patient queries, the existing studies show that trying to link laypeople query terms to medical professional vocabulary and documents might not be the optimal approach. There is a need for studies which focus on patients' information search, and therefore a need for evaluation datasets.

3. Patients' Information Needs

Patients have several information needs when attempting to understand their discharge summaries. First, they would need to be able to read it, which means that acronyms should be expanded and normalized. Following this, the question of what the summary means arises. Often, patients would like to get more information about their disease: what it is, if it is dangerous, if it can be cured and what kind of changes and possible consequences it brings to their everyday life. Moreover, patients would need information about the possible treatments and discharge medication: what are the effects, the side-effects, is there any alternative treatment, etc. They might also need to know about their rights as patients, as well as contacts for support groups (Heikkinen et al., 2007). Their information needs also vary depending on their disease: patients with long-term diseases will have considerable knowledge related to the disease after a few years, while patients newly diagnosed or with short term diseases will have a very limited knowledge (Nuutila and Salanterä, 2006). Based on previous knowledge about patients typical information needs, we identified a list of information that we will focus on in this study:

1. Main disease(s) diagnosed
    - General Information: symptoms, risks, factors
    - Complications
    - Lifestyle
2. Treatment
    - General information: list of possible treatments
    - Surgery: procedure description, possible complications
    - Medication: description, precaution, side effects

4. Dataset Description

In this section we describe the dataset provided by the ShARe project, on which our query creation is based.

4.1. Discharge Summaries

The discharge summaries for the study were drawn from de-identified clinical reports originating from the ShARe corpus[2], which has added layers of annotation over a subset of the clinical notes in version 2.5 of the MIMIC II database[3]. The corpus contains 200 documents consisting of discharge summaries, electrocardiogram reports, echocardiogram reports and radiology reports, as described in Table 1. They were authored in an intensive care setting.

Table 1: Distribution of document type in the MIMIC corpus.

Type	# docs	per cent (%)
Discharge summary	62	31
Electrocardiogram report	54	27
Echocardiogram report	42	21
Radiology report	42	21
Total	200	100

The dataset used in this paper is a subset of MIMIC II, consisting of the 62 discharge summaries. The discharge summaries are semi-structured reports, they contain fields such as "admission date", "discharge date", "service", "medication", "allergies", etc. However, these fields can be missing or empty. Moreover, the discharge summaries contain many acronyms (e.g. "ICU", "HCT", "EGD", etc.), and highly specialised vocabulary (e.g. "dysphagia", "dysarthria", "ankylosing spondylitis", etc.). They may also contain spelling errors and typing errors. These characteristics make the data challenging to process, both from the point of view of the patient and from the point of view of the computer.

4.2. Annotations on the discharge summaries

Annotation of disorder mentions was carried out as part of the ongoing ShARe project. For this task, the focus was on the annotation of disorder mentions only. There were two parts to the annotation: 1) identifying a span of text as a disorder mention and 2) mapping the span to a UMLS CUI. Each note was annotated by two professional coders trained for this task, followed by an open adjudication step. UMLS represents over 130 lexicons/thesauri with terms from a variety of languages. It integrates resources used world-wide in clinical care, public health, and epidemiology. It also provides a semantic network in which every concept is represented by its CUI and is semantically typed (Bodenreider and McCray, 2003). A disorder mention is defined as any span of text which can be mapped to a concept in the SNOMED-CT terminology and which belongs to the Disorder semantic group. A concept is in the Disorder semantic group if it belongs to one of the following UMLS semantic types: Congenital Abnormality; Acquired Abnormality; Injury or Poisoning; Pathologic Function; Disease or Syndrome; Mental or Behavioral Dysfunction; Cell or Molecular Dysfunction; Experimental Model of Disease; Anatomical Abnormality; Neoplastic Process; Signs and Symptoms. The annotations cover approximately 181,000 words.

[2] https://www.clinicalnlpannotation.org

[3] Multiparameter Intelligent Monitoring in Intensive Care, Version 2.5, http://mimic.physionet.org

5. Manual Query Generation

In this section, we describe the queries and how they were created for the previous eHealth evaluation campaign (2013) and for the current one (2014). An analysis of the 2013 query set is also provided, in the light of the patient needs highlighted in Section 3..

5.1. Topics from 2013

Our campaign provides an evaluation benchmark targeting patients search for medical information. An IR evaluation benchmark is generally composed of a document collection, a set of topics (extended queries), and relevance assessments (identifying for each topic which documents are relevant). Details on the 2013 benchmark can be found in (Goeuriot et al., 2013b).

Our interest here lies in the topics, which are extended queries. As we did not have access to any patient query logs, the decision was made to create queries from the discharge summaries (described in Section 4.1.), and especially to focus on disorders identified in them (described in Section 4.2.. For privacy reasons, we did not work with patients to generate the queries, but rather with experts in the domain, nursing researchers. These health professionals were provided with discharge summaries, in which one disorder had been randomly selected. Based on the contextual information in the report and the disorder picked, the health professionals created queries and additional fields forming the topics (following TREC standards):

Title: the text of the query;

Description: a longer description of what the query means;

Narrative: the expected content of the relevant documents;

Profile: a summary (such as age, gender, condition of the patient).

The task participants used these topics to evaluate their system performance, and were free to use external resources (such as medical terminologies, corpora, etc.), as well as the matching discharge summary as contextual information. The results and description of this 2013 task can be found in (Goeuriot et al., 2013a).

Although the same process was used to build each topic in the task, we observed differences among topics. These differences may be due to the fact that the topics are generated, from a highlighted disorder in a discharge summary, by a human estimating what the information need might be. Therefore, some topics may be directly related to a disease, while others enquire about the relationship between two disorders, or symptoms, for example. We thus categorize queries based on complexity, where complexity is derived from the number of concepts in a query. We define a concept as a specific medical entity (e.g. "diabetes

Table 2: Distribution of topic categories.

Category	# topics
1-concept	26
2-concepts	24
3-concepts	5
Total	55

mellitus" is a concept, "disease" is not). Based on this definition, we annotated the topics with the number of concepts they contained, based on the title (and if necessary, on the description).
The topic distribution for the annotated topics is shown in Table 2. We observe quite a balanced distribution between 1- and 2-concept topics.

5.2. Analysis of topics from 2013 in the light of the patients' information needs

In this section we describe a deep analysis of the queries and their content that has been conducted for this study, to gain a greater understanding of queries, towards automatic query generation. We annotated the queries according to the category of the concept they contained, i.e. disease, symptom, body part, treatment, procedure, care. Among the 1-concept ones, 96% belong to the disease category, which represents 25 queries, the remaining one relates to a body part. The 2-concepts categories distribution is shown in Table 3. Unsurprisingly, most of the multi-concept queries are centered on the link between a disease, and a concept from another category, the majority being another disease or a symptom. 3-concepts topics are more varied, and less numerous, so each of them contains a different combination of categories.
Single concept (1-concept) queries also quite often contain more general terms, that are not considered a concept, but more a facet of the main concept, for example "treatment", or "symptoms". We observed the distribution of these facets on the queries about a disease (N=25), reported in Table 4. When no facet was observed, we considered the patient to be looking for general information about the disease. The majority of the topics (15) have a facet with the disease. The facet can be the care, treatment or the symptoms of the disease. We also observed 4 queries relating to a disease in an acronym form. We considered this case independently, assuming that the patient would, besides general information, seek the meaning of the acronym.
While these queries have not been generated by patients, they give an idea of what a patients concern would be while reading their discharge summary. Moreover, they are very closely related to the information needs listed in Section 3.. While complex queries and relations between different entities have not been listed, we can see that, depending on the context and the knowledge the patient already has on his condition, it can be another information need.

5.3. Topics from 2014

Analysis of 2013 participating teams results showed that using the discharge summaries for contextual information did not improve IR system performance. One of the reasons explaining this is the way disorders within discharge summaries (from which the health professionals generated queries) were selected: they could either be one of the main disorders within the discharge summary, therefore be related to the discharge scenario; or be part of the patient history and mentioned in the discharge summary only for documentation, therefore not be related to the discharge scenario. Following discussion at the evaluation campaign workshop between participants and organizers, several solutions for building queries for the 2014 campaign were considered:

- Keep the topic creation based on a single disorder, therefore investigate ways to improve the selection of the disorder;
- Broaden the elements the topic could be based on: medication, surgery, care, etc.
- Consider the whole discharge summary as the topic/query.

The third option would make the task completely different from what it currently is, as it would require much more query processing: if the query is a full discharge summary, it would have to be processed first. Moreover, it would require a lot of preparation work, especially conducting the relevance assessment. As we would like the task to still be centered on the challenging improvement of health information retrieval, this option has not been further investigated.

Table 3: Distribution of categories in 2- and 3-concepts queries.

Categories	# topics
disease *AND* disease	10
disease *AND* symptom	5
disease *AND* body part	4
disease *AND* treatment	2
disease *AND* procedure	2
symptoms *AND* symptoms	1
disease *AND* disease *AND* disease	1
disease *AND* disease *AND* symptom	1
disease *AND* disease *AND* treatment	1
disease *AND* symptoms *AND* symptoms	1
symptom *AND* symptom *AND* symptom	1

Table 4: Distribution of facets in 1-concept queries relating to a disease category.

Facets	# topics
general information	10
acronym	4
care	4
treatment	4
symptoms	2
heredity	1

Table 5: Number of disorders listed in the discharge diagnosis field.

# Disorders	# Discharge Summaries	%
1	14	27
2	9	17.5
3	9	17.5
4	7	13
5+	13	25

Considering the first and second query generation methods, for time-related reasons, the first query generation method was selected for generation of the 2014 topics. A patient hospitalization is generally the cause of one or more disorders. These are often specified in the discharge summaries, or can be identified by health professionals. An analysis of the corpus described in Section 4.1. showed that among the 62 discharge summaries, 52 contained a field called *Discharge Diagnosis* or *Final Diagnosis*, containing a list of disorders the patient had been treated for. The number of disorders per discharge diagnosis is shown in Table 5. The topic creation has been made based on these identified diagnosis: if it contains only one disorder, the query is centered on it; if it contains more than one, the health professionals pick the disorder(s) they think the patient would first ask about. To avoid ending up with a biased set of topics, the health professionals were not given any specific guidelines for the disorder selection and quantity. As the task provides 5 training topics and 50 test topics, and only 52 discharge summaries contain a discharge diagnosis, it has been agreed with the health professionals that they would pick the main disorder from the remaining 3 summaries (in which all the annotated disorders were highlighted).
The structure of the topics remains the same as for 2013[4].
The second topic generation option, involving broadening the elements of the topic still has to be investigated. One approach to achieve this would be to combine our knowledge in patients' information needs and the information contained in the semi-structured and annotated discharge summaries to automatically generate patient queries. We present in the following section our primary investigations on this.

6. Towards Automatic Creation of Patients' Queries

6.1. Existing approaches

Ganguly et al. (2011) investigated generating queries automatically in the context of evaluation campaigns such as TREC 2010 session track, where the focus is the whole user session rather than single-queries. They reformulate existing queries, achieving either *specification*, where the reformulated query expresses a more precise information need, or *generalization*, where the reformulated query expresses a more general information need. They use a statistical corpus-based approach to retrieve more specialized or generalized terms to reformulate the query. As our main target here is to generate queries that are representative of patients information needs (over getting variants of existing queries), specification and generalization would not necessarily be the best approach. However, a corpus or ontology-based approach could allow identification of variants (e.g. synonyms, related disorders, or matching treatments) that could be used to generate query variants from an existing set. A risk would be the loss of the connection between the query and the discharge summary, and generating too homogeneous sets of queries.
Another strategy to automatically build queries from long documents involves summarizing them. Using classical summarization techniques based on statistical selection of segments in the text, a shorter version of a text (i.e. a query or document such as the discharge summary) can be generated to improve retrieval system performance (Arora et al., 2013). Using such a system in our case would require careful tuning of the summarization system, as classical summarization methods require adaptation to perform well on medical texts. This has been shown with scientific research articles (Nguyen and Leveling, 2013) and is even more the case with more condensed and less structured data such as medical reports.
A different approach would be based on the existing numerous work conducted on information extraction from electronic health records. Based on the existing annotations provided by the ShARe project (described in Section 4.2.), and the common semi-structure provided by the discharge summaries, information required to identify the main topics of interest of patients and generate queries from them. Based on the patients' information needs detailed in Section 3., the main fields that need to be identified in the documents are:

- the main disorders, often listed in the "discharge diagnosis" field of the discharge summaries;
- the treatment, often described in the "Discharge medication" field.

6.2. Evaluation Plan

We are planning on comparing these three methods. We will base our comparative analysis on two criteria:

- the quality of the generated queries, based on their relevance to the discharge summary, readability, and their usability;
- the quality of the results retrieved using standard IR systems: their relevance to the discharge summary, and the user satisfaction.

The quality of the generated queries will be manually assessed. Based on a given discharge summary, health professionals will rate generated queries according to their readability and relevance to the discharge summary. Information retrieval experts will judge how usable the queries for an IR evaluation task are, as our primary goal is to generate IR evaluation datasets.
As the set of queries for each discharge summary will be different depending on the method used, their relevance will be judged against the discharge summary (e.g. is that

[4]No example topic can be provided as the task is still running and only registered participants have access to the dataset.

query relevant to a patient receiving this discharge summary?). Similar to classical IR relevance assessment, health professionals will have to assess for each document its relevance to the discharge summary or a part of it.

7. Conclusion and Future Work

In this paper we investigated the generation of patient queries for an IR evaluation task. We first identified the patients' information needs, and then described how we manually created queries within CLEF eHealth. We then considered various ways to automatically generate queries, and how these methods could be adapted to medical IR: reformulation, summarization, and information extraction. We proposed a comparative evaluation plan to investigate these three approaches.

Our next step is to implement this comparative evaluation of the automatic patient query generation approaches. Another aspect worth exploring is the structure and the content of these queries, and the way this affects information retrieval performance. The evaluation campaign provides a privileged context to perform such experiments, with runs from various teams.

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) and the ESF project ELIAS.

8. References

Allvin, H., Carlsson, E., Dalianis, H., Danielsson-Ojala, R., Daudaravicius, V., Hassel, M., Kokkinakis, D., Lundgren-Laine, H., Nilsson, G., Nytrø, O., Salanterä, S., Skeppstedt, M., Suominen, H., and Velupillai, S. (2011). Characteristics of finnish and swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. *Journal of Biomedical Semantics*, 2(Suppl 3).

Arora, P., Foster, J., and Jones, G. J. F. (2013). DCU at FIRE 2013: Cross-language !ndian news story search. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE) 2013*.

Bodenreider, O. and McCray, A. (2003). Exploring semantic groups through visual approaches. *Journal of Biomedical Informatics*, 36:414–432.

Crain, S. P., Yang, S.-H., Zha, H., , and Jiao, Y. (2010). Dialect topic modeling for improved consumer medical search. In *Proceedings of AMIA Annual Symposium*, pages 132–136.

Ganguly, D., Leveling, J., and Jones, G. (2011). Automatic generation of query sessions using text segmentation. In *Proceedings of the Information Retrieval Over Query Sessions Workshop at ECIR 2011*.

Goeuriot, L., Jones, G. J. F., Kelly, L., Leveling, J., Hanbury, A., Mller, H., Salanter, S., Suominen, H., and Zuccon, G. (2013a). Share/clef ehealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In *CLEF online working notes*.

Goeuriot, L., Kelly, L., Jones, G. J. F., Zuccon, G., Suominen, H., Hanbury, A., Mller, H., and Leveling, J. (2013b). Creation of a new evaluation benchmark for information retrieval targeting patient information needs. In Song, R., Webber, W., Kando, N., and Kishida, K., editors, *Proceedings of the 5th International Workshop on Evaluating Information Access (EVIA), a Satellite Workshop of the NTCIR-10 Conference*, Tokyo/Fukuoka, Japan. National Institute of Informatics/Kijima Printing.

Heikkinen, K., Leino-Kilpi, H., Hiltunen, A., Johansson, K., Kaljonen, A., Rankinen, S., Virtanen, H., and Salanterä, S. (2007). Ambulatory orthopaedic surgery patients' knowledge expectations and perceptions of received knowledge. *Journal of Advanced Nursing*, 60(3):270–8.

McMullan, M. (2006). Patients using the internet to obtain health information: How this affects the patienthealth professional relationship. *Patient Education and Counseling*, 63:24–28.

Nguyen, D. T. and Leveling, J. (2013). Exploring domain-sensitive features for extractive summarization in the medical domain. In *Proceedings of the Conference on Application of Natural Language to Information Systems (NLDB 2013)*, pages 90–101.

Nuutila, L. and Salanterä, S. (2006). Children with a long-term illness: parents' experiences of care. *Journal of Pediatric Nursing*, 21(2):153–60.

Plovnick, R. M. and Zeng, Q. T. (2004). Reformulation of consumer health queries with professional terminology: a pilot study. *Journal of Medical Internet Research*, 6(3).

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G. K., Elhadad, N., Pradhan, S., South, B. R., Mowery, D., Jones, G. J. F., Leveling, J., Kelly, L., Goeuriot, L., Martínez, D., and Zuccon, G. (2013). Overview of the share/clef ehealth evaluation lab 2013. In Forner, P., Müller, H., Paredes, R., Rosso, P., and Stein, B., editors, *CLEF*, volume 8138 of *Lecture Notes in Computer Science*, pages 212–231. Springer.

White, R. W. and Horvitz, E. (2013). From health search to health care: Explorations of intention and utilization via query logs and user surveys. *Journal of the American Medical Informatics Association (JAMIA)*.

Zeng, Q. T., Crowell, J., Plovnick, R. M., Kim, E., Ngo, L., , and Dibble, E. (2006). Assisting consumer health information retrieval with query recommendations. *Journal of the American Medical Informatics Association*, 13(1):80–90.