EXPERIMENTAL EVALUATION OF QUERY REFORMULATION TECHNIQUES IN THE CONTEXT OF MEDICAL INFORMATION RETRIEVAL

[1]MOHAMMED MAREE*, [2]ISRA NOOR, [1]KHALID RABAYAH

[1]Department of Information Technology, Arab American University, Palestine

[2]Department of Computer Science, Arab American University, Palestine

[1]{mohammed.maree, khalid.rabayah}@aauj.edu, [2] e.noor@students.aauj.edu

ABSTRACT

With the proliferation of online medical information, a majority of laypeople (ordinary people with little medical background knowledge) and medical specialists now find the Web an indispensable tool for searching for medical information in various domains of interest. However, when using existing medical search engines, the precision of the retrieved results is governed by two main factors. First, users (be they laypeople or medical professionals) need to submit vocabularies that best describe their information needs. Second, the quality of the returned results is largely based upon the effectiveness of the techniques and medical knowledge resources that are exploited by such search engines. Several systems and approaches have been proposed to address problems associated with each of these factors independently; however, little attention has been paid to cooperatively address problems of both factors. In this article, we aim to investigate the impact of exploiting medical knowledge resources and information retrieval techniques on 1) reformulating medical queries through enriching them with semantically-related medical terms and 2) indexing documents in the medical domain to improve the matching process between the reformulated queries and their corresponding medical documents. A prototype of the proposed system has been instantiated and experimentally validated using CLEF2014 eHealth Dataset and state-of-the-art effectiveness indicators. The produced results by our system demonstrate that the quality of the returned results has improved compared to other similar medical information retrieval systems.

Keywords: *Query Reformulation, Medical Queries, CLEF eHealth Dataset, Medical Knowledge Bases, Precision/Recall Indicators*

1. INTRODUCTION

A wide range of users, including patients, medical researchers, general physicians, and professionals with specific expertise such as radiologists and oncology specialists are interested in medical information. The diversity of users, their information needs and their background knowledge in the medical domain have a large impact on the effectiveness of Medical Information Retrieval (MIR) systems [1, 2]. Each of these factors plays an important role in the way in which a medical query is formulated on the one hand, and has a direct impact on the quality of the retrieved results by MIR systems on the other [3-6]. Another important factor that affects the quality of the returned results by MIR systems is associated with the richness and domain coverage of their underlying medical semantic resources; which provide formal and explicit specifications of shared medical conceptualizations [7, 8]. Recently, several semantic resources and classification systems have been developed in the medical domain. Examples of such resources are the: Unified Medical Language System (UMLS) [9], Medical Subject Headings (MeSH) [10] terms, Systematized Nomenclature of Medicine - Clinical Terms SNOMED-CT [11], Logical Observation Identifier Names and Codes LOINC [12], DRUG, Gene and Human Disease ontologies [13-15], ICD-10 standard [16], and Pubmed [17]. For more details on existing medical semantic resources please refer to the bioportal[1] gate. Although the exploitation of medical semantic resources has proved to be more effective than conventional approaches that merely employ conventional information retrieval techniques, the quality of the result produced by medical resources-based approaches still needs further improvement as reported in [2, 6, 7]. This is mainly because existing medical resources suffer

[1] http://bioportal.bioontology.org/ontologies

from knowledge incompleteness and semantic heterogeneity issues [18]. Motivated by these observations, we present a semantics-based MIR system that aims at enhancing the quality of the produced medical search results through incorporating a semantics-based query reformulation and document indexing scheme. In this context, when a user submits a medical query, the proposed system reformulates the query through enriching it with semantically-related medical concepts using multiple external medical resources. The reformulated queries are then mapped to their corresponding medical documents. In our approach, for each medical document, a semantics-based inverted index is automatically constructed through utilizing the same medical resources that we exploit for reformulating users' queries.

The main contributions of our work are summarized as follows:

1. Employing multiple medical semantic resources for:
   a. Reformulating users' queries through enriching them with semantically-related medical concepts.
   b. Building a semantically-enhanced inverted index for indexing medical documents after capturing the hidden semantic dimensions that are encoded in the text of each medical document.
2. Classifying and re-weighting query terms based on the employed medical semantic resources. In this context, medical terms are identified and assigned higher weights against other less significant supportive query terms.

To evaluate the quality of our proposed system, we use CLEF e-Health 2014 medical dataset. We have developed a prototype of the proposed system with a medical search interface that facilitates users' access to medical documents in the dataset. A user in this context can submit her/his medical query in the form of natural language query. The system accordingly assigns relevance scores between each query-document pair based on their semantic similarity. As we present in the (Experimental Results) section, our system proved to produce promising results and outperformed similar state-of-the-art MIR systems that use the same dataset for evaluation purposes.

The rest of this paper is organized as follows. In Section 2, we review the literature and identify the main strengths and weaknesses of a number of works that are related to our proposed approach. Section 3, presents an overall description of the proposed system. The detailed features of the proposed system and its modules are discussed in section 4. Section 5 introduces the experimental setup for the proposed system, highlights the characteristics of the used dataset and discusses the details of the obtained results. Section 6 introduces the conclusions and presents the further improvements that we plan to incorporate in the next version of the system.

2. RELATED WORK

The interest in helping laypeople and medical professionals in accessing reliable medical resources has increased in the last few years [1-6, 8, 19-25]. In [1] the authors study the contexts in which a non-expert uses many words to describe a symptom instead of using the appropriate medical terms. The authors propose a supervised approach to link searched queries to medical concepts that can be mapped to specific disease/s. They use the professional definition of diseases from medical semantic resources to reformulate the user's query into a professional query that consists of medical concepts. Although the proposed approach achieved an improvement in mapping symptoms to the proper relevant disease/s, the authors ignored other query types (those that do not belong to symptoms and diseases) such as laboratory tests, medical devices,…etc. In [23] the authors propose a "bag of concepts" medical information retrieval model where they extract the medical concepts that exists in the user's query using medical semantic resources. The retrieval process is based on those selected concepts and their mappings to other concepts in the used medical resources. However, the proposed approach suffers from two main drawbacks. First, all non-medical terms that exist in the original query are ignored in this retrieval model. Second, some medical concepts that exist in the user's query are not recognized by the used medical semantic resources due to limited domain coverage problems. In [24] the authors propose a concept-based query expansion model using selective query concepts. To do this, the authors used CLEF eHealth14 dataset in their experiments where each query in the dataset has a related "discharge summary report" as an example of what the relevant results of each query should be like. The authors used UMLS to extract and expand medical concepts in the given query, and they ignored all concepts that exist in the user's query but not in its related discharge summary. The proposed system was not able to achieve except slight improvements on the quality of the produced results because of two reasons. First, the authors didn't consider tackling problems associated with compound terms and stop words that may exist in the input queries or in the medical documents. Second, they restricted the expansion scope to the content of the query-related discharge summary report provided in the dataset, while this report is provided as an example of the related results only. The system proposed in [25] focused on using local resources for query reformulation rather than using external medical semantic resources and NLP techniques. In this context, the authors used Pseudo Relevance Feedback (PRF) model for query reformulation based on terms occurring in the top-k documents retrieved by the system in its initial run. In addition, the exploited the medical concepts that exist in the discharge summary that is related to each query in the CLEF eHealth dataset. The main limitation of this approach is the utilization of local resources that suffer from a limited number of medical concepts to be mapped to their corresponding terms in the given queries. In [22] the authors analyzed the results retrieved by two commercial web search engines (Google and Bing) on a set of queries formulated by laypeople to describe medical symptoms. The authors found that three out of the top ten retrieved results by both search engines were relevant and obtained from trustworthy websites. The authors conclude that existing commercial search engines cannot perform well when they are used in specified domains such as the medical domain. In our proposed approach, we attempt to fill the semantic gap between medical queries and their corresponding medical documents using multiple medical semantic resources and query reformulation techniques. We use trusted and well-recognized external medical semantic resources (UMLS lexicon and UMLS metathesaurus) for medical concepts extraction and expansion. We also propose an approach for re-weighting medical terms in the given user's query through assigning higher weights to such terms against other non-medical query terms. In the next sections, we provide more details on the proposed techniques and experimentally demonstrate their effectiveness using CLEF eHealth dataset.

3. SYSTEM OVERVIEW

In this section, we present an overview of the proposed system's architecture and discuss its main components. As discussed in the previous section, we propose an approach that exploits query reformulation techniques and a set of external medical semantic resources to retrieve medical documents that best match the users' information needs.

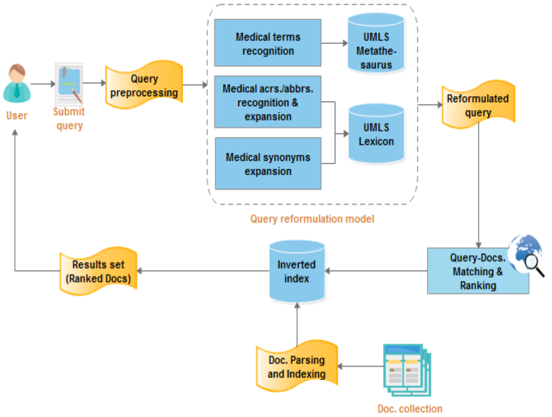
User
Submit query
Query preprocessing
Medical terms recognition
UMLS Metathe-saurus
Medical acrs./abbrs. recognition & expansion
UMLS Lexicon
Medical synonyms expansion
Query reformulation model
Reformulated query
Query-Docs. Matching & Ranking
Inverted index
Results set (Ranked Docs)
Doc. Parsing and Indexing
Doc. collection

Figure 1: Overall Architecture of the Proposed System

As depicted in Figure 1, when a user submits a medical query, the query is pre-processed using a set of Natural Language Processing (NLP) steps including tokenization, stop words removal and stemming (using Porter Stemmer). The output of the pre-processing phase is further processed to identify medical terms, medical acronyms, medical abbreviations, and medical synonyms. This step is carried out based on the exploited medical semantic resources. In other words, we first submit all uni-gram tokens to the UMLS lexicon to capture all acronyms, abbreviations, and synonyms. Then, we define a sliding window to identify all compound terms that are either bi or tri grams. We also use MetaMap to recognize medical terms from the user's query based on the UMLS metathesaurus. In our proposed approach, tokens (be they uni, bi, or tri grams) that correspond to medical terms, medical acronyms, medical abbreviations, or medical synonyms are assigned higher weights compared to the rest of the tokens (known henceforth as supportive terms) in the input query. On the other hand, we utilize the same NLP and query reformulation techniques to index each medical document in the documents dataset. In this context and unlike conventional approaches that use the bag-of-words model to index medical documents, we construct an inverted index that stores medical terms and their semantically-relevant terms that are obtained from the exploited medical semantic resources. It is important to point out that this step is carried out offline to reduce the complexity of the matching process between user queries and their corresponding medical documents.

The next section details the modules of our proposed approach.

4. DETAILED MODULES OF THE PROPOSED APPROACH

When a user submits a medical query, it may contain any of the following components:

- **Acronyms**: are terms that are formed from the initial letters of some longer names and are pronounced differently than the full representation (such as: ARV that stands for 'Adelaide River Virus', 'Average Rectified Value' or other medical terms).
- **Abbreviations**: are terms that are written differently from their full representations, but are pronounced the same (such as: Abd that stands for 'Abduction').
- **Medical Terms**: are terms that can be mapped to medical concepts in the exploited medical semantic resources (such as: aortic).
- **Other Supportive Terms**: any other terms in the user query that could not be classified as acronyms, abbreviations, or medical terms (such as: replacement, status, ..etc).

The following example illustrates these components:

Example 1- Given the following two queries (Uq_1 and Uq_2):

- Uq_1 : MRSA and wound infection, and its danger (QTRAIN2014.1 of CLEF e-health2014 dataset [26]).
- Uq_2 : Peptic Ulcer disease (qtest2014.35 of CLEF e-health2014 dataset [26]).

To process the queries in Example 1, we utilize the following modules:

4.1 Query Preprocessing

In the preprocessing phase, we first use conventional NLP techniques to process the user's query. We first remove all punctuation marks from the user’s query. Then, we remove stop words based on a pre-defined list such as: a, the, an,…etc. After that, the query terms are stemmed using Porter stemmer [27]. Finally, we use the NLP ngrams tokenization technique to tokenize the input text into n-gram tokens of lengths from 1 to 3. Accordingly, the output of the user’s query in Example 1 becomes as follows:

- **For** Uq_1 :
  - List of unigrams in (Ut1): [mrsa, wound, infect, danger]
  - List of bigrams (Bt1): [mrsa wound, wound infect, infect danger]
  - List of trigrams (Tt1): [mrsa wound infect, wound infect danger]
- **For** Uq_2 :
  - List of unigrams (Ut2): [peptic, ulcer, diseas]
  - List of bigrams (Bt2): [peptic ulcer, ulcer diseas]
  - List of trigrams (Tt2): [peptic ulcer diseas]

4.2 Medical Acronyms and Medical Terms Recognition

During this module, an automatic extraction of query components such as: medical acronyms, medical abbreviations, medical terms, and other supportive terms is carried out. An automatic extraction of the synonyms of medical query terms is also carried out during this module. To extract both medical acronyms and abbreviations, we use the UMLS lexicon [9] that is provided by the National Library of Medicine (NLM). We use the ACRONYM table from the lexicon to extract and expand medical acronyms and abbreviations (being uni, bi, or tri grams) given in the user’s query. We also use UMLS lexicon to find the synonyms of all medical query terms by using LEXSYNONYM table from the lexicon. The query is then reformulated by incorporating all of the full representations of the extracted acronyms and abbreviations, and also by adding the extracted synonyms. After applying this step, the following lists are be added to the output of the previous step:

- **For** Uq_1 :
  - List of medical acronyms (ACt_1): [mrsa]
  - List of medical abbreviations (ABt_1): []
  - List of full representations of both acronyms and abbreviations (Et): [methicillin resistant staphylococcus aureus]
  - List of synonyms (SYt_1): [vulnerat]
- **For** Uq_2 :
  - List of medical acronyms (ACt_2): []
  - List of medical abbreviations (ABt_2): []
  - List of full representations of both acronyms and abbreviations (Et_2): []
  - List of synonyms (SYt_2): [digest, ulcu, mal]

On the other hand, to extract medical terms from the user’s query, we use the MetaMap tool which maps biomedical texts to the UMLS Metathesaurus. It locates all the UMLS concepts associated with terms in biomedical texts using the knowledge intensive method that is based on symbolic, natural language processing and computational linguistic techniques detailed in [28]. The result of this step is the list of medical terms that is described below:

- **For** Uq_1:
  - List of medical terms (Mt_1): [wound, infect]
- **For** Uq_2:
  - List of medical terms (Mt_2): [peptic, ulcer, diseas, peptic ulcer, peptic ulcer diseas]

All remaining terms that are not recognized using the previous steps are considered as supportive terms. In Example 1, the supportive term lists contain the following:

- **For** Uq_1:
  - List of supportive terms (St_1): [danger]
- **For** Uq_2:
  - List of supportive terms (St_2): [].

In this context and after utilizing the previous modules, we are able to reformulate the given original query in the following form:

Example 1 – Reformulated Queries (Rq_1 and Rq_2):

- Rq_1 : mrsa, wound, infect, danger, methicillin resistant staphylococcus aureus.
- Rq_2: peptic, ulcer, diseas, digest, ulcu, mal, peptic ulcer, peptic ulcer diseas

4.3 Matching and Ranking

In this section, we discuss the proposed matching and ranking formulas that we use to improve the effectiveness of our proposed system. The key idea of our proposed ranking technique is to assign higher weights for medical terms (Mt, ACt, ABt, SYt) against other supportive terms St, and also against the full representations of acronyms and abbreviations Et that are automatically added to the original user's query. We use the vector space model (VSM) that is usually known as cosine similarity model [29] for finding the similarity between the user's query Uq and the document d in the document collection D. The cosine similarity model employs the $tf-idf$ weighting scheme to assign a weight for each term t in a document d. In our approach, we used the $Normalized-tf_{t,d}$ where the term occurrences are usually normalized to prevent a bias towards longer documents (which may have a higher term count regardless of the actual importance of that term in the document) to give a measure of the importance of the term t within the particular document d:

$$Normalized-tf_{t,d} = \begin{cases} tf_{t,d} / |d| & if \quad tf_{t,d} > 0 \\ 0, & Otherwise \end{cases} \qquad (1)$$

Where $tf_{t,d}$ is the number of occurrences for term t in d, and $|d|$ is the length of the document d.
We propose the following formula for calculating the occurrences of query terms $tf_{t,q}$ to give a highe r weight for medical terms Mt, medical acronyms ACt, medical abbreviations ABt, and medical synonyms SYt, against other supportive terms St and other semantically-related concepts Et that are added to the user's query:

$$tf_{t,q} = \begin{cases} tf_{t,q} / |Uq| & if \quad t \in [Mt, ACt, ABt, SYt] \\ 0.5 \times (tf_{t,q} / |Uq|) & if \quad t \in [St] \\ 1 / |Rq| & if \quad t \in [Et] \end{cases} \qquad (2)$$

where $|Uq|$ is the length of the original query and $|Rq|$ is the length of the reformulated query. In formula (2), we give the medical synonym SYt (that we add to the original user query), the same weight as its original form which is typed by the user (In Example1-Uq_2: the term 'peptic' and its synonym 'digest' have the same weight). But, we reduce the weight of all other terms Et that are semantically related to the original user query terms but with semantic relation other than synonymy (In Example1- Uq_1: the term 'mrsa' is given higher weight than its full representation 'methicillin resistant 'staphylococcus aureus').
The cosine similarity model deals with both document d and query Uq as vectors. Let $\vec{d}$ be the vector representation of d, and $\vec{Uq}$ is the vector representation of Uq. To find the similarity between these two vectors, the following formula is employed to assign the relevance score between a given document d and a user query Uq based on their dot product as follows:

$$sim(d, Uq) = \cos ine(d, Uq) = \frac{\vec{d} . \vec{Uq}}{|\vec{d}||\vec{Uq}|} \qquad (3)$$

The following algorithmic steps demonstrate the matching process between each reformulated query and its corresponding medical documents:

Algorithm 1. Matching between the reformulated queries and their corresponding medical documents

```
Input: Rq_terms_list [t1, t2, …,tn]
Output: list of relevant medical documents
temp_doc_list ←⟨⟩;
relevant_doc_list ←⟨⟩;
for i←0; i < Rq_terms_list.length; i++
    temp_doc_list ← GET_DOCS_FROM_INDEX
    (Rq_terms_list[i]);
    for j←0; j < temp_doc_list.length; j++
        if temp_doc_list[j] Not IN related_doc_list then
            ADD (related_doc_list, temp_doc_list[j]);
        end if
    end for
end for
Return relevant_doc_list;
```

As shown in Algorithm 1, the results of the matching function are returned as a list of relevant medical documents that are ordered in a descending manner starting from the most relevant document (the first result with the highest number of matching terms) moving downwards towards the least relevant document (with relevance score > 0). In the next section, we discuss in details the experiments that we have carried out to validate our proposal approach.

5. EXPERIMENTAL RESULTS

This section describes the experiments that we have carried out to evaluate the techniques of our proposed system. We start by describing the dataset used for evaluating our system (CLEF e-Health 2014 dataset). Next, we present the details of the conducted experiments and compare the results produced by our system to state-of-the-art MIR systems that have used the same dataset for assessing the quality of their proposed techniques.

5.1 Dataset

In order to evaluate the effectiveness of the proposed system, we used a subset of CLEF e-Health 2014 dataset that comprises 31,470 documents. The selected dataset is an evaluation collection for MIR with the following format:

- Document Collection: The document collection consists of automatically crawled web pages from various medical web sites, including pages certified by the Health On the Net[2] and other well-known medical web sites and databases [26]. The collection was provided by the Khresmoi[3] project and covers a broad set of medical topics. It contains around one million documents provided as semi-structured reports in raw HTML format and distributed over 8 .zip files; where each file contains multiple .dat files with different medical topics. Each .dat file contains multiple documents (where each document starts with '#UID: document _name' and ends with '#EOR', and the HTML content of the document lies between both tags).
- Queries: The queries in the dataset are 5 training queries and 50 test queries that have been created by experts involved in the CLEF e-Health consortium (registered nurses and clinical documentation researchers) and aims to model those used by laypeople (i.e., patients, their relatives or others). Queries are provided in a standard format consisting of a title, description, a narrative (expected content of the relevant documents), a profile (brief description of the patient), and a related discharge summary (sample of the information that should be included in relevant documents).
- Relevance Assessments: are collected from professional assessors (not medical experts) using Relevation[4] [3] which is a system for performing relevance judgments for the evaluation of Information Retrieval systems.

The relevance assessment is based on a four point scale. The relevance grades are:

- (0) where a document is irrelevant to a given query.
- (1) where a document is on topic of a given query but it is unreliable.
- (2) where a document is relevant to the given query.
- (3) where a document is highly relevant to the given query.

These relevance grades are mapped into a binary scale, with grades 0 and 1 corresponding to the binary grade 0 (irrelevant) and grades 2 and 3 corresponding to the binary grade 1 (relevant).

5.2 Indexing

In this section, we discuss the indexing process that we implemented to construct the inverted indexes for the medical documents in the dataset, as well as the experiments that we conducted using these documents.

Our prototype and experiments have been carried out on a PC with core i7 CPU (2.5GHz) and (8GB) RAM. For building the system's prototype, we used Java programming language with PrimeFaces framework for Java Server Faces (JSF) and we used Oracle 11g database to build our inverted indexes. We have downloaded a local copy of the exploited semantic resources (UMLS lexicon and UMLS

[2] http://www.hon.ch/
[3] http://khresmoi.eu/
[4] http://ielab.github.io/relevation/

metathesaurus) as the indexing processing was performed offline.

5.3 Document Processing

Documents in the collection are provided as raw web pages including all the HTML markup and also CSS style definitions, in addition to their accompanying Javascript codes. We used Jsoup[5] parser for cleaning and extracting text from the raw web pages to be able to start the indexing process. After extracting the text, we used natural language processing techniques to process the documents content. First, we converted all extracted texts into lower case. Then, we removed stop words based on predefined list that contains 566 stop words. Next, we utilized Porter stemmer to stem each term in the text. All stemmed terms including their frequencies ($Normalized - tf_{t,d}$) were added to our inverted index. If a term is related to a medical acronym or abbreviation (this is found using UMLS Lexicon), all full representations of the term are also added to the inverted index including their $Normalized - tf_{t,d}$; calculated as mentioned in previous section. To identify and add all compound terms (bi or tri grams) to the index, we have defined a sliding window of lengths between 2-3 tokens. Acronyms for each recognized compound term are also mapped to their UMLS correspondences and added to the inverted index. The values of document frequency D_f, inverse document frequency idf, and term frequency - inverse document frequency $tf - idf$ are all calculated and stored in our inverted index to reduce the complexity of the matching and ranking process between user queries and their relevant medical documents at run time. The indexing process of our system's prototype took around 1 month due to limited resources and huge document collection size (around 54 GB).

5.4 Runs

We have performed several runs to experimentally validate our proposal and evaluate the quality of the produced results. Accordingly, we have implemented the proposed techniques in the system's prototype to test and evaluate the quality of the query reformulation and query terms weighting. In addition, we aimed at testing the matching and retrieval steps, as well as the quality of the produced final search results. In this context, we evaluate the effectiveness of the proposed approach by comparing the results produced by our systems' prototype with:

1. Baseline run results that we achieve from our experiments using simple inverted index created without using any external semantic resources or term weighting techniques. The baseline is a measure of process functionality before any change occurs. In Information Retrieval it is a weighting model that counts as a run and allows comparison with the approach applied to verify if improvement was accomplished [30].
2. Three other proposed MIR systems in which the authors use the same dataset we use (CLEF e-Health2014) for testing and evaluation.

To evaluate our proposed approach step by step with the baseline run, we carried out five different runs using our prototype. The five runs are described as follows:

- RUN 1- baseline: The first run is the system baseline run. In this run we use only primitive inverted index and basic NLP techniques for both queries and documents processing (no query reformulation techniques, no external semantic resources, and no terms weighting).
- RUN 2: In this run, we reindexed the document collection by considering medical compound terms (either they are bi or tri grams), medical acronyms, and medical abbreviations using both UMLS lexicon and UMLS metathesaurus. During query processing we solve both acronyms and abbreviations (ACt, ABt) and neglect the compound terms. The main purpose of neglecting compound terms in this run, is that we need to measure the effect of each processing technique on the system's performance, as well as the effectiveness of the produced results. The full representations Et of both ACt and ABt are added to the reformulated query Rq. We would like to point out that the work and results of the following three runs (RUN3, RUN4, RUN5) are accumulated to the work and results of this RUN.
- RUN 3: Through run 3, we solve the medical compound terms that we ignored in processing users' queries in RUN2 (in addition to solving both acronyms and abbreviations).
- RUN 4: In this run, we use the UMLS Metathausurus via MetaMap tool to continue with classifying the users' queries into its components. We extract medical Mt terms

[5] https://jsoup.org/

and other supportive St terms, and weight query terms based on their classification as discussed in previous section.

- RUN 5: This is the last run in our experiments. In this run, we use the UMLS lexicon to extract the synonyms SYt of medical terms in the user's query. We add the extracted synonyms to the reformulated query by giving them higher wieghts against other semantically related concepts Et in the reformulated query as discussed in formula (2) in the previous section.

We discuss our findings for all runs in the next section. Precision@10 (P@10) evaluation metric was basically used to evaluate our proposed approach. We choose this metric since it is the most meaningful metric used in web-scale information retrieval systems, and there are multiple experiments done by different authors using the same dataset we used (CLEF e-heath 2014) calculating this measure which allow us to compare our findings with others. P@10 corresponds to the number of relevant results on the first 10 search results retrieved by the retrieval system. The next formula is used for calculating the P@10 (note: the maximum no. of results returned by our prototype is 10 documents - top 10 highest ranked documents)

$$P@10 = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \quad (4)$$

where the documents retrieved by our system compared with relevant assessments provided with CLEF e-health 2014 dataset to distinguish relevant and non-relevant documents.

5.5 Discussion

The results of the different runs are represented in Table 1. The P@10 values mentioned in this table represent the overall precision values for testing all training and testing queries provided in CLEF e-health 2014 dataset (5 training queries and 50 test queries).

TABLE 1: P@10 Matching Results

Runs	**Baseline**	**Run 2**	**Run 3**	**Run 4**	**Run 5**
P@10	0.721	0.734	0.751	0.776	0.794

The experimental results show that the methodologies proposed in our retrieval model improved the precision of our baseline run (traditional bag of words retrieval model) by around 0.0734. Our results indicate that exploiting integrated medical semantic resources for enriching laypeople queries in the medical domain increases the effectiveness of MIR systems. In Table 2 we compare the best results produced by our proposed approach (Run 5) with the results obtained by other researchers who used the same dataset for their systems evaluation. We compared our results with the best results obtained by the best 3 CLEF participating teams - as depicted in [26]- who used the same dataset that we used in our experiments. These teams are: GRIUM in their EN-Run.5, SNUMEDINFO in their EN-Run.2, and KISTI in their EN-Run.2.and their systems described in section 2 above [23-25].

TABLE 2: Comparison with other MIR systems

P@10	**Our system**	**GRIUM**	**SNUMEDINFO**	**KISTI**
Baseline Run	0.7211	0.7180	**0.7380**	0.7300
Best Run	**0.7945**	0.7560	0.7540	0.7400

As shown in Table 2, the results produced by our proposed system are slightly better than those produced by the three systems (GRIUM, SNUMEDINFO and KISTI). One of the major factors that resulted in this improvement is the exploitation of ngrams, medical acronyms and abbreviations (using the employed medical semantic resources) in the indexing process. While all of the other systems mentioned here use the basic existing indexing and retrieval systems (Indri and Lucene). The authors of GRIUM proposed a retrieval model using bag of concepts instead of traditional bag of words retrieval model. They used MetaMap tool for extracting medical concepts that exist in the user's query to be considered in their query-document matching process and ignored all other supportive terms in the original query. The main drawbacks of GRIUM system is that it ignores all query terms that are not identified as medical concepts by MetaMap, which leads to the possibility for ignoring some important medical concepts such as medical acronyms and abbreviations, and this explain the slight improvements in their system precision (0.7560) over the baseline run where the precision was (0.7180). In SNUMEDINFO, the authors used a simple inverted index (compound terms are not included) with UMLS metathesaurus for query expansion. The main drawbacks of SNUMEDINFO are: 1) Ignoring the compound terms that may occur in both the document collection or the user's query, and 2) ignoring the semantic type (synonym, mapped to, part of, .. etc) of the extracted concepts from UMLS metathesaurus and giving all extracted concepts the same importance as the original query terms. In addition, the authors didn't tackle problems associated with acronyms and abbreviations in the documents and queries. On the other hand, KISTI system achieved a little improvement in the quality of the produced results against the baseline. The main reason for this is that the authors used the related <discharge summary> information provided with each query in the dataset as their expansion resource instead of using external medical semantic resources. However, the discharge summary reports can't be considered as a trusted enriched medical semantic resource that can be used as effective expansion resource for medical queries. In our proposed system, we overcome the drawbacks of the existing MIR systems and we were able to achieve more precise results compared to similar works as depicted in Figure 2.

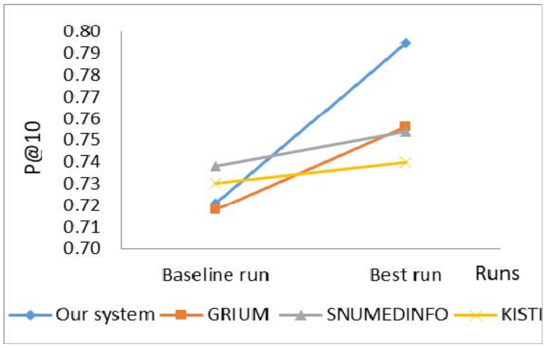

Figure 2: Comparison with other MIR System

It is important to point out that although our system was able to outperform the three systems, it still suffers from low performance issue. We plan to address this issue in the next version of the systems' prototype through incorporating the classification model wherein medical queries as well as their corresponding queries will be classified under their relevant medical topics. In this context, instead of matching each query with every document in the dataset, we aim to find matches between queries and medical documents that fall under the same medical topic/s.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a semantics-based medical information retrieval system that reformulates users' queries in an attempt to improve the quality of the returned search results and increase laypeople satisfaction about medical search engines. The key ideas of our proposed system are: 1) Building enhanced inverted index by considering NLP techniques and semantic relationships by using medical semantic resources. 2) Classifying terms in the original user query to medical and supportive terms and assigning higher weight to those medical terms against other supportive terms. 3) Enriching the user's query with additional medical terms (synonyms and the full representation of both medical acronyms and abbreviation) based on employing external medical semantic resources.

To evaluate the effectiveness of the proposed approach, we have developed a prototype of our proposed system through incorporating the underlying query processing and document indexing techniques. The various experimental runs, as well as the produced results by each run indicate that the employed techniques were able to produce more precise results than those produced by state-of-the-art medical retrieval systems, namely those that used the CLEF e-health2014 dataset. As we have pointed out in the previous section, the current version of our system's prototype suffers from performance issues (i.e. the run-time complexity issues) because we find attempt to find matches between each query-document pair in the dataset. However, we plan to eliminate this problem through the incorporation of a classification model that classifies each document/query under their corresponding medical topic/s. In this context, each medical query will be matched to medical documents that fall under the same medical topic/s.

ACKNOWLEDGMENT

We would to thank ELRA for providing the CLEFeHealth 2014 Task 3 Evaluation Package "CLEFeHealth 2014 Task 3 Evaluation Package, ELRA catalogue (http://catalog.elra.info), ISLRN: 725-020-897-275-7, ELRA ID: ELRA-E0043."

REFRENCES:

[1] I. Stanton, S. Ieong, and N. Mishra, "Circumlocution in diagnostic medical queries," in Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, 2014, pp. 133-142.

[2] L. Soldaini, A. Yates, E. Yom-Tov, O. Frieder, and N. Goharian, "Enhancing web search in the medical domain via query clarification," Information Retrieval Journal, vol. 19, pp. 149-173, 2016.

[3] R. W. White, S. T. Dumais, and J. Teevan, "Characterizing the influence of domain expertise on web search behavior," in Proceedings of the second ACM international conference on web search and data mining, 2009, pp. 132-141.

[4] J. Palotti, A. Hanbury, and H. Müller, "Exploiting health related features to infer user expertise in the medical domain," in Web Search Click Data workshop at WSCM, New York City, NY, USA, 2014.

[5] L. Goeuriot, G. J. Jones, L. Kelly, H. Müller, and J. Zobel, "Medical information retrieval: introduction to the special issue," Information Retrieval Journal, vol. 19, pp. 1-5, 2016.

[6] J. Palotti, A. Hanbury, H. Müller, and C. E. Kahn, "How users search and what they search for in the medical domain," Information Retrieval Journal, vol. 19, pp. 189-224, 2016.

[7] M. C. Díaz-Galiano, M. García-Cumbreras, M. T. Martín-Valdivia, A. Montejo-Ráez, and L. Urena-López, "Integrating mesh ontology to improve medical information retrieval," in Workshop of the Cross-Language Evaluation Forum for European Languages, 2007, pp. 601-606.

[8] B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley, "Information retrieval as semantic inference: A graph inference model applied to medical search," Information Retrieval Journal, vol. 19, pp. 6-37, 2016.

[9] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," Nucleic acids research, vol. 32, pp. D267-D270, 2004.

[10] C. E. Lipscomb, "Medical subject headings (MeSH)," Bulletin of the Medical Library Association, vol. 88, p. 265, 2000.

[11] D. Lee, R. Cornet, F. Lau, and N. De Keizer, "A survey of SNOMED CT implementations," Journal of biomedical informatics, vol. 46, pp. 87-96, 2013.

[12] D. J. Vreeman, C. J. McDonald, and S. M. Huff, "LOINC®: a universal catalogue of individual clinical observations and uniform representation of enumerated collections," International journal of functional informatics and personalised medicine, vol. 3, pp. 273-291, 2010.

[13] G. O. Consortium, "The gene ontology: enhancements for 2011," Nucleic acids research, vol. 40, pp. D559-D564, 2011.

[14] J. Hanna, E. Joseph, M. Brochhausen, and W. R. Hogan, "Building a drug ontology based on RxNorm and other sources," Journal of biomedical semantics, vol. 4, p. 44, 2013.

[15] W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, et al., "Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data," Nucleic acids research, vol. 43, pp. D1071-D1078, 2014.

[16] A. Coustasse and D. P. Paul III, "Adoption of the ICD-10 standard in the United States: The time is now," The health care manager, vol. 32, pp. 260-267, 2013.

[17] A. Malhotra, M. Gündel, A. M. Rajput, H.-T. Mevissen, A. Saiz, X. Pastor, et al., "Knowledge retrieval from PubMed abstracts and electronic medical records with the Multiple Sclerosis Ontology," PloS one, vol. 10, p. e0116718, 2015.

[18] M. Maree and M. Belkhatir, "Addressing semantic heterogeneity through multiple knowledge base assisted merging of domain-specific ontologies," Knowledge-Based Systems, vol. 73, pp. 199-211, 2015.

[19] H. Kondylakis, L. Koumakis, M. Psaraki, G. Troullinou, M. Chatzimina, E. Kazantzaki, et al., "Semantically-enabled Personal Medical Information Recommender," in International Semantic Web Conference (Posters & Demos), 2015.

[20] G. Luo, C. Tang, H. Yang, and X. Wei, "MedSearch: a specialized search engine for medical information retrieval," in Proceedings of the 17th ACM conference on Information and knowledge management, 2008, pp. 143-152.

[21] X. Zhang, M. Cole, and N. Belkin, "Predicting users' domain knowledge from search behaviors," in Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, 2011, pp. 1225-1226.

[22] G. Zuccon, B. Koopman, and J. Palotti, "Diagnose this if you can," in European Conference on Information Retrieval, 2015, pp. 562-567.

[23] W. Shen, J.-Y. Nie, X. Liu, and X. Liui, "An investigation of the effectiveness of concept-based approach in medical information retrieval GRIUM@ CLEF2014eHealthTask 3,"

Proceedings of the ShARe/CLEF eHealth Evaluation Lab, 2014.

[24] S. Choi and J. Choi, "Exploring Effective Information Retrieval Technique for the Medical Web Documents: SNUMedinfo at CLEFeHealth2014 Task 3," in CLEF (Working Notes), 2014, pp. 167-175.

[25] H.-S. Oh and Y. Jung, "A Multiple-stage Approach to Re-ranking Clinical Documents," in CLEF (Working Notes), 2014, pp. 210-219.

[26] L. Goeuriot, L. Kelly, W. Li, J. Palotti, P. Pecina, G. Zuccon, et al., "Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval," in Proceedings of CLEF 2014, 2014.

[27] P. Willett, "The Porter stemming algorithm: then and now," Program, vol. 40, pp. 219-223, 2006.

[28] A. R. Aronson and F.-M. Lang, "An overview of MetaMap: historical perspective and recent advances," Journal of the American Medical Informatics Association, vol. 17, pp. 229-236, 2010.

[29] R. R. Larson, "Introduction to information retrieval," Journal of the American Society for Information Science and Technology, vol. 61, pp. 852-853, 2010.

[30] I. Ruthven and D. Kelly, Interactive information seeking, behaviour and retrieval: Facet Publishing, 2011.