Summarizing Medical Texts for Effective Retrieval

by

Ganesh R Iyer
201311019

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
INFORMATION AND COMMUNICATION TECHNOLOGY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

May, 2015

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

ii) due acknowledgment has been made in the text to all the reference material used.

Ganesh R Iyer

Certificate

This is to certify that the thesis work entitled *Summarizing Medical Texts for Effective Retrieval* has been carried out by *Ganesh R Iyer* for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under my/our supervision.

PRASENJIT MAJUMDER
Thesis Supervisor

Acknowledgments

I would like to thank my supervisor *Prof. Prasenjit Majumder* for his support, encouragement and valuable guidance. The discussions I've had with my supervisor have played a very important role in shaping my research. I would also like to thank *Prof. Amit Sengupta* for providing medical insights. I am also grateful to my examiner *Prof. Bhaskar Chaudhury* for providing critical feedback without which the direction of my research would have gone astray. I am also grateful to my panel members: *Prof. Anil Roy*, *Prof. Gaurav Mishra* and *Prof. Nabin Kumar Sahu* for providing valuable feedback.

I would also like to thank *Mr. Harsh Thakkar* and *Ms. Ke sha Shah* for their contributions to our participation in the CLEF eHealth 2014 task 3. I am grateful to *Mr. Parth Mehta* for providing his inputs in text summarization. I would also like to thank the other members of *IR Lab DAIICT*, the discussions with them has often given me insights into different aspects of research.

Contents

Abstract

User centered health information retrieval is a challenging and important problem in information retrieval. In this work, we apply medical resources to bridge the vocabulary mismatch between lay-users and medical documents. We also applied text summarization techniques to reduce the document to relevant information while pruning irrelevant information. We provide a survey of medical resources and application of text summarization in information retrieval. The primary research goals were to investigate the use of medical resources in query expansion and text summarization in indexing. The experiments were performed as a part of a CLEF eHealth Task, overview of which is provided. From our experiments we observed that a summarized index can be used to replace a full collection index. Also a compression rate of 40-80% outperformed the baseline indicating that retrieval on the summarized collection can indeed improve performance. Using MeSH(Medical Subject Headings) as a thesaurus to supplement the query terms improved retrieval for certain queries. We obtained the best MAP score of 0.415, for all teams, using query expansion with discharge summaries.

List of Tables

List of Figures

CHAPTER 1
Introduction

There has been an in increase awareness amongst laypeople regarding health issues. Searching for health related information has also become an important practice for millions of users[1]. This has led to a dire need for expanding research on Information Retrieval(IR) in the biomedical domain. In this report we provide an overview of the field of Health Information Retrieval. We also provide the methodology that we used while participating in CLEF eHealth 2014[2]. We made use of Medical resources like MeSH and Discharge summaries for query expansion and used text summarization techniques for indexing. In this chapter we provide a background of the problem statement.

1.1 Motivation

Laypeople have different search behavior, which is exemplified in their formulation of queries and expectations from the retrieved documents. Searching for health advice is a common task performed by users. In a recent survey it was observed that more than 70% of search engine users in the US have used web search engines for information about specific diseases or health disorders[1]. Not just laypeople, doctors, clinicians and other medical personnel use such systems for their biomedical information needs too[3].
Health information retrieval is not only important for laypeople, or medical professionals like doctors, clinicians etc. It is an important part of the *knowledge discovery* process. IR techniques are required to reduce the vast body of literature to a narrow, focused set, that can in turn be used for knowledge discovery.

There has been a fast growing interest in the field of bio-medical research. MEDLINE, the primary database for biomedical literature, contained more than 17 million article citations in 2009[4](Fig. 1.1). Retrieving information from such a large pool of documents is challenging and of significant importance.

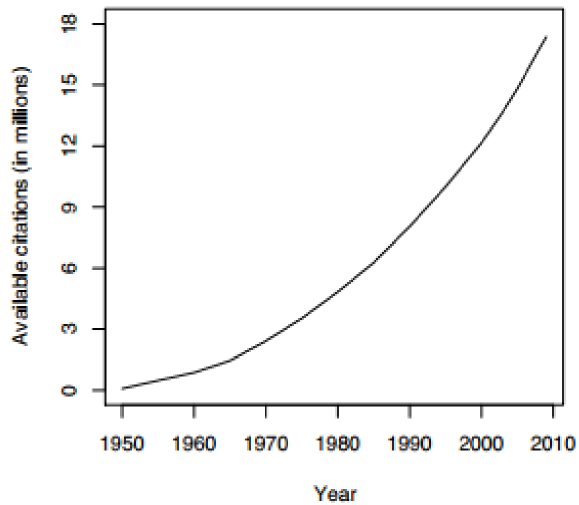

Figure 1.1: Number of available citations in MEDLINE

1.1.1 Use of Medical Resources

There are several well defined medical information structures avaiable as described in section 2.1. Our hypothesis was that such resources could be used to bridge the vocabulary gap between the layusers and the experts. The motivation between using discharge summaries was to supplement the query with contextual information about the patient's medical conditions.

1.1.2 Use of Text Summarization Techniques

The use of Summarization techniques was motivated by the need to retrieve documents whose central theme were in agreement with the users. By summarizing we can reduce the document to its crux while removing irrelevant information. Such techniques have proved effective in the past as described in section 2.3 but they haven't been explored in the medical domain. Also with the growing size of the corpus using a summarized collection can reduce the space requirements.

1.2 Health Information Retrieval

Health information retrieval resembles ad-hoc information retrieval but with a focus health related information. Information retrieval, in our context, is the retrieval of information from databases that contain textual information. The information is of unstructured nature. Our main challenge, in health IR is, given a health information need, formulated in terms of a *query*, retrieve *documents* that satisfy the information need.

Health Information retrieval is a domain-specific task that is performed by a large variety of users, including clinicians, healthcare workers and laypeople. Due to the broad variety of information seekers novel methods need to be devised to address the issues pertaining to different health information needs.

1.2.1 User-centered Health Information Retrieval

The Goal of *User-centered* Health Information Retrieval is to consider the information needs of laypeople(*users*), who do not have much knowledge about medical terminologies and concepts and seek health information on the web. The major challenge in user-centered health IR is that of vocabulary mismatch, wherein similar concepts are referred to by different terms by different users. This arises due to lack of knowledge of medical concepts in lay users.

To make eHealth easily accessible to lay users it is important to ensure that they have easy access to information relevant to their needs. To facilitate this searching information about the patient's background relevant to his/her information need should be used while retrieving.

1.2.2 CLEF eHealth 2014

The Cross Lingual Evaluation Forum(CLEF) community addressed the problem of Health IR by initiating the eHealth Tasks from 2013. The goal of this task was to form a framework for the evaluation of health information systems that support laypeople. These systems help laypeople in searching for and understanding their health information[3].

The CLEF eHealth lab's goal is to help patients understand eHealth information. eHealth documents are much easier to comprehend after processing them with: shorthand expansion, spelling correction and normalization of health conditions to standard terminology. Retrieving relevant documents pertaining to user queries and their medical condition is crucial to help their understanding.

The 2014 edition of the CLEF eHealth labs comprised of three tasks[1]. The use case considered was of Discharge summaries, which are reports that a patient receives after leaving a hospital. It describes the diagnosis and treatment received by the patient [5]. Following is a description of the 3 tasks:

[1] http://clefehealth2014.dcu.ie/

1. **First Task**: *Visual-Interactive Search and Exploration of eHealth Data*. This task aimed at visualization of medical information extracted from the discharge summaries to make it understandable to lay people.

2. **Second Task**: *Information extraction from clinical text*. This task was a NER task, in which the goal was to extract medically relevant entities from the text.

3. **Third Task**: *User centered health information retrieval*. In user centered health IR the goal is to address user queries by retrieving documents relevant to their health-related information needs.

We participated in Task 3 of 2014's lab. We provide a detailed overview of our approach in this report.

1.3 Problem Statement

Laypeople have varied information needs and searching behaviors. This can be seen in their formulation of queries and their expectations from the retrieved documents. Our challenge was to support these needs. Our aim is to develop a system for Information retrieval from the patients' perspective.

We were given:

1. A crawl of approximately 1.5 million biomedical documents made available by the Khresmoi project[2] in plain HTML format.

2. A set of English general public queries that lay users may pose based on the content of their discharge summaries.

Our goal was to retrieve the relevant documents for the user queries.

1.4 Thesis Organization

The subsequent portion of the thesis is divided into 5 chapters. In Chapter 2, we describe related work done in this field, focusing on medical resources for query expansion and text summarization in IR tasks. Chapter 3 provides the details of the dataset used. In Chapter 4 we describe our system, providing explanation about the various modules. Chapter 5 provides the details of the experiments performed. We provide an analysis and discussion of the results. In the last chapter we provide the conclusions drawn from our experiments.

[2]http://www.khresmoi.eu/

CHAPTER 2
Literature Survey

In this chapter we provide an overview of Health information retrieval. We first describe some health concept representation methods in Section 2.1. In Section 2.2 describe how these can be used for the purpose of query expansion. Later in Section 2.3 describe usage of Text summarization techniques in the context of Information retrieval, stressing on using summarization for indexing. We provide an overview of the methodologies used by some of the best teams in CLEF eHealth 2013 and 2014 in Section 2.4

2.1 Health concept representation

Well defined information structures are widely used in the field of IR. They can be used in indexing and retrieval processes. Knowledge organization in health information is highly detailed and is one of the earliest application of classification. Several resources are available which provide a well defined structure to cope with ambiguity and synonymy present in biomedical terminology. They vary both in coverage and purpose.

The concept representation systems described here can be applied to Information retrieval and natural language processing. In IR they may be for query expansion using the synonymous relations to improve the expression of information need.

2.1.1 Medical Subject Headings(MeSH)

The Medical Subject Headings(MeSH)[1] is the NLM's[2] thesaurus used to index most of NLM's databases. It has sets of terms naming descriptors that are arranged in a hierarchical manner.
Following is an example:

[1]MeSH - Online at www.ncbi.nlm.nih.gov/mesh
[2]NLM - United States National Library of Medicine

1. `Cardiomegaly [Descriptor]`
   (a) `Cardiomegaly [Concept, Preferred]`
      i. `Cardiomegaly [Term, Preferred]`
      ii. `Enlarged Heart [Term]`
      iii. `Heart Enlargement [Term]`
   (b) `Cardical Hypertrophy [Concept, Narrower]`
      i. `Cardiac Hypertrophy [Term, Preferred]`
      ii. `Heart Hypertrophy [Term]`

The structure of MeSH consists of the following components: *1. Descriptor Hierarchy 2. Descriptions 3. Qualifiers 4. Supplements* . There were a total of 25,186 subject headings in the 2009 version of MeSH.

2.1.2 Unified Medical Language System

The Unified Medical Language System(UMLS)[3] was started at the NLM in 1986. The aim of the project was towards effective retrieval of information. The UMLS primarily contains 3 main knowledge-sources that can be used together or separately:

1. **Metathesaurus**: Includes over one million medical concepts from more than 100 sources. It is called a Metathesaurus as it transcends the various thesauri it covers.

2. **Semantic Network**: It is a high level ontology for medical entities. It supports hierarchical and associative relationships.

3. **SPECIALIST Lexicon**: It is a set of *lexicon* and *tools*. The *lexicon* includes biomedical terms and the *tools* are programs that process these terms.

2.1.3 Other Ontologies

Ontologies are defined as *"hierarchical structuring of knowledge about concepts by subclassing them according to their properties and qualities"*[6]. The aim of developing ontologies in biomedical community is to represent the terminology in a common vocabulary. This vocabulary can then be shared and reused[7].

[3] www.nlm.nih.gov/research/umls/

GALEN GALEN(Generalised Architecture for Languages, Encyclopedia and Nomenclature in Medicine) is an European project which aims to represent medical concepts as a formal ontology[8]. To make the results of GALEN projects available freely to the world, OpenGALEN[4] was set up. The terminology for OpenGALEN is written in GRAIL(GALEN Representation And Integration Language), which is a domain specific formal language. It is also distributed in OWL.

SNOMED-CT SNOMED-CT(Systemized Nomenclature Of Medicine Clinical Terms) is developed in native description logic formalism. It contains a collection of medical terms with codes, terms, synonyms and definitions. It is considered to be the most comprehensive biomedical terminology[7].

WordNet is a lexical database that is widely used for application in Information Retrieval(IR) and Natural language processing(NLP). At core of its strcuture, is a set of synonyms(Synset) that represents an underlying concept. It provides short definitions and the number of relations between the synsets[9].

2.2 Using Medical Resources for Query Expansion

Medical resources have been used widely in Health information retrival to provide contextual information and to bridge the vocabulary mismatch.[10]. Use of MeSH in Pubmed was evaluated by Lu *et al*. They investigated the use of MeSH in query expansion. They used a process called: Automatic term mapping(ATM) for expanding the query and observed that query expansion using MeSH generally improved performance[11]. Use of UMLS and Wordnet has also been shown to improve precision in web based search for medical documents[12]. Synonym, hierarchical and term defination information was used by Hersh *et al.* to investigate usage of UMLS Metathesaurus in health information retrieval. They observed that overall performance was decreased, but for certain queries the performance was improved[13].

2.3 Text Summarization in IR Tasks

Text summarization is often used in tasks like question answering, text categorization etc[14]. Summarization is used to reduce the size of the text collection and also to obtain important information from the collection. In this section we restrict ourselves to the use of text-summarization in IR tasks.

[4]http://www.opengalen.org/

2.3.1 Summarization for Postprocessing

Much work has been done in summarization of retrieved documents. In these systems the output of the retrieval system is fed as input to the summarizer. Text summarization is applied to obtain important features from the retrieved documents. Often the retrieved documents are clustered according to topics, and finally summarization is done for each cluster. Kan *et al.* used automatic summarization techniques to post-process documents and provide the user with an overview of the results[15]. Tombros *et al.* investigated the use of *query biased* summaries to improve user relevance judgments[16]. Summarization is often used to improve the presentation of the retrieval results to the users[17].

2.3.2 Preprocessing and Indexing

Using summarization to aid in the retrieval process has not been explored much. Application of summarisation to pre-process the documents before indexing can be used was first suggested by Luhn in 1958[18].

Brandow *et al.* evaluated the effect of summarization on retrieval. They used it to enhance the effectiveness of Boolean based retrieval. Text summarization has been used for indexing, it was observed that summary index is as effective as the corresponding full text summary by Sakai *et al.*[19].They also used the summary index for pseudo-relevance feedback and observed an improvement of 5-10%. Their hypothesis was that an effective generic summary would retain the relevant contents of the original document while pruning irrelevant content. Text summarization has been applied to the task of geographical information retrieval[20]. The authors used generic and geographic summaries with varied compression rates to evaluate the effectiveness. They observed that geography-specific summaries had a better performance than the generic summaries. They also observed that compression rates between 60% to 80% obtained better results than the baseline.

Indexing has been done on summarized corpus to reduce the index size. It was shown by Lu *et al.* that pruning contents reduced storage costs by 54-93% while causing minor losses in accuracy [21]. They were motivated by the costs of storage of long documents and focused on reducing this cost. Summarization techniques have been applied to the tasks of Music Information Retrieval(MIR) as well where the authors observed that summarized datasets lead to a classification performance similar to that from using full songs[22].

2.3.3 In Psuedo Relevance feedback

Sakai *et al.* evaluated the effectiveness of using summarized phrases from ranked [19]. They observed that a compression ration of 5-10% worked best for precision-oriented search. Lam *et al.* hypothesized that query expansion from complete contents leads to addition of unrelated terms. They investigated the use of context-independent and query-biased summaries and observed that query-expansion using the summarized documents was more effective than full-document expansion[23].

2.4 Overview of best systems in CLEF eHealth 2013 and 2014

In this section we review the systems that were evaluated in the User centered Health Information Retrieval task of CLEF eHealth 2013 and 2014. We provide a brief overview of their methodology and their performance.

2.4.1 CLEF eHealth 2013

In the 2013 edition of the task there were 13 registered groups out of which 9 submitted runs. There were a total of 48 runs, with 9 baseline runs and 15 runs were using discharge summaries(DS). P@10 and NDCG@10 were the primary evaluation metrics for the task.

Team Mayo was the best performing team in the 2013 edition. They used query likelihood model for their retrieval. They transformed both queries and documents to UMLS concept space and performed reranking on concept space. For their best run they used Discharge summary with concept space reranking and obtained a P@10 of 0.52[24].

On observing results of other teams we observed that the baseline provided by the organizers: P@10 of 4860, was a strong baseline. Only 1 team(*Team Mayo*) obtained results above the baseline. The baseline system used BM25 for retrieval.

2.4.2 CLEF eHealth 2014

There were 91 registered groups for the 2014 edition of eHealth task. 14 groups managed to submit a total of 62 runs. The primary evaluation metrics were P@10 and NDCG@10. 4 teams used Discharge summaries to aid the retrieval.

Team SNUMEDINFO used language modeling for retrieval. For query expansion they used an intersection of preferred terms and terms from Discharge summaries. They used Metamap to obtain UMLS concepts. They obtained a P@10 of 0.8160 for their best run.

Team GRIUM used UMLS Metathesaurus to experiment with concept based retrieval. Their baseline was based on language modeling using Indri. Their best run obtained a P@10 of 0.75.

Two of the top ten runs used discharge summaries. For teams using discharge summaries for retrieval, the runs using discharge summaries proved to be more effective in retrieval.

CHAPTER 3

Data

In this chapter we provide information about the dataset we used for conducting our study. The dataset was provided by CLEF eHealth labs to participants. The dataset had 3 parts namely: *Corpus, Queries and Discharge summaries*. We describe the information available in the dataset, the format and examples of the dataset.

3.1 Corpus

The dataset was a crawl of 1.6 million documents covering various medical topics. The documents were a part of EU-FP7 Khresmoi[1] project. The collection was prepared from a variety of online sources like Health on the net Foundation websites[2], well-known medical sites and databses like: Genetics home reference, ClinicalTrial.gov, Diagnosia[3] etc.

The size of the corpus was *6.3 GB* compressed and **43.6 GB** uncompressed. The corpus was preprocessed by eliminating small documents and correcting errors in the markup using Jsoup. The document collection was split into group of .dat files, each containing a collection of web crawl from one domain.

The format of the data in the .dat files is described below: Each .dat file contains a collection of web pages and metadata (data about data i.e. keywords), where the data for one web page is organized as follows:

1. a unique identifier (#UID) for a web page in this document collection,
2. the date of crawl in the form YYYYMM (#DATE),
3. the URL (#URL) to the original web page, and
4. the raw HTML content (#CONTENT) of the web page

[1] http://khresmoi.eu/
[2] http://www.healthonnet.org
[3] http://www.diagnosia.com/

3.2 Queries

The query-set contained 50 test queries and 5 training queries for each year. Since we had the dataset for CLEF eHealth 2013 as well we had relevant judgements for the 50 queries used. We used them as our training set. The query set aims to model those used by laypeople (i.e. patients etc.) to know more about their medical condition. The medical conditions were identified within discharge summaries and linked to the matching UMLS (Unified Medical Language System.

A sample query from the official 5 training queries is shown below for reference:

```
<topic>
<id>QTRAIN2014.1</id>
<dischargeSummary>00211-DISCHARGESUMMARY.txt</dischargeSummary>
<title>Coronary artery disease.</title>
<desc>What does coronary artery disease mean?
</desc>
<profile>The documents should contain basic information about
coronary artery disease and its care.
</profile>
<narr>This positive 83 year old woman has had problems with her
heart with increased shortness of breath for a while.  She has now
received a diagnosis for these problems having visited a doctor.
She and her daughter are seeking information from the internet
related to the condition she has been diagnosed with.  They have
no knowledge about the disease.
</narr>
</topic>
```

The following is a description of the fields:

1. *Title*: Contains the actual text of the query.
2. *Description*: Detailed description of the query.
3. *Narrative*: Users expectation from the retrieved documents
4. *Profile*: Information about the patient e.g. Age, Sex etc.

5. *Discharge summary*: Filename of the associated discharge summary.

Along with English queries we were also provided with queryset of Czech, German and French queries. This was used to evaluate multilingual IR.

3.3 Discharge Summaries

We were also provided with discharge summaries describing the medical conditions of the users along with their queries. They were provided in order to obtain some contextual information about the user queries. They were obtained from the de-identified MIMIC-II database[4] (Multiparameter Intelligent Monitoring in Intensive Care, Version 2.5).

3.4 Relevance assessment

The relevant assessment related information is as follows:

- The official training set consisted of 5 queries and corresponding result set of documents generated from manual relevance assessment by experts on a shallow pool. This training set was provided by CLEF.
- Pooled sets of documents was generated using Vector space Model(VSM) and Okapi BM25 for the 5 training queries. Relevance assessments were formed based on these pooled sets.
- Pooled sets were created by taking the top 30 ranked documents returned by the two retrieval models with duplicates removed.
- Relevance is provided on a 2-point-scale: Non relevant (0); Relevant (1), and on a 4-point scale: Non relevant (0); On topic, but unreliable (1); Somewhat relevant (2); Highly relevant (3).

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

CHAPTER 4
System Overview

In this chapter, we provide an overview of our system. The Fig. 4.1 shows the block diagram of our retrieval system. The entire process can be divided into sub-parts namely:

- Corpus Summarization
- Indexing
- Query Expansion
- Retrieval
- Pseudo Relevance Feedback

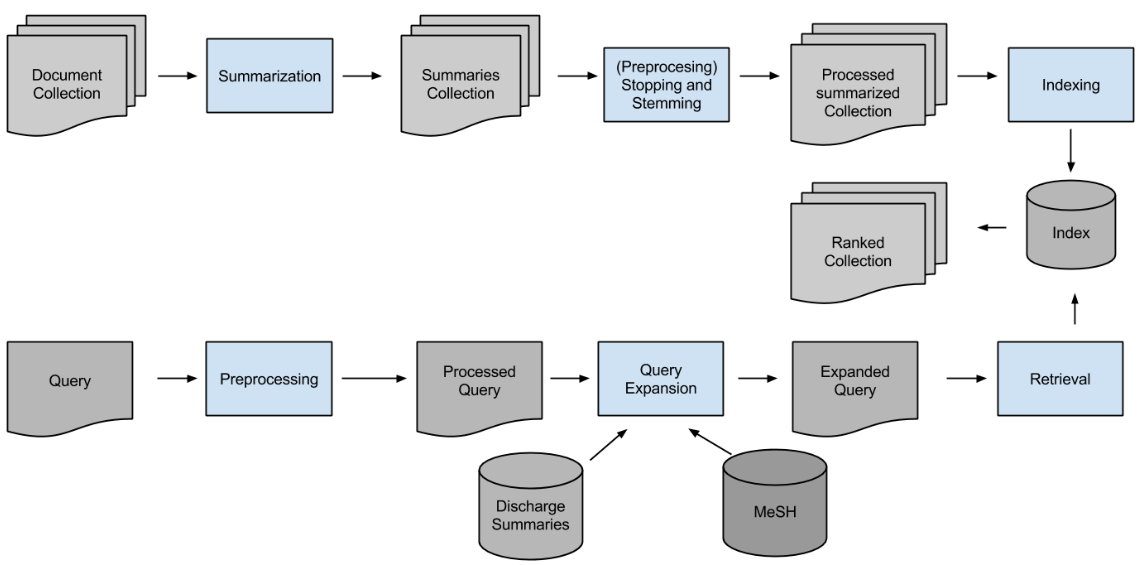

Figure 4.1: System Overview

We used text summarization and query expansion to supplement ad-hoc retrieval techniques.

4.1 Summarization of Document collection

We performed experiments with Textrank[25] and Lexrank[26] algorithms. Both of these are graph based summarization techniques and work by deciding the importance of a vertex within a graph using global information recursively drawn from the graph. Our hypothesis was that using Summarization we would be able to prune irrelevant information from the documents, while retaining salient information, thus improving the effectiveness of the retrieval models.

We describe the summarization algorithms used in the subsections.

4.1.1 TextRank

In TextRank the text is first converted into a weighted graph. It does so by identifying text units that can act as vertices(eg word collocations, sentences etc.) and then identifying the relations that connect the vertices. The edges are drawn based on the relation and it can be weighted or unweighted.

The similarity measure for 2 given senteces S_i and S_j is given as follows:

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i, w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)} \tag{4.1}$$

where each sentence is represented by a set of N_i words that appear in sentence $S_i = w^i_1, w^i_2, ..., w^i_{N_i}$.

For sentence scoring it uses PageRank[27]. If $In(V_i)$ is the set of *predecessors* of vertex and $Out(V_i)$ is the set of *successors* then the score of a vertex V_i is defined as:

$$S(V_i) = (1 - d) + d * \sum_{jinIn(V_i)} \frac{1}{|Out(V_j)} S(V_j) \tag{4.2}$$

4.1.2 LexRank

Lexrank is a graph-based technique for computing the relative salience of textual units. It uses the concept of eigenvector centrality after modeling text as a graph. It is a multi-document extractive summarization technique, that works by identifying the most *central* textual units(sentences etc.) from a multi-document cluster. The hypothesis is that the sentences that are similar to several other sentences in a cluster are more central to the overall topic. The cosine between the 2 corresponding vectors is used to compute the similarity between the sentences.

$$S(x,y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sqrt{\sum_{x_i \in x} (tf_{x_i,x} idf_{x_i})^2 \sum_{y_i \in y} (tf_{y_i,y} idf_{y_i})^2}} \tag{4.3}$$

where $tf_{w,s}$ is the term frequency of word w in sentence s.

For computing centrality LexRank uses Pagerank[27]. The idea of centrality is used to rank documents as done in Textrank. The centrality $p(u)$ of node u can be computed using:

$$p(u) = \frac{d}{N} + (1-d) \sum_{v \in adj(u)} \frac{p(v)}{deg(v)} \tag{4.4}$$

where $deg(v)$ is the degree of node v. The total number of nodes in the graph is represented by N and d is the damping factor.

4.2 Indexing

Once the document collection is summarized we perform indexing on it. Indexing is done on the cleaned and formatted document set using Indri with parameters including tokenizing, stop-word removal and stemming. For stop-word removal, we have used the PuBMed list of stop-words[1]. Standford Porter stemmer is used for stemming during indexing.

4.3 Query Expansion using Medical Resources

In the second phase, for query expansion stop-word removal from the same PuBMed list which is used during indexing is used. Spell-checking is also performed on the query terms. Two dictionaries are used for spell checking and correcting, one is the English US used in enchant and the second dictionary is specifically for medical domain.[2]

4.3.1 Using MeSH as a Thesaurus

We used Metamap to extract medical concepts from the query. Using MeSH synonyms of the queries were extracted. These synonyms were added to the query to obtain the expanded query. For example the query:

[1]Free Medical Dictionary at http://www.ncbi.nlm.nih.gov/books/NBK3827/table /pubmedhelp.T43/

[2]http://www.essex1.com/people/cates/free.htm

```
What is copd?
```

Was expanded(after stop-word removal and stemming) to:

```
copd chronic obstructive pulmonary disease
```

We used Metamap to perfrom a unigram named entity recognition(NER) on the query, to obtain medical concepts from the query. We then obtain the MeSH synonyms of these identified entities and add them to the query. In the above example `copd` was the identified medical entity. We added `chronic obstructive pulmonary disease` was the synonym.

4.3.2 Contextual Information from Discharge Summaries

We used medical history from the discharge summaries of users posing the query. This enabled us to supplement the user query with contextual information. We used Metamap to extract medical terms from the discharge summaries and expanded the query with them.

Our hypothesis was that there could be hidden medical concepts that could be extracted from the medical history that were not present in the query. This would prove to be a valuable context to obtain the patient's information need.

4.4 Documents Retrieval

We experimented with various retrieval models. We provide a brief overview of the models that performed well for us.

4.4.1 Okapi Model

We make use of the famous Okapi Model[28]. The weighting based documents score is calculated as in below formulae:

$$RSV_d = \sum_{t \epsilon q} \left[log \frac{N}{df_t} \right] . \frac{(k_1 + 1) tf_d}{k_1((1-b) + b(L_d/L_{ave})) + tf_d} . \frac{(k_3 + 1) tf_{tq}}{k_3 + tf_{tq}} \tag{4.5}$$

$$score(D, Q) = \sum_{i=1}^{n} IDF(q_i) \frac{f(q_i, D).(k_1 + 1)}{f(q_i, D) + k_1.(1 - b + b.\frac{|D|}{avgdl})} \tag{4.6}$$

$$IDF(q_i) = log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{4.7}$$

Here, t_f is term's frequency in the document.
qt_f is term's frequency in the query,
N is the total number of documents in the collection,
df is document frequency that contains the term,
dl is the document length (in bytes) and
$avdl$ is the average document length.

This formula has three components: The first is the id_f part which reflects the discriminative power of each word. Second part is tf component which is the number of documents in which that particular term is encountered. The value of tf generally increases, but reaches an asymptotic limit. Which implies that whether a term appears a 100 times or a 1000 times the function will weight it almost the same. Also, there is a correction for document weight. If a document is short, the tf for all its words is increased; if a document is long the tf for all its words is decreased. The count of each word is measured with respect to the document of average length in the collection. The third part is qtf component. If a word in the query appears more times than another it should be weighted higher.

4.4.2 Query Likelihood model

We construct from each document in the collection a language model M_d. Our goal is to rank documents d by $P(d|q)$, where the probability of a document is interpreted as the likelihood[29] that it is relevant to the query. Using Bayes rule, we have $P(d|q) = P(q|d) * P(d)/P(q)$

Since the probability of the query $P(q)$ is the same for all documents, this can be ignored. Further, it is typical to assume that the probability of documents is uniform. Thus, $P(d)$ is also ignored. Hence, $P(d|q) = P(q|d)$.

Documents are then ranked by the probability that a query is observed as a random sample from the document model. The multinomial unigram language model is commonly used to achieve this. We have:

$$P(q|M_d) = K_q \prod_{t \epsilon V} P(t|M_d)^{tf_{t,q}} \tag{4.8}$$

where the multinomial coefficient is :

$$K_q = L_q!/(tf_{t1,q}!tf_{t2,q}!...tf_{tM,q}!) \tag{4.9}$$

In practice the multinomial co-efficient is usually removed from the calculation. The calculation is repeated for all documents to create a ranking of all documents in the document collection.

CHAPTER 5

Experiments and Results

In this section we describe our experiments and explain our results. We compare our results with the best systems of CLEF eHealth 2014. Our system was the best system in terms of MAP performance in CLEF eHealth 2014. We investigate the use of MeSH, Discharge summaries and Summarization and perform failure analysis.

5.1 Runs submitted to CLEF eHealth 2014

We submitted six of the seven runs in the task that are described in table 5.1. Following is a brief description of each of the runs:

Run	QL	Okapi	MeSH	DS	Summary
1	✓				Baseline run
2	✓			✓	optional run WITH DS 1
3		✓		✓	optional run WITH DS 2
4	-	-	-	-	**NOT SUBMITTED** 3
5		✓			optional run WITHOUT DS 1
6	✓		✓		optional run WITHOUT DS 2
7		✓	✓		optional run WITHOUT DS 3

Table 5.1: Overview of runs submitted

1. **RUN 1:** The first run was is the system baseline run. In this run we use only the primitive blind relevance feedback mechanism for query feedback and query-likelihood model for query expansion. Use of external libraries and resources is exempted from in this run.

2. **RUN 2:** Run 2 and 3 are the runs WITH Discharge summaries. In run 2 we use the combination of query-likelihood model and discharge summaries, text extracted from $< desc >$ tags, for query expansion and document retrieval processes respectively. The medical words are extracted from dis-

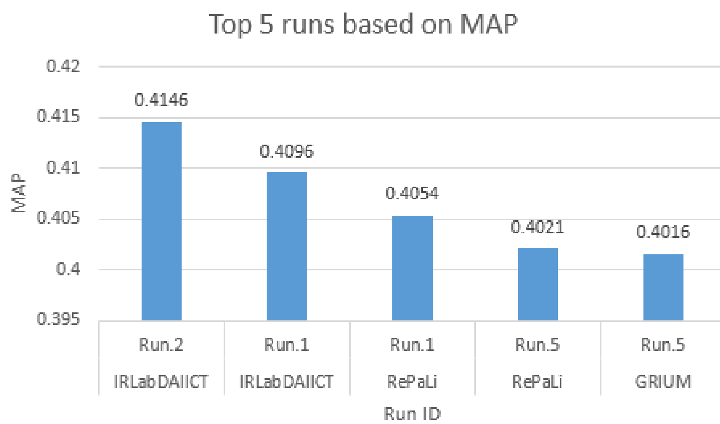

Figure 5.1: Comparison of MAP scores for top 5 runs

charge summaries and incorporated with the wordset obtained by query-likelihood model using a special weight function. The words obtained from the user query are given prominent weightage (0.7) while the words extracted from discharge summaries are given lesser weightage (0.3).

3. **RUN 3:** It is a variant of run 2, in which okapi model is introduced for retrieval process along-with discharge summaries for query expansion .

4. **RUN 4:** We did not submit run 4.

5. **RUN 5:** Run 5 through 7 are the runs WITHOUT Discharge summaries. Run 5 makes use of the Okapi model along with blind relevance feedback for retrieving documents.

6. **RUN 6:** We used the Query Likelihood model for retrieving documents. We also used MeSH for query expansion. Medical concepts are identified using MetaMap and their synonyms were used for query expansion following the same weighting strategy as with discharge summaries.

7. **RUN 7:** We used the Okapi model for retrieval along with MeSH for query expansion.

5.1.1 Comparison with Other Systems

Our system outperformed all other systems in terms of MAP score. Our Run2 and Run1 has MAP scores of 0.4146 and 0.4096 respectively. Figure 5.1 shows the MAP values for the top 5 teams that participated in CLEF eHealth 2014.

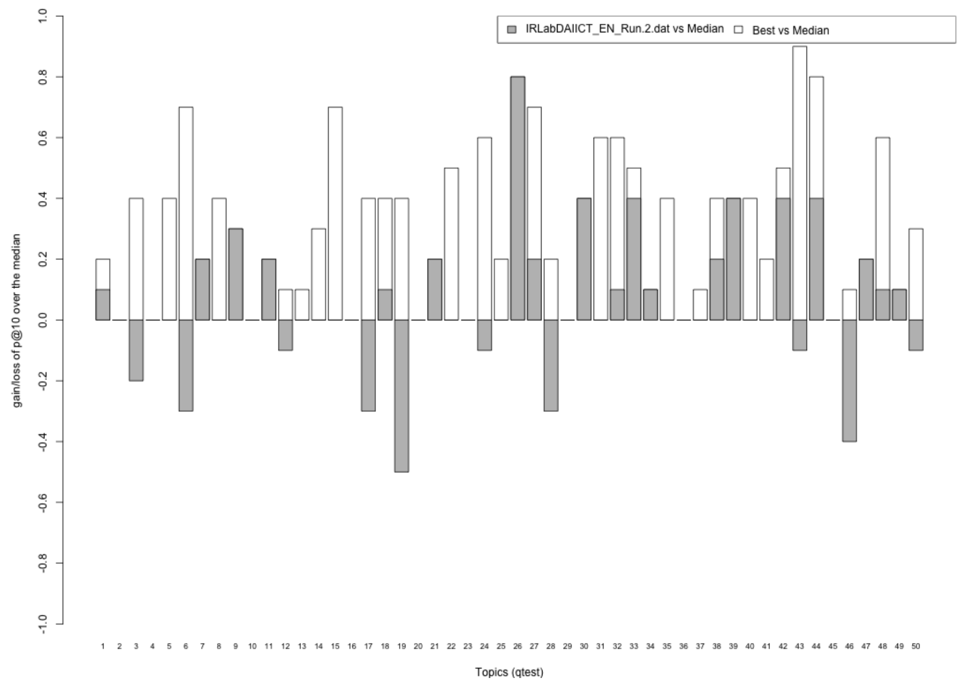

Figure 5.2: Querywise comparison of Run 2 with the best and median runs across all systems. The x-axis in the plot correspond to the query numbers. The y axis corresponds to gains and losses of our Run 2(and the best run) against the media

In Figure 5.2 we plotted the querywise comparison of our Run 2 against both the median and the best result. For 11 queries our system was better than the median result. For 8 queries our system was the best performing system. For 11 queries the system was worse than the median.

ID	P@5	P@10	NDCG@5	NDCG@10	MAP
Run.1	0.712	**0.706**	0.693	0.687	0.410
Run.2	0.704	0.702	0.686	**0.689**	**0.415**
Run.3	0.548	0.564	0.558	0.566	0.250
Run.5	0.668	0.654	0.652	0.636	0.303
Run.6	**0.732**	0.688	**0.717**	0.688	0.369
Run.7	0.316	0.294	0.311	0.294	0.174
Best	0.816	0.756	0.774	0.744	0.415
Median	0.672	0.631	0.652	0.634	0.318

Table 5.2: Comparison of performance metrics for our submitted runs vs the best and median off all submitted runs(by all teams)

5.1.2 Performance of various models

We experimented with 3 models, namely: tf-idf, Okapi and Query likelihood. We observed that the Query likelihood model out performed all other models. The

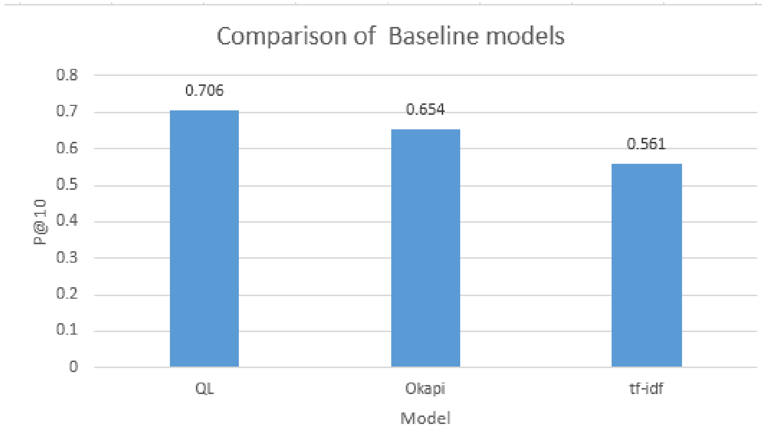

Figure 5.3: Comparison of P@10 for Query likelihood, Okapi and tf-idf

Okapi model performed better than the tf-idf model and hence we used it in Runs 3, 5 and 7. The results are shown in Figure 5.3

5.2 Summarizing Corpus Before Indexing

In our experiments we applied both Textrank and Lexrank to the document collection and observed the results. We tried different compression ratios and observed the effect of varying the compression ratios on the retrieval process. The compression ratio is defined as the ratio of the size of the summary(S) to the size of the original document(D):

$$Compressionratio = \frac{lengthS}{lengthD} \tag{5.1}$$

5.2.1 Effect of compression ratio on retrieval performance

We experimented with compression ratio of 20%, 40%, 60% and 80% and our goal was to find the optimum compression ratio that can be indexed to improve the retrieval performance. The effect of compression ratios is depicted in Fig.5.2.1

Table 5.3: Effect of LexRank with different compression ratios on retrieval performance.

Compression	**P@5**	**P@10**	**MAP**	**NDCG@5**	**NDCG@10**
20%	0.444	0.434	0.237	0.443	0.446
40%	**0.496**	0.454	0.270	**0.500**	0.475
60%	0.452	0.450	0.277	0.464	0.465
80%	0.492	**0.461**	0.295	0.492	**0.488**
Baseline	0.472	0.458	**0.321**	0.486	0.476

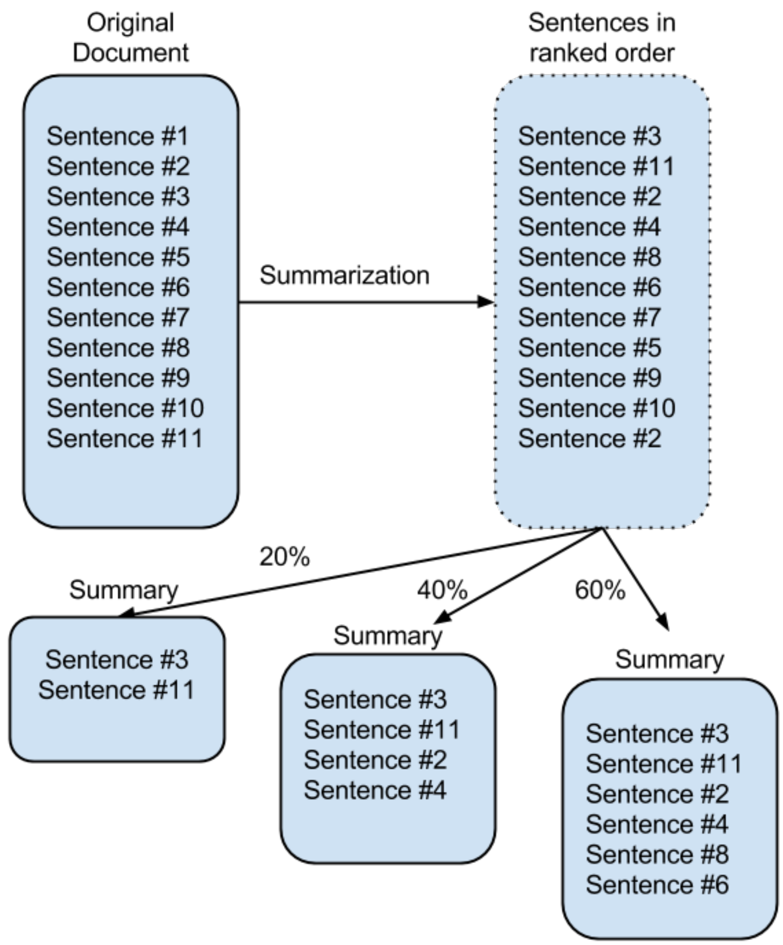

Figure 5.4: Effect of compression ratio on summaries

Table 5.4: Effect of TextRank with different compression ratios on retrieval performance.

Compression	P@5	P@10	MAP	NDCG@5	NDCG@10
20%	0.392	0.388	0.209	0.393	0.399
40%	0.432	0.406	0.242	0.445	0.430
60%	0.452	0.414	0.259	0.461	0.440
80%	**0.478**	0.456	0.280	**0.488**	0.475
Baseline	0.472	**0.458**	**0.321**	0.486	**0.476**

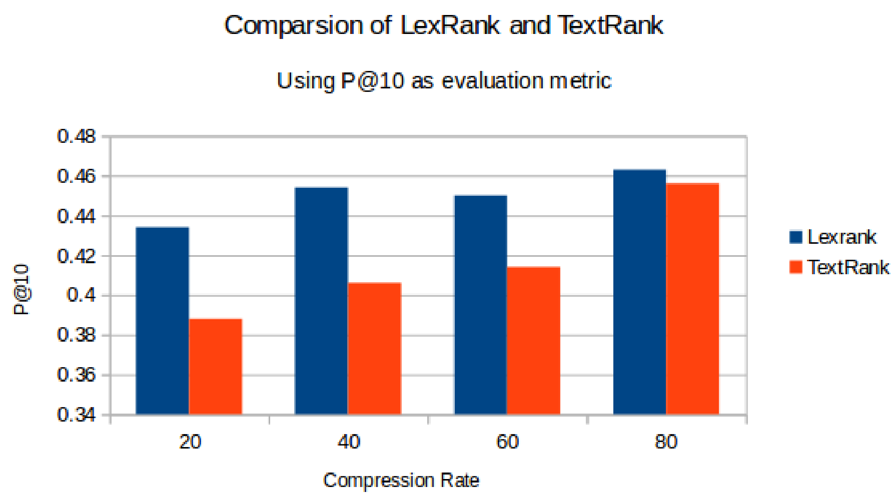

Figure 5.5: Comparison of LexRank vs Textrank on retrieval performance(in terms of P@10)

From the results obtained in Table 5.3 and Table 5.4 we observe that the summarization before indexing improves performance for certain compression ratios. It can also be observed that retrieval on summarized documents outperforms the baseline in terms of P@5. We also observe that in both Lexrank and Textrank the retrieval performance gradually increases as we increase the compression ratio, peaking at around 80% compression ratio, which is outperforming the baseline in terms of P@10. It should also be noted that the MAP score for the baseline is the highest.

5.2.2 Comparison of LexRank vs TextRank

In Fig. 5.5 we obsever that LexRank clearly outperforms TextRank in terms of P@10. On comparing Table 5.3 and Table 5.4 we observe that this is true for most metrics. We also observe the performance gap between TextRank and LexRank is very high for low compression ratios, the gap is much lower for higher compression ratios.

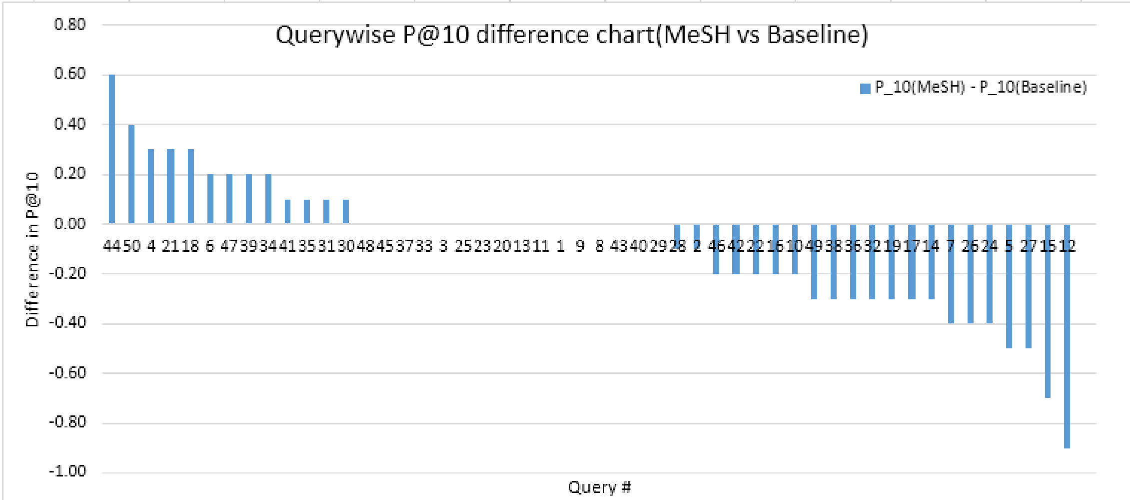

Figure 5.6: Querywise difference graph comparing MeSH vs Baseline run.

5.3 Result Analysis

5.3.1 Investigating use of MeSH

We observe that using MeSH as a thesaurus can boost the overall performance of the retrieval system. Although from the difference graph we observe that this is not true for all queries. We therefore need a mechanism to perform *selective query expansion*. From Fig 5.6 we observe that using MeSH improved performance on 13 queries. On inspecting the queries for which use of MeSH hampered the performance, we observed that for such queries the query was flooded with a lot of synonyms some of them being irrelevant or redundant. We therefore need to have a ranking mechanism to rank the query expansion terms and prune terms that are not highly ranked.

5.3.2 Investigating use of Discharge Summaries

Using Discharge Summaries and Query likelihood gave the best MAP score. From inspecting the queries for which use of Discharge summaries hampered the results we observed that several medical entities, irrelevant to the current query were added to the query thus hampering the performance.

5.3.3 Analysing effect of summarization before indexing

Text summarization was used to retain salient information from the documents while pruning irrelevant material. We observed that using summarization at 40-

80% works best for our purposes, often better than the baseline. For compression rates of less that 40% relevant information was lost leading to reduction in retrieval performance.

CHAPTER 6
Conclusion

We described the different approaches to user-centered health information retrieval. We observed that health information retrieval poses several challenges. To address the issue of vocabulary mismatch proper treatment of both the query and the document are required.

Using medical resources namely MeSH and Discharge summaries has shown promise. From the results we observed that query expansion resulted in improvement for only a set of of queries. Thus we need to develop mechanisms to perform selective query expansion.

A summarized index can be used to replace the full corpus index. Text summarization with a compression rate of 40-80% outperformed the baseline. This is because the documents in the summarized collection retain only salient information while pruning irrelevant information. Hence if there is an agreement between the query and the summarized document there is a high chance that the original document and the query are also in agreement.

References

[1] S. Fox, "Health topics: 80% of internet users look for health information online," *Technical Report, Pew Research Center*. [Online; accessed May-2015].

[2] H. Thakkar, G. Iyer, K. Shah, and P. Majumder, "Team irlabdaiict at share/-clef ehealth 2014 task 3: User-centered information retrieval system for clinical documents," *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*, 2014.

[3] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. Jones, *et al.*, "Overview of the share/clef ehealth evaluation lab 2013," in *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pp. 212–231, Springer Berlin Heidelberg, 2013.

[4] R. B. Trieschnigg, *Proof of concept: concept-based biomedical information retrieval*. No. 10-176, Centre for Telematics and Information Technology, University of Twente, 2010.

[5] L. Goeuriot, G. J. Jones, L. Kelly, J. Leveling, A. Hanbury, H. Müller, S. Salantera, H. Suominen, and G. Zuccon, "Share/clef ehealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports," *CLEF 2013 Online Working Notes*, vol. 8138, 2013.

[6] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," *International journal of human-computer studies*, vol. 43, no. 5, pp. 907–928, 1995.

[7] R. K. Saripalle, "Current status of ontologies in biomedical and clinical informatics," *Retrieved November*, vol. 16, p. 2013, 2013.

[8] A. L. Rector, W. A. Nowlan, G. Consortium, *et al.*, "The galen project," *Computer methods and programs in biomedicine*, vol. 45, no. 1, pp. 75–78, 1994.

[9] C. Fellbaum, *WordNet*. Wiley Online Library, 1998.

[10] C. T. Lopes, "Context features and their use in information retrieval," in *Third BCS-IRSG symposium on future directions in information access*, 2009.

[11] Z. Lu, W. Kim, and W. J. Wilbur, "Evaluation of query expansion using mesh in pubmed," *Information retrieval*, vol. 12, no. 1, pp. 69–80, 2009.

[12] V. Kakade and M. Sharangpani, "Improving the precision of web search for medical domain using automatic query expansion," *On-line]. http://www. stanford. edu/class/archive/cs/cs276a/cs276a*, vol. 1032, 2004.

[13] W. Hersh, S. Price, and L. Donohoe, "Assessing thesaurus-based query expansion using the umls metathesaurus.," in *Proceedings of the AMIA Symposium*, p. 344, American Medical Informatics Association, 2000.

[14] V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," *Journal of emerging technologies in web intelligence*, vol. 1, no. 1, pp. 60–76, 2009.

[15] M.-Y. Kan and J. L. Klavans, "Using librarian techniques in automatic text summarization for information retrieval," in *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pp. 36–45, ACM, 2002.

[16] A. Tombros and M. Sanderson, "Advantages of query biased summaries in information retrieval," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 2–10, ACM, 1998.

[17] A. Turpin, Y. Tsegay, D. Hawking, and H. E. Williams, "Fast generation of result snippets in web search," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 127–134, ACM, 2007.

[18] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[19] T. Sakai and K. Sparck-Jones, "Generic summaries for indexing in information retrieval," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 190–198, ACM, 2001.

[20] J. M. Perea-Ortega, E. Lloret, L. A. Ureña-López, and M. Palomar, "Application of text summarization techniques to the geographical information re-

trieval task," *Expert Systems with Applications*, vol. 40, no. 8, pp. 2966–2974, 2013.

[21] J. Lu and J. Callan, "Pruning long documents for distributed information retrieval," in *Proceedings of the eleventh international conference on Information and knowledge management*, pp. 332–339, ACM, 2002.

[22] F. Raposo, R. Ribeiro, and D. M. de Matos, "Using generic summarization to improve music information retrieval tasks," *arXiv preprint arXiv:1503.06666*, 2015.

[23] A. M. Lam-Adesina and G. J. Jones, "Applying summarization techniques for term selection in relevance feedback," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 1–9, ACM, 2001.

[24] D. Zhu, S. Wu, M. James, B. Carterette, and H. Liu, "Using discharge summaries to improve information retrieval in clinical domain," *Proceedings of the ShARe/-CLEF eHealth Evaluation Lab*, 2013.

[25] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts," Association for Computational Linguistics, 2004.

[26] G. Erkan and D. R. Radev, "Lexrank: graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, pp. 457–479, 2004.

[27] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," 1999.

[28] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," *Journal of documentation*, vol. 60, no. 5, pp. 503–520, 2004.

[29] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*, vol. 1. Cambridge university press Cambridge, 2008.