Team IRLabDAIICT at ShARe/CLEF eHealth 2014 Task 3: User-centered Information Retrieval system for Clinical Documents

Harsh Thakkar[1], Ganesh Iyer[2], Kesha Shah[3], Prasenjit Majumder[4]

IR Lab, DA-IICT, Gandhinagar, Gujarat, India
{harsh9t[1], lastlegion[2], kesha.shah1106[3], prasenjit.majumder[4]} @gmail.com

Abstract. In this paper we, Team IRLabDAIICT, describe our participation in the ShARe/CLEF ehealth 2014 task 3: Information Retrieval for addressing questions related to patients health based on clinical reports. We submitted a total of six runs out of the seven in this years task. In our approach we focus on examining the relevance between the documents and user generated query by conducting experiments through query analysis. Our major challenge is to bridge the conceptual gap between the user-generated queries (in-formal query) to biomedical specific terminology (formal query). We incorporate the MeSH (Medical Subject Headings) library , which is a medical thesaurus mapping layman terms to medical synonym terms in order to target the concept matching problem. We use blind relevance feedback model for relevance feedback and query-likelihood model for query expansion which performed the best in the experiments conducted by us. The retrieval system is evaluated based on various parameters as: mean average precision, precision (P@5), precision (P@10), NDCG@5 and NDCG@10, with P@10 and NDCG@10 being the primary and secondary evaluation measures. The experiments were conducted on the gigantic 43.6 GB ShARe/CLEF 2013 Task 3 dataset harvested using (a) EU-FP7 Khresmoi project and and (b) a new 2014 set of English general realistic public queries based on the discharge summary contents. We have obtained the highest result in our baseline run (run 1), with compared to our other five runs, which is 0.706 as declared by ShARe/CLEF organizing committee. We further propose to incorporate a machine learning based retrieval algorithm prediction model for further exploration.

1 Introduction

With the increase in awareness amongst people regarding health issues, a dire need for expanding the horizons of research on medical document retrieval has become mandatory these days. Patients now want answers to their health problems on the touch of a finger. Discharge summaries obtained from the physicians has attracted a lot of attention from patients. Thus, the concept of health information retrieval has become more popular[1]. The main challenge in this

area is to answer the patients questions[2] in a format which is understandable by the layman (i.e. the user/patient). The medical prescriptions and discharge summaries are written in professional medical terminology which makes no sense to the end user (patient). Taking this challenge as an opportunity the ShARe/CLEF (Cross Lingual Evaluation Forum) started an eHealth Task in year 2013, with a goal to develop such a system by attracting the young researchers from various organizations and universities of the world of computer science and biological domain and present a common platform for conducting the research. The ShARe/CLEF (Cross Lingual Evaluation Forum) community addressed this problem by initiating the eHealth Tasks from 2013[3] with a goal to evaluate systems that support laypeople in searching for and understanding their health information.

This ShARe/CLEF challenge ehealth 2014[4] comprises three tasks[1]. The specific use case considered is as follows: before leaving the hospital, a patient receives a discharge summary. This describes the diagnosis and the treatment that they received in the hospital. The **first task** considered in CLEF eHealth 2014[4] aims at delivering a visualization of the medical information extracted from the discharge summaries in manner which is conceivable by layman. While the **second task** requires normalization and expansion of abbreviations and acronyms present in the discharge summaries. The use case then postulates that, given the discharge summaries and the diagnosed disorders, patients often have questions regarding their health condition[5]. The goal of the **third task**[6, 7] is to provide valuable and relevant documents to patients, by developing a user-centered[2] or context-based[8] health information retrieval system so as to satisfy their health-related information needs.

In order to aid the evaluation process of the systems involved in task 3, the task organizers have provided us with the potential user queries and their relevant judgments which have been obtained from medical professionals and an enormous dataset consisting of a variety of health and biomedical documents. As is common in evaluation of information retrieval (IR), the test collection consists of documents, queries, and corresponding relevance judgments.

This paper describes our participation in Task 3 of 2014 ShARe/CLEF eHealth[4] Evaluation Lab. This paper is organized as follows: Section 2, discusses the collection of documents provided in the corpus, its characteristics, relevance assessment of the documents and guidelines for the submission of results. Section 3, presents the description of our proposed user-centered health information retrieval system. Section 4, discusses the conducted experimental runs and the analysis of harvested results. Section 5, concludes the paper with authors comments and future work.

2 Corpus

The corpus provided by ShARe/CLEF organizers[6] consists of a distributed medical web crawl of a large collection of files containing the documents from

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

the EU-FP7 Khresmoi (http://khresmoi.eu/) project's 2012 medical documents. This dataset consists of 1.6 million English documents covering a wide set of medical topics. This collection is prepared from a variety of sources (online) sources, including Health on the Net Foundation[2] certified websites, as well as well-known medical sites and databases (e.g. Genetics Home Reference, ClinicalTrial.gov, Diagnosia[3])

The corpus is divided into eight archive files (zip files) named *part_number.zip*. the size of the corpus is *6.3 GB* compressed and about **43.6 GB** uncompressed. This collection is the processed version of crawled documents after eliminating out the very small documents and correction of some errors in mark-up (i.e. applying Jsoup functions) from the original crawl. This document collection of CLEF eHealth 2014 Task 3 is split into a group of .dat files. Each of the .dat files in this collection contains the web crawl for one domain (where a domain is defined to be the root URL).

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

This crawled dataset is a result of work by: Health on the Net Foundation (HON), Switzerland and University of Applied Sciences Western Switzerland (HES-SO), Switzerland.

2.1 Relevance assessment

The information provided by the ShARe/CLEF on the relevance assessment is as follows:

- The official training query and result set for eHealth Task 3 consists of 5 queries and corresponding result set generated from manual relevance assessment (by medical professionals) on a shallow pool.
- Relevance assessments for these 5 training queries were formed based on pooled sets generated using the Vector Space Model and Okapi BM25.
- 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).

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

A sample query from the official 5 training queries[6] is shown below for reference:

```
<topic>
<id>QTRAIN2014.1</id>
<title>MRSA and wound infection</title>
<desc>What is MRSA infection and is it dangerous?
</desc>
<profile>This 60 year old lady has had coronary artery bypass grafting
surgery and during recovery her wound has been infected. She wants to
know how dangerous her infection is, where she got it and if she can
be infected again with it.
</profile>
<narr>Documents should contain information about sternal wound infection
by MRSA. They should describe the causes and the complications.
</narr>
</topic>
```

2.2 Submission guidelines

The guidelines stated for submission of the results of task 3 are as follows: Participants are asked to submit up to seven ranked runs for the English (EN) queries in Task 3. The runs to be submitted are described as following:

1. Run 1 (mandatory) is a baseline: only title and description in the query can be used, and no external resource (including discharge summary[5], corpora, ontology, etc.) can be used.
2. Runs 2-4 (optional) any experiment WITH the discharge summaries.
3. Runs 5-7 (optional) any experiment WITHOUT the discharge summaries.

One of the runs from 2-4 and one from 5-7 must use the IR technique in Run 1 as a baseline. The idea being to allow analysis of the impact of discharge summaries/other techniques on the performance of the baseline Run 1.

The optional runs must be ranked in order of priority (for Runs 2-4, 2 is the highest priority; for Runs 5-7, 5 is the highest priority).

3 Retrieval system

3.1 System overview

The figure 1 represents the block diagram of our proposed retrieval system. First we pre-process the dataset provided. We first convert the data of the collection to standard TREC format which is acceptable by Indri[4]. Indri is a software product of lemur for building indexing and retrieval systems based on language models. Then this pre-processed data goes through subsequent language processing steps as stemming (using porter-stemmer), stop-word removal (using a mixture of standard English and free medical dictionary comprising of 4000 medical words). Thus the data provided was first cleaned for the next (indexing) process.

[4] Indri - Online at http://www.lemurproject.org/indri.php

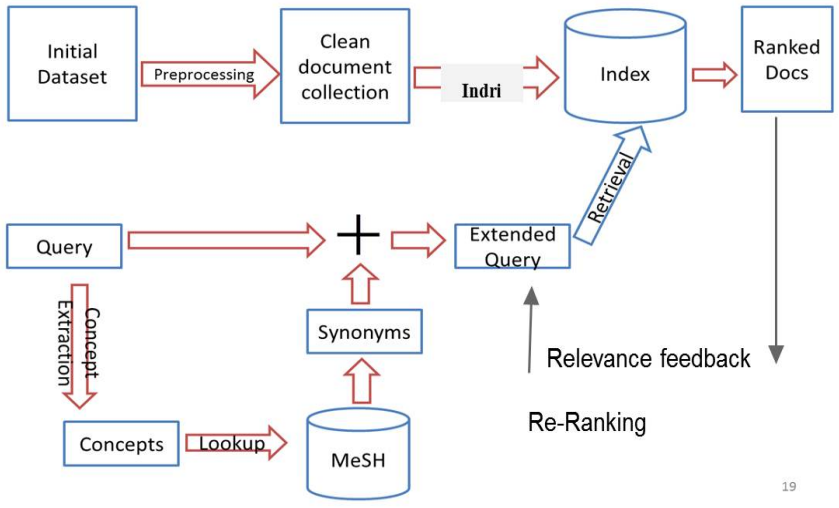

Fig. 1. A systematical block diagram of the system for eHealth Task 3.

3.2 Information retrieval process

The entire process of information/text retrieval is divided in to sub-parts namely:

- Indexing
- Query expansion
- Documents retrieval for every query
- Experiments and results

Indexing: In the first phase, 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[5]. Stanford Porter stemmer is used for stemming during indexing.

Query expansion: 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 general English (US) dictionary used in enchant and the second dictionary is specifically for medical domain.[6] Blind relevance feedback is also used to re-rank the documents. We use Metamap[9] to integrate MeSH[7] to extract medical nomenclature of layman terms, used in the subsequent runs 6 and 7 for query expansion. MeSH (stands for Medical Subject

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

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

[7] MeSH - Online at www.ncbi.nlm.nih.gov/mesh

Headings) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed. MeSH is used to map the query from a textual base (layman representation) to a concept base (medical terminology). Runs 6 and 7 are based on MeSH.

We make use of discharge summaries provided from the ShARe/CLEF ehealth task 2 dataset in runs 2 and 3 for the task of query expansion. We obtain the relevant discharge summary file for the user query from the documents indexed previously. Further, we extract medical terms from the tags like {`Major Surgical or Invasive Procedure, Past Medical History, Past Surgical History, Chief Complaint, Discharge Diagnosis`} from the information (clinical analysis data consisting of various other tags) provided in that specific discharge summary file. We use a combination of the the words extracted from user query and the discharge summaries with a ratio of 4:1 (i.e. 0.8 weightage is given to words extracted from user query whereas 0.2 weightage is given to those extracted from discharge summaries). Thus, we employ an alternative approach to using medical thesaurus is employed in our retrieval system.

Documents retrieval: In the third phase, scoring of documents is done for each query. The scores were calculated by running query on three retrieval models namely Okapi, tf-idf and Query Likelihood model. Okapi and tf-idf are non-language based models whereas Query Likelihood is language based model. However, after evaluating the lab results on last years test queries we decided discard the tf-idf runs due to its poor performance. The tf-idf model was reported to perform even worse than the okapi model. Hence, we have not compiled the tf-idf statistics in this report.

Thus we submitted six successful runs to the task which consists of okapi, MeSH and query-likelihood (baseline run).

Query-likelihood model The Query-likelihood is a language based model[10]. Using this, 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[11] 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:

$$P(q|M_d) = K_q \prod_{t \epsilon V} P(t|M_d)^{tf_{t,q}} \tag{1}$$

Where the multinomial coefficient is :

$$K_q = L_q!/(tf_{t1,q}!tf_{t2,q}!...tf_{tM,q}!) \tag{2}$$

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.

Okapi model We make use of the famous Okapi Model[12, 11]. The weighting based documents score is calculated as in below formulae:

$$RSV_d = \sum_{t \epsilon q} \left[log \frac{N}{df_t} \right] . \frac{(k_1 + 1) tf_d}{k_1((1 - b) + b(L_d / L_{ave})) + tf_d} . \frac{(k_3 + 1) tf_{tq}}{k_3 + tf_{tq}} \quad (3)$$

$$score(D, Q) = \sum_{i=1}^{n} IDF(q_i) \frac{f(q_i, D).(k_1 + 1)}{f(q_i, D) + k_1.(1 - b + b.\frac{|D|}{avgdl})} \quad (4)$$

$$IDF(q_i) = log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \quad (5)$$

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. This 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.

3.3 Experiments and results

Submitted Runs We submitted six of the seven runs in the task that are described as follows.

Run	Query-Likelihood	MeSH	Okapi	Discharge summaries	Summary
1	✓				mandatory Baseline run
2	✓			✓	optional run WITH discharge summary 1
3			✓	✓	optional run WITH discharge summary 2
5			✓		optional run WITHOUT discharge summary 1
6	✓	✓			optional run WITHOUT discharge summary 2
7		✓	✓		optional run WITHOUT discharge summary 3

1. **RUN 1:** the first run 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 {`Major Surgical or Invasive Procedure, Past Medical History, Past Surgical History, Chief Complaint, Discharge Diagnosis`} tags, for query expansion and document retrieval processes respectively. The medical words are extracted from discharge summaries and incorporated with the word set obtained by query-likelihood model using a special weight function. The words obtained from the 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 5:** Run 5 through 7 is the runs WITHOUT Discharge summaries. Run 5 make use of the Okapi model along with blind relevance feedback for retrieving documents.
5. **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.
6. **RUN 7:** we used the Okapi model for retrieval along with MeSH for query expansion.

We discuss the official findings released by ShARe/CLEF organizing committee in the following section along with their analysis.

4 Official results and Discussion

The precision at 10 (P@10) and normalized discounted cumulative gain were selected as the primary and secondary evaluation parameters for ehealth task 3 2014. Figure 2 shows the results of the submitted six runs. Figure 4 shows the variance in the performance of the six runs. It is clear from figure 2 that run 1 (baseline) is the best run yielding the highest values followed by run 2, 6 and 5 respectively. Whereas run 7 is the least performing run followed by run 3. It is

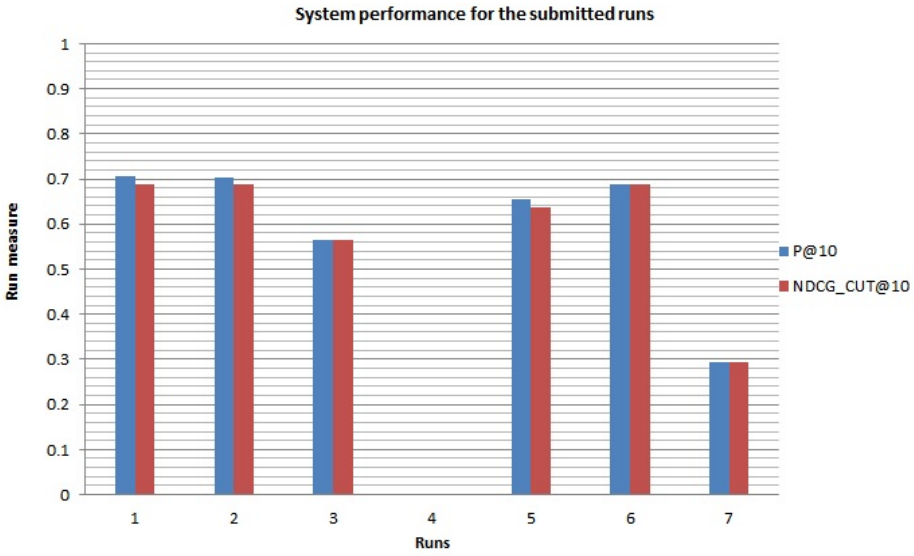

Fig. 2. official values of P@10 and NDCG_CUT@10 values for ehealth task 3 2014 released by CLEF.

observed that the best performing models are the query-likelihood model and its duo combination with discharge summaries or MeSH. Whereas the trio (i.e. the combination of query-likelihood, okapi and Mesh or discharge summaries) shows a drastic fall in the performance of the system. This is caused by the caching of irrelevant words extracted from the different texts. We observe that the query length is increased by leaps and bounds in the trio with respect to that in the duo. Thus, in the case of medical text retrieval, extensive use of keywords does not guarantee higher performance. Instead it suppresses the more relevant words and increases the vagueness in the query. It also observed from the comparison result of the runs 2 and 3 (with discharge summaries) that the use of combination of the okapi model and discharge summaries do not show any improvement over the baseline run rather a 12% downfall of the system performance. Whereas the

combination of query-likelihood and discharge summaries preserves the system performance when compared to the baseline run.

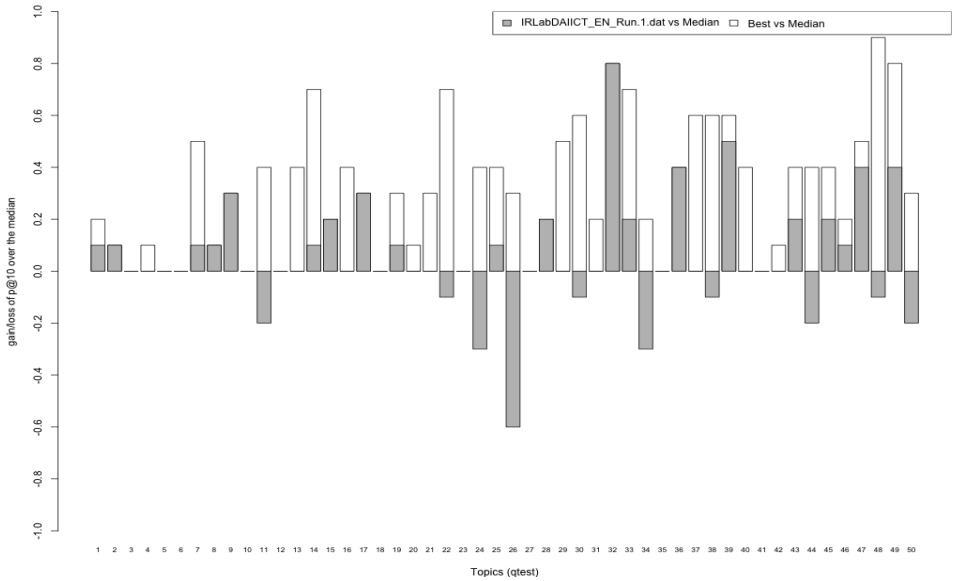

Fig. 3. Comparison graph of query wise evaluation of the baseline run (run 1) with other participating teams.

Query-wise comparison of the baseline run:
Figure 3 shows the query wise performance graph of the participating teams in the ehealth task 3 2014. Figure 4 is computed based on these given results for each run. For the baseline run it is observed that our system out-performs other systems in queries 2, 8, 9, 15, 17, 28, 32 and 36 respectively. On the other hand, our baseline system lags in queries 11, 22, 24, 26, 30, 34, 38, 44, 48,and 50 respectively. Figure 4 shows the query wise performance of other five runs. It is clear from the above figures that the query-likelihood model out-performs the okapi model in run 2 and run 6 as compared to in run 3 and run 5.

5 Conclusion & Future work

The nature of query varies from being very precise to being extremely vague. The model selected for a specific query type performs in accordance with the nature of the query. For the medical text retrieval task it is clear that the query-likelihood model works the best so far than the other models like okapi and tf-idf. We carried out experiments with the tf-idf model in the lab tests but its results were poor than that of the okapi model and thus were later excluded from the

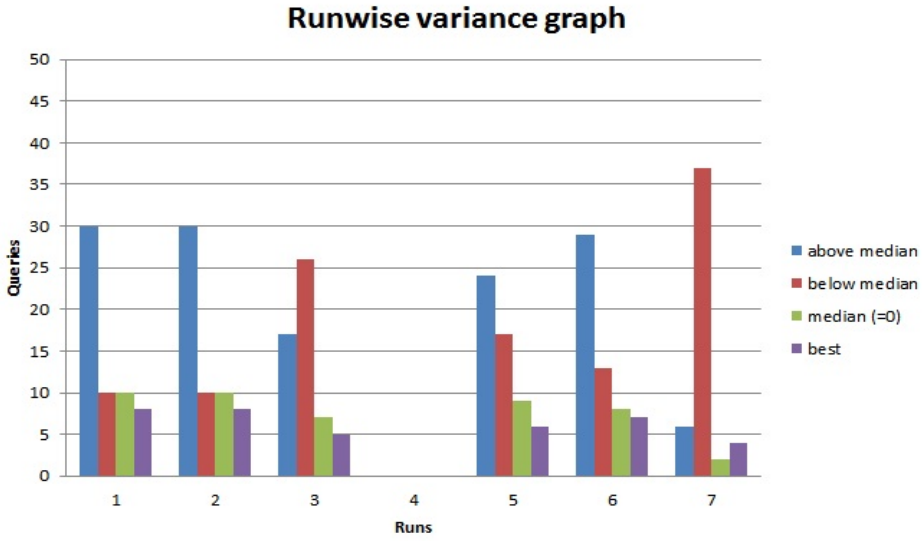

Fig. 4. Comparison of performance variance between the six runs of ehealth task 3 2014.

submitted runs (as mentioned in section 3.2). Hence, it can be concluded that tf-idf is not a suitable model for medical document information retrieval. Moreover, there is a need for developing a mechanism through which the deployment of these models could be predicted beforehand. By judging the nature of the query from the text, we can incorporate which model or which combination of models to use.

Keeping the current constraints in mind, we propose to develop a machine learning based retrieval algorithm prediction model for predicting query performance based on the features extracted from the query as future work. The features comprise of various factors like the combination of keywords/terms used in the query, length of query, query similarity score and etc. Incorporating such a mechanism promises to improve the evaluation score of our system by a factor of 8-10 %.

6 Acknowledgments

We would like to specially thank and acknowledge our faculty advisor Prof. Prasenjit Majumder for being around to provide quality inputs for the system. We would also like to convey our regards to the ShARe/CLEF team for organizing the eHealth Task enabling teams like ours to participate and give us a chance to contribute to the community to the best of our abilities.

References

1. Lopes, C.T.: Health information retrieval - a state of art report. Technical Report, Faculdade de Engenharia da Universidade do Porto (2013)
2. Burstein, F., Fisher, J., McKemmish, S., Manaszewicz, R., Malhotra, P.: User centred quality health information provision: benefits and challenges. Proceedings of the 38th Hawaii International Conference on System Sciences (2005)
3. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Pradhan, S., South, B.R., Mowery, D.L., Jones, G.J., et al.: Overview of the share/clef ehealth evaluation lab 2013. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization. Springer (2013) 212–231
4. Kelly, L., Goeuriot, L., Suominen, H., Schrek, T., Leroy, G., Mowery, D.L., Velupillai, S., Chapman, W.W., Martinez, D., Zuccon, G., Palotti, J.: Overview of the share/clef ehealth evaluation lab 2014. In: Proceedings of CLEF 2014. Lecture Notes in Computer Science (LNCS), Springer (2014)
5. Zhu, D., Wu, S., James, M., Carterette, B., Liu, H.: Using discharge summaries to improve information retrieval in clinical domain. Proceedings of the ShARe/-CLEF eHealth Evaluation Lab (2013)
6. Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G., Mueller, H.: Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval. In: Proceedings of CLEF 2014. (2014)
7. Goeuriot, L., Jones, G., Kelly, L., Leveling, J., Hanbury, A., Müller, H., Salanterä, S., Suominen, H., Zuccon, G.: Share/clef ehealth evaluation lab 2013, task 3: Information retrieval to address patients questions when reading clinical reports. Online Working Notes of CLEF, CLEF (2013)
8. Zhong, X., Xia, Y., Xie, Z., Na, S., Hu, Q., Huang, Y.: Concept-based medical document retrieval: Thcib at clef ehealth lab 2013 task 3. Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2013)
9. Aronson, A., Lang, F.M.: An overview of metamap: historical perspective and recent advances. Journal of the American Medical Informatics Association **17** (2010) 229–236
10. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM (1998) 275–281
11. Manning, C., Raghavan, P., Schütze, H.: Introduction to information retrieval. Volume 1. Cambridge university press Cambridge (2008)
12. Robertson, S.: Understanding inverse document frequency: on theoretical arguments for idf. Journal of documentation **60** (2004) 503–520