TeamHCMUS: A Concept-Based Information Retrieval Approach for Web Medical Documents

Nghia Huynh[1], Thanh Tuan Nguyen[2], Quoc Ho[1]

[1] Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
huynhnghiavn@gmail.com, hbquoc@fit.hcmus.edu.vn
[2] Faculty of Information Technology, HCMC University of Technology and Education, Vietnam
tuannt@fit.hcmute.edu.vn

Abstract. It's difficult for laypeople, even clinicians, to understand eHealth contents found in the web medical documents. With the objective to build a health search engine, task 2 of 2015 CLEF eHealth aims to detect levels of accuracy of information retrieval systems when searching for web medical documents. In this task, our approach is to integrate a retrieval of medical concepts into the preprocessing of corpora. This means that all terms in the documents that are not related to medicine are removed before indexing. We also expand queries for searching more effectively. In general, our results are not better than other participants' in doing task 2 except some queries. When using integration of extracting medical concepts and query expansion based on laypeople's queries, searching retrieval is also lower. It can be explained partly that laypeople's queries are not commonly included medical terms or only contain features painting their health situations. In addition, we also give a brief statement of the main points of an estimation of readability which is a significant assessment referred by CLEF eHealth near future.

Keywords: Concept-based, Medical Information Retrieval, Medical Documents, Language Model

1 Introduction

Laypeople as well as clinicians find it hard to comprehend the eHealth documents which are retrieved from searching for necessary information on the Internet. Their problems are how to understand professional terms more exactly. It's the third year when CLEF eHealth continues the purpose of promotion in doing research and finding out advanced methods to build a search engine system for meeting users' requirements of searching for medical information.

CLEF eHealth pointed out two tasks this year[1] . Task 1 is a mission statement of information extraction from clinical text. It is split into two sub-tasks: task 1a and task 1b. Specifically, task 1a is clinical speech recognition related to converting verbal nursing handover to written free-text records and task 1b is named entity recognition in clinical reports. Task 2 is user-centered health information retrieval [15-16]. In this paper, we proposed methods to meet some requirements of task 2.

There are some changes of type of queries and measure methods of relevance assessments this year. One of them is queries that should be made by laypeople do not come from experts in the heath domain because the laypeople want to find out information to help them to be clearer their related medical conditions. Another change is to apply two measures to assess results from participants' submissions.

As a further matter, CLEF eHealth has also considered a readability-biased assessment [14], the factor of understandability of information (or readability) within the evaluation of submissions. However, because the measure has been still at levels of experiment consideration, observations that were carried out might not be conclusive.

In this paper, our approach is to integrate retrieval of medical concepts from the web medical documents into the preprocessing of corpora which is based on a list of medical concepts built by [2]. Process of building a search engine system can be summarized as following descriptions: At first, we used some tools to remove tags of HTML files in the corpus provided by 2015 CLEF eHealth[2] for task 2. We collected a set of raw data from the corpus. We then removed stopwords and got stemming of terms in each document in the set. All data extracted from this process was indexed and called Index A. Another indexed corpus called Index B was also created from the data that only includes terms related to medical domain. Building the Index B is described in detail in Section 2.3.

We obtained a baseline run and other runs after doing experiments in searching for queries in Index A and B corpus with Dirichlet smooth coefficients [13].

We also expanded queries to get more information for searching. Techniques to do this are described in Section 3.

---

1 https://sites.google.com/site/clefehealth2015

2 https://sites.google.com/site/clefehealth2015/task-2

In general, most of our results are not better than other participants' in doing task 2 except some queries (see Figure 2, 3). In addition, when integrating the method of extracting medical concepts [2] in building Index B as well as expending laypeople's queries, searching results are also lower (see Figure 4, 5). This can be explained that their queries are not commonly included medical terms or only contain terms painting their health situations.

The rest of the paper is organized as follows: Section 2 outlines the CLEF eHealth dataset and methods of preprocessing. Section 3 describes the structure of a query and some techniques for query expansion. Section 4 presents relevance assessments. Section 5 demonstrates description of our runs. Section 6 explains experiments done by task participants. Finally, Section 7 concludes the paper.

2 Dataset and Preprocessing

The dataset for Task 2 is provided by Khresmoi project[3]. It has about one million documents, a set of documents in the HTML (Hyper Text Markup Language) format. All documents are collected from well-known health and medical sites and databases in 2012. The size of the dataset is about 6.3G in compressed status and approximately 43.6 GB after extracting.

Each file in the dataset is in the format of .dat files and contains a set of web pages and metadata where shows the original information of each web page as described below:

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

2.1 Parse HTML to Text

Majestic-12, Distributed Search Engine (DSearch) projects[4], built an open-source tool called HTML parser v3.1.4 for parsing tags in the HTML files. Basing on this tool, we extracted text contents from tags of HTML documents in the dataset. There are some tags in the documents that contain unnecessary

[3] http://khresmoi.eu/
[4] http://www.majestic12.co.uk/projects/

texts for building a search engine system in the medical field. Thus we tried to ignore all those.

2.2 Content Cleaning

The text contents extracted from HTML documents are not always good for building the system. Example, tags contain text of sitemap, update information and some items on the main and popup menu in the HTML documents. Therefore, all of them should be removed.

Next, we had a process of removing stopwords because they were common words in English language [4] and did not play an important role in determining the meaning of sentences. To get more efficient in preprocessing, we used the Porter algorithm [9] for stemming words. Finishing all above work, we had a corpus ready for indexing. We used Lucene-5.0.0 tool[5] to index this corpus and name Index A.

2.3 Extracting Concepts

We used a list of medical concepts built by [2] to extract medical concepts from the documents in the dataset by removing all their terms that are not in the list and also not in UMLS[6] (Unified Medical Language System).

After this processing, we had a dataset that contains terms related to medical field. We also used Lucene-5.0.0 to index this dataset and named Index B.

3 Queries and Query Expansion

Because of consideration in building a search engine system for English documents, we only concentrated on English queries. The number of queries for task 2 of 2015 CLEF eHealth includes 66 queries along with their narrative fields. The narrative fields are used to provide information to the assessors when performing relevance evaluations. Here is the typical structure of a query[7]:

---

[5] http://lucene.apache.org/

[6] http://www.nlm.nih.gov/research/umls/

[7] https://sites.google.com/site/clefehealth2015/task-2

<top>
<num>clef2015.training.1</num>
<query>loss of hair on scalp in an inch width round</query>
<narr>Documents should contain information allowing the user to understand they have alopecia</narr>
</top>

With purpose of getting more information for searching, with some tasks of CLEF eHealth before, participants or teams found out synonym of query's terms in the UMLS or MeSH[8] (Medical Subject Headings) for expanding their queries [10], [12]. Other participants used pseudo-relevance feedback (PRF) as a method for expansion [6], [11], or took Wikipedia[9] for making a semantic query expansion [1]. Because the set of queries of task 2 of 2015 CLEF eHealth is user-centered queries (i.e. they are made by laypeople rather than done by experts in the medical field), Terms of the queries are usually short and not much relative to medical concepts. So we lacked evidences to expand the queries. Thus, to get more information for every query expansion, we searched for queries in each Index (A and B) to get the relevant document at the top of each searching result. To get an expanded query, we connected that top document to the query.

4 Relevance Assessments

Methods of relevance assessments are provided by the Share/CLEF eHealth 2015 TASK 2 and described as follows:

- Result of runs is the top 1000 of relevant documents returned by searching for 66 queries that based on LM (language model) [8] with specification of Dirichlet smooth coefficients.
- Relevance is assessed as following descriptions:
  - Evaluation with standard trec_eval metrics[10]:
    - 2 point scale: non relevant (label 0); relevant (label 1)
  - Evaluation with nDCG: [3]
    - 3 point scale: gain 0 (label 0), gain 1 (label 1), gain 2 (label 2)
  - Readability-biased evaluation: [14]

[8] http://www.ncbi.nlm.nih.gov/mesh
[9] https://en.wikipedia.org
[10] http://trec.nist.gov/trec_eval/

- 4 point scale: very technical and difficult (label 0), somewhat technical and difficult (1), somewhat easy (label 2), very easy (label 3)

5 Description of Runs

With given 66 queries, we retrieved the top 1000 of relevant documents for each query when using LM in Lucene 5.0 for matching the queries with each document in the Index A or B corpus along with specific values of Dirichlet smooth coefficients [13]. We submitted 8 runs in the task that are described in summary as follows: (see Table 1)

Table 1. List of specification of submitted eight runs

Run	Index corpus		Query expansion	smooth coefficients	
	A	B		2000	10000
1	✓			✓	
2	✓				✓
3		✓		✓	
4		✓			✓
5	✓		✓	✓	
6	✓		✓		✓
7		✓	✓	✓	
8		✓	✓		✓

Run 1 (baseline run): We applied the default value (2000) of smooth coefficient to LM and search in Index A corpus.
Run 2: It is a variant of run 1 in which value of smooth coefficient is 10000.
Run 3: We used the corpus of medical concepts (i.e. Index B) for searching with the same smooth coefficient as run 1.
Run 4: It is a variant of run 3 in which value of smooth coefficient is 10000.
Run 5: Each query, we took the top 01 of relevant documents at Run 3 for expanding the query. Executing the same run 1, but for this expanded query.
Run 6: It is a variant of run 5 in which the top 01 of relevant documents at run 4 was used to expand the query. Executing the same Run 2, but for this expanded query.
Run 7: Each query, we took the top 01 of relevant documents at run 3 for expanding the query. Executing the same run 3, but for this expanded query.
Run 8: It is a variant of Run 7 in which the top 01 of relevant documents at run 4 was used for expanding the query. Executing the same run 4, but for this expanded query.

6 Experiments

6.1 Evaluation with standard trec_eval metrics and nDCG

Two primary evaluation parameters for task 2 of 2015 CLEF eHealth are the precision at 10 (P@10) and Normalized Discounted Cumulative Gain (NDCG) at rank 10. Figure 1 shows the results of the submitted eight runs. It can be seen clearly from Figure 1 that run 1 (baseline run) is the best run yielding the highest values followed by run 2 to 5, 7 respectively. Whereas run 8 is the least performing run and following it is run 6. It is observed that the values of nDCG are higher than P@10 in processing with integration of query expansion into the system (run 5 to 8).

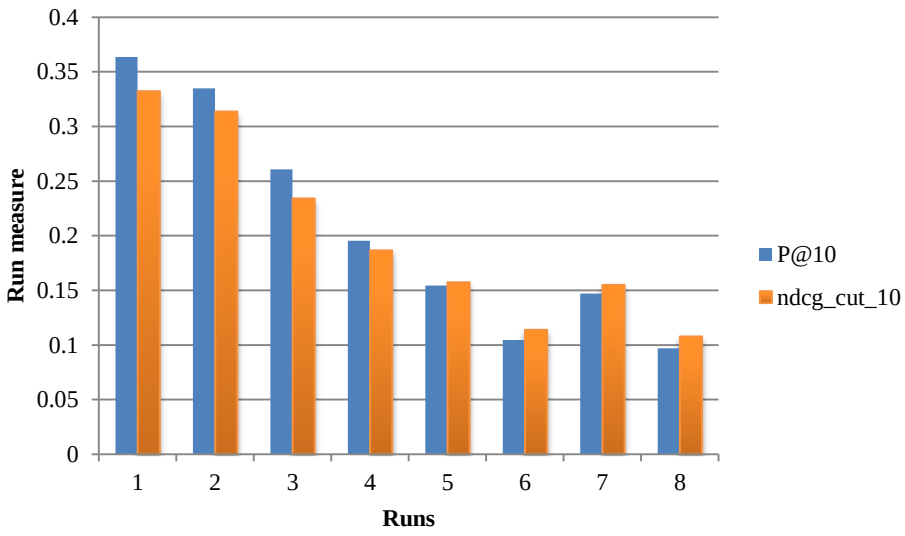

Fig. 1. P@10 and nDCG_cut_10 values for eHealth task 2 2015, released by CLEF

Demonstration of Figure 1 shows that our approach of integrating a retrieval of medical concepts [2] into the system (i.e. run 3, 4) does not perform better than the baseline run. This is because documents in Index B only contain medical concepts while queries have few terms related to the medical field.

The Figure 1 also indicates that run 5 to 8 with performance of query expansion techniques is worse than other runs. Thus it can be explained that laypeople's queries are not commonly included medical terms or only contain terms painting their health situations. In case of that the relevant document for

query expansion evaluated by CLEF eHealth is not related to the query, retrieval documents from searching for that query are not also in higher relevance assessments. Another reason for explanation of bad results is that query expansion is only a connection between a query and the relevant document at the top of the searching result of the query. As the result, new queries are too long to apply to LM.

Figure 2 shows the graph structure of the participants' performance of baseline run (run 1) in task 2 of 2015 CLEF eHealth. For the baseline run, it is observed that our experiment performed most of queries in under the best & median cases of all participated groups. But some queries, we reached out-performance more than other systems as queries: 13, 26, 29, and 65. A few queries are in lags such as 3, 32, 33, and 66 respectively.

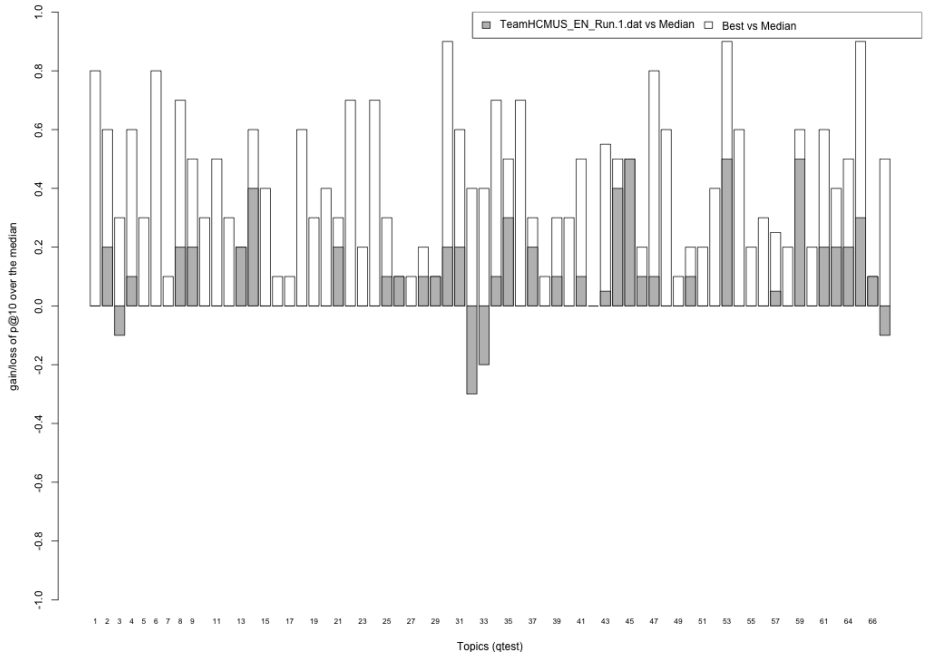

Fig. 2. Comparison graph of the baseline run (run 1) with other participants' systems, released by CLEF

Figure 3 shows the performance graph of the participants' run 2 in the task 2 of 2015 CLEF eHealth. For the run 2, it is observed that our experiment only out-performs more than other systems in queries 15, 16, and 25 respectively. On the other hand, our run 2 system lags in queries 5, 20, 30, 32, 34,

46, 51, 57, 61, and 65 respectively. Figure 3 points out that increasing value of smooth coefficient is not efficient.

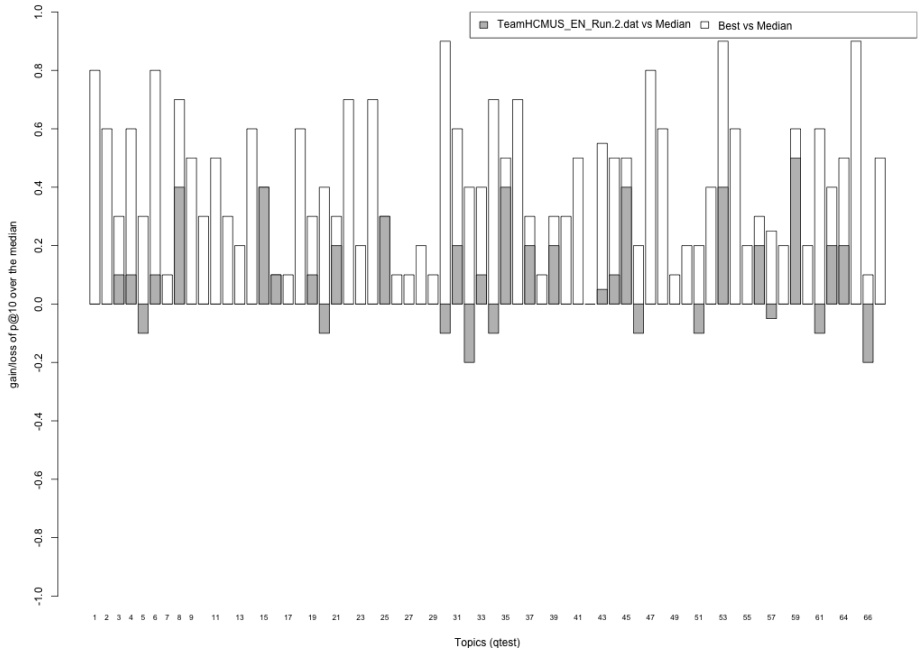

Fig. 3. Comparison graph of the run 2 with other participants' systems, released by CLEF

When applying some techniques as retrieval of medical concepts [2] (i.e. run 3, 4) and query expansion to do more experiments (i.e. run 5 to 8), we reached retrieval results that are not better than other participated groups' systems although there are still a few out-performed queries as indicated in Figure 4 (i.e. run 3) and Figure 5 (i.e. run 7). The reason of those can be explained that query expansion made new queries with large length and Index B only contains medical concepts. So treatments should be proposed and experimented on in future work.

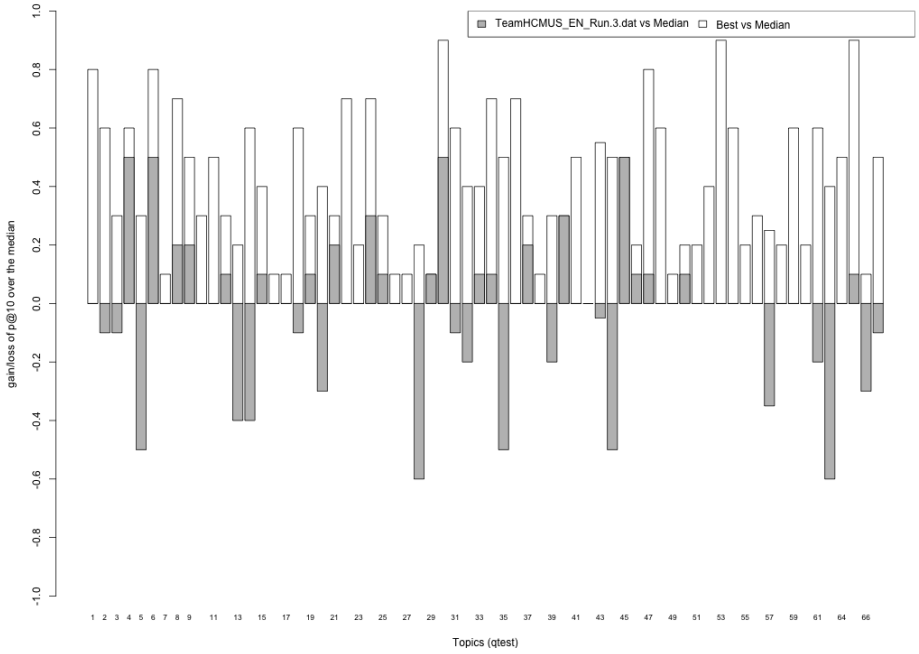

Fig. 4. Comparison graph of the run 3 with other participants' systems, released by CLEF

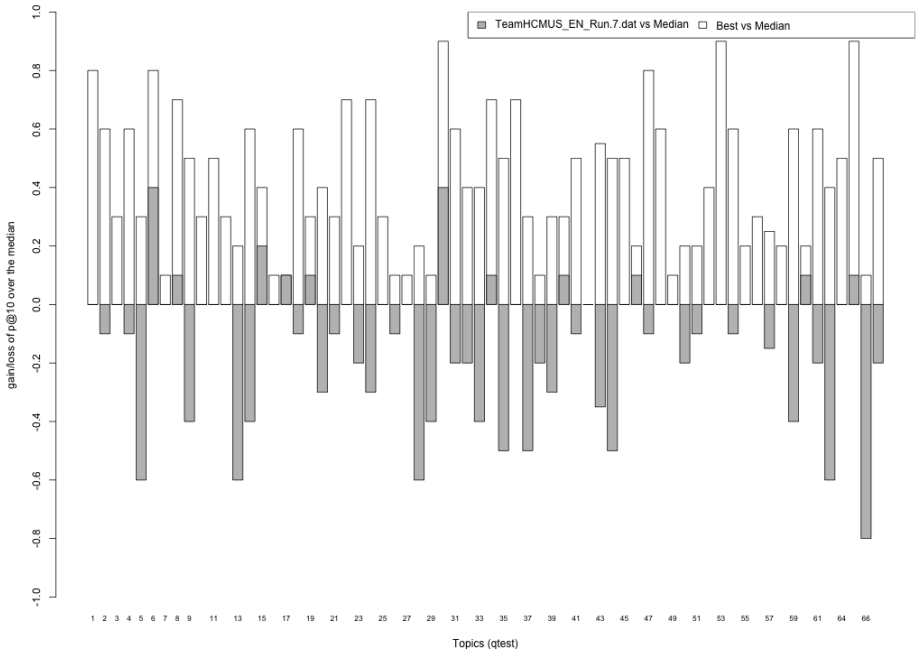

Fig. 5. Comparison graph of the run 7 with other participants' systems, released by CLEF

6.2 Readability-biased evaluation

For this year task, it's the first time CLEF eHealth has considered the factor of readability (understandability) within the evaluation of the submissions with assumption that readability is assessed independently of relevance assessments. To account for readability in the evaluation, we have computed an understandability biased measure, uRBP [14].

We used the ubire-v0.1.0 tool[11] to point out the values of RBP [5], and two versions of uRBP. The user persistence parameter p of RBP (and uRBP) was set to 0.8 [7]. Values of uRBP were computed by using user model 1 of [14] with threshold=2, i.e. documents with a understandability score of 0 or 1 where deemed unreadable and had P(U|k)=0, while documents with a understandability score of 2 or 3 where deemed readable and had P(U|k)=1. Values of uRBPgr were computed by mapping graded understandability scores to different probability values, in particular: readability of 0 was assigned P(U|k)=0, readability of 1 was assigned P(U|k)=0.4, readability of 2 was assigned P(U|k)=0.8, readability of 3 was assigned P(U|k)=1.

CLEF eHealth also notes that these readability-biased measures are still being experimented and observations that were made with the provided measures may not be conclusive.

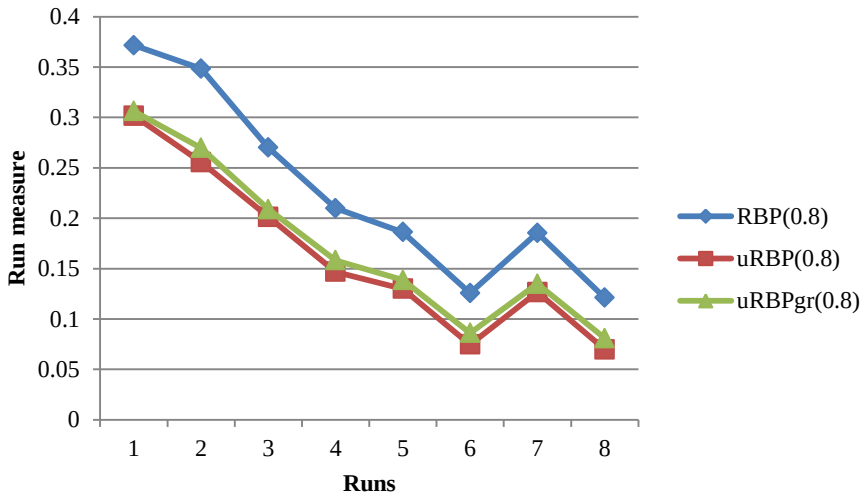

Fig. 6. Comparison graph of RBP, uRBP, uRBPgr values between the runs, released by CLEF

[11] https://github.com/ielab/ubire

Figure 6 indicates that RBP, uRBP, uRBPgr values have the same trend throughout runs. This is a reductive trend except a fluctuation at run 7. That the line of uRBP is under the uRBPgr line shows that mapping graded understandability scores to different probability values reaches a bit out-performance.

7 Conclusion

It's necessary to build an efficient retrieval system for searching laypeople's queries. However, this is not easy and still a challenge for participating groups in CLEF eHealth. This is partly because the features in laypeople's queries are not utterly medical terms or only terms of description of their health situations. Thus, searching of retrieval systems for laypeople's queries should be returned.

We carried out some experiments in applying our approach of retrieval of medical concepts [2] and query expansion techniques to establish a retrieval system. However, efficiency of the system is not still out-performance in general except few queries.

In future work, we will keep finding out advanced methods to refine corpora so that they only contain suitable features. Simultaneously, we will consider query expansion techniques, and experiment on various models of matching documents. In addition, we also do research on the results of other participating groups' systems to make them better.

References

1. M. Almasri J. Chevallet, C. Berrut: Exploiting Wikipedia Structure for Short Query Expansion in Cultural Heritage, CORIA (2014).
2. Nghia Huynh, Quoc Ho: TeamHCMUS: Analysis of Clinical Text, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages pp. 370–374 (2015).
3. K. Järvelin , J. Kekäläinen: Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems (TOIS) 20(4), pp. 422-446 (2002).
4. J. Leskovec, A. Rajaraman, J. D. Ullman: Mining of Massive Datasets. Cambridge University Press, chapter 1 (2011).
5. A. Moffat, J. Zobel: Rank-biased precision for measurement of retrieval effectiveness, ACM Transactions on Information Systems (TOIS), vol.27 no. 1, pp. 1–27 (2008).
6. E Noguera, F Llopis: Applying Query Expansion Techniques to Ad Hoc Monolingual tasks with the IR-n system, CEUR Workshop Proceedings, Vol-1173 (2007).
7. L. A. Park, Y. Zhang: On the distribution of user persistence for rank-biased precision, Proceedings of the 12th Australasian Document Computing Symposium (2007).

8. J. M. Ponte, W. B. Croft: A language modeling approach to information retrieval, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 275-281 (1998).
9. M. F. Porter: An algorithm for suffix stripping, Program, Vol. 14, no. 3, pp. 130-137 (1980).
10. H. Thakkar, G. Iyer, K. Shah, P. Majumder: Team IRLabDAIICT at ShARe/CLEF eHealth 2014 Task 3: User-centered Information Retrieval system for Clinical Documents, CEUR Workshop Proceedings, Vol-1180, pp. 248-259 (2014).
11. O. Thesprasith, C. Jaruskulchai: CSKU GPRF-QE for Medical Topic Web Retrieval, CEUR Workshop Proceedings, Vol-1180, pp. 260-268 (2007).
12. S. Verberne: A language-modelling approach to User-Centred Health Information Retrieval, CEUR Workshop Proceedings, Vol-1180, pp. 269-275 (2014).
13. C. Zhai and J. Lafferty: A study of smoothing methods for language models applied to ad hoc information retrieval, Proceedings of the ACM-SIGIR 2001, pp. 334-342 (2001).
14. G .Zuccon, B. Koopman: Integrating Understandability in the Evaluation of Consumer Health Search Engines, Proceedings of the SIGIR workshop on Medical Information Retrieval (2014).
15. Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Leif Hanlen, Aurélie Névéol, Cyril Grouin, João Palotti, Guido Zuccon. Overview of the CLEF eHealth Evaluation Lab 2015. CLEF 2015 - 6th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September 2015.
16. Palotti, João and Zuccon, Guido and Goeuriot, Lorraine and Kelly, Liadh and Hanbury, Allan and Jones, Gareth JF, and Lupu, Mihai and Pecina, Pavel. CLEF eHealth Evaluation Lab 2015, task 2: Retrieving Information about Medical Symptoms. CLEF 2015 Online Working Notes, CEUR-WS, 2015.