LABERINTO at ShARe/CLEF eHealth Evaluation Lab 2014

Juan Manuel Córdoba Malagón and Manuel Jesús Maña López

LABERINTO
Laboratorio de Recuperación de información y Minería de Texto y Datos
Universidad de Huelva
Carretera Palos de La Frontera s/n
21819 Palos de la Frontera (Huelva), Spain
juanmanuel.cordoba@gmail.com,manuel.mana@dti.uhu.es

Abstract. This paper describes the participation of LABERINTO team at the ShARe/CLEF eHealth Evaluation Lab 2014 task 3a. We perform four different experiments which consist of a baseline and three variants of the baseline model. The first was mandatory baseline system with only title and description in the query. Our baseline retrieval system used a Lucene Index scheme with traditional stopping and stemming, no external resources was used. We submitted three additional runs (without the discharge summaries), two from a Lucene-based system with MeSH query expansion and one of which made use of the National Library of Medicine's MetaMap tool to perform term boosting.

Keywords: Lucene, Solr, MetaMap, MeSH, query expansion

1 Introduction

ShARe/CLEF (Cross-Language Evaluation Forum) eHealth Evaluation Lab goal is to evaluate systems that support laypeople in searching for and understanding their health information [8]. It comprises three tasks: Visual-Interactive Search and Exploration of eHealth Data (Task 1), Information extraction from clinical text (Task 2) and User-centred health information retrieval (Task 3). Task 3 goal is to develop methods and resources for the evaluation of Information Retrieval (IR) from patients' perspective. Towards this, ShARe/CLEF eHealth Evaluation Lab 2014 task 3 is split into two parts: monolingual retrieval (task 3a) and multilingual retrieval (task 3b) [5]. In particular, LABERINTO team have focused on task 3a.

The LABERINTO group contributed 4 runs to this year's challenge. Our methods are based on Lucene retrieval engine. It's the first time that we are participating in the ShARe/CLEF eHealth Evaluation Lab and, for a first approximation, our baseline submission uses Lucene's default standard analyzer to process free-text title and description fields. The remaining submissions build upon this baseline approach. Specifically, we consider the contribution to retrieval effectiveness using boosting of Metamap identified terms and two versions of query expansion using MeSH.

This working notes are organised as follows. In Section 2 we describe some issues related to document preprocessing and indexing. Section 3 describes our approaches for task 3a. Section 4 lists the results of our work in comparison to median and best values obtained across all systems. Finally, we make conclusions and state the future work in Section 5.

2 Document collection preprocessing and indexing

The goal of the third task is to provide valuable and relevant documents to patients, so as to satisfy their health-related information needs. To evaluate systems that tackle this third task, the lab organizers provide potential patient queries and a document collection containing various health and biomedical documents for task participants to create their search system. As is common in evaluation of information retrieval (IR), the test collection consists of documents, queries, and corresponding relevance judgements [4]. Specifically, Task 3a uses an approximately one million medical documents made available by the EU-FP7 Khresmoi project[6][1] and a set of English general public queries that individuals may realistically pose[2].

This collection consists of web pages covering a broad range of health topics, targeted at both the general public and healthcare professionals. The crawled documents are provided in the dataset in their raw HTML (Hyper Text Markup Language) format along with their uniform resource locators (URL). In order to remove html tags ,raw webpages are preprocessed to extract main content by the Html parser Apache Tika. The Apache Tika toolkit detects and extracts metadata and structured text content from various documents using existing parser libraries including Html. Tika is a project of the Apache Software Foundation and was formerly a subproject of Apache Lucene.

After the data has been cleaned, we indexed all the documents. We used a very traditional IR system based on the Apache Lucene open-source toolkit that was essentially a successor to the system used by LABERINTO team for the 2011 TREC medical track [2]. Lucene is a powerful Java library that lets you easily add document retrieval to any application. In recent years, Lucene has become exceptionally popular and is now the most widely used information retrieval library.

Documents and fields are Lucene's fundamental units of indexing and searching. A document is Lucene's atomic unit of indexing and searching. It is a container that holds one or more fields, which in turn contain the "real" content. Each field has a name to identify it, a text or binary value, and a series of detailed options that describe what Lucene should do with the field value when you add the document to the index. To index our collection sources, we must first translate it into Lucene's documents and fields. Our indexing module takes every clean Html file from preprocessed collection into a single Lucene document.

[1] http://www.khresmoi.eu/
[2] http://clefehealth2014.dcu.ie/task-3

What we end up after running Lucene is a directory named index, which contains files used by Lucene to associate terms with documents. To accomplish this, a Lucene index was created with a specific analyzer model-dependent. An Analyzer takes a series of terms or tokens and creates the terms to be indexed. A unique kind of Lucene index has been used for all developed models, or in other words, all LABERINTO models for ShARe/CLEF eHealth Evaluation Lab 2014 share the same Lucene index.

3 Retrieval approaches

This section presents the different models developed for evaluation. Among the different test models developed, four have been selected for submission:

- Mandatory baseline run using only title and description fields[3].
- Model with Metamap boost terms[4].
- Baseline with Query expansion using MeSH[5].
- Baseline with query expansion using MeSH and adding narrative field[6].

All the proposed models no uses the discharge summaries. The differences between models are described in the following sections.

3.1 Baseline

This model has been designed to be the simplest approximation to the task. The Baseline model is based on the method of bag of words. In this model, topics text is represented as an unordered collection of words, disregarding grammar and even word order. Therefore, the model matches the words in the topic with the words contained in the index. The usefulness of the model is twofold: provides a basis results for comparing and, on the other hand, its code serves as a basis for implementing more complex models. In order to maintain the simplicity, the baseline model makes match the topics words only with the title and description fields.

To develop this model we used Lucene's default *StandardAnalyzer*. The analyzer takes only the text and provides a set of terms to be searched in the index. Our base analyzer discards stops words with little semantic value, such as "the", "a", "an","for",.... Cutting down on the number of terms indexed can save time and space in an index, but it can also limit accuracy.

[3] UHU_EN_Run1.dat run submission.
[4] UHU_EN_Run5.dat run submission.
[5] UHU_EN_Run6.dat run submission.
[6] UHU_EN_Run7.dat run submission.

3.2 Metamap terms Boost

Lucene provides the relevance level of matching documents based on the terms found. The higher the boost factor, the more relevant the term will be. Boosting allows you to control the relevance of a document by boosting its term[7].

In this model, we used medical main concepts identification by Metamap. Queries was built from the title and description fields. Next, UMLS concepts in queries are recognized using MetaMap [1].

Because our method don't need any special score mapping or disambiguation, Metamap's default options was used. Only medical term identification functionality was used. In other words, our method only need medical terms detection in order to know which topic part needs to be boosted. No overmatches are allowed and the highers score mappings are selected. For this UMLS detected mappings, a boost factor hits value is set. In particular, values from 1.25 to 2.25 (carried out with an interval of 0.25) were applied by an exploratory approach, centred mainly in the training topics and the ImageCLEFmed 2013 database and topics [3]. From this experimental test we selected a boost of 1.5 for submission. Since the set of training topics so small, we believed that no further tuning wasn't possible.

3.3 Query expansion with MeSH

Term expansion is one possible retrieval technique that can benefit from public accessibility of structured medical vocabularies. Applied at query-time usually it deals with the problem that real-world concepts are referred to using diferent terms. An information retrieval system can help users and also automatically refine their queries by exploiting the semantic relationships between terms [7].

MeSH (Medical Subject Headings) is a controlled vocabulary, produced and maintained by the U. S. National Library of Medicine [9]. There are currently over 26,000 descriptors or Main Headings and almost 180,000 alternative expressions (ENTRY TERMS), thus, MeSH offers many possibilities for expanding the query by MeSH tree structure and/or entry terms [10].

In this model, an open source implementation of SKOS-based[8] term expansion for Solr is used. For every term included in both title and description fields, term expansion through MeSH is performed with SKOS. In this approach, we expand the query terms with related terms from MeSH, duplicate related terms are removed. A default expansion terms weighting of 0.7 is used according to the previous work of Bernhard, Martins and Magalhães [7].

3.4 Query expansión with MeSH adding narrative field

As in the case mentioned above, this model use query expansion with MeSH with SKOS. The only difference lies in the use of the query expansion in addition of

[7] http://lucene.apache.org/core/2_9_4/queryparsersyntax.html
[8] https://github.com/behas/lucene-skos

narrative field terms. In this case, for every term included in title, description and narrative fields, term expansion through MeSH was performed and terms in narrative field were added.

It is worth mentioning at this point that narrative field hasn't an important relevance for query expansion. Major differences for information retrieval in this model comes from adding narrative field terms that MeSH concepts expansion determination.

In relation to this issue, table 1 shows a summary of MeSH concepts detected in topics for query expansion and some data for synonyms extracted from MeSH ontology in the performed expansion. For MeSH concepts, minimum, maximum and average number of MeSH concepts detected per topic has been collected (along with the standard deviation). In the same way, data from the synonyms entries used for expansion are gathered.

Table 1: Statistics data extracted from the query expansion performed.

	Min	Max	Average	Standard deviation
MeSH Concepts	1	3	1.45	0.88
Synonyms	0	25	3.23	4.24

4 Results

Two main metrics were taken account for Share/CLEF eHealth 2014 task 3: Precision at 10 (P@10) as primary measure, and Normalised Discounted Cumulative Gain at rank 10 (nDCG@10) as secondary measure. The Share/CLEF eHealth 2014 task 3 built result pools from participants submissions considering the top 10 documents ranked by baseline systems (run 1), and the two highest priority runs that used the discharge summaries (run 2 and 3) and the highest two priority runs that did not used the discharge summaries (run 5 and 6); thus runs 4 and 7 were not sampled to form the assessment pool.

Table 2: Results of the submitted runs to Share/CLEF eHealth 2014 task 3a.

Measure	RUN1	RUN5	RUN6	RUN7
P@10	0.8000	0.5860	0.5140	0.5100
nDCG@10	0.5530	0.5985	0.5163	0.5158
map	0.2624	0.3152	0.2588	0.3009

Table 2 shows the results of our submitted runs. In this table, we have taken account the two main metrics for 2014 edition and the mean average precision (MAP), a metric usually used in Information Retrieval, as a reference point.

First, we find that two of our submitted runs outperform the baseline model, one from Metamap boost model (in all metrics) and another one from query expansion model (taken MAP as reference). This shows that these approaches could be more effective than a simple baseline model. In contrast, the concept expansion approach based only in title and description fields decrease the retrieval performance in all metrics. We think that the improper query expansion settings may be the reason of this performance. Our query expansion models differ on fields use, RUN6 model uses the title and description fields, and RUN7 model uses the title, description and narrative fields to select terms to expand. Results show that refining the set of terms used for query expansion often prevents the query drift caused by blind expansion and yields substantial improvements in retrieval effectiveness. Although RUN7 improves map, and taking into account the reference values used at 2014 lab, our query expansion runs have proved that not all query expansion lead to improvements of retrieval.

By other hand, in RUN5, we consider positive the contribution of increase weights of query medical concepts for scoring a document. Empirical results show that considering boost medical concepts along with the original query concepts can improve retrieval effectiveness; which concepts to consider (with Metamap or other tool) and how to weight these is however a challenging issue.

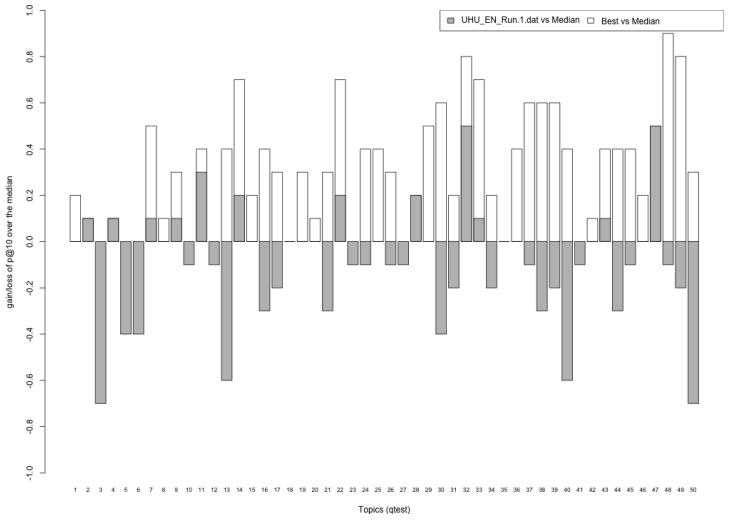
Plots comparing each of our runs against the median and best performance (p@10) across all systems submitted to CLEF for each query topic are shown in figure 1. In particular, for each query the height of a bar represents the gain/loss of our system and the best system (for that query) over the median system.

Per-topic comparison allows observe how performance varies in a important manner by model. Thus, we can see in Fig.1a that baseline has 12 queries perform better than the median, while 26 queries perform worse than the median, and other 12 queries perform in the median line. In fig.1b, Metamap based model has 14 queries perform better than the median, while 26 queries perform worse than the median, and other 10 queries perform in the median line. It means that comparing baseline and our best method, the UMLS concept-based method can do something better but heavily needs improvement to surpass baseline.

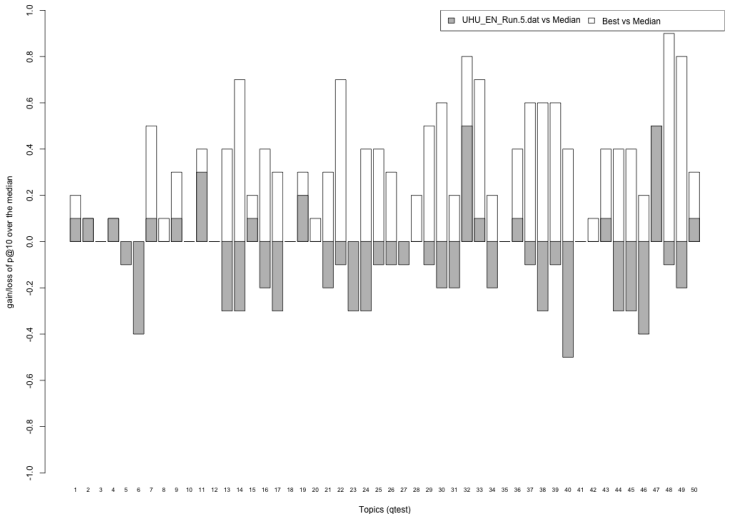
As regards query expansion models, we can see in Fig.1c that RUN6 has only 8 queries perform better than the median, while 32 queries perform worse than the median, and other 10 queries perform in the median line. In fig.1d, RUN7 has just 7 queries perform better than the median, while 33 queries perform worse than the median, and other 10 queries perform in the median line. Though one of the query expansion system has been better than baseline, per topic analysis shows a general poor performance for query expansion.

5 Conclusions and future work

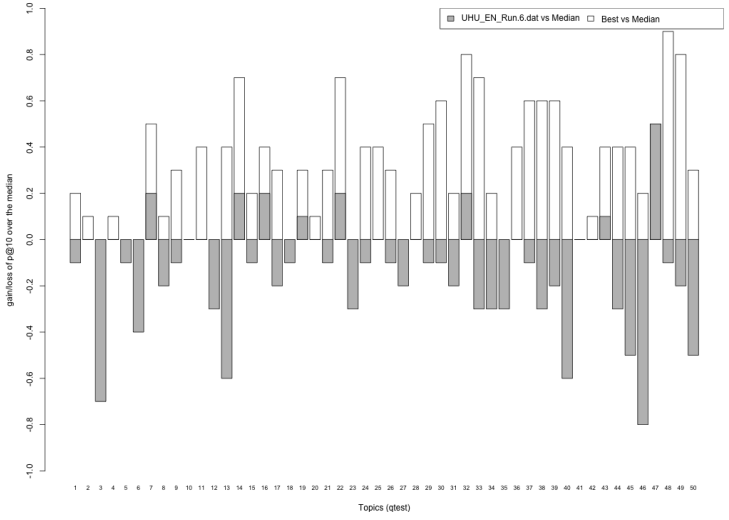
We have presented differents approachs to medical Information Retrieval from patients' perspective. Our models were based mainly on concept identification by Metamap and query expansion by MeSH. Both, the concept boosting and query expansion needs to be improved and refined. Some hints to improve, like



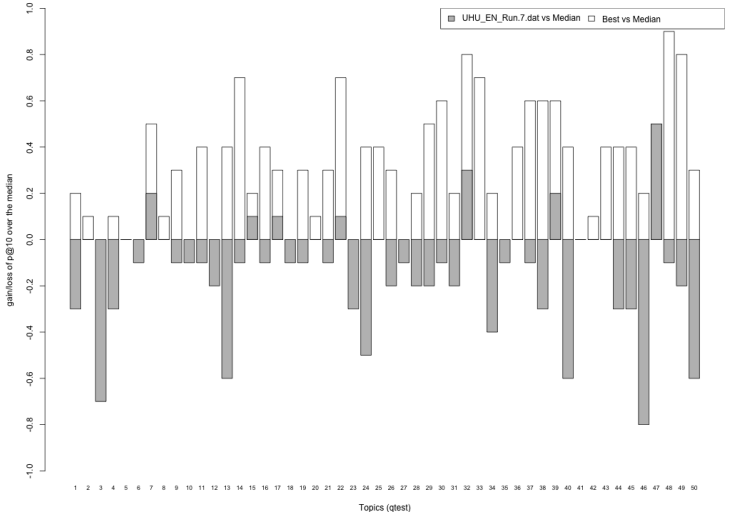
(a) Baseline model using only title and description fields.



(b) Model with Metamap boost terms.



(c) Baseline with Query expansion using MeSH.



(d) Baseline with query expansion using MeSH + narrative field.

Fig. 1: Per-topic comparison between submitted runs and the other systems (Best vs Median).

terms selection for expansion or tuning boosting parameters, has been exposed. Despite the inconspicuous results, we think that this first participation provides a platform for further development into medical concept based and query expansion retrieval systems for dealing with medical data from patients' perspective.

Acknowledgement

This work has been partially funded by the Andalusian Ministry of Economy, Innovation and Science (Bidamir project, TIC 07629).

References

1. Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.
2. Juan Manuel Córdoba, Manuel J Maña López, Noa P Cruz Díaz, Jacinto Mata Vázquez, Fernando Aparicio, Manuel de Buenaga Rodríguez, Daniel Glez-Peña, and Florentino Fdez-Riverola. Medical-miner at trec 2011 medical records track. In *TREC*, 2011.
3. A García Seco de Herrera, Jayashree Kalpathy-Cramer, D Demner Fushman, Sameer Antani, and Henning Müller. Overview of the imageclef 2013 medical tasks. *Working notes of CLEF*, 2013.
4. Lorraine Goeuriot, G Jones, Liadh Kelly, Johannes Leveling, Allan Hanbury, Henning Müller, Sanna Salanterä, Hanna Suominen, and Guido Zuccon. 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.
5. Lorraine Goeuriot, Liadh Kelly, Wei Li, Joao Palotti, Pavel Pecina, Guido Zuccon, Allan Hanbury, Gareth Jones, and Henning Mueller. Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval. In *Proceedings of CLEF 2014*, 2014.
6. A Hanbury and H Müller. Khresmoi–multimodal multilingual medical information search. *MIE Village of the Future*, 2012.
7. Bernhard Haslhofer, Flávio Martins, and João Magalhães. Using skos vocabularies for improving web search. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1253–1258. International World Wide Web Conferences Steering Committee, 2013.
8. Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Tobias Schrek, Gondy Leroy, Danielle L. Mowery, Sumithra Velupillai, Wendy W. Chapman, David Martinez, Guido Zuccon, and Joao Palotti. Overview of the share/clef ehealth evaluation lab 2014. In *Proceedings of CLEF 2014*, Lecture Notes in Computer Science (LNCS). Springer, 2014.
9. Henry J Lowe and G Octo Barnett. Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *Jama*, 271(14):1103–1108, 1994.
10. Jacinto Mata, Mariano Crespo, and Manuel J Maña. Laberinto at imageclef 2011 medical image retrieval task. *Working Notes of CLEF*, 2011, 2011.