Miracl at Clef 2018 : Consumer Health Search Task

Siwar ZAYANI, Nesrine KSENTINI, Mohamed TMAR, and Faiez GARGOURI
zayani.siouar@gmail.com,
ksentini.nesrine@ieee.org,
mohamed.tmar@isimsf.rnu.tn,
faiez.gargouri@isims.usf.tn

MIRACL Laboratory, City ons Sfax, University of Sfax, B.P.3023 Sfax TUNISIA

Abstract. This paper presents our participation in Consumer Health Search Task at the CLEFeHealth 2018 which is a continuation of the previous CLEF eHealth information retrieval (IR) tasks that ran in the period between 2013 and 2017.
This task focuses on improving access to medical information on the web. We have submitted four runs; two baseline runs with different weighting models and using no additional information or external resources. The two other runs present obtained results of our proposed approach which use the MeSH ontology, to perform query expansion with different ways by scope notes and by related terms.

Keywords: eHealth information retrieval · Mesh ontology · query expansion.

1 Introduction

Nowadays, medical information on the web grow with an etonnant and noticeable way. Indeed, medical content is becoming easily available electronically in a variety of forms ranging from patient records, scientific publications and health-related websites to medical-related topics.
Following the medical information overloaded today, it is increasingly difficult to retrieve and digest valid and relevant information to make health-centered decisions. In fact, clinicians and policy-makers need to easily retrieve, and make sense of medical content to support their decision making.
Information retrieval (IR) systems have been commonly used as a means to access health information available online in order to meet user's needs.
However, the reliability and the quality of returned results varies greatly between the diffrent information retrieval systems. Some systems tries to find high recall or coverage, that is, finding all relevant information for a user query, some others seek to obtain a high precision. Furthermore, web users in the health domain also experience difficulties in expressing their information needs as queries.
CLEF (Cross-Language Evaluation Forum) eHealth aims to bring together researchers working on related information access topics and provide them with

datasets to work with and validate the outcomes. This, the sixth year of the evaluation lab, offers the following three tasks [1]

- Task 1: Multilingual Information Extraction
- Task 2: Technologically Assisted Reviews in Empirical Medicine
- Task 3: Patient-centred information retrieval

The goal of the CLEF eHealth Evaluation Lab is to evaluate systems that support patients in searching for and understanding their health information.
In our case, our team MIRACL participated in the task 3 [2] in order to evaluate our own retrieval system.

2 Main objectives of experiments

Retrieving for health data and advice is an important task performed by individuals on the web. Thus, most of search engine users in recent years have conducted a web search for information about a specific disease or health problem.
The growing importance of health IR has provided the motivation for a number of evaluation campaigns focusing on health information. For example, the TREC (Text REtrieval Conference) [1] and the CLEF [2] which are present an international campaigns for assessment in a competitive context in order to evaluate several research systems of the various participants.
Our goal to participate this year to task 3 ("Patient-centred information retrieval") is like in the past years, to evaluate the effectiveness of our proposed information retrieval system to search health content on the web [3, 4].

3 Retrieval approaches used

This section presents the different search approaches developed for evaluation.
We have submitted 4 runs:

* Two baseline runs
* Two runs with automatic query expansion using MeSH ontology with different ways.

3.1 Baseline runs

For comparison, we created our own baseline experiments by implementing two information retrieval baselines with $TF.IDF$ and $Okapi\ BM25$ [5] methods.

[1] https://trec.nist.gov/
[2] http://www.clef-initiative.eu/

$TF.IDF$ is a weighting method often used in information retrieval and especially in text mining. This statistical measure makes it possible to evaluate the importance of a term contained in a document. The weight increases proportionally to the number of occurrences of the term in the document (TF).
IDF represent the inverse document frequency is a measure of the importance of the term throughout the collection. In the TF-IDF scheme, it aims to give greater weight to the less frequent terms considered more discriminating. It consists in calculating the logarithm (in base 10) of the inverse of the proportion of documents of the collection which contain the term (see equation 1).

$$idf_{term_i} = \frac{log|D|}{|d_j : term_i \in d_j|} \tag{1}$$

Where $|D|$: total number of documents in the collection.
$|dj : termidj|$: number of documents where the termi appears.

$OkapiBM25$ is a weighting method used in information retrieval. It is an application of the probabilistic model of relevance. The method is more simply called $BM25$, the term "$Okapi$" referring to the name of the research system of the University of London where it was initially implemented.
With each method, we calculate similarity (scores) between the user's query and documents in the collection and ranked the documents according to their scores in descending order. The top 1000 documents with the highest scores were returned as relevant documents for each query.
Each ligne in the result files (runs) contains the following fields :

$$\boxed{qid\ Q0\ docno\ rank\ score\ tag}$$

where:
qid: is the query number
$Q0$: is the literal Q0
docno: is the id of a document returned by our retrieval system for *qid*
rank: is the rank of this response for this *qid*
score: is a system-generated indication of the quality of the response
tag: is the identifier for our system for example $MIRACL$

3.2 Runs with automatic query expansion

In this sub-section, we describe our proposed retrieval approach used in submitted runs and based on query expansion. The idea is to use an external resource to ameliorate user's query and to automatically expand the original query without any user interaction.
As the provided collection of documents is medical, we proceed to use a domain ontology; the MeSH ontology.

MeSH (Medical Subject Headings) is a controlled vocabulary, produced and maintained by the U. S. National Library of Medicine [6]. There are currently over 26 million descriptors or Main Headings and almost 180,000 alternative expressions (ENTRY TERMS)[7].
Another definition of MeSH provided by [8], MeSH thesaurus is a controlled Vocabulary used for indexing, cataloging, identifying and searching biomedical database. MeSH thesaurus contains approximately 26 million terms and is updated time-to-time to reflect changes in medical terminology.
MeSH has a hierarchical structure with a set of terms and descriptors [8]; naming that allows various levels of searching. It will allow retrieving the document where the same concept is explained with different terminology.
The Hierarchy and the MeSH Structure are illustrated by the figure 1.

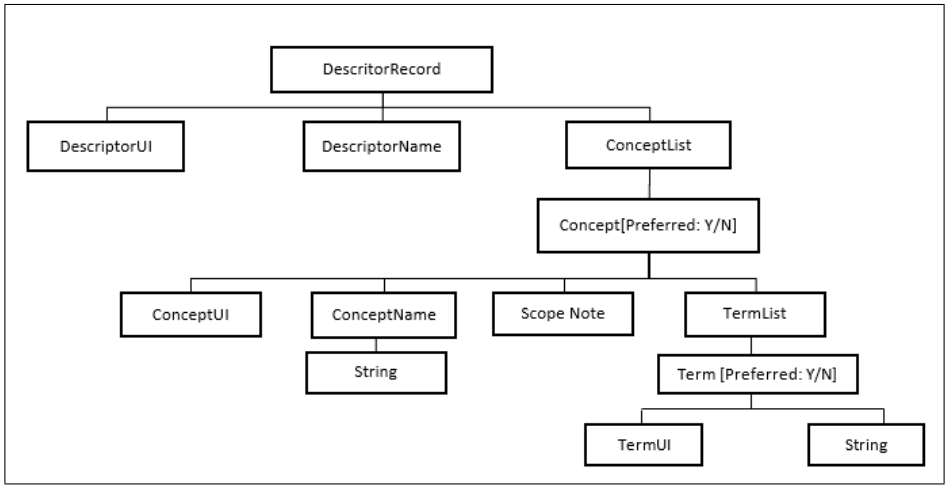

Fig. 1. MeSH ontolgy structure

Each MeSH record consists of one or more Concepts, and each Concept consists in one or more synonymous terms and Scope Note (i.e., a text description of the term).
Each of the subordinate concepts also will have a preferred term, as well as a labeled (e.g. narrower) relationship to the preferred concept. Terms meaning the same will be grouped in the same concept.
In figures (2,3), we illustrate two real examples of descriptors records. This Descriptors record consists of two Concepts and five terms. Each record has a Preferred Concept and each Concept has a Preferred Term, which is also said to be the name of the Concept.

In our case, we focus to use MeSH ontology to expand automatically user's queries, with two different methods:

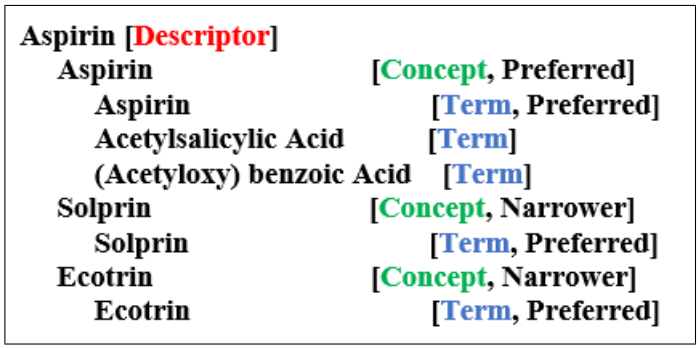

Fig. 2. Example 1 of descriptor record [9]

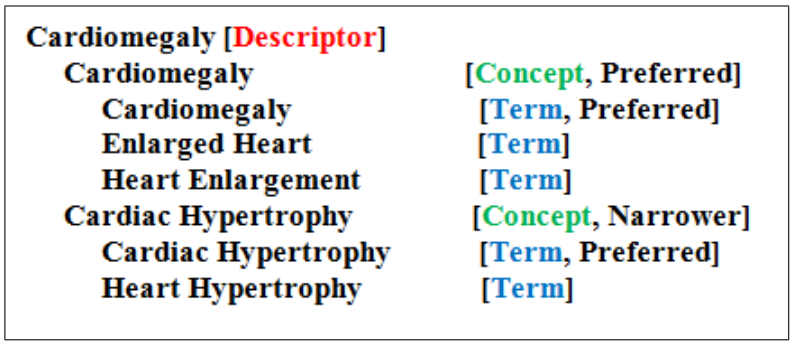

Fig. 3. Example 2 of descriptor record [10]

Query expansion method based on scope notes: This method is based on concepts extracted from the MeSH ontology.
Indeed, for each term in the initial query, if it's a MeSH concept, we add its scope notes which represent the medical definition of this term (by adding only the key terms).

Query expansion method based on related terms: This method is based on selecting terms semantically related to all terms in the initial query.
We have 2 cases when adding related terms:

- If a term in the initial query is a MeSH concept, then we add the list of synonymous terms related to this concept.
- If a term in the original query is a MeSH term, then we add their parent concepts.

4 Resources employed

4.1 Datasets

The document collection used in CLEFeHealth 2018 is composed of web pages acquired from the CommonCrawl.
An initial list of websites was identified for acquisition. The list was built by submitting the CLEF 2018 queries to the Microsoft Bing Apis (through the Azure Cognitive Services) repeatedly over a period of a few weeks by the CLEF team, and acquiring the URLs of the retrieved results. The list was further increased by including a number of known reliable health websites and other known unreliable health websites [1, 2].
The structure of the provided collection is as follows: the corpus is divided into folder by domain name. Each folder contains files which each one corresponds to a webpage from the domain. The document *id* for each webpage that is used in the collection (e.g. for the qrels) is the filename.
The full collection, named clefehealth2018, occupies about $480GB$ of space, uncompressed. With this gigantic base, the CLEF team made available to the participants a prepared index with different resources (ElasticSearch index, Indri index, and Terrier index).
Since we use the terrier platform to implement our information retrieval [11],[3], we exploit the provided terrier index to retrieve pertinent documents which have as size around $42GB$ compressed. This platform developed at the School of Computing Science at University of Glasgow, is efficient, effective and flexible open source search engine written in Java, easily deployable on large-scale collections of documents.
Indeed, terrier implements state-of-the-art of indexing functionalities in first step like tokenization, removing stop words, stemmatisation and storage of information with special structure called inverted file. In second step, it implements retrieval functionalities such as information retrieval models (Boolean, Tf-Idf, BM25) [11]. It is an open source, and a comprehensive and transparent platform for research and experimentation in text retrieval.

4.2 Queries

The query set for $CLEFeHealth2018$ consists of 50 queries issued by the general public to the HON and TRIP search services [12].
Queries are formatted one per line in the tab-separated query file, with the first string present the query *id*, and the second string present the query text (see figure 4). In Ad-hoc Search task which we have participated, we should use only the $< en > \cdots < /en >$ part of the query file.

5 Conclusion and future works

In our third participation at the CLEFeHealth competition, we try to evaluate our new proposed query expansion method based this time on an external re-

```
<queries>
  <query>
    <id> 151001 </id>
    <en> anemia diet therapy </en>
    <fr> anémie alimentation thérapie </fr>
    <de> anämie diet therapie </de>
    <cz> anémie dietoterapie </cz>
  </query>
  ¦
</queries>
```

Fig. 4. Query example

source; the MeSH ontology.
The aim of our participation is to test our research system with a new large collection of medical documents and to obtain competitive results with other participant teams.
For future work, we will try to combine this proposed method with our previous query expansion methods [13–16] described in previous participation [3, 4].

References

1. Suominen, H and Kelly, L and Goeuriot, L and Kanoulas, E and Azzopardi, L and Spijker, R and Li, D and Nvol, A and Ramadier, L and Robert, A and Zuccon, G and Palotti, J and Jimmy. Overview of the CLEF eHealth Evaluation Lab 2018. CLEF 2018 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September 2018.
2. Jimmy and Zuccon, G and Palotti, J. Overview of the CLEF 2018 Consumer Health Search Task. CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, September 2018.
3. Ksentini, N., Tmar, M., and Gargouri, F. Miracl at Clef 2014: eHealth Information Retrieval Task. In CLEF (Working Notes) (pp. 203-209). (2014).
4. Ksentini, N., Tmar, M., Boughanem, M., and Gargouri, F. Miracl at Clef 2015: User-Centred Health Information Retrieval Task. In CLEF (Working Notes).(2015).
5. Robertson, S. E., and Sprck Jones, K. Simple, proven approaches to text retrieval (No. UCAM-CL-TR-356). University of Cambridge, Computer Laboratory. (1994).
6. Lowe, H. J., and Barnett, G. O. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. Jama, 271(14), 1103-1108. (1994).

7. Mata, J., Crespo, M., and Maa, M. J. LABERINTO at ImageCLEF 2011 medical image retrieval task. Working notes of CLEF, (2011).
8. Rivas, A. R., Iglesias, E. L., and Borrajo, L. Study of query expansion techniques and their application in the biomedical information retrieval. The Scientific World Journal, (2014).
9. $https://www.nlm.nih.gov/pubs/techbull/ma00/ma00_mesh.html$
10. $https://www.nlm.nih.gov/mesh/concept_structure.html$
11. I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In Proceedings of the OSIR Workshop, pages 1825. Citeseer, (2006).
12. Goeuriot, L and Hanbury, A and Hegarty, B and Hodmon, J and Kelly, L and Kriewel, S and Lupu, M and Markonis, D and Pecina, P and Schneller, P (2014) D7.3 Meta-analysis of the second phase of empirical and user-centered evaluations. Public Technical Report, Khresmoi Project, August 2014.
13. Ksentini, N., Tmar, M., and Gargouri, F. Detection of Semantic Relationships between Terms with a New Statistical Method. In WEBIST (2) (pp. 340-343). (2014).
14. Ksentini, N., Tmar, M., and Gargouri, F. Controlled automatic query expansion based on a new method arisen in machine learning for detection of semantic relationships between terms. In Intelligent Systems Design and Applications (ISDA), 2015 15th International Conference on (pp. 134-139). IEEE. (2015, December).
15. Ksentini, N., Tmar, M., and Gargouri, F. Towards Automatic Improvement of Patient Queries in Health Retrieval Systems. Applied Medical Informatics, 38(2), 73-80. (2016).
16. Ksentini, N., Tmar, M., and Gargouri, F. The Impact of Term Statistical Relationships on Rocchios Model Parameters For Pseudo Relevance Feedback. International Journal of Computer Information Systems and Industrial Management Applications, 8, 135-44. (2016).