Miracl at Clef 2015 : User-Centred Health Information Retrieval Task

Nesrine KSENTINI[1], Mohamed TMAR[1], Mohand BOUGHANEM[2], and Faïez GARGOURI[1]

1 MIRACL Laboratory, City ons Sfax, University of Sfax, B.P.3023 Sfax TUNISIA
2 IRIT Laboratory, University of Toulouse, France
ksentini.nesrine@ieee.org,
mohamed.tmar@isimsf.rnu.tn,
bougha@irit.fr,
faiez.gargouri@isimsf.rnu.tn,

Abstract. This paper presents our second participation in user-centred health information retrieval task at the CLEFeHealth 2015. This task has as objective to retrieve pertinent documents that answer circumlocutory queries posed by users when they are faced with symptoms and signs of a disease.

We have submitted five runs, the first is the baseline system and the other runs use the pseudo relevance feedback technique to expand the original query in different ways.

This year, result pools are built from receiving submissions for all participants considering the top 10 documents ranked by every three highest priority runs (1, 2, 3). The obtained results are motivating, but we can be improved with P@10=0.3212 for the baseline system and $p@10$=0.2939 for the best P@10 in the other runs.

Keywords: information retrieval, least square method, medical documents, query expansion, blind feedback, circumlocutory queries.

1 Introduction

With the proliferation of documents on the web, searching pertinent documents to user's need becomes a difficult task, especially if the queries are not well expressed.

This process of searching documents in a corpus in order to meet a need is called Information Retrieval (IR). Many information retrieval systems are developed in order to provide all necessary functions to seek relevant information.

Evaluating these systems becomes important to know the quality of returned results. For that, International campaigns for assessment have intervened in a competitive context, such as TREC [1] and CLEF, in order to evaluate several research systems of the various participants in these competitions.

This paper presents our participation at clef ehealth 2015: User-Centred Health

---
[1] http://trec.nist.gov/

Information Retrieval task [8].

The goal of this task is to evaluate the effectiveness of information retrieval systems when searching health content on the web. The objective is to promote development of search engines and research adapted to health information.

This task is the further of previous CLEF eHealth Task 3 that held in 2013 and 2014, and adopts the TREC-style evaluation process, with a collection of documents and queries shared between participants [3, 4].

In this year's, we study queries which are different from queries of the previous years. In fact, Queries are expressed by not medical expert people whose are confronted with a symptom or a sign and try to find more about the disease they may have [10]. For example, non experts may use queries when confronted with signs of jaundice, like "white part of eye turned green" to search for information allowing them to diagnose themselves and to understand their health conditions. These queries are usually long with ambiguous words used in place of the actual name of the disease. Recent research has shown that these queries are used by health consumers, and that current web search engines fail to effectively support these queries [11].

In our participation, we used Vector Space Model that allows to compute a degree of similarity between documents and queries; also it returns ranking documents according to their relevance. We submitted five runs, the first was the baseline system and the others presents obtained results using query expansion techniques based on semantic relationships between terms.

2 Material and Methods

2.1 Database

Document Collection:
The data set is composed of a crawl of about one million documents, which have been made available to CLEF eHealth task through the Khresmoi project [2, 8]. This collection comprises a set of web pages covering a wide range of health subjects, addressed to both for the public and experts in the health sector. Web pages in this corpus are mainly medical and health-related websites that have been certified by the Health on the Net, as well as other commonly used health and medical websites such as diagnosia, Drugbank, and Trip Answers. The crawled documents are provided in the dataset in their raw HTML (Hyper Text Markup Language) format along with their URL (Uniform Resource Locators).

Topics:
The topic set contains 67 circumlocutory queries that users may pose when faced with symptoms and signs of a medical condition [8]. These queries contain the following fields:

- Num: Number of query.
- Query: longer description given by the user.

2.2 Methods

Our work was to index and search the top-1000 relevant medical documents for each topic. We used for the first run a very traditional information retrieval (IR) system based on the terrier platform [7]. For other runs, we used the same platform with expanded queries.

This platform used by MIRACL team for the 2014 Clef Ehealth task [6] is an 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).

It is an open source, and a comprehensive and transparent platform for research and experimentation in text retrieval. It is developed at the School of Computing Science, University of Glasgow.

3 Retrieval Approaches

This section presents the different models developed for evaluation:

* Mantadory baseline run.
* Baseline runs with automatic query expansion with different manners of adding related terms to the original query.

3.1 Mandatory baseline run

This model has been designed to be the simplest model to the task. It is based on Vector Space Model (VSM) developed by Salton [9] that uses a vector representation for documents as well as queries, and allows computing a degree of similarity between documents and queries and returning ranking documents according to their relevance.

The most widely used numeric similarity measures to calculate the relevance of a document is the cosine of the angle between the vector of the query and the document vector.

3.2 Baseline runs with automatic query expansion

In this section, we present submitted runs based on query expansion and using the same model (VSM) employed to mandatory submitted run.

The idea is to use the blind relevance feedback technique; called also pseudo relevance feedback; to automatically expand the original query without any user interaction [1].

In this method based on local analysis , the top k documents returned by

the baseline system are assumed as pertinent (relevant). Then we calculate semantic relationships between terms of these selected documents as referring to the statistical least square method (LSM) [5].

Indeed, we calculate for each term (t_i) its relation with other terms in a linear way (see equation 1).

$$t_i \approx \alpha_1 t_1 + \alpha_2 t_2 + \cdots + \alpha_{i-1} t_{i-1} + \alpha_{i+1} t_{i+1} + \cdots + \alpha_n t_n + \epsilon \qquad (1)$$

Where α are the real values of the model and present degrees of relationships between terms and ϵ represents the minimum associated error of the relation. In fact, we added to the original query only terms which have α that are above a certain threshold. We have submitted four runs that differ by the number of chosen documents (k) and by the fixed threshold.

4 Results

For this first year, the Share/CLEF eHealth 2015 TASK 2 built result pools from receiving submissions for all participants considering the top 10 documents ranked by every baseline system (run 1), and the two highest priority runs (run 2 and 3). This was because, unlike in last year, all submissions greatly differed between each other, and thus lead to very large (and different) pools.

Accordingly, principal measures proposed by the Share/CLEF eHealth 2015 TASK 2 to evaluate systems are:

- precision at 10 (P@10).
- Normalised Discounted Cumulative Gain at rank 10 ($ndcg - cut - 10$).

Metrics such as MAP (Mean Average Precision) may not be reliable for evaluation due to the limited pool depth.

Table 1. Obtained results for submitted runs

Run	P@10	ndgc@10	k	α
1	0.3212	0.2787		
2	0.2424	0.1965	250	> 0.0
3	0.2515	0.1833	100	> 0.0
4	0.1894	0.1572	100	> 0.05
5	0.2939	0.2465	100	> 0.3

In table 1, we show obtained results of the five submitted runs.

In the baseline system (run 1), we have obtained 0.32 of P@10 and 0.27 of $ndgc$@10. These values decrease slightly in the four other runs that uses a pseudo relevance feedback technique to expand the original query.

In fact, in the second run, we proposed that the top 250 documents returned by the baseline system as relevant. Then we calculated relations between terms that

appear in these selected documents using the statistical method: least square [5]. At last, we added only terms which have positive values of α with at least one term in the query.

Although this method has given results that are slightly lower compared to results of the first run, it allows to improve results of some queries. Take as an example the topic (query) 9:

$< top >$
$< num > clef2015.test.9 < /num >$
$< query >$red itchy eyes $< /query >$
$< /top >$

we took the top 250 documents returned by the baseline system as relevant. This set form a list of 8685 terms. Then we calculated relations between terms in the query with this set of terms and we expanded the original query by adding terms which have positive values of α. we have obtained this new query:

$< top >$
$< num > clef2015.test.9 < /num >$
$< query >$

red itchy eye see conjunct allerg nose drop gei zilink zaditor eyelid keratoconjunct lastacaft visin allergi relief antihistamin dose spray allergen tablet pollen claritin hydrochlorid olopatadin levocabastin bottl naphazolin liquid benzalkonium rxnorm formtitl inflamm flaki stye eyeexamin bedinghaus hive ige swollen pink bruis bump tear eyelash hay springtim slit dermat prescript rash sensit pharmaceut cream zyrtec blister section ketotifen western amerisourc fcab scream dismiss mart wal fbab mite flare urticaria ahref erysipela vital health zon tnss ectoin bhcsite mast overreact diphenhydramin azelastin cetirizin seborrh thrush hrcolor fluticason iannelli eczemadermat chkd prd beauti skindiseas outgrow respivert msfphover idiopath ledum vitamin tagid zpack eyebright emolli walgreen zyr butterbur hcl botrulelrulerrul allergicchild injector paradis euphrasia pheniramin nit mrhd enclos kmart ayala disk dolgencorp dryl personalhealthzon dbdc spoon teeter allina morri hydrochloridetablet bactrban
$< /query >$
$< /top >$

Where the three first terms present the original query and the other terms present the added terms in their root form.

We notice that almost all added terms are in strong relationship with the terms of the initial query. For example terms like conjunct, allerg, eyelid, inflamm, pink, keratoconjunct, dermat present the conjunctivitis disease also known as pink eye. It is an inflammation of the conjunctiva (the outer layer of the eye and the inner surface of the eyelids) due to an allergic reaction.

Terms like olopatadin, eyebright, zaditor, azelastin, levocabastin present some medication names for this disease. Some other terms present forums, sites and library of medicine such as :

  - rxnorm: National Library of Medicine.

- bhcsite: Better Health Channel which provides health and medical information.
- mast: Medical Academy for Science and Technology

This expanded query achieves the best performance on $p@10$ measure in this run ($p@10 = 0.5$) compared to the first run ($p@10 = 0.1$), see figures (1 and 2). This same process is used in (run 3) but relations are calculated just for the top 100 documents.

In (runs 4,5), we calculated relations for the top 100 documents and we added only terms that have α above a certain threshold with all terms in the original query. In these latter runs, we tried to keep the context of the initial query to not have a query drift problem.
We notice that measures of P@10 et ndgc@10 in (run 5), that expand query with strong related terms ($\alpha > 0.3$), are increased compared to runs 2 and 3.
Some queries in this run have been improved compared to the baseline system based on $p@10$ measure. For example query number 5:

$< top >$
$< num > clef2015.test.5 < /num >$
$< query >$whistling noise and cough during sleeping + children$< /query >$
$< /top >$

In the first run, we obtained -0.1 of $p@10$ by against this value becomes positive 0.1 in (run 5) due to the new added terms to the initial query which are in strong relationship with initial query terms. Indeed, the new request is the following:

$< top >$
$< num > clef2015.test.5 < /num >$
$< query >$
whistl nois cough sleep children ear wheez maud depress sound plug scream vitalhealthzon deafness research allergen chkd decibel otiti himend istictac pertussi toclevel maincat audiogram mcug wim opic ufa
$< /query >$
$< /top >$

The plots below in figures (1,2,3,4,5) compares our runs against the median and best performance ($p@10$) across all systems submitted to Ehealth task for each query.
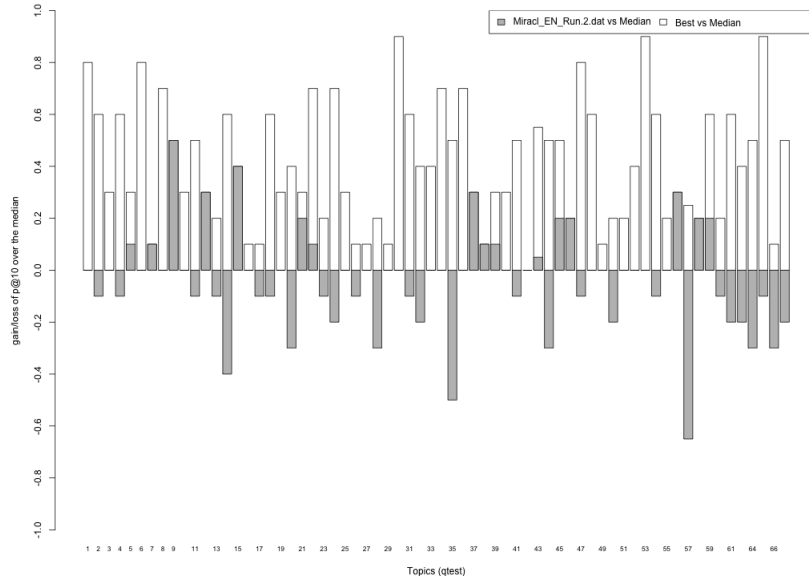In particular, for each query, the height of a bar represents the gain/loss of submitted system and the best system (for that query) over the median system. The height of a bar in then given by:
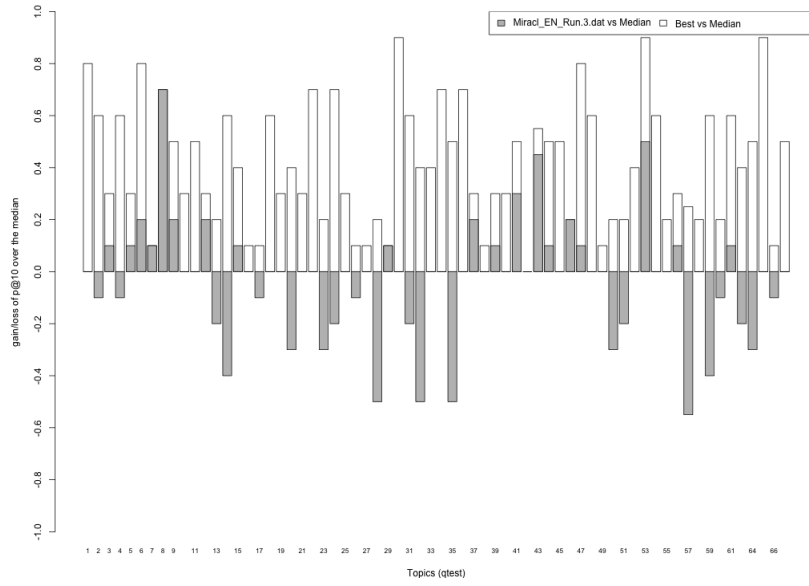
$$greybars : height(q) = our_p@10(q) - median_p@10(q)$$

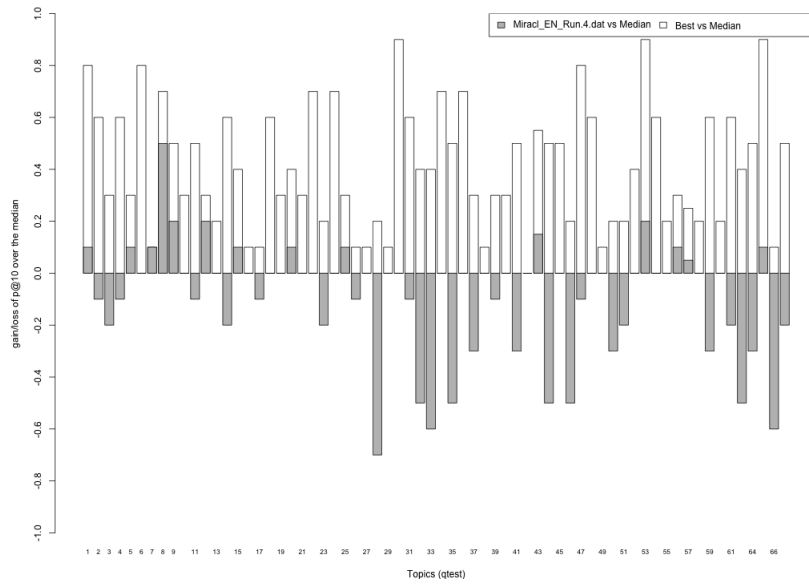$$whitebars : height(q) = best_p@10(q) - median_p@10(q)$$
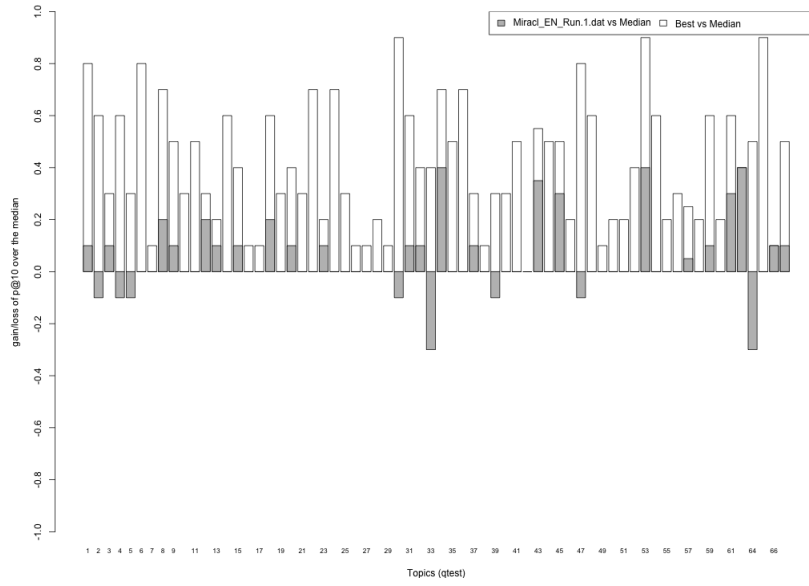
Fig. 1. Comparison between run 1 and the other systems against the median and best performance

Fig. 2. Comparison between run 2 and the other systems against the median and best performance
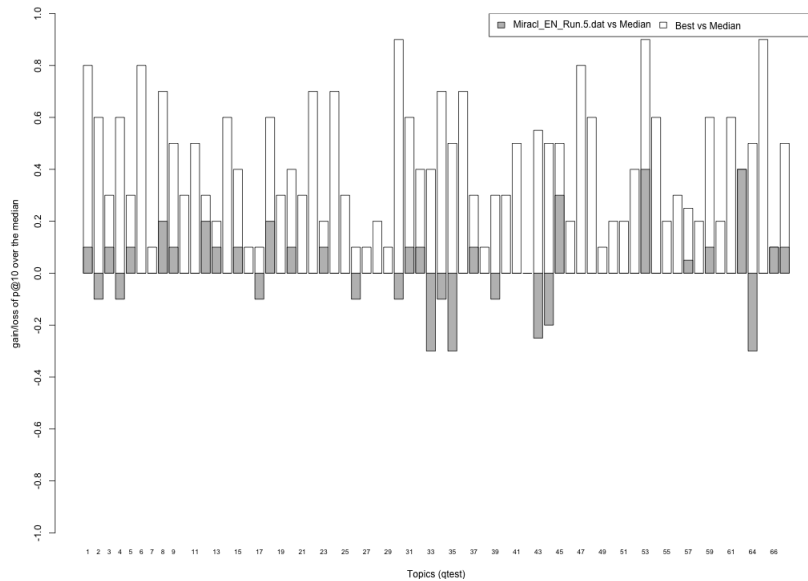
Fig. 3. Comparison between run 3 and the other systems against the median and best performance

Fig. 4. Comparison between run 4 and the other systems against the median and best performance

Fig. 5. Comparison between run 5 and the other systems against the median and best performance

We can see that (run 1) has one query that achieve the best performance, but in runs(2,3) we have ameliorated this number of achieved queries of best performance, we have obtained at respectively 8 and 3 (like query number 9 in run 2). In (run 5), we have obtained just one query that achieved this best, although it achieves 21 queries perform better than the median line, while 12 queries were worse than the median.

In comparison with (run 2) which have 16 queries perform better than the median line and 28 queries were worse than the median, we can conclude that proposed method to expand the original query taking into account the context of this latter car improve results for some queries but we need to improve general result of system.

5 Conclusion and future works

In our second participation in User-Centred Health Information Retrieval Task at the CLEF eHealth 2015 in order to evaluate our proposed expansion query method based on semantic relationships between terms defined with a new statistical method, we obtained motivating results when searching in a large collection of medical documents to answer circumlocutory queries that users may pose when faced with symptoms and signs of a medical condition.

For future work, we will try to improve these obtained results by modifying pa-

rameters of k and α and by using reduction dimensionality techniques to cope with problems of sparse and large dimension matrix.

References

1. C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1, 2012.
2. L. Goeuriot, A. Hanbury, G. J. Jones, L. Kelly, S. Kriewel, I. Martinez Rodriguez, H. Muller, and M. Tinte. Supporting collaborative improvement of resources in the khresmoi health information system. Springer, 2012.
3. 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. In *CLEF 2013 Online Working Notes*, volume 8138. CEUR-WS, 2013.
4. L. Goeuriot, L. Kelly, W. Li, J. Palotti, P. Pecina, G. Zuccon, A. Hanbury, G. J. Jones, and H. Mueller. Share/clef ehealth evaluation lab 2014, task 3: User-centred health information retrieval. In *CLEF 2014 Online Working Notes*. CEUR-WS, 2014.
5. N. Ksentini, M. Tmar, and F. Gargouri. Detection of semantic relationships between terms with a new statistical method. In *Proceedings of the 10th International Conference on Web Information Systems and Technologies*, pages 340–343, 2014.
6. N. Ksentini, M. Tmar, and F. Gargouri. Miracl at clef 2014: ehealth information retrieval task. *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*, 2014.
7. 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 18–25. Citeseer, 2006.
8. J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, A. Hanbury, G. J. Jones, M. Lupu, and P. Pecina. Clef ehealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In *CLEF 2015 Online Working Notes*. CEUR-WS, 2015.
9. G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
10. I. Stanton, S. Ieong, and N. Mishra. Circumlocution in diagnostic medical queries. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 133–142. ACM, 2014.
11. G. Zuccon, B. Koopman, and J. Palotti. Diagnose this if you can: On the effectiveness of search engines in finding medical self-diagnosis information. In *Advances in Information Retrieval (ECIR 2015)*, pages 562–567. Springer, 2015.