Health consumer-oriented information retrieval

Vincent CLAVEAU[a] and Thierry HAMON[b] and Sébastien LE MAGUER[a] and Natalia GRABAR[c,1]
[a] *IRISA - CNRS, Rennes, France*
[b] *LIMSI - CNRS, Orsay, France; Université Paris 13, Sorbonne Paris Cité, France*
[c] *STL UMR8163 CNRS, Université Lille 3, France*

Abstract. While patients can freely access their Electronic Health Records or online health information, they may not be able to correctly understand the content of these documents. One of the challenges is related to the difference between expert and non-expert languages. We propose to investigate this issue within the Information Retrieval field. The patient queries have to be associated with the corresponding expert documents, that provide trustworthy information. Our approach relies on a state-of-the-art IR system called Indri and on semantic resources. Different query expansion strategies are explored. Our system shows up to 0.6740 p@10 and up to 0.6793 NDCG@10.

Keywords. Information Retrieval, Natural Language Processing, Libraries, Digital, Consumer Health Information

Introduction

Patients can now freely access their Electronic Health Records (EHRs), although they may have difficulties with their understanding. This encourages patients in using Internet for searching health information [1-2] and modifies doctor-patient communication [3]. Hence, it becomes important that patients use information retrieval systems which are able to find trustworthy documents understandable by patients [4], and that the link between patient and medical doctors languages is possible. We propose a method that uses non-expert queries, such as those that can be submitted by patients after the reading of their EHRs, and that searches expert documents containing answers to patients' questions. Such documents provide trustworthy information usable by patients. More particularly, the objective of our work is to guarantee the semantic interoperability between the expert and non-expert languages. The existing work mainly addressed the aligning of expert and non-expert terms and expressions: Consumer Health Vocabulary (CHV) [5] or other experiments of the kind [6-8]. Currently, most of the CHV alignments are included in the UMLS [9]. Our experimental framework is the CLEF eHealth 2014's task 2 [10], for which queries are defined from real patient cases issued from clinical documents within the KRESMOI project [11]. We present first the material and the method used. We then present and discuss the results obtained, and conclude with some directions for future work.

1 Corresponding Author.

Material and Methods

The main material is the set of questions (5 in the training set, 50 in the test set) and of 976,249 documents (almost 200 M occurrences) as they are provided by the CLEF eHealth 2013 challenge. Besides, we use several types of semantic resources for query expansion: (1) the UMLS synonyms related to a given CUI; (2) 575 morpho-syntactic variants of terms from queries, acquired with FASTER [12], such as *cardiac disease/cardiac valve disease* (word insertion), *artery restenosis/arterial restenosis* (morphological derivation), *aorta coarctation/coarctation of the aorta* (permutation); (3) 1,114,959 pairs with lexical inclusions, like *muscle/muscle pain*, *cardiac disease/cardiac valve disease*; (4) 1,897 abbreviations for frequently used medical abbreviations. The acquisition of (2) and (3) resources is done on a part of the available dataset thanks to the use of TreeTagger POS-tagger [13] and YaTeA syntactic analyzer [14]. We also use a set of 627 English stopwords (*eg, for, under, amongst, indicate*) in order to reduce the noise that may be generated during the information search process.

Our method relies on the use of a state-of-the-art Information Retrieval (IR) system and on various strategies for query expansion with biomedical terms. The IR system is based on statistical language modelling as implemented by Indri [15]. This system has shown high performance in numerous IR tasks. We assume it may also offer interesting capabilities to express complex queries in biomedical context. Our method is composed of four steps:

1) *Pre-processing of documents and queries*. The pre-processing step is responsible for converting the documents in format processable by Indri. In particular, this is necessary for the indexing of documents and queries. Besides, the acquisition of semantic resources from corpora is also performed as part of this step;

2) *Setting the parameters*. The objective is to define the best parameters to be used with the documents to be processed and to maximize the evaluation measures. The setting of the method and its parameters is done with the document and query sets from 2013. For instance, we have set the smoothing and combination parameters lambda, and the corresponding binary relevance judgement using this 2013 dataset. We also tested various combinations of semantic resources and of the Indri parameters;

3) *Running the system*. On the basis of the previous step results, we applied different settings with the test set of queries: (a) running the Indri search engine with the best parameters estimated on the 2013 dataset, without use of semantic resources. This is our *baseline* setting; (b) running the best setting of Indri and the query expansion with the UMLS synonyms, which weight is set to 0.1. This setting will be referred to as *UMLS*; (c) running the best setting of Indri and the query expansion with other semantic resources (abbreviations and lexical inclusions). These two settings will be referred to as *expansion 1* and *expansion 2*;

4) *Evaluation of the system*. The evaluation of the results is done against the reference data with several evaluation metrics. The two major evaluation metrics assess the top ranked documents (at a cut-off of 5, 10 and up to 1000 top documents): i.e. P@5, P@10, etc. (precision for 5, 10, etc. top documents, respectively), NDCG@5, NDCG@10, etc. (normalized discounted cumulative gain for 5, 10, etc. top documents, respectively). In addition, the MAP (Mean Average Precision) is also used. Finally, the 1000 results are also evaluated with more recall-oriented measures Pprec and bpref.

Results

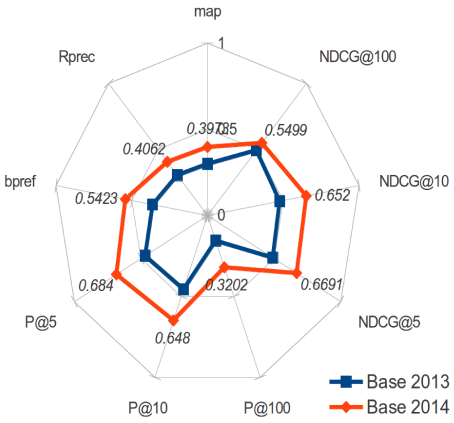
Table 1. Retrieval effectiveness of different settings.

	P@5	P@10	NDCG@5	NDCG@10	MAP
UMLS	0.6920	0.6740	0.6927	0.6793	0.4021
baseline	0.6980	0.6612	0.6691	0.6520	0.4054
expansion 1	0.6880	0.6600	0.6749	0.6590	0.3564
expansion 2	0.6720	0.6320	0.6615	0.6400	0.3453

In Table 1, we indicate the retrieval effectiveness of four settings. The results are sorted by the values of NDCG@10, which is the main evaluation metric. We can see that the best setting is when the UMLS [9] synonyms (including the Consumer Health Vocabulary [5]) are used: we get then 0.6793 points. The baseline is rated second in efficiency, with 0.6520 NDCG@10. The use of additional semantic resources (abbreviations, lexical inclusions) makes the results less efficient and slightly worse than the baseline. In 2014, 14 teams have participated in the challenge. The best NDCG@10 result is 0.7445, the worse result known is 0.0560 [10].

Discussion

Figure 1. Performance of the baseline setting, without application of the query expansion.



In Figure 1, we show a comparison of the baseline setting applied to the dataset from 2013 and 2014 challenges. We can see that the results obtained on the 2014 dataset are better. This fact is also acknowledged for the average results of all the participants [10]. One reason is that this increase of performances may be due to the fact that the topics are simpler in 2014, in the way that they correspond to main disorders, that are potentially more frequent and more searched in general. In Figure 1, we can observe that the performance is improved by several points in 2014 almost for all the metrics, although the setting has been done on dataset from 2013.

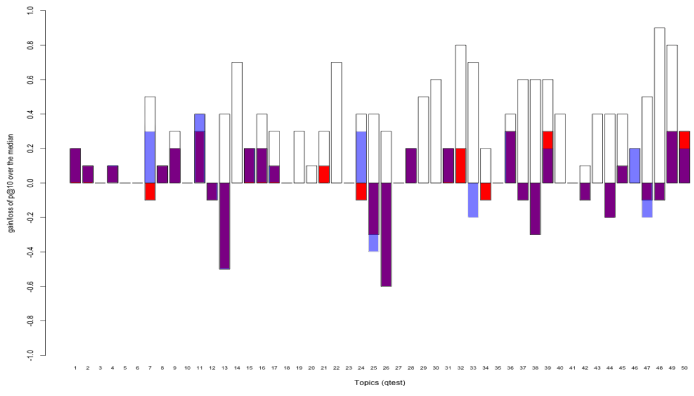
In our experimentations, we have put the accent on the acquisition and use of additional semantic resources expecting that these will improve the system performance. Nevertheless, it appears that such resources usually introduce noise in the results. Indeed, the distance between the semantics of a source word and its expansions

may be too loose. One example is expansion of *white* with common language synonyms, such as *caucasoid caucasian caucasians occidental*, that may be not suitable for the medical language. Such expansion may be even less correct within terms or phrases such as *white blood cell*. Concerning this issue, additional tests on the idiomatism of expressions can be performed before the expansion of their components with external resources is applied.

One type of additional resources is related to lexical inclusions, such as in *muscle/muscle pain* or *cardiac disease/cardiac valve disease*. In such pairs of terms, the short term conveys a more general meaning (it is hyperonym) while the long term has more specific meaning (it is hyponym) [16]. The underlying hypothesis was that non-expert users may use under-specified terms instead of more specific and precise terms, and that this kind of behaviour may provide the bridge between expert and non-expert users. Our results indicate that in the IR context, this kind of resources may not be suitable because there is a risk of over-generalisation. Our results indicate also that there is an instability of the impact of the resource across the queries: given semantic relations between terms from these resources may cause positive impact for some queries while they will have negative effect on other queries.

In Figure 2, we show a comparison between the impact of the baseline (in red) and of the UMLS setting (in blue) on the queries. In white are indicated the best results of the challenge obtained by any of the participating systems. We can see that for some queries (7, 11 , 24, 34, 46), the UMLS setting improves the precision, for other queries (21, 25, 32, 33, 39, 47, 50) it degrades the results. More interestingly, the expansion does not affect P@10 for the 38 remaining queries. Yet, for most of queries, some terms were actually added to the initial query, but they do not change the 10 first results.

Figure 2. Query expansion effect of setting UMLS (blue) vs. baseline (red) vs. best score (white) as gain or loss of P@10 compared with median result of the challenge.



Conclusion and Future Work

We presented an experience on information search in which expert and non-expert languages have to interact. The overall results are good, compared with other IR evaluation campaigns, with P@10 as high as 0.6740. Yet, our strategies to incorporate

external knowledge have yielded disappointing results. Indeed, the global benefits of the three query expansion strategies are limited, even though it appears as very interesting for particular queries. These mixed results are similar to the existing studies on query expansion for general language [17]. We plan to study further how to exploit the biomedical terminologies in IR tasks. A detailed analysis of the results may lead to better ways to choose which terms to consider in the queries, and which synonyms of these terms to add to the query. The incorporation of the terminological knowledge during the indexing step is also a promising avenue but raises computational issues.

Acknowledgments

This work has been partially funded by the Labex Comin'Labs platform.

References

[1] JA Diaz, RA Grith, JJ Ng, SE Reinert, PD Friedmann, AW Moulton. *Patients' use of the internet for medical information*. J Gen Intern Med **17(3),** 180-185 (2002).
[2] G Eysenbach, C Kohler, What is the prevalence of health-related searches on the world wide web? Qualitative and quantitative analysis of search engine queries on the internet. In: AMIA Annu Symp Proc. pp. 225-229 (2003).
[3] R Jucks, R Bromme. *Choice of words in doctor-patient communication: an analysis of health-related internet sites*. Health Commun **21**(3), 267-77 (2007).
[4] C Boyer, O Baujard, V Baujard, S Aurel, M Selby, RD Appel. *Health On the Net automated database of health and medical information*. Int J Med Inform **47(1-2),** 27-29 (1997).
[5] QT Zeng, T Tse. *Exploring and developing Consumer Health Vocabularies*. J Am Med Assoc **13,** 24-29 (2006).
[6] N Elhadad, K Sutaria. Mining a lexicon of technical terms and lay equivalents. In Proc BioNLP WS, 49-56 (2007).
[7] L Deléger, P Zweigenbaum, Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In Proc AMIA 2008, 146-150 (2008).
[8] N Grabar, T Hamon, Unsupervised method for the acquisition of general language paraphrases for medical compounds. 4th International Workshop on Computational Terminology Computerm (2014).
[9] UMLS Knowledge Sources Manual. National Library of Medicine, Bethesda, Maryland, USA. www.nlm.nih.gov/research/umls/ (2014).
[10] L Goeuriot, L Kelly, W Li, J Palotti, P Pecina, G Zuccon, A Hanbury, GJF Jones, H Mueller, *ShARe/CLEF eHealth evaluation lab 2014, task 3: user-centred health information retrieval.* In: Conference and Labs for the Evaluation Forum (CLEF 2014).
[11] http://www.khresmoi.eu
[12] C Jacquemin, Syntagmatic and paradigmatic representations of term variation. 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), 341-348 (1999).
[13] H Schmid, Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proc International Conference on New Methods in Language Processing, 44-49 (1994).
[14] S Aubin, T Hamon, Improving Term Extraction with Terminological Resources. In Proc FinTAL 2006, 380-387 (2006).
[15] T Strohman, D Metzler, H Turtle, WB Croft, Indri: a language-model based search engine for complex queries. In Proceedings of the International Conference on Intelligent Analysis (2005)
[16] G Kleiber, I Tamba. L'hyperonymie revisitée : inclusion et hiérarchie. Langages 98, 7-32 (1990).
[17] EM Voorhees, Query expansion using lexical-semantic relations. In: Proc of the 17th Annual International ACM SIGIR Conference on Research and Devel-opment in Information Retrieval. pp. 61-69