Query expansion using external resources for improving information retrieval in the biomedical domain

Khadim Dramé, Fleur Mougin, Gayo Diallo

ERIAS, INSERM U897, ISPED, University of Bordeaux
146 rue Leo Saignat 33076, Bordeaux

firstname.lastname@u-bordeaux.fr

Abstract. This paper presents the first participation of the ERIAS team in task 3 of the ShARe/CLEF eHealth Evaluation Lab 2014. The goal of this task is to evaluate the effectiveness of Information Retrieval systems to support patients in accessing easily relevant information. We propose a method which exploits external resources for improving information retrieval in the biomedical domain. The proposed approach is based on the well-known Vector Space Model and it uses two extensions of this model to enhance its performance. Specifically, the MeSH thesaurus is used for query expansion with different configurations. Experiments on a large collection of documents have shown that the use of these external resources can improve performance in medical in-formation retrieval.

Keywords: information retrieval, Lucene, Vector Space Model, n-gram extraction, query expansion, MeSH thesaurus.

1 Introduction

The role of an Information Retrieval (IR) system is to support users to access relevant information corresponding to their needs. In the medical domain, accessing useful information becomes increasingly important with the growing amount of available information. To tackle this challenging issue, different approaches have been proposed raising the challenge of assessing their performance. The ShARe/CLEF (Cross-Language Evaluation Forum) eHealth Lab [1] is an evaluation campaign in the biomedical domain which aims at easing patients (and their relatives) to understand their health-related information. Especially, the goal of the third task is to develop methods which facilitate the access to valuable information to patients regarding their health[2]. Indeed, the amount of biomedical information is growing rapidly with an abundant production of digital collection of documents. Accessing to useful information among this large amount of available data becomes a real challenge. To do so, controlled vocabularies, such as the Medical Subject Heading (MeSH) thesaurus, are widely used to improve the medical information retrieval (IR). In [3], the authors proposed the use of the MeSH thesaurus for expanding user queries. Terms associated

with a MeSH descriptor are considered as synonyms and used therefore to expand queries. They show, through their experiments, that queries expansion using MeSH thesaurus improves significantly the performances of IR in the medical domain. A similar queries expansion approach using the MeSH thesaurus is proposed in [4]. The hierarchical structure of the MeSH descriptors was exploited to improve the retrieval effectiveness. The popular MetaMap [5] tool is used to identify descriptors which are then expanded with their children. They show that queries expansion using the MeSH hierarchy yield significant improvements.

In this paper, we report the results achieved during our first participation in the Task 3 of ShARe/CLEF eHealth Evaluation Lab 2014. We have investigated the exploitation of controlled vocabularies and especially the MeSH (Medical Subject Heading) thesaurus in order to improve IR performance. First, we have explored the use of terms' synonyms to expand queries which contain vocabulary's entries. We have then investigated the use of semantic relations in MeSH for query expansion and their combination with synonyms' expansion. Evaluation on a large collection of over one million documents shows the use of these re-sources may improve performances in biomedical IR.

The remainder of the paper is organized as follows. First, we present our IR approach in Section 2. Then, experiments and results are described in Section 3 and discussed in Section 4. We conclude and give perspectives to this work in Section 5.

2 Methods

In this section, we describe the different techniques that we have used to address information access issue.

The proposed approach is based on the popular Vectorial Space Model (VSM) [6]. The principle of the VSM is to represent documents (and queries) by term vectors and then to use the cosine similarity measure to determine relevant documents according to a given query. In this work, instead of words, all unigrams and bigrams identified within documents are used as descriptors to index them. In a preprocessing phase, document contents are first tokenized and stop words are removed. Then, from these preprocessed documents, all the unigrams and bigrams are extracted. The latter, with their associated weights (term frequency – inverse document frequency), are used for indexing the documents. Our baseline system is based on this model using only the query title as the original query and does not use any external resources (Run 1). Indeed, each query can include a title, a description and additional fields.

Secondly, we have explored the usefulness of controlled vocabularies to enhance IR. Indeed, controlled vocabularies like MeSH thesaurus are widely used in medical IR [3][4][7]. However, since the use of solely vocabulary's terms for describing the queries was proven insufficient for this task last year [8], we additionally exploited the traditional keywords constituted by simple words. Thus, the n-grams model (our baseline) is combined with MeSH terms for query expansion. First, original queries

are extended using its synonymous terms, descendants (using hierarchical relations) and related MeSH terms (with the exploitation of the See Also relations in MeSH). For query terms not contained in the MeSH thesaurus, we have taken into account their synonyms retrieved in the Unified Medical Language System (UMLS) [7] Metathesaurus®. This corresponds to Run 5 of our method. In Run 6, we have investigated query expansion with only synonyms from the MeSH thesaurus or the UMLS, like in the previous Run 5. The difference is, in Run 6, only term synonyms are used to expand the queries while Run 5 combines term synonyms with related terms for query expansion. Finally, we have exploited query descriptions to estimate how they can improve the results. This last Run (7) is similar to the Run 6 but takes into account both title and query descriptions to build queries.

For identifying medical terms in queries, we have developed a simple method which focuses on the most specific terms; we consider terms that are entries of the vocabulary and that are not included in a longer entry in the query. So, for each query, only its medical terms not contained in other ones of the query are extracted in order to extend the query. After that, the original query is expanded by either terms' synonyms or other related terms or both. For example, for the query "Anoxic brain injury" (qtest2014.4), only the most specific term "Anoxic brain injury" (thus ignoring the more general terms brain, brain injury, injury) is extracted and its synonyms, such as "Anoxic Encephalopathy" and "Anoxic Brain Damage". The query "stroke and respiratory failure" (qtest2014.16) is constituted by the two specific terms "stroke" and "respiratory failure". The expanded query then includes synonyms like "cerebrovascular accident", "vascular accident, brain", "kidney failure", and "renal failure". For the query "aspiration pneumonia due to misplacement of dobhoff tube" (qtest2014.11), the terms "aspiration pneumonia" and "dobhoff tube" are extracted and the original query is extended by these terms' synonyms and/or related terms, such as "Acid Aspiration Syndrome" and "Respiratory Aspiration".

3 Experiments and Results

A document collection over one million documents (web pages from medical web sites) and 50 potential patient queries were provided by the organizers of ShARe/CLEF eHealth Evaluation Lab [2]. Beforehand, five queries with their corresponding relevance judgments were provided for the training. Each query includes a title, a description and other additional fields. The Lucene open-source search engine was used in our experiments for indexing and retrieving documents. For each query, we have used our different configurations to retrieve the top 1000 relevant documents The P@10 and NDCG@10 measures were primarily used to evaluate the submitted systems. Thus, for each system, the top ten documents retrieved were considered in the evaluation.

The results of our different runs are presented in Table 1. According to the evaluation measures, all the runs using external resources got better results than the baseline.

Therefore, using the controlled vocabularies improves the retrieval performance as expected. We also note that Run 6 which uses only query title as original query and terms' synonyms for query expansion obtained the best performance regarding P@10 and NDCG@10.

Compared with the global results in this task, our runs got reserved results according to queries. For some queries, they yielded good and even the best results. For example, our Run 5 got the best performance for the queries qtest2014.16 (P@10 =1), qtest2014.44 (P@10 = 0.9), qtest2014.47 (P@10 = 1) and greatly exceeded the median (+0.4, +0.4, +0.5 respectively). For these queries, almost all terms are entries of the MeSH thesaurus. In some cases, they just surpassed the median or got comparable performance. For example, for the queries qtest2014.7 (+0.3), qtest2014.11 (+0.3), Run 1 slightly surpassed the median and got the same P@10 as it for the queries qtest2014.3, qtest2014.12 and qtest2014.18. For other queries, our method got poorer performance. For example, for the queries qtest2014.4, all our runs yielded lower performance than the median: -0.6 for Run 5 and Run 6 and -0.7 for Run 7; for Run 1, the difference is less important (-0.1). This query is an entry of the vocabulary. So it is extended by synonyms and related terms of the term "Anoxic brain injury". Thus, documents containing the more general term "brain injury", which are considered as relevant, are ranked after the ones containing query term synonyms or related terms. The lowest P@10 compared with the median (-0.9) was yielded by Run 1 for query qtest2014.6. Here, the query is an entry of the vocabulary and therefore using keywords search, where the query is split into words, decreases the performances.

All our runs achieved a P@10 greater than or equal to 0.5 for half of the fifty queries. For five queries, Run 1 achieved a P@10 of 1 while Run 5 and Run 6 achieved this maximal precision in only four and two queries, respectively. But for five other queries, Run 1 had a P@10 equal to 0 while both Run 5 and Run 6 achieved this bad score for three queries. For queries qtest2014.37 and qtest2014.49, the top ten retrieved documents are irrelevant (P@10 = 0) for all our runs. The query qtest2014.37 is expressed with misspellings ("gynecolocigal" instead of "gynecological").

Table 1: Performance of our different systems in Task 3 of CLEF eHealth 2014

Runs	P@5	P@10	NDCG@5	NDCG@10
Run 1	0.5040	0.5080	0.4955	0.5023
Run 5	0.5440	0.5280	0.5470	0.5376
Run 6	0.5720	0.5460	0.5702	0.5574
Run 7	0.5960	0.5320	0.5905	0.5556

4 Discussion

The results presented in Table 1 show globally that query expansion using only terms' synonyms from MeSH improves the IR performance in our experiments (Run 6). Although the combination of synonyms and semantic relations for expanding queries enhance the results compared with the baseline, they did not increase the performance as expected. On the contrary, the use of related terms decreases the retrieval performance. In addition, while it got the best performance according to P@5 and NDCG@5, Run 7, which uses query description, provided worse results than Run 6 according to P@10 and NCDG@10. In other words, the use of query descriptions decreases the system performance if the top ten documents are considered but remains useful if we focus on the top 5 retrieved documents.

For several queries, Run 1 greatly exceeded the other runs even if it had lower performance overall: qtest2014.4, qtest2014.7, qtest2014.13, qtest2014.18, qtest2014.39 are some examples. The performance drop of the query expansion methods is generally caused by the negligence of some relevant terms not contained in the vocabulary (for example, "secondary" in "surviving rates after a secondary myocardial infarctus") or the decomposition of the query in simple terms ("dizziness" and "hypotension" in "dizziness and hypotension") which does not take into account the context. On the other hand, the results of Run 5 and Run 6 similarly vary but Run 6 often got a slightly better performance. This may be due to noise resulting from the use of related terms. For example the, from the query qtest2014.26, the term "gastrointestinal bleed" is extracted and expanded by its synonym "Gastrointestinal Hemorrhage" in Run 6 and additionally its related terms like "Peptic Ulcer Hemorrhage", "Endoscopic Hemostasis" and so on in Run 5. Consequently, the Run 6 got a higher P@10 of 0.9 while Run 5 yielded a P@10 of 0.1 which is even lower than the baseline result (P@10=0.4).

5 Conclusion and future work

In this work, we have investigated the usefulness of controlled vocabularies during query extension. We have proposed a vector space based method with several query expansion techniques and evaluated their impact in information retrieval effectiveness. Overall, using these controlled vocabularies improves the retrieval performance our experiments. However, the combination of terms' synonyms and semantic relations for query expansion surprisingly yield worse results than using only synonyms.

On the other hand, although our query expansion methods outperform our baseline, our results need to be improved compared with the global results of Task 3 participants. Overall, for our first participation to CLEF eHealth Lab, our method got promising results which we would like to improve.

In the future, to increase the performance of our retrieval method, we plan to explore other advanced retrieval models like the BM25 [9], which is used as the baseline in this challenge and its combination with controlled vocabularies. We also believe that the use of semantic similarity for query expansion may help to improve the information retrieval performance.

References

1. Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G., Mueller, H.: ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. Proceedings of CLEF 2014 (2014).
2. Kelly, L., Goeuriot, L., Suominen, H., Schrek, T., Leroy, G., Mowery, D.L., Velupillai, S., Chapman, W.W., Martinez, D., Zuccon, G., Palotti, J.: Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. Proceedings of CLEF 2014. Springer (2014).
3. Díaz-Galiano, M.C., Martín-Valdivia, M.T., Ureña-López, L.A.: Query expansion with a medical ontology to improve a multimodal information retrieval system. Comput. Biol. Med. 39, 396–403 (2009).
4. Azcárate, M.C., Vázquez, J.M., López, M.M.: Improving image retrieval effectiveness via query expansion using MeSH hierarchical structure. J. Am. Med. Inform. Assoc. amiajnl–2012–000943 (2012).
5. Aronson, A.R.: Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. (2001).
6. Salton, G., Wong, A., Yang, C.S.: A Vector Space Model for Automatic Indexing. Commun ACM. 18, 613–620 (1975).
7. Jimeno-Yepes, A.J., Plaza, L., Mork, J.G., Aronson, A.R., Díaz, A.: MeSH indexing based on automatically generated summaries. BMC Bioinformatics. 14, 208 (2013).
8. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G.K., Elhadad, N., Pradhan, S., South, B.R., Mowery, D., Jones, G.J.F., Leveling, J., Kelly, L., Goeuriot, L., Martínez, D., Zuccon, G.: Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Forner, P., Müller, H., Paredes, R., Rosso, P., and Stein, B. (eds.) Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings. pp. 212–231. Springer (2013).
9. Robertson, S.E., Jones, K.S.: Simple, Proven Approaches to Text Retrieval. (1997).