Query Reformulation for Clinical Decision Support Search

Luca Soldaini, Arman Cohan, Andrew Yates, Nazli Goharian, Ophir Frieder

Information Retrieval Lab
Computer Science Department
Georgetown University
{luca, arman, andrew, nazli, ophir}@ir.cs.georgetown.edu

Abstract

One of the tasks a Clinical Decision Support (CDS) system is designed to solve is retrieving the most relevant and actionable literature for a given medical case report. In this work, we present a query reformulation approach that addresses the unique formulation of case reports, making them suitable to be used on a general purpose search engine. Furthermore, we introduce five reranking algorithms designed to re-order a list of retrieved literature to better match the type of information needed for each case report.

1 Introduction

One of the tasks a Clinical Decision Support (CDS) system is designed to solve is retrieving the most relevant and informative literature for a given medical case report.

The unique formulation of case reports poses a serious challenge for general purpose search engines: case reports are much longer than traditional queries and present a narrative structure.

Our system, initially disclosed in [3], uses a combination of query expansion and query reduction techniques to address the unique formulation of medical case reports when used as queries. Because many of the medical and health-related terms in each case report have one or more domain-specific synonyms, we expand queries to ensure that the retrieval process would not suffer from the limited vocabulary coverage. Rather than relying on a thesaurus, we decided to use pseudo relevance feedback (PRF). The advantage of using such technique is that it is able to expand the case report not only by adding relevant medical terms, but also by incorporating many health-related expressions that are also used in related literature to the original query. That ensures that medical literature that is relevant to the original case report but uses different vocabulary is correctly retrieved. As PRF is not domain specific, we combine it with a health terms filter based on Wikipedia, which prevents query drift by removing terms that are not relevant in the context.

Our query reformulation approach does not directly take into account the generic question type (diagnosis, test, treatment) provided with each approach. To ameliorate that, we studied the task of providing the

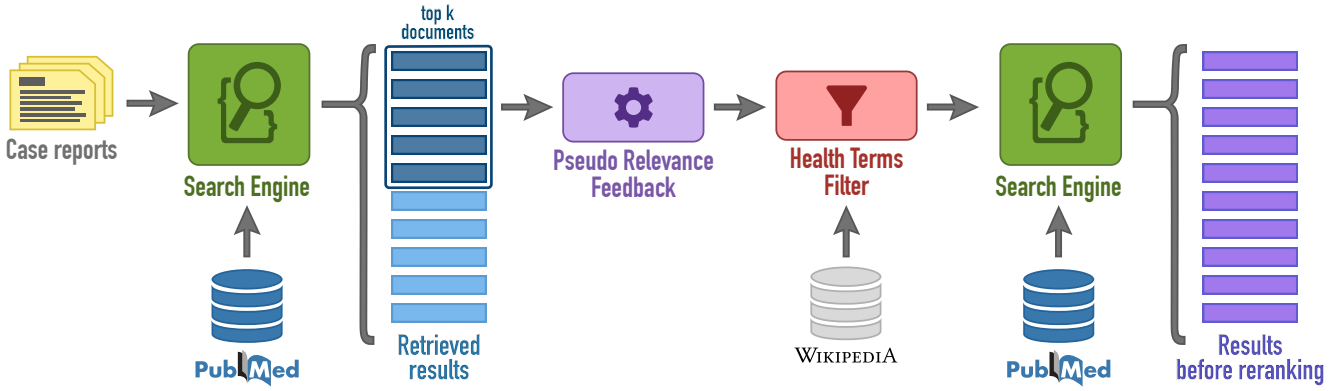

Figure 1: Query reformulation component (section 2.2) of the proposed system.

most relevant literature to a given type as a reranking task on the retrieved results. Different reranking methods were tested, both supervised and unsupervised.

In summary, our contributions are:

- A query reformulation technique that combines a domain independent approach with a technique designed for the health domain;
- A comparison of five different reranking techniques aimed to promote articles that match the type of information needed for each case report.

An improved and expanded system designed to address CDS search based on this work appeared in [6].

2 Methodology

2.1 Preprocessing

The dataset was indexed using ElasticSearch[1] v.1.2.1, a search server built on top of Lucene[2] v.4.8. The following fields were indexed and used for document retrieval (unless otherwise stated): article title, article abstract, and article text. Finally, the default implementation for the divergence from randomness retrieval model [2] was used.

2.2 Query Reformulation

Our query reformulation approach combines PRF with a health terms filter that removes non-medically related expressions from each case report before submitting it as a query to the search engine (Fig. 1). This strategy ensures that the original query formulation is expanded by adding many affine terms while preventing query drift.

[1] `http://www.elasticsearch.org`
[2] `http://lucene.apache.org`

Our PRF model is inspired by "IDF Query Expansion" (IDFQE) method proposed by [1]. Our system executes as follows: (*i*) a case report q of length n is issued as query to a search engine $\mathcal{S}$; (*ii*) of the set of documents $D_q = \{d_{q,1}, \ldots, d_{q,p}\}$ returned by $\mathcal{S}$, the top k are tokenized and used to build the root set $\mathcal{R}_q = \{t \mid t \in d_{q,i} \text{ for some } i \in \{1, \ldots, k\} \text{ or } t \in Q\}$, which consists of the set of all the terms in the top-k documents or in the case report; (*iii*) for each term $t_j \in \mathcal{R}_q$, a boost coefficient is determined according to the following formula:

$$b_j = \log_{10}(10 + w_j) \tag{1}$$

where w_j is computed as suggested in [1]:

$$w_j = \alpha \cdot I_q(t_j) \cdot tf_j + \frac{\beta}{k} \sum_{i=1}^{k} \tag{2}$$

where $I_q(t_j) = 1$ if and only if $t_j \in q$ (i.e., term t_j is in the case report q) and $I_{d_{q,i}}(t_j) = 1$ if and only if $t_j \in d_{q,i}$ (term t_j is in the case report $d_{q,i}$). Once the weights have been determined, the set of m terms $\{tc_1, \ldots, tc_m\}$ with the highest boosting coefficient are selected as candidates for expansion.

Rather than using a medical thesaurus to determine which terms are more suitable to expand the original query, we rely on Wikipedia. In detail, we estimate for each term tc_l its likelihood of being associated with a health-related page on Wikipedia by evaluating the odds ratio between the probability of tc_l appearing in a health-related Wikipedia page P over the probability of tc_l appearing in a non-health related Wikipedia page P:

$$\text{OR}(tc_l) = \frac{\Pr(P \text{ is health related} \mid tc_l \in P)}{\Pr(P \text{ is not-health related} \mid tc_l \in P)} \tag{3}$$

Each tc_l in the candidates list $\{tc_1, \ldots, tc_m\}$ is added to the original case report if its odds ratio $\text{OR}(tc_l)$ is greater than a threshold δ. Each term in the reformulated case report is finally boosted by its boosting coefficient b_j.

To compute the probabilities in equation (3), a Wikipedia dump from November 4, 2013 containing 2,794,145 unique entries was used. Pages with an information box containing at least one of the following medically-related fields were designated as health-related: DiseasesDB, eMedicine, MedlinePlus, MeSH, and OMIM (24,654 pages). The remaining pages were designated as not health-related.

2.3 Re-ranking

We approached the problem of identifying whether a retrieved document described a treatment, proposed a diagnosis or suggested a test as a reranking problem. In other words, given a list of search results, our system should rank higher those papers whose type aligns with the one in the query. We tested four different reranking strategies, as well as a fusion reranker that combines them via an unbiased voting scheme algorithm.

2.3.1 Supervised SVM Reranker

Upon observing that vocabulary in medical literature varies based on the goal of the paper (presenting a treatment, discussing a diagnosis, proposing a test), we decided to test whether such property could be exploited in a supervised setting.

To build a classier to distinguish between different types of biomedical articles, we asked three annotators to read 400 medical papers[3] each (1120 in total, as a 10% overlap was) and indicate whether they were discussing a treatment, diagnosis, test or none of them. Since it is not uncommon for a biomedical article to be more relevant to more than one category, we allowed the annotators to indicate more than one type for each paper. The agreement between the annotators on 40 overlapping question was 0.43 measured by Cohen's kappa [4].

We used a one-vs-rest SVM classifier with a linear kernel; for each paper the following features were considered:

- term stems, stopwords and numbers excluded;
- MetaMap[4] concepts in the title and abstract.

Only a portion of MetaMap concepts belonging to selected semantic categories were considered by the classifier[5]; this approach has been shown more effective by previous work [7].

Once the retrieved documents $\{d_1, \ldots, d_n\}$ are classified, the system groups them together in clusters $\{c_1, \ldots, c_h\}$ by using the affinity propagation algorithm [5]. Clusters are then ranked by number of articles that were classified in the same type category of the case report. In other words, given a case report q of type T, the following list in produced:

$$\{c_r \mid \forall t > r, \quad \mathit{typecount}(c_r, T) > \mathit{typecount}(c_t, T)\} \tag{4}$$

where *typecount* $(\cdot, \cdot)$ is defined as follows:

$$\mathit{typecount}(c_r, T) = |\{d_j \mid \text{type}(d_j) = T \wedge d_j \in c_r\}| \tag{5}$$

Finally, the reranked list of retrieved results is produced by concatenating the list of articles in each cluster.

2.3.2 Biographical Reranker

Since the appropriate treatment, diagnosis or test for a patient often depends on their age, biological sex or race, we hypothesized that articles that contain similar biographical information to those in the case report are more likely to be relevant. Therefore, for each case report q, we rerank the list of documents retrieved by the component of our system described in section 2.2 based on their biographical affinity with q.

[3] randomly selected from the entire dataset.

[4] `http://metamap.nlm.nih.gov/`

[5] `acab`, `anab`, `comd`, `cgab`, `dsyn`, `emod`, `fndg`, `inpo`, `mobd`, `neop`, `patf` and `sosy` for diagnoses, `topp` and `clnd`, for treatments, `lbpr`, `lbtr` and `diap` for tests.

The system proceeds as follows: (*i*) age, race and biological sex are extracted from the case report[6]; (*ii*) the same type of information is extracted from each retrieved biomedical article; a document receives a point for each matching biographical information[7]; (*iii*) for each paper, the original ranking score of each document is linearly combined with the points count; (*iv*) finally, documents are reranked based on the new score.

2.3.3 Seed Terms Reranker

In section 2.3.1 we mentioned how we observed that the goal of a biomedical article seems to affect its vocabulary. The reranker presented in this section exploits such observation; however, instead of a classifier, it uses a list of 55 selected seed terms (5 for treatments, 3 for diagnoses and 47 for tests) to determine whether a document matches the type of a case report.

For each retrieved document, the occurrences of terms in the document in the appropriate set of seed terms are counted; such value is then normalized by the length of the document and combined with the initial document score returned by the search engine. Finally, documents are reranked based on the new score.

2.3.4 MetaMap Similarity Reranker

This method reranks the retrieved documents based on their similarity with the query in terms of MetaMap concepts.

For each search result retrieved by submitting a case report q, the system uses MetaMap to extract UMLS concepts from its title and abstract. The concepts are matched with those extracted from the case report. Finally, documents are reranked based on their original score and the normalized number of detected matching concepts.

2.3.5 Fusion Reranker

We combined all the previously described approaches by building a voting-based fusion reranker. The rank assigned by each method to a document is considered as a vote; the final rank of the document consists of the average of the ranks determined by the other four rerankers.

[6] When such information is available.

[7] Ages were discretized in five buckets (*0-1*, *2-12*, *13-18*, *19-65*, *65+*); we say that a document matches the case report with respect of age if the two extracted ages fall in the same bucket.

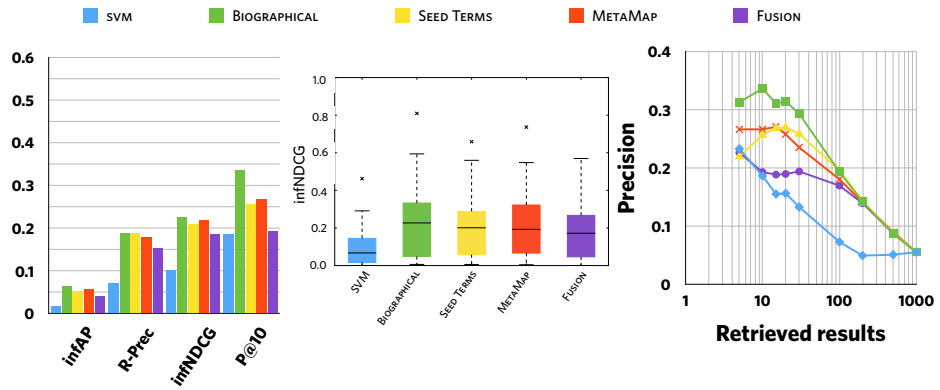

Figure 2: Overall results for each method.

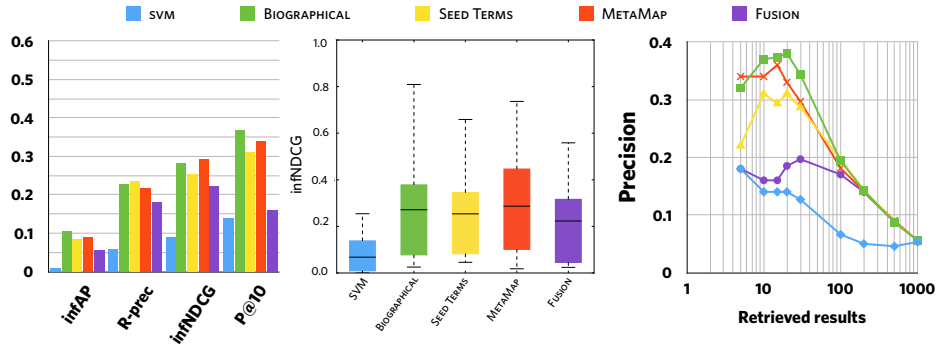

Figure 3: Results for diagnoses.

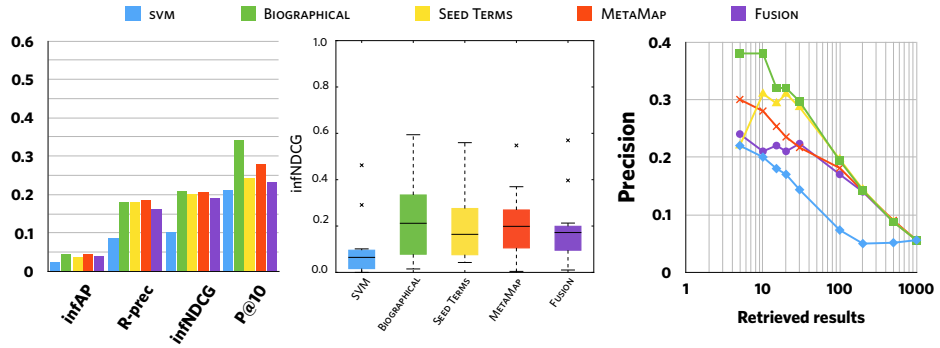

Figure 4: Results for treatments.

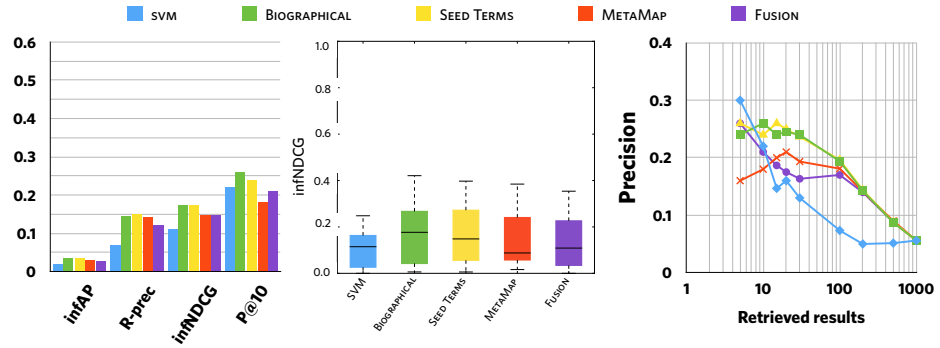

Figure 5: Results for tests.

Method	*inf-AP*	*R-prec*	*inf-nDCG*	*P@10*
SVM	0.0173	0.0691	0.1015	0.1867
Biographical	**0.0623**	**0.1871**	**0.2272**	**0.3367**
Seed Terms	0.0508	0.1884	0.2076	0.2567
MetaMap	0.0556	0.1792	0.2174	0.2667
Fusion	0.0394	0.1550	0.1847	0.1933

Table 1: Inferred average precision, R-precision, inferred nDCG, and precision at 10 documents retrieved for each reranking method. The best performing reranker is highlighted in bold.

3 Results

Results for each run are shown in table 1 and figures 2–4. We reported inferred average precision, inferred nDCG, R-precision, average precision after 10 results are retrieved (P@10), and precision when k documents are retrieved ($k = 5$ to 1000). We included results for each case report type (diagnosis, test, and treatment), as well as the overall performances of each method.

Combined results (figure 2) show that *Biographical* is the best performing reranking method, both in precision–oriented metrics and recall–oriented metrics. We attribute this outcome to the conservative nature of this reranker, as it modifies the rank of the results if and only if there is a strong biographical similarity between the case report and the retrieved documents. *SVM* performed poorly, underperforming all other methods. This is likely due to difficulty of the annotation task, as indicated by the only moderate agreement among annotators ($\kappa = 0.43$). *Seed terms* and *MetaMap* show a very similar behavior. The fusion reranker does not seem to offer any improvements over individual methods as its performance is likely to be affected by the worst performing method.

Some interesting behaviors are observable by looking at the performances of the reranking methods when grouped by case report type.

For diagnoses (figure 3), *MetaMap* shows comparable or better performances to *Biographica*. This suggests that the overlap between UMLS concepts in the case reports and the concepts in the query is a good indicator of the relevance of literature to a case report for diagnostic purposes. All the methods were found to perform similarly on treatments (figure 4).

For tests (figure 5), we noticed that *SVM* is capable of achieving solid performances in an high precision setting (e.g., P@5), yet its performances degrades quickly as the number of results retrieved increases. This might suggest that there are some language features exploited by *SVM* are strong indicators of the relevancy of a paper for this type of case reports. Further tuning of the parameters of the classifier will be conducted to better understand such behavior.

Finally, as shown in figure 6, we noticed that the performances for each topic vary greatly, suggesting that some inherit differences are present between topics. Moreover, we observed that, when *SVM* performed comparably to other methods, *Fusion* consistently outperformed all other methods. This observation validates the voting strategy adopted in this method.

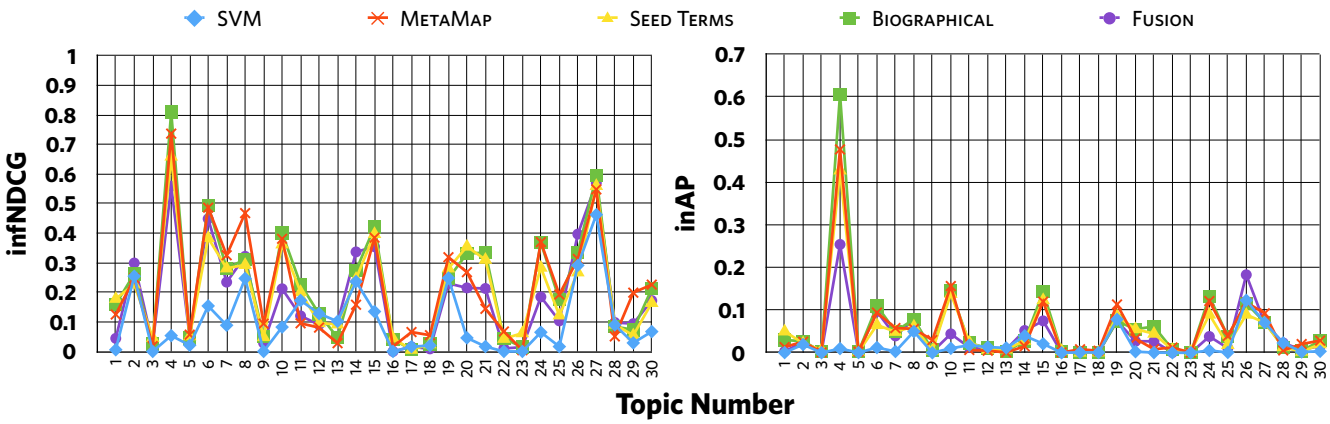

Figure 6: Inferred average precision and nDCG per topic across each run; topics 1–10 are diagnosis–related, topics 11–20 are test–related, and topics 21–30 are treatment–related.

4 Conclusions

We addressed search for clinical decision support, i.e., the task of retrieving relevant literature to a medical case report. We introduced a query reformulation technique that combines PRF with an effective domain specific approach. Furthermore, we studied five reranking algorithms that re-order a list of retrieved literature to better match the type of information needed for each case report.

5 Acknowledgments

This work was partially supported by the US National Science Foundation through grant CNS-1204347.

References

[1] S. Abdou and J. Savoy. Searching in medline: Query expansion and manual indexing evaluation. *Information Processing & Management*, 44(2):781–789, 2008.

[2] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002.

[3] A. Cohan, L. Soldaini, A. Yates, N. Goharian, and O. Frieder. On clinical decision support. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 651–652. ACM, 2014.

[4] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[5] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[6] L. Soldaini, A. Cohan, A. Yates, N. Goharian, and O. Frieder. Retrieving medical literature for clinical decision support. In *Advances in Information Retrieval*. Springer, 2015.

[7] S. Zhang and N. Elhadad. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics*, 46(6):1088–1098, 2013.