Retrieving Medical Literature for Clinical Decision Support

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

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

Abstract. Keeping current given the vast volume of medical literature published yearly poses a serious challenge for medical professionals. Thus, interest in systems that aid physicians in making clinical decisions is intensifying. A task of Clinical Decision Support (CDS) systems is retrieving highly relevant medical literature that could help healthcare professionals in formulating diagnoses or determining treatments. This search task is atypical as the queries are medical case reports, which differs in terms of size and structure from queries in other, more common search tasks. We apply query reformulation techniques to address literature search based on case reports. The proposed system achieves a statistically significant improvement over the baseline (29% – 32%) and the state-of-the-art (12% – 59%).

Keywords: medical literature search, medical query reformulation, query expansion, query reduction

1 Introduction

A Clinical Decision Support (CDS) system is a system designed to assist clinicians in providing patient care by offering timely and actionable health knowledge. One of tasks a CDS system could be designed to solve is the retrieval of key medical literature that can assist the practice of healthcare professionals given a medical case report (an example is shown in Fig. 1). We propose a system that addresses this need, which we refer to as CDS search.

CDS search presents some unique challenges: (*i*) compared to queries in traditional search domains, clinical case reports are substantially longer; (*ii*) although retrieval techniques for long queries have been widely studied in other domains (e.g., legal/patent search), case reports, unlike queries in those instances, have a narrative structure instead of being keyword based; (*iii*) most importantly, CDS search highly favors precision over recall, since healthcare professionals can only afford to spend limited time reading medical literature while practicing [4, 16].

Biomedical literature retrieval has been studied in the TREC genomics track[1]. CDS search, while sharing some aspects with it – descriptive queries, domain specific lexicon – is not limited to the genomics domain, but spans across multiple

[1] http://ir.ohsu.edu/genomics/

A 19-year-old African American student reports that he can "feel his heartbeat". It happens with exercise and is associated with some lightheadedness and shortness of breath. On examination, his heart has a regular rate and rhythm, but you hear a holosystolic murmur along his left sternal border. It increases with Valsalva maneuver.

Fig. 1. Example of a medical case report.

fields in medicine. Consequently, CDS search systems must process a variety of literature styles written with a wide domain specific vocabulary. Therefore, it is necessary to re-evaluate the effect of known IR techniques for this domain.

In this work we study the impact of query expansion and reduction methods that take advantage of medical domain knowledge, as well as general purpose IR techniques. Finally, we propose an approach that combines such methods, achieving a statistically significant improvement over the baseline (29%-32%) and an over all other approaches (12%-59%), including state-of-the-art.

Currently, no benchmark dataset containing case reports or medical publications can be used to evaluate a CDS search system. Clinical reports from last years' ShARe/CLEF eHealth Evaluation Lab [15, 10] are designed to test information extraction systems. OHSUMED [7] provides relevance annotations on medical literature, but its queries are considerably shorter than a case report (6 vs 67.6 terms on average) and are keyword based. NIST's TREC has added a CDS search track to the TREC 2014[2]; however, the system we propose was conceived and tested before the ground truth (q-rels) was publicly released. Thus, we developed an alternative, fully automated experimental framework for evaluating CDS search system based on the practice material for the United States Medical Licensing Examination (USMLE). Such dataset is publicly available[3] to other researchers; the performance obtained by our system on it were found to be comparable to TREC's [14].

In summary, our contributions are: (*i*) a system for retrieving highly relevant, and thus actionable, medical literature in support of clinical practice, (*ii*) an adaptation and evaluation of query reformulation techniques for CDS search, and (*iii*) publicly available experimental framework and benchmark for CDS search.

2 Related Work

Historically, search systems in the medical domain have focused on short and/or keyword-heavy queries. In PubMed, for example, the query is expanded by mapping each term to MeSH terms and then considered as a boolean conjunctive query. Such an approach is ill-suited when considering long, narrative case reports as queries. We approach CDS search as a reformulation problem. Many reduction and expansion approaches have been introduced over the years; here, we give an overview of domain-specific and domain-independent methodologies.

[2] `http://www.trec-cds.org/2014.html`
[3] `https://github.com/Georgetown-IR-Lab/CDS-search-dataset`

Query reduction algorithms have been extensively studied as a way to remove noisy terms from the original query. Their impact has mostly been tested in the web search domain. For example, Kumaran and Carvalho [11] used SVM^{rank} [9] to find the best sub-query using a series of clarity predictors and similarity measures as features. Balasubramanian et al. [3] also studied how to improve performance by reducing queries using quality predictors; however, their system only removes up to one term from the query. This approach is not viable when dealing with long, descriptive case reports. To the best of our knowledge, the only work that has adopted query reduction in the medical domain is by Luo et al. [12]. They built a search engine that performs query reduction by filtering non-important terms based on their tf-idf score. Unlike CDS search, their system is designed for lay people performing health search on the Web and does not focus on medical literature retrieval.

Over the past years, query expansion techniques were successfully employed in medical literature retrieval. Hersh et al. [8] expanded queries with terms manually selected from UMLS Metathesaurus relationships to enhance retrieval performance. Experimental results showed that thesaurus based query expansion did not necessarily improve search efficiency. Yu et al. [17] experimented with relevance feedback in PubMed; their system used RankSVM to re-arrange retrieved results based on explicit users' feedback. Abdou and Savoy [1] used pseudo relevance feedback methods to improve the retrieval of MEDLINE abstracts; their system was tested on manually crafted, keyword based queries substantially shorter than the case reports in our dataset (14 vs. 67.6 terms). In a preliminary version of this work, Cohan et al. [5] explored the use of pseudo relevance feedback for CDS search.

Another line of research related to CDS search is clinical question answering, given the shared goal of improving medical understanding. Demner-Fushman and Lin [6] focused on extracting medical concepts from MEDLINE abstracts that match the information need of the question. Sneiderman et al. [13] examined three knowledge-based methods to evaluate their efficiency in helping clinicians retrieve answers from MEDLINE. In contrast to our work, question answering search systems are designed to handle queries that are much shorter than a case report and are strictly formulated as query. Furthermore, they usually generate an answer rather than returning relevant resources.

3 Methodology

We approached CDS as a query reformulation problem. As such, we capitalized on query reduction (section 3.1) and expansion (section 3.2) techniques. For query reduction, we used a domain specific tool, MetaMap (*MMselect*), and Wikipedia (*HT*), to prune non-medical terms from the query. We also implemented one of the state-of-the-art techniques for domain-agnostic query reduction (*QQP*). Finally, we introduced a refined version of *QQP* that takes advantage of domain specific resources (*Fast QQP*). We then evaluated several query expansion techniques: one (*MMexpand*) takes advantage of a medical thesaurus,

another (PRF) uses pseudo relevance feedback to incorporate key terms in the original query. Finally, we introduced a new method (*HT-PRF*) that combines a domain specific approach with pseudo relevance feedback. As shown in section 5, this method outperforms all others, including *QQP* and *Fast QQP* (state-of-the-art and its derivative).

As a baseline, we considered an algorithm that submits the unmodified case report (after removing stopwords) to the search engine.

3.1 Query Reduction Techniques

UMLS Concepts Selection (*MMselect*)
We extract concepts from queries based on concepts defined in the Unified Medical Language System[4] (UMLS) to perform query reduction. For this extraction we utilize MetaMap[5], a tool designed for UMLS concept extraction. We reformulated the query by removing all the terms that did not have a mapping to any UMLS concepts.

Health-related Terms Selection (*HT*)
Rather than selecting health-related words based on a medical thesaurus, we leverage Wikipedia as an external resource. Specifically, for each word candidate c_l in the original query, we estimate its likelihood of being associated with a health-related Wikipedia entry by computing the odds ratio between the probability of a Wikipedia page P being health-related when $c_l \in P$ over the probability of P not being health-related over all the Wikipedia pages.

$$\mathrm{OR}(c_l) = \frac{\Pr\{P \text{ is health-related} \mid c_l \in P\}}{\Pr\{P \text{ is not health-related} \mid c_l \in P\}} \tag{1}$$

A word $c_l \in \{c_1, \ldots, c_m\}$ is kept as part of the reduced query if $\mathrm{OR}(c_l) \geqslant \delta$, where δ is a tuning parameter.

We used a Wikipedia dump from November 4, 2013 (2,794,145 unique entries). Those pages whose infobox[6] contain one or more of the following medically-related code entries were determined to be health-related: OMIM, eMedicine, MedlinePlus, DiseasesDB and MeSH (24,654 pages); the rest were considered to be not health-related. The optimal value for δ was empirically found to be 2.

Query Quality Predictors for Optimal Sub-query Identification (*QQP*)
We implemented the system suggested by Kumaran and Carvalho [11]. Their method uses quality predictors as features to rank sub-queries of the original query using $\mathrm{SVM}^{\mathrm{rank}}$. The following predictors are considered as features:

[4] `http://www.nlm.nih.gov/research/umls/`
[5] `http://metamap.nlm.nih.gov/`
[6] An infobox is template containing structured information that appear on the right of Wikipedia pages to improve concepts representation.

- *Mutual information*: each sub-query is represented as a fully connected weighted graph, where each vertex represents a term in the sub-query. Edges are weighted by mutual information. For each graph, the heaviest spanning tree is extracted; the average weight of the edge is used as query predictor.
- *Query clarity*: estimation of the divergence of the query model from the collection model using the top 500 documents retrieved per sub-query.
- *Simplified clarity score*: simplified version of clarity score that estimates the probability of a term in the language model by considering the likelihood of it appearing in the query.
- *Query scope*: measure of the size of the retrieved set of documents relative to the size of the collection. Sub-queries showing high query scope are expected to perform poorly since they contain terms that are too broad.
- *Similarity to orginal query*: *tf-idf* similarity is considered as one of the quality predictors under the hypothesis that the closer a sub-query is to the original query, the less likely it is to cause intent drift.

In addition to the previously listed features, *QQP* considers, for each sub-query, statistical measures[7] over the term frequency, document frequency and collection frequency of the terms in the sub-query as features for SVM^{rank}. The length of each sub-query is also considered as a feature. We refer the reader to the original paper for more details.

Since most of the query predictors are query dependent, they cannot be computed ahead of time, thus slowing the sub-query selection process. Therefore, as suggested by the authors, we implemented a set of heuristics to reduce the number of candidate sub-queries, which, prior to pruning, is exponential to the size of the original query: (*i*) select queries with length between three and six terms; (*ii*) select only the top twenty five sub-queries ranked by MI; (*iii*) select only the sub-queries containing name entities. The parameters for SVM^{rank} were set as suggested in [11].

Faster Query Quality Predictors with Medical Features (*Fast QQP*)

Since *QQP* was not designed specifically for CDS search, its performance is negatively affected by the greatly reduced length of the generated sub-queries and by the lack of domain-specific features. Because of the unique formulation of case reports, we implemented a set of sub-query candidates pruning heuristics that resulted in statistically significant improvements over the original formulation while reducing the processing time.

First, we increased the maximum length M_{subq} of a sub-query candidate from 6 to 16 terms (empirically determined). This is motivated by the fact that case reports are, on average, much longer than the queries in [11] (16.2 vs. 67.6 terms). The minimum length of a sub-query was not altered (i.e., $m_{\text{sub-q}} = 3$).

As the size of the candidates set grows exponentially when the maximum number of tokens increases linearly, *Fast QQP* prunes the list of candidates

[7] Maximum and minimum value; arithmetic, harmonic, and geometric mean; standard deviation and coefficient of variation.

after each increase in length of candidate sub-queries. In other words, for each $i \in \{m_{\text{subq}}, \ldots, M_{\text{subq}}\}$, the set of candidates C_i is ranked by MI; the top-k sub-queries are then extracted (set $C_{i,k}$) and used to build the set C_{i+1} accordingly with the following formula:

$$C_{i+1} = \{s_l \cup \{q_h\} \mid s_l \in C_{i+1} \wedge q_h \in Q\} \cup C_{i,k} \tag{2}$$

where Q is the original query. After empirical evaluation, we set $k = 50$.

We further improved *Fast QQP* by including some domain-specific features:

- number of UMLS concepts in the candidate sub-query,
- semantic type of the UMLS concepts in the candidate sub-query,
- statistical features[7] over the likelihood of each term in the candidate sub-query of being health related, as estimated by equation (1), and
- number of MeSH terms in the candidate sub-query.

3.2 Query Expansion Techniques

UMLS Concepts Extraction (*MMexpand*)

Similar to MM Select method, this method identifies UMLS Metathesaurus concepts that exist in the query using MetaMap. However, rather than filtering out terms, this method expands the query using new terms associated with the concepts identified. After detecting the concepts in the query, expansion terms were chosen by querying UMLS for new terms that were synonyms of the concepts in the query and were marked as preferred terms by UMLS; the query was expanded with all these terms. Given the extensive coverage of UMLS, we limited concept expansion to concepts containing drugs, diseases, and findings to prevent query drift.

Pseudo Relevance Feedback (*PRF*)

Pseudo relevance feedback was modeled after the "IDF Query Expansion" method proposed in [1]. We modified the algorithm to adapt it to our experimental setup: instead of directly altering term weights, our system determines a boosting coefficient for each term in the reformulated query. The query Q is expanded as follows: it tokenizes the top k retrieved documents retrieved for Q; it then builds the root set $\mathcal{R}_Q$, which consists of the union of the set containing all the terms in Q with the set of all the terms in the retrieved documents for Q. The boost coefficient b_j for each term $t_j \in \mathcal{R}_Q$ is calculated as:

$$\begin{aligned} b_j &= \log_{10}(10 + w_j) \\ w_j &= \alpha \cdot \mathrm{I}_Q(t_j) \cdot tf_j + \beta / k \sum_{i=1}^{k} \mathrm{I}_{D_i}(t_j) \cdot idf_j \end{aligned} \tag{3}$$

where t_j is the j-th term in the top Q documents, $\mathrm{I}_Q(t_j)$ is an indicator of the presence of term t_j in Q, $\mathrm{I}_{D_i}(t_j)$ is an indicator of the presence of term t_j in the document D_i, idf_j is the inverse document frequency of the j-th term in the top k documents. Finally, α and β are smoothing factors.

Once all the weights have been determined, the terms in $\mathcal{R}_Q$ are ranked by their boost coefficient; the top m terms not in the original query are added to Q; each term in the reformulated query is boosted by its boosting factor. Tuning parameters were set as suggested in [1]: $\alpha = 2$, $\beta = .75$, $k = 10$, $m = 20$.

Health Terms Pseudo Relevance Feedback (*HT-PRF*)

We explored the effect of combining a pure IR approach – pseudo relevance feedback – with domain specific knowledge (health terms). *HT-PRF* operates similarly to *PRF*– it retrieves the top k documents, builds the root set $\mathcal{R}_Q$ of the query, scores each term in the root set using the equation (3) – but instead of always expanding with top m candidates, it calculates, for each term, the odds of it being health related using equation (1), retaining only those whose odds ratio is greater or equal to δ', where δ' is a tuning parameter of the system. Because of this, the number of terms m'_q added to each query varies.

Finally, we would like to stress the fact that, despite taking advantage of *HT*, *HT-PRF* is not a reduction method: non-health specific terms are only pruned off the list of candidates for query expansion; the original query is left untouched.

4 Experimental Setup

As stated in the introduction, the lack of datasets designed to evaluate a CDS search system required us to create our own. To create a benchmark for evaluation, we developed an approach to automatically identify relevant documents to case reports by making use of external information about each case report (the correct diagnosis, treatment or test associated with each one as well as explanations about the correctness of such relations). Our dataset contains two components: medical papers and medical case reports. The medical literature was obtained from Open Access Subset of PubMed central[8], a free full-text archive of health journals (728,455 documents retrieved January 1, 2014).

495 medical case reports were obtained from three USMLE preparation books[9] Each case report contains a description of a patient followed by a question asking for the correct diagnosis, treatment, or test that should be executed. Case reports from USMLE are modeled after real clinical situations with goal of assessing the ability of future physicians in applying clinical knowledge, concepts and principles for effective patient care[10].

Given a case report, our goal is to retrieve documents (medical publications) that can help a physician diagnose the patient, treat the patient's condition, or request a test relevant to the case; the content of three USMLE prep books were used to determine which documents in our collection were relevant. In detail, we took advantage of the multiple answer choices associated with the case reports as well as the explanation of why an answer is correct. To determine relevant documents for each case report, we separately issued as queries the explanation paragraph (q_{E}) and each answer choice individually ($q_{a_0}, \ldots, q_{a_3}$). Documents retrieved by the correct answer $q_{a_{\text{corr}}}$ and q_{E} received a relevance score of two, while documents retrieved by q_{E} and any incorrect answer choice were given a score of one. By using this approach, we were able to take into account that not only the correct documents retrieved by querying the correct answer contribute

[8] `http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist`

[9] `https://github.com/Georgetown-IR-Lab/CDS-search-dataset`

[10] Bulletin of Information, `http://www.usmle.org/pdfs/bulletin/2012bulletin.pdf`

to determine the right treatment/test/diagnosis, but also those related to the incorrect options. Any answer choice query ($q_{a_i}, \in \{0, \ldots, 3\}$) that contained more than 200 documents was discarded under the assumption that the query was too broad. A case report was discarded if its correct answer choice query was discarded. This process left us with 195 valid queries (i.e., case reports).

Three human assessors were then instructed to read each of these case reports and determine their validity. Specifically, they were asked to categorize each one as invalid or as asking for a diagnosis, treatment, or test. Invalid queries were those that were primarily quantitative (i.e., contained only numeric values about some tests or vital signs e.g. blood pressure, heart rate, body temperature, etc). The three assessors' inter-rater agreement was 0.56 as measured by Fleiss' kappa[11]. Any query deemed invalid by at least two assessors was discarded. This left us with 85 case reports; of those, 17 were reserved for parameters tuning, while the remaining 68 were used for testing.

We used ElasticSearch v1.2.1, a search server built on top of Lucene v4, to index the medical documents in our dataset and to retrieve results. The default tokenizer and the divergence from randomness retrieval model [2] were used.

5 Results and Discussion

We validate our query reformulation approach for CDS search by running two experiments. First, we compare the performance of each method introduced in section 3; second, we describe the tuning process for the best performing method. In both experiments, we retrieve 1000 documents for each test query.

5.1 Comparison of Reformulation Methods

As previously mentioned, CDS search is a precision oriented task; it is meant to support healthcare professionals who are looking for findings that could help them determine the next action in the care of a patient. For this reason, performance at the first ten points of precision (Fig. 2) is key to assert the quality of a reformulation method. We focus our analysis on precision at five documents retrieved (P@5), as the performance of each method is consistent throughout the first ten points (Fig. 2, left) of precision and show no significant difference up to P@100 (Fig. 2, right). Recall and nDCG are also reported (Table 1); these metrics, albeit less key to the task, are still useful indicators to assert the overall quality of each method. We use a paired Student's t-test to measure whether the difference between any two methods is statistically significant ($p < 0.01$).

MMselect performed significantly worse than the baseline. We attribute such difference to the fact that, while it successfully identifies most medical concepts in the query, it often discards terms that have a key role connecting domain specific expression. For example, for the case report in Fig. 1, *MMselect* fails to

[11] The moderate level of agreement between assessors is attributable to the hardness of the task. The evaluators reported that many reports laid in the spectrum between fully quantitative and fully qualitative, thus representing a noteworthy challenge.

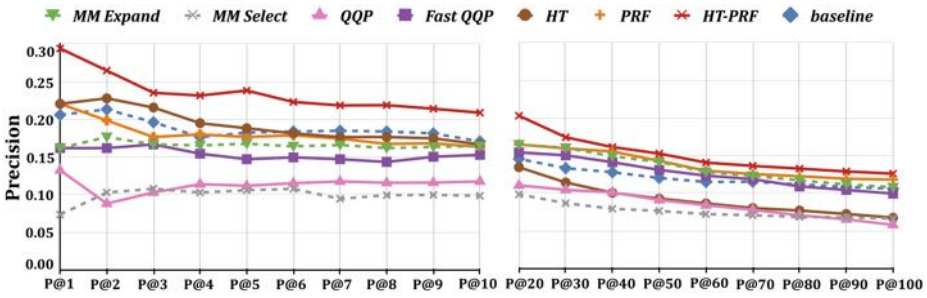

Fig. 2. Points of precision for each method. The best performing method, *HT-PRF*, achieves a 43% increase over the baseline for P@1.

identify "increases" as relevant term (last sentence), which is key in understanding the outcome of the "Valsalva maneuver" on the patient. *MMexpand* showed a minor but significant gain in terms of nDCG and recall over the baseline, but it performed worse (although not significantly) than the baseline in terms of P@5. We attribute the modest difference to the limited coverage of the portion of the synonym map in UMLS *MMexpand* uses with respect to the size of our dataset. This tradeoff was necessary to prevent query drift.

QQP performed very poorly. Its limited performance is due to its aggressive reduction algorithm, which reduces the original query to at most six terms. As result, the reduced query loses most of the information content of the case report.

Fast QQP showed substantially better nDCG and recall results, but fell short in terms of P@5. We attribute the improvement to the fact that the inclusion of domain specific features and a more conservative approach lead to a more effective reduction. On the other hand, the worsening in terms of P@5 is likely due to the insufficient coverage of medical terms in the query: in medical literature, the same concept is often expressed using different terms and expression; thus a method that only performs reduction is likely to miss documents that are relevant to the case report, but differ from it in terms of vocabulary.

Both *HT* and *PRF* methods showed a statistically significant improvement over the baseline in terms of nDCG and recall; *HT* removes common non-health-

Table 1. Each method's performance ($\circ$ for query reduction, $\bullet$ for expansion). A $\triangle/\triangledown$ indicate a significant improvement/worsening ($p < 0.01$) over the baseline. $\blacktriangle$ indicates a significant improvement over Simple and methods marked with $\triangle$.

	nDCG		Recall		P@5	
baseline	0.2855	–	0.2741	–	0.1824	–
MMselect$^{\circ}$	$0.1622^{\triangledown}$	(−43.2%)	$0.1486^{\triangledown}$	(−45.8%)	$0.1059^{\triangledown}$	(−41.9%)
MMexpand$^{\bullet}$	$0.3020^{\triangle}$	(+5.8%)	$0.2958^{\triangle}$	(+7.9%)	0.1676	(−8.1%)
QQP$^{\circ}$	$0.2557^{\triangledown}$	(−10.4%)	$0.2494^{\triangledown}$	(−9.0%)	$0.1118^{\triangledown}$	(−38.7%)
Fast QQP$^{\circ}$	$0.3177^{\triangle}$	(+11.3%)	$0.3129^{\triangle}$	(+14.2%)	$0.1471^{\triangledown}$	(−19.4%)
HT$^{\circ}$	$0.3328^{\triangle}$	(+16.5%)	$0.3262^{\triangle}$	(+19.0%)	0.1882	(+3.2%)
PRF$^{\bullet}$	$0.3390^{\triangle}$	(+16.5%)	$0.3263^{\triangle}$	(+19.0%)	0.1765	(−3.4%)
HT-PRF$^{\bullet}$	$\mathbf{0.3768}^{\blacktriangle}$	**(+32.0%)**	$\mathbf{0.3520}^{\blacktriangle}$	**(+28.9%)**	$\mathbf{0.2382}^{\blacktriangle}$	**(+30.5%)**

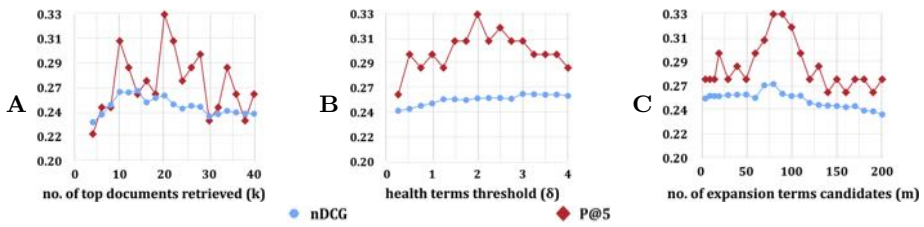

Fig. 3. Effect of different parameter values for *HT-PRF* in terms of nDCG and P@5 (other precision levels exhibit similar behavior). The best precision performances are achieved when $k = 20$, $\delta' = 2$, $m = 90$.

related terms, whereas *PRF* reweights the entire query, increasing the importance of health-related terms, which naturally have a high IDFQE coefficient given the domain of the dataset. In *HT* some improvement is expected, as it keeps more generalized medical concepts in comparison with the UMLS concept selection method. Neither *HT* nor *PRF* showed significant improvement in terms of P@5. *HT* is likely to suffer from the same limitation in terms of vocabulary coverage *Fast QQP* has, while *PRF* is partially affected by query drift.

We achieved the most noteworthy results by using the *HT-PRF*. The nDCG and recall values shown in Table 1 are statistically significant not only with respect to the baseline but also over simple *PRF* and *HT*. Moreover, *HT-PRF* consistently improves over the baseline for each precision level shown in Fig. 2 ($p < 0.01$). The substantial increase in performances of *HT-PRF* is due to the fact that it combines two very effective techniques: by expanding the query using the most relevant document, it is able to broad its vocabulary; on the other side, filtering the list of candidate terms for expansion prevents query drifting.

5.2 Parameter Tuning for *HT-PRF*

In this section we detail the tuning process for *HT-PRF*. We studied the outcome of varying the number k of the top ranking documents used by pseudo relevance feedback to build the list of candidate terms for query expansion (Fig. 3A), the value δ' of the conditional probability threshold used to select expansion terms from the list of candidate terms (Fig. 3B), as well as the number m of candidate terms for query expansion (Fig. 3C).

The results we present were obtained on a subset of 17 separate case reports we reserved for tuning purposes. For all three tuning parameters, we preferred those values that yielded better performance in terms of P@5. As in section 5.1, we chosen to report the performances in terms of P@5, as we observed comparable behavior at all the other precision levels between one and ten (the differences between methods are not statistically significant after ten results).

Fig. 3A shows that the highest performance in terms of P@5 is obtained when the number of top documents k is equal to 20. However, we also noticed an ample variation in terms of P@5 for small differences in the number of retrieved documents. This variation clearly depends on which terms are used to expand the original query. Since the terms picked for expansion are the most representative

terms of the top k documents retrieved, their effectiveness in improving the retrieval performance depends on whether the top k documents are relevant or not. Given the fact that the top document set is small, each time a new document is added (i.e. k increases) the set of terms picked for expansion varies substantially. In other words, when a non-relevant document is added to the set of top documents, the relevance of the terms selected for expansion decreases, thus causing query drift. Similarly, when a relevant document is included in the top k documents, the relevance of terms selected for expansion increases, leading to better performance. Nevertheless, we observed that the retrieval performance decreases as the number of top documents increases past 20. This outcome is expected, since the more documents the system considers, the more likely it is to suffer from query drift, as less relevant terms are picked for expansion.

With health terms' threshold (Fig. 3B) we noticed a much more defined trend: the best precision is achieved when $\delta' = 2$. The bigger δ' is, the more aggressive the filter is. And for higher values of δ', precision starts to decrease. That is, because bigger values of δ' result in selection of more focused and specific medical terms, many more general key terms for optimal retrieval are being discarded. In fact, the lower performance of thesaurus based methods further reveals the fact that considering only highly focused medical terms decreases P@5. On the other side, when δ' is smaller the method is more likely to consider all sorts of terms for query expansion, which eventually results in query drift.

Finally, we recorded the best retrieval performance when the number of candidates for expansion m is set to 90 (Fig. 3C). Different values of m tend to cause query drift when they are larger than the optimal and cause key terms to be removed from the when they are smaller than the optimal.

6 Conclusions

We described CDS search based on medical case reports, which is a search task intended to help medical practitioners retrieve relevant publications to clinical case reports. We used query reformulation to perform CDS search, and found that the best methods for this task are a query reduction method retaining only health-related terms and a pseudo relevance feedback query expansion method. Both methods independently improved performance significantly (as measured by nDCG and recall), yet showed limited improvements in terms of precision. However, when combined, the resulting method outperformed each individual method and greatly improved precision. We conclude that while this method decisively improved retrieval performance, there is still room for improvement; this stresses that CDS search is significantly different than other types of health-related search, making it a novel search task worthy of further study.

7 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*, 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)*, 2002.
3. N. Balasubramanian, G. Kumaran, and V. R. Carvalho. Exploring reductions for long web queries. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010.
4. D. T. Burke, M. C. DeVito, J. C. Schneider, S. Julien, and A. L. Judelson. Reading habits of physical medicine and rehabilitation resident physicians. *American journal of physical medicine & rehabilitation*, 2004.
5. 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.
6. D. Demner-Fushman and J. Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 2007.
7. W. Hersh, C. Buckley, T. Leone, and D. Hickam. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *SIGIR'94*. Springer, 1994.
8. W. Hersh, S. Price, and L. Donohoe. Assessing thesaurus-based query expansion using the umls metathesaurus. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2000.
9. T. Joachims. Training linear svms in linear time. In *Proceedings of the 12^{th} SIGKDD international conference on Knowledge Discovery and Data Mining*. ACM, 2006.
10. L. Kelly, L. Goeuriot, H. Suominen, T. Schreck, G. Leroy, D. L. Mowery, S. Velupillai, W. W. Chapman, D. Martinez, G. Zuccon, et al. Overview of the share/clef ehealth evaluation lab 2014. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*. Springer, 2014.
11. G. Kumaran and V. R. Carvalho. Reducing long queries using query quality predictors. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009.
12. G. Luo, C. Tang, H. Yang, and X. Wei. Medsearch: a specialized search engine for medical information retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008.
13. C. A. Sneiderman, D. Demner-Fushman, M. Fiszman, N. C. Ide, and T. C. Rindflesch. Knowledge-based methods to help clinicians find answers in medline. *Journal of the American Medical Informatics Association*, 2007.
14. L. Soldaini, A. Cohan, A. Yates, N. Goharian, and O. Frieder. Query reformulation for clinical decision support search. In *The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014)*, 2015.
15. H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. Jones, et al. Overview of the share/clef ehealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*. Springer, 2013.
16. C. Tenopir, D. W. King, M. T. Clarke, K. Na, and X. Zhou. Reading patterns and preferences of pediatricians. *Journal of the Medical Library Association*, 2007.
17. H. Yu, T. Kim, J. Oh, I. Ko, and S. Kim. Refmed: relevance feedback retrieval system fo pubmed. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009.